Advancing models of the visual system using spiking neural networks optimized for temporal prediction

Luke Taylor

Hertford College, Oxford University

Department of Physiology Anatomy & Genetics

Thesis submitted for the degree of Doctor of Philosophy

Hilary 2023

Abstract

Spikes are thought to provide a fundamental unit of computation in the nervous system. The retina is known to use the relative timing of spikes to encode visual input, whereas primary visual cortex (V1) exhibits sparse and irregular spiking activity – but what do these different spiking patterns represent about sensory stimuli? To address this question, I set out to model the retina and V1 using a biologically-realistic spiking neural network (SNN), exploring the idea that temporal prediction underlies the sensory transformation of natural inputs.

Firstly, I trained a recurrently-connected SNN of excitatory and inhibitory units to predict the sensory future in natural movies under metabolic-like constraints. This network exhibited V1-like spike statistics, simple and complex cell-like tuning, and - advancing prior studies - key physiological and tuning differences between excitatory and inhibitory neurons.

Secondly, I modified this spiking network to model the retina to explore its role in visual processing. I found the model optimized for efficient prediction to capture retina-like receptive fields and - in contrast to previous studies - various retinal phenomena, such as latency coding, response omissions, and motion-tuning properties. Notably, the temporal prediction model also more accurately predicts retinal ganglion cell responses to natural images and movies across various animal species.

Lastly, I developed a new method to accelerate the simulation and training of SNNs, obtaining a $\sim 10-50\times$ speedup, with performance on a par with the standard training approach on supervised classification benchmarks and for fitting electrophysiological recordings of cortical neurons.

The retina and V1 models lay the foundation for developing normative models of increasing biological realism and link sensory processing to spiking activity, suggesting that temporal prediction is an underlying function of visual processing. This is complemented by a new approach to drastically accelerate computational research using SNNs.

Acknowledgement

The last three years have undoubtedly been an incredible journey. My gratitude is due to many who have guided and supported me throughout this time. Foremost, I am grateful to my supervisors for their guidance and the academic freedom they have granted me to explore my interests. A special thank you to Nicol and Andy for the many hours you have put into this work and the many lessons you have taught me. Your attention to detail has instilled in me a greater sense of scientific rigour. I would also like to thank Friedemann for drawing me into the intriguing world of spiking neural networks. I am thankful to the Clarendon Fund, without whose generous funding my DPhil would not have been possible; and to Sarah for helping me navigate all of the university guidelines.

I am immensely grateful to my family and friends. To my parents and brother for their love and encouragement over the years. To my grandparents and granny Eva for your kind messages and support. To friends old and new, who bring much joy and vibrance to my life. Especially to Adi and Chris for our continued adventures over the years – no matter where we are. And most importantly to Melissa, for your love and companionship. As scientists say, the greatest discoveries are those that you least expect and you undoubtedly mark such an occurrence in my life.

Contents

1 Introduction

1.1 Creating an artificial visual system

Our brains dictate our everyday experiences, what we think, what we feel, and how we perceive and interact with the world. The oldest known records that highlight the significance of the brain date back to the ancient Egyptians approximately ca. $3000 - 2500$ B.C., in which they outline a series of medical injuries, in particular, how damage to the head can result in limb paralysis or the inability to speak [1]. Ancient Greek physician Hippocrates (ca. $460 - 370$ B.C.) further refined such observations, suggesting the brain to be central to our intelligence: it is what makes us, us [2]. The notion of the brain being at the centre of our intelligence was not generally accepted at that time, as evident by the views of the ancient Greek philosopher Aristotle (ca. $384 - 322$ B.C), who rather believed the heart to be the centre of our intelligence [3]. However, a few hundred years later the Greek physician Galen (ca. $129 - 216$ A.D.) made the same observations as those of the ancient Egyptians, noting how damage to the heads of Roman gladiators resulted in impaired mental faculties [4].

Since then scientists have been trying to decipher exactly how the brain works. Two observations over the last few centuries have arguably underpinned the start of modern neuroscience (the study of the brain). Firstly, Emil du Bois-Reymond - a pioneer of electrophysiology - noted how nerve fibers operate by means of electrical currents [5]; and secondly, Santiago Ramón y Cajal - a pioneer of neuroanatomy - discovered that the brain is comprised of individual cells, called neurons, which form the brain's key units of computation [6]. It is these neurons, communicating by electrical and chemical signals, which through their collective interactions give rise to our brain's intelligence and our everyday experience. How they do so is a question that continues to perplex and challenge neuroscientists.

It is now known that the human brain consists of many billions of neurons and trillions of synaptic connections [7]. Making sense of their collective interactions - especially at such a large scale - remains a challenging problem. One approach to simplifying such complexity is to view the brain as being modular, comprised of many distinct subregions, and to make sense of each of their particular functions. Making an analogy to programming: to understand how a program

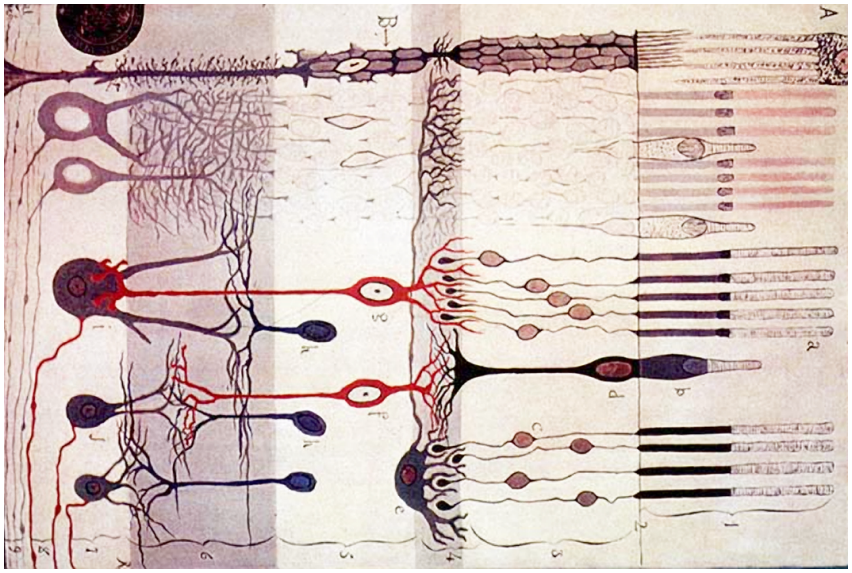

Fig. 1.1: Classic drawing of the retina by Santiago Ramón y Cajal [8]. Ellipse-like shapes represent retinal neurons, with the different colours denoting different neuron types. The tree-like structure illustrates a neuron forming connections with other neurons.

works, try to understand each of the unique subroutines that collectively form the program. Such a reductionist view has, for example, been adopted in the scientific investigation of the brain's sensory systems, by studying how sensory information is processed by particular circuits in the brain to give rise to our distinct sensations - like touch, hearing, sight, smell, and taste.

One of the earliest senses to grasp the attention of neuroscientists was vision. Some of the first illustrations of neurons by Santiago Ramón y Cajal are from the retina (Figure 1.1), a thin line of tissue at the back of the eye. Decades later, it was discovered that these retinal neurons respond to flashes of light [9, 10]. Photoreceptors in the retina convert incoming light into electrical signals, which are propagated via other neurons in the retina to certain regions within the brain, which collectively form the visual system. In the mid-twentieth century it was determined that many higher visual brain areas respond to the sight of oriented bars and edges [11, 12, 13]. To date, neuroscientists have obtained a large amount of information about the visual system, including: the different anatomical regions involved and their organization [14, 15, 16, 17, 18]; the different cell types found in the different regions [19, 20, 21, 22]; how such neurons connect to each other [23]; their physiological modes of operations [24, 25, 26]; their spike statistics [27, 28]; their tuning properties [11, 12, 13, 29, 30, 31, 32, 28]; and their developmental trajectories [33, 34, 35]. Although we have amassed a great deal of knowledge about the visual system, we arguably still do not understand what the system as a whole is performing. Just like the heart supports life, the visual system enables vision. Yet we understand that the heart supports life

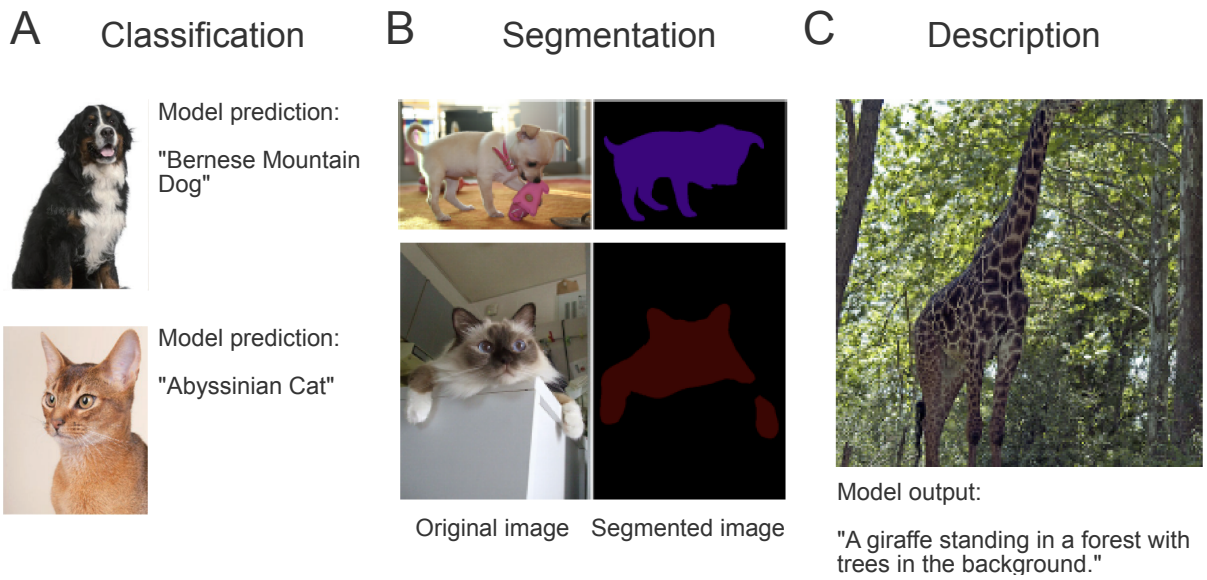

Fig. 1.2: **A.** Deep learning vision model classifying different breeds of cats and dogs (adapted from [39]). **B.** Deep learning vision model performing image segmentation and extracting the outline of a cat and dog (adapted from [40]). **C.** Deep learning vision model describing the photograph of a giraffe (adapted from [41]).

by pumping oxygenated blood through the body. Contrary to the heart, we are still trying to determine how the visual system enables visual perception.

One avenue to deciphering the function of the visual system is by viewing the system as a mathematician or an engineer might, that is, building a model of the visual processes in the brain. Models have been instrumental in understanding our physical reality. For example, Newton's laws of motion have allowed us to understand how the forces acting on an object dictate an object's motion [36]; Maxwell's equations provide us with the means to understand how charges and currents give rise to magnetic and electrical fields [37]; and Einstein's theory of general relativity allows us to understand even more perplexing relations, such as the effect of gravity on time and space [38]. Models provide us with an understanding of particular relations in the world, yet it may not be immediately clear what a model of the visual system may look like - and what relations it should explain.

The last decade (2012 onwards), has marked the rise of a new age of artificial intelligence (AI), the so-called deep learning revolution [42]. During this period we have witnessed computers matching and sometimes surpassing human capabilities on various tasks. This includes mastering complex video and board games like Starcraft and Go [43, 44]; autonomously navigating challenging environments in self-driving cars [45], and even surpassing law students on the

bar exam [46]. Inspired by the brain, deep learning models structurally resemble the brain, consisting of a network of connected units, which - like neurons - give rise to the network functionality. These models, in particular vision-based deep learning models, have demonstrated remarkable capabilities in processing and interpreting visual scenes, such as distinguishing objects in photographs [47]; precisely isolating the outline of an object from its background in a photo [48]; and generating detailed descriptions of arbitrary visual scenes [41] (Figure 1.2). Furthermore, these models exhibit brain-like processing, with certain units being tuned to oriented edges and bars [47], just like neurons in the visual system [11, 12, 13]; and these models excel at predicting neural responses across the visual pathway, ranging from lower regions like the retina [49] to higher cortical areas [50, 51].

Although they mimic the structure of the visual system and emulate brain-like computations, these deep learning models are inconsistent with many aspects of the biology. Neurons typically operate using all-or-nothing binary outputs known as spikes, whereas deep learning models employ continuous unit activation. Neurons also exhibit temporal dynamics and adhere to specific connectivity constraints, characteristics that are not accounted for in conventional deep-learning vision models. Lastly, unlike the brain, deep learning models mostly depend on supervised training signals to learn network functionality, whereas the brain learns through a combination of unsupervised, supervised, and reinforcement learning signals [52]. The objective of this thesis is to address these shortcomings, in an attempt to create a more biologically realistic model of the visual system.

1.2 Overview of the visual system

The visual system is comprised of many different brain regions, each processing visual input in distinct stages. In a simplified description, light is detected in the retina, which relays the visual input to the lateral geniculate nucleus in the thalamus, which further passes the visual information to the primary visual cortex (V1) - the first stage of visual processing in the cortex (Figure 1.3A). V1 relays the visual information to higher cortical areas, which are categorized into two pathways: the ventral stream and the dorsal stream [16, 17]. The main brain regions of the ventral stream are V2, V4 and inferior temporal cortex (IT) [18]; and the main brain regions of the dorsal stream are the middle temporal area (MT) [54], the medial superior temporal area (MST) and the posterior parietal area (PP) [18]. Both streams are typically characterized by different functional roles, with the ventral stream - also sometimes known as the what-stream -

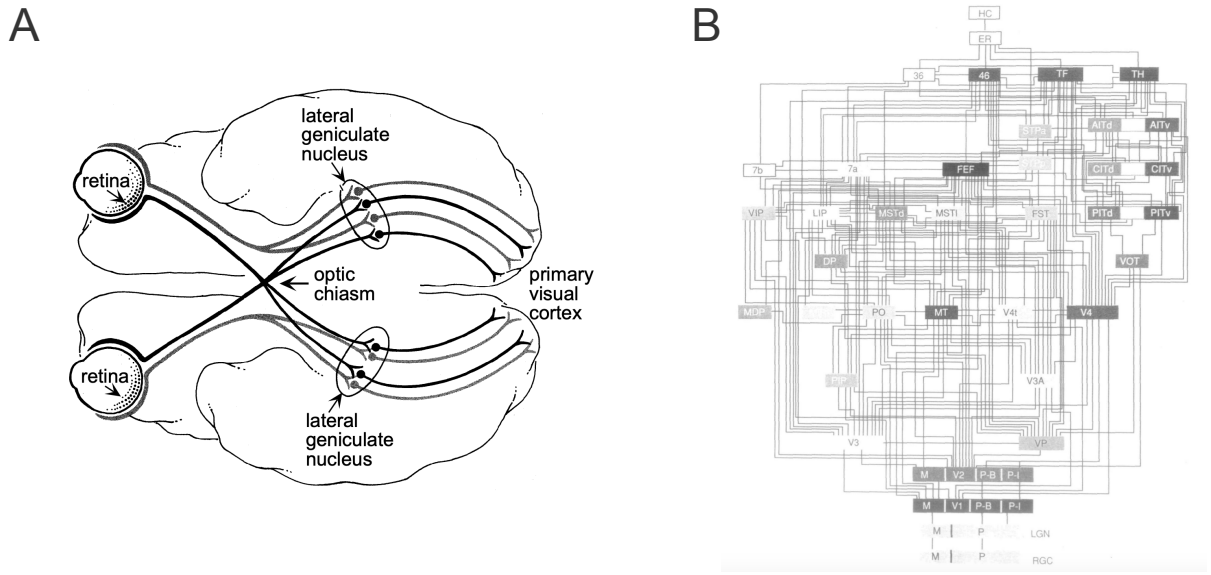

Fig. 1.3: **A.** Pathways in the visual system from the retina, through to the lateral geniculate nucleus in the thalamus, and further to the first cortical stage of visual processing: the primary visual cortex (adapted from [53]). **B.** Complex connectivity scheme of feedforward and feedback connections between the different brain regions throughout the visual system (adapted from [14]).

processing features related to object identity and form; and the dorsal stream - also known as the where-stream - processing features related to object motion and location [16].

Throughout the visual system, visual information propagates in both the forward and backward directions across the ventral and dorsal pathways, where such bi-directional transmission is facilitated by intricate connectivity among the various brain regions within each visual stream [14] (Figure 1.3B). Traditionally the visual system is viewed to process information in a hierarchical fashion along the feedforward connectivity from the lower (*e.g.* retina) to higher (*e.g.* MT and IT) regions [14] (although see [55] for a non-hierarchical interpretation of the visual system). The exact role of the feedback connectivity remains elusive, and is thought to play a modulatory role, permitting higher brain regions to sharpen the sensitivity and tuning of neurons in lower regions [56, 57]. While the visual system is characterized by extensive feedback connectivity [58, 59], it could be argued that excluding this connectivity in models of visual processing may be justified. This is because visual recognition happens rapidly, possibly too fast for feedback connections to play a role in many cases [60, 57].

The scope of investigation of this thesis is focused on the visual processes in the retina and in primary visual cortex. Before introducing these different anatomical regions, I briefly introduce

the different ways by which neuroscientists traditionally characterize the tuning properties of neurons in the visual system.

1.2.1 Characterizing neural tuning

Neurons in the visual system are typically characterized by the relation to their visual input. A neuron is said to be tuned to a particular visual stimulus if that stimulus evokes a large response, with the optimal (or preferred) visual stimulus eliciting the largest response. The earliest investigations into the visual system adopted a trial-and-error approach to determine what stimulus a neuron is tuned to. Such examinations revealed the retinal neurons to be tuned to flashes of dots of light [9, 10] and V1 neurons to be tuned to orientated bars and edges [11, 12, 13]. Searching for a neuron's preferred stimulus using such a trial-and-error approach is, however, time-consuming.

A more efficient way to determine a neuron's preferred stimulus is to adopt more sophisticated quantitative methods for mapping its receptive field (RF). The notion of an RF is used in two different ways by neuroscientists, and either refers to a neuron's spatial extent in visual space [9] (*e.g.* a V1 neuron only views a small spatial portion of the entire visual input); or more commonly, it refers to a linear mapping between visual input and a neuron's response. The linear mapping, when visualized, corresponds to a neuron's preferred stimulus. Typically the RF is determined by measuring a neuron's response to many different samples of white or binary input-noise. The RF can then be calculated by either taking a spike-triggered-average (STA), by averaging over all the visual inputs which elicited a spike; or by linear regression between the visual input and neuron's responses [61, 62]. The RF is not only characterized by a spatial structure (*e.g.* a bar-like RF shape as found in V1 neurons), but also by a temporal profile (*e.g.* a sequence of bar-like shapes), which are jointly referred to as a spatiotemporal RF [34, 62]. The temporal dimension outlines how a neuron is tuned to a changing spatial structure, where retinal and V1 neurons often prefer stimuli that change polarity over time (dark regions becoming light, and light regions becoming dark). Although RFs have been useful in determining different neural tuning properties, this method is not applicable to all neurons. An RF describes the linear dependence of a neuron's response on the visual input. This works well for early-stage neurons like those in the retina and some neurons in V1. However, neurons higher up in the visual pathway exhibit more non-linear processing and are thus less suited for RF characterizations (although see [63, 64] for latest developments in determining neural tuning).

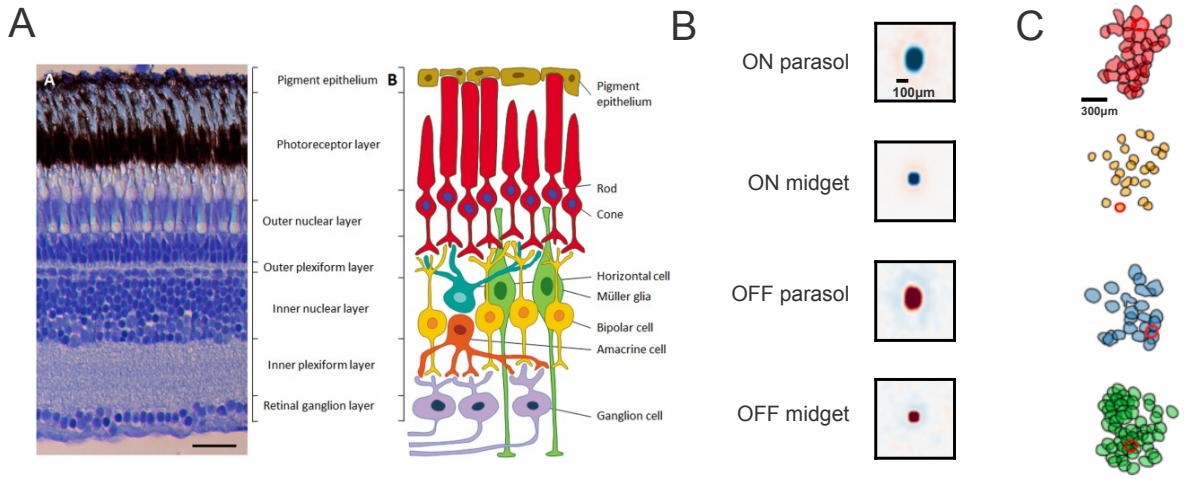

Fig. 1.4: **A.** Cross-section photomicrograph through the retina of an adult zebrafish (scale bar measures $25\mu m$), showcasing different cellular layers (left) and their corresponding cell types (right) (adapted from [67]). **B.** The RFs of the different major retinal ganglion cell types in the macaque (adapted from [68]; blue - excitatory region, red - inhibitory region). **C.** Mosaic-like organization across the retina for each major retinal cell type (adapted from [69]). Cell types correspond to those labelled in **B**.

In addition to the RF, neurons in the visual system - in particular the retina and V1 - are also often characterized by their motion-tuning properties to moving bar-like shapes. Various types of measures have been proposed to quantify a neuron's motion preference along a certain orientation and in a certain direction [35, 65, 66, 28]. These measures are usually calculated from neural responses to drifting sinusoidal gratings, presented at different spatial frequencies and orientations, and moving at various speeds.

1.2.2 The retina

The retina is a thin tissue at the back of the eye, and forms the first stage of visual processing, converting incoming light into a neural representation. The retina is comprised of three cellular layers (Figure 1.4A). The first layer consists of the photoreceptor cells, which convert incoming light into neural responses. Counterintuitively, these cells are located at the back of the retina and light has to pass through the other neural layers before it can reach the photoreceptors. Photoreceptors are divided into two classes of cells, the rods and the cones. The rods outnumber the cones, with an estimated 92 million rods compared to the approximately 6 million cones in the primate retina [70]. Within a visual scene, the rods convey the light intensity and the cones

convey the colour information [71]. These cells convey the information to an intermediate layer of neurons, the bipolar cells, and the horizontal cells. The horizontal cells deliver inhibitory feedback to both the photoreceptor and bipolar cells, whereas the bipolar cells further connect to the last layer of retinal neurons, the ganglion and amacrine cells. Like the bipolar cells, the amacrine cells provide an inhibitory mechanism, regulating the activity of both the bipolar and ganglion cells. The ganglion cells act as the final output of the retina and signal the visual world to the rest of the brain [72].

Ganglion cells - including the amacrine cells - mark the first cells in the visual pathway to employ spikes rather than graded potentials, whereas other retinal cells - like the photoreceptors, bipolar, and horizontal cells - use graded potentials, whereby a neuron's response continuously varies in relation to its stimulus input. There are various types of ganglion cells in the mammalian retina [73]. Traditionally, these cells are classified into two primary groups, the midget and parasol cells. These cells account for 95% of the ganglion cell types in the central primate retina [74]. Both cell types respond to flashes of light within an ellipse-like region within the visual field, where such region is conventionally quantified as a two-dimensional Gaussian [75, 76], or a difference of two-dimensional Gaussians [77]. Parasol and midget cells either respond to the onset of light, or the offset of light, and are respectively referred to as ON-type and OFF-type cells, with the retina containing more OFF-type than ON-type cells [78, 75] (Figure 1.4B). Parasol cells differ from the midget cells, in that they are less plentiful [74, 79] and exhibit a larger RF [75]. Furthermore, parasol cells tend to respond with a burst of spikes to sudden changes in light intensity, whereas midget cells tend to exhibit sustained responses [80, 81]. Another characteristic of the ganglion cells is that their respective types (*e.g.*, OFF-type parasol or ON-type midget), tile the visual space with a mosaic-like organization [82, 83, 84] (Figure 1.4C).

The retina has traditionally been viewed to be like a camera, capturing incoming light, and transmitting a neural representation thereof to the brain [85, 86]. It is now known that the retina supports a wider range of tasks and functionality. Apart from signalling visual information, the retina also acts as a circadian clock, regulating biological rhythms [87]. Furthermore, the ganglion cells in the retina also exhibit many complex tuning properties to visual stimuli. For example, certain cells respond to the motion of moving objects, including the direction and the orientation of motion [88, 89, 90]; differential motion (motion within a neuron's RF moving differentially to the motion outside of the RF) [91]; and even the anticipation of motion [92]. Moreover, certain cells also respond to the violation of a periodic signal, like a sequence of flashing lights with random light-flash omissions [93]. These different tuning characteristics hint at the retina performing more complex transformations on visual stimuli than simply performing

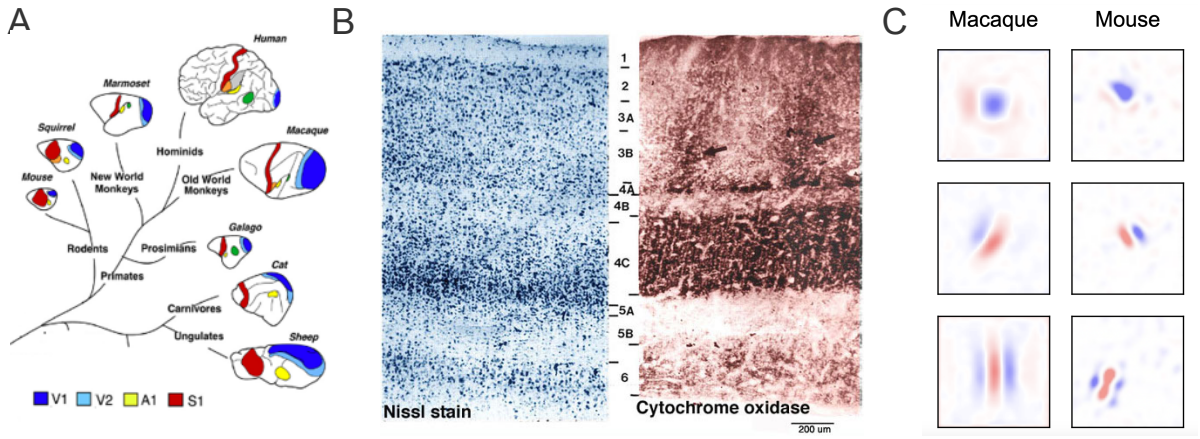

Fig. 1.5: **A.** Illustration of V1 (blue) and other primary sensory cortical areas (non-blue colours) in different mammalian species (adapted from [95] and originally from [96]). **B.** Different layers of V1 of a macaque monkey. Nissl stain (left) and cytochrome oxidase (right) are both labelling techniques for visualising the different cortical layers (adapted from [19]). **C.** Example V1 RFs in macaque and mouse (adapted from [97] and [28]). Red denotes excitatory regions and blue denotes inhibitory regions.

a camera-like projection. Lastly, the activity of neurons in the retina also varies with the level of arousal and locomotion of an animal [94].

1.2.3 Primary visual cortex

The primary visual cortex, often simply referred to as V1, also known as the striate cortex or Brodmann area 17 in primates, is located in the occipital lobe at the back of the brain [98]. This brain region consists of an estimated 140 million neurons per hemisphere in humans [99] and marks the first stage of visual processing within the cortex. V1 is also present in other mammalian species [96] (Figure 1.5A), and has primarily been examined in cats, non-human primates, and, more recently, mice. As in the rest of the cortex, V1 is traditionally viewed to consist of six distinct layers, with particular connectivity statistics across the different layers [15] (Figure 1.5B). The exact number of layers can be further refined, for example in primates, layer 4 is usually further subdivided into layers 4A, 4B, and 4C. V1 receives visual input through feedforward connections into layer 4 from the lateral geniculate nucleus in the thalamus [14]. It takes approximately a minimum of $\sim 30-40$ms in mice [100, 101] and monkeys [102, 103] for visual information to reach V1 from the time light impinges on the photoreceptors. Input to V1 also comes from feedback connections from higher cortical areas and connections from other

subcortical structures, including the claustrum, and the nucleus basalis. However, in contrast to the thalamocortical projections, the exact role of these connections remains less well understood [95].

The discovery of V1 was made by anatomist Panizza, who in 1855 published his findings on patients who became blind after a stroke in the posterior part of their brain [104]. It took more than a century thereafter to begin to understand how V1 was supporting vision. Around the 1960s, neurophysiologists Hubel and Wiesel discovered that neurons in V1 are tuned to the orientation and motion of moving edges and shapes. These neurons are referred to as the simple and complex cells [11, 12, 13]. The tuning characteristics of these cells are usually measured in response to drifting sinusoidal gratings, where simple and complex cells are tuned to the orientation, spatial frequency (number of oscillations), and temporal frequency (drifting speed) of these drifting gratings. Simple cells have stereotyped RFs consisting of oriented parallel bars of excitatory and inhibitory regions [31, 32], which are usually mathematically quantified using a Gabor function [105] (Figure 1.5C). These cells exhibit a maximum response when exposed to a bar-shaped pattern, wherein the dark and bright areas align precisely with the inhibitory and excitatory regions of their RFs. Simple cells also exhibit stereotypical temporal dynamics, being most sensitive to visual stimuli near the present with the sensitivity decaying into the past [106]. The spatial structure within the simple cell's spatiotemporal RF may also switch polarity over time, where monophasic RFs exhibit no polarity reversal; biphasic RFs exhibit a single polarity switch; and triphasic RFs exhibit two polarity switches [34, 107, 108]. The spatial configuration in a spatiotemporal RF can also shift over time, as in space-time inseparable RFs, or remain fixed in time, as in space-time separable RFs. Space-time inseparable RFs predominantly respond to specific directions and speeds of grating motion, whereas space-time separable RFs primarily react to particular speeds of grating motion [34, 108]. Unlike simple cells, complex cells have a less defined RF spatial structure and will respond to moving bars (or drifting sinusoidal gratings) regardless of their precise location within a neuron's RF. Classifying a neuron as simple-cell or complex-cell is usually achieved by quantifying its responses to drifting sinusoidal gratings. Simple cell responses oscillate to drifting gratings substantially more than complex cell responses, with simple cells exhibiting a higher phase dependence with larger variations in their responses to the moving gratings [29, 30].

Neurons in V1 can also be categorized into inhibitory and excitatory neurons. Upon firing, inhibitory neurons decrease the spiking probability of the neurons they connect to, by lowering their membrane potential. Conversely, excitatory neurons increase the spiking probability of the neurons they connect to, by raising their membrane potential. Such differences are established by excitatory and inhibitory neurons releasing different neurotransmitters at their connections

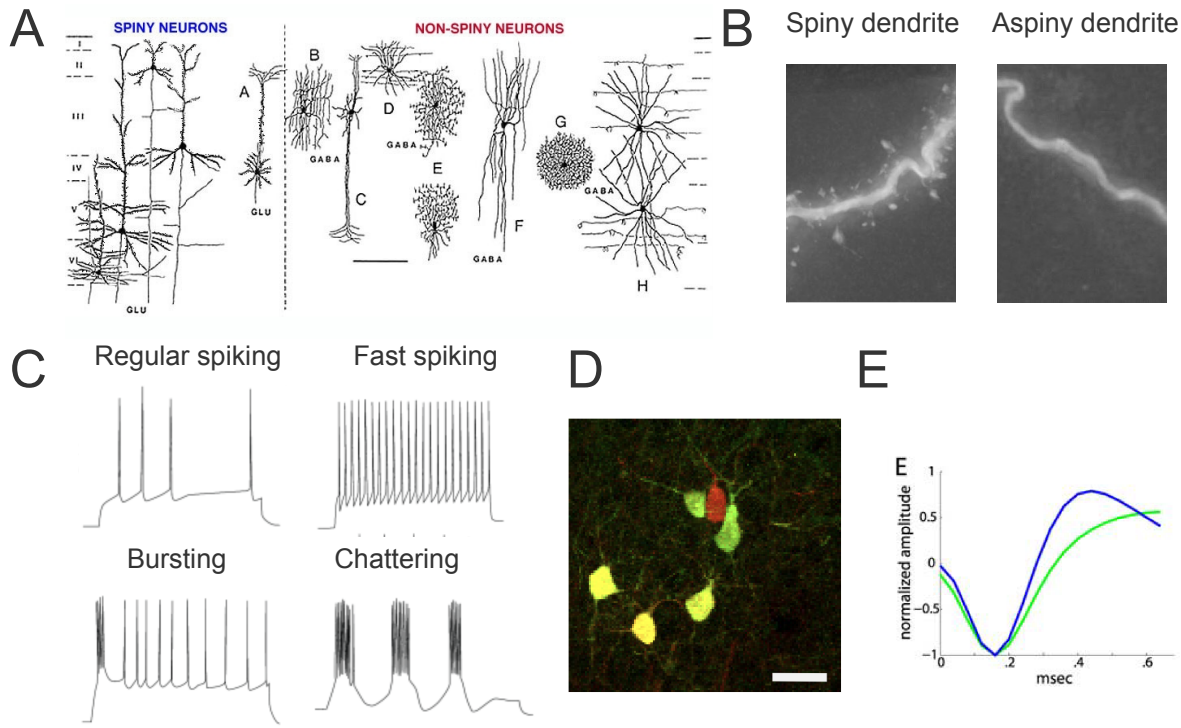

Fig. 1.6: **A.** Example morphological structures of different excitatory neurons (left) and inhibitory neurons (right) from monkey cortex: A. pyramidal cells and stellate cells; B. cell with local axon arcades; C. double bouquet cell; D. and H. basket cells; E. chandelier cells; F. bitufted cell; G. neurogliaform cell (adapted from [19]). **B.** Zoomed-in dendrites with spines (left) and without spines (right) (adapted from [109]). **C.** Different classes of neuron spike responses to sustained current injection in the soma (adapted from [110]). **D.** Immuno-labeling of neurons in transgenic mouse V1, where inhibitory neurons are shown in red, while excitatory neurons are green. Scale bar measures $20\mu m$ (adapted from [23]). **E.** Average spike-waveforms of mouse V1 neurons, where inhibitory neurons exhibit a narrow-spiking waveform (blue), whilst excitatory neurons exhibit a broad-spiking waveform (green) (adapted from [28]).

(synapses) onto other neurons. Namely, excitatory neurons mainly release the neurotransmitter glutamate, and inhibitory neurons mainly release the neurotransmitter γ-aminobutyric acid (GABA) [62]. The commonly observed principle that at all the synapses from a neuron release, the same neurotransmitter or set of neurotransmitters is known as Dale's law [111] - thus individual neurons' synaptic outputs are either excitatory (*e.g.* glutamate) or inhibitory (*e.g.* GABA) but are not a mixture. Approximately $15 - 20\%$ of neurons in V1 across mice, cats, and monkeys are inhibitory [112, 113, 114, 23].

Apart from differences in their effects, excitatory and inhibitory neurons also differ in their morphological structure, where anatomists have named different neural subtypes based on their different shapes, sizes, and dendritic branching [19]. For example, pyramidal cells exhibit a pyramid-shaped cell body (soma) with a long projection at their apex (apical dendrite); stellate cells exhibit a star-shaped soma and dendrites; and basket cells exhibit basket-like shapes in the branching of its output projection (axon) (Figure 1.6A). Most excitatory neurons in mouse V1 are pyramidal cells [20], and most inhibitory neurons in mouse V1 are basket cells [21]. However, the prevalence of the cell types can vary depending on the cortical region, since, for example, stellate cells are the most common excitatory neuron in the somatosensory cortex of rats [22] and mice [20]. A key morphological difference distinguishing the excitatory from the inhibitory neurons is that excitatory neurons generally have many spines (small knobs) on their dendrites, in comparison to the inhibitory neurons, which generally exhibit smooth dendritic surfaces [19, 109] (Figure 1.6B).

Excitatory and inhibitory neurons also exhibit different response properties, with characteristic spike responses to depolarizing step current pulses [26, 115, 116]. Most excitatory neurons exhibit a regular spiking response: a sequence of a few spikes separated by an interspike interval (*i.e.* the temporal difference between consecutive spikes) which increases in duration over time (*i.e.* adapts). Most inhibitory neurons exhibit a fast-spiking response: a sequence of many spikes separated by very short interspike intervals that remain fixed in time [115]. Although less common, some excitatory and inhibitory neurons display other types of spiking characteristics, such as bursting or chattering responses. Bursting neurons fire a burst of spikes followed by a repeating sequence of single spikes, whilst chattering neurons fire a repeating sequence of bursts of spikes [116, 110] (Figure 1.6C). Neuroscientists can identify excitatory and inhibitory neurons during experimental recordings using various methods. For example by using genetically modified animal models in which the different neural populations fluoresce with different colours [23] (Figure 1.6D) or by examining the neuron spike waveforms, where inhibitory neurons exhibit a more narrow waveform than the excitatory neurons [28] (Figure 1.6E).

Excitatory and inhibitory neurons in V1 are also characterized by various other physiological, connectivity, and tuning differences. Inhibitory V1 neurons tend to operate on faster timescales than their excitatory counterparts, as measured by differences in their membrane time constants [25]. Excitatory V1 neurons also tend to make stronger connections with inhibitory V1 neurons than with other excitatory V1 neurons [23]. Although V1 spike activity is generally sparse, irregular, and uncorrelated between different neurons [27], inhibitory V1 neurons tend to fire at higher rates than their excitatory counterparts [28]. Lastly, V1 inhibitory neurons tend to be more complex-cell than simple-cell, and more broadly tuned for orientation, than the excitatory V1 neurons [28].

A hallmark of modern electrophysiology is the notion that excitatory and inhibitory currents into cortical neurons are generally balanced in time [117, 118, 24, 119, 120]. Balance is referred to as global when the mean excitatory and inhibitory currents are equal over time, and referred to as detailed when these currents are temporally coupled over time, with each excitatory input to a neuron coinciding with a corresponding inhibitory counterpart [121]. The balance of excitatory and inhibitory currents is important, as it ensures stability in the activity over the neural population [122], and thus supports proper cortical function [123].

1.3 Modelling the visual system

Modelling the brain, or specific parts like the visual system, can be addressed at various levels of abstraction, depending on the scientific question of investigation. Neuroscientist David Marr outlined three levels of abstraction to modelling the brain [124]: the computational level, the goal of the circuit (*e.g.* the retina perhaps acting like a camera, capturing incoming light and sending this to the brain); the algorithmic level, how the circuit computation is performed (*e.g.* the retina may perform a camera-like transformation using the different types of RFs, capturing different details in the visual stimulus); and the level of implementation, how the computation is implemented in the neural circuit (*e.g.* the different cell types of the retina, connecting in stereotyped ways). Throughout the literature, different terminologies are used to refer to the same ideas proposed by David Marr. For example, the computational level might be referred to as interpretive or normative modelling; the algorithmic level might be referred to as descriptive modelling; and the implementation level might be referred to as mechanistic modelling [62, 125].

These different levels of modelling are not necessarily mutually exclusive, where the grand goal of modelling a particular neural circuit would be to understand the circuit at all the levels

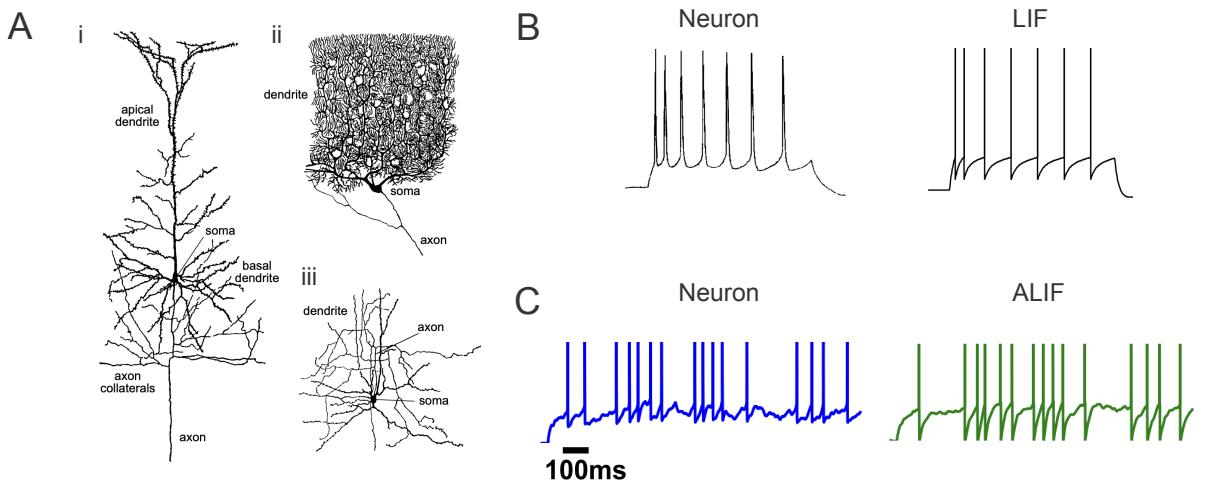

Fig. 1.7: **A.** Illustration of three neurons with labels identifying the dendrites, the soma and the axon of a cortical pyramidal cell (i), a Purkinje cell of the cerebellum (ii), and a stellate cell of the cerebral cortex (iii) (adapted from [62] and originally from [126]). **B.** Example cortical neuron recording to a constant current injection (left) and the corresponding LIF model prediction (right) to the same input current (adapted from [62] and originally from [127]). **C.** Example cortical neuron recording to a noisy current injection (left) and the corresponding ALIF model prediction (right) to the same input current (adapted from [128]).

of abstraction: its goal, the mechanisms underlying its goal, and how those mechanisms are implemented by the neurons. In this thesis, I set out to model the retina and V1 in alignment with this grand goal. That is, implement models using architectures that align with the known biological physiology (implementation level), that implement a goal (computational level), and that learn to do so through an optimization process (algorithmic level). In the next subsections, I briefly introduce different model architectures, normative goals of vision, and model learning algorithms.

1.3.1 Model architecture

The architecture of a model should ideally mirror the neural circuit it aims to replicate, characterized by a distinct arrangement of interconnected neurons. As a brief overview, a neuron consists of four key morphological features: the dendrites, the soma, the axon, and the synaptic boutons (Figure 1.7A). The dendrites and soma collect incoming signals from other neurons, the soma integrates all these incoming signals, and the axon transmits a neuron's response to the synaptic boutons which act upon the post-synaptic neurons by neurotransmitter release. As previously mentioned, transmission along axons is generally achieved using brief electrical pulses known as

spikes which cause the release of neurotransmitters from the synaptic boutons. However, a few neurons, including some of those found in the retina, transmit signals to their synapses using graded electrical signals.

A neuron emits a spike if its membrane potential reaches the so-called firing threshold. A neuron is encapsulated in a lipid bilayer, where the membrane potential measures the difference in electrical potential between the inside and the outside of the cell. A neuron's membrane potential usually sits at around -70mV in the resting state [62]. The membrane potential rises if a neuron is positively stimulated, which is referred to as depolarization (contrary, lowering the membrane potential is referred to as hyperpolarization). If the membrane potential at the base of the axon reaches a large enough value (the neuron's firing threshold), the neuron emits a spike, which travels down the axon and is broadcasted to all the neurons which the axon synapses upon.

The membrane potential arises from the different ion concentrations inside and outside of the neuron, where the ion charges give rise to the electrical gradient of a neuron. The primary ions that determine the membrane potential are K^+ (potassium ions) and Na^+ (sodium ions). At rest, a neuron has a larger intracellular concentration of K^+ and a lower concentration of Na^+ than that found outside the neuron. These resting-state ion concentrations arise when the electrical forces between the ions are equal to the diffusive forces of the ions across the cell membrane [129, 62]. Thus, the membrane potential of a neuron changes depending on the relative concentration of ions inside vs outside of the neuron and the relative permeability of the membrane to those ions. The membrane potential depolarizes in response to pre-synaptic inputs from excitatory neurons, and conversely, hyperpolarizes in response to pre-synaptic inputs from inhibitory neurons. These de- and hyperpolarizing membrane potentials result from certain neurotransmitters being released by the pre-synaptic neurons which bind to the post-synaptic neuron. This neurotransmitter binding has the net effect of opening certain ion channels in the membrane of the neuron and permitting the flow of ions across the membrane. This changes ion concentrations inside of the neuron and thus alters the membrane potential. See [62] for a more in-depth discussion.

Needless to say, modelling a single neuron is challenging, let alone modelling an entire network of neurons. Neuroscientists have developed models of varying complexity to capture the response properties of individual neurons. The simplest models are the linear model (the RF) and the linear-nonlinear model. These models usually map a sensory or neural input $s \in \mathbb{R}^N$, like a pixel-valued image (with N pixels) or set of neural inputs (from N input neurons), using a linear map $W \in \mathbb{R}^N$ to a real number which represents the neural output $r \in \mathbb{R}$. The number is often interpreted as a neuron's firing rate (spikes per second) over a certain time window.

$$
\begin{aligned}
r &= Ws \quad \text{(linear model)} \\
r &= f(Ws) \quad \text{(linear-nonlinear model)}
\end{aligned} \tag{1.1}
$$

The linear-nonlinear model (a linear mapping, followed by a non-linear transformation f, for example, like a rectified linear function) has the advantage of ensuring that the model's firing rate does not drop below zero (where a negative firing rate is biologically implausible). These types of models have the same form as the units commonly used deep learning models. However, these linear-nonlinear models omit two key physiological phenomena of neurons: the evolution of their membrane potential over time, and their spiking output. The leaky integrate-and-fire (LIF) model captures these dynamics [110].

$$
\tau \frac{dV(t)}{dt} = -V(t) + V_{rest} + RI(t)
$$

The LIF model integrates the incoming spikes, or rather the input current $I(t)$ which the spikes induce, into the model's membrane potential $V(t)$ (for a given input resistance R; see [62]). Like in neurons, this membrane potential decays in time when no input is provided, with the leakiness of this decay defined by the membrane time constant τ. If the artificial membrane potential reaches the model's firing threshold, a spike $S(t)$ is emitted and the artificial membrane potential is reset to rest value V_{rest}. Conventionally, these continuous-time dynamics are discretized in time (*e.g.* using the forward Euler method), to simulate the LIF model on a computer [62, 110] (simulating the dynamics using a discretization time step of Δt). Additionally, and without loss of generality, these dynamics are normalized, by setting the artificial resting and reset state to a value of zero, and the artificial firing threshold (and the input resistance R) to a value of one [130]. The normalized and discretized LIF model outlines how to update the membrane potential from one time step to the next, and usually appears in two common forms in the literature, v1, and v2 [131, 132, 133, 128]:[1]

$$
\begin{aligned}
V[t+1] &= \beta V[t] + (1-\beta) I[t] - S[t] \quad \text{(v1)} \\
V[t+1] &= \big(\beta V[t] + (1-\beta) I[t]\big)\big(1 - S[t]\big) \quad \text{(v2)}
\end{aligned}
$$

[1] Note, I use square brackets [] to refer to discrete time, and rounded brackets () to refer to continuous time, for a given function.

