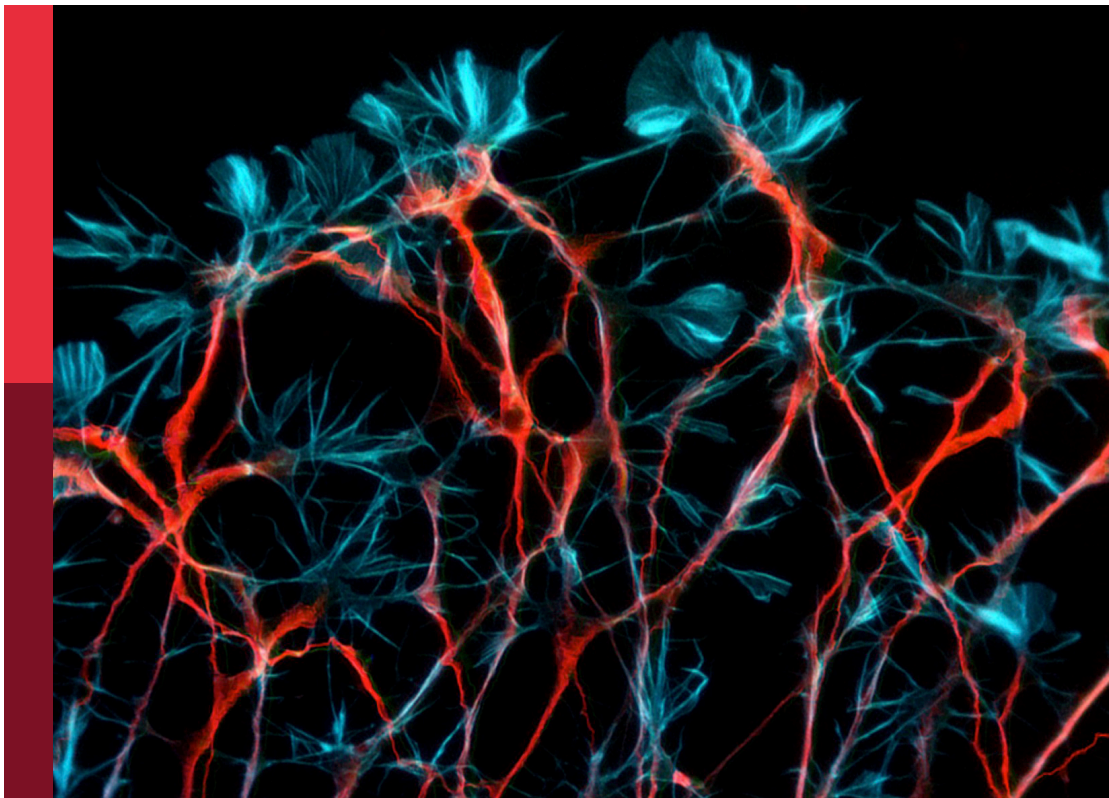
frontiers | Research Topics

Insights in computational neuroscience

Edited by
Si Wu and Misha Tsodyks

Published in
Frontiers in Computational Neuroscience



ISSN 1664-8714
ISBN 978-2-83252-050-5
DOI 10.3389/978-2-83252-050-5

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Insights in computational neuroscience

Topic editors

Si Wu — Peking University, China

Misha Tsodyks — Weizmann Institute of Science, Israel

Citation

Wu, S., Tsodyks, M., eds. (2023). *Insights in computational neuroscience*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83252-050-5

Table of contents

frontiers | Frontiers in Computational Neuroscience

PERSPECTIVE
published: 08 April 2022
doi: 10.3389/fncom.2022.836532

Physics Clues on the Mind Substrate and Attributes

Joaquin J. Torres and Joaquín Marro*

Institute Carlos I for Theoretical and Computational Physics, University of Granada, Granada, Spain

The last decade has witnessed a remarkable progress in our understanding of the brain. This has mainly been based on the scrutiny and modeling of the transmission of activity among neurons across lively synapses. A main conclusion, thus far, is that essential features of the mind rely on collective phenomena that emerge from a willful interaction of many neurons that, mediating other cells, form a complex network whose details keep constantly adapting to their activity and surroundings. In parallel, theoretical and computational studies developed to understand many natural and artificial complex systems, which have truthfully explained their amazing emergent features and precise the role of the interaction dynamics and other conditions behind the different collective phenomena they happen to display. Focusing on promising ideas that arise when comparing these neurobiology and physics studies, the present perspective article shortly reviews such fascinating scenarios looking for clues about how high-level cognitive processes such as consciousness, intelligence, and identity can emerge. We, thus, show that basic concepts of physics, such as *dynamical phases* and *non-equilibrium phase transitions*, become quite relevant to the brain activity while determined by factors at the subcellular, cellular, and network levels. We also show how these transitions depend on details of the processing mechanism of stimuli in a noisy background and, most important, that one may detect them in familiar electroencephalogram (EEG) recordings. Thus, we associate the existence of such phases, which reveal a brain operating at (non-equilibrium) criticality, with the emergence of most interesting phenomena during memory tasks.

Keywords: collective brain phenomena, adaptive complex networks, dynamic synapses, non-equilibrium phase transitions, EEG oscillations, intelligence, identity, consciousness

OPEN ACCESS

Edited by:
Si Wu,
Peking University, China

Reviewed by:
Haiping Huang,
Sun Yat-sen University, China

****Correspondence:***
Joaquin J. Torres
jtorres@onsager.ugr.es

Received: *15 December 2021*
Accepted: *07 February 2022*
Published: *08 April 2022*

Citation:
Torres JJ and Marro J (2022) Physics Clues on the Mind Substrate and Attributes. Front. Comput. Neurosci. 16:836532. doi: 10.3389/fncom.2022.836532

INTRODUCTION

As humans we are interested in the age-old question *What are we?,* perhaps now rephrased *Can one identify guidelines to understand our intimate being?* The doubt is not banal. Looking into this requires involving the mind that, for a very long time, has been an ambiguous entity, and therefore source of misunderstandings and unfortunate hypotheses. However, by developing new means of observation and computation, science has uncovered details, and paths are now open on which to begin to walk confidently. So much so that we may rationally precise, for instance: *What makes us be the way we are*? *Where is located our own identity*? *Could it be manipulated*? Even more, it has been uncovered that probing solutions to these queries equals looking for the keys of our *identity* and *consciousness* and, in trying to do so, it has been realized that the relevant scenarios closely relate to *intelligence*.

According to thesauruses, intelligence is "ability to acquire and apply knowledge and skills," and relate this term to "understand," "solve problems," "experience," and "competence." We stuck to this. Nevertheless, we are not interested in meanings such as "purely spiritual substance," and forget for the moment the so-called *social intelligence*—the one ascribed to *groups* of ants, bees, birds, fish, and humans. In any case, this eclectic vision is insufficient for us, as it hides the essence or content of what we could name intelligence "function," i.e., how those capacities unfold as an essential part of our being. This is what most interests us here.

We manage to go one-step farther by exploring what in this connection distinguishes the animal species. Perhaps it then surprises that the sperm whale has a bigger brain than we have, and that the shrew's one weights more relative to the total. To have humans leading a list in this context, we must ponder the neurons connectivity—in which case however dolphins follow us, not primates as one might have expected (Roth and Dicke, 2005). Endorsing a popular identification, this associates intelligence with *gray matter* that, leaving aside other structures, consists of near 100 billion *somas* or neural bodies. These extend through long eager-to-connect filamentous extensions that end in terminals with a complicated internal structure and one may call *synapses*, the name of one of its parts. All this stuff forms the cerebral *cortex* that, just under the skull, is the best-organized part of our very well-organized nervous system. The 10^{11} neurons (Azevedo et al., 2009) thus continually interact with any part of our body, including muscles, organs, and senses, through about 10^{15} synapses mainly in the neocortex (DeFelipe et al., 2002).

This picture suggests that properties of the cortex, such as its thickness, may be related and perhaps aid to estimate the intellectual capacity of an individual. However, today we understand this, together with some of the mechanisms that improve the brain operation and eventually may induce its malfunction, in more detail. For example, we have a harmonious framework (Marro, 2014; Marro and Torres, 2021), coherent with what we learned from experiments, which allows one to quantitatively exploring intriguing phenomena associated to the concept of intelligence and the mechanisms that seem to favor it up. This is a simple though rigorous scheme, kind of "mathematical metaphor" that includes, together with other details, a realistic description of synaptic cooperation between neurons. That is, it does not presume a passive participation of the synapses, but it specifies how these, constantly using both intrinsic and external information, actively modulate the interrelation between neurons, including its network effective topology, which affects, even dramatically, the current result of that collaboration. In short, synapses should now be viewed as effective processors responsible for achieving certain, fruitful, mutual influence between neurons at every time, and they do so with the mediation of several relatively complex biophysical mechanisms. Insomuch that these show up to the observer (using suitable techniques) as "noises" or fluctuations along several time scales propagating through an adapting complex network. These "noises" carry relevant information that characterizes some brain activity states (Lendner et al, 2020), and may be important to understand brain interrelations (Waschke et al., 2021), so that the resultant scene is very subtle.

This perspective article shortly reviews the framework supporting the above scenarios, namely, we follow here a statistical physics point of view looking for indications about how high-level cognitive processes such as consciousness, intelligence, and identity can emerge. In particular, we, thus, illustrate how basic concepts of physics, such as *dynamical phases* and *non-equilibrium phase transitions*, are quite relevant for the emergence of intriguing synchronization phenomena and for the understanding of the dynamical features of actual brain activity which can be related with different brain cognitive functions. In addition, we explore the factors at the subcellular, cellular, and network levels that seem to induce the non-equilibrium phases that happen to show up, and remark the important role that synaptic mechanisms and network development, and refinement processes such as synaptic pruning, have on the observed phenomena. We also show here that the nature of the relevant non-equilibrium transitions depends on how incoming stimuli are processed in a noisy background, which might provide a useful and plausible tool to detect them in actual electroencephalogram (EEG) recordings. Additionally, we, thus, associate the existence of such phases, which reveal a brain operating at (non-equilibrium) criticality, with the emergence of the most interesting phenomena during memory tasks.

A FIRST MESOSCOPIC VIEW

An important and widely accepted fact here is that the synaptic actions connecting neurons are conditioned by "memories," namely, patterns that previously stored within during a kind of *learning process* and which are constantly adapted throughout the individual life due to new acquired information (Hebb, 1949; Amit, 1989). That is, by means of biophysical processes, we plastically store pieces of information (sensory perceptions, behavioral procedures, etc.) in our synapses, and we continually update that data while undergoing new circumstances (Marro and Torres, 2021). In practice, this happens to determine a very changing agenda of neuronal collaborations, which constantly conditions most high-level mental functions. On the other hand, in association to each mental process, there is now clear evidence that sets of synapses organize themselves into specific dynamics whose objective is to achieve constantly operating economy and proximity between different, even quite distant regions (Muñoz, 2018). Thus, by a proper combination of all this—mainly, continuous modulation of the synaptic interactions between neurons as well as eventual efficient coordination among groups of them—most elaborated mental properties emerge, including human intelligence and associated high-level functions such as working memory (Mongillo et al., 2008) or episodic memoires (Takeuchi et al., 2013). In particular, definite correlations between the familiar intelligence coefficient IQ and properties of the underlying neuron network, including its topology, effectiveness in transmitting information throughout, and the synaptic links dynamic activity have been reported (Li et al., 2009). This

complex scenario is celebrated at the light of the relative components simplicity producing it.

The fact is that the mind functions and how the brain manages to structure itself just result from cooperation among (very many) rather humble neurons mediating continuous dynamic actions of their synaptic links, which may eventually (a few or many) abstain from acting (Marro and Torres, 2021). A crucial aspect of this image is that synaptic fluctuations, especially those on short time scales, are determined to induce and constantly maintain a situation that physics describes as "*critical*" (Muñoz, 2018), which is in all similar to the one that characterizes the so-called *critical points* in condensed matter phenomena such as, for instance, *condensation* and *ferromagnetism* (Stanley, 1987). This intriguing situation is characterized, for instance, by the appearance of *avalanche* dynamics for neural population activity showing power law distributions (Beggs and Plenz, 2003), as well as by the existence of long-range correlations in space and time that have recently been associated with the sense of identity (Sugimura et al., 2021). Therefore, neuroscientists have today the possibility to explore the emergence of such "critical" conditions during brain activity as many theoretical and experimental works, including neuronal cultures, functional MRI (fMRI) data, EEG time series, have already revealed (Beggs and Plenz, 2003; Tagliazucchi et al., 2012; Yaghoubi et al., 2018; Fontenele et al., 2019). This notion of criticality, brought to these non-equilibrium settings from the study of equilibrium physical systems, can be extended to the concept of the Griffiths phase (Griffiths, 1969) in the brain with structural heterogeneity. In fact, the existence of *critical* zones has been computationally verified for humans and *Caenorhabditis elegans* connectomes (Moretti and Muñoz, 2013). This clarifies our understanding of how some high-level cognitive functions can emerge during brain operation, as well as possible clinical applications to some neurological disorders (Zimmern, 2020). Furthermore, particular features of such emerging critical state can be important to understand our capacity to solve problems and make decisions, thus conforming our intelligence and identity (Ezaki et al., 2020; Jiang et al., 2021) as we explore next.

INTELLIGENCE AND IDENTITY

One may highlight now two main aspects that concern intelligence. Firstly that we, as humans, essentially are kind of mixture of neuron collaborations and time variations of the intensities at which synapses happen to relate them. Even popular newspapers long ago recognized that "*Brainpower May Lie in Complexity of Synapses*" (Wade, 2008), then properly explaining that "*synapses get considerably more complex going up the evolutionary scale [. . .] It is likely this is one of the design principles by which the human brain is constructed.*" Indeed, it is sensible to say that our most important part as humans is the whole of our about 10^{15} synapses (DeFelipe et al., 2002). It is this mesh what likely houses our identity. In this way, each of us is uniquely—as a matter of probability—identified by information contained in all this enormous wired set of filaments (see **Figure 1**). This is the main of our identity, namely, all the data there plastically stored, which is a mixture of genetic inheritance and information frequently acquired and updated. Thanks to this immense and continually renewed data warehouse, the whole of processes that we associate with our intelligence are able, at any time and quickly, of remembering, combining, contrasting, and making decisions, computing, etc., making it possible what we call consciousness. Hence, the identity relying on this can diminish in any measure, due to loss or deterioration of part or that entire network or of the mechanisms that make it to correctly work and be efficiently useful, but we do not imagine how it could be transferred to another human being or exchanged with actual technology. In this sense, it is crucial understanding how the brain wiring network develops from conception until its mature form. In particular, this will help to understand the origin of brain network disorders, such as Autism Spectrum Disorder (Tang et al., 2014), schizophrenia (Keshavan et al., 1994), epilepsy (Andoh et al., 2019), and its Alzheimer deterioration.

In fact, understanding dynamical principles of how our brain structure develops has attracted some attention (Tetzlaff et al., 2010; Millán et al., 2018, 2019, 2021). It was reported, for example, that a suitable mathematical framework to study brain development is the master equation for the neuron degree probability distribution $p(k)$ (Johnson et al., 2009, 2010; Millán et al., 2018):

$$\frac{\boldsymbol{dp}\,(\boldsymbol{k},\boldsymbol{t})}{\boldsymbol{dt}} = \boldsymbol{T_{gain}}\,(\kappa, \boldsymbol{k}-\boldsymbol{1},\ldots)\,\boldsymbol{p}\,(\boldsymbol{k}-\boldsymbol{1},\boldsymbol{t}) + \boldsymbol{T_{loss}}\,(\kappa, \boldsymbol{k}+\boldsymbol{1},\ldots)\,\boldsymbol{p}\,(\boldsymbol{k}+\boldsymbol{1},\boldsymbol{t}) - \left[\boldsymbol{T_{gain}}\,(\kappa, \boldsymbol{k},\ldots) + \boldsymbol{T_{loss}}\,(\kappa, \boldsymbol{k},\ldots)\right]\boldsymbol{p}\,(\boldsymbol{k},\boldsymbol{t}) \quad (1)$$

where it is considered different type of microscopic mechanisms to add and remove synapses with time, here represented, respectively, by the transition rates $T_{gain}(\kappa,\ k, \ldots)$ to increase the number of neighbors of a given neuron from k *to* $k+1$, and $T_{loss}\left(\kappa,\ k,\ \ldots\right)$ to decrease the number of neighbors of a given neuron from k *to* $k-1$ (see **Figure 2** on top for a graph interpretation of the model). Assuming these depend on global topological aspects related with homeostatic considerations, such as the mean degree $\kappa = \langle k \rangle$ in the neural population, and local dependencies as the neuron degree k, this model explains the synaptic pruning curves observed in actual brains (Johnson et al., 2010; Millán et al., 2018). See the bottom left panel of **Figure 2** for typical synaptic pruning profiles generated with this model. The functions $T_{gain}(\kappa,\ k, \ldots)$ and $T_{loss}\left(\kappa,\ k,\ \ldots\right)$ have a neurophysiological justification, as it is well known that neuron electrical activity regulates neural connectivity inducing axonal branch formation (Uesaka et al., 2006) and synaptic refinement (Vonhoff and Keshishian, 2017). Since neuron activity depends on the net current the neuron receives from its neighbors—being larger for increasing number of neighbors—ultimately it depends on its degree k. Also synaptic growth and death depends on the concentration of molecules that can diffuse through the whole neural medium, and given that calcium ions activate proteins involved in regulation of synaptic growth and pruning (Jourdain et al., 2003; Cornelia Koeberle et al., 2017), such processes cannot be considered only local. On the other hand, one may consider explicitly not only topological factors but also neurophysiological influences

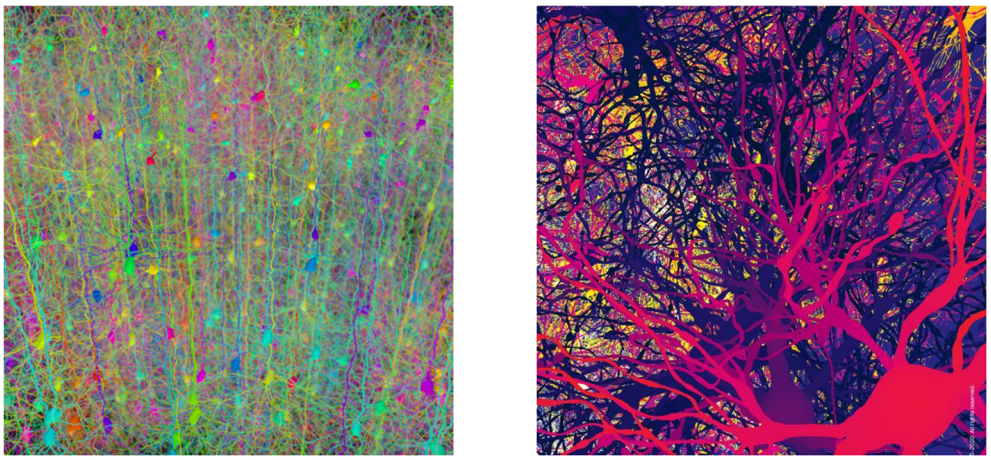


FIGURE 1 | Forests of somas and synapses. The image on the left is a forest of synthetic pyramidal dendrites grown using Cajal's laws of neuronal branching (Image Credit: Hermann Cuntz, licensed under Creative Commons Attribution License, PLoS Comput Biol 6(8): ev06.i08. https://doi.org/10.1371/image.pcbi.v06.i08). The image on the right is a visualization of neurons in a digitally reconstructed thalamus model from the "Blue Brain Project" in École Polytechnique Fédérale de Lausanne (Image Credit: ©Blue Brain Project/EPFL 2005 – 2020. All rights reserved).

such as the synaptic currents arriving to each neuron I_{syn}. This has allowed to investigate the interplay between brain function and network topology during development (Millán et al., 2018, 2019, 2021). More plausible realistic assumptions could be included within this framework, for example, the interplay between subcellular mechanisms such as intracellular calcium dynamics and astrocytes function since it has been recently reported that astrocytes actively contribute to synaptic pruning and developmental refinement of neural circuits in the brain (Lee et al., 2021).

LOOKING INTO OUR MINDS

Another important aspect of the mind that is worth to be highlighted here—since it has practical and conceptual relevance, though one might at first glance feel it is just a technical aspect—is that transitions among mental states of qualitatively different properties often can be appropriately interpreted, using physics terminology, as *non-equilibrium phase transitions*. In fact, this concept (Marro and Dickman, 2005) in general concerns a transition among different macroscopic well-defined states (phases) in a system that, due to any fluxes or other interactions with the outside, cannot be in thermodynamic equilibrium, so that one cannot characterize in practice it by any Hamiltonian function. Concerning the brain, there are many facts that prevent equilibrium and from writing such a function, e.g., the different type of currents and fields that affect the neural excitability at the subcellular, cellular, and network level and time-dependent external fields such as the stimuli currents arriving from the senses. In spite of this and in analogy with a thermodynamic phase transition, the brain shows some relatively sharp changes at certain values of relevant parameters where its response to quite small perturbations exhibits a very large susceptibility and propagates in time and space without damping as in the familiar equilibrium criticality in, for instance, condensation. Under this particular condition, the brain is able to develop its characteristic cognitive functions such as decision tasks, attention breaks, optimal processing of information from the outside, or the processing of episodic like memories.

In other words, what we casually name "gray matter" is nothing but "condensed matter," in the sense that it undergoes (possibly dynamic) changes formally similar to those shown by other materials, so much so that they are described with precisely the same mathematical structures (Marro and Dickman, 2005). That is, there are qualitative changes in the mind, whether they are dynamic while brain function or structural along the individual's evolution, that happen to be essentially equivalent in a formal sense (Marro and Torres, 2021) to *phase transitions* in physics, such as ferromagnetism, superconductivity, and superfluidity. The difference is conceptual more than practical in the sense that the later ones are of a thermodynamic nature, as they affect isolated systems in the state known as *thermodynamic equilibrium*. On the contrary, the mind and the nervous system are *open systems*, which experience stationary flows of matter, energy, and/or information with its surroundings and typically exhibit different types of inhomogeneity, so that they constantly are far from that equilibrium state. A practical consequence of the otherwise similarity between phases and mental states is that both, structure and dynamics of the brain are likely to be studied with very powerful methods and concepts developed in the study of matter and radiation (Marro and Dickman, 2005; Marro and Torres, 2021).

This important fact happens to offer a solid, conceptual and mathematical, support thus washing out a certain mystery and consequent misgivings initially affecting to reports that the mind shows avalanches, seismicity, and critical and chaotic dynamics, which for a time was not considered befitting of the brain. Now it is quite clear why mental states often entail long-range

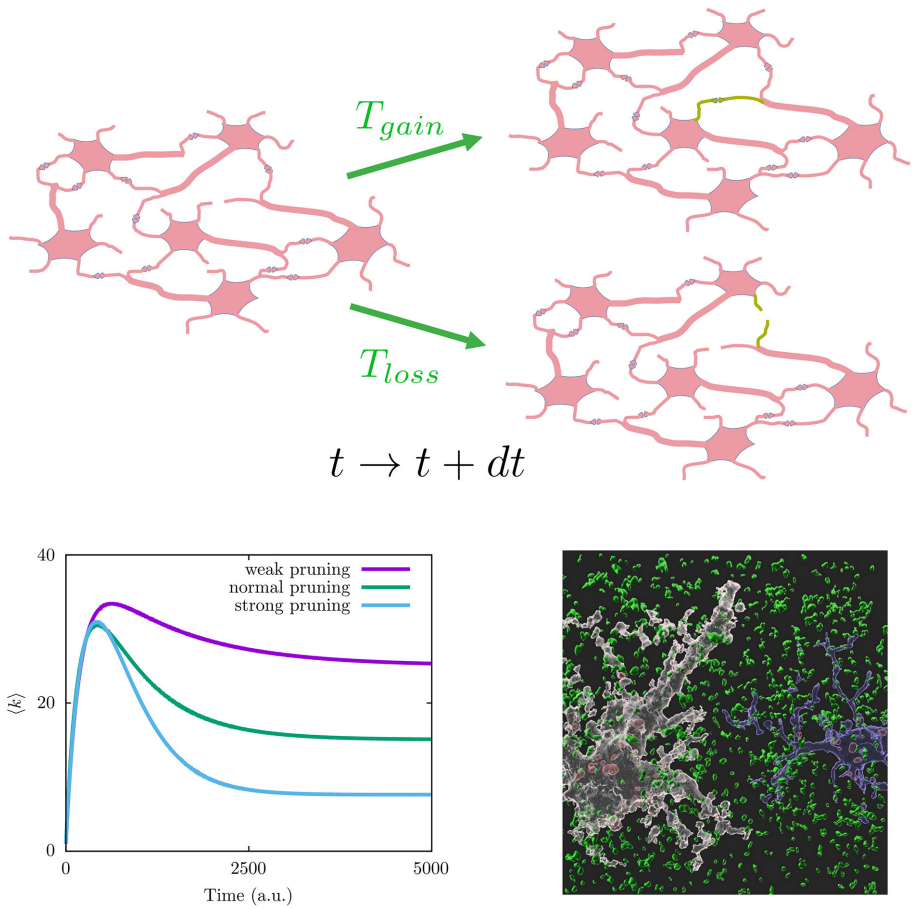


FIGURE 2 | (Top) Graph interpretation of the mathematical model in Equation (1). Given a neuronal network configuration at time t (left), a new synapse can be generated (lost) with probability $T_{gain}(T_{loss})$ at time $t + dt$ (*green connections on the right neuronal network configurations*). (Bottom left) Different synaptic pruning curves produced by Equation (1) (details for particular choices for transition probabilities T_{gain} *and* T_{loss} can be found in Johnson et al., 2010). (Bottom right) Image showing synapse phagocytosis by astrocytes reported in mouse hippocampus (Lee et al., 2021). Presynapses are colored in green, astrocytes in white, and microglia in blue. Phagocytosed presynapses by glia were shown in red. Image has been taken from the Korea Advanced Institute of Science and Technology (KAIST) (Image credit: Won-Suk Chung, https://www.kaistglia.org/).

correlations, as this occurs at critical points in physics where it is well known it allows any part of the system to be strongly receptive to what happens in any other, and vice versa. This is surely very important, since physics shows how the most extraordinary phenomenology emerges in this way. Moreover, the phenomena that are associated to such strong extensive correlations and large susceptibility in the mind, given the prevailing inhomogeneity and flows with the environment in this case, turn out to be even more varied and bizarre than in systems at thermodynamic equilibrium. Even more, by adding without reluctance the brain to the set of physical or, say, condensed matter systems, a new and exciting world becomes accessible to experiments. For example, it has been described how one may detect transitions between mental states by simple, e.g., "psychophysics" experiments that, studying the propagation of signals through the brain, report on the existence of *stochastic resonances* (Manjarrez et al., 2002, 2003; Yasuda et al., 2008; Torres et al., 2011). And it has thus been shown how these turn out to correspond precisely with *phase* transitions clearly denouncing very significant changes of the mind dynamics (Torres and Marro, 2015; Marro and Torres, 2021). Particularly, this has allowed to interpreting the celebrated cerebral "rhythms of activity"—those named *alpha*, *beta*, and *gamma waves* and *ultrafast oscillations*—whose existence was first revealed by the time series of EEGs long ago. Actually, a simple model has recently shown (Galadí et al., 2020; Pretel et al., 2021) how the changes between these types of oscillatory behavior are just transitions between *phases* or mental states that one can classify and decipher in neuroscience performing appropriate EEG and magnetoencephalography (MEG) experimental setups (see **Figure 3**).

At the light of parallel situations in physics, one expects that a main detail within this scenario of emerging (non-equilibrium) mental phases will be the topology of the brain, which is certainly expected to condition its functions and interactions with the environment. In particular, one may anticipate that different species will exhibit a different chart of characteristic non-equilibrium mental phases, as illustrated in **Figure 4** using a

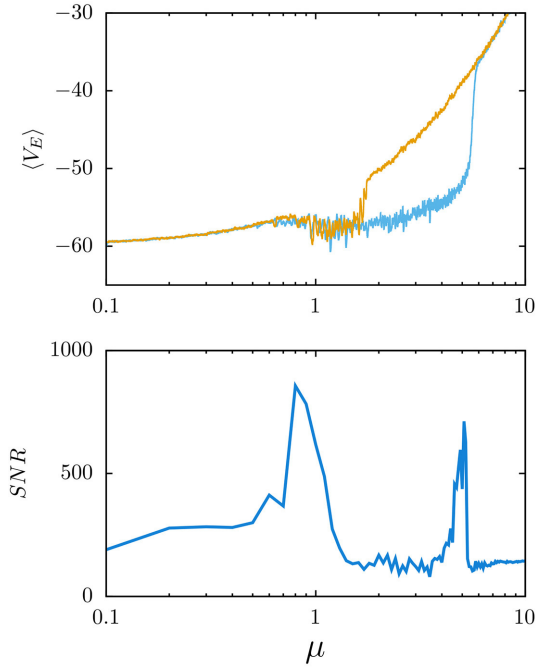


FIGURE 3 | Electroencephalogram (EEG) waves in stochastic resonance phenomena. (Top) Model generation of waves, similar to the ones recorded in actual EEGs. This is for an excitation/inhibition balanced neural population, and the mean membrane voltage of all excitatory neurons is plotted as a function of the level of arriving excitatory uncorrelated inputs from outside of the population—measured by the parameter μ (the mean value of random depolarizing inputs that an excitatory neuron receives during 4 ms from other regions outside of the population)—and in the presence of short-term synaptic depression (Pretel et al., 2021). Note that the waves emerge around $\mu \approx 0.6$ and disappear in an explosive phase transition for stronger noise levels ($\mu \approx 5$). This is a consequence of depression, which reduces the effect of the strong activity of excitatory neurons over the inhibitory ones thus reducing the firing of inhibitory neurons and increasing further the activity of excitatory neurons. (Bottom) Signal-to-noise ratio (SNR) vs. μ when adding a weak sinusoidal input to each neuron while processing the weak signal, which allows detecting relevant phase transitions because of an enhancement of the SNR around the phase transition points.

simple model implemented with a number of connectomes data (see Torres and Marro, 2015 for model details).

Concerning mind phases, it was just shown (Calim et al., 2020) emergence of *chimera states*, namely, oscillatory phases where some neurons of the population are coherently oscillating in synchrony and the rest are oscillating out of phase in an asynchronous regime (see **Figure 5**). This seems robust in many situations, including spiking and bursting neuronal populations and complex network topologies including hybrid synaptic schemes with chemical and electrical connections. Therefore, they may possibly occur also in actual cases, where they might be related, for instance, with neural activity during uni-hemispheric sleep in dolphins and with the emergence of bumps of activity in working memory tasks (Compte et al., 2000). Since current studies have demonstrated their occurrence in parameter's regions between traveling wave phases and coherent synchronous phases, these states could be important to increase synchrony in some brain areas or to prevent epileptic seizures associated to traveling wave behavior.

DISCUSSION AND PERSPECTIVES

Summing up, this perspective article presents some promising ideas and research lines for the study of the brain function based on simple biologically motivated neural population models that have been analyzed using statistical physics methods. This uncovers intriguing emerging phenomena due to cooperation of the systems basic elements, namely, neurons and synapses that are related by a complex network topology. In particular, we emphasize here the possibility that critical phenomena, similar to those in condensed matter, occur in the brain, which is quite sensible given the universality of the basic nature laws within such scenarios. We, thus, describe emerging phase transitions separating non-equilibrium phases that seem to characterize mental states as well as brain functions that seemingly involved by human identity and intelligence. No need to say that our (statistical physics) point of view here has limitations. For example, it is difficult to include all the microscopic details that can affect neuronal and synaptic dynamics, and some of them, including also the connection networks details, are yet to be fully described by neurobiologists. On the other hand, it yet needs to be clarified whether some brain functions are just emergent collective phenomena. For example, it is still difficult to fully understand and quantify how our subjective experience is coupled to the emergent process arising from the complex interrelation of the elements, mainly neurons and synapses. In any case, our approach here—also followed by many other colleagues—may be seen as a first meaningful analysis in which one may easily incorporate additional details as provided by new experiments.

On the other hand, it is remarkable how the whole of the framework above (see also Marro and Torres, 2021, and references therein) naturally leads to motivating extensions. In particular, returning to the notion of intelligence, it makes sense to assume within this scenario that, as a substantial part of the cognitive process, the mind constantly and quickly simulates relevant events and alternatives. That is, the mind would be producing, mostly unconsciously and for its own and immediate use, kind of well-informed "short films" including a variety of data, feelings and emotions pertinent to each case. This means that we instinctively imagine options that help us to decide at every moment what could be the most convenient in view of the "maximum" of available information (Wang, 2012). We can imagine that this maximum includes, in addition to all the data stored in our synapses relevant for the task in question, sensory data on the spatial and temporal environment and other predictions perhaps generated on the fly. In fact, an essential deficit in today's computers imitating intelligence would be this intimate relationship "back and forth" between memories and current processing that surely determines our decision-making and characterizes our brain functions.

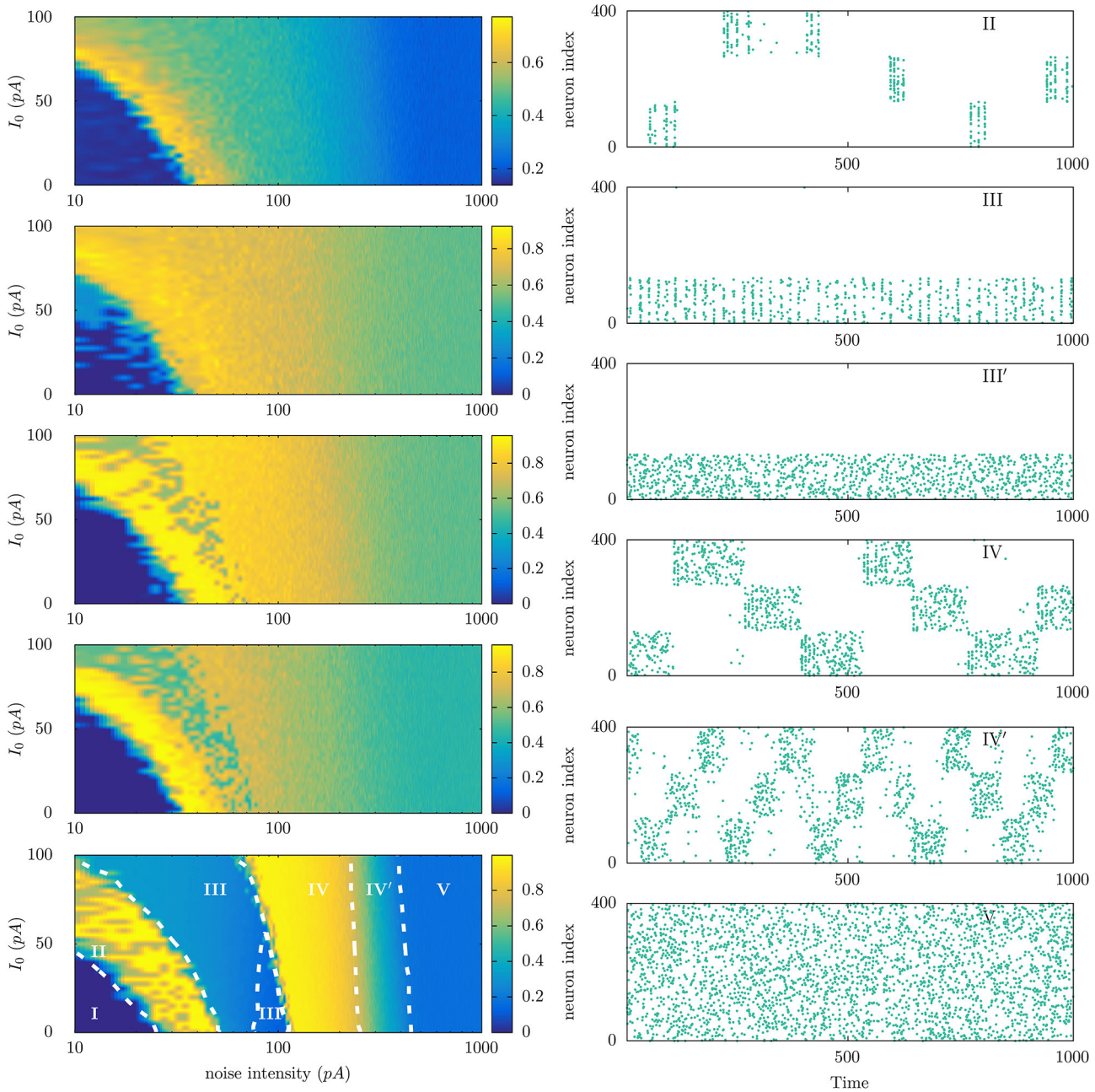


FIGURE 4 | (Left) Phase diagrams showing the emergence of different non-equilibrium phases in neural populations that are connected with each other using different connectome data corresponding to different animal species. For each phase diagram, the vertical axis corresponds to the level of the constant current I_0 injected to each neuron in the population to increase its excitability level, and the horizontal axis corresponds to the intensity of a noisy current also injected in each neuron to account for the possibility of random uncorrelated noises—arising from different sources—affecting neuron dynamics. Model details are described in Ref. Torres and Marro, 2015. From top to bottom connectomes are of *Caenorhabditis elegans* worm, pigeon, macaque, human, and a fully connected network. The color code represents the level of the time averaged overlap (ranged between 0 and 1) between the neuron population activity and any of the three particular patterns of activity stored at the synaptic maximal conductances. (Right) Raster plots illustrating the dynamical features of the different mental states observed in the parameter's regions marked with roman letters in the left bottom panel.

Consequently, it seems that we may perceive *intelligence* as the result of all this, combined with the ability with which each individual is able to handle it. This is a human quality that shows to us as a kind of multifaceted device able of attending, perceiving, relating, and predicting. All this thanks to that critical, very effective dynamics described above commissioned to establish broad and rapid correlations between any part of our brain and the rest of our nervous system, including the senses as "windows" to the outside. In addition, our mind is so dynamic and adaptive that, not only does it generate new information

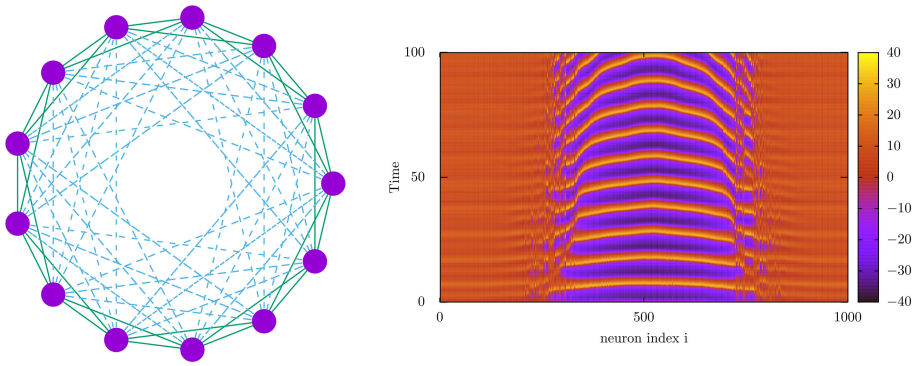


FIGURE 5 | (Left) Network architecture used for the study of the emergence of *chimera* behavior using nearest neighbor electrical connections (solid lines) and long-range chemical synapses (dashed lines). (Right) Density plot depicting the time-dependent voltage traces of a 1,000 neuron population interconnected with the topology in the left panel (Calim et al., 2020). This shows the emergence of two chimeras separating different neuron subpopulations with different dynamical regimes, one (the center of the image) with high-amplitude normal spiking activity and the others (the non-centered regions) with high-voltage low amplitude oscillations, constituting a so-called *chaotic amplitude chimera* (Calim et al., 2020). Voltage membrane dynamics of the neurons (purple circles in the right panel) has been computed using a Morris–Lecar neuron model (Morris and Lecar, 1981).

using its warehouse to make predictions about situations that arise, but also reinforces or weakens our memories by adapting them to the success or failure of those predictions. Thus, apart from its relationship with the individual's ability to correlate, reason, resolve, etc., intelligence brings learning effectiveness. This increases our ability to anticipate threats and visualize even the most remote possibilities while reviewing the past, so that intelligence also seeks and, at best, achieves a better forecast of the future in our environment.

In this scenario, we may wonder about the control we have on our intelligence, besides circumstances that develop and activate (or not) throughout our lives under the rule of genetics, which probably determines, at least in part, aspects such as the intelligence capacity, disease propensity, or gender identity. In any case, it is not at all realistic to imagine that determinism governs the mind, e.g., a noise brain level is clearly noted during the state of consciousness (Lendner et al., 2020) and influence our behavior (Waschke et al., 2021). Moreover, the mechanisms involved are quite adaptive—meaning that they may be conditioned by eventual and rather random interactions with the outside, including social ones hanging on others—, and there are clear indications of defects and contradictions in relation to our decision-making processes. One source for these is surely the prejudices, traumas, and "manias" that, possibly hidden from ourselves, will induce biases not necessarily consistent with the reality in those "short films" we make. This, in addition to lessening our mind finesse and perhaps spoiling our best forecasts, would ruin a hypothetical determinism. Nor we should rule out that, as it is characteristic of natural phenomena, the aforementioned processes of memory, contrast, simulation, prediction, and decision include some randomness to its unconscious character. In short, decisions are probably made with certain autonomy, that is, without explicit real consent, that perhaps would only be explicitly communicated to us an instant after being taken without our will.

Endorsing the above, there is evidence that voluntary acts are preceded of subtle electrical changes in the brain, which would reflect a preparation process before the individual realizes it. In particular, perfecting experiments decades ago, recent EEGs and magnetic resonance studies show that the frontal cortex displays indications of the action to be performed a few seconds before the subject happens to "know" it (see, for instance, Soon et al., 2008; which confirms previous experiments of the neurologist Benjamin Libet in the 1980s), which suggests that the control we have of our own mind is limited. Actually, it seems that there are a few seconds in which we do not have a conscious supervision of each of our acts. This delay is compatible with the image of intelligence given above, according to which we evaluate—consciously, partially consciously or unconsciously—our options, which takes a finite time.

Also interesting are some consequences of the above on the concept of consciousness. Imagine that, at the request of an emergency phone operator, we have to distinguish whether a person just injured in an accident is "conscious or not." Could we conclude with confidence? Back home, we check for consciousness in the thesaurus. It will say something like: "*our spontaneous knowledge, more or less vague and reflective, of the surrounding reality*." OK but insufficient; it is important to note that it is not an eventual passive act, but a *perception*, where knowledge needs to be "perfected" with an intuitive element that constantly conditions us, to the point that it not only integrates us into the near context but also makes that we recognize ourselves in it. It follows that, as intelligence and identity, consciousness rests in memory, that is, in how we do to maintain huge stores of information and, quickly and automatically, we are able to recovering any specific portion that we might need. In short,

these are human activities that require both interaction with the environment and ability to experience subjective sensations, so that they rest in the mind. According to what we have seen above, today one can say that consciousness and identity are a global property of the nervous system, especially in relation to the whole of its synapses.

We end up noticing that a main conclusion here may be that, as compared to other cooperative—natural or artificial—systems, an adequate activity of the intimate neural relationships is essential for the superiority of our minds. More than a century ago, when the matter was still believed to be a continuous medium, Santiago Ramón y Cajal noted the existence of those synapses that for him were "*mysterious butterflies of the soul whose beating of wings who knows if one day will clarify the secret of mental life.*" In a way, this is fully confirmed. We know that the versatility and power of the mind is inherent to the modulation on several time scales—kind of breathing, from calm to anxious—that these butterflies make of neuronal cooperation. They house indeed our intelligence, identity, and consciousness.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

JT and JM contributed to conception and design of the study and wrote the manuscript. Both authors contributed to manuscript revision, read, and approved the submitted version.

ACKNOWLEDGMENTS

This study is part of the Project of I+D+i Ref. PID2020-113681GB-I00, financed by MICIN/AEI/10.13039/501100011033 and FEDER A way to make Europe and also financed by FEDER/*Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades*/Project Ref. P20_00173. We thank M.A. Muñoz and J. Pretel for fruitful discussions.

REFERENCES

Amit, D. (1989). *Modeling Brain Function: The World of Attractor Neural Networks*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511623257

Andoh, M., Ikegaya, Y., and Koyama, R. (2019). Synaptic pruning by microglia in epilepsy. *J. Clin. Med.* 8, 2170. doi: 10.3390/jcm8122170

Azevedo, F. A., Carvalho, L. R., Grinberg, L. T., Farfel, J. M., Ferretti, R. E., Leite, R. E., et al. (2009), Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol.* 513, 532–541. doi: 10.1002/cne.21974

Beggs, J. M., and Plenz, D. (2003). Neuronal avalanches in neocortical circuits. *J. Neurosci.* 23, 11167–11177. doi: 10.1523/JNEUROSCI.23-35-11167.2003

Calim, A., Torres, J. J., Ozer, M., and Uzuntarla, M. (2020). Chimera states in hybrid coupled neuron populations. *Neural Netw.* 126, 108–117. doi: 10.1016/j.neunet.2020.03.002

Compte, A., Brunel, N., Goldman-Rakic, P. S., and Wang, X.-J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex.* 10, 910–23. doi: 10.1093/cercor/10.9.910

Cornelia Koeberle, S., Tanaka, S., Kuriu, T., Iwasaki, H., Koeberle, A., Schulz, A., et al. (2017). Developmental stage-dependent regulation of spine formation by calcium-calmodulin-dependent protein kinase IIα and Rap1. *Sci. Rep.* 7, 13409. doi: 10.1038/s41598-017-13728-y

DeFelipe, J., Alonso-Nanclares, L., and Arellano, J. I. (2002). Microstructure of the neocortex: comparative aspects. *J. Neurocytol.* 31, 299–316.

Ezaki, T., Fonseca dos Reis, E., Watanabe, T., Sakaki, M., and Masuda, N. (2020). Closer to critical resting-state neural dynamics in individuals with higher fluid intelligence. *Commun. Biol.* 3, 52. doi: 10.1038/s42003-020-0774-y

Fontenele, A. J., de Vasconcelos, N. A. P., Feliciano, T., Aguiar, L. A. A., Soares-Cunha, C., Coimbra, B., et al. (2019). Criticality between cortical states. *Phys. Rev. Lett.* 122, 208101. doi: 10.1103/PhysRevLett.122.208101

Galadí, J. A., Torres, J. J., and Marro, J. (2020). Emergence and interpretation of oscillatory behaviour similar to brain waves and rhythms. *Commun. Nonl. Sci. Numer. Simul.* 83, 105093. doi: 10.1016/j.cnsns.2019.105093

Griffiths, R. B. (1969). Nonanalytic behavior above the critical point in a random ising ferromagnet. *Phys. Rev. Lett.* 23, 17–19. doi: 10.1103/PhysRevLett.23.17

Hebb, D. O. (1949). *The Organisation of Behaviour*. New York, NY: John Wiley and Sons.

Jiang, L., Qiao, K., and Li, C. (2021). Distance-based functional criticality in the human brain: intelligence and emotional intelligence. *BMC Bioinform.* 22: 32. doi: 10.1186/s12859-021-03973-4

Johnson, S., Marro, J., and Torres, J. J. (2010). Evolving networks and the development of neural systems. *J. Stat. Mech.* 2010, P03003. doi: 10.1088/1742-5468/2010/03/P03003

Johnson, S., Torres, J. J., and Marro, J. (2009). Nonlinear preferential rewiring in fixed-size networks as a diffusion process. *Phys. Rev.* E79, 050104(R). doi: 10.1103/PhysRevE.79.050104

Jourdain, P., Fukunaga, K., and Muller, D. (2003). Calcium/calmodulin-dependent protein kinase II contributes to activity-dependent filopodia growth and spine formation. *J. Neurosci.* 23, 10645–9. doi: 10.1523/JNEUROSCI.23-33-10645.2003

Keshavan, M. S., Anderson, S., and Pettergrew, J. W. (1994). Is schizophrenia due to excessive synaptic pruning in the prefrontal cortex? The feinberg hypothesis revisited. *J. Psychiatr. Res.* 28, 239–265. doi: 10.1016/0022-3956(94)90009-4

Lee, J.H., Kim, J.Y., Noh, S., Lee, H., Lee, S.Y., Mun, J.Y., et al. (2021). Astrocytes phagocytose adult hippocampal synapses for circuit homeostasis. *Nature.* 590, 612–617. doi: 10.1038/s41586-020-03060-3

Lendner, J.D., Helfrich, R.F., Mander, B.A., Romundstad, L., Lin, J.J., Walker, M.P., et al. (2020). An electrophysiological marker of arousal level in humans. *eLife.* 9, e55092. doi: 10.7554/eLife.55092

Li, Y., Liu, Y., Li, J., Qin, W., Li, K., et al. (2009). Brain Anatomical Network and Intelligence. *PLOS Comput. Biol.* 5, e1000395. doi: 10.1371/journal.pcbi.1000395

Manjarrez, E., Diez-Martínez, O., Mendez, I., and Flores, A. (2002). Stochastic resonance in human electroencephalographic activity elicited by mechanical tactile stimuli. *Neurosci. Lett.* 324, 213–216. doi: 10.1016/S0304-3940(02)00212-4

Manjarrez, E., Rojas-Piloni, G., Méndez, I., and Flores, A. (2003). Stochastic resonance within the somatosensory system: effects of noise on evoked field potentials elicited by tactile stimuli. *J. Neurosci.* 23, 1997–2001. doi: 10.1523/JNEUROSCI.23-06-01997.2003

Marro, J. (2014). *Physics, Nature and Society – A Guide to Order and Complexity in our World*. Springer, Berlin.

Marro, J., and Dickman, R. (2005). *Nonequilibrium Phase Transitions in Lattice Models.* Cambridge University Press, Cambridge.

Marro, J., and Torres, J. J. (2021), *Phase Transitions in Grey Matter – Brain Architecture and Mind Dynamics.* American Institute of Physics Pub., New York. doi: 10.1063/9780735421769

Millán, A. P., Torres, J. J., Johnson, S., and Marro, J. (2018). Concurrence of form and function in developing networks and its role in synaptic pruning. *Nature Comm.* 9, 2236. doi: 10.1038/s41467-018-04537-6

Millán, A. P., Torres, J. J., Johnson, S., and Marro, J. (2021). Growth strategy determines the memory and structural properties of brain networks. *Neural Netw.* 142, 44–56. doi: 10.1016/j.neunet.2021.04.027

Millán, A. P., Torres, J. J., and Marro, J. (2019). How memory conforms to brain development. *Front. Comput. Neurosci.* 13. doi: 10.3389/fncom.2019.00022

Mongillo, G., Barak, O., and Tsodyks, M. (2008). Synaptic theory of working memory. *Science.* 319, 1543–6. doi: 10.1126/science.1150769

Moretti, P., and Muñoz, M. (2013). Griffiths phases and the stretching of criticality in brain networks. *Nat. Commun.* 4, 252. doi: 10.1038/ncomms3521

Morris, C., and Lecar, H. (1981). Voltage Oscillations in the barnacle giant muscle fiber. *Biophys. J.* 35, 193–213. doi: 10.1016/S0006-3495(81)84782-0

Muñoz, M. A. (2018). Criticality and dynamical scaling in living systems. *Rev. Mod. Phys.* 90, 031001. doi: 10.1103/RevModPhys.90.031001

Pretel, J., Torres, J. J., and Marro, J. (2021). EEGs disclose significant brain activity correlated with synaptic fickleness. *Biology.* 10, 647. doi: 10.3390/biology10070647

Roth, G., and Dicke, U. (2005), Evolution of the brain and intelligence. *Trend Cogn. Sci.* 9, 250–257. doi: 10.1016/j.tics.2005.03.005

Soon, C.S., Brass, M., Heinze, H.J., and Haynes, J.D. (2008). Unconscious determinants of free decisions in the human brain. *Nat. Neurosci.* 11, 543–545. doi: 10.1038/nn.2112

Stanley, E. (1987). *Introduction to Phase Transitions and Critical Phenomena*. Oxford University Press; Reprint edition.

Sugimura, K., Iwasa, Y., Kobayashi, R., et al. (2021). Association between long-range temporal correlations in intrinsic EEG activity and subjective sense of identity. *Sci. Rep.* 11, 422. doi: 10.1038/s41598-020-79444-2

Tagliazucchi, E., Balenzuela, P., Fraiman, D., and Chialvo, D.R. (2012). Criticality in large-scale brain fMRI dynamics unveiled by a novel point process analysis. *Front. Physiol.* 3, 15. doi: 10.3389/fphys.2012.00015

Takeuchi, T., Duszkiewicz, A. J., and Morris, R. G. (2013). The synaptic plasticity and memory hypothesis: encoding, storage and persistence. *Philosoph. Transac. R. Soc. London. B.* 369, 20130288. doi: 10.1098/rstb.2013.0288

Tang, G., Gudsnuk, K., Kuo, S.-H., Cotrina, M. L., Rosoklija, G., Sosunov, A., et al. (2014). Loss of mtor-dependent macroautophagy causes autistic-like synaptic pruning deficits. *Neuron.* 83, 1131–1143. doi: 10.1016/j.neuron.2014.07.040

Tetzlaff, C., Okujeni, S., Egert, U., Wörgötter, F., and Butz, M. (2010). Self-organized criticality in developing neuronal networks. *PLoS Comput. Biol.* 6, e1001013. doi: 10.1371/journal.pcbi.1001013

Torres, J. J., and Marro, J. (2015). Brain performance versus phase transitions. *Sci. Rep.* 5, 12216. doi: 10.1038/srep12216

Torres, J. J., Marro, J., and Mejias, J. F. (2011). Can intrinsic noise induce various resonant peaks?" *New. J. Phys.* 13, 053014. doi: 10.1088/1367-2630/13/5/053014

Uesaka, N., Ruthazer, E. S., and Yamamoto, N. (2006). The role of neural activity in cortical axon branching. *Neuroscientist.* 12, 102–106. doi: 10.1177/1073858405281673

Vonhoff, F., and Keshishian, H. (2017). Activity-dependent synaptic refinement: new insights from drosophila. *Front. Syst. Neurosci.* 11 doi: 10.3389/fnsys.2017.00023

Wade, N. (2008). *Brainpower May ◦lie in Complexity of synapses*. The New York Times.

Wang, X.-J. (2012). Neural dynamics and circuit mechanisms of decision-making. *Curr. Op. Neurobiol.* 22, 1–8. doi: 10.1016/j.conb.2012.08.006

Waschke, L., Kloosterman, N. A., Obleser, J., and Garrett, D. D. (2021). Behavior needs neural variability. *Neuron.* 109, 751–766. doi: 10.1016/j.neuron.2021.01.023

Yaghoubi, M., de Graaf, T., Orlandi, J.G., Girotto, F., Colicos, M.A., and Davidsen, J. (2018). Neuronal avalanche dynamics indicates different universality classes in neuronal cultures. *Sci. Rep.* 8, 3417. doi: 10.1038/s41598-018-21730-1

Yasuda, H., Miyaoka, T., Horiguchi, J., Yasuda, A., Hänggi, P., and Yamamoto, Y. (2008). Novel class of neural stochastic resonance and error-free information transfer. *Phys. Rev. Lett.* 100, 118103. doi: 10.1103/PhysRevLett.100.118103

Zimmern, V. (2020). Why brain criticality is clinically relevant: a scoping review. *Front. Neural. Circ.* 26. doi: 10.3389/fncir.2020.00054

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.

frontiers | Frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 27 April 2022
doi: 10.3389/fncom.2022.876315

Dynamics and Information Import in Recurrent Neural Networks

Claus Metzner[1*†] *and Patrick Krauss*[1,2,3†]

[1] Neuroscience Lab, University Hospital Erlangen, Erlangen, Germany, [2] Cognitive Computational Neuroscience Group, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany, [3] Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany

Recurrent neural networks (RNNs) are complex dynamical systems, capable of ongoing activity without any driving input. The long-term behavior of free-running RNNs, described by periodic, chaotic and fixed point attractors, is controlled by the statistics of the neural connection weights, such as the density d of non-zero connections, or the balance b between excitatory and inhibitory connections. However, for information processing purposes, RNNs need to receive external input signals, and it is not clear which of the dynamical regimes is optimal for this information import. We use both the average correlations C and the mutual information I between the momentary input vector and the next system state vector as quantitative measures of information import and analyze their dependence on the balance and density of the network. Remarkably, both resulting phase diagrams $C(b, d)$ and $I(b, d)$ are highly consistent, pointing to a link between the dynamical systems and the information-processing approach to complex systems. Information import is maximal not at the "edge of chaos," which is optimally suited for computation, but surprisingly in the low-density chaotic regime and at the border between the chaotic and fixed point regime. Moreover, we find a completely new type of resonance phenomenon, which we call "Import Resonance" (IR), where the information import shows a maximum, i.e., a peak-like dependence on the coupling strength between the RNN and its external input. IR complements previously found Recurrence Resonance (RR), where correlation and mutual information of successive system states peak for a certain amplitude of noise added to the system. Both IR and RR can be exploited to optimize information processing in artificial neural networks and might also play a crucial role in biological neural systems.

Keywords: recurrent neural networks (RNNs), dynamical system, edge of chaos, information processing, resonance phenomena

OPEN ACCESS

Edited by:
Joaquín J. Torres, University of Granada, Spain

Reviewed by:
Ana P. Millan, Amsterdam University Medical Center, Netherlands
Jorge F. Mejias, University of Amsterdam, Netherlands

***Correspondence:**
Claus Metzner
claus.metzner@gmail.com

†These authors have contributed equally to this work

Received: *15 February 2022*
Accepted: *04 April 2022*
Published: *27 April 2022*

Citation:
Metzner C and Krauss P (2022) Dynamics and Information Import in Recurrent Neural Networks. Front. Comput. Neurosci. 16:876315. doi: 10.3389/fncom.2022.876315

INTRODUCTION

At present, the field of Machine Learning is strongly dominated by feed-forward neural networks, which can be optimized to approximate an arbitrary vectorial function $\mathbf{y} = \mathbf{f}(\mathbf{x})$ between the input and output spaces (Funahashi, 1989; Hornik et al., 1989; Cybenko, 1992). Recurrent neural networks (RNNs) however, are a much broader class of models, which encompass the feed-forward architectures as a special case, but which also include partly recurrent systems, such as contemporary LSTMs (long short-term memories) (Hochreiter and Schmidhuber, 1997) and classical Jordan or Elman networks (Cruse, 2006), up to fully connected systems without any

layered structure, such as Hopfield networks (Ilopfield, 1982) or Boltzmann machines (Hinton and Sejnowski, 1983). Due to the feedback built into these systems, RNNs can learn robust representations (Farrell et al., 2019), and are ideally suited to process sequences of data such as natural language (LeCun et al., 2015; Schilling et al., 2021a), or to perform sequential-decision tasks such as spatial navigation (Banino et al., 2018; Gerum et al., 2020). Furthermore, RNNs can act as autonomous dynamical systems that continuously update their internal state $\mathbf{s}_t$ even without any external input (Gros, 2009), but it is equally possible to modulate this internal dynamics by feeding in external input signals $\mathbf{x}_t$ (Jaeger, 2014). Indeed, it has been shown that RNNs can approximate any open dynamical system $\mathbf{s}_{t+1} = \mathbf{g}(\mathbf{s}_t, \mathbf{x}_t)$ to arbitrary precision (Schäfer and Zimmermann, 2006).

It is therefore not very surprising that biological neural networks are also highly recurrent in their connectivity (Binzegger et al., 2004; Squire et al., 2012), so that RNN models play an important role in neuroscience research as well Barak (2017) and Maheswaranathan et al. (2019). Modeling natural RNNs in a realistic way requires the use of probabilistic, spiking neurons, but even simpler models with deterministic neurons already have highly complex dynamical properties and offer fascinating insights into how structure controls function in non-linear systems (Krauss et al., 2019b,c). For example, we have demonstrated that by adjusting the density d of non-zero connections and the balance b between excitatory and inhibitory connections in the RNN's weight matrix, it is possible to control whether the system will predominantly end up in a periodic, chaotic, or fixed point attractor (Krauss et al., 2019b). Understanding and controlling the behavior of RNNs is of crucial importance for practical applications (Haviv et al., 2019), especially as meaningful computation, or information processing, is believed to be only possible at the "edge of chaos" (Bertschinger and Natschläger, 2004; Natschläger et al., 2005; Legenstein and Maass, 2007; Schrauwen et al., 2009; Büsing et al., 2010; Toyoizumi and Abbott, 2011; Dambre et al., 2012).

In this paper, we continue our investigation of RNNs with deterministic neurons and random, but statistically controlled weight matrices. Yet, the present work focuses on another crucial precondition for practical RNN applications: the ability of the system to store information, i.e., to "take up" external information and to incorporate it into the ongoing evolution of the internal system states. For this purpose, we first set up quantitative measures of information import, in particular the input-to-state correlation $C(\mathbf{x}_t, \mathbf{s}_{t+1})$, which is defined as the root-mean-square (RMS) average of all pairwise neural correlations between the momentary input $\mathbf{x}_t$ and the subsequent system state $\mathbf{s}_{t+1}$. Furthermore, we compute the input-to-state mutual information $I(\mathbf{x}_t, \mathbf{s}_{t+1})$, an approximation for the mean pairwise mutual information between the same two quantities. We then compute these measures for all possible combinations of the structural parameters b (balance) and d (density) on a grid, resulting in high-resolution phase diagrams $C(b, d)$ and $I(b, d)$. This reveals that the regions of phase space in which information storage (memory capacity) and information import (representation) are optimal, surprisingly do not coincide, but nevertheless have a small area of phase space in common. We speculate that this overlap region, where both crucial functions are simultaneously possible, may represent a "sweet spot" for practical RNN applications and might therefore be exploited by biological nervous systems.

RESULTS

Free-Running Network

In the following, we are analyzing networks composed of $N_{neu} = 100$ deterministic neurons with arctangent activation functions. The random matrix of connection weights is set up in a controlled way, so that the density d of non-zero connections as well the balance b between excitatory and inhibitory connections can be pre-defined independently (for details see Section 4). Visualizations of typical weight matrices for different combinations of the statistical control parameters d and b are shown in **Figure 1**.

We first investigate free-running networks without external input and compute a dynamical phase diagram $C_{ss}(b, d)$ of the average correlation $C_{ss} = C(\mathbf{s}_t, \mathbf{s}_{t+1})$ between subsequent system states (**Figure 2a**, for details see Section 4). The resulting landscape is mirror-symmetric with respect to the line $b = 0$, due to the symmetric activation functions of our model neurons, combined with the definition of the balance parameter. Apart from the region of very low connection densities with $d \leq 0.1$, the phase space consists of three major parts: the oscillatory regime in networks with predominantly inhibitory connections ($b \ll 0$, left green area in **Figure 2a**), the chaotic regime with approximately balanced connections ($b \approx 0$, central blue and red area in **Figure 2a**), and the fixed point regime with predominantly excitatory connections ($b \gg 0$, right green area in **Figure 2a**).

It is important to note that $C(\mathbf{s}_t, \mathbf{s}_{t+1})$ is a root-mean-square (RMS) average over all the $N_{neu} \times N_{neu}$ pairwise correlations between subsequent neural activations (so that negative and positive correlations are not distinguished), and that these pairwise correlations are properly normalized in the sense of a Pearson coefficient (each ranging between –1 and +1 before the RMS is computed). For this reason, $C(\mathbf{s}_t, \mathbf{s}_{t+1})$ is close to one (green) both in the oscillatory and in the fixed point regimes, where the system is behaving regularly. By contrast, $C(\mathbf{s}_t, \mathbf{s}_{t+1})$ is close to zero (blue) in the high-density part of the chaotic regime, where the time-evolution of the system is extremely irregular. Medium-level correlations (red) are therefore expected in the transition region between these two extreme dynamical regimes, and they are indeed found in the correlation phase diagram for densities larger than ≈0.3 in the form of narrow stripes at the border of the chaotic "valley." It is however surprising that medium-level correlations also exist across the whole chaotic valley for relatively low densities $d \in [0.1, 0.3]$. Since medium-level correlations are thought to be optimally suited for information processing (Bertschinger and Natschläger, 2004; Natschläger et al., 2005; Legenstein and Maass, 2007; Schrauwen et al., 2009; Büsing et al., 2010; Toyoizumi and Abbott, 2011; Dambre et al., 2012), it is remarkable that this can take place not only at the classical "edge of chaos" (between the

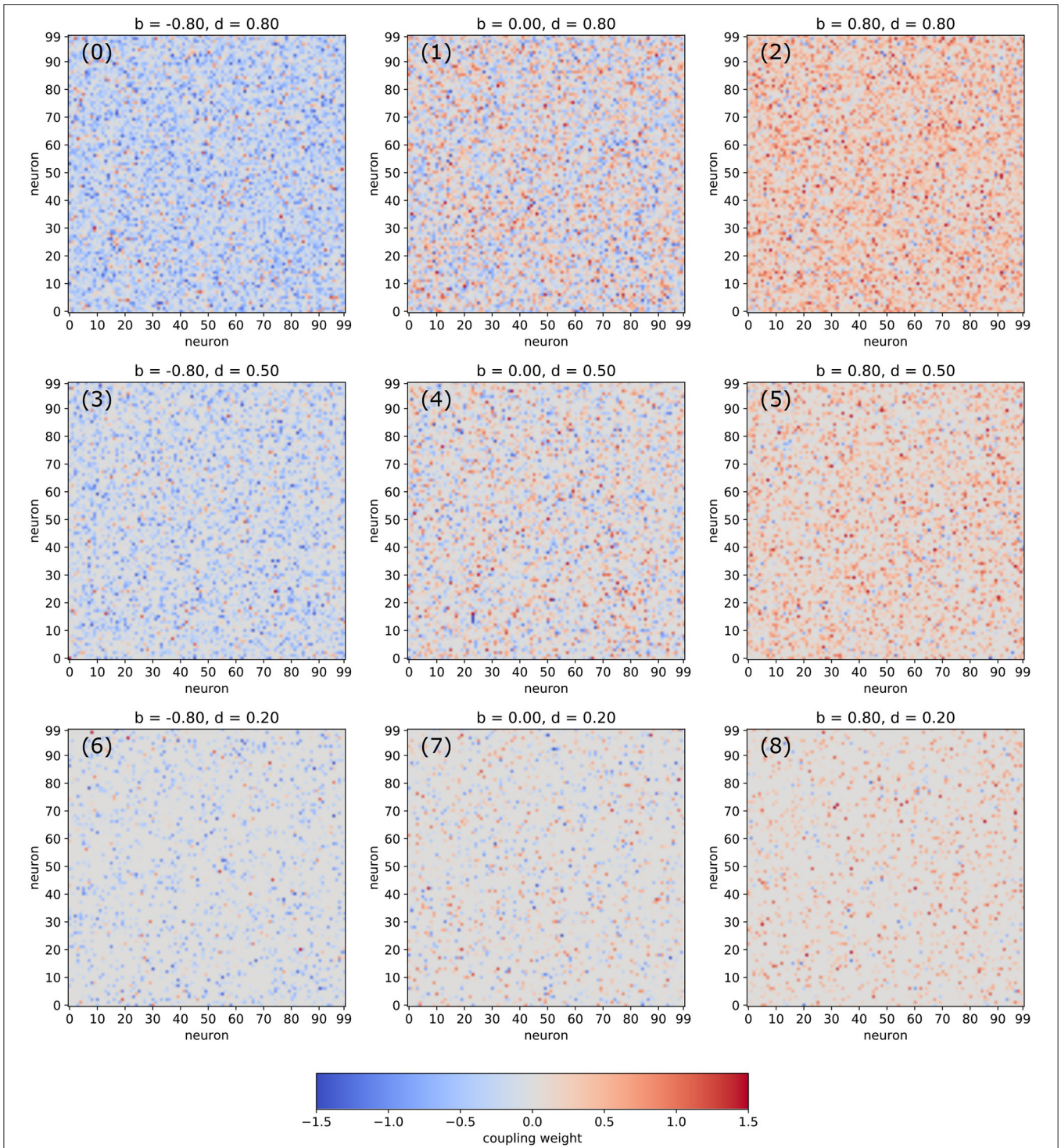


FIGURE 1 | Examples of weight matrices for selected combinations of the balance b between excitatory and inhibitory connections and the density d of non-zero connections in an RNN.

oscillatory and the chaotic regime), but also in other (and less investigated) regions of the network's dynamical phase space.

In order to verify the nature of the three major dynamical regimes, we investigate the time evolution of the neural activations for selected combinations of the control parameters b and d. In particular, we fix the connection density to $d = 0.5$ and gradually increase the balance from $b = -0.5$ to $b = +0.5$ in five steps (**Figures 2b–f**). As expected, we find

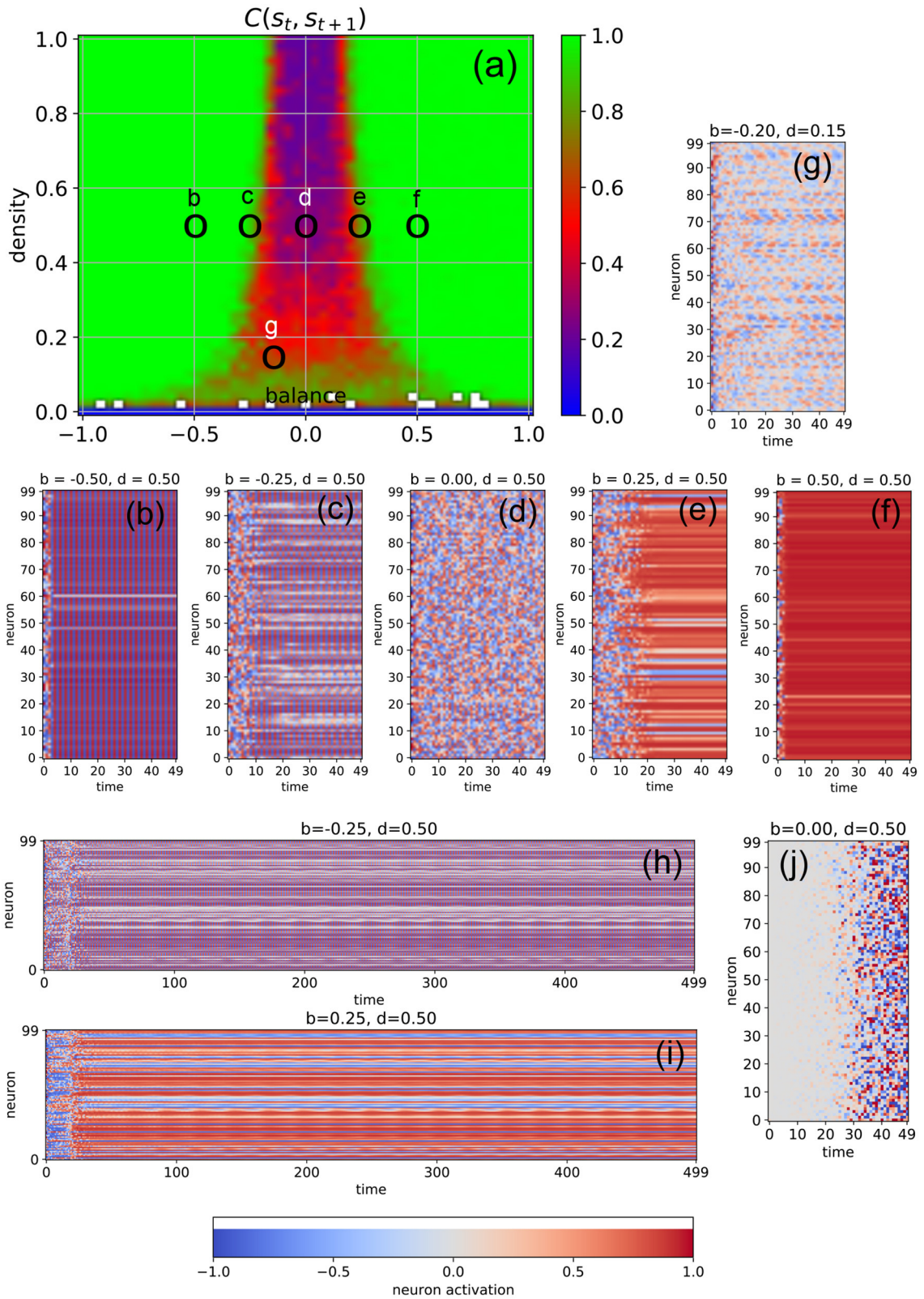


FIGURE 2 | Dynamical phases of a free-running RNN, controlled by the structural parameters *b* (balance) and *d* (density). **(a)** Phase diagram of the correlation $C(\mathbf{s}_t, \mathbf{s}_{t+1})$ between successive neuron activations, as defined in the methods section. The three basic regimes are the oscillatory phase for negative balances (large correlations), the chaotic phase for balances close to zero (small correlations), and the fixed point phase for positive balances (large correlations).

(Continued)

FIGURE 2 | (b–f) Typical time dependence of the neural activations for fixed density ($d = 0.5$) and balances increasing from $b = -0.5$ to $b = +0.5$. The system behavior evolves from almost homogeneous oscillations **(b)**, to a heterogeneous oscillatory state **(c)**, to fully chaotic behavior **(d)**, to a heterogeneous fix point state state with a sub-group of slowly oscillating neurons **(e)**, and finally to an almost global fixed point attractor **(f)**. The low-density example **(g)** shows out-of-phase, imperfect oscillations with a period larger than 2, with phase differences between the neurons. Longer state sequences of the cases **(c,d)** are shown in **(h,i)**. **(j)** Shows the difference of neural activations between the chaotic state sequence **(d)** and a second run, where the initial activation of only one neuron (with index 0) was changed by a value of 0.1.

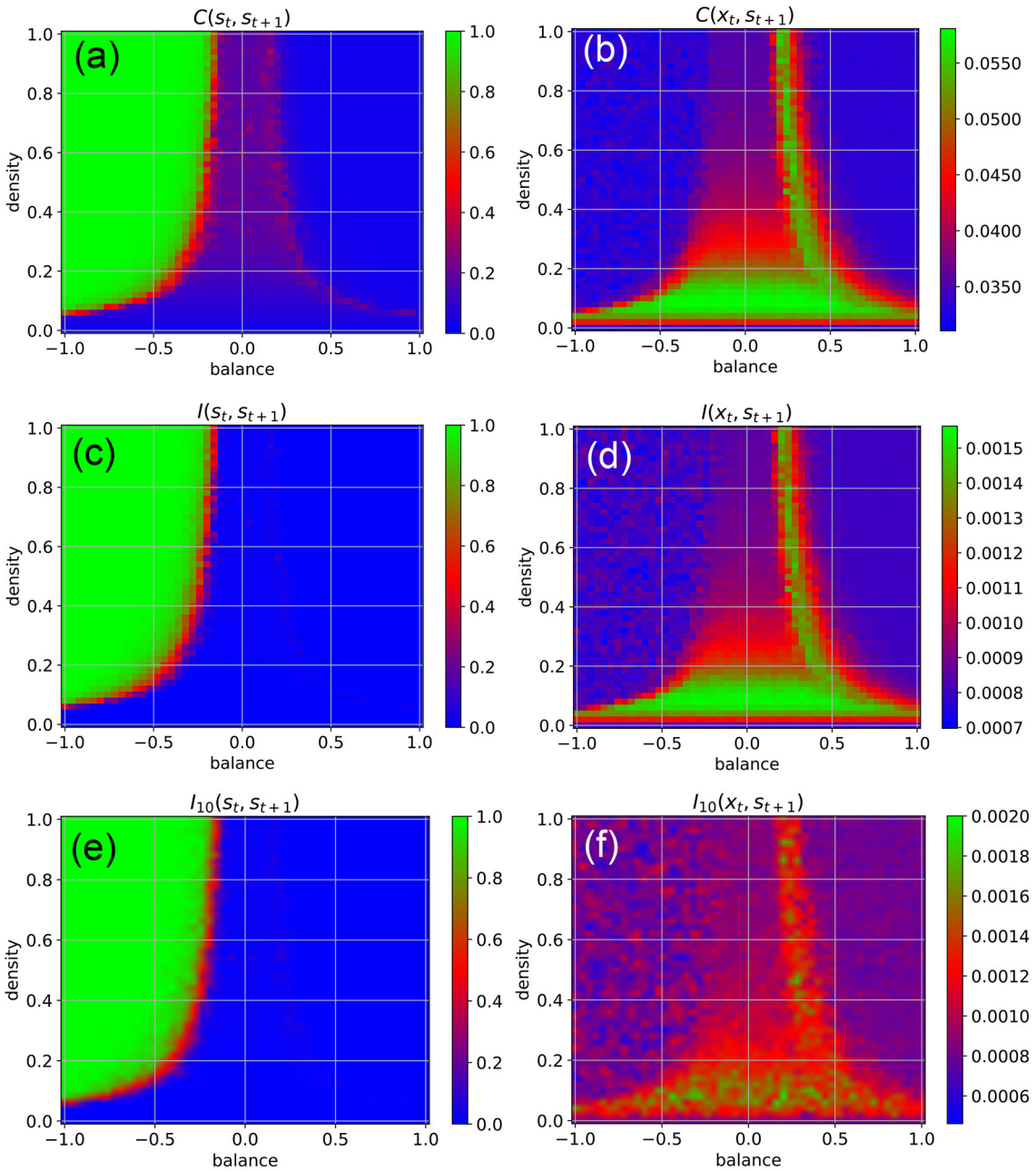


FIGURE 3 | Dynamical phases of a RNN driven by external input in the form of continuous random signals that are coupled independently to all neurons with a coupling constant of $\eta = 0.5$. The suitability of the system for information processing is characterized by the statistical dependency between subsequent states (left column), the suitability for information import by the statistical dependency between the input $\mathbf{x}_t$ and the subsequent state $\mathbf{s}_{t+1}$ (right column). First row **(a,b)** Root-mean-square of correlations. Second row **(c,d)** Mean pairwise mutual information. Information import is optimal in the low-density chaotic regime and at the border between the chaotic and fixed point regime (red and green color in right column). Third row **(e,f)** Approximation of the mean pairwise mutual information, where only a sub-population of 10 neurons is included to the evaluation.

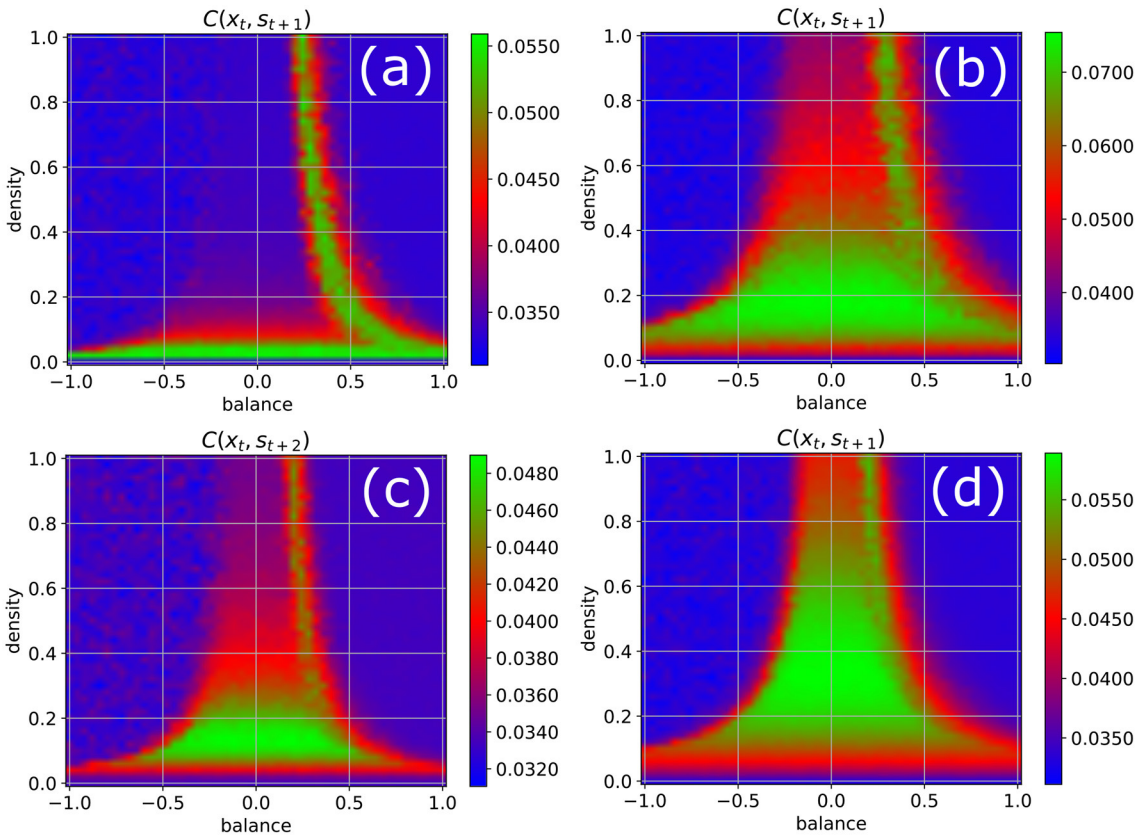


FIGURE 4 | Phase diagram of information import as in **Figure 3b**, but with one parameter changed in each of the four panels. **(a)** Width of the Gaussian distribution of weight magnitudes increased from $w = 0.5$ to $w = 1$. **(b)** Number of neurons reduced from $N = 100$ to $N = 50$. **(c)** Time delay between input signal and system state increased from 1 to 2. **(d)** Width of the Gaussian distribution of weight magnitudes decreased from $w = 0.5$ to $w = 0.25$. The results are similar to **Figure 3b** in all cases except for reduced weight fluctuations **(d)**, where both edges of chaos become available for information import.

almost perfect oscillations (here with a period of two time steps) for $b = -0.5$ (case **Figure 2b**), at least after the transient period in which the system is still carrying a memory of the random initialization of the neural activations. At $b = 0$ (case **Figure 2d**), we find completely irregular, chaotic behavior, and at $b = +0.5$ (case **Figure 2d**) almost all neurons reach the same fixed point. However, the cases close to the two edges of the chaotic regime reveal an interesting intermediate dynamic behavior: For $b = -0.25$ (cases **Figures 2c,h**), most neurons are synchronized in their oscillations, but some are out of phase. Others show a long-period regular "beating"-like behavior superposed on the oscillations of period two (see the longer time trace in **Figure 2h**). For $b = +0.25$ (cases **Figures 2e,i**), most neurons reach (approximately) a shared fixed point, but some end up in a different, individual fixed point, thus resembling a state of quenched disorder. However, a sub-group of neurons is simultaneously engaged in long-period oscillations (see the longer time trace in **Figure 2i**).

The apparent irregularity of the neural activations in case **Figure 2d** does not necessarily imply chaotic behavior. To demonstrate the sensitive dependence of the neural trajectories on the initial condition, we change the activation of only a single neuron at $t = 0$ by a small amount of 0.1 and re-run the simulation. We find that drastic, system-spanning differences appear between the two time evolutions after about 30 time steps (see **Figure 2j**).

Moreover, we observe that the memory time τ of the system for the information imprinted by the initialization (that is, the duration of the transient phase) depends systematically on the balance parameter: Deep within the oscillatory regime ($b = -0.5$, case **Figure 2b**), τ is short. As we approach the chaotic regime ($b = -0.25$, case **Figure 2c**), τ increases, finally becoming "infinitely" long at $b = 0$ (case **Figure 2d**). Indeed, from this viewpoint the chaotic dynamics may be interpreted as the continuation of the transient phase. As we move deeper into the fixed point regime (cases **Figures 2e,f**), the memory time τ is decreasing again.

In the medium and high-density regime of the phase diagram, we find for negative values of the bias parameter mainly oscillations of period two, as the large number of negative weights causes the neurons to switch the sign of their sigmoidal outputs from one time step to the next. However, in the low-density regime, the magnitude of the neuron's total input is reduced and we then find also oscillations with larger periods (case **Figure 2g**).

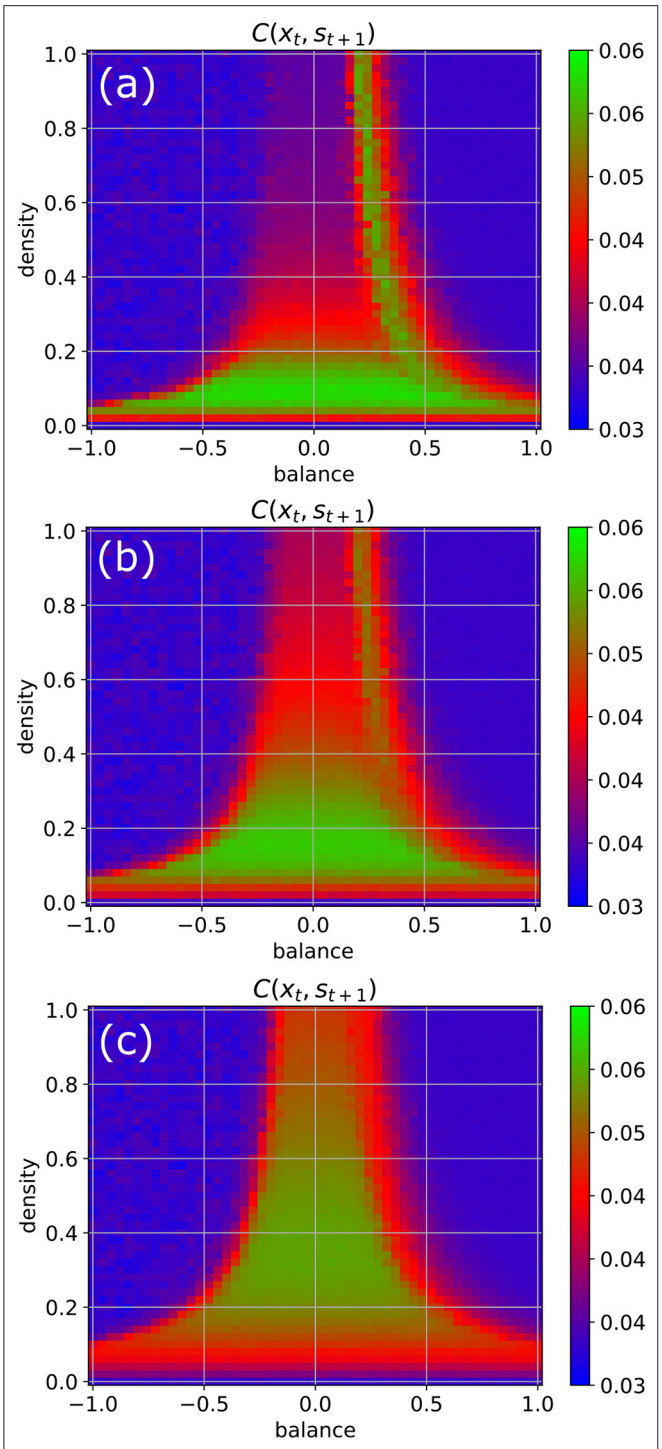


FIGURE 5 | Information import as a function of the coupling strength η between the RNN neurons and the external input signals. For weak coupling [$\eta = 0.5$ in **(a)**], only the low-density chaotic regime and the border between the chaotic and fixed point regime are suitable for information import. As the coupling in increases from $\eta = 1$ in **(b)** to $\eta = 2$ in **(c)**, the correlations between input $\mathbf{x}_t$ and subsequent RNN states $\mathbf{s}_{t+1}$ become gradually large throughout the complete chaotic regime.

Network Driven by Continuous Random Input

Next, we feed into the network a relatively weak external input (with a coupling strength of $\eta = 0.5$), consisting of independent normally distributed random signals that are continuously injected to each of the neurons (for details see Section 4).

We find that the external input destroys the medium-level state-to-state correlations $C(\mathbf{s}_t, \mathbf{s}_{t+1})$ in most parts of the chaotic regime, except at the classical edge of chaos (**Figure 3a**, red). Moreover, the input also brings the state-to-state correlations in the fixed point regime down to a very small value, as now the external random signals are superimposed onto the fixed points of the neurons.

Another important practical factor is the ability of neural networks to store information, i.e., to take up external information at any point in time and to incorporate it into their system state. We quantify this ability of information import by the RMS-averaged correlation $C(\mathbf{x}_t, \mathbf{s}_{t+1})$ between momentary input and subsequent system state. Surprisingly, we find that information import is best, i.e., $C(\mathbf{x}_t, \mathbf{s}_{t+1})$ is large, in the low-density part of the chaotic regime, including the lowest part of the classical edge of chaos (region between chaotic and oscillatory regimes), but also at the opposite border between the chaotic and fixed point regimes (**Figure 3b**, green and red). We thus come to the conclusion that (at least for weak external inputs with $\eta = 0.5$) our network model is simultaneously capable of information import and information processing only in the low-density part of the classical edge of chaos.

To backup this unexpected finding, we also quantify information storage and information import by the average pairwise state-to-state mutual information $I(\mathbf{s}_t, \mathbf{s}_{t+1})$ (**Figure 3c**), and the mutual information between the momentary input and the subsequent system state $I(\mathbf{x}_t, \mathbf{s}_{t+1})$ (**Figure 3d**), respectively. These mutual-information-based measures can also capture possible non-linear dependencies, but are computationally much more demanding (for details see Section 4).

Despite of these drastic differences between the two measures, we obtain practically the same phase diagrams for information import and information storage/processing when we use the RMS-averaged pairwise correlations (**Figures 3a,b**) and when we use the mutual information (**Figures 3c,d**). This congruence may simply indicate the absence of higher-order statistical dependencies between subsequent states in our specific RNN system. However, in the context of adaptive stochastic resonance, we have already reported a surprisingly close relation between linear correlation and mutual information for a large range of model systems (Krauss et al., 2017). Taken together, these findings suggest a possible link between information-processing and dynamical approaches to complexity science (Mediano et al., 2021).

Furthermore, we compare the results to a computationally more tractable approximation of the mean pairwise mutual information, where only a sub-population of 10 neurons is included to the evaluation. It also shows the same basic

characteristics (**Figures 3e,f**), implicating the possibility to approximate mutual information in large dynamical systems, where an exhaustive sampling of all joint probabilities necessary to calculate entropy and mutual information is impractical or impossible.

Effect of Other System Parameters

In order to test the robustness of the above results on information import, we re-compute the phase diagram of the correlations between the input and a later system state (**Figure 4**), now however varying some of the parameters that have been kept at their standard values ($w = 0.5$, $N = 100$, $\Delta t = 1$, $\eta = 0.5$) so far. We obtain results similar to **Figure 3b** when the fluctuation width w of the Gaussian weight distribution is increased to $w = 1$ (**Figure 4a**), when the number of neurons in reduced to $N = 50$ (**Figure 4b**), and when the lag-time between input signal and system state is increased to $\Delta t = 2$ (**Figure 4c**). However, when the fluctuation width of the weight distribution is reduced to $w = 0.5$, which decreases the total neural inputs and therefore brings the system closer to the linear regime, we find that now both edges of chaos become available for information uptake (**Figure 4d**).

Effect of Increasing Coupling Strength

We return to our standard parameters ($w = 0.5$, $N = 100$, $\Delta t = 1$), but now increase the coupling strength to the random input signals step-wise from $\eta = 0.5$ to $\eta = 1$ and finally to $\eta = 2$ (**Figure 5**). We observe that by this way also the higher density parts of the chaotic regime become eventually available for information import (green color).

Import Resonance (IR) and Recurrence Resonance (RR)

Next, we increase the coupling strength η gradually from zero to a very large value of 20, at which the random input already dominates the system dynamics. For this numerical experiment, we keep the balance and density parameters fixed at $b = -0.5$, $d = 0.5$ (oscillatory regime), $b = 0$, $d = 0.5$ (chaotic regime), and $b = 0.5$, $d = 0.5$ (fixed point regime), respectively.

When in the fixed point regime (**Figure 6f**), we find that the dependence of the state-to-state correlation $C(\mathbf{s}_t, \mathbf{s}_{t+1})$ on the coupling strength η has the shape of a "resonance peak." Since η effectively controls the amplitude of "noise" (used by us as pseudo input) added to the system, this corresponds to the phenomenon of "Recurrence Resonance" (RR), which we have previously found in three-neuron motifs (Krauss et al., 2019a): At small noise levels η, the system is stuck in the fixed point attractor, but adding an optimal amount of noise (so that $C(\mathbf{s}_t, \mathbf{s}_{t+1})$ becomes maximal) is freeing the system from this attractor and thus makes recurrent information "flux" possible, even in the fixed point regime. Adding too much noise is however counter-productive and leads to a decrease of $C(\mathbf{s}_t, \mathbf{s}_{t+1})$, as the system dynamics then becomes dominated by noise. We do not observe recurrence resonance in the other two dynamic regimes (**Figures 6b,d**).

Interestingly, we find very pronounced resonance-like curves also in the dependence of the input-to-state correlation $C(\mathbf{x}_t, \mathbf{s}_{t+1})$ on the coupling strength η, for all dynamical regimes (**Figures 6a,c,e**). Since $C(\mathbf{x}_t, \mathbf{s}_{t+1})$ is a measure of information import, we call this novel phenomenon "Import Resonance" (IR).

Network Driven by Continuous Sinusoidal Input

Next, we investigate the ability of the system to import more regular input signals with built-in temporal correlations, as well as inputs that are identical for all neurons. For this purpose, we feed all neurons with the same sinusoidal input signal, using an amplitude of $a_{sin} = 1$, an oscillation period of $T_{sin} = 25$ time steps, and a coupling strength of $\eta = 2$ (**Figure 7**). The density parameter is again fixed at $d = 0.5$, while the balance increases from $b = -0.6$ to $b = +0.6$ in five steps. We find that the input signal does not affect the evolution of neural states when the system is too far in the oscillatory phase or too far in the fixed point phase (c,g). Only systems where excitatory and inhibitory connections are approximately balanced are capable of information import (d-f). For $b = -0.3$ (d), most of the neurons are still part of the periodic attractor, but a small sub-population of neurons is taking up the external input signal (d). Interestingly, the system state is reflecting the periodic input signal even in the middle of the chaotic phase (e).

Correlations for Longer Lagtimes

So far, we have analyzed input-to-state and state-to-state correlations mainly for a lag-time $\Delta t = 1$. We finally extend this analysis to larger lag-times up to 50 time steps (**Figure 8**), however only for three selected RNNs in the oscillatory, chaotic and fixpoint regime, using again our standard parameters ($w = 0.5$, $N = 100$, $\eta = 0.5$). Since our correlation measures $C(\mathbf{x}_t, \mathbf{s}_{t+1})$ (left column) and $C(\mathbf{s}_t, \mathbf{s}_{t+1})$ (right column) are defined as RMS averages, these values never fall below a certain noise level, which is in our case about 0.034. Another consequence of the RMS-average is that perfectly oscillatory RNN states with a period of two show up as $C(\mathbf{s}_t, \mathbf{s}_{t+1}) = 1$ (**Figure 8b**).

In the oscillatory regime, we find that input-to state correlations (as a measure of information import) remain at the noise level for all lag-times (a), while the system states are bound in a perfectly periodic attractor (b). Also in the fixpoint regime, both types of correlation are negligible for all non-zero lag-times. But remarkably, information can be imported (c) and stored (d) to a small but significant extent even in the middle of the chaotic regime, although the correlations decay back to noise level after about 20 time steps for this specific point in phase space ($b = 0$, $d = 0.5$). Future work will analyze how this correlation decay time depends on the statistical system parameters b and d.