The membrane potential dissipates by a decay factor $0 \leq \beta \leq 1$ at every time step, which is related to the membrane time constant τ as $\beta = \exp(-\Delta t/\tau)$; and a spike is emitted if the membrane potential value reaches the firing threshold, which is implemented using the Heaviside step function H.

$$S[t] = H(V[t] - 1) = \begin{cases} 1, & \text{if } V[t] > 1 \\ 0, & \text{otherwise} \end{cases}$$

The biological accuracy of the LIF model can be further enhanced by adding an adaptation component, which reflects the adaptive behaviour of the many neurons that exhibit increasing interspike-intervals in response to a steady current injection. This extended model is referred to as the adaptive leaky integrate-and-fire (ALIF) model [110, 128]. Adaptation is incorporated by raising the model firing-threshold $\theta[t]$ following every spike $S[t-1]$, which decays exponentially to baseline $\theta = 1$ in the absence of any spikes (using decay factor $0 \leq p \leq 1$ and adaptation scalar d).

$$\begin{aligned} \theta[t] &= 1 + da[t] \\ a[t] &= pa[t-1] + S[t-1] \end{aligned} \tag{1.2}$$

Both the LIF and ALIF provide qualitatively good fits to spikes of cortical neurons in response to the same input current injections [127, 128] (Figure 1.7B and C). However, both of these models abstract away how the input currents to a neuron are generated, namely, by the changes in the permeability of the membrane to different ions. The Hodgkin-Huxley model [134] and the Connor-Stevens model [135] extend upon the LIF and ALIF models by incorporating the evolution of different ion concentrations, and their effect on a neuron's membrane potential. Additionally, the realism of all spiking models can be further enhanced by considering the morphology of a neuron and modelling the membrane potential at different points on a neuron's surface [136, 62] - which are referred to as multi-compartment models. However, models of increasing complexity require more computational resources and time to simulate. Furthermore, integrate-and-fire-type models approximate the dynamics of more biologically detailed neural models, like the Hodgkin-Huxley model [137, 138]. Thus, LIF and ALIF models have become the status quo for modelling neural circuits and studying their dynamics [139, 132, 131, 133, 128, 121].

The model architecture of a neural circuit does not only depend on the chosen neuron model but also on how those neuron models connect and interact with each other. A neuron model is typically referred to as a neuron or a unit within a model of a neural circuit. Circuit models (also known as population models) can consist of a single layer of units, processing input in parallel; or a hierarchy of layers, which process input sequentially. Processing does not strictly need to be feedforward (*i.e.* from one layer to the next), but can also be more complex, where units in a layer could be recurrently connected, or form feedback connections to units in earlier layers. Lastly, constraints can also be applied to the population of units, segregating them into distinct subpopulations of inhibitory and excitatory units (just like the excitatory and inhibitory neurons in the brain). Now that we know how to construct a model architecture of a neural circuit, we next look at some popular normative objectives that a model of the visual system might employ.

1.3.2 Model objective: theories of visual processing

Once we have a model architecture, the next goal is to define its normative objective: what task should the model fulfil? For example, for a given sensory input, the model should produce some desired and quantifiable output. Typically, normative objectives of visual processing are decomposed into sub-objectives: the primary objective usually defines the transformation on the sensory input and the secondary objectives place any additional constraints on the model, such as metabolic-like constraints. Various theories of visual processing have been proposed and examined over the last decades.

A classic theory of visual processing is the so called efficient coding hypothesis [147, 85]. This theory stipulates that sensory systems, like the visual system, employ a neural code that efficiently represents incoming sensory input, with the least number of required spikes. Although efficient coding has been an influential theory, it does not specifically outline how features within the sensory stimulus are encoded - apart from stating that the encoding is done efficiently so as to extract maximum sensory information. Arguably, we are more interested in understanding what goal a sensory system is performing, rather than knowing that it is performing some goal efficiently - which in the context of evolution would be assumed. For example, making an analogy to the heart: we can measure the number of beats of the heart (like the spikes of neurons) and argue that the heart is beating in an efficient manner. Such explanation, however, does not inform us why the heart is performing its beats. In some sense, all normative theories of the brain are descendants of the efficient coding hypothesis - the sensory transformation is done

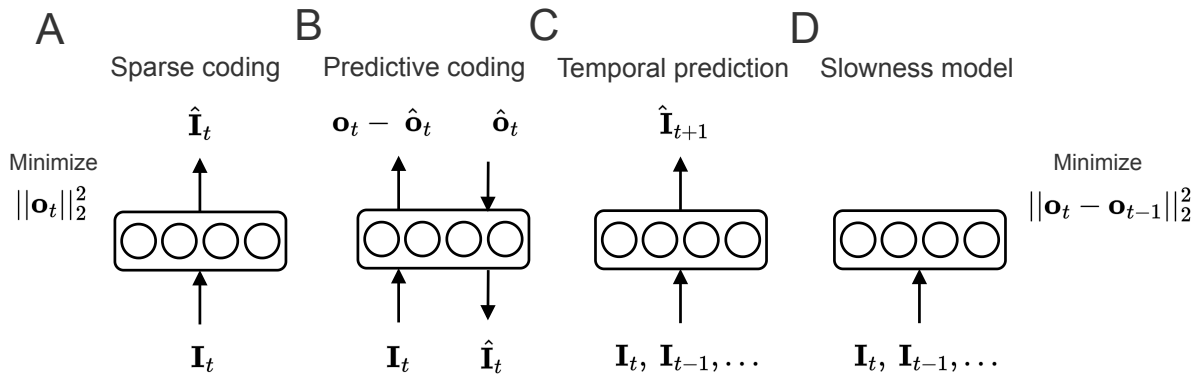

Fig. 1.8: **A.** The sparse coding model [140] consumes input $\mathbf{I}_t$, and reconstructs the input as $\hat{\mathbf{I}}_t$ from the unit activities $\mathbf{o}_t$, with a sparseness constraint on the unit activities (*i.e.* with $||\mathbf{o}_t||_2^2$ being minimized). **B.** The predictive coding model [141] - here represented as a single layer of units - consumes input $\mathbf{I}_t$ and outputs prediction errors $\mathbf{o}_t - \hat{\mathbf{o}}_t$, between the unit activities $\mathbf{o}_t$ and the predicted unit activities $\hat{\mathbf{o}}_t$, where the predicted unit activities are obtained from the layer above via feedback connections. In the case of a single layer, the predicted unit activities are equivalent to the reconstructed input $\hat{\mathbf{I}}_t$. **C.** The temporal prediction model [142, 106] consumes a sequence of inputs $\ldots, \mathbf{I}_{t-1}, \mathbf{I}_t$ and generates a prediction of future input $\hat{\mathbf{I}}_{t+1}$. **D.** The slowness model [143, 144, 145, 146] consumes a sequence of inputs $\ldots, \mathbf{I}_{t-1}, \mathbf{I}_t$ and does not generate any output. Rather, a slowness constraint is placed on the unit activities $\mathbf{o}_t$, which ensures the unit activities vary slowly between consecutive time steps (by minimizing $||\mathbf{o}_t - \mathbf{o}_{t-1}||_2^2$). Additional modelling constraints are also applied, ensuring that the unit activities have a zero mean and are of unit variance (to avoid learning trivial solutions *i.e.* a constant output); unit activities are also enforced to be decorrelated (such that units encode different features within a sensory stimulus). For **A-D**: spatial input $I_t \in \mathbb{R}^{H \times W}$ (of height H and width W at time step t); predicted spatial input $\hat{\mathbf{I}}_t \in \mathbb{R}^{H \times W}$; population unit activities $\mathbf{o}_t \in \mathbb{R}^N$ (for N units at time step t); and predicted population unit activities $\hat{\mathbf{o}}_t \in \mathbb{R}^N$.

efficiently - but these different theories differ in exactly what is to be done efficiently and how efficiency is defined.

The sparse coding [140] and predictive coding [148, 141, 149] hypotheses extend the efficient coding hypothesis, outlining a primary functional objective for the sensory transformation, and implementing these ideas in a model. The sparse coding model posits that the visual system efficiently encodes current sensory input like a brief glimpse of a scene projected onto the retina [150] (Figure 1.8A). This efficiency is achieved by only having a few neurons active at any point in time (*i.e.* sparse activity) from which the current sensory input is reconstructed. In the model, the sensory reconstruction is implemented by learning a set of bases, which linearly combine, using weighting amplitudes, to reconstruct the sensory input [140]. In addition, when optimising the model for sensory reconstruction, a constraint is applied of using the least number of bases as possible. That is, ensuring that most of the weighting amplitudes are zero and only a few are non-zero. When visualized, the sparse coding model's bases reveal structure and spatial tuning, reminiscent of those observed in the simple cells of V1 [140]. Similar ideas to the sparse coding model have also been explored in other modelling studies, such as the independent component analyses (ICA) framework [151]. These models differ in their number of bases, where ICA employs a number of bases equivalent to the number of input dimensions (a critically complete code); and sparse coding assumes a number of bases that outnumber the number of input dimensions (an overcomplete code). Like sparse coding, ICA models also develop simple cell-like spatial RFs [151].

The predictive coding hypothesis stipulates that incoming sensory information is efficiently processed, where neurons only signal unpredictable elements within sensory input (Figure 1.8B). Neurons do not signal the elements predictable by the model within the sensory stimulus and thus reduce statistical redundancies, yielding an efficient neural code [149]. Models employing predictive coding have been shown to exhibit retinal-like and V1-like RFs. Specifically, predicting the centre pixel value from the surrounding region pixel values in an image patch using linear filters, yields filters (RFs) which closely resemble the RFs of retinal ganglion cells [148]. In a more complex instantiation, the predictive coding framework has also been extended into a multi-layered network: units in a layer l predict the activity of the units in the lower layer $l-1$. Units in layer l signal the prediction errors (*i.e.* the difference in the predicted and actual unit activities from layer l) to higher layers $l+1$ along feedforward connections, and signal predicted unit activities to the lower layer $l-1$ along feedback connections, where the lower layer then subtracts these predictions from the activity of its units. Such multi-layered predictive coding yields unit filters which somewhat resemble the spatial RFs of V1 simple cells [141]. Further explorations of predictive coding in multi-layered networks revealed additional V1-like tuning

properties, such as orientation tuning, spatial and temporal frequency tuning [152], and aspects of higher visual areas, such as object identification [153].

Both the sparse coding and early predictive coding investigations have primarily focused on efficiently coding the current sensory stimulus. Another branch of normative theories argues that the brain efficiently predicts future sensory stimuli (or related factors), rather than efficiently encoding the present. The notion of the brain performing prediction dates at least as far back to the German physicist and physiologist Hermann von Helmholtz [154], who hypothesized that the brain takes into account its own actions when creating a perception of the world. Self-motion, such as the rapid saccades of our eyes, does not produce jittering imagery in our perception. This phenomenon, as Helmholtz suggested, is because the brain accounts for predictable motion, resulting in a perception that primarily reflects unpredictable elements within the world. Thus, the brain needs to maintain a record of its own actions, and contain an internal model capable of predicting the effect of these actions on the visual experience, in order to cancel out the visual perturbations resulting from self-motion. In support of these ideas, a computational study simulated an agent in a 3D visual environment and trained a deep neural network to predict its self-motion parameters from the visual stimuli, where the authors found the model units to account for the selectivity of neurons from the dorsal visual stream of non-human primates [155].

A brain capable of prediction has an evolutionary advantage, as it embodies a notion of cause-and-effect in the world. For example, you can likely extrapolate that a fast-moving flying object heading towards your head will impact your head unless you move out of the way. Thus, by prediction, you avoid a possibly fatal head injury by simply stepping aside. More generally, a system that can predict future inputs from its past is likely to contain representations reflective of the underlying causes in the world (*e.g.* the speed and direction of fast-moving flying objects). In addition, such sensory processing likely reduces the load on sensory processing, as it would discard any information that is irrelevant to predicting the future [156]. There is also a physiological reason why the brain may perform prediction of future input. There is a minimum transduction and transmission latency of approximately $\sim 30 - 40$ms from the photoreceptors to V1 [102, 103, 100, 101]. Thus, a system capable of overcoming these latencies using prediction can respond more rapidly to a potentially threatening situation than a system that is optimized to encode the present.

Predictive coding models have also been explored in the temporal domain of prediction [157, 158] - in fact, original predictive coding models were proposed in the context of temporal prediction [159, 160]. Predictive coding models have also demonstrated neural-like tuning

properties when extended to predicting future sensory input from past sensory input. A multi-layered predictive coding network optimized for such sensory prediction has been shown to develop spatial and temporal tuning properties found throughout the visual pathway, and visual-cortex-like response dynamics to perceptual motion illusions [161]. Even when predicting the current sensory input - and not the future sensory input - using a linear weighting of past sensory input, yields filters with similar spatial and temporal profiles to those of retinal ganglion cells [148].

Non-predictive-coding models which predict the future sensory input in natural movies also exhibit many V1-like tuning properties. Single-layer networks trained to predict future patches of movie frames from the past exhibit units with V1-like spatial [142, 106] and temporal RF properties [106] (Figure 1.8C). Furthermore, such temporal prediction objective can be extended from a single-layer to a multi-layered network, where layer l predicts the future unit activities from layer $l-1$. Such multi-layered model exhibits units with both simple and complex cell-like tuning properties, capable of predicting V1 macaque response to natural images and movies [162]. Temporal prediction has also been explored in the latent space (*i.e.*, predicting future hidden unit activity rather than predicting future pixel observations). These latent prediction models have been able to account for the ventral and dorsal tuning properties of the visual pathway in mice [163], as well as primate frontal cortex neural dynamics whilst playing a pong-like ball interception game [164].

Other normative theories that are related to predicting future sensory stimuli are the information bottleneck and temporal slowness models. The information bottleneck method [165] compresses one random variable (like past sensory input) to maximize information [166] about another variable (like future sensory input). In the context of a sensory system, like the visual system, this suggests that features with maximum mutual information between the past and future stimulus might be encoded [167, 168]. Models optimized for such an objective have also been shown to account for V1-like RFs [168]. The temporal slowness models posit that the brain encodes features within the sensory world that vary slowly over time, such as the motion of a moving ball, instead of registering every varying light intensity value in a scene [143, 144, 145, 146] (Figure 1.8D). These models place a slow-varying constraint on the activation of the model units by optimizing their temporal derivatives to be as close to zero as possible. Furthermore, unit activities are enforced to be a zero mean and of unit variance (to avoid learning trivial solutions *i.e.* a constant output), and are also enforced to be decorrelated (such that units encode different features within sensory stimulus). Past explorations of these models exhibit units with simple cell-like [145] and complex cell-like [144] tuning properties. It has also been suggested that models employing the slowness constraint encode features within a sensory stimulus that - like

the temporal prediction models - are predictive of the future [146]. Now that we know what sensory transformations a model of the visual system may employ, we next look at how such transformations can be instilled in the model's weights and parameters.

1.3.3 Model training: learning the neural parameters and connectivity

Once the model architecture and normative objective are established, setting the model's weights and parameters to achieve the normative goal is the next step. One way to do this would be to hardcode the model weights, which used to be a mainstream method for developing artificial vision models [169, 170, 171, 172]. However, determining how to properly configure the model weights to perform some normative goal is difficult, and a more efficient technique - that is reminiscent of the brain - is to learn the model weights. In the 1980s, a key learning algorithm, called backpropagation [173], was developed, which enables models to learn their own weights in an experience-driven manner. Here, a dataset of various inputs and desired outputs is required, where the backpropagation algorithm specifies how to adjust the model weights, such that the model correctly maps the input data to the desired output data. This is achieved by minimization of a loss function $\mathcal{L}$, which quantitatively specifies how close the model outputs are to the desired output data. Backpropagation iteratively calculates how to adjust the model weights by calculating their partial derivatives with respect to the loss function and updating the weights in the direction of their partial derivatives by a small perturbation η - known as the learning rate. This algorithm has underpinned the rise of deep learning models over the last decade, with the efficacy of the algorithm arising from the large amounts of available training data, and the ability to quickly process such training data using hardware accelerators such as Graphics Processing units (GPUs) [42].

The applicability of the backpropagation algorithm - also often referred to as backprop - does not directly extend to training models of increasing biological realism, such as spiking neural networks (SNNs). That is because backprop requires that the system it is applied to is differentiable, but SNNs violate this prerequisite. This is because the spiking non-linearity, the Heaviside step function, has a derivative of zero, except at the transition point, where the derivative is undefined. The development of SNN training algorithms remains an active area of research. However, a few SNN training methods have been developed that employ backprop in some form, with promising results. [2]

[2] It is worth pointing out that modern neural networks routinely employ non-differentiable activation functions such as the rectified linear unit $f(x) = \max(0, x)$ whose derivative is undefined when $x = 0$, thus violating the differentiability criteria of backprop. However, this is not a problem in practice, as the non-linearity has a valid

One such method, known as shadow training, trains a differentiable non-spiking artificial neural network (ANN), which can then be mapped to an SNN, enforcing the firing rates of the SNN neurons to approximate the activations of the ANN units [174, 175, 176, 177]. While this method permits SNNs to be trained on challenging large-scale image classification datasets, it endures various shortcomings: firstly, network mappings require sampling from long simulation durations (which is slow) [178]; secondly, the conversion process results in a loss in network performance (although see ANN-SNN training coupling methods [179, 180, 181]); and thirdly - and most relevant to neuroscience investigations - this method imposes a rate code on SNNs. This is problematic as the brain likely makes some use of the timing of individual spikes for sensory processing [182, 183].

Another approach is to directly train SNNs using approximated gradients. Gradients can be estimated by randomly perturbing weights [184, 185]. However, this method requires many repeated weight perturbations in order to estimate gradients, which is slow and does not scale for growing network sizes. The non-differentiable problem of SNNs can be overcome by employing latency learning: estimate gradients at the timing of spikes, as unlike spikes, time is continuous [186, 187, 188, 189]. This approach, however, enforces the constraint that units can spike at most once and the method experiences training instabilities (although see [190, 191] for the latest developments). A popular approach to enabling gradient calculation using backprop in SNNs is by replacing the undefined derivative of the spike function with a surrogate function [192, 193, 194, 195] - known as surrogate gradients. This training approach has been shown to be robust to a variety of surrogate functions, such as the piece-wise linear, sigmoid, and sigmoid-like functions [133]. Notably - and of applicability to neuroscience - this method allows networks to learn temporal, rather than rate coded, spike dynamics [196, 197, 198]. Although this method has become the status quo for directly training SNNs, it is very slow, as it requires the sequential simulation of network dynamics through time [128].

1.3.4 Model validation

Once we have developed a model of the visual system - that is, we have picked a model architecture and trained it with respect to some normative objective - the next step is to verify how visual-system-like the model is: what relations can the model account for in the visual

derivative when x is not zero, permitting the flow of gradients in the network; and the function's derivative can be set to zero when $x = 0$ to circumvent the issue of non-differentiability. Like the rectified linear unit, the Heaviside step function also has a derivative when x is not zero. However, this derivative is zero, and thus no gradients are propagated in the spiking network.

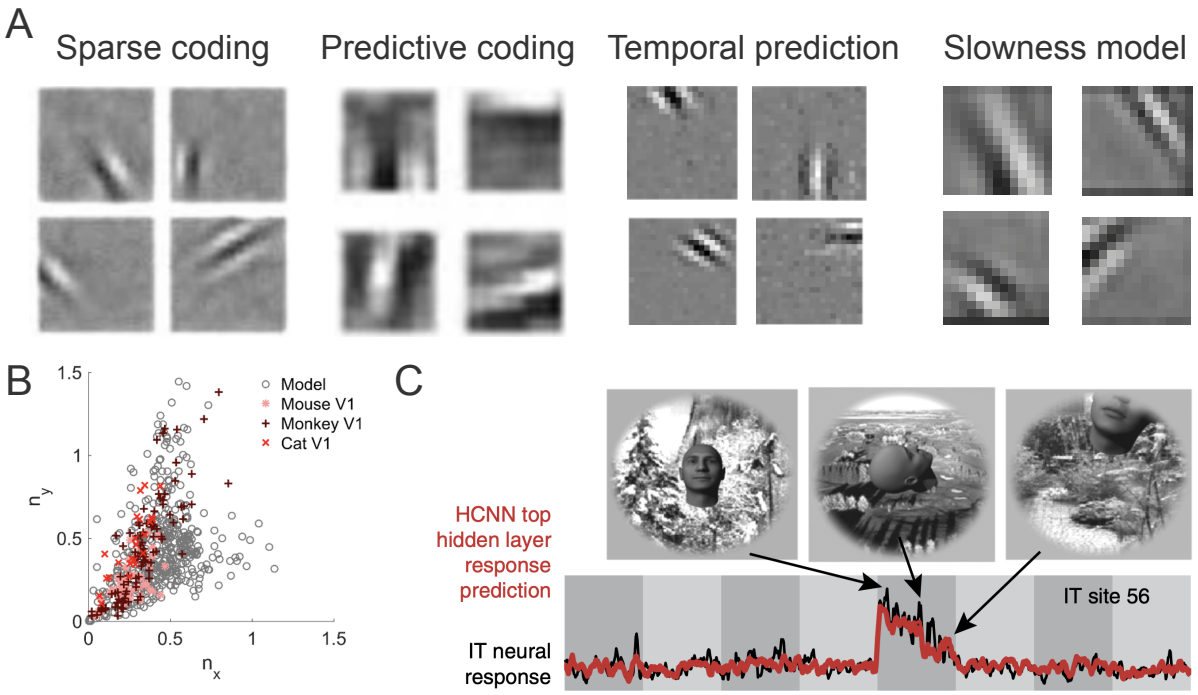

Fig. 1.9: **A.** Similar spatial V1-like RFs from different normative models (sparse coding RFs adapted from [140]; predictive coding RFs adapted from [141], temporal prediction RFs adapted from [106]; and, slowness model RFs adapted from [145]). **B.** Quantitative comparison between V1 RFs from different animal species and model unit RFs (trained using temporal prediction). The x-axis is a measure related to the number of subfields in an RF and the y-axis is a measure related to the length of those subfields (adapted from [106]). **C.** Neural response for a single neural site in IT (black line) and respective deep learning model (HCNN) predictions (red trace). Different images (like the faces) elicit different neural responses (adapted from [199]).

system? One popular approach is to compare the model unit RFs to those found in the visual system. For example, a model of V1 would be compared to the RFs of V1 neurons. While such qualitative examinations have proven useful, they pose a problem. Various normative models - *e.g.* sparse coding [140], predictive coding [148, 141, 149], temporal coding [142, 106] and slowness models [143, 144, 145, 146] - have been able to account for V1-like simple cell spatial RFs, consisting of orientated parallel bars of excitatory and inhibitory regions (Figure 1.9A). Thus, the question is raised: how do we determine which model is most V1-like?

A more technical approach to determining the V1-likeness of a model is to quantify the model unit RF properties and contrast these to the quantified RF properties of neurons - rather than qualitatively comparing model unit RFs to neuron RFs. As previously mentioned, the tuning properties of retinal RFs can be quantified using a two-dimensional Gaussian [75, 76], or a difference of two-dimensional Gaussians [77], and the tuning properties of V1 simple cells can be quantified using a Gabor function [105]. The resulting function parameters can then be contrasted between the population of model unit RFs and the population of corresponding visual neuron RFs (Figure 1.9B), which can answer questions like: do the RF shapes match? do their sizes match? does their orientation match? do their temporal and spatial frequencies match?

Most models - in particular models of V1 - have been able to account for both the qualitative and quantitative spatial tuning properties of simple cells. Thus, to further distinguish which model is most brain-like, requires explaining additional neural phenomena - where the model which is able to account for most phenomena is arguably the best model of the particular neural circuit being modelled. For example, the temporal prediction model has been able to account for not only the spatial but also the temporal tuning properties of V1 neurons [106] - a neural property not readily captured by other models. It could therefore be argued, that such a model is more V1-like than other models, given its ability to account for additional neural phenomena. Furthermore, models of increasing biological realism are arguably a better model of the brain compared to models that overlook the biological realism - in particular if both models are able to account for the same neural phenomena. For example, a recurrent model of V1 that poses no constraints on the recurrent connectivity, is less biologically plausible than a recurrent model of V1, which places connectivity constraints, such as Dale's law [111], on the recurrent connectivity. If both models account for the same neural phenomena, then the model with the connectivity constraints would arguably be a better model of V1. Additionally, models of increasing biological realism - like the model with the connectivity constraints - can be contrasted to more neural phenomena than models that overlook the biological constraints (discussed in the next section).

Another means of evaluating the similarity of a model to the visual system is by validating its ability to predict neural responses to unseen arbitrary and natural stimuli. Models with better neural prediction capabilities are arguably a better representation of the visual system than models with worse prediction capabilities [200]. Historically, early models of the visual system (like the linear-nonlinear model) have been able to account for responses to artificial stimuli, such as moving gratings. These early models, however, fail to accurately predict neural responses to more complex and arbitrary stimuli [51]. It has been argued the visual system evolved to process complex natural stimuli, hence any model that purports to explain the visual system should be able to explain responses to such stimuli [200]. In the last few years, we have witnessed a big improvement in neural response predictions to complex natural stimuli throughout the visual system. These prediction improvements have been made possible through the use of mostly large-scale and goal-driven ANNs. Best performance is usually obtained by fitting a readout model (like a linear or linear-nonlinear model) from the intermediate layer representations of an ANN, pre-trained on a large supervised image classification task [199]. These models have been able to fairly well predict the neural responses to natural images in V1 [51], V4, and IT [50] (Figure 1.9B).

1.3.5 Model shortcomings and new frontiers

The visual system is a complex system within the brain, comprised of many idiosyncratic phenomena, like how neurons operate, connect, and respond to certain sensory signals. Deep learning models of vision have proven to be powerful, reaching uncanny performance on image-classification, image-segmentation, and image-description tasks [47, 48, 41]. Although our visual system also supports such functionality, it arguably does so through different mechanisms than deep learning models. In particular, deep learning models largely abstract away the finer operational details of the brain; are largely trained using supervised signals (*e.g.* learning to classify images); and employ non-biologically constrained model architectures.

Normative models of the visual system are more aligned with how the brain may have evolved. Instead of learning model weights and parameters using a deep-learning-like supervised signal, these models rather instantiate model connectivity through the means of an unsupervised signal (*e.g.* efficiently encoding the current or future sensory stimulus). Unlike deep learning models, these models also try to explain what information neurons are signalling to each other, such as error signals or predictive signals, as stipulated in the predictive coding and temporal prediction frameworks respectively [141, 106]. However, to date, normative models of the visual system

omit key physiological constraints, namely, the all-or-nothing spike output of neurons and their excitatory and inhibitory connectivity constraints.

Increasing the biological realism of models is important. Apart from the models being more brain-like, additional biological modelling constraints allow models to be compared to an increasing number of experimental phenomena. For example, employing spiking units in a model of the retina would permit the model to be contrasted to the latency-like code of retinal ganglion cells [201]; or including spiking units in a model of V1 would permit the spike-statistics of the model to be contrasted to those of V1, where V1 exhibits a sparse and irregular spike code [27]. Such comparisons are not directly possible in non-spiking models. In addition, including separate populations of excitatory and inhibitory units in a model of V1 creates many more possibilities for comparison to the biology, namely contrasting the model units to the distinct physiological and tuning differences between the excitatory and inhibitory neurons in V1 (see section 1.2.3).

Predicting neural responses to arbitrary natural stimuli has played an important role in validating models of vision over the last few years [199]. However, most of these natural stimulus datasets have been recordings of visual neurons in response to natural images. Arguably, it would be better to validate how well models can predict neural responses to natural movies, given the abundance of temporal tuning and motion tuning properties of visual neurons.

1.4 Thesis overview

- **Chapter 2: Beyond deep learning: a neuro-inspired approach to modelling primary visual cortex.** A rich interplay between the spiking activity of excitatory and inhibitory neurons within primary visual cortex (V1) allows us to perceive the visual world. Understanding the mechanisms of visual processing thus requires modelling the spiking dynamics of neurons in their excitatory and inhibitory subgroups - both of which mainstream deep learning and normative models of vision exclude. Recent studies have produced models that exhibit V1-like tuning when optimized to predict the sensory future. Inspired by these investigations, we trained a model of excitatory and inhibitory spiking units to predict the sensory future on the timescale of the latencies of V1 responses, under metabolic-like constraints. We found our model to exhibit V1-like spike statistics, simple and complex cell-like tuning, and key physiological and tuning differences between excitatory and inhibitory neurons. These findings suggest our model to be an advance in achieving a more detailed and precise representation of V1's visual processes *in silico* (unpublished).

- **Chapter 3: Crystal balls as eyes: retina optimised for prediction across animal species.** The retina's role in visual processing has traditionally been viewed as two extremes: an efficient compressor of incoming visual stimuli akin to a camera [68, 202, 203, 204, 205], or as a predictor of future stimuli [206, 207, 208, 209]. Addressing this dichotomy, we developed a biologically-detailed spiking retinal model trained on natural movies under metabolic-like constraints to either encode the present or predict future scenes. Our findings reveal that when optimized for efficient prediction, the model not only captures retina-like receptive fields and their mosaic-like organizations, but also exhibits complex retinal processes such as latency-coding [201], motion-anticipation [92], differential tuning [91], and stimulus omission responses [93, 210]. Notably, the predictive model also more accurately predicts retinal ganglion responses across various animal species to natural images and movies. Our findings demonstrate that the retina is not merely a light compressor, but rather provides the brain foresight into the visual world (unpublished).
- **Chapter 4: Addressing the speed-accuracy simulation trade-off for adaptive spiking neurons.** The adaptive leaky integrate-and-fire (ALIF) model is fundamental within computational neuroscience and has been instrumental in studying our brains *in silico*. Due to the sequential nature of simulating these neural models, a commonly faced issue is the speed-accuracy trade-off: either accurately simulate a neuron using a small discretisation time-step (DT), which is slow, or more quickly simulate a neuron using a larger DT and incur a loss in simulation accuracy. Here we provide a solution to this dilemma, by algorithmically reinterpreting the ALIF model, reducing the sequential simulation complexity and permitting a more efficient parallelisation on GPUs. We computationally validate our implementation to obtain over a $50\times$ training speedup using small DTs on synthetic benchmarks. We also obtained a comparable performance to the standard ALIF implementation on different supervised classification tasks - yet in a fraction of the training time. Lastly, we showcase how our model makes it possible to quickly and accurately fit real electrophysiological recordings of cortical neurons, where very fine sub-millisecond DTs are crucial for capturing exact spike timing (published in NeurIPS 2023).
- **Chapter 5: Discussion.** This chapter contains general high-level discussions regarding the modelling work, its shortcomings and possible research avenues for future work.

2 Beyond deep learning: a neuro-inspired approach to modelling primary visual cortex

2.1 Introduction

Our visual system grants us a crucial sense, providing evolutionary advantages in identifying food and predators, and allowing us to perceive even the most unusual of scenes, such as the perplexing illusions in Escher's art, or the sight of a toy brachiosaurus perched atop a brick wall. Neuroscience has endowed us with decades' worth of experimental findings underlying the visual processes throughout the brain. Neurons in primary visual cortex (V1), the first cortical stage of visual processing, are tuned to the orientation and motion of moving edges and shapes [11, 12, 13]. They appear to implement such visual processing through a sparse and irregular spike code, carefully balanced by the interactions between the cortex's excitatory and inhibitory neural subpopulations. These neural phenomena underly our brain's ability to process diverse and complex natural stimuli. Nonetheless, despite our efforts to understand these processes, we still lack a comprehensive theory or model that can explain all these phenomena.

One approach to deciphering the visual processing in V1 involves normative modeling [125], whereby a model is constructed and optimized for a certain goal presumed to underly V1's functionality. Various normative models have been developed, exploring a broad spectrum of objectives, ranging from efficient encoding to forms of prediction of sensory stimuli. These models capture some aspects of V1, such as its spatial [140, 151, 141, 152, 142] and temporal [211, 106, 168] simple cell-like RF properties, or complex cell-like tuning [212, 143, 144, 162]. Although these models loosely mimic V1's biological circuitry, they omit the brain's key physiological unit of computation: the spike. Studies on V1 spiking models have successfully replicated aspects of V1-like spike statistics and responses to motion [213, 214, 215, 27, 216]. However, in contrast to normative models, these models rely on hard-coded and manually-tuned parameters, and do not endeavor to explore the underlying function implemented by V1. Therefore, there remains a significant gap in developing biologically realistic spiking models of

30

V1 that are trained for specific functions, which is crucial for a comprehensive understanding of V1's computational role.

In this study, we developed a normative model comprised of recurrently connected spiking units. In line with Dale's law [111], units were separated into distinct excitatory and inhibitory subpopulations. Inspired by previous works [106, 162], we optimized our model to predict the sensory future on the timescale of the transmission latencies from retina to V1, using past population spike activity. Similar to earlier studies, our model demonstrates simple and complex cell-like tuning. Furthermore, our investigation uniquely reveals key V1-like differences between excitatory and inhibitory neurons. Specifically, the inhibitory units in our model are characterized by higher firing rates, broader orientation tuning, more complex cell-like responses, and stronger connections from excitatory units, and they operate on faster timescales than their excitatory counterparts. Additionally, akin to neurons in V1, the model units display sparse and irregular spike responses with low correlations between units, maintaining a detailed balance of excitatory and inhibitory input currents in response to natural movie stimuli. Our model not only makes several predictions about V1 processing, but also sets the stage for further unravelling the visual code through simulation.

2.2 Results

2.2.1 The spiking V1 model

We implemented our model as a single hidden layer of recurrently-connected spiking units and adopted the leaky integrate-and-fire (LIF) model to mimic real neural spiking dynamics [110] (Figure 2.1A). Similar to the ratio of excitatory-to-inhibitory neurons in V1 [113, 114, 112, 23], we set 85% of these units as excitatory and the remaining 15% as inhibitory. The model's input consists of a sequence of 20×20px patches obtained from retina-filtered natural movies recorded at 120 frames-per-second. At each time step, the model linearly translates a 168ms (20 frames) span of stimulus history into a distinct input current to each unit. To emulate the transmission latencies from retina to V1 [102, 103, 100, 101], we applied a mask to exclude the most recent 42ms (5 frames) of stimulus in each of these mappings - the temporal prediction span. This effectively reduces the input span to 126ms (15 frames), where the model was trained to predict as to overcome this latency by linearly reconstruct the most recent movie patch using a 16ms (2 frames) span of past population spike activity (Figure 2.1B). This training process determined the model's connectivity weights and the membrane time constant of each spiking unit.

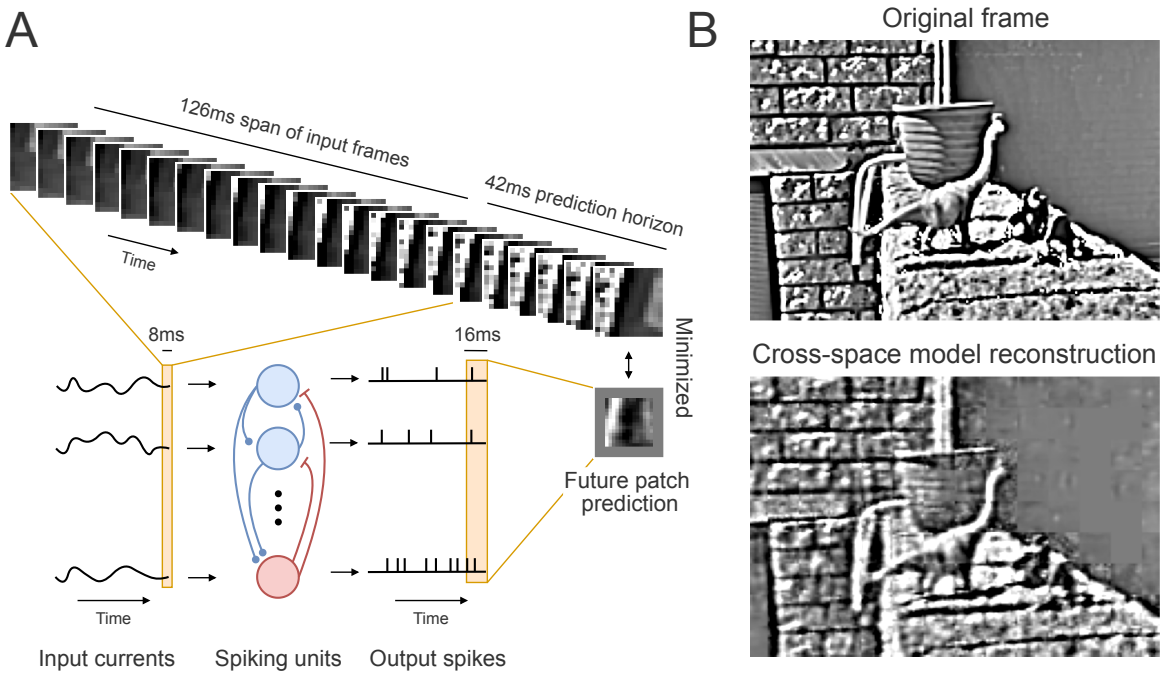

Fig. 2.1: **A.** Schematic of our model, which at every time step linearly maps a 126ms span of natural movie frame patches of the past into a distinct input current value to every spiking unit. A 16ms span of population output spikes is linearly mapped to create a prediction of the spatial movie patch 42ms into the future. Blue units are excitatory and red units are inhibitory. **B.** An example movie frame of a toy dinosaur sitting on a brick wall, which was not included in the training data (top), with the model's reconstruction using multiple patch predictions across different spatial locations (bottom).

To emulate the metabolic energy constraints of the brain, we also trained our model to reduce a metabolic-like cost whilst performing the frame prediction (see Methods). This approximated the activity-induced cost at the model weights, similarly to the energy expenditure of neural synapses. We made an assumption about the brain in our metabolic-like cost function, namely that inhibitory V1 neurons consume less energy than their excitatory counterparts per unit work (see Discussion). We found this assumption to be important for obtaining all the V1-like tuning differences between the excitatory and inhibitory model units. Lastly, all training was performed using an adaptation of the backpropogation algorithm [173], known as surrogate gradient learning [198], which permits gradient descent in non-differentiable spiking networks.

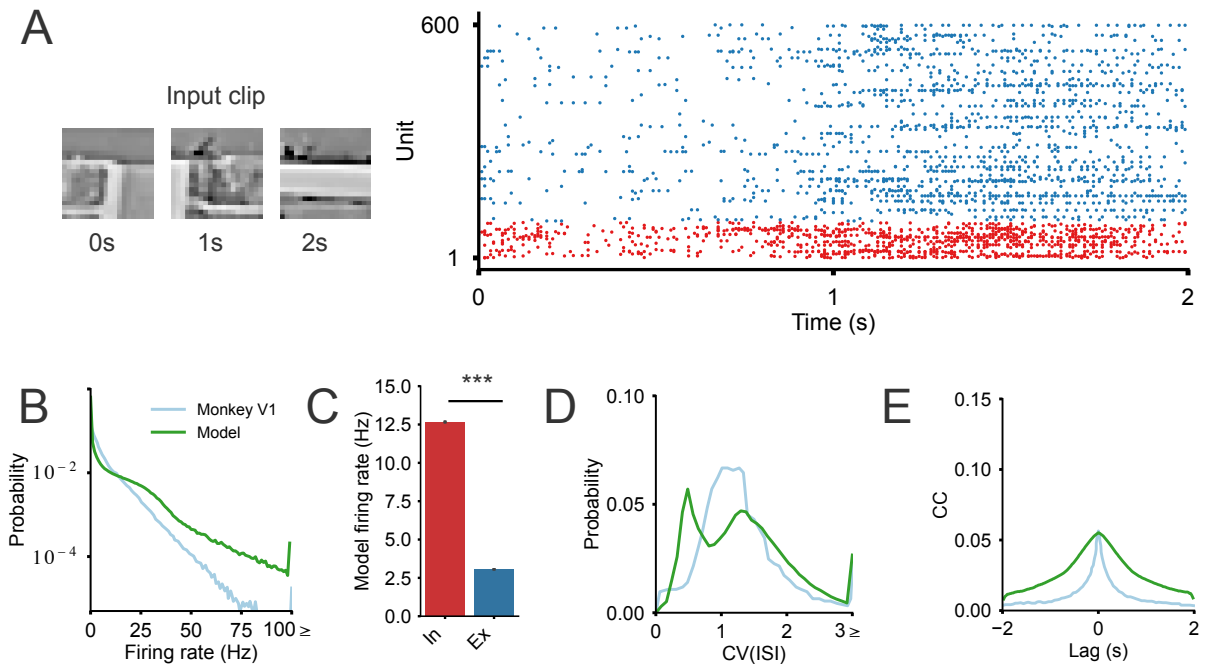

Fig. 2.2: **A.** Input movie clip (left) and resulting model spike raster (right). Blue and red dots correspond to the excitatory and inhibitory units, respectively. **B.** Firing rate distribution of the model units and monkey V1 neurons. **C.** Difference in the model's firing rates between the inhibitory (In) and excitatory (Ex) units. Bars plot the mean and standard error over units and input clips, with statistical significance assessed using the Mann-Whitney U test ($^{***}p < 0.001$). **D.** Coefficient of variation of the interspike intervals CV(ISI) distribution of the model units and monkey V1 neurons. **E.** Cross-correlogram of the model units and monkey V1 neurons. Monkey V1 data in all plots was obtained from [27] in response to monkey's watching segments of the Star Wars movies.

2.2.2 Emergence of a V1-like spike code

We examined the model's spike activity in response to held-out natural stimulus clips not used for training, and found a number of V1-like response characteristics to emerge. We observed the model's spike responses to be sparse and diverse to the natural stimulus input, with variability in the number and timing of spikes across units (Figure 2.2A). Further quantifying the spike responses revealed a firing rate distribution with an exponential fall-off (Figure 2.2B), as has similarly been reported in monkey [27] and cat V1 [217]; and the average model's firing rate was comparable to mean V1 firing rates observed across different animal species in response to natural stimuli (model 4.50 ± 10.11; anesthetized monkey V1: 5.06 ± 0.75Hz [27]; anesthetized cat V1: 3.96 ± 3.61Hz [217] and awake cat V1:8.9 ± 7Hz [218]). Additionally, we found the inhibitory units to exhibit a higher mean firing rate compared to the excitatory units (Figure 2.2C). Such differences in firing rates between excitatory and inhibitory neurons to visual stimulation have also been reported in mouse V1 [28].

We also found the model's spike responses to be highly variable, exhibiting a low degree of correlation between the units. We examined the variability in the spike responses using the coefficient of variation of the interspike intervals CV(ISI) (Figure 2.2D), which is defined as the ratio of the ISI standard deviation to the ISI mean. A spike train with a CV(ISI) below unity is more regular than a spike train with a CV(ISI) above unity. We observed a large range of variability in the spike trains, with sampled spike trains exhibiting more irregular (62% with CV(ISI)≥ 1) than regular (38% with CV(ISI)< 1) firing patterns, reminiscent of the spike variability observed in cat [219] and monkey V1 [220, 27].

We quantified the correlation between the different unit spike trains, by calculating the spike train cross-correlogram, which measures the correlation between randomly-sampled unit pairs at different time lags (Figure 2.2E). We found the model units to have a low peak correlation ($CC = 0.05$ at lag 0s), with the correlation symmetrically decaying for increasing and decreasing lag. Similar correlation statistics have also been reported in anesthetized monkey [27] and in awake mouse V1 [221]. However, we found the strength of the correlation to vary depending on the specific movie stimulus input, aligning with similar reports in monkey V1 [27]. We observed that movie stimuli that elicited more extensive population activity resulted in higher correlations between the model units (see Supplementary material).

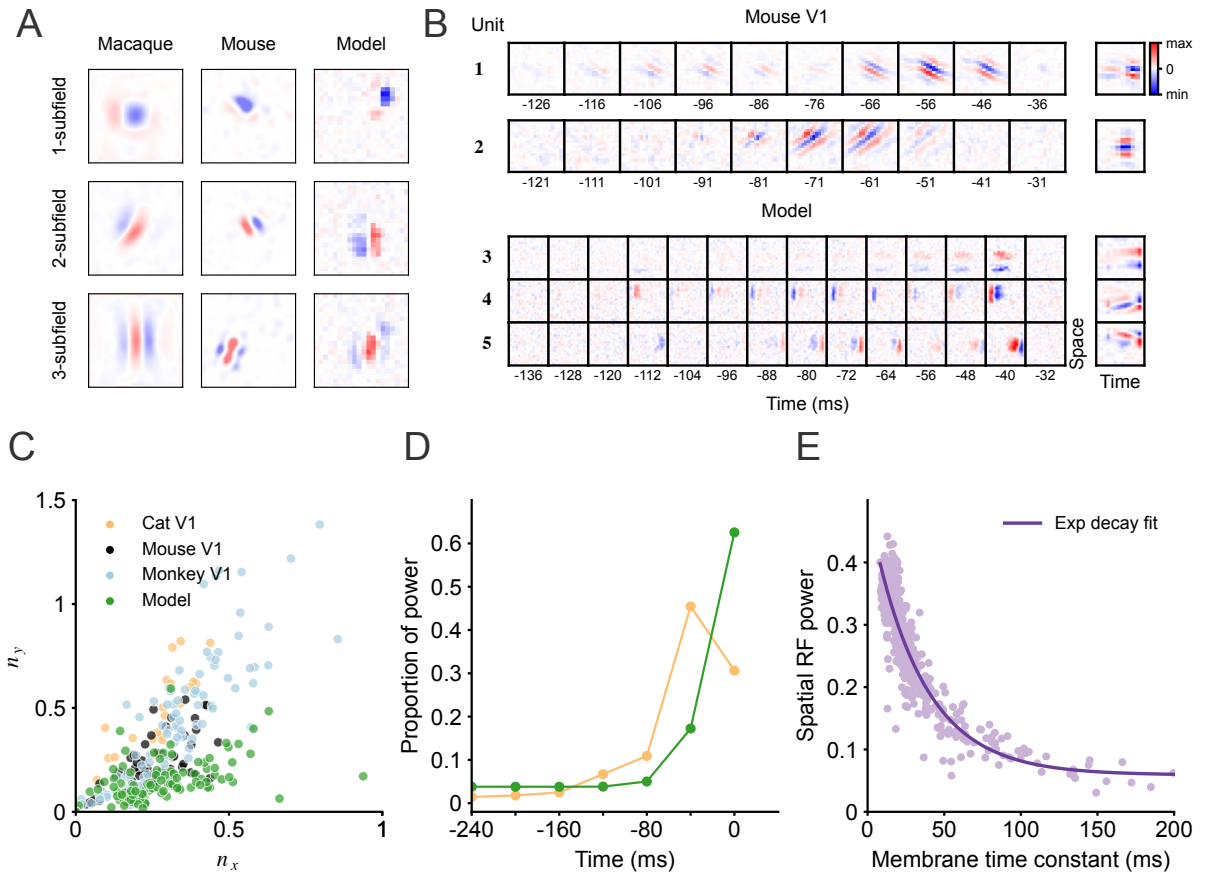

Fig. 2.3: **A.** Example spatial RFs with varying number of subfields of macaque V1 neurons [97], mouse V1 neurons [28] and model units. **B.** $1-2$: example space-time separable spatiotemporal RFs of mouse V1 neurons [222]. $3-5$: example spatiotemporal RFs of model units (3: space-time separable and $4-5$: space-time inseparable). Right column: corresponding spatiotemporal RFs obtained by summing along the RFs axis of orientation (see Methods). **C.** Distribution of RF shapes from V1 neurons in cats [31], mice [28], monkeys [32] and model units. The x-axis (n_x) and y-axis (n_y) are a proportional measure of the number of RF subfields and their respective lengths [32]. **D.** Temporal RF power profile of model units and V1 cat neurons [223], quantified as the sum of squared weights over space and averaged across the population over ~ 40ms time spans. **E.** Scatter plot of the spatial RF power and membrane time constant of the model units. Dark purple line is the fitted exponential curve ($R^2 = 0.84$).

2.2.3 Emergence of simple and complex cell-like tuning

We obtained model unit RFs using a spike-triggered average to spatiotemporal white-noise input clips, and observed a considerable number of similarities between the model unit RFs and the RFs of V1 neurons across different animal species. Qualitatively, we found many units that resemble simple cells [11, 13], with stereotyped RFs consisting of varying number of excitatory and inhibitory subfields tuned to a particular orientation and spatial frequency [31, 32] (Figure 2.3A). The temporal structure of these RFs also resemble those of V1 neurons, as evident in their decay into the past (Figure 2.3B); their polarity profile, with RF polarity either changing (Figure 2.3B plot $1, 4$ and 5) or remaining fixed over time (Figure 2.3B plot 2 and 3) [34, 107, 108]; and their spatiotemporal structure, with RFs either being space-time separable (Figure 2.3B plot $1, 2$ and 3) or space-time inseparable (Figure 2.3B plot 4 and 5) [34].