DISCUSSION

In this study, we investigate the ability of RNNs to import and store information as a function of the weight statistics, a

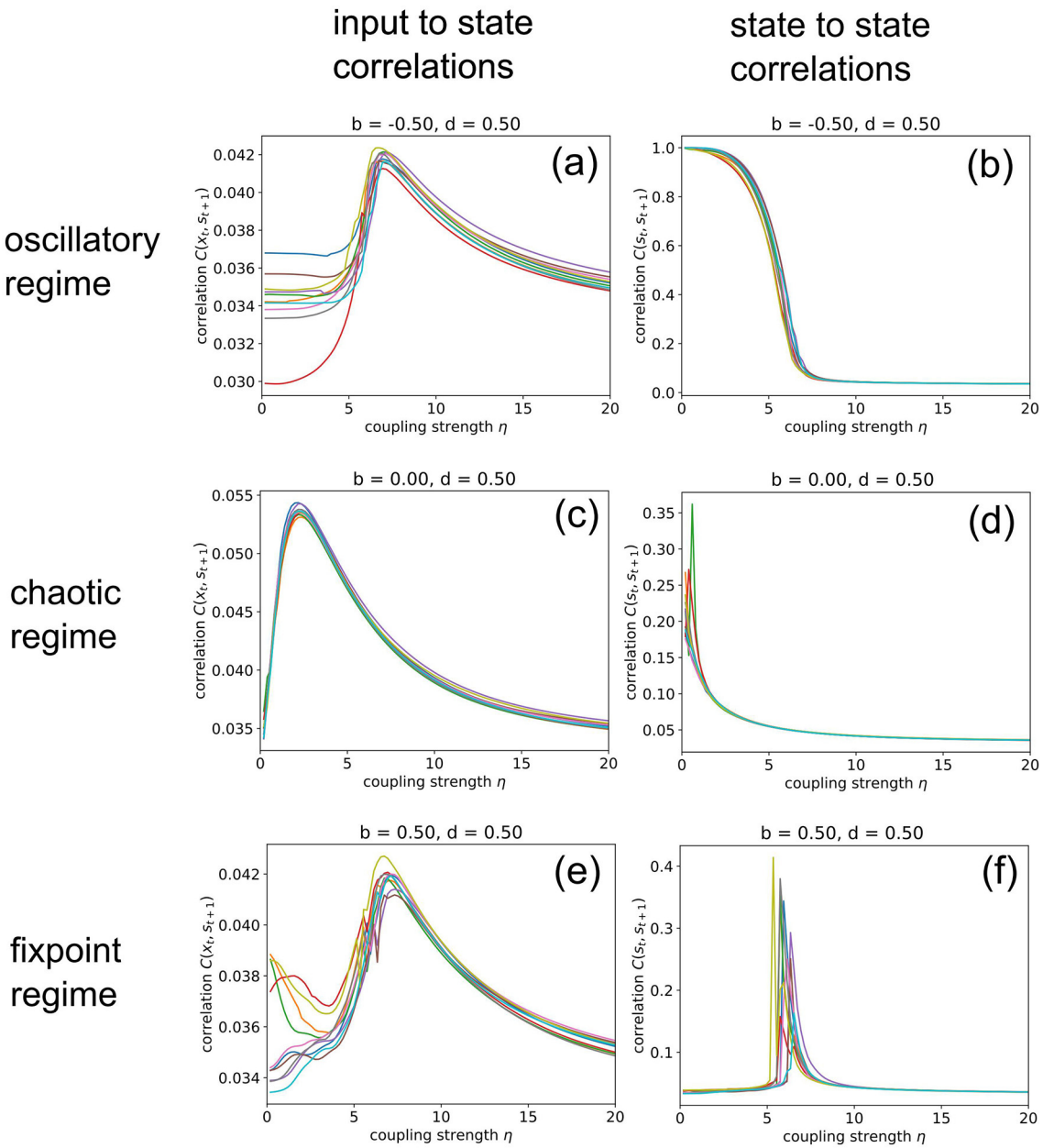


FIGURE 6 | Import resonance and recurrence resonance in RNNs. We compute the input-to-state correlation $C(\mathbf{x}_t, \mathbf{s}_{t+1})$ (left column) and the state-to-state correlation $C(\mathbf{s}_t, \mathbf{s}_{t+1})$ (right column) for RNNs in the oscillatory (top row), chaotic (middle row) and fixed point regimes (bottom row), as the coupling strength to the random (noise) input $\mathbf{x}_t$ is gradually increased from zero to 20. The computation has been repeated for 10 different realizations (colors) of RNNs with the given control parameters b (balance) and d (density). We find the phenomenon of import resonance in all three dynamical regimes **(a,c,e)** and the phenomenon of recurrence resonance in the fixed point regime **(f)**. No resonance is found in cases **(b,d)**.

problem that has been met with considerable interest during the past years (Bässler, 1986; Derrida et al., 1987; Gutfreund et al., 1988; Langton, 1990; Wang et al., 1990, 2011; Molgedey et al., 1992; Crisanti et al., 1993; Kaneko and Suzuki, 1994; Solé and Miramontes, 1995; Greenfield and Lecar, 2001; Jaeger, 2001; Bertschinger and Natschläger, 2004; Rajan et al., 2010; Toyoizumi and Abbott, 2011; Boedecker et al., 2012; Wallace et al., 2013; Kadmon and Sompolinsky, 2015; Brunel, 2016; Folli et al., 2018; Schuecker et al., 2018; Grigoryeva and Ortega, 2019; Grigoryeva et al., 2021). We specialize on discrete-time, deterministic RNNs with an arctan activation function and describe the weight statistics by the *density of non-zero weights* and on the *balance of excitatory and inhibitory connections*, as introduced in our previous studies (Krauss et al., 2019b,c). In contrast to the human brain, where the vast majority of neurons is either purely excitatory or purely inhibitory (Dale's principle),

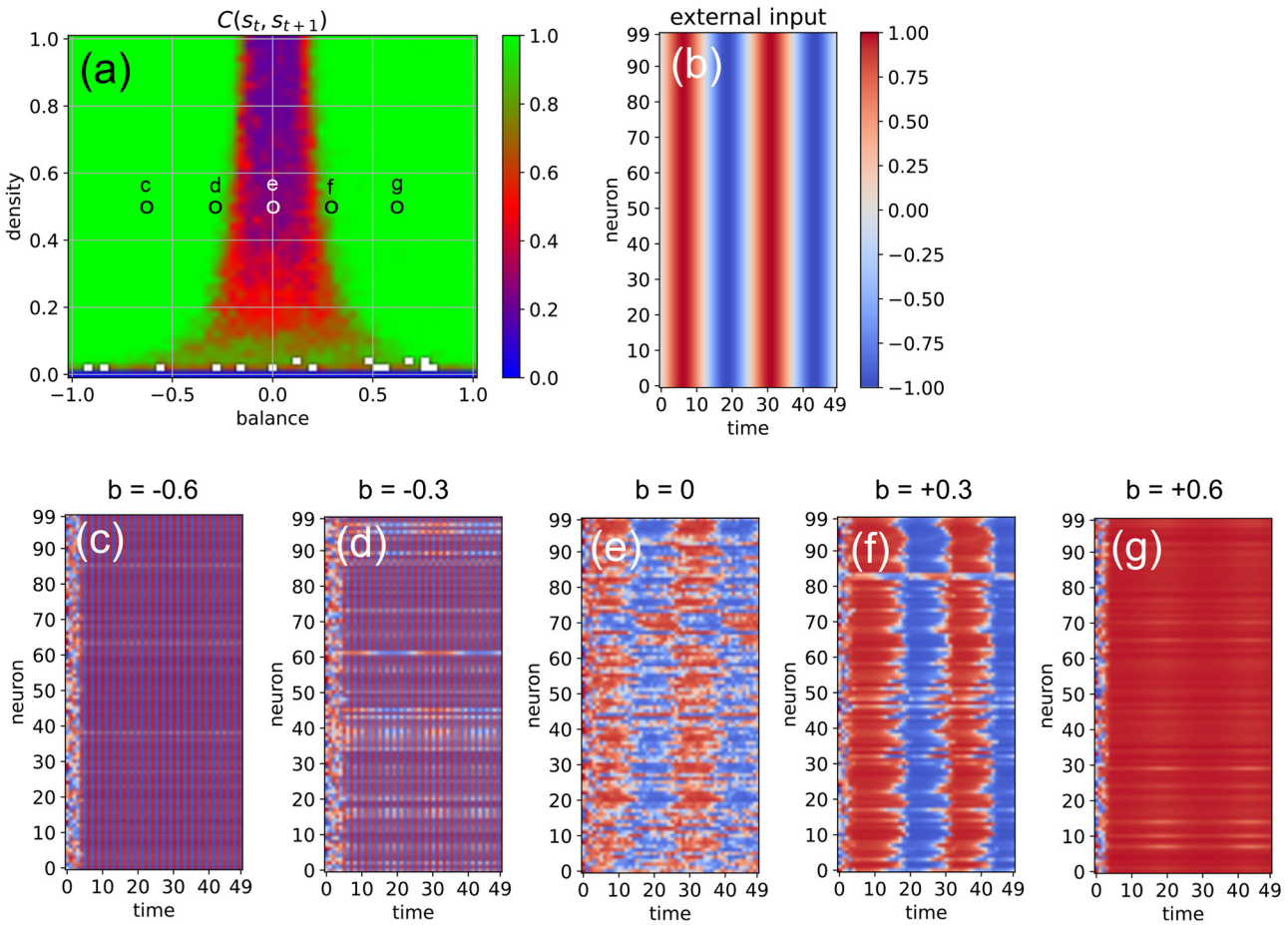


FIGURE 7 | Effect of a "sinusoidal" input **(b)** on the activations of the RNN neurons **(c–g)** at five different points in the system's dynamic phase space **(a)**. For all cases **(c–g)**, the density parameter is $d = 0.5$, while the balance increases from –0.6 to +0.6. Only for balances sufficiently close to zero **(d,e,f)** the input is able to affect the system state.

each given neuron can simultaneously have positive and negative output weights in our simplified model system.

It turned out that our RNN model is simultaneously capable of both information import and information storage only in the low-density, i.e., sparse, part of the classical edge of chaos. Remarkably, this region of the phase space corresponds to the connectivity statistics known from the brain, in particular the cerebral cortex (Song et al., 2005; Sporns, 2011; Miner and Triesch, 2016). In line with previous findings, i.e., that sparsity prevents RNNs from overfitting (Narang et al., 2017; Gerum et al., 2020) and is optimal for information storage (Brunel, 2016), we therefore hypothesize that cortical connectivity is optimized for both information import and processing. In addition, it seems plausible that there might be distinct networks in the brain that are either specialized to import and to represent information, or to process information and perform computations.

Furthermore, we found a completely new resonance phenomenon which we call *import resonance*, showing that the correlation or mutual information between input and the subsequent network state depends on certain control parameters (such as coupling strength) in a peak-like way. Resonance phenomena are ubiquitous not only in simplified neural network models (Ikemoto et al., 2018; Krauss et al., 2019a; Bönsel et al., 2021), but also in biologically more realistic systems (McDonnell and Abbott, 2009), where they show up in diverse variants such as coherence resonance (Lindner and Schimansky-Geier, 2000; Gu et al., 2002; Lindner et al., 2002), finite size resonance (Toral et al., 2003), bimodal resonance (Mejias and Torres, 2011; Torres et al., 2011), heterogeneity-induced resonance (Mejias and Longtin, 2012, 2014), or inverted stochastic resonance (Buchin et al., 2016; Uzuntarla et al., 2017). They have been shown to play a crucial role for neural information processing (Moss et al., 2004; Krauss et al., 2018; Schilling et al., 2020). In particular with respect to the auditory system, it has been argued that resonance phenomena like stochastic resonance are actively exploited by the brain to maintain optimal information processing (Krauss et al., 2016, 2017, 2018; Schilling et al., 2021b). For instance, in a theoretical study it could be demonstrated that stochastic resonance improves speech recognition in an artificial neural network as a model of the auditory pathway (Schilling

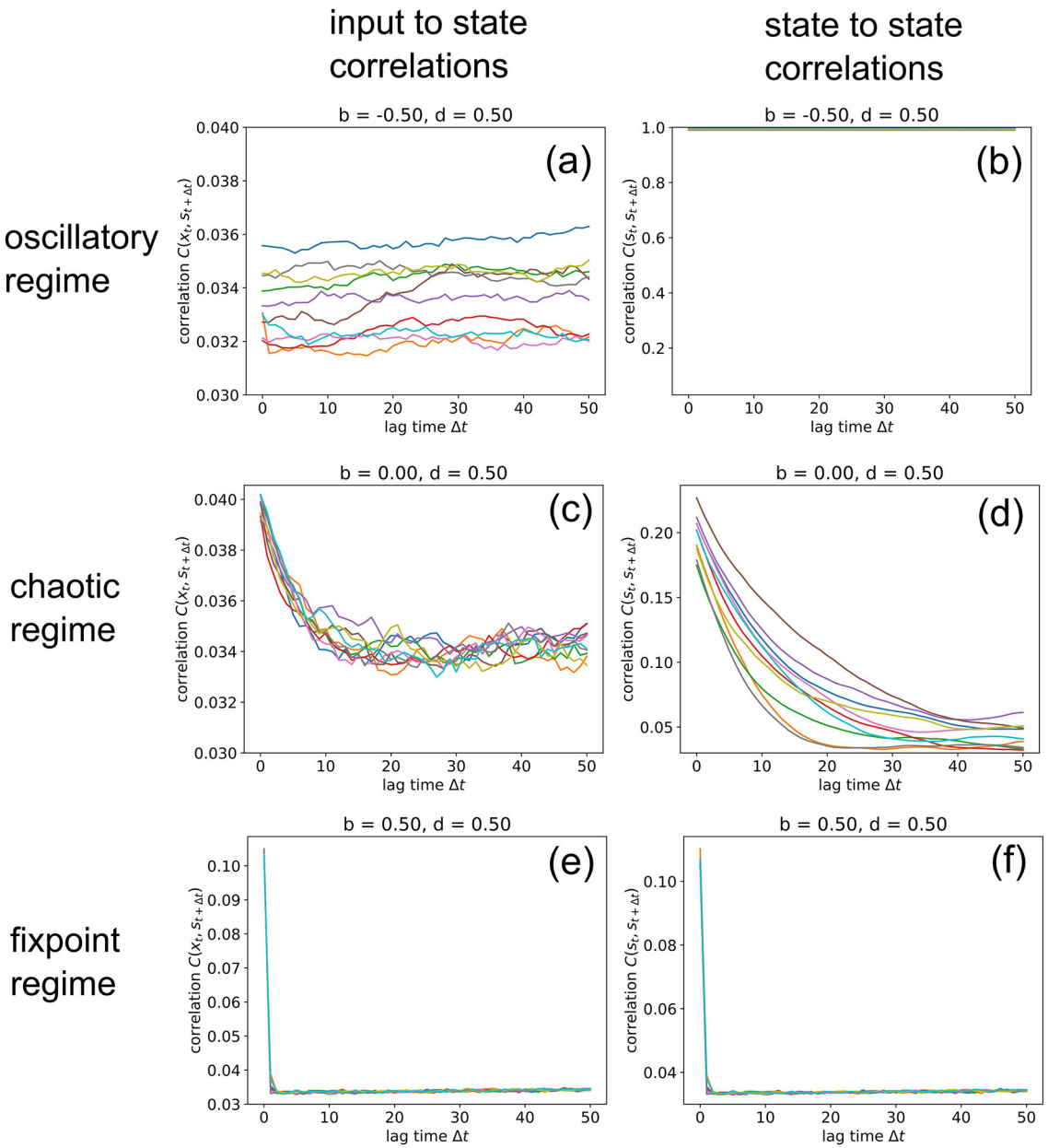


FIGURE 8 | Information import and storage for longer lagtimes. We compute the input-to-state correlation $C(\mathbf{x}_t, \mathbf{s}_{t+1})$ (left column) and the state-to-state correlation $C(\mathbf{s}_t, \mathbf{s}_{t+1})$ (right column) for RNNs in the oscillatory (top row), chaotic (middle row) and fixed point regimes (bottom row), for increasing lagtimes between zero and 50. The computation has been repeated for 10 different realizations (colors) of RNNs with the given control parameters b (balance) and d (density). Note that correlations C never become lower than a noise level of about 0.034, because we compute C as an RMS average. Due to this RMS, the signature of an oscillatory state is $C(\mathbf{s}_t, \mathbf{s}_{t+1}) = 1$, as found in **(b)**. Import and storage of information, above the noise level (and at non-zero lagtimes), is observed only in the cases **(c,d)**, even though the RNN is deeply in the chaotic regime at $b = 0, d = 0.5$. In the oscillatory and fixpoint regimes **(a,b,e,f)**, this is not possible.

et al., 2020). Very recently, we were even able to show that stochastic resonance, induced by simulated transient hearing loss, improves auditory sensitivity beyond the absolute threshold of hearing (Krauss and Tziridis, 2021). The extraordinary importance of resonance phenomena for neural information processing indicates that the brain, or at least certain parts of the brain, do also actively exploit other kinds of resonance phenomena besides classical stochastic resonance. Whereas, stochastic resonance is suited to enhance the detection of weak signals from the environment in sensory brain systems (Krauss et al., 2017), we speculate that parts of the brain dealing with sensory integration and perception might exploit import resonance, while structures dedicated to transient information storage (short-term memory) (Ichikawa and Kaneko, 2020) and

processing might benefit from recurrence resonance (Krauss et al., 2019a). Similarly, the brain's action and motor control systems might also benefit from a hypothetical phenomenon of *export resonance*, i.e., the maximization of correlation or mutual information between a given network state and a certain, subsequent readout layer.

Finally, our finding that both, correlation- and entropy-based measures of information import and storage yield almost identical phase diagrams (**Figures 3a,b** compare with **Figures 3c,d**), is in line with previously published results, i.e., that mutual information between sensor input and output can be replaced by the auto-correlation of the sensor output in the context of stochastic resonance (SR) (Krauss et al., 2017). However, in this study we find that the equivalence of measures based on correlations and mutual information even extends to the phenomena of recurrence resonance (RR) (Krauss et al., 2019a) and import resonance (IR), thereby bridging the conceptual gap (as described in Mediano et al., 2021) between the information-processing perspective and the dynamical systems perspective on complex systems.

METHODS

Weight Matrices With Pre-defined Statistics

We consider a system of N_{neu} neurons without biases, which are mutually connected according to a weight matrix $\{w_{mk}\}$, where w_{mk} denotes the connection strength from neuron k to neuron m. The weight matrix is random, but controlled by three statistical parameters, namely the *"density"* d of non-zero connections, the excitatory/inhibitory *"balance"* b, and the *"width"* w of the Gaussian distribution of weight magnitudes. The density ranges from $d = 0$ (isolated neurons) to $d = 1$ (fully connected network), and the balance from $b = -1$ (purely inhibitory connections) to $b = +1$ (purely excitatory connections). The value of $b = 0$ corresponds to a perfectly balanced system.

In order to construct a weight matrix with given parameters (b, d, w), we first generate a matrix $M_{mn}^{(magn)}$ of weight magnitudes, by drawing the N_{neu}^2 matrix elements independently from a zero-mean normal distribution with standard deviation w and then taking the absolute value. Next, we generate a random binary matrix $B_{mn}^{(nonz)} \in \{0, 1\}$, where the probability of a matrix element being 1 is given by the density d, i.e., $p_1 = d$. Next, we generate another random binary matrix $B_{mn}^{(sign)} \in \{-1, +1\}$, where the probability of a matrix element being $+1$ is given by $p_{+1} = (1 + b)/2$ where b is the balance. Finally, the weight matrix is constructed by elementwise multiplication $w_{mn} = M_{mn}^{(magn)} \cdot B_{mn}^{(nonz)} \cdot B_{mn}^{(sign)}$. Note that throughout this paper, the width parameter is set to $w = 0.5$.

Time Evolution of System State

The momentary state of the RNN is given by the vector $\mathbf{s}(t) = \{s_m(t)\}$, where the component $s_m(t) \in [-1, +1]$ is the activation of neuron m at time t. The initial state $\mathbf{s}(t=0)$ is set by assigning to the neurons statistically independent, normally distributed random numbers with zero mean and a standard deviation of $\sigma_{ini} = 1$.

We then compute the next state vector by simultaneously updating each neuron m according to

$$s_m(t+1) = \frac{2}{\pi} \arctan\left(\eta\, x_m(t) + \sum_{k=1}^{N} w_{mk}\, s_k(t)\right). \quad (1)$$

Here, $x_m(t)$ are the external inputs of the RNN and η is a global *"coupling strength"*. Note that the input time series $x_m(t)$ can, but must not be different for each neuron. In one type of experiment, we set the $x_m(t)$ to independent, normally distributed random signals with zero mean and unit variance. In another experiment, we set all $x_m(t)$ to the same oscillatory signal $x(t) = a_{sin} \cdot \sin(2\pi t/T_{sin})$.

After simulating the sequence of system states for $N_{stp} = 1000$ time steps, we analyze the properties of the state sequence (see below). For this evaluation, we disregard the first $N_{tra} = 100$ time steps, in which the system may still be in a transitory state that depends strongly on the initial condition. The simulations are repeated $N_{run} = 10$ times for each set of control parameters (b, d, η).

Root-Mean-Squared Pairwise Correlation $C(\mathbf{u}_t, \mathbf{v}_{t+1})$

Consider a vector $\mathbf{u}(t)$ in M dimensions and a vector $\mathbf{v}(t)$ in N dimensions, both defined at discrete time steps t. The components of the vectors are denoted as $u_m(t)$ and $v_n(t)$. In order to characterize the correlations between the two time-dependent vectors by a single scalar quantity $C(\mathbf{u}_t, \mathbf{v}_{t+1})$, we proceed as follows:

First, we compute for each vector component m the temporal mean,

$$\mu_{um} = \langle u_m(t) \rangle_t \quad (2)$$

and the corresponding standard deviation

$$\sigma_{um} = \sqrt{\left\langle \left(u_m(t) - \mu_{um}\right)^2 \right\rangle_t}. \quad (3)$$

Based on this, we compute the $M \times N$ pairwise (Pearson) correlation matrix,

$$C_{mn}^{(uv)} = \frac{\left\langle \left[u_m(t) - \mu_{um}\right] \cdot \left[v_n(t+1) - \mu_{vn}\right] \right\rangle_t}{\sigma_{um}\, \sigma_{vn}}, \quad (4)$$

defining $C_{mn}^{(uv)} = 0$ whenever $\sigma_{um} = 0$ or $\sigma_{vn} = 0$.

Finally we compute the root-mean-squared average of this matrix,

$$C(\mathbf{u}_t, \mathbf{v}_{t+1}) = \mathrm{RMS}\left\{C_{mn}^{(uv)}\right\}_{mn} = \sqrt{\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\left|C_{mn}^{(uv)}\right|^2} \quad (5)$$

This measure is applied in the present paper to quantify the correlations $C(\mathbf{s}_t, \mathbf{s}_{t+1})$ between subsequent RNN states, as well as the correlations $C(\mathbf{x}_t, \mathbf{s}_{t+1})$ between the momentary input and the subsequent RNN state.

Mean Pairwise Mutual Information $I(\mathbf{u}_t, \mathbf{v}_{t+1})$

In addition to the linear correlations, we consider the mutual information between the two vectors $\mathbf{u}(t)$ and $\mathbf{v}(t)$, in order to capture also possible non-linear dependencies. However, since the full computation of this quantity is computationally extremely demanding, we binarize the continuous vector components and then consider only the pairwise mutual information between these binarized components.

For the binarization, we first subtract the mean values from each of the components,

$$u_m(t) \longrightarrow \Delta u_m(t) = u_m(t) - \mu_{um}. \tag{6}$$

We then map the continuous signals $\Delta u_m(t) \in [-\infty, +\infty]$ onto two-valued bits $\hat{u}_m(t) \in \{0, 1\}$ by defining $\hat{u}_m(t) = 0$ if $\Delta u_m(t) < 0$ and $\hat{u}_m(t) = 1$ if $\Delta u_m(t) > 0$. In the case of a tie, $\Delta u_m(t) = 0$, we set $\hat{u}_m(t) = 0$ with a probability of $1/2$.

We next compute the pairwise joint probabilities $P(\hat{u}_m, \hat{v}_n)$ by counting how often each of the four possible bit combinations occurs during all available time steps. From that we also obtain the marginal probabilities $P(\hat{u}_m)$ and $P(\hat{v}_n)$.

The matrix of pairwise mutual information is then defined as

$$I_{mn}^{(uv)} = \sum_{\hat{u}_m=0,1} \sum_{\hat{v}_n=0,1} P(\hat{u}_m, \hat{v}_n) \log\left[\frac{P(\hat{u}_m, \hat{v}_n)}{P(\hat{u}_m) \cdot P(\hat{v}_n)}\right], \tag{7}$$

defining all terms as zero where $P(\hat{u}_m) = 0$ or $P(\hat{v}_n) = 0$.

Finally we compute the mean over all matrix elements (each ranging between 0 and 1 bit),

$$I(\mathbf{u}_t, \mathbf{v}_{t+1}) = \text{MEAN}\left\{I_{mn}^{(uv)}\right\}_{mn} = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} I_{mn}^{(uv)} \tag{8}$$

This measure is applied in the present paper to quantify the mutual information $I(\mathbf{s}_t, \mathbf{s}_{t+1})$ between subsequent RNN states, as well as the mutual information $I(\mathbf{x}_t, \mathbf{s}_{t+1})$ between the momentary input and the subsequent RNN state.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): grant KR 5148/2-1 (project number 436456810) to PK.

REFERENCES

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., et al. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature* 557, 429–433. doi: 10.1038/s41586-018-0102-6

Barak, O. (2017). Recurrent neural networks as versatile tools of neuroscience research. *Curr. Opin. Neurobiol.* 46, 1–6. doi: 10.1016/j.conb.2017.06.003

Bässler, U. (1986). On the definition of central pattern generator and its sensory control. *Biol. Cybern.* 54, 65–69. doi: 10.1007/BF00337116

Bertschinger, N., and Natschläger, T. (2004). Real-time computation at the edge of chaos in recurrent neural networks. *Neural Comput.* 16, 1413–1436. doi: 10.1162/089976604323057443

Binzegger, T., Douglas, R. J., and Martin, K. A. (2004). A quantitative map of the circuit of cat primary visual cortex. *J. Neurosci.* 24, 8441–8453. doi: 10.1523/JNEUROSCI.1400-04.2004

Boedecker, J., Obst, O., Lizier, J. T., Mayer, N. M., and Asada, M. (2012). Information processing in echo state networks at the edge of chaos. *Theory Biosci.* 131, 205–213. doi: 10.1007/s12064-011-0146-8

Bönsel, F., Krauss, P., Metzner, C., and Yamakou, M. E. (2021). Control of noise-induced coherent oscillations in time-delayed neural motifs. *arXiv preprint* arXiv:2106.11361. doi: 10.1007/s11571-021-09770-2

Brunel, N. (2016). Is cortical connectivity optimized for storing information? *Nat. Neurosci.* 19, 749–755. doi: 10.1038/nn.4286

Buchin, A., Rieubland, S., Häusser, M., Gutkin, B. S., and Roth, A. (2016). Inverse stochastic resonance in cerebellar purkinje cells. *PLoS Comput. Biol.* 12, e1005000. doi: 10.1371/journal.pcbi.1005000

Büsing, L., Schrauwen, B., and Legenstein, R. (2010). Connectivity, dynamics, and memory in reservoir computing with binary and analog neurons. *Neural Comput.* 22, 1272–1311. doi: 10.1162/neco.2009.01-09-947

Crisanti, A., Falcioni, M., and Vulpiani, A. (1993). Transition from regular to complex behaviour in a discrete deterministic asymmetric neural network model. *J. Phys. A Math. Gen.* 26, 3441. doi: 10.1088/0305-4470/26/14/011

Cruse, H. (2006). Neural networks as cybernetic systems. *Brains Minds* 2, 114.

Cybenko, G. (1992). Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.* 5, 455–455. doi: 10.1007/BF02134016

Dambre, J., Verstraeten, D., Schrauwen, B., and Massar, S. (2012). Information processing capacity of dynamical systems. *Sci. Rep.* 2, 1–7. doi: 10.1038/srep00514

Derrida, B., Gardner, E., and Zippelius, A. (1987). An exactly solvable asymmetric neural network model. *EPL* 4, 167. doi: 10.1209/0295-5075/4/2/007

Farrell, M., Recanatesi, S., Moore, T., Lajoie, G., and Shea-Brown, E. (2019). Recurrent neural networks learn robust representations by dynamically balancing compression and expansion. *bioRxiv* 564476.

Folli, V., Gosti, G., Leonetti, M., and Ruocco, G. (2018). Effect of dilution in asymmetric recurrent neural networks. *Neural Netw.* 104, 50–59. doi: 10.1016/j.neunet.2018.04.003

Funahashi, K.-I. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Netw.* 2, 183–192. doi: 10.1016/0893-6080(89)90003-8

Gerum, R. C., Erpenbeck, A., Krauss, P., and Schilling, A. (2020). Sparsity through evolutionary pruning prevents neuronal networks from overfitting. *Neural Netw.* 128, 305–312. doi: 10.1016/j.neunet.2020.05.007

Greenfield, E., and Lecar, H. (2001). Mutual information in a dilute, asymmetric neural network model. *Phys. Rev. E* 63, 041905. doi: 10.1103/PhysRevE.63.041905

Grigoryeva, L., Hart, A., and Ortega, J.-P. (2021). Chaos on compact manifolds: Differentiable synchronizations beyond the takens theorem. *Phys. Rev. E* 103, 062204. doi: 10.1103/PhysRevE.103.062204

Grigoryeva, L., and Ortega, J.-P. (2019). Differentiable reservoir computing. *J. Mach. Learn. Res.* 20, 1–62.

Gros, C. (2009). Cognitive computation with autonomously active neural networks: an emerging field. *Cognit. Comput.* 1, 77–90. doi: 10.1007/s12559-008-9000-9

Gu, H., Yang, M., Li, L., Liu, Z., and Ren, W. (2002). Experimental observation of the stochastic bursting caused by coherence resonance in a neural pacemaker. *Neuroreport* 13, 1657–1660. doi: 10.1097/00001756-200209160-00018

Gutfreund, H., Reger, J., and Young, A. (1988). The nature of attractors in an asymmetric spin glass with deterministic dynamics. *J. Phys. A Math. Gen.* 21, 2775. doi: 10.1088/0305-4470/21/12/020

Haviv, D., Rivkind, A., and Barak, O. (2019). "Understanding and controlling memory in recurrent neural networks," in *International Conference on Machine Learning* (Long Beach: PMLR), 2663–2671.

Hinton, G. E., and Sejnowski, T. J. (1983). "Optimal perceptual inference," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Vol. 448* (Washington, DC: Citeseer).

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Netw.* 2, 359–366. doi: 10.1016/0893-6080(89)90020-8

Ichikawa, K., and Kaneko, K. (2020). Short term memory by transient oscillatory dynamics in recurrent neural networks. *arXiv preprint* arXiv:2010.15308. doi: 10.1103/PhysRevResearch.3.033193

Ikemoto, S., DallaLibera, F., and Hosoda, K. (2018). Noise-modulated neural networks as an application of stochastic resonance. *Neurocomputing* 277, 29–37. doi: 10.1016/j.neucom.2016.12.111

Ilopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* 79, 2554. doi: 10.1073/pnas.79.8.2554

Jaeger, H. (2001). *The "echo state" approach to analysing and training recurrent neural networks-with an erratum note*. Bonn, Germany: German National Research Center for Information Technology GMD Technical Report 148, 13.

Jaeger, H. (2014). Controlling recurrent neural networks by conceptors. *arXiv preprint* arXiv:1403.3369.

Kadmon, J., and Sompolinsky, H. (2015). Transition to chaos in random neuronal networks. *Phys. Rev. X* 5, 041030. doi: 10.1103/PhysRevX.5.041030

Kaneko, K., and Suzuki, J. (1994). "Evolution to the edge of chaos in an imitation game," in *Artificial life III* (Santa Fe: Citeseer).

Krauss, P., Metzner, C., Schilling, A., Schütz, C., Tziridis, K., Fabry, B., et al. (2017). Adaptive stochastic resonance for unknown and variable input signals. *Sci. Rep.* 7, 1–8. doi: 10.1038/s41598-017-02644-w

Krauss, P., Prebeck, K., Schilling, A., and Metzner, C. (2019a). Recurrence resonance in three-neuron motifs. *Front. Comput. Neurosci.* 13, 64. doi: 10.3389/fncom.2019.00064

Krauss, P., Schuster, M., Dietrich, V., Schilling, A., Schulze, H., and Metzner, C. (2019b). Weight statistics controls dynamics in recurrent neural networks. *PLoS ONE* 14, e0214541. doi: 10.1371/journal.pone.0214541

Krauss, P., and Tziridis, K. (2021). Simulated transient hearing loss improves auditory sensitivity. *Sci. Rep.* 11, 1–8. doi: 10.1038/s41598-021-94429-5

Krauss, P., Tziridis, K., Metzner, C., Schilling, A., Hoppe, U., and Schulze, H. (2016). Stochastic resonance controlled upregulation of internal noise after hearing loss as a putative cause of tinnitus-related neuronal hyperactivity. *Front. Neurosci.* 10, 597. doi: 10.3389/fnins.2016.00597

Krauss, P., Tziridis, K., Schilling, A., and Schulze, H. (2018). Cross-modal stochastic resonance as a universal principle to enhance sensory processing. *Front. Neurosci.* 12, 578. doi: 10.3389/fnins.2018.00578

Krauss, P., Zankl, A., Schilling, A., Schulze, H., and Metzner, C. (2019c). Analysis of structure and dynamics in three-neuron motifs. *Front. Comput. Neurosci.* 13, 5. doi: 10.3389/fncom.2019.00005

Langton, C. G. (1990). Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D* 42, 12–37. doi: 10.1016/0167-2789(90)90064-V

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Legenstein, R., and Maass, W. (2007). Edge of chaos and prediction of computational performance for neural circuit models. *Neural Netw.* 20, 323–334. doi: 10.1016/j.neunet.2007.04.017

Lindner, B., and Schimansky-Geier, L. (2000). Coherence and stochastic resonance in a two-state system. *Phys. Rev. E* 61, 6103. doi: 10.1103/PhysRevE.61.6103

Lindner, B., Schimansky-Geier, L., and Longtin, A. (2002). Maximizing spike train coherence or incoherence in the leaky integrate-and-fire model. *Phys. Rev. E* 66, 031916. doi: 10.1103/PhysRevE.66.031916

Maheswaranathan, N., Williams, A. H., Golub, M. D., Ganguli, S., and Sussillo, D. (2019). Universality and individuality in neural dynamics across large populations of recurrent networks. *Adv. Neural Inf. Process. Syst.* 2019, 15629. doi: 10.48550/arXiv.1907.08549

McDonnell, M. D., and Abbott, D. (2009). What is stochastic resonance? definitions, misconceptions, debates, and its relevance to biology. *PLoS Comput. Biol.* 5, e1000348. doi: 10.1371/journal.pcbi.1000348

Mediano, P. A., Rosas, F. E., Farah, J. C., Shanahan, M., Bor, D., and Barrett, A. B. (2021). Integrated information as a common signature of dynamical and information-processing complexity. *arXiv preprint* arXiv:2106.10211. doi: 10.1063/5.0063384

Mejias, J., and Longtin, A. (2012). Optimal heterogeneity for coding in spiking neural networks. *Phys. Rev. Lett.* 108, 228102. doi: 10.1103/PhysRevLett.108.228102

Mejias, J. F., and Longtin, A. (2014). Differential effects of excitatory and inhibitory heterogeneity on the gain and asynchronous state of sparse cortical networks. *Front. Comput. Neurosci.* 8, 107. doi: 10.3389/fncom.2014.00107

Mejias, J. F., and Torres, J. J. (2011). Emergence of resonances in neural systems: the interplay between adaptive threshold and short-term synaptic plasticity. *PLoS ONE* 6, e17255. doi: 10.1371/annotation/3c57af7b-02a6-4267-b586-8b5a437fa5ba

Miner, D., and Triesch, J. (2016). Plasticity-driven self-organization under topological constraints accounts for non-random features of cortical synaptic wiring. *PLoS Comput. Biol.* 12, e1004759. doi: 10.1371/journal.pcbi.1004759

Molgedey, L., Schuchhardt, J., and Schuster, H. G. (1992). Suppressing chaos in neural networks by noise. *Phys. Rev. Lett.* 69, 3717. doi: 10.1103/PhysRevLett.69.3717

Moss, F., Ward, L. M., and Sannita, W. G. (2004). Stochastic resonance and sensory information processing: a tutorial and review of application. *Clin. Neurophysiol.* 115, 267–281. doi: 10.1016/j.clinph.2003.09.014

Narang, S., Elsen, E., Diamos, G., and Sengupta, S. (2017). Exploring sparsity in recurrent neural networks. *arXiv preprint* arXiv:1704.05119.

Natschläger, T., Bertschinger, N., and Legenstein, R. (2005). At the edge of chaos: Real-time computations and self-organized criticality in recurrent neural networks. *Adv. Neural Inf. Process Syst.* 17, 145–152.

Rajan, K., Abbott, L., and Sompolinsky, H. (2010). Stimulus-dependent suppression of chaos in recurrent neural networks. *Phys. Rev. E* 82, 011903. doi: 10.1103/PhysRevE.82.011903

Schäfer, A. M., and Zimmermann, H. G. (2006). "Recurrent neural networks are universal approximators," in *International Conference on Artificial Neural Networks* (Athens: Springer), 632–640.

Schilling, A., Gerum, R., Zankl, A., Schulze, H., Metzner, C., and Krauss, P. (2020). Intrinsic noise improves speech recognition in a computational model of the auditory pathway. *bioRxiv*. doi: 10.1101/2020.03.16.993725

Schilling, A., Maier, A., Gerum, R., Metzner, C., and Krauss, P. (2021a). Quantifying the separability of data classes in neural networks. *Neural Netw.* 139, 278–293. doi: 10.1016/j.neunet.2021.03.035

Schilling, A., Tziridis, K., Schulze, H., and Krauss, P. (2021b). The stochastic resonance model of auditory perception: a unified explanation of tinnitus development, zwicker tone illusion, and residual inhibition. *Prog. Brain Res.* 262, 139–157. doi: 10.1016/bs.pbr.2021.01.025

Schrauwen, B., Buesing, L., and Legenstein, R. (2009). "On computational power and the order-chaos phase transition in reservoir computing," in *22nd Annual Conference on Neural Information Processing Systems (NIPS 2008), Vol. 21* (Vancouver, BC: NIPS Foundation), 1425–1432.

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

Solé, R. V., and Miramontes, O. (1995). Information at the edge of chaos in fluid neural networks. *Physica D* 80, 171–180. doi: 10.1016/0167-2789(95)90075-6

Song, S., Sjöström, P. J., Reigl, M., Nelson, S., and Chklovskii, D. B. (2005). Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol.* 3, e68. doi: 10.1371/journal.pbio.0030068

Sporns, O. (2011). The non-random brain: efficiency, economy, and complex dynamics. *Front. Comput. Neurosci.* 5, 5. doi: 10.3389/fncom.2011.00005

Squire, L., Berg, D., Bloom, F. E., Du Lac, S., Ghosh, A., and Spitzer, N. C. (2012). *Fundamental Neuroscience*. Oxford: Academic Press.

Toral, R., Mirasso, C., and Gunton, J. (2003). System size coherence resonance in coupled fitzhugh-nagumo models. *EPL* 61, 162. doi: 10.1209/epl/i2003-00207-5

Torres, J., Marro, J., and Mejias, J. (2011). Can intrinsic noise induce various resonant peaks? *New J. Phys*. 13, 053014. doi: 10.1088/1367-2630/13/5/053014

Toyoizumi, T., and Abbott, L. (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

Uzuntarla, M., Torres, J. J., So, P., Ozer, M., and Barreto, E. (2017). Double inverse stochastic resonance with dynamic synapses. *Phys. Rev. E* 95, 012404. doi: 10.1103/PhysRevE.95.012404

Wallace, E., Maei, H. R., and Latham, P. E. (2013). Randomly connected networks have short temporal memory. *Neural Comput*. 25, 1408–1439. doi: 10.1162/NECO_a_00449

Wang, L., Pichler, E. E., and Ross, J. (1990). Oscillations and chaos in neural networks: an exactly solvable model. *Proc. Natl. Acad. Sci. U.S.A*. 87, 9467–9471. doi: 10.1073/pnas.87.23.9467

Wang, X. R., Lizier, J. T., and Prokopenko, M. (2011). Fisher information at the edge of chaos in random boolean networks. *Artif. Life* 17, 315–329. doi: 10.1162/artl_a_00041

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.

frontiers | Frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 09 May 2022
doi: 10.3389/fncom.2022.854218

The Face Inversion Effect in Deep Convolutional Neural Networks

Fang Tian[1†], *Hailun Xie*[2†], *Yiying Song*[2*], *Siyuan Hu*[2] *and Jia Liu*[3]

[1] State Key Laboratory of Cognitive Neuroscience and Learning and IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China, [2] Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, Beijing Normal University, Beijing, China, [3] Department of Psychology & Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing, China

The face inversion effect (FIE) is a behavioral marker of face-specific processing that the recognition of inverted faces is disproportionately disrupted than that of inverted non-face objects. One hypothesis is that while upright faces are represented by face-specific mechanism, inverted faces are processed as objects. However, evidence from neuroimaging studies is inconclusive, possibly because the face system, such as the fusiform face area, is interacted with the object system, and therefore the observation from the face system may indirectly reflect influences from the object system. Here we examined the FIE in an artificial face system, visual geometry group network-face (VGG-Face), a deep convolutional neural network (DCNN) specialized for identifying faces. In line with neuroimaging studies on humans, a stronger FIE was found in VGG-Face than that in DCNN pretrained for processing objects. Critically, further classification error analysis revealed that in VGG-Face, inverted faces were miscategorized as objects behaviorally, and the analysis on internal representations revealed that VGG-Face represented inverted faces in a similar fashion as objects. In short, our study supported the hypothesis that inverted faces are represented as objects in a pure face system.

Keywords: face inversion effect, deep convolutional neural network, VGG-Face, face system, AlexNet

OPEN ACCESS

Edited by:
Yu-Guo Yu,
Fudan University, China

Reviewed by:
Jiedong Zhang,
Institute of Biophysics (CAS), China
Jun Li,
Baidu, United States
Yuanyuan Mi,
Chongqing University, China

***Correspondence:**
Yiying Song
songyiying@bnu.edu.cn

[†] *These authors have contributed equally to this work*

Received: *13 January 2022*
Accepted: *22 March 2022*
Published: *09 May 2022*

Citation:
Tian F, Xie H, Song Y, Hu S and Liu J (2022) The Face Inversion Effect in Deep Convolutional Neural Networks. Front. Comput. Neurosci. 16:854218. doi: 10.3389/fncom.2022.854218

INTRODUCTION

Faces are an important type of visual stimulus in human social life and interaction, conveying a wealth of characteristic information (e.g., identity, age, and emotion) (Bahrick et al., 1975; O'Toole et al., 1998; Rhodes et al., 2011). Previous studies have found that humans processed faces differently from ordinary objects (e.g., Tanaka and Sengco, 1997). A classic manifestation of face specificity was the face inversion effect (FIE) (Yin, 1969; Valentine, 1988), in which humans are disproportionately less likely to recognize a face correctly when it is inverted than when an object (e.g., a cup) is inverted. However, the underlying mechanism of the FIE remains unclear.

Neuroimaging studies have been conducted to investigate how face-selective regions respond to upright and inverted faces. They found that the fusiform face area (FFA) is activated more highly when processing upright faces than inverted faces (Kanwisher et al., 1998; Yovel and Kanwisher, 2005; Epstein et al., 2006; Mazard et al., 2006). Further, the neural FIE observed in the FFA is positively correlated with behavioral FIE, suggesting that the FFA is likely the neural basis of the FIE (Yovel and Kanwisher, 2005; Zhu et al., 2011). In contrast, the activation of lateral occipital cortex (LOC), which is specialized for processing objects (Malach et al., 1995; Epstein, 2005), is greater during processing inverted faces than upright faces (Haxby et al., 1999; Yovel and Kanwisher, 2005). Taken together, the double dissociation of upright and inverted faces is considered as evidence that

they are processed by the face system and object system, respectively. However, the findings are inconclusive; for example, the FFA is still responsive to inverted faces (Kanwisher et al., 1998; Yovel and Kanwisher, 2005) and the LOC is still responsive to upright faces (Haxby et al., 1999; Yovel and Kanwisher, 2005). It is possibly because the FFA is interacted with the LOC (Haxby et al., 1999; Yovel and Kanwisher, 2005; Epstein et al., 2006) and cannot completely rule out the influences from the object processing system.

Deep convolutional neural network (DCNN), which is inspired by biological visual systems, is used to simulate human vision recently (Kriegeskorte, 2015; Parkhi et al., 2015; Simonyan and Zisserman, 2015; Krizhevsky et al., 2017; Liu et al., 2020; Song et al., 2020; Huang et al., 2021; Tian et al., 2021; Zhou et al., 2021). Here we used a representative DCNN for face recognition, VGG-Face (Parkhi et al., 2015), which is pretrained to identify faces only. In recent years, various deep learning methods have been used in face recognition systems (Fuad et al., 2021). Among the various methods, DCNN is the most popular deep learning method for face recognition (Fuad et al., 2021). Further, visual geometry group network-face (VGG-Face) is one of the most commonly used CNN models for face recognition (e.g., Ghazi and Ekenel, 2016; Karahan et al., 2016; Grm et al., 2017) and has shown successful performance of face recognition under various conditions (Ghazi and Ekenel, 2016). Therefore, we selected VGG-Face in the present study as representative of face recognition models. VGG-Face provides an ideal model for human face system, completely insulated from the interference of the object system. Here we asked how the artificial face system, VGG-Face, represented inverted faces.

METHODS

Deep Convolutional Neural Networks

As a pure face system, VGG-Face (available in https://www.robots.ox.ac.uk/~vgg/software/vgg_face/) is pretrained with the VGG Face Dataset. The architecture of VGG-Face includes 13 convolutional layers and 3 fully connected layers (i.e., FC1, FC2, and FC3), and the FC3 is a 2,622-dimensional classifier, corresponding to the 2,622 face identities to be identified during pretraining (Parkhi et al., 2015).

To compare the FIE between face system and object system, we used AlexNet (available in https://pytorch.org/) as an object system, which was pretrained for classifying objects with the ImageNet data set (Krizhevsky et al., 2017). AlexNet has an eight-layer architecture; the first five layers are the convolutional layers, and the last three layers are fully connected layers (i.e., FC1, FC2, and FC3); the FC3 layer is a classifier of 1,000 units.

To examine the effects of network architecture and pretraining task on the FIE, we also used VGG-16 (Simonyan and Zisserman, 2015), which has the same network architecture as VGG-Face but the same pretraining experience as AlexNet. In addition, we used an AlexNet with the same pretraining experience of face recognition as VGG-Face using the VGG Face Dataset (Grm et al., 2017).

Experiment Settings

The Face and Object Data Sets

We used a data set of 60 groups of images, of which 30 groups were face images and 30 groups were object images (**Figure 1**). Each group of face images contained images of one individual in different scenes, all selecting from CASIA-WebFace database (Yi et al., 2014). To rule out the effect of the pretrained face identities on the transfer training, the face identities in our data set did not overlap with those in VGG-Face pretraining data set. Each group of object images contained images of one specific cup in different scenes, and all object images were selected from the Internet. All the images were evaluated to ensure that the face or object in each image was complete. In each group, there were 75 images for transfer training, 25 images for validation, and 50 upright images and 50 inverted images for testing. The inverted stimuli were obtained by rotating the upright images 180 degrees. Thus, a total of 12,000 images were used in this study, with 4,500 images used for training, 1,500 images used for validation, and 6,000 images used for testing (1,500 upright faces, 1,500 inverted faces, 1,500 upright objects, and 1,500 inverted objects). Before training, the input images were normalized to a uniform size of 224 $\times$ 224 and normalized according to the mean and standard deviation of the ImageNet database (mean = [0.481,0.457,0.398], std = [0.237, 0.232, 0.231]).

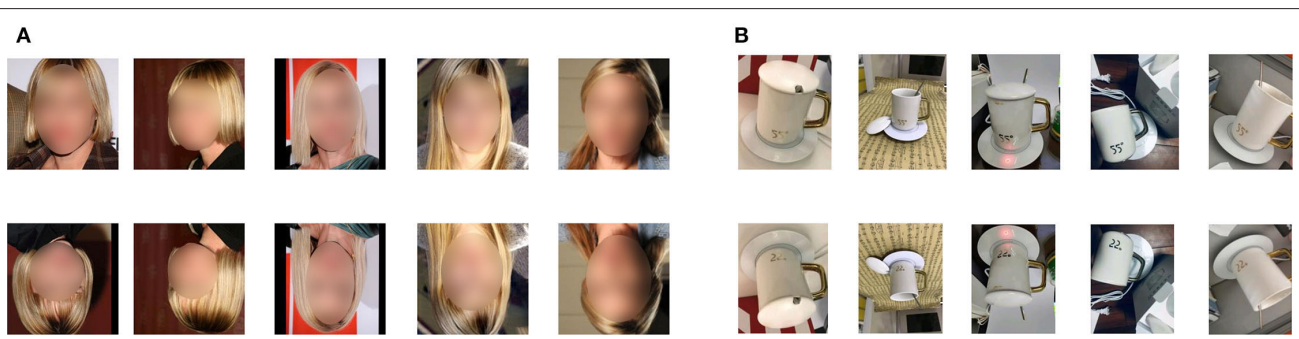


FIGURE 1 | Example stimuli in our study. **(A)** Top, upright faces; bottom, inverted faces. **(B)** Top, upright objects; bottom, inverted objects.

Transfer Learning

The transfer learning included a training period and a validation period. During the training period, the DCNNs were trained to classify images of 30 face identities and 30 cup identities. All network parameters of the pretrained DCNNs were frozen except for the last FC3 layer, using DNNBrain (Chen et al., 2020) based on the PyTorch. The FC3 layer of the DCNNs was changed to an FC classifier of 60 units to fit the training task. The 60-group data set was presented to DCNNs, and the classifiers of the FC3 layers were trained. A total of 100 epochs were performed in the training period, and the loss value of the network was generated after each epoch. The loss fluctuates within a stable range when the training is finished. After the training period, other exemplars of the 60 faces and cup identities were presented to the DCNNs in the validation period, and the recognition accuracy of validation was evaluated.

Testing Experiment

After transfer learning, we presented upright faces, inverted faces, upright objects, and inverted objects to the DCNNs in the testing experiment. The recognition accuracy was obtained by comparing the output and input identities of each image. Further, for classification error analysis, we examined the errors the DCNNs made in different conditions (i.e., whether upright/inverted faces were classified as objects and whether upright/inverted objects were classified as faces).

To further explore how the DCNNs represented upright and inverted images, we used representational similarity (RS) analysis (Kriegeskorte et al., 2008) to examine the RS of different stimulus identities in the DCNNs. We used DNNBrain (Chen et al., 2020) to extract the activation values of the 6,000 testing images in the three FC layers of the DCNNs. For both networks, the activation values of 4,096 units in FC1, 4,096 units in FC2, and 60 units in FC3 were extracted. The activation values of the 50 upright images and 50 inverted images of each identity were averaged, respectively, and for each identity, the activation patterns of upright and inverted conditions were obtained in each FC layer. Then, Pearson's correlation was calculated to obtain the representation similarity of different stimulus identities in each FC layer.

RESULTS

Transfer Learning of VGG-Face

VGG-Face was trained to classify images of 30 face identities and 30 cup identities in transfer learning. The training performance reached stability after 50 epochs. The validation accuracy of VGG-Face was 67.8%, significantly higher than the random level (random accuracy = 1.67%), which indicated that the transfer learning of VGG-Face was successful.

FIE in VGG-Face

We first examined whether there was an FIE in VGG-Face as a pure face system. We performed a two-way ANOVA analysis on recognition accuracy with orientation (upright, inverted) and stimuli category (faces, objects) as factors (**Figure 2A**). The main effects of both orientation [$F_{(1, 116)} = 824.76$, $p < 0.001$] and stimulus category [$F_{(1, 116)} = 55.90$, $p < 0.001$] were significant. There was an interaction between stimulus category and orientation [$F_{(1, 116)} = 228.58$, $p < 0.001$]. The accuracy of inverted images decreased more in face condition [$F_{(1, 116)} = 960.8$, $p < 0.001$] than in object condition [$F_{(1, 116)} = 92.47$, $p < 0.001$], indicating that there was an FIE in VGG-Face.

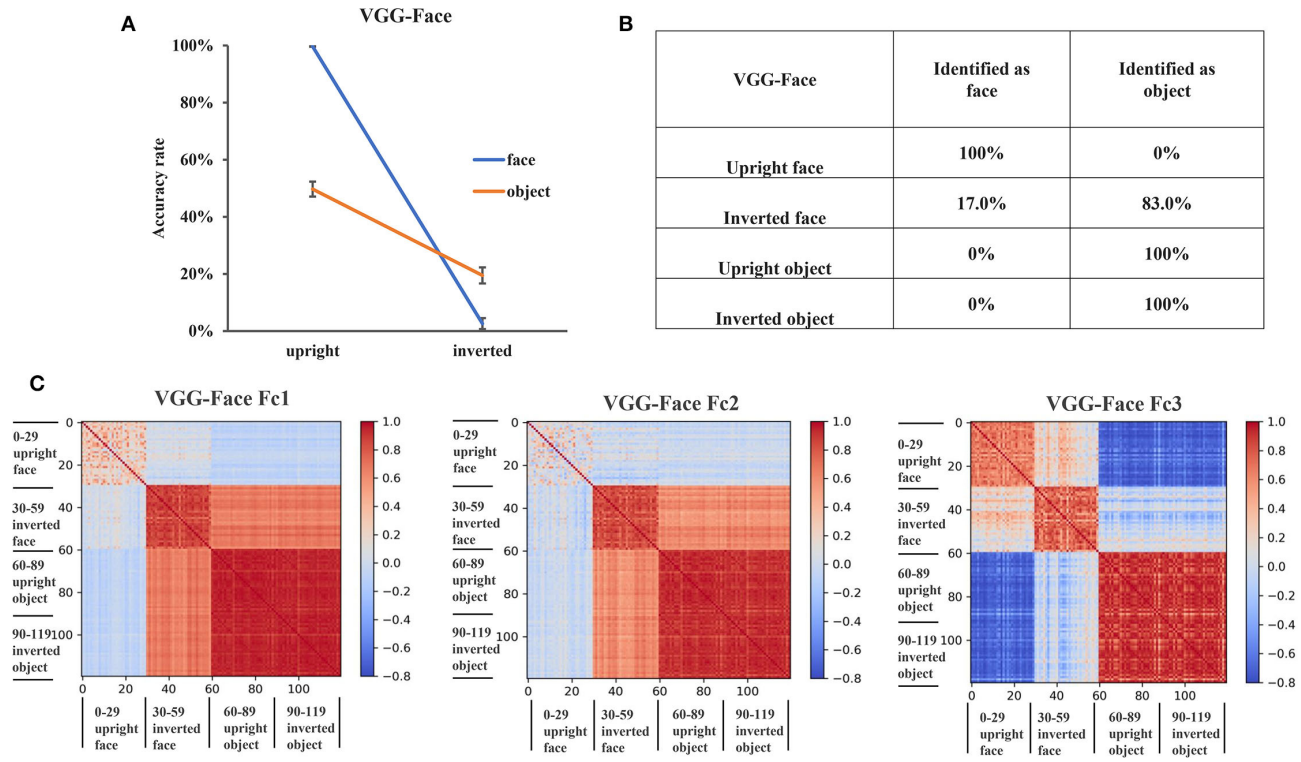
To further investigate why the VGG-Face showed an FIE, we examined the classification errors of VGG-Face in different conditions. While all upright faces were classified as faces and all upright and inverted objects were classified as objects, only 17% of the inverted faces were classified as faces and 83% were classified as objects in VGG-Face (**Figure 2B**). This result suggested that VGG-Face showed an FIE because it tended to classify inverted faces as objects.

VGG-Face Represented Inverted Faces Similarly as Objects

The misclassification of inverted faces as objects behaviorally suggested that inverted faces might be represented more similarly to objects in VGG-Face. To test this intuition, we performed the RS analysis in the three FC layers of VGG-Face. We found that in FC1 and FC2, the representation of inverted faces was clustered with that of the objects, rather than with that of upright faces (**Figure 2C**). In FC1 layer, the RS within upright faces (0.27) was much lower than that within inverted faces (0.84) and within objects (0.94). Importantly, the RS between inverted faces and objects was 0.66, while the RS between inverted and upright faces was only −0.037, and that between upright faces and objects was −0.07. The results in FC2 layer showed a similar pattern as in FC1. That is, the RS within upright faces (0.12) was much lower than that within inverted faces (0.82) and within objects (0.91). The RS between inverted faces and objects was 0.59, while the RS between inverted and upright faces was only −0.002, and that between upright faces and objects was −0.013. In FC3 layer, we observed that the representations of upright faces, inverted faces, and objects were clustered into three clusters (**Figure 2C**). The RS within upright faces was 0.66, the RS within inverted faces was 0.71, and the RS within objects was 0.87. The RS between upright faces and inverted faces was 0.21, the RS between upright faces and objects was −0.68, and the RS between inverted faces and objects was −0.05. These results indicated that inverted faces were represented more similarly as objects than as upright faces in the FC layers of VGG-Face, providing representational basis for the behavioral results that VGG-Face tended to classify inverted faces as objects.

AlexNet Did Not Show an FIE

Having shown the FIE and revealed its internal representations in a pure face system, VGG-Face, we next examined whether the FIE was specific to the pure face system or would also be observed in an object system. Here we used AlexNet (Krizhevsky et al., 2017), which was pretrained for object



VGG-Face	Identified as face	Identified as object
Upright face	100%	0%
Inverted face	17.0%	83.0%
Upright object	0%	100%
Inverted object	0%	100%

FIGURE 2 | Recognition performance and representations of VGG-Face. **(A)** The recognition accuracy of the upright and inverted faces and objects of VGG-Face. The error bars denote the standard error of the mean across the 30 groups of images in each condition. **(B)** The classification confusion matrix of VGG-Face. The percentage in the matrix denotes the classification errors of VGG-Face in each condition. **(C)** The representational similarity matrix of the three FC layers in VGG-Face. The color in the matrix indicates correlation values between activation patterns of different stimulus identities, with cool color in the matrix indicating low correlation and the warm color indicating high correlation.

categorization with ImageNet. The same procedure of transfer learning was applied for AlexNet as for VGG-Face, the training performance reached stability after 40 epochs. After training, the validation accuracy of AlexNet was 69%.

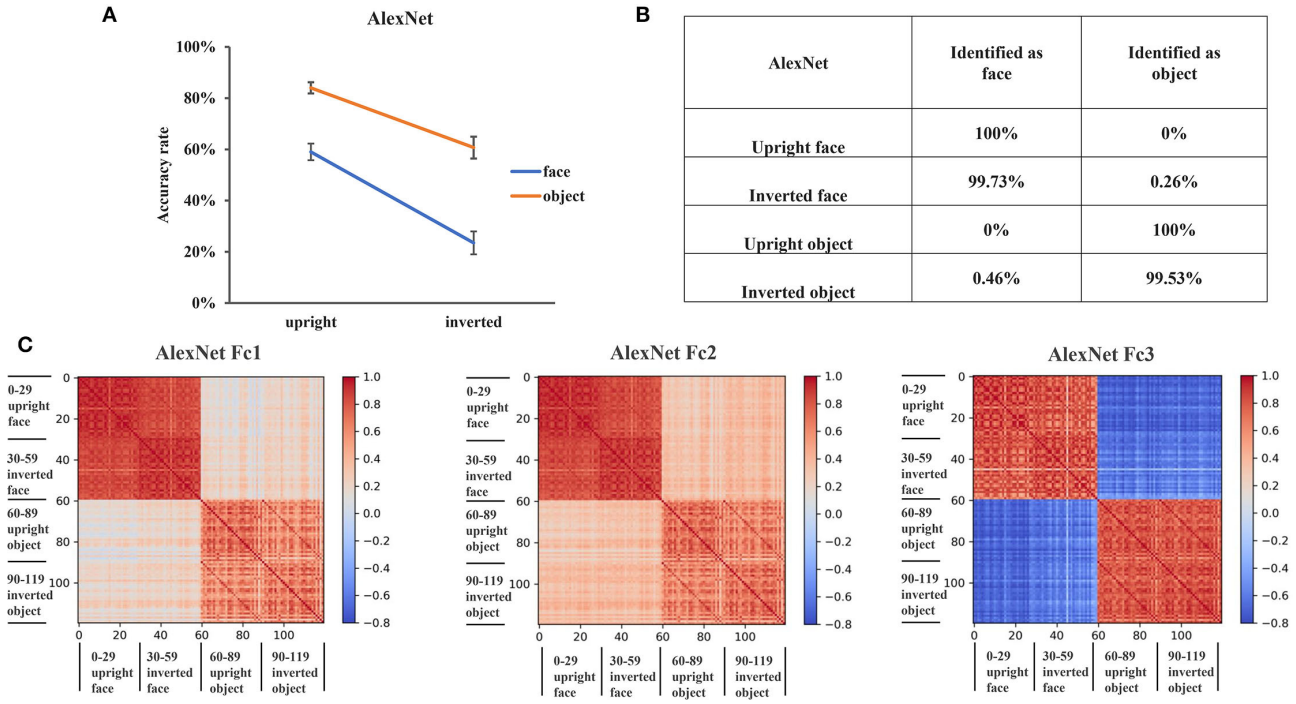
To examine whether there was an FIE in AlexNet, we performed a two-way ANOVA of orientation (upright, inverted) by stimuli category (faces, objects) on recognition accuracy. The main effects of both orientation [$F_{(1, 116)} = 62.27$, $p < 0.001$] and stimulus category [$F_{(1, 116)} = 70.43$, $p < 0.001$] were significant, but the interaction between stimulus category and orientation was not significant [$F_{(1, 116)} = 2.57$, $p = 0.11$] (**Figure 3A**). This result indicated that there was no FIE in AlexNet. We also performed a three-way ANOVA of orientation (upright, inverted), stimuli category (faces, objects), and network (AlexNet, VGG-Face), and the significant three-way interaction [$F_{(1, 232)} = 40.41$, $p < 0.001$] indicated that VGG-Face showed a greater FIE than AlexNet.

Then, we examined the classification errors of AlexNet in different conditions. In contrast to VGG-Face where most inverted faces were classified as objects, 99.7% of the inverted faces were classified as faces in AlexNet (**Figure 3B**). Besides, all upright faces were classified as faces, and all upright objects and 99.5% inverted objects were classified as objects in AlexNet. These results suggested that inverted faces were represented similarly as upright faces, rather than objects, in AlexNet.

To test this hypothesis, we performed the RS analysis in the three FC layers of AlexNet. We found that the representations of faces and objects were grouped into two clusters in AlexNet, regardless of the upright and inverted orientations (**Figure 3C**). In all FC layers, the within-category RS was greater than the between-category RS. That is, the RS between upright and inverted faces (FC1 layer was 0.86; FC2 layer was 0.86; FC3 layer was 0.78) and the RS between upright and inverted objects (FC1 layer was 0.62; FC2 layer was 0.64; FC3 layer was 0.79) were greater than the RS between faces and objects (FC1 layer was 0.19; FC2 layer was 0.34; FC3 layer was −0.63). These results indicated that upright and inverted faces were similarly represented in AlexNet.

VGG-16 Pretrained With Object Classification and AlexNet Pretrained With Face Recognition

The different FIEs observed in VGG-Face and AlexNet might be accounted for either by their different network architectures or by different pretraining tasks (face recognition vs. object classification). In order to explore the effects of pretraining



AlexNet	Identified as face	Identified as object
Upright face	100%	0%
Inverted face	99.73%	0.26%
Upright object	0%	100%
Inverted object	0.46%	99.53%

FIGURE 3 | Recognition performance and representations of AlexNet. **(A)** The recognition accuracy of the upright and inverted faces and objects of AlexNet. The error bars denote the standard error of the mean across the 30 groups of images in each condition. **(B)** The classification confusion matrix of AlexNet. The percentage in the matrix denotes the classification errors of AlexNet in each condition. **(C)** The representational similarity matrix of the three FC layers in AlexNet. The color in the matrix indicates correlation values between activation patterns of different stimulus identities, with cool color in the matrix indicating low correlation and the warm color indicating high correlation.

experience and network architecture on FIE, two more experiments were conducted. First, we used VGG-16 (Simonyan and Zisserman, 2015), which has the same network architecture as VGG-Face but the same pretraining task of object classification as AlexNet. Second, we used an AlexNet trained from scratch with the same pretraining experience of face recognition as VGG-Face (Grm et al., 2017).

The same procedure of transfer learning was applied, and the training performance reached stability after 50 epochs. After training, the validation accuracy was 65.4% for VGG-16 and 64.7% for AlexNet.

For VGG-16, we performed a two-way ANOVA of orientation (upright, inverted) by stimuli category (faces, objects) on recognition accuracy. The main effects of both orientation [$F_{(1, 116)} = 51.73$, $p < 0.001$] and stimulus category [$F_{(1, 116)} = 150.98$, $p < 0.001$] were significant, but the interaction between stimulus category and orientation was not significant [$F_{(1, 116)} = 0.96$, $p = 0.32$] (**Figure 4A**). That is, the VGG-16 pretrained with object classification did not show an FIE.

Similar analysis was performed for the AlexNet pretrained with face recognition. The main effect of orientation [$F_{(1, 116)} = 223.96$, $p < 0.001$] was significant and the main effect of stimulus category was not significant [$F_{(1, 116)} = 0.002$, $p = 0.96$]. Importantly, there was an interaction between stimulus category and orientation [$F_{(1, 116)} = 17.29$, $p < 0.001$] (**Figure 4B**). The accuracy of inverted images decreased more in face condition [$F_{(1, 116)} = 182.85$, $p < 0.001$] than in object condition [$F_{(1, 116)} = 58.37$, $p < 0.001$], indicating that there was an FIE in the AlexNet pretrained for face recognition.

Taken together, the two network architectures showed similar FIE after pretrained with face recognition task, but showed no FIE after pretrained with object classification task. These results suggested that the observed FIE in DCNNs may result from pretraining experience of face recognition, rather than particular DCNN network architectures.

DISCUSSION

In this study, we used VGG-Face to examine whether there was an FIE in an artificial pure face system and how upright and inverted faces were represented in this system. We found that there was an FIE in VGG-Face and the FIE was stronger than that in AlexNet which was pretrained for processing objects. Further classification error analysis revealed that in VGG-Face, inverted faces were misclassified as objects behaviorally, and the analysis on internal representations revealed that the VGG-Face represented inverted faces in a similar fashion as objects. These findings supported the hypothesis that inverted faces are represented as objects

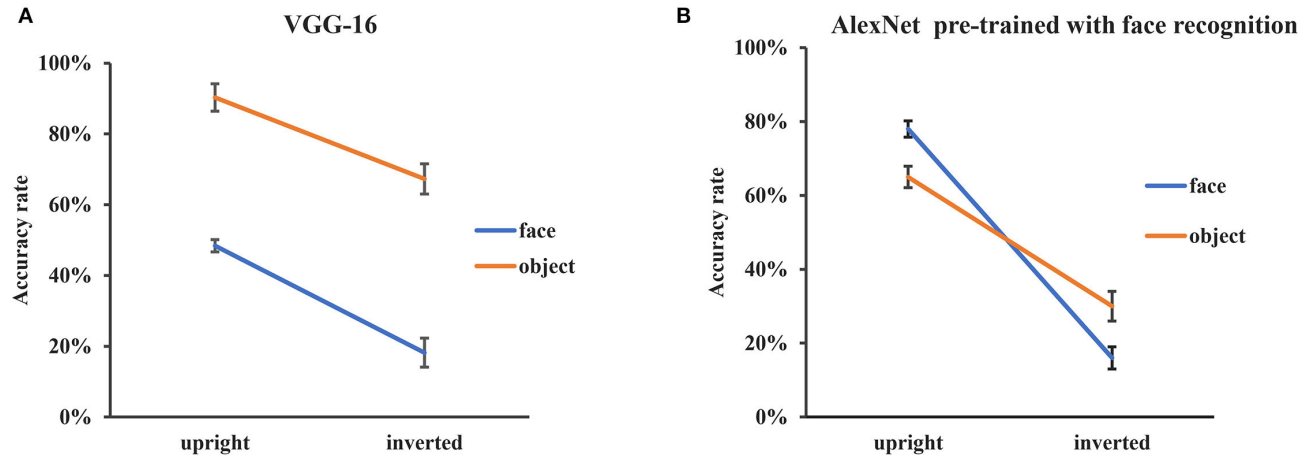


FIGURE 4 | Recognition performance of VGG-16 and AlexNet pre-trained with face recognition. **(A)** The recognition accuracy of the upright and inverted faces and objects of VGG-16 pretrained with object classification. **(B)** The recognition accuracy of the upright and inverted faces and objects of AlexNet pretrained with face recognition. The error bars denote the standard error of the mean across the 30 groups of images in each condition.

in the face system. Although fMRI studies have revealed some neural basis of FIE, especially in the FFA (Yovel and Kanwisher, 2005), the results are inconclusive, which may be due to the fact that the face system is not completely insulated from object system in human brain. By using an artificial pure face system as well as a pure object system, our study provides a clearer account for the representations underlying FIE.

The FIE found in VGG-Face as a pure face system is consistent with previous human fMRI findings showing an FIE (i.e., higher response to upright than inverted faces) in the face-selective FFA, and the FIE in the FFA correlates with the behavioral FIE (Yovel and Kanwisher, 2005). Further, fMRI adaptation results provide a possible neural basis for the behavioral FIE by showing that the FFA was more sensitive to identity differences between upright faces than inverted faces (Yovel and Kanwisher, 2005). This finding fits nicely with our results that the RS within upright faces was much lower than that within inverted faces and objects in VGG-Face, indicating that different identities were more uniquely represented in upright faces than inverted faces. Our results are also consistent with a previous study which showed a similar FIE using pretrained VGG-Face (Elmahmudi and Ugail, 2019).

More importantly, we extended previous finding by revealing representations underlying the observed FIE. First, we found that VGG-Face misclassified the inverted faces as objects behaviorally. This result is in line with neuropsychological finding that a patient with object recognition impairment was severely impaired in recognition of inverted faces, but normal at recognition of upright faces (Moscovitch et al., 1997). Moreover, RS analysis showed that in VGG-Face, inverted faces were represented similarly as objects, while representation of upright faces was separate from those of inverted faces and objects. Together, these results provide novel and clear evidence for an account of human FIE that inverted faces are represented by general object mechanisms whereas upright faces are represented by mechanisms specialized for faces (Yin, 1969; Pitcher et al., 2011).

In contrast, the AlexNet and VGG-16 pretrained for object categorization did not show an FIE, and upright and inverted faces were similarly represented in AlexNet. This result is consistent with human fMRI results that the object-selective LOC shows similar sensitivity to face identities for upright and inverted faces (Yovel and Kanwisher, 2005). Notably, although the AlexNet did not show a behavioral FIE in our study, it is reported that responses of face-selective units in untrained AlexNet responded more highly to upright faces than inverted faces (Baek et al., 2021). The discrepancy may be caused by different analysis levels (behavioral level vs. single unit response level) or different layers (FC layers vs. convolution layers). It will be interesting to examine whether untrained AlexNet will show an FIE behaviorally.

In sum, the present study showed an FIE in an artificial pure face system. Our study highlighted the important role of pretraining of face identification for a system to show the FIE; future studies are awaited to examine whether other DCNN networks or other types of deep learning models pretrained with face identification tend to show a similar FIE and whether the exposure of face stimuli or the task of face identification is more critical. Additionally, our study provided evidence for a possible mechanism of the FIE that inverted faces are represented as objects while the upright faces are represented differently from objects and inverted faces. Human behavioral studies have suggested that processing of upright faces is special in that they are processed in a holistic manner, while processing of inverted faces and non-face objects is based on featural information (Young et al., 1987; Tanaka, 1993; Farah et al., 1995; Tanaka and Sengco, 1997; Maurer et al., 2002; Tanaka and Farah, 2006). Future studies are invited to examine in what manners

upright and inverted faces are represented in artificial face system. Finally, our study may inspire more researchers to use DCNNs to explore the cognitive mechanisms of face recognition, especially the problems that cannot be solved with human subjects because of some limitations (such as ethics, experience, and workload).

DATA AVAILABILITY STATEMENT

All codes for analyses and examples of datasets are available on https://github.com/Helen-Xie-Nep/DCNN-Face-inversion.

AUTHOR CONTRIBUTIONS

JL and SH designed the research. HX collected the data. HX and FT analyzed the data. FT, JL, and YS wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (31861143039, 31872786).

REFERENCES

Baek, S., Song, M., Jang, J., Kim, G., and Paik, S. B. (2021). Face detection in untrained deep neural networks. *Nat. Commun.* 12:7328. doi: 10.1038/s41467-021-27606-9

Bahrick, H. P., Bahrick, P. O., and Wittlinger, R. P. (1975). Fifty years of memory for names and faces: a cross-sectional approach. *J. Exp. Psychol.-Gen.* 104, 54–75. doi: 10.1037/0096-3445.104.1.54

Chen, X., Zhou, M., Gong, Z., Xu, W., Liu, X., Huang, T., et al. (2020). DNNBrain: a unifying toolbox for mapping deep neural networks and brains. *Front. Comput. Neurosci.* 14:580632. doi: 10.3389/fncom.2020.580632

Elmahmudi, A., and Ugail, H. (2019). Deep face recognition using imperfect facial data. *Futur. Gener. Comp. Syst.* 99, 213–225. doi: 10.1016/j.future.2019.04.025

Epstein, R. A. (2005). The cortical basis of visual scene processing. *Vis. Cogn.* 12, 954–978. doi: 10.1080/13506280444000607

Epstein, R. A., Higgins, J. S., Parker, W., Aguirre, G. K., and Cooperman, S. (2006). Cortical correlates of face and scene inversion: a comparison. *Neuropsychologia* 44, 1145–1158. doi: 10.1016/j.neuropsychologia.2005.10.009

Farah, M. J., Tanaka, J. W., and Drain, H. M. (1995). What causes the face inversion effect? *J. Exp. Psychol.-Hum. Percept. Perform.* 21, 628–634. doi: 10.1037/0096-1523.21.3.628

Fuad, M. T. H., Fime, A. A., Sikder, D., Iftee, M. A. R., Rabbi, J., Al-Rakhami, M. S., et al. (2021). Recent advances in deep learning techniques for face recognition. *IEEE Access.* 9, 99112–99142. doi: 10.1109/ACCESS.2021.3096136

Ghazi, M. M., and Ekenel, H. K. (2016). "A comprehensive analysis of deep learning based representation for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Las Vegas, NV: IEEE) 2016, 102–109.

Grm, K., Štruc, V., Artiges, A., Caron, M., and Ekenel, H. K. (2017). Strengths and weaknesses of deep learning models for face recognition against image degradations. *IET Biom.* 7, 81–89. doi: 10.1049/iet-bmt.2017.0083

Haxby, J. V., Ungerleider, L. G., Clark, V. P., Schouten, J. L., Hoffman, E. A., and Martin, A. (1999). The effect of face inversion on activity in human neural systems for face and object perception. *Neuron* 22, 189–199. doi: 10.1016/S0896-6273(00)80690-X

Huang, T., Zhen, Z., and Liu, J. (2021). Semantic relatedness emerges in deep convolutional neural networks designed for object recognition. *Front. Comput. Neurosci.* 15, 625804. doi: 10.3389/fncom.2021.625804

Kanwisher, N., Tong, F., and Nakayama, K. (1998). The effect of face inversion on the human fusiform face area. *Cognition* 68, B1–B11. doi: 10.1016/S0010-0277(98)00035-3

Karahan, S., Yildirum, M. K., Kirtac, K., Rende, F. S., Butun, G., and Ekenel, H. K. (2016). "How image degradations affect deep CNN-based face recognition?," in *International Conference of the Biometrics Special Interest Group (BIOSIG)* (Darmstadt: IEEE) 2016, 1–5. doi: 10.1109/BIOSIG.2016.7736924

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* 1, 417–446. doi: 10.1146/annurev-vision-082114-035447

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2, 4. doi: 10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM.* 60, 84–90. doi: 10.1145/3065386

Liu, X., Zhen, Z., and Liu, J. (2020). Hierarchical sparse coding of objects in deep convolutional neural networks. *Front. Comput. Neurosci.* 14:578158. doi: 10.3389/fncom.2020.578158

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., et al. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proc. Natl. Acad. Sci. U.S.A.* 92, 8135–8139. doi: 10.1073/pnas.92.18.8135

Maurer, D., Le Grand, R., and Mondloch, C. J. (2002). The many faces of configural processing. *Trends Cogn. Sci.* 6, 255–260. doi: 10.1016/S1364-6613(02)01903-4

Mazard, A., Schiltz, C., and Rossion, B. (2006). Recovery from adaptation to facial identity is larger for upright than inverted faces in the human occipito-temporal cortex. *Neuropsychologia* 44, 912–922. doi: 10.1016/j.neuropsychologia.2005.08.015

Moscovitch, M., Winocur, G., and Behrmann, M. (1997). What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *J. Cogn. Neurosci.* 9, 555–604. doi: 10.1162/jocn.1997.9.5.555

O'Toole, A. J., Deffenbacher, K. A., Valentin, D., McKee, K., Huff, D., and Abdi, H. (1998). The perception of face gender: the role of stimulus structure in recognition and classification. *Mem. Cogn.* 26, 146–160. doi: 10.3758/BF03211378

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). "Deep face recognition," in *Proceedings of the British Machine Vision Conference 2015* (Swansea: British Machine Vision Association), 41.1–41.12. doi: 10.5244/C.29.41

Pitcher, D., Duchaine, B., Walsh, V., Yovel, G., and Kanwisher, N. (2011). The role of lateral occipital face and object areas in the face inversion effect. *Neuropsychologia* 49, 3448–3453. doi: 10.1016/j.neuropsychologia.2011.08.020

Rhodes, G., Calder, A., Johnson, M., and Haxby, J. V. (2011). The Oxford handbook of face perception. *Perception* 41, 1410–1411. doi: 10.1093/oxfordhb/9780199559053.001.0001

Simonyan, K., and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv [Preprint]*. arXiv: 1409.1556. Available online at: http://arxiv.org/abs/1409.1556.pdf

Song, Y., Qu, Y., Xu, S., and Liu, J. (2020). Implementation-independent representation for deep convolutional neural networks and humans in processing faces. *Front. Comput. Neurosci.* 14, 601314. doi: 10.3389/fncom.2020.601314

Tanaka, J. W. (1993). Parts and wholes in face recognition. *Q. J. Exp. Psychol.* 46, 225–245. doi: 10.1080/14640749308401045

Tanaka, J. W., and Farah, M. J. (2006). "The holistic representation of faces," in *Perception of Faces, Objects and Scenes: Analytic and Holistic Processes*, eds M. A. Peterson and G. Rhodes (New York, NY: Oxford University Press), 53–74. doi: 10.1093/acprof:oso/9780195313659.003.0003

Tanaka, J. W., and Sengco, J. A. (1997). Features and their configuration in face recognition. *Mem. Cogn.* 25, 583–592. doi: 10.3758/BF03211301

Tian, J., Xie, H., Hu, S., and Liu, J. (2021). Multidimensional face representation in a deep convolutional neural network reveals the mechanism underlying AI racism. *Front. Comput. Neurosci.* 15, 620281. doi: 10.3389/fncom.2021.620281

Valentine, T. (1988). Upside-down faces: a review of the effect of inversion upon face recognition. *Br. J. Psychol.* 79, 471–491. doi: 10.1111/j.2044-8295.1988.tb02747.x

Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). Learning face representation from scratch. *arXiv [Preprint]*. arXiv: 1411.7923. Available online at: https://arxiv.org/pdf/1411.7923.pdf

Yin, R. K. (1969). Looking at upside-down faces. *J. Exp. Psychol.* 81, 141–145. doi: 10.1037/h0027474

Young, A. W., Fau, H. D., and Hay, D. C. (1987). Configurational information in face perception. *Perception* 16, 747–759. doi: 10.1068/p160747

Yovel, G., and Kanwisher, N. (2005). The neural basis of the behavioral face-inversion effect. *Curr. Biol.* 15, 2256–2262. doi: 10.1016/j.cub.2005.10.072

Zhou, C., Xu, W., Liu, Y., Xue, Z., Chen, R., Zhou, K., et al. (2021). Numerosity representation in a deep convolutional neural network. *J. Pac. Rim Psychol.* 15, 1–11. doi: 10.1177/18344909211012613

Zhu, Q., Zhang, J., Luo, Y. L., Dilks, D. D., and Liu, J. (2011). Resting-state neural activity across face-selective cortical regions is behaviorally relevant. *J. Neurosci.* 31, 10323–10330. doi: 10.1523/JNEUROSCI.0873-11.2011

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.

frontiers | Frontiers in Computational Neuroscience

OPINION
published: 11 May 2022
doi: 10.3389/fncom.2022.865805

Computational Psychiatry and Computational Neurology: Seeking for Mechanistic Modeling in Cognitive Impairment and Dementia

Ludmila Kucikova[1], *Samuel Danso*[2], *Lina Jia*[3] *and Li Su*[1,4]*

[1] Department of Neuroscience, Sheffield Institute for Translational Neuroscience, Insigneo Institute for in silico Medicine, University of Sheffield, Sheffield, United Kingdom, [2] Edinburgh Dementia Prevention and Centre for Clinical Brain Sciences, Edinburgh Medical School, University of Edinburgh, Edinburgh, United Kingdom, [3] Beijing Anding Hospital, Capital Medical University, Beijing, China, [4] Department of Psychiatry, School of Clinical Medicine, University of Cambridge, Cambridge, United Kingdom

Keywords: dementia, computational psychiatry, computational neurology, computational neuroscience, computational modeling, machine learning, artificial intelligence

INTRODUCTION

The prevalence of dementia is increasing globally and carries a growing personal and societal burden (Guerchet et al., 2013). Multimodal and longitudinal neuroimaging provides biomarkers about disease progression and informs early detection of dementia (Ten Kate et al., 2018). However, current empirical data is still insufficient to infer the underlying mechanisms of the disorder necessary for developing targeted therapeutics. Equally important to the lack of empirical data, there is an absence of sufficient theoretical tools to investigate the relationships among the genetic risks, neuropathophysiology, clinical symptoms and environmental factors in neurodegenerative diseases.

We argue that the recent advances in computational psychiatry and computational neurology offer a promising translational neuroscience framework for integrating multiple levels of abstractions and investigating neurobiological and pathological mechanisms of dementia. In addition, they can derive mechanistic models that predict disease trajectory and treatment effects. Here, we extend historical discussions on this topic (Adams et al., 2016; Paulus et al., 2016; Hitchcock et al., 2022) by discussing the potential of integrative computational modeling for dementia research. We will discuss the potential translational benefits and how it might account for some of the current limitations in dementia research.

COMPUTATIONAL PSYCHIATRY AND COMPUTATIONAL NEUROLOGY

With the rise of computational and data sciences applications in biological, medical, biomedical, and psychological disciplines since the early 2010s, computational psychiatry and computational neurology have demonstrated the potential to help account for some of the limitations of traditional techniques (Montague et al., 2012). Computational psychiatry and neurology are interwind and overlap in disorders like dementia, so in the context of this paper, we do not distinguish them.

There are several different dichotomies of computational psychiatry and neurology models (e.g., descriptive vs. predictive, discriminative vs. generative, exploratory vs. confirmatory models). We recognize that all these dichotomies have their value in

OPEN ACCESS

Edited by:
Si Wu,
Peking University, China

Reviewed by:
Yashar Zeighami,
McGill University, Canada

***Correspondence:**
Li Su
l.su@sheffield.ac.uk

Received: *30 January 2022*
Accepted: *25 April 2022*
Published: *11 May 2022*

Citation:
Kucikova L, Danso S, Jia L and Su L (2022) Computational Psychiatry and Computational Neurology: Seeking for Mechanistic Modeling in Cognitive Impairment and Dementia. Front. Comput. Neurosci. 16:865805. doi: 10.3389/fncom.2022.865805

describing specific classes of computational models. However, in this paper, we would like to make a distinction between two types of scientific research approaches by their different objectives: data-driven vs. theory-driven approaches. We argue that the former primarily aims to explain patterns in novel data or information "about" the diseased brain, while the latter primarily aims to develop, validate, or falsify theories which describe information of the brain.

Data-Driven Computational Approaches

The primary objective of data-driven approaches is to "label" experimental data based on multivariate patterns or statistical regularities in the data (e.g., by using machine learning). An advantage of data-driven approaches is that they require minimal prior assumptions about the data (Magoulas and Pentza, 1999). However, this very characteristic also makes the interpretation of data-driven models challenging as discussed by Goecks and colleagues (Goecks et al., 2020).

The current applications of machine learning in dementia research focus on disease detection and prediction. For example, by using neuroimaging, biological, and clinical data based on recurrent neural networks, support vector machines, decision trees, Naïve Bayes classifiers, clustering, or other methods (Cui and Liu, 2019; Kuan et al., 2021; Skolariki et al., 2021). Moreover, the efforts were made to develop a personalized dementia risk model that could predict the onset of dementia years before patients develop symptoms by using ensemble learning from demographic and medical history data (Danso et al., 2021). Data-driven applications were also used to discriminate different types of dementia (Dauwan et al., 2016; Bougea et al., 2021) or to decrease the number of measures necessary for diagnosis (Weakley et al., 2015). Other computer-aided diagnosis systems that automatically detect neurological abnormalities have been developed for the identification of dementia from neuroimaging data (Siuly and Zhang, 2016). However, translating these efforts into clinical practice is still problematic and need more of easily used real-time methods that can be incorporated into everyday clinical practice.

Although, some examples that directly aim to model disease mechanisms, neuropathology, or subtypes exist (Young et al., 2014; Oxtoby et al., 2018; Su et al., 2018, 2021), majority of mainstream data-driven approaches do not explicitly intend to capture the neurobiological and neuropathological mechanisms underlying dementia. While applicable for disease prediction and diagnosis, data-driven approaches alone are still limited to inform novel treatments and capture the underlying complexities of dynamic nature of dementia (i.e., interactions of multiple disease factors on different levels that can evolve in complex ways over time).

Theory-Driven Computational Approaches

Theory-driven approaches for computational psychiatry and neurology are used to describe the mechanisms of altered pathology or information processing related to the "cause" of psychiatric or neurological conditions. They are used as tools for characterizing what nervous systems do (i.e., descriptive models), determining how they function (i.e., mechanistic models), and understanding why they operate in particular ways (i.e., interpretive models) at multiple levels of abstraction (Dayan and Abbott, 2001). Thus, their goal is fundamentally different from data-driven approaches as they "force" us to seek mechanisms and causality (**Figure 1A**).

First, biophysical models of synaptic, cellular, and neural circuits aim to describe the association between psychiatric symptoms and abnormal information processing intrinsic to assemblies of neurons and microcircuit dysfunctions. These models aim to investigate mechanisms underlying cognitive decline and dementia. For instance, early work like the "synaptic deletion and compensation" model (Horn et al., 1993; Ruppin and Reggi, 1995) demonstrated that synaptic connections in Alzheimer's disease are associated with memory loss and learning difficulties. Second, large-scale neural network models address the links between psychiatric problems and information processing dysfunction intrinsic to large circuit functions (e.g., Raj and Itturia-Medina, 2019). Third, normative models address how the nervous system should behave and how certain behavior or neural activity deviates from those standards (e.g., "Perception and Attention Deficit model"; Collerton et al., 2005; Makin et al., 2013).

A concern with many theory-driven models is that they are often based upon mechanisms that are not directly accessible from experimental data (Moran et al., 2011) and provide very specific assumptions that do not lead to empirically testable predictions (Baker et al., 2018). Hence, methods that can bridge the gap between modeling the clinically relevant symptoms (at macro-level) and modeling the brain where the neurobiological mechanisms are implemented (at micro-level) are urgently needed. Here, we argue that intermedium level models would complement macro- and micro-level models, being specifically targeted at "meso-level" modeling. This allows for directly represented distributed control of neural mechanisms and neurobiologically detailed cellular functions.

At this meso-level, the model has sufficient complexity to prescribe the hierarchical architecture of the brain (i.e., a layer of units rather a single unit representing a brain region); while each layer still can implement macro-level distributed representations for perception, action, and language. Each unit in the model also includes micro-level biological details (e.g., membrane potentials, ion channels, neurotransmitters) allowing empirical validation such as by using neuroimaging. This has the potential to extend traditional models that are predominantly informative on only one level of abstraction (**Figure 1B**).

Examples of such models include our work on modeling attentional impairments in Alzheimer's disease and making predictions about possible electrophysiological features in the relevant neural circuits (Mavritsaki et al., 2019). These models can be seen as "virtual" patients capturing the cognitive dysfunctions on computer simulations. By testing the models with neuroimaging, biological and clinical data from real patients, we can obtain mechanistic understanding and develop new drug treatments *in-silico* before they are experimented on animals and humans. This can further speed up translational effects and drug development in an area of great unmet need,

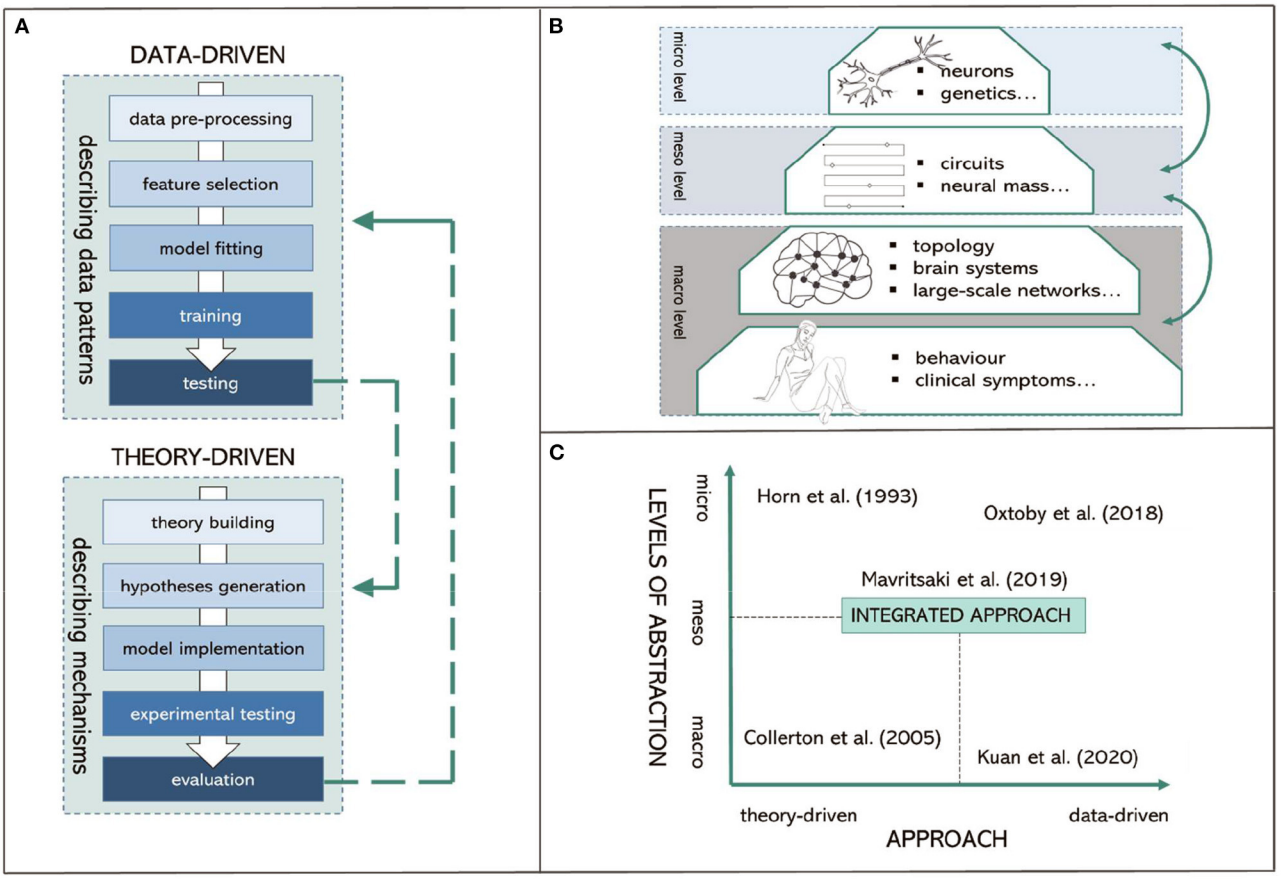


FIGURE 1 | Visual summary of proposed integrated computational modeling. **(A)** Data-driven and theory-driven pipelines. **(B)** Different levels of abstraction. **(C)** Proposed integrated computational modeling with examples.

reduce socio-economic impact, and add to sustainability efforts in dementia research.

Combining Data-Driven and Theory-Driven Computational Approaches

Combining theory-driven and data-driven approaches in a single modeling framework has the potential to account for the complexity of different forms of dementia while considering overlapping pathologies and clinical symptoms (**Figure 1C**). This is particularly crucial for: (I.) providing the understanding of the underlying mechanisms and "causal" interactions obtained by theoretical models, and for (II.) overcoming the current scalability limitations of theoretical models (Baker et al., 2018). For example, Maia and Frank (2011) used reinforcement learning to quantify the learning ability of individuals with Parkinson's disease based on the underlying dopaminergic mechanisms. Pinaya et al. (2021) used deep learning on normative models of brain structure to detect Alzheimer's disease progression. Bayesian models such as Dynamic Causal Modeling (DCM; Friston et al., 2016) can be applied to neuroimaging data to describe effective connectivity within and among neural circuits providing a principled data-driven way to fit subsets of model parameters in theory-driven models.

Theory-driven models provide prior knowledge and context for estimating features specifically relevant to disorders. This enables data-driven models to derive parameters for further modeling (e.g., biophysically realistic recurrent neural network models, algorithmic reinforcement learning models, Bayesian models) with increased efficiency and reliability (Huys et al., 2016). Hence, combining these approaches and linking between the levels of abstraction have the potential to increase translational benefits by relating symptoms and cognitive functions to clinically traceable entities such as cellular processes.

For instance, AI models based on recurrent neural networks are approaching human-level performance in many domains. Thus, if implemented with plausible biological details, recurrent neural networks can be "meso-level" models trained to simulate patients' clinical symptoms while the model parameters are simultaneously fitted to patients' neuroimaging and biological data. These models often contain tens of thousands of artificial neurons, making them large enough to model complex symptomology while remaining tractable to study neural mechanisms in unprecedented detail. Previous work summarized

the application of such models to complex psychiatric disorders where sufficient information about the relevant circuits exist, such as in schizophrenia (Huys et al., 2016). We argue that there is potential to extend these applications to study underlying mechanisms of different forms of dementia. Additionally, by integrating computational models with neuroimaging, neuronal dysfunction underlying psychosis symptoms in dementia such as hallucinations, delusions, paranoia could be explained by impairments in multiple neurotransmitter systems (Marreiros et al., 2013). This could link clinical symptoms with biological details more comprehensively than was previously available.

DISCUSSION

Computational psychiatry and neurology endeavor a biopsychological and mechanistic perspective by showing how each level of abstraction ranging from molecular to circuit levels can provide a context for the human brain's hierarchical architecture, functioning and disorders. Computational approaches provide a whole new lexicon for understanding neural processes (Montague et al., 2012). Machine learning techniques can detect complex and subtle mental and brain dysfunctions and their neurobiological underpinnings that are otherwise difficult to uncover.

Computational approaches are a valuable tool moving forward in research, but the current implementation introduces several challenges. First, the availability of good quality data is crucial to create reliable, accurate, and robust data-driven and theory-driven models of mental health illnesses and brain disorders. This includes the need for widely generalisable, open access and reproducible data of different dementia types with large sample sizes (Pellegrini et al., 2018). Second, the differences in the standardization in the dementia care pathway across clinical practices, assessment centers and research might pose a further challenge in appropriate data digitalisation, curation, and integration (Wong-Lin et al., 2020). Third, researchers should additionally ensure that data pre-processing steps do not introduce unrealistic attributes to general healthcare datasets when used in modeling less-common types of dementia.

When interpreting models, researchers need to be conscious of potential challenges. For instance, predictive models in psychiatry still suffer from overfitting and lack of generalisability and validation (Meehan et al., 2022). Finding a model with the appropriate complexity, therefore, requires finding a suitable balance between bias and variance (Lever et al., 2016). Yet, testing and falsifying current models and subsequent ability to develop accurate predictions of both common and rare types of dementia are still challenging due to the limited knowledge of their neurobiology and neuropathology.

Traditional AI modeling techniques and biophysical models often consider the macro-level constraints on the brain and mind for very specific cognitive phenomena. However, existing neural models do not satisfactorily provide an architectural-level explanation for how symptoms experienced by patients were mechanistically produced by genetic, molecular, and circuit abnormalities in different inter-connected brain regions. Moving forward, we need models that do not only describe data but can also manipulate data while preserving the integrity of the data. In other words, we argue to "treat the models as if they are data".

If present limitations are considered thoroughly, computational psychiatry and neurology models for dementia have the potential to establish as experimental medicine platform for personalized medicine and development of novel treatments and to advance the predictive health systems that would support clinicians in their decision-making process (Miotto et al., 2017). Importantly, current challenges are likely to be minimized by fast-evolving computational fields such as AI models based on deep-learning, which can serve as the basis of "virtual" patients. This allows for testing mechanistic hypotheses, while maintaining the informative value obtained from real life data.