We further quantified each model unit's RF shape, orientation and spatial frequency using a fitted Gabor function. The distribution over RF shapes, as defined by the width n_x and height n_y of the RF relative to its spatial frequency, matches the neural data (Figure 2.3C). We found the RF shapes to extend along a 1D-curve, from blob-like RFs at the origin (Figure 2.3A first row) to elongated bars with multiple subfields, which are located further away from the origin (Figure 2.3A last row). Our model captures a large portion of the blob-like RFs, which constitute a substantial part of the neural data (with n_x and $n_y <\sim 0.25$). We found the model RFs to be most similar to the mouse V1 RFs, as measured by the Euclidean distance between the model RF-shape centroid and the different animal RF-shape centroids (cat [31] centroid $(0.26, 0.46)$ with distance 0.33; mouse [28] centroid $(0.25, 0.24)$ with distance 0.10; monkey [32] centroid $(0.33, 0.43)$ with distance 0.31; model centroid $(0.25, 0.13)$). The temporal profile across all model RFs also exhibited clear similarities to the neural data, as characterized by a decaying power profile (mean squared RF values over space, units and ~ 40ms time spans) (Figure 2.3D).

We made an additional observation in our model, for which we found no prior reports in V1, by identifying an exponentially-decaying relationship between the maximum spatial RF power and the membrane time constant of a unit (which is not hard-coded but rather learnt) (Figure 2.3E). The membrane time constant determines the operational timescale of a neuron, where neurons with smaller membrane time constants operate on shorter timescales than neurons with larger membrane time constants. Thus, translating our finding to the biology, suggests that V1 neurons with a greater sensitivity to visual input operate on shorter timescales than neurons that exhibit a weaker stimulus sensitivity.

In addition to simple cells, V1 is also distinguished by another significant category of cells known as complex cells [12]. Like simple cells, complex cells are tuned for oriented edges and gratings.

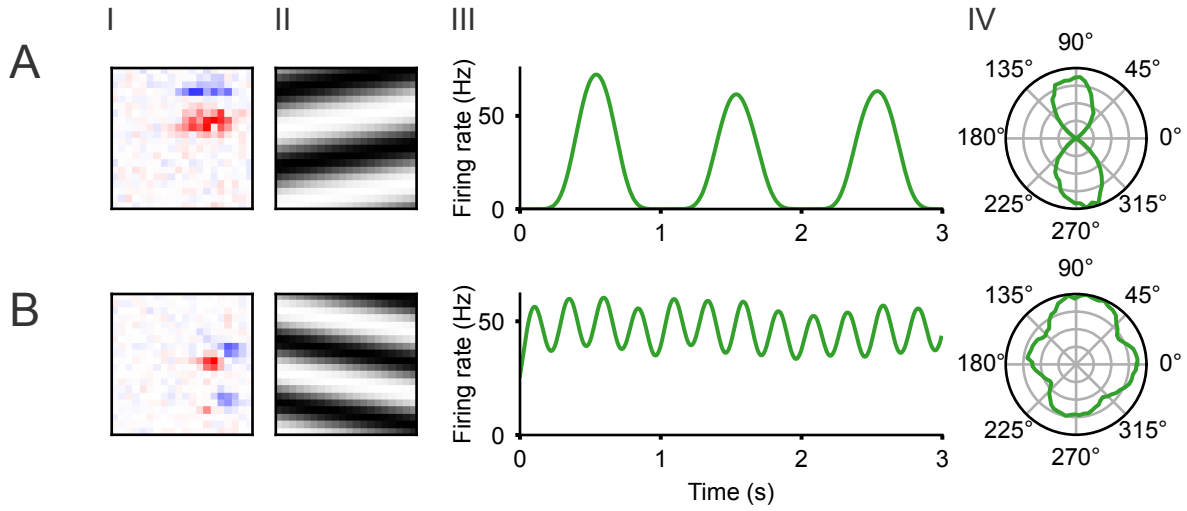

Fig. 2.4: **A.** Example simple cell-like unit, including its RF (I); optimal grating (II); response to its optimal grating (III); and the average response of the unit over time plotted in relation to grating orientation (in degrees) for gratings presented at the unit's preferred spatial and temporal frequency (IV). **B.** Example complex cell-like unit.

Unlike simple cells, however, complex cells are spatially invariant and respond to an oriented pattern regardless of its precise location in the neuron's RF. Using full-field drifting sinusoidal gratings, we qualitatively identified units with response characteristics reminiscent of both simple and complex cells. We observed unit responses to oscillate over time to the movement of their optimal grating. Like simple cells, some model units displayed a high phase dependence with pronounced variations in their response to grating movement [29] (Figure 2.4A). In contrast, other units exhibited a larger mean response over time and were less influenced by the grating phase, consistent with the description of complex cells [30] (Figure 2.4B).

We further quantified the distribution of simple and complex cell-like units in our model by calculating their modulation ratio F_1/F_0. This ratio is defined as the quotient between the amplitude of the first harmonic (F_1) and the mean value (F_0) of a unit's response to its optimal grating, with a value below one classifying a response as complex cell-like [224, 225]. The distribution of modulation ratios in our model is bimodal (Figure 2.5A), which has similarly been reported in V1 neurons of monkeys [226] and mice [28, 227]. Intriguingly, we found the inhibitory units (median $F_1/F_0 = 0.20$) to exhibit more complex cell-like responses compared to the excitatory units (median $F_1/F_0 = 1.62$) ($p = 6.92 \times 10^{-28}$; Mann-Whitney U test), a difference also found between inhibitory and excitatory V1 neurons in mice [28].

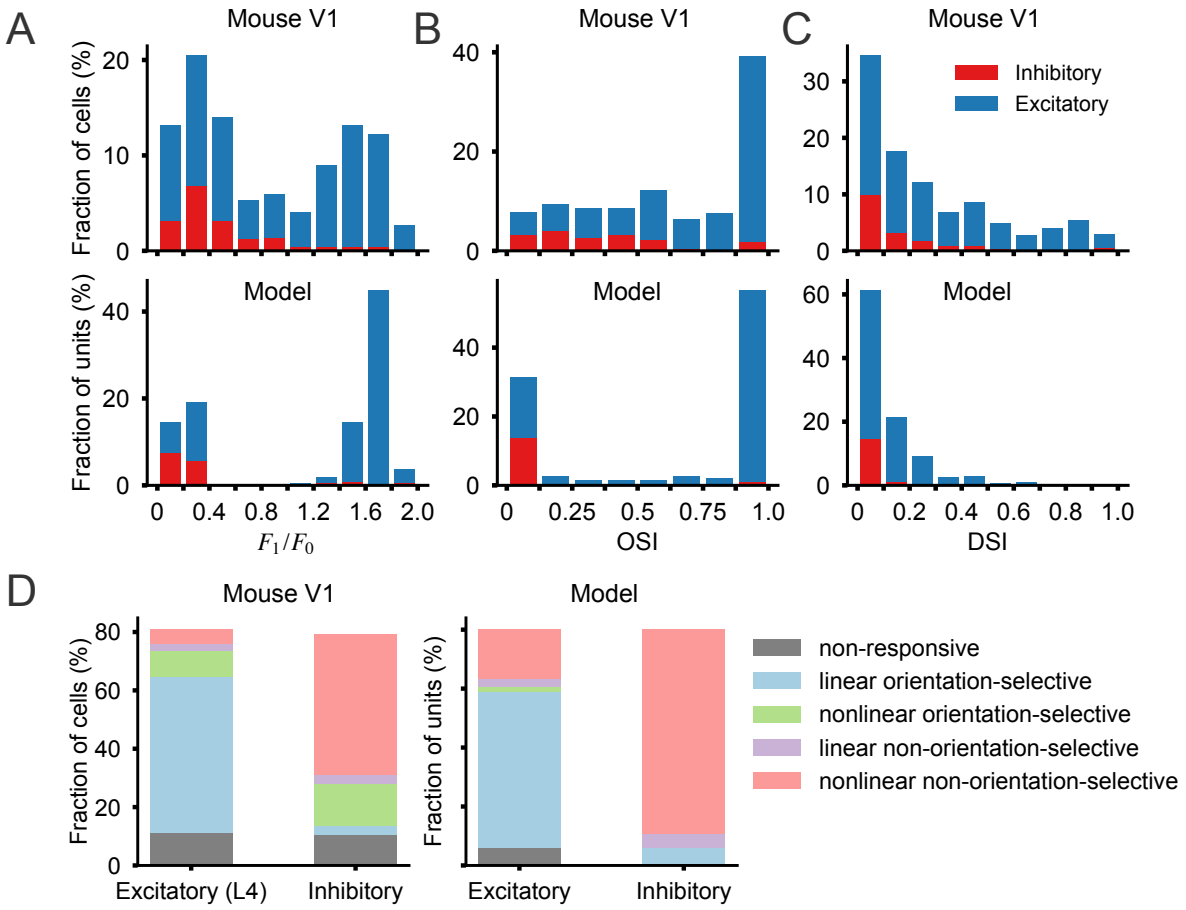

Fig. 2.5: **A.** Distribution of modulation ratios F_1/F_0. **B.** Distribution of orientation selectivity. **C.** Distribution of direction selectivity. **A-C** showing tuning properties of mouse V1 neurons [28] (top) and model units (bottom) to drifting gratings. **D.** Distribution of functional response categories for excitatory layer 4 and inhibitory neurons in mouse V1 [28] (left) and for excitatory and inhibitory model units (right).

We also observed the model units to exhibit a similar orientation and direction preference to those observed in V1 neurons in mice, as measured by the orientation selectivity index (OSI) and direction selectivity index (DSI), respectively [28, 35] (see Methods). Both measures range between zero and one, with a value of one corresponding to maximum selectivity. We found most units to have strong preference for particular orientations ($OSI > 0.5$: 63%), with many units exhibiting near complete selectivity, as has similarly been reported in mouse V1 neurons ($OSI > 0.5$: 74%) [28] (Figure 2.5B). Capturing another difference between inhibitory and excitatory mouse V1 neurons [28], we found the inhibitory units (median $OSI = 0.03$) to be more broadly tuned than the excitatory units (median $OSI = 1.00$) ($p = 2.79 \times 10^{-30}$; Mann-Whitney U test). We found the model units to be less selective for direction ($DSI < 0.5$: 91%), as has similarly been reported for mouse V1 neurons ($DSI < 0.5$: 77%) [28] (Figure 2.5C).

We classified units into different functional response categories, classifying a unit as linear ($F_1/F_0 \geq 1$) or non-linear ($F_1/F_0 < 1$), and as orientation selective ($OSI \geq 0.5$) or non-orientation selective ($OSI < 0.5$) using the modulation ratio and OSI, respectively. The largest functional response categories in the model for both the inhibitory and excitatory units were the same as the largest functional response categories for inhibitory and excitatory neurons in mouse V1 [28] (Figure 2.5D). Specifically, most inhibitory model units are non-linear and non-orientation selective (86%), as are most inhibitory neurons in mouse V1 (60%), while most excitatory model units are linear and orientation selective (66%), as are most excitatory neurons in mouse V1 (layer 4) (67%).

2.2.4 Mirroring V1-cell physiology: membrane dynamics, connectivity and EI balance

V1 is characterised by distinct physiological differences between inhibitory and excitatory neurons, such as the timescales they operate at [25]; their connectivity statistics [23]; and their balanced contribution of input currents received by other neurons [24, 119, 120]. We found our model units to mirror these key differences between the inhibitory and excitatory neurons. Inhibitory neurons tend to operate at shorter timescales than excitatory neurons, as captured by differences in their membrane time constants. Mouse V1 inhibitory neurons exhibit a median membrane time constant of 9.76ms, whilst excitatory mouse V1 neurons have a notably longer median membrane time constant of 21.10ms [25] ($p = 3.43 \times 10^{-47}$; Mann-Whitney U test). Similar to mouse V1, we found the inhibitory and excitatory model units to also operate at different timescales, with the inhibitory and excitatory units having a median membrane time

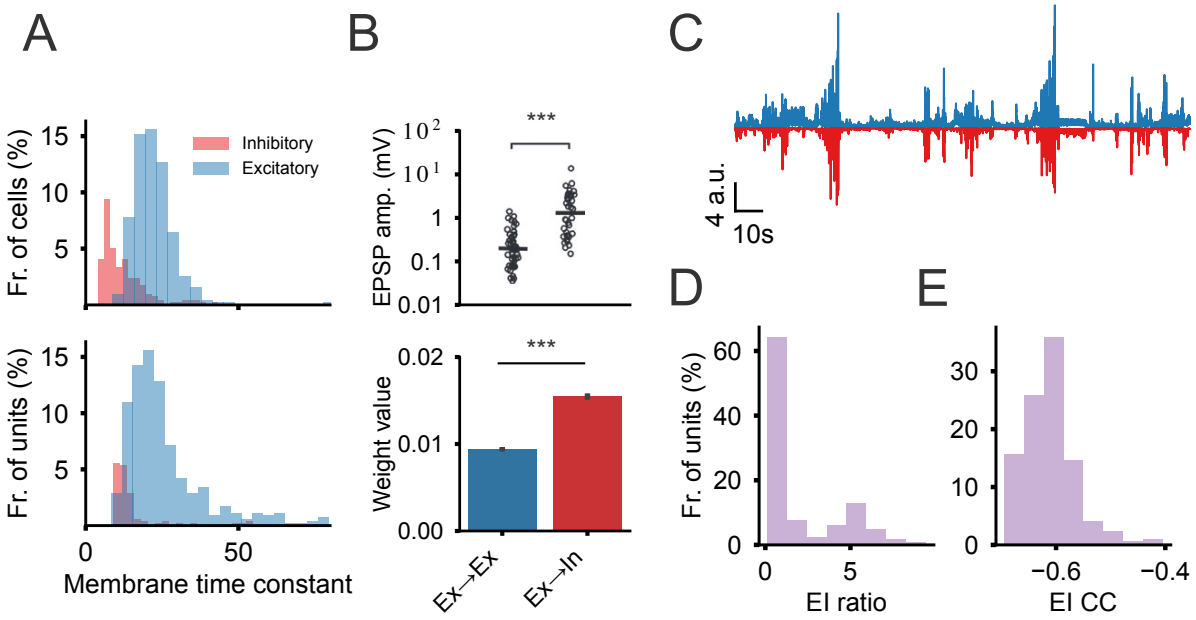

Fig. 2.6: **A.** Membrane time constant distribution of inhibitory and excitatory neurons in mouse V1 (layer 4) [25] (top), and of inhibitory and excitatory model units (bottom). Membrane time constants from 4 inhibitory and 76 excitatory model units with a value above 80ms are not shown. **B.** Top: EPSP amplitudes between excitatory neurons (Ex→Ex) and from excitatory neurons to inhibitory neurons (Ex→In) in mouse V1 (black lines plot median amplitudes) [23]. Bottom: model weight values between excitatory units and from excitatory to inhibitory units. Bars plot the median and bootstrapped standard error, with statistical significance assessed using the Mann-Whitney U test ($***p < 0.001$). **C.** Net-excitatory (top) and net-inhibitory (bottom) currents of an example model unit whilst inputting natural movie clips into the model. **D.** Distribution of model global population EI balance over input clips (left) and distribution of detailed EI balance over model units (right).

constant of 11.97ms and 21.83ms, respectively ($p = 1.31 \times 10^{-25}$; Mann–Whitney U test) (Figure 2.6A).

Excitatory V1 neurons have been shown to synapse more strongly onto inhibitory V1 neurons than other excitatory neurons in mice [23] and rats [228], as measured by their excitatory postsynaptic potential (EPSP) amplitude. We made a similar observation in our model, with the excitatory units connecting more strongly with the inhibitory units than with other excitatory units, as measured by their absolute connection weight ($p = 0$; Mann-Whitney U test) (Figure 2.6B).

A fundamental aspect in neurophysiology is the concept of cortical neurons receiving a balance of excitatory and inhibitory (EI) synaptic currents, as reported in somatosensory cortex [117], primary auditory cortex [118] and V1 [24, 119, 120]. This equilibrium is crucial for ensuring stability in neural activity [123] and proper cortical function [122]. We qualitatively observed a balance between the excitatory and inhibitory currents in the model units to held-out natural input movie clips (Figure 2.6C). EI balance can be quantified as global when the average excitatory and inhibitory currents are equal over time, and as detailed, when the excitatory and inhibitory currents are coupled in time, in which each excitatory input to a neuron arrives simultaneously with an inhibitory counterpart [121]. We found the model units to exhibit a median global EI ratio of 0.77 over the held-out natural input movie clips (Figure 2.6D), where an EI ratio below unity has similarly been reported in mouse V1 [120, 229, 230]. Furthermore, we found the model units to exhibit a detailed EI balance, as indicated by a negative correlation between the net-excitatory and net-inhibitory currents over clips and time across the model units (median: -0.61) (Figure 2.6E).

2.3 Discussion

We have demonstrated that a spiking neural network model, when optimized to efficiently predict future frames in natural movies using past population spike activity, captures several well-known characterizations of V1 neurons. Namely, it captures their stereotypic RFs; their simple and complex cell-like tuning; and - advancing upon previous modelling work - their spike statistics, and key differences between excitatory and inhibitory neurons. Just like V1 neurons, we found the model units to exhibit sparse and irregular spike activity with low correlation between units in response to natural movies. Furthermore, capturing key differences between excitatory and inhibitory V1 neurons, we found that - compared to the excitatory model units - the inhibitory model units had a higher mean firing rate to natural movies; more complex than simple cell-like

responses to drifting gratings; broader orientation tuning; operated on faster timescales, as characterized by shorter membrane time constants; and received stronger connections from excitatory units than the median connection strength between excitatory units.

2.3.1 Comparison to other visual models

Numerous normative models of V1 have previously been developed. We contrast these different models to our work by their normative objective and their architectural implementation. There exists a large body of prior work exploring the normative objective of V1, which attempts to understand the functional role of V1 [125]: just as our livers have evolved to detoxify various metabolites, has V1 evolved to perform a particular functional task? Examples of such normative objectives are: efficient [147, 85] and sparse coding [150, 140, 231], which suggests that V1 efficiently encodes sensory features under metabolic constraints; independent component analysis (ICA) [232], which - applied to neuroscience - stipulates that V1 minimizes redundancies in the stimulus by encoding independent features within sensory input [151, 211]; predictive coding, which posits that neurons remove statistical redundancies in the stimulus by only signalling unpredictable features within sensory input [148, 141, 152, 161]; temporal coherence [233], slow subspace analysis [212] and slow feature analysis (SFA) [143, 146], which hypothesize that V1 encodes slowly varying features within the sensory stimulus; information bottleneck [165], which - applied to neuroscience - suggests that V1 encodes sensory features with maximum mutual information about the future, whilst minimizing information about the past [167]; and temporal prediction, which posits that V1 predicts the immediate sensory future from its past [106, 142, 168] (as explored in our work).

These different normative investigations have all accounted for certain properties of V1 neurons, such as the Gabor-like RFs and simple and complex cell-like tuning. Direct comparisons between these studies are, however, challenging, as they differ in their architectural implementation: some models are based on spatial image input, whereas others have used spatiotemporal movie input; some models are constructed as a hierarchy of layers, whereas others are constructed using a single layer; and some models employ recurrent connections between units, whereas others do not. What these studies do share in common, and in contrast to our model, is that they omit key biological properties of V1 neurons, namely that neurons spike and are separated into distinct subpopulations of excitatory and inhibitory neurons. Thus, these studies were unable to contrast their models to V1 spike statistics and phenomena specific to excitatory and inhibitory V1 neurons.

Certain studies have developed spiking models of V1 with distinct subpopulations of excitatory and inhibitory neurons. These models are able to capture V1-like spike statistics, motion tuning properties like orientation and direction selectivity, and units that exhibit simple and complex cell-like responses [213, 214, 215, 27, 216]. However, unlike our work, these investigations did not learn the model parameters, but rather hard-coded them, and they do not capture the tuning and physiological differences between the excitatory and inhibitory V1 neurons. Furthermore, they do not make any claim regarding the normative objective of V1. Few works have implemented the normative objective of sparse coding in a biologically realistic model of spiking neurons [234, 97], with extensions into a network of excitatory and inhibitory spiking neurons [235]. Unlike our work, these studies focus solely on the spatial RF tuning characteristics of V1 neurons, yet fall short of demonstrating V1-like spike statistics, tuning properties, and associated physiological phenomena.

2.3.2 Biological implementation and extensions of our model

The potential instantiation of our model in V1 can be understood through its biological mapping onto the cortical circuit and the mechanisms by which the brain could acquire the necessary synaptic and neural parameters to facilitate the objective of sensory prediction. Early studies of the cortex have noted its laminar organization, espousing the view of the canonical cortical circuit [236, 237] based on the presence within all neocortical areas of six different layers, each with distinct types of neurons. Within this architecture, layer 4 is the main target for sensory inputs from the thalamus. Similar to layer 4, our model consists of a single layer of spiking units that receive direct sensory input. Thus, the units in our model most likely resemble the neurons in layer 4 of V1. Other features reported for certain sensory cortical areas include their columnar organization, with each column consisting of a small patch of neurons spanning the cortical layers [238, 12]. In V1, this is supported by the existence of dense connections across the cortical layers, with neurons in each layer all tuned to a similar region within the visual field, as well as exhibiting other common properties, such as their orientation selectivity [239]. This hints at the cortex implementing some form of distributed computation, where each cortical patch may be performing a similar function. However, it remains unclear what that function may be [239]. Like the putative columns of V1, our model also operates on a small patch of visual input. Given that our model architecture resembles V1, and accounts for numerous V1 properties, we suggest that V1 may be performing some form of distributed sensory prediction - with each V1 patch predicting its own sensory future - and that all V1 patches in unison predict the future sensory input across the entire visual field.

We could further explore the idea of V1 performing sensory prediction, by extending our model to incorporate additional cortical circuit motifs. Extending our model to be multi-layered would permit us to explore differences in unit tuning properties across layers, and contrast these properties to layer-specific tuning characteristics in mouse V1 [28]. Furthermore, it would be interesting to modify our model to be convolutional with lateral connectivity between groups of units across space, like the horizontal connections that exist between groups of neurons in V1. This would allow us to study how V1 might generate predictions of the sensory future over the entire visual field, and what computational role horizontal connections might serve.

How the visual system instantiates its synaptic and neural parameters is a question that Hubel and Wiesel had asked themselves many decades ago [240]: is the visual circuit hard-wired and encoded in DNA, or rather determined by sensory experience and neural plasticity? We remained agnostic to this question and set our model's parameters using the backpropagation algorithm [173], which incrementally adjusted the parameters, minimizing our model's predictive-and-metabolic cost function. Despite backpropagation sharing similarities with biological learning, certain assumptions in the algorithm have been deemed as biologically incompatible [241, 242]. The reasons for this include the requirement of propagating error backwards through the same synaptic weights used for the forward processing, a characteristic generally not exhibited by real neurons [243]; additionally, backpropagation requires propagating global error representations through layers, in contrast to neurons, which seem to primarily learn through local interactions [244]. Several studies have since attempted to reconcile these issues, showing how models can still learn using backpropagation without using the same weights for forward and backwards propagation [245, 246, 247], and outlining theories of how global error representations could be propagated within the brain [248, 249]. Although backpropagation is considered to be biologically implausible, learning our model's synaptic and neural parameters using this algorithm is not necessarily an issue, as what we were interested in was the end result of our model being optimized for sensory prediction, rather than how the model learns to predict the sensory future.

It would be interesting to extend our work to incorporate biologically-plausible learning rules. Certain aspects of V1 appear to be hard-wired [250, 251, 252, 253, 254], yet different visual experiences (monocular deprivation [33], dark rearing [255, 256, 257] and frame-scrambled input [258]) at early stages of life can lead to drastic changes in neural tuning in V1. Biologically plausible learning rules may allow us to study such phenomena in our model. Furthermore, they may allow us to better compare the artificial development of the model units with other developmental characteristics of V1 neurons, such as the emergence of their spatial and temporal tuning properties [34]. A starting point for such biological plausible learning rules might be to

shift the predictive hypothesis from a systems-level to an individual neuron-level: rather than optimizing population model units for prediction of future sensory stimulus, each model unit could be optimized to predict its own future activity. Such learning objective has been argued to be biologically implementable, and has been shown to result in realistic visual processing in artificial neural networks [259, 260, 162].

Lastly, we could further enhance the biological realism of our model through additional modeling extensions. V1 neurons are known to adapt, by decreasing their activity in response to a constant input stimulus [261]. The adaptive leaky integrate-and-fire (ALIF) model captures the adaptive nature of cortical neurons [110], providing more realistic neural dynamics over the LIF model. Furthermore, inhibitory V1 neurons have been reported to be less adaptive than their excitatory counterparts [262, 263]. Thus, implementing our model units to be adaptive would increase the biological realism of our model and provide another mechanism of comparison between our model and the biology. In our model, we assumed the effect of input current from one unit to another to be instantaneous. However, in reality, synapses conduct current with a temporally-decaying profile, where different types of synapses are governed by a fast or slow decay [264]. Extending our model to include learnable synaptic-like temporal profiles would not only enhance its biological realism, but also offer more opportunities to contrast the model with the biology.

Further modeling improvements could be obtained by training our model at a finer temporal resolution. We trained our model using natural movies recorded at 120Hz, resulting in a ~ 8.33ms discretization step for emulating the LIF dynamics. This may not necessarily pose a problem, as V1 neurons rarely spike more than once during a ~ 8.33ms time window [218, 217, 27]. However, the neural simulation accuracy and biological realism of our model would be improved by adopting a smaller discretization step of around ~ 1ms - in line with the duration of the absolute refractory period of a neuron [265, 266, 267] - in which a V1 neuron would spike at most once. Training our model at such small discretization steps is not straightforward, due to the increase in memory requirements and training times. Recent algorithmic advances may, however, make such training possible [132, 128]. Finally, unlike V1, our model takes filtered movie pixel values as input rather than spikes. It would be an interesting modelling extension to transform the movie input into a spiking representation, and rather use this as input to the model.

2.3.3 Predictions of our model

Our model makes a number of predictions regarding V1:

RF sensitivity related to its timescale of operation: We identified an exponentially-decaying relationship between the preferred-spatial-RF power and the membrane time constant across the model units, suggesting that V1 neurons that are increasingly sensitive to the input stimulus tend to also operate at faster timescales.

Physiological phenomena supports sensory prediction: The exact role of particular V1 physiological observations remains unclear. Neurons in V1 are heterogeneous, exhibiting a wide range of membrane time constant values among neurons, including differences between their excitatory and inhibitory subtypes. Why such heterogeneity and differences exist remains unclear [268], and has been hypothesized to support a range of objectives, such as efficient coding [269], working memory [270] and robust learning [131]. Furthermore, neurons in V1 operate under a balance of excitatory and inhibitory input currents, known to support network stability and sensory processing [123, 122, 121], yet it remains unclear what functional role this might serve. We predict that EI balance and heterogenous neural timescales in V1 may support the functional objective of sensory prediction.

Different metabolic cost for excitatory and inhibitory V1 neurons: The emergence of the tuning differences between the excitatory and inhibitory units in our model was hinged on the assumption that inhibitory units had a lower metabolic cost than the excitatory units. The exact metabolic costs of neurons [271, 272, 273] and their excitatory and inhibitory subtypes [274] is still an ongoing area of investigation. We predict that V1 inhibitory neurons consume less energy than their excitatory counterparts, per unit spike and per equal amounts of depolarization or hyperpolarization, which will need to be tested in future experimental studies.

Visual processing is shallow: A long-standing view is that visual processing is hierarchical, performed sequentially through a stack of layers. This view was inspired by original V1 anatomical connectivity diagrams [14] and by findings of higher-level visual areas extracting more abstract features than lower-level areas, such as V1 (e.g. extracting faces rather than edges) [275]. Contrary to this belief, the shallow brain hypothesis posits that the cortex is less hierarchical and more shallow and parallel in design, with rich visual processing emerging via recurrent rather than hierarchical processing [55]. Our model is consistent with the shallow brain hypothesis, where we found a single hidden layer of recurrently-connected spiking units to capture non-linear phenomena like complex cell-like responses, which have otherwise been argued to emerge from hierarchical processing [276]. These observations may inspire a new breed of computer vision models, which favour a recurrent over a hierarchical architecture, where the latter is considered the current status quo [277].

2.3.4 Applications of our model

Our model's similarity to V1 makes it a promising candidate for studying vision on computers. Traditional exploration of vision relies on experimental methods, which often involve significant costs, extensive time commitments, and ethical concerns regarding animal use. Given that our model captures many V1-like phenomena, it may exhibit other V1 processes yet to be discovered. Thus, neuroscientists may utilize our model for probing and examining questions about V1 at both the system and circuit level, so as to better refine the questions to be asked in animal experiments and reduce the number needed.

Functional role of excitatory and inhibitory neurons: It would be interesting to further explore the differences in the functional roles of excitatory and inhibitory V1 neurons in the context of sensory prediction. While inhibitory neurons are recognized to contribute towards network stability [278], they come at the significant cost of consuming approximately 25% of the cortex's energy budget [274]. Such large energy demands would hint at inhibitory neurons supporting more than a homeostatic role, as is also suggested by their tuning differences compared to the excitatory neurons [28]: but what may this be? Further exploration of our work could inspect the future-frame prediction made from the excitatory and from the inhibitory units separately, revealing possible qualitative differences in their sensory predictions. Additionally, new questions could be explored if the ratio of excitatory to inhibitory units were shaped by experience rather than hard-coded: Would inhibitory units naturally emerge? Would the learned ratio align with that of V1? How would this ratio adapt to variations in our model's metabolic-like regularization?

Translational medicine: Our model might provide theoretical insights underlying certain vision-related medical conditions. Some individuals experience photosensitive epilepsy, a condition where specific visual triggers, like flashing lights, can induce seizures. Seizures are characterized by abnormally high neural activity, presenting as a range of symptoms, such as loss of consciousness and uncontrollable movements [279]. Photosensitive epilepsy is thought to originate in visual cortex [280], resulting from either excessive activity of excitatory neurons or a lack of activity in inhibitory neurons. However, the precise mechanisms underlying this condition remain largely unclear [281]. It would be interesting to manually adjust our model's connectivity to either favour excitation or lessen inhibition, and see if seizure-like activity emerges in response to flashes of light. Such experimentation may provide a starting point in the search for more detailed explanations underlying the cause of seizure activity.

2.4 Methods

2.4.1 Spiking V1 model

Training data

We used a publicly available dataset of natural movie recordings to train and validate our model [282]. This dataset consists of 39 clips, of 9.6 second duration, of everyday objects and scenes (*e.g.* a dog walking, a fish swimming, or a human's view walking through nature), recorded using different techniques (*e.g.* camera moving through space, panning through space or fixed in space). We used 31 clips for training and 8 clips for validation. All clips are grayscale, with a temporal resolution of 120Hz and a spatial resolution of 140×240px. To emulate the transformation of the retina and thalamus, we bandpass filtered all clips, as done in other studies [283, 106]. We also normalized the training and validation data by subtracting the pixel mean and dividing by the pixel standard deviation of the training data, and finally clipped the absolute pixel values to be no larger than 3.5 standard deviations. The training was performed using randomly sampled 20×20px spatial patches of temporally non-overlapping subsamples of 350ms duration (*i.e.* 42 frames). Clips were also randomly flipped around the vertical axis to artificially increase the training data [284].

Spiking neurons

We modeled the V1 cortical neurons as a single-hidden layer of $N = 600$ recurrently-connected excitatory and inhibitory spiking units. The input current $I_i[t] \in \mathbb{R}$ to unit i at time step t is derived from the bandpass-filtered input movie stimulus $x \in \mathbb{R}^{T \times H \times W}$ (of $T = 42$ frames, and spatial height $H = 20$px and spatial width $W = 20$px); model output spikes $S[t-1] \in \mathbb{R}^N$ from the previous time step; and a bias term $b_i \in \mathbb{R}$.

$$I_i[t] = b_i + \underbrace{\sum_{t'=1}^{T_E} \sum_{h=1}^{H} \sum_{w=1}^{W} W_{it'hw} x_{hw}[t - t' + 1]}_{\text{Feedforward current}} + \underbrace{\sum_{j=1, j \neq i}^{N} W^{\text{rec}}_{ij} S_j[t-1]}_{\text{Recurrent current}} \tag{2.1}$$

The feedforward connectivity $W \in \mathbb{R}^{N \times T_E \times H \times W}$ (of $T_E = 15$ temporal frame span) linearly maps the input movie stimulus into a feedforward current contribution (capturing the transformation of the retina and thalamus); the recurrent connectivity $W^{\text{rec}} \in \mathbb{R}^{N \times N}$ maps previous output spikes

into a recurrent current contribution. We implemented all spiking units using the normalized and discretized leaky integrate-and-fire (LIF) model, which evolves the model membrane potential $V_i[t] \in \mathbb{R}$ of unit i at time step t using the following difference equation:

$$V_i[t] = \Big(\beta_i V_i[t-1] + (1-\beta_i) I_i[t]\Big)\Big(1 - S_i[t]\Big) \tag{2.2}$$

During each time step, the membrane potential undergoes a decay governed by a factor β_i. We let each neuron learn its decay factor, and bounded values into the range $(0, 1)$ to enforce correct LIF dynamics [131]. The membrane potential resets to zero in the event of a spike occurring in the preceding time step, which happens when the membrane potential reaches the firing threshold (equal to one).

$$S_i[t] = \begin{cases} 1 & \text{if } V_i[t] > 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

Finally, at every time step t, a prediction of the future spatial movie frame $\hat{y}[t] \in \mathbb{R}^{H \times W}$ was generated (with $\hat{y}_{hw}[t] \in \mathbb{R}$ denoting the predicted pixel value). This was achieved by linearly mapping a span of $T_D = 2$ temporal frames of proceeding spike activity using projection weights $P \in \mathbb{R}^{N \times T_D \times H \times W}$, plus a bias $b^{\text{proj}} \in \mathbb{R}$.

$$\hat{y}_{hw}[t] = b^{\text{proj}} + \sum_{i=1}^{N} \sum_{t'=1}^{T_D} P_{it'hw} S_i[t - t' + 1] \tag{2.4}$$

V1 latencies

There is a minimum neural response latency from the retina to V1, of at least $\sim 30 - 40$ms in mice and monkeys [102, 103, 100, 101], where latencies can significantly vary between neurons [285]. We modelled the transduction and transmission latency to V1, by zero-masking the first 5 spatial frames (*i.e.* 42ms) within the feedforward connectivity tensor W.

Neural noise

Cortical V1 neurons do not respond with identical spike trains to repeated visual stimulus presentations. One of the reasons for this is neural noise. We include two stereotypical neural

noise sources within our model. Firstly, we modeled the noisy photoreceptors in the retina [286], by adding Gaussian sampled noise to the bandpass-filtered movie clips $\epsilon_p \sim \mathcal{N}(0, 0.2^2)$. Secondly, we modeled the synaptic noise of cortical V1 neurons [287], by multiplicatively perturbing the input current $I_i[t]$ of every unit i at time step t using Gaussian sampled noise $\epsilon_g \sim \mathcal{N}(0, 0.6^2)$.

$$\tilde{I}_i[t] = I_i[t](1 + \epsilon_g) \tag{2.5}$$

Dale's law: Enforcing units to be inhibitory or excitatory

Approximately $15 - 20\%$ of cortical V1 neurons are inhibitory in mice, cats, and monkeys [112, 113, 114, 23]. We assigned 15% of the model unit population to be inhibitory, resulting in $N_I = 90$ inhibitory units and $N_E = 510$ excitatory units. We enforced a neuron to be excitatory (or inhibitory), by taking the absolute value of every weight in the recurrent weight matrix W^{rec}, and appropriately negating connection weights from unit j to i. All outwards connection weights from unit j are either positive (*i.e.* excitatory) or negative (*i.e.* inhibitory).

$$W_{ij}^{\text{rec}} = \begin{cases} -|W_{ij}^{\text{rec}}| & \text{if } j < N_I \\ |W_{ij}^{\text{rec}}| & \text{otherwise} \end{cases} \tag{2.6}$$

Weight initialisation

The initial weights for feedforward (W), recurrent (W^{rec}), and projection (P - *i.e.* the readout weights) connections were randomly generated from a uniform distribution $U(-k, k)$, where k was determined as $k = \frac{1}{\sqrt{T_E HW}}$ for feedforward weights, $k = \frac{0.1}{\sqrt{N}}$ for recurrent weights, and $k = \frac{1}{\sqrt{NT_D}}$ for projection weights. Additionally, all initial membrane time constants were uniformly set to 20ms (denoted as $\beta_i \approx 0.81$), and the initial bias terms were uniformly initialized to zero.

Loss function

We trained the model by minimising loss $\mathcal{L}_{\text{total}}$, comprised of prediction loss $\mathcal{L}_{\text{normative}}$ and metabolic loss $\mathcal{L}_{\text{metabolic}}$. The prediction loss measures the similarity between the predicted and target future movie frames, and the metabolic loss measures the approximated physiological

energy utilization of the network. We weighted this loss by hyperparameter $\lambda = 10^{-2.75}$, which we selected to produce the most V1-like receptive fields.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{prediction}} + \lambda \underbrace{\left(\gamma_{\text{transmission}} \mathcal{L}_{\text{input}} + (1 - \gamma_{\text{transmission}}) \mathcal{L}_{\text{spiking}}\right)}_{\text{Metabolic loss } \mathcal{L}_{\text{metabolic}}} \tag{2.7}$$

We computed the prediction loss $\mathcal{L}_{\text{prediction}}$ as the mean squared error (MSE) between all predicted spatial movie frames $\hat{y}$ and target spatial movie frames $y \in \mathbb{R}^{H \times W}$. The calculation was performed by taking the average over all batch samples B (omitted here for brevity), simulation steps T (starting from simulation step t_0 to allow unit membrane potentials to depolarize sufficiently, *i.e.* to enable network warm-up), and spatial dimensions H (height) and W (width), both which were cropped by c pixels on all sides.

$$\mathcal{L}_{\text{prediction}} = \underbrace{\frac{1}{(T - t_0)(H - c)(W - c)} \sum_{t=t_0}^{T} \sum_{h=c}^{H-c} \sum_{w=c}^{W-c}}_{\text{Avg. over time and space}} \underbrace{(\hat{y}_{hw}[t] - y_{hw}[t])^2}_{\text{MSE}} \tag{2.8}$$

We defined our metabolic-like loss $\mathcal{L}_{\text{metabolic}}$ as the average synaptic-like transmission over all the model weights, where synaptic transmission has been noted to require significant energy within cerebral cortex [274]. We defined the synaptic-like transmission of a single model weight as its absolute value, multiplied by the absolute value of every input value, summed over all time steps. To account for the large energy utilization of spikes [288], we weighted the synaptic-like transmission resulting from the spikes ($\mathcal{L}_{\text{spiking}}$) differently to the synaptic-like transmission resulting from the input ($\mathcal{L}_{\text{input}}$), using hyperparameter $\gamma_{\text{transmission}} = 0.3$. We also assumed the metabolic-like cost of the excitatory and inhibitory units to differ, with the inhibitory units consuming more energy than excitatory units, captured by hyperparameter $\gamma_{\text{type}} = 0.1$.

$$\mathcal{L}_{\text{input}} = \gamma_{\text{type}}\mathcal{L}_{\text{inhibitory-input}} + (1-\gamma_{\text{type}})\mathcal{L}_{\text{excitatory-input}} \tag{2.9}$$

$$\mathcal{L}_{\text{spiking}} = \gamma_{\text{type}}\mathcal{L}_{\text{inhibitory-spiking}} + (1-\gamma_{\text{type}})\mathcal{L}_{\text{excitatory-spiking}} \tag{2.10}$$

$$\mathcal{L}_{\text{inhibitory-input}} = \underbrace{\frac{1}{N_I T}\sum_{i=1}^{N_I}\sum_{t=1}^{T}}_{\text{Avg. over inhibitory units and time}} \underbrace{\Big(|b_i| + \sum_{t'=1}^{T_E}\sum_{h=1}^{H}\sum_{w=1}^{W}|W_{it'hw}||x_{hw}[t-t'+1]|\Big)}_{\text{Afferent synaptic cost}} \tag{2.11}$$

$$\mathcal{L}_{\text{excitatory-input}} = \underbrace{\frac{1}{N_E T}\sum_{i=N_I}^{N}\sum_{t=1}^{T}}_{\text{Avg. over excitatory units and time}} \underbrace{\Big(|b_i| + \sum_{t'=1}^{T_E}\sum_{h=1}^{H}\sum_{w=1}^{W}|W_{it'hw}||x_{hw}[t-t'+1]|\Big)}_{\text{Afferent synaptic cost}} \tag{2.12}$$

$$\mathcal{L}_{\text{inhibitory-spiking}} = \underbrace{\frac{1}{N_I T}\sum_{i=1}^{N_I}\sum_{t=1}^{T}}_{\text{Avg. over inhibitory units and time}} \underbrace{\Big(\sum_{j=1,j\neq i}^{N}|W_{ij}^{\text{rec}}|S_j[t-1]\Big)}_{\text{Recurrent synaptic cost}} \tag{2.13}$$

$$+ \underbrace{\frac{1}{THW}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{w=1}^{W}}_{\text{Avg. over time and space}} \underbrace{\Big(\sum_{i=1}^{N_I}\sum_{t'=1}^{T_D}|P_{it'hw}|S_i[t-t'+1]\Big)}_{\text{Efferent synaptic cost}} \tag{2.14}$$

$$\mathcal{L}_{\text{excitatory-spiking}} = \underbrace{\frac{1}{N_E T}\sum_{i=N_I}^{N}\sum_{t=1}^{T}}_{\text{Avg. over excitatory units and time}} \underbrace{\Big(\sum_{j=1,j\neq i}^{N}|W_{ij}^{\text{rec}}|S_j[t-1]\Big)}_{\text{Recurrent synaptic cost}} \tag{2.15}$$

$$+ \underbrace{\frac{1}{THW}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{w=1}^{W}}_{\text{Avg. over time and space}} \underbrace{\Big(\sum_{i=N_I}^{N}\sum_{t'=1}^{T_D}|P_{it'hw}|S_i[t-t'+1]\Big)}_{\text{Efferent synaptic cost}} \tag{2.16}$$

(2.17)

Surrogate gradient descent

We trained the spiking models using a variation of the backpropagation algorithm [173], called surrogate gradient descent [198]. Here, to enable smooth gradient flow during training, a well-behaved function was employed as a replacement for the gradient of the non-differentiable

Heaviside step function. We adopted the fast sigmoid function, which has been shown to perform well in practice [194, 133].

$$\frac{\partial S_i[t]}{\partial V_i[t]} = (10|V_i[t] - 1| + 1)^{-2} \tag{2.18}$$

We conducted all training using the Adam optimizer [289], employing its default parameters. The training spanned 1200 epochs with an initial learning rate set to 10^{-4}, which was subsequently reduced by a factor of 0.2 at epoch count 200, 600, and 800. We performed training using a batch size of 1024 and saved model weights whenever a new minimum training loss was achieved.

2.4.2 Spike analysis

For all the spike analysis, we used the model output spike trains resulting from 5000 randomly sampled input movie clips from the held-out test set not used for training. To be comparable with the analysis of [27], we calculated all spike statistics using a time window of 2s with a step size of 0.2s. In the main-text results, we only included model activity resulting from a clip if its mean activity (averaged over units and time) was within 1.5 standard deviations of the mean population activity (over all test clips and units). This criterion resulted in 3252 clips being included in the analyses. See Supplementary material for the spike results calculated using all of the test clips.

Firing rate

Every spike train was mapped to a firing rate, by counting the total number of spikes and dividing by the spike train duration.

Coefficient of variation

We quantified the variability of the spike pattern in each spike train by calculating the coefficient of variation of the interspike interval CV(ISI), which is defined as the ratio between the standard deviation and the mean of the interspike intervals [110]. A $CV(ISI) < 1$ corresponds to a more regular firing pattern than a $CV(ISI) > 1$. As done in [27], we only calculated CV(ISI) values for a given spike train if it had at least three spikes.

Neuron synchronization

We quantified the correlation between different unit spike trains by calculating the mean cross-correlogram between the spike trains of randomly-sampled unit pairs. This measures the average correlation $\hat{\rho}[\tau]$ between the spike trains s_1 and s_2 of different units for varying temporal lag τ, defined as:

$$\hat{\rho}[\tau] = \mathbb{E}_{s_1,s_2}\Big[\frac{\mathbb{E}[(s_1[t]-\mu_{s_1})(s_2[t-\tau]-\mu_{s_2})]}{\sigma_{s_1}\sigma_{s_2}}\Big] \tag{2.19}$$

We sampled 400 random unit pairs for this calculation. Furthermore, we binned the spike trains with a bin size of 25ms before calculating the correlations, in order for our analyses to be comparable to [27].

2.4.3 Receptive field analysis

Spike-triggered average

We estimated the spatiotemporal receptive field of every unit using a spike-triggered average, using responses to 1000 spatiotemporal white-noise input clips of 100-frame duration. Clips were generated by sampling each pixel from a Gaussian distribution with a standard deviation of $\sigma = 10$.

Fitting Gabors

We fitted a Gabor function [105] to the RF of every unit to quantify its tuning properties. The Gabor function has been shown to provide a good approximation for most RF spatial aspects [31, 32] and provides a quantitative measure to compare the RFs of the model with the RFs of V1 neurons. The 2D Gabor function is defined as:

$$G(\hat{x},\hat{y}) = A\exp\left(-\left(\frac{\hat{x}}{\sqrt{2}\sigma_x}\right)^2-\left(\frac{\hat{y}}{\sqrt{2}\sigma_y}\right)^2\right)\cos(2\pi f\hat{x}+\phi) \tag{2.20}$$

where spatial coordinates $(\hat{x},\hat{y})$ are obtained by translating the centre of the RF (x_0,y_0) to $(0,0)$ and rotating the RF by its orientation θ:

$$\begin{aligned}\hat{x} &= (x-x_0)\cos(\theta)+(y-y_0)\sin(\theta)\\ \hat{y} &= -(x-x_0)\sin(\theta)+(y-y_0)\cos(\theta)\end{aligned} \tag{2.21}$$

The amplitude of the Gaussian envelope is defined by A, whereas parameters σ_x and σ_y define the Gaussian envelope width along the $\hat{x}$ and $\hat{y}$ axes, respectively. Parameters f and ϕ define the spatial frequency (in cycles/pixel) and phase of the sinusoidal grating along $\hat{x}$. We optimized the Gabor parameters for each model unit RF by minimizing the mean squared error between the corresponding Gabor function and its respective model RF. To avoid local minima during fitting, we performed each Gabor fit in the spectral domain, by transforming the model RFs and Gabors using a 2D Fourier transform [31, 106]. We then used the resulting parameters to continue the fitting procedure in the spatial domain. We excluded units which had a poor fit (correlation coefficient < 0.7, 445 units), whose fit was based on very few pixel values (Gaussian envelopes $\sigma_x < 0.5$ or $\sigma_y < 0.5$, 367 units), and whose Gabor's tail was only being fitted to ((x_0, y_0) outside of the RF, 6 units). This resulted in 111 out of 600 units with a sufficient fit, having a median pixel-wise correlation coefficient of 0.85.

Spatial receptive field properties

Power. The power of a RF is defined as the mean of the squared values of the RF over a span of time (as in Figure 2.3D) or at a point in time (as in Figure 2.3E) [106].

Structure. The structure of a RF is characterized by two variables $n_x = \sigma_x f$ and $n_y = \sigma_y f$, which give a measure of the number of oscillations of its sinusoidal component and the length of the bars in the RF, respectively [32]. Blob-like RFs lie close to the origin in the $n_x - n_y$ plane, the number of bars increases along the n_x axis, and the RF bar-length increases along the n_y axis.

Phase. The phase of the sinusoidal component of a RF, mapped into the range $[0 - \frac{\pi}{2}]$ using $\hat{\phi} = arg(|\cos\phi| + i|\sin\phi|)$, gives a measure of the RF symmetry, with an RF having an even symmetry when $\hat{\phi} = 0$ and an odd symmetry when $\hat{\phi} = \frac{\pi}{2}$ [32].

Estimating space-time separability

Spatiotemporal RFs are either space-time separable or inseparable, with space-time separable RFs tuned to flashing and moving bars, and space-time inseparable RFs tuned to direction of movement. Space-time separable spatiotemporal RFs can be represented as a product of a spatial and temporal function and space-time inseparable spatiotemporal RFs cannot [62]. We classified the space-time separability of model unit RFs using the singular values of a singular value decomposition (SVD), as done in [106]. We flattened the spatial dimensions and removed the transmission latency time steps for each model unit RF, and took the SVD of the resulting

matrix. If the ratio between the second and first singular value was < 0.5, the RF was deemed space-time separable; otherwise, the RF was deemed space-time inseparable.

2D spatiotemporal receptive fields

The 2D spatiotemporal RF illustrates the temporal evolution of the spatiotemporal RF [34]. This view was constructed by translating and rotating every spatial RF by the position and rotation that centres the spatial RF of the largest power, placing all oriented bars parallel to the y-axis. Next, all spatial RFs were flattened along the y-axis and concatenated, resulting in a 2D spatiotemporal RF. This new view can qualitatively inform us about the temporal phase (through polarity changes in the flattened spatial structure) and separability (through shifts in the flattened spatial structure) of a spatiotemporal RF.

2.4.4 Motion tuning

Virtual physiology

We constructed model unit responses to drifting full-field sinusoidal gratings, varying in orientation (from $0°$ to $360°$ in $5°$ increments), spatial frequency (ten values evenly spaced from 0.01 to 0.2 cycles per pixel), and temporal frequency (1, 2, 4, and 8Hz), each with an amplitude of ± 1. Each grating clip, lasting 3 seconds, was presented to the model four times. We calculated the firing rate of each model unit by averaging responses over clip repeats and convolving with a Gaussian kernel ($\sigma = 72$ms) [62]. For each model unit, we computed its mean firing rate over each grating clip, and constructed a 3D mean-firing tuning space (over orientation, spatial, and temporal frequency), with the optimal grating (defined by its orientation, spatial and temporal frequency) eliciting the largest mean firing rate. We only included responsive units in the analysis (552 out of 600), which had an optimal response no less than 5% of the mean optimal response over all model units.

Modulation ratio

We measured the modulation ratio F_1/F_0 of each model unit by taking the Fast Fourier Transform (FFT) of each unit's response to its optimal grating and calculating the ratio between the amplitude of the first harmonic (F_1) and DC component (F_0). A unit with a modulation ratio

below one is considered complex cell-like, whereas a unit with a modulation above one is considered simple cell-like [224, 225, 28].

Orientation and direction selectivity

We calculated the orientation selectivity index (OSI) and direction selectivity index (DSI) for each unit, which quantify a unit's preference for stimulus orientation and direction of motion, respectively [28, 35]:

$$\text{OSI} = \frac{R^{\text{orient}}_{\text{pref}} - R^{\text{orient}}_{\text{orth}}}{R^{\text{orient}}_{\text{pref}} + R^{\text{orient}}_{\text{orth}}} \tag{2.22}$$

$$\text{DSI} = \frac{R^{\text{dir}}_{\text{pref}} - R^{\text{dir}}_{\text{non-pref}}}{R^{\text{dir}}_{\text{pref}} + R^{\text{dir}}_{\text{non-pref}}} \tag{2.23}$$

Both metrics are calculated using the mean unit responses to gratings at a unit's preferred spatial and temporal frequency. $R^{\text{orient}}_{\text{pref}} = R^{\text{dir}}_{\text{pref}}$ is a unit's mean response to a stimulus in its preferred orientation and direction, and $R^{\text{orient}}_{\text{orth}}$ and $R^{\text{dir}}_{\text{non-pref}}$ are the mean responses in the orthogonal orientation and opposite direction to the preferred direction, respectively. A value close to zero corresponds to less selective responses, and a value closer to one corresponds to greater selectivity for both measures.

2.4.5 Physiology analysis

Calculating membrane time constants

The membrane time constant τ of every model unit is obtained from the decay factor β in the LIF Equation using the following relation:

$$\tau = \frac{-\Delta t}{\ln(\beta)} \tag{2.24}$$

EI balance

Using the held-out test dataset of natural movie stimuli, we calculated the net-excitatory (Ex_i) and net-inhibitory (In_i) input current to each model unit i at time step t as:

$$\text{Ex}_i[t] = \max(0, I_i[t]) + \sum_{j=N_I, j\neq i}^{N} W_{ij}^{\text{rec}} \, S_j[t-1] \tag{2.25}$$

$$\text{In}_i[t] = \min(0, I_i[t]) + \sum_{j=1, j\neq i}^{N_I} W_{ij}^{\text{rec}} \, S_j[t-1] \tag{2.26}$$

The net-excitatory and net-inhibitory currents were derived from the excitatory and inhibitory units, respectively, with the feedforward current $I_i[t]$ being classed as either excitatory or inhibitory at a given time step, depending on its sign value. From this, we calculated the global EI balance of a model unit as:

$$\frac{\sum_t \text{Ex}_i[t]}{\sum_t |\text{In}_i[t]|} \tag{2.27}$$

with the excitatory and inhibitory current time series for each input clip concatenated over time. We calculated the detailed EI balance of every model unit by first convolving each current time series using a Gaussian kernel ($\sigma = 42$ms) and then calculating the resulting correlation between the smoothed inhibitory and excitatory current time series (with both time series concatenated over time for each input clip). We smoothed the current time series to emulate synaptic conductance [110], finding the non-smoothed excitatory and inhibitory current time series of units to otherwise be uncorrelated (median CC of 0.02).

Supplementary

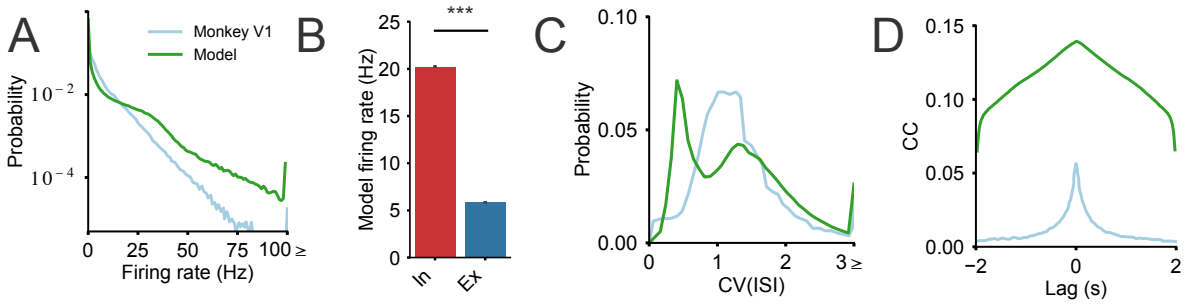

Fig. 2.7: Spike statistics from Figure 2 recalculated using all of the input clips without filtering. **A.** Firing rate distribution of the model units and monkey V1 neurons. **B.** Difference in the model's firing rates between the inhibitory (In) and excitatory (Ex) units. Bars plot the mean and standard error over units and input clips, with statistical significance assessed using the Mann-Whitney U test ($^{***}p < 0.001$). **C.** Coefficient of variation of the interspike intervals CV(ISI) distribution of the model units and monkey V1 neurons. **D.** Cross-correlogram of the model units and monkey V1 neurons. Monkey V1 data in all plots was obtained from [27] in response to monkey's watching segments of the Star Wars movies.

3 Crystal balls as eyes: retina optimised for prediction across animal species

3.1 Introduction

Retinal ganglion cells (RGCs) respond to different patterns of light [9, 10] and send a ~ 100-fold downsampled representation of the visual world to the brain [86]. It remains a contentious question what information this pattern of activity represents. One hypothesis posits that the retina efficiently compresses incoming light into a downsampled neural code [68, 202, 203, 204, 205] - like a camera capturing incoming light. Another hypothesis is that the retina efficiently signals features within sensory input that are predictive of the future [206, 207, 208, 209] - like a crystal ball foretelling the future.

Normative models provide a commonly used approach for exploring such coding questions [62, 125]. Here, a model of the retina is optimized for a particular goal (*e.g.* efficient compression or prediction), and the resulting phenomena within the model are compared for similarity to the biology. Previous work has explored the efficient compression of natural images in models employing units that output firing rates [68, 202, 203, 204, 205]. These models have successfully captured the different types of retina-like spatial RFs and their mosaic-like organizations. However, these models lack particular biological realism, such as the spiking non-linearities of RGCs and their recurrent interactions [72]. Modelling spikes are important as they play an important role in sensory transmission through the retina encoding stimuli using the relative timing of spikes [201]. Furthermore, these models have not been shown to capture more complex retinal processes, such as stimulus-omission [93, 210] and motion-anticipation [92] responses or differential motion tuning [91]. Most importantly, and pivotal to the compression versus prediction dichotomy, it remains unclear to what extent a model optimized for prediction compares to the retina.

To address these shortcomings, we developed a spiking retinal model, trained on natural movies under metabolic-like constraints, to either encode the present or to predict the future, from recent

60

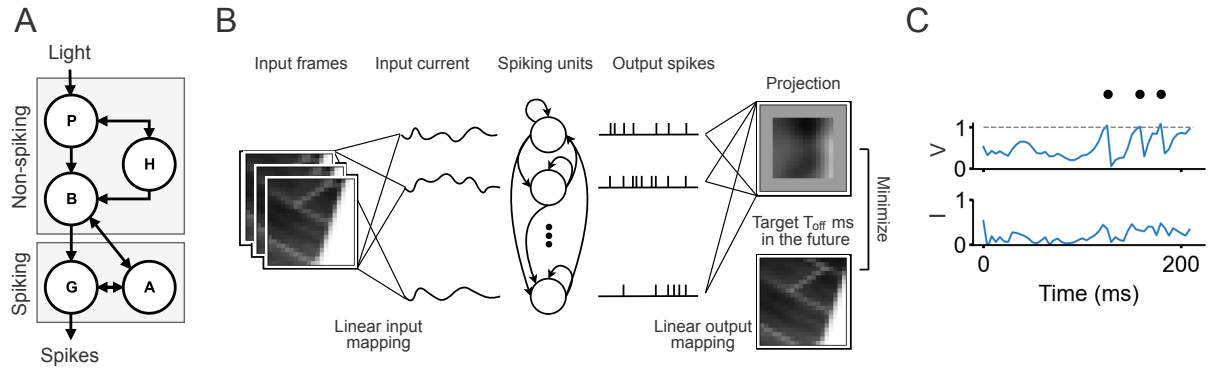

Fig. 3.1: **A.** Schematic of the retinal circuit with its different cell types (P=Photoreceptor, H=Horizontal cell, B=Bipolar cell, A=Amacrine cell, G=Ganglion cell) and respective connections, with the arrow heads indicating the flow of visual information. **B.** Schematic of the retinal model. The model takes a sequence of movie frames as input and linearly transforms these into an input current fed into a single hidden layer of recurrently-connected spiking LIF units. Photoreceptor-like noise was added to the input image and ganglion-cell-like noise was added to the input current. The spike output of these units was used to linearly construct a projection, either of the present or the future, of the input stimulus. The linear transformation approximates the transformation of the pre-ganglionic cells in the top grey box in **A.**, and the recurrently-connected spiking units model the ganglion cells, and their interactions with the amacrine cells, in the bottom grey box. **C.** Example activity traces of one of the LIF units over time: input current I charges the membrane potential V and elicits a spike if the firing threshold (dotted line) is crossed.

spike activity. Like prior normative investigations, we found our spiking model optimized for prediction to capture retina-like RFs and their spatial organizations. Extending previous studies, we found that our model also exhibited retina-like latency coding [201] and spike statistics in response to natural images and movies. Furthermore, the predictive model captured complex retinal processes of motion tuning and stimulus omission. In contrast, the model optimized for efficient compression did not account for complex retinal processes like stimulus omission responses [93, 210]. Notably, we found the model optimized for prediction to more accurately predict RGC responses across different animal species to both natural images and movies. Our findings suggest that the retina in mammals and amphibians has evolved to supply the brain with temporally-predictive features about the visual world.

3.2 Results

3.2.1 The spiking retinal model

The retina consists of three cellular layers: the photoreceptors (which transduce light), the bipolar cells (an intermediate cell layer), and the ganglion cells (the spiking output layer), interspersed by the horizontal and amacrine cell inhibitory interneurons [72] (Figure 3.1A). We modelled this circuit, using a single hidden layer of recurrently-connected spiking units (Figure 3.1B), where the spiking units are akin to the ganglion cells. All spiking units were modelled using the leaky integrate-and-fire model [110], which captures the key biophysical nature of real neurons: integrate incoming current, and fire a spike when the firing threshold is reached (Figure 3.1C). Pre-ganglionic cells (the photoreceptors, horizontal and bipolar cells) appear to mostly respond to incoming light in a linear fashion [290]. We thus opted to model their transformation using a linear mapping, whose output was fed as input current to the ganglion-like spiking units. Similar to prior modelling work [291], we included recurrent connectivity to capture the interactions between ganglion cell types, arising from electrical synapses between ganglion cells [292, 293] and electrical and chemical synaptic connections between ganglion cells via amacrine cells [292, 72].

We also noise-perturbed the input stimulus and each unit's membrane potential with Gaussian noise, to capture the noise characteristics of the photoreceptors and ganglion cells [286, 294]. We chose these noise sources so that the model's spike train variability matched that of the ganglion cells across repeated stimulus presentations. As in the retina, the model's spike trains were almost identical in response to uniform full-field illumination with randomly varying intensity over multiple repeats [295, 86] (see Supplementary material), and the model's response variability was sub-Poisson with Fano factors mostly below unity [295, 296, 76] (see Supplementary material).

The model was trained on diverse natural movies, with the objective of estimating a spatial frame T_{off}ms into the future, or at the present, using a linear readout from past spike activity. We explored different prediction offset values T_{off} to examine if an encoding ($T_{\text{off}} = 0$ms) or predictive ($T_{\text{off}} > 0$ms) objective better captures the sensory processing of the retina. Lastly, we regularized the model's connectivity during training using a metabolic-like cost function, which penalized individual synaptic-like connections by their cost in energy, approximated as the absolute value of a connection weighted by its incoming activity over time (see Methods).

Unless otherwise mentioned, we report our results using the model optimized for prediction of $T_{\text{off}} = 128$ms into the future.

The model, metabolic regularization function, training dataset and training procedure in this chapter are similar to that of the previous chapter. However, the results in this chapter differ from those of the previous chapter. This is primarily due to the chosen level of regularization and movie filtering. In contrast to the V1 model, the retinal model employed a stronger metabolic regularization and no movie filtering. In addition, there are some minor modelling differences between the spiking models, where the V1 model has separate populations of excitatory and inhibitory units - a constraint that was not imposed on the retinal model. Furthermore, the input data to the retinal model was also perturbed with some added Gaussian noise to emulate the noisy photoreceptors, which the V1 model did not include.

3.2.2 Retina-like spike statistics to natural images and movies

We compared the spike activity of the model units to the RGC activity across different animal species in response to natural images and movies. Here, we used publicly available recordings from the retina in macaque, marmoset, mouse and salamander [297, 69, 76]. We found our model to exhibit similar spike activity to the RGCs across the different animal species. The retina qualitatively exhibits more sparse and regular firing activity in response to natural images [76] (Figure 3.2A) compared to natural movies [69] (Figure 3.2B). We found our model to also qualitatively exhibit similar firing patterns to the different types of input.

We quantified the sparsity and variability in the spike responses to the natural images and movies across the different animal species and the model. We measured the spike response sparsity and variability by respectively calculating the mean firing rate and the mean coefficient of variation of the interspike interval CV(ISI). The CV(ISI) measures the variability in a spike train, where $CV(ISI) < 1$ corresponds to a more regular firing pattern than $CV(ISI) > 1$ [110] (see Methods). Both the mean firing rate and CV(ISI) are lower across the animal species in response to the natural images (Figure 3.2C), than to the natural movies (Figure 3.2D). Similarly, we found the model units to exhibit a lower firing rate (images 8.27Hz vs movies 17.27Hz; $p = 6.51 \times 10^{-14}$, Mann-Whitney U test) and CV(ISI) (images 0.38 vs movies 1.91; $p = 2.19 \times 10^{-166}$, Mann-Whitney U test) to the images compared to the movies. We used the same image dataset from the salamander recordings [76] to quantify the model's response statistics to natural images, and used our held-out movie test set to calculate the model's response statistics to natural movies.

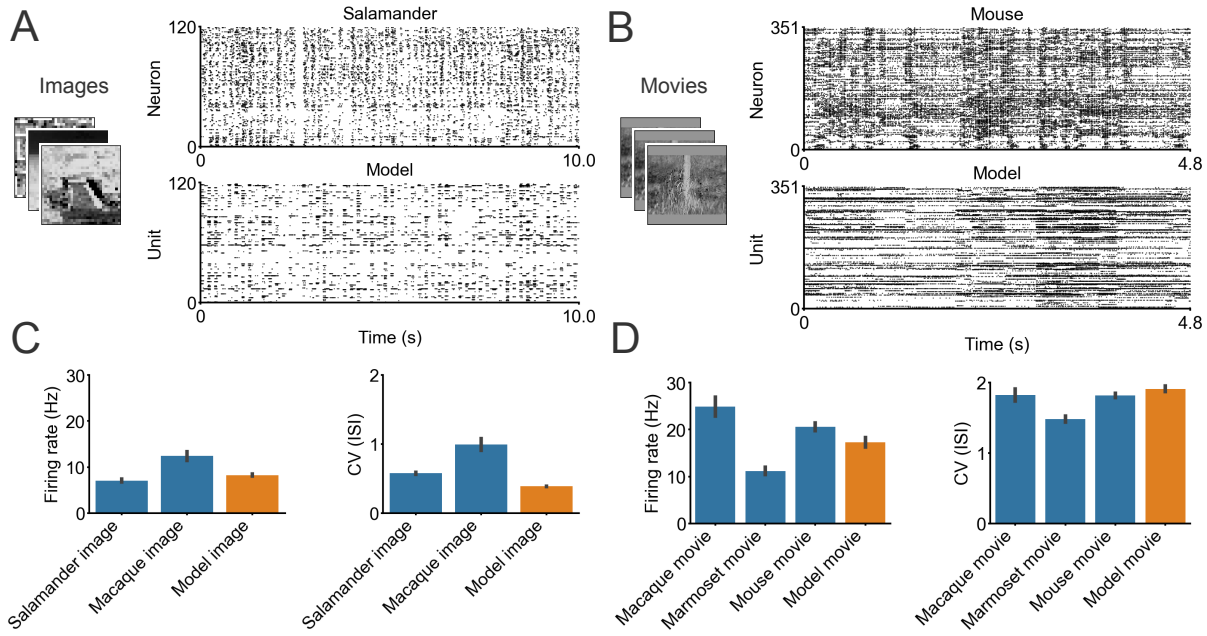

Fig. 3.2: **A.** Raster plot from the salamander retina (top) and the model (bottom) to the same natural images (left). **B.** Raster plot from the mouse retina (top) and the model (bottom) to the same natural movies (left). **C.** Left: Mean firing rate of RGCs from different animals (blue) and model units (orange) to natural images. Right: Corresponding mean coefficient of variation of the interspike intervals CV(ISI) of the neurons and model units to the natural images. **D.** Same as **C.** but with the statistics instead calculated using the responses to natural movies. The salamander data was obtained from [76]; the macaque image and movie data was obtained from [297], and the marmoset and mouse movie data was obtained from [69]. The number of model units was randomly sampled for plots **A.** and **B.** to match the number of neurons.

3.2.3 Emergence of major retina-like cells

We analyzed the functional type of each model unit, just like one would categorize the functional types of RGCs [76], by mapping the spatiotemporal receptive fields (RFs) of the model units using a spike-triggered average to white-noise. We fitted a 2D Gaussian function to each unit's spatial RF and projected the spatial RF size and dimensionality-reduced RF temporal profile onto a 2D scatter plot. We then separated the units into functional groups using k-means clustering (see Methods).

We found the model units to separate into four functional classes (Figure 3.3A), each resembling one of the major cell types found in the primate retina, which together constitute over 95% of the ganglion cells in the central retina [74]: ON parasol and midget cells and their OFF type counterparts [298, 74] (Figure 3.3B). Just like in biology, we found the parasol-like units to have a larger spatial structure and a more biphasic temporal profile than the midget-like units [299, 75]. We also found the model parasol- and midget-like units to respond similarly to their biological counterparts in response to flashes of light (Figure 3.3C): the ON and OFF parasol-like units responding with a transient burst of spikes to changes in light intensity, and the ON and OFF midget-like units responding with a sustained train of spikes (with the ON-type cells responding to the light onset, and the OFF-type cells responding to the light offset) [80, 81]. As reported in the retina across different animal species [83, 82, 84], we found the RFs of the different unit types in the model to tile visual space, forming mosaic-like organizations (Figure 3.3D). We also found fewer parasol- than midget-like units, and fewer ON- than OFF-type units within the model (Figure 3.3E). This aligns with experimental reports of midget cells being more plentiful than parasol cells [74, 79], and OFF-type cells occurring more frequently than ON-type cells [78, 75]. Lastly, the model units exhibit a heterogeneous membrane time constant distribution on the scale of tens of milliseconds, which has similarly been reported for cat RGCs [300] (Figure 3.3F).

3.2.4 Texture, differential and anticipatory motion selectivity

We quantified the tuning characteristics of the model units and found the units to exhibit various well-known motion-tuning characteristics of RGCs. Various ganglion cells across animal species have been reported to be selectively tuned to the orientation and direction of moving objects, with a preference for motion along the cardinal axes [301, 88, 89, 302, 303, 90, 304, 305]. Similar to the biology, we found some units to be selective for orientation (11.75% of units with $OSI > 0.5$) and some selective for direction (2.5% of units with $DSI > 0.5$), with a preference

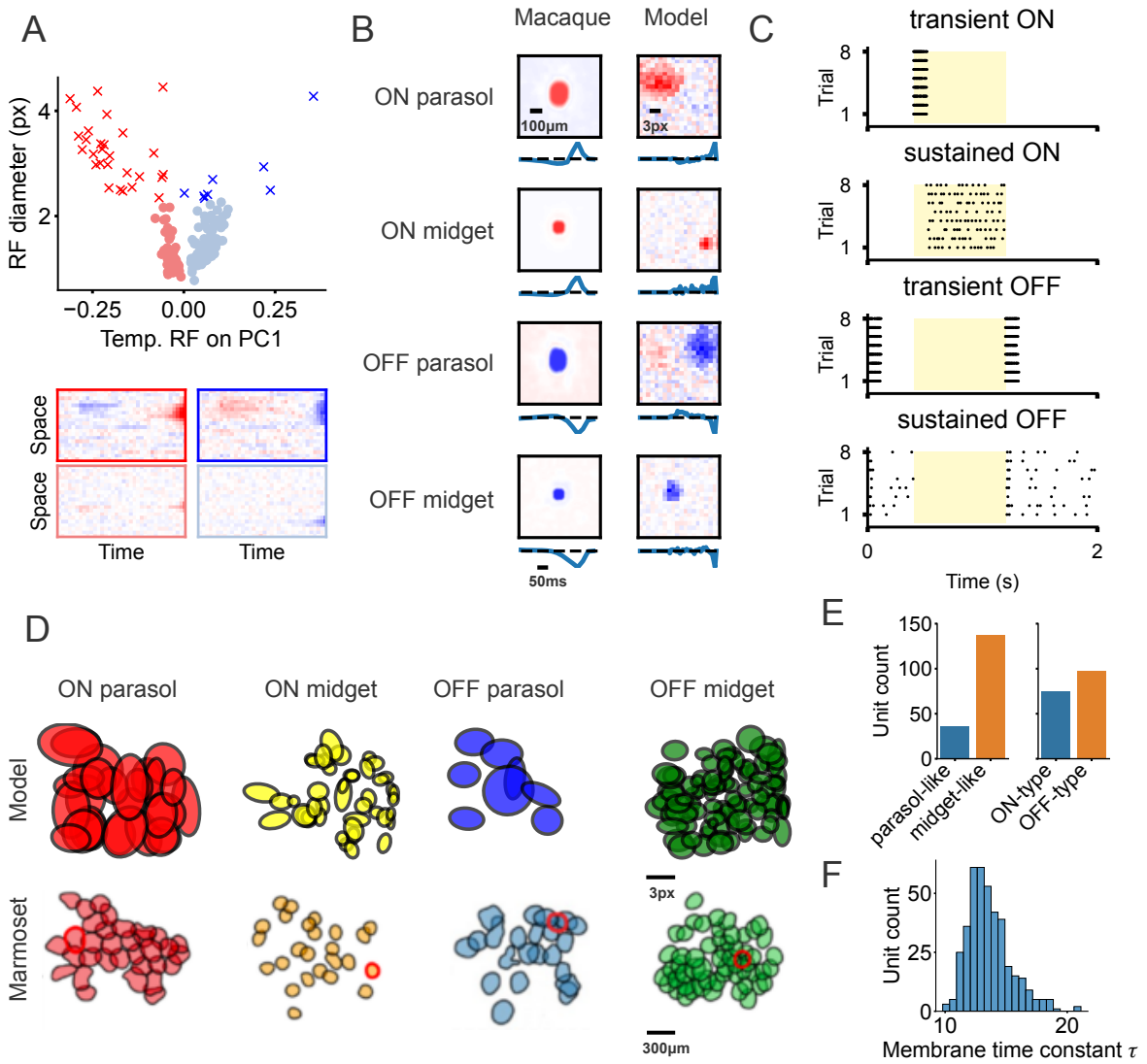

Fig. 3.3: **A.** Scatter plot of model unit RF diameter (obtained from a 2D Gaussian fit) vs the first principal component of the dimensionality-reduced RF temporal profile, with the spatiotemporal RFs obtained from a spike-triggered average to white-noise. Separate colours and shapes denote distinct functional groups (obtained using k-means clustering) with red and blue colours denoting ON- and OFF-type units, respectively, and crosses and dots denoting parasol- and midget-like units, respectively. Sample 2D spatiotemporal RFs from each group are plotted below: left are ON-type units, right are OFF-type units, top are parasol-like units and bottom are midget-like units. **B.** Example spatial RFs, with corresponding temporal profile plotted below, from the macaque retina [68] (left) and the model (right). **C.** Light-flash responses of the model units depicted in **B.** (yellow rectangle denotes the duration of the flash of light). **D.** RF mosaics of the different model unit types (top) with the corresponding RF mosaics of the different marmoset ganglion cell types [69] (bottom). **E.** Left: number of model units that are parasol- and midget-like. Right: number of units which are ON- and OFF-type. **F.** Membrane time constant distribution over all model units.

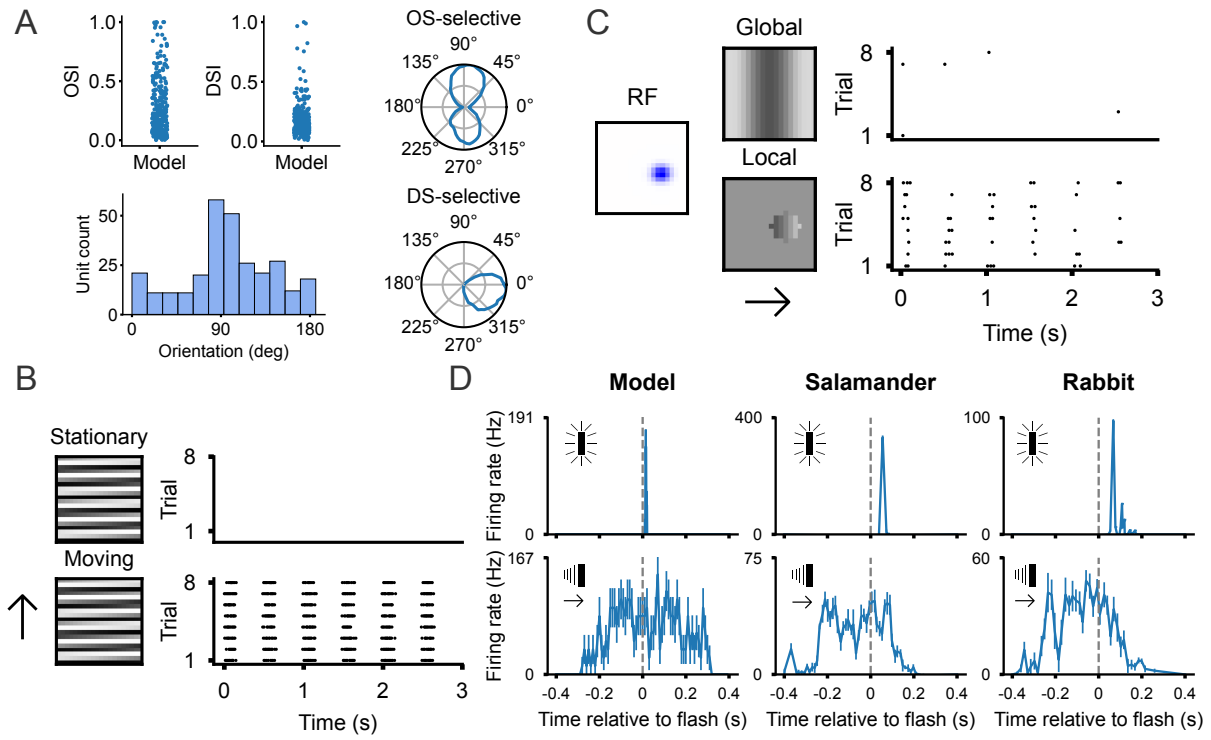

Fig. 3.4: **A.** Top left: model unit orientation- and direction-selectivity distributions (values closer to unity denote stronger tuning). Bottom left: distribution of orientation tuning for the model units. Right: sample tuning curves of an orientation-selective and direction-selective unit. **B.** Spike raster of a model motion-sensitive unit, with no response to a stationary grating of high spatial frequency (top), and a response to the grating moving upwards (bottom). **C.** Spike raster of an object-motion-sensitive model unit, with no clear response to the grating moving over the entire visual space (top), and a stronger response to the same grating moving spatially localized within the model unit's RF (bottom). **D.** Top: responses of a single model unit, and a single salamander and rabbit retinal ganglion cell to a briefly flashed bar (~ 15ms) within their RFs. Bottom: Anticipation-like responses in the same model unit and RGCs to a rightward moving bar (~ 0.44mm s^{-1} - see Methods) before the bar crosses the RFs. The bars were aligned at time zero and error bars denote the s.e. over repeated stimulus presentations, with the salamander and rabbit data adapted from [92].

for motion along the horizontal and vertical axes (Figure 3.4A). A well-known class of ganglion cells, known as Y-type cells ($\sim 3\%$ of ganglion cells in cats [306]), are tuned to the general motion of textures of high spatial frequency, but not their static presentation [307, 308, 309, 72]. Similar to these neurons, we found units in the model (5% of units) that did not respond to a stationary grating of high spatial frequency (for each unit's preferred orientation), yet fired as soon as the grating moved (Figure 3.4B). Some ganglion neurons, known as object-motion-sensitive (OMS) cells, exhibit differential motion selectivity, where a neuron remains silent when an image (or grating) moves across the retina, yet fires if the motion in the RF centre differs from the wider surround (*e.g.* by masking out the wider surround) [89, 310, 91, 311]. We also found such OMS tuning in some of the model units (Figure 3.4C). Certain ganglion cells have been characterized to have an anticipation-like response to moving objects (like a bar moving left or right), where the neuron fires even before the object crosses the neuron's RF, whereas there is a delayed response to simply flashing the bar onto the RF [92]. Similarly, we found certain units in our model to also have an anticipation-like response to a moving bar (Figure 3.4D).

3.2.5 Emergence of a relative spike latency code and stimulus omission responses

A contentious and open question in neuroscience is how neurons transmit information using their spikes [264]. This has historically been cast as a code of two extremes, where neurons either communicate using a number of spikes (*i.e.* a rate code) or the timing of spikes (*i.e.* a temporal code). Most acknowledge that these strategies are not necessarily mutually exclusive [312], and that different neurons likely employ different coding strategies depending on their location within the nervous system. In the retina, there is strong evidence that certain RGCs employ a temporal-like code, encoding the onset of a new visual scene using the relative timing of the first spikes. Reconstructing natural images from such a differential spike latency of fast OFF ganglion cells, rather than their spike counts, qualitatively results in a clearer image reconstruction [201]. We examined these coding strategies in our model and similarly found a rate-based reconstruction to qualitatively be more blurry, noisy and with erroneously overemphasized edges than a latency-based reconstruction (Figure 3.5A). A differential spike latency code can transmit new visual information using very few spikes, which is energy efficient and fast. This is particularly important, as our gaze only remains fixed for a fraction of a second, before rapid saccadic eye movements, induce a new visual scene onto our retina [313, 314].

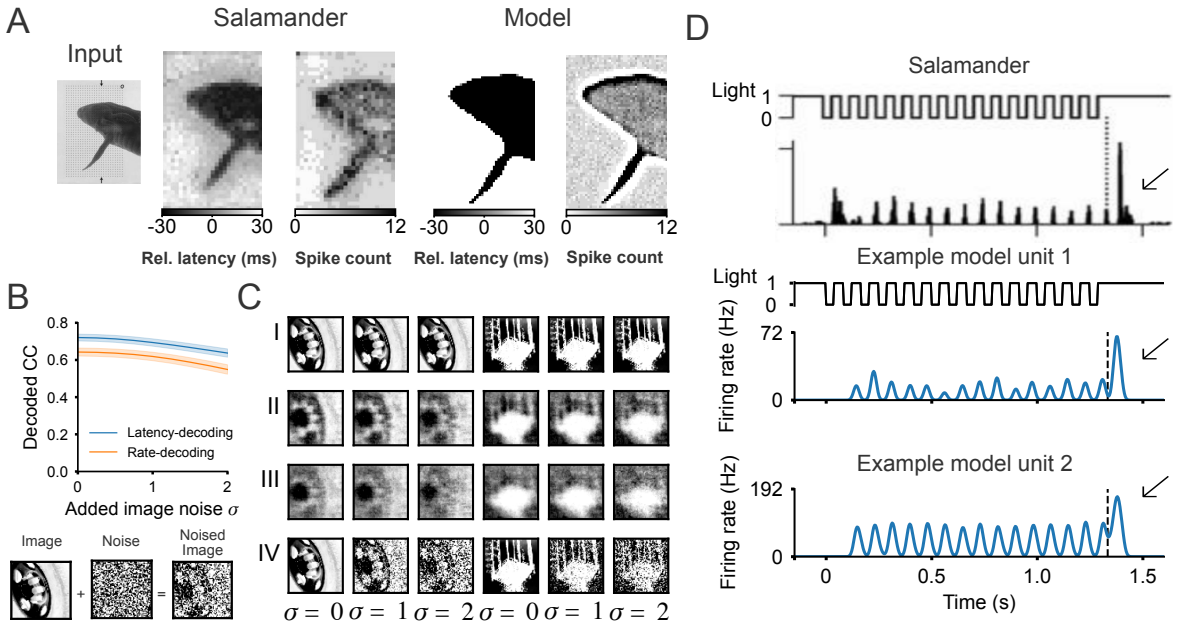

Fig. 3.5: **A.** Reconstruction of an input image using the relative spike latencies and the spike counts from a single OFF-type neuron from the salamander retina [201] and from a single OFF-type model unit. **B.** Top: mean decoding accuracy (Pearson correlation coefficient) in reconstructing natural images from noise-corrupted inputs using a linear transformation of the population model activity with a latency code (blue line) and a rate code (orange line) (the shaded area is s.e.). Bottom: illustration of noise-corrupting the input images. **C.** Example image reconstructions from **B.**. I: target image, II: latency-decoding, III: rate-decoding, IV: noised-image. **D.** Firing rate of a single salamander RGC [93] (top) and of two example model units (bottom) during a sequence of 16 40ms-flashes presented at 12Hz. The dashed line marks the time when the flash sequence was halted with the arrow indicating the omitted stimulus response.

Most of our understanding of the retina's neural code is at a single-cell level based on responses to artificial stimuli, like drifting gratings [72, 315]. It is less well understood how the different neural coding strategies encode natural stimuli from the population activity. Additionally, it remains to be understood how the different coding strategies compare to each other in the presence of varying levels of noise. This is important, as the retina likely employs a noise-robust coding strategy, given the ubiquity of noise in the retina and the visual environment [288]. To explore this, we tested how well we could reconstruct natural images, by inputting their noise-corrupted counterparts into the model, and linearly decoding the images from the resulting rate and latency codes of the population model spike activity. We found the latency-decoding strategy to more robustly decode the noise-corrupted natural images than the rate-decoding strategy (Figure 3.5B), qualitatively producing clearer image reconstructions, even under high levels of noise corruption (Figure 3.5C).

Certain ganglion cells in the salamander and mouse retina have been reported to exhibit an omitted-stimulus response (OSR), generally thought to encode prediction or prediction errors [316]. Ganglion cells exhibiting OSRs fire strongly to a violation in a periodic temporal pattern, such as an omitted flash within a steady dark flashing sequence [93]. Such neurons either respond weakly (20% of neurons) or not at all (22% of neurons) during the flash sequence, as reported in the salamander retina [210]. We found some of the model units (6%) to produce an OSR to a violation in a periodic flash sequence (see Methods). These units' responses were sustained throughout the flash sequence, and greatly increased at the end of the sequence, around the expected time of the response to the omitted flash (Figure 3.5D). Interestingly, OSR responses were only present in the model optimized for temporal prediction and not in the model optimized for compression.

3.2.6 A predictive retinal model captures neural responses across animal species

We assessed whether the retinal model, optimized to encode the present or the future, best predicts the activity of RGCs across different animal species in response to natural images and movies. We used five publicly available datasets of RGCs recordings in macaque, marmoset, mouse, and salamander [297, 69, 76]. For every dataset and retinal model, we fitted a linear-nonlinear readout to predict the neural responses [51] and quantified the performance of each neuron's fit as the normalized correlation coefficient (CC_{norm}) between the predicted and the target neuron response on a held-out validation set [317]. As commonly done [318, 319,

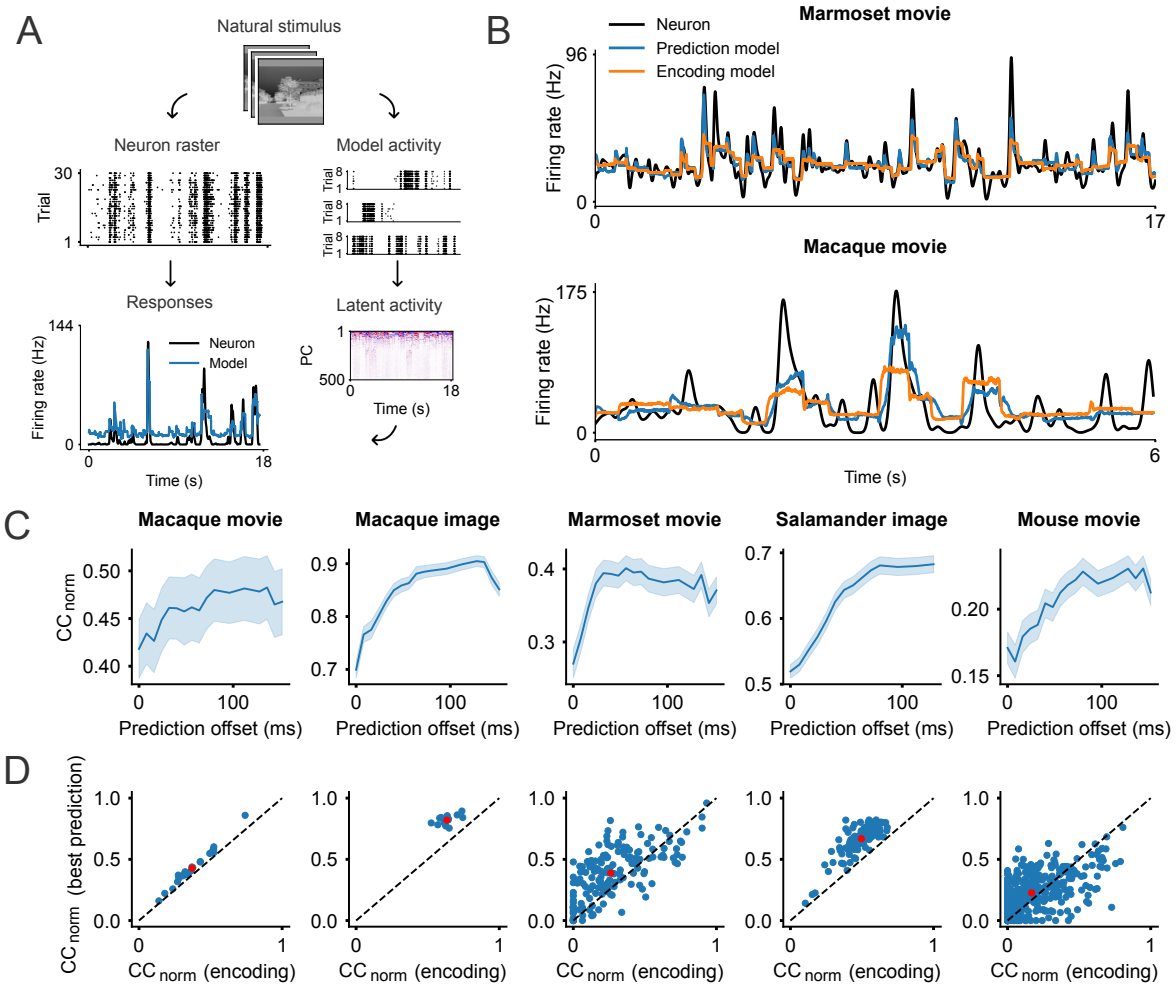

Fig. 3.6: **A.** Each retinal model (trained to encode the present or predict different offsets into the future) was fitted to predict the neural responses across multiple natural stimuli in different retinal datasets. Model unit activity was obtained for each stimulus set used in these studies and downsampled into a latent activity space using principal component analyses, from which the neural responses were predicted using a regularized linear-nonlinear readout (see Methods). All results report the normalized correlation coefficient between the predicted and the target neural responses, computed on a held-out validation set. **B.** Sample neural response (black line) of a single neuron in the marmoset retina (top) and macaque retina (bottom) to natural movie stimuli. Predicted responses of the predictive model (blue line) and encoding model (orange line) are superimposed. **C.** Neural prediction scores across the different retinal datasets and retinal models, as a function of the prediction offset (graphs plot the mean and s.e. over neurons). **D.** Neuron prediction scores of the best predictive model (y-axis) against the encoding model (x-axis) across the different datasets (with the red dot plotting the centroid).

155, 320, 321], we performed the fits from a dimensionality-reduced latent representation of the model's hidden-unit activity in response to the different dataset inputs (Figure 3.6A; see Methods).

Across all animal species and stimulus types, we found the retinal model optimized for prediction to most accurately predict the neural responses. The predictive model appeared to qualitatively match the neural responses, particularly at the peak firing rates (Figure 3.6B). Notably, the mean CC_{norm} steadily increased across all datasets for increasing model prediction offset (*i.e.* the further the retinal model predicted into the future, the better the neural fits), up until a maximum score was reached, after which the curves declined (Figure 3.6C). Across the datasets, the maximum CC_{norm} was obtained using the retinal model optimized to predict $56-128$ms into the future. Lastly, a larger CC_{norm} was obtained across all datasets for the majority of neurons - and not some small subset - using the prediction, rather than the encoding retinal model (Figure 3.6D).

3.3 Discussion

The retina forms the first stage of visual processing within the visual system. In this study, we set out to uncover the sensory transformation performed by the retina, which is typically thought to efficiently compress incoming input [68, 202, 203, 204, 205] or to efficiently extract sensory-features predictive of the future [206, 207, 208, 209]. To examine this, we implemented a biologically-detailed spiking model of the retina, which we separately optimized for efficient compression or efficient prediction using natural movie stimuli. We found the model optimized for prediction to exhibit various retina-like phenomena, not all of which were captured by the model optimized for efficient compression. Notably, the predictive model more accurately predicted RGC responses to both natural images and movies across animal species. These findings suggest an evolutionary conserved role of the retina in informing the brain about temporally predictive features in the visual world.

Comparison to prior models Normative models are the status quo for examining the sensory transformation of a biological circuit like the retina [62, 125]. Here, a model of the retina is constructed and optimized for a particular goal, such as compression or prediction. Prior studies have developed models of the retina using non-spiking units optimized for efficient coding, by maximizing the mutual information between the sensory input and model unit firing rates [202, 204, 205, 203] or by linearly reconstructing the present sensory input [68]. These models have

accounted for the retina's spatial RFs [68, 202, 203, 204, 205], their temporal profiles [68, 203], and their mosaic-like organization [202, 203, 204, 205].

We developed a model of greater biological realism. In contrast to prior work [68, 202, 204, 205, 203], our model units output individual spikes rather than firing rates. This is particularly important, as the retina more likely encodes sensory input using spike times rather than firing rates [201]. Our model also includes recurrent connectivity to capture the interplay known to exist between ganglion cells, whereas other models omit recurrence [68, 203, 205, 204, 202]. Like prior studies [203, 204, 205, 202], our model incorporates input and output noise to model the noisy phototransduction and the noisy spike generation of ganglion cells. We trained our model on natural movie recordings, whereas previous work has omitted temporal dynamics by training on natural images [202, 204, 205] (although see [203]). Our model was also trained to minimize a detailed metabolic-like loss, approximating synapse-like energy requirements, unlike previous studies, which have simply constrained model firing rates [68, 203, 205, 204, 202].

We explored the longstanding hypothesis of the retina performing prediction, in contrast to studies focusing exclusively on the retina encoding the present [68, 203, 205, 204, 202]. Like prior work, our model captures the spatial RFs, their temporal dynamics, and mosaic-like spatial organization. Unique to our work, we contrasted the spiking dynamics of our model to the spiking dynamics of RGCs. We found our model to exhibit retina-like spike statistics to natural images and movies, as quantified using firing rates and measures of spike variability. Notably, our model more accurately encoded sensory input using the relative spike latencies rather than firing rates, as has also been reported in the retina [201].

Our model exhibited several retinal phenomena unaccounted for by other normative studies of the retina. This includes finding certain units tuned to the direction and orientation of moving gratings, as has similarly been reported in the retina [322]; units that exclusively fire to high-spatial-frequency gratings only if they move, like Y-type cells in the retina [72]; units tuned for the differential motion of objects versus their background [91]; units tuned for the anticipation of moving objects [92]; and units tuned for stimulus omission within a temporal sequence of flashing lights [93, 210].

Salisbury and Palmer [207] hypothesized that certain nonlinear retinal processes may be connected to prediction. Supporting this view, we found the stimulus omission responses to only emerge when our model was optimized for efficient prediction and not for efficient compression of the present. Lastly, we addressed the challenging problem of predicting RGC responses to natural images and movies across different animal species. We found that the model optimized for efficient sensory prediction more accurately predicted neural responses than the model

optimized for efficient sensory encoding of the present. Interestingly, the best performing model was trained to predict on a timescale of $\sim 50 - 100$ms, similar to the latency of the photoreceptor responses [323, 324]. Furthermore, this duration also corresponds to the duration at which retinal spikes contain the most information about the sensory future, as estimated in a prior study [206].

Model limitations We attempted to model the retina with increased biological realism over prior models. However, we likely still fall short of capturing all of the retina's biological complexities and processes. RGCs are known to adapt to stimulus contrast [325], the variability of light intensities within the stimulus. RGCs exhibit a faster and more biphasic temporal profile with lower sensitivity to high-contrast stimuli [326, 290]. Our model might be able to capture such dynamic processes by implementing the units as adaptive leaky integrate-and-fire neurons, which extend the LIF units with an adaptive firing threshold [110, 128].

Unlike prior normative models, we included recurrent connectivity in our model, similar to the coupling filters in encoding models [291]. It would be more realistic to include constraints on the connectivity, such as the inhibitory connections from amacrine to ganglion cells [292, 72]. Additionally, it would be more realistic to include gap-like junctions between the spiking units to emulate the electrical coupling between the ganglion cells [292, 293].