AUTHOR CONTRIBUTIONS

LK and LS wrote the manuscript. SD and LJ contributed to the writing and provided feedback for the manuscript. LS secured the funding and oversaw the study. All authors have made a substantial contribution to this work and approved it for publication.

FUNDING

LK was the recipient of the University of Sheffield Flagship Scholarship. LS's participation was funded by Alzheimer's Research UK Senior Research Fellowship (ARUK-SRF2017B-1).

REFERENCES

Adams, R. A., Huys, Q. J. M., and Roiser, J. P. (2016). Computational psychiatry: towards a mathematically informed understanding of mental illness. *J. Neurol. Neurosurg. Psychiatry* 87, 53–63. doi: 10.1136/jnnp-2015-310737

Baker, R. E., Peña, J. M., Jayamohan, J., and Jérusalem, A. (2018). Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biol. Lett*. 14, 1–4. doi: 10.1098/rsbl.2017.0660

Bougea, A., Efthymiopoulou, E., Spanou, I., and Zikos, P. (2021). A novel machine learning algorithm predicts dementia with lewy bodies versus Parkinson's disease dementia based on clinical and neuropsychological scores. *J. Geriatr. Psychiatry Neurol.* 35, 4–7. doi: 10.1177/0891988721993556

Collerton, D., Perry, E., and McKeith, I. (2005). Why people see things that are not there: a novel perception and attention deficit model for recurrent complex visual hallucinations. *Behav. Brain Sci.* 28, 737–757. doi: 10.1017/S0140525X05000130

Cui, R., and Liu, M. (2019). RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. *Comput. Med. Imag. Graph* 73, 1–10. doi: 10.1016/j.compmedimag.2019.01.005

Danso, S. O., Zeng, Z., Muniz-Terrera, G., and Ritchie, C. W. (2021). Developing an explainable machine learning-based personalised dementia risk prediction model: a transfer learning approach with ensemble learning algorithms. *Front. Big Data* 4, 1–14. doi: 10.3389/fdata.2021.613047

Dauwan, M., van der Zande, J. J., van Dellen, E., Sommer, I. E. C., Scheltens, P., Lemstra, A. W., et al. (2016). Random forest to differentiate dementia with Lewy bodies from Alzheimer's disease. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* 4, 99–106. doi: 10.1016/j.dadm.2016.07.003

Dayan, P., and Abbott, L. F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. Computational Neuroscience Series*. Cambridge, MA: MIT Press.

Friston, K. J., Litvak, V., Oswal, A., Razi, A., Stephan, K. E., Van Wijk, B. C. M., et al. (2016). Bayesian model reduction and empirical Bayes for group (DCM) studies. *Neuroimage* 128, 413–431. doi: 10.1016/j.neuroimage.2015.11.015

Goecks, J., Jalili, V., Heiser, L. M., and Gray, J. W. (2020). How machine learning will transform biomedicine. *Cell* 181, 92–101. doi: 10.1016/j.cell.2020.03.022

Guerchet, M., Prina, M., and Prince, M. (2013). "Policy brief for heads of government: the global impact of dementia 2013–2050," in *Policy Br Heads Gov Glob Impact Dement 2013–2050* (London: Alzheimer's Dis Int (ADI)), 1–8.

Hitchcock, P. F., Fried, E. I., and Frank, M. J. (2022). Computational psychiatry needs time and context. *Annu. Rev. Psychol.* 73, 243–270. doi: 10.1146/annurev-psych-021621-124910

Horn, D., Ruppin, E., Usher, M., and Herrmann, M. (1993). Neural network modeling of memory deterioration in Alzheimer's disease. *Neural Comput.* 5, 736–749. doi: 10.1162/neco.1993.5.5.736

Huys, Q. J. M., Maia, T V., and Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* 19, 404–413. doi: 10.1038/nn.4238

Kuan, V., Fraser, H. C., Hingorani, M., Denaxas, S., Gonzalez-Izquierdo, A., Direk, K., et al. (2021). Data-driven identification of ageing-related diseases from electronic health records. *Sci. Rep.* 11, 1–17. doi: 10.1038/s41598-021-82459-y

Lever, J., Krzywinski, M., and Altman, N. (2016). Points of significance: model selection and overfitting. *Nat. Methods* 13, 703–704. doi: 10.1038/nmeth.3968

Maia, T. V., and Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nat. Neurosci.* 14, 154–162. doi: 10.1038/nn.2723

Makin, S. M., Redman, J., Mosimann, U. P., Dudley, R., Clarke, M. P., Colbourn, C., et al. (2013). Complex visual hallucinations and attentional performance in eye disease and dementia: a test of the perception and attention deficit model. *Int. J. Geriatr. Psychiatry* 28, 1232–1238. doi: 10.1002/gps.3947

Marreiros, A. C., Cagnan, H., Moran, R. J., Friston, K. J., and Brown, P. (2013). Basal ganglia-cortical interactions in Parkinsonian patients. *Neuroimage* 66, 301–310. doi: 10.1016/j.neuroimage.2012.10.088

Mavritsaki, E., Bowman, H., and Su, L. (2019). "Attentional deficits in Alzheimer's disease: investigating the role of acetylcholine with computational modelling," in *Multiscale Models of Brain Disorders. Springer Series in Cognitive and Neural Systems, Vol 13*, ed V. Cutsuridis (Cham: Springer), 13–126. doi: 10.1007/978-3-030-18830-6_11

Meehan, A., Stephanie, L. J., Fazel, S., Fusar-Poli, P., Steyerberg, E. W., Stahl, D., et al. (2022). Clinical prediction models in psychiatry: a systematic review of progress and limitations to date. *Mol. Psychiatry.* 1, 1–9. doi: 10.1038/s41380-022-01528-4

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2017). Deep learning for healthcare: Review, opportunities and challenges. *Brief Bioinform.* 19, 1236–1246. doi: 10.1093/bib/bbx044

Montague, P. R., Dolan, R. J., Friston, K. J., and Dayan, P. (2012). Computational psychiatry. *Trends Cogn. Sci.* 16, 72–80. doi: 10.1016/j.tics.2011.11.018

Moran, R. J., Symmonds, M., Stephan, K. E., Friston, K. J., and Dolan, R. J. (2011). An *in vivo* assay of synaptic function mediating human cognition. *Curr. Biol.* 21, 1320–1325. doi: 10.1016/j.cub.2011.06.053

Oxtoby, N. P., Young, A. L., Cash, D. M., Benzinger, T. L. S., Fagan, A. M., Morris, J. C., et al. (2018). Data-driven models of dominantly-inherited Alzheimer's disease progression. *Brain* 141, 1529–1544. doi: 10.1093/brain/awy050

Paulus, M. P., Huys, Q. J. M., and Maia, T. V. (2016). A roadmap for the development of applied computational psychiatrymodels for better diagnosis, prognosis and treatment. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 1, 1–22. doi: 10.1016/j.bpsc.2016.05.001

Pellegrini, E., Ballerini, L., Hernandez M del, C. V., Chappell, F. M., González-Castro, V., Anblagan, D., et al. (2018). Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* 10, 519–535. doi: 10.1016/j.dadm.2018.07.004

Pinaya, W. H. L., Scarpazza, C., Dias, R. G., Vieira, S., Baecker, L., Costa, P. F., et al. (2021). Using normative modelling to detect disease progression in mild cognitive impairment and Alzheimer ' s disease in a cross - sectional multi - cohort study. *Sci. Rep.* 11, 1–13. doi: 10.1038/s41598-021-95098-0

Raj, A., and Itturia-Medina, Y. (2019). Editorial: network spread models of neurodegenerative diseases Ashish. *Front. Neurol.* 9, 1159. doi: 10.3389/fneur.2018.01159

Ruppin, E., and Reggi, J. A. (1995). A neural model of memory impairment in diffuse cerebral atrophy. *Br. J. Psychiatry* 166, 19–28. doi: 10.1192/bjp.166.1.19

Siuly, S., and Zhang, Y. (2016). Medical big data: neurological diseases diagnosis through medical data analysis. *Data Sci. Eng.* 1, 54–64. doi: 10.1007/s41019-016-0011-3

Skolariki, K., Terrera, G. M., and Danso, S. O. (2021). Predictive models for mild cognitive impairment to Alzheimer's disease conversion. *Neural Regen. Res.* 16, 1766–1767. doi: 10.4103/1673-5374.306071

Su, L., Huang, Y., Wang, Y., Rowe, J., and O'Brien, J. (2018). Predict disease progression with reaction rate equation modeling of multimodal MRI and PET. *Front. Aging Neurosci.* 10, 1–5. doi: 10.3389/fnagi.2018.00306

Su, L., Surendranathan, A., Huang, Y., Bevan-Jones, W. R., Passamonti, L., Hong, Y. T., et al. (2021). Relationship between tau, neuroinflammation and atrophy in Alzheimer's disease: the NIMROD study. *Inf. Fus.* 67, 116–124. doi: 10.1016/j.inffus.2020.10.006

Ten Kate, M., Ingala, S., Schwarz, A. J., Fox, N. C., Chételat, G., Van Berckel, B. N. M., et al. (2018). Secondary prevention of Alzheimer's dementia: Neuroimaging contributions. *Alzheimers Res. Ther.* 10, 1–21. doi: 10.1186/s13195-018-0438-z

Weakley, A., Williams, J. A., Schmitter-Edgecombe, M., and Cook, D. J. (2015). Neuropsychological test selection for cognitive impairment classification: a machine learning approach. *J. Clin. Exp. Neuropsychol.* 37, 899–916. doi: 10.1080/13803395.2015.1067290

Wong-Lin, K. F., McClean, P. L., McCombe, N., Kaur, D., Sanchez-Bornot, J. M., Gillespie, P., et al. (2020). Shaping a data-driven era in dementia care pathway through computational neurology approaches. *BMC Med.* 18, 1–10. doi: 10.1186/s12916-020-01841-1

Young, A. L., Oxtoby, N. P., Daga, P., Cash, D. M., Fox, N. C., Ourselin, S., et al. (2014). A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain* 137, 2564–2577. doi: 10.1093/brain/awu176

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.

frontiers | Frontiers in Computational Neuroscience

REVIEW
published: 26 May 2022
doi: 10.3389/fncom.2022.866517

Whole-Brain Network Models: From Physics to Bedside

Anagh Pathak[1], *Dipanjan Roy*[2] *and Arpan Banerjee*[1]*

[1] National Brain Research Centre, Gurgaon, India, [2] Centre for Brain Science and Applications, School of Artificial Intelligence and Data Science, Indian Institute of Technology, Jodhpur, India

Computational neuroscience has come a long way from its humble origins in the pioneering work of Hodgkin and Huxley. Contemporary computational models of the brain span multiple spatiotemporal scales, from single neuronal compartments to models of social cognition. Each spatial scale comes with its own unique set of promises and challenges. Here, we review models of large-scale neural communication facilitated by white matter tracts, also known as whole-brain models (WBMs). Whole-brain approaches employ inputs from neuroimaging data and insights from graph theory and non-linear systems theory to model brain-wide dynamics. Over the years, WBM models have shown promise in providing predictive insights into various facets of neuropathologies such as Alzheimer's disease, Schizophrenia, Epilepsy, Traumatic brain injury, while also offering mechanistic insights into large-scale cortical communication. First, we briefly trace the history of WBMs, leading up to the state-of-the-art. We discuss various methodological considerations for implementing a whole-brain modeling pipeline, such as choice of node dynamics, model fitting and appropriate parcellations. We then demonstrate the applicability of WBMs toward understanding various neuropathologies. We conclude by discussing ways of augmenting the biological and clinical validity of whole-brain models.

Keywords: whole brain model, neural mass, neural field, network, neuroimaging, DTI, connectome

OPEN ACCESS

Edited by:
Si Wu,
Peking University, China

Reviewed by:
Thomas Bolton,
Advanced Telecommunications Research Institute International (ATR), Japan
Daqing Guo,
University of Electronic Science and Technology of China, China

***Correspondence:**
Arpan Banerjee
arpan@nbrc.ac.in

Received: *31 January 2022*
Accepted: *02 May 2022*
Published: *26 May 2022*

Citation:
Pathak A, Roy D and Banerjee A (2022) Whole-Brain Network Models: From Physics to Bedside. Front. Comput. Neurosci. 16:866517. doi: 10.3389/fncom.2022.866517

PHYSICAL MODELS OF THE BRAIN

Billions of years of evolution have invested the nervous system with tremendous complexity. Modern neuroscience has sought to understand this complexity as a hierarchical ladder that spans multiple spatial and temporal scales, starting from the interaction of biomolecules through to more complex structures like neurons and neural networks. Building on the pioneering work of Hodgkin and Huxley, significant progress took place toward the elucidation of the function of the single neuron. However, in spite of all the remarkable achievements at the single neuron level, relatively little is known about how populations of neurons coordinate with one another to facilitate cognitive processes. While it is fair to characterize neurons, or even individual dendrites as the canonical units of computation in the brain, it is evident that complex cognitive processes rely on interactions between several neural ensembles (McIntosh, 2004; Bressler and Tognoli, 2006)[1] that are distributed across the cortex (Deco et al., 2008). Gaining an understanding of neural circuitry assumes vital importance not only for explaining cognition, but also for the treatment of various neurological diseases.

[1] Several thousand neurons exhibiting coordinated firing patterns.

Early efforts toward bridging the gap between single-neuron activity and circuit operation led to the formulation of neural mass models (Beurle, 1956; Wilson and Cowan, 1972) which conceptualized cortical activity as arising from the dynamic interplay of multiple neural populations (or masses) with excitatory-inhibitory feedback (**Figure 1**). Such models leverage the fact that while the spiking of individual neurons is highly irregular (even chaotic), the mean activity of neural ensembles obeys fairly low-dimensional dynamics (Deco et al., 2008) [2]. Furthermore, mean-field descriptions of cortical tissue may be extended in space and endowed with spatial gradients in neural parameters that follow mathematically defined connectivity (Amari, 1977). Field theory, with its deep roots in physics, provides analytically tractable solutions and has been employed extensively in neuroscience to explain wide-ranging phenomena. A classic field-theoretic model is the one proposed by Amari, which considered lateral-inhibition to explain oscillatory waves and input-evoked transients (Amari, 1977). Two-dimensional field models support diverse phenomena such as spiral and target waves that are organized into complex checkerboard patterns, reminiscent of neural activity observed during different brain states (Ermentrout and Cowan, 1979; Jirsa and Haken, 1996).

By the turn of the century, neuroimaging modalities like PET and fMRI were being increasingly used to study cognition. The abstract nature of existing large-scale models made it difficult to exploit the rich datasets which were being churned in such experiments (Tagamets and Horwitz, 1998; Horwitz et al., 1999, 2000). Additionally, despite their success in providing theoretical accounts for neural phenomena such as traveling waves (Amari, 1977) or resting-state dynamics, Robinson et al. (2021) continuum field models had limited applicability in the clinical setting since crucial medical observables such as anatomical connectivity, functional correlations between brain areas or distribution of various cell types cannot be expressed in terms of mathematical expressions which can then be analytically solved within the field-theoretic framework. Therefore, it was deemed desirable to setup the neurodynamic model so that patient-specific neuroimaging data (e.g., DTI, fMRI connectivity) could be fused with simulations in order to facilitate precision medicine (Ritter et al., 2013; Deco and Kringelbach, 2014).

Within this framework, anatomical connectivity derived from diffusion MRI is used as a structural scaffold to simulate mesoscopic neural interactions (Horwitz et al., 2000; Honey et al., 2009). Nodes, representing mean-field activity of individual brain areas, evolve according to differential equations under the influence of coupling from other brain regions, external input and noise (Deco et al., 2009). Parameters representing biological or phenomenological properties of the nodes and edges are systematically varied, and time-series obtained for each run. For fMRI data, a further hemodynamic convolution is applied to the time-series and functional correlations (FC) are estimated from the resulting data (Deco and Kringelbach, 2014). Alternatively, for EEG/MEG studies, FC is estimated from amplitude envelopes that are extracted for each frequency band of interest and downsampled to correspond with BOLD time-scales (Hipp et al., 2012). Model fitting techniques are then utilized to obtain working points and regimes that best capture the corresponding empirical data. After model fitting, the researcher can ask how this system responds to various perturbations like external inputs (e.g., stimuli), noise or structural insults (e.g., lesions) (**Figure 2**).

Whole-brain models provide actionable insights into various neurological deficits (e.g., identifying optimal resection zone in epilepsy), while also retaining a link to fundamental dynamical and graph theoretic concepts like attractors, metastability, stochastic dynamics, chaos and modularity (Popovych et al., 2019). In the following we outline the major variables that need to be considered before establishing a successful WBM pipeline.

MODELING CONSIDERATIONS

Whole-brain modeling has tremendous clinical applicability as it provides prognostic tools and predictive insights for a host of neurological diseases (Deco and Kringelbach, 2014). However, since not all brain pathologies have the same origin or mechanism, the models seeking to understand them are also customized according to the specific etiology of the disease (**Table 1**). Following E.P Box's adage- "all models are wrong, but some are useful," system equations are set up keeping in mind the specific properties of the underlying clinical context at the expense of biological realism. For example, it may be unnecessary to include conduction delays in models seeking to fit fMRI data due to the widely differing time-scales between BOLD activity (seconds) and axonal propagation (milliseconds). On the other hand, delays assume vital importance when the object of study involves electrophysiological spectral coherence between neural oscillators. Broadly, establishing an effective whole- brain modeling pipeline essentially comes down to the following choices- parcellation scheme, node dynamics, model fitting technique and type of perturbation applied.

Structural Connectivity Matrices and the Role for Parcellation

Firstly, thousands of voxels are reduced to only a few relevant areas of interest. Diffusion imaging is performed to extract anatomical connectivity matrices (**Box 1**). Connectivity matrices specify fiber density across various white matter tracts. A crucial decision at this stage is the choice of a parcellation scheme for obtaining an adjacency matrix. In the absence of a general consensus on what constitutes a "good" parcellation, one must consider carefully how the parcellation scheme may affect the WBM pipeline. Parcellation dictates the spatial resolution and topology of the model. Topological properties are known to be affected by the spatial scale of the parcellation used (Zalesky et al., 2010). Zalesky et al. (2010) demonstrate that while the basic properties of network topology such as scale-freeness or small-worldness remain invariant across spatial scales, the extent of these properties significantly varies between parcellations.

Abbreviations: MRI, Magnetic resonance imaging; WBM, Whole-Brain Model; ROI, Region of Interest; BOLD, Blood oxygen level dependent; EEG/MEG, Electro/Magneto encephalography; FC, Functional connectivity.

[2] In mean field models, coarse-grained variables representing ensemble activity such as the population firing rate are used to track the evolution of the system.

Modeling has further corroborated that significant variability exists in graph-theoretical attributes of network dynamics as a function of the parcellation scheme (Domhof et al., 2021). Significant inter-parcellation variability also exists in resting-state dynamical models (Fornito et al., 2010). Additionally, since the computation time for whole-brain models scales with the number of coupled differential equations (same as the nodes in the network), finer-grained parcellations may be computationally cumbersome to solve. Further, diffusion MRI (dMRI) techniques are biased against short-range and intracortical connectivity, which may have significant ramifications for simulated dynamics (Proix et al., 2016). Highly granular parcellation schema may also lead to redundancy and rank-deficiency during source reconstruction (Tait et al., 2021b).

Broadly, atlases bin brain areas on the basis of either anatomical or functional similarities. Commonly used anatomical criteria for parcellating brain regions include gross anatomy, cytoarchitecture, myeloarchitecture, chemoarchitecture and gene expression profiles (Nowinski, 2021). By contrast, functional atlases utilize resting state or task-related functional correlations to allocate ROIs (Craddock et al., 2012; James et al., 2016). Anatomical atlases are known to fare poorly in comparison to functional atlases when it comes to reproducing FC patterns at the voxel scale (Craddock et al., 2012). Since functional homogeneity is a crucial precondition for modeling ROI dynamics, this would argue for the superiority of functional over anatomical parcellations for whole-brain modeling (Craddock et al., 2012). On the other hand, with anatomically defined ROIs it is easier to interpret results in the light of extant neuroscience literature. Therefore, multimodal atlases which integrate anatomical and functional criteria may offer a suitable tradeoff to ensure functional homogeneity while retaining anatomical specificity in whole-brain analysis (Glasser et al., 2016).

Ultimately, the scope of the study dictates the choice of parcellation. For example, it may be crucial to include sub-cortical nodes where the primary pathology may involve subcortical structures like the thalamus (Ji et al., 2016; Bazin et al., 2020).

BOX 1 | Estimating anatomical connectivity

In-vivo estimation of white matter structural connectivity is enabled by diffusion magnetic resonance imaging (dMRI). Broadly, dMRI approaches measure the preferential direction of diffusion of water molecules in brain tissue. Computational algorithms estimate fiber orientations (streamlines) from dMRI data using a process known as tractography. Streamlines are counted and averaged according to pre-defined brain parcellations to yield adjacency graphs, which can then be submitted as input for whole-brain models. Computational libraries that perform tractography include FSL (Jenkinson et al., 2012), MRtrix (Tournier et al., 2012), BrainSUITE (Shattuck and Leahy, 2002), and DSI studio (Yeh et al., 2013). Considerable variability may exist in the output of different libraries due to differences in the choice of diffusion models, model parameters and tractographic algorithms. Thus, optimal algorithm selection remains an active area of research (Bastiani et al., 2012; Zhan et al., 2015; Petrov et al., 2017).

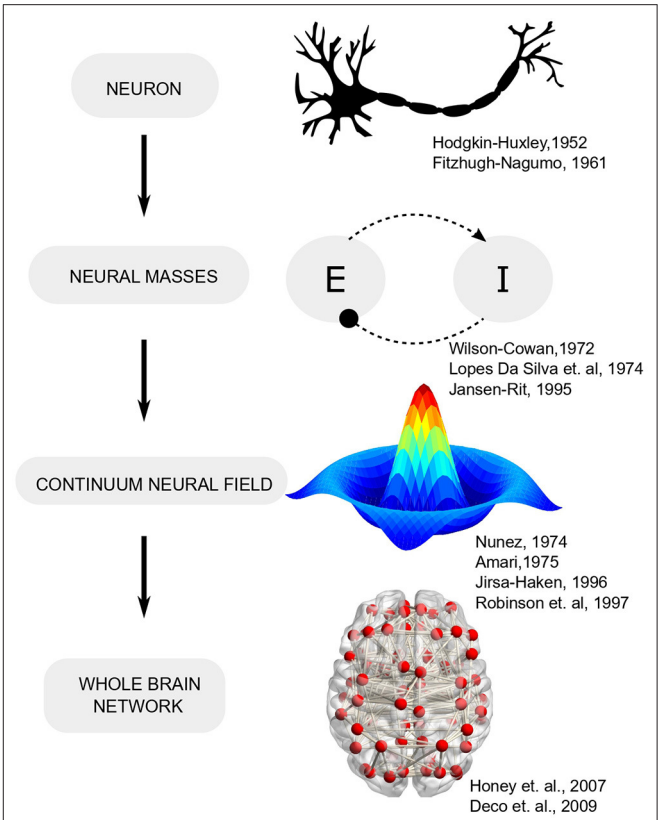


FIGURE 1 | Evolution of computational neuroscience models from single neurons to network models.

Node Dynamics

Node dynamics consist of differential equations specifying the temporal evolution of the population activity of each region of interest (ROI). Each anatomically defined node may potentially consist of thousands of neurons and therefore, the dynamics of the ensemble is reduced to a low-dimensional description using mean-field formalisms. For example, Deco et al. reduce a spiking neuron model with synaptic conductance to yield a dynamic mean-field model that is subsequently used to specify node dynamics (Deco et al., 2013; Roy et al., 2014). Examples of node dynamics may range from the simple phenomenological ones, such as the Kuramoto model (Breakspear et al., 2010) or the normal form of the supercritical Hopf bifurcation (Lord et al., 2017) to the more biologically inspired ones such as the Wilson-Cowan model (Wilson and Cowan, 1972) or thalamocortical motifs (Griffiths et al., 2020) (see **Figure 3** and **Table 1**). For example, the Kuramoto model reduces node dynamics to a phase variable, which evolves according to a natural frequency and a sinusoidal interaction term (Breakspear et al., 2010). On the other hand, both asynchronous and synchronous dynamics can be captured in the same set of equations in bifurcation models that can possess a relaxation solution (damped oscillations) or limit cycle solution (self-sustained oscillations) depending on the value of the bifurcation parameter (**Figure 3**) (Lord et al., 2017). Most computational studies model average functional connectivity,

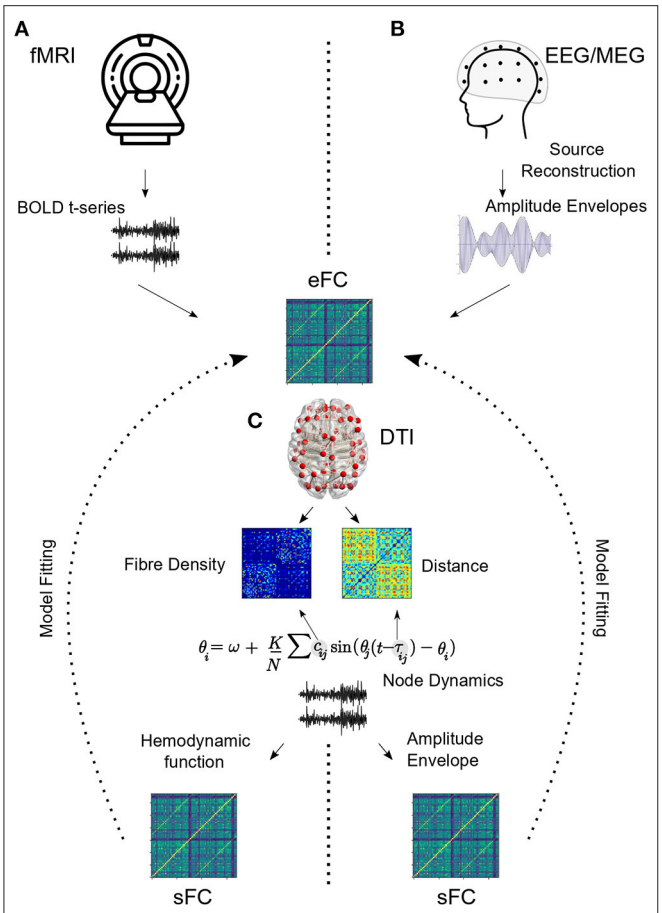


FIGURE 2 | Model pipeline. **(A)** Inter-regional functional correlations (empirical functional connectivity or eFC) are extracted from resting state BOLD time-series. **(B)** For EEG/MEG, raw electrode/sensor time series are used to estimate source level activity, which is then bandpass filtered and hilbert transformed to obtain amplitude envelopes. Band specific functional connectivity is estimated from amplitude envelopes. DTI connectivity provides information about inter-regional white matter fiber density. Euclidean distances may be scaled to obtain temporal delay information. **(C)** Kuramoto model is used here to demonstrate how DTI information is incorporated into the node dynamics. For fMRI, an additional hemodynamic convolution is performed (Balloon-Windkessel) to project model output to BOLD time scales before estimating simulated FC (sFC). For EEG/MEG, amplitude envelopes are extracted and simulated FC is estimated. Model fitting procedures estimate model parameters based on eFC and sFC.

however, brain dynamics is also marked by transitions in the patterns of functional connectivity with time. Switching between FC configurations (FC state) requires node dynamics to possess multistable solutions which may be imparted through the addition of non-linear terms in model equations (Deco et al., 2013; Hansen et al., 2015).

Heterogeneity in node dynamics may be introduced by assigning multiple oscillatory frequencies. For example, Deco et al. (2017) show that models utilizing multiple natural frequencies confirm better to empirical rsMEG networks. Similarly, Roberts et al. devise a principled approach for natural frequency allocation by scaling frequencies by the topological degree of each node (Gollo et al., 2017). For non-oscillatory node dynamics, temporal heterogeneity may be introduced by modulating exponential decay rates or synaptic time constants (**Figure 3**).

Another decision to be made at this step is the inclusion of transmission delays (Nakagawa et al., 2014). Computational models have demonstrated the value of including transmission delays, particularly in explaining oscillatory activity at electrophysiological time scales (Banerjee and Jirsa, 2007; Deco et al., 2009). Neural delays are known to play a crucial role in motor control, particularly in bimanual coordination (Banerjee and Jirsa, 2007) and in explaining perceptual variability in multi-sensory integration (Thakur et al., 2016). Conduction delays, on the order of a few milliseconds, can flip the phase relationship between two gamma oscillators from in-phase to out-of-phase (Pajevic et al., 2014). Network delays crucially dictate oscillation frequency (Niebur et al., 1991; Petkoski and Jirsa, 2019; Pathak et al., 2021) and propagation of cortical traveling waves (Ermentrout and Kleinfeld, 2001). Delays can even cause the complete cessation of self-sustained oscillations (amplitude death) in networks of coupled limit-cycle oscillators (Reddy et al., 1998). On the other hand, under certain conditions, time delays may also enhance neural synchrony (Dhamala et al., 2004).

Conduction delays may be estimated by scaling cortico-cortical tract lengths by conduction velocity, which is usually parametrically varied between 1-30 m/s, in accordance with experimental studies (Swadlow, 1982). However, given the millisecond scale of delays involved, it may be redundant to include delays in cases where the object of interest is fMRI BOLD time scales.

Model Fitting

Typically, whole-brain models aim to explain data collected at fMRI BOLD or electrophysiological (EEG, MEG, sEEG) time scales. Functional time series collected from fMRI experiments are used to estimate inter-areal functional connectivity. For EEG or MEG, pre-processed signals are bandpass filtered in various frequency bands of interest and Hilbert-transformed to extract amplitude envelopes, which are then used to estimate functional connectivity (Hipp et al., 2012; Deco et al., 2017). For both fMRI and EEG/MEG, typically static correlations (presuming stationarity) are employed to estimate model fit. However, recent work has strongly argued that static measures fail to capture the rich, higher-order dynamics inherent in neuroimaging data, and therefore, have advocated the use of dynamic measures of functional connectivity (dFC) (Hutchison et al., 2013; Preti et al., 2017). dFC may be estimated through a windowed manner, or through techniques not requiring arbitrarily chosen temporal windows (Cabral et al., 2017). For dFC analysis, every time step (or time window) has a characteristic FC pattern associated with it. One way to perform model fitting for dFC is by collapsing this 3D data structure (ROI*ROI*time) to a 2D matrix (time*time) consisting of correlations between the leading eigenvectors at each time point or window; the resulting matrix may be considered as the object of model fitting (Cabral et al., 2017).

TABLE 1 | List of studies employing WBMs to understand neuropathologies.

References	Clinical context	Node dynamics	Model fitted to	Parcellation(N)	FC (dynamic or static)
Alstott et al. (2009)	Lesion	Neural mass model	BOLD FC	Hagmann (998)	Static
Demirtaş et al. (2017)	AD	Hopf Normal Form (Stuart-Landau)	BOLD FC	78 Cortical	Static
Vattikonda et al. (2016)	Stroke	Dynamic Mean Field (DMF)	BOLD FC	Desikan Killainy (68), Hagmann (998)	Static
Jirsa et al. (2017)	Epilepsy	Epileptor	SEEG spectral power	EZ/PZ	Static
Nakagawa et al. (2014)	Aging	Dynamic Mean Field (DMF)	BOLD FC	Modified CoCoMac (74)	Static
Deco et al. (2017)	Psychadelics	Dynamic Mean Field (DMF)	BOLD FC	Automatic Anatomical Labelling (90)	Dynamic
Griffiths et al. (2020)	Stimulation	Thalamocortical Motif	AEC MEG	Lausanne Scale 1 (68)	Static
López-González et al. (2021)	Disorders of Consciousness (DOC)	Hopf Normal Form (Stuart-Landau)	BOLD phase synchrony	Shen (214)	Dynamic
Tait et al. (2021a)	Seizure Propensity in AD	Theta Model	EEG phase locked FC	Brainnetome (40)	Static
Hellyer et al. (2015)	Traumatic Brain Injury	Kuramoto Oscillator	BOLD FC	Desikan-Killainy (68)	Static
Cabral et al. (2013)	Schizophrenia	Linear relaxation process	BOLD FC	AAL (90), Hagmann (66)	Static
Yang et al. (2014)	Schizophrenia	Dynamic Mean Field (DMF)	BOLD FC	Hagmann (66)	Static

Model parameters are systematically varied and simulated FCs (static or dynamic) are estimated for each parametric set. Estimation of the optimal parameter set (often referred to as the dynamic working point of the system), offering closest concordance with empirical FC may be achieved by minimizing an error function or by maximizing correlation between empirical and simulated FCs (Deco and Kringelbach, 2014). Bayesian modeling is often employed to estimate parameters associated with the underlying generative models (Vattikonda et al., 2016; Hashemi et al., 2020). Here, models are initialized with a randomly chosen parameter set; stochastic gradient descent is then used to update model parameters.

Instead of FC, one could alternatively perform model fitting against other empirical features of the data. For example, Jirsa et al. (2017) develop a personalized epileptic brain model by estimating model parameters from the spectral distribution of stereotactic (SEEG) electrodes. Indeed, it is even possible to do away with model fitting altogether when the research question is of a qualitative nature. For example, Mejias and Wang (2022) simulate a large-scale model of primate neo-cortex to elucidate the emergence of distributed attractor states subserving various internal processes. In such studies, explaining the salient aspects of the underlying system takes precedence over precise model fitting (Mišić et al., 2015; Mejias and Wang, 2022).

Perturbation

After successfully fitting the model to relevant empirical data, it is desired to introduce various perturbations to the model in order to understand the fallout of various pathological scenarios (Deco et al., 2015). For example, in the case of stroke or TBI, one would like to induce partial or complete lesions at various network nodes and study the differential contribution of node topology in disease progression (Alstott et al., 2009; Vattikonda et al., 2016). Since the thrust here is to understand recoverability, individual node dynamics can be endowed with plasticity mechanisms that homeostatically regulate firing rates (Vattikonda et al., 2016; Abeysuriya et al., 2018; Páscoa Dos Santos and Verschure, 2021). Similarly, stimulation protocols require providing current input to specific nodes in the network to study network response (Griffiths et al., 2020). Epilepsy models require altering node dynamics such as channel properties or neurotransmitter concentrations to model seizure spread from the seizure onset zone (SOZ) (Jirsa et al., 2017). Levels of consciousness in whole-brain models can be manipulated by adjusting neural gain, say, mediated by subcortical structures (Shine, 2021).

CLINICAL APPLICATIONS

Modeling Seizure Propagation

Epilepsy is marked by the occurrence of frequent seizures, which often spread from an onset zone to other distal areas along white matter tracts. In some cases, this necessitates the surgical resection of epileptogenic tissue. Due to the obvious role of network dynamics and structural topology, epilepsy is particularly well-suited for whole-brain modeling, as described in previous sections (Engel Jr et al., 2013; Taylor et al., 2014; Jirsa et al., 2017). Since surgery carries obvious risks, it is desirable to minimize the extent of resected tissue. Jirsa et al. show that personalized whole-brain modeling can be used to aid medical decision making for optimal surgery

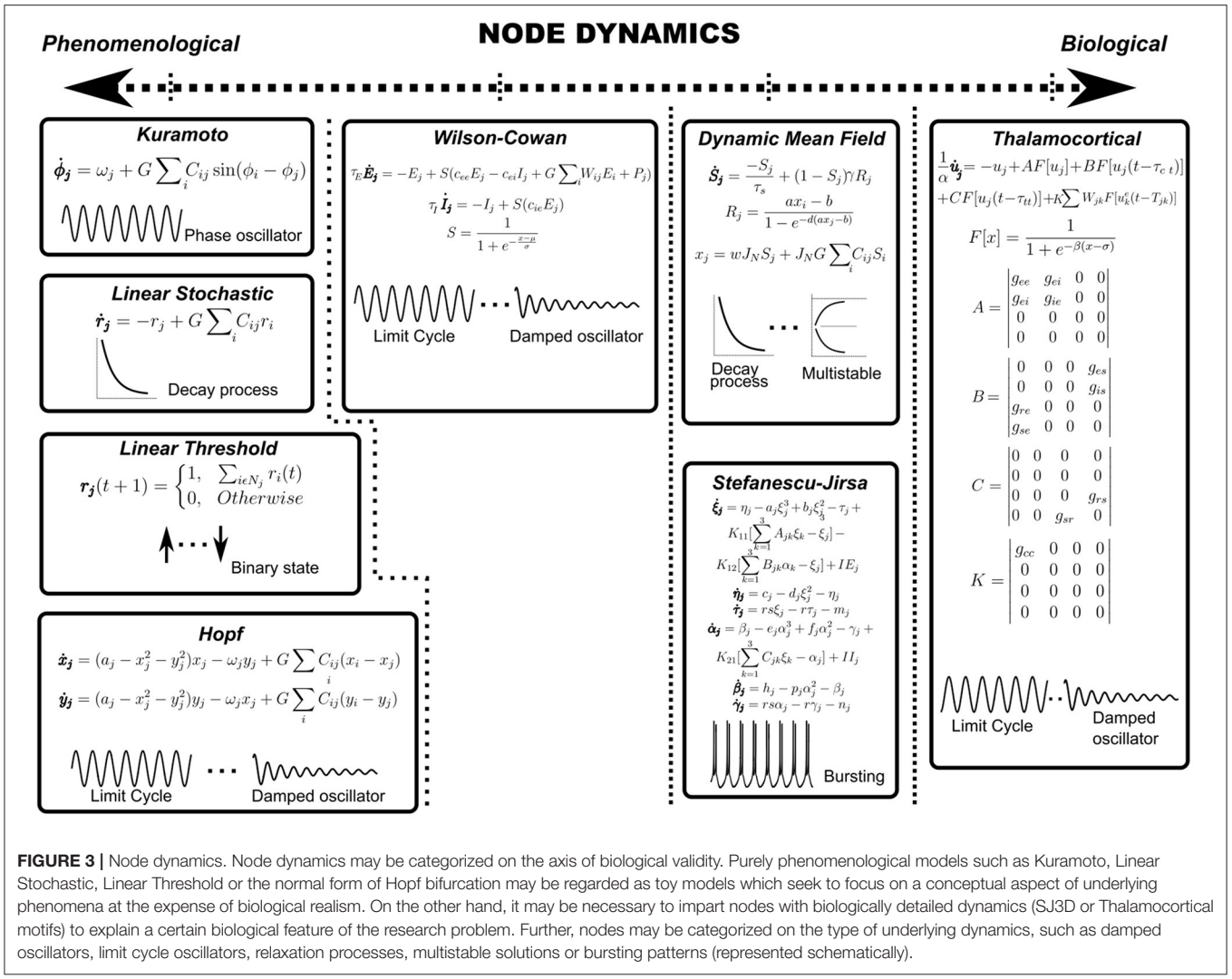


FIGURE 3 | Node dynamics. Node dynamics may be categorized on the axis of biological validity. Purely phenomenological models such as Kuramoto, Linear Stochastic, Linear Threshold or the normal form of Hopf bifurcation may be regarded as toy models which seek to focus on a conceptual aspect of underlying phenomena at the expense of biological realism. On the other hand, it may be necessary to impart nodes with biologically detailed dynamics (SJ3D or Thalamocortical motifs) to explain a certain biological feature of the research problem. Further, nodes may be categorized on the type of underlying dynamics, such as damped oscillators, limit cycle oscillators, relaxation processes, multistable solutions or bursting patterns (represented schematically).

(Vattikonda et al., 2016; Jirsa et al., 2017). Patient-specific brain connectivity is integrated to model empirical EEG data for the identification of the epileptogenic zone (EZ). This technique is particularly useful for instances where conventional methods for EZ identification provide sub-optimal results due to a lack of a clear MRI lesion (Hashemi et al., 2020). Recently, whole-brain modeling has also been used to explain seizures in non-epileptic conditions as well. For example, it is known that patients with Alzheimer's disease are about 6-10 times more likely to develop seizures as compared to the normal population (Pandis and Scarmeas, 2012). Tait et al. (2021a) using a whole-brain pipeline, find that functional connectomes of AD patients show a greater propensity to transition into seizure states as compared to healthy connectomes. Here individual nodes in the network are modeled as phase oscillators capable of producing neuronal spiking in response to inputs. By systematically varying the excitability parameter of individual nodes, the authors show that AD connectomes are more ictogenic as compared to control connectomes for a wide range of excitatory input (Tait et al., 2021a).

Lesions

Neural tissue undergoes lesioning due to various factors like traumatic brain injury, stroke or neurodegenerative diseases (Alstott et al., 2009). Focal lesions can cause disruptions in large-scale functional connectivity, leading to severe cognitive and behavioral impairment. Alstott et al. (2009) demonstrate that the extent and severity of functional deterioration depends on the topological profile of the lesioned nodes, with nodes occupying the most central position causing the greatest network deficit upon lesioning. Vattikonda et al. (2016) extend this idea to gauge potential recoverability from stroke induced lesions by endowing node dynamics with an inhibitory plasticity mechanism that can rescue neural firing rates in response to structural insult (**Figure 4**). Recently, Good et al. used whole-brain modeling to predict the chronic outcomes following traumatic brain injury. Their approach, which utilizes the Virtual Brain simulation platform (Sanz Leon et al., 2013), is able to distinguish semiacute mild to moderate TBI patients from a control group (Good et al., 2022). The effect of lesions on segregative and integrative tendencies can be quantified using

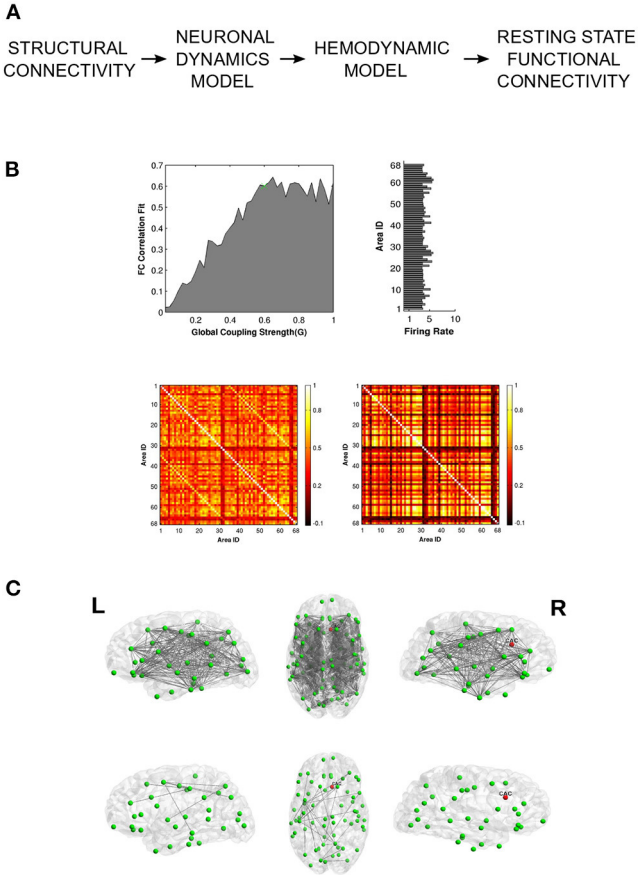


FIGURE 4 | Vattikonda and colleagues use a whole-brain pipeline to elucidate network recovery after lesional insult. **(A)** DTI structural connectivity is used in conjunction with mean field node dynamics (with feedback inhibition) to generate ROI time series which are then convolved with hemodynamic function to obtain simulated BOLD time-series. Resting state FC is calculated using pearson correlation. **(B)** Global working point of the system is estimated by comparing simulated FC with empirical FC. Histogram shows the ROI-wise average firing rate for optimal global coupling. Heat Maps show empirical and optimal simulated FC (averaged across subjects). **(C)** Network recovery takes place due to local plasticity implemented through feedback inhibition. Here lesion was introduced in the right caudal anterior cingulate node. The top 3 figures display connections that have significantly changed before re-establishing local E-I balance. The bottom 3 figures represent the connections that have significantly changed after re-establishing local E-I balance. Adapted from Vattikonda et al. (2016), with permission from the authors.

WBMs. For example, Hellyer et al. (2015) estimate metastability- a measure of segregation and integration, and find disrupted metastable dynamics in patients with traumatic brain injury (TBI). By simulating a network of phase oscillators on topology specified by connectomes obtained from TBI patients, the authors demonstrate how structural disconnection can lead to a reduction in metastable brain dynamics. These observations provide a mechanistic explanation for the significant reductions in cognitive flexibility and information processing, often seen in patients recovering from TBI lesions. Váša et al. (2015) highlight the usefulness of computational lesion studies by demonstrating how graph theoretic properties of network nodes such as modularity determine synchrony and metastability in response to virtual lesioning. The authors find that lesions to nodes with high eigenvector centrality or to nodes which connect segregated modules lead to a decrease in global synchrony along with an increase in global metastability (Váša et al., 2015).

Alzheimer's Disease

Alzheimer's Disease (AD) has traditionally been regarded as a disease of the gray matter, however, recent neuroimaging studies have implicated white matter abnormalities in the pathogenesis of AD (Sachdev et al., 2013). This is reflected in aberrant functional connectivity patterns observed in preclinical populations. Demirtaş et al. (2017) fit a whole-brain model to healthy controls; the model parameters thus obtained are then systematically varied to generate FCs which match empirical FCs seen in preclinical AD, Mild Cognitive Impairment and AD. The authors find that simulated FCs mimic pathological FCs as the individual node dynamics is shifted toward damped oscillations by altering the bifurcation parameter (Demirtaş et al., 2017). Stefanovski et al. (2019) fuse PET-derived Amyloid beta levels with averaged healthy connectomes to shed light on possible pathogenetic mechanims of AD. In this model, Amyloid beta levels modulate the regional Excitation/Inhibition balance, providing a mechanistic explanation for EEG alterations in AD. Further, their whole-brain approach provides therapeutic insights by accounting for large-scale functional reversibility of EEG alterations by modeling the effect of memantine (NMDA receptor antagonist) on local neural populations (Stefanovski et al., 2019). Recently, whole-brain network models have also been utilized for virtual data completion to augment multimodal AD datasets such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (Arbabyazd et al., 2021).

Schizophrenia

The dysconnection hypothesis posits that symptoms of Schizophrenia are best characterized as emerging from functional, rather than anatomical disconnection (Friston, 1998). In line with this assertion, several studies have observed extensive decrease in resting state functional connectivity of patients, pointing to disrupted integration between segregated brain areas (Lynall et al., 2010). Cabral et al. (2013) employ structural connectivity matrices obtained from adolescent patients with early onset schizophrenia and show that functional disruptions associated with Schizophrenia are better explained by reductions in global coupling rather than structural differences, in line with the dysconnection hypothesis. Yang et al. (2014) use whole-brain modeling to show that widely reported differences in global brain signal (GBS) in resting state fMRI of patients may be explained by changes in the net strength of overall brain connectivity in schizophrenia, further corroborating dysconnection. Anticevic et al. (2012) use whole-brain models to identify the role for glutamate in establishing large-scale functional patterns associated with Schizophrenia. Their whole-brain approach, which allows for the introduction of pharmacological manipulations, provides a framework for

understanding the role of NMDA-mediated disruption of cortical excitation/inhibition balance and its role in producing the cognitive symptoms of schizophrenia.

Disorders of Consciousness

Loss of consciousness is either temporary, like in deep sleep or anesthesia- or permanent- like in brain injury or other Disorders of Consciousness (DoC). Efficient classification of brain states as either reversibly or irreversibly unconscious is needed to advance therapeutics. One way to gauge whether a certain unconscious state is transient or permanent is from the response of that state to externally provided perturbation. Recently, whole-brain models have been used to characterize brain states in terms of their stability toward perturbation. Sanz Perl et al. (2021) demonstrate that perturbational analysis can complement machine-learning based algorithms which classify different states of consciousness. López-González et al. (2021) use structural connectivity from healthy and injured subjects to show that low-level states of consciousness are associated with decreased network interactions, leading to segregation of synchronization patterns in fMRI brain dynamics. Segregative tendencies are found to be associated with the global coupling parameter that scales the weights of the SC matrix.

PROMISES AND PITFALLS

Since whole-brain approaches leverage neuroimaging modalities for modeling neural dynamics, future improvement is contingent upon parallel advances in diffusion imaging, functional imaging and signal processing techniques. Here, we discuss a few directions that can significantly augment current neurocomputational models.

One potential avenue for enriching current whole-brain models is by improving the estimation of structural adjacency matrices. For example, DTI derived structural matrices are bidirectional, whereas actual white matter fibers have a well-defined point of origin and termination which imparts directionality and has obvious consequences for the emerging dynamics. Additionally, current protocols for structural estimation rely on the number of streamline (NOS) methods which reconstruct structure by counting the number of streamlines between ROI pairs. Although showing concordance with tract-tracing, the NOS method has inherent limitations as it does not consider other biologically crucial parameters like conduction speeds. Thus, estimation of structural connectivity matrices can be further improved by inclusion of myeloarchitecture, since myelin plays a crucial role in determining conduction speeds across axons (Boshkovski et al., 2021). One way to achieve this is by weighting the connectome with longitudinal relaxation rate (R1), which is sensitive to myelin. Boshkovski et al. (2021) show how including myelin weighted structural connectomes is successful at separating transmodal regions from unimodal regions . Inclusion of myelin in network simulations has particular application at electrophysiological time-scales where phase lags often arise due to finite conduction delays (Petkoski and Jirsa, 2019). g-ratio, which quantifies the ratio between axon diameter and myelin thickness, has recently been shown to be estimable through MRI protocols (Berman et al., 2019; Drakesmith et al., 2019). *In vivo* g-ratio mapping has the potential to provide novel insights into cortical conduction speeds (Berman et al., 2019; Drakesmith et al., 2019). Another method being currently explored for the estimation of cortical conduction velocity uses direct electrical stimulation to measure the propagation of electrophysiological responses across the cortex in patients implanted with intracranial electrodes for seizure monitoring (David, 2021). Harnessing signal propagation information has far-reaching applications, especially toward understanding various demyelinating disorders such as multiple sclerosis.

Further augmentation of whole-brain connectomes comes from incorporating neuromodulatory information (Deco et al., 2018; Kringelbach et al., 2020; Naskar et al., 2021). Multi-modal integration between diffusion imaging (structural connectivity) and PET (receptor density) allows for the infusion of dynamic information to static network models. Kringelbach et al. (2020) have employed a similar pipeline to model the bidirectional interaction of neuronal and neurotransmitter systems that sheds light on the action of psilocybin on human resting state activity. Understanding large-scale functional impact of neuromodulation is of primary importance to computational neuropsychiatry given the therapeutic potential of psychedelics in the treatment of anxiety and depression (Deco et al., 2018).

Another limitation of most current large-scale models is the absence of sub-cortical nodes in the network. This is partly due to inadequate resolution offered by most atlases at the sub-cortical level. Additionally, various sub-cortical structures (e.g., thalamus) possess unique network architecture, requiring the development of specialized node dynamics (see thalamocortical motifs, **Figure 3**). Here we direct the interested reader to some recent efforts toward addressing this lacuna (see Shine et al., 2018; Griffiths et al., 2020; Shine, 2021). Future developments in high field strength imaging, sub-cortical node dynamics and parcellations offer the possibility of having truly whole-brain models.

Despite substantial progress in the field, most successful whole-brain models are limited to either BOLD (fMRI) or BOLD time-scale (amplitude envelopes) functional correlations. Lacunae exist about the extent to which whole-brain models may explain phenomena at electrophysiological time-scales, especially since neural oscillations are so well-linked to the underlying white matter structure and are crucial to cognition (Chu et al., 2015; Hindriks et al., 2015). Signal processing techniques that circumvent or correct for volume/field spread effects which tend to contaminate electrophysiological data would go a long way toward informing whole-brain modeling (Hipp et al., 2012).

Similarly, the present thrust of whole-brain approaches is oriented toward modeling recordings while participants are not engaged in overt cognition, aka resting-state (Biswal et al., 1995; Deco et al., 2011; Popovych et al., 2019). Going forward, whole-brain models could also be explored for explaining various tasks and learning paradigms, requiring richer node dynamics with neuromodulatory and plasticity properties (Abel et al., 2013; Maniglia and Seitz, 2018; Zhang et al., 2021). Finally, foundational discoveries in graph theory and non-equilibrium

physics will continue to offer new insights into the mechanistic underpinnings of large-scale brain dynamics.

CONCLUSION

Computational neuroscience aims to understand the biophysical principles underlying brain function. Many cognitive phenomena crucial for understanding the brain in health and disease evolve at the mesoscopic scale, where the firing patterns of individual neurons get averaged out, thereby offering an opportunity for radical dimensionality reduction. Whole-brain models leverage new advances in neuroimaging techniques to simulate white matter-mediated large-scale brain networks that underlie cognitive and behavioral processes in health and disease. In this article, we provided a brief outline of how coarsely grained models of brain dynamics may be employed to gain insights into the mechanistic underpinnings of brain dynamics, an endeavor central to the emerging field of computational psychiatry. We summarized the various choices at hand for the successful implementation of whole-brain pipelines and discussed those in the context of relevant case studies. Researchers must be mindful of how the choices of parcellation, node dynamics, model fitting procedure and perturbation impact the modeling pipeline and relate to the underlying scientific objective of the study.

We discussed how large-scale modeling has provided crucial insights into the biology of various neuropathologies like Epilepsy, Stroke, Traumatic Brain Injury, Alzheimer's Disease, Schizophrenia and Disorders of Consciousness. Like any emerging field, whole-brain modeling also requires further developments to tap into its full potential and we provided methodological and technical recommendations for the growth of large-scale modeling. Improvements in structural brain imaging and signal processing techniques can significantly enhance the accuracy of neurocomputational models. Similarly, the inclusion of sub-cortical, neurotransmitter and myelination information can lead the field toward truly whole-brain models. Going forward, the continued development of new computational platforms like the Virtual Brain simulator (Sanz Leon et al., 2013) is likely to bridge the gap between theory and implementation, making whole-brain modeling more accessible to medical professionals and biologists alike. In closing, whole-brain models are the newest addition to the rich arsenal of computational neuroscience techniques and promise to usher in a new era in personalized medicine.

AUTHOR CONTRIBUTIONS

AP and AB conceived the study. AP wrote the first draft. AP, DR, and AB wrote and revised manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

DR, Ramalingaswami Fellowship, Department of Biotechnology, Government of India, Award ID: BT/RLF/Re-entry/07/2014. DR, Department of Science and Technology (DST), Ministry of Science and Technology, Government of India, Award ID: SR/CSRI/21/2016. AB, Ministry of Youth Affairs and Sports, Government of India, Award ID: F.NO.K-15015/42/2018/SP-V. AB, NBRC Flagship program, Department of Biotechnology, Government of India, Award ID: BT/MED-III/NBRC/Flagship/Flagship2019.

ACKNOWLEDGMENTS

We acknowledge the generous support of NBRC Core funds.

REFERENCES

Abel, T., Havekes, R., Saletin, J. M., and Walker, M. P. (2013). Sleep, plasticity and memory from molecules to whole-brain networks. *Curr. Biol.* 23, R774–R788. doi: 10.1016/j.cub.2013.07.025

Abeysuriya, R. G., Hadida, J., Sotiropoulos, S. N., Jbabdi, S., Becker, R., Hunt, B. A., et al. (2018). A biophysical model of dynamic balancing of excitation and inhibition in fast oscillatory large-scale networks. *PLoS Comput. Biol.* 14, e1006007. doi: 10.1371/journal.pcbi.1006007

Alstott, J., Breakspear, M., Hagmann, P., Cammoun, L., and Sporns, O. (2009). Modeling the impact of lesions in the human brain. *PLoS Comput. Biol.* 5, e1000408. doi: 10.1371/journal.pcbi.1000408

Amari, S.-I. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol. Cybern.* 27, 77–87. doi: 10.1007/BF00337259

Anticevic, A., Gancsos, M., Murray, J. D., Repovs, G., Driesen, N. R., Ennis, D. J., et al. (2012). NMDA receptor function in large-scale anticorrelated neural systems with implications for cognition and schizophrenia. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16720–16725. doi: 10.1073/pnas.1208494109

Arbabyazd, L., Shen, K., Wang, Z., Hofmann-Apitius, M., Ritter, P., McIntosh, A. R., et al. (2021). Virtual connectomic datasets in alzheimer's disease and aging using whole-brain network dynamics modelling. *Eneuro* 8, ENEURO.0475-20.2021. doi: 10.1523/ENEURO.0475-20.2021

Banerjee, A., and Jirsa, V. K. (2007). How do neural connectivity and time delays influence bimanual coordination? *Biol. Cybern.* 96, 265–278. doi: 10.1007/s00422-006-0114-4

Bastiani, M., Shah, N. J., Goebel, R., and Roebroeck, A. (2012). Human cortical connectome reconstruction from diffusion weighted mri: the effect of tractography algorithm. *Neuroimage* 62, 1732–1749. doi: 10.1016/j.neuroimage.2012.06.002

Bazin, P.-L., Alkemade, A., Mulder, M. J., Henry, A. G., and Forstmann, B. U. (2020). Multi-contrast anatomical subcortical structures parcellation. *Elife* 9, e59430. doi: 10.7554/eLife.59430

Berman, S., Filo, S., and Mezer, A. A. (2019). Modeling conduction delays in the corpus callosum using MRI-measured g-ratio. *Neuroimage* 195, 128–139. doi: 10.1016/j.neuroimage.2019.03.025

Beurle, R. L. (1956). Properties of a mass of cells capable of regenerating pulses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 240, 55–94. doi: 10.1098/rstb.1956.0012

Biswal, B., Zerrin Yetkin, F., Haughton, V. M., and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.* 34, 537–541. doi: 10.1002/mrm.1910340409

Boshkovski, T., Kocarev, L., Cohen-Adad, J., Mišić, B., Lehéricy, S., Stikov, N., et al. (2021). The R1-weighted connectome: complementing brain networks with a myelin-sensitive measure. *Netw. Neurosci.* 5, 358–372. doi: 10.1162/netn_a_00179

Breakspear, M., Heitmann, S., and Daffertshofer, A. (2010). Generative models of cortical oscillations: neurobiological implications of the kuramoto model. *Front. Hum. Neurosci.* 4, 190. doi: 10.3389/fnhum.2010.00190

Bressler, S. L., and Tognoli, E. (2006). Operational principles of neurocognitive networks. *Int. J. Psychophysiol.* 60, 139–148. doi: 10.1016/j.ijpsycho.2005.12.008

Cabral, J., Fernandes, H. M., Van Hartevelt, T. J., James, A. C., Kringelbach, M. L., and Deco, G. (2013). Structural connectivity in schizophrenia and its impact on the dynamics of spontaneous functional networks. *Chaos* 23, 046111. doi: 10.1063/1.4851117

Cabral, J., Vidaurre, D., Marques, P., Magalh aes, R., Moreira, P. S., Soares, J. M., et al. (2017). Cognitive performance in healthy older adults relates to spontaneous switching between states of functional connectivity during rest. *Sci. Rep.* 7, 1–13. doi: 10.1038/s41598-017-05425-7

Chu, C. J., Tanaka, N., Diaz, J., Edlow, B. L., Wu, O., Hämäläinen, M., et al. (2015). EEG functional connectivity is partially predicted by underlying white matter connectivity. *Neuroimage* 108, 23–33. doi: 10.1016/j.neuroimage.2014.12.033

Craddock, R. C., James, G. A., Holtzheimer, III, P. E., Hu, X. P., and Mayberg, H. S. (2012). A whole brain fmri atlas generated via spatially constrained spectral clustering. *Hum. Brain Mapp.* 33, 1914–1928. doi: 10.1002/hbm.21333

David, O. (2021). Functional brain tractography. *Brain Stimul.* 14, 1729. doi: 10.1016/j.brs.2021.10.466

Deco, G., Cabral, J., Woolrich, M. W., Stevner, A. B., Van Hartevelt, T. J., and Kringelbach, M. L. (2017). Single or multiple frequency generators in on-going brain activity: a mechanistic whole-brain model of empirical MEG data. *Neuroimage* 152, 538–550. doi: 10.1016/j.neuroimage.2017.03.023

Deco, G., Cruzat, J., Cabral, J., Knudsen, G. M., Carhart-Harris, R. L., Whybrow, P. C., et al. (2018). Whole-brain multimodal neuroimaging model using serotonin receptor maps explains non-linear functional effects of LSD. *Curr. Biol.* 28, 3065–3074. doi: 10.1016/j.cub.2018.07.083

Deco, G., Jirsa, V., McIntosh, A. R., Sporns, O., and Kötter, R. (2009). Key role of coupling, delay, and noise in resting brain fluctuations. *Proc. Natl. Acad. Sci. U.S.A.* 106, 10302–10307. doi: 10.1073/pnas.0901831106

Deco, G., Jirsa, V. K., and McIntosh, A. R. (2011). Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nat. Rev. Neurosci.* 12, 43–56. doi: 10.1038/nrn2961

Deco, G., Jirsa, V. K., Robinson, P. A., Breakspear, M., and Friston, K. (2008). The dynamic brain: from spiking neurons to neural masses and cortical fields. *PLoS Comput. Biol.* 4, e1000092. doi: 10.1371/journal.pcbi.1000092

Deco, G., and Kringelbach, M. L. (2014). Great expectations: using whole-brain computational connectomics for understanding neuropsychiatric disorders. *Neuron* 84, 892–905. doi: 10.1016/j.neuron.2014.08.034

Deco, G., Ponce-Alvarez, A., Mantini, D., Romani, G. L., Hagmann, P., and Corbetta, M. (2013). Resting-state functional connectivity emerges from structurally and dynamically shaped slow linear fluctuations. *J. Neurosci.* 33, 11239–11252. doi: 10.1523/JNEUROSCI.1091-13.2013

Deco, G., Tononi, G., Boly, M., and Kringelbach, M. L. (2015). Rethinking segregation and integration: contributions of whole-brain modelling. *Nat. Rev. Neurosci.* 16, 430–439. doi: 10.1038/nrn3963

Demirtaş, M., Falcon, C., Tucholka, A., Gispert, J. D., Molinuevo, J. L., and Deco, G. (2017). A whole-brain computational modeling approach to explain the alterations in resting-state functional connectivity during progression of alzheimer's disease. *Neuroimage* 16, 343–354. doi: 10.1016/j.nicl.2017.08.006

Dhamala, M., Jirsa, V. K., and Ding, M. (2004). Enhancement of neural synchrony by time delay. *Phys. Rev. Lett.* 92, 074104. doi: 10.1103/PhysRevLett.92.074104

Domhof, J. W., Jung, K., Eickhoff, S. B., and Popovych, O. V. (2021). Parcellation-induced variation of empirical and simulated brain connectomes at group and subject levels. *Netw. Neurosci.* 5, 798–830. doi: 10.1162/netn_a_00202

Drakesmith, M., Harms, R., Rudrapatna, S. U., Parker, G. D., Evans, C. J., and Jones, D. K. (2019). Estimating axon conduction velocity in vivo from microstructural MRI. *Neuroimage* 203, 116186. doi: 10.1016/j.neuroimage.2019.116186

Engel Jr, J., Thompson, P. M., Stern, J. M., Staba, R. J., Bragin, A., and Mody, I. (2013). Connectomics and epilepsy. *Curr. Opin. Neurol.* 26, 186. doi: 10.1097/WCO.0b013e32835ee5b8

Ermentrout, G. B., and Cowan, J. D. (1979). A mathematical theory of visual hallucination patterns. *Biol. Cybern.* 34, 137–150. doi: 10.1007/BF00336965

Ermentrout, G. B., and Kleinfeld, D. (2001). Traveling electrical waves in cortex: insights from phase dynamics and speculation on a computational role. *Neuron* 29, 33–44. doi: 10.1016/S0896-6273(01)00178-7

Fornito, A., Zalesky, A., and Bullmore, E. T. (2010). Network scaling effects in graph analytic studies of human resting-state fMRI data. *Front. Syst. Neurosci.* 4, 22. doi: 10.3389/fnsys.2010.00022

Friston, K. J. (1998). The disconnection hypothesis. *Schizophr. Res.* 30, 115–125. doi: 10.1016/S0920-9964(97)00140-0

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., et al. (2016). A multi-modal parcellation of human cerebral cortex. *Nature* 536, 171–178. doi: 10.1038/nature18933

Gollo, L. L., Roberts, J. A., and Cocchi, L. (2017). Mapping how local perturbations influence systems-level brain dynamics. *Neuroimage* 160, 97–112. doi: 10.1016/j.neuroimage.2017.01.057

Good, T., Schirner, M., Shen, K., Ritter, P., Mukherjee, P., Levine, B., et al. (2022). Personalized connectome-based modeling in patients with semi-acute phase tbi: Relationship to acute neuroimaging and 6 month follow-up. *Eneuro* 9, ENEURO.0075-21.2022. doi: 10.1523/ENEURO.0075-21.2022

Griffiths, J. D., McIntosh, A. R., and Lefebvre, J. (2020). A connectome-based, corticothalamic model of state-and stimulation-dependent modulation of rhythmic neural activity and connectivity. *Front. Comput. Neurosci.* 14, 575143. doi: 10.3389/fncom.2020.575143

Hansen, E. C., Battaglia, D., Spiegler, A., Deco, G., and Jirsa, V. K. (2015). Functional connectivity dynamics: modeling the switching behavior of the resting state. *Neuroimage* 105, 525–535. doi: 10.1016/j.neuroimage.2014.11.001

Hashemi, M., Vattikonda, A., Sip, V., Guye, M., Bartolomei, F., Woodman, M. M., et al. (2020). The bayesian virtual epileptic patient: a probabilistic framework designed to infer the spatial map of epileptogenicity in a personalized large-scale brain model of epilepsy spread. *Neuroimage* 217, 116839. doi: 10.1016/j.neuroimage.2020.116839

Hellyer, P. J., Scott, G., Shanahan, M., Sharp, D. J., and Leech, R. (2015). Cognitive flexibility through metastable neural dynamics is disrupted by damage to the structural connectome. *J. Neurosci.* 35, 9050–9063. doi: 10.1523/JNEUROSCI.4648-14.2015

Hindriks, R., Woolrich, M., Luckhoo, H., Joensson, M., Mohseni, H., Kringelbach, M. L., et al. (2015). Role of white-matter pathways in coordinating alpha oscillations in resting visual cortex. *Neuroimage* 106, 328–339. doi: 10.1016/j.neuroimage.2014.10.057

Hipp, J. F., Hawellek, D. J., Corbetta, M., Siegel, M., and Engel, A. K. (2012). Large-scale cortical correlation structure of spontaneous oscillatory activity. *Nat. Neurosci.* 15, 884–890. doi: 10.1038/nn.3101

Honey, C. J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J.-P., Meuli, R., et al. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proc. Natl. Acad. Sci. U.S.A.* 106, 2035–2040. doi: 10.1073/pnas.0811168106

Horwitz, B., Friston, K. J., and Taylor, J. G. (2000). Neural modeling and functional brain imaging: an overview. *Neural Netw.* 13/, 829–846. doi: 10.1016/S0893-6080(00)00062-9

Horwitz, B., Tagamets, M., and McIntosh, A. R. (1999). Neural modeling, functional brain imaging, and cognition. *Trends Cogn. Sci.* 3, 91–98. doi: 10.1016/S1364-6613(99)01282-6

Hutchison, R. M., Womelsdorf, T., Allen, E. A., Bandettini, P. A., Calhoun, V. D., Corbetta, M., et al. (2013). Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage* 80, 360–378. doi: 10.1016/j.neuroimage.2013.05.079

James, G. A., Hazaroglu, O., and Bush, K. A. (2016). A human brain atlas derived via n-cut parcellation of resting-state and task-based fmri data. *Magn. Reson. Imaging* 34, 209–218. doi: 10.1016/j.mri.2015.10.036

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). Fsl. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Ji, B., Li, Z., Li, K., Li, L., Langley, J., Shen, H., et al. (2016). Dynamic thalamus parcellation from resting-state fMRI data. *Hum. Brain Mapp.* 37, 954–967. doi: 10.1002/hbm.23079

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., Proix, T., Perdikis, D., Woodman, M. M., Wang, H., Gonzalez-Martinez, J., et al. (2017). The virtual epileptic patient: individualized whole-brain models of epilepsy spread. *Neuroimage* 145, 377–388. doi: 10.1016/j.neuroimage.2016.04.049

Kringelbach, M. L., Cruzat, J., Cabral, J., Knudsen, G. M., Carhart-Harris, R., Whybrow, P. C., et al. (2020). Dynamic coupling of whole-brain neuronal

and neurotransmitter systems. *Proc. Natl. Acad. Sci. U.S.A.* 117, 9566–9576. doi: 10.1073/pnas.1921475117

López-González, A., Panda, R., Ponce-Alvarez, A., Zamora-López, G., Escrichs, A., Martial, C., et al. (2021). Loss of consciousness reduces the stability of brain hubs and the heterogeneity of brain dynamics. *Commun. Biol.* 4, 1–15. doi: 10.1038/s42003-021-02537-9

Lord, L.-D., Stevner, A. B., Deco, G., and Kringelbach, M. L. (2017). Understanding principles of integration and segregation using whole-brain computational connectomics: implications for neuropsychiatric disorders. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 375, 20160283. doi: 10.1098/rsta.2016.0283

Lynall, M.-E., Bassett, D. S., Kerwin, R., McKenna, P. J., Kitzbichler, M., Muller, U., et al. (2010). Functional connectivity and brain networks in schizophrenia. *J. Neurosci.* 30, 9477–9487. doi: 10.1523/JNEUROSCI.0333-10.2010

Maniglia, M., and Seitz, A. R. (2018). Towards a whole brain model of perceptual learning. *Curr. Opin. Behav. Sci.* 20, 47–55. doi: 10.1016/j.cobeha.2017.10.004

McIntosh, A. R. (2004). Contexts and catalysts. *Neuroinformatics* 2, 175–181. doi: 10.1385/NI:2,2:175

Mejias, J. F., and Wang, X.-J. (2022). Mechanisms of distributed working memory in a large-scale network of macaque neocortex. *Elife* 11, e72136. doi: 10.7554/eLife.72136

Mišić, B., Betzel, R. F., Nematzadeh, A., Goni, J., Griffa, A., Hagmann, P., et al. (2015). Cooperative and competitive spreading dynamics on the human connectome. *Neuron* 86, 1518–1529. doi: 10.1016/j.neuron.2015.05.035

Nakagawa, T. T., Woolrich, M., Luckhoo, H., Joensson, M., Mohseni, H., Kringelbach, M. L., et al. (2014). How delays matter in an oscillatory whole-brain spiking-neuron network model for MEG alpha-rhythms at rest. *Neuroimage* 87, 383–394. doi: 10.1016/j.neuroimage.2013.11.009

Naskar, A., Vattikonda, A., Deco, G., Roy, D., and Banerjee, A. (2021). Multi-scale dynamic mean field model (MDMF) relates resting-state brain dynamics with local cortical excitatory-inhibitory neurotransmitter homeostasis. *Netw. Neurosci.* 5, 757–782. doi: 10.1162/netn_a_00197

Niebur, E., Schuster, H. G., and Kammen, D. M. (1991). Collective frequencies and metastability in networks of limit-cycle oscillators with time delay. *Phys. Rev. Lett.* 67, 2753. doi: 10.1103/PhysRevLett.67.2753

Nowinski, W. L. (2021). Evolution of human brain atlases in terms of content, applications, functionality, and availability. *Neuroinformatics* 19, 1–22. doi: 10.1007/s12021-020-09481-9

Pajevic, S., Basser, P. J., and Fields, R. D. (2014). Role of myelin plasticity in oscillations and synchrony of neuronal activity. *Neuroscience* 276, 135–147. doi: 10.1016/j.neuroscience.2013.11.007

Pandis, D., and Scarmeas, N. (2012). Seizures in alzheimer disease: clinical and epidemiological data: Seizures in alzheimer disease. *Epilepsy Curr.* 12, 184–187. doi: 10.5698/1535-7511-12.5.184

Páscoa Dos Santos, F., and Verschure, P. F. (2021). Excitatory-inhibitory homeostasis and diaschisis: tying the local and global scales in the post-stroke cortex. *Front. Syst. Neurosci.* 15, 806544doi: 10.3389/fnsys.2021.806544

Pathak, A., Sharma, V., Roy, D., and Banerjee, A. (2021). Preservation of neural synchrony at peak alpha frequency via global synaptic scaling compensates for white matter structural decline over adult lifespan. *bioRxiv*. doi: 10.1101/2021.10.24.465613

Petkoski, S., and Jirsa, V. K. (2019). Transmission time delays organize the brain network synchronization. *Philos. Trans. R. Soc. A* 377, 20180132. doi: 10.1098/rsta.2018.0132

Petrov, D., Ivanov, A., Faskowitz, J., Gutman, B., Moyer, D., Villalon, J., et al. (2017). "Evaluating 35 methods to generate structural connectomes using pairwise classification," in *International Conference on medical Image Computing and Computer-Assisted Intervention* (Quebec: Springer), 515–522.

Popovych, O. V., Manos, T., Hoffstaedter, F., and Eickhoff, S. B. (2019). What can computational models contribute to neuroimaging data analytics? *Front. Syst. Neurosci.* 12, 68. doi: 10.3389/fnsys.2018.00068

Preti, M. G., Bolton, T. A., and Van De Ville, D. (2017). The dynamic functional connectome: state-of-the-art and perspectives. *Neuroimage* 160, 41–54. doi: 10.1016/j.neuroimage.2016.12.061

Proix, T., Spiegler, A., Schirner, M., Rothmeier, S., Ritter, P., and Jirsa, V. K. (2016). How do parcellation size and short-range connectivity affect dynamics in large-scale brain network models? *Neuroimage* 142, 135–149. doi: 10.1016/j.neuroimage.2016.06.016

Reddy, D. R., Sen, A., and Johnston, G. L. (1998). Time delay induced death in coupled limit cycle oscillators. *Phys. Rev. Lett.* 80, 5109. doi: 10.1103/PhysRevLett.80.5109

Ritter, P., Schirner, M., McIntosh, A. R., and Jirsa, V. K. (2013). The virtual brain integrates computational modeling and multimodal neuroimaging. *Brain Connect* 3, 121–145. doi: 10.1089/brain.2012.0120

Robinson, P. A., Henderson, J. A., Gabay, N. C., Aquino, K. M., Babaie-Janvier, T., and Gao, X. (2021). Determination of dynamic brain connectivity via spectral analysis. *Front. Hum. Neurosci.* 15, 655576. doi: 10.3389/fnhum.2021.655576

Roy, D., Sigala, R., Breakspear, M., McIntosh, A. R., Jirsa, V. K., Deco, G., et al. (2014). Using the virtual brain to reveal the role of oscillations and plasticity in shaping brain's dynamical landscape. *Brain Connect.* 4, 791–811. doi: 10.1089/brain.2014.0252

Sachdev, P. S., Zhuang, L., Braidy, N., and Wen, W. (2013). Is alzheimer's a disease of the white matter? *Curr. Opin Psychiatry* 26, 244–251. doi: 10.1097/YCO.0b013e32835ed6e8

Sanz Leon, P., Knock, S. A., Woodman, M. M., Domide, L., Mersmann, J., McIntosh, A. R., et al. (2013). The virtual brain: a simulator of primate brain network dynamics. *Front. Neuroinform.* 7, 10. doi: 10.3389/fninf.2013.00010

Sanz Perl, Y., Pallavicini, C., Pérez Ipi na, I., Demertzi, A., Bonhomme, V., Martial, C., et al. (2021). Perturbations in dynamical models of whole-brain activity dissociate between the level and stability of consciousness. *PLoS Comput. Biol.* 17, e1009139. doi: 10.1371/journal.pcbi.1009139

Shattuck, D. W., and Leahy, R. M. (2002). Brainsuite: an automated cortical surface identification tool. *Med. Image Anal.* 6, 129–142. doi: 10.1016/S1361-8415(02)00054-3

Shine, J. M. (2021). The thalamus integrates the macrosystems of the brain to facilitate complex, adaptive brain network dynamics. *Progr. Neurobiol.* 199, 101951. doi: 10.1016/j.pneurobio.2020.101951

Shine, J. M., Aburn, M. J., Breakspear, M., and Poldrack, R. A. (2018). The modulation of neural gain facilitates a transition between functional segregation and integration in the brain. *Elife* 7, e31130. doi: 10.7554/eLife.31130

Stefanovski, L., Triebkorn, P., Spiegler, A., Diaz-Cortes, M.-A., Solodkin, A., Jirsa, V., et al. (2019). Linking molecular pathways and large-scale computational modeling to assess candidate disease mechanisms and pharmacodynamics in Alzheimer's disease. *Front. Comput. Neurosci.* 13, 54. doi: 10.3389/fncom.2019.00054

Swadlow, H. A. (1982). Impulse conduction in the mammalian brain: physiological properties of individual axons monitored for several months. *Science* 218, 911–913. doi: 10.1126/science.7134984

Tagamets, M., and Horwitz, B. (1998). Integrating electrophysiological and anatomical experimental data to create a large-scale model that simulates a delayed match-to-sample human brain imaging study. *Cereb. Cortex* 8, 310–320. doi: 10.1093/cercor/8.4.310

Tait, L., Lopes, M. A., Stothart, G., Baker, J., Kazanina, N., Zhang, J., et al. (2021a). A large-scale brain network mechanism for increased seizure propensity in alzheimer's disease. *PLoS Comput. Biol.* 17, 1–21. doi: 10.1101/2021.01.19.427236

Tait, L., Özkan, A., Szul, M. J., and Zhang, J. (2021b). A systematic evaluation of source reconstruction of resting MEG of the human brain with a new high-resolution atlas: Performance, precision, and parcellation. *Hum. Brain Mapp.* 42, 4685–4707. doi: 10.1002/hbm.25578

Taylor, P. N., Kaiser, M., and Dauwels, J. (2014). Structural connectivity based whole brain modelling in epilepsy. *J. Neurosci. Methods* 236, 51–57. doi: 10.1016/j.jneumeth.2014.08.010

Thakur, B., Mukherjee, A., Sen, A., and Banerjee, A. (2016). A dynamical framework to relate perceptual variability with multisensory information processing. *Sci. Rep.* 6, 1–13. doi: 10.1038/srep31280

Tournier, J.-D., Calamante, F., and Connelly, A. (2012). Mrtrix: diffusion tractography in crossing fiber regions. *Int. J. Imaging Syst. Technol.* 22, 53–66. doi: 10.1002/ima.22005

Váša, F., Shanahan, M., Hellyer, P. J., Scott, G., Cabral, J., and Leech, R. (2015). Effects of lesions on synchrony and metastability in cortical networks. *Neuroimage* 118, 456–467. doi: 10.1016/j.neuroimage.2015.05.042

Vattikonda, A., Surampudi, B. R., Banerjee, A., Deco, G., and Roy, D. (2016). Does the regulation of local excitation-inhibition balance aid in recovery of

functional connectivity? a computational account. *Neuroimage* 136, 57–67. doi: 10.1016/j.neuroimage.2016.05.002

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

Yang, G. J., Murray, J. D., Repovs, G., Cole, M. W., Savic, A., Glasser, M. F., et al. (2014). Altered global brain signal in schizophrenia. *Proc. Natl. Acad. Sci. U.S.A.* 111, 7438–7443. doi: 10.1073/pnas.1405289111

Yeh, F.-C., Verstynen, T. D., Wang, Y., Fernández-Miranda, J. C., and Tseng, W.-Y. I. (2013). Deterministic diffusion fiber tracking improved by quantitative anisotropy. *PLoS ONE* 8, e80713. doi: 10.1371/journal.pone.0080713

Zalesky, A., Fornito, A., Harding, I. H., Cocchi, L., Yücel, M., Pantelis, C., et al. (2010). Whole-brain anatomical networks: does the choice of nodes matter? *Neuroimage* 50, 970–983. doi: 10.1016/j.neuroimage.2009.12.027

Zhan, L., Zhou, J., Wang, Y., Jin, Y., Jahanshad, N., Prasad, G., et al. (2015). Comparison of nine tractography algorithms for detecting abnormal structural brain networks in alzheimer's disease. *Front. Aging Neurosci.* 7, 48. doi: 10.3389/fnagi.2015.00048

Zhang, G., Cui, Y., Zhang, Y., Cao, H., Zhou, G., Shu, H., et al. (2021). Computational exploration of dynamic mechanisms of steady state visual evoked potentials at the whole brain level. *Neuroimage* 237, 118166. doi: 10.1016/j.neuroimage.2021.118166

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.

frontiers | Frontiers in Computational Neuroscience

REVIEW
published: 16 June 2022
doi: 10.3389/fncom.2022.859874

Toward Reflective Spiking Neural Networks Exploiting Memristive Devices

Valeri A. Makarov[1,2], Sergey A. Lobov[2,3,4], Sergey Shchanikov[2,5], Alexey Mikhaylov[2] and Viktor B. Kazantsev[2,3,4]*

[1] *Instituto de Matemática Interdisciplinar, Universidad Complutense de Madrid, Madrid, Spain,* [2] *Department of Neurotechnologies, Research Institute of Physics and Technology, Laboratory of Stochastic Multistable Systems, Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia,* [3] *Neuroscience and Cognitive Technology Laboratory, Center for Technologies in Robotics and Mechatronics Components, Innopolis University, Innopolis, Russia,* [4] *Center For Neurotechnology and Machine Learning, Immanuel Kant Baltic Federal University, Kaliningrad, Russia,* [5] *Department of Information Technologies, Vladimir State University, Vladimir, Russia*

The design of modern convolutional artificial neural networks (ANNs) composed of formal neurons copies the architecture of the visual cortex. Signals proceed through a hierarchy, where receptive fields become increasingly more complex and coding sparse. Nowadays, ANNs outperform humans in controlled pattern recognition tasks yet remain far behind in cognition. In part, it happens due to limited knowledge about the higher echelons of the brain hierarchy, where neurons actively generate predictions about what will happen next, i.e., the information processing jumps from reflex to reflection. In this study, we forecast that spiking neural networks (SNNs) can achieve the next qualitative leap. Reflective SNNs may take advantage of their intrinsic dynamics and mimic complex, not reflex-based, brain actions. They also enable a significant reduction in energy consumption. However, the training of SNNs is a challenging problem, strongly limiting their deployment. We then briefly overview new insights provided by the concept of a high-dimensional brain, which has been put forward to explain the potential power of single neurons in higher brain stations and deep SNN layers. Finally, we discuss the prospect of implementing neural networks in memristive systems. Such systems can densely pack on a chip 2D or 3D arrays of plastic synaptic contacts directly processing analog information. Thus, memristive devices are a good candidate for implementing in-memory and in-sensor computing. Then, memristive SNNs can diverge from the development of ANNs and build their niche, cognitive, or reflective computations.

Keywords: spiking neural networks (SNNs), memristors and memristive systems, high-dimensional brain, plasticity, reflective systems

OPEN ACCESS

Edited by:
Misha Tsodyks,
Weizmann Institute of Science, Israel

Reviewed by:
Davide Zambrano,
Swiss Federal Institute of Technology Lausanne, Switzerland
Wenrui Zhang,
University of California, Santa Barbara, United States

****Correspondence:***
Valeri A. Makarov
vmakarov@ucm.es

Received: *21 January 2022*
Accepted: *10 May 2022*
Published: *16 June 2022*

Citation:
Makarov VA, Lobov SA, Shchanikov S, Mikhaylov A and Kazantsev VB (2022) Toward Reflective Spiking Neural Networks Exploiting Memristive Devices. Front. Comput. Neurosci. 16:859874. doi: 10.3389/fncom.2022.859874

INTRODUCTION

Brief History of Artificial Neural Networks

Since the early steps of artificial intelligence (AI) ,there have been several moments in history when it approached neuroscience in searching for bio-inspiration. In the middle of the twentieth century, the biomimetic approach was critical for developing artificial neural networks (ANNs). In 1943, W. McCulloch and W. Pitts proposed a model of the first artificial neuron (McCulloch and Pitts, 1943).

The neuron received several synaptic-like inputs and generated an output if the number of activated synapses exceeded a threshold, thus mimicking the "all-or-none" principle of action potentials. Later, F. Rosenblatt further developed this idea and coined the term perceptron (Rosenblatt, 1958). Subsequent studies have shown the unreasonable effectiveness of artificial neurons coupled into networks in a constantly growing number of AI applications.

Mathematically speaking, an ANN is a function $y = f(x)$ that maps input data, x, into output, y. Thus, it can emulate reflex responses. For example, to decide if there is a cat or a dog on an image, we can designate the image as the input x, and the animal category as y. Then, by presenting a photo to the ANN, we could quickly determine which animal appears in the image. But are we sure that there is such an ANN? In other words, can ANNs approximate arbitrarily complex functions?

The universal approximation theorem provides the answer. In 1989, G. Cybenko showed that an ANN with sigmoid activation could approximate any continuous function (Cybenko, 1989). Later, this result was extended into Lebesgue integrable functions and the rectified linear unit (ReLU) activation function (Lu et al., 2017; Hanin, 2019). In practical terms, no matter what $f(x)$ is, there is an ANN approximating it with an arbitrary degree of accuracy.

However, mere existence is not enough for AI applications. The input and output sets and the function $f(x)$ can be arbitrarily complex and high-dimensional. It makes impracticable, the use of human-tailored standard techniques like extracting a set of predictive features by principal component analysis, Fourier transform, etc. The main advantage of ANNs is the ability to learn from data, although such learning is data-hungry.

The synaptic weights of each neuron (i.e., the parameters of function $f(x)$) can be adjusted by a training process aiming at minimizing the prediction error. The optimization is usually done by a version of the stochastic gradient descent method (Robbins and Monro, 1951) based on the derivatives of the loss function evaluated by the backpropagation algorithm (Rumelhart et al., 1986). Thus, an ANN can automatically identify the hidden features essential for the classification. Then, the trained ANN can predict the output for unseen inputs taken from the same distribution, i.e., it gains the generalizing capability.

While feedforward fully connected ANNs were achieving solid results in many areas, until recently, they have been ineffective in tasks that are comparatively easy for humans (Schmidhuber, 2015). In 1997, Deep Blue, a chess machine developed by IBM, defeated the world champion, Garry Kasparov, while those days computers could not compete with kids in recognizing faces. This problem was utterly complex for AI systems, much harder than chess. To process an RGB photo with a rather mediocre resolution of 1 Mpx, the input layer of an ANN must have 3×10^6 neurons. If the second layer has only 1,000 neurons, we approach 10^{10} synaptic weights to train. Thus, we rapidly get numbers prohibitive for modern computers, databases, and algorithms.

A critical breakthrough has been achieved by copying the converging architecture of the visual system of the brain (Olshausen and Field, 2004). In a seminal work, LeCun et al. (1989) reported a new class of ANNs, convolutional neural networks (CNNs; Goodfellow et al., 2016). The CNN architecture mimics the primate's visual cortex (Hubel and Wiesel, 1968; Laskar et al., 2018). The V1 and V2 cortex regions are similar to convolutional and subsampling layers of a CNN, whereas the inferior temporal area resembles the higher layers (Grill-Spector et al., 2018; Khan et al., 2020).

Different filters using convolution have been known for a long time in image processing. But CNNs offered the possibility of learning these filters from the data automatically. As in the visual cortex, the first CNN layers detect simple shapes, such as lines and circles, and combine them into more complex features at each successive layer. The detected features stop making sense for a human observer at some point, but they encapsulate the essence of images (Altenberger and Lenz, 2018).

Thus, the rise of CNNs provided a methodology that allowed for outperforming humans in object recognition. In 2012, the AlexNet was the first CNN that beat traditional geometric approaches in the object recognition contest (Krizhevsky et al., 2012; Russakovsky et al., 2015). Since then, CNNs have constantly improved the results of the state-of-the-art (Altenberger and Lenz, 2018). The current winner, CoAtNet-7 (Dai et al., 2021), provides 90.88% of the Top 1 accuracy in image classification on the ImageNet benchmark (Imagenet, 2022). Although CNNs achieved superhuman performance in the visual pattern recognition in controlled competitions, humans are still much better in general recognition tasks (Schmidhuber, 2015).

From Reflex to Reflective Neural Networks Aiming at Cognition

Artificial neural networks are already widely used in the first domestic robots, cars with an increasing autonomy to make in-driving decisions, and apps that manage our data and anticipate our actions and desires in daily life (Mackenzie, 2013; Lee et al., 2016; Bogue, 2017; Hussain and Zeadally, 2018). However, along with these successes, there emerges an awareness of fundamental limitations, mainly associated with the reflex nature of ANNs and shortcuts in simulating deep cognition, i.e., the mental process ruling our interactions with the environment. In the late 80s, Moravec (1988), Minsky, and others anticipated the unexpected slow progress in deep artificial cognition. The Moravec paradox says: *It will be much easier to create a robot capable of talking with us than a robot ready to move among us.* After almost 40 years, we can only confirm the prophecy.

Experimental studies on rats have shown that reinforcement is not necessary for learning (Tolman and Honzik, 1930). Rats actively process information rather than operate on a stimulus-response relationship as most contemporary ANNs do. Based on these data, in 1948, Edward Tolman coined the term cognitive map, which is an internal representation of one's environment. Such an internal representation emerges as a reflective (thinking) processing of external information. Recent advances provided evidence of time compaction performed by the human brain when dealing with dynamic situations (Villacorta-Atienza et al., 2021). Theoretically, such compaction occurs through an active wave propagating in a neuronal lattice (Villacorta-Atienza et al., 2010, 2015). Thus, the fundamental difference between the

biological neural networks in higher brain stations and modern ANNs is reflective vs. reflex information processing.

Future ANNs will also address the issue of energy efficiency. In modern ANNs, the information flow occurs continuously, and usually, all neurons are active and consume energy. Implementations of ANNs on GPUs are "hot ovens," much hungrier for energy than the biological brains. Current trends in chip building go by increasing the power density (now about 100 W/cm^2 vs. 0.01 W/cm^2 for the brain) and the clock frequency (about 10 GHz vs. 10 Hz; Merolla et al., 2014). In the brain, only a tiny fraction of neurons are active at a given time. Neurons efficiently communicate by brief spikes and often remain quiet. Therefore, spiking neural networks (SNNs) mimicking real neurons progressively gain importance. However, the intrinsic complexity of SNNs slows down their expansion. In practical applications, current SNNs trained by supervised learning algorithms have already caught up with ANNs in recognition tasks (Shrestha and Orchard, 2018; Zambrano et al., 2019; Panda et al., 2020; Yin et al., 2021; Zenke and Vogels, 2021; Chen et al., 2022). However, the use of SNNs within the reflex paradigm limits the range of tasks to be solved and our understanding of the brain. We foresee that the future of SNNs will concentrate on the development of cognitive devices based on novel mathematical paradigms beyond standard ANN applications.

The recent experimental discovery of concept cells (Quian Quiroga, 2012) and the associated mathematical concept of a high-dimensional brain (Gorban et al., 2019) can boost the use of SNNs in tasks related to reflective information processing. Reflective SNNs can take advantage of their intrinsic dynamics and emulate complex, not reflex-based, brain actions, such as generating new abilities from previously learned skills. SNNs can be implemented as analog computational systems. We foresee that such systems will use the emerging memristive hardware paradigm for this purpose.

Memristors are passive elements of electrical circuits that can be densely packed in 2D or 3D matrices on a chip and emulate plastic changes in synaptic contacts in ANNs (Strukov and Williams, 2009; Jo et al., 2010). This enables a natural implementation of the synaptic integration of information in neurons. Thus, memristive crossbars are good candidates for building in-memory calculations for future reflective neural networks. The latter may open new horizons for deploying compact, low-power wearable devices that will provide a next-level cognitive experience to a user.

SPIKING NEURAL NETWORKS AS AN ALTERNATIVE FOR BUILDING REFLECTIVE ARTIFICIAL INTELLIGENCE

Models of Spiking Neurons

The synergy between neuroscience and novel mathematical approaches can be a solution for building novel systems exhibiting reflective AI. In contrast to formal neurons used in ANNs, biological cells exchange information by brief pulses, called action potentials or spikes. Then, complex internal dynamics of neurons can significantly affect the processing and transmission of information, and the spike times matter.

Many mathematical models of spiking neurons have been proposed (Hodgkin and Huxley, 1952; FitzHugh, 1961; Koch and Segev, 1999; Izhikevich, 2003). They differ in the degree of biological realism. The most complete models use the Hodgkin–Huxley (HH) formalism. However, such models are computationally demanding, and their analytical analysis is complicated. In many practical applications, one can use an HH model only with the leaky current and assume that a neuron fires a spike if its membrane potential crosses a threshold. This reduction yields the simplest leaky integrate-and-fire model of spiking neurons (Abbott, 1999). Its most significant disadvantage is a reduced repertoire of dynamic behaviors, e.g., the absence of neuronal adaptation. However, if biological relevance is of no concern, integrate-and-fire models are attractive for large-scale simulations (Delorme et al., 1999). The Izhikevich model provides a balance between the computational cost and the variety of behaviors it can reproduce (Izhikevich, 2005). Besides modeling the neuronal membrane, there is a class of models, called multicompartmental, that also simulate the neuron's morphology (Bower and Beeman, 1998; Koch and Segev, 1999). Such models are essential for studying complex processes occurring in a neuronal tissue, e.g., the spreading of depression or migraines (Makarova et al., 2010; Dreier et al., 2017).

The Challenge of Training Spiking Neural Networks

Spiking neural networks are arguably more biologically realistic than ANNs, and the only viable option if one wants to simulate brain computations. Nevertheless, the reverse side of the coin is intrinsic complexity. The output of a neuron is no longer a univocal function of the input, which in turn is a fundamental property of a reflective system. Training SNNs usually employs diverse forms of supervised, unsupervised, or reinforcement learning. Different versions of Hebbian learning, particularly spike-timing dependent plasticity (STDP), have shown significant potential in a variety of cognitive tasks. Being experimentally supported, STDP strengthens a connection if the postsynaptic neuron generates a spike after the presynaptic one and weakens in the opposite case (Markram et al., 1997; Bi and Poo, 1998; Sjöström et al., 2001). We note that this type of plasticity has inherent elements of synaptic competition, which makes the "success" of the synapse dependent on the spike timings (Song et al., 2000).

Most modern attempts in training SNNs are still based on algorithmic approaches working well in ANNs, e.g., the minimization of loss (error) functions (Taherkhani et al., 2020). The so-called ANN-to-SNN conversion adopts methods already existing in deep ANNs. First, a corresponding ANN is trained, and then, taking into account some restrictions, the obtained synaptic weights are transferred to a similar SNN (Cao et al., 2015; Esser et al., 2016). Under this approach, the firing rates of spiking neurons should match the graded activations of formal neurons. Various optimization techniques and theoretical generalizations

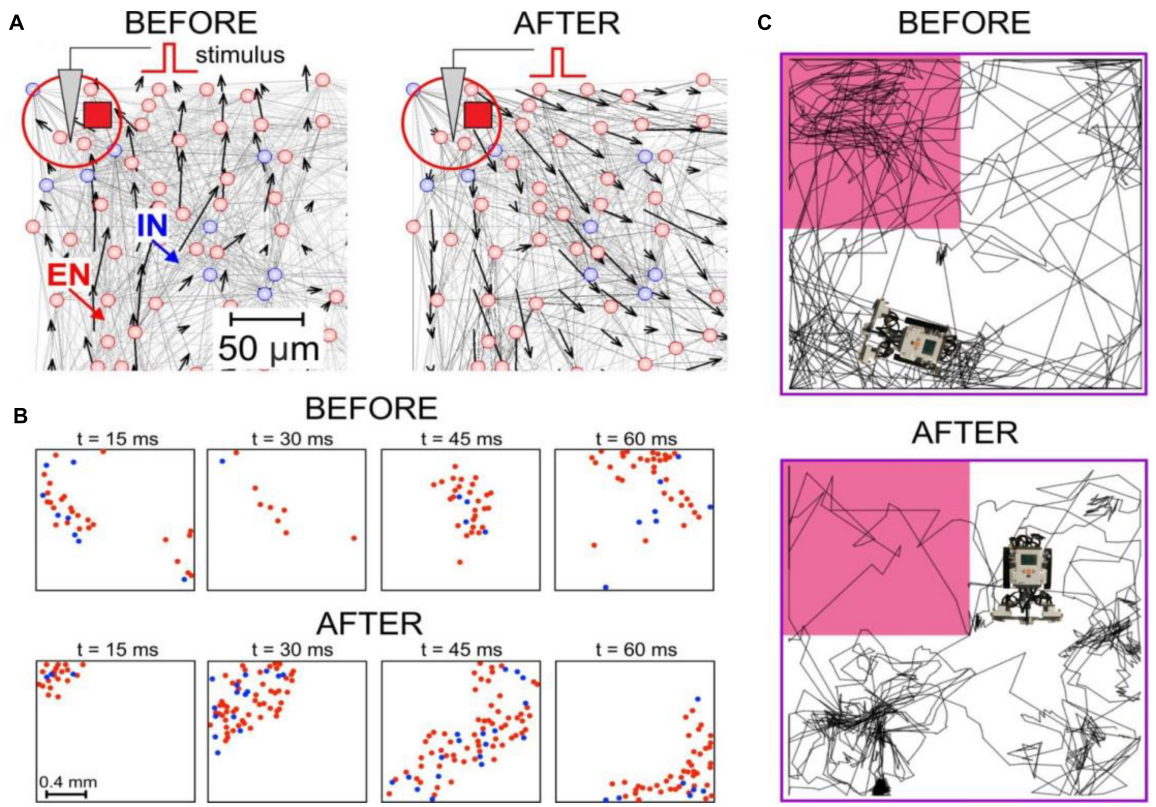


FIGURE 1 | Spike-timing dependent plasticity (STDP)-driven spatial computing in spiking neural networks (SNNs). **(A)** The synaptic vector field of an SNN reveals the potentiation of centrifugal connections after local stimulation. **(B)** Connectome rearrangements lead to the transformation of patches of spike activity into traveling waves. **(C)** A neurorobot driven by an SNN avoids the dangerous zone (marked by pink) after learning.

of this approach have been proposed (Diehl et al., 2015; Ruckauer et al., 2017).

In image processing, ANN-to-SNN methods allow for obtaining high accuracy, close to the performance of classical deep learning in ANNs (Neil et al., 2016; Tavanaei et al., 2019). When using event-based input data, e.g., from dynamic vision sensors, and energy-efficient hardware implementation, such SNN-based solutions can compete with deep ANNs (Cao et al., 2015; Esser et al., 2016).

Another approach to training SNNs relies on adapting the backpropagation algorithm to the temporal coding scheme in which input and output data are represented by relative spikes' times (or delays). Several backpropagation-like algorithms for multilayer SNNs have been proposed, such as SpikeProp (Bohte et al., 2002), backpropagation with momentum (Xin and Embrechts, 2001), Levenberg–Marquardt algorithm for SNNs (Silva and Ruano, 2005), QuickProp and Resilient propagation (RProp) versions of SpikeProp (McKennoch et al., 2006; Ghosh-Dastidar and Adeli, 2007), and SpikeProp based on adaptive learning rate (Shrestha and Song, 2015). Mostafa (2018) used a transformation of variables in a feedforward SNN and showed that the input–output relation is differentiable and piecewise linear in a temporal coding scheme. Thus, methods of training ANNs can be used in SNNs. In the proposed back propagation-based algorithm, the performance of the SNN was slightly inferior to ANN. Still, it showed a much shorter time in the network response to a pattern presented to the input.

These approaches use only the first spike of each neuron during SNN learning and operating (the so-called time-to-first-spike or TTFS coding). Such limitation is overcome by methods using neurons capable of learning to fire a precise temporal spike pattern in response to a particular sequence of spike trains at the input: ReSuMe (Ponulak, 2005; Ponulak and Kasiński, 2010), tempotron (Gütig and Sompolinsky, 2006), chronotron (Florian, 2012), and SPAN (Mohemmed et al., 2012). These algorithms use biological-like elements in learning rules (such as STDP and anti-STDP window), but they work only with one (output) layer of spiking neurons or even with a single neuron. Further development of this idea offered supervised learning methods for multilayer networks with hidden neurons (Sporea and Grüning, 2013; Taherkhani et al., 2018).

Recently, the concept of surrogate gradients in SNNs has addressed the problem of the discontinuous derivative of the spike functions (Neftci et al., 2019). In particular, the spike function was approximated by a continuous one that served as the surrogate for the gradient. This approach enables direct training of deep SNNs using input spikes both in the temporal and rate coding schemes. The effectiveness of surrogate gradients

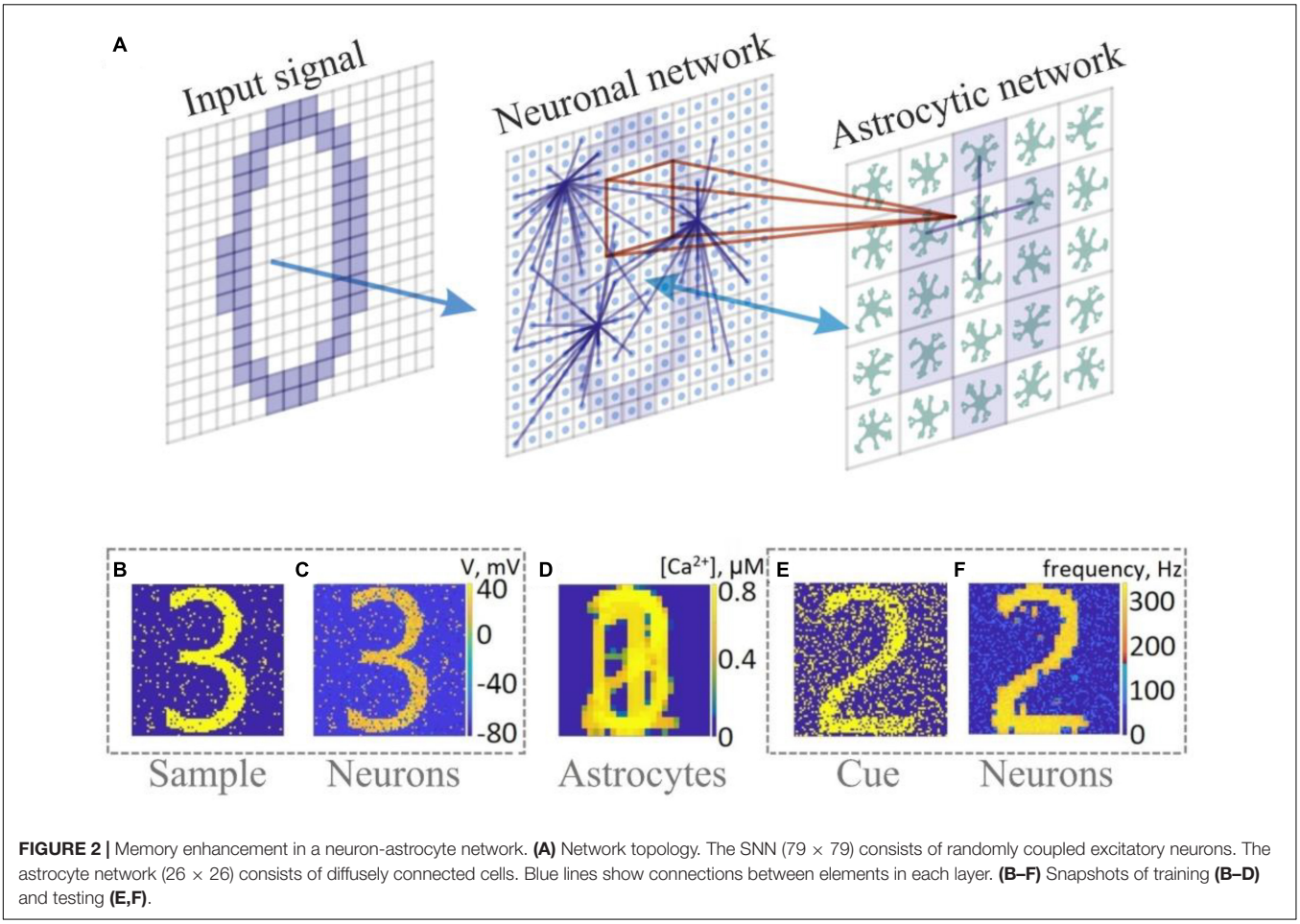


FIGURE 2 | Memory enhancement in a neuron-astrocyte network. **(A)** Network topology. The SNN (79 × 79) consists of randomly coupled excitatory neurons. The astrocyte network (26 × 26) consists of diffusely connected cells. Blue lines show connections between elements in each layer. **(B–F)** Snapshots of training **(B–D)** and testing **(E,F)**.

in training deep SNNs achieved the state-of-the-art performance for an ANN in a significant number of standard tests (Shrestha and Orchard, 2018; Lee C. et al., 2020; Panda et al., 2020; Yin et al., 2021; Zenke and Vogels, 2021).

Collective Dynamics in Spiking Neural Networks: Architecture vs. Function

Communication by spikes *via* plastic synaptic contacts provides different encoding modalities, including successive excitation, number of spikes in a train, spike timings (or phases) relative to a clock signal, rate encoding, etc. Various learning schemes for SNNs employ a binary categorization of processed information (Taherkhani et al., 2020; Dora and Kasabov, 2021). In other words, SNNs are initially thought of as biological neuron-like analog processing units that operate with digital computing tasks and tools.

Analog units with theoretically unlimited degrees of freedom are hardly controllable. Therefore, existing SNNs frequently lose in competition with modern ANNs originally constructed as algorithmic digitized networks solving logic computational tasks. However, reflecting SNNs mimicking structural and functional features of brain circuits have untapped the potential in exploring cognitive tasks, thereby bringing un closer to "intelligent" AI. To explore this potential, one should try to employ concepts of modern neuroscience from cellular and molecular to cognitive levels. Let us now have a short excursion into the concepts of structural and functional plasticity, which might be helpful in training SNNs to process data in a biologically relevant way.

In the brain, neuronal plasticity plays a crucial role in establishing functions. In common words, plasticity is an activity-dependent change in the dynamics of neurons and synapses at cellular (local) and circuit (global) levels. Features of the local synaptic plasticity typically appear as a change of synaptic strength depending on the local activity of corresponding neurons. The Hebbian learning rule is represented by STDP which corrects the synaptic strengths depending on spiking times between the pre- and postsynaptic neurons (Morrison et al., 2008). These changes may facilitate or depress particular signal transmission pathways in an unsupervised manner. In other words, STDP results in the formation of specific synaptic network architecture reflecting current activity patterns and, hence, may be specific to input data.

At the circuit level, plastic changes can lead to different behaviors. The vector field method can be used to visualize the network architecture and functionality (Ponulak and Hopfield, 2013; Lobov et al., 2016, 2017). **Figure 1A** shows an example of the network reorganization provoked by a stimulus. Recently, it

has been shown that there is an interplay between the anatomic architecture and functionality, and functional changes can drive the rebuilding of the network and vice versa (Lobov et al., 2021b).

Experimental data suggested that propagating patches of spike activity (**Figure 1B**) can play the role of basic functional units in brain information processes (Gong and Van Leeuwen, 2009; Muller et al., 2018). Based on this hypothesis, the concept of spatial computing was proposed, which can be defined as computations in neural networks mediated by the interaction of waves and patches of propagating excitation. This coding principle enables the detection of different signals and performing various stimulus transformations, for example, signal frequency reduction (Villacorta-Atienza and Makarov, 2013). One of the implementations of this concept can be considered a learning model in a neural network based on the STDP association of interacting traveling waves (Alexander et al., 2011; Palmer and Gong, 2014). Spatial computing in small neural circuits and modular SNNs can simulate Pavlovian conditioning and operant learning in neurorobots (Lobov et al., 2020b, 2021a). Another possible way to implement spatial computations is cognitive maps and spatial memory with positive (Ponulak and Hopfield, 2013) or negative (Lobov et al., 2021b) environmental stimuli (**Figure 1C**). Note that due to the presence of spontaneous activity in SNNs (unlike ANNs), they can "live" without external input, determining the "behavior" of neurorobots (Lobov et al., 2020b, 2021a,b).

The formation of cognitive maps and extraction of information from them can be based on the dependence of wave propagation on the connectom (Keane and Gong, 2015; Naoumenko and Gong, 2019; Lobov et al., 2021c). On the other hand, there are mechanisms for rapid switching of wave dynamics based on the balance of inhibition and excitation (Heitmann et al., 2012). Generalized cognitive maps provide another example of wave computations. In particular, the propagation of a wave of excitation in an SNN generates a cognitive map of a dynamic situation observed by a subject in the environment (Villacorta-Atienza et al., 2010; Makarov and Villacorta-Atienza, 2011).

Several studies have used unsupervised Hebbian learning in the STDP form to solve classification problems (Querlioz et al., 2013; Diehl and Cook, 2015; Tavanaei and Maida, 2015). Elements of reward in SNNs can force learning in a desirable direction (Izhikevich, 2007; Chou et al., 2015; Mozafari et al., 2018). Methods of supervised SNN learning are also proposed based on both temporal and frequency coding by stimulating target neurons (Legenstein et al., 2005; Lobov et al., 2020a). Another way to implement supervised learning is feedback from output neurons and element associative learning (Lebedev et al., 2020).

Interplay of Neurons and Glial Cells in Spiking Neural Networks

In recent decades, experimental findings in cellular and molecular neuroscience revealed that glial cells also participate in information processing (Santello et al., 2019). Glial cells, specifically astrocytes accompanying neural networks, can effectively modulate local synaptic transmission (Perea and Araque, 2007; Durkee and Araque, 2019). Neurotransmitters diffusing from the synaptic cleft and bounding to specific receptors expressed in the plasma membrane activate astrocytes. In turn, the latter release neuroactive chemicals, called gliotransmitters, that activate specific receptors on both pre- and postsynaptic neurons. Such an interplay changes the efficacy of synaptic transmission on neighboring synapses. The modulation may last for dozens of seconds and have bidirectional influence, either facilitating or depressing synapses.

Interacting with both pre- and postsynaptic neurons, astrocytes form a so-called tripartite synapse (Araque et al., 1999). In terms of information processing, astrocytes may enhance the learning capability of the network. Gordleeva et al. (2021) showed the possibility of memory enhancement by exploring an SNN interacting with astrocytes that served as reservoir preserving information patterns independently on neurons within dozens of seconds (**Figure 2**).

Local changes of synaptic strength, due to, e.g., short-term plasticity, regulate signal transmission in a synapse depending on its activity, often referred to as homosynaptic regulations. There is also another type of regulation called heterosynaptic plasticity when other inactive synapses change their efficiency (Chater and Goda, 2021). Heterosynaptic plasticity has different forms working at a similar time scale as the Hebbian plasticity. It can lead to both long-term potentiation and depression (LTP and LTD) of synapses. Thus, it can also play a crucial role in learning-related changes. Understanding of molecular and cellular mechanisms of heterosynaptic plasticity remains fragmentary. At the functional level, astrocytes may provide coordination between different signal transmission pathways (Gordleeva et al., 2019). Being activated by one of the synapses, astrocytes may release gliotransmitters back to the active synapses and inactive ones located at different spatial sites.

Homeostatic Plasticity Is Relevant for Learning in Spiking Neural Networks

In living neural networks, homeostatic plasticity sustains the physiological conditions of functioning and balance (Keck et al., 2017). It prevents neurons from hyper- and hypo excitations. Thus, it acts mainly opposite the Hebbian learning rule that potentiates synapses in an activity-dependent manner.

Learning of complex patterns by an SNN requires neuronal competition (i.e., competition of the "outputs" of the network) similar to the "winner-takes-all" rule: the winning neuron should selectively recognize the pattern that caused its activation (refer to Section "Novel Mathematical Principles for Spiking Neural Networks: Concept Cells and High-Dimensional Brain"). It can be achieved by lateral inhibition (Quiroga and Panzeri, 2013; Lobov et al., 2020a,b). In addition to the neuronal competition, it is necessary to implement synaptic competition (i.e., competition of the "inputs" to the network). It can be achieved directly (Bhat et al., 2011; Lobov et al., 2020b) or indirectly *via* the homeostatic plasticity and synaptic scaling (Keck et al., 2017; Turrigiano, 2017) or synaptic forgetting (Panda et al., 2018; Lobov et al., 2020a).

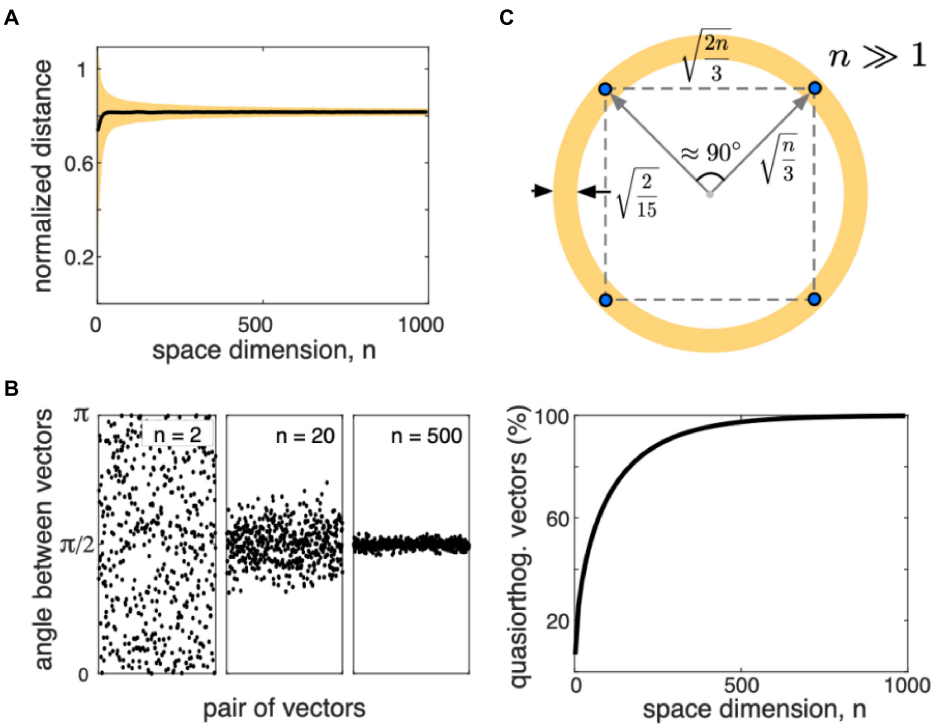


FIGURE 3 | Some properties of the distributions of random vectors in high-dimensional (HD) spaces (n is the space dimension). **(A)** The mean (black) and standard deviation (yellow) of the distance between two randomly chosen points sampled from a uniform distribution in the hypercube $[-1, 1]^n$, normalized to $\sqrt{n}$. Distances concentrate around $\sqrt{2n/3}$. **(B)** *Left:* Examples of the angles between pairs of randomly chosen vectors for n 2, 20, and 500. For high n, angles concentrate around $\pi/2$. *Right:* Proportion of quasi-orthogonal (with the tolerance 0.1 rad) vectors vs. the space dimension. For high enough dimensions, almost all vectors are quasi-orthogonal. **(C)** Sketch of the distribution of random points (blue dots) in a high-dimensional space.

Mechanisms of homeostatic plasticity were thoroughly studied, including synaptic scaling, changing postsynaptic density, control of excitation/inhibition balance, sliding thresholds for LTP, and LTD induction in Hebbian plasticity (Keck et al., 2017). An interesting point in the homeostatic changes concerns the activity of the brain extracellular matrix (ECM; Dityatev et al., 2010). The ECM is an activity-dependent environment for SNNs affecting synaptic transmission by synaptic scaling on the postsynaptic side and ECM receptors on the presynaptic one. At the functional level, the ECM works at much longer time scales (hours or days) and may serve as a long-term reservoir containing memory traces (Kazantsev et al., 2012; Lazarevich et al., 2020).

At the network level, changes in network architecture are driven by structural plasticity (Yin and Yuan, 2015). It accounts for structural changes in the number of synaptic receptors expressed in the dendritic spines, the number of synapses, and the number of neurons. The structural plasticity implements two essential strategic functions: (1) Sustain homeostasis. For example, the number of inhibitory synapses can increase to compensate for hyperexcitation (Keck et al., 2017). (2) Enhance learning capabilities (Hellrigel et al., 2019; Calvo Tapia et al., 2020a; Rentzeperis et al., 2022). For instance, synaptic receptors and synapses can be additionally generated to extend a specific signal-transmitting channel. In other words, the network architecture becomes dynamic. A network can change its dimension depending on the activity and entrust tasks. Thus, structural plasticity is crucial for learning, and network robustness compensates for injuries and ill-functioning states.

NOVEL MATHEMATICAL PRINCIPLES FOR SPIKING NEURAL NETWORKS: CONCEPT CELLS AND HIGH-DIMENSIONAL BRAIN

How Does the Brain Encode Complex Cognitive Functions?

As we mentioned in the Introduction, the brain is not inert but actively generates predictions about what will happen next. Such predictions presumably occur in higher brain stations that summarize and process converging information from different sensory pathways. An intriguing question concerns the role of individual neurons in complex cognitive functions and, in the end, in conciseness. This question is as old as Neuroscience itself. It yielded many significant results, such as discovering efficient or sparse coding (Barlow, 1961; Field, 1987; Olshausen and Field, 1997), and is far from being satisfactorily answered (Valdez et al., 2015).

A somewhat extended opinion says that complex intellectual phenomena result from a perfectly orchestrated collaboration between many cells (Bowers, 2009). This idea, known as the "million-fold democracy," was put forward by C. Sherrington (1940). Our actions are driven by the joint activity of millions of neurons in an "election" in which some neurons vote more often than others. It yielded the concept of population coding: The brain encodes information by populations or clusters of cells rather than by single neurons (Pouget et al., 2000). For example, in the primate primary motor cortex, individual neurons are tuned to the direction of arm movement, and populations of such neurons have to be pooled together to compute the direction a monkey is about to move its arm (Georgopoulos et al., 1986). This finding prompted the development of brain–machine interfaces using population coding (Lebedev and Nicolelis, 2017).

In 1890, even before the pioneering works on neuroanatomy by S. Ramon y Cajal, W. James (1890) proposed that neurons have individual consciousness and that there is one "pontifical" cell to which our consciousness is attached. Although such an idea sounds absurd nowadays, it may not be so far from the truth. According to J. Edwards (2005), to combine different flows of information into a smoothly unrolling, multi-modal experience of reality, the relevant bits of information must come together in one unit somewhere. A brain or its parts are too big, but a single neuron may be just about right. Gnostic (i.e., single-cell) coding may also provide metabolic efficiency. The high cost of spiking drives the brain to use codes that minimize the number of active neurons (Lennie, 2003).

Individual Concept Cells Can Be Responsible for Cognitive Phenomena

The "degree" of consciousness in gnostic cells may depend on the spatial pattern a neuron receives (Sevush, 2006; Cook, 2008). The conscious activity of neurons in the initial relay stations is simple and cannot directly affect the animal's macroscopic behavior. However, at higher brain stations, neurons operate with complexity and diversity sufficient to account for complex conscious experiences. Converging experimental evidence confirms that small neuron groups or single cells can implement complex cognitive functions, such as generating abstract concepts.

Some pyramidal neurons in the medial temporal lobe (MTL) can exhibit remarkable selectivity and invariance to complex stimuli (Quian Quiroga et al., 2005; Mormann et al., 2011). It has been shown that the so-called concept cells (or grandmother cells) can fire when a subject sees one of seven different pictures of Jennifer Aniston but not the other 80 pictures of other persons and places. Concept cells can also fire to the spoken or written name of the same person (Quian Quiroga, 2012). Thus, a single concept cell responds to an abstract concept but not to the sensory features of the stimuli. Moreover, concept cells are relatively easily recorded in the hippocampus (Quian Quiroga, 2019). Thus, they must be abundant, at least in the MTL, contrary to the common opinion that their existence is highly unlikely (Bowers, 2009). Kutter et al. (2018) have found that single neurons in MTL encode numbers. They suggested that number neurons provide the neuronal basis of human number representations that ultimately give rise to number theory and mathematics.

Spiking Neural Networks Can Take Advantage of the Blessing of Dimensionality

The discovery of concept cells has stimulated theoretical research, which led to the theory of a high-dimensional brain (Tyukin et al., 2019). It uses fundamental properties of high-dimensional (HD) data. On the one hand, in HD-spaces, we can observe the curse of dimensionality, the term coined by Bellman (1957). It highlights, for instance, the combinatorial explosion. To sample n Boolean features, we must check 2^n cases. Even for a relatively low dimensional space with $n = 30$, this number goes to almost 10^{10}, prohibitive for modern computers. Another example is the concentration of the distances between randomly selected points. If n increases, the pairwise distances concentrate around the mean value (**Figure 3A**). Then, the distance-based methods, such as k-nearest neighbors work poorly (Beyer et al., 1999; Pestov, 2013; Lin et al., 2014).

On the other hand, it turns out that rather general stochastic processes can generate HD signals with relatively simple geometric properties (Gorban et al., 2019, 2020). In 2000, D. Donoho introduced the term "blessing of dimensionality," with which the curse of dimensionality are two sides of the same coin (Donoho, 2000). As a system becomes more complex, it has been observed that its analysis can be complicated at first, but then it becomes simpler (Kreinovich and Kosheleva, 2021). A good example is the Central Limit Theorem. A statistical analysis of a few random variables can be highly complicated. However, a mixture of many random variables follows a Gaussian distribution and can be easily described by the mean and the standard deviation.

Both the curse and the blessing of dimensionality are the consequences of the measure concentration phenomena (Ledoux, 2005; Gorban et al., 2016; Gorban and Tyukin, 2018). **Figure 3B** illustrates examples of the angle between two randomly chosen vectors (sampled from a uniform distribution in a hypercube $[-1, 1]^n$). Together with the distribution of the inter-point distances, we can conclude that all vectors having approximately equal length, are nearly orthogonal, and the distances between data points are roughly equal (**Figure 3B**). **Figure 3C** illustrates a sketch of how random data points appear in the HD space. Tyukin et al. (2019) hypothesized that neurons could take advantage of such a notorious simplicity of the distribution and use simple mathematical mechanisms for processing complex, HD data.

High-Dimensional Neurons Can Exhibit Unexpected Properties

According to Sevuch and Cook (see, e.g., Sattin et al., 2021), the synaptic connections within a neural network could represent the substrate of cognition. The pattern complexity plays a key role, and conscious human behavior requires the processing of complex multidimensional data. Recent empirical evidence

shows that a variation in the dendrite length and hence in the number of synapses n can explain up to 25% of the variance in IQ scores between individuals (Gorioununova et al., 2018).

Figure 4A illustrates a theoretical model of an HD neuron (Tyukin et al., 2019). The neuron receives as an input an HD vector pattern $\boldsymbol{x} = (x_1, \ldots, x_n)^T \in [-1, 1]^n$, such that the number of individual inputs or the neuronal dimension $n \gg 1$. The inputs are connected to the neuronal membrane through synaptic contacts. The output of the neuron is given by a transfer function, e.g., ReLU. During operation, the synaptic weights of the neuron change by a Hebbian-type rule (Calvo Tapia et al., 2020a).

Given that the distribution of the input patterns in a big but finite set of stimuli has no strong clots, it has been shown that the neuron can accurately learn a single pattern from the entire set. An important consequence is that no a *priori* assumptions on the structural organization of neuronal ensembles are essential for explaining the fundamental concepts of static and dynamic memories. Cognitive functionality develops with the dimension of single neurons in a series of steps (Gorban et al., 2019; Tyukin et al., 2019).

The neuronal selectivity emerges when the dimension exceeds some critical value, around $n = 30$ (**Figure 4A**). At this crucial transition, single neurons become selective to single information items. The second critical transition occurs at significantly larger dimensions, around $n = 300$. At this second stage, the neuronal selectivity to multiple uncorrelated stimuli develops. The ability to respond selectively to a given set of numerous uncorrelated information items is crucial for rapid learning "by temporal association" in such neuronal systems.

Single High-Dimensional Neurons in Deep Spiking Neural Network Layers May Provide Cognition

Remarkably, a simple generic model offers a clear-cut mathematical explanation of a wealth of empirical evidence related to *in vivo* recordings of "grandmother" cells and rapid learning at the level of individual neurons. It also sheds light on the question of why Hebbian learning may give rise to neuronal selectivity in the prefrontal cortex (Lindsay et al., 2017) and explain why adding single neurons to deep layers of ANNs is an efficient tool to acquire novel information while preserving previously trained data representations (Draelos et al., 2017).

Calvo Tapia et al. (2020b) extended results into the problem of building abstract concepts by binding individual items of the same kind. **Figure 4B** illustrates the model mimicking primary signaling pathways in the hippocampus. It considers the stratified structure of the hippocampus that facilitates ramification of axons, leaving multiple buttons in the passage and conveying the same HD input to multiple pyramidal cells (Teyler and Discenna, 1984). The latter has been supported by electrophysiological observations showing that Schaffer collaterals create modules of coherent activity with a large spatial extension in the CA3 region (Benito et al., 2014, 2016). Thus, the hippocampal formation possesses rather exclusive anatomical and functional properties required for the emergence of concept cells.

In the beginning, the receptive fields of all neurons (areas in the sensory domain evoking a response) in both strata form a disordered mixture of random regions (see the cartoon in **Figure 4C**, left). Thus, the output of the concept stratum is random, and the system cannot follow the music. The purpose of learning is to organize the receptive fields so that the concept cells become note-specific (**Figure 4C**, right). In this case, each concept cell will not be stimulus-specific but represent a set of associated stimuli or a concept, e.g., note A.

The network has been tested on the perception of the 9th Symphony by Beethoven (**Figure 4D**). The selective stratum detects individual sound waves, while the concept stratum puts them together and forms the note-specific output (Calvo Tapia et al., 2020a). Thus, concept cells respond to particular notes regardless of the phase of sound waves, and the "brain" now does follow the music. This result supports the hypothesis of a strong correlation between the level of neuronal connectivity in living organisms, and different cognitive behaviors such organisms can exhibit (Herculano-Houzel, 2012).

THE MEMRISTIVE ARCHITECTURE ENABLES THE IMPLEMENTATION OF REFLECTIVE SPIKING NEURAL NETWORKS

What Is a Memristor?

In 1971, Leon Chua (1971) discovered the memristor as a hypothetical fourth passive element of electrical circuits. A memristor relates a change in the magnetic flux with a variation of the electric charge flowing through this element. Mathematically, it is equivalent to a nonlinear resistor that changes its resistance depending on the history of the electric current. Therefore, it was called a memristor, i.e., a memory resistor.

In 2008, Strukov et al. (2008) associated the memristive effect with resistive switching in thin-film metal-oxide-metal structures. Such films were actively studied as early as the middle of the twentieth century (Dearnaley, 1970). Starting in 2008, the current wave of interest in memristors began to rise. Although memristors have been thoroughly studied, there are still debates and doubts about the existence of an ideal memristor satisfying the original definition and the validity of its correlation with resistive switching (Vongehr and Meng, 2015; Demin and Erokhin, 2016; Kim et al., 2020). Despite that, the generalized definition of a memristor as a dynamical system, which Chua and Kang (1976) proposed in 1976, remains valid. According to it, a memristor is a system described by the following equations:

$$I(t) = \frac{V(t)}{R(x, V)} \tag{1a}$$

$$\frac{dx}{dt} = f(x, V), \tag{1b}$$

where $I(t)$ is the current flowing through the system, $V(t)$ is the voltage drop, $R(x, V)$ is the resistance with memory or

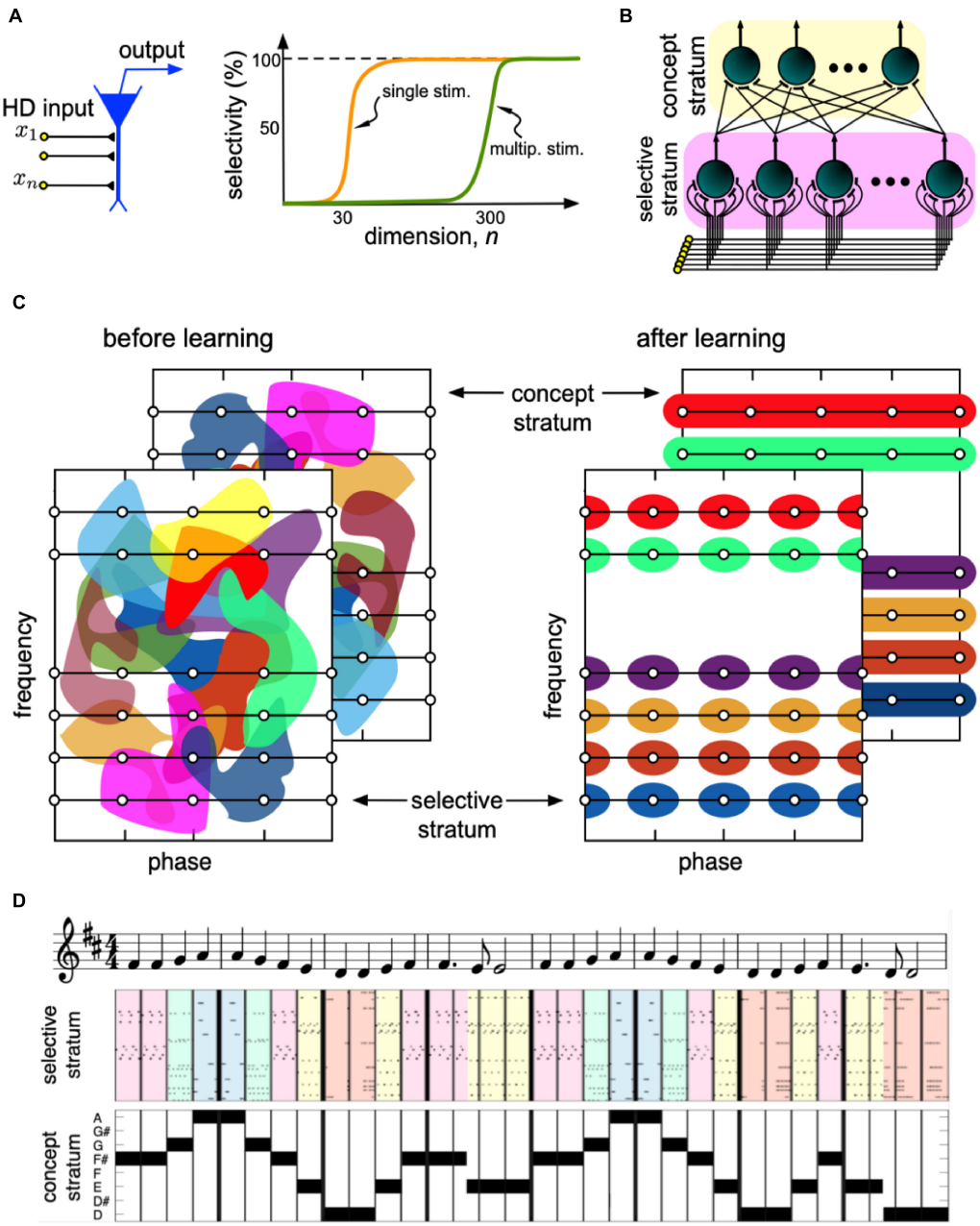


FIGURE 4 | The concept of an HD brain. **(A)** A neuron receives an HD pattern x and produces output if the pattern matches the pattern of synaptic weights. The cognitive functionality of a neuron develops in steps. At dimensions n 30, the neuron becomes capable of selectively detecting a single stimulus from a large set (orange curve). At n 300, a new ability appears. The neuron can now detect multiple uncorrelated stimuli. **(B)** Model mimicking the information flow in the hippocampus. A stimulus, e.g., a sound wave, goes to HD neurons in the selective stratum. Neurons learn different sound waves and send collaterals to neurons in the concept stratum. Concept cells extract concepts of musical notes. **(C)** Under learning, the rearrangement of the neuronal receptive fields leads to the formation of note-specific concept cells (different colors correspond to the receptive fields of different neurons). **(D)** Binding of sounds into notes for a fragment of the 9th Symphony by Beethoven "Ode to joy".

memristance, $x(t) \in \mathbb{R}^m$ is an m-dimensional dynamic variable describing the internal state of the system, and $f : \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}^m$ is a nonlinear function.

From a physical point of view, Equation 1 is Ohm's law, which describes any nonlinear memory resistor, regardless of the nature of the nonlinearity and the mechanism of resistance change. Thus, the generalized definition (1) of the memristive effect applies to the description of resistive switching in any materials: inorganic (Ohno et al., 2011), organic (Demin et al., 2014), molecular (Goswami et al., 2020), etc. Various

physical and chemical phenomena, including ion migration and redox reactions, ferroelectric and magnetoresistive effects, and phase transitions, can be responsible for the change in the resistance of inorganic materials and structures. Wang et al. (2020) provided a detailed comparison of different resistive switching mechanisms and concluded that resistive random-access memory (RRAM) are superior in terms of dimension (nanometer-scale), number of distinguishable resistive states (>64), switching speed (picoseconds), endurance (10^{12} cycles), and retention (10^3 years).

The metal-oxide-metal structures of the RRAM type (**Figure 5A**) are the most compatible materials to be integrated into the conventional CMOS process (Ielmini and Waser, 2016). Such devices can store Boolean values given by the conductivity and allow it to be changed in the same physical place, implementing new "non-von Neumann" paradigms of in-memory computation (Papandroulidakis et al., 2017; Erokhin, 2020; Lee S. H. et al., 2020). It is provided by the typical current-voltage characteristics with a pinched hysteresis (**Figure 5B**). It exhibits a wide range of resistances, as well as the pronounced and inherent stochastic nature of the conductance switching in memristors. The change in conductivity of a memristive device in response to spiking activity is analogous to the plasticity of a biological synapse and is usually described by the STDP rule (**Figure 5C**; Zamarreño-Ramos et al., 2011; Emelyanov et al., 2019; Demin et al., 2021).

The simple two-terminal structure of the memristor enables the building of superdense and, in future, three-dimensional "crossbar" arrays (**Figure 5A**). Based on Ohm's and Kirchhoff's laws, such arrays naturally implement analog operations of matrix-vector and matrix-matrix products, underlying the massive computations in traditional ANNs (Xia and Yang, 2019; Mehonic et al., 2020). Recently, using an analog-digital platform, it has been shown that a memristive crossbar can perform analog operations while digital circuits control the crossbar and enable writing synaptic weights into them (Bayat et al., 2018; Cai et al., 2019; Wang et al., 2019; Yao et al., 2020; Zahari et al., 2020). Thus, hardware-based ANN algorithms for learning and operating have been implemented based on memristors. They can significantly improve the parameters of neuromorphic computing systems, which have been actively developed in recent years due to new applications, algorithms, and element base (Indiveri et al., 2011; Schuman et al., 2022).

Memristor as a Key Element in Building Reflective Spiking Neural Networks

Let us now discuss the rich dynamics of memristive systems and present some examples within the framework of the above-mentioned conceptual approaches. The universal description of the memristor system expressed in Equation 1, hides a plethora of sound effects that yield various functional applications of memristors (**Figure 6**). The function $f(x, V)$ plays a central role in the dynamics of a memristive system and determines the complexity of the internal state of the system (Pershin and Slipko, 2019a). Moreover, the function $f(x, V)$ can include both internal and external noise, making it possible to describe a memristor as a stochastic system (Agudov et al., 2020).

Contrary to a popular belief and the standard approach focused on studying the state equation with a linear function, $f(x, V)$ plays a decisive role in achieving the complex dynamics of a memristive system. Moreover, for complex dynamics, the memristance $R(x, V)$ should be a nonlinear and a nonseparable function of its variables (Guseinov et al., 2021a). It yields the condition $R(x, V) \neq g(x)p(V)$. Different combinations of these desired properties enable simple or arbitrarily complex behaviors of real memristive devices.

Among the simple examples that can be obtained by using the first-order linear models, we can mention the widespread Hebbian plasticity described by the STDP rule (**Figure 5C**). It can be achieved by overlapping signals from pre- and postsynaptic neurons applied to a memristor (Zamarreño-Ramos et al., 2011; Emelyanov et al., 2019; Demin et al., 2021).More complex versions of plasticity, for example, frequency-dependent, require at least two dynamic variables operating at different time scales (Du et al., 2015; Kim et al., 2015; Matsukatova et al., 2020). Three state variables yield a neuron-like activity of a memristive device based on a volatile type of resistive switching (Kumar et al., 2020).

It is worth noting that the rich dynamics of memristive devices allows for going beyond the conditional rules of plasticity. We can build neural networks from the first principles based on the self-organization of adaptive memristive connections and the synchronization of neurons coupled by memristive devices. The coupling of neurons by the stochastic plasticity in memristive connections has been illustrated experimentally for several neurons in an ensemble (Ignatov et al., 2016, Ignatov et al., 2017; Gerasimova et al., 2017). The experimentally observed complex dynamics of memristively connected neurons requires description using high-order dynamical models to design larger brain-like cognitive systems (Gerasimova et al., 2021).

The use of a nonlinear potential function describing the state of memristors leads to the appearance of different types of attractors in the state space, which drives the dynamic characteristics of the memristors (Pershin and Slipko, 2019a,b). The multidimensionality of this space, combined with nonlinear and nonseparable memristance, provides the necessary and sufficient conditions for observing the complex dynamics of the memristor response to external periodic stimulation (Guseinov et al., 2021a). The corresponding transition from periodic response modes to intermittency and chaos can partially explain the variability in the parameters of real memristive devices (Guseinov et al., 2021b).

Stochasticity is an intrinsic property of a memristor (Carboni and Ielmini, 2019). Noise can be used both to study the multistable nature and to control the behavior, thanks to such well-known effects as stochastic resonance and enhancement of the stability of metastable states (Mikhaylov et al., 2021), resonant activation (Ryabova et al., 2021), etc. These and other phenomena related to the constructive role of noise can be described within the framework of analytical stochastic models (Agudov et al., 2020, 2021). They are well suited for design at the circuit level.

Thus, the presented range of functional capabilities of memristive systems already makes it possible to implement SNN architectures in hardware. Although memristor-based SNNs

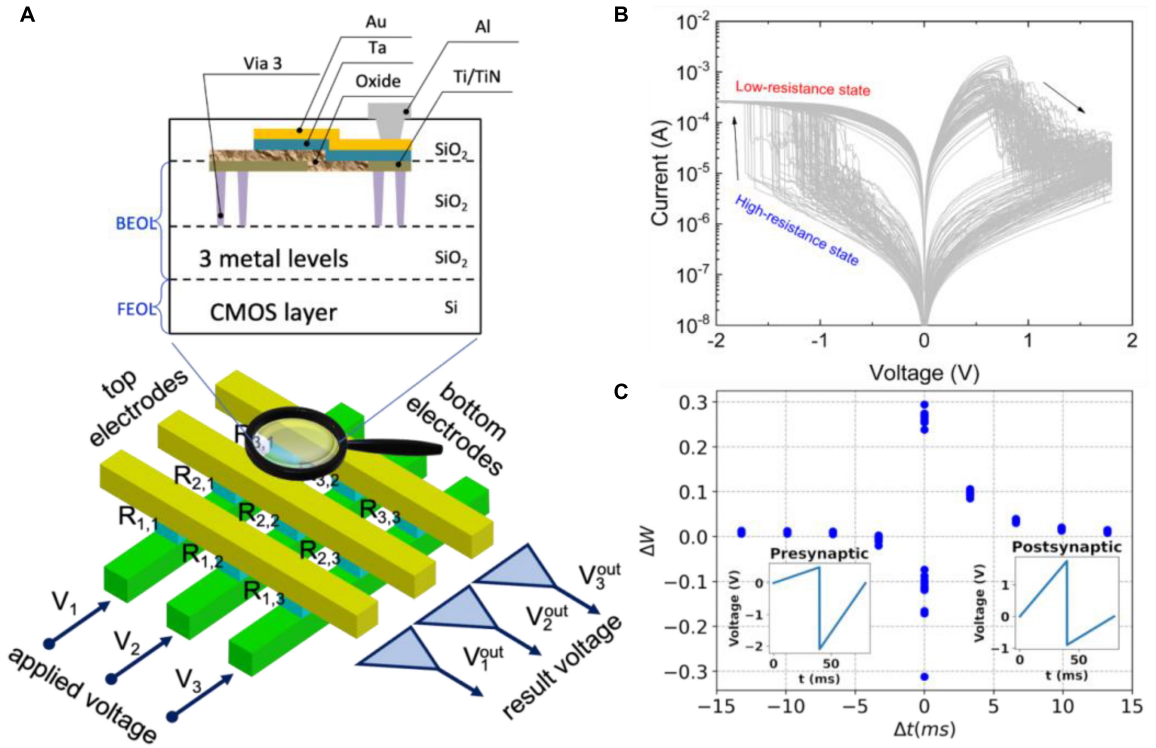


FIGURE 5 | Memristive systems. **(A)** An array of crossbar metal-oxide-metal memristive devices integrated into the top metallization layers back-end-of-line (BEOL) of the CMOS layer front-end-of-line (FEOL). **(B)** Typical current-voltage characteristics of a memristive device with stochastic switching between low- and high-resistance states. **(C)** Synaptic functionality of the memristive device mimicking the STDP rule.

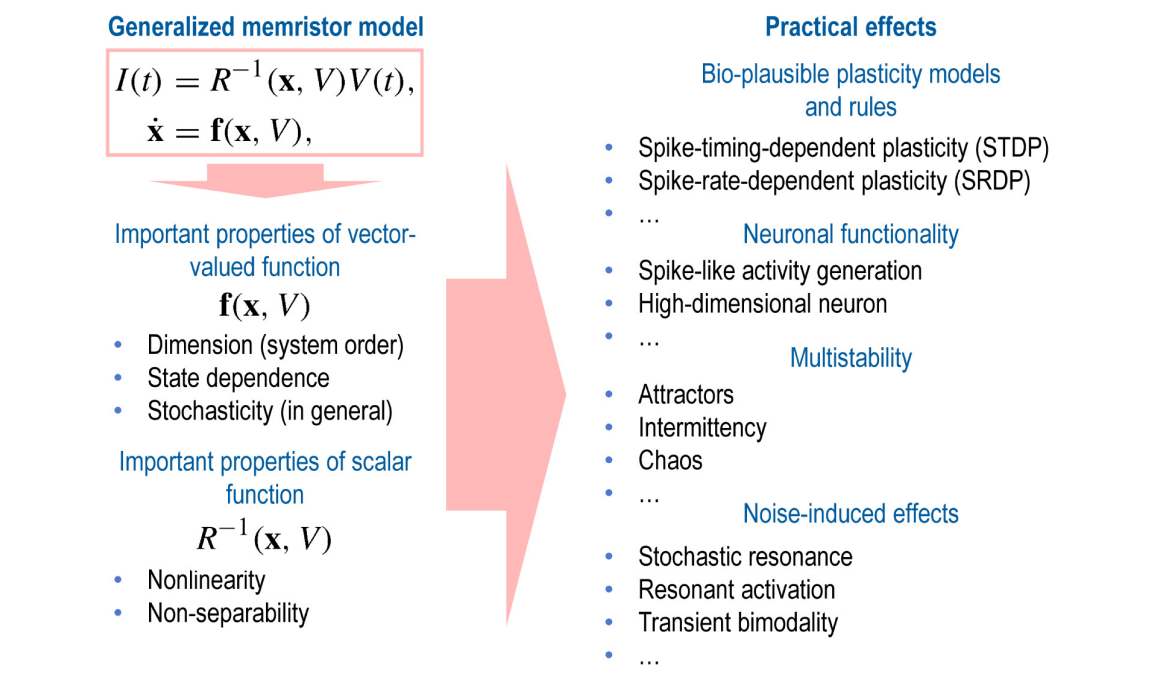


FIGURE 6 | Schematic illustration of the variety of practical effects hidden in the basic properties of the generalized memristor model.

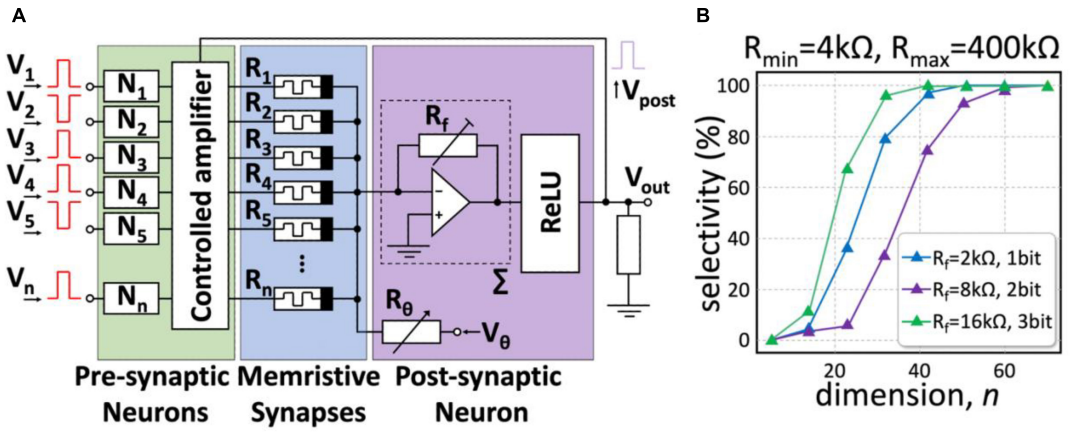


FIGURE 7 | The concept of a memristive HD neuron. **(A)** Hardware implementation of an HD neuron. A neuronal membrane (purple) is an inverting adder receiving input from the synaptic connections made of memristors $R_1, ..., R_n$ (blue). The feedback loop enhances couplings of pre- and post-neurons under learning through the controlled amplifier (green). **(B)** The result of simulations of the electric circuit. As the HD-theory predicts, the neuronal selectivity steeply increases with the dimension and attains 100% for $n = 50 - 70$.

have already been developed and even tested in crossbars (Ankit et al., 2017; Prezioso et al., 2018; Demin et al., 2021), they roughly simulate STDP to implement local learning rules. However, STDP does not cover the whole variety of biochemical processes and describes only one of the mechanisms determining synaptic plasticity (Feldman, 2012). Moreover, the memristive STDP models use a simplified algorithm based on a temporal overlap of pre- and postsynaptic spikes at a millisecond time scale (Demin et al., 2021). They essentially can be reduced to the direct programming of the memristor resistive state. Such an approach significantly complicates the electric circuits of the developed SNNs and compromises their energy efficiency and performance. However, it is still relevant for building small-sized demonstration prototypes. We foresee further implementations of various mechanisms of synaptic plasticity, multistability, and stochasticity at the complexity level critical to building perfect systems from imperfect elements. At the same time, the immature memristive technology cannot currently meet the constantly growing requirements for ANNs from developing digital services. Large-scale crossbar arrays suffer from several parasitic effects.

In the section below, we overview two approaches that should reveal the potential of memristive devices in reflective ("thinking") information and computing systems. They are being developed as alternatives to the standard "digital" approach based on programming the states of memristive devices as customarily done in traditional electronics. The first approach aims at creating self-learning SNNs based on the rich dynamics of memristive devices and simple architectures, using elegant and efficient solutions prompted by nature and corresponding to the well-known principle of simplicity in neurosciences (refer to Section "Novel Mathematical Principles for Spiking Neural Networks: Concept Cells and High-Dimensional Brain"). The second approach proceeds by completely rejecting digital algorithms and implies direct (on-site or at the edge) processing of analog information from outside. It aims to effectively implement such perception functions as vision, hearing, etc.

Memristive High-Dimensional Neurons as Building Blocks for Artificial Cognitive Systems

The possibility of a mathematical description and hardware implementation of synaptic functions based on a memristive device enables implementation of even the most daring mathematical concepts in hardware. Recent advances use simplified architectures of neurons and include the concept of the high-dimensional brain (Section "Novel Mathematical Principles for Spiking Neural Networks: Concept Cells and High-Dimensional Brain"), which explains the unreasonable efficiency of single cognitively specialized neurons. The system consists of software and hardware parts and is controlled by a microcontroller (Shchanikov et al., 2021). Memristive devices based on a metal-oxide-metal thin-film structure, where yttrium-stabilized zirconium dioxide acts as a switching medium, can implement adaptable synaptic weights of a high-dimensional neuron (Mikhaylov et al., 2020).

Figure 7A illustrates an electric circuit implementing a high-dimensional (HD) neuron. An HD input vector pattern encoded by bipolar pulses $v = (V_1, ..., V_n)^T$ is fed to the circuit input. The resistances of the memristive devices $R_1, ..., R_n$ determine the weights of the synaptic connections, and their combination for a particular neuron determines the neuron selectivity as discussed in Section "Novel Mathematical Principles for Spiking Neural Networks: Concept Cells and High-Dimensional Brain". Then, an inverting adder implements the main functionality of the neuronal membrane by integrating n informational channels and the membrane threshold, V_θ. Such a neuron performs mathematical operations of multiplication and addition following Ohm's and Kirchhoff's laws.

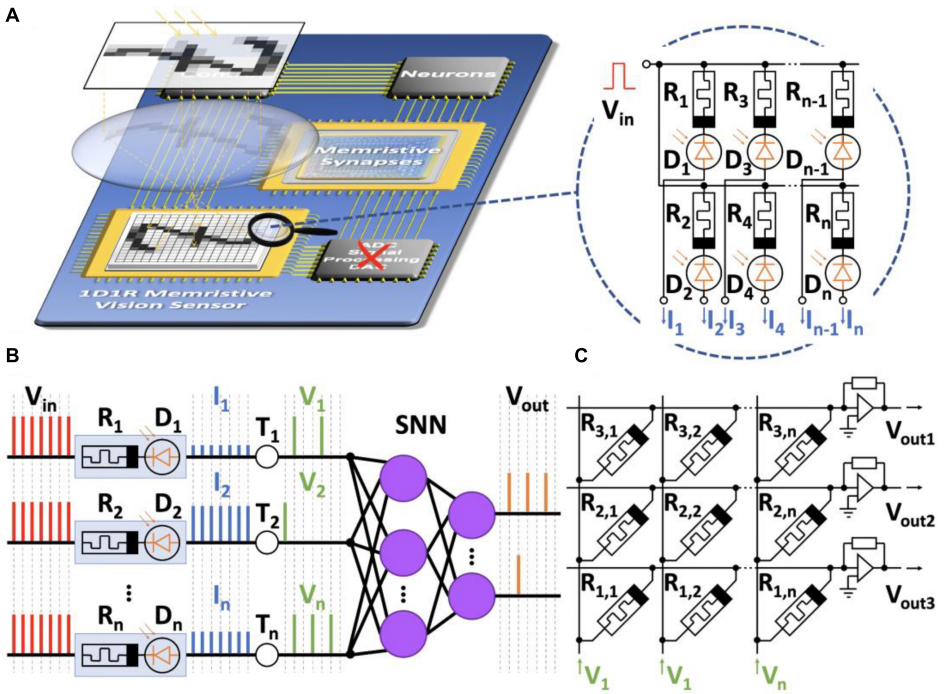


FIGURE 8 | The concept of analog memristive vision. **(A)** Schematic representation of the vision system. It includes a 1D1R memristive sensor capturing images, synaptic connections (memristive crossbar), and an SNN. There is no need for an analog-to-digital converter (red cross). **(B)** Process of transformation of spikes in the memristive SNN. **(C)** A layer of memristive neurons implemented by an array of memristors in a crossbar.

The proposed analog circuit implementation of an HD neuron is simple but, at the same time, allows for simulating the concept described in Section "Novel Mathematical Principles for Spiking Neural Networks: Concept Cells and High-Dimensional Brain". **Figure 7B** shows the result of simulations of the operational performance of a memristive HD neuron with memristive devices working in the resistance range of $4 - 400$ kOhm. At high dimensions, the neuron exhibits absolute selectivity to the input stimuli. Such neurons can be used to implement a variety of cognitive behaviors (Tyukin et al., 2019; Calvo Tapia et al., 2020b; see also Section "Novel Mathematical Principles for Spiking Neural Networks: Concept Cells and High-Dimensional Brain"). The switching dimension depends on the circuit components and starts at $n = 50$. By adjusting the value of R_f, we can achieve absolute selectivity even for the relatively low resolution of neuron weights (**Figure 7B**).

Learning the proposed HD neuron is automatic and goes the following way (Calvo Tapia et al., 2020b). At the beginning, the memristors have arbitrary initial resistances, which means that the neuron has some combination of the synaptic weights. Then, we supply to the neuron a sequence of input data vectors encoded by the inverted voltage amplitudes of the input signal v. The maximal amplitude must not exceed the switching voltage of the memristors V_{th} to maintain the original combination of the resistances. At the presentation of a certain input vector, the ReLU output will become positive, which means the neuron has detected the vector. Then, following the Hebbian rule, the coupling of pre- and postsynaptic neurons is strengthened. It is achieved by setting voltages at the inputs of the neuron required to increase or decrease the resistance of the corresponding memristors in the range $[R_{min}, R_{max}]$ for positive and negative inputs, respectively.

To train an HD neuron at the hardware level, we can use noninverting operational amplifiers with controlled gain (controlled amplifier in **Figure 7A**). The gain is adjusted by adding a load to the feedback of the operational amplifier when the voltage-controlled switch is opened. In the feedback loop, the voltage V_{post} is set at the ReLU output only when the neuron detects an input pattern, and this pulse opens the keys and increases the amplitude of the input pulses $V_1, \ldots, V_n$. In turn, it causes a change in the resistance of $R_1, \ldots, R_n$, and in the strength of the synaptic connections.

Bio-Inspired Analog Signal Processing Enabled by Memristors

According to the second alternative approach, memristive devices and SNNs may also facilitate the implementation of neuromorphic analog machine vision systems. HP Labs and the University of Berkeley have shown one of the first implementations of an ANN with memristive devices used for pattern recognition. Bayat et al. (2018) described a device based on passive crossbars with 20×20 memristors, which implements a multilayer feed-forward perceptron capable of recognizing Latin alphabet letters with 97% accuracy.

A publicly available simulator of the human retina (Eshraghian et al., 2019; Baek et al., 2020) can be used to develop advanced analog vision systems. Based on computing systems with memristor chips, a Hopfield ANN and a convolutional ANN were implemented and tested in pattern recognition tasks and associative memory (Zhou et al., 2019; Li et al., 2020; Yao et al., 2020). It has been shown that the implementation of ANNs on memristive devices of the size 128 × 64 is several times faster than graphics and signal processors in terms of speed and lower power consumption (Li et al., 2018a,b). In general, the results of comparing memristive devices with modern systems of a hardware implementation of ANNs show their advantages in accuracy, speed, power consumption, etc. (Xia and Yang, 2019; Amirsoleimani et al., 2020; Lee S. H. et al., 2020; Qin et al., 2020).

At the same time, the need for analog-to-digital and digital-to-analog conversions minimizes the potential energy gain from using memristors in traditional architectures (Amirsoleimani et al., 2020). Memristive devices allow for creating neuromorphic systems in which all processing takes place in an analog form. Thus, it seems reasonable to exclude analog-to-digital and digital-to-analog conversions from machine vision systems. The signals from the photosensor can be fed to an SNN without digitization. Then, the conductivities of the memristors will shape the model of visual information processing and simultaneously perform this processing (in-sensor computing).

The first steps have already been taken to combine memristive devices with photosensors. The described architecture of a 1D1R sensor for machine vision is a 20 × 20 or 32 × 32 matrix of SiN_x memristive devices coupled to a photodiode or a phototransistor (Vasileiadis et al., 2021a,b). The coupling of memristors with photosensors shows that this approach can simulate some retinal functions (Chen et al., 2018; Eshraghian et al., 2018). Adding such photosensors to layers of SNNs based on memristors may allow for the implementation of the concept of analog machine vision.

Figure 8 illustrates the concept of analog memristive vision exploiting coupled memristors and photodiodes (Vasileiadis et al., 2021a,b). The 1D1R memristive sensor receives visual information (**Figure 8A**). The sensor is a photodetector consisting of photodiodes $D_1, ..., D_n$ connected to a voltage source V_{in} and memristors of resistances $R_1, ..., R_n$. The voltage source forms spikes at the input of the SNN. After exposure, the memristors change their resistances depending on the illumination. Therefore, a different voltage drop will occur in each input channel when voltage pulses are applied. Then, the first SNN layer consisting of integrate-and-fire neurons fires spikes with frequencies depending on the resistance of the memristors and the thresholds $T_1, ..., T_n$ (**Figure 8B**).

Thus, visual information can be encoded by analog spikes without analog-to-digital conversions and transmitted directly to the input of the memristive SNN. The main element of the SNN is the memristive crossbar (**Figure 8C**). Memristors in the crossbar can change their conductivities and play the role of synapses. Since spikes come at different frequencies at the input, the STDP model can be used in the SNN to implement local learning rules.

We note that the concept of a high-dimensional brain and analog machine vision complement each other and may bring this area to a qualitatively new level. Although we have described only the simplest selective effect emerging in HD neurons, more complex architecture (see, e.g., Calvo Tapia et al., 2020b) are ready to be implemented in memristive architectures and SNNs.

CONCLUSION

In recent decades, SNNs have increasingly gained attention. This study has provided an overview of current theoretical, computational, and hardware approaches to building reflective SNNs. Some of the discussed problems, such as learning in SNNs, are unsolved and require new efforts from the scientific community. The synergy between neuroscience and mathematical approaches can be a solution for building novel systems demonstrating reflective AI.

Current neural networks usually deal with the abstraction of "static" stimuli (objects, persons, landscapes, or even speech). The abstraction of actions and behaviors is a great challenge that should be addressed in the future. Some of the proposals argue that it can be done through a specific type of internal representation (Calvo Tapia et al., 2020c) or through building motor motifs (Calvo Tapia et al., 2018). Now our knowledge about higher echelons of information processing in the brain is limited. There is no clear evidence on how biological neurons represent spatiotemporal concepts and end up with cognition. However, it likely happens in an active manner through a constant interplay between the intrinsic brain dynamics and external input.

The theory of the HD brain, based on the measure concentration phenomena, suggests that individual neurons can become "intelligent" through a series of quantum leaps if the complexity of information they process grows. It helps explain that a cognitive phenomenon is not a linear combination of component functions. Adding up components increases the system dimension, and at some key points, novel faculties emerge. These advances suggest that learning in higher brain stations can be majorly local, and different versions of Hebbian rules, e.g., STDP, can be behind various cognitive phenomena.

The hardware friendliness of SNNs has stimulated the search for methods of their implementation in low-power hardware devices. We foresee that memristive technology is a strong candidate for a breakthrough in this area. The review has discussed recent successful attempts to reproduce synaptic plasticity and implement in-memory/in-sensor computations. Together with SNNs and the theory of the high-dimensional brain, the latter can produce novel approaches to neuromorphic computing. Then, SNNs can diverge from the development of ANNs and build their niche, cognitive, or reflective computations. The energetic efficiency and computational speed of future devices will be significantly improved. In turn, it may allow for overcoming the heat and memory walls that the current CMOS technology is facing.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported by the Russian Science Foundation (grant No. 21-12-00246, learning in SNNs, grant No. 21-11-00280, memristive high-dimensional brain, grant No. 21-71-00136, analog computer vision), by the Russian Foundation for Basic Research (grant No. 20-01-00368, spatial neurocomputing concept), by the Ministry of Education and Science of Russia (project 075-15-2021-634 concept cells; project 074-02-2018-330 (2), Sec. 4.2), by the scientific program of the National Center for Physics and Mathematics (project "Artificial intelligence and big data in technical, industrial, natural and social systems") and by the Santander-UCM grant PR44/21.

REFERENCES

Abbott, L. F. (1999). Lapicque's introduction of the integrate-and-fire model neuron (1907). *Brain Res. Bull.* 50, 303–304. doi: 10.1016/S0361-9230(99)00161-6

Agudov, N. V., Dubkov, A. A., Safonov, A. V., Krichigin, A. V., Kharcheva, A. A., Guseinov, D. V., et al. (2021). Stochastic model of memristor based on the length of conductive region. *Chaos Solitons Fract.* 150:111131. doi: 10.1016/j.chaos.2021.111131

Agudov, N. V., Safonov, A. V., Krichigin, A. V., Kharcheva, A. A., Dubkov, A. A., Valenti, D., et al. (2020). Nonstationary distributions and relaxation times in a stochastic model of memristor. *J. Stat. Mech. Theory Exp.* 2020:24003. doi: 10.1088/1742-5468/ab684a

Alexander, D. M., Trengove, C., Sheridan, P. E., and van Leeuwen, C. (2011). Generalization of learning by synchronous waves: from perceptual organization to invariant organization. *Cogn. Neurodyn.* 5, 113–132. doi: 10.1007/s11571-010-9142-9

Altenberger, F., and Lenz, C. (2018). A non-technical survey on deep convolutional neural network architectures. *arXiv* [preprint]. Available online at: https://arxiv.org/abs/1803.02129 (accessed May 26, 2022).

Amirsoleimani, A., Alibart, F., Yon, V., Xu, J., Pazhouhandeh, M. R., Ecoffey, S., et al. (2020). In-memory vector-matrix multiplication in monolithic complementary metal–oxide–semiconductor-memristor integrated circuits: design choices, challenges, and perspectives. *Adv. Intell. Syst.* 2:2000115. doi: 10.1002/aisy.202000115

Ankit, A., Sengupta, A., Panda, P., and Roy, K. (2017). "RESPARC: a reconfigurable and energy-efficient architecture with memristive crossbars for deep spiking neural networks," in *Proceedings of the 54th Annual Design Automation Conference*, (Austin, TX: IEEE).

Araque, A., Parpura, V., Sanzgiri, R. P., and Haydon, P. G. (1999). Tripartite synapses: glia, the unacknowledged partner. *Trends Neurosci.* 22, 208–215. doi: 10.1016/s0166-2236(98)01349-6

Baek, S., Eshraghian, J. K., Thio, W., Sandamirskaya, Y., Iu, H. H., and Lu, W. D. (2020). "Live demonstration: video-to-spike conversion using a real-time retina cell network simulator," in *Proceedings of the 2020 2nd IEEE Int. Conf. Artif. Intell. Circuits System (AICAS)*, (Piscataway. NJ: IEEE), 131–131.

Barlow, H. B. (1961). "Possible principles underlying the transformation of sensory messages," in *Sensory Communication*, ed. S. Ferne (Cambridge, MA: MIT Press).

Bayat, F. M., Prezioso, M., Chakrabarti, B., Nili, H., Kataeva, I., and Strukov, D. (2018). Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits. *Nat. Commun.* 9, 1–7. doi: 10.1038/s41467-018-04482-4

Bellman, R. (1957). *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Benito, N., Fernandez-Ruiz, A., Makarov, V. A., Makarova, J., Korovaichuk, A., and Herreras, O. (2014). Spatial modules of coherent activity in pathway-specific LFPs in the hippocampus reflect topology and different modes of presynaptic synchronization. *Cereb. Cortex* 24, 1738–1752. doi: 10.1093/cercor/bht022

Benito, N., Martin-Vazquez, G., Makarova, J., Makarov, V. A., and Herreras, O. (2016). The right hippocampus leads the bilateral integration of gamma-parsed lateralized information. *eLife* 5:e16658. doi: 10.7554/eLife.16658.001

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). "When is "nearest neighbor" meaningful?," in *Proceedings of the 7th International Conference Database Theory (ICDT)*, (Princeton, NJ: IEEE), 217–235.

Bhat, A. A., Mahajan, G., and Mehta, A. (2011). Learning with a network of competing synapses. *PLoS One* 6:e25048. doi: 10.1371/journal.pone.0025048

Bi, G. Q., and Poo, M. M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* 18, 10464–10472. doi: 10.1523/JNEUROSCI.18-24-10464.1998

Bogue, R. (2017). Domestic robots: has their time finally come? *Ind. Robot Intern. J.* 44, 129–136.

Bohte, S. M., Kok, J. N., and Poutré, H. L. (2002). Error-backpropagation in temporally encoded networks of spiking neurons. *Neurocomputing* 48, 17–37.

Bower, J. M., and Beeman, D. (1998). *The Book of GENESIS: Exploring Realistic Neural Models with the General NEural SImulation System*, 2nd Edn. New York, NY: Springer Verlag.

Bowers, J. (2009). On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience. *Psychol. Rev.* 116, 220–251. doi: 10.1037/a0014462

Cai, F., Correll, J. M., Lee, S. H., Lim, Y., Bothra, V., Zhang, Z., et al. (2019). A fully integrated reprogrammable memristor–CMOS system for efficient multiply – accumulate operations. *Nat. Electron.* 2, 290–299. doi: 10.1038/s41928-019-0270-x

Calvo Tapia, C., Makarov, V. A., and van Leeuwen, C. (2020a). Basic principles drive self-organization of brain-like connectivity structure. *Commun. Nonlinear Sci. Numer.* 82:105065. doi: 10.1016/j.cnsns.2019.105065

Calvo Tapia, C., Tyukin, I., and Makarov, V. A. (2020b). Universal principles justify the existence of concept cells. *Sci. Rep.* 10:7889. doi: 10.1038/s41598-020-64466-7

Calvo Tapia, C., Villacorta-Atienza, J. A., Diez-Hermano, S., Khoruzhko, M., Lobov, S. A., Potapov, I., et al. (2020c). Semantic knowledge representation for strategic interactions in dynamic situations. *Front. Neurorobot.* 4:4. doi: 10.3389/fnbot.2020.00004

Calvo Tapia, C., Tyukin, I. Y., and Makarov, V. A. (2018). Fast social-like learning of complex behaviors based on motor motifs. *Phys. Rev. E* 97:052308. doi: 10.1103/PhysRevE.97.052308

Cao, Y., Chen, Y., and Khosla, D. (2015). Spiking deep convolutional neural networks for energy-efficient object recognition. *Int. J. Comput. Vis.* 113, 54–66.

Carboni, R., and Ielmini, D. (2019). Stochastic memory devices for security and computing. *Adv. Electron. Mater.* 5, 1–27. doi: 10.1002/aelm.201900198

Chater, T. E., and Goda, Y. (2021). My neighbour hetero-deconstructing the mechanisms underlying heterosynaptic plasticity. *Curr. Opin. Neurobiol.* 67, 106–114. doi: 10.1016/j.conb.2020.10.007

Chen, S., Lou, Z., Chen, D., and Shen, G. (2018). An artificial flexible visual memory system based on an UV-motivated memristor. *Adv. Mater.* 30:1705400. doi: 10.1002/adma.201705400

Chen, Y., Mai, Y., Feng, R., and Xiao, J. (2022). An adaptive threshold mechanism for accurate and efficient deep spiking convolutional neural networks. *Neurocomputing* 469, 189–197. doi: 10.1016/j.neucom.2021.10.080

Chou, T.-S., Bucci, L., and Krichmar, J. (2015). Learning touch preferences with a tactile robot using dopamine modulated STDP in a model of insular cortex. *Front. Neurorobot.* 9:6. doi: 10.3389/fnbot.2015.00006

Chua, L. O. (1971). Memristor-The missing circuit element. *IEEE Trans. Circ. Theory* 18, 507–519. doi: 10.1109/TCT.1971.1083337

Chua, L. O., and Kang, S. M. (1976). Memristive devices and systems. *Proc. IEEE* 64, 209–223. doi: 10.1109/PROC.1976.10092

Cook, N. (2008). The neuron-level phenomena underlying cognition and consciousness: synaptic activity and the action potential. *Neuroscience* 153, 556–570. doi: 10.1016/j.neuroscience.2008.02.042

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Contr. Signals Syst.* 2, 303–314.

Dai, Z., Liu, H., Le, Q. V., and Tan, M. (2021). CoAtNet: marrying convolution and attention for all data sizes. *arXiv* [preprint]. Available online at: https://arxiv.org/abs/2106.04803 (accessed May 26, 2022).

Dearnaley, G. (1970). Electrical phenomena in amorphous oxide films. *Rep. Progr. Phys.* 33:1129.

Delorme, A., Gautrais, J., van Rullen, R., and Thorpe, S. (1999). SpikeNET: a simulator for modeling large networks of integrate and fire neurons. *Neurocomputing* 26–27, 989–996. doi: 10.1016/S0925-2312(99)00095-8

Demin, V. A., and Erokhin, V. V. (2016). Hidden symmetry shows what a memristor is. *Int. J. Unconv. Comput.* 12, 433–438.

Demin, V. A., Erokhin, V. V., Kashkarov, P. K., and Kovalchuk, M. V. (2014). Electrochemical model of the polyaniline based organic memristive device. *J. Appl. Phys.* 116:064507. doi: 10.1063/1.4893022

Demin, V. A., Nekhaev, D. V., Surazhevsky, I. A., Nikiruy, K. E., Emelyanov, A. V., Nikolaev, S. N., et al. (2021). Necessary conditions for STDP-based pattern recognition learning in a memristive spiking neural network. *Neural Netw.* 134, 64–75. doi: 10.1016/j.neunet.2020.11.005

Diehl, P., and Cook, M. (2015). Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Front. Comput. Neurosci.* 9:99. doi: 10.3389/fncom.2015.00099

Diehl, P. U., Neil, D., Binas, J., Cook, M., Liu, S.-C., and Pfeiffer, M. (2015). "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," in *Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN)*, (Piscataway, NJ: IEEE), 1–8.

Dityatev, A., Schachner, M., and Sonderegger, P. (2010). The dual role of the extracellular matrix in synaptic plasticity and homeostasis. *Nat. Rev. Neurosci.* 11, 735–746. doi: 10.1038/nrn2898

Donoho, D. L. (2000). High-dimensional data analysis: the curses and blessings of dimensionality. *AMS Math Challeng. Lecture* 1:32.

Dora, S., and Kasabov, N. (2021). Spiking neural networks for computational intelligence: an overview. *Big Data Cogn. Comput.* 5:67.

Draelos, T. J., Miner, N. E., Lamb, C. C., Vineyard, C. M., Carlson, K. D., James, C. D., et al. (2017). Neurogenesis deep learning. *arXiv* [preprint]. Available online at: https://arxiv.org/abs/1612.03770 (accessed May 26, 2022).

Dreier, J. P., Fabricius, M., Ayata, C., Sakowitz, O. W., Shuttleworth, C. W., Dohmen, C., et al. (2017). Recording, analysis, and interpretation of spreading depolarizations in neurointensive care: review and recommendations of the COSBID research group. *J. Cereb. Blood Flow Metab.* 37, 1595–1625. doi: 10.1177/0271678X16654496

Du, C., Ma, W., Chang, T., Sheridan, P., and Lu, W. D. (2015). Biorealistic implementation of synaptic functions with oxide memristors through internal ionic dynamics. *Adv. Funct. Mater.* 25, 4290–4299. doi: 10.1002/adfm.201501427

Durkee, C. A., and Araque, A. (2019). Diversity and specificity of astrocyte–neuron communication. *Neuroscience* 396, 73–78. doi: 10.1016/j.neuroscience.2018.11.010

Edwards, J. (2005). Is consciousness only a property of individual cells? *J. Conscious. Stud.* 12, 60–76.

Emelyanov, A. V., Nikiruy, K. E., Demin, V. A., Rylkov, V. V., Belov, A. I., Korolev, D. S., et al. (2019). Yttria-stabilized zirconia cross-point memristive devices for neuromorphic applications. *Microelectron. Eng.* 215:110988. doi: 10.1016/j.mee.2019.110988

Erokhin, V. (2020). Memristive devices for neuromorphic applications: comparative analysis. *Bionanoscience* 10, 834–847. doi: 10.1007/s12668-020-00795-1

Eshraghian, J. K., Baek, S., Thio, W., Sandamirskaya, Y., Iu, H. H., and Lu, W. D. (2019). A real-time retinomorphic simulator using a conductance-based discrete neuronal network. *arXiv* [preprint]. Available online at: https://arxiv.org/abs/2001.05430 (accessed May 26, 2022).

Eshraghian, J. K., Cho, K., Zheng, C., Nam, M., Iu, H. H. C., Lei, W., et al. (2018). Neuromorphic vision hybrid RRAM-CMOS architecture. *IEEE Trans. Very Large Scale Integrat. Syst.* 26, 2816–2829.

Esser, S. K., Merolla, P. A., Arthur, J. V., Cassidy, A. S., Appuswamy, R., Andreopoulos, A., et al. (2016). Convolutional networks for fast, energy-efficient neuromorphic computing. *Proc. Natl. Acad. Sci. U.S.A.* 113, 11441–11446. doi: 10.1073/pnas.1604850113

Feldman, D. E. (2012). The spike-timing dependence of plasticity. *Neuron* 75, 556–571. doi: 10.1016/j.neuron.2012.08.001

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A* 4, 2379–2394. doi: 10.1364/josaa.4.002379

FitzHugh, R. (1961). Impulses and physiological states in theoretical models of nerve membrane. *Biophys. J.* 1, 445. doi: 10.1016/s0006-3495(61)86902-6

Florian, R. V. (2012). The Chronotron: a neuron that learns to fire temporally precise spike patterns. *PLoS One* 7:e40233. doi: 10.1371/journal.pone.0040233

Georgopoulos, A. P., Schwartz, A. B., and Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science* 233, 1416–1419.

Gerasimova, S. A., Belov, A. I., Korolev, D. S., Guseinov, D. V., Lebedeva, A. V., Koryazhkina, M. N., et al. (2021). Stochastic memristive interface for neural signal processing. *Sensors* 21, 1–12. doi: 10.3390/s21165587

Gerasimova, S. A., Mikhaylov, A. N., Belov, A. I., Korolev, D. S., Gorshkov, O. N., and Kazantsev, V. B. (2017). Simulation of synaptic coupling of neuron-like generators via a memristive device. *Tech. Phys.* 62, 1259–1265. doi: 10.1134/S1063784217080102

Ghosh-Dastidar, S., and Adeli, H. (2007). Improved spiking neural networks for EEG classification and epilepsy and seizure detection. *Integr. Comput. Aided. Eng.* 14, 187–212.

Gong, P., and Van Leeuwen, C. (2009). Distributed dynamical computation in neural circuits with propagating coherent activity patterns. *PLoS Comput. Biol.* 5:e1000611. doi: 10.1371/journal.pcbi.1000611

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: The MIT Press.

Gorban, A. N., Makarov, V. A., and Tyukin, I. Y. (2019). The unreasonable effectiveness of small neural ensembles in high-dimensional brain. *Phys. Life Rev.* 29, 55–88. doi: 10.1016/j.plrev.2018.09.005

Gorban, A. N., Makarov, V. A., and Tyukin, I. Y. (2020). High-dimensional brain in a high-dimensional world: blessing of dimensionality. *Entropy* 22:82. doi: 10.3390/e22010082

Gorban, A. N., Tyukin, I., Prokhorov, D., and Sofeikov, K. (2016). Approximation with random bases: pro et contra. *Inf. Sci.* 364–365, 129–145.

Gorban, A. N., and Tyukin, I. Y. (2018). Blessing of dimensionality: mathematical foundations of the statistical physics of data. *Philos. Trans. R. Soc. A* 376:20170237. doi: 10.1098/rsta.2017.0237

Gordleeva, S. Y., Ermolaeva, V. A., Kastalskiy, I. A., and Kazantsev, V. B. (2019). Astrocyte as spatiotemporal integrating detector of neuronal activity. *Front. Physiol.* 10:294. doi: 10.3389/fphys.2019.00294

Gordleeva, S. Y., Tsybina, Y. A., Krivonosov, M. I., Ivanchenko, M. V., Zaikin, A. A., Kazantsev, V. B., et al. (2021). Modeling working memory in a spiking neuron network accompanied by astrocytes. *Front. Cell. Neurosci.* 15:631485. doi: 10.3389/fncel.2021.631485

Goriounova, N. A., Heyer, D. B., Wilbers, R., Verhoog, M. B., and Giugliano, M. (2018). Large and fast human pyramidal neurons associate with intelligence. *eLife* 7:e41714. doi: 10.7554/eLife.41714

Goswami, S., Deb, D., Tempez, A., Chaigneau, M., Rath, S. P., Lal, M., et al. (2020). Nanometer-scale uniform conductance switching in molecular memristors. *Adv. Mater.* 32, 1–11. doi: 10.1002/adma.202004370

Grill-Spector, K., Weiner, K. S., Gomez, J., Stigliani, A., and Natu, V. S. (2018). The functional neuroanatomy of face perception: from brain measurements to deep neural networks. *Interface Focus* 8:20180013. doi: 10.1098/rsfs.2018.0013

Guseinov, D. V., Matyushkin, I. V., Chernyaev, N. V., Mikhaylov, A. N., and Pershin, Y. V. (2021a). Capacitive effects can make memristors chaotic. *Chaos Solitons Fract.* 144:110699. doi: 10.1016/j.chaos.2021.110699

Guseinov, D. V., Mikhaylov, A. N., and Pershin, Y. V. (2021b). The rich dynamics of memristive devices with non-separable nonlinear response. *IEEE Trans. Circ. Syst. II Express Briefs* 7747, 1–5. doi: 10.1109/TCSII.2021.3115111

Gütig, R., and Sompolinsky, H. (2006). The tempotron: a neuron that learns spike timing–based decisions. *Nat. Neurosci.* 9, 420–428. doi: 10.1038/nn1643

Hanin, B. (2019). Universal function approximation by deep neural nets with bounded width and ReLU activations. *Mathematics* 7:992.

Heitmann, S., Gong, P., and Breakspear, M. (2012). A computational role for bistability and traveling waves in motor cortex. *Front. Comput. Neurosci.* 6:67. doi: 10.3389/fncom.2012.00067

Hellrigel, S., Jarman, N., and van Leeuwen, C. (2019). Adaptive rewiring in weighted networks. *Cogn. Syst. Res.* 55, 205–218.

Herculano-Houzel, S. (2012). The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proc. Natl Acad. Sci. U.S.A.* 109, 10661–10668. doi: 10.1073/pnas.1201895109

Hodgkin, A. L., and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* 117, 500–544.

Hubel, D. H., and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195, 215–243.

Hussain, R., and Zeadally, S. (2018). Autonomous cars: research results, issues and future challenges. *IEEE Comm. Surv. Tutor.* 21, 1275–1313. doi: 10.1109/COMST.2018.2869360

Ielmini, D., and Waser, R. (2016). *Resistive Switching: From Fundamentals of Nanoionic Redox Processes to Memristive Device Applications*. Weinheim: WILEY-VCH.

Ignatov, M., Hansen, M., Ziegler, M., and Kohlstedt, H. (2016). Synchronization of two memristively coupled van der Pol oscillators. *Appl. Phys. Lett.* 108. doi: 10.1063/1.4942832

Ignatov, M., Ziegler, M., Hansen, M., and Kohlstedt, H. (2017). Memristive stochastic plasticity enables mimicking of neural synchrony: memristive circuit emulates an optical illusion. *Sci. Adv.* 3:e1700849. doi: 10.1126/sciadv.1700849

Imagenet (2022). Available online at: https://paperswithcode.com/sota/image-classification-on-imagenet (accessed July 01, 2022).

Indiveri, G., Linares-Barranco, B., Hamilton, T. J., van Schaik, A., Etienne-Cummings, R., Delbruck, T., et al. (2011). Neuromorphic silicon neuron circuits. *Front. Neurosci.* 5:73. doi: 10.3389/fnins.2011.00073

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Trans. Neural Networks* 14, 1569–1572.

Izhikevich, E. M. (2005). *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. Cambridge, MA: MIT press.

Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cereb. Cortex* 17, 2443–2452. doi: 10.1093/cercor/bhl152

James, W. (1890). "The mind-stuff theory," in *The Principles of Psychology*, ed. W. James (New York, NY: Henry Holt and Co), 145–182.

Jo, S. H., Chang, T., Ebong, I., Bhadviya, B. B., Mazumder, P., and Lu, W. (2010). Nanoscale memristor device as synapse in neuromorphic systems. *Nano Lett.* 10, 1297–1301. doi: 10.1021/nl904092h

Kazantsev, V., Gordleeva, S., Stasenko, S., and Dityatev, A. (2012). A homeostatic model of neuronal firing governed by feedback signals from the extracellular matrix. *PLoS One* 7:e41646. doi: 10.1371/journal.pone.0041646

Keane, A., and Gong, P. (2015). Propagating waves can explain irregular neural dynamics. *J. Neurosci.* 35, 1591–1605. doi: 10.1523/JNEUROSCI.1669-14.2015

Keck, T., Hübener, M., and Bonhoeffer, T. (2017). Interactions between synaptic homeostatic mechanisms: an attempt to reconcile BCM theory, synaptic scaling, and changing excitation/inhibition balance. *Curr. Opin. Neurobiol.* 43, 87–93. doi: 10.1016/j.conb.2017.02.003

Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.* 53, 5455–5516.

Kim, J., Pershin, Y. V., Yin, M., Datta, T., and Di Ventra, M. (2020). An experimental proof that resistance-switching memory cells are not memristors. *Adv. Electron. Mater.* 6, 1–6. doi: 10.1002/aelm.202000010

Kim, S., Du, C., Sheridan, P., Ma, W., Choi, S., and Lu, W. D. (2015). Experimental demonstration of a second-order memristor and its ability to biorealistically implement synaptic plasticity. *Nano Lett.* 15, 2203–2211. doi: 10.1021/acs.nanolett.5b00697

Koch, C., and Segev, I. (1999). *Methods in Neuronal Modeling: From Ions to Networks*, 2nd Edn. Cambridge, MA: MIT Press.

Kreinovich, V., and Kosheleva, O. (2021). Limit theorems as blessing of dimensionality: neural-oriented overview. *Entropy* 23:501. doi: 10.3390/e23050501

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neur. Inf. Proces. Syst.* 25, 1097–1105.

Kumar, S., Williams, R. S., and Wang, Z. (2020). Third-order nanocircuit elements for neuromorphic engineering. *Nature* 585, 518–523. doi: 10.1038/s41586-020-2735-5

Kutter, E. F., Bostroem, J., Elger, C. E., Mormann, F., and Nieder, A. (2018). Single neurons in the human brain encode numbers. *Neuron* 100, 753–761. doi: 10.1016/j.neuron.2018.08.036

Laskar, M. N. U., Giraldo, L. G. S., and Schwartz, O. (2018). Correspondence of deep neural networks and the brain for visual textures. *arXiv* [preprint]. Available online at: https://arxiv.org/abs/1806.02888 (accessed May 26, 2022).

Lazarevich, I., Stasenko, S., Rozhnova, M., Pankratova, E., Dityatev, A., and Kazantsev, V. (2020). Activity-dependent switches between dynamic regimes of extracellular matrix expression. *PLoS One* 15:e0227917. doi: 10.1371/journal.pone.0227917

Lebedev, A. E., Solovyeva, K. P., and Dunin-Barkowski, W. L. (2020). "The large-scale symmetry learning applying Pavlov principle," in *Proceedings of the International Conference on Neuroinformatics*, (Cham: Springer), 405–411.

Lebedev, M. A., and Nicolelis, M. A. L. (2017). Brain-machine interfaces: from basic science to neuroprostheses and neurorehabilitation. *Physiol. Rev.* 97, 767–837. doi: 10.1152/physrev.00027.2016

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1, 541–551.

Ledoux, M. (2005). "The concentration of measure phenomenon," in *Mathematical Surveys & Monographs*, ed. N. H. Mustafa (Providence, RI: AMS).

Lee, C., Sarwar, S. S., Panda, P., Srinivasan, G., and Roy, K. (2020). Enabling spike-based backpropagation for training deep neural network architectures. *Front. Neurosci.* 14:119. doi: 10.3389/fnins.2020.00119

Lee, E. K., Gerla, M., Pau, G., Lee, U., and Lim, J. H. (2016). Internet of vehicles: from intelligent grid to autonomous cars and vehicular fogs. *Int. J. Distrib. Sensor Netw.* 12:1550147716665500.

Lee, S. H., Zhu, X., and Lu, W. D. (2020). Nanoscale resistive switching devices for memory and computing applications. *Nano Res.* 13, 1228–1243.

Legenstein, R., Naeger, C., and Maass, W. (2005). What can a neuron learn with spike-timing-dependent plasticity? *Neural Comput.* 17, 2337–2382. doi: 10.1162/0899766054796888

Lennie, P. (2003). The cost of cortical computation. *Curr. Biol.* 13, 493–497. doi: 10.1016/S0960-9822(03)00135-0

Li, C., Hu, M., Li, Y., Jiang, H., Ge, N., Montgomery, E., et al. (2018a). Analogue signal and image processing with large memristor crossbars. *Nat. Electron.* 1, 52–59. doi: 10.1038/s41928-017-0002-z

Li, C., Belkin, D., Li, Y., Yan, P., Hu, M., Ge, N., et al. (2018b). Efficient and self-adaptive in-situ learning in multilayer memristor neural networks. *Nat. Commun.* 9, 7–14. doi: 10.1038/s41467-018-04484-2

Li, X., Tang, J., Zhang, Q., Gao, B., Yang, J. J., Song, S., et al. (2020). Power-efficient neural network with artificial dendrites. *Nat. Nanotechnol.* 15, 776–782. doi: 10.1038/s41565-020-0722-5

Lin, Y., Li, J., Lin, M., and Chen, J. (2014). A new nearest neighbor classifier via fusing neighborhood information. *Neurocomputing* 143, 164–169.

Lindsay, G. W., Rigotti, M., Warden, M. R., Miller, E. K., and Fusi, S. (2017). Hebbian learning in a random network captures selectivity properties of prefrontal cortex. *J. Neurosci.* 37, 11021–11036. doi: 10.1523/JNEUROSCI.1222-17.2017

Lobov, S., Simonov, A., Kastalskiy, I., and Kazantsev, V. (2016). Network response synchronization enhanced by synaptic plasticity. *Eur. Phys. J. Spec. Top.* 225, 29–39.

Lobov, S. A., Chernyshov, A. V., Krilova, N. P., Shamshin, M. O., and Kazantsev, V. B. (2020a). Competitive learning in a spiking neural network: towards an intelligent pattern classifier. *Sensors* 20:500. doi: 10.3390/s20020500

Lobov, S. A., Mikhaylov, A. N., Shamshin, M., Makarov, V. A., and Kazantsev, V. B. (2020b). Spatial properties of STDP in a self-learning spiking neural network enable controlling a mobile robot. *Front. Neurosci.* 14:88. doi: 10.3389/fnins.2020.00088

Lobov, S. A., Mikhaylov, A. N., Berdnikova, E. S., Makarov, V. A., and Kazantsev, V. B. (2021a). Spatial computing in structured spiking neural networks with a robotic embodiment. *arXiv* [preprint]. Available online at: https://arxiv.org/abs/2112.07150 doi: 10.48550/arXiv.2112.07150 (accessed May 26, 2022).

Lobov, S. A., Zharinov, A. I., Makarov, V. A., and Kazantsev, V. B. (2021b). Spatial memory in a spiking neural network with robot embodiment. *Sensors* 21:2678. doi: 10.3390/s21082678

Lobov, S. A., Zharinov, A. I., Semenova, O., and Kazantsev, V. B. (2021c). "Topological classification of population activity in spiking neural network,"

in *Proceedings of the Saratov Fall Meeting 2020: Computations and Data Analysis: from Molecular Processes to Brain Functions (SPIE)*, ed. D. E. Postnov (Bellingham: SPIE).

Lobov, S. A., Zhuravlev, M. O., Makarov, V. A., and Kazantsev, V. B. (2017). Noise enhanced signaling in STDP driven spiking-neuron network. *Math. Model. Nat. Phenom.* 12, 109–124.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. (2017). The expressive power of neural networks: a view from the width. *Int. Adv. Neural Inf. Proc. Syst.* 30, 6231–6239.

Mackenzie, A. (2013). Programming subjects in the regime of anticipation: software studies and subjectivity. *Subjectivity* 6, 391–405.

Makarov, V. A., and Villacorta-Atienza, J. A. (2011). "Compact internal representation as a functional basis for protocognitive exploration of dynamic environments," in *Recurrent Neural Networks for Temporal Data Processing*, ed. H. Cardot (London: Intech).

Makarova, J., Makarov, V. A., and Herreras, O. (2010). Generation of sustained field potentials by gradients of polarization within single neurons: a macroscopic model of spreading depression. *J. Neurophysiol.* 103, 2446–2457. doi: 10.1152/jn.01045.2009

Markram, H., Lübke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* 275, 213–215. doi: 10.1126/science.275.5297.213

Matsukatova, A. N., Emelyanov, A. V., Minnekhanov, A. A., Nesmelov, A. A., Vdovichenko, A. Y., Chvalun, S. N., et al. (2020). Resistive switching kinetics and second-order effects in parylene-based memristors. *Appl. Phys. Lett.* 117:243501. doi: 10.1063/5.0030069

McCulloch, W. S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5, 115–133.

McKennoch, S., Liu, D., and Bushnell, L. G. (2006). "Fast modifications of the SpikeProp algorithm," in *Proceedings of the 2006 IEEE International Joint Conference on Neural Networks*, (Vancouver, BC: IEEE), 3970–3977.

Mehonic, A., Sebastian, A., Rajendran, B., Simeone, O., Vasilaki, E., and Kenyon, A. J. (2020). Memristors-from in-memory computing, deep learning acceleration, and spiking neural networks to the future of neuromorphic and bio-inspired computing. *Adv. Intell. Syst.* 2:2000085. doi: 10.1002/aisy.202000085

Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., et al. (2014). A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* 345, 668–673. doi: 10.1126/science.1254642

Mikhaylov, A., Pimashkin, A., Pigareva, Y., Gerasimova, S., Gryaznov, E., Shchanikov, S., et al. (2020). Neurohybrid memristive CMOS-integrated systems for biosensors and neuroprosthetics. *Front. Neurosci.* 14:358. doi: 10.3389/fnins.2020.00358

Mikhaylov, A. N., Guseinov, D. V., Belov, A. I., Korolev, D. S., Shishmakova, V. A., Koryazhkina, M. N., et al. (2021). Stochastic resonance in a metal-oxide memristive device. *Chaos Solitons Fract.* 144:110723. doi: 10.1016/j.chaos.2021.110723

Mohemmed, A., Schliebs, S., Matsuda, S., and Kasabov, N. (2012). SPAN: spike pattern association neuron for learning spatio-temporal spike patterns. *Int. J. Neural Syst.* 22:1250012. doi: 10.1142/S0129065712500128

Moravec, H. (1988). *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA: Harvard University Press.

Mormann, F., Dubois, J., Kornblith, S., Milosavljevic, M., Cerf, M., Ison, M., et al. (2011). A category-specific response to animals in the right human amygdala. *Nat. Neurosci.* 14, 1247–1249. doi: 10.1038/nn.2899

Morrison, A., Diesmann, M., and Gerstner, W. (2008). Phenomenological models of synaptic plasticity based on spike timing. *Biol. Cybern.* 98, 459–478. doi: 10.1007/s00422-008-0233-1

Mostafa, H. (2018). Supervised learning based on temporal coding in spiking neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* 29, 3227–3235. doi: 10.1109/TNNLS.2017.2726060

Mozafari, M., Ganjtabesh, M., Nowzari-Dalini, A., Thorpe, S. J., and Masquelier, T. (2018). Bio-inspired digit recognition using reward-modulated spike-timing-dependent plasticity in deep convolutional networks. *arXiv* [preprint]. Available online at: https://arxiv.org/abs/1804.00227 (accessed May 26, 2022).

Muller, L., Chavane, F., Reynolds, J., and Sejnowski, T. J. (2018). Cortical travelling waves: mechanisms and computational principles. *Nat. Rev. Neurosci.* 19, 255–268. doi: 10.1038/nrn.2018.20

Naoumenko, D., and Gong, P. (2019). Complex dynamics of propagating waves in a two-dimensional neural field. *Front. Comput. Neurosci.* 13:50. doi: 10.3389/fncom.2019.00050

Neftci, E. O., Mostafa, H., and Zenke, F. (2019). Surrogate gradient learning in spiking neural networks: bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Process. Mag.* 36, 51–63. doi: 10.1109/MSP.2019.2931595

Neil, D., Pfeiffer, M., and Liu, S.-C. (2016). "Learning to be efficient: algorithms for training low-latency, low-compute deep spiking neural networks," in *Proceedings of the 31st Ann. ACM Symp. Appl. Comp. SAC '16*, (New York, NY: Association for Computing Machinery), 293–298.

Ohno, T., Hasegawa, T., Tsuruoka, T., Terabe, K., Gimzewski, J. K., and Aono, M. (2011). Short-term plasticity and long-term potentiation mimicked in single inorganic synapses. *Nat. Mater.* 10, 591–595. doi: 10.1038/nmat3054

Olshausen, B. A., and Field, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vis. Res.* 37, 3311–3325. doi: 10.1016/s0042-6989(97)00169-7

Olshausen, B. A., and Field, D. J. (2004). Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* 14, 481–487. doi: 10.1016/j.conb.2004.07.007

Palmer, J. H. C., and Gong, P. (2014). Associative learning of classical conditioning as an emergent property of spatially extended spiking neural circuits with synaptic plasticity. *Front. Comput. Neurosci.* 8:79. doi: 10.3389/fncom.2014.00079

Panda, P., Aketi, S. A., and Roy, K. (2020). Toward scalable, efficient, and accurate deep spiking neural networks with backward residual connections, stochastic softmax, and hybridization. *Front. Neurosci.* 14:653. doi: 10.3389/fnins.2020.00653

Panda, P., Allred, J. M., Ramanathan, S., and Roy, K. (2018). ASP: learning to forget with adaptive synaptic plasticity in spiking neural networks. *IEEE J. Emerg. Sel. Top. Circ. Syst.* 8, 51–64. doi: 10.1109/JETCAS.2017.2769684

Papandroulidakis, G., Vourkas, I., Abusleme, A., Sirakoulis, G. C., and Rubio, A. (2017). Crossbar-based memristive logic-in-memory architecture. *IEEE Trans. Nanotechnol.* 16, 491–501.