We modelled the transformation of the photoreceptors and bipolar cells as a linear transformation, given reports of their linear-like integration properties [290]. However, it has also been shown that the input and output of bipolar cells can exhibit non-linear integration properties [327]. It would thus be more realistic to include an additional hidden layer of graded units in our model to emulate the bipolar cells. We trained our model on gray-scale normalized movie patches, where close to half of the pixel values are negative. It might be more realistic to train our model using positive-only pixel values to more accurately capture the range of light intensities (*e.g.* zero - no light; one - light). Furthermore, it would be interesting to train our model using colour movies, to further examine the retina's visual processing of stimuli of different wavelengths [328].

Lastly, we trained our model on randomly-sampled spatial patches of movie clips. Future extensions could develop and train a spiking model that operates over the entire visual scene. This could be a convolutional model [47], or more realistically, a set of unique patch models, each operating at a different spatial location within the visual frame. This model may capture the different cross-space tuning properties of ganglion cells, such as their differences in direction-selectivity [329]. However, it remains a challenging problem to train spiking neural networks over space and time, due to speed and memory constraints (although see [128, 132]).

Extended validations We found our model optimized for temporal prediction to exhibit various retinal properties across different species. Future work could further examine whether our model exhibits other well-known retinal phenomena. For example, RGCs have been reported to switch their polarity from OFF- to ON-type, resulting from large shifts within sensory input [330]. RGCs also respond to the sudden motion reversal of a moving bar, which, like stimulus omission responses, are thought to signal violations in predictions [331]. It would be insightful to further quantify the different types of tuning properties in the model. For example, this could include counting the number of units that exhibit motion anticipation or differential motion tuning; or further characterizing more specific types of stimulus omission responses, as done in [210].

Lastly, it would be interesting to further explore the effect of noise in the model. The RFs in our model lack a clear surround, whereas the RFs of RGCs have stereotypically been characterized as center-surround [72]. Perhaps learning the noise sources in the model, instead of hardcoding them, would result in RFs with surrounds [204]. However, it remains unclear to what degree RGCs exhibit RF surrounds, as several studies report RFs with little to no surrounds [332, 75]. Perhaps, these differences result from the way the RFs are mapped, where spike-triggered averages - as employed in our study - have been shown to result in RFs with less defined surrounds [333].

Speculations and implications Finally, we hypothesize that models of increasing biological realism and capturing an increasing number of biological phenomena will play an important role in generating more neuroscience theories. On this front, we used our spiking retinal model to make an experimentally verifiable prediction regarding the retina's neural code: that a population differential-latency spike code can more robustly encode incoming stimuli than a population rate code under varying levels of noise. Furthermore, our model might have clinical implications in the development of retinal prosthetics. Current prosthetics use standard computer vision algorithms to translate camera signals into a stimulatory code [334, 335], which fall short of delivering a normal visual experience [336]. We speculate that the sensory transformation of our model could aid in delivering a more ganglion-like stimulatory signal.

3.4 Methods

3.4.1 Spiking retinal model

Training data

We recorded a natural visual stimulus dataset for model training. Various types of scenes were obtained, categorized as *flow* (camera moving through space), *pan* (camera panning through space), *still* (fixed camera recording a still scene), and *still moving* (fixed camera recording an active scene). A total of 39 clips of 9.6 second duration were recorded at a frame rate of 120Hz and a spatial extent of 720×1280px. The dataset was divided into a training and test set, both containing a balanced number of the different recording types (training set: 12 *flow*, 8 *pan*, 3 *still* and 8 *still moving* clips; test set: 3 *flow*, 2 *pan*, 1 *still* and 2 *still moving* clips). All clips were pre-processed by converting the RGB channels to greyscale (using a weighted average of the colour channels of R:G:B = 30:59:11 [76]); bilinearly downsampling the spatial dimensions to 144×256 and centrally cropping the spatial dimensions to 140×240; normalizing all clips using the pixel mean and standard deviation of the training set; and artificially increasing the temporal resolution to 240Hz by duplicating every frame. The model was trained on randomly sampled 20×20 pixel spatial patches of temporally non-overlapping subsamples of 320ms (*i.e.* 77 frames). We artificially doubled the training data using data augmentation, by flipping each clip around the vertical axis [284]. Lastly, in contrast to the V1 model, no bandpass filtering was applied to the movies.

Spiking neurons

We modelled the ganglion neurons in the retina using a single hidden layer of $N = 400$ recurrently connected spiking units, comprised of a series of computational steps:

Step 1, photoreceptor noise: to emulate the noisiness of the photoreceptors [286], we added Gaussian sampled noise $\epsilon_p \sim \mathcal{N}(0, 0.01)$ to every pixel within the input movie stimulus $x \in \mathbb{R}^{T \times H \times W}$ (for $T = 77$ simulation time steps of $H = 20$ and $W = 20$ pixel spatial height and width), to produce noisy input movie stimulus $\tilde{x}$. This stands in contrast to the V1 model which did not include added Gaussian noise to the input movie stimulus.

Step 2, input current: the input current $I_i[t] \in \mathbb{R}$ to unit i at time t is derived from the noisy-input movie stimulus $\tilde{x} \in \mathbb{R}^{T \times H \times W}$, model output spikes $S[t-1] \in \mathbb{R}^N$ and a bias term $b_i \in \mathbb{R}$.

$$I_i[t] = \Big(b_i + \underbrace{\sum_{t'=1}^{T_E} \sum_{h=1}^{H} \sum_{w=1}^{W} W_{it'hw} \tilde{x}_{hw}[t - t' + 1]}_{\text{Feedforward current}} + \underbrace{\sum_{j=1, j \neq i}^{N} W_{ij}^{\text{rec}} \, S_j[t-1]}_{\text{Recurrent current}} \Big) \tag{3.1}$$

The feedforward connectivity $W \in \mathbb{R}^{N \times T_E \times H \times W}$ linearly maps the noisy-input movie stimulus to a feedforward current contribution and captures the transformation of the pre-RGC cells. We set the span of latencies to $T_E = 30$ time steps (125ms) in the feedforward connectivity, consistent with the potentially wide integration times of the photoreceptors [324]. The recurrent connectivity $W^{\text{rec}} \in \mathbb{R}^{N \times N}$ (excluding autapse connections) maps the output spikes into a recurrent current contribution to capture the ganglion-to-ganglion cell and ganglion-to-amacrine-to-ganglion cell interactions. Most of the connections between the amacrine cells and ganglion cells are synaptic [72], with the majority of the amacrine cells using spikes, rather than graded potentials for communication [337, 338], hence the modelling decision for constructing the recurrent connectivity from the spiking unit output, rather than the membrane potential via gap-like junctions. Unlike in the V1 model, units in the retinal model were not constrained to be excitatory or inhibitory in their recurrent connectivity.

Step 3, ganglionic noise: to emulate the ganglionic noise [294], we multiplicatively perturbed the input current $I_i[t]$ of unit i at every time step t using Gaussian sampled noise $\epsilon_g \sim \mathcal{N}(0, 0.6)$.

$$\tilde{I}_i[t] = I_i[t](1 + \epsilon_g) \tag{3.2}$$

Step 4, ganglionic transformation: We modelled all ganglion spiking units using the normalized and discretized leaky integrate-and-fire model [128]. This evolves the model membrane potential $V_i[t] \in \mathbb{R}$ of unit i at time t using the difference equation:

$$V_i[t] = \Big(\beta_i V_i[t-1] + (1 - \beta_i) I_i[t]\Big)\Big(1 - S_i[t]\Big) \tag{3.3}$$

The membrane potential decays by a learnt factor β_i at every simulation step and resets to zero if a spike occurred in the previous simulation step. A spike is emitted if the membrane potential reaches the firing threshold (equal to one).

$$S_i[t] = \begin{cases} 1 & \text{if } V_i[t] > 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

Step 5, stimulus estimation: finally, at every simulation step t, a span of $T_D = 2$ steps of proceeding spike activity, plus a bias $b^{\text{proj}} \in \mathbb{R}$, was linearly mapped using projection weights $P \in \mathbb{R}^{N \times T_D \times H \times W}$ to generate an estimation of the stimulus movie frame $\hat{x}[t] \in \mathbb{R}^{H \times W}$ (either of the present or the future, with $\hat{x}_{hw}[t] \in \mathbb{R}$ denoting the projected pixel value).

$$\hat{x}_{hw}[t] = b^{\text{proj}} + \sum_{i=1}^{N} \sum_{t'=1}^{T_D} P_{it'hw} S_i[t - t' + 1] \tag{3.5}$$

Bounding membrane time constants

We bound the β_i value of every spiking unit to an appropriate range to ensure that the network employed biologically realistic membrane potentials [131].

$$\beta_i = \begin{cases} 0.001 & \text{if } \beta_i < 0.001 \\ 0.999 & \text{if } \beta_i > 0.999 \end{cases} \tag{3.6}$$

Weight initialisation

The initial feedforward W, recurrent W^{rec} and projection P weights were all sampled from a uniform distribution $U(-k, k)$, with $k = \frac{1}{\sqrt{T_E \cdot H \cdot W}}$, $k = \frac{0.1}{\sqrt{N}}$ and $k = \frac{1}{\sqrt{N \cdot T_D}}$ for the feedforward, recurrent and projection weights, respectively. All initial membrane time constants were set to 20ms (*i.e.* $\beta_i \approx 0.81$). All initial bias terms were set to zero.

Loss function

The model was trained, by minimizing loss $\mathcal{L}_{\text{total}}$, to generate a single-frame projection of the natural movie stimulus at every simulation step t (corresponding to loss $\mathcal{L}_{\text{projection}}$) under a metabolic-like constraint (corresponding to loss $\mathcal{L}_{\text{metabolic}}$), weighted by hyperparameter $\lambda = 10^{-2.5}$ (which by qualitative inspection we determined to produce the most retina-like receptive fields across all models). This chosen level of regularization is the most important factor for the differences in results between the V1 and retinal models, where the V1 model was trained with a lower metabolic regularization.[1]

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{projection}} + \lambda \underbrace{\left(\gamma \mathcal{L}_{\text{graded}} + (1-\gamma)\mathcal{L}_{\text{spiking}}\right)}_{\text{Metabolic loss } \mathcal{L}_{\text{metabolic}}} \tag{3.7}$$

The projection loss $\mathcal{L}_{\text{projection}}$ is the mean squared error (MSE) between every pixel in the generated single-frame projection $\hat{x}[t] \in \mathbb{R}^{H \times W}$ and the corresponding natural stimulus target frame $x[t + T_{\text{off}}^{\text{steps}}] \in \mathbb{R}^{H \times W}$, where $T_{\text{off}}^{\text{steps}}$ specifies if the frame is of the present (*i.e.* $T_{\text{off}}^{\text{steps}} = 0$) or of the future (*i.e.* $T_{\text{off}}^{\text{steps}} > 0$). This was computed as the average over all batch samples B (which we omit from the notation for brevity); simulation steps T (starting from time step t_0 onwards, to allow the unit membrane potentials to sufficiently depolarize, *i.e.* to permit the network to *warm up*); and spatial height dimension H and width dimension W, both cropped by c pixels on all sides.

$$\mathcal{L}_{\text{projection}} = \underbrace{\frac{1}{(T-t_0)(H-c)(W-c)} \sum_{t=t_0}^{T} \sum_{h=c}^{H-c} \sum_{w=c}^{W-c}}_{\text{Avg. over time and space}} \underbrace{(\hat{x}_{hw}[t] - x_{hw}[t + T_{\text{off}}^{\text{steps}}])^2}_{\text{MSE}} \tag{3.8}$$

The metabolic-like loss $\mathcal{L}_{\text{metabolic}}$ is an approximation of the biological energy utilization of the network. We took this to be a proxy of the energy expenditure at every synapse, calculated as the absolute input value to a synaptic weight, multiplied by its respective absolute value. We weighted the metabolic loss of the graded pre-ganglionic-like units $\mathcal{L}_{\text{graded}}$ and the spiking

[1] Although the $\lambda = 10^{-2.75}$ used in the V1 model is similar to the $\lambda = 10^{-2.5}$ used in the retinal model, the retinal model was trained using an artificially increased frame rate of 240Hz compared to the 120Hz used for the V1 model. Both models used an input weight kernel of the same temporal duration, thus the retinal model has twice as many input weights compared to the V1 model. Employing a similar penalty thus penalizes the retinal model more.

ganglionic-like units $\mathcal{L}_{\text{spiking}}$ separately using hyperparameter $\gamma = 0.3$, as spiking neurons consume more energy than graded neurons [339, 340].

$$\mathcal{L}_{\text{graded}} = \underbrace{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}}_{\text{Avg. over units and time}} \underbrace{\Big(|b_i| + \sum_{t'=1}^{T_E}\sum_{h=1}^{H}\sum_{w=1}^{W}|W_{it'hw}||x_{hw}[t-t'+1]|\Big)}_{\text{Pre-ganglionic synaptic cost}} \tag{3.9}$$

$$\mathcal{L}_{\text{spiking}} = \underbrace{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}}_{\text{Avg. over units and time}} \underbrace{\Big(\sum_{j=1,j\neq i}^{N}|W_{ij}^{\text{rec}}|S_j[t-1]\Big)}_{\text{Recurrent synaptic cost}} \tag{3.10}$$

$$+ \underbrace{\frac{1}{THW}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{w=1}^{W}}_{\text{Avg. over time and space}} \underbrace{\Big(\sum_{i=1}^{N}\sum_{t'=1}^{T_D}|P_{it'hw}|S_i[t-t'+1]\Big)}_{\text{Post-ganglionic synaptic cost}} \tag{3.11}$$

$$\tag{3.12}$$

Surrogate gradient descent

All spiking retinal models were optimized using a modification of the backpropagation algorithm [173], called surrogate gradient descent [198]. To permit the flow of gradient, the gradient of the non-differentiable Heaviside step function was replaced with a well-behaved function. We adopted the fast sigmoid function, which has shown to work well in practice [194, 133].

$$\frac{\partial S_i[t]}{\partial V_i[t]} = (10|V_i[t]-1|+1)^{-2} \tag{3.13}$$

All training was performed using the Adam optimizer [289] (with default parameters) over 100 epochs, with an initial learning rate of 10^{-4} (decayed to 10^{-5} after 50 epochs) and batch size of 1024. Model weights were saved during training whenever a new minimum training loss was obtained.

3.4.2 Receptive field analysis

Spike-triggered average

We estimated the spatiotemporal receptive field of every unit using a spike-triggered average, using responses to 200 white-noise input clips of 100-frame duration. Clips were generated by sampling each pixel from a Gaussian distribution with a standard deviation of $\sigma = 10$. We estimated the RF temporal profile by averaging over the spatial pixel values for every frame in time [341].

Fitting 2D Gaussians

We fitted a two-dimensional Gaussian function [75, 76] to the spatial RF using gradient descent, with the function defined as:

$$G(x,y) = \frac{A}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\Big(-\frac{1}{2(1-\rho^2)}\Big(\Big(\frac{x-\mu_X}{\sigma_X}\Big)^2 - 2\rho\Big(\frac{x-\mu_X}{\sigma_X}\Big)^2\Big(\frac{y-\mu_Y}{\sigma_Y}\Big)^2 + \Big(\frac{y-\mu_Y}{\sigma_Y}\Big)^2\Big)\Big) \tag{3.14}$$

with (x, y) denoting the spatial position in pixel space; A the amplitude; (μ_X, μ_Y) the RF centre position; σ_X and σ_Y the standard deviation along the x- and y-axis, respectively, and ρ the RF orientation. For each RF, the parameters of the two-dimensional Gaussian were fitted by minimizing the mean squared error between the two-dimensional Gaussian and the RF. For the RF analyses, we only included units $(173/400)$ that had a good fit (those with a correlation coefficient > 0.7 and those with standard deviation σ_X and $\sigma_Y > 0.5$ pixels).

Classifying unit types

Similarly to classifying the different retinal cell types [76], we classified the functional type of every model unit that had a valid 2D Gaussian fit. For every such unit, we obtained the spatial RF with the largest power in time, and projected its RF diameter (calculated as $\sqrt{\sigma_X\sigma_Y}$) against the first principal component of its temporal profile. We classified the different functional unit types, by grouping each unit into four distinct groups using k-means clustering with four fixed centroids.

3.4.3 Spike analysis

Firing rate

We calculated the firing rate of a neuron (and model unit), by counting the number of spikes over a given stimulus presentation and dividing this by the stimulus duration, to obtain a firing rate.

Coefficient of variation

We computed the coefficient of variation of the interspike interval CV(ISI), to quantify the variability of the timing of spikes for a given neuron (and model unit). This is defined as the ratio between the standard deviation and mean of the interspike intervals [110], where a $CV(ISI) < 1$ corresponds to a regular firing pattern and a $CV(ISI) > 1$ corresponds to an irregular firing pattern. We dropped CV(ISI) values from the analysis if there were less than three spikes for a given stimulus presentation [27].

Fano factor

We reported the Fano factor for each neuron (and model unit) over each stimulus presentation, to estimate the spike count variability over the stimulus-repeat trials. The Fano factor is defined as the variance of the spike count divided by the mean spike count over stimulus-repeat trials.

Image decoding from single cell responses

We adopted the experimental paradigm of [201], and obtained the activity of an OFF-type model unit in response to a 150ms-flash of a gray-scale image of a swimming salamander larva. Similar to the experimental paradigm, we convolved the model unit along the x- and y-direction (using a stride of 2 pixels), to obtain the model unit spike activity at different spatial locations. The image was then reconstructed - with every pixel denoting a unique spatial location - using either the number of spikes or the relative timing of the first spikes.

Image decoding from population responses to noise-corrupted inputs

We performed all analyses using a selected set of 300 natural photographs from the McGill Calibrated Colour Image Database [342]. The spatial dimensions were rescaled to 64×64 pixels using bicubic interpolation, with 240 images used for training and cross-validation and the remaining 60 images used for testing. For every experiment, we appended different sampled Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma_s)$ to every image pixel, and varied the distribution standard deviation σ_s from 0 to 2.0 in 0.25 increments between the different experiments. We passed the set of noise-corrupted images into the model (using a stride of 4 pixels) and transformed the resulting spike outputs into a rate code and a relative spike latency code. We then fit a separate linear readout model from each of these respective spike codes, to reconstruct the original images. We penalized the readout weights using an L1 penalty, with the optimal penalty weighting found using five-fold cross-validation. Lastly, we fitted each model using its optimal regularization penalty on all of the training data and reported the mean pixel correlation coefficient of the reconstructed images on the held-out test set.

Classifying units with an omitted stimulus response

Ganglion cells exhibit OSRs at the end of a flash sequence around the expected time of the next flash. These OSRs have been characterized to be stronger than the responses during the flash sequence [93]. Prior studies classified OSRs via qualitative inspection [93, 210]. We developed a set of conditions using the description of OSRs and classified units as exhibiting OSRs if they satisfied the following constraints:

$$1. \quad \frac{R_{OM}}{R_{FL}} > 1.05 \tag{3.15}$$

$$2. \quad \frac{R_{OM}}{R_{BA}} > 1.2 \tag{3.16}$$

$$3. \quad R_{OM} > 24\text{Hz} \tag{3.17}$$

Here, R_{OM} was the maximum response during the 80ms duration following the termination of the flash sequence; R_{FL} was the maximum response during the flash sequence; and R_{BA} was the baseline response, which we took to be the maximum value during the constant 150ms light flash proceeding the flash sequence. Condition one means that the response directly following the flash sequence was larger than the responses during the flash sequence; condition two means that the response directly following the flash sequence was larger than the maximum response to the non-flashing sequence; and condition three means that the response directly following the

flash sequence was not the result of possible noise fluctuations (threshold chosen via qualitative inspection).

3.4.4 Motion tuning

Virtual physiology

We measured the model unit response properties to full-field sinusoidal gratings (of amplitude one) of varying orientation ($0°$ to $360°$ in $5°$ increments), spatial frequency (10 evenly spaced values between 0.01 to 0.2 cycles per pixel) and temporal frequency ($1, 2, 4$ and 8Hz). We presented each grating to each unit for 3s and computed the resulting PSTH by averaging over 8 repeats and convolving with a Gaussian kernel ($\sigma = 37$ms). Next, we computed the mean firing rate over time (ignoring the first 41ms for network warmup, as done in training), resulting in a 3D mean-firing tuning space for each unit (with the preferred stimulus eliciting the highest mean response). We dropped all non-responsive units ($113/400$ units) from the analysis, defined as those units which had a preferred response less than 1% of the mean response to the preferred stimulus over all units.

Orientation and direction selectivity

We measured the orientation selectivity index (OSI) and direction selectivity index (DSI) for each unit, which respectively quantify the tuning preference for motion across a particular orientation and direction:

$$\text{OSI} = \frac{R^{\text{orient}}_{\text{pref}} - R^{\text{orient}}_{\text{orth}}}{R^{\text{orient}}_{\text{pref}} + R^{\text{orient}}_{\text{orth}}} \tag{3.18}$$

$$\text{DSI} = \frac{R^{\text{dir}}_{\text{pref}} - R^{\text{dir}}_{\text{non-pref}}}{R^{\text{dir}}_{\text{pref}} + R^{\text{dir}}_{\text{non-pref}}} \tag{3.19}$$

with $R^{\text{orient}}_{\text{pref}} = R^{\text{dir}}_{\text{pref}}$ being a unit's response to a stimulus moving in the preferred direction, and $R^{\text{orient}}_{\text{orth}}$ and $R^{\text{dir}}_{\text{non-pref}}$ being the responses to the orthogonal orientation and opposite direction, respectively (all using a unit's preferred spatial and temporal frequency). For both metrics, a

value close to zero corresponds to weaker tuning, and a value close to unity corresponds to stronger tuning.

Counting Y-type-like units

We presented full-field sinusoidal gratings with a spatial frequency of 0.3 cycles per pixel, at each unit's preferred orientation. We then constructed each unit's firing rate for the stationary grating R_{SG} and the grating moving at each unit's preferred temporal frequency R_{MG}. We considered a unit to be Y-type-like if the following constraints were satisfied:

$$1. \quad R_{MG} > R_{SG} \tag{3.20}$$

$$2. \quad R_{SG} < 24\text{Hz} \tag{3.21}$$

$$3. \quad R_{MG} > 24\text{Hz} \tag{3.22}$$

Condition one means that the moving grating responses were larger than the stationary grating responses; and conditions two and three mean that the stationary and moving grating responses were respectively below and above a firing threshold (chosen via qualitative inspection) to potentially avoid counting units as Y-type-like due to noise fluctuations.

Anticipation code

We implemented the experimental protocol of [92] as closely as possible, where the authors projected a dark bar on a white background moving $6.7\mu m$ across the retina every 15ms, or where they flashed the bar directly over a unit's RF for 15ms. We replicated this experiment for the model by projecting a dark bar (pixel value of -1) on a white background (pixel value of 1), moving one pixel every 150ms (36 frames). We calculated this to match the movement speed of the bar in the experiment of [92], assuming a ganglion RF diameter of $300\mu m$ [343] and a unit RF diameter of 5 pixels, which corresponds to moving the bar by ≈ 0.11 pixel every 4 frames (16.66ms) (or one pixel every 36 frames).

3.4.5 Predicting neural responses from model activity

We used publicly available and spike-sorted recordings of RGCs of different animal species in response to natural images and movies. For all datasets, we used 80% of the data for training

and the remaining 20% for testing (with splits separated by unique clips or images). As the datasets were recorded at different sampling rates (*i.e.* the frame rate of the projector or the sampling rate of the spike responses), we resampled all data to be 240Hz (the same sampling rate of the natural movie stimulus with which the spiking retinal model was trained). We also normalized all natural stimuli (clips and images), by subtracting the mean and dividing by the standard deviation of the training set. We averaged the recorded spike responses over trials and convolved the resulting PSTH with a Gaussian kernel (using $\sigma = 37$ms) to obtain a smoothed spike firing probability for every neuron [69].

Macaque image and movie dataset

We used recordings from ON- and OFF-parasol cells in the peripheral retina of macaque monkeys in response to natural images and movies (made available by [297]). This consisted of 16 cells recorded in response to 48 250ms-flashed natural images over 5 trials; and 24 cells recorded in response to 7 6s-long natural movies over 4 trials, with the movies consisting of the natural images spatially shifted in time using eye movement trajectories of human observers [344]. All stimuli were projected at a frame rate of 60Hz onto the receptive field center of each neuron using a circular aperture.

Marmoset movie dataset

We used recordings from ON- and OFF-midget and ON- and OFF-parasol cells in the retina of marmoset monkeys in response to natural movies (made available by [69]). This includes 423 cells recorded in response to 22 1s-long natural movies over 30 trials, with the movies consisting of the natural images spatially shifted in time using eye movement trajectories of awake, head-fixed marmoset monkeys. All stimuli were projected at a frame rate of 85Hz. Lastly, to keep the fitting procedure computationally manageable, we only fitted a subset of the most responsive neurons, as measured by the maximum correlation coefficient (see Section 3.4.5), resulting in 164 neurons (using $CC_{\text{max}} \geq 0.93$).

Mouse movie dataset

We used recordings from ON- and OFF-midget and ON- and OFF-parasol cells in the mouse retina in response to natural movies (made available by [69]). These 1791 cells were recorded in response to 22 1s-long natural movies over 30 trials, with the movies consisting of the

natural images spatially shifted in time using the horizontal gaze component of freely-moving mice. All stimuli were projected at a frame rate of 75Hz. Again, to keep the fitting procedure computationally manageable, we only fitted a subset of the most responsive neurons, as measured by the maximum correlation coefficient (see Section 3.4.5), resulting in 351 neurons (using $CC_{\text{max}} \geq 0.93$).

Salamander image

We used recordings of various ON- and OFF-type cells in the salamander retina in response to natural images (made available by [76]). This includes 215 cells recorded in response to 300 natural images flashed for 200ms each over 12 trials. There was an 800ms inter-stimulus interval between image presentations, with spike responses to each image recorded between image onset and 100ms after image offset. The images came from the Van Hateren database [211] over 12 trials. All stimuli were projected at a frame rate of 30Hz.

Obtaining latent model activity

We trained a set of 14 retinal models, one for every temporal offset T_{off} into the future ($T_{\text{off}} \in \{0, 8, 16, 24, 32, 40, 48, 56, 64, 72, 80, 96, 112, 128\}$ms, where $T_{\text{off}} = 0$ corresponds to the present). We convolved each retinal model (using a stride of 4 pixels) over the spatial dimensions of each dataset, to obtain the unit spike activity $D \in \mathbb{R}^{\tilde{T} \times \tilde{H} \times \tilde{W}}$ (where the number of simulation steps $\tilde{T}$, and spatial height $\tilde{H}$ and width $\tilde{W}$ dimensions varied between the different datasets). This was repeated 8 times, and averaged over repeats, to obtain a new activity dataset $D^{\text{prob}} \in \mathbb{R}^{\tilde{T} \times \tilde{H} \times \tilde{W}}$, containing the spike probability of every unit over space and time. We further compressed this dataset to a latent activity dataset $D^{\text{latent}} \in \mathbb{R}^{\tilde{T} \times 500}$, using the first $N_{\text{PC}} = 500$ principal components over space. This technique is routinely used [318, 319, 155, 320], as it reduces overfitting [318] and keeps the fitting procedure computationally tractable [321].

Fitting procedure

For every retinal model, we fitted a linear-nonlinear readout, to predict the response $\hat{r}_i \in \mathbb{R}^{\tilde{T}}$ of every neuron i in every dataset from the latent model activity.

$$\hat{r}_i[t] = \exp\Big(b_i^{\text{read}} + \sum_{t'=1}^{T_R}\sum_{j=1}^{N_{\text{PC}}} W_{ij}^{\text{read}}[t-t'+1]D^{\text{latent}}[t-t'+1]\Big) \tag{3.23}$$

with readout weights $W_i^{\text{read}} \in \mathbb{R}^{N_{\text{PC}}\times T_R}$ using a $T_R = 5$ temporal frame span equivalent to ≈ 21ms (except for the mouse movie dataset where we used a span of 46ms). Each readout was jointly fitted across all $\tilde{N}$ neurons of a dataset, by minimizing the negative Poisson log-likelihood of the readout predictions $\hat{r}_i[t]$ and the neuron's responses $r_i[t]$, and an L1 penalty on the readout weights (weighted by hyperparameter λ^{read}) to avoid overfitting.

$$\mathcal{L}^{\text{out}} = -\frac{1}{\tilde{N}\tilde{T}}\sum_{i=1}^{\tilde{N}}\sum_{t=1}^{\tilde{T}}\Big(r_i[t]\ln(\hat{r}_i[t]) + \hat{r}_i[t]\Big) + \lambda^{\text{read}}\Big(\sum_{i=1}^{\tilde{N}}\Big|\Big|W_i^{\text{read}}\Big|\Big|_1\Big) \tag{3.24}$$

We trained each readout using the Adam optimizer (with default parameters) using a learning rate of 10^{-4} over 2000 epochs with a batch size of 8 subsamples of at most 600ms duration (except the salamander dataset where we used a batch size of 32). We found the optimal regularization hyperparameter λ^{read} using five-fold cross-validation (over a set of five different log-spaced values, which we varied between the different datasets). Lastly, we also explored different spatial scales of the input stimuli, to account for differences in spatial scale between the model units and neurons [51, 155]. To determine the optimal scale for each dataset, we ran all fits using three different spatial scales ($\times 0.66$, $\times 1$, and $\times 1.5$) using the retinal model optimized for efficient compression of the present (see Supplementary). We then chose the spatial scale which performed the best on each dataset to fit the remaining models.

Normalised and maximum correlation coefficient

We used the normalized correlation coefficient CC_{norm} to quantify the neural fits, expressed as the Pearson correlation coefficient CC between the model's prediction and target neural response, normalized by the maximum obtainable correlation coefficient CC_{max} to take into account neural noise [317, 345]. A value of zero indicates no correlation between the predicted and target neural response, whereas a value of one indicates the best possible score.

Supplementary

Dataset	Scale	CC
Macaque movie	× 0.66	0.35
	× **1.0**	**0.37**
	× 1.5	0.35
Macaque image	× 0.66	0.55
	× **1.0**	**0.63**
	× 1.5	0.47
Marmoset movie	× 0.66	0.24
	× 1.0	0.25
	× **1.5**	**0.26**
Salamander image	× 0.66	0.46
	× **1.0**	**0.50**
	× 1.5	0.49
Mouse movie	× 0.66	0.13
	× 1.0	0.16
	× **1.5**	**0.17**

Tab. 3.1: Prediction score using the compression retinal model across the different datasets at different spatial scales. Bold values indicate the best performing scale, which was chosen to fit the remaining predictive retinal models in each respective dataset.

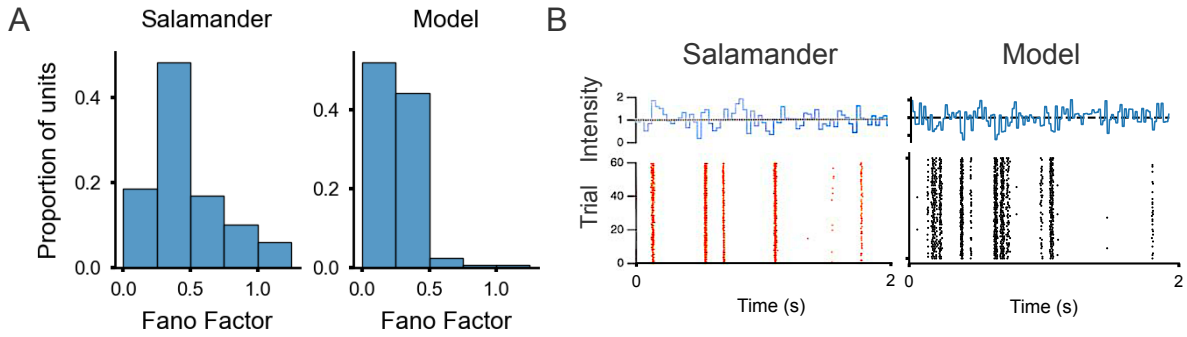

Fig. 3.7: **A.** Histogram of Fano factors in the salamander retina (left) (data from [76]) and the model (right) in response to the same natural images. **B.** Spike raster over 60 trials in response to a uniform full-field illumination of randomly-varying intensity in the salamander retina [86] (left) and the model (right).

4 Addressing the speed-accuracy simulation trade-off for adaptive spiking neurons

4.1 Introduction

The surge of progress in artificial neural networks (ANNs) over the last decade has advanced our understanding of the potential computational principles underlying the processing of sensory information [346, 162, 51, 347, 199, 163, 155, 68, 320]. Although these networks architecturally bear a resemblance to the brain [348], they tend to omit a key physiological constraint: the spike. With their increased biological realism, spiking neural networks (SNNs) have shown great promise in bridging the gap between experimental data and computational models. SNNs can be fitted to real neural data [139, 349, 350, 351, 352], or used to simulate the brain, offering a new level of understanding of the complex workings of the nervous system [353, 121, 354, 355]. They also have engineering applications in energy-efficient machine learning [356].

An established class of spiking models in computational neuroscience is the leaky integrate-and-fire (LIF) neuron, with origins dating back to 1907 [357]. Just like in real neurons, input current (resulting from presynaptic input) charges the membrane potential of the model neurons, which then output binary signals (*i.e.* spikes) as a form of communication (Figure 4.1a). The adaptive leaky integrate-and-fire (ALIF) model [110] is a modern extension of the LIF. It more closely mimics the biology, capturing a key property of neurons, which is their adaptive firing threshold (*i.e.* spikes become less frequent in response to a steady input current [7]). ALIF neurons have been shown to accurately fit real neural recordings [139, 110, 358, 359] and to outperform the simpler LIF neurons on various machine learning benchmarks [121, 196, 360, 361, 362].

Despite these modelling advances, a major shortcoming of LIF and ALIF neurons is their slow inference and training times. Unlike real neurons, modelling neuron dynamics involves sequential computation over discretised time. This leads to a problematic trade-off between speed and accuracy when simulating SNNs. A small DT enables accurate modelling of dynamics, but is slow to stimulate and train on computer systems such as GPUs. A large DT obtains less accurate

91

dynamics, but at the benefit of being faster to simulate and train [363] (Figure 4.1b). This raises the important question of capturing the best of both worlds: **is there a way to accelerate the inference and training of spiking LIF and ALIF neurons without sacrificing simulation accuracy?**

In this work, we address the speed-accuracy trade-off when simulating and training LIF and ALIF SNNs, and present a solution that is both fast and accurate. We take advantage of a fundamental property of neurons that is sometimes not modelled - the absolute refractory period (ARP). This is a short period of time following spike initiation, during which a neuron cannot fire again. As a result, a neuron can spike at most once within such a period (Figure 4.1c). We leverage this observation to develop a novel algorithmic reformulation of the ALIF model which reduces the sequential simulation complexity of the standard ALIF algorithm. Specifically, we outline how ALIF recurrent networks can be simulated with a constant sequential complexity $O(1)$ over the ARP simulation length T_R, and how this approach can be extended to longer simulation lengths to obtain identical dynamics to the standard ALIF network algorithm. Faster simulation and training are theoretically obtained for growing length T_R, by either employing physiologically plausible ARPs of ~ 2ms and decreasing the DT to a very fine value ~ 0.1ms (for realistic neural modelling), or setting the DT to a coarser value and adopting larger non-physiological ARPs (for machine learning tasks). Our main contributions are the following:

- We develop a novel algorithmic reformulation of the ALIF model with an $O(T/T_R)$ - rather than $O(T)$ - sequential complexity, for simulation length T and ARP length T_R.[1]
- We find that our model achieves substantial inference (up to $40\times$) and training speedup (up to $53\times$) over the standard ALIF SNN for increasing ARP time steps, and find this to hold over different numbers of simulation lengths, batch sizes, number of neurons and layers.
- We demonstrate the feasibility of our model to be trained using surrogate gradient descent, with our accelerated ALIF SNN achieving comparable accuracies to the standard ALIF SNN on temporal spiking classification datasets - yet in a fraction of the training time.
- Finally, we showcase how our ALIF implementation makes it possible to quickly fit real electrophysiological recordings of cortical neurons, where very fine sub-millisecond discretisation steps are important for accurately capturing exact spike timing.

[1] To avoid ambiguity, simulation length T and ARP length T_R are number of time steps (dimensionless) and not unit time. Additionally, in a slight abuse of notation we include the constant T_R in the complexity expression.

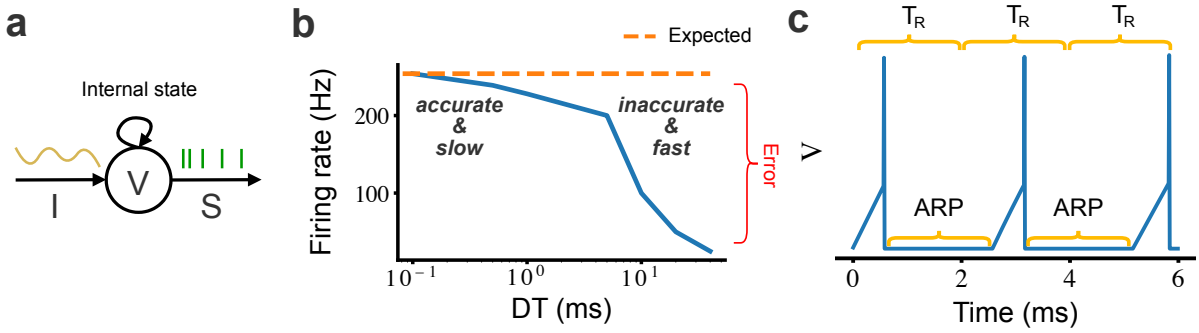

Fig. 4.1: Problem overview. a. Schematic of an ALIF neuron: input current I charges membrane potential V and outputs spikes S if firing threshold is reached (with the neuron's internal state evolving over time). **b.** An example of the simulation trade-off problem when simulating a single ALIF neuron with fixed synaptic weights receiving Poisson spike input. The simulation error and the speed grow for increasing discretisation time (DT). **c.** Observation for our solution: a neuron emits at most a single spike during a simulation span T_R equal in length to the neuron's absolute refractory period (ARP).

4.2 Background and related work

Standard ALIF model We introduce the recurrent SNN of ALIF neurons with fixed ARP and latency of recurrent transmission, defined by the following set of equations [196, 364].

$$S_i^{(l)}[t] = f(V_i^{(l)}[t]) = \mathbb{1}_{V_i^{(l)}[t] > \theta_i^{(l)}[t]} \quad \text{(Output spike)} \tag{4.1}$$

$$V_i^{(l)}[t] = (\beta_i^{(l)} V_i^{(l)}[t-1] + (1-\beta_i^{(l)}) I_i^{(l)}[t])(1 - S_i^{(l)}[t-1]) \quad \text{(Membrane potential)} \tag{4.2}$$

$$\tilde{I}_i^{(l)}[t] = \Big(b_i^{(l)} + \underbrace{\sum_{j=1}^{N^{(l-1)}} W_{ij}^{(l)} S_j^{(l-1)}[t]}_{\text{Feedforward current}} + \underbrace{\sum_{j=1}^{N^{(l)}} W_{ij}^{\text{rec }(l)} S_j^{(l)}[t-D]}_{\text{Recurrent current}}\Big) \quad \text{(Input current)} \tag{4.3}$$

$$I_i^{(l)}[t] = \begin{cases} \tilde{I}_i^{(l)}[t] & \text{if } C_i^{(l)} \geq T_R \\ 0 & \text{otherwise} \end{cases} \quad \text{(Absolute refractory period)} \tag{4.4}$$

$$\begin{aligned} \theta_i^{(l)}[t] &= 1 + d_i^{(l)} a_i^{(l)}[t] \\ a_i^{(l)}[t] &= p_i^{(l)} a_i^{(l)}[t-1] + S_i^{(l)}[t-1] \end{aligned} \quad \text{(Adaptation)} \tag{4.5}$$

At time t, neuron i within layer l (consisting of $N^{(l)}$ neurons) receives input current $I_i^{(l)}[t]$ and outputs a binary value $S_i^{(l)}[t] \in \{1, 0\}$ (*i.e.* spike or no spike) if a neuron's membrane potential

$V_i^{(l)}[t]$ reaches firing threshold $\theta_i^{(l)}$ (Equation 4.1).[2] The evolution of the membrane potential is described by the normalised discretised LIF equation (Equation 4.2), in which the membrane potential dissipates by a learnable factor $0 \leq \beta_i^{(l)} \leq 1$ at every time step and resets to zero if a spike occurred at the previous time step. The input current is comprised of a constant bias source $b_i^{(l)}$ and derived from incoming spikes reflecting feedforward $W^{(l)} \in \mathbb{R}^{N^{(l)} \times N^{(l-1)}}$ and recurrent connectivity $W^{\text{rec }(l)} \in \mathbb{R}^{N^{(l)} \times N^{(l)}}$ (Equation 4.3).

Here, we assume the recurrent transmission latencies D to be of fixed length and equal in length to the ARP, that is $T_R = D$. With a simple modification however, our methods can work for longer D, or different D on each connection, so long as $D \geq T_R$. The ARP is enforced by only allowing input current to enter the neuron if the number of time steps $C_i^{(l)}$ following the last spike equals or exceeds the ARP length (Equation 4.4). Lastly, adaptation is implemented by raising the firing threshold $\theta_i^{(l)}[t]$ following each spike $S_i^{(l)}[t-1]$ (Equation 4.5), which decays exponentially to baseline $\theta_i^{(l)} = 1$ in the absence of any spikes (using learnt decay factor $0 \leq p_i^{(l)} \leq 1$ and adaptation scalar $d_i^{(l)}$).

The ARP in biological neurons is typically about $\sim 1-2$ms [265, 266, 267]. Our method takes advantage of the fact that the monosynaptic connection latency between neurons in local circuits is typically also often around $\sim 1-2$ms [365]. LIF and ALIF neuronal models are typically run with a DT of about ~ 0.1ms in the computational neuroscience literature and ~ 1ms in machine learning literature [363]. Furthermore, the firing rate of neurons is often less than the reciprocal of the ARP, suggesting that a higher T_R than the ARP may still provide a reasonable approximation of neural behaviour for some purposes.

SNN training The main problem with training SNNs is the non-differentiable nature of their activation function in Equation 4.1 (whose derivative is undefined at $V_i^{(l)}[t] = \theta_i^{(l)}[t]$ and zero otherwise). This precludes the direct use of the backprop algorithm [173], which has underpinned the successful training of ANNs. A popular solution is to replace the undefined spike derivative with a well-behaved function, referred to as a surrogate gradient [192, 193, 194, 195], and training the network with backprop-through-time [198]. This method also supports the training of neural parameters other than synaptic connectivity, such as membrane time constants [131, 361] and adaptive firing thresholds [196, 360, 361], both of which we include in our work, as they have been shown to improve performance (and increase biological realism). It is worth mentioning that other SNN training methods exist such as mapping trained ANNs to SNNs by transforming ANN neuron activations to SNN neuron firing rates [174, 175, 176, 177].

[2] Here notation $\mathbb{1}_{\text{condition}}$ denotes the indicator function, which is equal to one if the condition is true and zero otherwise.

These methods are, however, of less interest to computational neuroscientists as they discard all temporal spike precision.

Related work Recent work has provided new theoretical insights for increasing the simulation accuracy when employing a larger DT $\geq$ 1ms [363]. We are not aware of any work that accelerates the simulation and training times on GPUs when employing a smaller DT $\leq$ 1ms (although see the NEST simulation library for simulating SNNs on CPUs [366, 367, 368]). There are, however, different methods for speeding up the backward pass (*i.e.* how gradients are computed within the SNN), which consequentially speeds up training times. Rather than being viewed as competing approaches, these methods could further augment the speed of our solution. In sparse gradient descent, gradients are only passed through neurons whose membrane potential is close to firing threshold, which can significantly accelerate the backward pass when neurons are mostly silent [132]. Inspired by work on training non-spiking networks [369, 370], other methods completely bypass the backward pass and adjust weights online [371, 360, 362]. Another approach is to propagate gradient information through spikes [186, 187, 188, 189, 372, 373, 374], which - similar to the idea of sparse gradient descent - is fast when neurons are mostly silent. This method, however, enforces neurons to spike at most once and can suffer from training instabilities (although see [190, 191]).

4.3 Theoretical results

We outline a novel reformulation of the ALIF model, which theoretically reduces the sequential simulation complexity from $O(T)$ to $O(T/T_R)$ (for simulation length T and ARP length T_R). Using the observation that a neuron can spike at most once over simulation length T_R (equal to the ARP; Figure 1c), we propose simulating network dynamics sequentially in blocks of length T_R - as opposed to simulating every time step individually (Figure 4.2a) - as we show that these blocks can be simulated with a constant sequential complexity $O(1)$. Our reformulated ALIF model is mathematically the same as the standard ALIF model, but substantially faster to simulate and train. All the proofs for the propositions can be found in the Supplementary material.

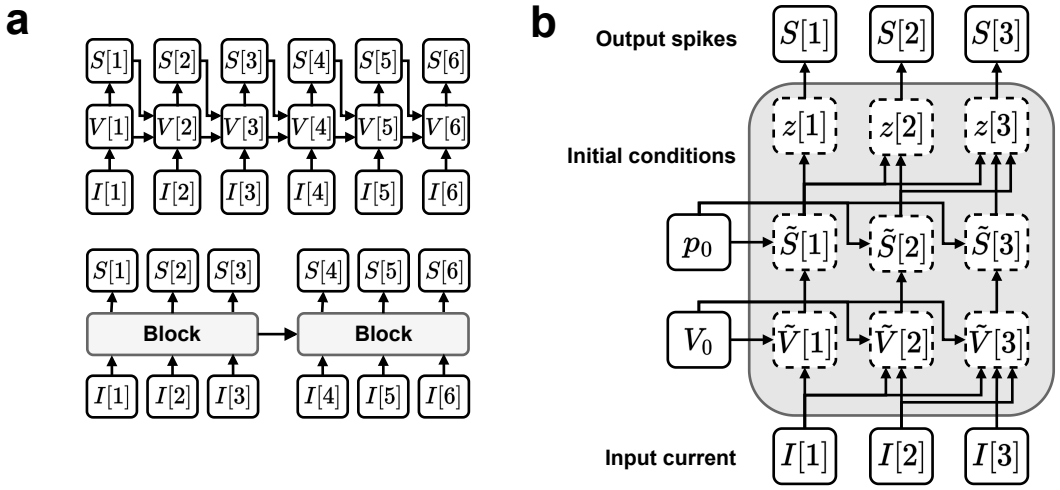

Fig. 4.2: Solution overview. **a.** Our proposed solution: instead of simulating network dynamics one time step after another (top), we sequentially simulate blocks of time equal in length to the neuron ARP (bottom), in which a neuron can spike at most once. **b.** A schematic of a Block: our proposed solution for emulating ALIF dynamics with a constant sequential complexity $O(1)$ over a short duration in which a neuron spikes at most once.

4.3.1 Block: simulating single-spike ALIF dynamics with a constant sequential complexity

A SNN exhibits a sequential dependence due to the spike reset mechanism, where a neuron's membrane potential $V_i[t]$ can be reset based on its previous output $S_i[t-1]$. Consequently, the simulation of a SNN necessitates a sequential approach. This restriction can, however, be alleviated if we assume a neuron to spike at most once over a particular simulation length (such as the ARP). The following steps - grouped together into a module referred to as a Block (Figure 4.2b) - compute ALIF dynamics (assuming at most one spike) without any sequential operations.

$$\tilde{V}_i[t] = (I_i * \tilde{\beta}_i)[t] \quad \text{(No-reset membrane potential)} \tag{4.6}$$

$$\tilde{S}_i[t] = f(\tilde{V}_i[t]) \quad \text{(Faulty output spikes)} \tag{4.7}$$

$$z_i[t] = \phi(\tilde{S}_i)[t] \quad \text{(Latent timing of spikes)} \tag{4.8}$$

$$S_i[t] = \mathbb{1}_{z_i[t]=1} \quad \text{(Correct output spikes)} \tag{4.9}$$

1. Calculate membrane potentials without reset The first step in the Block (Equation 4.6) converts input current $I_i[t]$ to membrane potentials $\tilde{V}_i[t]$ without spike reset (*i.e.* excluding the reset mechanism in Equation 4.2). This transformation is achieved using a convolution (Proposition 1), thus avoiding any sequential operations.