Perea, G., and Araque, A. (2007). Astrocytes potentiate transmitter release at single hippocampal synapses. *Science* 317, 1083–1086. doi: 10.1126/science.1144640

Pershin, Y. V., and Slipko, V. A. (2019a). Bifurcation analysis of a TaO memristor model. *J. Phys. D. Appl. Phys.* 52:505304. doi: 10.1088/1361-6463/AB4537

Pershin, Y. V., and Slipko, V. A. (2019b). Dynamical attractors of memristors and their networks. *Europhys. Lett.* 125, 1–6. doi: 10.1209/0295-5075/125/20002

Pestov, V. (2013). Is the k-NN classifier in high dimensions affected by the curse of dimensionality? *Comput. Math. Appl.* 65, 1427–1437.

Ponulak, F. (2005). *ReSuMe-New Supervised Learning Method for Spiking Neural Networks*. Pozna?: Poznan University.

Ponulak, F., and Hopfield, J. (2013). Rapid, parallel path planning by propagating wavefronts of spiking neural activity. *Front. Comput. Neurosci.* 7:98. doi: 10.3389/fncom.2013.00098

Ponulak, F., and Kasiński, A. (2010). Supervised learning in spiking neural networks with ReSuMe: sequence learning, classification, and spike shifting. *Neural Comput.* 22, 467–510. doi: 10.1162/neco.2009.11-08-901

Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. *Nat. Rev. Neurosci.* 1, 125–132. doi: 10.1038/35039062

Prezioso, M., Mahmoodi, M. R., Bayat, F. M., Nili, H., Kim, H., Vincent, A., et al. (2018). Spike-timing-dependent plasticity learning of coincidence detection with passively integrated memristive circuits. *Nat. Commun.* 9, 1–8. doi: 10.1038/s41467-018-07757-y

Qin, Y. F., Bao, H., Wang, F., Chen, J., Li, Y., and Miao, X. S. (2020). Recent progress on memristive convolutional neural networks for edge intelligence. *Adv. Intell. Syst.* 2:2000114.

Querlioz, D., Bichler, O., Dollfus, P., and Gamrat, C. (2013). Immunity to device variations in a spiking neural network with memristive nanodevices. *IEEE Trans. Nanotechnol.* 12, 288–295.

Quian Quiroga, R. (2012). Concept cells: the building blocks of declarative memory functions. *Nat. Rev. Neurosci.* 13:587. doi: 10.1038/nrn3251

Quian Quiroga, R. (2019). Akakhievitch revisited. Comment on "The unreasonable effectiveness of small neural ensembles in high-dimensional brain" by Alexander N. Gorban et al. *Phys. Life Rev.* 28, 111–114.

Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature* 435, 1102–1107. doi: 10.1038/nature03687

Quiroga, R. Q., and Panzeri, S. (2013). *Principles of Neural Coding*. Boca Raton, FL: CRC Press.

Rentzeperis, I., Laquitaine, S., and van Leeuwen, C. (2022). Adaptive rewiring of random neural networks generates convergent–divergent units. *Commun. Nonlin. Sci. Numer. Simul.* 107:106135.

Robbins, H., and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* 22:400.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 386–408. doi: 10.1037/h0042519

Ruckauer, B., Lungu, I.-A., Hu, Y., Pfeiffer, M., and Liu, S.-C. (2017). Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Front. Neurosci.* 11:682. doi: 10.3389/fnins.2017.00682

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comp. Vis.* 115, 211–252.

Ryabova, M. A., Antonov, D. A., Kruglov, A. V., Antonov, I. N., Filatov, D. O., and Gorshkov, O. N. (2021). In situ investigation of individual filament growth in conducting bridge memristor by contact scanning capacitance microscopy. *J. Phys. Conf. Ser.* 2086:012205. doi: 10.1088/1742-6596/2086/1/012205

Santello, M., Toni, N., and Volterra, A. (2019). Astrocyte function from information processing to cognition and cognitive impairment. *Nat. Neurosci.* 22, 154–166. doi: 10.1038/s41593-018-0325-8

Sattin, D., Magnani, F. G., Bartesaghi, L., Caputo, M., Fittipaldo, A. V., Cacciatore, M., et al. (2021). Theoretical models of consciousness: a scoping review. *Brain Sci.* 11, 535. doi: 10.3390/brainsci11050535

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neur. Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Schuman, C. D., Kulkarni, S. R., Parsa, M., Mitchell, J. P., Date, P., and Kay, B. (2022). Opportunities for neuromorphic computing algorithms and applications. *Nat. Comput. Sci.* 2, 10–19. doi: 10.1038/s43588-021-00184-y

Sevush, S. (2006). Single-neuron theory of consciousness. *J. Theor. Biol.* 238, 704–725. doi: 10.1016/j.jtbi.2005.06.018

Shchanikov, S., Bordanov, I., Belov, A., Korolev, D., Shamshin, M., Gryaznov, E., et al. (2021). "Memristive concept of a high-dimensional neuron," in *Proceedings of the 2021 Third IEEE Inter. Conf. Neurotechn. Neurointerf. (CNN)*, (Piscataway, NJ: IEEE), 96–99.

Sherrington, C. (1940). *Man on His Nature*. Cambridge, MA: Cambridge Univ. Press.

Shrestha, S. B., and Orchard, G. (2018). "SLAYER: spike layer error reassignment in time," in *Advances in Neural Information Processing Systems*, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Red Hook, NY: Curran Associates, Inc.).

Shrestha, S. B., and Song, Q. (2015). Adaptive learning rate of SpikeProp based on weight convergence analysis. *Neur. Netw.* 63, 185–198. doi: 10.1016/j.neunet.2014.12.001

Silva, S. M., and Ruano, A. E. (2005). "Application of Levenberg-Marquardt method to the training of spiking neural networks," in *Proceedings of the 2005 Int. Conf. Neur. Netw. Brain*, (Piscataway, NJ: IEEE), 1354–1358.

Sjöström, P. J., Turrigiano, G. G., and Nelson, S. B. (2001). Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron* 32, 1149–1164. doi: 10.1016/s0896-6273(01)00542-6

Song, S., Miller, K. D., and Abbott, L. F. (2000). Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nat. Neurosci.* 3:919.

Sporea, I., and Grüning, A. (2013). Supervised learning in multilayer spiking neural networks. *Neural Comput.* 25, 473–509.

Strukov, D. B., Snider, G. S., Stewart, D. R., and Williams, R. S. (2008). The missing memristor found. *Nature* 453, 80–83. doi: 10.1038/nature06932

Strukov, D. B., and Williams, R. S. (2009). Four-dimensional address topology for circuits with stacked multilayer crossbar arrays. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20155–20158. doi: 10.1073/pnas.0906949106

Taherkhani, A., Belatreche, A., Li, Y., Cosma, G., Maguire, L. P., and McGinnity, T. M. (2020). A review of learning in biologically plausible spiking neural networks. *Neur. Netw.* 122, 253–272. doi: 10.1016/j.neunet.2019.09.036

Taherkhani, A., Belatreche, A., Li, Y., and Maguire, L. P. (2018). A supervised learning algorithm for learning precise timing of multiple spikes in multilayer spiking neural networks. *IEEE Trans. Neural Networks Learn. Syst.* 29, 5394–5407. doi: 10.1109/TNNLS.2018.2797801

Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., and Maida, A. (2019). Deep learning in spiking neural networks. *Neur. Netw.* 111, 47–63.

Tavanaei, A., and Maida, A. S. (2015). A minimal spiking neural network to rapidly train and classify handwritten digits in binary and 10-digit tasks. *Int. J. Adv. Res. Artif. Intell.* 4, 1–8.

Teyler, T. J., and Discenna, P. (1984). The topological anatomy of the hippocampus: a clue to its function. *Brain Res. Bull.* 12, 711–719. doi: 10.1016/0361-9230(84)90152-7

Tolman, E. C., and Honzik, C. H. (1930). Introduction and removal of reward, and maze performance in rats. *Univ. Calif. Public Psychol.* 4, 257–275.

Turrigiano, G. G. (2017). The dialectic of Hebb and homeostasis. *Philos. Trans. R. Soc. B Biol. Sci.* 372:20160258. doi: 10.1098/rstb.2016.0258

Tyukin, I. Y., Gorban, A. N., Calvo, C., Makarova, J., and Makarov, V. A. (2019). High-dimensional brain: a tool for encoding and rapid learning of memories by single neurons. *Bull. Math. Biol.* 81, 4856–4888. doi: 10.1007/s11538-018-0415-5

Valdez, A. B., Papesh, M. H., Treiman, D. M., Smith, K. A., Goldinger, S. D., and Steinmetz, P. N. (2015). Distributed representation of visual objects by single neurons in the human brain. *J. Neurosci.* 35, 5180–5186.

Vasileiadis, N., Ntinas, V., Sirakoulis, G. C., and Dimitrakis, P. (2021a). In-memory-computing realization with a photodiode/memristor based vision sensor. *Materials* 14, 1–15. doi: 10.3390/ma14185223

Vasileiadis, N., Ntinas, V., Fyrigos, I., Karamani, R., Ioannou-Sougleridis, V., Normand, P., et al. (2021b). "A new 1p1r image sensor with in-memory computing properties based on silicon nitride devices," in *Proceedings of the 2021 IEEE Int. Symp. Circuits and Systems (ISCAS)*, (Piscataway, NJ: IEEE).

Villacorta-Atienza, J. A., Calvo, C., and Makarov, V. A. (2015). Prediction-for-CompAction: navigation in social environments using generalized cognitive maps. *Biol. Cybern.* 109, 307–320. doi: 10.1007/s00422-015-0644-8

Villacorta-Atienza, J. A., Calvo Tapia, C., Diez-Hermano, S., Sanchez-Jimenez, A., Lobov, S., Krilova, N., et al. (2021). Static internal representation of dynamic situations reveals time compaction in human cognition. *J. Adv. Res.* 28, 111–125. doi: 10.1016/j.jare.2020.08.008

Villacorta-Atienza, J. A., and Makarov, V. A. (2013). Wave-processing of long-scale information by neuronal chains. *PLoS One* 8:e0057440. doi: 10.1371/journal.pone.0057440

Villacorta-Atienza, J. A., Velarde, M. G., and Makarov, V. A. (2010). Compact internal representation of dynamic situations: neural network implementing the causality principle. *Biol. Cybern.* 103, 285–329. doi: 10.1007/s00422-010-0398-2

Vongehr, S., and Meng, X. (2015). The missing memristor has not been found. *Sci. Rep.* 5, 1–7. doi: 10.1038/srep11657

Wang, Z., Li, C., Song, W., Rao, M., Belkin, D., Li, Y., et al. (2019). Reinforcement learning with analogue memristor arrays. *Nat. Electron.* 2, 115–124. doi: 10.1038/s41928-019-0221-6

Wang, Z., Wu, H., Burr, G. W., Hwang, C. S., Wang, K. L., Xia, Q., et al. (2020). Resistive switching materials for information processing. *Nat. Rev. Mater.* 5, 173–195. doi: 10.1038/s41578-019-0159-3

Xia, Q., and Yang, J. J. (2019). Memristive crossbar arrays for brain-inspired computing. *Nat. Mater.* 18, 309–323.

Xin, J., and Embrechts, M. J. (2001). "Supervised learning with spiking neural networks," in *Proceedings of the IJCNN'01. Int. Joint Conf. Neur. Network (Cat. No.01CH37222)*, (Piscataway, NJ: IEEE), 1772–1777.

Yao, P., Wu, H., Gao, B., Tang, J., Zhang, Q., Zhang, W., et al. (2020). Fully hardware-implemented memristor convolutional neural network. *Nature* 577, 641–646. doi: 10.1038/s41586-020-1942-4

Yin, B., Corradi, F., and Bohté, S. M. (2021). Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks. *Nat. Mach. Intell.* 3, 905–913. doi: 10.1038/s42256-021-00397-w

Yin, J., and Yuan, Q. (2015). Structural homeostasis in the nervous system: a balancing act for wiring plasticity and stability. *Front. Cell. Neurosci.* 8:439. doi: 10.3389/fncel.2014.00439

Zahari, F., Pérez, E., Mahadevaiah, M. K., Kohlstedt, H., Wenger, C., and Ziegler, M. (2020). Analogue pattern recognition with stochastic switching binary CMOS-integrated memristive devices. *Sci. Rep.* 10:14450. doi: 10.1038/s41598-020-71334-x

Zamarreño-Ramos, C., Camuñas-Mesa, L. A., Perez-Carrasco, J. A., Masquelier, T., Serrano-Gotarredona, T., and Linares-Barranco, B. (2011). On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex. *Front. Neurosci.* 5:26. doi: 10.3389/fnins.2011.00026

Zambrano, D., Nusselder, R., Scholte, H. S., and Bohté, S. M. (2019). Sparse computation in adaptive spiking neural networks. *Front. Neurosci.* 12:987. doi: 10.3389/fnins.2018.00987

Zenke, F., and Vogels, T. P. (2021). The remarkable robustness of surrogate gradient learning for instilling complex function in spiking neural networks. *Neural Comput.* 33, 899–925. doi: 10.1162/neco_a_01367

Zhou, Y., Wu, H., Gao, B., Wu, W., Xi, Y., Yao, P., et al. (2019). Associative memory for image recovery with a high-performance memristor array. *Adv. Funct. Mater.* 29:1900155. doi: 10.1002/adfm.201900155

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.

frontiers | Frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 23 June 2022
doi: 10.3389/fncom.2022.862126

Insertion Guidance Based on Impedance Measurements of a Cochlear Electrode Array

Enver Salkim[1,2]*, *Majid Zamani*[1], *Dai Jiang*[1], *Shakeel R. Saeed*[3] and *Andreas Demosthenous*[1]

[1] Department of Electronic and Electrical Engineering, University College London (UCL), London, United Kingdom, [2] Department of Electronic and Electrical Engineering, Biomedical Device Technology Group, Muş Alparslan University, Muş, Turkey, [3] UCL Ear Institute, London, United Kingdom

The cochlear implantable neuromodulator provides substantial auditory perception to those with severe or profound impaired hearing. Correct electrode array positioning in the cochlea is one of the important factors for quality hearing, and misplacement may lead to additional injury to the cochlea. Visual inspection of the progress of electrode insertion is limited and mainly relies on the surgeon's tactile skills, and there is a need to detect in real-time the electrode array position in the cochlea during insertion. The available clinical measurement presently provides very limited information. Impedance measurement may be used to assist with the insertion of the electrode array. Using computational modeling of the cochlea, and its local tissue layers merging with the associated neuromodulator electrode array parameters, the impedance variations at different insertion depths and the proximities to the cochlea walls have been analyzed. In this study, an anatomical computational model of the temporal region of a patient is used to derive the relationship between impedance variations and the electrode proximity to the cochlea wall and electrode insertion depth. The aim was to examine whether the use of electrode impedance variations can be an effective marker of electrode proximity and electrode insertion depth. The proposed anatomical model simulates the quasi-static electrode impedance variations at different selected points but at considerable computation cost. A much less computationally intensive geometric model (∼1/30) provided comparative impedance measurements with differences of <2%. Both use finite element analysis over the entire cross-section area of the scala tympani. It is shown that the magnitude of the impedance varies with both electrode insertion depth and electrode proximity to the adjacent anatomical layers (e.g., cochlea wall). In particular, there is a 1,400% increase when the electrode array is moved very close to the cochlea wall. This may help the surgeon to find the optimal electrode position within the scala tympani by observation of such impedance characteristics. The misplacement of the electrode array within the scala tympani may be eliminated by using the impedance variation metric during electrode array insertion if the results are validated with an experimental study.

Keywords: cochlear implant, computational models, electrode proximity, impedance variation, parameterization

OPEN ACCESS

Edited by:
Ursula van Rienen,
University of Rostock, Germany

Reviewed by:
Daniel Baumgarten,
Private University for Health Sciences, Medical Informatics and Technology (UMIT), Austria
Martin Han,
University of Connecticut, United States

***Correspondence:**
Enver Salkim
e.salkim@ucl.ac.uk

Received: *25 January 2022*
Accepted: *18 May 2022*
Published: *23 June 2022*

Citation:
Salkim E, Zamani M, Jiang D, Saeed SR and Demosthenous A (2022) Insertion Guidance Based on Impedance Measurements of a Cochlear Electrode Array. Front. Comput. Neurosci. 16:862126. doi: 10.3389/fncom.2022.862126

1. INTRODUCTION

The cochlea has a vital role in generating a sense of hearing. It transforms the sound waves into mechanical vibrations of the hair cells and subsequently into electrical pulses. The pulses are transmitted to the brain through the auditory nerve to provide hearing sensation. Sensorineural hearing loss is caused by damage to the inner ear, especially the hair cells, or the dysfunction of the auditory nerve (Svirsky et al., 2004). This is a socioeconomic burden and has led to substantial constraints globally. Over 5% of the world's population suffers from hearing loss (432 million adults and 34 million children) (Kushalnagar, 2019). The available solutions are varied depending on the type of hearing loss and include hearing aids, cochlear implants (CIs), and other assistive devices (Kushalnagar, 2019). The CI is a neural prosthesis designed to restore hearing loss by electrical stimulation of the auditory nerve. Using an electrode array inserted in the scala tympani of the cochlea, the implant can deliver modulated electric stimuli directly to the residual auditory nerve fibers, thus replacing the function of the damaged hair cells (Dang, 2017; Dhanasingh and Jolly, 2017).

Over the decades, conventional surgery using CIs remains essentially unchanged and is generally considered safe and effective (Caversaccio et al., 2019). Although advancements in CI design have been reported (Hajioff, 2016; Dazert et al., 2017; Dhanasingh and Jolly, 2017), the quality of restored hearing sensation is strongly related to the quality of the CI surgery, the design of the electrode structure, and the insertion tools and techniques (Tan et al., 2013). As the electrode array is inserted mainly guided by touch, it has been reported that partial insertion, deformation of the electrode array, and even penetration of the basilar membrane can occur which prejudices the performance of hearing after implantation (Rebscher et al., 2008). Obtaining the optimum positioning of the electrode array during cochlear electrode implantation is essential for the preservation of residual hearing and improved clinical outcomes (Finley and Skinner, 2008). Misplacement of the electrode array may lead to further hearing loss and insertion trauma if the electrode array touches the cochlea wall (Holden et al., 2013; Min et al., 2013). Furthermore, if the electrode array touches sensitive layers of the cochlea such as a basilar membrane or osseous spiral lamina layers due to significant variability in the size of the human scala tympani (Skinner et al., 2002; O'Connell et al., 2016), it may lead to severe trauma. Also, it has been shown that there is a significant correlation between hearing outcomes and the correct placement of electrode arrays entirely in the scala tympani (Wanna et al., 2014). It is, therefore, important that the electrode array should be positioned accurately within the scala tympani to minimize such consequences and improve hearing outcomes (O'Connell et al., 2016).

There are some emerging concepts, such as careful surgical techniques and training, new designs of the electrode structure, and novel insertion tools (electrode arrays with softer material, pre-curved perimodiolar arrays, and Advance Off-Stylet insertion technique are some examples), that may help reduce insertion mishaps and intracochlear trauma. Surgeons have no real-time feedback about electrode status while inserting the electrode array into the cochlea (Jethanamest et al., 2010; Tan et al., 2013). It has been shown that a magnetically guided system (Clark et al., 2012) and robotic insertion can help control insertion forces by varying insertion speed (Zhang et al., 2010). These systems may reduce trauma, but real-time local position information of the electrode array in the cochlea during insertion is required. The electrode position can be monitored using medical imaging (e.g., computer tomography). While this method may help to accurately place the electrodes, it is not suitable due to the danger of radiation risk on the patient, and it is rarely done intra-operatively (Giardina et al., 2017). Alternatively, the electrode array position can be rapidly assessed from the implant at the time of implantation by electrically-evoked neural responses, electric field imaging, or impedance variations (Mens, 2007). The first method may not be reliable due to the highly variable results reported (Miller et al., 2008; Mittmann et al., 2015). Although the major error position of the electrodes in the scala tympani could be registered using electric field imaging, it was not utilized to predict the positions of the electrodes in the scala tympani (Vanpoucke et al., 2011).

Using impedance measurements can be a safer and more reliable method to help determine the relation of the electrodes' position to the cochlea wall during surgery (Mens, 2007; Tan et al., 2013; Newbold et al., 2014). It has been shown that perilymph (fluid in the scala tympani) has relatively higher conductivity than bone and cochlea wall, leading to the hypothesis that the measured electrode impedance to ground should be higher when an electrode approaches the cochlea wall compared to when the electrode is in the middle of the scala (Frijns et al., 1995). Thus, impedance measurement can be an option to monitor the proximity of the electrode array to the cochlea wall in real-time to prevent any damage, and find the optimum position for the electrode during the insertion process (Mens, 2007; Tan et al., 2013; Newbold et al., 2014; Giardina et al., 2017).

As the human cochlea is embedded deep inside the temporal bone and there is geometrical variation in the size of the scala tympani, direct measurements of electrical potential or impedance may not be readily feasible (Bai et al., 2019). Also, using conventional techniques it may not be feasible to conduct systematic comparison across individuals to examine the precise position of the electrode (Pile et al., 2017). Computational cochlea models have been utilized to simulate the current spread in the cochlea and provided useful insights (Malherbe et al., 2016; Salkim et al., 2020). Such models are implemented using the finite element method (FEM). The models consist of a volume conductor that accounts for various anatomical structures and the inserted electrode array by their respective conductivities and appropriate boundary conditions. This study examines the relationship between electrode impedances to the ground and their proximity to adjacent layers, and their insertion depth using accurate FEM computation models. A multi-layered anatomical three-dimensional (3D) volume conductor model of the human cochlea was generated using micro-CT (μCT) datasets as shown in **Figures 1A,B**. An electrode array was generated based on the Advanced Bionics HiFocus^{TM} SlimJ electrode (Hannover, Germany) and combined with the anatomical volume conductor

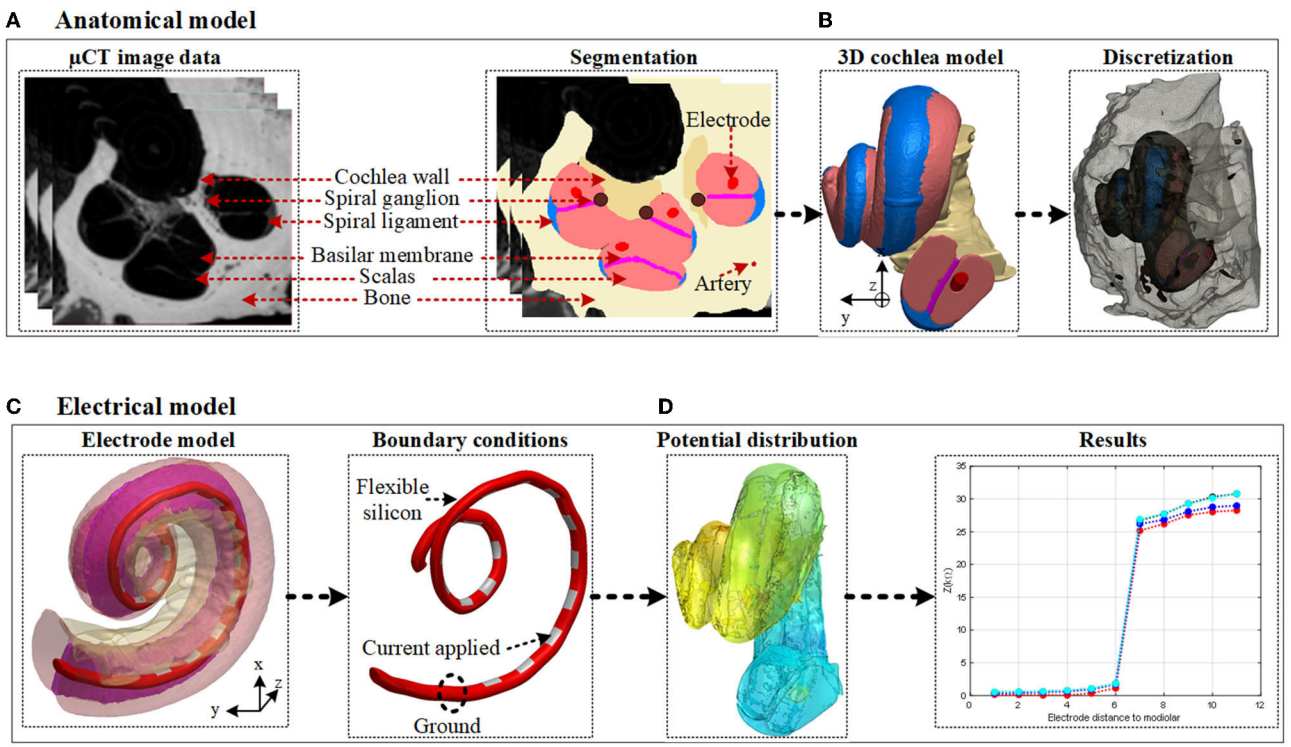


FIGURE 1 | Human cochlea modeling and electrical potential simulation. **(A)** The anatomical layers are segmented based on the image dataset. **(B)** 3D finite element method (FEM) domains are constructed based on captured individual image sets using associated labeling and smoothing filters in ScanIP software. **(C)** The electrode model is generated and embedded in the constructed 3D cochlea. The model is then discretized, and a current source is applied using appropriate boundary conditions. **(D)** The electrical potential distribution in the volume conductor is resolved by FEM and the impedance variation is calculated.

of the cochlea as shown in **Figure 1C**. Twelve different models were generated in X and Y proximity and insertion depth in the z direction used for all electrode insertions. The models were simulated, and the results were analyzed (as shown in **Figure 1D**) to examine if the impedance variation can be used as a marker for electrode position guidance.

The computation cost using the anatomical cochlea model limits the quantity of information that can be examined. More detailed information about different electrode proximities can be investigated using an adequately accurate and simpler geometric model at a much lower computation cost. A 3D geometrical FEM model was generated by imitating the anatomical model of the cochlea and neuromodulator parameters to readily parameterize the electrode array proximity to the cochlea (shown in **Figure 4**). The impact of the different proximities and insertion depths of the electrodes in the scala tympani was evaluated using impedance distribution analysis to determine whether the safe position of the electrode array could be predicted from impedance measurements. The electrode array proximity was parameterized in x and y directions for each insertion depth in the z direction. Different models were generated by selecting samples in X and Y positions. This resulted in 144 different electrode proximity models for 16 different electrode insertion combinations. The impedance variation was simulated and recorded for all significant electrode array proximities in the scala tympani using the geometrical FEM model. Useful information was obtained using a multi-layered anatomical model but at high computation cost and time. Using a geometrical cochlea model enabled multiple detailed measurements of impedance variation vs. proximity of the electrode array to the adjacent layers at reasonable computation cost and time. The results showed that the magnitude of the impedance significantly varied with both electrode insertion depth and proximity to the cochlea wall.

The rest of the article is organized as follows. Section 2 describes methods to generate anatomical and geometric volume conductors of the cochlea, electrode array design, and quasi-static electrical potential simulation. The results of electrode proximity to the anatomical 3D cochlea wall and insertion depth based on impedance variation are presented in Section 3. Discussion on the results is reported in Section 4, and conclusions in Section 5.

2. METHODS

For all simulations, a computer with an Intel Core i7-6700 CPU at 3.4 GHz with 64 GB RAM was used.

2.1. Cochlea Anatomical Model Development

The process of image data segmentation involves the construction of the cochlea volume conductor and its associated

TABLE 1 | Tissue conductivities of cochlea structures used in finite element method (FEM) models of the cochlea.

Tissue layer	Conductivity (S/m)	References
Scalas	1.43	Finley et al., 1990
Cochlea wall	0.3	Finley et al., 1990
Basilar membrane	0.0125	Hanekom and Hanekom, 2016
Spiral ligament	1.67	Frijns et al., 1995
Stria vascularis	0.0053	Frijns et al., 1995
Spiral ganglion	0.33	Hanekom and Hanekom, 2016
Artery	0.32	Gabriel et al., 1996
Bone	0.0156	Finley et al., 1990; Hanekom and Hanekom, 2016
Silicone	1e-7	Hanekom and Hanekom, 2016

anatomical layers. These include the scalas, cochlea wall, basilar membrane, spiral ganglion, spiral ligament, artery, and bone.

2.1.1. Micro-CT Data and Segmentation Process

To obtain accurate FEM results, it is important to develop a 3D anatomical model of the inner ear within the cochlea. The volume conductor of the cochlea and the layers in its vicinity were generated based on a high-resolution ($2.24 \times 2.24 \times 5$ μm) voxel size μCT image stack of a human cochlea (Avci et al., 2014) as shown in **Figure 1A**. Due to limited computation memory, the effective operative field of the scans was rescaled to include only the cochlea and its immediate surroundings and was later down-sampled to an isotropic resolution of 9.6 μm with a spatial resolution of $930 \times 930 \times 1{,}014$ voxels.

The μCT data was imported to Simpleware ScanIP v2016.09 (Synopsys, Mountain View, USA) for image processing and data segmentation by defining regions in the image data that belong to the same anatomical layers. In this way, it becomes possible to construct 3D models that represent the anatomical layers. The detailed cochlea volume conductor was composed of the scala tympani, scala vestibuli, cochlea wall, basilar membrane, spiral ligament, stria vascularis, spiral ganglion, and associated arteries as listed in **Table 1**. The outermost layer that surrounds the cochlea was designated as the bony layer. Both automatic and manual segmentation processes were used to obtain a highly efficient and reliable model for simulation (Salkim et al., 2019). Smoothing filters utilizing recursive Gaussian, median, and mean filters were employed to allocate each tissue layer in a specific grayscale range. Each tissue layer was then generated based on an automatic segmentation process using this grayscale. Manual segmentation was used when editing the morphology or filling cavities (i.e., dilate, erode, open, and close functions) were used in ScanIP software. To obtain appropriate boundaries and remove any overlapping sections between the tissue layers, Boolean operations were applied.

2.1.2. Generation of a 3D Model of the Cochlea

After labeling all tissue layers and their edges on image data, the 3D model of each tissue layer in the cochlea was generated as shown in **Figure 1B** to enable simulation of the electrical potential distribution generated by a given electrode setting. The added computation time due to sharp edges were reduced by applying 3D editing filters in ScanIP software. Spiral ganglion and nerves are distributed throughout the tunnel spiral in the modiolus called the Rosenthal's canal. Since they possess similar conductivities, they were considered as one layer for potential distribution analysis. The basilar membrane and the osseous spiral lamina layers were combined and modeled as one layer due to the discontinuity of the osseous spiral lamina. The thin membranes between the scala vestibuli and the vestibule were excluded when modeling as they cannot be identified in the μCT scans. Since the stria vascularis layer is comparatively thin, it was modeled for all models as "contact impedance" during simulations. Practical computation times were attained using these adjustments.

2.2. Electrode Array Design in the Anatomical Model

To conduct stimulation currents to different parts of the cochlea, an electrode array model was based on a commercially available electrode [Advanced Bionics HiFocus^{TM} SlimJ electrode (Hannover, Germany)] with 16 platinum electrodes. The electrodes provide adequate quality of cochlea stimulation (Dhanasingh and Jolly, 2017). The 16 platinum electrodes are supported by flexible silicone and are designed to face the inner cochlea wall (**Figure 1C**).

First, the electrode array was considered to be placed at the center of the scala tympani. The centerline of the scala tympani was manually generated by calculating the variable cross-section of the scala tympani along with the spiral shape of the cochlea and stored as x, y, and z coordinates in ScanIP software. The electrode array was modeled inside the cochlea by interpolating the center points of the scala tympani and using the sweep function in COMSOL Multiphysics® v5.2a (COMSOL, Ltd, Cambridge, UK). Since the electrode's plates are relatively thin, they were designed as boundary surfaces and combined with the electrode array in COMSOL. The 16 electrode array was inserted into the scala tympani at the midscale position. The electrodes were numbered from E1 at the apical end to E16 at the basal end. After placing the 3D model of the electrode in the scala tympani, the electrode model was relocated to evaluate the impact of the proximity to the cochlea wall on the impedance variation. This resulted in six models in the x and six in the y directions as samples shown in **Figure 2**. Note that the models are equally spaced in both the x and y directions. It was not possible to generate more different electrode array samples due to the shape of the cochlea. The electrode model eventually touched the scala tympani's wall (in both x and y directions) as shown in **Figure 3**.

In the following subsections, the impact of the electrode array's proximity to the cochlea wall and the depth of the electrode array penetration are investigated based on impedance variations.

2.2.1. Electrode Proximity

In **Figure 2A** since the cochlea wall is at one surface of the scala tympani, the effect of the proximity to this layer of the electrode impedance in the y direction through the cochlea wall

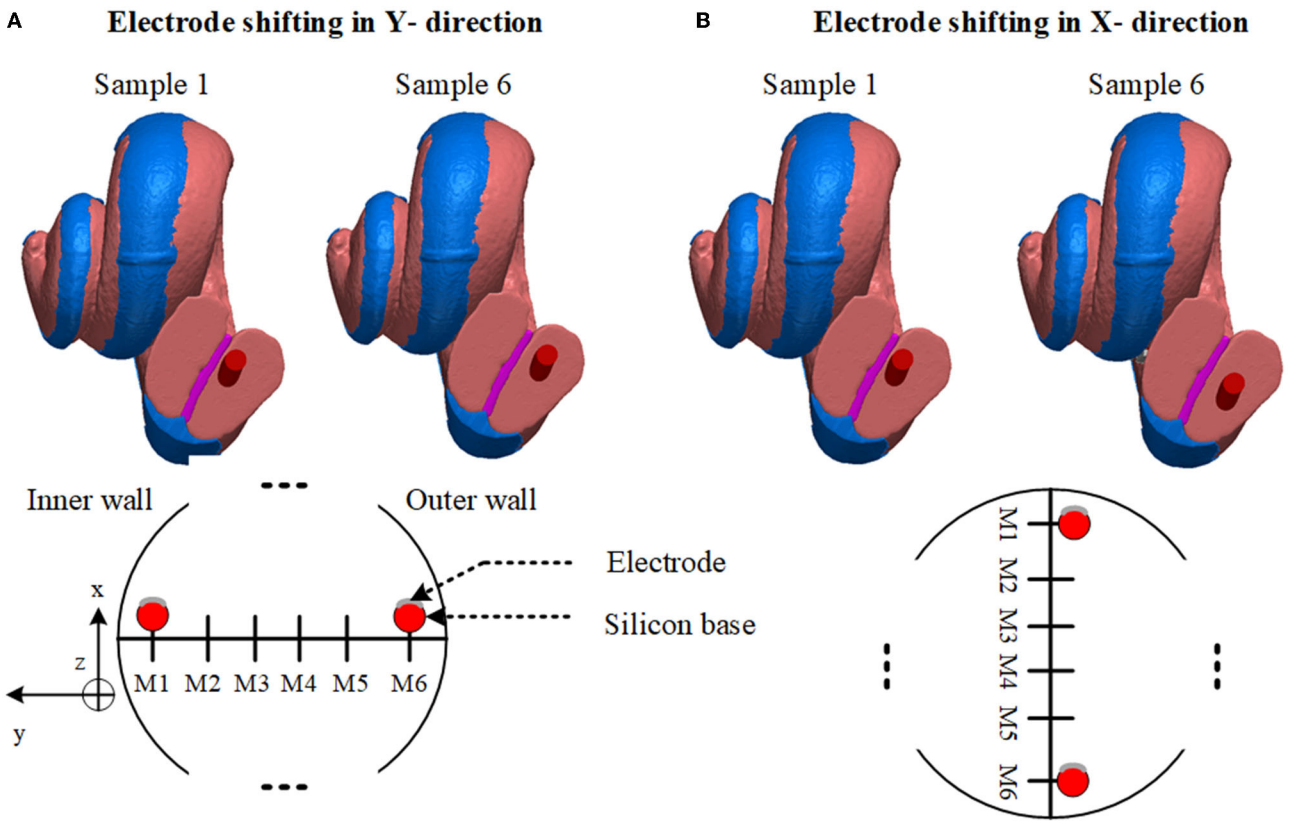


FIGURE 2 | Anatomical layers and electrode array positions. **(A)** Sample 1 shows the electrode array nearly touching the basilar membrane. Sample 6 shows near the outer wall. Lower: diagram of electrode sweeps in the in-y direction. **(B)** Sample 1 shows a relatively closer model to the cochlea wall. Sample 6 shows near the lower wall. Lower: Diagram of electrode sweep in the x direction. Models are equally spaced.

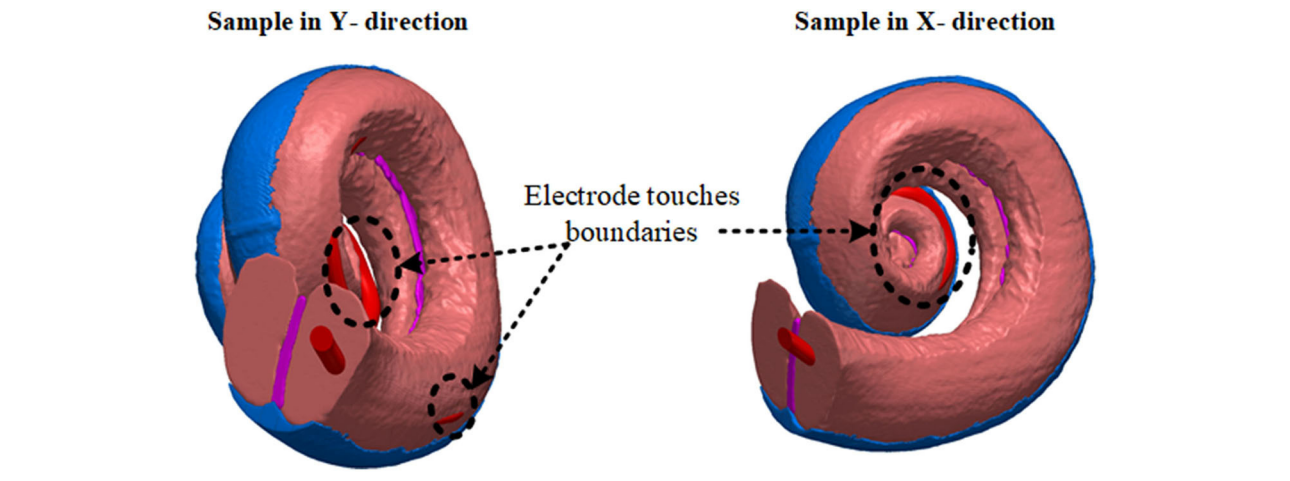


FIGURE 3 | The distance between the anatomical layer and electrode array is not the same for all electrode contacts along the cochlea. This limits the number of samples in the x and y directions. Electrode arrays exit from the scala tympani when they are shifted beyond a certain distance.

is investigated. The electrode array was shifted in the y direction in incremental steps until the silicone base of the electrode array nearly touched the wall of the scala tympani. The process was repeated in the x direction, **Figure 2B**, from one outer wall to the opposite outer wall. The step distance between any two models was equal and was defined to obtain significant impedance

differences (at least 2%) between adjacent positions. Since the cochlea has a helical shape tapering down from the base to the apex, this limits the generation of more samples due to the electrode array exiting from the scala tympani after a certain distance in both x and y directions as shown in **Figure 3**. Six different measurements were obtained in each x and y direction. Each electrode position (M1 to M6) was merged in the 3D cochlea volume conductor, discretized and the electrical potential field due to a current input at the electrode was simulated and the impedance was measured.

2.2.2. Electrode Insertion Depth

The simulated 16 electrode array was positioned along the center of the 3D scala tympani. The electrode contacts were designed and combined with the silicone carrier in COMSOL to form an array model where E1 and E16 represent the initial insertion and full insertion depth, respectively. The 3D electrode model was imported into ScanIP to combine with the 3D cochlea model. It was assumed that the electrode array has been inserted in the optimal place (center of the scala tympani) of the cochlea. The electrode array was inserted into the cochlea wall from the apical electrode (E1) until full insertion (E16). The electrical potential distribution was simulated for each electrode insertion and impedance variation was assessed for each electrode accordingly as shown in **Figure 8**.

2.3. Model Validation

Very detailed electrode proximity parametrization based on impedance variation employing the accurate anatomical cochlea model is impractical due to its complexity requiring very long computation times (refer below). Also, it was shown that it was not possible to parameterize the electrode array position within the scala tympani due to the helical shape of the cochlea as shown in **Figure 2**. The electrode array touched the nearby anatomical layers (as shown in **Figure 3**), and this limited generation of more samples using an anatomical model. An alternative simplified and sufficiently accurate model of the cochlea and adjacent tissue layers can be represented by geometries (i.e., ellipsoids, cylinders) that describe only the regions of interest (Salkim et al., 2019). This significantly reduces computation times at the cost of some minor added error, allowing practical multiple measurements. Note that the same electrode dimensions and current source were used for all generated models.

The two models were compared based on each electrode impedance variation (as shown in **Figure 8**) for the full electrode array insertion. The resulting impedance variation with electrode depth was recorded (from E1 to E16) for both anatomical and geometrical models. The resulting error was 1 (minimum) to 2% (maximum) when compared to impedance measurements for the same distance to the cochlea wall as shown in **Figure 8**. The computation time per measurement for the anatomical model was ∼5 h but it was 10 min for the geometrical model. This significantly reduces computation time but still has sufficient accuracy and enables a more detailed parametrization of the proximity of the electrode array based on impedance measurements.

2.4. Detailed Electrode Proximity Impedance Parametrization Using a Geometrical Model

To generate the geometrical model the bony layer, scala tympani, vestibular and basilar membrane layer, spiral ligament, and electrode array were constructed based on ellipsoids and cylinders as shown in **Figure 4A** in COMSOL with relatively larger element dimensions compared to the anatomical model. The stria vascularis layer was modeled as "contact impedance" during simulations. The electrode array was initially inserted into the scala tympani and impedance variation was measured to assess the difference between the initial and full insertion of the electrode. Each electrode, in turn, was activated, and impedance was calculated for electrode contacts. As shown in **Figure 5** the impedance varies for each electrode contact in agreement with (Vanpoucke et al., 2011). The parametrization process was assessed based on the full insertion of the electrode within the scala tympani. Measurements of electrode impedances were sequentially made between each contact and the ground ([E1 to the ground], [E1 to the ground, E2 to the ground], … [E1 to ground,...E16 to ground]). For each electrode insertion from E1 to E16, impedances were again assessed across those contacts that were already in the scala tympani. This resulted in 136 impedance recordings for specified proximities to the cochlea wall (e.g., X_n, Y_0) until a certain distance approached the tympani border (e.g., X_0). The measurements were limited to these areas to reduce computation costs.

The electrode position parametrization was based on an 80 × 60 matrix of cross points overlaying the oval-shaped cross-section shown in **Figure 4B**, resulting in somewhat fewer than 144 positions by ignoring insignificant variation in impedances (as shown in **Figure 6**). A reduction in computation time was made by selecting samples in X and Y positions (**Figure 4B**). First, a sample in the x direction (e.g., X_0) was kept constant and the electrode position was swept up to 12 increments in the y direction to very close to the tympani border or basilar membrane. This process was repeated for the remaining samples (from X_0 to X_n). The same procedure was repeated with X and Y interchanged for $-Y_n$ to Y_n. At each point, the electrode electrical potential was simulated and its impedance to the ground was recorded. Note that the impedance was not calculated for the models that touched the cochlea wall or border of any adjacent layers. This provided sufficient detail for analysis. Since the variation when approaching the cochlea wall is vital, the impedance variation was calculated for all electrode contacts in the x direction but it was recorded for E1, E6, E11, and E16 in the y direction as discussed in Section 3.

2.5. Tissue Conductivity and Boundary Conditions

Once the anatomical cochlea volume conductor model and electrode array settings were completed, the electrical characteristics for each tissue layer were assigned using the parameters in **Table 1** to perform the impedance measurements. The simulations were solved based on Dirichlet boundary conditions using (1) which approximates to ground at the

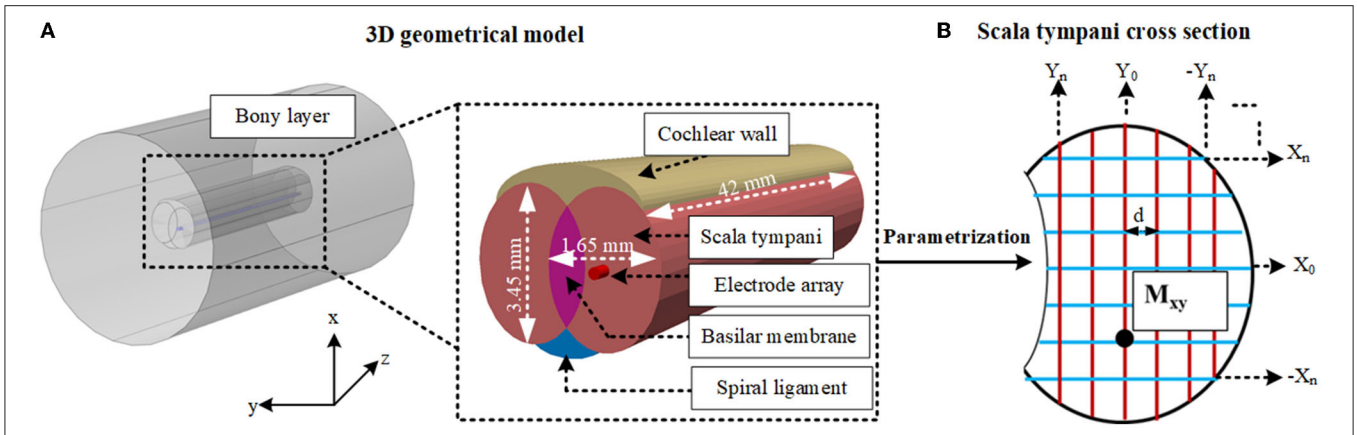


FIGURE 4 | (A) Shows the layers of the 3D geometrical model and dimensions of the cochlea. **(B)** Shows the cross-section of the scala tympani and the parametrization of the proximity of the electrode array in this layer. Different samples are highlighted in both directions (X and Y). M_{xy} shows one of the samples, and d shows the equal distance between each sample for both x and y directions. Other layers are not shown.

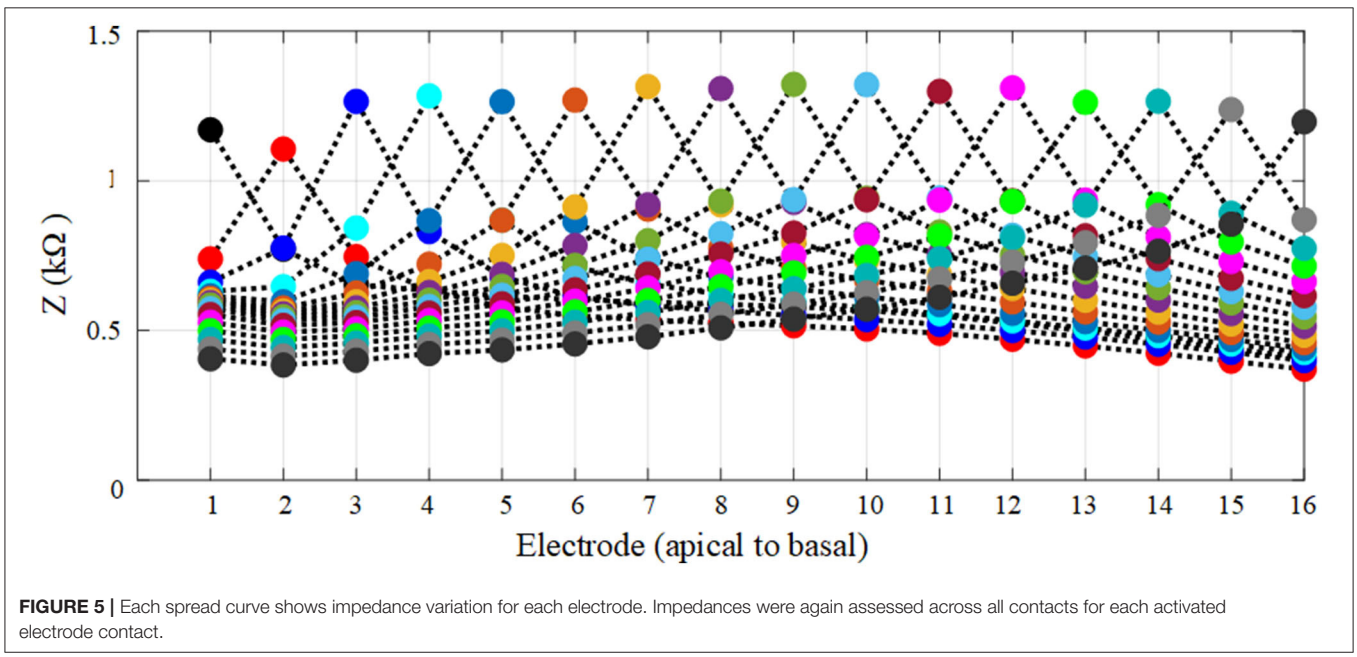


FIGURE 5 | Each spread curve shows impedance variation for each electrode. Impedances were again assessed across all contacts for each activated electrode contact.

infinity boundary condition.

$$V(\delta\Omega) = 0 \tag{1}$$

where V shows the electrical potential and $\delta\Omega$ represents the outermost surface layer of the model. The conductivities of tissue layers within the volume conductor are listed in **Table 1**. The electrical features of the cochlea layers have been reported in numerous studies (Frijns et al., 1995; Hanekom and Hanekom, 2016), and the conductivity values that are currently used in the computational modeling of the cochlea used in this study are shown in **Table 1**. They are assumed to be isotropic as there is no data in the literature on the anisotropy of the layer conductivities of the cochlea except for the bone layer. The quasi-static approximation was used as detailed in the following subsection. The conductivities of the scala tympani and scala vestibuli were assumed to be the same since both layers are composed of the same fluid (perilymph) and possess similar electrical characteristics. The thin anatomical layers around the scala (veins, nerve trunk) were not considered in the final volume conductor model; it was assumed they have a negligible effect on impedance when typical CI current pulses are applied. The external surface of the membrane was insulated to prohibit current flow from the scala tympani into the non-conductive middle ear air space. Finally, the surface electrode remaining external to the scala tympani was grounded to represent the current sink.

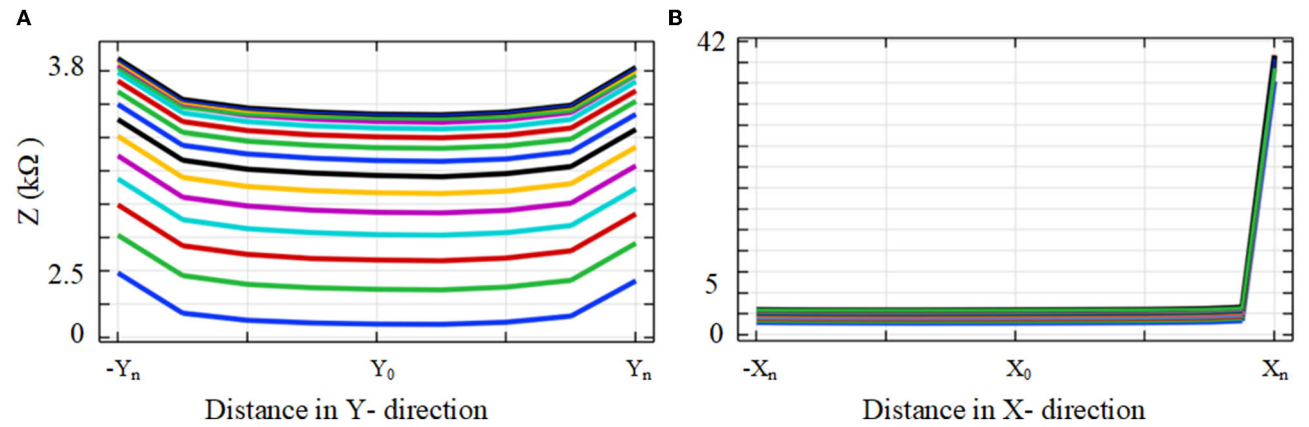


FIGURE 6 | (A) Electrode array proximity was parameterized for the scala tympani border in the y direction and **(B)** cochlea wall in the x direction. Y_0 was kept constant and electrode proximity to the cochlea wall in the x direction was parameterized. The same process was repeated for the y direction.

2.6. Computing the Electrical Potentials in a Volume Conductor

Each model was imported into COMSOL for finite element analysis. Models were then discretized using tetrahedral finite elements for numerical solutions of partial differential equations in COMSOL. Each simulation was solved iteratively on a 64-bit multicore processor using the conjugate gradients method. The accuracy of the simulation is proportional to the volume conductor mesh resolution. The scalas and the tissue layers near the scalas were meshed using a minimum element size of 1 μm and a relatively lower growth rate (1.1) and the remaining tissue layers were meshed with relatively larger minimum element sizes (e.g., 0. 1–1 mm) to obtain sufficient accuracy while reducing excessive computation time. Mesh settings for the electrode (electrodes) were adaptively adjusted to different sizes and growth rates in different models. Since the outermost layer (bone) was far from the region of interest, the discretization element size was selected to be larger (i.e., known as normal tetrahedral setting) than the cochlea layers. The number of elements varied approximately between three and five million during the discretization process, depending on the model.

In this study, simulations calculated the electrical potential distribution within the volume conductor using the quasi-static approximation of the Laplace equation:

$$\nabla \cdot (\sigma \nabla V) = 0 \tag{2}$$

where σ is the tissue conductivity (as shown in **Table 1**), and V is the electrical potential in the representative geometry. The electrical potential variation for each model was simulated by applying a 34 μA current to calculate impedance measurements as shown in (3). The impedance Z_{el} to ground for each model was derived from (3)

$$Z_{el} = V_{el}/I_{el} \tag{3}$$

where V_{el} is the resulting electrode potential, and I_{el} is the applied quasi-static current (chosen to be unity). Since the study is based on quasi-static approximation due to the lack of the dielectric parameters of the cochlea, the electrode-tissue interface contact impedance was assumed to be zero. The appropriate continuity conditions were implemented at the boundary of the different domains to provide a unique solution.

3. RESULTS

In this section, impedance variation was initially analyzed based on the anatomical cochlea model. After comparing the results for both the anatomical and geometrical models, parametrization results were generated based on the geometrical model.

3.1. Impedance Measurements in the Anatomical Model

The impedance to ground variations of the electrode (of the fully inserted array) with electrode proximity to the cochlea wall for x and y directions is shown in **Figure 7**. Each measurement was color-coded and labeled with the proximity to the cochlea wall or border of the nearby tissues. For the x direction a red circle indicates its relatively closer position to the cochlea wall but not touching, a blue circle indicates a mid-scalar position and a cyan circle is when the electrode is at the furthest position from the cochlea wall relatively close to (but not touching) the outer wall of scala tympani; similarly for the y direction.

As shown in **Figure 7A**, there is a direct relationship between impedance magnitude and electrode proximity to the scala tympani border in the y direction. The impedance changes increase with the electrode proximity to the basilar membrane. They increased by about 12% when the closer position of the electrode is compared to the mid-scalar one. There is a notable change in impedance magnitude when the electrode is moved away from the sensitive layer (basilar membrane). The impedance varies from 2.6 to 1.4 kΩ. **Figure 7B** shows the results for the electrode proximity to the cochlea wall in the x direction. There is significant variability in the impedance when the electrode is placed

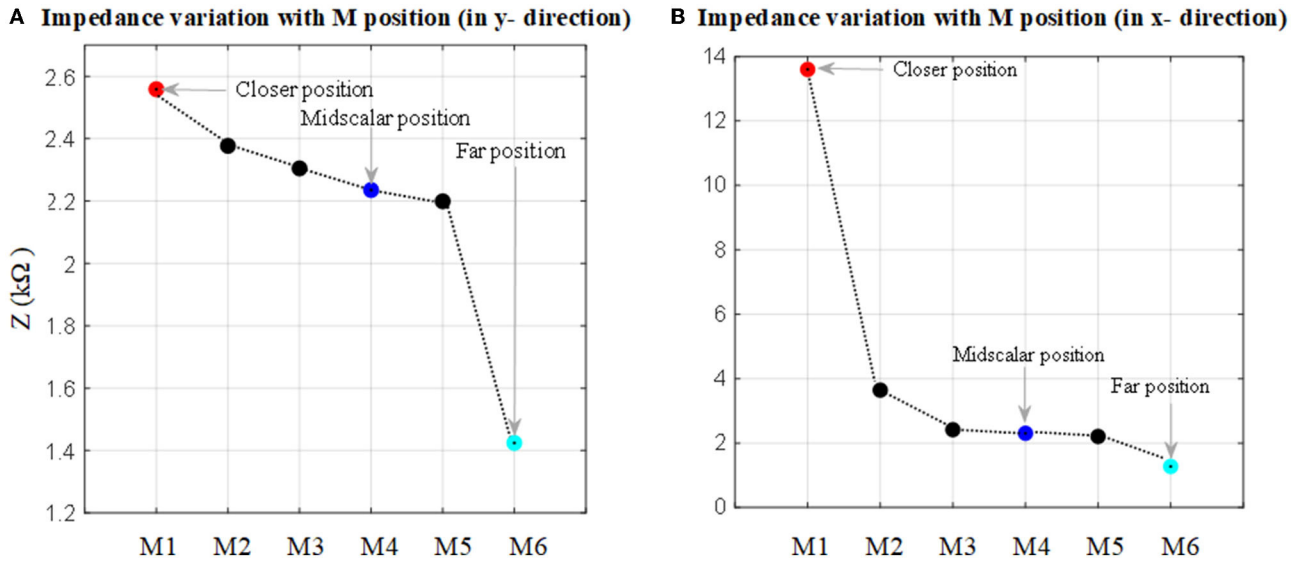


FIGURE 7 | Impedance to ground variation with the position in the cochlea for an electrode when the 16 electrode array is fully inserted. **(A)** Electrode impedance variation with proximity to the basilar layer in the y direction (refer to **Figure 2A**). **(B)** Electrode impedance variation with proximity to the cochlea wall in the x direction (refer to **Figure 2B**).

closer to the cochlea wall, compared to the mid-scalar and furthest position. The impedance varies from 2 to 13.5 kΩ.

3.2. Model Validation

The electrode array is fully inserted into the cochlea and is assumed to be at the mid-scalar position. The anatomical and geometrical models were compared based on electrode insertion from the apex (E1) to basal (E16) electrode impedance variation. The impedance variation between these models based on a fully inserted electrode array is highlighted in different colors and shown in **Figure 8**. The impedance of electrodes at different insertion depths shows an approximately linear increasing change with electrode depth for both models. The impedance measurement is slightly higher using the geometrical model compared to the anatomical model. The impedance difference between the two models for different electrodes varies between 1 and 2%, providing sufficient accuracy with far less computation cost.

3.3. Measurements of Impedance Variation With Electrode Proximity Based on a Geometrical Model

The impedance variation with different proximities of the electrode to the cochlea wall in the y and the x directions are shown in **Figures 9**, **10**, respectively.

Figure 9 shows the impedance variation for the same samples of the electrode contacts such as E1, E6, E11, and E16. Examination of **Figure 9** shows that there is a correlation between the electrode impedance and its distance to the basilar membrane or scala tympani outer wall for all positions. It shows that the magnitude of the impedance increases when the electrode is closer to these boundary layers for all samples. In particular, the results highlighted in red indicate that the magnitude of the impedance is considerably changed in the y direction for a certain sample in the x direction. As the electrode array is placed closer to the cochlea wall (distance in the x direction), this resulted in the highest magnitude difference in impedance which is in agreement with **Figure 10**. On the other hand, the results for the electrode array that is placed toward the center of the scala tympani show the lowest impedance difference for different distances.

Figure 10 shows impedance variation for sequential electrode insertion (E1 insertion, E1 to E2 insertion... E1 to E16 insertion) for different proximities (d to 12d) of the cochlea wall. As shown in the subplots in **Figure 10**, there is a relationship between the magnitude impedance and distance in the x direction. In particular, there is a significant impedance variation when the electrode array is placed at a certain distance to the cochlea wall (5d) for all electrode contacts compared to the other distances. The magnitude of the impedance is approximately increased from 2 to 28 kΩ for all electrode insertion samples. Although the magnitude of the impedance is increased with closer proximity (<5d), this is not significant. The remaining distances (from 12d to 6d) do not show notable variation in impedance being relatively far from the cochlea wall.

4. DISCUSSION

One of the key requirements of the CI is the positioning, or geometry, of the electrode array relative to cochlea anatomy (Finley and Skinner, 2008). The experimental visual inspection

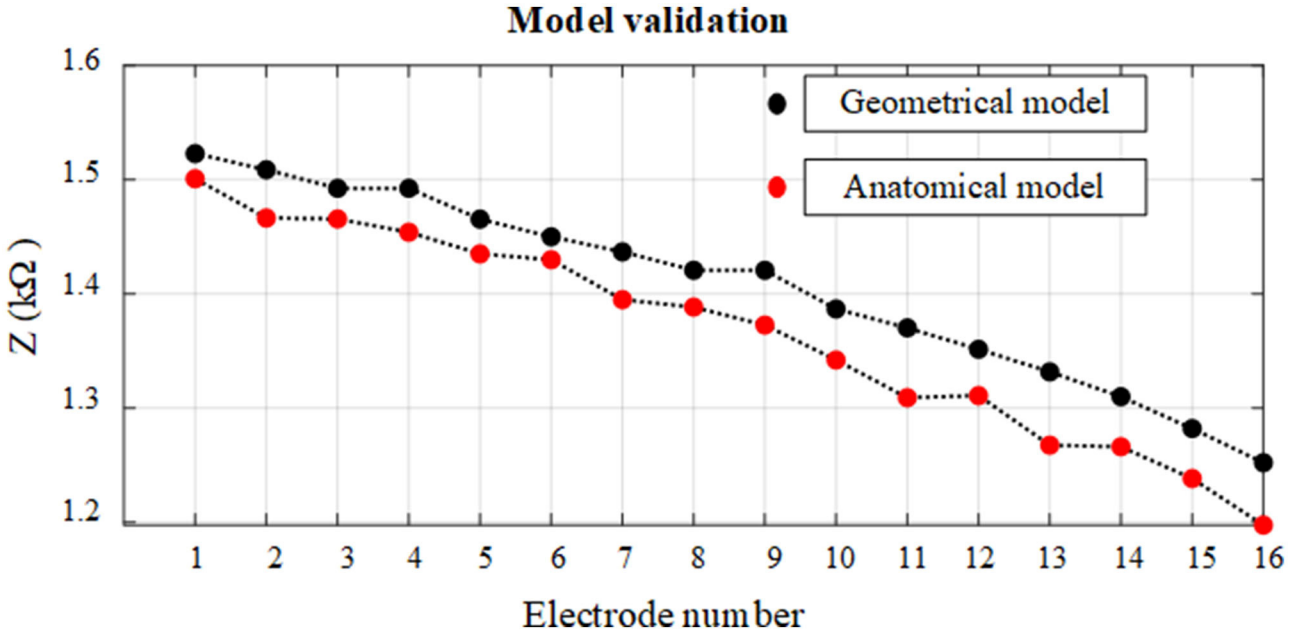


FIGURE 8 | Relationship between impedance variation and electrode insertion depth for anatomical and geometrical cochlea models. The electrodes on the fully inserted array are numbered from E1 to E16.

of the implant is limited (Kratchman et al., 2016). Bio-modeling is increasingly becoming an alternative option to the design and optimization of biomedical devices (Hanekom and Hanekom, 2016; Salkim et al., 2019). Specifically, the electrode array positioning within the anatomical layer can be readily investigated using these models. In such models, the electrical potential is simulated within the volume conductor using appropriate boundary conditions in relation to the associated tissue and electrode electrical parameters.

In this article, a detailed 3D anatomical model of the human cochlea was generated using an individual image dataset. Different set models of the electrode array were generated based on a commercial electrode array and each model was merged with a 3D model of the cochlea to examine the impact of the electrode proximity to the cochlea wall. Using a detailed anatomical model may not be an optimal method to accomplish such an investigation due to its computation cost and the limitation of the model samples. As an alternative, a 3D geometrical model was constructed based on the anatomical model to readily parameterize the proximity of the electrode. Thus, the impact of electrode proximity to the cochlea wall and the electrode insertion depth based on impedance measurements were examined to investigate whether the impedance variation can be a guide of the electrode positioning during surgery. Different electrode array models, from far to close to the cochlea wall, within the 3D cochlea were developed.

The results showed that the impedance varied with both proximity and insertion depth as shown in **Figures 7–10**. As shown in **Table 2**, these results are in line with other clinical real-time and computational measurements (Tan et al., 2013; Giardina et al., 2017; Pile et al., 2017). The results for the anatomical model (**Figure 7**) showed that there was a significant impedance increase (350%, comparing M1-M2 to M2-M3) when the electrode was placed closer to the cochlea wall in the x direction. This may be due to the lower conductivity of the cochlea wall which is much lower than the center of the scala tympani. Since the 3D model of the basilar membrane was not exactly perpendicular to the y-axis, M1 is closer to the basilar membrane compared to M6 in the y direction. This leads to the observed higher impedance variation in M1 when compared to the remaining ones in the y direction. This may be due to the basilar membrane resistivity which is much higher than the center of the scala tympani (Frijns et al., 1995). The strong impedance dependence of the electrode proximity near the cochlea wall is a useful characteristic.

The impedance variation of electrode insertion depth was compared based on anatomical and geometrical models to examine the use of the geometrical model for the further assessment of the impedance variation as shown in **Figure 8**. The results showed that the geometrical model can be used to parameterize electrode array in the scala tympani with a maximum error of 2%.

The results based on geometrical model simulations in **Figures 9**, **10** for different points in the x and y directions in the scala tympani indicated that the variation in impedance can be correlated with the proximity of the electrode array to the cochlea wall in agreement with the clinical results (Tan et al., 2013; Pile et al., 2017). This difference can be readily observed in the impedance variation when the electrode array was placed closer to the cochlea wall in the x direction and basilar

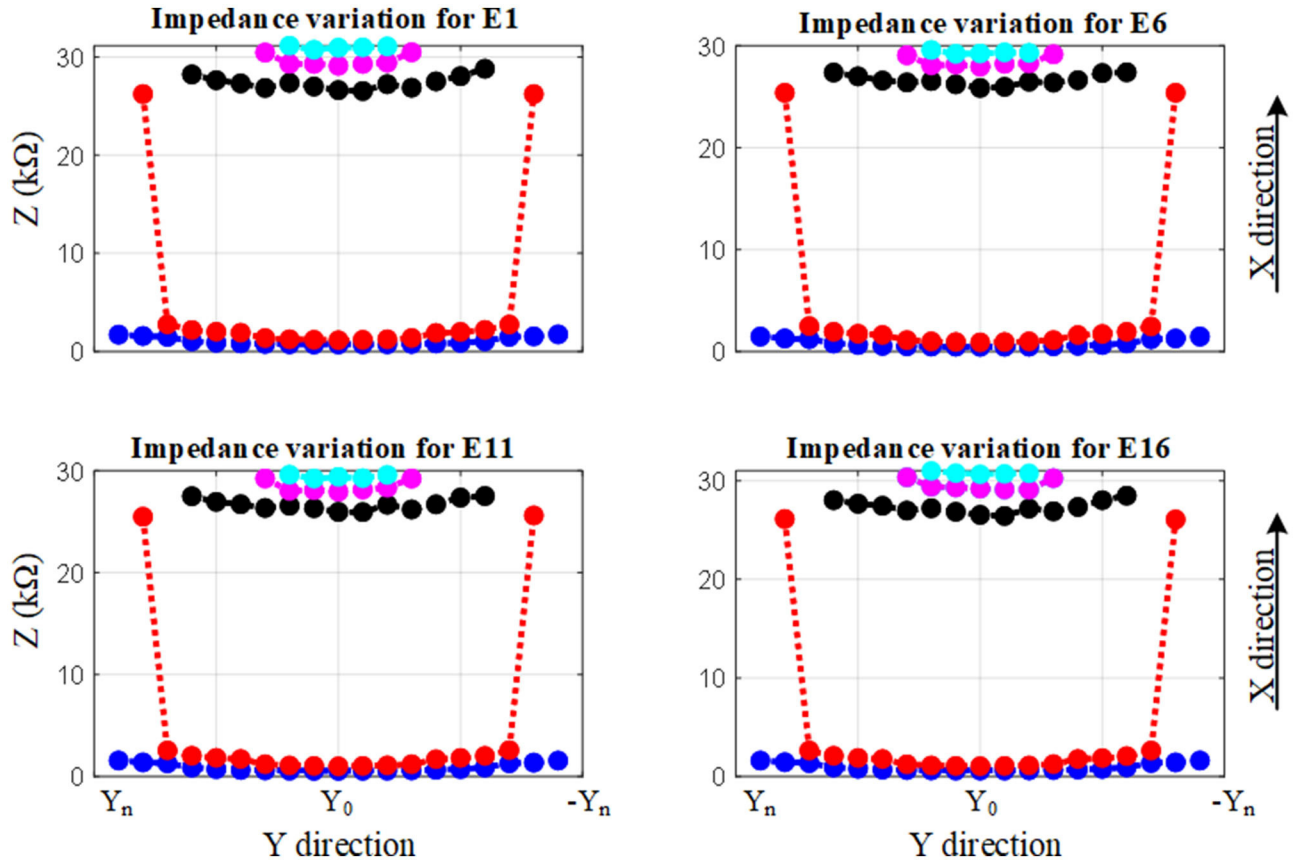


FIGURE 9 | Impedance variation for different electrode proximities to the basilar membrane and scala tympani. The electrode was swept in the y direction for each sample in the x direction.

TABLE 2 | Comparison of this study and published results on the electrode impedance variation range.

Study	Impedance range (kΩ)	Electrode	References
Experimental	5–22	Contour advance	Tan et al., 2013
Experimental	3–25	Contour advance	Pile et al., 2017
Experimental	2–8	Flex	Bruns et al., 2021
Modeling (Computational)	2–5	Flex	Bruns et al., 2021
Modeling (Phantom)	1–25	HiFocus^{TM} SlimJ	Giardina et al., 2017
Modeling (Computational)	3–4.5	HiFocus^{TM} SlimJ	Salkim et al., 2020
Modeling (Computational)	1–38	HiFocus^{TM} SlimJ	This work

membrane in the y direction. Although there is no significant change in impedance when the electrode array is swept in the y direction for a certain distance in the x direction, a notable variation was observed for a larger distance in the x direction as shown in **Figure 9**. It is noted that impedance increases when the electrode array is shifted toward the inner and outer scala tympani as shown in **Figures 6**, **9**. This may be a guide the surgeon to safely place the electrode in the scala tympani without touching the borders in the y direction. The same variation trend was partially observed for anatomical model results in **Figure 7A**. This is due to the generation of the limited samples in the y direction.

There was a considerable change in the magnitude of the impedance for all insertions based on different proximities to the cochlea wall in the x direction as shown in **Figure 10** subplots. The impedance increased by 1,400% when the electrode was moved from 6d to 5d (relative distance). Note that there was a small impedance increment when the electrode was placed further away from the cochlea wall in the x direction. It has been shown that the majority of CI current is confined to the scala timpani due to a relatively higher current pathway compared to the transversal current pathways toward the cochlea wall (Vanpoucke et al., 2004). As the current is limited to the scala tympani, the return path to the ground becomes longer and

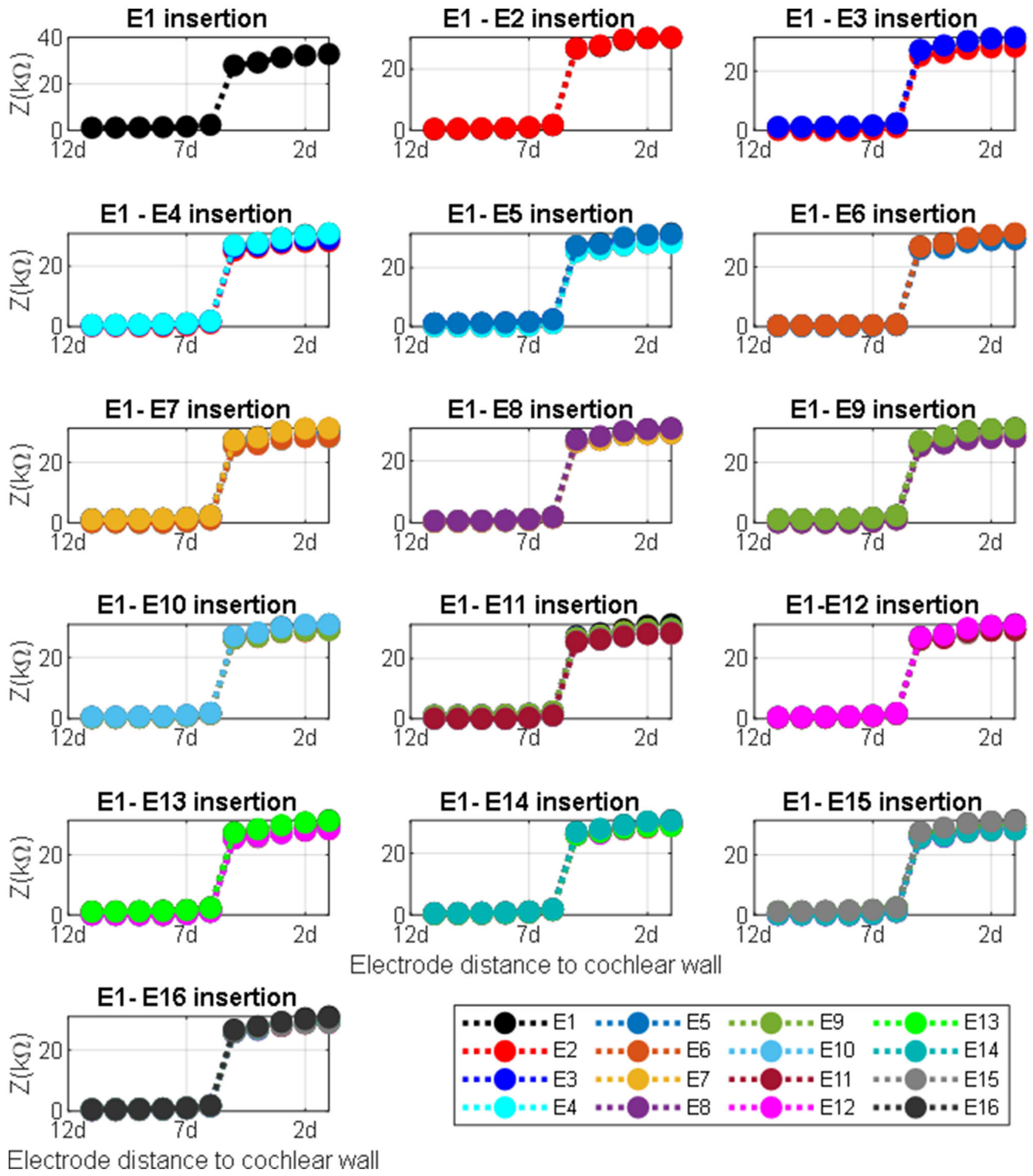


FIGURE 10 | Impedance variation for different electrode proximities to the cochlea wall. The electrode was swept in the x direction for each sample in the y direction for defined electrode insertion. Distance d = proximity.

the cochlea conductive space becomes narrower as the electrode array is inserted deeper into the scala tympani. This may explain why the total impedance increases with both insertion depth and proximity to the cochlea wall (Tan et al., 2013; Pile et al., 2017), consistent with the results in this study. Thus, the electrode array proximity sample (relative distance 6d) is much more

sensitive and specific to detecting which electrodes are in very close proximity to the cochlea wall. Each model provides discrete but complementary information regarding the position of the electrode relative to the cochlea wall or the borders of the scala tympani, which may be clinically valuable in assessing the electrode positioning. In this way, the surgeon could adjust the position of the electrode in the scala tympani during the insertion process if the results showed around this threshold.

The conductivity values that are most commonly used in current computational modeling were used in this study. The impact of the conductivity variation on the impedance variation was investigated. Since the most important layer is scala tympani for the electrode insertion guidance, the conductivity of this layer was changed in $\pm 5\%$ steps and the resulting simulated electric potential was recorded to investigate the impact of the conductivity on the impedance variation. It was shown that there is no significant change in impedance variation (error $< 1\%$) for $\pm 10\%$ variation in the conductivity.

Although various impedance values were recorded when the electrode array was placed relatively close to the cochlea wall, it has been shown in modeling and experimental studies that there is a significant increment in the impedance variation when the electrode array is placed very close to the cochlea wall. This may help to alert the surgeon to further action.

A limitation of this study is the assumption that all tissue layers are purely conductive and isotropic without considering dielectric properties. Also, it was assumed that each contact of the electrode array has equal proximity to the cochlea wall for each design.

The results of this study demonstrate that impedance variation can be a guidance marker for the positioning of the electrode array. The method could be used to develop a real-time guidance tool for the surgeon to prevent hearing loss by avoiding the electrode array touching the cochlea wall and delicate tissue layers (e.g., basilar membrane, hair cells) during insertion.

5. CONCLUSION

Accurate anatomical and geometrical volume conductor models of a human cochlea provide useful tools for studying the relationship between electrode impedance and electrode position in the scala tympani. Using the geometrical model of the cochlea and combined with adequate electrical parameters of CI, the parametrization processes were applied to construct an impedance variation map based on both electrode array insertion depth and electrode proximity to the anatomical layers at the vicinity (e.g., cochlea wall). The method has been shown to identify the impedance variation levels for the electrode proximity position and electrode insertion. The results of this study suggest it may be clinically applicable and lead to optimal electrode array positioning if they are validated with the experimental study. Future study will involve an experimental study of the electrode array positioning in temporal bone and cadaveric tests to further validate the relationship between impedance and electrode position and compare it with computational results.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

ES designed the models, performed the simulations, analyzed the data, and wrote the manuscript. MZ, DJ, and SS reviewed the manuscript and contributed to the research. AD supervised the research and reviewed and edited the manuscript. All authors approved the final manuscript.

FUNDING

This study was supported by the Engineering and Physical Sciences Research Council under grant EP/R511638/1.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Patrick Boyle from Advanced Bionics for providing the electrode array used in this study.

REFERENCES

Avci, E., Nauwelaers, T., Lenarz, T., Hamacher, V., and Kral, A. (2014). Variations in microanatomy of the human cochlea. *J. Compar. Neurol.* 522, 3245–3261. doi: 10.1002/cne.23594

Bai, S., Encke, J., Obando-Leitón, M., Weiß, R., Schäfer, F., Eberharter, J., et al. (2019). Electrical stimulation in the human cochlea: a computational study based on high-resolution micro-CT scans. *Front. Neurosci.* 13:1312. doi: 10.3389/fnins.2019.01312

Bruns, T. L., Riojas, K. E., Labadie, R. F., and Webster III, R. J. (2021). Real-time localization of cochlear-implant electrode arrays using bipolar impedance sensing. *IEEE Trans. Biomed. Eng.* 69, 718–724. doi: 10.1109/TBME.2021.3104104

Caversaccio, M., Wimmer, W., Anso, J., Mantokoudis, G., Gerber, N., Rathgeb, C., et al. (2019). Robotic middle ear access for cochlear implantation: first in man. *PLoS ONE* 14:e0220543. doi: 10.1371/journal.pone.0220543

Clark, J. R., Leon, L., Warren, F. M., and Abbott, J. J. (2012). Magnetic guidance of cochlear implants: Proof-of-concept and initial feasibility study. *J. Med. Devices* 6:035002. doi: 10.1115/1.4007099

Dang, K. (2017). *Electrical conduction models for cochlear implant stimulation* (Ph.D. thesis). Université Côte d'Azur, Nice, France.

Dazert, S., Thomas, J. P., Büchner, A., Müller, J., Hempel, J. M., Löwenheim, H., et al. (2017). Off the ear with no loss in speech understanding: comparing the rondo and the opus 2 cochlear implant audio processors. *Eur. Arch. Oto-Rhino-Laryngol.* 274, 1391–1395. doi: 10.1007/s00405-016-4400-z

Dhanasingh, A., and Jolly, C. (2017). An overview of cochlear implant electrode array designs. *Hear. Res.* 356, 93–103. doi: 10.1016/j.heares.2017.10.005

Finley, C. C., and Skinner, M. W. (2008). Role of electrode placement as a contributor to variability in cochlear implant outcomes. *Otol. Neurotol.* 29:920. doi: 10.1097/MAO.0b013e318184f492

Finley, C. C., Wilson, B. S., and White, M. W. (1990). "Models of neural responsiveness to electrical stimulation," in *Cochlear Implants*, eds J.

M. Miller, A. Arbor, and M. I. A. Francis (Seattle: Springer), 55–96. doi: 10.1007/978-1-4612-3256-8_5

Frijns, J., De Snoo, S., and Schoonhoven, R. (1995). Potential distributions and neural excitation patterns in a rotationally symmetric model of the electrically stimulated cochlea. *Hear. Res.* 87, 170–186. doi: 10.1016/0378-5955(95)00090-Q

Gabriel, C., Gabriel, S., and Corthout, Y. E. (1996). The dielectric properties of biological tissues: I. literature survey. *Phys. Med. Biol.* 41:2231. doi: 10.1088/0031-9155/41/11/001

Giardina, C. K., Krause, E. S., Koka, K., and Fitzpatrick, D. C. (2017). Impedance measures during *in vitro* cochlear implantation predict array positioning. *IEEE Trans. Biomed. Eng.* 65, 327–335. doi: 10.1109/TBME.2017.2764881

Hajioff, D. (2016). Cochlear implantation: a review of current clinical practice. *Br. J. Hosp. Med.* 77, 680–684. doi: 10.12968/hmed.2016.77.12.680

Hanekom, T., and Hanekom, J. J. (2016). Three-dimensional models of cochlear implants: a review of their development and how they could support management and maintenance of cochlear implant performance. *Network* 27, 67–106. doi: 10.3109/0954898X.2016.1171411

Holden, L. K., Finley, C. C., Firszt, J. B., Holden, T. A., Brenner, C., Potts, L. G., et al. (2013). Factors affecting open-set word recognition in adults with cochlear implants. *Ear Hear.* 34:342. doi: 10.1097/AUD.0b013e3182741aa7

Jethanamest, D., Tan, C.-T., Fitzgerald, M. B., and Svirsky, M. A. (2010). A new software tool to optimize frequency table selection for cochlear implants. *Otol. Neurotol.* 31:1242. doi: 10.1097/MAO.0b013e3181f2063e

Kratchman, L. B., Schuster, D., Dietrich, M. S., and Labadie, R. F. (2016). Force perception thresholds in cochlear implantation surgery. *Audiol. Neurotol.* 21, 244–249. doi: 10.1159/000445736

Kushalnagar, R. (2019). "Deafness and hearing loss," in *Web Accessibility*, eds Y. Yesilada, Güzelyurt, and S. Harper (Manchester: Springer), 35–47. doi: 10.1007/978-1-4471-7440-0_3

Malherbe, T., Hanekom, T., and Hanekom, J. (2016). Constructing a three-dimensional electrical model of a living cochlear implant user's cochlea. *Int. J. Numer. Methods Biomed. Eng.* 32:e02751. doi: 10.1002/cnm.2751

Mens, L. H. (2007). Advances in cochlear implant telemetry: evoked neural responses, electrical field imaging, and technical integrity. *Trends Amplif.* 11, 143–159. doi: 10.1177/1084713807304362

Miller, C. A., Brown, C. J., Abbas, P. J., and Chi, S.-L. (2008). The clinical application of potentials evoked from the peripheral auditory system. *Hear. Res.* 242, 184–197. doi: 10.1016/j.heares.2008.04.005

Min, K. S., Jun, S. B., Lim, Y. S., Park, S.-I., and Kim, S. J. (2013). Modiolus-hugging intracochlear electrode array with shape memory alloy. *Comput. Math. Methods Med.* 2013:250915. doi: 10.1155/2013/250915

Mittmann, P., Ernst, A., and Todt, I. (2015). Intraoperative electrophysiologic variations caused by the scalar position of cochlear implant electrodes. *Otol. Neurotol.* 36, 1010–1014. doi: 10.1097/MAO.0000000000000736

Newbold, C., Mergen, S., Richardson, R., Seligman, P., Millard, R., Cowan, R., et al. (2014). Impedance changes in chronically implanted and stimulated cochlear implant electrodes. *Cochlear Implants Int.* 15, 191–199. doi: 10.1179/1754762813Y.0000000050

O'Connell, B. P., Hunter, J. B., and Wanna, G. B. (2016). The importance of electrode location in cochlear implantation. *Laryngosc. Invest. Otolaryngol.* 1, 169–174. doi: 10.1002/lio2.42

Pile, J., Sweeney, A. D., Kumar, S., Simaan, N., and Wanna, G. B. (2017). Detection of modiolar proximity through bipolar impedance measurements. *Laryngoscope* 127, 1413–1419. doi: 10.1002/lary.26183

Rebscher, S. J., Hetherington, A., Bonham, B., Wardrop, P., Whinney, D., and Leake, P. A. (2008). Considerations for the design of future cochlear implant electrode arrays: electrode array stiffness, size and depth of insertion. *J. Rehabil. Res. Dev.* 45:731. doi: 10.1682/JRRD.2007.08.0119

Salkim, E., Shiraz, A., and Demosthenous, A. (2019). Impact of neuroanatomical variations and electrode orientation on stimulus current in a device for migraine: a computational study. *J. Neural Eng.* 17:016006. doi: 10.1088/1741-2552/ab3d94

Salkim, E., Zamani, M., Jiang, D., and Demosthenous, A. (2020). "Detection of electrode proximity to the cochlea wall based on impedance variation: a preliminary computational study," in *Proceedings of UKSim-AMSS 22nd International Conference on Modelling & Simulation* (Cambridge, UK). doi: 10.5013/IJSSST.a.21.02.13

Skinner, M. W., Ketten, D. R., Holden, L. K., Harding, G. W., Smith, P. G., Gates, G. A., et al. (2002). CT-derived estimation of cochlear morphology and electrode array position in relation to word recognition in nucleus-22 recipients. *J. Assoc. Res. Otolaryngol.* 3, 332–350. doi: 10.1007/s101620020013

Svirsky, M. A., Teoh, S.-W., and Neuburger, H. (2004). Development of language and speech perception in congenitally, profoundly deaf children as a function of age at cochlear implantation. *Audiol. Neurotol.* 9, 224–233. doi: 10.1159/000078392

Tan, C.-T., Svirsky, M., Anwar, A., Kumar, S., Caessens, B., Carter, P., et al. (2013). Real-time measurement of electrode impedance during intracochlear electrode insertion. *Laryngoscope* 123, 1028–1032. doi: 10.1002/lary.23714

Vanpoucke, F. J., Boermans, P.-P. B., and Frijns, J. H. (2011). Assessing the placement of a cochlear electrode array by multidimensional scaling. *IEEE Trans. Biomed. Eng.* 59, 307–310. doi: 10.1109/TBME.2011.2173198

Vanpoucke, F. J., Zarowski, A. J., and Peeters, S. A. (2004). Identification of the impedance model of an implanted cochlear prosthesis from intracochlear potential measurements. *IEEE Trans. Biomed. Eng.* 51, 2174–2183. doi: 10.1109/TBME.2004.836518

Wanna, G. B., Noble, J. H., Carlson, M. L., Gifford, R. H., Dietrich, M. S., Haynes, D. S., et al. (2014). Impact of electrode design and surgical approach on scalar location and cochlear implant outcomes. *Laryngoscope* 124, S1–S7. doi: 10.1002/lary.24728

Zhang, J., Wei, W., Ding, J., Roland, J. T. Jr, Manolidis, S., and Simaan, N. (2010). Inroads toward robot-assisted cochlear implant surgery using steerable electrode arrays. *Otol. Neurotol.* 31, 1199–1206. doi: 10.1097/MAO.0b013e3181e7117e

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.

frontiers | Frontiers in Computational Neuroscience

TYPE Hypothesis and Theory
PUBLISHED 18 July 2022
DOI 10.3389/fncom.2022.849323

OPEN ACCESS

EDITED BY
Petia D. Koprinkova-Hristova,
Institute of Information and Communication Technologies (BAS), Bulgaria

REVIEWED BY
David M. Devilbiss,
Rowan University School of Osteopathic Medicine, United States
Shivakeshavan Ratnadurai Giridharan,
Burke Medical Research Institute, United States

*CORRESPONDENCE
Florence Véronneau-Veilleux
florence.veronneau-veilleux@umontreal.ca

RECEIVED 05 January 2022
ACCEPTED 28 June 2022
PUBLISHED 18 July 2022

CITATION
Véronneau-Veilleux F, Robaey P, Ursino M and Nekka F (2022) A mechanistic model of ADHD as resulting from dopamine phasic/tonic imbalance during reinforcement learning.
Front. Comput. Neurosci. 16:849323.
doi: 10.3389/fncom.2022.849323

A mechanistic model of ADHD as resulting from dopamine phasic/tonic imbalance during reinforcement learning

Florence Véronneau-Veilleux[1]*, Philippe Robaey[2], Mauro Ursino[3] and Fahima Nekka[1,4,5]

[1]Faculté de Pharmacie, Université de Montréal, Montreal, QC, Canada, [2]Children's Hospital of Eastern Ontario, University of Ottawa, Ottawa, ON, Canada, [3]Department of Electrical, Electronic and Information Engineering "Guglielmo Marconi," University of Bologna, Bologna, Italy, [4]Centre de Recherches Mathématiques, Université de Montréal, Montreal, QC, Canada, [5]Centre for Applied Mathematics in Bioscience and Medicine, McGill University, Montreal, QC, Canada

Attention deficit hyperactivity disorder (ADHD) is the most common neurodevelopmental disorder in children. Although the involvement of dopamine in this disorder seems to be established, the nature of dopaminergic dysfunction remains controversial. The purpose of this study was to test whether the key response characteristics of ADHD could be simulated by a mechanistic model that combines a decrease in tonic dopaminergic activity with an increase in phasic responses in cortical-striatal loops during learning reinforcement. To this end, we combined a dynamic model of dopamine with a neurocomputational model of the basal ganglia with multiple action channels. We also included a dynamic model of tonic and phasic dopamine release and control, and a learning procedure driven by tonic and phasic dopamine levels. In the model, the dopamine imbalance is the result of impaired presynaptic regulation of dopamine at the terminal level. Using this model, virtual individuals from a dopamine imbalance group and a control group were trained to associate four stimuli with four actions with fully informative reinforcement feedback. In a second phase, they were tested without feedback. Subjects in the dopamine imbalance group showed poorer performance with more variable reaction times due to the presence of fast and very slow responses, difficulty in choosing between stimuli even when they were of high intensity, and greater sensitivity to noise. Learning history was also significantly more variable in the dopamine imbalance group, explaining 75% of the variability in reaction time using quadratic regression. The response profile of the virtual subjects varied as a function of the learning history variability index to produce increasingly severe impairment, beginning with an increase in response variability alone, then accumulating a decrease in performance and finally a learning deficit. Although ADHD is certainly a heterogeneous disorder, these results suggest that typical features of ADHD can be explained by a phasic/tonic imbalance in dopaminergic activity alone.

KEYWORDS

attention deficit hyperactivity disorder, tonic and phasic dopamine, neurocomputational model, basal ganglia, reinforcement learning

1. Introduction

Attention Deficit Hyperactivity Disorder (ADHD) is a complex neurodevelopmental disorder characterized by pervasive inattention, impulsivity, and restlessness that is inconsistent with the patient's age (American Psychiatric Association, 2013). The origin of ADHD is largely genetic, and for a smaller part environmental, mostly specific to each individual (Burt, 2009; Wood et al., 2010). The first genome-genome wide meta-analysis identified twelve loci in regions containing enhancers and promoters of expression in central nervous system tissues (Demontis et al., 2019). None of these loci were linked to the dopamine system, despite the fact that dopamine genes have been associated with ADHD in candidate gene approaches (Li et al., 2006; Faraone and Larsson, 2019). Other converging evidence supports a role for dopaminergic dysfunction in ADHD. To briefly list them, most animal models used in ADHD research show some type of dopamine dysfunction (van der Kooij and Glennon, 2007). Stimulants such as methylphenidate, which are the first line of treatment, block more than 50% of dopamine transporters (DAT) in the striatum when given in therapeutic doses (Volkow et al., 1998). ADHD patients are vulnerable to drug dependence, which may be explained by an overlap of ADHD with the dopamine deficiency syndrome (Blum et al., 2008). In functional brain imaging, the most consistent findings are deficits in activity in fronto-striatal circuits where dopamine supports reinforcement learning (Dickstein et al., 2006; Norman et al., 2016). The clearest and most reproducible structural abnormalities in ADHD are located in the basal ganglia and can be normalized by the use of stimulant medications (Nakao et al., 2011). There appears to be a 5- to 10-year lag in the pruning of fronto-striatal circuits in ADHD patients compared to their typically developing peers (Dickstein, 2018). Functional magnetic resonance and diffusion tensor imaging modalities consistently indicate disrupted connectivity in regions and tracts involving fronto-striatal-thalamic loops in ADHD (Saad et al., 2020).

Different models have been proposed to account for a dopaminergic dysfunction. In the basal ganglia, dopamine release may be sustained (tonic) and regulated by prefrontal cortical afferents, or transient (phasic), caused by bursts of firing of dopaminergic neurons (Grace, 1991). The dynamic developmental theory (DDT) of ADHD proposed a hypo-dopaminergic cause. Blunted phasic dopamine bursts impair reinforcement learning (Sagvolden et al., 2005; Volkow et al., 2005), while a hypoactive tonic firing rate results in impaired extinction of previously reinforced behaviors (Sagvolden et al., 2005). A neural network developed by Frank et al. (Frank, 2005; Frank and Claus, 2006) instantiated key properties of cortico-striatal-thalamocortical loops, including direct and indirect basal ganglia pathways. These authors used this basal ganglia model to test the plausibility of the DDT of ADHD with reduced tonic and phasic dopamine levels in the striatum (Frank et al., 2007). While they showed that dopamine modulates the Go and NoGo pathways in the striatum, as well as average reaction time, they were unable to reproduce the increased variability in reaction time, a key feature of ADHD (Kofler et al., 2013), with this hypodopaminergic model alone.

As an alternative we here tested the plausibility of a model that combines a decrease in tonic dopamine activity with an increase in phasic responses (Grace, 2001). In Grace's model, this imbalance is the result of impaired presynaptic regulation of dopamine at the terminal level, and not a central decrease in DA tonic activity that is associated with other conditions, such as chronic stress (Belujon and Grace, 2015; Douma and de Kloet, 2020). This imbalance produces abnormally large reward reinforcements, which explains impulsivity, as well as the preference for smaller immediate rewards over larger delayed rewards (Jackson and MacKillop, 2016; Patros et al., 2016). This model received some support in a PET study showing reduced tonic release and increased phasic release of dopamine in the right caudate in adults with ADHD (Badgaiyan et al., 2015).

In the present study, we used a mechanistic model of the basal ganglia dopaminergic system that we previously developed to help rationally improve pharmacological interventions in Parkinson's disease (Véronneau-Veilleux et al., 2020). The model is a combination of a neurocomputational model of the basal ganglia (Baston and Ursino, 2015; Baston et al., 2016) and a model of dopamine dynamics (Dreyer, 2014) that includes dopamine release and reuptake by DAT. In addition, we included the tonic and phasic release of dopamine as well as the negative regulation of dopaminergic neuron activity by autoreceptors. We used phasic dopamine release as a reward prediction error signal (RPE) for a correct response and a phasic decrease in tonic dopamine activity as a punishment prediction error signal for a false response (Schultz, 2002). Considering that ADHD results from transactions between at-risk individuals and their specific environment (Burt, 2010; Burt et al., 2012), we used this computational model to test the hypothesis that the phasic/tonic imbalance of DA release would lead, during reinforcement learning, to the development in some individuals of ADHD characteristics, in particular response variability.

As dopamine in basal ganglia is primarily involved in learning reinforcement, we considered dopamine phasic vs. tonic release imbalance as a risk factor, and created two groups of virtual participants: one with a phasic/tonic imbalance and the other with the normal balance. We trained all of them to learn responses to 4 stimuli presented in a random sequence, using a forced-choice probabilistic task with a fixed reinforcement learning schedule and fully informative reinforcement feedback. Next, we assessed the outcome of learning reinforcement process in a test phase to determine whether or not ADHD characteristics would be present more frequently in the dopamine imbalance group than in the control

group. Finally, we identified the characteristics of the learning phase that were associated with the development of these ADHD features in the dopamine imbalance group.

2. Methods

The mechanistic model herein developed can be divided into two parts: the dopamine dynamics model and the neurocomputational model of basal ganglia. Synaptic learning in the basal ganglia is modeled with the Hebb's rule. This rule allows the value of synaptic weights to be modified according to tonic and phasic dopamine concentrations. The simulations comprise a learning and a test phase.

2.1. Dopamine dynamics model

The dopamine dynamics model describes the temporal dopamine concentration, both tonic and phasic, autoreceptors occupancy and dopaminergic receptors occupancy. It was adapted from previously published models (Dreyer et al., 2010; Dreyer and Hounsgaard, 2013; Dreyer, 2014; Fuller et al., 2019).

The main mechanisms of dopamine regulation are outlined in the equations of the model and are represented in Figure 1. Dopamine is synthesized in the dopaminergic neurons and then released in the synaptic cleft. Sustained dopamine release refers to tonic dopamine, while transient dopamine release generated by bursts refers to phasic dopamine. The release of phasic dopamine is a reward prediction error signal (RPE) (Waelti et al., 2001; Marinelli and McCutcheon, 2014), whereas a drop in dopamine levels is a punishment prediction error signal. In the synaptic cleft, dopamine can be recaptured by DATs into the presynaptic neuron or be removed from the synaptic cleft by different mechanisms such as diffusion or inactivation by the Catechol-O-methyltransferase.

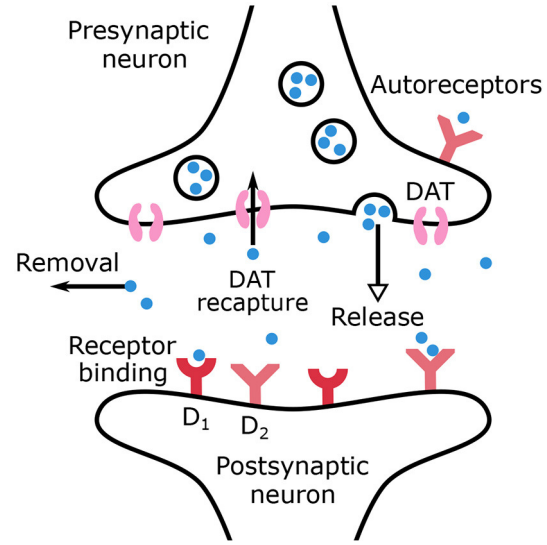


FIGURE 1
Schematic representation of dopamine release, recapture, removal and binding to receptors in the synaptic cleft.

The remaining dopamine molecules can bind to dopaminergic autoreceptors located on the presynaptic neurons or to receptors on the postsynaptic neurons. In the present work, only dopaminergic receptors D_1 and D_2 are considered. All the above mentioned mechanisms are accounted for by the dopamine dynamics model, formulated in Equations (1) and (2), where $C_{DA}(t)$ is the dopamine concentration ($\mu M/L$) in the synaptic cleft and $AR(t)$ the autoreceptors occupancy.

$$\underbrace{\frac{dC_{DA}(t)}{dt}}_{\text{Dopamine concentration}} = \underbrace{(I_{DA}^{tonic} + I_{DA}^{phasic}(t))}_{\text{Dopamine Release}} - \underbrace{\frac{V_{max}C_{DA}(t)}{(k_m + C_{DA}(t))}}_{\text{Recapture by DATs}} - \underbrace{k_{rem}C_{DA}(t)}_{\text{Removal}}, \quad (1)$$

$$\underbrace{\frac{dAR(t)}{dt}}_{\text{Autoreceptor occupancy}} = \underbrace{C_{DA}(t)k_{on}(1 - AR(t))}_{\text{Binding to autoreceptors}} - \underbrace{k_{off}AR(t)}_{\text{Unbinding to autoreceptors}}. \quad (2)$$

As indicated in Equation (1), the release of dopamine is divided into two terms to account for both tonic and phasic release. The recapture by DATs is a saturable process described by a Michaelis-Menten equation. All other mechanisms contributing to dopamine removal are assumed to be linear (Budygin et al., 2002; Dreyer, 2014) and are schematized through the last term in the right-hand member of Equation (1). The binding to autoreceptors is proportional to dopamine concentration and free autoreceptors, while unbinding is proportional only to bound autoreceptors.

Autoreceptors have a regulatory effect on dopamine concentration. Indeed, they provide a negative feedback to adjust dopamine concentration through firing rate, synthesis, and release (Benoit-Marand et al., 2001; Beaulieu and Gainetdinov, 2011). Prolonged dopamine agonist exposure desensitizes autoreceptors in dopamine neurons (Robinson et al., 2017). Loss of inhibition influence facilitates further dopamine release and has been linked to drug abuse. Desensitization was not included in the model which is focused on the short-term effect of dopamine on autoreceptors. If tonic dopamine level decreases (in our ADHD model through increased dopamine reuptake), the temporary decrease in autoreceptor-mediated inhibition would mainly increase phasic dopamine release following the model developed by Grace (Grace, 1991, 2016). Autoinhibition of the presynaptic neurons

is included in the model through the phasic release term only which is associated with the reward prediction error, while the tonic term is not here modified by autoreceptors occupancy (Grace, 1991).

The tonic dopamine release term is given by:

$$I_{DA}^{tonic} = \rho \frac{P_r^{tonic} n_0}{\alpha_{vf} N_A} \upsilon_{tonic}, \tag{3}$$

where ρ is the terminal density, P_r^{tonic} the tonic release probability, n_0 the number of molecules released per vesicles fusion, α_{vf} the extracellular volume fraction, N_A the Avogadros constant and υ_{tonic} the tonic firing rate. The tonic release is independent of autoreceptors occupancy, as explained above.

The phasic release term at time t is given by:

- when there is no response yet, and no prediction error signal:

$$I_{DA}^{phasic}(t) = 0, \tag{4}$$

- when there is a reward prediction error signal at time t_{reward}:

$$I_{DA}^{phasic}(t) = \rho \frac{\left(P_r^{phasic} \cdot \frac{0.334}{AR(t)}\right) n_0 |RPE|}{\alpha_{vf} N_A} \left(\upsilon_{phasic} \cdot \frac{0.334}{AR(t)}\right), \tag{5}$$

$$\text{for } t_{reward} + 0.1 \leq t \leq t_{reward} + 0.1 + 0.05, \tag{6}$$

- when there is a punishment prediction error at time $t_{punishment}$:

$$C_{DA}(t) = 0, \tag{7}$$

$$\text{for } t_{punishment} + 0.1 \leq t \leq t_{punishment} + 0.1 + 0.05. \tag{8}$$

The terminal density (ρ), the number of molecules released per vesicles fusion (n_0), the extracellular volume fraction (α_{vf}) and the Avogadros constant (N_A) parameters are not modified by autoreceptors occupancy. Since vesicular release probability (P_r^{phasic}) and phasic firing rate (υ_{phasic}) are decreased by autoreceptors (Grace, 1991), they are assumed to be inversely proportional to autoreceptors occupancy (Beaulieu and Gainetdinov, 2011; Dreyer and Hounsgaard, 2013). The exact relationship is not known but assumed here as inversely proportional for simplicity. The value 0.334, used to normalize the equation for the control case, corresponds to autoreceptors occupancy. Therefore, Equation (5) indicates that the activation of autoreceptors reduces phasic dopamine release. The values 0.1*s* (Bamford et al., 2018) and 0.05*s* represent the latency and duration of the reward or punishment error prediction signal, respectively. Phasic dopamine release is also proportional to the reward prediction signal (RPE). This issue will be discussed in more details in Section 2.3.

In the occurrence of a punishment, the activity of the dopamine neuron is temporarily suppressed (both tonic and phasic firing rate fall to zero). According to Equations (1) and (3), this can be simulated in the model assuming $\upsilon_{tonic} = 0$ which corresponds to the following differential equation:

$$\frac{dC_{DA}(t)}{dt} = -\frac{V_{max} C_{DA}(t)}{k_m + C_{DA}(t)} - k_{rem} C_{DA}(t). \tag{9}$$

With the parameters we used, this equation requires about 500 ms to reach the new equilibrium with $C_{DA} = 0$, which is close to the duration of dopamine neuron activity suppression after the absence of an expected reward (Schultz et al., 1997). However, the time to reach this equilibrium may vary as a function of the previous discharge rate, tonic dopamine level, or reuptake. To simplify the model, the value $C_{DA} = 0$ was directly applied at the same time as for the phasic dopamine discharge associated with a reward, as shown in Equations (7) and (8). Setting the dopamine concentration at zero instantaneously when a punishment occurs is a simplification of the physiologic mechanisms and the pause in the firing rate was defined as in Dreyer et al. (2010). This simplification was used since the purpose of this work was to study the behaviors in a qualitative manner. In future work, we will implement more physiologic parameters with their variability.

In the model, autoreceptors occupancy depends on the overall dopamine concentration (tonic and phasic). It could be argued that, due to diffusion, only a fraction of phasic dopamine reaches autoreceptors and thus alters the release. Simulations were performed to integrate this concentration gradient on phasic dopamine reaching autoreceptors, but the results were not significantly different (not shown here), therefore the version presented here was chosen for simplicity.

Finally, dopamine molecules can bind to dopaminergic receptors, corresponding to D_1 and D_2 receptors in the current work. The occupancy of receptors of type $i \in \{1, 2\}$ in time is given by the following equation:

$$D_i(t) = \frac{B_{max}^{D_i} C_{DA}(t)}{k_D^{D_i} + C_{DA}(t)} \tag{10}$$

where $B_{max}^{D_i}$ and $k_D^{D_i}$ are the maximal concentration and dissociation constant of type i receptors, respectively. Receptors occupancy will be used in the neurocomputational model of basal ganglia as the postsynaptic effect of dopamine on the neurons in the different neurotransmission pathways (Hille, 1992).

The parameter values for the dopamine dynamics model are given in Table 1. As mentioned in this Table, the dopaminergic terminal density was adapted. As this density is inhomogeneous (Dreyer, 2014; Fuller et al., 2019), its value was set to obtain a tonic dopamine concentration in the control group of 0.02

TABLE 1 Parameters value in the dopamine dynamic model.

Parameters	Description	Value	Literature value	Reference
V_{max}	Maximal reuptake rate by DATs	Control : 1.2 $\mu M/Ls$ Dopamine imbalance : 1.8 $\mu M/Ls$	[0.2 4.3]	May et al., 1988; Nicholson, 1995; Schönfuss et al., 2001; Fuller et al., 2019
k_m	apparent Michaelis-Menten constant	0.15 $\mu M/L$	[0.1, 0.2]	May et al., 1988; Horn, 1990; John et al., 2006; Fuller et al., 2019
k_{rem}	Removal rate	0.04 s^{-1}	0.04	Dreyer and Hounsgaard, 2013
k_{on}	On-rate for DA binding to presynaptic autoreceptors	10 $\mu M^{-1}s^{-1}$	10	Dreyer and Hounsgaard, 2013
k_{off}	Off-rate for DA binding to presynaptic autoreceptors	0.4 s^{-1}	0.4	Dreyer and Hounsgaard, 2013
ρ	Density of dopamine terminals in striatum	0.025 · 10^{15} terminals/L	adapted	
α_{vf}	Volume fraction of extracellular space	0.21	[0.19, 0.22]	Syková and Nicholson, 2008
n_0	Number of dopamine molecules released during vesicle fusion	3,000 molecules/terminal	3,000	Pothos et al., 1998; Dreyer, 2014
N_A	Avogadros constant	6.02214076 · 10^{23} M^{-1}	6.02214076 · 10^{23}	
P_r^{phasic}	Vesicle release probability	0.06	[0.025, 0.15]	Dreyer and Hounsgaard, 2013
P_r^{tonic}	Vesicle release probability	0.06	[0.025, 0.15]	Dreyer and Hounsgaard, 2013
υ_{tonic}	Average tonic firing rate	4 s^{-1}	[4,5]	Fennell et al., 2020
υ_{phasic}	Average phasic firing rate	40 s^{-1}	[20,100]	Fennell et al., 2020
B_{max}^{D1}	D_1 receptor maximal density	1.6 $\mu M/L$	1.6	Hunger et al., 2020
k_D^{D1}	D_1 receptor dissociation constant	1 $\mu M/L$	1	Rice and Cragg, 2008
B_{max}^{D2}	D_2 receptor maximal density	0.08 $\mu M/L$	0.08	Hunger et al., 2020
k_D^{D2}	D_2 receptor dissociation constant	0.01 $\mu M/L$	0.01	Rice and Cragg, 2008

$\mu M/L$ as reported in the literature (Wanat et al., 2009; Hunger et al., 2020).

Using the developed model, two groups of virtual individuals were created: control and dopamine imbalance individuals. The difference between the two groups lies in the modification of the V_{max} parameter of Equation (1). From a mathematical standpoint, the parameter k_m could also have been decreased to obtain similar results.

2.2. Neurocomputational model of basal ganglia

Tonic and phasic dopamine are coding prediction error signals in the basal ganglia (Schultz, 2017). ADHD is associated with dopamine dysfunctions in the cortex and the basal ganglia (Giedd et al., 2001; Seidman et al., 2005; Nakao et al., 2011; Cubillo et al., 2012; Frodl and Skokauskas, 2012; Oldehinkel et al., 2016). Hence, a neurocomputational model of basal ganglia with a learning procedure was added to the dopamine dynamics model.

The neurocomputational model presented here is an adaptation from the model developed in Baston et al. (2016). It involves the temporal neural activity in the cortex, the thalamus and the different regions of the basal ganglia (striatum, globus pallidus pars interna and pars externa, and subthalamic nucleus), with a representation of the external stimulus *S*. The neuronal activities are normalized to obtain a value between 0 and 1. The connection between each region follows three neurotransmission pathways: direct, indirect and hyperdirect. The direct pathway promotes movement, the indirect inhibits it, and the hyperdirect pathway suppresses erroneous movements. D_1 and D_2 receptors occupancy have an excitatory effect in the direct pathway and an inhibitory effect in the indirect pathway, respectively. Both pathways are potentiated by the effect of cholinergic interneurons, also included in the model.

A representation of the neurocomputational model of basal ganglia is given in Figure 2. Each region of the model is divided into four action channels, representing different alternative choices. This division allow investigating the response of basal ganglia to various target stimuli. Neural activity in each action channel is computed through an ordinary differential equation, simulating neural dynamics, and a sigmoidal relationship, which mimics the typical non-linear phenomena of the neurons (lower threshold and upper saturation). The input to each differential equation is calculated by summing all the upstream activities converging to that neuron, weighted by the synaptic strength. The synaptic weight matrices correspond to the weight of connections between the regions for all four action channels.

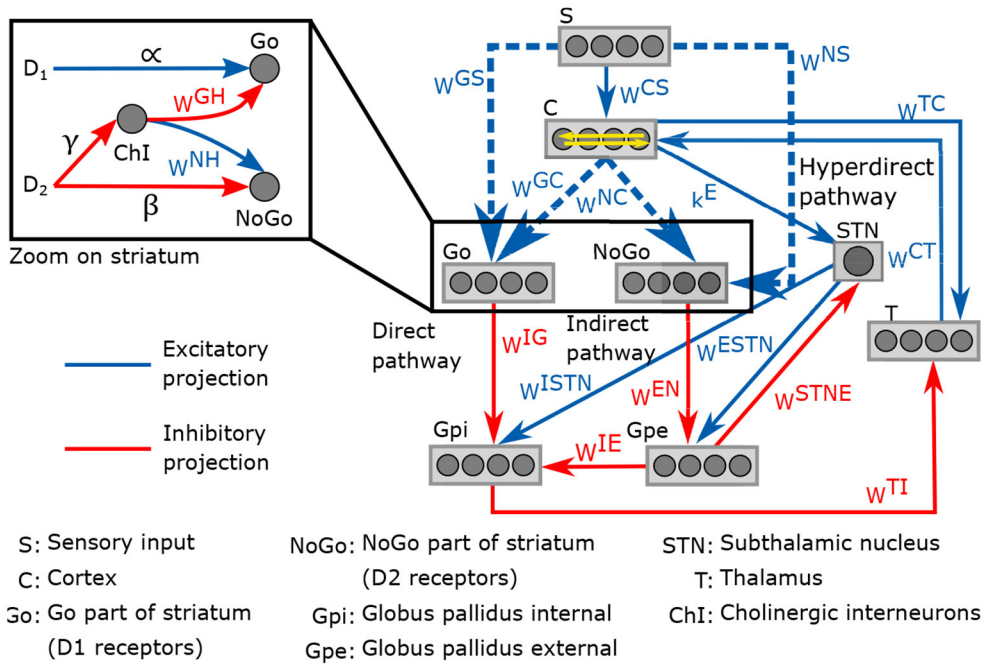


FIGURE 2
Schematic representation of the neurocomputational model of basal and its four action channels ganglia.

Equations and parameter values of the model can be found in the Supplementary Material.

2.3. Learning in the basal ganglia

Impairments in reinforcement learning are thought to be involved in ADHD (Sagvolden et al., 2005; Tripp and Wickens, 2008; Alexander and Farrelly, 2018). Therefore, we included a reinforcement learning process with reward and punishment prediction error signals in the model. The strength of connections between each region of basal ganglia is given by synaptic weight matrices noted w^{ij}, where i and j are the postsynaptic and presynaptic regions, respectively. The values of these weights can be modified by the learning process. For simplicity, only matrices related to striatum, w^{GS}, w^{NS}, w^{GC}, w^{NC}, were considered to be plastic; these connections are represented by dashed lines in Figure 2. The matrices w^{GC} and w^{NC} are diagonal while w^{GS} and w^{NS} are full matrices. At the beginning of the learning process, these weight matrices are in a naive state, with no differentiation between the actions channels. Here are the initial value of the matrices:

$$w^{GC} = w^{NC} = \begin{bmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}, \tag{11}$$

$$w^{GS} = w^{NS} = \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 \end{bmatrix}. \tag{12}$$

We here give the details of a typical trial of the learning process. A stimulus representation S is sent for 800 ms to each action channel. One channel will receive a strong stimulus of value 1, another one receives a weaker stimulus of value 0.2, while the two others receive even weaker stimuli with a value of 0.1 each. In the present work, we used an input vector with the same dimension as the number of possible actions, with a higher value (close to 1) at the same position of the rewarded action, and a smaller value at the positions of the punished actions, just to simplify the final analysis of the synapses. An input vector with different dimensions and with different values could be used as well, resulting in a more complex pattern of synapses. The idea here is to simply associate an input vector to a "winner takes all" output vector, considered as the selected response. The possible considered vectors for S are $S = \begin{bmatrix} 1 & 0.2 & 0.1 & 0.1 \end{bmatrix}$, $S = \begin{bmatrix} 0.2 & 1 & 0.1 & 0.1 \end{bmatrix}$, $S = \begin{bmatrix} 0.1 & 0.1 & 0.2 & 1 \end{bmatrix}$ and $S = \begin{bmatrix} 0.1 & 0.1 & 1 & 0.2 \end{bmatrix}$. Neuronal activity in all regions of basal ganglia are computed for 800 ms. An action is considered to have been performed or chosen if the activity in its related action channel in the cortex is above 0.9, while the activity in all other channels is close to zero, using the winner-takes-all dynamics implemented in the cortex.

We used a fixed scale of prediction error values throughout learning. The prediction error is the discrepancy between observed and expected outcome, and a naive subject cannot predict whether the response would be correct or not. If the chosen action is in the action channel with the highest value of S, a reward prediction error of 1 is attributed. If however the second highest value (0.2) is chosen, a smaller reward prediction error of 0.1 is attributed. A punishment prediction error is given when the lowest value (0.1) is chosen. Rewards prediction errors are signaled by phasic dopamine peaks governed by Equation (6). When a punishment prediction error occurs, dopamine concentration drops to zero. This is equivalent to providing the virtual subjects with rewards and punishments, but we delivered directly the reward/punishment prediction error dopamine signals. These prediction errors are the differences between received and predicted rewards (Schultz, 2016), where here the virtual patient always predicts a reward when an action is chosen. This process is repeated over 1,000 trials (epochs). Once the learning procedure is complete, the model is expected to effectively differentiate between weak and strong stimuli, so that responses occur only when strong stimuli are applied.

The resulting rewards/punishments prediction error signal will lead to a modification of the synaptic weights contained in the matrices. These weights modifications during the learning process are dictated by the Hebb Rule, which states two neurons having both high activity will strengthen their connection, whereas connection will weaken in case of neurons with opposite activity. The Hebb rule describes how much the weights are increased or decreased at each step of the training procedure. In particular, the following equation holds at each step to assign a new synaptic value, Baston and Ursino (2015):

$$w^{AB} \leftarrow w^{AB} + \Delta w^{AB}, \tag{13}$$

where w^{AB} represents the matrix containing the weights from the presynaptic region B to the postsynaptic region A, with A being either S or C in Figure 2 and B being either G (Go) or N (NoGo) in the same figure, and Δw^{AB} is the synaptic change computed at that step. Each row in these matrices represent the synapses entering the postsynaptic neuron, and each column those emerging from the presynaptic one. Hence, all matrices have 4×4 dimensions in the work presented here. This modification of the synaptic weights happens once every epoch between a latency period of $0.1s$ and for a duration of $0.05s$ once an action is chosen. The latency and duration are the same as the ones for the reward/punishment error prediction signal. The individual elements at position ij in the array Δw^{AB} are computed through the following equation (Hebb rule):

$$\Delta w_{ij}^{AB} = \phi \cdot (y_j^B - \vartheta_{presynaptic})^+ (y_i^A - \vartheta_{postsynaptic}), \tag{14}$$

where y_j^B is the activity of the presynaptic neuron in the action channel j of the region B, y_i^A is the activity of the postsynaptic neuron in the action channel i of the region A and $\vartheta_{presynaptic}$, $\vartheta_{postsynaptic}$ the pre- and postsynaptic thresholds. The positive part function $([]^+)$ ensures that learning occurs only if the presynaptic neurons are excited and their activity is above the threshold. Dopamine is thought to have the ability to modulate synaptic plasticity, although the exact relationship does not seem to be established (Reynolds and Wickens, 2002; Frémaux and Gerstner, 2015; Madadi Asl et al., 2019). From previous work, it seemed reasonable to assume a proportional relationship with dopamine ratio and RPE. Of course, in case of diagonal matrices (w^{GC} and w^{NC}), only the elements with $i = j$ are trained, compared to non-diagonal matrices w^{GS} and w^{NS} where all elements are trained. The gain parameter ϕ is proportional to the reward prediction error since, for example, a large reward prediction error will lead to a larger variation in the synaptic value than a small reward prediction error. The gain parameter is also proportional to the ratio of phasic peak and tonic dopamine. This ratio is calculated beforehand and considered as a constant. The equation is the following:

$$\phi = 0.0013 \cdot \mid RPE \mid \cdot \text{DA ratio}, \tag{15}$$

$$\text{DA ratio} = \left(\frac{C_{DA}^{phasic} - C_{DA}^{tonic}}{C_{DA}^{tonic}} \right). \tag{16}$$

The dopamine ratio is higher in the dopamine imbalance group (with a value of ~ 8.3) compared to the control one (with a value of ~ 3), so the gain parameter ϕ is higher.

2.4. Simulation of virtual patients groups

The control and dopamine imbalance groups, with 10 virtual subjects each, were created with the model. The only difference between the two groups is in the value of V_{max}. A higher rate of dopamine recapture is expected to lower the dopamine tonic concentration which in turn is expected to increase the phasic dopamine concentration, and thus in the tonic phasic dopamine ratio, through a lower occupancy of autoreceptor. The steps of the learning procedure of a subject are summarized below.

1. The synaptic weight matrices w^{GS}, w^{NS}, w^{GC}, w^{NC} start in a naive state.
2. Out of the four choices ($S = \begin{bmatrix}1 & 0.2 & 0.1 & 0.1\end{bmatrix}$, $S = \begin{bmatrix}0.2 & 1 & 0.1 & 0.1\end{bmatrix}$, $S = \begin{bmatrix}0.1 & 0.1 & 0.2 & 1\end{bmatrix}$ and $S = \begin{bmatrix}0.1 & 0.1 & 1 & 0.2\end{bmatrix}$), a stimulus S in sent to the cortex for 800 ms. The process will be repeated for the other 3 stimuli in a random order. Noise was added in the cortex, derived from a uniform distribution and ranging from 0 to 0.2. The seed of the noise differentiates between individuals

within a group but not between groups, while the V_{max} value differentiates between the two groups. For example, the control individual #1 has the same noise's seed as dopamine imbalance individual #1, but a different value of V_{max}. At the end of the 800 *ms*, the subject receives either a large or a small reward prediction error signal according to his choice of the action that corresponds to the highest or the second strongest stimulus, respectively. Otherwise, the patient receives a punishment prediction error signal. Transient peaks of phasic dopamine are given accordingly and the Hebb rule is applied to modify the value of synapses. This process is repeated with the three other choices of *S*.

3. Step 2 is repeated 250 times for a total of $250 \times 4 = 1,000$ epochs.
4. Once the training phase is over, the performance of the virtual subjects in each group was assessed in a testing phase. For each individual, the weight matrices were fixed to the values found at the end of the training process to assess their performance.

During the test phase, we also used a four-choice reaction time task. A series of stimuli are presented to the virtual individuals in the different action channels through a signal *S* of the neurocomputational model of BG to the cortex. The stimulus in the targeted action channel has a value of 1 with the addition of noise. Noise is also added in the other action channels directly in the cortex. Each stimulus is presented for 1,800 *ms* with a 500 *ms* pause in between each stimulus. The criterion for a response is an activity in one of the four action channels in the cortex *C*, which constitute the output of the model, greater than 0.9. Due to the winner-takes-all dynamics, the other three channels will then have activity close to zero. For simplicity purposes, a response in the same action channel as the target stimulus is considered as a success. Otherwise it constitutes a failure. Or course successes and failures could have been defined in different ways. The idea here is simply to associate to an input vector, an output vector considered as the correct responses.

During the test phase, there is always a response after a stimulus, being a success or a failure. The number of correct answers or successes represents the performance of the virtual individuals. Each individual is presented 100 stimuli. The mean and standard deviation of the percentage of successes and of the reaction times are computed in each simulated group. Stimulus of different amplitudes were also sent in the first action channel and the responses were recorded to study the differentiation between weak and strong signals. In order to compare the ability of differentiating between weak and strong signals, we repeated the task and computed the cortex activity for different values of noise added to the input signal (*S*).

During the test phase, reaction times were also computed. The reaction time is here defined by the difference between the time at which the neuronal activity in one of the action channels reaches a value of 0.9 and the time at which the stimulus was sent in the sensory representation *S*.

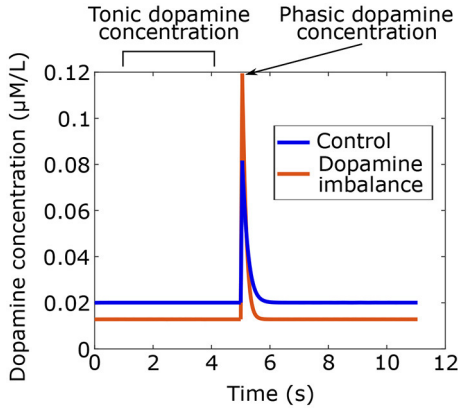


FIGURE 3
Tonic and phasic dopamine concentrations in time simulated with the model for the dopamine imbalance and the control group. In the dopamine imbalance group, tonic dopamine levels are lower due to increased recapture, which leads to decreased autoreceptor occupancy. Reduced autoreceptor occupancy causes higher peak of phasic dopamine because of autoregulation.

3. Results

3.1. Tonic and phasic dopamine release

Using the model, dopamine concentrations were simulated for the two groups as shown in Figure 3. Phasic peaks were created by a burst lasting 0.05 s.

As seen in Figure 3, dopamine imbalance individuals have lower tonic dopamine concentration due to higher dopamine recapture. In turn, autoreceptors regulation causes higher phasic dopamine concentration. This dopamine imbalance will have different impact on the learning process in the basal ganglia.

3.2. Performance during the training phase

During the training phase, we computed the number of trials to obtain 5 successful responses over 10 successive trials. All participants in the normal group reached the learning criterion, but 2 participants in the dopamine imbalance group failed to do so even after 1,000 trials. The number of trials to reach criterion was on average 65.1 ($SD = 52.6$) in the control group, but 20% higher in the dopamine imbalance group, with an average of 85.5 ($SD = 67.8$), excluding those who never reached the criterion.

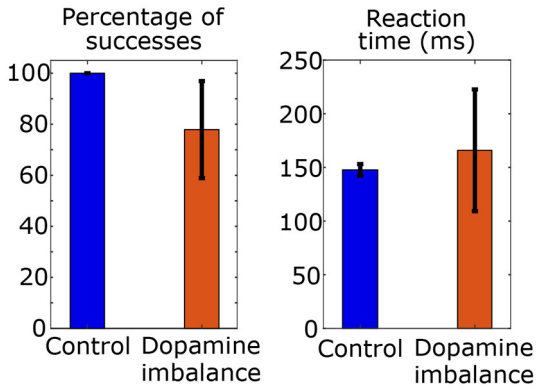


FIGURE 4
Mean and standard deviation of reaction time and percentage of success of choices in a series of 100 stimulus in each group.

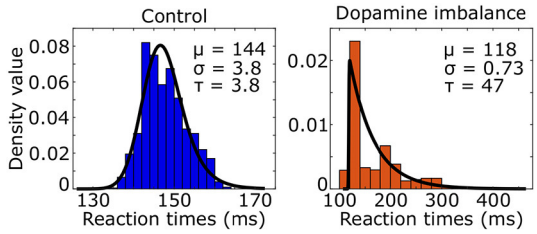


FIGURE 5
Histogram (colored boxes) and fitted density function (black line) of simulated reaction times of the virtual individuals in the control and dopamine imbalance groups.

3.3. Performance during the test phase

In the first task, the mean and standard deviation of the percentage of successes to a series of 100 stimuli and of reaction times are computed in each simulated group and shown in Figure 4.

The mean reaction time in the control group is 148 ms and the standard deviation is 5 ms. The mean percentage of successes is 100 with a standard deviation of 0. In the dopamine imbalance group, the mean reaction time is 166 ms with a standard deviation of 57 ms. The mean percentage of successes is 78 with a standard deviation of 19. As shown in Figure 4, the rate of successes was lower and more variable in the dopamine imbalance group, as compared to the control group. Moreover, the simulated mean reaction times was slower in the dopamine imbalance group than in the control group. In our simulations, the mean and standard deviation of reaction times are respectively, 1.12 and 11.4 times larger in the dopamine imbalance group than in the control group. The significance of the reaction time difference was not evaluated because only 10 patients were simulated in this study to present the model. Also, as described further, the patients in the dopamine imbalance group are heterogeneous and can be divided into three subgroups with different mean reaction times.

We used the ex-Gaussian distribution to estimate the reaction time distribution by combining a normal and an exponential distribution. Three parameters characterized the ex-Gaussian distribution: the mean μ and standard deviation σ of the normal distribution, and τ representing the mean and standard deviation of the exponential part. An ex-Gaussian distribution was fitted to the simulated reaction times of the virtual individuals as seen in Figure 5.

The τ parameter was 12 times larger in the dopamine imbalance group than in the control group (47 vs. 3.8) while the μ parameter was 0.82 times smaller (118 vs. 144).

3.4. Performance with increasing noise

We assessed the performance of the individuals in each group described in the above section by increasing the standard deviation of the noise added to the input signal *S*. A series of 100 stimuli was again presented with noise directly added to the stimulus representation in the cortex *S*, with a mean of 1 and a standard deviation ranging from 0 to 1. As the standard deviation of the noise increases, the probability of having high intensity noise increases which further complicates decision making for the virtual patients and therefore affects the percentage of successes. Figure 6 shows that in the dopamine imbalance group the mean percentage of successes (orange solid line) quickly dropped while the variability (orange shaded area) increased with increasing noise variability. By contrast, in the control group, the performance remained optimal, with no variability, until the noise variability was greater than 0.6.

3.5. Input and output of basal ganglia

During the test phase, we also computed the output activity in the cortex related to the response as a function of the input value of the stimulus. A stimulus of different amplitudes, ranging from 0.1 to 1, is sent in the first action channel while all three other channels receive noise of small amplitude. The mean, the 5th and the 95th output curves of the cortex neuronal activity in the first action channel as a function of the input signal value for each group are shown in Figure 7.

By comparing neural activity at basal ganglia input and output, it is clear that in control subjects, the basal ganglia have a high neural gain. Response-related activity is suppressed until stimulus-related cortical activity reaches 0.5 in the control

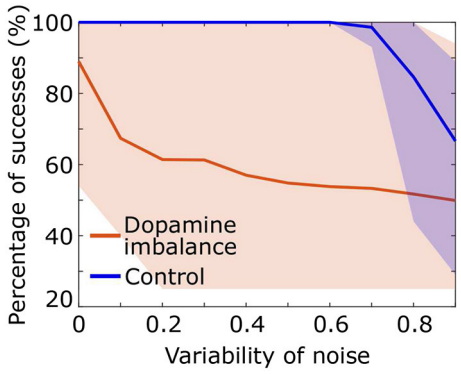


FIGURE 6
Percentage of success as a function of noise standard deviation for each individual in the dopamine imbalance group. The curve of each individual are contained in the shaded area.

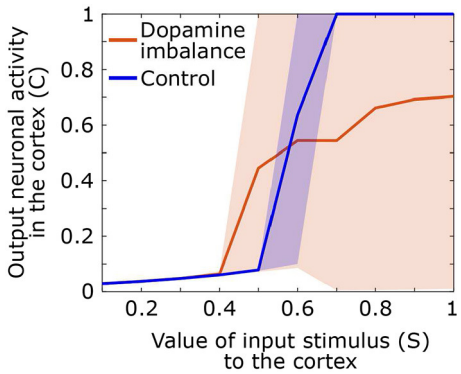


FIGURE 7
Output neuronal activity in the cortex as a function of different input stimuli. Solid line: mean neuronal activity of the individuals in each group, shaded area: 5th and 95th percentiles of neuronal activity of the individuals in the group.

group. Output activity then increases rapidly for an input between 0.5 and 0.7 at which point it remains maximal. In contrast, in the dopamine imbalance group, activity is suppressed up to an input of 0.4, after which the gain increases rapidly but only for stimulus-related activity between 0.4 and 0.5. For stimulus-related activity values between 0.5 and 1, the gain is strongly attenuated as response-related activity increases from 4.5 to 7. However, the most striking aspect of the gain is the extreme variability of the output in the dopamine imbalance group, which ranges from 0 to 1 in response to stimulus-related activity values between 0.7 and 1. In this group, some individuals respond correctly and others have wrong responses which will lead to an output activity close to zero due to the winner-takes-all dynamic, thus inducing high variability. In contrast, in the controls, the variability is almost zero, except for the amplification phase, especially around the inflection point.

3.6. Evolution of synaptic weights

Four synaptic weights matrices were modified during training: w^{GS}, w^{NS} (stimulus-related synaptic weights) and w^{GC}, w^{NC} (response-related synaptic weights). These matrices start in a naive configuration, with no differentiation between the four action channels. They are modified during the training by using the Hebb Rule, with a gain parameter that is proportional to the phasic vs. tonic dopamine ratio.

Over the course of the 1,000 trials in the training phase, the matrix weights changed differently between the two groups, and between individual subjects within each group. Indeed, the trends of synaptic weight evolution were the same for the control and dopamine imbalance groups, but inter-individual differences in synaptic weights and their evolution during learning were much larger in the dopamine imbalance group. Hence, inter-individual differences were much larger at the end of the learning phase in the dopamine imbalance than in the control group. More details on the evolution of the synaptic weight matrices are given in the Supplementary Material.

3.7. History of rewards and punishments prediction errors during training

In the present section, a metric is developed to differentiate the performance in the test phase of the dopamine imbalance group from the control one based on their history during the training phase. During the training process, the history of rewards and punishments is stored in a vector with value 1 for a large reward, 0.1 for a small reward, −1 for a punishment and 0 for no response. It is therefore possible to study the history of each individual and to relate it to his performance in the test phase.

Figure 8 shows the cumulative sum of the history vector for each action channel of the first 5 individuals in each group. A negative cumulative sum results from a series of failures overcoming successes, while a positive cumulative sum would indicate the opposite.

There seems to be an initial phase in which there is an excess of errors. The virtual individuals start in a naive state, meaning no differentiation between the action channels. Hence, the initial responses have a random success rate of 25% and can lead to an excess of errors. In the second phase (> 500 epoch), rewards prediction errors dominate over punishments for all actions.

Individuals from the control group seem to learn each action in a proportional way for all action channels. The individuals in the dopamine imbalance group had a higher number of rewards for some action channels at the expense of the others. In order to quantify the inter-individual differences in learning, a weighted standard deviation (*weighted std*) for the cumulative sum of

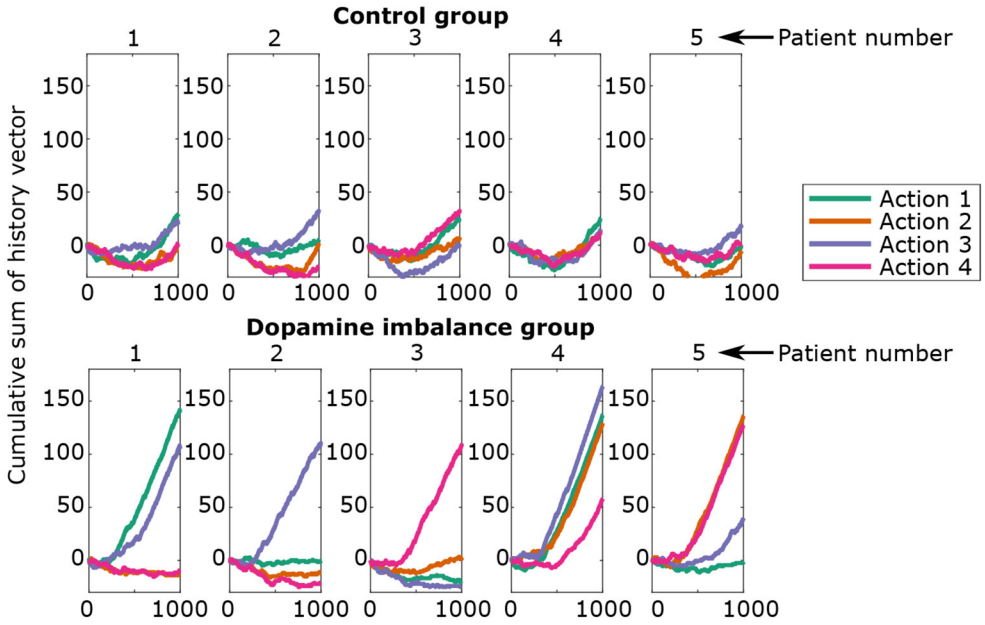


FIGURE 8
Cumulative sum of history vector at each epoch for the first five individuals in each group.

history was computed for each individual, and expressed by the following equations:

$$ratio = \frac{1}{1,000} \frac{\sum_{i=1}^{1,000} \sum_{j=1}^{4} \#\text{negative } cumsum_{action_j}(i)}{\sum_{i=1}^{1,000} \sum_{j=1}^{4} \#\text{positive } cumsum_{action_j}(i)}, \tag{17}$$

$$std_{history} = \frac{1}{1,000} \sqrt{\sum_{i=1}^{1,000} \left(\sum_{j=1}^{4} (cumsum_{action_j}(i) - mean(i))^2 \right)}, \tag{18}$$

$$weighted\ std_{history} = ratio \cdot std_{history} \tag{19}$$

where i is the epoch number, j the action number, $cumsum_{action_j}(i)$ the cumulative sum of history vector for action j at epoch i and $mean(i)$ is the mean of cumulative history at epoch i for all action channels. The standard deviation of the history ($std_{history}$) is weighted by a ratio to take into account the fact that the cumulative sum of history is either positive or negative. The ratio is the sum of negative cumulative sum of history divided by the sum of positive cumulative sum of history, leading to a larger ratio when the negative cumulative sum exceeds the positive one. Division by 1,000 is for scaling. The *weighted* $std_{history}$ was larger in the dopamine imbalance group than the control one. In order to assess the relationship between the training and test phase, a plot of the standard deviation of the reaction times as a function of the *weighted* $std_{history}$ value is depicted in Figure 9. A linear regression (dashed line) and a quadratic function (dashed curve) between the *weighted* $std_{history}$ and the standard deviation of reaction times were applied to the control group and the imbalance group, respectively. The individuals in the dopamine imbalance group could be divided into three subgroups (a, b, and c) along the quadratic regression as seen in Figure 9. Group a contained the individuals with a perfect performance, low μ, low σ and low τ, which explains their proximity to the individuals in the control group. The individuals less than perfect performance were divided into groups b (75% of successes) and c (60% of successes). The distribution of reaction times in the group b is closer to an exponential distribution than to a normal one with low μ and σ but very high τ. These individuals have both fast and very slow reaction times, driving thus the mean to a high value. As the *weighted* $std_{history}$ increases for individuals in group c, the performance further decreased with fewer correct responses, the μ parameters increased, and the σ and τ had intermediate values and were quite similar.

4. Discussion

In the current work, we investigated the effect of phasic vs. tonic dopamine imbalance during reinforcement learning on overt responses and on synaptic weights in the basal ganglia. We altered the phasic vs. tonic ratio by increasing the rate of maximal dopamine reuptake by DATs. As the rate of dopamine reuptake increases, the tonic level of dopamine decreases, which results in a decrease in autoreceptor binding, and in turn in an

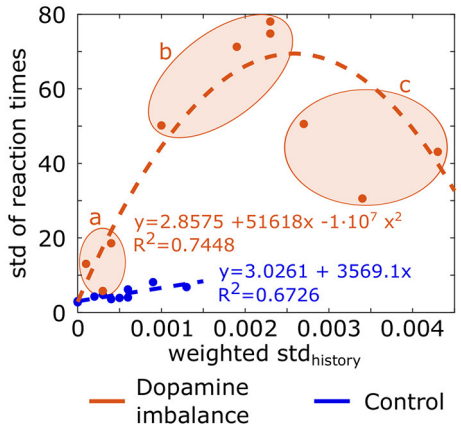


FIGURE 9
Standard deviation (std) of reaction times for each individual in each group as a function of the *weighted std$_{history}$*. Equation of the linear regression and quadratic regression performed, respectively in the control group and dopamine imbalance group are shown. Individuals in the dopamine imbalance group are divided into three sub-groups, a, b, and c.

increase in the phasic response (Ford, 2014). This modification increased the phasic response by about 40%. The values of simulated dopamine concentrations that we found are consistent with those reported in the literature, with a tonic concentration between 0.005 and $0.02\mu M/L$ (Wanat et al., 2009; Hunger et al., 2020), and a phasic concentration ranges between 0.01 and 1 $\mu M/L$ (Wickham et al., 2013). More precisely, phasic dopamine concentrations were estimated to be $\sim$ 0.1 $\mu M/L$ in Bamford et al. (2018).

Clinically, subjects with ADHD consistently show a typical response pattern on a variety of tasks. They generally make more errors than controls and their reaction times are paradoxically both faster and slower, and more variable overall, as compared with healthy controls (Hervey et al., 2006; Huang-Pollock et al., 2012). This variability is primarily due to an excess of slow responses that can be detected by the τ component of an ex-Gaussian distribution (Kofler et al., 2013). This τ parameter best discriminates ADHD subjects (Leth-Steensen et al., 2000) from controls and appears to be a reliable endophenotype, as unaffected siblings showed intermediate values between ADHD subjects and healthy controls (Lin et al., 2015). In the present simulations, the group with dopamine imbalance also showed more variable reaction times, including an excess of very slow responses, as compared with the control group. Specifically, the μ parameter was smaller, reflecting impulsive responses, but the τ was much larger, due to a greater proportion of very slow responses, with a decrease of the σ parameter overall, which reflects the Gaussian variance. Thus, shifting the phasic/tonic dopamine ratio reproduced a response pattern typically seen in ADHD subjects, whereas a model incorporating only a decrease in both phasic and tonic dopamine release did not (Frank et al., 2007). We observed this response pattern in a simple reinforcement learning task while it has been observed in a wide variety of experimental tasks with ADHD subjects. Future studies will need to test whether this response pattern generalizes to other tasks, but it is a possibility insofar as any experimental task has a learning component. Indeed, data are typically collected after participants have reached a performance threshold during a training phase.

The change in reaction time distribution, although most typical of ADHD, is not the only difference we observed. The subjects with a dopamine imbalance also showed a lower and more variable success rate on average. Within the signal detection theory (Stanislaw and Todorov, 1999), the sensory discrimination ability is termed d'. In our simulation, the test phase used a force choice task in which d' is the percentage of successes (Stanislaw and Todorov, 1999). The control group obtained perfect results, but the success rate was decreased by 22% in the dopamine imbalance group. Subjects with ADHD also showed decreased d' in a meta-analysis of continuous performance test (CPT) performance (Huang-Pollock et al., 2012). Furthermore, we tested the effect of noise, matching each individual in the dopamine imbalance group with one individual in the control group for the seed of noise. In both groups, the success rate degraded and became more variable with increasing noise, but the dopamine imbalance group was more sensitive and showed a drop in success and a large variability for low noise levels that did not affect the performance of control subjects. Similarly, children with ADHD have been shown to have lower auditory discrimination ability than controls in the presence of background noise (Tien et al., 2019).

In order to further characterize the response pattern to stimuli of varying intensity we computed the neural gain between the input and the output of the system. A strong gain is associated with a stable attractor (Hauser et al., 2016) in which the system quickly converges to a stable activity pattern. In contrast, a weak gain is characterized by variable attractors that can lead to different unstable and shallow activity patterns. In the present simulation, for stimulus-related input values that always produced a stable response in controls ($\geq$ 0.7), response-related output activity was much more variable in the group with dopamine imbalance. In this group, the more random responses reflected a more exploratory approach where different responses could be produced even for high stimulus-related inputs in the cortex. In experimental situations, subjects with ADHD demonstrated the same type of exploratory approach. In a probabilistic reversal learning task (Hauser et al., 2014), ADHD subjects did not choose their response strictly on the basis of their belief in the value of the stimulus, but more often took an exploratory approach. When the neural gain was estimated by a sigmoidal function, this exploratory approach also resulted in a less steep decision function. The phasic response may reinforce the response to low-intensity sensory events, which could lead to

a more prolonged phase of discovery of new actions in a learning situation (Redgrave et al., 2008).

But the most significant result of the simulation, consistent with our original hypothesis, is that while all at-risk subjects had the same dopamine release imbalance, the ADHD response pattern developed to different degrees depending on the individual learning experience. On average, during this probabilistic learning task with 100% valid feedback, subjects in the dopaminergic imbalance group required more learning trials than controls to reach a success criterion. Again, this replicates a result obtained with ADHD children (Luman et al., 2020). However, the sequence of stimuli was random with a unique seed of noise for each individual within a group, which ultimately resulted in a unique learning environment for each individual within each group. This unique environment was shared with the matched individual in the other group. When we examined separately for each individual the cumulative changes in synaptic weights between cortex and basal ganglia over the course of learning, we found that individuals in the control group showed a similar history regardless of response. In contrast, in the dopamine imbalance group, individuals showed a larger increase in synaptic weight for one or more actions, with onset at different times in the first half of the training phase. As a consequence, the intraindividual differences were much larger in the dopamine imbalance group than in the control group. We computed the weighted standard deviation of the cumulative sum of history to estimate the intraindividual differences during learning. In the control group, using a linear model, we could explain 67% of the variability of individual reaction times during the test phase with the weighted cumulative sum of history. In the control group, however, we had to use a quadratic model to explain the variability between these two measures. Three subgroups of individuals could be distinguished in the dopamine imbalance group (Figure 9). Within a similar range of weighted history variability as the controls, individuals in this subgroup a showed the same perfect performance as the controls. However, the initial slope of the parabola was much steeper than in controls, reflecting the excessive reinforcement for some responses, and the variability of their reaction time was much higher than in controls, but still lower than in the rest of the dopamine imbalance group. This combination of perfect accuracy but high variability in response could define a subthreshold ADHD subgroup, where features of ADHD are already present but do not affect overt accuracy. Closer to the vertex of the parabola, we distinguish a second subgroup b of individuals with weighted history variability larger than the controls (with some negative cumulative weights), and whose accuracy was impaired though not dramatically. The distribution of reaction times contained both fast and very slow responses. Their performance most closely resembled that observed in most of the subjects diagnosed with ADHD as their functioning is clearly impaired. Individuals with extreme weighted history variability (with mostly negative cumulative weights) were hardly learned the stimulus-response association and their performance was even poorer. Their reaction time distribution looked more gaussian with a large variability and very slow mean reaction time. Individuals in this subgroup c could be compared to subjects with a severe ADHD leading to a learning disability.

In conclusion, variability in response history is much greater in subjects with dopamine imbalance, although they were exposed on average to the same learning environment as controls. Intraindividual variability in response times is related to intraindividual variability in experience with the learning environment. It increases when certain responses are reinforced at the expense of other responses during learning, making response selection more difficult in a test phase. But this variability in experience, and therefore also in response times, is much more pronounced in subjects with an imbalance in dopamine release. For subjects in subgroups a and b, the increase in response time variability as a function of weighted learning history variability is approximately linear, but the slope is much steeper than for controls. In these subjects, the increase in phasic dopamine release at the expense of tonic release can excessively strengthen or weaken cortico-striatal synapses associated with different responses and strengthen some responses at the expense of others. These imbalances lead first to an increase in response time variability with a mixture of fast and slow responses, and as these imbalances increase during learning to a decrease in performance in the test phase. In contrast, healthy controls show little variation in the history vector during learning. Consequently, they exhibited a small normal variation in reaction time that was also predicted by the weighted variability of the history with a linear function, but with a smaller slope that reflects a more balanced reinforcement of responses. To the extent that functional connectivity between the striatum and cortex reflects changes in their synaptic connections, our model is consistent with the observed correlation between inattention and hyperactivity/impulsivity scores in networks involving the striatum (Oldehinkel et al., 2016). As these changes are marked by the strengthening of some connections at the expense of others, this also explains the contradictory results in studies comparing an ADHD group with a control group that report either hypoconnectivity (Cao et al., 2009; Posner et al., 2013) or hyperconnectivity (Tian et al., 2006; Costa Dias et al., 2013) within the cortico-striato-thalamo-cortical loops in ADHD.

This qualitative agreement we observed between simulations and experimental findings is remarkable because it is achieved by altering a single parameter of dopaminergic terminal functioning, which results in phasic-tonic imbalance in dopamine release. Frank's model (Frank et al., 2007), which implemented a reduction in both phasic and tonic dopamine levels, needed to incorporate a noradrenergic component with an increased tonic vs. phasic ratio in order to mimic the increase in reaction time variability observed in ADHD subjects. These

authors did not further analyze the distribution of reaction time as a function of noradrenaline release imbalance, so we do not know whether this model reproduces the typical ex-gaussian distribution that we found. Obviously, our results do not rule out noradrenergic dysfunction in ADHD. There is strong evidence of it. Drugs modulating norepinephrine transmission by blocking the NET such as atomoxetine (Schwartz and Correll, 2014) or the alpha2-adrenergic agonists such as clonidine or guanfacine (Arnsten et al., 2007) are effective treatments for ADHD. Methylphenidate significantly occupies NET at clinically relevant doses in humans (Hannestad et al., 2010) and atomoxetine showed a dose-dependent occupancy of NET in monkeys (Ding et al., 2014). NET availability was decreased in a group of adult ADHD subjects in attention-relevant regions (frontal, parietal, thalamic, cerebellar), especially in the right hemisphere (Ulke et al., 2019). The shift from exploitation to exploration behavior has been proposed to be mediated by the firing mode of norepinephrine neurons in the locus coeruleus (Aston-Jones and Cohen, 2005). However, the results of our model suggest that norepinephrine is not necessary to reproduce the typical ADHD response pattern observed in experimental reaction time tasks, which may be accounted for by a phasic/tonic imbalance in dopaminergic activity alone. This reinforces the concept of ADHD as a heterogeneous disorder, in which the same response patterns may be produced by different dysfunctions, whether or not interacting.

Grace's model locates the mechanism of phasic/tonic imbalance of dopamine release at the level of presynaptic regulation, and not at the level of neuron activity itself (Grace, 2001). In our modeling, this presynaptic imbalance may be caused by changes in DAT reuptake (Equation 1), DA removal (Equation 1), autoreceptor occupancy (Equation 2), or a combination of these factors. We chose to increase V_{max}. Yet, it is known that the binding potential of DAT, like that of D2/3 receptors, decreased in adults with ADHD (Volkow et al., 2009), and increased with long-term stimulant treatment (Fusar-Poli et al., 2012; Wang et al., 2013). DAT binding potential may reflect the density of dopamine terminals, but it is also regulated over the long term by dopamine tone, decreasing when extracellular dopamine is decreased and increasing when extracellular dopamine is increased (Zahniser and Doolen, 2001). The decrease in DAT density in ADHD adults could thus be the consequence of a long-term adaptation to a chronic low tonic dopamine level, and its increase during chronic treatment related to the restoration of a higher level. Our model does not consider these long-term changes, but only evaluates the short-term effects of the dopamine release imbalance on learning. Changes in DAT binding potential in these studies (Volkow et al., 2009; Fusar-Poli et al., 2012; Wang et al., 2013) are thus not incompatible with our choice of increasing V_{max}. Moreover, in Equation (1), V_{max} or K_m could have been modified to obtain similar results. Beyond its density, the functional dynamics of DAT (characterized by its K_m) may be altered by other changes (such as ion dependence, or conformational balance) that may themselves be related to genetic mutations. For example, a variable number tandem repeat (VNTR) in the 3' regulatory region of the DAT gene results in two main forms (long 10R and short 9R). The10R form has been found to be associated with ADHD, at least in children and youth (Grünblatt et al., 2019), and can combine with another VNTR to produce haplotypes (Gizer et al., 2009; Franke et al., 2010), susceptible to be modulated by epigenetic factors (Xu et al., 2015; Lambacher et al., 2020; Tonelli et al., 2020). Genetic and epigenetic changes may ultimately affect DAT dynamics. Instead of increasing V_{max}, we could have also increased K_{rem} in the removal part of Equation (1). Catechol-O-methyltransferase (COMT) regulates dopamine level by degrading it, mainly in the prefrontal cortex (PFC). COMT haplotypes showed different level of activity (Diatchenko et al., 2005; Nackley et al., 2006) and it has been proposed that a decrease in COMT activity in the PFC could increase firing of pyramidal neurons and glutamate transmission in basal ganglia, leading to an increase in tonic dopamine, which in turn results in a decrease in phasic dopamine (Bilder et al., 2004). However, this model has yet to be convincingly proven (Nolan et al., 2004; Rosa et al., 2010), as the association of genetic variants of COMT with ADHD (Kang et al., 2020). In our model, dopamine phasic release decreases with autoreceptor occupancy (Benoit-Marand et al., 2001). However, the interactome governing dopamine release is much more complex and includes transporters, G-protein-coupled receptors, ion channels, intracellular signaling modulators, and protein kinases. The phasic/tonic ratio of dopamine release is thus a complex trait that varies along a continuum whose regulation is still poorly understood, but where DAT plays a key role. Increasing V_{max} was not proposed as a unique cause for a complex trait such as ADHD, but rather as a means to shift the dopamine release to a more unbalanced phasic/tonic ratio that can lead to an ADHD-like phenotype through interactions with specific learning experiences. In this perspective, we believe that our model has sound biological and clinical plausibility.

The present model has limitations. Some parameters in the model might not be identifiable and the exact value of some others is not known. The values assigned to parameters is the same for all the subjects within each group and does not reflect the interindividual variability found in control and clinical groups, but support the proof-of-concept approach. The task we used does not require inhibitory processes, which will have to be tested in further studies. Also, in further studies the dysfunctions in the noradrenergic system should also be included to better simulate the pathophysiology of ADHD. Nevertheless, our model is a first step to investigate the implication of the dopaminergic system in ADHD with a mechanistic approach.

To conclude, our model opens perspectives to be used as a platform to generate and test hypothesis regarding the

dopaminergic system in ADHD. The effect of medication on performance, the impact of different patterns of noise, the difference in commission and omission errors and the continuum in the severity of ADHD symptoms could be explored with this model. The effect of gradual changes in the tonic and phasic dopamine ratio will be simulated in further studies to see if the effects on the associated behavior are continuous or discontinuous with a threshold. The model could also be used to simulate a no-response task where the patient is asked to withhold the response when a certain stimulus is sent like in the go/no-go task performed in clinical practice. This modeling approach is a promising step toward the development of an integrative model of the dopaminergic system in basal ganglia for the elucidation of its associated pathologies.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

This work makes up a portion of the doctoral thesis of FV-V. FV-V, PR, MU, and FN: construction of the model and writing of the paper. FV-V: numerical simulations. All authors contributed to the article and approved the submitted version.

Funding

FV-V received a scholarship from the Natural Sciences and Engineering Research Council (NSERC), Canada through the PGS-D program. Support was also provided by NSERC-Industrial Chair in Pharmacometrics funded by Novartis, Pfizer and Syneos, as well as FRQNT Projet d'équipe (FN).

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2022.849323/full#supplementary-material

References

Alexander, L., and Farrelly, N. (2018). Attending to adult adhd: a review of the neurobiology behind adult adhd. *Ir. J. Psychol. Med.* 35, 237–244. doi: 10.1017/ipm.2017.78

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders*. Paris: Elsevier Masson.

Arnsten, A. F., Scahill, L., and Findling, R. L. (2007). alpha2-adrenergic receptor agonists for the treatment of attention-deficit/hyperactivity disorder: emerging concepts from new data. *J. Child Adolesc. Psychopharmacol.* 17, 393–406. doi: 10.1089/cap.2006.0098

Aston-Jones, G., and Cohen, J. D. (2005). Adaptive gain and the role of the locus coeruleus-norepinephrine system in optimal performance. *J. Comp. Neurol.* 493, 99–110. doi: 10.1002/cne.20723

Badgaiyan, R. D., Sinha, S., Sajjad, M., and Wack, D. S. (2015). Attenuated tonic and enhanced phasic release of dopamine in attention deficit hyperactivity disorder. *PLoS ONE* 10, e0137326. doi: 10.1371/journal.pone.0137326

Bamford, N. S., Wightman, R. M., and Sulzer, D. (2018). Dopamine's effects on corticostriatal synapses during reward-based behaviors. *Neuron* 97, 494–510. doi: 10.1016/j.neuron.2018.01.006

Baston, C., Contin, M., Calandra Buonaura, G., Cortelli, P., and Ursino, M. (2016). A mathematical model of levodopa medication effect on basal ganglia in parkinson's disease: an application to the alternate finger tapping task. *Front. Hum. Neurosci.* 10, 280. doi: 10.3389/fnhum.2016.00280

Baston, C., and Ursino, M. (2015). A biologically inspired computational model of basal ganglia in action selection. *Comput. Intell. Neurosci.* 2015, 187417. doi: 10.1155/2015/187417

Beaulieu, J.-M., and Gainetdinov, R. R. (2011). The physiology, signaling, and pharmacology of dopamine receptors. *Pharmacol. Rev.* 63, 182–217. doi: 10.1124/pr.110.002642

Belujon, P., and Grace, A. A. (2015). Regulation of dopamine system responsivity and its adaptive and pathological response to stress. *Proc. Biol. Sci.* 282, 2516. doi: 10.1098/rspb.2014.2516

Benoit-Marand, M., Borrelli, E., and Gonon, F. (2001). Inhibition of dopamine release via presynaptic d2 receptors: time course and functional characteristics *in vivo*. *J. Neurosci.* 21, 9134–9141. doi: 10.1523/JNEUROSCI.21-23-09134.2001

Bilder, R. M., Volavka, J., Lachman, H. M., and Grace, A. A. (2004). The catechol-o-methyltransferase polymorphism: relations to the tonic-phasic dopamine hypothesis and neuropsychiatric phenotypes. *Neuropsychopharmacology* 29, 1943–1961. doi: 10.1038/sj.npp.1300542

Blum, K., Chen, A. L.-C., Braverman, E. R., Comings, D. E., Chen, T. J. H., Arcuri, V., et al. (2008). Attention-deficit-hyperactivity disorder and reward deficiency syndrome. *Neuropsychiatr. Dis. Treat.* 4, 893–918. doi: 10.2147/NDT.S2627

Budygin, E. A., John, C. E., Mateo, Y., and Jones, S. R. (2002). Lack of cocaine effect on dopamine clearance in the core and shell of the nucleus

accumbens of dopamine transporter knock-out mice. *J. Neuroscie*. 22, RC222. doi: 10.1523/JNEUROSCI.22-10-j0002.2002

Burt, S. A. (2009). Rethinking environmental contributions to child and adolescent psychopathology: a meta-analysis of shared environmental influences. *Psychol. Bull.* 135, 608–637. doi: 10.1037/a0015702

Burt, S. A. (2010). Are there shared environmental influences on attention-deficit/hyperactivity disorder? reply to wood, buitelaar, rijsdijk, asherson, and kuntsi [corrected] (2010). *Psychol. Bull.* 136, 341–343. doi: 10.1037/a0019116

Burt, S. A., Larsson, H., Lichtenstein, P., and Klump, K. L. (2012). Additional evidence against shared environmental contributions to attention-deficit/hyperactivity problems. *Behav. Genet*. 42, 711–721. doi: 10.1007/s10519-012-9545-y

Cao, X., Cao, Q., Long, X., Sun, L., Sui, M., Zhu, C., et al. (2009). Abnormal resting-state functional connectivity patterns of the putamen in medication-naïve children with attention deficit hyperactivity disorder. *Brain Res*. 1303, 195–206. doi: 10.1016/j.brainres.2009.08.029

Costa Dias, T. G., Wilson, V. B., Bathula, D. R., Iyer, S. P., Mills, K. L., Thurlow, B. L., et al. (2013). Reward circuit connectivity relates to delay discounting in children with attention-deficit/hyperactivity disorder. *Eur. Neuropsychopharmacol*. 23, 33–45. doi: 10.1016/j.euroneuro.2012.10.015

Cubillo, A., Halari, R., Smith, A., Taylor, E., and Rubia, K. (2012). A review of fronto-striatal and fronto-cortical brain abnormalities in children and adults with attention deficit hyperactivity disorder (adhd) and new evidence for dysfunction in adults with adhd during motivation and attention. *Cortex* 48, 194–215. doi: 10.1016/j.cortex.2011.04.007

Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., et al. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet*. 51, 63–75. doi: 10.1038/s41588-018-0269-7

Diatchenko, L., Slade, G. D., Nackley, A. G., Bhalang, K., Sigurdsson, A., Belfer, I., et al. (2005). Genetic basis for individual variations in pain perception and the development of a chronic pain condition. *Human Mol. Genet.* 14, 135–143. doi: 10.1093/hmg/ddi013

Dickstein, D. P. (2018). Paying attention to attention-deficit/hyperactivity disorder. *JAMA Netw. Open* 1, e181504. doi: 10.1001/jamanetworkopen.2018.1504

Dickstein, S. G., Bannon, K., Castellanos, F. X., and Milham, M. P. (2006). The neural correlates of attention deficit hyperactivity disorder: an ale meta-analysis. *J. Child Psychol. Psychiatry* 47, 1051–1062. doi: 10.1111/j.1469-7610.2006.01671.x

Ding, Y.-S., Naganawa, M., Gallezot, J.-D., Nabulsi, N., Lin, S.-F., Ropchan, J., et al. (2014). Clinical doses of atomoxetine significantly occupy both norepinephrine and serotonin transports: Implications on treatment of depression and adhd. *Neuroimage* 86, 164–171. doi: 10.1016/j.neuroimage.2013.08.001

Douma, E. H., and de Kloet, E. R. (2020). Stress-induced plasticity and functioning of ventral tegmental dopamine neurons. *Neurosci. Biobehav. Rev*. 108, 48–77. doi: 10.1016/j.neubiorev.2019.10.015

Dreyer, J. K. (2014). Three mechanisms by which striatal denervation causes breakdown of dopamine signaling. *J. Neurosci.* 34, 12444–12456. doi: 10.1523/JNEUROSCI.1458-14.2014

Dreyer, J. K., Herrik, K. F., Berg, R. W., and Hounsgaard, J. D. (2010). Influence of phasic and tonic dopamine release on receptor activation. *J. Neurosci.* 30, 14273–14283. doi: 10.1523/JNEUROSCI.1894-10.2010

Dreyer, J. K., and Hounsgaard, J. (2013). Mathematical model of dopamine autoreceptors and uptake inhibitors and their influence on tonic and phasic dopamine signaling. *J. Neurophysiol.* 109, 171–182. doi: 10.1152/jn.00502.2012

Faraone, S. V., and Larsson, H. (2019). Genetics of attention deficit hyperactivity disorder. *Mol. Psychiatry* 24, 562–575. doi: 10.1038/s41380-018-0070-0

Fennell, A. M., Pitts, E. G., Sexton, L. L., and Ferris, M. J. (2020). Phasic dopamine release magnitude tracks individual differences in sensitization of locomotor response following a history of nicotine exposure. *Sci. Rep*. 10, 173. doi: 10.1038/s41598-019-56884-z

Ford, C. P. (2014). The role of d2-autoreceptors in regulating dopamine neuron activity and transmission. *Neuroscience* 282, 13–22. doi: 10.1016/j.neuroscience.2014.01.025

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated parkinsonism. *J. Cogn. Neurosci*. 17, 51–72. doi: 10.1162/0898929052880093

Frank, M. J., and Claus, E. D. (2006). Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychol. Rev*. 113, 300–326. doi: 10.1037/0033-295X.113.2.300

Frank, M. J., Santamaria, A., O'Reilly, R. C., and Willcutt, E. (2007). Testing computational models of dopamine and noradrenaline dysfunction in attention deficit/hyperactivity disorder. *Neuropsychopharmacology* 32, 1583–1599. doi: 10.1038/sj.npp.1301278

Franke, B., Vasquez, A. A., Johansson, S., Hoogman, M., Romanos, J., Boreatti-Hümmer, A., et al. (2010). Multicenter analysis of the slc6a3/dat1 vntr haplotype in persistent adhd suggests differential involvement of the gene in childhood and persistent adhd. *Neuropsychopharmacology* 35, 656–664. doi: 10.1038/npp.2009.170

Frémaux, N., and Gerstner, W. (2015). Neuromodulated spike-timing-dependent plasticity, and theory of three factor learning rules. *Front. Neural Circ.* 9, 85. doi: 10.3389/fncir.2015.00085

Frodl, T., and Skokauskas, N. (2012). Meta-analysis of structural mri studies in children and adults with attention deficit hyperactivity disorder indicates treatment effects. *Acta Psychiatr. Scand*. 125, 114–126. doi: 10.1111/j.1600-0447.2011.01786.x

Fuller, J. A., Burrell, M. H., Yee, A. G., Liyanagama, K., Lipski, J., Wickens, J. R., et al. (2019). Role of homeostatic feedback mechanisms in modulating methylphenidate actions on phasic dopamine signaling in the striatum of awake behaving rats. *Progr. Neurobiol.* 182, 101681. doi: 10.1016/j.pneurobio.2019.101681

Fusar-Poli, P., Rubia, K., Rossi, G., Sartori, G., and Balottin, U. (2012). Striatal dopamine transporter alterations in adhd: pathophysiology or adaptation to psychostimulants? a meta-analysis. *Am. J. Psychiatry* 169, 264–272. doi: 10.1176/appi.ajp.2011.11060940

Giedd, J. N., Blumenthal, J., Molloy, E., and Castellanos, F. X. (2001). Brain imaging of attention deficit/hyperactivity disorder. *Ann. N. Y. Acad. Sci.* 931, 33–49. doi: 10.1111/j.1749-6632.2001.tb05772.x

Gizer, I. R., Ficks, C., and Waldman, I. D. (2009). Candidate gene studies of adhd: a meta-analytic review. *Hum. Genet*. 126, 51–90. doi: 10.1007/s00439-009-0694-x

Grace, A. A. (1991). Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: a hypothesis for the etiology of schizophrenia. *Neuroscience* 41, 1–24. doi: 10.1016/0306-4522(91)90196-U

Grace, A. A. (2001). "Psychostimulant actions on dopamine and limbic system function: Relevance to the pathophysiology and treatment of adhd," in *Stimulant Drugs and ADHD: Basic and Clinical Neuroscience* (Oxford: Oxford University Press), 134–157.

Grace, A. A. (2016). Dysregulation of the dopamine system in the pathophysiology of schizophrenia and depression. *Nat. Rev. Neurosci*. 17, 524–532. doi: 10.1038/nrn.2016.57

Grünblatt, E., Werling, A. M., Roth, A., Romanos, M., and Walitza, S. (2019). Association study and a systematic meta-analysis of the vntr polymorphism in the 3'-utr of dopamine transporter gene and attention-deficit hyperactivity disorder. *J. Neural Trans.* 126, 517–529. doi: 10.1007/s00702-019-01998-x

Hannestad, J., Gallezot, J.-D., Planeta-Wilson, B., Lin, S.-F., Williams, W. A., van Dyck, C. H., et al. (2010). Clinically relevant doses of methylphenidate significantly occupy norepinephrine transporters in humans *in vivo*. *Biol. Psychiatry* 68, 854–860. doi: 10.1016/j.biopsych.2010.06.017

Hauser, T. U., Fiore, V. G., Moutoussis, M., and Dolan, R. J. (2016). Computational psychiatry of adhd: neural gain impairments across marrian levels of analysis. *Trends Neurosci*. 39, 63–73. doi: 10.1016/j.tins.2015.12.009

Hauser, T. U., Iannaccone, R., Ball, J., Mathys, C., Brandeis, D., Walitza, S., et al. (2014). Role of the medial prefrontal cortex in impaired decision making in juvenile attention-deficit/hyperactivity disorder. *JAMA Psychiatry* 71, 1165–1173. doi: 10.1001/jamapsychiatry.2014.1093

Hervey, A. S., Epstein, J. N., Curry, J. F., Tonev, S., Eugene Arnold, L., Keith Conners, C., et al. (2006). Reaction time distribution analysis of neuropsychological performance in an adhd sample. *Child Neuropsychol*. 12, 125–140. doi: 10.1080/09297040500499081

Hille, B. (1992). G protein-coupled mechanisms and nervous signaling. *Neuron* 9, 187–195. doi: 10.1016/0896-6273(92)90158-A

Horn, A. S. (1990). Dopamine uptake: a review of progress in the last decade. *Progr. Neurobiol.* 34, 387–400. doi: 10.1016/0301-0082(90)90033-D

Huang-Pollock, C. L., Karalunas, S. L., Tam, H., and Moore, A. N. (2012). Evaluating vigilance deficits in adhd: a meta-analysis of cpt performance. *J. Abnorm Psychol.* 121, 360–371. doi: 10.1037/a0027205

Hunger, L., Kumar, A., and Schmidt, R. (2020). Abundance compensates kinetics: similar effect of dopamine signals on d1 and d2 receptor populations. *J. Neurosci.* 40, 2868–2881. doi: 10.1523/JNEUROSCI.1951-19.2019

Jackson, J. N. S., and MacKillop, J. (2016). Attention-deficit/hyperactivity disorder and monetary delay discounting: A meta-analysis of case-control studies. *Biol. Psychiatry* 1, 316–325. doi: 10.1016/j.bpsc.2016.01.007

John, C. E., Budygin, E. A., Mateo, Y., and Jones, S. R. (2006). Neurochemical characterization of the release and uptake of dopamine in ventral tegmental

area and serotonin in substantia nigra of the mouse. *J. Neurochem.* 96, 267–282. doi: 10.1111/j.1471-4159.2005.03557.x

Kang, P., Luo, L., Peng, X., and Wang Y. (2020). Association of val158met polymorphism in comt gene with attention-deficit hyperactive disorder: an updated meta-analysis. *Medicine* 99, e23400. doi: 10.1097/MD.0000000000023400

Kofler, M. J., Rapport, M. D., Sarver, D. E., Raiker, J. S., Orban, S. A., Friedman, L. M., et al. (2013). Reaction time variability in adhd: a meta-analytic review of 319 studies. *Clin. Psychol. Rev.* 33, 795–811. doi: 10.1016/j.cpr.2013.06.001

Lambacher, G., Pascale, E., Pucci, M., Mangiapelo, S., D'Addario, C., and Adriani, W. (2020). Search for an epigenetic biomarker in adhd diagnosis, based on the dat1 gene 5'-utr methylation: a new possible approach. *Psychiatry Res.* 291, 113154. doi: 10.1016/j.psychres.2020.113154

Leth-Steensen, C., Elbaz, Z. K., and Douglas, V. I. (2000). Mean response times, variability, and skew in the responding of adhd children: a response time distributional approach. *Acta Psychol.* 104, 167–190. doi: 10.1016/S0001-6918(00)00019-6

Li, D., Sham, P. C., Owen, M. J., and He, L. (2006). Meta-analysis shows significant association between dopamine system genes and attention deficit hyperactivity disorder (adhd). *Hum. Mol. Genet.* 15, 2276–2284. doi: 10.1093/hmg/ddl152

Lin, H.-Y., Hwang-Gu, S.-L., and Gau, S. S.-F. (2015). Intra-individual reaction time variability based on ex-gaussian distribution as a potential endophenotype for attention-deficit/hyperactivity disorder. *Acta Psychiatr. Scand.* 132, 39–50. doi: 10.1111/acps.12393

Luman, M., Janssen, T. W. P., Bink, M., van Mourik, R., Maras, A., and Oosterlaan, J. (2020). Probabilistic learning in children with attention-deficit/hyperactivity disorder. *J. Attent. Disord.* 25, 1407–1416. doi: 10.1177/1087054720905094

Madadi Asl, M., Vahabie, A. H., and Valizadeh, A. (2019). Dopaminergic modulation of synaptic plasticity, its role in neuropsychiatric disorders, and its computational modeling. *Basic Clin. Neurosci.* 10, 1–12. doi: 10.32598/bcn.9.10.125

Marinelli, M., and McCutcheon, J. E. (2014). Heterogeneity of dopamine neuron activity across traits and states. *Neuroscience* 282, 176–197. doi: 10.1016/j.neuroscience.2014.07.034

May, L. J., Kuhr, W. G., and Wightman, R. M. (1988). Differentiation of dopamine overflow and uptake processes in the extracellular fluid of the rat caudate nucleus with fast-scan in vivo voltammetry. *J. Neurochem.* 51, 1060–1069. doi: 10.1111/j.1471-4159.1988.tb03069.x

Nackley, A. G., Shabalina, S. A., Tchivileva, I. E., Satterfield, K., Korchynskyi, O., Makarov, S. S., et al. (2006). Human catechol-o-methyltransferase haplotypes modulate protein expression by altering mrna secondary structure. *Science* 314, 1930–1933. doi: 10.1126/science.1131262

Nakao, T., Radua, J., Rubia, K., and Mataix-Cols, D. (2011). Gray matter volume abnormalities in adhd: voxel-based meta-analysis exploring the effects of age and stimulant medication. *Am. J. Psychiatry* 168, 1154–1163. doi: 10.1176/appi.ajp.2011.11020281

Nicholson, C. (1995). Interaction between diffusion and michaelis-menten uptake of dopamine after iontophoresis in striatum. *Biophys. J.* 68, 1699–1715. doi: 10.1016/S0006-3495(95)80348-6

Nolan, K. A., Bilder, R. M., Lachman, H. M., and Volavka, J. (2004). Catechol o-methyltransferase val158met polymorphism in schizophrenia: differential effects of val and met alleles on cognitive stability and flexibility. *Am. J. Psychiatry* 161, 359–361. doi: 10.1176/appi.ajp.161.2.359

Norman, L. J., Carlisi, C., Lukito, S., Hart, H., Mataix-Cols, D., Radua, J., et al. (2016). Structural and functional brain abnormalities in attention-deficit/hyperactivity disorder and obsessive-compulsive disorder: a comparative meta-analysis. *JAMA Psychiatry* 73, 815–825. doi: 10.1001/jamapsychiatry.2016.0700

Oldehinkel, M., Beckmann, C. F., Pruim, R. H. R., van Oort, E. S. B., Franke, B., Hartman, C. A., et al. (2016). Attention-deficit/hyperactivity disorder symptoms coincide with altered striatal connectivity. *Biol. Psychiatry* 1, 353–363. doi: 10.1016/j.bpsc.2016.03.008

Patros, C. H. G., Alderson, R. M., Kasper, L. J., Tarle, S. J., Lea, S. E., and Hudec, K. L. (2016). Choice-impulsivity in children and adolescents with attention-deficit/hyperactivity disorder (adhd): a meta-analytic review. *Clin. Psychol. Rev.* 43, 162–174. doi: 10.1016/j.cpr.2015.11.001

Posner, J., Rauh, V., Gruber, A., Gat, I., Wang, Z., and Peterson, B. S. (2013). Dissociable attentional and affective circuits in medication-naïve children with attention-deficit/hyperactivity disorder. *Psychiatry Res.* 213, 24–30. doi: 10.1016/j.pscychresns.2013.01.004

Pothos, E. N., Davila, V., and Sulzer, D. (1998). Presynaptic recording of quanta from midbrain dopamine neurons and modulation of the quantal size. *J. Neurosci.* 18, 4106–4118. doi: 10.1523/JNEUROSCI.18-11-04106.1998

Redgrave, P., Gurney, K., and Reynolds, J. (2008). What is reinforced by phasic dopamine signals? *Brain Res. Rev.* 58, 322–339. doi: 10.1016/j.brainresrev.2007.10.007

Reynolds, J. N. J., and Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Netw.* 15, 507–521. doi: 10.1016/s0893-6080(02)00045-x

Rice, M. E., and Cragg, S. J. (2008). Dopamine spillover after quantal release: rethinking dopamine transmission in the nigrostriatal pathway. *Brain Res. Rev.* 58, 303–313. doi: 10.1016/j.brainresrev.2008.02.004

Robinson, B. G., Bunzow, J. R., Grimm, J. B., Lavis, L. D., Dudman, J. T., Brown, J., et al. (2017). Desensitized d2 autoreceptors are resistant to trafficking. *Sci. Rep.* 7, 4379. doi: 10.1038/s41598-017-04728-z

Rosa, E. C., Dickinson, D., Apud, J., Weinberger, D. R., and Elvevåg, B. (2010). Comt val158met polymorphism, cognitive stability and cognitive flexibility: an experimental examination. *Behav. Brain Funct.* 6, 53. doi: 10.1186/1744-9081-6-53

Saad, J. F., Griffiths, K. R., and Korgaonkar, M. S. (2020). A systematic review of imaging studies in the combined and inattentive subtypes of attention deficit hyperactivity disorder. *Front. Integr. Neurosci.* 14, 31. doi: 10.3389/fnint.2020.00031

Sagvolden, T., Johansen, E. B., Aase, H., and Russell, V. A. (2005). A dynamic developmental theory of attention-deficit/hyperactivity disorder (adhd) predominantly hyperactive/impulsive and combined subtypes. *Behav. Brain Sci.* 28, 397–419; discussion 419-68. doi: 10.1017/S0140525X05000075

Schönfuss, D., Reum, T., Olshausen, P., Fischer, T., and Morgenstern, R. (2001). Modelling constant potential amperometry for investigations of dopaminergic neurotransmission kinetics *in vivo*. *J. Neurosci. Methods* 112, 163–172. doi: 10.1016/S0165-0270(01)00465-4

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron* 36, 241–263. doi: 10.1016/S0896-6273(02)00967-4

Schultz, W. (2016). Dopamine reward prediction error coding. *Dial. Clin. Neurosci.* 18, 23–32. doi: 10.31887/DCNS.2016.18.1/wschultz

Schultz, W. (2017). Reward prediction error. *Curr. Biol.* 27, 369-R371. doi: 10.1016/j.cub.2017.02.064

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599. doi: 10.1126/science.275.5306.1593

Schwartz, S., and Correll, C. U. (2014). Efficacy and safety of atomoxetine in children and adolescents with attention-deficit/hyperactivity disorder: results from a comprehensive meta-analysis and metaregression. *J. Am. Acad. Child Adolesc. Psychiatry* 53, 174–187. doi: 10.1016/j.jaac.2013.11.005

Seidman, L. J., Valera, E. M., and Makris, N. (2005). Structural brain imaging of attention-deficit/hyperactivity disorder. *Biol. Psychiatry* 57, 1263–1272. doi: 10.1016/j.biopsych.2004.11.019

Stanislaw, H., and Todorov, N. (1999). Calculation of signal detection theory measures. *Behav. Res. Methods Instrument. Comput.* 31, 137–149. doi: 10.3758/BF03207704

Syková, E., and Nicholson, C. (2008). Diffusion in brain extracellular space. *Physiol. Rev.* 88, 1277–1340. doi: 10.1152/physrev.00027.2007

Tian, L., Jiang, T., Wang, Y., Zang, Y., He, Y., Liang, M., et al. (2006). Altered resting-state functional connectivity patterns of anterior cingulate cortex in adolescents with attention deficit hyperactivity disorder. *Neurosci. Lett.* 400, 39–43. doi: 10.1016/j.neulet.2006.02.022

Tien, Y.-M., Chen, V. C.-H., Lo, T.-S., Hsu, C.-F., Gossop, M., and Huang, K.-Y. (2019). Deficits in auditory sensory discrimination among children with attention-deficit/hyperactivity disorder. *Eur. Child Adolescent Psychiatry* 28, 645–653. doi: 10.1007/s00787-018-1228-7

Tonelli, E., Pascale, E., Troianiello, M., D'Addario, C., and Adriani, W. (2020). Dat1 gene methylation as an epigenetic biomarker in attention deficit hyperactivity disorder: a commentary. *Front. Genet.* 11, 444. doi: 10.3389/fgene.2020.00444

Tripp, G., and Wickens, J. R. (2008). Research review: dopamine transfer deficit: a neurobiological theory of altered reinforcement mechanisms in adhd. *J. Child Psychol. Psychiatry* 49, 691–704. doi: 10.1111/j.1469-7610.2007.01851.x

Ulke, C., Rullmann, M., Huang, J., Luthardt, J., Becker, G.-A., Patt, M., et al. (2019). Adult attention-deficit/hyperactivity disorder is associated with reduced norepinephrine transporter availability in right attention networks: a (s,s)-o-[, javax.xml.bind.jaxbelement@32a363f0, c]methylreboxetine positron emission tomography study. *Transl. Psychiatry* 9, 301. doi: 10.1038/s41398-019-0619-y

van der Kooij, M. A., and Glennon, J. C. (2007). Animal models concerning the role of dopamine in attention-deficit hyperactivity disorder. *Neurosci. Biobehav. Rev*. 31, 597–618. doi: 10.1016/j.neubiorev.2006.12.002

Véronneau-Veilleux, F., Robaey, P., Ursino, M., and Nekka, F. (2020). An integrative model of parkinson's disease treatment including levodopa pharmacokinetics, dopamine kinetics, basal ganglia neurotransmission and motor action throughout disease progression. *J. Pharmacokinet. Pharmacodyn*. 48, 133–148. doi: 10.1007/s10928-020-09723-y

Volkow, N. D., Wang, G.-J., Fowler, J. S., and Ding, Y.-S. (2005). Imaging the effects of methylphenidate on brain dopamine: new model on its therapeutic actions for attention-deficit/hyperactivity disorder. *Biol. Psychiatry* 57, 1410–1415. doi: 10.1016/j.biopsych.2004.11.006

Volkow, N. D., Wang, G. J., Fowler, J. S., Gatley, S. J., Logan, J., Ding, Y. S., et al. (1998). Dopamine transporter occupancies in the human brain induced by therapeutic doses of oral methylphenidate. *Am. J. Psychiatry* 155, 1325–1331. doi: 10.1176/ajp.155.10.1325

Volkow, N. D., Wang, G. J., Kollins, S. H., Wigal, T. L., Newcorn, J. H., Telang, F., et al. (2009). Evaluating dopamine reward pathway in adhd: clinical implications. *JAMA* 302, 1084–1091. doi: 10.1001/jama.2009.1308

Waelti, P., Dickinson, A., and Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature* 412, 43–48. doi: 10.1038/35083500

Wanat, M. J., Willuhn, I., Clark, J. J., and Phillips, P. E. M. (2009). Phasic dopamine release in appetitive behaviors and drug addiction. *Curr. Drug Abuse Rev*. 2, 195–213. doi: 10.2174/1874473710902020195

Wang, G.-J., Volkow, N. D., Wigal, T., Kollins, S. H., Newcorn, J. H., Telang, F., et al. (2013). Long-term stimulant treatment affects brain dopamine transporter level in patients with attention deficit hyperactive disorder. *PLoS ONE* 8, e63023. doi: 10.1371/journal.pone.0063023

Wickham, R. J., Solecki, W., Rathbun, L. R., Neugebauer, N. M., Wightman, R. M., and Addy, N. A. (2013). Advances in studying phasic dopamine signaling in brain reward mechanisms. *Front. Biosci*. 5, 678. doi: 10.2741/E678

Wood, A. C., Buitelaar, J., Rijsdijk, F., Asherson, P., and Kuntsi, J. (2010). Rethinking shared environment as a source of variance underlying attention-deficit/hyperactivity disorder symptoms: comment on burt (2009). *Psychol. Bull*. 136, 331–340. doi: 10.1037/a0019048

Xu, Y., Chen, X.-T., Luo, M., Tang, Y., Zhang, G., Wu, D., et al. (2015). Multiple epigenetic factors predict the attention deficit/hyperactivity disorder among the chinese han children. *J. Psychiatr. Res*. 64, 40–50. doi: 10.1016/j.jpsychires.2015.03.006

Zahniser, N. R., and Doolen, S. (2001). Chronic and acute regulation of na+/cl−-dependent neurotransmitter transporters: drugs, substrates, presynaptic receptors, and signaling systems. *Pharmacol. Therapeut*. 92, 21–55. doi: 10.1016/s0163-7258(01)00158-9

frontiers | Frontiers in Computational Neuroscience

TYPE Original Research
PUBLISHED 15 August 2022
DOI 10.3389/fncom.2022.933818

OPEN ACCESS

EDITED BY
Yu-Guo Yu,
Fudan University, China

REVIEWED BY
Minpeng Xu,
Tianjin University, China
Du Mengmeng,
Shaanxi University of Science and Technology, China

*CORRESPONDENCE
Xiuling Liu
liuxiuling121@hotmail.com
Licong Li
lilicong16@163.com

RECEIVED 01 May 2022
ACCEPTED 25 July 2022
PUBLISHED 15 August 2022

CITATION
Wei J, Li L, Song H, Du Z, Yang J, Zhang M and Liu X (2022) Response of a neuronal network computational model to infrared neural stimulation. *Front. Comput. Neurosci.* 16:933818. doi: 10.3389/fncom.2022.933818

Response of a neuronal network computational model to infrared neural stimulation

Jinzhao Wei[1,2], Licong Li[1,2]*, Hao Song[1,2], Zhaoning Du[1,2], Jianli Yang[1,2], Mingsha Zhang[3,4,5] and Xiuling Liu[1,2]*

[1]Key Laboratory of Digital Medical Engineering of Hebei, Hebei University, Baoding, China, [2]College of Electronic and Information Engineering, Hebei University, Baoding, China, [3]State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, China, [4]IDG/McGovern Institute for Brain Research at BNU, Beijing Normal University, Beijing, China, [5]Division of Psychology, Beijing Normal University, Beijing, China

Infrared neural stimulation (INS), as a novel form of neuromodulation, allows modulating the activity of nerve cells through thermally induced capacitive currents and thermal sensitivity ion channels. However, fundamental questions remain about the exact mechanism of INS and how the photothermal effect influences the neural response. Computational neural modeling can provide a powerful methodology for understanding the law of action of INS. We developed a temperature-dependent model of ion channels and membrane capacitance based on the photothermal effect to quantify the effect of INS on the direct response of individual neurons and neuronal networks. The neurons were connected through excitatory and inhibitory synapses and constituted a complex neuronal network model. Our results showed that a slight increase in temperature promoted the neuronal spikes and enhanced network activity, whereas the ultra-temperature inhibited neuronal activity. This biophysically based simulation illustrated the optical dose-dependent biphasic cell response with capacitive current as the core change condition. The computational model provided a new sight to elucidate mechanisms and inform parameter selection of INS.

KEYWORDS

computational model, infrared neural stimulation, neuronal network, photothermal effect, ionic channel, membrane capacitance

Introduction

Many forms of external physical stimulation (electric, optical, ultrasound, and magnetic stimulation) can regulate brain functions and the treatment of brain disorders (Darmani et al., 2022). Compared with electrical stimulation, known as the gold standard (Barborica et al., 2022), optical stimulation techniques have an extremely high value in the neuromodulation field due to their high spatial accuracy and positional targeting. Among the optical stimulation techniques, the ability of infrared neural stimulation (INS) to activate or inhibit nerve cells without any genetic or chemical tissue modification provides better safety and clinical feasibility when compared with the other types of optical techniques (Rajguru et al., 2011). This form of neuromodulation has potential

applications in diagnosing and treating many neurological and psychiatric disorders, such as dementia (Iaccarino et al., 2016), Parkinson's disease (Darlot et al., 2016), and depression (Tanaka et al., 2011). Nevertheless, the rational design and optimization of INS are hampered by the limited understanding of its neural effects.

Understanding how infrared (IR) stimulation affects neuronal activity and the mechanisms of interaction between the influences generated by IR light and neural tissue is necessary to address the question regarding the interaction between INS and neurons. A growing body of *in vitro* and *in vivo* evidence strongly suggests that laser is mediated by absorption of the local aqueous medium surrounding the heated cell to produce a thermal transient (Liu et al., 2009). INS regulates neuronal membrane capacitance and ion channel conductivity through the temperature-dependent mechanism generated by this thermal transient effect (Shapiro et al., 2012; Singh et al., 2019) and then modulates neuronal excitability. It is important to investigate the effects of physical field modulation on the neural network by considering the complexities of neuron types and their connections. However, only a few studies on the efficiency of INS at the neural network level were reported (Xia and Nyberg, 2019). Hence, the mechanism of interaction between the thermal effects produced by IR neural stimulation and the activity of neuronal populations is still not clearly elucidated, especially how the changes at one cell level can affect the network. Indeed, the network effect of neuromodulation has been shown to exist in other physical stimulations through experiments and computational models (Miyawaki et al., 2012; Di Lazzaro et al., 2018).

Computational modeling is a powerful tool for investigating the mechanisms of INS and for helping bridge research scales from a single cell to the network. A wealth of theoretical and numerical models on the interaction of INS with neural tissue exists. Most of them used spiral ganglion neurons (SGNs) as potential stimulation targets to explore the effects of IR stimulation on the level of isolated individual neurons. For example, acute *in vivo* experiments using gerbils to record optically evoked compound action potentials in the cochlea demonstrated that the auditory nerve could be stimulated by optical radiation (Littlefield et al., 2010). Some researchers accurately simulated neuronal responses by building a modified Hodgkin-Huxley (HH)-type model to predict the action potential threshold generated by SGN stimulation (Brown et al., 2021). Optical stimulation techniques can significantly improve cochlear implants hampered by a lack of spatial selectivity (Richardson et al., 2020). The above results showed that most studies were performed at the level of individual neuronal cells and did not address the dynamic activity of neuronal networks exposed to IR light. Therefore, considering the specific effects of photothermal effects on the complex neuronal network and illustrating the interaction between the photothermal effect and the neuronal network through the simulation results of the computational model are necessary.

In this study, we pursued a mixed strategy and developed a cortical neuronal network model by lumping both microscopic and macroscopic aspects to quantify the process of neuronal network response to IR light stimulation. The model combined excitatory and inhibitory neurons and synaptic structures, all of which were essential to accurately model the effects of IR neural stimulation. The present study aimed to investigate whether laser irradiation could regulate network activity. With this more complete model, we illustrated that the thermal effect in optical modulation affected the activity in individual neurons, as well as neuronal networks in a biphasic dose-response manner, thus providing a reasonable reference for biological experiments.

Materials and methods

Based on neurophysiological features and experimental observations, the typical neuronal network model includes excitatory and inhibitory neurons, which are connected by excitatory and inhibitory synapses, respectively, forming a feedback circuit (Ocker et al., 2015). We focused on two levels, ion channels and membrane capacitance, and extended to the network structure to investigate the process of IR regulation on neurons. The model construction and its dynamics analysis from individual neurons to the neuronal network were as follows.

Neuron model

As the basic element of a neuronal network, the neuron plays a fundamental role in modeling. Neuron modeling is the primary step in developing neuronal networks. The well-known Hodgkin-Huxley (HH) model was used in our neural network (Hodgkin and Huxley, 1952). The corresponding dynamic equation is as follows:

$$C_m \frac{dv_m}{dt} = -\bar{g}_{leak}(v_m - E_{leak}) - \bar{g}_{Na} m^3 h(v_m - E_{Na}) - \bar{g}_K n^4 (v_m - E_K) + I_{ext} \quad (1)$$

where C_m is the membrane capacitance, v_m is the membrane potential, $\bar{g}_{leak}$, $\bar{g}_{Na}$, and $\bar{g}_K$ are maximal conductance of the leak, sodium, and potassium channels, respectively. E_{leak}, E_{Na}, and E_K are the reversal potentials, and I_{ext} is the external current injected into the membrane, i.e., background current.

The environment to which the neuronal cells are exposed generates temperature changes according to the rapid thermal transients generated by IR radiation in biological tissues. Therefore, a modified model of HH neurons was proposed through this temperature-dependent process. The improved

model could visualize the kinetics process of neurons under the photothermal effect.

We extended the temperature influence factor $\phi(T)$ (Chandler and Meves, 1970), which affected neuronal activity by modulating the conductance and gating kinetics of ion channels. Thus, the firing process and dynamic changes of neurons under the premise of the thermal effect can be effectively simulated. Additionally, neuronal excitability is acutely affected by temperature through the changes in Nernst equilibrium potential (Kim and Connors, 2012). According to the original hypothesis of Hodgkin and Huxley, the activation m, n and inactivation h gating variables could be combined with the temperature coefficient $\phi(T)$, thus introducing temperature variables into the opening and closing rates of ion channels. Thus, the model of modulation of neuronal ion channels by photothermal effects is described by the following equations:

$$\alpha_n = \frac{0.032\phi(T)5}{\exp[(-48 - v_m)/5]} \tag{2}$$

$$\beta_n = 0.5\phi(T)\exp[(-53 - v_m)/40] \tag{3}$$

$$\alpha_m = \frac{0.32\phi(T)4}{\exp[(-50 - v_m)/4]} \tag{4}$$

$$\beta_m = \frac{0.28\phi(T)5}{\exp[(-103 - v_m)/5]} \tag{5}$$

$$\beta_h = \frac{4\phi(T)}{1 + \exp[(-23 - v_m)/5]} \tag{6}$$

$$\phi(T) = 3^{(T-6.3)/10} \tag{7}$$

where α_n and β_n are the opening and closing rates of the K^+ channel, $\alpha_m$and β_m are the opening and closing rates for the activation gates of the Na^+ channel, and α_h and β_h are the opening and closing rates for the inactivation gates of the Na^+ channel, respectively.

The IR radiation not only thermally modulates the ion channel but also produces a temperature-dependent effect on the membrane capacitance. Early experimental studies demonstrated a correlation between capacitance (C_m) and temperature (T) (Santos-Sacchi and Huang, 1998). Based on the ferroelectric Curie-Weiss law, the temperature-dependent effect of membrane capacitance could be represented visually and the temperature-capacitance relationship could be effectively fitted experimentally (Leuchtag, 1995). The fitting equation is described as follows:

$$C_m = C_0 + \frac{k}{T_c - T} \tag{8}$$

where k is capacitance constant; C_0 is a constant membrane capacitance; and T_c is the Curie temperature of membrane capacitance. The Curie temperature of the membrane capacitor varied depending on the type of squid. Therefore, based on the data obtained from the HH model, the Curie temperature range was 31–50°C.