Proposition 1. *Membrane potentials without spike reset are computed as a convolution* $\tilde{V}_i[t] = (I_i * \tilde{\beta}_i)[t]$ *between input current* $I_i[t]$ *and kernel* $\tilde{\beta}_i[t] = (1-\beta_i)\beta_i^t$ *with the initial membrane potential encoded as* $I_i[0] = \frac{V_i[0]}{1-\beta_i}$.

2. Faulty output spikes No-reset membrane potentials $\tilde{V}_i[t]$ are mapped to erroneous output spikes $\tilde{S}_i[t] = f(\tilde{V}_i[t])$ (Equation 4.7) using spike function f (Equation 4.1). This output can contain more than one spike, but only the first spike complies with the standard model dynamics, due to the omission of the reset mechanism and ARP constraint. Thus, to ensure that the Block only emits a single spike, all spikes succeeding the first spike occurrence are removed using the following steps.

3. Latent timing of spikes Erroneous spike output $\tilde{S}_i[t]$ is mapped to a latent representation $z_i[t] = \phi(\tilde{S}_i)[t]$ (Equation 4.8), encoding the timing of spikes. Function $\phi(\cdot)$ (taking vector $\tilde{S}_i$ as input; Proposition 2) is constructed to map all erroneous spikes $\tilde{S}_i[t]$, besides the first spike occurrence, to a value other than one (*i.e.* $z_i[t] \neq 1$ for all t except for the smallest t satisfying $\tilde{S}_i[t] = 1$ if such t exists).

Proposition 2. *Function* $\phi(\tilde{S}_i)[t] = \sum_{k=1}^{t} \tilde{S}_i[k](t-k+1)$ *acting on* $\tilde{S}_i \in \{0,1\}^T$ *contains at most one element equal to one* $\phi(\tilde{S}_i)[t] = 1$ *for the smallest* t *satisfying* $\tilde{S}_i[t] = 1$ *(if such* t *exists).*

4. Correct output spikes Lastly, the correct spike output $S_i[t] = \mathbb{1}_{z_i[t] = 1}$ is obtained by setting every value in $z_i[t]$, besides the value one (*i.e.* the first spike), to zero (Equation 4.9).

4.3.2 Blocks: simulating ALIF SNN dynamics with a $O(T/T_R)$ sequential complexity

The standard ALIF SNN model can be reformulated using a chaining of Blocks, which reduces the sequential simulation complexity (as each Block is simulated in $O(1)$). For a given ARP of length T_R, we observed that a neuron spikes at most once over simulation length T_R (Figure

4.1c). Thus, a simulation length T can be simulated using $N = \frac{T}{T_R}$ Blocks, each of length T_R. [3] Next, we outline how to simulate ALIF dynamics across Blocks to emulate the dynamics of the standard ALIF SNN. We introduce new notation to index Block $1 \leq n \leq N$ using subscript n (*e.g.* input current to neuron i simulated in Block n is expressed as $I_{i,n}[t]$) with time steps indexed between $[1, T_R]$ (as opposed to $[1, T]$) within a Block.

Input current and the ARP of a Block The input current in the standard model (Equation 4.3 and 4.4) is modified for the Block model (Proposition 3). The feedforward and recurrent current to Block $n+1$ are derived from the presynaptic and postsynaptic spikes from Block $n+1$ and Block n respectively. In addition, the ARP is enforced by applying a mask derived from the latent timing of spikes $z_{i,n}[t]$ (Equation 4.8).

Proposition 3. *The input current $I_{i,n+1}[t]$ of neuron i simulated in Block $n+1$ (of length T_R) is defined as follows, and enforces an absolute refractory period of length T_R and a monosynaptic transmission latency of $D = T_R$.*

$$I_{i,n+1}[t] = \Big(b_i + \underbrace{\sum_{j=1}^{N^{in}} W_{ij} S_{j,n+1}[t]}_{\textit{Feedforward current}} + \underbrace{\sum_{j=1}^{N^{out}} W_{ij}^{rec} S_{j,n}[t]}_{\textit{Recurrent current}}\Big) \underbrace{\mathbb{1}_{z_{i,n}[t] \geq \max_t S_{i,n}[t]}}_{\textit{ARP mask}}$$

Evolving membrane potentials between Blocks Two cases are distinguished to correctly emulate the evolution of the membrane potentials between Blocks (Proposition 4): 1) if neuron i did not spike in Block n (*i.e.* $\max_t S_{i,n}[t] = 0$), then its initial membrane potential $V_{i,n+1}[0]$ in Block $n+1$ is set to its final membrane potential $V_{i,n}[T_R]$ in Block n. Otherwise 2), the initial membrane potential is set to zero to emulate a spike reset (and no state needs to be transferred between Blocks as the neuron is in a refractory state).

Proposition 4. *The initial membrane potential $V_{i,n+1}[0]$ of neuron i simulated in Block $n+1$ (of length T_R) is equal to the last membrane potential in Block n if no spike occurred and zero otherwise.*

$$V_{i,n+1}[0] = \begin{cases} V_{i,n}[T_R] & \textit{if } \max_t S_{i,n}[t] = 0 \\ 0 & \textit{otherwise} \end{cases}$$

[3] With appropriate zero padding to the simulation length if it is not divisible by T_R.

Evolving adaptive firing thresholds between Blocks The adaptive firing threshold $\theta_{i,n+1}[t]$ of neuron i in Block $n+1$ is derived from the initial adaptive parameter $a_{i,n+1}[0]$ (Proposition 5). Two cases are distinguished for deriving this parameter: 1) if the neuron did not spike during the previous Block n, this parameter is set to its last value, $a_{i,n}[T_R]$, in the prior Block; otherwise 2), the effect of the spike needs to be taken into account, for which the initial adaptive parameter is expressed as $p_i^m(a_s + p_i^{-1})$. Here, m is the number of time steps remaining in Block n following the spike and a_s is the adaptive parameter value at the time of spiking.

Proposition 5. *The adaptive firing threshold $\theta_{i,n+1}[t]$ of neuron i simulated in Block $n+1$ (of length T_R) is constructed from the initial adaptive parameter $a_{i,n+1}[0]$, which is equal to its last value in the previous Block if no spike occurred, and otherwise equal to an expression which accounts for the effect of the spike on the adaptive threshold.*

$$\theta_{i,n+1}[t] = 1 + d_i p_i^t a_{i,n+1}[0]$$

$$a_{i,n+1}[0] = \begin{cases} a_{i,n}[T_R] & \text{if } \max_t S_{i,n}[t] = 0 \\ p_i^m(a_s + p_i^{-1}) & \text{otherwise} \end{cases}$$

$$m = \sum_k^{T_R} \mathbb{1} z_{i,n}[k] > 1, \quad a_s = \sum_k^{T_R} a_{i,n}[k] S_{i,n}[k]$$

4.3.3 Theoretical speedup for simulating ALIF SNNs using Blocks

Method	Computational Complexity	Sequential Operations
Standard	$O(N^{\text{in}} N^{\text{out}} T)$	$O(T)$
Blocks	$O(N^{\text{in}} N^{\text{out}} T_R^2 N)$	$O(T/T_R)$

Tab. 4.1: Computational and sequential complexity of simulating a layer of N^{out} neurons with N^{in} input neurons for T time steps using the standard method and our method (with N Blocks each of length T_R).

Simulating an ALIF SNN using our Blocks, rather than the conventional approach, requires fewer sequential operations (Table 4.1). Although the computational complexity of our Blocks approach is larger than the standard approach, the number of sequential operations is less. If we assume that the sequential steps in both methods are executed in an equal amount of time (as all the non-sequential operations can be run in parallel on a GPU), we obtain a theoretical speedup equal to the length of the ARP $\frac{T}{N} = T_R$.

4.4 Experiments

We evaluated the training speedup of our accelerated ALIF SNN relative to the standard ALIF SNN and explored the capacity of our model to be fitted using different spiking classification datasets and real electrophysiological recordings of cortical neurons. Implementation details can be found in the Supplementary material and the code at `https://github.com/webstorms/Blocks`.

4.4.1 Training speedup scales with an increasing ARP

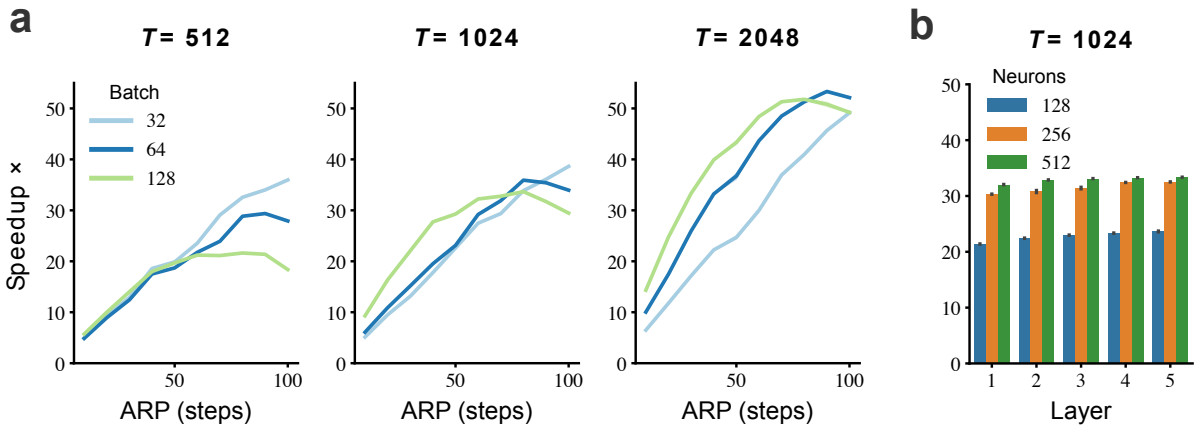

Fig. 4.3: Training speedup of our model. a. Training speedup of our accelerated ALIF model compared to the standard ALIF model for different simulation lengths T, ARP time steps and batch sizes. **b.** Training speedup over different number of layers and hidden units (with an ARP of $T_R = 40$ time steps, $T = 1024$ time steps and batch size $= 64$; bars plot the mean and standard error over ten runs). Assuming DT$= 0.1$ms, then 10 time steps $= 1$ms.

To determine how much faster our model is, we benchmarked the required training duration (forward and backward pass) of our accelerated ALIF SNN to the standard ALIF SNN for varying ARP and simulation lengths using a synthetic spike dataset, with the task of minimizing the number of final-layer output spikes (see Supplementary material). We found the training speedup of our model to increase for a growing ARP (Figure 4.3a). This speedup was more pronounced for longer simulation durations ($53\times$ speedup for $T = 2048$) than shorter simulation durations ($36\times$ speedup for $T = 512$). These speedups were also present when just running the forward pass (*i.e.* inference using the network; see Supplementary material). Furthermore, we found the speedup of our model to be robust over varying numbers of neurons and layers (Figure 4.3b). Lastly, we also found our method to perform the forward pass more than an order of magnitude faster than other publicly available SNN implementations [375, 376] (see Supplementary material).

4.4.2 Accelerated training on spiking classification tasks

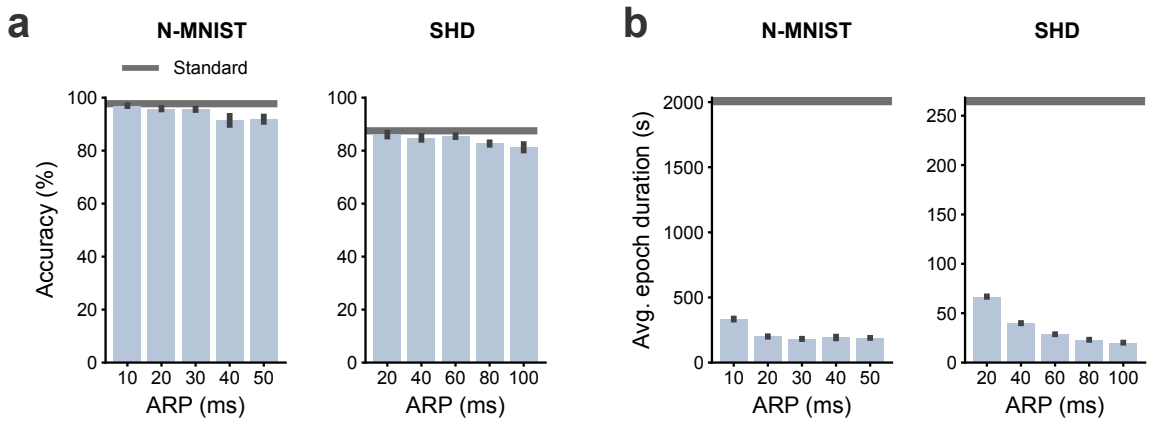

Fig. 4.4: Performance of our model on spiking datasets. a. Classification accuracy of our model and the standard ALIF SNN on the N-MNIST and SHD datasets over different (non-biological) ARPs (N-MNIST 1ms=1 time step; SHD 2ms=1 time step). **b.** Training durations of our model and the standard ALIF SNN on the N-MNIST and SHD datasets. Horizontal gray lines plot the standard model's performance (using no ARP) and bars plot the mean and standard error over three runs.

To establish whether our model can learn using backprop with surrogate gradients, and perform on a par with the standard model, we trained our accelerated ALIF SNN and the standard ALIF SNN on the Neuromophic-MNIST (N-MNIST) (using DT= 1ms) [377] and Spiking Heidelberg Digits (SHD) (using DT= 2ms) [378] spiking classification datasets (both commonly used for benchmarking SNN performance [195, 197, 362, 131, 132]). The task of the N-MNIST dataset is to classify spike representations of handwritten digits, and the task of the SHD dataset is to classify spike representations of spoken digits. In all experiments we employed identical model architectures consisting of two hidden layers (of 256 neurons each) and an additional integrator readout layer, with predictions taken from the readout neurons with maximal summated membrane potential over time (as commonly done [378, 133, 362, 131, 132]; see Supplementary material). We adopted the multi-Gaussian surrogate-gradient function [362] to overcome the non-differentiable discontinuity of the spike function (although we found that other surrogate-gradient functions also work, see Supplementary material). Furthermore (as suggested by [133] for conventional SNNs), we only permitted surrogate gradients to flow through the non-recurrent connectivity (which significantly improved performance; see Supplementary material).

We trained our model with different non-biological ARPs on each dataset and contrasted the accuracy and training duration to the standard model (with no ARP). We found a favourable trade-off between classification accuracy and training time for our model. Compared to the

standard ALIF SNN model, our model achieved a very similar, albeit slightly lower, accuracy on the N-MNIST (96.97% for ARP$=10$ms vs standard 97.71%) and SHD dataset (86.07% for ARP$=20$ms vs standard 87.48%), with the accuracy declining only for the largest ARPs tested (Figure 4.4a). However, our model only required a fraction of the training time, with an average training epoch of 181s and 20s for the N-MNIST and SHD datasets, respectively, when using the largest tested ARP, compared to the respective standard model training times of 2034s and 268s (Figure 4.4b).

4.4.3 Quickly fitting real neural recordings on sub-millisecond timescales

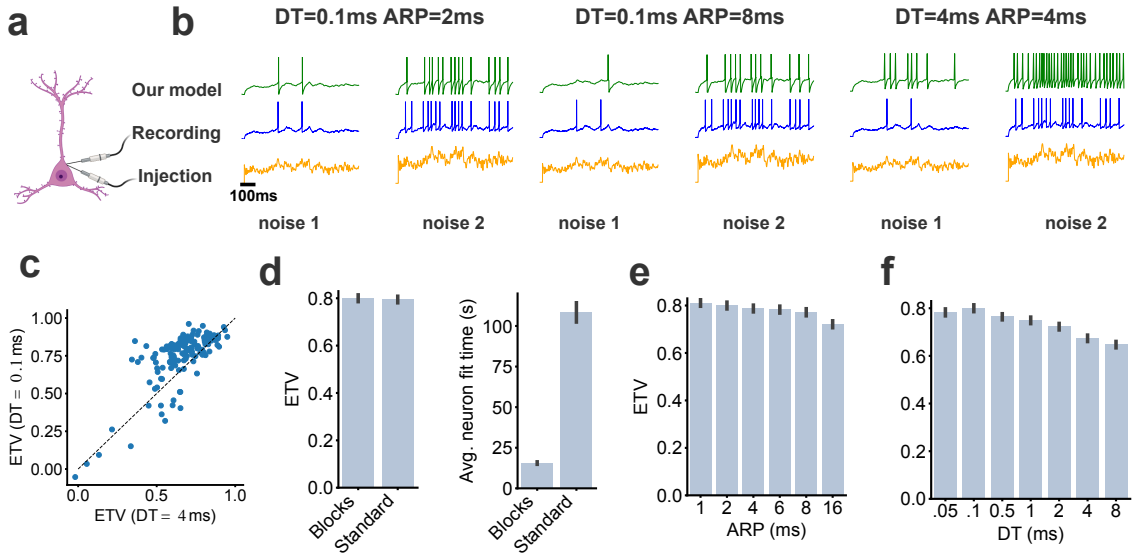

Fig. 4.5: Fitting cortical electrophysiological recordings. **a.** An illustration of a cortical neuron in mouse V1 being recorded whilst stimulated with a noisy current injection. **b.** For held-out data not used for fitting, an example current injection (bottom) and recorded membrane potential (middle) with corresponding fitted model predictions (top). **c.** Comparison of neuron fit accuracy of our model for DT$=0.1$ms (y-axis) against DT$=4$ms (x-axis). Explained temporal variance (ETV) measures the goodness-of-fit (one is a perfect fit and zero is a chance-level fit). **d.** Comparison of the fit accuracy (left) and duration (right) of our model and the standard model (both using DT$=0.1$ms and ARP$=2$ms). **e.** Our model's fit accuracy for increasing ARP (with DT$=0.1$ms). **f.** Our model's fit accuracy for increasing DT (with ARP$=\max(2, \text{DT})$ms). **d.** to **f.** plots the median and standard error over neurons, except that **d.** (right) plots the mean fit time.

We explored the ability of our model to fit *in vitro* electrophysiological recordings from 146 inhibitory and excitatory neurons in mouse primary visual cortex (V1) (provided by the Allen Institute [379, 380]). In these recordings, a variable input current was repeatedly injected at

various amplitudes into each real neuron, and the resulting alteration in the membrane potential was recorded (Figure 4.5a). For each neuron, we used half of the recordings for fitting and the other half for testing, and report all the qualitative and quantitative results on the held-out test dataset. Fitting was achieved by minimising the van Rossum distance between the model and real spike trains [381]. To quantitatively compare the fits of our model using different DTs and ARPs, we employed the explained temporal variance (ETV) measure (used on the Allen Institute website; see Supplementary material). This metric quantifies how well the temporal dynamics of the neurons are captured by the model and takes into account trial-to-trial variability due to intrinsic neural noise. It has a value of one for a perfect fit and a value of zero if the model predicts at chance.

We found that our accelerated ALIF model captured the spike timing of the neurons. Pertinent to the speed-accuracy trade-off, this fit strongly depended on the chosen DT, and less so on the chosen ARP. Qualitatively, using a DT$= 0.1$ms and an ARP$= 2$ms, we found that our model captured the spike timings of the neurons for current injections of varying amplitude. This still seemed to be the case when we used a larger ARP$= 8$ms, but the model was worse at capturing spike timings when we used a larger DT$= 4$ms (with ARP$= 4$ms; Figure 4.5b; see Supplementary material for zoomed-in neural traces).

Quantitatively comparing the neuron fits one by one, we found that nearly all of the neuron fits were better when using a DT$= 0.1$ms (ETV$= 0.80$) than a DT$= 4$ms (ETV$= 0.66$; Figure 4.5c; using an ARP$= 4$ms). We examined how accurate and fast our fits were compared to the standard model using an ARP$= 2$ms and a DT$= 0.1$ms. Both models achieved a similar ETV of ~ 0.8, yet our model only required 15.5s on average per neuron fit time compared to the 108.4s of the standard model (Figure 4.5d). We further investigated whether a larger ARP could still reasonably fit the data for DT$= 0.1$ms (to benefit from a faster fit). Consistent with our qualitative observations, we found a less marked reduction in fit accuracy when using larger ARPs (Figure 4.5e) compared to the drop in performance when using larger DTs (Figure 4.5f; with an ARP of max$(2, \text{DT})$ms).

4.5 Discussion

ALIF neurons are a popular model for studying the brain *in silico*. Training these neurons is, however, slow due to their sequential nature [198, 382]. We overcome this by algorithmically reinterpreting the ALIF model. We found that our method permits faster simulation of SNNs, which will likely play an important role in future research into neural dynamics through

simulation [121, 353]. We also confirmed the validity of this approach for modelling neural data. Firstly - of interest to computational neuroscientists - we fitted a multilayered network of neurons using two common spike classification datasets. We found that our model achieved a similar classification accuracy to that of the standard model, even if we increased the ARP to large non-physiological values, which drastically reduced the training duration required for both datasets. Secondly - of interest to computational and experimental neuroscientists - we explored the applicability of our method to fit real electrophysiological recordings of neurons in mouse V1 using sub-millisecond DTs.

We found that our method accurately captured the spike timing of real neurons, fitting their activity patterns in a fraction of the time required by the standard model (although we note that other, more recent spiking-models, might improve fit accuracies [383]). This is particularly important as datasets become larger with advances in recording techniques, requiring faster fitting solutions to characterise computational properties of neurons. As an example of the potential insights provided by our model, we found the fitted V1 neurons to have a heterogenous membrane time constant distribution, suggesting that V1 processes visual information on multiple timescales [131] (see Supplementary material).

Our work will likely also be of interest to neuromorphic engineers and researchers, developing energy-efficient hardware to emulate SNNs [356]. These systems - like real neurons - run in continuous time and thus require training on extremely fine DTs off-chip [384, 385]. Our method could help to accelerate off-chip training times. Furthermore, the ARP hyperparameter in our model can limit high spike rates and thus reduce energy consumption on neuromorphic systems, as this consumption scales approximately proportionally with the number of emitted spikes [386].

A limitation of our method is that the training speedup scales sublinearly - as opposed to linearly - with an increasing ARP simulation length (see Section 4.3 and 4.4.1). This is likely due to GPU overheads and employed cudnn routines [387], which further improvements to our code implementation could overcome. An additional limitation is the requirement to define the ARP hyperparameter, whose value influences the training speed and test accuracy in our method and relates to biological realism. However, we found a beneficial trade-off - by using large non-biological ARPs on the artificial spiking classification dataset and small physiological ARPs for the neural fits (although we found larger values to also perform reasonably well) the model achieved comparable accuracy to the standard ALIF model in both cases, while also having greatly increased speed of training and simulation. In particular, the capacity of our model to accurately and quickly fit the data from real neurons is crucial in terms of the biological applicability of this approach.

Supplementary

4.5.1 Theoretical proofs

Proofs for all the propositions in the paper, outlining the mathematical equivalence of our Block model to the standard ALIF SNN.

Proposition 6. *Membrane potentials without spike reset are computed as a convolution* $\tilde{V}_i[t] = (I_i * \tilde{\beta}_i)[t]$ *between input current* $I_i[t]$ *and kernel* $\tilde{\beta}_i[t] = (1-\beta_i)\beta_i^t$ *with the initial membrane potential encoded as* $I_i[0] = \frac{V_i[0]}{1-\beta_i}$.

Proof. We proceed our proof in two steps. In step 1, we unroll the discretized LIF difference equation (without reset) in time and in step 2, we show how this is equivalent to the proposed convolution.

Step 1, we prove the equivalence between the following equations

$$\tilde{V}_i[t] = \beta_i \tilde{V}_i[t-1] + (1-\beta_i) I_i[t] \tag{4.10}$$

$$\tilde{V}_i[t] = \beta_i^t \tilde{V}_i[0] + (1-\beta_i) \sum_{j=1}^{t} \beta_i^{t-j} I_i[j] \tag{4.11}$$

We proceed by induction. For $t = 1$ in Equation 2 we obtain

$$\begin{aligned} \tilde{V}_i[1] &= \beta_i^1 \tilde{V}_i[0] + (1-\beta_i) \sum_{j=1}^{1} \beta_i^{1-j} I_i[j] \\ &= \beta_i^1 \tilde{V}_i[0] + (1-\beta_i) I_i[1] \end{aligned} \tag{4.12}$$

Hence the relation holds true for the base case $t = 1$. Assume the relation holds true for $t = k \geq 1$, then for $t = k+1$ we derive

$$\begin{aligned} \tilde{V}_i[k+1] &= \beta_i \tilde{V}_i[k] + (1-\beta_i) I_i[k+1] \\ &= \beta_i \Big(\beta_i^k \tilde{V}_i[0] + (1-\beta_i) \sum_{j=1}^{k} \beta_i^{k-j} I_i[j] \Big) + (1-\beta_i) I_i[k+1] \\ &= \beta_i^{k+1} \tilde{V}_i[0] + (1-\beta_i) \sum_{j=1}^{k} \beta_i^{(k+1)-j} I_i[j] + (1-\beta_i) I_i[k+1] \\ &= \beta_i^{k+1} \tilde{V}_i[0] + (1-\beta_i) \sum_{j=1}^{k+1} \beta_i^{(k+1)-j} I_i[j] \end{aligned} \tag{4.13}$$

This implies equivalence between Equations 1 and 2 for $t = k + 1$ assuming equivalence between Equations 1 and 2 holds true for $t = k$. By the principle of induction, equivalence is established given that both the base case and inductive step hold true.

Step 2, as per the proposition, we have

$$\begin{aligned}
\tilde{V}_i[t] &= (I_i * \tilde{\beta}_i)[t] \\
&= \sum_{j=0}^{t} \tilde{\beta}_i[t-j] I_i[j] \\
&= (1-\beta_i) \sum_{j=0}^{t} \beta_i^{t-j} I_i[j] \\
&= (1-\beta_i)\beta_i^t I_i[0] + (1-\beta_i) \sum_{j=1}^{t} \beta_i^{t-j} I_i[j] \\
&= \beta_i^t \tilde{V}_i[0] + (1-\beta_i) \sum_{j=1}^{t} \beta_i^{t-j} I_i[j]
\end{aligned} \tag{4.14}$$

This is identical to Equation 4.11 and (by step 1) identical to Equation 4.10. Thus, the proposed convolution in the Proposition computes the membrane potentials without reset. □

Proposition 7. *Function $\phi(\tilde{S}_i)[t] = \sum_{k=1}^{t} \tilde{S}_i[k](t-k+1)$ acting on $\tilde{S}_i \in \{0,1\}^T$ contains at most one element equal to one $\phi(\tilde{S}_i)[t] = 1$ for smallest t satisfying $\tilde{S}_i[t] = 1$ (if such t exists).*

Proof. Firstly, if $\tilde{S}_i^{(l)}[t] = 0$ for all $t \in [1, T]$ then $\phi(\tilde{S}_i^{(l)})[t] = 0$ for all $t \in [1, T]$ (follows from substitution). Secondly, if $\tilde{S}_i^{(l)}[t_1] = 1$ for smallest $t_1 \in [1, T]$ then $\phi(\tilde{S}_i^{(l)})[t_1] = 1$ (follows from substitution) and there can exist no $t_2 > t_1$ such that $\phi(\tilde{S}_i^{(l)})[t_2] = 1$ as

$$\begin{aligned}
\phi(\tilde{S}_i^{(l)})[t+1] &= \sum_{k=1}^{t+1} \tilde{S}_i^{(l)}[k]((t+1)-k+1) \\
&= \sum_{k=1}^{t} \tilde{S}_i^{(l)}[k]((t+1)-k+1) + \tilde{S}_i^{(l)}[t+1] \\
&= \sum_{k=1}^{t} \tilde{S}_i^{(l)}[k](t-k+1) + \sum_{k=1}^{t} \tilde{S}_i^{(l)}[k] + \tilde{S}_i^{(l)}[t+1] \\
&= \phi(\tilde{S}_i^{(l)})[t] + \sum_{k=1}^{t+1} \tilde{S}_i^{(l)}[k]
\end{aligned} \tag{4.15}$$

Thus $\phi(\tilde{S}_i^{(l)})[t_2] > \phi(\tilde{S}_i^{(l)})[t_1]$ for all $t_2 > t_1$ as $\sum_{k=1}^{t_2} \tilde{S}_i^{(l)}[k] \geq \sum_{k=1}^{t_1} \tilde{S}_i^{(l)}[k] = 1 > 0$. □

Proposition 8. *The input current $I_{i,n+1}[t]$ of neuron i simulated in Block $n+1$ (of length T_R) is defined as follows, and enforces an absolute refractory period of length T_R and a monosynaptic transmission latency of $D = T_R$.*

$$I_{i,n+1}[t] = \Big(b_i + \underbrace{\sum_{j=1}^{N^{in}} W_{ij} S_{j,n+1}[t]}_{\textit{Feedforward current}} + \underbrace{\sum_{j=1}^{N^{out}} W_{ij}^{rec} S_{j,n}[t]}_{\textit{Recurrent current}}\Big) \underbrace{\mathbb{1}_{z_{i,n}[t] \geq \max_t S_{i,n}[t]}}_{\textit{ARP mask}}$$

Proof. We proceed by showing that 1. the transmission latency is the same as in the standard model and 2. the ARP mask enforces an ARP of identical length to the standard model.

1. Identical transmission latency: The relation between the time step $1 \leq t \leq T$ and the time step $1 \leq t_b \leq T_R$ in Block $n \geq 1$ can be expressed as:

$$t = (n-1)T_R + t_b \tag{4.16}$$

As we assumed the transmission latency $D = T_R$ to be equal to the ARP length, we have

$$t - T_R = ((n-1)-1)T_R + t_b \tag{4.17}$$

and hence the input current $I_{i,n}[t]$ to Block n at time t (*i.e.* time step t_b in the Block) is computed from the output spikes from the prior Block $n-1$ at the same Block time step t_b.

2. Identical ARP length: Case one, if neuron i emitted no spikes during Block n, then neuron i should receive input current at every time step in Block $n+1$, which the ARP mask permits (as $z_{i,n}[t] \geq \max_t S_{i,n}[t] = 0$ for all t). Case two, if neuron i spiked during Block n, then the ARP mask should appropriately mask out the input current to Block $n+1$ (to enforce an ARP of length T_R). The number of elements in $z_{i,n}$ which are smaller than one and equal to or larger than one is equal to T_R (can be shown by proof by contradiction):

$$\underbrace{\sum \mathbb{1}_{z_{i,n}[t] < 1}}_{\text{Number of elements masked in Block } n+1} + \underbrace{\sum \mathbb{1}_{z_{i,n}[t] \geq 1}}_{\text{Number of time steps from the spike onwards in Block } n} = T_R \tag{4.18}$$

The construction of the Block ensures that at most one spike is emitted within a Block (where $z_{i,n}[t] = 1$) and no further spikes are emitted thereafter (where $z_{i,n}[t] > 1$). The ARP mask ensures that no spikes are emitted in the next Block for the time steps where $z_{i,n}[t] < 1$ (as it only permits input current to flow into the Block for time steps where $z_{i,n}[t] \geq 1$). Thus, the mask ensures enforces the correct ARP length of T_R steps. □

Proposition 9. *The initial membrane potential $V_{i,n+1}[0]$ of neuron i simulated in Block $n+1$ (of length T_R) is equal to the last membrane potential in Block n if no spike occurred and zero otherwise.*

$$V_{i,n+1}[0] = \begin{cases} V_{i,n}[T_R], & \textit{if } \max_t S_{i,n}[t] = 0 \\ 0, & \textit{otherwise} \end{cases}$$

Proof. Two cases are distinguished for correctly evolving the membrane potential of a neuron over time. Case one, if no spike occurred in neuron i in Block n (*i.e.* $\max_t S_{i,n}[t] = 0$), then the initial membrane potential of neuron i in Block $n+1$ is equal to the final membrane potential value of neuron i in Block n. Otherwise, case two, the initial membrane potential is set to zero (as no state needs to be transferred as the neuron is in an absolute refractory state). □

Proposition 10. *The adaptive firing threshold $\theta_{i,n+1}[t]$ of neuron i simulated in Block $n+1$ (of length T_R) is constructed from the initial adaptive parameter $a_{i,n+1}[0]$, which is equal to its last value in the previous Block if no spike occurred, and otherwise equal to an expression which accounts for the effect of the spike on the adaptive threshold.*

$$\theta_{i,n+1}[t] = 1 + d_i p_i^t a_{i,n+1}[0]$$

$$a_{i,n+1}[0] = \begin{cases} a_{i,n}[T_R], & \textit{if } \max_t S_{i,n}[t] = 0 \\ p_i^m (a_s + p_i^{-1}) & \textit{otherwise} \end{cases}$$

$$m = \sum_k^{T_R} \mathbb{1}_{z_{i,n}[k] > 1}, \quad a_s = \sum_k^{T_R} a_{i,n}[k] S_{i,n}[k]$$

Proof. We proceed our proof in two steps. In step 1, we show how the dynamic firing threshold of neuron i in Block $n+1$ can be computed using initial adaptive parameter $a_{i,n+1}[0]$; and in step 2, we show how this initial adaptive parameter is derived.

Step 1, we prove the equivalence between the Block adaptive threshold (Equation 4.19) and the standard adaptive firing threshold (Equations 4.20 and 4.21).

$$\theta_{i,n+1}[t] = 1 + d_i p_i^t a_{i,n+1}[0] \tag{4.19}$$

$$\theta_{i,n+1}[t] = 1 + d_i a_{i,n+1}[t] \tag{4.20}$$

$$a_{i,n+1}[t] = p_i a_{i,n+1}[t-1] + S_{i,n+1}[t-1] \tag{4.21}$$

The spike term in Equation 4.21 can be dropped, as only the spike occurrence in Block n (and not Block $n+1$) can affect the firing threshold in Block $n+1$ (due to the single spike constraint). Thus, we rewrite Equation 4.20 as:

$$\begin{aligned}\theta_{i,n+1}[t] &= 1 + d_i a_{i,n+1}[t] \\ &= 1 + d_i p_i a_{i,n+1}[t-1] \\ &= 1 + d_i p_i^2 a_{i,n+1}[t-2] \\ &\dots \\ &= 1 + d_i p_i^t a_{i,n+1}[0]\end{aligned}$$

Step 2, to derive the initial adaptive parameter $a_{i,n+1}[0]$, we distinguish two cases. Case one, neuron i did not spike in Block n, in which case the initial adaptive parameter is set to the final adaptive parameter $a_{i,n}[T_R]$ in Block n. Case two, neuron i did spike in Block n, and we need to account for the affect of this on the firing threshold in Block $n+1$. If the spike occurred at Block time step $1 \leq T_R - m \leq T_R$ for $0 \leq m < T_R$ we have

$$\begin{aligned}a_{i,n}[T_R] &= p_i a_{i,n}[T_R - 1] \\ &= p_i^2 a_{i,n}[T_R - 2] \\ &\dots \\ &= p_i^{m-1} a_{i,n}[T_R - (m-1)] \\ &= p_i^{m-1}\Big(p_i a_{i,n}[T_R - m] + 1\Big) \\ &= p_i^{m}\Big(a_{i,n}[T_R - m] + p_i^{-1}\Big) \\ &= p_i^{m}\Big(a_s + p_i^{-1}\Big) \end{aligned} \tag{4.22}$$

with $a_s = \sum_k^{T_R} a_{i,n}[k] S_{i,n}[k] = a_{i,n}[T_R - m]$ (as $S_{i,n}[k]$ is one for $k = T_R - m$ and zero otherwise). Lastly, $m = \sum_k^{T_R} \mathbb{1}_{z_{i,n}[k] > 1}$, as $z_{i,n}[k] > 1$ at ever Block time step $k > T_R - m$ (see Proposition 7 proof showing $z_{i,n}$ to be a strictly increasing sequence if a spike occurred in Block n) and thus $\sum_k^{T_R} \mathbb{1}_{z_{i,n}[k] > 1} = \sum_{k=T_R-m+1}^{T_R} 1 = T_R - (T_R - m) = m$. □

4.5.2 Experimental details

All models were implemented using PyTorch [388] (although nothing prohibits the use of other auto differentiation frameworks [388, 389]). The speedup benchmarks and neural fits were

done on an NVIDIA GeForce RTX 3090, and the training on the spiking classification datasets was done on an NVIDIA GeForce GTX 1080 Ti. Following details apply to both our Block model and standard SNN model which we used as a control.

Speed benchmarks

Synthetic spike dataset We generated binary input spike tensors of shape $B \times N \times T$ (B being the batch size, N the number of input neurons and T the number of time steps). For every batch dimension b a firing rate $r_b \sim \mathbf{U}(u_{\text{min}}, u_{\text{max}})$ was uniformly sampled (with $u_{\text{min}} = 0$Hz and $u_{\text{max}} = 200$Hz – Assuming 1 time step $= 1$ms), from which a random binary spike matrix of shape $N \times T$ was constructed as a homogenous Poisson process, such that every input neuron in this matrix had a firing rate of r_bHz.

Supervised learning

Datasets We tested our model (and control) on two common spike classification datasets, the Neuromophic-MNIST (N-MNIST) [377] and Spiking Heidelberg Digits (SHD) [378] dataset (both released under the Creative Commons Attribution 4.0 International License). The N-MNIST dataset is the classical MNIST dataset mapped onto a spike code using a neuromorphic vision sensor and the SHD dataset comprises spoken digit waveforms converted into spikes using a model of the auditory bushy neurons in the cochlear nucleus.

Weight initialisation All network connectivity weights were sampled from a uniform distribution $\mathbf{U}(-\sqrt{N^{-1}}, \sqrt{N^{-1}})$ with N number of afferent connections. All biases were initialised as 0. The hidden neurons were initialised with a membrane time constant of 20ms (*i.e.* $\beta_i^{(l)} = \exp(\frac{-DT}{20})$), an adaptive time constant of 150ms (*i.e.* $p_i^{(l)} = \exp(\frac{-DT}{150})$) and adaptive parameter of $d_i^{(l)} = 1.8$. The readout neurons were initialised with a membrane time constant of 20ms.

Clamping time constants To enforce correct neuron dynamics, we clamped the values of $\beta_i^{(l)}$ into the range $[0.01, 0.99]$ and the values of $p_i^{(l)}$ into the range $[0.0, 0.999]$

$$\beta_i^{(l)} = \begin{cases} 0.99, & \text{if } \beta_i^{(l)} > 0.99 \\ 0.01, & \text{if } \beta_i^{(l)} < 0.01 \end{cases} \tag{4.23}$$

$$p_i^{(l)} = \begin{cases} 0.999, & \text{if } p_i^{(l)} > 0.999 \\ 0.0, & \text{if } p_i^{(l)} < 0.0 \end{cases} \tag{4.24}$$

Readout neurons Every network had an output layer of readout neurons (containing the same number of neurons as the number of classes within the dataset trained on), where we removed the spike and reset mechanism (as done in [133]). The output of the readout neuron c in response to input sample b was taken to be the summated membrane potential over time $o_{b,c} = \sum_t V_{b,c}^L[t]$ (L being the readout layer).

Supervised training loss We trained all networks to minimise a cross-entropy loss (with B and C being the number of batch samples and dataset classes respectively)

$$\mathcal{L} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{c=1}^{C} y_{bc}\log(\hat{y}_{bc}) \tag{4.25}$$

Variable $y_{bc} \in \{0,1\}^C$ is the one hot target vector and $\hat{y}_{bc}$ are the network prediction probabilities, which were obtained by passing the readout neuron outputs o_{bc} through the softmax function.

$$\hat{y}_{bc} = \frac{\exp(o_{bc})}{\sum_{k=1}^{C}\exp(o_{bk})} \tag{4.26}$$

Surrogate gradient We tested training on the SHD dataset using three different surrogate gradient functions, including the multi-Gaussian [364], fast sigmoid [194] and the boxcar [390, 391] function - all which have shown to perform well in training SNNs.

$$\frac{\partial S_i^{(l)}[t]}{\partial V_i^{(l)}[t]} = 1.15\mathcal{N}(V_i^{(l)}[t] \mid 0, 0.5^2) - 0.15\mathcal{N}(V_i^{(l)}[t] \mid 3, 3^2) \tag{4.27}$$

$$- 0.15\mathcal{N}(V_i^{(l)}[t] \mid -3, 3^2) \quad \text{(multi-Gaussian)} \tag{4.28}$$

$$\frac{\partial S_i^{(l)}[t]}{\partial V_i^{(l)}[t]} = (10|V_i^{(l)}[t]| + 1)^{-2} \quad \text{(fast sigmoid)} \tag{4.29}$$

$$\frac{\partial S_i^{(l)}[t]}{\partial V_i^{(l)}[t]} = \begin{cases} 0.5, & \text{if } |V_i^{(l)}[t] - \theta_i^{(l)}| \leq 0.5 \\ 0, & \text{otherwise} \end{cases} \quad \text{(boxcar)} \tag{4.30}$$

$$\tag{4.31}$$

We chose the multi-Gaussian function as this obtained the best performance (Supplementary Figure 4.6a).

Detaching gradients Detaching gradients from flowing through the spike reset and recurrent connections has been shown to improve classification accuracies when training SNNs [133]. We thus explored 1. allowing gradients to flow through all elements within the computational graph (*i.e.* attached) and 2. detaching the surrogate gradient from all elements within the computational graph, except for the feedforward connections to other neurons (*i.e.* detached). We found that detaching improved test accuracies across all the tried surrogate gradient functions on the SHD dataset (Supplementary Figure 4.6b).

Training procedure We used the Adam optimiser (with default parameters) [289] for all training, starting with an initial learning rate of 0.001, which was decayed by a factor of 10 every time the number of epochs reached a new milestone. Model weights were saved if the training error at the end of each epoch was lowered. The hyperparameters are found in Supplementary Table 4.2.

Additional results Forward (*i.e.* inference using the network) speedup can be found in Supplementary Figure 4.7. Here we used the same simulation setup as the one we used for benchmarking the training speedups. We also compared the forward pass of our model to other publicly available SNN implementations, the Norse library [375] and the Spiking Jelly library [376], and found our method to run considerably faster (Supplementary Table 4.3). We benchmarked a simple one layer SNN network of 100 units on the forward pass, over 1000 time steps with 200 input units (with a batch size of 128). As expected, we found the Norse (0.30s) and SpikingJelly (0.18s) implementations took a similar time to our standard SNN implementation (0.33s) with the SpikingJelly implementation running slightly faster (as they are all governed by the same sequential time complexity). In contrast, our Blocks model (using a 50 steps ARP) ran over an order of magnitude faster (0.016s) than the other implementations.

Neural fits

Dataset preprocessing We fitted the model to *in vitro* whole-cell patch clamp electrophysiological recordings from neurons in layer 4 in mouse primary visual cortex, provided by the Allen institute [379, 380]. These neurons were injected with a varying current at different amplitudes.

For each neuron, the training and test datasets were those stipulated by the Allen Institute (50% for training and 50% for testing). We only fitted and reported our results on neurons which had four repeats for each unique current injection. We processed the input data by 1. removing all the long periods in the recordings during which no current was injected, 2. resampling the data to have a DT$= 0.1$ms, and lastly 3. normalising the data by subtracting the mean and by dividing by the standard deviation of the training dataset. We used the spike times provided by the Allen Institute for fitting.

Weight initialisation The neuron input weight was initialised to a constant value of $\frac{s}{100}$ for scale value s, ensuring that the input weight was correctly rescaled for different simulation resolutions (*e.g.* $s = 1$ corresponds to simulating with DT$= 0.1$ms and $s = 2$ corresponds to simulating with DT$= 0.2$ms). The bias was initialized to 0. The membrane time constant was initialized to 20ms, the adaptive time constant was initialized to 100ms, and the adaptive parameter was initialized to $d = \frac{0.1}{s}$.

Supervised training loss Each ALIF neuron model was optimised to produce the same spike times of the real neuron being fit to. This was achieved by minimising the van Rossum distance D_R [381] between the predicted model spike train x and real neuron spike train y, defined as

$$D_R(\tau_R) = \sqrt{\frac{1}{\tau_R} \int_0^T \left(k * x(t) - k * y(t)\right)^2 dt} \tag{4.32}$$

with exponential kernel $k = H(t) \exp(\frac{-t}{\tau_R})$ and H being the heaviside function. We used $\tau_R = 100$ms for all our reported results.

Training procedure We used the Adam optimiser (with default parameters) [289] for every neuron fitted, with a learning rate of 0.0001. Training was carried out over 200 epochs in full batch mode (*i.e.* estimating gradients from the entire training dataset). Model parameters were saved whenever the training score improved, and training was halted if there was no improvement over the last five epochs.

Explained temporal variance We used the explained temporal variance (ETV) metric to assess the goodness-of-fit of the fitted models to the neural data (used on the Allen Institute website).

The metric has a value of zero when the model predicts at chance and a value of one for a perfect fit to the data, and is defined as:

$$\text{ETV} = \frac{\text{ETV}_{\text{raw}}}{\text{ETV}_{\text{max}}} \tag{4.33}$$

ETV_{raw} measures the pairwise explained variance of the data with the model and ETV_{max} measures the upper limit on how well the model can perform (*i.e.* how much of the neuron's response variability from repeat to repeat can be accounted for).

$$\text{ETV}_{\text{raw}} = \sum_r \frac{\text{var}\left(g * x\right) + \text{var}\left(g * y_r\right) - \text{var}\left(g * x - g * y_r\right)}{\text{var}\left(g * x\right) + \text{var}\left(g * y_r\right)} \tag{4.34}$$

$$\text{ETV}_{\text{max}} = \sum_r \frac{\text{var}\left(g * \bar{y}\right) + \text{var}\left(g * y_r\right) - \text{var}\left(g * \bar{y} - g * y_r\right)}{\text{var}\left(g * \bar{y}\right) + \text{var}\left(g * y_r\right)} \tag{4.35}$$

Here x is the predicted model spike train, y_r is recorded neuron spike train for repeat r, g is a Gaussian kernel (with mean $\mu = 0$ and standard deviation $\sigma = 150$ms) and $\bar{y}[t] = \frac{1}{R}\sum_r^R (g * y_r)[t]$ is the mean recorded (and smoothed) neuron response. Convolving the spike trains with the Gaussian kernel converts them into peristimulus time histograms.

Additional results Zoomed-in plots of the neural traces can be found in Supplementary Figure 4.8, and the membrane time constant distribution of the fitted V1 neurons can be found in Supplementary Figure 4.9.

Tab. 4.2: Dataset and corresponding training parameters.

	N-MNIST	**SHD**
Dataset (train/test)	60k/10k	8156/2264
Input neurons	1156	700
Dataset classes	10	20
Epochs	30	40
Batch size B	64	64
Time steps T	300	600
Time resolution Δt (ms)	1	2
LR decay epoch milestones	N/A	$(15, 30)$

Tab. 4.3: Comparison to other SNN libraries. Forward pass simulation duration of our model compared to other publicly available SNN implementations.

	Duration (s)
SpikingJelly [375]	0.18s
Norse [376]	0.30s
Standard SNN (our implementation)	0.33s
Block SNN (our model)	0.016s

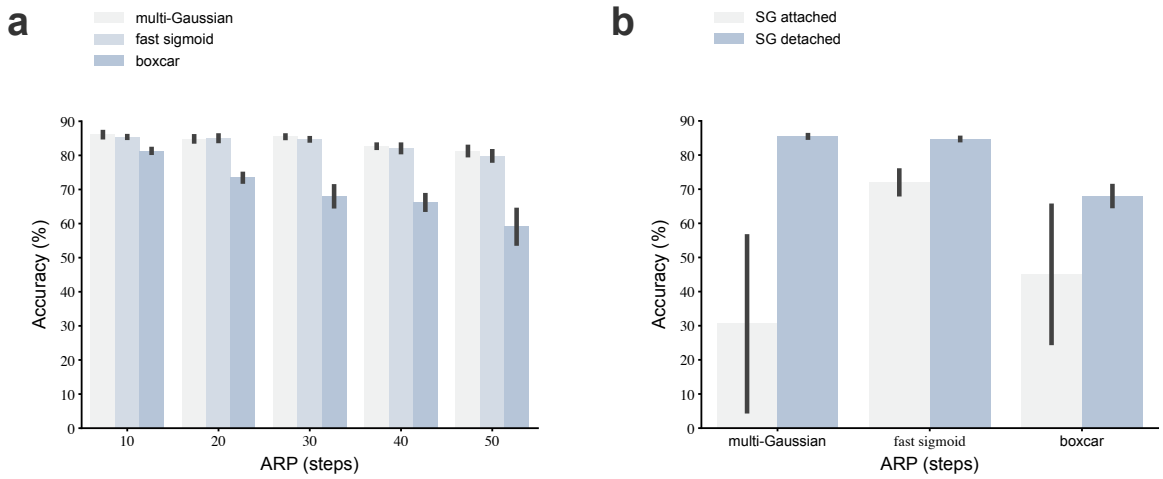

Fig. 4.6: Surrogate gradient search. a. Accuracy on the SHD dataset using our model for different surrogate gradient functions. **b.** Accuracy on the SHD dataset using our model when attaching the surrogate gradient to all elements within the computational graph vs detaching it from everything besides the connections to efferent neurons. Bars plot the mean and standard error over three runs.

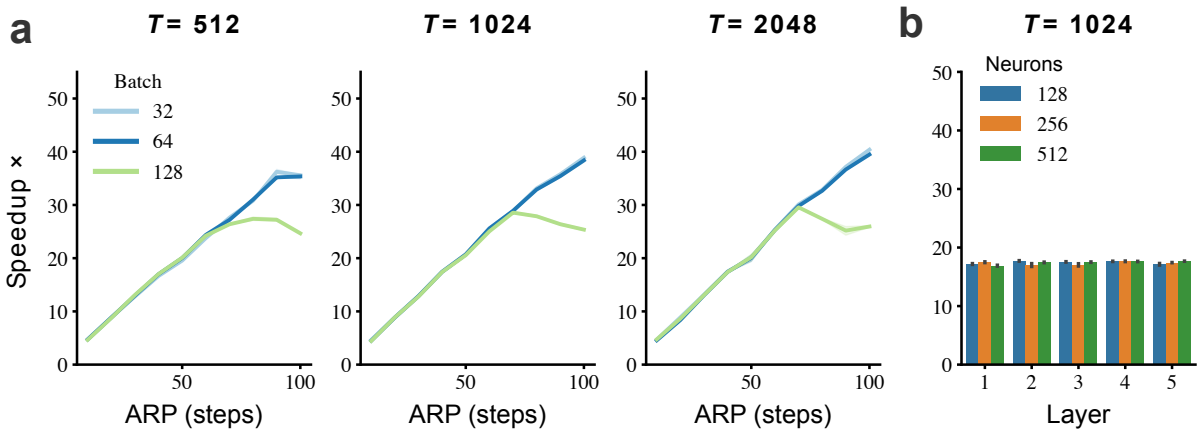

Fig. 4.7: Forward speedup of our model. a. Forward speedup of our accelerated ALIF model compared to the standard ALIF model for different simulation lengths T, ARP time steps and batch sizes. **b.** Forward speedup over different number of layers and hidden units (with ARP time steps$= 40$, $T = 1024$ and batch$= 64$; bars plot the mean and standard error over ten runs). Assuming DT$= 0.1$ms, then 10 time steps $= 1$ms.

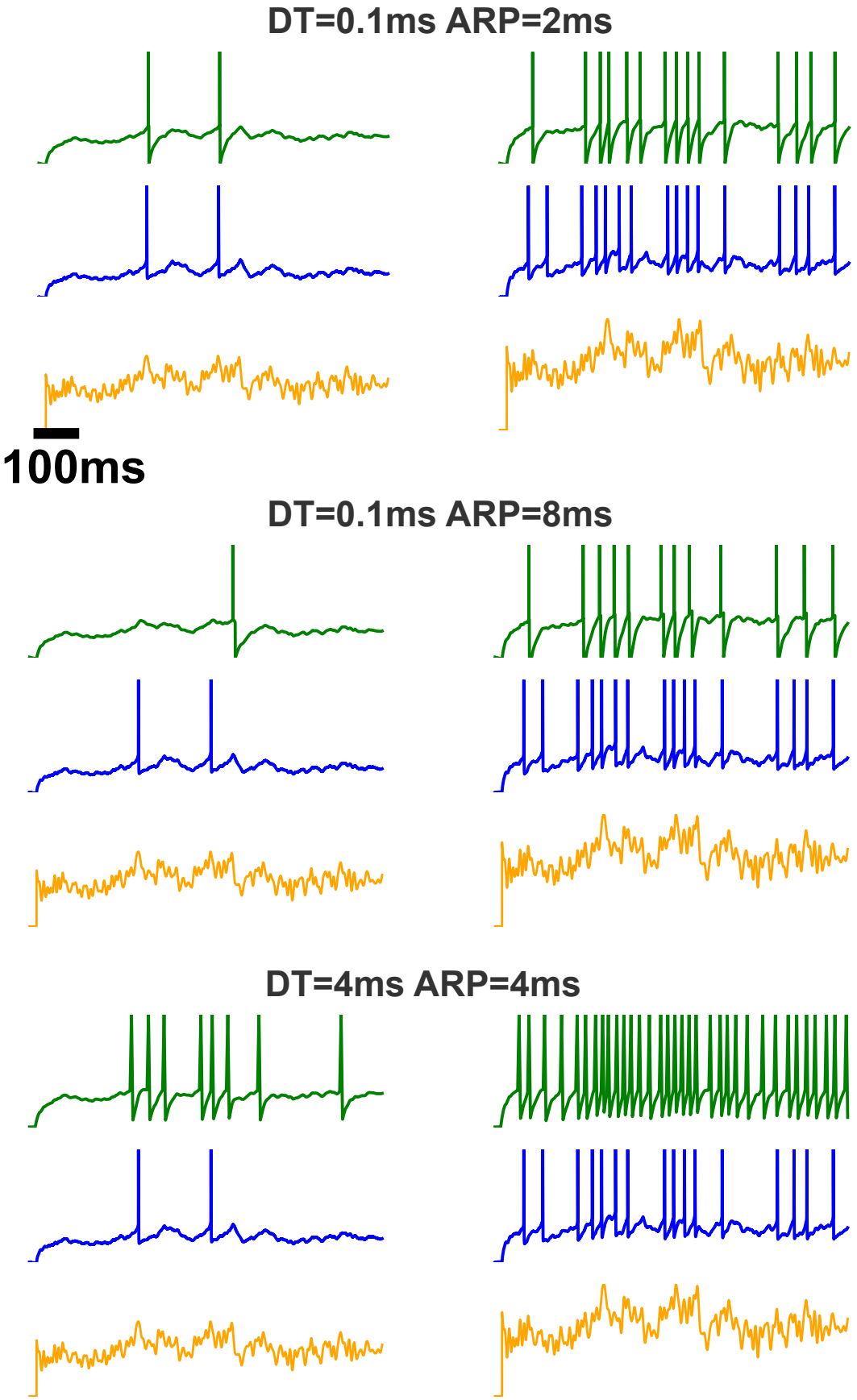

Fig. 4.8: Zoomed-in versions of the neural traces from Figure 5b.

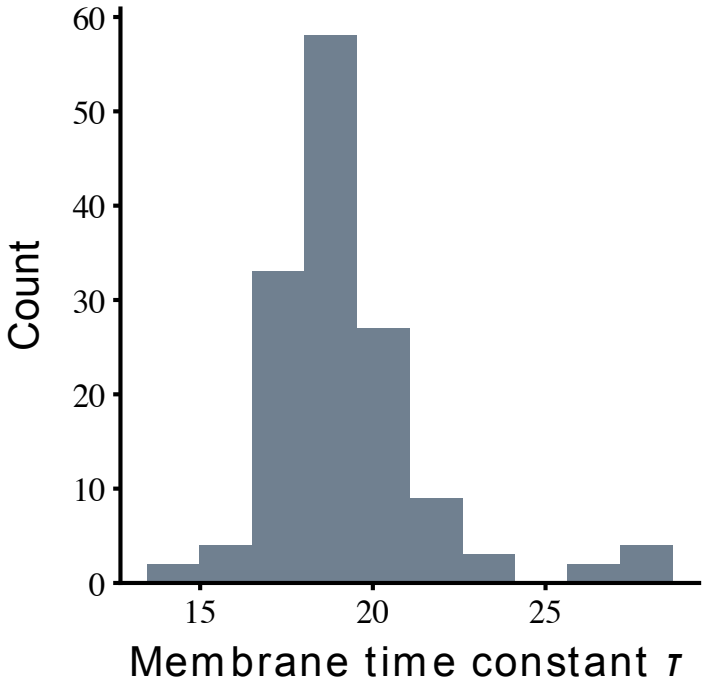

Fig. 4.9: **Membrane time constant distribution of the neurons fitted in Figure 5.**

5 Discussion

5.1 Summary of results

A central aim in visual neuroscience is to understand the operations of the visual system: how is visual information represented in and transformed by visual circuits? A popular modelling approach to investigating the functional principles underlying neural circuits is normative modelling. Here, a model architecture is developed to approximate the neural architecture of the visual system or a neural region therein (*e.g.* the retina or V1). This model architecture is then optimized for a normative goal, which is thought to govern the way the system operates (*e.g.* compressing or predicting incoming sensory stimuli). After optimization, the model's brain-like characteristics are evaluated by contrasting its tuning properties with biological data (*e.g.* the RFs of recorded neurons), or by assessing its neural response predictive accuracy.

Normative investigations have been able to account for certain neural tuning properties found in the visual system [148, 232, 140, 141, 143, 167, 152, 146, 168, 106]. In particular, models based on the normative goal of temporal prediction have been able to account for the spatial and temporal tuning properties of RFs of V1 neurons [106]; visual-cortex-like response dynamics to perceptual motion illusions [161]; and predict V1 responses in monkeys to natural movies on par with other state-of-the-art models [162]. However, these normative temporal prediction models - and normative models of the visual system in general - omit key biological constraints of the brain, namely, the ubiquitous spiking dynamics of neurons and the stereotypical excitatory and inhibitory neural subpopulations of the cortex. In this thesis, I set out to address this gap.

In Chapter 2, I developed a normative model of V1 comprised of recurrently-connected spiking units separated into distinct excitatory and inhibitory subpopulations. This model was trained to predict future sensory stimuli from past spiking activity, with a prediction timescale similar to the transmission latencies from the retina to V1. This model exhibited V1-like spike statistics and accounted for key neural tuning and physiological differences between the excitatory and inhibitory subpopulations in V1.

In Chapter 3, I extended the spiking model from V1 to the retina, using a few modifications and by increasing the metabolic-like regularization. I explored whether the retina is optimized for

119

efficient compression or efficient temporal prediction, by training multiple spiking retinal models using different prediction offsets. I found the retinal models optimized for prediction to better predict the neural responses of retinal ganglion cells from different animal species to natural images and movie stimuli. Furthermore, the predictive spiking retinal model also exhibited retinal-like RF structures, tuning properties and spiking characteristics, like latency coding.

In Chapter 4, I designed a new algorithm to more rapidly train and simulate spiking neural networks on hardware accelerators, such as GPUs. This method reduces the sequential simulation complexity of LIF and ALIF models by algorithmically reinterpreting their computational graphs. In particular, the spiking dynamics are more efficiently simulated during the absolute refractory period of neurons using non-sequential tensor operations. This method achieves a $10 - 50\times$ training speedup over the standard training method, with comparable performance to the standard ALIF implementation on different supervised classification benchmarks and on fitting electrophysiological recordings.

5.2 Project development summary

The V1 spiking model underwent significant development before obtaining units that reproduced all the biological properties of V1 neurons. I started the development of the model by extending the ANN temporal prediction model [106] with a binary activation function (*i.e.* an output of one or zero). Units in this model exhibited V1-like RFs. Next, I increased the biological realism of this model by incorporating an evolving membrane potential. Again, units in this model exhibited V1-like RFs. Thereafter, I made another two modelling additions: imposing Dale's law in the model by hardcoding units to be excitatory or inhibitory, and extending the model from operating on a patch of movie input to operating over the entire spatial extent of a movie frame using convolutions. I had to decay the learning rate of this new model after a certain number of epochs to avoid drastic spikes in the loss function. Once again, units in this model exhibited V1-like RFs. This early-stage model appeared to capture the spiking variability of V1 neurons, however, the log-scale firing rate distribution was slightly bimodal and did not agree with that of V1 neurons.

I used the natural movie training data used by [106]. However, this training data was recorded at 25Hz which posed a problem for accurately simulating neuron dynamics as neurons evolve on shorter timescales than 40ms (the duration of every movie frame). To overcome this issue I created a new natural stimulus dataset recorded at 120Hz (8.33ms duration for every movie frame). This slightly alleviated the issue faced in the firing rate distribution.

I designed various metabolic regularization functions in an attempt to obtain more V1-like spike statistics. This included having separate terms penalizing the synaptic connectivity, firing rate, and membrane potential. However, testing the effect of these different terms was time-consuming, due to the slow training times of SNNs and the convolutional implementation that I was using, where training a single model could take $\sim 2-3$ days. To accelerate research efforts I reverted the model to patch-mode, reducing model training to ~ 1 day. After more trial-and-error I employed the metabolic regularization function as reported in the chapters that jointly penalizes synaptic connectivity and firing rates - this did not yet penalize the excitatory and inhibitory units separately. Although the model at this point was capturing the V1-like RFs and spike statistics, it did not capture the nuance differences in tuning between the excitatory and inhibitory V1 neurons.

I was able to capture the differences between the excitatory and inhibitory V1 neurons by regularizing the excitatory and inhibitory units with separate penalties - under the assumption that inhibitory neurons consume less energy than excitatory neurons per unit work. Here, I again trained various models in search of the most V1-like model. In addition to inspecting the RFs, I also inspected the membrane time constant distributions of the model as well as the motion tuning statistics. This finally resulted in the model for which the results are presented in Chapter 2.

The retinal model was a direct extension of the V1 model, yet with no Dale's law on the recurrent connectivity and using a different metabolic regularization strength. Unlike the V1 model, I obtained the results reported in Chapter 3 without performing many trial-and-error experiments.

5.3 Extending the temporal prediction objective

In this thesis, I trained the cortical (Chapter 2) and retinal (Chapter 3) models to predict the most likely sensory input at a fixed time into the future. This was achieved by minimizing the MSE between the future movie frames and the corresponding model predictions. This, however, has two potential shortcomings: firstly, the brain might predict a span of offsets into the future (*e.g.* 8ms, 16ms, 24ms, etc. into the future); and secondly, the brain might predict a distribution over possible future outcomes, rather than predicting the single most likely outcome.

Understanding the probability distribution of potential future events is crucial for effective action selection, particularly for balancing survival against energy conservation. Consider a surfer

anticipating an oncoming wave: the optimal decision might be to paddle towards and over the wave to avoid injury. However, if the wave is likely to pass by harmlessly, conserving energy by staying put might be preferable. Furthermore, if the wave has a chance of breaking to the left or right, the surfer must decide which way to paddle. A simplistic prediction might suggest that the wave will hit directly head-on, but this is an example of the golden mean fallacy — the erroneous conclusion that the truth lies in between two extremes [392]. The surfer could thus make an informed decision by knowing the likelihood of the different future outcomes of the wave - and not just predicting a single outcome.

The models of the retina and V1 presented in this thesis could be trained to generate various possible future outcomes using recent advancements in generative machine learning. Unlike the fixed single-frame prediction employed during the training process of the retina and V1 models, generative models could instead be used to generate different possible future frames. These models, in particular conditional generative models, consume a condition-vector and a noise-vector as input. The condition-vector provides the necessary contextual information from which an image can be generated - for example, the past spiking activity in the retina and V1 models. The noise-vector specifies a random starting seed for the generative process, whereby different noise-vectors produce different generated samples [393].

Possible starting points to extend the models from predicting the single most likely frame to a distribution of possible frames would be to include a generative adversarial network (GAN) or a diffusion model in the training setup. A GAN is comprised of two neural networks, a generator and a discriminator network. These are simultaneously trained, such that the generator creates data aiming to mimic real data, while the discriminator tries to distinguish between the real and generated data [394]. Although these networks can generate high-resolution images [395], they are typically challenging to train and can suffer from so-called mode-collapse, a lack of diversity in generated samples [396]. Diffusion models overcome the shortcomings of GANs and synthesize images of higher quality [397]. These models learn to generate data by gradually adding random noise to a data sample and then learning to reverse this process [398]. The reversal process yields a generative model capable of generating high-resolution images [399] and even movies [400]. The movie generation capabilities of diffusion models would be a good starting point for extending the spiking models to predict different sequences of possible future sensory outcomes.

It remains unclear how the temporal prediction objective might be implemented and supported by the neural circuit. An active area of research is to determine which aspects of the visual system are hard-wired vs shaped via visual experience [254]. If the brain is performing temporal prediction, one might assume that it is continuously tweaking its connectivity to best predict the

sensory future. For example, raccoons live in various environments, including forests, mountains, and urban areas [401], which each exhibit different visual statistics (*e.g.* urban areas have driving cars, which are less likely to be seen in mountains and forests). Thus, the brain might reconfigure certain connections to best adapt to particular environmental statistics [402] (*e.g.* a raccoon moving from the mountains to an urban area). How a brain region could be optimised (or fine-tuned) for the temporal prediction objective by experience-dependent neural plasticity is unclear, as neurons would require a learning signal, such as the error between the predicted and future sensory states. The brain would thus need to keep a record of its sensory predictions to compare these to the sensory futures. Future investigations could try to understand how such sensory prediction errors might be computed, and how the synaptic connectivity might adapt to reduce such errors.

5.4 Modelling extensions

There are various avenues for extending the V1 and retinal models developed in this thesis. One modelling extension would be to make the models operate over an entire spatial extent of the visual input. Like prior normative models of the visual system [106, 140, 141], the models developed in this thesis consume a patch of visual input, instead of processing an entire visual scene. One extension would be to make the models convolutional, jointly processing multiple patches at different spatial locations [47]. However, a convolutional model would likely not differ drastically from convolving a pre-trained patch-model over space. This is due to the patch-models being trained using movie patches uniformly sampled from different spatial locations within the entire movie frame. Thus, similar to the convolutional model, the patch-models were trained on the movie statistics spanning the spatial dimensions.

A better approach than a convolutional model might be to learn a unique patch-model at every spatial location covered by the visual input. One reason for doing so is that the input statistics vary over different spatial regions within the visual field. Such differences in the input statistics are also reflected in the visual system, *e.g.* by the spatial layout of the retina across different animal species. For example, mice mostly see blue skies above (top visual hemisphere) and the ground and bushes below (bottom visual hemisphere). Their eyes are also slightly tilted outwards, which creates a visual flow over their retinas when moving forwards. The retina of mice exhibit different cellular arrangements corresponding to different parts of the visual field, as illustrated by the types of photoreceptors, variations in the densities of ganglion cells, as well as varying levels of ganglion-cell direction-selectivity [329]. These different properties

likely reflect the retina's evolutionary adaptation to process visual input of different statistics at different spatial locations. Exploring such modelling extensions is, however, challenging, due to the large memory requirements and long training times when simulating biologically-detailed spiking networks over time and space.

Another avenue for extending this work would be to capture a larger portion of the visual system in a single model, rather than modelling its individual subregions (like the retina and V1) separately. A next step could be to use the spiking output of the retinal model as input to the cortical V1 model and examine whether V1-like tuning properties still emerge within the V1 model. Here, the modified V1 model would be trained to predict future retinal spiking activity instead of future movie frames. Such a normative objective, of higher layers predicting the activity of lower layers, has been shown to produce visual-system-like tuning properties in hierarchical non-spiking networks [162]. Furthermore, additional layers could be included, in an attempt to capture the tuning properties of the LGN or higher cortical areas. In particular, it would be interesting to explore whether different visual streams corresponding to the dorsal and ventral streams found in the primate cortex emerge in a hierarchical spiking network optimized for temporal prediction. Functional specification has been shown to emerge in non-spiking hierarchical networks trained using a prediction-like objective [163], or by using a self-supervised learning objective with spatial connectivity constraints [403].

The developed spiking models could also be examined using other sensory modalities, like auditory input. Instead of training the spiking models to predict future movie frames, the models could instead be trained to predict the future values in cochleagrams or raw waveforms of natural sounds. Prior investigations found that optimizing shallow non-spiking networks to predict the future of natural sound spectrograms from the past resulted in model units with spectrotemporal RFs resembling those of A1 neurons [106]. Additionally, optimizing a simple non-spiking network to predict the immediate future of raw waveforms of natural sounds from the past results in units with cochlear-like tuning properties [404]. It would be interesting to explore whether similar tuning properties would emerge in the spiking temporal prediction models. Another possibility would be to develop a spiking model of the visual and auditory pathway, which jointly predicts the visual and auditory future from the recent past of both modalities. Various studies have described how neurons in the auditory cortex are modulated by visual input [405], or how neurons in the visual cortex are modulated by auditory input [406]. The function of such modulation remains elusive. Perhaps the different sensory streams support each other in predicting their respective sensory futures.

5.5 The challenges of normative modelling

Can normative modelling lead to an understanding of the brain? What constitutes an understanding of the brain remains a contentious question. There are billions of neurons connected through trillions of synapses, making it a difficult problem to conceptualize how these all interact to form a functioning brain. Arguably, we would still fall short of understanding the brain if we were to map all the brain's connections and henceforth be able to predict all its spiking dynamics, because we would still not understand what computation these connections and spiking dynamics are performing. Normative models bridge this gap through the development of models that attempt to explain why the brain looks and behaves the way it does.

Models can help us to comprehend various interactions and observations in the world (and beyond). For example, models allow us to calculate the trajectory of motion of two colliding bodies [36], and - more abstractly - they allow us to derive how electrical charges give rise to magnetic fields [37]. Models thus provide us with an understanding of the world. As Nobel laureate and theoretical physicist Richard Feynman puts it: "What I cannot create, I do not understand" [407]. Normative models embody such ethos, with the aspiration of understanding why the neurons in the brain behave the way they do. Although these models have successfully managed to account for various neural tuning properties throughout the visual system [106, 140, 141], normative models face several limitations:

1. Scaling constraints. The primate brain consists of billions of neurons and trillions of synapses [7]. In comparison, the largest ANNs typically consist of orders of magnitude fewer trainable parameters. In particular, biologically-detailed networks of spiking neurons are typically comprised of a few hundred (or thousand) units, which is drastically less than the number of neurons in most brains. This raises the question: how important is it to match the scale of the brain when modelling the brain *in silico*? Some research endeavours argue that such a scale is pivotal in understanding the brain. For example, the Blue Brain Project aims to create a detailed digital reconstruction of the mouse brain, to further examine its function [408]. Such large-scale investigations, however, do not train the model architecture with respect to a normative goal, due to limitations in software (*e.g.* training and simulation algorithms) and hardware. New software developments are making it possible to train biologically-plausible SNNs more rapidly (Chapter 4) [128, 132] and of larger scale [132]. Furthermore, new developments in hardware - aiming to re-build computers to be more like the brain - might make it possible to emulate larger brain-like networks [356]. These developments are paving the way towards training large-scale, biologically-detailed models. Although it remains unclear to what extent the network size

is important for understanding the brain, one may argue that larger networks more closely resemble the brain, and thus form a better basis for studying its functional organization *in silico*.

2. Setting hyperparameters. The emergent phenomena of a normative model rely on several hyperparameters. These parameters are not learnt but rather hard-coded. For example, in the spiking V1 temporal prediction model explored in Chapter 2, one hyperparameter selects how far into the future the model should predict, and another hyperparameter determines how many of the units should be excitatory vs inhibitory. The values for these hyperparameters can be informed by biology. For example, the prediction horizon could be on the timescale of the transmission latency from the retina to V1 [100, 101, 102, 103], and the fraction of excitatory to inhibitory units can match the fraction of excitatory to inhibitory neurons found in V1 [113, 114, 112, 23]. However, certain hyperparameters are more challenging to infer, such as the hyperparameter values in the metabolic-like cost functions used to train the V1 and retinal models. No direct criterion is available to derive these hyperparameters. Rather, through trial-and-error, various models were trained using different hyperparameter values. Hyperparameter values were selected which gave the most neural-like tuning properties. This is arguably a major shortcoming of normative models, as model development requires many computational resources and long training times. This results in a slow scientific process, and a large carbon footprint due to the energy requirements of the hardware used to simulate and train the models [409, 410]. Unfortunately, certain hyperparameters, like those in the metabolic-like cost functions, are challenging to derive from theoretical measures and biological observations. In particular, the calculations regarding the energy expenditure of neurons - and their different excitatory and inhibitory subtypes - is still an ongoing topic of investigation [271, 272, 273, 274]. Further developments in this area will hopefully allow for the metabolic-like cost function hyperparameters to be derived, rather than having to determine these using a brute-force search of the hyperparameter space.

3. To learn or not to learn. The goal in normative modelling is for brain-like phenomena to emerge within a model during the optimization process. The number of emergent phenomena is, however, bound by the design of the model architecture. For example, training a network of LN units with respect to a particular normative hypothesis will not result in spiking network dynamics. Clearly, the spiking mechanism must be included within the model architecture for the model to exhibit spiking dynamics. Some criteria are, however, less well defined as to whether they should be included in the model architecture, or emerge during the optimization process. For example, the spiking V1 temporal prediction

model in Chapter 2 has a fixed ratio of excitatory to inhibitory units, matching the ratio of excitatory to inhibitory neurons in V1. Instead of hard-coding this ratio, it could also be learned. However, models with an increasing number of learnable parameters become more difficult to develop, as they are more challenging to implement, and require more memory and time to train. Thus, a trade-off is reached between hard-coding and learning the parameters within a normative model.

4. One objective to rule them all? Different normative models are contrasted to each other, with the aspiration of determining which normative theory is most plausible. For example, efficiently predicting future sensory input from the past produces model units with temporal-like tuning of V1 neurons, whereas efficiently encoding the present sensory input does not [106]. Thus, it could be argued that prediction, rather than compression, is more likely to be the computational principle underlying the computation of V1. However, as geneticist Theodosius Dobzhansky puts it: "Nothing in biology makes sense except in the light of evolution" [411]. Natural selection is the overarching principle governing the design and function of the brain [85]. It may therefore be myopic to assume that a single principle, like temporal prediction, governs the computations underlying sensory circuits, as various normative theories are reconcilable in the context of evolution. Rather, sensory circuits may have evolved to support a variety of computational principles. Future investigations may thus want to investigate an ensemble of normative theories, rather than treating them as mutually exclusive.

5. Likely unexplainable phenomena. Normative models can account for many neural properties, such as the RFs of neurons; their spike statistics; their tuning preferences to motion; and their different timescales of operations (Chapters 2 and 3). However, it seems improbable that all biological phenomena can be accounted for by training a model architecture with respect to a normative objective. For example, it seems unlikely that a normative model of the retina could account for its backward-wiring, where light first passes through the higher-layer nerve cells before reaching the photoreceptors. Additionally, it seems unlikely that a normative model of the visual system could account for the crossover of visual information, whereby light entering the left retina projects to the right cortical hemisphere and the light entering the right retina projects to the left cortical hemisphere. These wiring phenomena likely emerged from particular developmental or evolutionary pressures and contingencies unrelated to normative theories.

Although normative modelling endures various shortcomings, this does not mean that normative models are not useful for studying the brain. As per the aphorism: "All models are wrong, but

some are useful". Normative models may not account for the backwards wiring of the retina, but they can, for example, account for the neural tuning and spike statistics of the retina and V1 (Chapters 2 and 3).

One approach to broadening the applicability of normative models would be to incorporate the notion of embodiment: allowing models to interact with a 3D world through sensorimotor-like skills. A long-standing test of brain-like intelligence is the so-called "imitation game" proposed by Alan Turing, in which an interrogator must determine whether they are communicating with a human or a machine via some form of written dialogue [412]. With the advent of large language models (LLMs) in the last years, it has been questioned whether this is still a valid test of brain-like intelligence, given that LLMs possess human-level language abilities [413]. Recently, neuroscience and AI researchers have proposed a new Turing-like test, the "embodied Turing test" [414]. The measure of intelligence in this new test is whether a model with sensorimotor-like capabilities exhibits behavior indistinguishable from that of living counterparts (*e.g.* mice, squirrels, or humans). The model could be interacting in a physical or virtual environment, with the objective of exhibiting species-specific behaviour. As neuroscientist Gordon Shepherd puts it: "Nothing in neurobiology makes sense except in the light of behavior" [415]. Casting the normative framework from the sensory to the sensorimotor setting permits the models to exhibit behaviour and would thus increase their biological realism, and create more opportunities for comparison to biology.

A starting point for developing models for the "embodied Turing test", would be to train an agent in a realistic virtual environment, with the ability to see (visual input) and act (motor output). Simulated agent interactions are more accessible than physical interactions, where the latter requires complex and expensive robots [416]. The agent could be a mouse-like model interacting in a forest-like virtual environment. This environment could contain food sources (for energy), predators (which need to be avoided), and underground holes (for protection). The agent could then be trained using a method like reinforcement learning [417] or evolutionary algorithms [418], to maximize chances of survival (eat food, avoid predators, and find shelter). A biologically-plausible model architecture could be employed, like the spiking network of excitatory and inhibitory units outlined in Chapter 2. It would be interesting to observe whether distinct motor-like and visual-like pathways or motor-visual dependencies emerge within the model, where particular behaviours have been linked to neural responses within V1 of mice [419].

5.6 Validation extensions

The spiking V1 and retinal model were contrasted to the biology using common measures, like RFs, spike statistics, and motion-tuning properties, between the model units and real neurons. The spiking retinal model was also used to predict neural responses from the retina of different animal species to natural movies and images. These different measures allow us to assess how similar the models are to their respective biological counterparts. Future investigations might want to explore additional measures to assess the similarity of the models to the biology.

It would be interesting to further investigate the similarity between the population responses between the model and the biology to different forms of stimuli. A standard approach for doing such a comparison is representational similarity analysis (RSA) [420]. This approach uses a response matrix that contains the model (or brain) responses over the units (or neurons) along the columns, in response to different stimuli listed as rows. RSA calculates the similarity between two response matrices - one from the model and one from the brain - as a single number, ranging between zero and one. A value of zero implies dissimilarity and a value of one implies equivalence between the matrices. Other measures, such as the canonical correlation analysis (CCA) [421] and centered kernel alignment (CKA) [422], similarly measure the similarity between the responses of models and brains to different stimuli. Although these measures have been popular in assessing the similarity of models to the brain, they suffer from various shortcomings. For example, these measures have been shown to violate the properties of a proper metric, as defined in mathematics (*e.g.* violation of the triangle inequality [423]). Recent developments are addressing these shortcomings [423]. However, what constitutes a good measure of similarity between the population responses of a model and the brain remains an active area of investigation.

Another opportunity for future analyses would be to investigate the retinal and V1 model spike dynamics using neural manifolds [424]. A neural manifold represents a low-dimensional latent space, along which high-dimensional neural spike trains evolve. Such neural manifolds have been shown to govern the evolution of neural dynamics in the dorsal premotor cortex of monkeys during a delayed center-out reach task [425], and the evolution of neural dynamics in V1 of monkeys to various input stimuli [426, 427]. It would be interesting to analyse whether the spiking model dynamics also evolve in a lower-dimensional space like a neural manifold; and if they do, do these match the neural manifolds observed in the retina and V1? For example, it has been shown that drifting gratings of different orientations elicit high-dimensional spike responses in V1 of monkeys, which evolve along a lower-dimensional torus-like shape. Spike

responses corresponding to a particular grating orientation evolve along a unique trajectory on the torus [428]. Perhaps the same results hold true for the spiking V1 model from Chapter 2.

Lastly, to assess the similarity of a model to a particular region within the visual system (like V1), it has become a common practice to predict that region's neural responses from the model unit activations in response to the same visual stimulus [199]. However, it remains a contentious question of how best to assess the goodness-of-fit between the predicted and recorded neural responses. When neural responses are represented as firing rates, a common measure is the normalized correlation coefficient, which measures the Pearson correlation between the predicted and target neural responses, taking into account the maximal explainable correlation of each neuron due to their response variabilities [317]. However, it remains less clear which measure to adopt when assessing the goodness-of-fit for the prediction of individual spiking events, as opposed to predicting firing rates. In Chapter 4, I employed the explained temporal variance (ETV) (outlined on the Allen Institute website) to assess the goodness-of-fit in predicting the spike trains of cortical neurons to various noisy input current injections. However, this measure smoothes the spike trains using a Gaussian kernel, which results in a signal similar to a firing rate - thus potentially omitting the information encoded by the individual spiking events. Various other measures have been proposed to assess the goodness-of-fit when predicting spike trains, such as likelihood approaches [110], the van Rossum distance [381], and the Victor-Purpura metric [429]. However, these measures either suffer the same shortcomings as the ETV measure (of spike smoothing) or are more challenging to interpret than the normalized correlation coefficient or ETV. Further developments of new metrics, particularly metrics that assess the goodness-of-fit between spike trains, will likely play an important role in understanding the brain's functionality using computational models.

Bibliography

[1] Charles A Elsberg. “The Edwin Smith surgical papyrus: and the diagnosis and treatment of injuries to the skull and spine 5000 years ago”. In: *Annals of Medical History* 3.3 (1931), p. 271 (cit. on p. 1).

[2] Tomislav Breitenfeld et al. “Hippocrates: the forefather of neurology”. In: *Neurological Sciences* 35.9 (2014), pp. 1349–1352 (cit. on p. 1).

[3] Stanley Finger. *Origins of neuroscience: a history of explorations into brain function*. Oxford University Press, 2001 (cit. on p. 1).

[4] Frank R Freemon. “Galen’s ideas on neurological function”. In: *Journal of the History of the Neurosciences* 3.4 (1994), pp. 263–271 (cit. on p. 1).

[5] Gabriel Finkelstein. *Emil du Bois-Reymond: Neuroscience, self, and society in nineteenth-century Germany*. MIT Press, 2013 (cit. on p. 1).

[6] Stanley Finger. *Minds behind the brain: A history of the pioneers and their discoveries*. Oxford University Press, 2000 (cit. on p. 1).

[7] Eric R Kandel et al. *Principles of neural science*. Vol. 4. McGraw-hill New York, 2000 (cit. on pp. 1, 91, 125).

[8] Seth Blackshaw et al. “Genomic analysis of mouse retinal development”. In: *PLoS biology* 2.9 (2004), e247 (cit. on p. 2).

[9] Haldan Keffer Hartline. “The response of single optic nerve fibers of the vertebrate eye to illumination of the retina”. In: *American Journal of Physiology-Legacy Content* 121.2 (1938), pp. 400–415 (cit. on pp. 2, 6, 60).

[10] Stephen W Kuffler. “Discharge patterns and functional organization of mammalian retina”. In: *Journal of neurophysiology* 16.1 (1953), pp. 37–68 (cit. on pp. 2, 6, 60).

[11] David H Hubel et al. “Receptive fields of single neurones in the cat’s striate cortex”. In: *The Journal of physiology* 148.3 (1959), pp. 574–591 (cit. on pp. 2, 4, 6, 10, 30, 36).

[12] David H Hubel et al. “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex”. In: *The Journal of physiology* 160.1 (1962), p. 106 (cit. on pp. 2, 4, 6, 10, 30, 36, 43).

[13] David H Hubel et al. “Receptive fields and functional architecture of monkey striate cortex”. In: *The Journal of physiology* 195.1 (1968), pp. 215–243 (cit. on pp. 2, 4, 6, 10, 30, 36).

[14] Daniel J Felleman et al. “Distributed hierarchical processing in the primate cerebral cortex.” In: *Cerebral cortex (New York, NY: 1991)* 1.1 (1991), pp. 1–47 (cit. on pp. 2, 5, 9, 46).

[15] Rudolf Nieuwenhuys. “The neocortex: an overview of its evolutionary development, structural organization and synaptology”. In: *Anatomy and embryology* 190 (1994), pp. 307–337 (cit. on pp. 2, 9).

[16] Leslie G Ungerleider. “Two cortical visual systems”. In: *Analysis of visual behavior* 549 (1982), chapter–18 (cit. on pp. 2, 4, 5).

[17] Melvyn A Goodale et al. “Separate visual pathways for perception and action”. In: *Trends in neurosciences* 15.1 (1992), pp. 20–25 (cit. on pp. 2, 4).

[18] David C Van Essen et al. “Neural mechanisms of form and motion processing in the primate visual system”. In: *Neuron* 13.1 (1994), pp. 1–10 (cit. on pp. 2, 4).

[19] Helga Kolb et al. “Webvision: the organization of the retina and visual system [Internet]”. In: (1995) (cit. on pp. 2, 9, 11, 12).

[20] Federico Scala et al. “Layer 4 of mouse neocortex differs in cell types and circuit organization between sensory areas”. In: *Nature communications* 10.1 (2019), p. 4174 (cit. on pp. 2, 12).

[21] Edward M Callaway. “Inhibitory cell types, circuits and receptive fields in mouse visual cortex”. In: *Micro-, Meso-and Macro-Connectomics of the Brain* (2016), pp. 11–18 (cit. on pp. 2, 12).

[22] Jochen F Staiger et al. “Functional diversity of layer IV spiny neurons in rat somatosensory cortex: quantitative morphology of electrophysiologically characterized and biocytin labeled cells”. In: *Cerebral Cortex* 14.6 (2004), pp. 690–701 (cit. on pp. 2, 12).

[23] Sonja B Hofer et al. “Differential connectivity and response dynamics of excitatory and inhibitory neurons in visual cortex”. In: *Nature neuroscience* 14.8 (2011), pp. 1045–1052 (cit. on pp. 2, 11–13, 31, 39–41, 50, 126).

[24] Jeffrey S Anderson et al. “Orientation tuning of input conductance, excitation, and inhibition in cat primary visual cortex”. In: *Journal of neurophysiology* 84.2 (2000), pp. 909–926 (cit. on pp. 2, 13, 39, 41).

[25] Nathan W Gouwens et al. “Classification of electrophysiological and morphological neuron types in the mouse visual cortex”. In: *Nature neuroscience* 22.7 (2019), pp. 1182–1195 (cit. on pp. 2, 13, 39, 40).

[26] David A McCormick et al. “Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex”. In: *Journal of neurophysiology* 54.4 (1985), pp. 782–806 (cit. on pp. 2, 12).

[27] Malte J Rasch et al. “Statistical comparison of spike responses to natural stimuli in monkey area V1 with simulated responses of a detailed laminar network model for a patch of V1”. In: *Journal of neurophysiology* 105.2 (2011), pp. 757–778 (cit. on pp. 2, 13, 28, 30, 33, 34, 43, 45, 53, 54, 59, 82).

[28] Cristopher M Niell et al. “Highly selective receptive fields in mouse visual cortex”. In: *Journal of Neuroscience* 28.30 (2008), pp. 7520–7536 (cit. on pp. 2, 7, 9, 11–13, 34–39, 44, 47, 57).

[29] J Anthony Movshon et al. “Receptive field organization of complex cells in the cat’s striate cortex.” In: *The Journal of physiology* 283.1 (1978), pp. 79–99 (cit. on pp. 2, 10, 37).

[30] J Anthony Movshon et al. “Spatial summation in the receptive fields of simple cells in the cat’s striate cortex.” In: *The Journal of physiology* 283.1 (1978), pp. 53–77 (cit. on pp. 2, 10, 37).

[31] Judson P Jones et al. “An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex”. In: *Journal of neurophysiology* 58.6 (1987), pp. 1233–1258 (cit. on pp. 2, 10, 35, 36, 54, 55).

[32] Dario L Ringach. “Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex”. In: *Journal of neurophysiology* (2002) (cit. on pp. 2, 10, 35, 36, 54, 55).

[33] David H Hubel et al. “The period of susceptibility to the physiological effects of unilateral eye closure in kittens”. In: *The Journal of physiology* 206.2 (1970), pp. 419–436 (cit. on pp. 2, 44).

[34] Gregory C DeAngelis et al. “Spatiotemporal organization of simple-cell receptive fields in the cat’s striate cortex. I. General characteristics and postnatal development”. In: *Journal of neurophysiology* 69.4 (1993), pp. 1091–1117 (cit. on pp. 2, 6, 10, 36, 44, 56).

[35] Nathalie L Rochefort et al. “Development of direction selectivity in mouse cortical neurons”. In: *Neuron* 71.3 (2011), pp. 425–432 (cit. on pp. 2, 7, 39, 57).

[36] Isaac Newton et al. *Newton’s Principia: the mathematical principles of natural philosophy*. Geo. P. Putnam, 1850 (cit. on pp. 3, 125).

[37] James Clerk Maxwell. “VIII. A dynamical theory of the electromagnetic field”. In: *Philosophical transactions of the Royal Society of London* 155 (1865), pp. 459–512 (cit. on pp. 3, 125).

[38] Albert Einstein et al. “The foundation of the general theory of relativity”. In: *Annalen Phys* 49.7 (1916), pp. 769–822 (cit. on p. 3).

[39] I-Hung Wang et al. “Images classification of dogs and cats using fine-tuned VGG models”. In: *2020 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)*. IEEE. 2020, pp. 230–233 (cit. on p. 3).

[40] Liang-Chieh Chen et al. “Rethinking atrous convolution for semantic image segmentation. arXiv 2017”. In: *arXiv preprint arXiv:1706.05587* 2 (2019) (cit. on p. 3).

[41] Kelvin Xu et al. “Show, attend and tell: Neural image caption generation with visual attention”. In: *International conference on machine learning*. PMLR. 2015, pp. 2048–2057 (cit. on pp. 3, 4, 27).

[42] Yann LeCun et al. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444 (cit. on pp. 3, 23).

[43] David Silver et al. “Mastering the game of Go with deep neural networks and tree search”. In: *nature* 529.7587 (2016), pp. 484–489 (cit. on p. 3).

[44] Oriol Vinyals et al. “Grandmaster level in StarCraft II using multi-agent reinforcement learning”. In: *Nature* 575.7782 (2019), pp. 350–354 (cit. on p. 3).

[45] Mariusz Bojarski et al. “End to end learning for self-driving cars”. In: *arXiv preprint arXiv:1604.07316* (2016) (cit. on p. 3).

[46] Daniel Martin Katz et al. “Gpt-4 passes the bar exam”. In: *Available at SSRN 4389233* (2023) (cit. on p. 4).

[47] Alex Krizhevsky et al. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012) (cit. on pp. 4, 27, 74, 123).

[48] Shervin Minaee et al. “Image segmentation using deep learning: A survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.7 (2021), pp. 3523–3542 (cit. on pp. 4, 27).

[49] Lane McIntosh et al. “Deep learning models of the retinal response to natural scenes”. In: *Advances in neural information processing systems* 29 (2016) (cit. on p. 4).

[50] Daniel LK Yamins et al. “Performance-optimized hierarchical models predict neural responses in higher visual cortex”. In: *Proceedings of the national academy of sciences* 111.23 (2014), pp. 8619–8624 (cit. on pp. 4, 27).

[51] Santiago A Cadena et al. “Deep convolutional models improve predictions of macaque V1 responses to natural images”. In: *PLoS computational biology* 15.4 (2019), e1006897 (cit. on pp. 4, 27, 70, 88, 91).

[52] Takafumi Sasakawa et al. “A brainlike learning system with supervised, unsupervised, and reinforcement learning”. In: *Electrical Engineering in Japan* 162.1 (2008), pp. 32–39 (cit. on p. 4).

[53] NICHOLLS JG. “A cellular and molecular approach to the function of the nervous system”. In: *From Neuron to Brain* (1992), pp. 90–120 (cit. on p. 5).

[54] Semir M Zeki. “Functional organization of a visual area in the posterior bank of the superior temporal sulcus of the rhesus monkey”. In: *The Journal of physiology* 236.3 (1974), pp. 549–573 (cit. on p. 4).

[55] Mototaka Suzuki et al. “How deep is the brain? The shallow brain hypothesis”. In: *Nature Reviews Neuroscience* (2023), pp. 1–14 (cit. on pp. 5, 46).

[56] Farran Briggs et al. “Emerging views of corticothalamic function”. In: *Current opinion in neurobiology* 18.4 (2008), pp. 403–407 (cit. on p. 5).

[57] Farran Briggs. “Role of feedback connections in central visual processing”. In: *Annual review of vision science* 6 (2020), pp. 313–334 (cit. on p. 5).

[58] Paul-Antoine Salin et al. “Corticocortical connections in the visual system: structure and function”. In: *Physiological reviews* 75.1 (1995), pp. 107–154 (cit. on p. 5).

[59] Henry J Alitto et al. “Corticothalamic feedback and sensory processing”. In: *Current opinion in neurobiology* 13.4 (2003), pp. 440–445 (cit. on p. 5).

[60] Simon Thorpe et al. “Speed of processing in the human visual system”. In: *nature* 381.6582 (1996), pp. 520–522 (cit. on p. 5).

[61] EJ Chichilnisky. “A simple white noise analysis of neuronal light responses”. In: *Network: computation in neural systems* 12.2 (2001), p. 199 (cit. on p. 6).

[62] Peter Dayan et al. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Computational Neuroscience Series, 2001 (cit. on pp. 6, 12–17, 55, 56, 60, 72).

[63] Edgar Y Walker et al. “Inception loops discover what excites neurons most using deep predictive models”. In: *Nature neuroscience* 22.12 (2019), pp. 2060–2065 (cit. on p. 6).

[64] Rudi Tong et al. “The feature landscape of visual cortex”. In: *bioRxiv* (2023), pp. 2023–11 (cit. on p. 6).

[65] Peter H Schiller et al. “Quantitative studies of single-cell properties in monkey striate cortex. I. Spatiotemporal organization of receptive fields”. In: *Journal of neurophysiology* 39.6 (1976), pp. 1288–1319 (cit. on p. 7).

[66] Martin S Gizzi et al. “Selectivity for orientation and direction of motion of single neurons in cat striate and extrastriate visual cortex”. In: *Journal of neurophysiology* 63.6 (1990), pp. 1529–1543 (cit. on p. 7).

[67] E Gramage et al. “The expression and function of midkine in the vertebrate retina”. In: *British journal of pharmacology* 171.4 (2014), pp. 913–923 (cit. on p. 7).

[68] Samuel Ocko et al. “The emergence of multiple retinal cell types through efficient coding of natural movies”. In: *Advances in Neural Information Processing Systems* 31 (2018) (cit. on pp. 7, 29, 60, 66, 72, 73, 91).

[69] Dimokratis Karamanlis et al. “Natural stimuli drive concerted nonlinear responses in populations of retinal ganglion cells”. In: *bioRxiv* (2023), pp. 2023–01 (cit. on pp. 7, 63, 64, 66, 70, 86).

[70] Christine A Curcio et al. “Human photoreceptor topography”. In: *Journal of comparative neurology* 292.4 (1990), pp. 497–523 (cit. on p. 7).

[71] Robert S Molday et al. “Photoreceptors at a glance”. In: *Journal of cell science* 128.22 (2015), pp. 4039–4045 (cit. on p. 8).

[72] Tim Gollisch et al. “Eye smarter than scientists believed: neural computations in circuits of the retina”. In: *Neuron* 65.2 (2010), pp. 150–164 (cit. on pp. 8, 60, 62, 68, 70, 73–75, 77).

[73] Heinz Wässle. “Parallel processing in the mammalian retina”. In: *Nature Reviews Neuroscience* 5.10 (2004), pp. 747–757 (cit. on p. 8).

[74] Dennis M Dacey. “The mosaic of midget ganglion cells in the human retina”. In: *Journal of Neuroscience* 13.12 (1993), pp. 5334–5355 (cit. on pp. 8, 65).

[75] Alexandra Kling et al. “Functional organization of midget and parasol ganglion cells in the human retina”. In: *BioRxiv* (2020), pp. 2020–08 (cit. on pp. 8, 26, 65, 75, 81).