In addition, the photothermal effect also affects the size of the lipid bilayer, which in turn leads to changes in the membrane capacitance of the neurons. In conventional models, the capacitance would be assumed to be constant. However, recent studies demonstrated that under the condition of IR radiation, the change in capacitance caused a part of displacement current, with temperature dependence (Peterson and Tyler, 2012). As a result, we introduced the capacitive current component (Brown et al., 2021), which could be expressed as the time derivative of the membrane capacitance charge $C_m(v_m - V_s)$:

$$I_m = (v_m - V_s)\frac{dC_m}{dt} \tag{9}$$

where dC_m/dt denotes the laser-induced dT/dt as a function of the relational gradient dC_m/dT. V_s is the asymmetric surface charge potential. The capacitive current component was well-fitted to the equation .

For the heat transfer effect of continuous wave laser, the increased temperature varied for different wavelengths, but a similar trend occurred in the case of the temperature change rate. As the irradiation time increased, the temperature gradient decreased significantly with respect to the initial value. Therefore, under the specified laser pulse conditions, dC_m/dt was linearly proportional to dT/dt with a temperature-dependent capacitance factor $dT/dt = 0.313\%^{\circ}C^{-1}$ (Plaksin et al., 2018). Thus, the capacitor current equation is read as follows:

$$I_m = 3.13 \times 10^{-3}\frac{dT}{dt}(v_m - V_s) \tag{10}$$

The schematic illustration of IR regulation on ion channels and membrane capacitance is shown in Figure 1. The parameters (Hodgkin and Huxley, 1952) used in the neuronal model are listed in Table 1.

Synapse model

In neurophysiology, synapses are the sites where connections between neurons occur functionally and are also the key players in constituting models of complex neuronal networks. We used the synapse model originally proposed by Tsodyks and Markram to describe the dynamics of the synaptic terminal (Tsodyks et al., 1998; Barak and Tsodyks, 2007). The synaptic release process was achieved by the product of the variables u_s and x_s, in which u_s represents the fraction of available neurotransmitter resources "docked" for release, and x_s is related to the proportion of total neurotransmitters that could be released. Upon the arrival of an action potential, u_s decayed to 0 at the $1/\tau_f$ rate while x_s reinstated to 1 at the $1/\tau_r$ rate. The process mimicked neurotransmitter

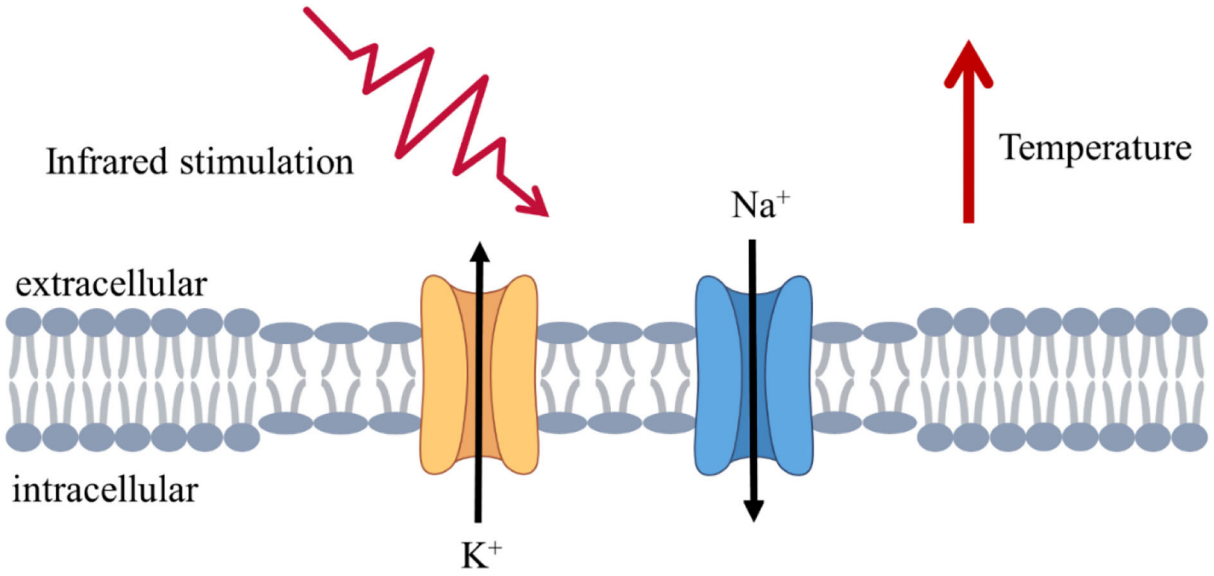


FIGURE 1
Schematic illustration of infrared stimulation at the cellular and molecular levels. Infrared irradiation caused changes in ion channel and membrane capacitance, increases in the rate of ion channel opening and closing, and changes in the thinning and larger area of the lipid bilayer, leading to changes in the response current.

TABLE 1 Parameters used in the neuron model.

Parameter	Description	Value
g_{leak}	Leak channel conductance	0.05 mS
g_{Na}	Sodium channel conductance	50 mS
g_K	Potassium channel conductance	30 mS
E_{Leak}	Reversal potential of leakage channel	−60 mV
E_{Na}	Reversal potential of Na^+	90 mV
E_K	Reversal potential of K^+	−85 mV
V_s	Asymmetric surface charge potential	28 mV
V_{th}	Firing threshold	−63 mV
V_{rest}	Resting potential	−70 mV
I_{ext}	External current	1 pA
k	Capacitance constant	2.2
C_0	Constant membrane capacitance	0.824 μF

depletion and reintegration and can be read by the following set of equations:

$$\frac{du_s}{dt} = \frac{-u_s}{\tau_f} + U_0 \cdot (1 - u_s) \cdot \delta(t - t_K) \tag{11}$$

$$\frac{dx_s}{dt} = \frac{1 - x_s}{\tau_r} - r_s \cdot \delta(t - t_K) \tag{12}$$

where U_0 is initial synaptic release probability at rest; released neurotransmitter resources from the presynaptic terminal can be calculated as follows:

$$r_s = u_s \cdot x_s \tag{13}$$

Then, the neurotransmitter concentration G_s in the synaptic cleft is given by De Pitta and Brunel (2016):

$$\frac{dG_s}{dt} = -\Omega_c \cdot G_s + r_s \cdot Q_c \cdot Y_T \cdot \delta(t - t_K) \tag{14}$$

where in neurotransmitter clearance rate, Q_c is vesicular vs. mixing volume ratio and Y_T is total vesicular neurotransmitter concentration. When a presynaptic action potential occurred, the postsynaptic neuron was responded by increasing corresponding excitability or inhibition conductance and then gave rise to postsynaptic currents. The fraction of postsynaptic receptors in the open state r can be described by the following first-order dynamic equation:

$$\frac{dr}{dt} = \alpha \cdot G_s \cdot (1 - r) - \beta \cdot r \tag{15}$$

where α and β are the forward and backward rate constants, respectively. Finally, α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA)- and N-methyl D-aspartate (NMDA)-mediated excitatory postsynaptic currents (EPSCs) are expressed by the following equations:

TABLE 2 Parameters used in the synapse model.

Parameter	Description	Value
U_0	Resting synaptic release probability	0.6
τ_f	Facilitation time constant	0.3 s^{-1}
τ_r	Recovery time constant	0.5 s^{-1}
Y_T	Total vesicular neurotransmitter concentration	500 mM
Ω_c	Neurotransmitter clearance rate	40 s^{-1}
Q_c	Vesicular vs. mixing volume ratio	0.005
α_{AMPA}	AMPA forward rate constant	1.1 $\mu M^{-1} \cdot s^{-1}$
β_{AMPA}	AMPA backward rate constant	190 s^{-1}
α_{NMDA}	NMDA forward rate constant	0.072 $\mu M^{-1} \cdot s^{-1}$
β_{NMDA}	NMDA backward rate constant	6.6 s^{-1}

$$I_{AMPA} = \overline{g}_{AMPA} \cdot r(t) \cdot (v_m - E_{AMPA}) \quad (16)$$

$$I_{NMDA} = \overline{g}_{NMDA} \cdot Mg(v_m) \cdot r(t) \cdot (v_m - E_{NMDA}) \quad (17)$$

$$Mg(v_m) = \frac{1}{1 + \exp(-0.062 * v_m)[Mg^{2+}]/3.57} \quad (18)$$

where $\overline{g}$ is the maximum synaptic conductance with $\overline{g}_{AMPA}$ = 0.35 nS, $\overline{g}_{NMDA}$ = 0.026 mS. E is the synaptic reversal potential with $E_{AMPA} = E_{NMDA} = 0$ mV. Noteworthily, NMDA receptor channels contain a voltage-dependent term representing magnesium (Mg^{2+}) block with $[Mg^{2+}] = 1$ mM (Jahr and Stevens, 1990). The parameters (De Pitta and Brunel, 2016) used in the synapse model are listed in Table 2.

Neuronal network

The cerebral cortex is a multi-scale structure with local circuits interwoven to form a global network of remote connections. Within this complex network structure, neural activity propagates widely across temporal and spatial scales. The network model constructed in this study started from the microscale and took excitatory and inhibitory neurons as the basic components to respond to the photothermal effect of INS through synaptic interactions. Based on the neuroanatomical ratio of excitatory to inhibitory neurons (4:1) (Manos et al., 2021), the network model we designed comprised 3,200 excitatory neurons and 800 inhibitory neurons. The excitatory neurons with 5% of the connected weight enhanced signals, and the inhibitory neurons with 20% of the connected weight transmitted suppression signals.

Further, the cell populations were distributed in the Euclidean space to visualize and analyze the neuronal network. In the 2D map of the network, red represents excitatory neurons and blue represents inhibitory neurons, as shown in Figure 2. In cortical neuronal networks, excitatory inputs and inhibitory equivalents entered the cell together, allowing targeted transient or sustained opening of signal receptors. This tight coupling of excitatory and inhibitory signals exhibited a more intuitive state of network equilibrium (Jirsa, 2004). The model we developed was implemented in the Brian 2.0 simulator (Goodman and Brette, 2008; Stimberg et al., 2017).

Results

IR neural stimulation-induced temperature rise

The IR radiation is absorbed by the cellular tissue and converted into thermal energy (Wells et al., 2007a; Thompson et al., 2013). The temperature of the stimulation target increases, leading to temperature-dependent neuronal stimulation. The focus of this study was to investigate the spike activity and coding processes of neurons and neuronal networks in terms of the thermal effects generated by the action of IR light. Since the majority of the material in biological tissues is water, we first considered the process of temperature change produced by the irradiation of IR light in water. The data obtained from the experiments provided support for the simulation of neuronal networks.

The schematic illustration of a laser irradiation detection device is shown in Figure 3A. Phosphate-buffered solution (PBS; Solarbio, China) of 0.5 ml was irradiated using IR laser 1,550 nm (Changchun New Industries Optoelectronics, China) in 24-well plates (Corning, USA) at room temperature. The optical stimulation was performed at the bottom of each well with a temperature probe (Fluke 17B+, USA) 10 mm away from the well. The power of the laser over the beam region was monitored by a Thorlabs (Thorlabs PM100D, USA). The temperature variation induced by laser irradiation is shown in Figure 3B. The increased temperature distribution ranged from 0.9 to 30°C at the different laser powers, in which the trends showed a rapid increase and gradual stabilization of temperatures (Xia and Nyberg, 2019). Consequently, the temperature increases in the INS computational model were primarily 10, 20, and 30°C.

Spiking rhythms exposed to IR neural stimulation

The equations were inserted in editable format from the equation editor. We described the firing behavior of the neurons to verify whether IR-induced temperature changes could evoke neuron depolarization. A constant offset current with an amplitude of I_{ext} = 1 pA was injected into the neuronal model to induce tonic spikes in the neuronal action potential. We used neuronal spikes without photothermal effects as the original reference and the variation of the neuronal

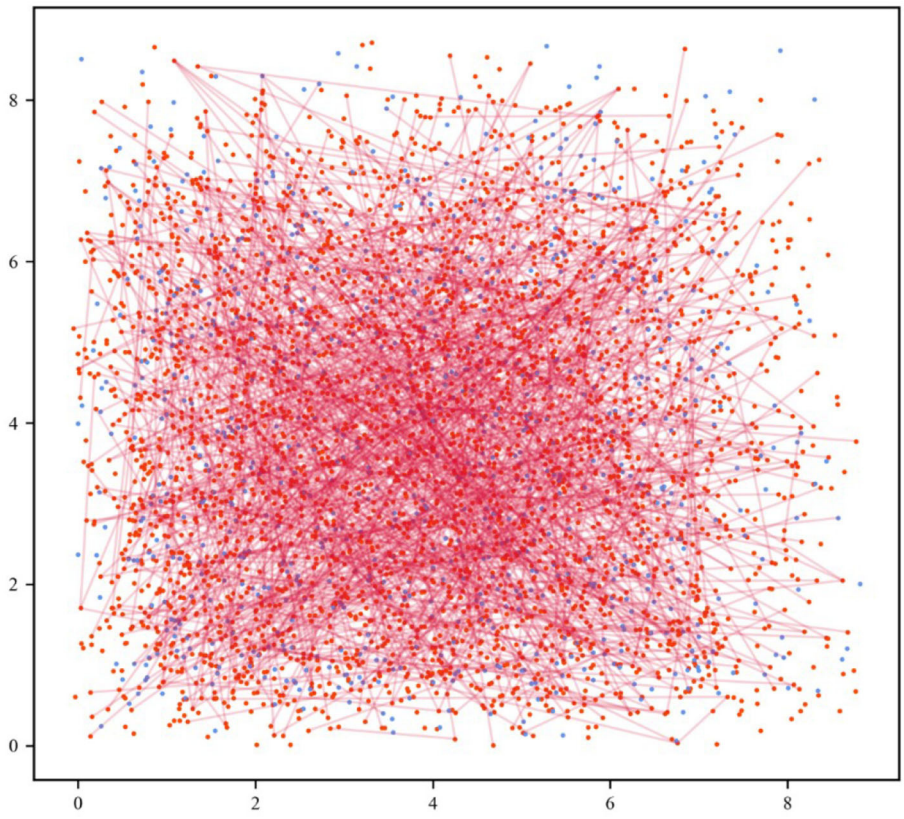


FIGURE 2
Neuronal network visualization in the Euclidean space. The 2D network diagram shows excitatory neurons in the red dots and inhibitory neurons in the blue dots. For the sake of clarity, the connection of 5% out of all excitatory synapses was selected at random (red lines).

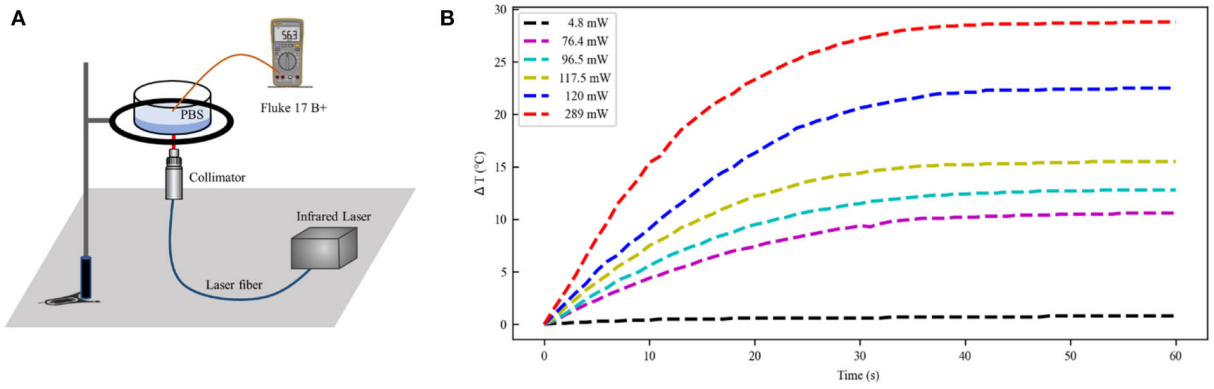


FIGURE 3
Temperature distribution at the different laser power. **(A)** The experimental setup of laser irradiation detection. **(B)** Temperature change curve over time caused by infrared light irradiation at 1,550 nm (initial temperature 21°C). PBS, phosphate buffered solution.

membrane potential in the experimentally probed temperature range was characterized.

The thermal effect produced by IR light is interfered with the spike timing of neurons, as shown in Figure 4. The numbers of neuronal spikes from the recording time were 22, 52, 34, and 9 with the temperature increase of 0, 10, 20, and 30°C, respectively. Compared with no change in temperature, the neuronal spikes at 10 and 20°C were increased by 136.4

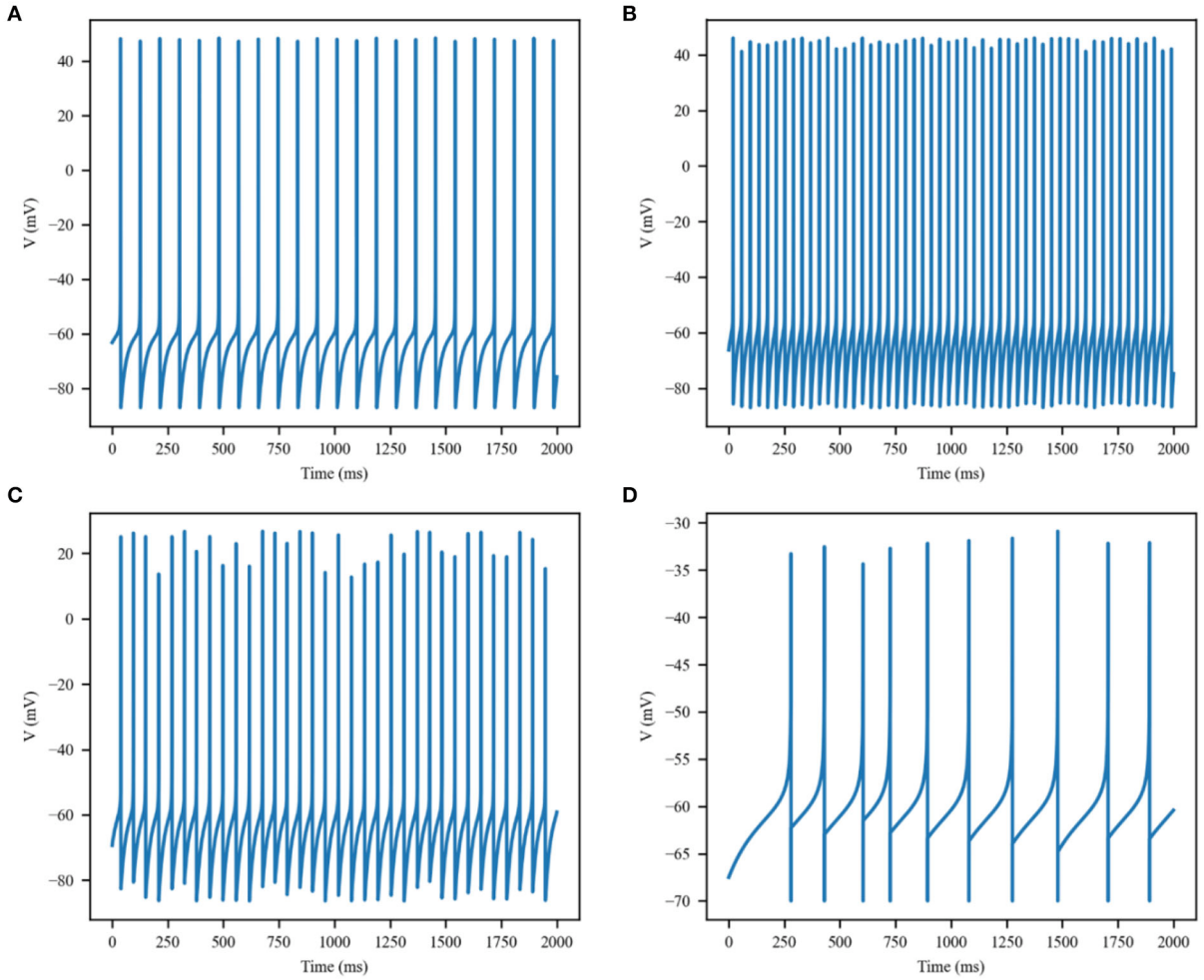


FIGURE 4
Membrane potential of single excitatory neurons under the action of increase in temperature by **(A)** 0, **(B)** 10, **(C)** 20, and **(D)** 30°C.

and 45.4%, respectively, and the spike count at 30°C was decreased by 59.1%. These results revealed an optical dose-dependent biphasic cell response. Figure 5 shows the inter-spike intervals (ISIs) changes in neuronal spiking trains evoked by different temperature conditions. ISIs were equally distributed with approximately the same value (84.5 ms) in the absence of optical stimulation. The variations in ISIs were related to the changes in neuronal spike time. With an increase in temperature ($0°C < \Delta T < 20°C$), the lower ISIs values represented a high neuronal spike count and the data presented irregularity. With increasing temperature ($20°C < \Delta T < 30°C$), the higher value of ISIs referred to sparse firing of neurons, indicating that the neuronal activity was inhibited, which was in tune with the results shown in Figure 4D. The large range of ISIs showed irregular neuron firing. Furthermore, we calculated and analyzed the Coefficient of Variation (CV) of neurons under different temperature changes, as shown in Figure 6. The results visually displayed the increasing trend of CV value with the increased temperature. In this process, the irregularity of interspike time became larger, and the neuronal activity became more active. When the temperature continued to rise, the CV value began to decrease, and the neuron activity decreased. Overall, the result of CV is consistent with the change of spike rate and shape of the single neuron during treatment with increased temperature. Except that the time course of an action potential characterized by ISIs and CV was affected, the result also shows the decrease of amplitudes of action potentials changing with temperature increase (Hodgkin and Katz, 1949).

Overall, these results indicated that the temperature changes of different intensities powerfully influenced neuronal spiking rhythms by the capacitive current and voltage-gated ion channels, and this effect was increased with increasing stimulus intensity.

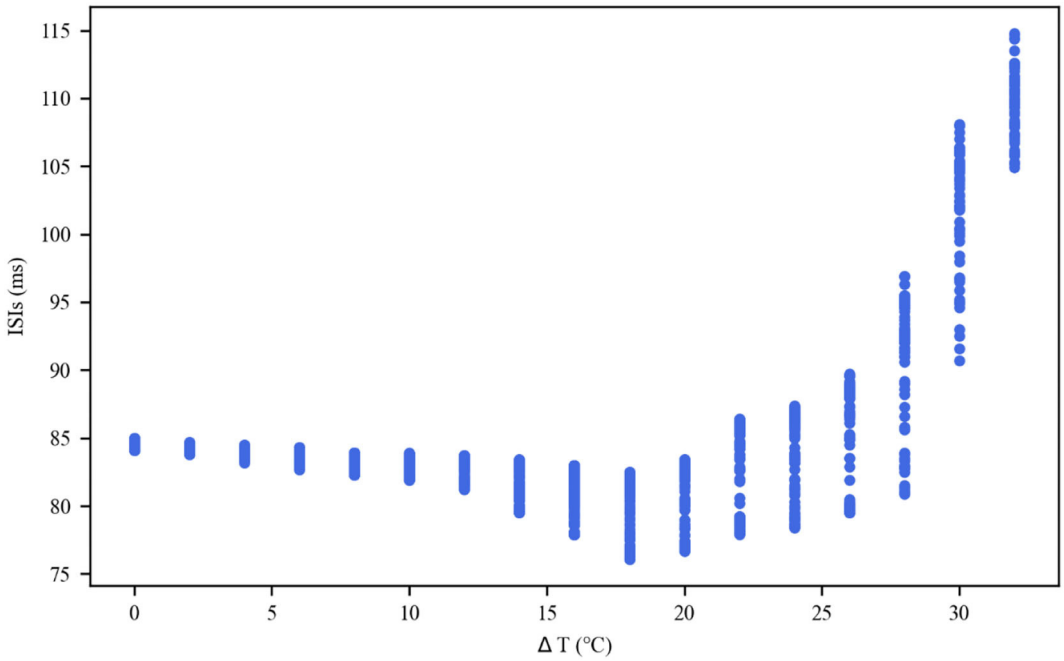


FIGURE 5
Inter-spike interval (ISI) sequences of neuronal spiking trains exposed to different temperatures.

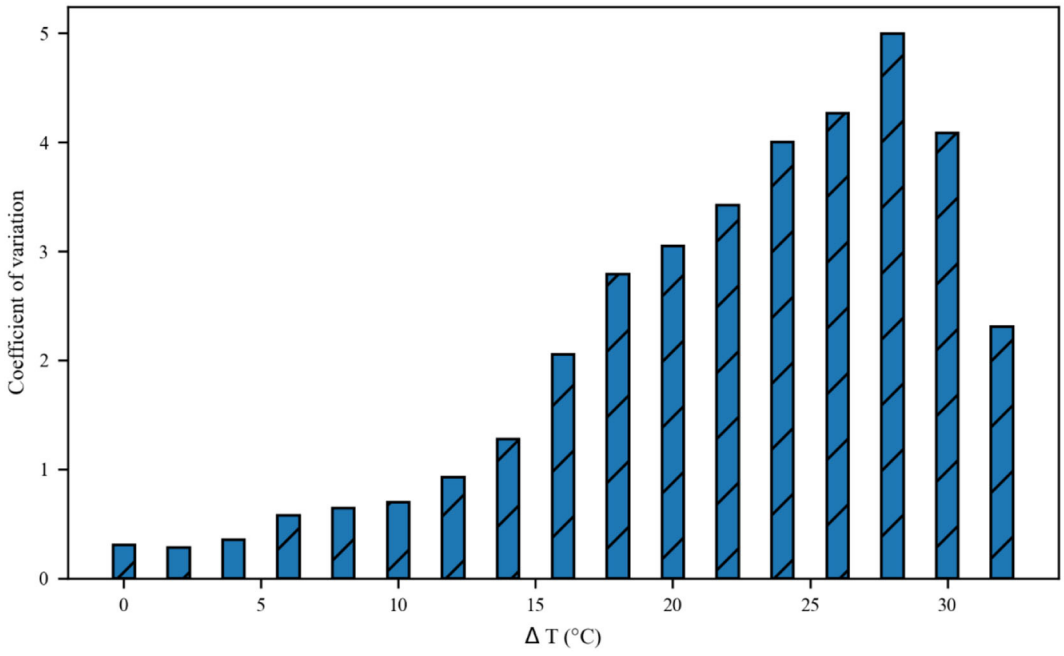


FIGURE 6
The Coefficient of Variance (CV) value of neurons inter-spike intervals exposed to different temperatures.

Modulation of presynaptic release and postsynaptic currents with IR neural stimulation

The aforementioned analysis revealed the impact of IR neural stimulation on neuronal activity. The neurotransmitters released from presynaptic terminals were changed with neuronal spike activity. Thus, we investigated the synaptic transmission response to the photothermal effects of IR neural stimulation. As previously reported in the literature, IR stimulation has the potential ability to modulate glutamate release by stimulating glutamatergic nerve endings (Amaroli et al., 2018). This scenario is illustrated in Figure 7, depicting the response of the synaptic model to a series of action potential changes induced by the photothermal influence. Increased temperature (10 and 20°C) robustly enhanced excitatory presynaptic release probability (Figure 7, blue and green lines). A slight temperature increase generated by IR laser light could stimulate vesicular neurotransmitter release. In the presence of a higher temperature at 30°C, the scenario was reversed (Figure 7, orange line), that is, photoinduced hyperthermia inhibited excitatory presynaptic release. Figure 8 shows that different temperatures affect the EPSCs under photothermal action with a running time of 10 s. At different temperatures, the discrepancy in postsynaptic activity was observed, which was in line with the dose of photothermal effect and the quantal size variability in presynaptic neurotransmitter release (Figure 7). The increased temperature induced by INS affected synaptic activity to a large extent when compared with the absence of photothermal stimulation. It could increase or decrease the frequency of neuronal spikes and affect synaptic efficacy and neural information processing.

Impacts of IR neural stimulation exposure on a neuronal network

The connectivity of excitatory and inhibitory neurons as the basic ingredients is specified by synapses, which ultimately make up the interaction and co-regulation of a complex network structure. Akin to simulate the network structure in cortical neurons, the network is capable of displaying complex dynamics analysis behaviors.

The simulation of the neuronal network in Figure 9 shows a raster plot of the firing activity of 25% of the excitatory neurons (red) and inhibitory neurons (blue) in the network and in response to temporarily increasing external stimuli with 10, 20, and 30°C (rectangular stimulus change in the top panel). Prior to stimulus onset ($t < 3$ s, increased 0°C), the neuronal ambient temperature was in a moderate situation. Therefore, the model was in a state of network equilibrium that included a network-averaged firing rate (bottom panel). For $3 < t < 6$ s and $9 < t < 12$ s (increasing to 10 and 20°C, respectively), all neurons were affected with increased temperature. The neuronal activity was significantly enhanced, as reflected by a denser raster plot and high-frequency population activity during this period. The external stimulus returned to its original value (at $t = 6$ and 12 s), and the neuronal firing returned to normal accordingly. With a temperature rise to 30°C ($15 < t < 18$ s), the presented network raster plot and consequently the dynamic characteristics of the total firing rate were observed to show low-frequency population activity. The increased temperature and excitatory and inhibitory spike counts are shown in Figure 10. Thus, it was inferred that small temperature increases enhanced neuronal network activity, whereas higher temperatures inhibited the neuronal spike activity of the neuronal network. These results matched well with previous experimental observations (Xia and Nyberg, 2019), that is, beneficial at a low dose and harmful at a high dose.

Discussion and conclusion

The potential utility of IR stimulation has been demonstrated by numerous experimental research studies. However, the lack of understanding of its underlying mechanisms has hindered its scientific and clinical applications. Based on neurophysiological findings, we designed a biophysical neuronal network model to mimic the interaction between the photothermal effect of IR neural stimulation and neurons. The simulation results of our study provided new insights to explore the response of neurons to optical stimulation.

Optical modulation can stimulate discrete groups of nerve fibers in a contact-free, damage-free, and artifact-free way. Regardless of the application, the interaction between the laser and the biological tissue results in light distribution and absorption, leading to photobiological effects. The generation of the photothermal effect is related to the transient irradiation process by IR light, which is a temperature dependent and transient mechanism. Our experiments are consistent with this conclusion by measuring the temperature increase of the PBS caused by the laser. After laser irradiation, the temperature increases exponentially, which reflects not only the temperature change but also the continuous transformation of temperature change rate with laser irradiation time. Combined with research and experiments (Ebtehaj et al., 2018; Ganguly et al., 2019), the thermo-induced capacitive current and the modulation of thermal-sensitive ion channels were modified in response to IR stimulation. The results suggested that IR light-induced thermal effects could regulate neuronal spikes, displaying the characteristics of optical dose dependence, that is, low-level laser enhances neuronal activity and high-level laser inhibits neuronal activity. These findings were akin to the expected optical dose-dependent biphasic cell response (Huang et al., 2009). Though the excitability of neurons depends on synaptic connections,

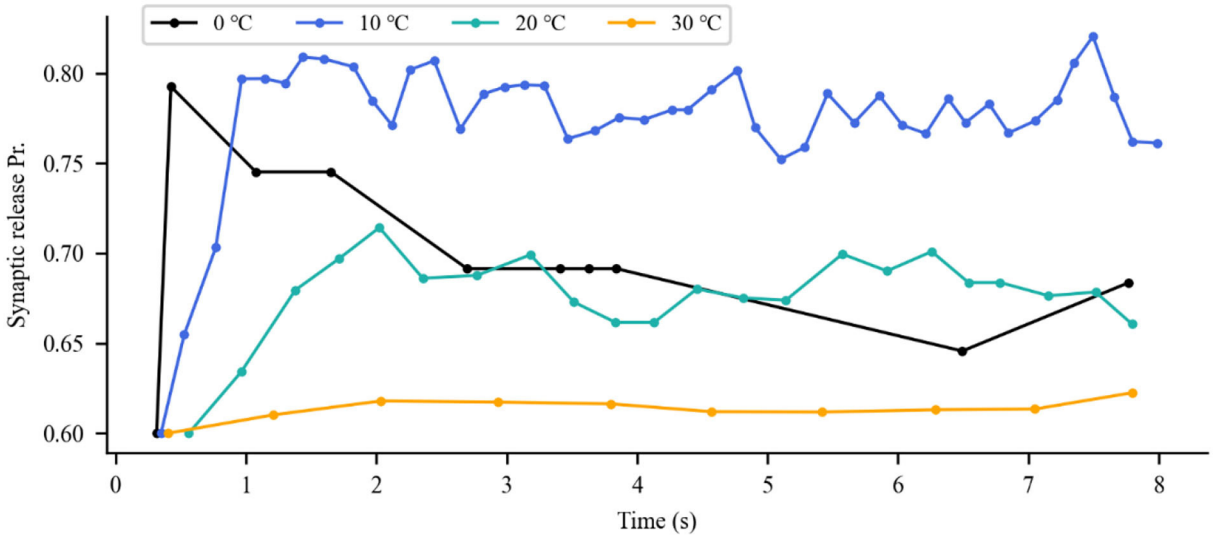


FIGURE 7
Variations in presynaptic glutamate release probability (Pr) in response to different temperatures. The dots represent each presynaptic release event.

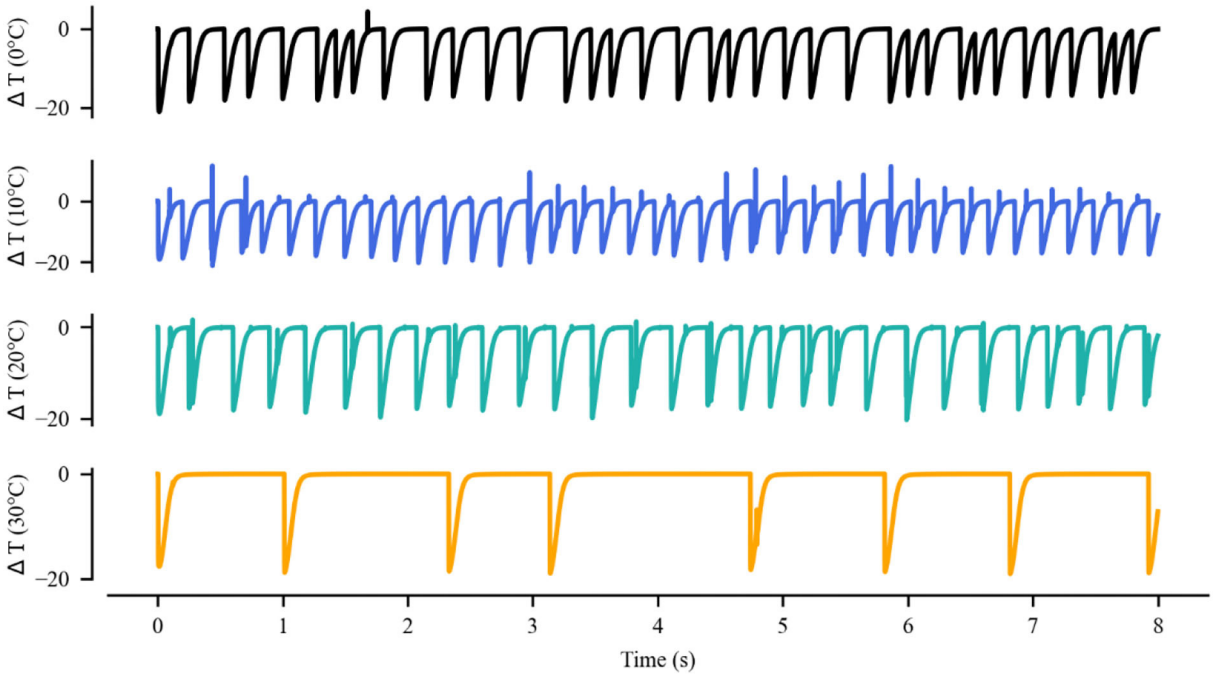


FIGURE 8
Variations in excitatory postsynaptic currents (EPSCs) evoked by glutamate released from the presynaptic terminal during treatment with different temperatures.

our study here focused on individual neurons, suggesting that the laser can activate or inhibit the activity of neurons even without synaptic interactions or neuronal network properties. The result proves that the INS mechanism is mediated by temperature transients induced by IR absorption, and neural activation with laser light results from the thermal transient. In fact, mounting evidence indicates that many types of mild stresses, such as hyperthermia, hypothermia, and an altered pH,

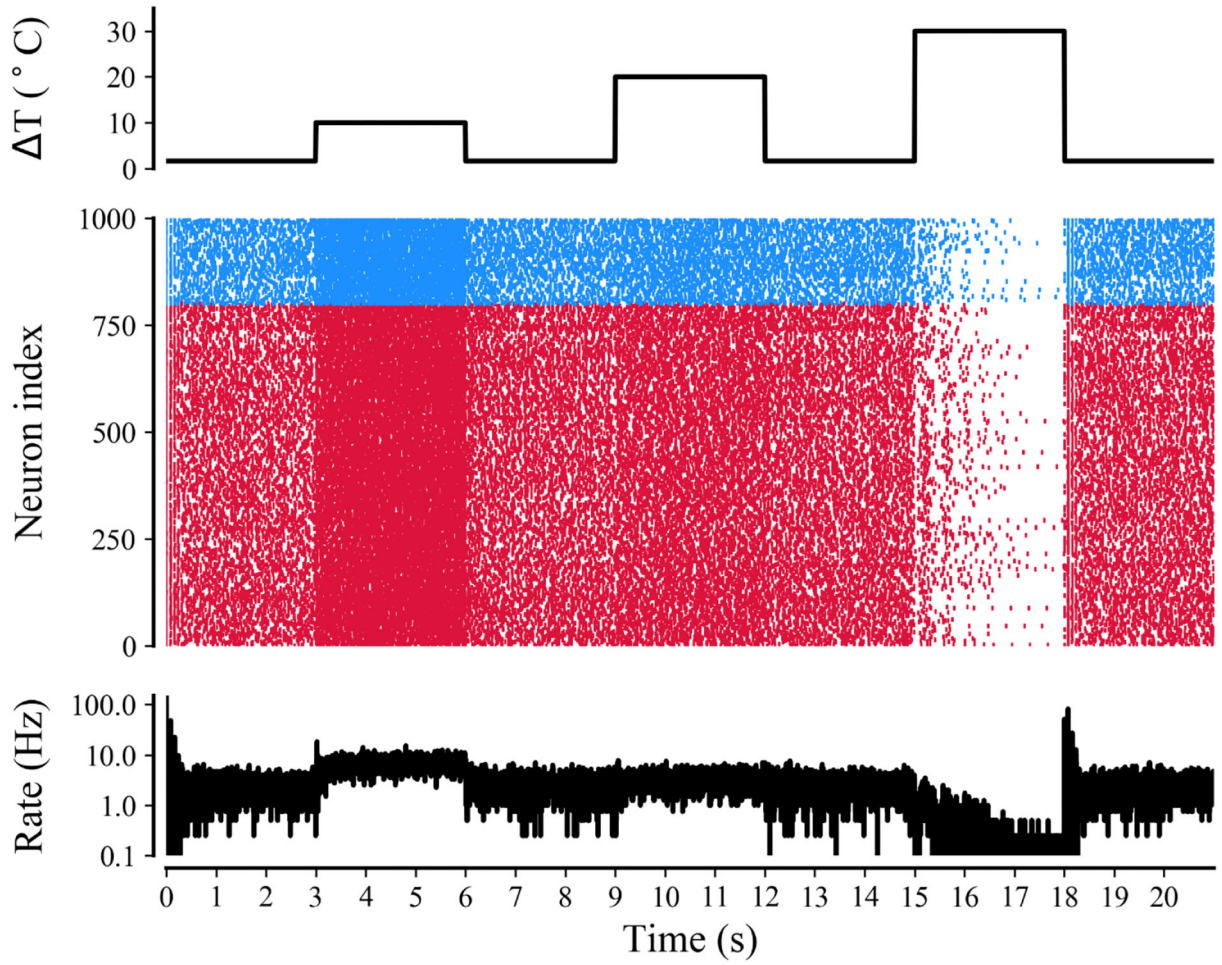


FIGURE 9
Raster plot and mean firing rate of neural activity during treatment with different temperatures. Simulations of neuronal network for a rectangular-pulse increase in temperature by 10, 20, and 30°C (top panel). The raster plot (middle panel) shows the spike activity of 25% of all excitatory (red) and inhibitory neurons (blue) of the network. The mean firing rate of the network is shown in the bottom panel.

can directly or indirectly interfere with protein functions and signaling pathways in cells and then affect cell activity (Chen and Chiao, 2020). It is worth noting that the temperature change caused by the photothermal effect not only has an effect on voltage gating but also affects the reversal potential of sodium and potassium current based on the application of the Nernst equation (Yu et al., 2012).

A further interesting prediction of this model was the modulation of synaptic and network activity by IR neural stimulation. The results of the present study showed that moderately increased temperatures indeed enhanced neurotransmitter release probability and neuronal network activity in an optical dose-dependent manner. Irradiation-induced photothermal effect somehow stimulated the release of neurotransmitters at the synaptic level, thus facilitating the transmission of neural information. The efficiency of exocytosis is higher at a slightly increased temperature than without stimulation, while laser above a certain energy threshold reduced synaptic activity. This was in agreement with existing literature and experimental observations that nerve endings were sensitive to light, and the IR light was shown to induce amino acid neurotransmitters to release by stimulating glutamatergic or GABAergic nerve endings (Nouvian, 2007; Wells et al., 2007b; Ahmed et al., 2008; Feng et al., 2010; Amaroli et al., 2018), leading to the transmission or inhibition of nerve excitation. However, the specific neuron types and corresponding stimulus parameters have not been integrated into a unified framework. The reasons for this variability may be the different neuronal types in the brain, such as excitatory neurons and inhibitory neurons, or their subtypes vary in response to INS (Ahmed et al., 2008; Feng et al., 2010). Though precise stimulus parameters have not been validated, and the mechanism by which optical energy causes changes in synaptic function remains unclear, our simulation result showed that

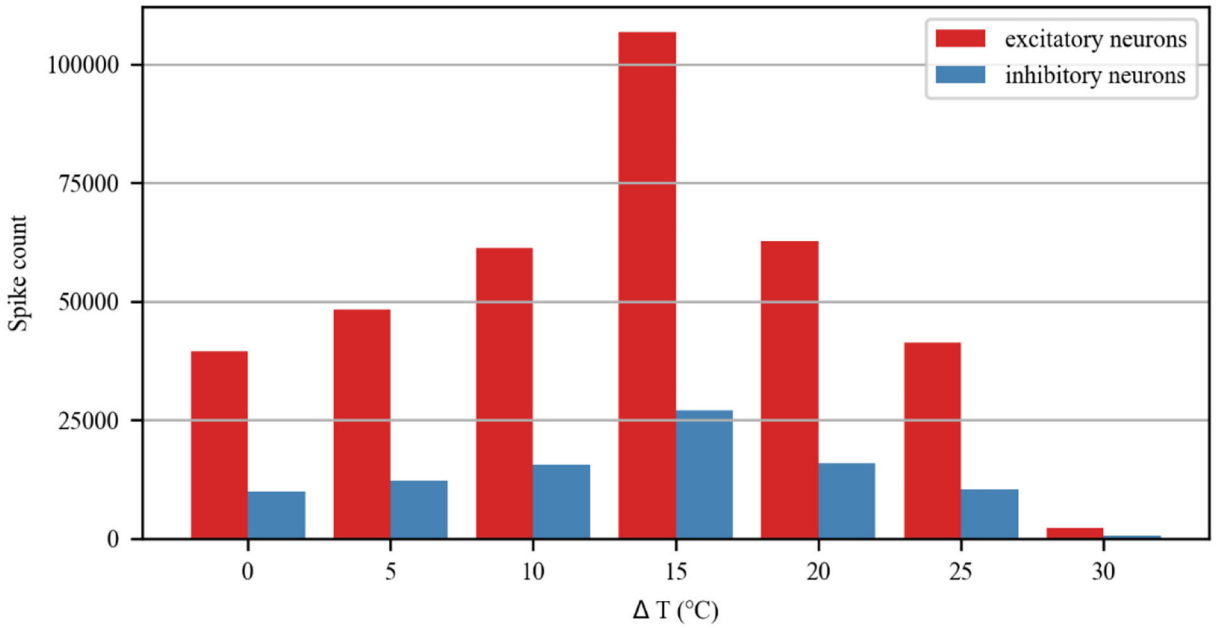


FIGURE 10
Spike count of excitatory and inhibitory neurons in the neuronal network during treatment with different temperatures.

synapses could act as filters through a release-decreasing or release-increasing response to the external stimulus (Abbott and Regehr, 2004). The changes in synaptic structure and function often represent plasticity, which is the candidate mechanism for the change in brain function. Numerous studies on transcranial magnetic stimulation and transcranial electrical stimulation demonstrated that these two types of stimulations had effects on synaptic plasticity (Fritsch et al., 2010; Tang et al., 2017). However, optical stimulation working through direct or indirect effects on synapsis requires further exploration to elucidate the exact mechanism. In addition to the cell level, the photothermal effect of optical stimulation could regulate neural network firing rhythms in the manner of dose-dependent biphasic cell response. It is well-known that brain activity depends largely on collective phenomena, which arise from the complex networks connected through synapses. The network model structure can be adapted and adjusted by sensing external stimuli, which is a scale between the macroscopic brain and the microscopic neuron (Sporns et al., 2005). Numerical simulation results of our model show that the characteristics of changes in the network match well with the response of individual neurons and synaptic activity to temperature increase. The most important was that these scenarios were reversible, not permanent. Apart from photothermal effects, photochemical and optoacoustic effects or some other effects could also potentially contribute the neuronal response to IR neural stimulation (Kramer et al., 2009; Shi et al., 2022). Further experiments are needed to explore these possibilities.

In conclusion, this study developed a computational model for simulating the response of cortical neurons to IR neural stimulation and enabled the quantification of the effects of photothermal effects on individual neurons, synapses, and networks. The numerical simulation results demonstrated the importance of the photothermal effects of INS. This model will be optimized and integrated into a multi-scale model in the future to guide non-invasive brain stimulation programs.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

JW and LL contributed to the conception of the study and wrote an initial draft of the manuscript. HS and ZD organized the database and prepared the figures. JY, MZ, and XL reviewed and edited the manuscript and performed the statistical analysis. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

Financial support for this work was provided by the National Natural Science Foundation of China (62006067), the Regional Innovation and Development Joint Fund of National Natural Science Foundation of China (U20A20224), and the Natural Science Foundation of Hebei Province (F2021201008) is gratefully acknowledged.

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.

References

Abbott, L. F., and Regehr, W. G. (2004). Synaptic computation. *Nature* 431, 796–803. doi: 10.1038/nature03010

Ahmed, N., Radwan, N. M., Ibrahim, K. M., Khedr, M. E., El Aziz, M. A., and Khadrawy, Y. A. (2008). Effect of three different intensities of infrared laser energy on the levels of amino acid neurotransmitters in the cortex and hippocampus of rat brain. *Photomed. Laser Surg.* 26, 479–488. doi: 10.1089/pho.2007.2190

Amaroli, A., Marcoli, M., Venturini, A., Passalacqua, M., Agnati, L. F., Signore, A., et al. (2018). Near-infrared laser photons induce glutamate release from cerebrocortical nerve terminals. *J. Biophotonics* 11:e201800102. doi: 10.1002/jbio.201800102

Barak, O., and Tsodyks, M. (2007). Persistent activity in neural networks with dynamic synapses. *PLoS Comput. Biol.* 3, 323–332. doi: 10.1371/journal.pcbi.0030104

Barborica, A., Oane, I., Donos, C., Daneasa, A., Mihai, F., Pistol, C., et al. (2022). Imaging the effective networks associated with cortical function through intracranial high-frequency stimulation. *Hum. Brain Mapp.* 43, 1657–1675. doi: 10.1002/hbm.25749

Brown, W. G. A., Needham, K., Begeng, J. M., Thompson, A. C., Nayagam, B. A., Kameneva, T., et al. (2021). Response of primary auditory neurons to stimulation with infrared light *in vitro*. *J. Neural Eng.* 18:046003. doi: 10.1088/1741-2552/abe7b8

Chandler, W. K., and Meves, H. (1970). Rate constants associated with changes in sodium conductance in axons perfused with sodium fluoride. *J. Physiol.* 211, 679–705. doi: 10.1113/jphysiol.1970.sp009299

Chen, G. H., and Chiao, C. C. (2020). Mild stress culture conditions promote neurite outgrowth of retinal explants from postnatal mice. *Brain Res.* 1747:147050. doi: 10.1016/j.brainres.2020.147050

Darlot, F., Moro, C., El Massri, N., Chabrol, C., Johnstone, D. M., Reinhart, F., et al. (2016). Near-infrared light is neuroprotective in a monkey model of Parkinson disease. *Ann. Neurol.* 79, 59–75. doi: 10.1002/ana.24542

Darmani, G., Bergmann, T. O., Butts Pauly, K., Caskey, C. F., De Lecea, L., Fomenko, A., et al. (2022). Non-invasive transcranial ultrasound stimulation for neuromodulation. *Clin. Neurophysiol.* 135, 51–73. doi: 10.1016/j.clinph.2021.12.010

De Pitta, M., and Brunel, N. (2016). Modulation of synaptic plasticity by glutamatergic gliotransmission: a modeling study. *Neural Plast.* 2016:7607924. doi: 10.1155/2016/7607924

Di Lazzaro, V., Rothwell, J., and Capogna, M. (2018). Noninvasive stimulation of the human brain: activation of multiple cortical circuits. *Neuroscientist* 24, 246–260. doi: 10.1177/1073858417717660

Ebtehaj, Z., Hatef, A., Malekmohammad, M., and Soltanolkotabi, M. (2018). Computational modeling and validation of thermally induced electrical capacitance changes for lipid bilayer membranes irradiated by pulsed lasers. *J. Phys. Chem. B.* 122, 7319–7331. doi: 10.1021/acs.jpcb.8b02616

Feng, H. J., Kao, C., Gallagher, M. J., Jansen, E. D., Mahadevan-Jansen, A., Konrad, P. E., et al. (2010). Alteration of GABAergic neurotransmission by pulsed infrared laser stimulation. *J. Neurosci. Methods* 192, 110–114. doi: 10.1016/j.jneumeth.2010.07.014

Fritsch, B., Reis, J., Martinowich, K., Schambra, H. M., Ji, Y. Y., Cohen, L. G., et al. (2010). Direct current stimulation promotes BDNF-dependent synaptic plasticity: potential implications for motor learning. *Neuron* 66, 198–204. doi: 10.1016/j.neuron.2010.03.035

Ganguly, M., Jenkins, M. W., Jansen, E. D., and Chiel, H. J. (2019). Thermal block of action potentials is primarily due to voltage-dependent potassium currents: a modeling study. *J. Neural Eng.* 16:036020. doi: 10.1088/1741-2552/ab131b

Goodman, D., and Brette, R. (2008). Brian: a simulator for spiking neural networks in python. *Front. Neuroinform.* 2:5. doi: 10.3389/neuro.11.005.2008

Hodgkin, A. L., and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* 117, 500–544. doi: 10.1113/jphysiol.1952.sp004764

Hodgkin, A. L., and Katz, B. (1949). The effect of temperature on the electrical activity of the giant axon of the squid. *J. Physiol.* 109, 240–249. doi: 10.1113/jphysiol.1949.sp004388

Huang, Y. Y., Chen, A. C. H., Carroll, J. D., and Hamblin, M. R. (2009). Biphasic dose response in low level light therapy. *Dose Resp.* 7, 358–383. doi: 10.2203/dose-response.09-027.Hamblin

Iaccarino, H. F., Singer, A. C., Martorell, A. J., Rudenko, A., Gao, F., Gillingham, T. Z., et al. (2016). Gamma frequency entrainment attenuates amyloid load and modifies microglia. *Nature* 540, 230–235. doi: 10.1038/nature20587

Jahr, C. E., and Stevens, C. F. (1990). Voltage dependence of NMDA-activated macroscopic conductances predicted by single-channel kinetics. *J. Neurosci.* 10, 3178–3182. doi: 10.1523/JNEUROSCI.10-09-03178.1990

Jirsa, V. K. (2004). Connectivity and dynamics of neural information processing. *Neuroinformatics* 2, 183–204. doi: 10.1385/NI:2:2:183

Kim, J. A., and Connors, B. W. (2012). High temperatures alter physiological properties of pyramidal cells and inhibitory interneurons in hippocampus. *Front. Cell. Neurosci.* 6:27. doi: 10.3389/fncel.2012.00027

Kramer, R. H., Fortin, D. L., and Trauner, D. (2009). New photochemical tools for controlling neuronal activity. *Curr. Opin. Neurobiol.* 19, 544–552. doi: 10.1016/j.conb.2009.09.004

Leuchtag, H. R. (1995). Fit of the dielectric anomaly of squid axon membrane near heat-block temperature to the ferroelectric Curie-Weiss law. *Biophys. Chem.* 53, 197–205. doi: 10.1016/0301-4622(94)00103-Q

Littlefield, P. D., Vujanovic, I., Mundi, J., Matic, A. I., and Richter, C. P. (2010). Laser stimulation of single auditory nerve fibers. *Laryngoscope* 120, 2071–2082. doi: 10.1002/lary.21102

Liu, T., Yang, Z. M., and Xu, S. H. (2009). Analytical investigation on transient thermal effects in pulse end-pumped short-length fiber laser. *Opt. Express* 17, 12875–12890. doi: 10.1364/OE.17.012875

Manos, T., Diaz-Pier, S., and Tass, P. A. (2021). Long-term desynchronization by coordinated reset stimulation in a neural network model with synaptic and structural plasticity. *Front. Physiol.* 12:716556. doi: 10.3389/fphys.2021.716556

Miyawaki, Y., Shinozaki, T., and Okada, M. (2012). Spike suppression in a local cortical circuit induced by transcranial magnetic stimulation. *J. Comput. Neurosci.* 33, 405–419. doi: 10.1007/s10827-012-0392-x

Nouvian, R. (2007). Temperature enhances exocytosis efficiency at the mouse inner hair cell ribbon synapse. *J. Physiol.* 584, 535–542. doi: 10.1113/jphysiol.2007.139675

Ocker, G. K., Litwin-Kumar, A., and Doiron, B. (2015). Self-organization of microcircuits in networks of spiking neurons with plastic synapses. *PLoS Comput. Biol.* 11:e1004458. doi: 10.1371/journal.pcbi.1004458

Peterson, E. J., and Tyler, D. J. (2012). "Activation using infrared light in a mammalian axon model," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (San Diego, CA: IEEE Engineering in Medicine and Biology Society), 1896–1899.

Plaksin, M., Shapira, E., Kimmel, E., and Shoham, S. (2018). Thermal transients excite neurons through universal intramembrane mechanoelectrical effects. *Phys. Rev. X.* 8, 011043. doi: 10.1103/PhysRevX.8.011043

Rajguru, S. M., Richter, C. P., Matic, A. I., Holstein, G. R., Highstein, S. M., Dittami, G. M., et al. (2011). Infrared photostimulation of the crista ampullaris. *J. Physiol.* 589, 1283–1294. doi: 10.1113/jphysiol.2010.198333

Richardson, R. T., Ibbotson, M. R., Thompson, A. C., Wise, A. K., and Fallon, J. B. (2020). Optical stimulation of neural tissue. *Healthc. Technol. Lett.* 7, 58–65. doi: 10.1049/htl.2019.0114

Santos-Sacchi, J., and Huang, G. (1998). Temperature dependence of outer hair cell nonlinear capacitance. *Hear. Res.* 116, 99–106. doi: 10.1016/S0378-5955(97)00204-9

Shapiro, M. G., Homma, K., Villarreal, S., Richter, C. P., and Bezanilla, F. (2012). Infrared light excites cells by changing their electrical capacitance. *Nat. Commun.* 3:736. doi: 10.1038/ncomms1742

Shi, L., Jiang, Y., Zheng, N., Cheng, J.-X., and Yang, C. (2022). High-precision neural stimulation through optoacoustic emitters. *Neurophotonics.* 9, 032207. doi: 10.1117/1.nph.9.3.032207

Singh, A. K., Mcgoldrick, L. L., Demirkhanyan, L., Leslie, M., Zakharian, E., and Sobolevsky, A. I. (2019). Structural basis of temperature sensation by the TRP channel TRPV3. *Nat. Struct. Mol. Biol.* 26, 994–998. doi: 10.1038/s41594-019-0318-7

Sporns, O., Tononi, G., and Kotter, R. (2005). The human connectome: a structural description of the human brain. *PLoS Comput. Biol.* 1, 245–251. doi: 10.1371/journal.pcbi.0010042

Stimberg, M., Goodman, D. F. M., Brette, R., and De Pittà, M. (2017). "Modeling neuron-glia interactions with the Brian 2 simulator," in *Computational Glioscience*, eds M. De Pittà and H. Berry (Cham: Springer; CRC Press), 471–505. doi: 10.1007/978-3-030-00817-8_18

Tanaka, Y., Akiyoshi, J., Kawahara, Y., Ishitobi, Y., Hatano, K., Hoaki, N., et al. (2011). Infrared radiation has potential antidepressant and anxiolytic effects in animal model of depression and anxiety. *Brain Stimul.* 4, 71–76. doi: 10.1016/j.brs.2010.04.001

Tang, A., Thickbroom, G., and Rodger, J. (2017). Repetitive transcranial magnetic stimulation of the brain: mechanisms from animal and experimental models. *Neuroscientist* 23, 82–94. doi: 10.1177/1073858415618897

Thompson, A. C., Wade, S. A., Cadusch, P. J., Brown, W. G. A., and Stoddart, P. R. (2013). Modeling of the temporal effects of heating during infrared neural stimulation. *J. Biomed. Opt.* 18:035004. doi: 10.1117/1.JBO.18.3.035004

Tsodyks, M., Pawelzik, K., and Markram, H. (1998). Neural networks with dynamic synapses. *Neural Comput.* 10, 821–835. doi: 10.1162/089976698300017502

Wells, J., Kao, C., Konrad, P., Milner, T., Kim, J., Mahadevan-Jansen, A., and Jansen, E. D. (2007a). Biophysical mechanisms of transient optical stimulation of peripheral nerve. *Biophys. J.* 93, 2567–2580. doi: 10.1529/biophysj.107.104786

Wells, J., Konrad, P., Kao, C., Jansen, E. D., and Mahadevan-Jansen, A. (2007b). Pulsed laser versus electrical energy for peripheral nerve stimulation. *J. Neurosci. Methods* 163, 326–337. doi: 10.1016/j.jneumeth.2007.03.016

Xia, Q. L., and Nyberg, T. (2019). Inhibition of cortical neural networks using infrared laser. *J. Biophoton.* 12:e201800403. doi: 10.1002/jbio.201800403

Yu, Y. G., Hill, A. P., and Mccormick, D. A. (2012). Warm body temperature facilitates energy efficient cortical action potentials. *PloS Comput. Biol.* 8:e1002456. doi: 10.1371/journal.pcbi.1002456

frontiers | Frontiers in Computational Neuroscience

TYPE Hypothesis and Theory
PUBLISHED 14 October 2022
DOI 10.3389/fncom.2022.869268

OPEN ACCESS

EDITED BY
Misha Tsodyks,
Weizmann Institute of Science, Israel

REVIEWED BY
Karl Friston,
University College London,
United Kingdom
Si Wu,
Peking University, China

*CORRESPONDENCE
James Joseph Wright
jj.w@xtra.co.nz

RECEIVED 02 March 2022
ACCEPTED 27 September 2022
PUBLISHED 14 October 2022

CITATION
Wright JJ and Bourke PD (2022) Unification of free energy minimization, spatiotemporal energy, and dimension reduction models of V1 organization: Postnatal learning on an antenatal scaffold.
Front. Comput. Neurosci. 16:869268.
doi: 10.3389/fncom.2022.869268

Unification of free energy minimization, spatiotemporal energy, and dimension reduction models of V1 organization: Postnatal learning on an antenatal scaffold

James Joseph Wright[1,2]* and Paul David Bourke[3]

[1]Centre for Brain Research, University of Auckland, Auckland, New Zealand, [2]Department of Psychological Medicine, School of Medicine, University of Auckland, Auckland, New Zealand, [3]Faculty of Arts, Business, Law and Education, School of Social Sciences, University of Western Australia, Perth, WA, Australia

Developmental selection of neurons and synapses so as to maximize pulse synchrony has recently been used to explain antenatal cortical development. Consequences of the same selection process—an application of the Free Energy Principle—are here followed into the postnatal phase in V1, and the implications for cognitive function are considered. Structured inputs transformed *via* lag relay in superficial patch connections lead to the generation of circumferential synaptic connectivity superimposed upon the antenatal, radial, "like-to-like" connectivity surrounding each singularity. The spatiotemporal energy and dimension reduction models of cortical feature preferences are accounted for and unified within the expanded model, and relationships of orientation preference (OP), space frequency preference (SFP), and temporal frequency preference (TFP) are resolved. The emergent anatomy provides a basis for "active inference" that includes interpolative modification of synapses so as to anticipate future inputs, as well as learn directly from present stimuli. Neurodynamic properties are those of heteroclinic networks with coupled spatial eigenmodes.

KEYWORDS

free energy principle, spatiotemporal energy, dimension reduction, visual cortex, synchronous oscillation, apoptosis, cortical self-organization

Introduction

Explication of the stimulus filter characteristics of neurons has been a major theme in neuroscience for more than 50 years and studied in greatest detail in cortical area V1. This analysis has contributed significantly to the field of artificial neural networks, as well as visual processing. Yet puzzles in the organization of the filter characteristics have

persisted, entwined with other puzzles—particularly the functional relevance of cortical columns and their variable definition in different cortical sites and species (Horton and Adams, 2005; Molnàr, 2013)—leaving uncertain the ways real cortical mechanisms differ from the simplified solutions applied in deep learning backpropagation networks (Domingos, 2015; Marblestone et al., 2016). Selective filtering in real neurons has been carried over into artificial neural networks, but is this the only essential property? How much of the orderliness in mesoscopic cortical anatomy has functional importance? Is this order, or its lack, a co-incidental manifestation of growth processes and just a metabolically efficient arrangement, or is it essential to information processing *per se*?

Latterly, theoretical developments, based upon fundamentals of information processing, computation, and predictive coding, suggest that, *via* the Free Energy Principle and its concept of "active inference" (Friston, 2005, 2010; Clark, 2013), a deeper unification of brain structure and cognitive function may be possible. This abstract concept requires explication in cellular specifics—literally and metaphorically, flesh on its bones—but offers clarification of the goal to be achieved by models of the brain. Hopefully moving toward that goal, this paper extends our earlier "minimum free energy" (Wright and Bourke, 2021a,b) account of antenatal mesoscopic neocortical development into the postnatal period. We will show that, in this extension, further functional relations between the meso-anatomy of the cortex, the filter characteristics of cortical neurons, and of the storage and manipulation of sensory images become apparent.

Feature filter models and problems encountered

From the foundational studies of Hubel and Wiesel, 1962, 1963, 1968, it was apparent that individual neurons responded to afferent pulses preferentially, as if filtering for selected characteristics, and were shown to exhibit anatomical order on the basis of these filter characteristics (Bonhoeffer and Grinvald, 1991 and subsequent)—whether OP, ocular dominance (OD), or the later emphasized SFP and TFP. Explaining how the selective characteristics developed was, and is, the central theoretical problem. A definitive review of early models, comparative and in historical order, is provided by Swindale (1996, 2008). A distinction may be drawn between models emphasizing feed-forward connections from the visual pathways, vs. those emphasizing contextual intracortical connections. The former class of models has been recently reviewed by Vidyasagar and Eysel (2015).

The initial feed-forward Hebbian models for OP share in common a conception of OP as consequent to the generation of a shaped field of excitation in small cortical areas and draw upon common assumptions of patterned retinal activity, Hebbian synapses, radially symmetric short-range excitatory and longer-range inhibitory lateral connections, and normalization of input strengths. Beginning from the work of von der Malsburg (1973), subsequent models in the family (von der Malsburg and Willshaw, 1976, 1977; Swindale, 1980, 1981a,b, 1982; Linsker, 1986a,b,c; Miller et al., 1989; Obermayer et al., 1990, 1992; Goodhill, 1993) varied in learning rule details and updating, range of lateral correlation and inhibitory surround, nature of synaptic competition, distribution of synaptic terminals from afferents, correlation of binocular inputs, etc. All produced, to a varying degree, good accounts of the topology of OP, columnar order, and OD but did not easily explain why ordered OP emerged in the antenatal period without structured visual input (Wiesel and Hubel, 1974). Internally generated retinal waves were then supposed to provide the needed stimulus (Galli and Maffei, 1988; Burgi and Grzywacz, 1994). Yet the converse finding that visual stimuli are required to maintain the OP order in post-natal life (Hubel and Wiesel, 1970; Blakemore and Van Sluyters, 1974) even to the extent of requiring lines at particular orientations for the development of normal dendritic structure (Tieman and Hirsch, 1982) seemed contradictory if a simple stimulus were sufficient. A crucial assumption—that of a symmetric inhibitory surround extending beyond each zone of excitation—was not justified anatomically. Further, this class of models treated OP as a fixed filter property—not a property interactive with other stimulus contexts—and this was to prove problematic.

Another suggestion made early by Hubel and Wiesel was directed not to the origin of the filter properties, but their spatial ordering, and led to the development of another major idea (Kohonen, 1982; Mitchison and Durbin, 1986; Durbin and Willshaw, 1987; Durbin and Mitchison, 1990; Swindale et al., 2000)—that all combinations of different feature responses should be equally well represented over all positions in visual space. This would necessarily involve conflict at all points between continuity and completeness of all types of filtered representations—yet would favor minimization of axon and dendrite distances of connection between the cells. Conflict resolution required a packing of cells of different categories, constrained so as to fit all features closely together in the best approximation possible. This accounted well for the organization of OP about pinwheel singularities, linear zones, and saddle points, and could be seen to be operating to good effect at the margins of OD columns, and also at elevation/azimuth lines, in variants of V1 organization (Yu et al., 2005; Farley et al., 2007). It even accounted for extremes of either high space frequency preference (HFSP) or low space frequency preference (LSFP) about OP pinwheels (Issa et al., 2000), since this produces the best general matching of all OPs with all SFPs because of conflicts in attaining best continuity (Issa et al., 2008). As well as introducing a "small world" notion of cortical connections, the concept implied "dimension reduction," since a higher dimensional feature space was being compressed onto

the two-dimensional cortical surface. Associated ideas from information theory suggested that information coming from the retina is projected to the cortex with minimization of redundancy (Barlow, 1959) and conservation of maximum mutual information (Linsker, 1989). The dimension reduction model was compatible with feed-forward and Hebbian accounts but did not depend upon them, since it could be argued that the Hebbian group of models had been successful in the reproduction of OP and OD simply because they had each provided non-unique conditions imposing continuity and completeness on the outcomes. Similar issues are now emerging in machine learning in the form of disentangling representations in deep (convolutional) neuronal networks (Higgins et al., 2021).

Hebbian feedforward models then encountered another need for revision. The separate filter characteristics were interdependent, not independent. SFP, TFP, and stimulus velocity were interrelated because TFP was the optimum combination of stimulus space frequency and velocity (Baker, 1990). The OP preferences of neurons were not, as they had initially been assumed, fixed, simple responses to a single line. A neuron's OP had been traditionally measured for slowly drifting stimulus lines oriented orthogonally to their direction of motion. However, OP varied systematically with speed of stimulus motion for all angles of attack other than that strictly orthogonal to motion, varying up to an OP orthogonal to that of the lowest speed (Basole et al., 2003, 2006). Prompted by this finding, the spatiotemporal energy model was advanced. This treated the individual neurons' responses as a combination of their OP, SFP, and TFP responses, with spatiotemporal energy defined as the product of stimulus space frequency and speed. The individual cell's responses could be predicted by summing feature preferences obtained from feature preference maps in the locale of the neuron (Zhang et al., 2007; Issa et al., 2008). When combined with the dimension reduction model, most problems seemed solved, but the origin of filter selectivity remained mysterious, and it was not entirely clear how OP and spatiotemporal energy were associated. An oddity not accounted for was that concurrent stimulation using stimuli with different orientation, yet all at optimum SFP, resulted in antagonistic blockade of responses, rather than independence or summation (Benevento et al., 1972; Blakemore and Tobin, 1972).

Arising from a rather different line of enquiry but motivated in part by the above problems, an account of the antenatal development of the neocortex was proposed by the present authors (Wright and Bourke, 2013, 2016, 2021a). The development of both columnar and of non-columnar cortex, the nature of superficial patch-to-patch connectivity, the organization of OP around singularities, OP linear zones and saddle points, like-to-like superficial patch/OP connections, and differences between monocular V1 and OD columns were explained. The model accounts for the emergence of ultra-small-world organization, and the generation of a transformed map of the visual input field—but does not depend upon structured input other than as diffuse noise. So, although consistent with continuity and completeness requirements, it is not a dimension reduction model in the usual sense. Synaptic competition and Hebbian learning are assumed, but the concept of an inhibitory surround is not required. The antenatal structure is considered a scaffold upon which postnatal organization can begin. The variation of OP with stimulus speed and angle of attack are explained, not as a consequence of combinations of features, but consequent to lag conduction within the superficial patch system. However, considerations of SFP and TFP were otherwise ignored. Consistent findings in non-columnar somatosensory cortex further supported the account (Wright et al., 2014) and it was later shown that the same principles can be extended from mesoscopic scale to inter-areal cortico-cortical connectivity (Wright and Bourke, 2021b). A link emerged to the very general, abstract, approach to learning proposed in the Free Energy Principle and related concepts of prediction error minimization and cortical computation, supplementing the earlier interpretations of continuity and completeness as redundancy minimization and maximization of mutual information.

Summary of antenatal model

Our model is applied in the very sparse one-to-many connectivity of cortical neurons under unified fast and slow synaptic learning rules (Izhikevich and Desai, 2003) and neural dynamics,[1] as summarized in Wright and Bourke (2021a,b). It has been observed that during embryogenesis synchronous firing of neurons protects them against apoptosis (Heck et al., 2008; Sang et al., 2021), as they form into small-world assemblies (Downes et al., 2012). This led us to propose that selection of developing neurons and synapses by apoptosis operates to maximize synchronous cell firing, thus shaping the outcome of genetically regulated cell numbers, patterns of cell migration, and differentiation into cell phenotypes (Rakic, 2009; Geschwind and Rakic, 2013). Synchronous oscillation is the "ground state" of equilibrium pulse exchanges among mixed excitatory and inhibitory cells (Chapman et al., 2002), so that, while constantly seeking equilibrium, the developing neurons also maximize their uptake of growth stimulation factors and thus tend to survive. Minimum resource consumption requires an approach to ultra-small-world configuration, further favoring avoidance of apoptosis early in embryogenesis.

1 The neural field models upon which our arguments are based follow from the work of Freeman (1975) and Liljenstrom (1991), and their development is recounted in Wright (2016). They model the power spectrum, frequency wavenumber content, evoked responses, and synchrony of electrocortical waves (Wright and Liley, 1996; Robinson et al., 1997, 2001; Rennie et al., 2002; Chapman et al., 2002).

Extension of these principles would also regulate the generation and pruning of synapses at later stages.

In the developing cortex, early spontaneous synchrony comes under the influence of the sensory periphery as soon as afferents reach the cortex (Schmidt et al., 1999; Espinosa and Stryker, 2012; Molnàr et al., 2020), and there is no clear transition from an antenatal to a postnatal state—merely an early phase and later stages through to adulthood. However, for purposes of convenience in the following account, we have referred to all later development once sensory inputs become structured as "postnatal," although no definite time of transition between antenatal to postnatal is clear.

The early selection process is followed in a population of short and long-axon excitatory intracortical cells mixed with short-axon inhibitory partners. Polysynaptic flow in the one-to-many sparse connectivity of neurons leads to multi-stable equilibria of pulse exchange between all cells even though few are initially monosynaptically connected. This explains how long-range correlation of firing of developing neurons appears even before long-range connections are established (Smith et al., 2018).

Equilibrium requires the excitatory and inhibitory populations each fire in phase with cells of the same type, and in inverse phase between the two populations, so that early in development

$$\varphi_{ij}(t) = \varphi_{ji}(t) \tag{1}$$

where φ_{ij} and φ_{ji} represent the exchanged pre-synaptic fluxes between *i-th* and *j-th* neurons over all pathways of connection. Competition and feedbacks inherent in synaptic learning rules lead toward bidirectional symmetry of gains along the prolific pathways, so a trend develops such that

$$\rho_{ij} g_{ij} \varepsilon_{ij} = \rho_{ji} g_{ji} \varepsilon_{ji} \tag{2}$$

where $\rho_{ij,ji}$ is the net structural synaptic connectivity between the two cells over all paths of connection, $g_{ij,ji}$ is their slowly consolidated synaptic gain, and $\varepsilon_{ij,ji}$ is fast transient synaptic efficacy. Each of the three factors converges on a separate time scale toward symmetry. Neurons unsuccessful in these competitive processes are eliminated, and initial, almost entirely unidirectional excitatory synaptic links become supplemented by an increased proportion of bidirectional monosynaptic connections, emerging from the polysynaptic background. Consistent with the Free Energy Principle, development follows a governing equation

$$F = A - C \tag{3}$$

where A is the population sum of pulse autocorrelations, C is the sum of pulse cross-correlations, and F, the analog of thermodynamic free energy, is continuously minimized as bidirectional monosynaptic connections increase in number. This formulation of self-organization of the functional architecture of visual cortex reflects a key fact of the Free Energy Principle: many self-organizing systems move toward generalized synchrony and minimization of prediction errors, until all interactions have become established and reliable, and provide a complementary interpretation of Equation 3.

$$\text{Freeenergy} = \text{accuracy minus complexity.}$$

Accuracy (under the free energy principle) is the expected log likelihood of some observable outcome (e.g., presynaptic strengths), while complexity scores the divergence between posterior and prior representations of the latent causes of observable inputs. This can be read as the degrees of freedom that are induced by presynaptic inputs to cause a change in internal representations stored in a neuronal population.

The geometrical consequences for cell organization are indicated in **Figure 1**. Symmetry of the formation of synapses in small-world configuration requires the longer-axon cells to become superficial patch cells, forming patch-to-patch connections with other long-axon cells, while the short-axon cells form local clusters. Short and long-axon cells connect reciprocally at a range at which the population density of their axonal trees are similar, creating in the process an approach to classical "like to like" OP connections [although in this model, and in reality, patch cells communicate more broadly than strictly "like to like" (Martin et al., 2014)]. This results in the formation of "global to local maps," where the "global map" is defined as the topography of an extended part of the cortical surface surrounding a local short-axon cluster, and the "local map" is the projection of the global map onto excitatory neurons of the local cluster. This leads to column formation, or to diffuse, apparently formless, connectivity, depending on the relative lengths of short and long axons—yet with the same pattern of small world organization—an order based on inverse synchrony-vs.-distance relations, synaptic competition, and local self-stabilization of pulse frequencies.

The emergent system provides lateral contextual information to neurons, determining their pattern of activation when they are also directly triggered by their extra-areal inputs. With regard to OP, coverage is both continuous and complete. This involves dimension reduction in a second sense, since, as bidirectional monosynaptic connections increase in number and free energy is minimized, system dimension falls.

The global to local maps are not simple Euclidean maps. Instead, they require projections to the inter-winding and cross-connected networks of local neurons and can be represented in the following mathematical form. P is a complex number position on the cortical surface and p is a complex number position within a local map with map center origin, p_0. The global map projection to any of many neighboring local maps takes the approximate form of projection of a Euclidean plane to intersecting Mobius strips, as

$$P \to \left\{ p = \pm p' \frac{(P-p_0)^n}{|P-p_0|^{n-1}} + p_0 \right\} \tag{4}$$

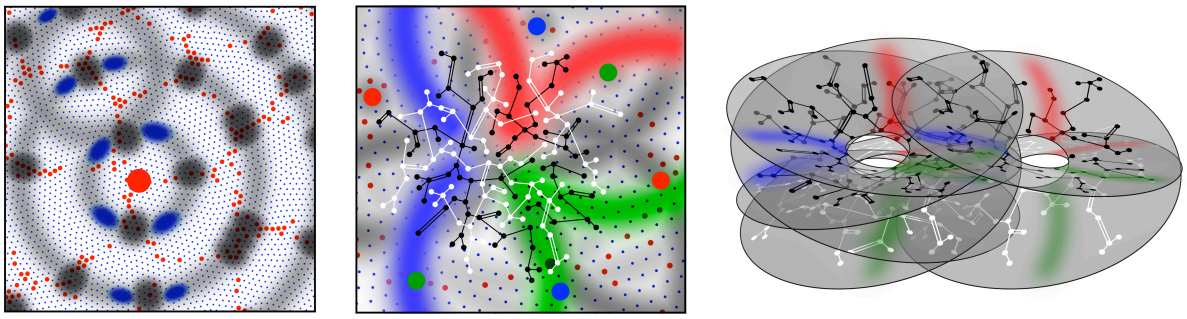


FIGURE 1
Patterns of synaptic connectivity seen in outcomes of growth simulations. **(Left)** Superficial patch cells. A representative long-axon (patch) cell (large red central spot) and patch connections. Surrounding zones of potential connection with other patch cells have been delineated in light gray concentric circles. Dark gray patches occur where other clusters of patch cells are positioned and able to make reciprocal connections, regularly spaced, patch-to-patch. Patch cell connections to short-axon local cells, maximizing resonance under stimulation of a particular angular domain are shown as darkened blue areas of "like to like" connections. **(Middle)** Local connectivity. Sparse short-axon cell connections have been marked in black or white, showing how interweaving networks occur. Some connections result in partial closure rather than complete independence of the interpenetrating networks. Fields of synaptic connections from patch cells to local cells are colored red, green, and blue according to their origins from diametrically opposite patch cell clusters These oppositely placed cell groups establish synapses on interpenetrating, distinct parts of the local cell network in a pattern best maximizing synchronous resonance, and creating local maps. **(Right)** A representation of intermingled networks of short-axon local cells conceptualized as cross-connected systems analogous to Mobius strips. Red, blue, and green bands indicate synaptic connections to/from the surrounding patch cell network. The degree of overlap of closed local cell connections can vary from clearly columnar to blurring with the apparent absence of columnar order.

where $p' = \sqrt{-1}k$ defines the rotation and scale of the local map, $\pm$ indicates map chirality, and $p_0 = p_0(1), p_0(2), p_0(3)$, are the local map centers. Symmetric reciprocal connections develop between superficial patch cells and local cells in arcs radiating from each map center, while maximum synchronous resonance requires the interpenetrating local networks are cross-linked into closed loops. Consequently n must take even integer values—the simplest case, $n = 2$, being that of projection to a single Mobius strip, or to multiple cross-linked Mobius strip-like networks. Other cases representing more complicated patterns of higher n may also be embedded and cross-linked with each other, but all appear similar to the simplest case, $n = 2$, when describing the appearance on the cortical surface, as if it were two dimensional.

Each inverse map, describing the return of reciprocal monosynaptic connections from local to patch cells, is given by

$$\mp P \leftarrow \pm \frac{1}{p'}\left(\mp \mathrm{p}\text{-}\mathrm{p}_0\right)^{\frac{1}{n}}\left|\mp \mathrm{p}\text{-}\mathrm{p}_0\right|^{n-1} - p_0 \qquad (5)$$

The $\mp$ sign (distinct from the use of $\pm$ for map chirality) is here introduced because the input map results in coincident mappings (as viewed in two dimensions) to the $0-\pi$ and $\pi-2\pi$ (i.e., $+$ or $-$) "limbs" in the Mobius representation from the Euclidean global positions at angles $0-2\pi$, thus creating the typical form of OP about a singularity at each map center.

Further maximizing synchronous resonance, adjacent local maps are arrayed in an approximately mirror-image formation, with cross-links between homologous map positions between neighbors. This results in the formation of linear zones and saddle points with, to greater or lesser degrees, interpenetration with other maps.

In proposing that developmental self-organization is based upon synchrony, we do not intend to exclude the possible relevance of alternative or complementary effects—as examples, organization of patch cell connectivity in a chemical diffusion model (Bauer et al., 2014), or recent revision of the retinal wave hypothesis (Kim et al., 2020)—and as a model of contextual interactions, there is some overlap with the model of Grabska-Barwinska and von der Malsburg (2008). However, the range of anatomical features explained by synchronous selection is so extensive that this model appears sufficient in itself.

Requirements for approach to minimum prediction error

Overarching rules for adaptation included in the Free Energy Principle (Friston, 2005, 2010; Friston et al., 2012; Clark, 2013; Ramstead et al., 2018) help define goals for the outcome of the present model. The Free Energy Principle requires that, as learning progresses, the states of lower neural subsystems are precisely predicted, and their perturbing effects minimized, by subsystems higher in the sensory hierarchy. Zero prediction error requires that for any cortical area (with V1 representative) as external signals are input to the intracortical cells, signals later return from their distributions to the local maps to the sites of input in a precisely required match.

In the antenatal model, the bidirectional intra-areal exchange of signals can be represented as

$$O\left(P, t - \frac{|P-p|}{v}\right) \rightarrow \{o\,(p, t)\} \tag{6}$$

and

$$\{o\,(p, t)\} \rightarrow O\left(P, t + \frac{|P-p|}{v}\right) \tag{7}$$

where v is the speed of intracortical signal conduction, $\{o\,(p, t)\}$ are sets of synchronous pulse activity generated in the local maps, and $O\left(P, t \pm \frac{|P-p|}{v}\right)$ are the patterns of activity generating forward and backward pulse trains between sites of arrival of the input signals and the laterally distributed local maps. At an asymptotic limit of fully completed learning the difference in forward and backward signals must be minimized to zero in the face of ongoing perturbation by the inputs. That is

$$\{o\,(p, t)\} \leftrightarrow O\,(P, t)\ \forall\,(P, p) \tag{8}$$

At that idealized limit, the input field and stored representations would exchange complete mutual information. This requires the exchanges must take place with group modes (eigenvectors of a delay matrix) that are invariant and bidirectionally symmetrical. There is an exact physical analogy to transmission without distortion of signals in fiber-optic cables, where absence of distortion (i.e., invariant group modes) requires continuous coupled interaction of spatial eigenmodes (Carpenter et al., 2016). So, our model may be expected to exhibit an analogous physiological expression of coupled spatial eigenmodes.

There are further demands to be made for a reasonably complete account. How will the newly induced selective-filter topography differ from the old? The stored information must enable association over both short and long ranges within the cortex. It should be seen how the antenatal organization provides a template for later development better than a random connectivity, and learning must converge more rapidly than a random walk.

Postnatal development

Spatiotemporal energy mapping *via* patch cells to local cells

We next consider the way in which inputs from the global field are conveyed to each local map.

Positions $P\,(1)$ and $P\,(2)$ on the cortical surface are crossed by a stimulus representation projected to the cortex, and convey pulses *via* superficial patch cells to a pair of closely situated local cells at positions $p\,(1)$ and $p\,(2)$ within any one of several local maps. We consider initially only the simplest cases, in which $p\,(1)$ and $p\,(2)$ pairs are always in the same limb of the same map. We need to determine conditions for arrival of synchronous, and near-synchronous, pulses at $p\,(1)$ and $p\,(2)$, since these will favor ongoing synaptic development within each local map.

Equivalent to a single space frequency in the representation of a moving object, consider a sinusoidal grating, with grating spacing L, and space frequency $K = 1/L$, moving over the surface at speed V, and angle θ to the line $P\,(1)\,P\,(2)$ —itself oriented at an angle ϕ in the P plane relative to the local map in which $p\,(1)$ and $p\,(2)$ lie (**Figure 2**, left). For simplicity, assume one action potential pulse is generated each time a grating line crosses $P\,(1)$ or $P\,(2)$. Pulses will be generated at a rate KV at $P\,(1)$ and $P\,(2)$, and KV is spatiotemporal energy.

Inputs to $p\,(1)$ and $p\,(2)$ travel a distance ΔS further from $P\,(1)$ than from $P\,(2)$

$$\Delta S = \left|P\,(1) - p\,(1)\right| - \left|P\,(2) - p\,(2)\right| \tag{9}$$

so the difference in time of pulse travel to $p\,(1)$ and $p\,(2)$ from the respective source is

$$\Delta T = \frac{\Delta S}{v} \tag{10}$$

where v is the speed of axonal conduction.

As grating lines cross $P\,(1)$ and $P\,(2)$, each grating line will traverse along $P\,(1)\,P\,(2)$ at a velocity $V sin\theta$, so the same grating line will generate pulses at $P\,(1)$ and then $P\,(2)$ after an interval δT

$$\delta T = \frac{|P(1)-P(2)|}{V sin\theta} \tag{11}$$

$|P\,(1) - P\,(2)|$ is necessarily some multiple of L, so

$$\delta T = \frac{mL}{V sin\theta} \tag{12}$$

Pairs of pulses must arrive at $p\,(1)$ and $p\,(2)$ with a time separation, λ

$$\lambda = \Delta T - \delta T = \frac{\Delta S}{v} - \frac{mL}{V sin\theta} \tag{13}$$

In the case that $\lambda = 0$, synchronous pulse-pairs arrive simultaneously at p (1) and p (2), and do so at a rate, ω, the rate of generation of synchronous pairs

$$\omega = \frac{1}{\Delta T} = \frac{v}{\Delta S} = \frac{KV sin\theta}{m} \tag{14}$$

The relative length, m, of $P\,(1)\,P\,(2)$, has an effect equivalent to alteration of the spatial frequency, so writing $K' = K/m$

$$\omega = \frac{v}{\Delta S} = K' V sin\theta \tag{15}$$

Since ω is a fixed function of ΔS, for any given $P\,(1)$ and $P\,(2)$, synchronous pairs can be created only for specific triplet combinations of $\{K, V, sin\theta\}$ (see **Figure 2**, middle). This means that as synchronous pair arrivals stimulate synchrony and encourage bidirectional synaptic connections among local neurons, they are also tuning these cells to specific combinations of spatiotemporal energy and the direction of object movement.

In all cases in which $\lambda \neq 0$, pulse pairs reach $p\,(1)$ and $p\,(2)$ asynchronously, with either a lead or lag. Unidirectional

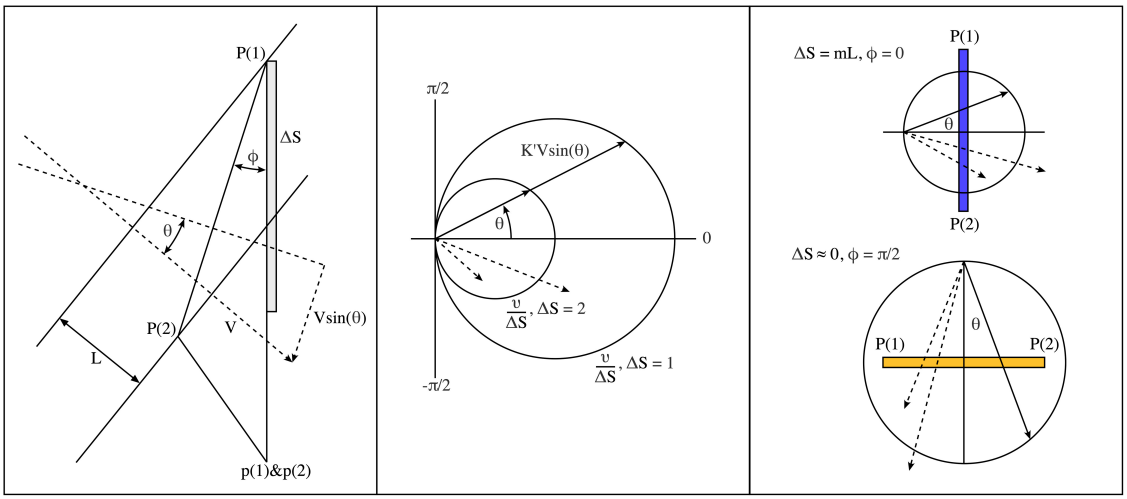


FIGURE 2
Geometric considerations determining the evolution of post-natal connectivity as structured external signals are imposed on the cortical field. **(Left)** Grating lines cross a pair of points, P (1) and P (2), in the global field, and axonal pulses are then relayed to closely situated local map cells, p (1) and p (2) *via* patch cell connectivity. **(Middle)** Polar diagram showing parameter combinations leading to synchronous pair arrival at p (1) and p (2). Circles show combinations resulting in synchronous pairs. Dashed vectors indicate a few of many possible asynchronous pulse arrivals at p (1) and p (2). **(Right)** Limiting cases of spatiotemporal orientation. Top: P (1) and P (2) are arranged radially to the local map singularity within which p (1) and p (2) lie. Bottom: P (1) and P (2) are arranged circumferentially.

monosynaptic connections will thus be promoted between $p\,(1)$ and $p\,(2)$, permitting the development of recurrent chains of connections promoting self-excitation among local cells. Self-exciting chains provide a basis for "winnerless competition" in synapse formation—an essential requirement of heteroclinic neural dynamics (Rabinovich et al., 2008). These considerations indicate that learned synaptic modifications will be capable of storing information about moving stimulus objects.

Spatiotemporal orientation

It can be seen from Equations 13 and 14–15 that λ and ω vary continuously for small changes of ΔS and θ, and all four terms are dependent upon the alignment, ϕ, of $P\,(1)\,P\,(2)$. Comparing the pairing of pulses generated from two closely situated pairs of cortical positions, $P\,(1)\,P\,(2)$ vs. $P\,(3)\,P\,(4)$, their difference in synchronous frequency is greatest when one pair of cortical positions is circumferential and one radially aligned (**Figure 2**, right). Conversely, afferent pulse pairs can approach concurrent synchrony when $P\,(1)\,P\,(2)$ and $P\,(3)\,P\,(4)$ are closely aligned and positioned. This relation of synchrony/asynchrony to alignment and position makes it helpful to define ϕ as the *Spatiotemporal Orientation* (STO).

Table 1 shows the bounds of the STO-related parameters, enabling these to be referred to as equivalents, according to the context.

The effects of STO on the organization of connections among local neurons are discussed in the following section, "Concurrent evolution of local cell connections: Dimension reduction, minimized prediction error, and eigenmode dynamics," and the implications for experimental findings in the sections subsequent, "Space frequency preference and temporal frequency preference experimental characteristics" and "Space frequency preference topographic order."

TABLE 1 Bounds of STO.

ϕ (STO)	0	$\frac{\pi}{2}$
ΔS	mL	0
$\omega = \frac{\nu}{\Delta S}$	$\frac{\nu}{mL}$	∞

Concurrent evolution of local cell connections: Dimension reduction, minimized prediction error, and eigenmode dynamics

The impact of synchronous pair arrivals upon synaptic organization in the local map can be anticipated from the same neurodynamic principles applied in the antenatal model (Chapman et al., 2002; Wright and Bourke, 2021a). Cooperative processes of excitatory synaptic connection generation, and antagonistic excitatory/inhibitory interactions, must each be considered.

From the considerations in sections "Spatiotemporal energy mapping *via* patch cells to local cells" and "Spatiotemporal

orientation," the induction of synchrony among local cells by the arrival of synchronous pulse pairs must bring about synapse formation such that the spatial arrangement of $P(1)$ and $P(2)$, vs. $P(3)$ and $P(4)$ pairs becomes mapped onto the density of synaptic connectivity between corresponding $p(1)$ and $p(2)$ vs. $p(3)$ and $p(4)$ pairs—inducing a shift from the antenatal radial arrangement of "like to like" connections toward a revised system in which a new circumferential order is imposed upon the prior radial arrangement, as shown in **Figure 3**, left and middle. Thus, STO becomes an imposed local map property, as a continuous variable distributed over the complete antenatal small world representation of stimulus space.

As a secondary effect, some connectivity will also emerge between cells at $p(1)$ and $p(2)$ and those at $p(3)$ and $p(4)$ because of local interactions. The amplitude of synchrony between any two neurons reflects the "in-phase" (even) components received by each, with dissipation of odd components (Chapman et al., 2002) and in case of four cells, the degree to which all four achieve co-synchrony achieves the highest magnitude where the cells share a common resonance frequency—so locally generated synchrony between $p(1)$ and $p(2)$ and $p(3)$ and $p(4)$ bring about a partial merging of their STO responsivity, also weighted, in accord with their ultra-small-world organization, by their squared separation distance, r^2. Certain $p(1)$ and $p(2)$ cell pairs that achieve early establishment of STO, $\{\phi_i\}$, by achieving the most stable pattern of co-synchrony over the local map, will force preliminary STO upon the more slowly developing connectivity. Provisional STO, $\{\phi_{int}\}$ thus imposed, can be approximated by interpolation,[2] as

$$\phi_{int} = arg \sum_{i=1}^{i=n} \frac{\phi_i}{1+Cr_i^2} \tag{16}$$

where C scales the range of interactions, but is without qualitative effects on the ordering of STO. The outcome in an example is shown in **Figure 3**, middle and right.

By this mechanism, local linkage between neurons of disparate STO will remain small compared to neurons of similar STO, achieving the compromise of continuity vs. completeness as well as smoothing and dimension reduction of the STO map. The modification and smoothing of STO at longer ranges has an important implication for learning because the property of spatiotemporal continuity at the level of the external stimulus world can be thus transferred to spatiotemporal continuity within the final STO organization of the local map—so as learning progresses, the interpolated circumferential/radial order must approximate more closely than chance the definitive order that will ultimately be attained. This constitutes a form of anticipatory prediction, minimizing future error, and facilitating Bayesian minimization during learning, not only minimizing prediction error on the basis of already-experienced inputs but anticipating aspects of the stimulus field not yet encountered.

As well as this cooperative organization of local connections, dynamic antagonism of radial and circumferential organizations must also arise, since these organizations share relatively few excitatory cross-links. Synchronous oscillation arises from equilibrium of exchange between both excitatory and inhibitory cells, with phase inversion between excitatory and inhibitory components (Wright and Bourke, 2021a), and it can be shown that where neurons lack strong excitatory cross-links, yet share interaction *via* intervening inhibitory short-axon cells, then equilibria can be reached by suppression of firing in either group by the other. Radially and circumferentially connected systems of neurons, engaging in crossed-inhibitory interactions, provide the anticipated analogy to coupled eigenmode dynamics, able to mediate the complicated time-sequences anticipated in heteroclinic dynamics. It may be noticed that this excitatory/inhibitory arrangement is not equivalent to the older concept of inhibitory surround.

Space frequency preference and temporal frequency preference experimental characteristics

Comparisons can be made with experimental observations, where synaptic connections have formed as described above.

(i) A neuron driven from any global position $P(1)$ by a drifting grid will respond by emitting pulses at frequency KV and will achieve maximum response at the frequency, ω, that best elicited synchronous resonance among the assembly of locally connected $p(1)$ and $p(2)$ pairs to which the stimulated neuron belongs—so exhibiting its TFP. That is, TFP = ω. TFP is more easily approached for cells with low TFP, given the relatively low stimulus speeds and wavenumbers that are experimentally practicable, compared to the high spatiotemporal energy required to approach TFP for cell with broad bandwidth, where $\Delta S \rightarrow 0$. So, for HSFP cells, their TFP will generally be outside the experimental range.

(ii) $SFP = K$ for given V, where $KV = \omega$

(iii) Cells with broad bandwidth and thus high SFP will respond better to stimuli with a broad space frequency spectrum, and therefore more strongly to square waves than single sinusoidal inputs.

These properties can account for findings reported in Zhang et al. (2007) and Issa et al. (2008) in support of the spatiotemporal energy model. They explained response curves

2 Values of ϕ_i were specified as unit vectors—either 1, 0 to represent a radial STO, or 0, 1, to represent a circumferential STO. These seeding values for interpolation were then used to compute interpolated STOs, $\{\phi_{int}\}$, throughout the local map, as arguments of weighted vector sums, in accord with Equation 16. This method utilizes the property that STO maps to positions in the local map, but is not a function of position—only of radial/circumferential orientation.