[76] Jian K Liu et al. “Simple model for encoding natural images by retinal ganglion cells with nonlinear spatial integration”. In: *PLoS Computational Biology* 18.3 (2022), e1009925 (cit. on pp. 8, 26, 62–65, 70, 76, 81, 87, 90).

[77] Robert W Rodieck. “Quantitative analysis of cat retinal ganglion cell response to visual stimuli”. In: *Vision research* 5.12 (1965), pp. 583–601 (cit. on pp. 8, 26).

[78] Charles P Ratliff et al. “Retina is structured to process an excess of darkness in natural scenes”. In: *Proceedings of the National Academy of Sciences* 107.40 (2010), pp. 17368–17373 (cit. on pp. 8, 65).

[79] Dennis M Dacey. “Physiology, morphology and spatial densities of identified ganglion cell types in primate retina”. In: *Ciba Foundation Symposium 184-Higher-Order Processing in the Visual System: Higher-Order Processing in the Visual System: Ciba Foundation Symposium 184*. Wiley Online Library. 2007, pp. 12–34 (cit. on pp. 8, 65).

[80] Barry B Lee. “Receptive field structure in the primate retina”. In: *Vision research* 36.5 (1996), pp. 631–644 (cit. on pp. 8, 65).

[81] Greg D Field et al. “Spatial properties and functional organization of small bistratified ganglion cells in primate retina”. In: *Journal of Neuroscience* 27.48 (2007), pp. 13261–13272 (cit. on pp. 8, 65).

[82] H Wässle et al. “Morphology and topography of on-and off-alpha cells in the cat retina”. In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 212.1187 (1981), pp. 157–175 (cit. on pp. 8, 65).

[83] Steven H Devries et al. “Mosaic arrangement of ganglion cell receptive fields in rabbit retina”. In: *Journal of neurophysiology* 78.4 (1997), pp. 2048–2060 (cit. on pp. 8, 65).

[84] Greg D Field et al. “Information processing in the primate retina: circuitry and coding”. In: *Annu. Rev. Neurosci.* 30 (2007), pp. 1–30 (cit. on pp. 8, 65).

[85] Horace B Barlow et al. “Possible principles underlying the transformation of sensory messages”. In: *Sensory communication* 1.01 (1961), pp. 217–233 (cit. on pp. 8, 18, 42, 127).

[86] Markus Meister et al. “The neural code of the retina”. In: *neuron* 22.3 (1999), pp. 435–450 (cit. on pp. 8, 60, 62, 90).

[87] Gianluca Tosini et al. “The circadian clock system in the mammalian retina”. In: *Bioessays* 30.7 (2008), pp. 624–633 (cit. on p. 8).

[88] Clyde W Oyster et al. “Direction-selective units in rabbit retina: distribution of preferred directions”. In: *Science* 155.3764 (1967), pp. 841–842 (cit. on pp. 8, 65).

[89] Norma Krystyna Kühn et al. “Joint encoding of object motion and motion direction in the salamander retina”. In: *Journal of Neuroscience* 36.48 (2016), pp. 12203–12216 (cit. on pp. 8, 65, 68).

[90]Amurta Nath et al. “Cardinal orientation selectivity is represented by two distinct ganglion cell types in mouse retina”. In: *Journal of Neuroscience* 36.11 (2016), pp. 3208–3221 (cit. on pp. 8, 65).

[91]Bence P Ölveczky et al. “Segregation of object and background motion in the retina”. In: *Nature* 423.6938 (2003), pp. 401–408 (cit. on pp. 8, 29, 60, 68, 73).

[92]Michael J Berry et al. “Anticipation of moving stimuli by the retina”. In: *Nature* 398.6725 (1999), pp. 334–338 (cit. on pp. 8, 29, 60, 67, 68, 73, 85).

[93]Greg Schwartz et al. “Detection and prediction of periodic patterns by the retina”. In: *Nature neuroscience* 10.5 (2007), pp. 552–554 (cit. on pp. 8, 29, 60, 61, 69, 70, 73, 83).

[94]Sylvia Schröder et al. “Arousal modulates retinal output”. In: *Neuron* 107.3 (2020), pp. 487–495 (cit. on p. 9).

[95]Matteo Carandini. “Area V1”. In: *Scholarpedia* 7.7 (2012), p. 12105 (cit. on pp. 9, 10).

[96]Leah Krubitzer. “The magnificent compromise: cortical field evolution in mammals”. In: *Neuron* 56.2 (2007), pp. 201–208 (cit. on p. 9).

[97]Joel Zylberberg et al. “Sparse coding models can exhibit decreasing sparseness while learning sparse codes for natural images”. In: *PLoS computational biology* 9.8 (2013), e1003182 (cit. on pp. 9, 35, 43).

[98]Brian A Wandell et al. “Visual cortex in humans”. In: *Encyclopedia of neuroscience* 10 (2009), pp. 251–257 (cit. on p. 9).

[99]Genevieve Leuba et al. “Changes in volume, surface estimate, three-dimensional shape and total number of neurons of the human primary visual cortex from midgestation until old age”. In: *Anatomy and embryology* 190 (1994), pp. 351–366 (cit. on p. 9).

[100]R Land et al. “Response properties of local field potentials and multiunit activity in the mouse visual cortex”. In: *Neuroscience* 254 (2013), pp. 141–151 (cit. on pp. 9, 21, 31, 49, 126).

[101]Lisa Kirchberger et al. “Contextual drive of neuronal responses in mouse V1 in the absence of feedforward input”. In: *Science advances* 9.3 (2023), eadd2498 (cit. on pp. 9, 21, 31, 49, 126).

[102]Steven E Raiguel et al. “Response latencies of visual cells in macaque areas V1, V2 and V5”. In: *Brain research* 493.1 (1989), pp. 155–159 (cit. on pp. 9, 21, 31, 49, 126).

[103]LG Nowak et al. “Visual latencies in areas V1 and V2 of the macaque monkey”. In: *Visual neuroscience* 12.2 (1995), pp. 371–384 (cit. on pp. 9, 21, 31, 49, 126).

[104]Charles G Gross. *Brain, vision, memory: Tales in the history of neuroscience*. MIT Press, 1999 (cit. on p. 10).

[105]Dennis Gabor. “Theory of communication. Part 1: The analysis of information”. In: *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering* 93.26 (1946), pp. 429–441 (cit. on pp. 10, 26, 54).

[106] Yosef Singer et al. “Sensory cortex is optimized for prediction of future input”. In: *elife* 7 (2018), e31557 (cit. on pp. 10, 19, 22, 25–27, 30, 31, 42, 48, 55, 119, 120, 123–125, 127).

[107] Gregory C DeAngelis et al. “Receptive-field dynamics in the central visual pathways”. In: *Trends in neurosciences* 18.10 (1995), pp. 451–458 (cit. on pp. 10, 36).

[108] Judith McLean et al. “Contribution of linear mechanisms to the specification of local motion by simple cells in areas 17 and 18 of the cat”. In: *Visual neuroscience* 11.2 (1994), pp. 271–294 (cit. on pp. 10, 36).

[109] Mozak Science. *Neuron Types*. `https://www.mozak.science/guide/neuron-types`. Accessed: December 11, 2023 (cit. on pp. 11, 12).

[110] Wulfram Gerstner et al. *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014 (cit. on pp. 11, 12, 16, 17, 31, 45, 53, 58, 62, 63, 74, 82, 91, 130).

[111] John Carew Eccles. “From electrical to chemical transmission in the central nervous system: the closing address of the sir henry dale centennial symposium cambridge, 19 september 1975”. In: *Notes and records of the Royal Society of London* 30.2 (1976), pp. 219–230 (cit. on pp. 12, 26, 31).

[112] Yuri Gonchar et al. “Multiple distinct subtypes of GABAergic neurons in mouse visual cortex identified by triple immunostaining”. In: *Frontiers in neuroanatomy* 2 (2008), p. 93 (cit. on pp. 12, 31, 50, 126).

[113] AM Sillito. “The contribution of inhibitory mechanisms to the receptive field properties of neurones in the striate cortex of the cat.” In: *The Journal of physiology* 250.2 (1975), pp. 305–329 (cit. on pp. 12, 31, 50, 126).

[114] D Fitzpatrick et al. “Distribution of GABAergic neurons and axon terminals in the macaque striate cortex”. In: *Journal of Comparative Neurology* 264.1 (1987), pp. 73–91 (cit. on pp. 12, 31, 50, 126).

[115] Eugene M Izhikevich. “Simple model of spiking neurons”. In: *IEEE Transactions on neural networks* 14.6 (2003), pp. 1569–1572 (cit. on p. 12).

[116] Henry Markram et al. “Interneurons of the neocortical inhibitory system”. In: *Nature reviews neuroscience* 5.10 (2004), pp. 793–807 (cit. on p. 12).

[117] Michael Okun et al. “Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities”. In: *Nature neuroscience* 11.5 (2008), pp. 535–537 (cit. on pp. 13, 41).

[118] Michael Wehr et al. “Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex”. In: *Nature* 426.6965 (2003), pp. 442–446 (cit. on pp. 13, 41).

[119] Cyril Monier et al. “Orientation and direction selectivity of synaptic inputs in visual cortical neurons: a diversity of combinations produces spike tuning”. In: *Neuron* 37.4 (2003), pp. 663–680 (cit. on pp. 13, 39, 41).

[120] Mingshan Xue et al. “Equalizing excitation–inhibition ratios across visual cortical neurons”. In: *Nature* 511.7511 (2014), pp. 596–600 (cit. on pp. 13, 39, 41).

[121]Tim P Vogels et al. “Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks”. In: *Science* 334.6062 (2011), pp. 1569–1573 (cit. on pp. 13, 17, 41, 46, 91, 104).

[122]Carl Van Vreeswijk et al. “Chaos in neuronal networks with balanced excitatory and inhibitory activity”. In: *Science* 274.5293 (1996), pp. 1724–1726 (cit. on pp. 13, 41, 46).

[123]Marc A Dichter et al. “Cellular mechanisms of epilepsy: a status report”. In: *Science* 237.4811 (1987), pp. 157–164 (cit. on pp. 13, 41, 46).

[124]D Man et al. “A computational investigation into the human representation and processing of visual information”. In: *WH San Francisco: Freeman and Company, San Francisco* 1 (1982) (cit. on p. 13).

[125]Daniel Levenstein et al. “On the role of theory and modeling in neuroscience”. In: *arXiv preprint arXiv:2003.13825* (2020) (cit. on pp. 13, 30, 42, 60, 72).

[126]John E Dowling. *Neurons and networks: an introduction to behavioral neuroscience*. Harvard University Press, 2001 (cit. on p. 14).

[127]David A McCormick. “Membrane properties and neurotransmitter actions”. In: *The synaptic organization of the brain* (1990) (cit. on pp. 14, 17).

[128]Luke Taylor et al. “Addressing the speed-accuracy simulation trade-off for adaptive spiking neurons”. In: *arXiv preprint arXiv:2311.11390* (2023) (cit. on pp. 14, 16, 17, 24, 45, 74, 77, 125).

[129]Henry Clavering Tuckwell. *Introduction to theoretical neurobiology: linear cable theory and dendritic structure*. Vol. 1. Cambridge University Press, 1988 (cit. on p. 15).

[130]Eric Hunsberger. “Spiking deep neural networks: Engineered and biological approaches to object recognition”. In: (2018) (cit. on p. 16).

[131]Nicolas Perez-Nieves et al. “Neural heterogeneity promotes robust learning”. In: *Nature communications* 12.1 (2021), p. 5791 (cit. on pp. 16, 17, 46, 49, 78, 94, 101, 104).

[132]Nicolas Perez-Nieves et al. “Sparse spiking gradient descent”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 11795–11808 (cit. on pp. 16, 17, 45, 74, 95, 101, 125).

[133]Friedemann Zenke et al. “The remarkable robustness of surrogate gradient learning for instilling complex function in spiking neural networks”. In: *Neural computation* 33.4 (2021), pp. 899–925 (cit. on pp. 16, 17, 24, 53, 80, 101, 111, 112).

[134]Alan L Hodgkin et al. “A quantitative description of membrane current and its application to conduction and excitation in nerve”. In: *The Journal of physiology* 117.4 (1952), p. 500 (cit. on p. 17).

[135]JA Connor et al. “Prediction of repetitive firing behaviour from voltage clamp data on an isolated neurone soma”. In: *The Journal of physiology* 213.1 (1971), p. 31 (cit. on p. 17).

[136] Wilfrid Rall. “Branching dendritic trees and motoneuron membrane resistivity”. In: *Experimental neurology* 1.5 (1959), pp. 491–527 (cit. on p. 17).

[137] Renaud Jolivet et al. “Predicting spike times of a detailed conductance-based neuron model driven by stochastic spike arrival”. In: *Journal of Physiology-Paris* 98.4-6 (2004), pp. 442–451 (cit. on p. 17).

[138] Werner M Kistler et al. “Reduction of the Hodgkin-Huxley equations to a single-variable threshold model”. In: *Neural computation* 9.5 (1997), pp. 1015–1045 (cit. on p. 17).

[139] Renaud Jolivet et al. “The quantitative single-neuron modeling competition”. In: *Biological Cybernetics* 99 (2008), pp. 417–426 (cit. on pp. 17, 91).

[140] Bruno A Olshausen et al. “Emergence of simple-cell receptive field properties by learning a sparse code for natural images”. In: *Nature* 381.6583 (1996), pp. 607–609 (cit. on pp. 19, 20, 25, 26, 30, 42, 119, 123, 125).

[141] Rajesh PN Rao et al. “Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects”. In: *Nature neuroscience* 2.1 (1999), pp. 79–87 (cit. on pp. 19, 20, 25–27, 30, 42, 119, 123, 125).

[142] Rasmus Berg Palm. “Prediction as a candidate for learning deep hierarchical models of data”. In: *Technical University of Denmark* 5 (2012) (cit. on pp. 19, 22, 26, 30, 42).

[143] Laurenz Wiskott et al. “Slow feature analysis: Unsupervised learning of invariances”. In: *Neural computation* 14.4 (2002), pp. 715–770 (cit. on pp. 19, 22, 26, 30, 42, 119).

[144] Pietro Berkes et al. “Slow feature analysis yields a rich repertoire of complex cell properties”. In: *Journal of vision* 5.6 (2005), pp. 9–9 (cit. on pp. 19, 22, 26, 30).

[145] Pietro Berkes et al. “A structured model of video reproduces primary visual cortical organisation”. In: *PLoS computational biology* 5.9 (2009), e1000495 (cit. on pp. 19, 22, 25, 26).

[146] Björn Weghenkel et al. “Slowness as a proxy for temporal predictability: An empirical comparison”. In: *Neural computation* 30.5 (2018), pp. 1151–1179 (cit. on pp. 19, 22, 23, 26, 42, 119).

[147] Fred Attneave. “Some informational aspects of visual perception.” In: *Psychological review* 61.3 (1954), p. 183 (cit. on pp. 18, 42).

[148] Mandyam Veerambudi Srinivasan et al. “Predictive coding: a fresh view of inhibition in the retina”. In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 216.1205 (1982), pp. 427–459 (cit. on pp. 20, 22, 26, 42, 119).

[149] Yanping Huang et al. “Predictive coding”. In: *Wiley Interdisciplinary Reviews: Cognitive Science* 2.5 (2011), pp. 580–593 (cit. on pp. 20, 26).

[150] David J Field. “What is the goal of sensory coding?” In: *Neural computation* 6.4 (1994), pp. 559–601 (cit. on pp. 20, 42).

[151] Anthony J Bell et al. “The “independent components” of natural scenes are edge filters”. In: *Vision research* 37.23 (1997), pp. 3327–3338 (cit. on pp. 20, 30, 42).

[152]Michael W Spratling. “Predictive coding as a model of response properties in cortical area V1”. In: *Journal of neuroscience* 30.9 (2010), pp. 3531–3543 (cit. on pp. 21, 30, 42, 119).

[153]S Dora et al. “Deep predictive coding accounts for emergence of complex neural response properties along the visual cortical hierarchy”. In: *bioRxiv* (2020), pp. 2020–02 (cit. on p. 21).

[154]Hermann von Helmholtz. “Concerning the perceptions in general, 1867.” In: (1948) (cit. on p. 21).

[155]Patrick Mineault et al. “Your head is there to move you around: Goal-driven models of the primate dorsal pathway”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 28757–28771 (cit. on pp. 21, 72, 87, 88, 91).

[156]William Bialek et al. “Predictability, complexity, and learning”. In: *Neural computation* 13.11 (2001), pp. 2409–2463 (cit. on p. 21).

[157]Mufeng Tang et al. “Sequential memory with temporal predictive coding”. In: *Advances in Neural Information Processing Systems* 36 (2024) (cit. on p. 21).

[158]Beren Millidge et al. “Predictive coding networks for temporal prediction”. In: *bioRxiv* (2023), pp. 2023–05 (cit. on p. 21).

[159]Rajesh Rao. “Correlates of attention in a model of dynamic visual recognition”. In: *Advances in neural information processing systems* 10 (1997) (cit. on p. 21).

[160]Rajesh PN Rao et al. “Dynamic model of visual recognition predicts neural response properties in the visual cortex”. In: *Neural computation* 9.4 (1997), pp. 721–763 (cit. on p. 21).

[161]William Lotter et al. “A neural network trained for prediction mimics diverse features of biological neurons and perception”. In: *Nature machine intelligence* 2.4 (2020), pp. 210–219 (cit. on pp. 22, 42, 119).

[162]Yosef Singer et al. “Hierarchical temporal prediction captures motion processing along the visual pathway”. In: *Elife* 12 (2023), e52599 (cit. on pp. 22, 30, 31, 45, 91, 119, 124).

[163]Shahab Bakhtiari et al. “The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 25164–25178 (cit. on pp. 22, 91, 124).

[164]Aran Nayebi et al. “Neural Foundations of Mental Simulation: Future Prediction of Latent Representations on Dynamic Scenes”. In: *arXiv preprint arXiv:2305.11772* (2023) (cit. on p. 22).

[165]Naftali Tishby et al. “The information bottleneck method”. In: *arXiv preprint physics/0004057* (2000) (cit. on pp. 22, 42).

[166]Claude Elwood Shannon. “A mathematical theory of communication”. In: *The Bell system technical journal* 27.3 (1948), pp. 379–423 (cit. on p. 22).

[167]William Bialek et al. “Efficient representation as a design principle for neural coding and computation”. In: *2006 IEEE international symposium on information theory*. IEEE. 2006, pp. 659–663 (cit. on pp. 22, 42, 119).

[168] Matthew Chalk et al. “Toward a unified theory of efficient, predictive, and sparse coding”. In: *Proceedings of the National Academy of Sciences* 115.1 (2018), pp. 186–191 (cit. on pp. 22, 30, 42, 119).

[169] Nick Kanopoulos et al. “Design of an image edge detection filter using the Sobel operator”. In: *IEEE Journal of solid-state circuits* 23.2 (1988), pp. 358–367 (cit. on p. 23).

[170] Rajiv Mehrotra et al. “Gabor filter-based edge detection”. In: *Pattern recognition* 25.12 (1992), pp. 1479–1494 (cit. on p. 23).

[171] Rainer Lienhart et al. “An extended set of haar-like features for rapid object detection”. In: *Proceedings. international conference on image processing*. Vol. 1. IEEE. 2002, pp. I–I (cit. on p. 23).

[172] Ben Willmore et al. “The berkeley wavelet transform: a biologically inspired orthogonal wavelet transform”. In: *Neural computation* 20.6 (2008), pp. 1537–1564 (cit. on p. 23).

[173] David E Rumelhart et al. “Learning representations by back-propagating errors”. In: *nature* 323.6088 (1986), pp. 533–536 (cit. on pp. 23, 32, 44, 52, 80, 94).

[174] Peter O’Connor et al. “Real-time classification and sensor fusion with a spiking deep belief network”. In: *Frontiers in neuroscience* 7 (2013), p. 178 (cit. on pp. 24, 94).

[175] Steve K Esser et al. “Backpropagation for energy-efficient neuromorphic computing”. In: *Advances in neural information processing systems* 28 (2015) (cit. on pp. 24, 94).

[176] Bodo Rueckauer et al. “Theory and tools for the conversion of analog to spiking convolutional neural networks”. In: *arXiv preprint arXiv:1612.04052* (2016) (cit. on pp. 24, 94).

[177] Bodo Rueckauer et al. “Conversion of continuous-valued deep networks to efficient event-driven networks for image classification”. In: *Frontiers in neuroscience* 11 (2017), p. 682 (cit. on pp. 24, 94).

[178] Simon Davidson et al. “Comparison of artificial and spiking neural networks on digital hardware”. In: *Frontiers in Neuroscience* 15 (2021), p. 651141 (cit. on p. 24).

[179] Jibin Wu et al. “A tandem learning rule for effective training and rapid inference of deep spiking neural networks”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2021) (cit. on p. 24).

[180] Jibin Wu et al. “Progressive tandem learning for pattern recognition with deep spiking neural networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021) (cit. on p. 24).

[181] Saeed Reza Kheradpisheh et al. “Spiking neural networks trained via proxy”. In: *arXiv preprint arXiv:2109.13208* (2021) (cit. on p. 24).

[182] Rudy Guyonneau et al. “Temporal codes and sparse representations: a key to understanding rapid processing in the visual system”. In: *Journal of Physiology-Paris* 98.4-6 (2004), pp. 487–497 (cit. on p. 24).

[183] Peter Cariani. “Temporal coding of periodicity pitch in the auditory system: an overview”. In: *Neural plasticity* 6.4 (1999), pp. 147–172 (cit. on p. 24).

[184] Ronald J Williams. “Simple statistical gradient-following algorithms for connectionist reinforcement learning”. In: *Machine learning* 8.3 (1992), pp. 229–256 (cit. on p. 24).

[185] H Sebastian Seung. “Learning in spiking neural networks by reinforcement of stochastic synaptic transmission”. In: *Neuron* 40.6 (2003), pp. 1063–1073 (cit. on p. 24).

[186] Sander M Bohte et al. “Error-backpropagation in temporally encoded networks of spiking neurons”. In: *Neurocomputing* 48.1-4 (2002), pp. 17–37 (cit. on pp. 24, 95).

[187] Hesham Mostafa. “Supervised learning based on temporal coding in spiking neural networks”. In: *IEEE transactions on neural networks and learning systems* 29.7 (2017), pp. 3227–3235 (cit. on pp. 24, 95).

[188] Iulia M Comsa et al. “Temporal coding in spiking neural networks with alpha synaptic function”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 8529–8533 (cit. on pp. 24, 95).

[189] Saeed Reza Kheradpisheh et al. “Temporal backpropagation for spiking neural networks with one spike per neuron”. In: *International Journal of Neural Systems* 30.06 (2020), p. 2050027 (cit. on pp. 24, 95).

[190] Julian Göltz et al. “Fast and energy-efficient neuromorphic deep learning with first-spike times”. In: *Nature machine intelligence* 3.9 (2021), pp. 823–835 (cit. on pp. 24, 95).

[191] Yaoyu Zhu et al. “Training spiking neural networks with event-driven backpropagation”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 30528–30541 (cit. on pp. 24, 95).

[192] Steven K. Esser et al. “Convolutional networks for fast, energy-efficient neuromorphic computing”. In: *Proceedings of the National Academy of Sciences* 113.41 (2016), pp. 11441–11446 (cit. on pp. 24, 94).

[193] Eric Hunsberger et al. “Spiking deep networks with LIF neurons”. In: *arXiv preprint arXiv:1510.08829* (2015) (cit. on pp. 24, 94).

[194] Friedemann Zenke et al. “Superspike: Supervised learning in multilayer spiking neural networks”. In: *Neural computation* 30.6 (2018), pp. 1514–1541 (cit. on pp. 24, 53, 80, 94, 111).

[195] Jun Haeng Lee et al. “Training deep spiking neural networks using backpropagation”. In: *Frontiers in neuroscience* 10 (2016), p. 508 (cit. on pp. 24, 94, 101).

[196] Guillaume Bellec et al. “Long short-term memory and learning-to-learn in networks of spiking neurons”. In: *Advances in neural information processing systems* 31 (2018) (cit. on pp. 24, 91, 93, 94).

[197] Sumit B Shrestha et al. “Slayer: Spike layer error reassignment in time”. In: *Advances in neural information processing systems* 31 (2018) (cit. on pp. 24, 101).

[198] Emre O Neftci et al. “Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks”. In: *IEEE Signal Processing Magazine* 36.6 (2019), pp. 51–63 (cit. on pp. 24, 32, 52, 80, 94, 103).

[199] Daniel LK Yamins et al. “Using goal-driven deep learning models to understand sensory cortex”. In: *Nature neuroscience* 19.3 (2016), pp. 356–365 (cit. on pp. 25, 27, 28, 91, 130).

[200] Matteo Carandini et al. “Do we know what the early visual system does?” In: *Journal of Neuroscience* 25.46 (2005), pp. 10577–10597 (cit. on p. 27).

[201] Tim Gollisch et al. “Rapid neural coding in the retina with relative spike latencies”. In: *science* 319.5866 (2008), pp. 1108–1111 (cit. on pp. 28, 29, 60, 61, 68, 69, 73, 82).

[202] Na Young Jun et al. “Scene statistics and noise determine the relative arrangement of receptive field mosaics”. In: *Proceedings of the National Academy of Sciences* 118.39 (2021), e2105115118 (cit. on pp. 29, 60, 72, 73).

[203] Na Young Jun et al. “Efficient coding, channel capacity, and the emergence of retinal mosaics”. In: *Advances in neural information processing systems* 35 (2022), pp. 32311–32324 (cit. on pp. 29, 60, 72, 73).

[204] Yan Karklin et al. “Efficient coding of natural images with a population of noisy linear-nonlinear neurons”. In: *Advances in neural information processing systems* 24 (2011) (cit. on pp. 29, 60, 72, 73, 75).

[205] Suva Roy et al. “Inter-mosaic coordination of retinal receptive fields”. In: *Nature* 592.7854 (2021), pp. 409–413 (cit. on pp. 29, 60, 72, 73).

[206] Stephanie E Palmer et al. “Predictive information in a sensory population”. In: *Proceedings of the National Academy of Sciences* 112.22 (2015), pp. 6908–6913 (cit. on pp. 29, 60, 72, 74).

[207] Jared M Salisbury et al. “Optimal prediction in the retina and natural motion statistics”. In: *Journal of Statistical Physics* 162.5 (2016), pp. 1309–1323 (cit. on pp. 29, 60, 72, 73).

[208] Belle Liu et al. “Predictive encoding of motion begins in the primate retina”. In: *Nature neuroscience* 24.9 (2021), pp. 1280–1291 (cit. on pp. 29, 60, 72).

[209] Michael J Berry et al. “The retina as embodying predictions about the visual world”. In: *Predictions in the brain: Using our past to generate a future* (2011), p. 295 (cit. on pp. 29, 60, 72).

[210] Greg Schwartz et al. “Sophisticated temporal pattern recognition in retinal ganglion cells”. In: *Journal of neurophysiology* 99.4 (2008), pp. 1787–1798 (cit. on pp. 29, 60, 61, 70, 73, 75, 83).

[211] J Hans Van Hateren et al. “Independent component filters of natural images compared with simple cells in primary visual cortex”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 265.1394 (1998), pp. 359–366 (cit. on pp. 30, 42, 87).

[212] Christoph Kayser et al. “Extracting slow subspaces from natural videos leads to complex cells”. In: *International conference on artificial neural networks*. Springer. 2001, pp. 1075–1080 (cit. on pp. 30, 42).

[213] Wei Zhu et al. “A neuronal network model of primary visual cortex explains spatial frequency selectivity”. In: *Journal of computational neuroscience* 26 (2009), pp. 271–287 (cit. on pp. 30, 43).

[214] David McLaughlin et al. “A neuronal network model of macaque primary visual cortex (V1): Orientation selectivity and dynamics in the input layer 4Cα”. In: *Proceedings of the National Academy of Sciences* 97.14 (2000), pp. 8087–8092 (cit. on pp. 30, 43).

[215] David Cai et al. “An effective kinetic representation of fluctuation-driven neuronal networks with application to simple and complex cells in visual cortex”. In: *Proceedings of the National Academy of Sciences* 101.20 (2004), pp. 7757–7762 (cit. on pp. 30, 43).

[216] Logan Chariker et al. “A computational model of direction selectivity in Macaque V1 cortex based on dynamic differences between ON and OFF pathways”. In: *Journal of Neuroscience* 42.16 (2022), pp. 3365–3380 (cit. on pp. 30, 43).

[217] Roland Baddeley et al. “Responses of neurons in primary and inferior temporal visual cortices to natural scenes”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 264.1389 (1997), pp. 1775–1783 (cit. on pp. 34, 45).

[218] CR Legendy et al. “Bursts and recurrences of bursts in the spike trains of spontaneously active striate cortex neurons”. In: *Journal of neurophysiology* 53.4 (1985), pp. 926–939 (cit. on pp. 34, 45).

[219] Gary R Holt et al. “Comparison of discharge variability in vitro and in vivo in cat visual cortex neurons”. In: *Journal of neurophysiology* 75.5 (1996), pp. 1806–1814 (cit. on p. 34).

[220] William R Softky et al. “The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs”. In: *Journal of neuroscience* 13.1 (1993), pp. 334–350 (cit. on p. 34).

[221] Sergio Arroyo et al. “Correlation of synaptic inputs in the visual cortex of awake, behaving mice”. In: *Neuron* 99.6 (2018), pp. 1289–1301 (cit. on p. 34).

[222] Patrick E Williams et al. “A dynamic nonlinearity and spatial phase specificity in macaque V1 neurons”. In: *Journal of Neuroscience* 27.21 (2007), pp. 5706–5718 (cit. on p. 35).

[223] Izumi Ohzawa et al. “Encoding of binocular disparity by simple cells in the cat’s visual cortex”. In: *Journal of Neurophysiology* 75.5 (1996), pp. 1779–1805 (cit. on p. 35).

[224] Bernt C Skottun et al. “Classifying simple and complex cells on the basis of response modulation”. In: *Vision research* 31.7-8 (1991), pp. 1078–1086 (cit. on pp. 37, 57).

[225] Ferenc Mechler et al. “On the classification of simple and complex cells”. In: *Vision research* 42.8 (2002), pp. 1017–1033 (cit. on pp. 37, 57).

[226] Dario L Ringach et al. “Orientation selectivity in macaque V1: diversity and laminar dependence”. In: *Journal of neuroscience* 22.13 (2002), pp. 5639–5651 (cit. on p. 37).

[227] Gert Van den Bergh et al. “Receptive-field properties of V1 and V2 neurons in mice and macaque monkeys”. In: *Journal of Comparative Neurology* 518.11 (2010), pp. 2051–2070 (cit. on p. 37).

[228] Carl Holmgren et al. “Pyramidal cell communication within local networks in layer 2/3 of rat neocortex”. In: *The Journal of physiology* 551.1 (2003), pp. 139–153 (cit. on p. 41).

[229] Abhishek Banerjee et al. “Jointly reduced inhibition and excitation underlies circuit-wide changes in cortical processing in Rett syndrome”. In: *Proceedings of the National Academy of Sciences* 113.46 (2016), E7287–E7296 (cit. on p. 41).

[230] Hillel Adesnik. “Layer-specific excitation/inhibition balances during neuronal synchronization in the visual cortex”. In: *The Journal of physiology* 596.9 (2018), pp. 1639–1657 (cit. on p. 41).

[231] Ben DB Willmore et al. “Sparse coding in striate and extrastriate visual cortex”. In: *Journal of neurophysiology* 105.6 (2011), pp. 2907–2919 (cit. on p. 42).

[232] Pierre Comon. “Independent component analysis, a new concept?” In: *Signal processing* 36.3 (1994), pp. 287–314 (cit. on pp. 42, 119).

[233] Peter Földiák. “Learning invariance from transformation sequences”. In: *Neural computation* 3.2 (1991), pp. 194–200 (cit. on p. 42).

[234] Joel Zylberberg et al. “A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of V1 simple cell receptive fields”. In: *PLoS computational biology* 7.10 (2011), e1002250 (cit. on p. 43).

[235] Paul D King et al. “Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of V1”. In: *Journal of Neuroscience* 33.13 (2013), pp. 5475–5485 (cit. on p. 43).

[236] Rodney J Douglas et al. “A functional microcircuit for cat visual cortex.” In: *The Journal of physiology* 440.1 (1991), pp. 735–769 (cit. on p. 43).

[237] Rodney J Douglas et al. “Neuronal circuits of the neocortex”. In: *Annu. Rev. Neurosci.* 27 (2004), pp. 419–451 (cit. on p. 43).

[238] Vernon B Mountcastle. “Modality and topographic properties of single neurons of cat’s somatic sensory cortex”. In: *Journal of neurophysiology* 20.4 (1957), pp. 408–434 (cit. on p. 43).

[239] Jonathan C Horton et al. “The cortical column: a structure without a function”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1456 (2005), pp. 837–862 (cit. on p. 43).

[240] Torsten N Wiesel et al. “Effects of visual deprivation on morphology and physiology of cells in the cat’s lateral geniculate body”. In: *Journal of neurophysiology* 26.6 (1963), pp. 978–993 (cit. on p. 44).

[241] Stephen Grossberg. “Competitive learning: From interactive activation to adaptive resonance”. In: *Cognitive science* 11.1 (1987), pp. 23–63 (cit. on p. 44).

[242] Francis Crick et al. “The recent excitement about neural networks”. In: *Nature* 337.6203 (1989), pp. 129–132 (cit. on p. 44).

[243] Sen Song et al. “Highly nonrandom features of synaptic connectivity in local cortical circuits”. In: *PLoS biology* 3.3 (2005), e68 (cit. on p. 44).

[244] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005 (cit. on p. 44).

[245] Timothy P Lillicrap et al. “Random synaptic feedback weights support error backpropagation for deep learning”. In: *Nature communications* 7.1 (2016), pp. 1–10 (cit. on p. 44).

[246] Theodore H Moskovitz et al. “Feedback alignment in deep convolutional networks”. In: *arXiv preprint arXiv:1812.06488* (2018) (cit. on p. 44).

[247] Qianli Liao et al. “How important is weight symmetry in backpropagation?” In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 30. 1. 2016 (cit. on p. 44).

[248] Timothy P Lillicrap et al. “Backpropagation and the brain”. In: *Nature Reviews Neuroscience* 21.6 (2020), pp. 335–346 (cit. on p. 44).

[249] James CR Whittington et al. “Theories of error back-propagation in the brain”. In: *Trends in cognitive sciences* 23.3 (2019), pp. 235–250 (cit. on p. 44).

[250] Barbara Chapman et al. “Development of orientation selectivity in ferret visual cortex and effects of deprivation”. In: *Journal of Neuroscience* 13.12 (1993), pp. 5251–5262 (cit. on p. 44).

[251] Barbara Chapman et al. “Development of orientation preference maps in ferret primary visual cortex”. In: *Journal of Neuroscience* 16.20 (1996), pp. 6443–6453 (cit. on p. 44).

[252] Jianhua Cang et al. “Development of precise maps in visual cortex requires patterned spontaneous activity in the retina”. In: *Neuron* 48.5 (2005), pp. 797–809 (cit. on p. 44).

[253] Cory Pfeiffenberger et al. “Ephrin-As and patterned retinal activity act together in the development of topographic maps in the primary visual system”. In: *Journal of Neuroscience* 26.50 (2006), pp. 12873–12884 (cit. on p. 44).

[254] Andrew D Huberman et al. “Mechanisms underlying development of visual maps and receptive fields”. In: *Annu. Rev. Neurosci.* 31 (2008), pp. 479–509 (cit. on pp. 44, 122).

[255] Maria M Carrasco et al. “Visual experience is necessary for maintenance but not development of receptive fields in superior colliculus”. In: *Journal of neurophysiology* 94.3 (2005), pp. 1962–1970 (cit. on p. 44).

[256] Lupeng Wang et al. “Visual receptive field properties of neurons in the superficial superior colliculus of the mouse”. In: *Journal of Neuroscience* 30.49 (2010), pp. 16573–16584 (cit. on p. 44).

[257] Ye Li et al. “The development of direction selectivity in ferret visual cortex requires early visual experience”. In: *Nature neuroscience* 9.5 (2006), pp. 676–681 (cit. on p. 44).

[258] Giulio Matteucci et al. “Causal adaptation to visual input dynamics governs the development of complex cells in V1”. In: *biorxiv* (2019), p. 756668 (cit. on p. 44).

[259] Artur Luczak et al. “Neurons learn by predicting future activity”. In: *Nature machine intelligence* 4.1 (2022), pp. 62–72 (cit. on p. 45).

[260] Manu Srinath Halvagal et al. “The combination of Hebbian and predictive plasticity learns invariant object representations in deep sensory networks”. In: *Nature Neuroscience* (2023), pp. 1–10 (cit. on p. 45).

[261] Tristan G Heintz et al. “Opposite forms of adaptation in mouse visual cortex are controlled by distinct inhibitory microcircuits”. In: *Nature communications* 13.1 (2022), p. 1031 (cit. on p. 45).

[262] Lionel G Nowak et al. “Electrophysiological classes of cat primary visual cortical neurons in vivo as revealed by quantitative analyses”. In: *Journal of neurophysiology* 89.3 (2003), pp. 1541–1566 (cit. on p. 45).

[263] Jessica A Cardin et al. “Stimulus feature selectivity in excitatory and inhibitory neurons in primary visual cortex”. In: *Journal of Neuroscience* 27.39 (2007), pp. 10333–10344 (cit. on p. 45).

[264] Peter Dayan et al. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press, 2005 (cit. on pp. 45, 68).

[265] P Andersen et al. “Functional characteristics of unmyelinated fibres in the hippocampal cortex”. In: *Brain research* 144.1 (1978), pp. 11–18 (cit. on pp. 45, 94).

[266] Michael Avissar et al. “Refractoriness enhances temporal coding by auditory nerve fibers”. In: *Journal of Neuroscience* 33.18 (2013), pp. 7681–7690 (cit. on pp. 45, 94).

[267] Edmund T Rolls. “Absolute refractory period of neurons involved in MFB self-stimulation”. In: *Physiology & Behavior* 7.3 (1971), pp. 311–315 (cit. on pp. 45, 94).

[268] Julijana Gjorgjieva et al. “Computational implications of biophysical diversity and multiple timescales in neurons and synapses for circuit performance”. In: *Current opinion in neurobiology* 37 (2016), pp. 44–52 (cit. on p. 46).

[269] Fleur Zeldenrust et al. “Efficient and robust coding in heterogeneous recurrent networks”. In: *PLoS computational biology* 17.4 (2021), e1008673 (cit. on p. 46).

[270] Zachary P Kilpatrick et al. “Optimizing working memory with heterogeneity of recurrent cortical excitation”. In: *Journal of neuroscience* 33.48 (2013), pp. 18999–19011 (cit. on p. 46).

[271] David Attwell et al. “An energy budget for signaling in the grey matter of the brain”. In: *Journal of Cerebral Blood Flow & Metabolism* 21.10 (2001), pp. 1133–1145 (cit. on pp. 46, 126).

[272] Peter Lennie. “The cost of cortical computation”. In: *Current biology* 13.6 (2003), pp. 493–497 (cit. on pp. 46, 126).

[273] Jeremy E Niven. “Neuronal energy consumption: biophysics, efficiency and evolution”. In: *Current opinion in neurobiology* 41 (2016), pp. 129–135 (cit. on pp. 46, 126).

[274] Clare Howarth et al. “Updated energy budgets for neural computation in the neocortex and cerebellum”. In: *Journal of Cerebral Blood Flow & Metabolism* 32.7 (2012), pp. 1222–1232 (cit. on pp. 46, 47, 51, 126).

[275]Malcolm P Young et al. “Sparse population coding of faces in the inferotemporal cortex”. In: *Science* 256.5061 (1992), pp. 1327–1331 (cit. on p. 46).

[276]Joshua P Van Kleef et al. “Complex cell receptive fields: evidence for a hierarchical mechanism”. In: *The Journal of physiology* 588.18 (2010), pp. 3457–3470 (cit. on p. 46).

[277]Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on p. 46).

[278]Nicolas Brunel. “Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons”. In: *Journal of computational neuroscience* 8 (2000), pp. 183–208 (cit. on p. 47).

[279]Varun Padmanaban et al. “Clinical advances in photosensitive epilepsy”. In: *Brain Research* 1703 (2019), pp. 18–25 (cit. on p. 47).

[280]Arnold J Wilkins et al. “Physiology of human photosensitivity”. In: *Epilepsia* 45 (2004), pp. 7–13 (cit. on p. 47).

[281]Dora Hermes et al. “Gamma oscillations and photosensitive epilepsy”. In: *Current Biology* 27.9 (2017), R336–R338 (cit. on p. 47).

[282]Luke Taylor. “Natural movies”. In: (Oct. 2023) (cit. on p. 48).

[283]Bruno A Olshausen et al. “Sparse coding with an overcomplete basis set: A strategy employed by V1?” In: *Vision research* 37.23 (1997), pp. 3311–3325 (cit. on p. 48).

[284]Luke Taylor et al. “Improving deep learning with generic data augmentation”. In: *2018 IEEE symposium series on computational intelligence (SSCI)*. IEEE. 2018, pp. 1542–1547 (cit. on pp. 48, 76).

[285]Enquan Gao et al. “Parallel input channels to mouse primary visual cortex”. In: *Journal of Neuroscience* 30.17 (2010), pp. 5912–5926 (cit. on p. 49).

[286]D_A Baylor et al. “Two components of electrical dark noise in toad retinal rod outer segments.” In: *The Journal of physiology* 309.1 (1980), pp. 591–621 (cit. on pp. 50, 62, 76).

[287]Dmitri A Rusakov et al. “Noisy synaptic conductance: Bug or a feature?” In: *Trends in Neurosciences* 43.6 (2020), pp. 363–372 (cit. on p. 50).

[288]Peter Sterling et al. *Principles of neural design*. MIT press, 2015 (cit. on pp. 51, 70).

[289]Diederik P Kingma et al. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on pp. 53, 80, 112, 113).

[290]Stephen A Baccus et al. “Fast and slow contrast adaptation in retinal circuitry”. In: *Neuron* 36.5 (2002), pp. 909–919 (cit. on pp. 62, 74).

[291]Jonathan W Pillow et al. “Spatio-temporal correlations and visual signalling in a complete neuronal population”. In: *Nature* 454.7207 (2008), pp. 995–999 (cit. on pp. 62, 74).

[292]David I Vaney. “Patterns of neuronal coupling in the retina”. In: *Progress in retinal and eye research* 13.1 (1994), pp. 301–355 (cit. on pp. 62, 74).

[293]Anastasiia Vlasiuk et al. “Feedback from retinal ganglion cells to the inner retina”. In: *PLoS One* 16.7 (2021), e0254611 (cit. on pp. 62, 74).

[294]Mark CW van Rossum et al. “Effects of noise on the spike timing precision of retinal ganglion cells”. In: *Journal of neurophysiology* 89.5 (2003), pp. 2406–2419 (cit. on pp. 62, 77).

[295]Michael J Berry et al. “The structure and precision of retinal spike trains”. In: *Proceedings of the National Academy of Sciences* 94.10 (1997), pp. 5411–5416 (cit. on p. 62).

[296]Michael Berry et al. “Refractoriness and neural precision”. In: *Advances in neural information processing systems* 10 (1997) (cit. on p. 62).

[297]Julian Freedland et al. “Systematic reduction of the dimensionality of natural scenes allows accurate predictions of retinal ganglion cell spike outputs”. In: *Proceedings of the National Academy of Sciences* 119.46 (2022), e2121744119 (cit. on pp. 63, 64, 70, 86).

[298]Dennis Dacey. “20 origins of perception: Retinal ganglion cell diversity and the creation of parallel visual pathways”. In: (2004) (cit. on p. 65).

[299]Florentina Soto et al. “Efficient coding by midget and parasol ganglion cells in the human retina”. In: *Neuron* 107.4 (2020), pp. 656–666 (cit. on p. 65).

[300]Brendan J O’Brien et al. “Intrinsic physiological properties of cat retinal ganglion cells”. In: *The Journal of physiology* 538.3 (2002), pp. 787–802 (cit. on p. 65).

[301]Shai Sabbah et al. “A retinal code for motion along the gravitational and body axes”. In: *Nature* 546.7659 (2017), pp. 492–497 (cit. on p. 65).

[302]Vadim Maximov et al. “Direction selectivity in the goldfish tectum revisited”. In: *Annals of the New York Academy of Sciences* 1048.1 (2005), pp. 198–205 (cit. on p. 65).

[303]DB Bowling. “Light responses of ganglion cells in the retina of the turtle”. In: *The Journal of Physiology* 299.1 (1980), pp. 173–196 (cit. on p. 65).

[304]William R Levick. “Receptive fields and trigger features of ganglion cells in the visual streak of the rabbit’s retina”. In: *The Journal of physiology* 188.3 (1967), p. 285 (cit. on p. 65).

[305]Franklin R Amthor et al. “Morphologies of rabbit retinal ganglion cells with complex receptive fields”. In: *Journal of comparative neurology* 280.1 (1989), pp. 97–121 (cit. on p. 65).

[306]Ari Rosenberg et al. “The primate retina contains distinct types of Y-like ganglion cells”. In: *Journal of Neuroscience* 29.16 (2009), pp. 5048–5050 (cit. on p. 68).

[307]Christina Enroth-Cugell et al. “The contrast sensitivity of retinal ganglion cells of the cat”. In: *The Journal of physiology* 187.3 (1966), pp. 517–552 (cit. on p. 68).

[308]JH Caldwell et al. “New properties of rabbit retinal ganglion cells.” In: *The Journal of Physiology* 276.1 (1978), pp. 257–276 (cit. on p. 68).

[309]Dumitru Petrusca et al. “Identification and characterization of a Y-like primate retinal ganglion cell type”. In: *Journal of Neuroscience* 27.41 (2007), pp. 11019–11027 (cit. on p. 68).

[310] Jerome Y Lettvin et al. “What the frog’s eye tells the frog’s brain”. In: *Proceedings of the IRE* 47.11 (1959), pp. 1940–1951 (cit. on p. 68).

[311] Stephen A Baccus et al. “A retinal circuit that computes object motion”. In: *Journal of Neuroscience* 28.27 (2008), pp. 6807–6817 (cit. on p. 68).

[312] Garrett B Stanley. “Reading and writing the neural code”. In: *Nature neuroscience* 16.3 (2013), pp. 259–263 (cit. on p. 68).

[313] Stuart J Judge et al. “Implantation of magnetic search coils for measurement of eye position: an improved method.” In: *Vision research* (1980) (cit. on p. 68).

[314] Susana Martinez-Conde et al. “The role of fixational eye movements in visual perception”. In: *Nature reviews neuroscience* 5.3 (2004), pp. 229–240 (cit. on p. 68).

[315] Dimokratis Karamanlis et al. “Retinal encoding of natural scenes”. In: *Annual Review of Vision Science* 8 (2022), pp. 171–193 (cit. on p. 70).

[316] Ryszard Auksztulewicz et al. “Omission responses in local field potentials in rat auditory cortex”. In: *BMC biology* 21.1 (2023), p. 130 (cit. on p. 70).

[317] Oliver Schoppe et al. “Measuring the performance of neural models”. In: *Frontiers in computational neuroscience* 10 (2016), p. 10 (cit. on pp. 70, 88, 130).

[318] Martin Schrimpf et al. “Brain-score: Which artificial neural network for object recognition is most brain-like?” In: *BioRxiv* (2018), p. 407007 (cit. on pp. 70, 87).

[319] Chengxu Zhuang et al. “Unsupervised neural network models of the ventral visual stream”. In: *Proceedings of the National Academy of Sciences* 118.3 (2021), e2014196118 (cit. on pp. 70, 87).

[320] Colin Conwell et al. “Neural regression, representational similarity, model zoology & neural taskonomy at scale in rodent visual cortex”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 5590–5607 (cit. on pp. 72, 87, 91).

[321] Katherine R Storrs et al. “Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting”. In: *Journal of cognitive neuroscience* 33.10 (2021), pp. 2044–2064 (cit. on pp. 72, 87).

[322] Daniel Kerschensteiner. “Feature detection by