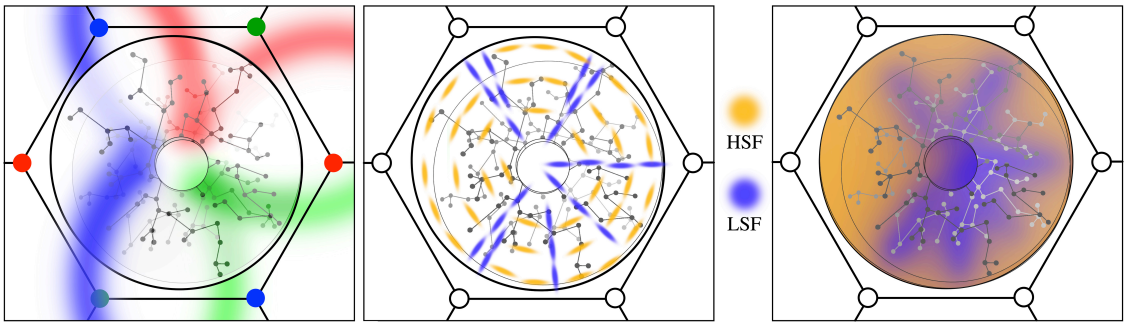


FIGURE 3
(Left) The antenatal local map. Like-to-like patch connections impose position and orientation on the global field onto the local map in radial form. **(Middle)** The early imposition of new patterns of local cell synaptic connection, based upon STO. Amber: local cells with circumferential orientation. Blue: radial orientation. **(Right)** Subsequent smoothing and dimension reduction *via* local synaptic connections. Zones of high and low spatial-frequency preference merge as local cell links provide smoothed interpolation between radial and circumferential extremes.

as composites of specific responses according to a combination of SFP and TFP. A strong TFP component (available by closer approach to ω in LSFP cells) explained the stronger recruitment of LSFP cells by increasing drift speed, compared to cells with HSFP, and the broader spatial bandwidth of square waves recruited HSFP neurons more than LSFP neurons.

Issa et al. (2008) also accounted for the change in OP with increasing stimulus speed and changes in stimulus angle of attack in spatiotemporal energy terms. As previously remarked, we have explained the same findings in terms of Doppler shift of lateral waves generated by a moving stimulus, without reference to OP/SFP linkage as such (Wright and Bourke, 2013). However, since both accounts successfully match the experiment, they can be considered equivalent time-series vs. Fourier explanations of the phenomena—or put in other terms, the Doppler-shifted spatiotemporal energy of input signals affects the SF and TFP of the local cells.

The crossed inhibition exerted each upon the other by circumferential and radially arrayed linked groups explains why response to concurrent presentation of stimulus grids of similar space frequency and speed, but differing orientations, produces not summation but cancelation of response—a property not otherwise explained in earlier models. This effect has further consequences for the topography of SFP.

Space frequency preference topographic order

Figure 4 shows the way in which SFP becomes topographically ordered in the way found experimentally and shows that the topological order reflects the degree of synergy or conflict between STO and OP responses in different situations. $\triangle S \rightarrow 0$ isolines can be constructed circling the singularity at all distances, and orthogonal to the antenatal radial like-to-like lines, for which $\triangle S \approx mL$. The antagonistic cross-inhibitory interactions of circumferential and radial arrangements induce conflicts near the singularity. For stable synaptic consolidations to be attained, distinct domains of either high, or of low, SFP, must appear randomly located around OP singularities, consequent to conflict resolution one way or the other. Conversely, the association of HSFP areas with OP linear zones also follows, as there is minimal conflict far from the singularity, where OP is itself essentially circumferential in relationship to the adjacent singularities. This can be seen in the form of the curved like-to-like connections shown in **Figure 3** left, and the same effect is suggested by dashed curved lines in **Figure 4**, left and center. This joint alignment at the map periphery causes STO and OP to be synergic, both connection systems arising with low ΔS, and therefore SFP high.

With these extensions from the antenatal to the postnatal situation, the properties of SFP and TFP order are added to those of OP order. The present model thus incorporates the properties of both the dimension-reduction and spatiotemporal energy models.

The storage of correlations at long range

Developing local connections permits association over short global distances, but how can learning of short-range correlations be generalized to association over the wider field of a cortical area? To answer this question, we consider the general case, in which cell pairs $\mp p\,(1)$ and $\mp p\,(2)$ are closely physically proximate in the cortical surface, but instead of being only close neighbors in the same limb of a single map as we first considered above—the two cells may be located in the same, or different limbs, within a single map, or in different intertwined maps, which may be of the same, or of different, chirality.

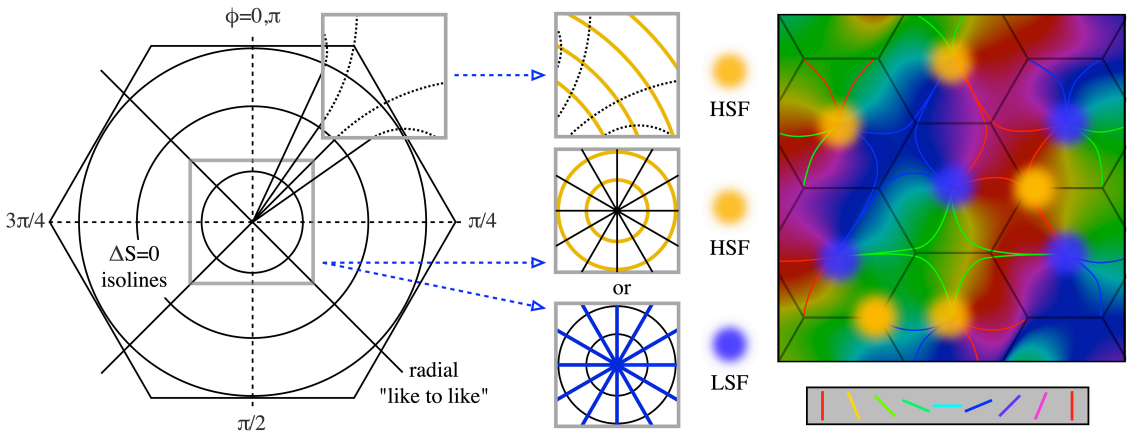


FIGURE 4
Synaptic competition and the emergence of experimentally observed spatial-frequency preference. **(Left)** Diagram shows the circumferential arrangement of p (1) and p (2) pairs responding to higher spatial frequencies, and the radial (along "like-to-like" lines) arrangement of pairs responding to lower spatial frequencies. At the periphery of the local map, like-to-like connections curve into a more circumferential array, as indicated by the dashed continuation of the radial lines. **(Middle)** Cut-out sections (top) show that on the local map periphery, where OP is normally continuous with that in the adjacent local map, circumferential pairs can be arranged contiguously with low conflict with radial arrangement. In contrast, near the singularity (lower cut-out sections) conflicts of radial and circumferential arrangement can lead to one or other of alternate HSFP or LSFP outcomes. **(Right)** Consequently, OP and high SFP are found together in OP linear zones, while SFP around the singularity must be either HSFP or LSFP (cp Issa et al., 2008).

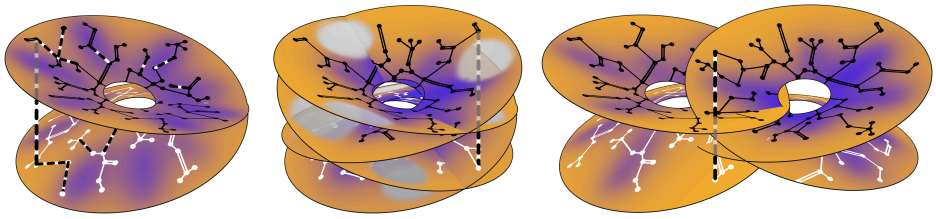


FIGURE 5
Establishment of long-range correlations. The antenatal neural connections are shown as in **Figure 1** (right), and postnatal connection modifications are shown in dashed black and white. **(Left)** Within a single Mobius- like connection system, some antenatal connections are preferentially reinforced, while new post-natal connections also form, bridging the limbs of the earlier system. **(Middle)** Postnatal bridging connections establish further longer-range correlations within a multiplicity of such systems all surrounding a single singularity, each with only a partially complete representation of SFP. **(Right)** Further extending the possible range of association, overlapping local maps, surrounding separate singularities, are similarly brought into association by further postnatal bridges.

The way in which $\triangle S$ in the global field is related to distances within different local maps, or different limbs of the same local map, follows from the inverse maps (Equation 5)

$$\triangle S = \left| \pm \tfrac{1}{p}\left(\mp p\,(1) - p_0\,(1)\right)^{\frac{1}{n}} \left|\mp p\,(1) - p_0\,(1)\right|^{n-1} - p_0(1)\right| - \left| \pm \tfrac{1}{p}\left(\mp p\,(2) - p_0\,(2)\right)^{\frac{1}{n}} \left|\mp p\,(2) - p_0\,(2)\right|^{n-1} - p_0\,(2)\right| \quad (17)$$

and similar considerations apply to the creation of synchronous pair inputs as in the simpler case.

Therefore the general case includes the possibility of long-range associations by local synaptic linkages—of disparate inputs from widely separated positions, differing relative orientations, and translations in the global field—all created by further cross-connections, breaking the antenatal Mobius-like order. The richness and range of cross connections that can be made in this way depend on overlap of local maps and suggest why the apparently random non-columnar order in most cortical areas may be functionally advantageous.

The positioning of cell positions with regard to breakdown of the Mobius order is shown in **Figure 5**.

Conclusion

The goals for the extension of our antenatal model to postnatal development appear to have been met. Without the introduction of new assumptions, we have shown that the change from random noise inputs to

structured inputs can transform the small world model of spatial positions and their short-range correlations to a finer grain of association at short and long ranges, and in temporal sequences. In this way, the antenatal structure acts as a scaffold able to guide finer resolution of spatiotemporal information within the pre-existing antenatal local maps.

This model conforms to expectation from the Free Energy Principle, with reduction of variational free energy and dimension reduction accompanying continually increasing mutual information between external inputs and the synaptic order, and provides a mechanism (coupled spatial eigenmodes) for asymptotic approach to zero "surprisal." The previously unexplained observation that space-frequency-tuned responses delivered at multiple orientations block one another, seems to be of crucial significance, since this effect underlies the interaction of spatial eigenmodes.

It appears that the antenatal scaffold promotes later learning by "active inference" in ways that go beyond back-propagation in random networks, as usually conceived. First, the antenatal scaffold arising because of the declining synchrony-vs.-distance relationship general among cortical neurons establishes initial connections that conveniently approximate the topological order of generally declining cross-correlation-vs.-distance relationships of the sensory world in space and time. The initial antenatal order then gives way to postnatal connections that progressively represent ever more detailed partial correlations in the sensory world, superimposed upon, and given order by, the basic framework. Second, the establishment of later learning on the antenatal framework further exploits the cross-correlated structure of space and time to fill in tentative synaptic connections by the extrapolation mechanism described in section "Concurrent evolution of local cell connections: Dimension reduction, minimized prediction error, and eigenmode dynamics," in advance of receipt of later inputs. That is, the general statistical order of the known is used to continuously update anticipation of the likely structure of the unknown. Generalizing to all subsequent exchanges within the cortical hierarchy, this would contribute a degree of flexible creativity to the brain's self-supervision.

The long-standing interpretation of feature preferences as inherent filter properties of individual neurons is further qualified, and we have introduced a new concept, STO. Our account explains all the data incorporated in the spatiotemporal energy and dimension reduction models, and provides an explanatory mechanism for both, unifying this with other anatomical features explained by the earlier antenatal model. It does not purport to be an exhaustive model, of course. Discrepancies include the occurrence of OP fractures, and the occurrence of direction preference in some species, alluded to in Issa et al. (2008). Properties of the input pathways lying outside our consideration may account for these discrepancies.

Further testing of this model is within the realm of existing technologies. Further single cell testing using the methods described in Zhang et al. (2007) could test whether the postulated link between strength and preferred frequency of synchrony among cells in a small locale, and their individual TFP, is in fact the case. Detailed neurodynamic simulations are required to further demonstrate that the evolution of connections follows the paths we have here indicated. Synaptic architectonics at the micro- and meso-scales could be analyzed to see that the proposed general organization of local cells and patch cells follows the same form, whether the cortex is columnar or non-columnar, and accords with a Mobius-like organization. A relatively simple test would be to confirm or deny that superficial patch synapses from neurons on opposite sides of an OP singularity terminate on different "limbs" of the Mobius-like sheafs of local neuron connections.

As is emphasized in the Free Energy Principle, systems that learn—or develop—to minimize variational free energy are simply those in which members of an ensemble can predict each other accurately and with minimum complexity cost (i.e., maximum information and thermodynamic efficiency; Jarzynski, 1997). Since this phenomenon can be seen in *in vitro* cell cultures exposed to a structured input (Isomura and Friston, 2018) there is a possibility of testing the model by using structured inputs in cell culture preparations, to see what extent epigenetic scaffolding emerges and is necessary.

Although presented in terms of V1, there is reason to believe the model sufficiently general to apply throughout neocortex. Provisional extension to inter-areal interactions, and to computation mediated by inter-areal interactions (Perlovsky et al., 2011), as well as the necessary additional role of brain-stem mediated reward-based learning, have been discussed in association with other commentators in Wright and Bourke (2021b). If ultimately shown to be valid, this model may have implications for the further development of artificial intelligence, since it differs considerably from current orthodox deep learning networks, and, in its alliance with aspects of the Free Energy Principle, suggests the capacity for unsupervised learning.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JW devised and wrote the manuscript. PB was responsible for software and graphics. Both authors contributed to the article and approved the submitted version.

Funding

This work received long-term support including the Frank Hixon Fund of the California Institute of Technology, the MHRF and Wellcome Trust in the UK, the United Kingdom, New Zealand, and Australian Medical Research Councils, the Oakley, and Pratt Foundations of Australasia.

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

References

Baker, C. L. (1990). Spatial- and temporal-frequency selectivity as a basis for velocity preference in cat striate cortex neurons. *Vis. Neurosci.* 4, 101–113. doi: 10.1017/S0952523800002273

Barlow, H. B. (1959). "Sensory mechanisms, the reduction of redundancy, and intelligence," in *Proceedings of the NPL symposium on the mechanisation of thought processes No 10* (London: HMSO).

Basole, A., Kreft-Kerekes, V., White, L. E., and Fitzpatrick, D. (2006). Cortical cartography revisited: A frequency perspective on the functional architecture of visual cortex. *Prog. Brain Res.* 154, 121–134. doi: 10.1016/S0079-6123(06)54006-3

Basole, A., White, L. E., and Fitzpatrick, D. (2003). Mapping multiple features in the population response of visual cortex. *Nature* 423, 986–990.

Bauer, R., Zubler, F., Hauri, A., Muir, D. R., and Douglas, R. J. (2014). Developmental origin of patchy axonal connectivity in neocortex: A computational model. *Cereb. Cortex* 24, 487–500. doi: 10.1093/cercor/bhs327

Benevento, L. A., Creutzfeldt, O. D., and Kuhnt, U. (1972). Significance of intracortical inhibition in the visual cortex. *Nat. New Biol.* 238, 124–126.

Blakemore, C., and Tobin, E. A. (1972). Lateral inhibition between orientation detectors in the cat's visual cortex. *Exp. Brain Res.* 15, 439–444. doi: 10.1007/BF00234129

Blakemore, C., and Van Sluyters, R. C. (1974). Reversal of the physiological effects of monocular deprivation in kittens: Further evidence for a sensitive period. *J. Physiol.* 237, 195–216. doi: 10.1113/jphysiol.1974.sp010478

Bonhoeffer, T., and Grinvald, A. (1991). ISO-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature* 353, 429–431.

Burgi, P.-Y., and Grzywacz, N. M. (1994). Model for the pharmacological basis of spontaneous synchronous activity in developing retinas. *J. Neurosci.* 14, 7426–7439.

Carpenter, J., Benjamin, J., and Eggleton, J. S. (2016). Comparison of principle modes and spatial eigenmodes in multimode optical fibre. *Laser Photonics Rev.* 11:1600259. doi: 10.1002/lpor.201600259

Chapman, C. L., Bourke, P. D., and Wright, J. J. (2002). Spatial eigenmodes and synchronous oscillation: Coincidence detection in simulated cerebral cortex. *J. Math. Biol.* 45, 57–78. doi: 10.1007/s002850200141

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–253. doi: 10.1017/S0140525X12000477

Domingos, P. (2015). *The master algorithm*. New York, NY: Basic Books.

Downes, J. H., Hammond, M. W., Xydas, D., Spencer, M., Becerra, V. M., Warwick, K., et al. (2012). Emergence of a small-world functional network in cultured neurons. *PLoS Comput. Biol.* 8:e1002522. doi: 10.1371/journal.pcbi.1002522

Durbin, R., and Mitchison, G. (1990). A dimension reduction framework for understanding cortical maps. *Nature* 343, 644–647. doi: 10.1038/343644a0

Durbin, R., and Willshaw, D. J. (1987). An analogue approach to the traveling salesman problem using an elastic net method. *Nature* 326, 689–691. doi: 10.1038/326689a0

Espinosa, J. S., and Stryker, M. P. (2012). Development and plasticity of the primary visual cortex. *Neuron* 75, 230–249. doi: 10.1016/j.neuron.2012.06.009

Farley, B. J., Yu, H., Jin, D. Z., and Sur, M. (2007). Alteration of visual input results in a coordinated reorganization of multiple visual cortex maps. *J. Neurosci.* 27, 10299–10310. doi: 10.1523/JNEUROSCI.2257-07.2007

Freeman, W. J. (1975). *Mass action in the nervous system*. Cambridge, MA: Academic Press.

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. B* 360, 815–836.

Friston, K. (2010). The free energy principle: A unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138.

Friston, K., Thornton, C., and Clark, A. (2012). Free energy minimization and the dark room problem. *Front. Psychol.* 3:130. doi: 10.3389/fpsyg.2012.00130

Galli, L., and Maffei, L. (1988). Spontaneous impulse activity of rat retinal ganglion cells in prenatal life. *Science* 242, 90–91. doi: 10.1126/science.3175637

Geschwind, D. H., and Rakic, P. (2013). Cortical evolution: Judge the brain by its cover. *Neuron* 80, 633–647. doi: 10.1016/j.neuron.2013.10.045

Goodhill, G. J. (1993). Topography and ocular dominance: A model exploring positive correlations. *Biol. Cybern.* 69, 109–118. doi: 10.1007/BF00226194

Grabska-Barwinska, A., and von der Malsburg, C. (2008). Establishment of a scaffold for orientation maps in primary visual cortex of higher mammals. *J. Neurosci.* 28, 249–257. doi: 10.1523/JNeurosci.5514-06.2008

Heck, N., Golbs, A., Riedemann, T., Sun, J.-J., Lessmann, V., and Luhmann, H. J. (2008). Activity dependent regulation of neuronal apoptosis in neonatal mouse cerebral cortex. *Cereb. Cortex* 18, 1335–1349.

Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsa, D., et al. (2021). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nat. Commun.* 12:6456. doi: 10.1038/s41467-021-26751-5

Horton, C. H., and Adams, D. L. (2005). The cortical column: A structure without a function Philos. *Trans. R. Soc. Lond. B. Biol. Sci.* 360, 837–862.

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture of cat striate cortex. *J. Physiol.* 160, 106–154. doi: 10.1615/CritRevBiomedEng.2017020607

Hubel, D. H., and Wiesel, T. N. (1963). Receptive fields of cells in striate cortex of very young, visually inexperienced kittens. *J. Neurophysiol.* 26, 994–1002.

Hubel, D. H., and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195, 215–243.

Hubel, D. H., and Wiesel, T. N. (1970). The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *J. Physiol.* 206, 419–436. doi: 10.1113/jphysiol.1970.sp009022

Isomura, T., and Friston, K. (2018). In vitro neural networks minimize variational free energy. *Sci. Rep.* 8:16926. doi: 10.1038/s41598-018-35221-w

Issa, N. P., Rosenberg, A., and Hussson, T. R. (2008). Models and measurements of functional maps in V1. *J. Neurophysiol.* 99, 2754–2754.

Issa, N. P., Trepel, C., and Stryker, M. P. (2000). Spatial frequency maps in cat visual cortex. *J. Neurosci.* 20, 8504–8514.

Izhikevich, E. M., and Desai, N. S. (2003). Relating STDP to BCM. *Neural Comput.* 15, 1511–1523.

Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* 78:2690. doi: 10.1103/PhysRevLett.78.2690

Kim, J., Song, M., Jang, J., and Paik, S.-B. (2020). Spontaneous retinal waves can generate long-range horizontal connectivity in visual cortex. *J. Neurosci.* 40, 6584–6599. doi: 10.1523/JNeurosci.0649-20.2020

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biol. Cybern.* 43, 59–69.

Liljenstrom, H. G. (1991). Modelling the dynamics of olfactory cortex using simplified network units and realistic architecture. *Int. J. Neural Syst.* 2, 1–15. doi: 10.1142/S0129065791000029

Linsker, R. (1986a). From basic network principles to neural architecture: Emergence of spatial-opponent cells. *Proc. Natl Acad. Sci. U.S.A.* 83, 7508–7512. doi: 10.1073/pnas.83.19.7508

Linsker, R. (1986b). From basic network principles to neural architecture: Emergence of orientation selective cells. *Proc. Natl. Acad. Sci. U.S.A.* 83, 8390–8394.

Linsker, R. (1986c). From basic network principles to neural architecture: Emergence of orientation columns. *Proc. Natl. Acad. Sci. U.S.A.* 83, 8779–8783. doi: 10.1073/pnas.83.22.8779

Linsker, R. (1989). How to generate ordered maps by maximizing the mutual information between input and output signals. *Neural Comput.* 1, 402–411.

Marblestone, A. H., Wayne, G., and Kording, K. (2016). Toward an integration of deep learning and neuroscience. *Front. Comp. Neurosci.* 10:94. doi: 10.3389/fncom.2016.00094

Martin, K. A. C., Roth, S., and Rusch, E. S. (2014). Superficial layer pyramidal cells communicate heterogeneously between multiple functional domains of cat primary visual cortex. *Nat. Commun.* 5:5252. doi: 10.1038/ncomms6252

Miller, K. D., Keller, J. B., and Stryker, M. P. (1989). Ocular dominance column development: Analysis and simulation. *Science* 245, 605–615.

Mitchison, G., and Durbin, R. (1986). Optimal numberings of an N × N array SIAM. *J. Alg. Disc. Methods* 7, 571–581.

Molnaìr, Z. (2013). "Cortical columns," in *Comprehensive developmental neuroscience: Neural circuit development and function in the brain*, Vol. 3, eds J. L. R. Rubenstein and P. Rakic (Amsterdam: Elsevier), 109–129.

Molnaìr, Z., Luhmann, H. J., and Kanold, P. O. (2020). Transient cortical circuits match spontaneous and sensory driven activity during development. *Science* 370:eabb2153. doi: 10.1126/science.abb2153

Obermayer, K., Ritter, H., and Schulten, K. (1990). A principle for the formation of the spatial structure of cortical feature maps. *Proc. Natl. Acad. Sci. U.S.A.* 87, 8345–8349.

Obermayer, K., Ritter, H., and Schulten, K. (1992). A model for the development of the spatial structure of retinotopic maps and orientation columns. *IEICE Trans. Fundamentals* E75-A, 537–545.

Perlovsky, L. I., Deming, R. W., and Ilin, R. (2011). *Emotional cognitive neural algorithms with engineering applications, dynamic logic: From vague to crisp*. Heidelberg: Springer.

Rabinovich, M. I., Huerta, R., Varona, P., and Afraimovich, V. S. (2008). Transient cognitive dynamics, metastability, and decision making. *PLoS Comput. Biol.* 4:e1000072. doi: 10.1371/journal.pcbi.1000072

Rakic, P. (2009). Evolution of neocortex: Perspective from developmental biology. *Nat. Rev. Neurosci.* 10, 724–735. doi: 10.1038/nrn2719

Ramstead, M. J. D., Badcock, P. B., and Friston, K. (2018). Answering Schroedinger's question: A free energy formulation. *Phys. Life Rev.* 24, 1–16. doi: 10.1016/j.plrev.2017.09.001

Rennie, C. J., Robinson, P. A., and Wright, J. J. (2002). Unified neurophysical model of EEG spectra and evoked potentials. *Biol. Cybern.* 86, 457–471. doi: 10.1007/s00422-002-0310-9

Robinson, P. A., Rennie, C. J., and Wright, J. J. (1997). Propagation and stability of waves of electrical activity in the cerebral cortex. *Phys. Rev. E* 56, 826–841.

Robinson, P. A., Rennie, C. J., Wright, J. J., Bahramali, H., Gordon, E., and Rowe, D. L. (2001). Prediction of electroencephalographic spectra from neurophysiology. *Phys. Rev. E* 63:021903.

Sang, I. E. W. F., Schroer, J., Halbhuber, L., Warm, D., Yang, J.-W., Luhmann, H. J., et al. (2021). Optogenetically controlled activity pattern determines survival rate of developing neocortical neurons. *Int. J. Mol. Sci.* 22:6575. doi: 10.3390/ijms22126575

Schmidt, K. E., Galuske, R. A., and Singer, W. (1999). Matching the modules: Cortical maps and long-range intrinsic connections in visual cortex during development. *J. Neurobiol.* 41, 10–17. doi: 10.1002/(sici)1097-4695(199910)41:1<10::aid-neu3>3.0.co;2-l

Smith, G. B., Hein, B., Whitney, D. E., Fitzpatrick, D., and Kaschube, M. (2018). Distributed network interactions and their emergence in developing neocortex. *Nat. Neurosci.* 21, 1600–1608. doi: 10.1038/s41593-018-0247-5

Swindale, N. V. (1980). A model for the formation of ocular dominance stripes. *Proc. R. Soc. B* 208, 243–264. doi: 10.1098/rspb.1980.0051

Swindale, N. V. (1981a). Rules for pattern formation in mammalian visual cortex. *Trends Neurosci.* 4, 102–104.

Swindale, N. V. (1981b). Absence of ocular dominance patches in dark reared cats. *Nature* 290, 332–333. doi: 10.1038/290332a0

Swindale, N. V. (1982). A model for the formation of orientation columns. *Proc. R. Soc. B* 215, 211–230.

Swindale, N. V. (1996). The development of topography in the visual cortex: A review of models. *Network* 7, 161–247.

Swindale, N. V. (2008). Visual map. *Scholarpedia* 3:4607.

Swindale, N. V., Shoham, D., Grinvald, A., Bonhoeffer, T., and Hubener, M. (2000). Visual cortical maps are optimized for uniform coverage. *Nat. Neurosci.* 3, 822–826. doi: 10.1038/77731

Tieman, S. B., and Hirsch, H. V. (1982). Exposure to lines of only one orientation modifies dendritic morphology of cells in the visual cortex of the cat. *J. Comp. Neurol.* 211, 353–362. doi: 10.1002/cne.902110403

Vidyasagar, T. R., and Eysel, U. T. (2015). Origins of feature selectivities and maps in the mammalian primary visual cortex. *Trends Neurosci.* 38, 475–485.

von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik* 14, 85–100.

von der Malsburg, C., and Willshaw, D. J. (1976). A mechanism for producing continuous neural mappings: Ocularity dominance stripes and ordered retino-tectal projections. *Exp. Brain Res.* (Suppl 1), 463–469.

von der Malsburg, C., and Willshaw, D. J. (1977). How to label nerve cells so that they can interconnect in an ordered fashion. *Proc. Natl. Acad. Sci. U.S.A.* 74, 5176–5178. doi: 10.1073/pnas.74.11.5176

Wiesel, T. N., and Hubel, D. H. (1974). Ordered arrangement of orientation columns in monkeys lacking visual experience. *J. Comput. Neurol.* 158, 307–318. doi: 10.1002/cne.901580306

Wright, J. J. (2016). *Work toward a theory of brain function*. DSc dissertation. Dunedin: University of Otago. Available online at: https://ourarchive.otago.ac.nz/handle/10523/6400

Wright, J. J., and Bourke, P. D. (2013). On the dynamics of cortical development: Synchrony and synaptic self-organization. *Front. Comput. Neurosci.* 7:4. doi: 10.3389/fncom.2013.00004

Wright, J. J., and Bourke, P. D. (2016). Further work on the shaping of cortical development and function by synchrony and metabolic competition. *Front. Comput. Neurosci.* 10:127. doi: 10.3389/fncom.2016.00127

Wright, J. J., and Bourke, P. D. (2021a). The growth of cognition: Free energy minimization and the embryogenesis of cortical computation. *Phys. Life Rev.* 36, 83–99. doi: 10.1016/j.plrev.2020.05.004

Wright, J. J., and Bourke, P. D. (2021b). Combining inter-areal, mesoscopic and neurodynamical models of cortical function: Response to commentary on The growth of cognition: Free energy minimization and the embryogenesis of cortical computation. *Phys. Life Rev.* 39, 88–95. doi: 10.1016/j.plrev.2021.07.004

Wright, J. J., Bourke, P. D., and Favorov, O. V. (2014). Mobius-strip-like columnar functional connections are revealed in somato-sensory receptive field centroids. *Front. Neuroanat.* 8:119. doi: 10.3389/fnana.2014.00119

Wright, J. J., and Liley, D. T. J. (1996). Dynamics of the brain at global and microscopic scales: Neural networks and the EEG. *Behav. Brain Sci.* 19, 285–295.

Yu, H., Farley, B. J., Jin, D. Z., and Sur, M. (2005). The coordinated mapping of visual space and response features in visual cortex. *Neuron* 47, 267–280.

Zhang, J. X., Rosenberg, A., Mallik, A. K., Husson, T. R., and Issa, N. P. (2007). The representation of complex images in spatial frequency domains of primary visual cortex. *J. Neurosci.* 27, 9310–9318.

frontiers | Frontiers in Computational Neuroscience

TYPE Original Research
PUBLISHED 13 January 2023
DOI 10.3389/fncom.2022.1060757

OPEN ACCESS

EDITED BY
Si Wu,
Peking University, China

REVIEWED BY
Guido Marco Cicchini,
National Research Council (CNR), Italy
Yuanyuan Mi,
Chongqing University, China

*CORRESPONDENCE
Ning Qian
✉ nq6@columbia.edu

RECEIVED 03 October 2022
ACCEPTED 21 December 2022
PUBLISHED 13 January 2023

CITATION
Qian N, Goldberg ME and Zhang M (2023) Tuning curves vs. population responses, and perceptual consequences of receptive-field remapping.
Front. Comput. Neurosci. 16:1060757.
doi: 10.3389/fncom.2022.1060757

Tuning curves vs. population responses, and perceptual consequences of receptive-field remapping

Ning Qian[1,2]*, Michael E. Goldberg[1,3] and Mingsha Zhang[4]

[1]Department of Neuroscience and Zuckerman Institute, Columbia University, New York, NY, United States, [2]Department of Physiology and Cellular Biophysics, Columbia University, New York, NY, United States, [3]Departments of Neurology, Psychiatry, and Ophthalmology, Columbia University, New York, NY, United States, [4]State Key Laboratory of Cognitive Neuroscience and Learning, IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China

Sensory processing is often studied by examining how a given neuron responds to a parameterized set of stimuli (tuning curve) or how a given stimulus evokes responses from a parameterized set of neurons (population response). Although tuning curves and the corresponding population responses contain the same information, they can have different properties. These differences are known to be important because the perception of a stimulus should be decoded from its population response, not from any single tuning curve. The differences are less studied in the spatial domain where a cell's spatial tuning curve is simply its receptive field (RF) profile. Here, we focus on evaluating the common belief that perrisaccadic forward and convergent RF shifts lead to forward (translational) and convergent (compressive) perceptual mislocalization, respectively, and investigate the effects of three related factors: decoders' awareness of RF shifts, changes of cells' covering density near attentional locus (the saccade target), and attentional response modulation. We find that RF shifts *alone* produce either no shift or an opposite shift of the population responses depending on whether or not decoders are aware of the RF shifts. Thus, forward RF shifts do not predict forward mislocalization. However, convergent RF shifts change cells' covering density for aware decoders (but not for unaware decoders) which may predict convergent mislocalization. Finally, attentional modulation adds a convergent component to population responses for stimuli near the target. We simulate the combined effects of these factors and discuss the results with extant mislocalization data. We speculate that perisaccadic mislocalization might be the flash-lag effect unrelated to perisaccadic RF remapping but to resolve the issue, one has to address the question of whether or not perceptual decoders are aware of RF shifts.

KEYWORDS

predictive remapping, forward expansion, LIP, FEF, transsaccadic visual stability, corollary discharge, space perception

Introduction

Tuning curves and population responses are among the most useful concepts in sensory studies. Consider, for example, a neuron with preferred orientation x_p, and write its response to stimulus orientation x_s as $f(x_p, x_s)$. If we fix the preferred orientation x_p of a neuron and plot response f as a function of a range of stimulus orientations x_s, we obtain a tuning curve. On the other hand, if we fix the stimulus orientation x_s and plot f as a function of a set of cells' preferred orientations x_p, we obtain a population response. Thus, a collection of tuning curves of different cells and the corresponding collection of population responses for different stimuli contain the same information; they just slice the same function $f(x_p, x_s)$ along the different axes of the independent variables.

Despite their close relationship, tuning curves and population responses can be different in important ways. For example, it is known that when tuning curves shift in one direction, the corresponding population responses shift in the opposite direction (Gilbert and Wiesel, 1990; Suzuki and Cavanagh, 1997; Yao and Dan, 2001; Teich and Qian, 2003, 2010) (This is under the assumption that the decoders are unaware of the tuning shifts, a point we will elaborate below.). In the domain of stereovision, binocular phase-shifts and position-shifts between cells' RFs in the two eyes produce similarly unreliable disparity tuning curves, but the former generate more reliable population responses than do the latter (Chen and Qian, 2004; Tsang and Shi, 2004; Li and Qian, 2015). The differences between tuning curves and population responses are particularly important when perception is studied. Our perception of a stimulus must depend on relevant cells' population responses to that stimulus, instead on any single cell's responses to different stimuli (tuning curve). If a condition or manipulation changes population responses and tuning curves differently, then one must use population responses, not tuning curves, to predict the perceptual consequences.

In the spatial domain, cells' spatial tuning curves are simply their RF profiles. Around saccade onset, two types of RF changes, known as forward and convergent remapping, have been found in lateral intraparietal area (LIP), frontal eye fields (FEF), and other brain areas (Duhamel et al., 1992; Umeno and Goldberg, 1997; Kusunoki and Goldberg, 2003; Zirnsak et al., 2014; Neupane et al., 2016; Wang et al., 2016). Forward remapping is the shift of a cell's perisaccadic RF (pRF) from current (pre-saccadic) RF (cRF) toward its future (post-saccadic) RF (fRF) in the direction of the pending saccade (Figures 1A, B) whereas convergent remapping is the pRF shift toward the saccade target (Figure 1C). Further studies suggest that the two types of remapping originate from corollary discharge (CD) of saccade commands and attention at the target, respectively (Sommer and Wurtz, 2006; Neupane et al., 2016; Yang et al., 2019). Here we focus on the perceptual consequences, instead of the origins, of the remapping. There are also two types of perisaccadic perceptual mislocalization reported in the literature: forward (translational) in the direction of the pending saccade and convergent (compressive) toward the saccade target (Matin and Pearce, 1965; Honda, 1991; Ross et al., 1997; Lappe et al., 2000; Schlag and Schlag-Rey, 2002). Given such apparent correspondence between physiology and perception, it is often assumed that the two types of RF remapping generate the two types of perceptual mislocalization, respectively (Ross et al., 2001; Zirnsak et al., 2014). We will call this the *same-direction assumption* as it posits that RF shifts in a direction produce perceptual mislocalization in the same direction. In this paper, we evaluate this assumption in great detail. The above-mentioned differences between tuning curves and population responses should already cast some doubts on the assumption, but as we will see, the problem is further complicated by other factors including decoders' awareness of RF shifts, the RF-convergence induced change of cells' covering density for aware decoders, and attentional modulation of responses around the saccade target.

Results

We first examine the relationship between tuning curves and population responses in a simple case to develop intuition. If a stimulus attribute (orientation, direction, spatial location, etc.) can be parameterized by x, then let $f(x_p, x_s)$ represent how much a cell preferring stimulus x_p responds to input stimulus x_s. In the case of spatial RFs, x_p and x_s are two-dimensional (2D) vectors representing the preferred position and stimulus position on the retina, respectively. The following discussion holds regardless of whether x is a scalar or 2D vector. For simplicity, simulations in this paper consider only one spatial dimension. Assume

$$f(x_p, x_s) = f(x_p - x_s) \quad (1)$$

namely that the response depends only on the difference between x_p and x_s. This is the commonly assumed translational invariance. It is a good approximation if we view $f(x_p, x_s)$ as representing the average response of all cells with the same x_p, and the parameter range is limited so that, for example, we do not need to consider the difference between fovea and periphery (We will relax this assumption later.). Then the tuning curve of a cell preferring $x_p = x_o$ [i.e., $f(x_o - x_s)$ as a function of x_s] and the population response to a stimulus $x_s = x_o$ [i.e., $f(x_p - x_o)$ as a function of x_p] are exact mirror images of each other with respect to x_o. If $f(x_p - x_s)$ is even symmetric with respect to $x_p = x_s$ (as is often the case for some commonly used functions such as Gaussian), then the tuning curve of a cell preferring x_o and the population response to stimulus x_o are the same (Figure 2, left column). Perhaps for this reason, tuning curves and population responses are often viewed as the

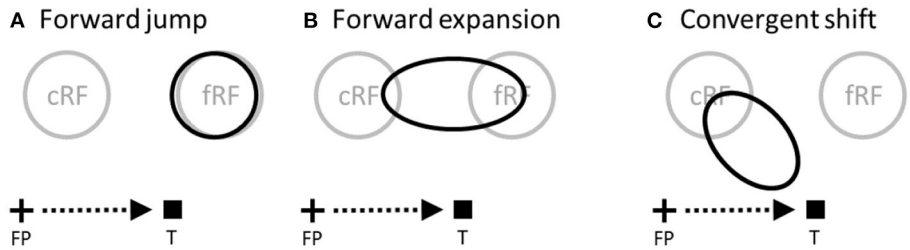


FIGURE 1
Perisaccadic RF remapping, drawn on the display screen for the stimuli. The cross, square and arrow represent the fixation point (FP), saccade target (T), and saccade vector, respectively. cRF and fRF refer to a cell's current (pre-saccadic) and future (post-saccadic) RFs, respectively. In each panel, the region(s) enclosed by black curve(s) represent perisaccadic RF (pRF). **(A)** Forward jump to fRF. **(B)** Forward expansion toward fRF. Both **(A)** and **(B)** will be referred to as forward shift. **(C)** Convergent shift toward the target.

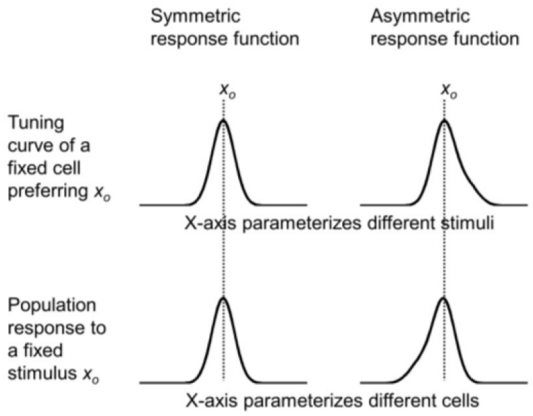


FIGURE 2
Simulations of the mirror relationship between the tuning curve of a cell preferring $x_p = x_o$ **(top row)** and the population response to a stimulus $x_s = x_o$ **(bottom row)**, under the assumption of translational invariance. The mirror relationship is hidden when a symmetric response function is used **(left column)**, but is revealed with an asymmetric function **(right column)**.

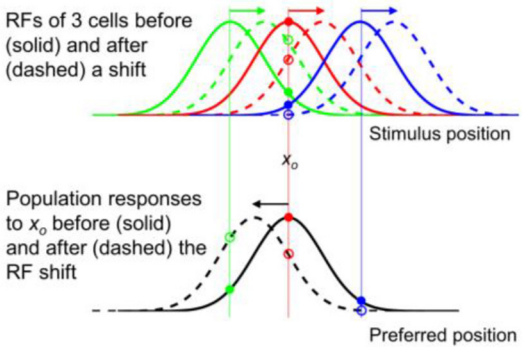


FIGURE 3
Opposite shifts of tuning curves (here RFs) and the population response for unaware decoders (after figure 6 of Yao and Dan, 2001). **(Top)** RFs of three arbitrary cells before (solid) and after (dashed) a rightward translation (rightward arrows), similar to the forward pRF jump in Figure 1A. x_o indicates a specific stimulus position which evokes responses from the cells before (filled dots) and after (open dots) the shift. **(Bottom)** The population responses of all cells to x_o before (solid black) and after (dashed black) the RF shifts, plotted here at the cells' pre-shift preferred positions (unaware decoders). The three cells' responses from the top panel are indicated. If the cells' post-shift responses are plotted at their post-shift preferred positions (aware decoders), then the pre- and post-shift population responses are identical (both solid black).

same. However, their mirror relationship becomes obvious with an asymmetric response function (Figure 2, right column).

Now consider the situation where all the tuning curves translate by an amount d. This means that the independent variable x_s in the tuning function $f(x_p - x_s)$ should be replaced by $(x_s - d)$ to produce the new function $f[x_p - (x_s - d)]$. However, since

$$f\left(x_p - (x_s - d)\right) = f\left((x_p + d) - x_s\right) \quad (2)$$

shifting tuning curves (as a function of x_s) by d is equivalent to shifting the corresponding population response (as a function of x_p) by negative d. This is an algebraic demonstration of the known result that when tuning curves shift in one direction, the corresponding population responses shift in the opposite direction (Gilbert and Wiesel, 1990; Suzuki and Cavanagh, 1997; Yao and Dan, 2001; Teich and Qian, 2003, 2010). A related algebraic demonstration appeared in the appendix of Teich and Qian (2010). The opposite shifts between tuning curves and population responses are a consequence of their mirror relationship, and can be seen regardless of whether the response function $f(x_p - x_s)$ is symmetric or not; the simulations in Figure 3 use a symmetric (Gaussian) function.

There is an implicit assumption in the above demonstration, namely that the decoder is unaware of the tuning shift so that

the post-shift population response is plotted against the cells' *pre-shift* preferred parameter x_p. If, instead, the decoder *is* aware of the tuning shift so that the post-shift population response is plotted against the cells' *post-shift* preferred parameter $x'_p = x_p + d$, then Eq. 2 becomes:

$$f\left((x_p + d) - x_s\right) = f\left(x'_p - x_s\right) \quad (3)$$

Therefore, the post-shift population response $f\left(x'_p - x_s\right)$ as a function of post-shift preferred parameter x'_p is identical to the pre-shift population response $f(x_p - x_s)$ as a function of pre-shift preferred parameter x_p. That is, for decoders aware of the tuning shift, the corresponding population responses, and hence perception, do not shift. [See Teich and Qian (2003) for related work on plotting post-adaptation population response as a function of pre- and post-adaptation preferred orientations.]

These results are simulated in Figure 3 for spatial tuning (RFs). The top panel shows RFs of three example cells (colored green, red, and blue) before (solid) and after (dashed) a rightward shift. The bottom panel shows the population response of all cells to stimulus position x_o before (solid black) and after (dashed black) the RF shift, as a function of the pre-shift preferred positions (unaware decoders). Note the mirror relationship between the dashed red curve and the dashed black curve with respect to x_o. If the cells' post-shift responses are plotted at their post-shift preferred positions (aware decoders), then the pre- and post-shift population responses will be identical (both solid black). Thus, for a tuning (RF) translation, the aware and unaware decoders should report no mislocalization and an opposite mislocalization, respectively, contradicting the same-direction assumption.

The above conclusion can be understood intuitively. Consider, for example, the "red" cell in the top panel of Figure 3 whose pre- and post-shift RFs are represented by the solid and dashed red curves, respectively. If the decoder is "unaware" of the shift, then whenever the cell fires, it is evidence that a stimulus appears at the peak position of the solid red curve. Now after the RF shift, the cell fires maximally to a stimulus at the peak position of the dashed red curve but the decoder will still view that as strong evidence for a stimulus at the peak position of the solid red curve. Thus a rightward RF shift contributes to a leftward shift of the decoded position. If, on the other hand, the decoder is "aware" of the RF shift, then a cell's firing is evidence for stimulation at the peak position of its current RF. So after the RF shift, the decoder will view the red cell's maximal firing to a stimulus at the peak position of the dashed red curve as strong evidence for a stimulus at the same position, and thus no perceptual mislocalization.

With the above basic understanding of the relationship between tuning curves and population responses, we now turn to the perceptual consequences of various types RF remapping. The simplest type is perisaccadic forward jump (Duhamel et al., 1992; Kusunoki and Goldberg, 2003): around saccade onset, cells respond to stimuli in their post-saccadic RFs (future RFs or fRFs) with a concurrent reduction of their responses to stimuli in their current RFs (cRFs). This can be approximated as a translation of perisaccadic RFs (pRFs) in the saccade direction by the amount of the saccade amplitude (Figure 1A). Therefore, the above analysis of tuning shift applies, with d equal to the saccade amplitude. We conclude that population responses (and thus perception) should show either no shift or a backward shift against the saccade direction depending on whether or not decoders are aware of the forward RF jump. The conclusion is the same even if the RF shift d is not equal to the saccade amplitude.

Next we consider perisaccadic forward RF expansion (Wang et al., 2016): a closer examination shows that LIP perisaccadic remapping is a progressive shift of a cell's cRF toward its fRF over several tens of msec. When the shifting pRF is integrated over this time window, it appears as a forward expansion covering the region between the cRF and fRF (Figure 1B). If perceptual decoders are fast enough to resolve pRFs' progressive shift over time, then the conclusion is basically a time-varying version of Figure 3: the population responses (and thus perception) should show either no shift or a progressive backward shift against the saccade direction depending on whether or not decoders are aware of the RF progression.

If, on the other hand, perceptual decoders operate at a time scale significantly longer than that of pRF progression, then they effectively integrate a cell's shifting pRF as a spatial expansion, with a center between the cRF and fRF (Figure 1B). In this case, we can express the pRFs by replacing the tuning function $f(x_p - x_s)$ by $f([x_p - (x_s - d)]/k)$ where center shift d is less than the saccade amplitude and $k > 1$ is the RF expansion factor. Since, similar to Eq. 2, we have:

$$f\left([x_p - (x_s - d)]/k\right) = f\left([(x_p + d) - x_s]/k\right) \quad (4)$$

the above conclusions for the forward RF jump hold for the forward RF expansion, the only difference being that the RF expansion factor k increases the width of the population responses for both the aware (dashed magenta) and unaware (dashed black) decoders in the simulations of Figure 4, as expected from Eq. 4. For aware decoders, the RF expansion does not change the peak location of the population responses (cf. solid black and dashed magenta). We conclude that for both types of forward RF shifts, the population response (and thus perception) shows either no shift or a backward shift, depending on whether or not decoders are aware of the RF shifts.

We can remove the translational-invariance assumption used above by considering the dependence of RF size on eccentricity. So instead of Eq. 1, we now assume:

$$f\left(x_p, x_s\right) = f\left([x_p - x_s]/[a\left|x_p\right| + 1]\right) \quad (5)$$

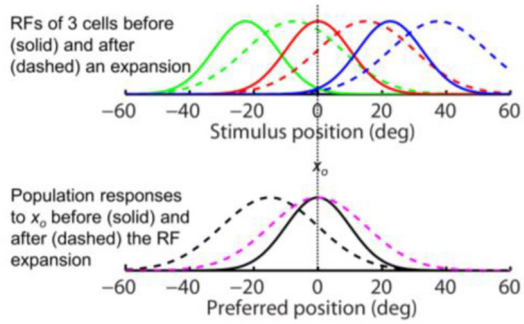


FIGURE 4
The same-direction assumption is false for forward RF expansion as is for forward RF jump in Figure 3. **(Top)** RFs of three arbitrary cells before (solid) and after (dashed) a perisaccadic rightward expansion. **(Bottom)** The population responses of all cells to $x_o = 0$ before (solid) and after (dashed) the RF expansion. The dashed black and magenta curves are the post-expansion responses plotted against the pre- and post-expansion preferred positions, for the unaware and aware decoders, respectively.

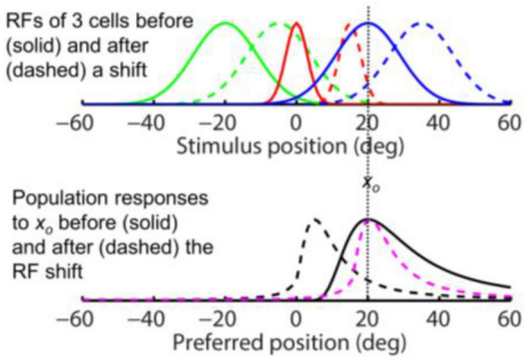


FIGURE 5
Removing translational invariance does not change the conclusion that the same-direction assumption is false for forward RF shifts. This figure assumes that a cell's shifted RF size is determined by its *pre*-shift eccentricity. The *x*-axes measure position as eccentricity from fovea (0). **(Top)** RFs of three arbitrary cells before (solid) and after (dashed) a perisaccadic rightward jump. **(Bottom)** the population responses of all cells to a stimulus at $x_o = 20$ deg before (solid) and after (dashed) the RF shift. The dashed black and magenta curves are the post-shift responses plotted against the pre- and post-shift preferred positions, for the unaware and aware decoders, respectively.

where position vectors x's are all measured from the fovea as the origin, and $a > 0$ is a constant. [$a = 0$ would reduce Eq. 5 to Eq. 1.] To understand Eq. 5, first note that for a cell with preferred position vector x_p from the fovea, its eccentricity is given by the norm $|x_p|$. Thus, the factor $(a|x_p| + 1) \geq 1$ simply scales up the RF size with its distance $|x_p|$ from the fovea, and f without the scaling (the factor equals 1) determines the RF size at the fovea where $|x_p| = 0$.

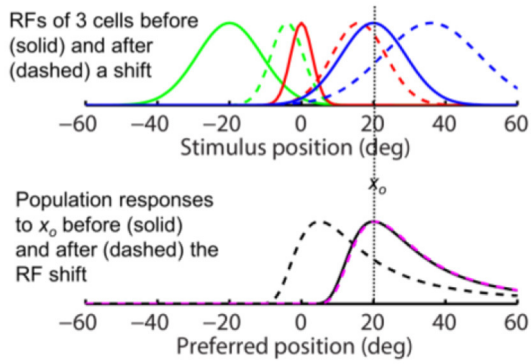


FIGURE 6
Removing translational invariance does not change the conclusion that the same-direction assumption is false for forward RF shifts. This figure assumes that a cell's shifted RF size is determined by its *post*-shift eccentricity. The format is identical to that for Figure 5.

Now consider the forward RF jump in the context of eccentricity dependence of RF size (the forward expansion case can be similarly treated). When a cell's RF shifts to a new position by d, its RF size may be determined by either the pre-shift eccentricity $|x_p|$ or the post-shift eccentricity $|x_p+d|$. Since we do not know which case is true, we consider both. We can represent the post-shift response function as $f([x_p - (x_s - d)]/[a|x_p| + 1])$ and $f([x_p - (x_s - d)]/[a|x_p + d| + 1])$, for the two cases respectively. Since we have

$$f\left([x_p - (x_s - d)]/[a\,|x_p| + 1]\right) = f([(x_p + d) - x_s]/[a|x_p| + 1]) \quad (6)$$

and

$$f\left([x_p - (x_s - d)]/\,[a\,|x_p + d| + 1]\right) = f([(x_p + d) - x_s]/[a\,|x_p + d| + 1]) \quad (7)$$

for the two cases, the above conclusions on the differences between RF shifts and population-response shifts remain valid. This is confirmed by simulations in Figures 5, 6 for the two cases, respectively. In the top panel of Figure 5, since a cell's shifted RF size is determined by its *pre*-shift eccentricity, its size does not change with the shift (the dashed and solid RF curves of the same color have the same width). In the top panel of Figure 6, in contrast, a cell's shifted RF size is determined by its *post*-shift eccentricity. For example, the "green" cell shifts to a smaller eccentricity and thus has a smaller post-shift size (the dashed green curve has a smaller width than the solid green curve). In both cases, the RF shifts produce either no shift or an opposite shift of the population response, depending on whether

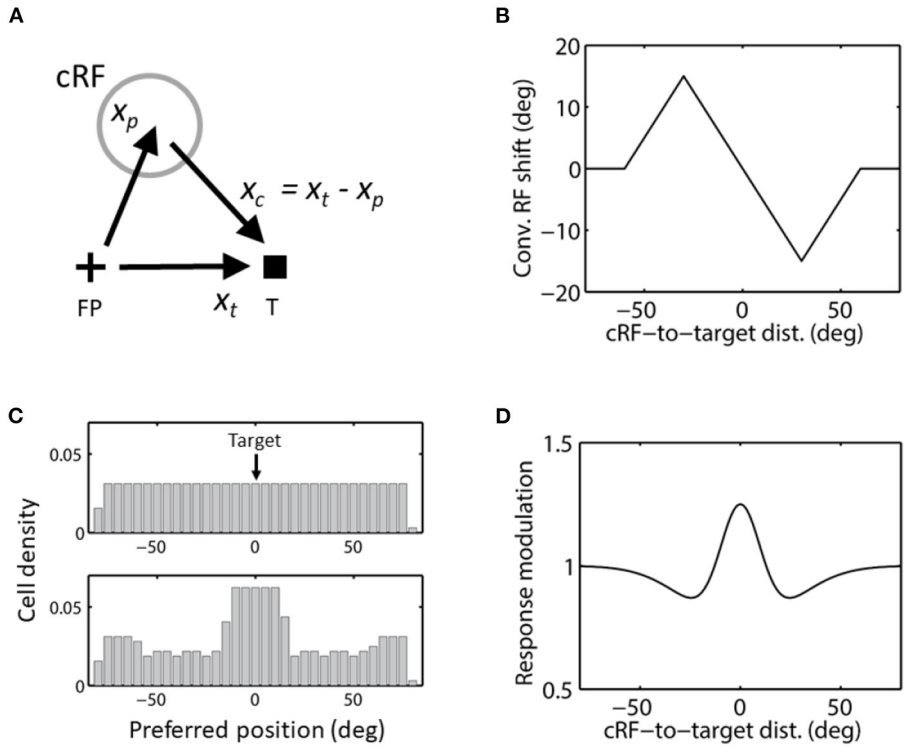


FIGURE 7
Convergent RF shift. **(A)** The direction of convergent RF shift. FP and T indicate the initial fixation point and saccade target, respectively. **(B)** Convergent RF shift as a function of the cRF-to-target distance used in the simulations of Figures 8–10. cRFs on the left and right side of the target (at 0) shift to the right (positive) and left (negative), respectively, as they converge to the target. **(C)** The cells' covering density before (top) and after (bottom) the convergent RF shifts toward the target in c, for aware decoders. The density stays the same (top) for unaware decoders. **(D)** The center/surround attentional modulation as a function of the cRF-to-target distance, used in the simulations of Figure 9. This curve is scaled by a factor of 4 in the simulations of Figure 10.

or not decoders are aware of the RF shifts. Thus, removing translational invariance does not change the conclusion that the same direction assumption is false.

Note that in Figures 5, 6, even though the RFs are symmetric (with respect to the preferred positions), the population responses can be asymmetric because of the eccentricity dependence of the RF sizes. Also note that for decoders aware of the RF shift, the pre- and post-shift population responses are identical in Figure 6 (solid black and dashed magenta) but not in Figure 5. This can be understood *via* the corresponding Eqs 7 and 6. When Eq. 7 is plotted as a function of the post-shift preferred position $x'_p = x_p + d$, it is identical to the pre-shift response $f([x_p - x_s]/[a|x_p| + 1])$ as a function of the pre-shift preferred position x_p. This is not true for Eq. 6.

We finally consider convergent RF remapping (Figure 1C). Let x_t represent the target position vector, then for a cell whose cRF prefers position vector x_p, the convergent shift is in the direction of the vector (Figure 7A):

$$x_c = x_t - x_p \quad (8)$$

Since the convergence is likely due to the attention at the target (Connor et al., 1996; Zirnsak et al., 2014; Neupane et al., 2016; Wang et al., 2019; Yang et al., 2019), the magnitude of the convergence must depend on the cRF-to-target distance $|x_c|$. Thus the convergence shift vector can be represented by $c(|x_c|)\ x_c$, where the function $c(|x_c|)$ satisfies $0 \leq c(|x_c|) \leq 1$ to ensure that the shift is between the cRF and target. For simplicity, we assume translational invariance (Eq. 1) before the shift. Then the response function after the shift can be obtained by replacing $f(x_p - x_s)$ by $f(x_p - [x_s - c(|x_c|)\ x_c])$. Since

$$f(x_p - [x_s - c(|x_c|)\ x_c]) = f([x_p + c(|x_c|)x_c\ ] - x_s) \quad (9)$$

we see once again the familiar pattern: for decoders unaware of the RF convergence, the population response shifts in the opposite, divergent directions, away from the target. If, on the other hand, decoders are aware of the RF convergence, the population response plotted as a function of the new preferred position $x'_p = x_p + c(|x_c|)x_c$ is $f(x'_p - x_s)$, identical to the pre-shift population response $f(x_p - x_s)$ as a function of the pre-shift preferred position x_p.

Before we draw conclusions on perceptual consequences of convergent RF shifts, we need to consider a new factor: for aware decoders, convergent RF shifts change the density distribution of cells covering different positions. First note that this is not an issue for unware decoders which, by definition, do not "know" the RF shifts and always attribute a cell's response to its original (pre-shift) preferred position. As such, from the perspective of unware decoders, there is no change of cells' preferred positions and thus no change of the cells' covering density. In contrast, aware decoders attribute a cell's response to its new (post-shift) preferred position, and from their perspective, convergent RF shifts toward the target must change the cells' covering density. Also note that in the above discussions of the *forward* RF shifts, a uniform translation of RFs (with or without an expansion) in the saccade direction does not change the cells' covering density (Obviously, if future experiments find a non-uniform forward-shift pattern across space, then we will have to consider the change of cells' covering density for aware decoders.).

How a change of cells' covering density affects perceptual decoding depends on whether or not a given decoder uses the covering density. We showed above that for aware decoders, convergent RF shifts do not change the functional form of population responses. If a *specific* aware decoder uses the center-of-mass (mean) of a population response to represent perception, then the increased cell density tuned to the target must bias perception toward the target. In other words, even without any change to the shape of the population response, a decoder that take into account the changed sampling from the population response will generate a convergent (compressive) mislocalization toward the target. In contrast, if we use another specific aware decoder that identifies the peak (mode) of a population response as perception, then the cell-density change does not matter and convergent RF shifts do not produce perceptual mislocalization.

We ran simulations using a size of convergent RF shifts that first increases linearly with the cRF-to-target distance up to 30 deg and then decreases linearly to 0 up to 60 deg (Figure 7B). This is based on a circuit model for convergence RF shifts (Wang et al., 2019), and is consistent with the available data (Zirnsak et al., 2014; Yang et al., 2019). Intuitively, convergent RF shift size must be small at both small and large cRF-to-target distances, with a maximum at an intermediate distance: when the distance is small, there is not much room for the cRF to shift to the target and when the distance is large, the attentional effect at the cRF is diminished. For aware decoders, we show the cell densities covering different locations before and after the convergent shifts in Figure 7C.

Figure 8A, top panel, shows RFs of a few cells before (solid) and after (dashed) converging toward the target at 0 deg. Each RF is a Gaussian with $\sigma = 10$ deg. The "blue" cell tuned to the target location has no shift. The bottom panel shows the population responses to a stimulus at -10 deg before (solid) and after (dashed) the RF convergence. The dashed magenta and dashed black curves are population responses for aware and unaware decoders, respectively, showing no shift and a divergent shift (away from the target), respectively. We use both the center-of-mass and peak decoders to determine perceptual mislocalization for both the aware (Figure 8A, dashed magenta) and unaware (Figure 8A, dashed black) population responses, relative to the pre-shift baseline (Figure 8A, solid black). The results are shown in Figure 8B. As expected, the center-of-mass aware decoder (solid magenta) predicts convergent mislocalization: stimuli to the left and right of the target (over a range of about 40 deg) have positive and negative mislocalization, respectively. The predicted convergent mislocalization will be even stronger in two-dimensional space because the change of cells' covering density will be greater. The peak aware decoder (dashed magenta) predicts no mislocalization. The center-of-mass and peak unaware decoders (solid and dashed black) both predict divergent mislocalization. The maximum 15 deg divergent mislocalization at 15 deg distance predicted by the peak unaware decoder (dashed black in Figure 8B) can be understood: stimuli at this distance activate cells originally tuned to 30 deg distance but converged toward the target by 15 deg. Thus the unaware decoder mistakes stimuli at 15 deg as stimuli at 30 deg.

We now consider yet another new factor: attention at the target may not only produce convergent RF shifts but also modulate neuronal response strength. In fact, in a period before saccades, both LIP and FEF neurons tuned near the target location have enhanced visual responses while those tuned to locations further away from the target have suppressed visual responses (Schall et al., 1995; Falkner et al., 2010). In the above, we only considered the effect of RF shifts on population responses. We need to include the effect of the response modulation as well.

Based on the experimental data (Schall et al., 1995; Falkner et al., 2010), we consider an attentional modulation factor $g(|x_c|)$ as a function of the cRF-to-target distance $|x_c|$ (Figure 7D). To combine the effects of both the convergent RF shift and response modulation, we now replace the response function $f(x_p - x_s)$ by the product:

$$f(x_p - [x_s - c(|x_c|)\, x_c])g(|x_c|) \quad (10)$$

The $f(.)$ part is the same as before (Eq. 9). The attentional modulation factor $g(|x_c|)$ is peaked at the target $|x_c| = 0$, and is greater and less than 1, respectively, for small and large $|x_c|$, and stays at 1 (no modulation) for very large $|x_c|$. Since $g(|x_c|)$ is not a function of stimulus position x_s, the RFs as a function of x_s will just be scaled by $g(|x_c|)$ without changing their shapes (including preferred positions). On the other hand, $g(|x_c|)$ will change the shapes of the population responses as a function of x_p, because $|x_c|$ depends on x_p (Eq. 8). In particular, for a stimulus close to and away from

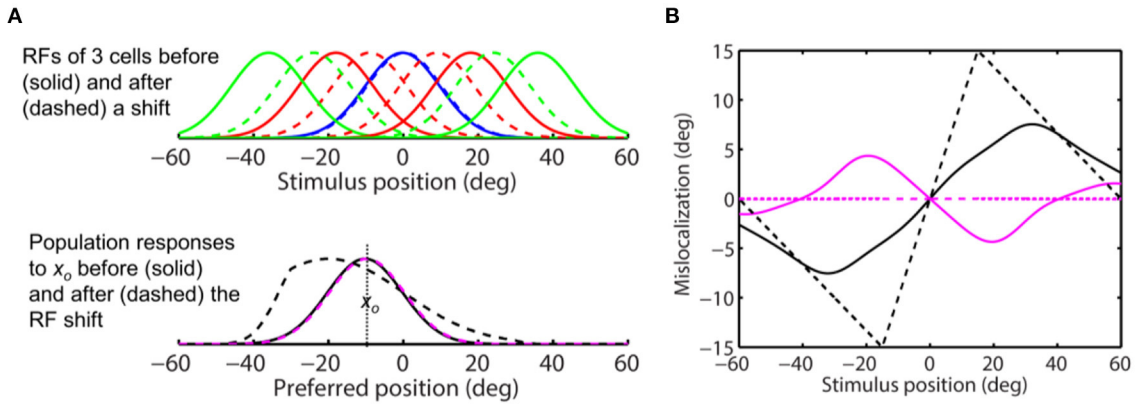


FIGURE 8
Perceptual consequence of convergent RF shifts. The target is at 0 deg. **(A)** (Top) RFs of five arbitrary cells before (solid) and after (dashed) the convergence. The "blue" cell tune to the target (0 deg) does not shift. (Bottom) the population responses of all cells to stimulus position $x_o = -10$ deg before (solid) and after (dashed) the RF convergence. The dashed black and magenta curves are the post-convergence responses plotted against the pre- and post-convergence preferred positions, for the unaware and aware decoders, respectively. **(B)** Perceptual mislocalization as a function of stimulus position relative to the target (0 deg). The results depend on whether the decoder is aware (magenta) or unaware (black) of the RF convergence, and whether the decoder uses the peak (dashed) or center-of-mass (solid) of population responses. The aware center-of-mass decoder (solid magenta) considers the change of the cells' covering density (see text); it predicts convergent mislocalization as stimuli to the left and right of the target (over a range of about 40 deg) have positive and negative mislocalization, respectively. The aware peak decoder (dashed magenta) predicts no mislocalization. The unaware center-of-mass and peak decoders (solid and dashed black) both predict divergent mislocalization.

the target, its population response will be "pulled" toward and "pushed" away from the target by $g(|x_c|)$, respectively, compared with the no-modulation case, generating convergent and divergent mislocalization, respectively. Since the center excitation of attention is stronger than surround inhibition (Figure 7D), the main effect is convergent mislocalization for stimuli near the target.

Thus for stimuli close to the target, attentional modulation introduces a convergent component to population responses, increasing the convergent (compressive) mislocalization predicted by aware decoders. If the modulation is extremely strong, it may even make population responses of *unaware* decoders converge toward the target.

To get a better sense of all the effects together, we ran simulations using a difference of Gaussians as the modulation factor:

$$g(|x_c|) = 1 + s\,(\exp[-|x_c|^2/(2\sigma_e^2)] - b\,\exp[-|x_c|^2/(2\sigma_i^2)]) \quad (11)$$

where σ_e, σ_i, and b determine the shape of the function, and s scales the function to determine the modulation strength. In Figure 7D, we let $\sigma_e = 10$ deg, $\sigma_i = 25$ deg, and $b = 0.5$ to produce a $g(|x_c|)$ similar in shape to the measured one in LIP (Falkner et al., 2010), and $s = 0.5$ so that the maximum attentional enhancement of responses (at the target) is 25% (Bushnell et al., 1981; Goldberg and Bushnell, 1981; Maunsell, 2015). We ran simulations with this modulation factor and the same convergent RF shift pattern as in Figure 8. The results are shown in Figure 9. As expected, the response modulation generates a convergent shift of the aware population response (Figure 9A, dashed magenta) and reduces the divergent shift of the unaware population response (dashed black) although the effects are relatively small. We again applied the center-of-mass and peak decoders to calculate perceptual mislocalization as a function of stimulus position (Figure 9B). Now for stimuli close to the target, both the center-of-mass and peak aware decoders (solid and dashed magenta), and the peak unaware decoder (dashed black), predict convergent mislocalization. The center-of-mass unaware decoder (solid black) still predicts divergent mislocalization. For stimuli further away from the target, the peak unaware decoder (dashed black) also predict divergent mislocalization.

We then repeated the simulation with $s = 2$ in Eq. 11 so that the peak attentional enhancement of responses (at the target) is 100%. The results in Figure 10A show sizeable convergent shifts for both the aware and unaware population responses (bottom panel, dashed magenta and black). With such large response modulation, all four decoders predict convergent mislocalization for stimuli close to the target (Figure 10B). This is consistent with a previous model (Hamker et al., 2008) which appeared to use an even larger response increase at the target location (300% in the first layer of the model) to generate convergent mislocalization. However, unaware decoders (solid and dashed black) still predict divergent mislocalization for stimuli away from the target.

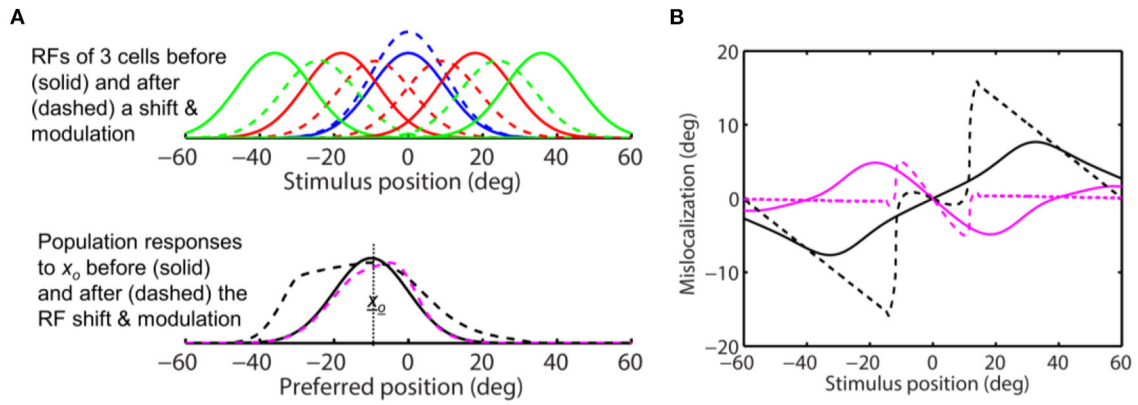


FIGURE 9
Perceptual consequence of convergent RF shifts and response modulation. The target is at 0 deg. The maximum response enhancement (at the target) is 25%. The format is identical to that of Figure 8. For stimuli close to the target, the aware center-of-mass and peak decoders (solid and dashed magenta) and the unaware peak decoder (dashed black) all predict convergent mislocalization. The unaware center-of-mass decoder predicts divergent mislocalization. The unaware peak decoder also predicts divergent mislocalization for stimuli away from the target.

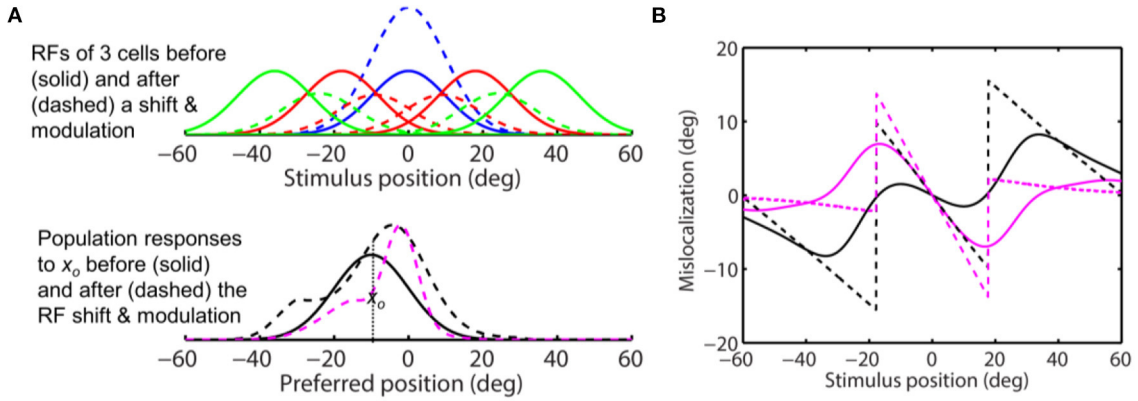


FIGURE 10
Perceptual consequence of convergent RF shifts and response modulation. The target is at 0 deg. The maximum response enhancement (at the target) is 100%. The format is identical to that of Figures 8, 9. For stimuli near the target, all the four decoders predict convergent mislocalization. Unaware decoders (solid and dashed black) still predict divergent mislocalization for stimuli away from the target.

We conclude that convergent RF shifts may predict convergent, divergent, or no mislocalization, depending on multiple factors including whether or not decoders are aware of the RF shift, whether or not aware decoders take cells' covering density into account, the strength of attentional modulation, and the stimulus-to-target distance. The same-direction hypothesis is correct for aware decoders that consider cells' covering density. It is also likely to be correct for aware decoders when stimuli are near the target and attentional modulation is present. Unaware decoders tend to predict a divergent mislocalization unless attentional modulation is extremely strong and stimuli are near the target.

Discussions

In this paper, we analyzed differences between tuning curves and the corresponding population responses, and applied the analysis to determine how population responses change with two types of spatial tuning-curve shifts, namely the forward and convergent RF remapping (Duhamel et al., 1992; Umeno and Goldberg, 1997; Kusunoki and Goldberg, 2003; Zirnsak et al., 2014; Wang et al., 2016; Yang et al., 2019). We found that in the absence of response modulation, forward/convergent RF shifts alone produce either no shift or backward/divergent shifts of the population responses, depending on whether or not decoders are aware of the RF shifts. Since forward RF shifts, whether

the forward jump or expansion, are assumed to be a uniform translation which does not change the density of cells covering different locations, the perceptual consequence simply follows the population responses: there should be no mislocalization and backward mislocalization for the aware and unaware decoders, respectively. This conclusion holds even when the increasing RF size with eccentricity, which breaks translational invariance, is considered.

The perceptual consequence of convergent RF shifts, however, is more complicated. For unaware decoders, the perception also simply follows the population response which predicts a divergent mislocalization. In contrast, for aware decoders, convergent RF shifts increase the density of cells covering the target area. Even though convergent RF shifts do not change the shape of the population response for aware decoders, perceptual decoding of the population response depends on whether a specific decoder takes into account the change of the cells' covering density (density-sensitive) or not (density-insensitive). As examples, we considered the center-of-mass aware decoder and the peak aware decoder. The former is density sensitive and predicts a convergent mislocalization whereas the latter is density insensitive and predicts no mislocalization. Thus, without response modulation, the same-direction assumption is correct for convergent RF shifts *only* when aware and density-sensitive decoders are used (Figure 8B, solid magenta). In all other cases, the same-direction assumption is false.

We then included the effect of attentional modulation which enhances responses at and near the saccade target (the attentional locus), and suppresses responses away from the target (Schall et al., 1995; Falkner et al., 2010). The effect of this modulation *alone* is straightforward: it produces convergent and divergent mislocalization for stimuli near and away from the target, respectively. When attentional modulation and RF shifts are combined, they predict a variety of mislocalization patterns, as demonstrated by our simulations, which depend on the parameters of the response modulation and RF shifts (including the strengths and ranges), decoders' awareness of the RF shifts, aware decoders' sensitivity to cells' covering density, and the stimulus-to-target distance. Using physiologically plausible parameters for convergent RF shifts and response modulation in Figures 7B, C, we found that aware and unaware decoders predict mostly convergent and divergent mislocalization, respectively (Figure 9B). Not surprisingly, when the attentional modulation is extremely strong, all decoders predict convergent mislocalization for stimuli near the target but unaware decoders still predict divergent mislocalization for stimuli away from the target (Figure 10B). Most studies in the literature seem to assume (often implicitly) unaware decoders. Since for unaware decoders convergent RF shifts do not change cells' covering density (Figure 7C), convergent mislocalization can occur only through strong attentional response modulation and only for stimuli near the attentional locus.

Although we only simulated the combined effect of attentional modulation and convergent RF shifts, it is easy to image the combined effect of attentional modulation and forward RF shifts. The forward RF shifts alone generate either no mislocalization (aware decoders) or backward mislocalization (unaware decoders). Attentional modulation, through its center excitation and surround inhibition, will add a convergent and divergent mislocalization component for stimuli close to and away from the target (or any attentional locus), respectively.

We now briefly summarize psychophysical data of perisaccadic perceptual mislocalization, which has been interpreted as reflecting imperfections of the mechanisms for transsaccadic visual stability (TSVS) (Matin and Pearce, 1965; Honda, 1991), and after the discovery of perisaccadic RF shifts (a specific mechanism for TSVS), as a perceptual consequence of the RF shifts (Ross et al., 2001; Kaiser and Lappe, 2004; Zirnsak et al., 2014). In a typical experiment, a probe stimulus (e.g., a dot or line) is flashed at various times around saccade onset, and subjects report the remembered stimulus location after the saccade. Over a window of about 150 ms around saccade onset, the probe is mislocalized with a translational component along the saccade axis and a compressive component toward the saccade target (Honda, 1991; Ross et al., 1997). The translational and compressive components are stronger, respectively, in the absence and presence of a post-saccadic visual reference, such as a ruler (Lappe et al., 2000). The translational mislocalization is in the saccade direction (forward) around the saccade onset, and disappears, or reverses the direction (backward), around the saccade offset (Honda, 1991; Lappe et al., 2000; Schlag and Schlag-Rey, 2002).

As we mentioned in the Introduction, the commonly held same-direction assumption posits that the forward and convergent RF shifts are responsible for the forward (translational) and convergent (compressive) perceptual mislocalization, respectively. Our analysis and simulations, however, cast some doubts on this assumption. First, the forward RF shifts predict either no mislocalization (aware decoders) or backward mislocalization (unaware decoders). Since forward RF shifts occur around the saccade onset, they cannot explain the observed forward mislocalization in that period. An exception to the forward mislocalization around the saccade onset is the study of Jeffries et al. (2007) who found backward mislocalization in monkeys. This is consistent with unaware decoders' prediction. However, that study differed from others in one aspect (in addition to monkey vs. human subjects): visual feedback of the veridical stimulus position was provided at the end of each trial. Further studies are needed to sort out the impact, if any, of this difference. Second, with reasonable parameters and for stimuli near the target, convergent RF shifts and attentional modulation together predict convergent mislocalization *only* for aware decoders (Convergent RF shifts *alone* predict convergent mislocalization only for aware *and* density-sensitive decoders.). Unaware decoders predict little

or divergent mislocalization. Even with 100% attentional enhancement at the target, unaware decoders still predict divergent mislocalization for stimuli further away from the target. Without independent information on the brain's choice of aware vs. unaware decoders, we cannot determine whether or not convergent RF shifts explain convergent mislocalization.

There are other reasons to doubt the same-direction assumption. Convergent mislocalization depends on a post-saccadic visual reference such as a ruler or any visible background (Lappe et al., 2000). It is unknown whether or not convergent RF shifts depend on such reference but since convergent shifts in FEF and LIP appear around saccade onset or even in the delay period well before saccades (Zirnsak et al., 2014; Yang et al., 2019), they are unlikely to be dependent on post-saccadic references. A study showed that perceptual compression can occur without a post-saccadic reference if the stimuli are weak with near threshold luminance and observers dark adapt (Georg et al., 2008). However, convergent RF shifts are measured with supra-threshold stimuli (as weak stimuli would evoke too few spikes to measure the shifts efficiently). Moreover, forward remapping has been measured with both briefly flashed stimuli (Wang et al., 2016) and stimuli persisting to the end of trials (Duhamel et al., 1992). In contrast, mislocalization measurements appear to require brief stimuli (more on this later). Additionally, it is unclear how forward RF shifts may explain both the forward mislocalization at saccade onset *and* the backward mislocalization at saccade offset. Finally, perisaccadic RF remapping is regarded as a physiological mechanism for TSVS. Perisaccadic mislocalization, in contrast, is visual distortion or instability. The same-direction assumption has the conceptual difficulty of asserting that the stability mechanism directly causes instability. It would be more reasonable to assert that imperfect aspects of RF remapping generates residual instability. For example, forward RF remapping may start too early, before the saccade onset (Duhamel et al., 1992). However, this would simply predict an early start of either backward mislocalization (unaware decoders) or no mislocalization (aware decoders), still contradicting the same-direction assumption and the observed forward mislocalization before the saccade onset.

If the same-direction assumption is at least questionable, what then could be responsible for transsaccadic perceptual mislocalization? Traditional models assume that translational mislocalization results from the brain's poor estimate of eye position used to compensate for saccade-induced retinal shifts of stimuli (Matin and Pearce, 1965; Honda, 1991). Specifically, a slow-changing estimate that first leads but then lags the actual eye position during a saccade explains the forward and backward mislocalization around the saccade onset and offset, respectively. Pola (2004) argues that when latency and persistence of visual responses to flashed stimuli are considered, a delayed but otherwise veridical eye-position estimate can account for the translational mislocalization. Interestingly, Teichert et al. (2010) show that when temporal characteristics of visual responses to different stimuli are considered, the eye-position estimate that eliminates mislocalization for *persistent* stimuli produces the observed translational mislocalization for *flashed* stimuli. However, these models do not explain convergent mislocalization. Note that these models focus on eye-position estimates whereas our study focuses on RF remapping. Eye-position estimates rely on extraretinal signals but may also be influenced by retinal stimuli (Teichert et al., 2010). Conversely, RFs process retinal inputs but their remapping depends on extraretinal signals such as CD (Sommer and Wurtz, 2006). So the two approaches are not completely independent and may be combined in future research. Also note that temporal characteristics of visual responses to stimuli (such as latency and persistence) do not change our analysis on the perceptual consequences of RF remapping as long as the stimuli are timed to produce the remapping (such as the perisaccadic stimuli we consider). Similarly, early or late decoding does not change our results as long as the decoders act on remapped RFs. Clearly, our results are irrelevant if a stimulus does not produce RF remapping or perceptual decoders act on RFs at a time without remapping.

Cicchini et al. (2013) found that a perisaccadically presented bar is attracted to another bar presented either pre- or post-saccadically. The interaction occurs over an oriented region of spatial and temporal separations between the bars, characteristic of motion detectors (Adelson and Bergen, 1985). Interestingly, the forward RF expansion can also be interpreted as a spatiotemporal orientation because around saccade onset, stimuli closer to a cell's cRF and fRF evoke visual responses with shorter and longer latencies, respectively (Wang et al., 2016). Thus, forward expanded RFs may be viewed as CD-enabled high-speed motion detectors that measure spatiotemporal correlation in retinal image motion across saccades, not for perceiving the motion, but for linking pre- and post-saccadic retinal images to achieve TSVS. Perisaccadic mislocalization could then occur if this motion detection is imperfect (Cicchini et al., 2013). However, this possibility again cannot explain convergent mislocalization because saccade-induced retinal motion is largely uniform across the retina at a given time.

Surprisingly, patterns similar to perisaccadic mislocalization, with both the translational and convergent components, have been produced by simulating saccade-like retinal motion without the actual saccade (Ostendorf et al., 2006; Shim and Cavanagh, 2006). Ostendorf et al. argue that earlier experiments that failed to find convergent mislocalization with simulated motion either did not match simulated motion and saccade-induced motion, or were not designed to measure compression. Just like perisaccadic mislocalization which starts before the saccade onset, the motion induced mislocalization starts before the motion

onset. Such motion induced mislocalization of flashed stimuli in the absence of eye movements is known as the flash-lag effect (Brenner et al., 2006; Watanabe and Yokoi, 2006). Although the mechanism of the flash-lag effect itself is still debated (Nijhawan, 2002; Khoei et al., 2017), it likely involves uncertainty and delays in processing brief stimuli whose noisy memory representations interact with other visual references. Indeed, most demonstrations of both the flash-lag and perisaccadic mislocalization effects use flashes of less than 10 ms. The flash-lag effect greatly decreases or largely disappears for flash durations longer than 100 ms (Lappe and Krekelberg, 1998; Cantor and Schor, 2007). This makes sense because longer stimuli produce more reliable neural representations which are less vulnerable in memory. Similarly, although perisaccadic mislocalization of brief stimuli can be many degrees of visual angle, we never notice it in our daily life suggesting that it may not exist for persistent stimuli (Teichert et al., 2010). [Alternatively, one could argue that saccades suppress persistent objects in daily life much more strongly than brief stimuli in mislocalization studies. This, however, contradicts the fact that saccadic suppression is stronger for magnocellular stimuli such as flashes (Ross et al., 2001)]. Thus, perisaccadic mislocalization might be a version of the flash-lag effect (Teichert et al., 2010) unrelated to saccades per se or mechanisms of TSVS.

Since saccades generate retinal motion which then produces mislocalization by itself, how can we determine perceptual consequences of RF shifts without the confound of the saccade-induced retinal motion? Total darkness would eliminate retinal motion but it is hard to achieve when initial fixation points, targets, and probes have to be visible. Fortunately, convergent RF shifts can be generated by attention in time periods well separated from saccades (Neupane et al., 2016; Yang et al., 2019). One can thus measure attention induced mislocalization without saccades and the associated retinal motion. Suzuki and Cavanagh (1997) did exactly such a study. They used a Vernier task to measure attentional mislocalization, and found that probe stimuli were repelled away from, not attracted toward, the attentional locus. This result contradicts the same-direction assumption, and is consistent with unaware decoders' prediction of divergent mislocalization (Figures 8B, 9B, black curves). Indeed, they discussed a model which uses unaware decoders without attentional modulation. However, the maximum mislocalization they measured was only about 0.3 deg, much smaller than typical perisaccadic mislocalization at similar stimuli-to-target distance. Perhaps the model could accommodate the small mislocalization by using the fact that convergent RF remapping is relatively weak (Neupane et al., 2016) and only a small fraction of cells show convergent RF shifts (Yang et al., 2019). Another issue is that since they measured mislocalization by comparing two flashed Vernier lines, the result must be the difference between the two lines' mislocalizations (which may also contribute to the small effect). To explain the data, the model has to assume that attention repels the near line more than the far line, or attract the near line less than the far line, in the experiment. This is possible (Figure 9B) but requires independent verification.

In conclusion, our work suggests that there is no strong theoretical support for the commonly-held same-direction assumption, which links perisaccadic forward and convergent RF shifts to perisaccadic forward (translational) and convergent (compressive) perceptual mislocalization. Perisaccadic mislocalization might be a version of the flash-lag effect caused by saccade-induced retinal motion instead of by saccades per se or by mechanisms of TSVS. However, although forward RF shifts cannot explain forward mislocalization, convergent RF shifts, together with attentional response modulation, may contribute to convergent mislocalization particularly for aware decoders. To resolve this issue, one has to address the key open question of whether or not perceptual decoders in the brain are aware of the RF shifts.

Data availability statement

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author/s.

Author contributions

NQ, MZ, and MG conceived and discussed the project and edited the manuscript. NQ did the simulations and wrote the first draft of the manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

We thank Dr. Vince Ferrera for helpful discussions. Supported by NIH (R01 EY032938), NSF (1754211), and National Natural Science Foundation of China (32030045 and 32061143004).

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

References

Adelson, E. H., and Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* 2, 284–299. doi: 10.1364/JOSAA.2.000284

Brenner, E., van Beers, R. J., Rotman, G., and Smeets, J. B. (2006). The role of uncertainty in the systematic spatial mislocalization of moving objects. *J. Exp. Psychol. Hum. Percept. Perform*. 32, 811. doi: 10.1037/0096-1523.32.4.811

Bushnell, M. C., Goldberg, M. E., and Robinson, D. L. (1981). Behavioral enhancement of visual responses in monkey cerebral cortex. I. Modulation in posterior parietal cortex related to selective visual attention. *J. Neurophysiol.* 46, 755–772. doi: 10.1152/jn.1981.46.4.755

Cantor, C. R., and Schor, C. M. (2007). Stimulus dependence of the flash-lag effect. *Vision Res.* 47, 2841–2854. doi: 10.1016/j.visres.2007.06.023

Chen, Y., and Qian, N. (2004). A coarse-to-fine disparity energy model with both phase-shift and position-shift receptive field mechanisms. *Neural Comput.* 16, 1545–1577. doi: 10.1162/089976604774201596

Cicchini, G. M., Binda, P., Burr, D. C., and Morrone, M. C. (2013). Transient spatiotopic integration across saccadic eye movements mediates visual stability. *J. Neurophysiol.* 109, 1117–1125. doi: 10.1152/jn.00478.2012

Connor, C. E., Gallant, J. L., Preddie, D. C., and Van Essen, D. C. (1996). Responses in area V4 depend on the spatial relationship between stimulus and attention. *J. Neurophysiol.* 75, 1306–1308. doi: 10.1152/jn.1996.75.3.1306

Duhamel, J. R., Colby, C. L., and Goldberg, M. E. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* 255, 90–92. doi: 10.1126/science.1553535

Falkner, A. L., Krishna, B. S., and Goldberg, M. E. (2010). Surround suppression sharpens the priority map in the lateral intraparietal area. *J. Neurosci.* 30, 12787–12797. doi: 10.1523/JNEUROSCI.2327-10.2010

Georg, K., Hamker, F. H., and Lappe, M. (2008). Influence of adaptation state and stimulus luminance on peri-saccadic localization. *J Vision* 8, 15–15. doi: 10.1167/8.1.15

Gilbert, C. D., and Wiesel, T. N. (1990). The influence of contextual stimuli on the orientation selectivity of cells in primary visual cortex of the cat. *Vision Res.* 30, 1689–1701. doi: 10.1016/0042-6989(90)90153-C

Goldberg, M. E., and Bushnell, M. C. (1981). Behavioral enhancement of visual responses in monkey cerebral cortex. II. Modulation in frontal eye fields specifically related to saccades. *J. Neurophysiol.* 46, 773–787. doi: 10.1152/jn.1981.46.4.773

Hamker, F. H., Zirnsak, M., Calow, D., and Lappe, M. (2008). The peri-saccadic perception of objects and space. *PLoS Comput. Biol.* 4, e31. doi: 10.1371/journal.pcbi.0040031

Honda, H. (1991). The time courses of visual mislocalization and of extraretinal eye position signals at the time of vertical saccades. *Vision Res.* 31, 1915–1921. doi: 10.1016/0042-6989(91)90186-9

Jeffries, S. M., Kusunoki, M., Bisley, J. W., Cohen, I. S., and Goldberg, M. E. (2007). Rhesus monkeys mislocalize saccade targets flashed for 100 ms around the time of a saccade. *Vision Res.* 47, 1924–1934. doi: 10.1016/j.visres.2007.02.021

Kaiser, M., and Lappe, M. (2004). Perisaccadic mislocalization orthogonal to saccade direction. *Neuron* 41, 293–300. doi: 10.1016/S0896-6273(03)00849-3

Khoei, M. A., Masson, G. S., and Perrinet, L. U. (2017). The flash-lag effect as a motion-based predictive shift. *PLoS Comput. Biol.* 13, e1005068. doi: 10.1371/journal.pcbi.1005068

Kusunoki, M., and Goldberg, M. E. (2003). The time course of perisaccadic receptive field shifts in the lateral intraparietal area of the monkey. *J. Neurophysiol.* 89, 1519–1527. doi: 10.1152/jn.00519.2002

Lappe, M., Awater, H., and Krekelberg, B. (2000). Postsaccadic visual references generate presaccadic compression of space. *Nature* 403, 892–895. doi: 10.1038/35002588

Lappe, M., and Krekelberg, B. (1998). The position of moving objects. *Perception* 27, 1437–1449. doi: 10.1068/p271437

Li, Z., and Qian, N. (2015). Solving stereo transparency with an extended coarse-to-fine disparity energy model. *Neural Comput.* 27, 1058–1082. doi: 10.1162/NECO_a_00722

Matin, L., and Pearce, D. G. (1965). Visual perception of direction for stimuli flashed during voluntary saccadic eye movements. *Science* 148, 1485–1488. doi: 10.1126/science.148.3676.1485

Maunsell, J. H. (2015). Neuronal mechanisms of visual attention. *Ann. Rev. Vis. Sci.* 1, 373. doi: 10.1146/annurev-vision-082114-035431

Neupane, S., Guitton, D., and Pack, C. C. (2016). Two distinct types of remapping in primate cortical area V4. *Nat. Commun.* 7, 10402. doi: 10.1038/ncomms10402

Nijhawan, R. (2002). Neural delays, visual motion and the flash-lag effect. *Trends Cogn. Sci.* 6, 387–393. doi: 10.1016/S1364-6613(02)01963-0

Ostendorf, F., Fischer, C., Gaymard, B., and Ploner, C. (2006). Perisaccadic mislocalization without saccadic eye movements. *Neuroscience* 137, 737–745. doi: 10.1016/j.neuroscience.2005.09.032

Pola, J. (2004). Models of the mechanism underlying perceived location of a perisaccadic flash. *Vision Res.* 44, 2799–2813. doi: 10.1016/j.visres.2004.06.008

Ross, J., Morrone, M. C., and Burr, D. C. (1997). Compression of visual space before saccades. *Nature* 386, 598. doi: 10.1038/386598a0

Ross, J., Morrone, M. C., Goldberg, M. E., and Burr, D. C. (2001). Changes in visual perception at the time of saccades. *Trends Neurosci.* 24, 113–121. doi: 10.1016/S0166-2236(00)01685-4

Schall, J., Hanes, D., Thompson, K., and King, D. (1995). Saccade target selection in frontal eye field of macaque. I. Visual and premovement activation. *J. Neurosci.* 15, 6905–6918. doi: 10.1523/JNEUROSCI.15-10-06905.1995

Schlag, J., and Schlag-Rey, M. (2002). Through the eye, slowly; delays and localization errors in the visual system. *Nat. Rev. Neurosci.* 3, 191–191. doi: 10.1038/nrn750

Shim, W. M., and Cavanagh, P. (2006). Bi-directional illusory position shifts toward the end point of apparent motion. *Vision Res.* 46, 3214–3222. doi: 10.1016/j.visres.2006.04.001

Sommer, M. A., and Wurtz, R. H. (2006). Influence of the thalamus on spatial visual processing in frontal cortex. *Nature* 444, 374–377. doi: 10.1038/nature05279

Suzuki, S., and Cavanagh, P. (1997). Focused attention distorts visual space: an attentional repulsion effect. *J. Exp. Psychol. Hum. Percept. Perform*. 23, 443. doi: 10.1037/0096-1523.23.2.443

Teich, A. F., and Qian, N. (2003). Learning and adaptation in a recurrent model of V1 orientation selectivity. *J. Neurophysiol.* 89, 2086–2100. doi: 10.1152/jn.00970.2002

Teich, A. F., and Qian, N. (2010). V1. orientation plasticity is explained by broadly tuned feedforward inputs and intracortical sharpening. *Vis. Neurosci.* 27:57–73. doi: 10.1017/S0952523810000039

Teichert, T., Klingenhoefer, S., Wachtler, T., and Bremmer, F. (2010). Perisaccadic mislocalization as optimal percept. *J. Vis.* 10, 19–19. doi: 10.1167/10.8.19

Tsang, E. K., and Shi, B. E. (2004). A preference for phase-based disparity in a neuromorphic implementation of the binocular energy model. *Neural Comput.* 16, 1579–1600. doi: 10.1162/089976604774201604

Umeno, M. M., and Goldberg, M. E. (1997). Spatial processing in the monkey frontal eye field.1. Predictive visual responses. *J. Neurophysiol.* 78, 1373–1383. doi: 10.1152/jn.1997.78.3.1373

Wang, X., Fung, C. C. A., Guan, S., and Wu, S., Goldberg Michael, E., Zhang, M. (2016). Perisaccadic receptive field expansion in the lateral intraparietal area. *Neuron* 90, 400–409. doi: 10.1016/j.neuron.2016.02.035

Wang, X., Zhang, C., Yang, L., Goldberg, M. E., Zhang, M., and Qian, N. (2019). "Modeling circuit mechanisms of receptive field remapping in LIP and FEF in non-human primates," in: *Program No. 226.23, 2019. Abstract Viewer/Itinerary Planner*. Washington, DC: Society for Neuroscience. Online.

Watanabe, K., and Yokoi, K. (2006). Object-based anisotropies in the flash-lag effect. *Psychol. Sci.* 17, 728–735. doi: 10.1111/j.1467-9280.2006.01773.x

Yang, L., Zhang, C., Wang, X., Goldberg, M. E., Qian, N., and Zhang, M. (2019). "Comparison of receptive field remapping around saccadic onset between lateral intraparietal area and frontal eye field in macaques," in: *Program No. 226.21, 2019. Abstract Viewer/Itinerary Planner*. Washington, DC: Society for Neuroscience. Online.

Yao, H., and Dan, Y. (2001). Stimulus timing-dependent plasticity in cortical processing of orientation. *Neuron* 32, 315–323. doi: 10.1016/S0896-6273(01)00460-3

Zirnsak, M., Steinmetz, N. A., Noudoost, B., Xu, K. Z., and Moore, T. (2014). Visual space is compressed in prefrontal cortex before eye movements. *Nature* 507, 504. doi: 10.1038/nature13149

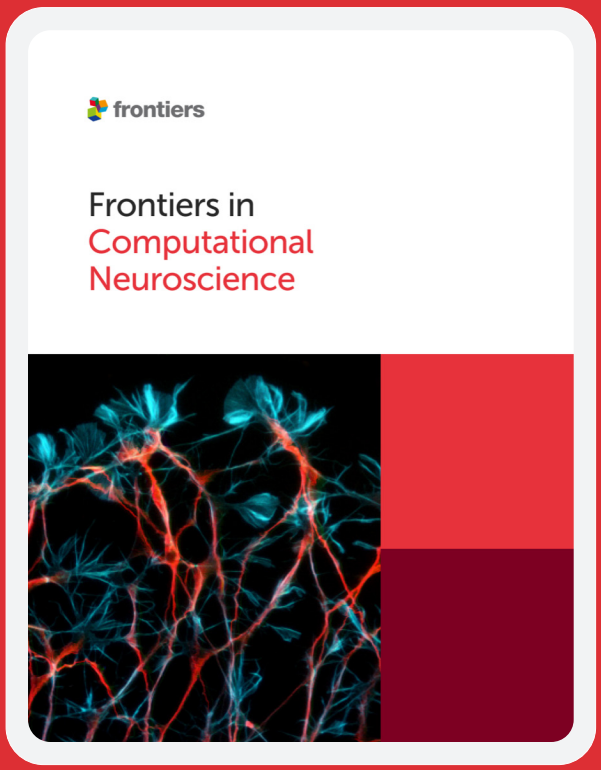
Frontiers in
Computational Neuroscience

Fosters interaction between theoretical and experimental neuroscience

Part of the world's most cited neuroscience series, this journal promotes theoretical modeling of brain function, building key communication between theoretical and experimental neuroscience.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact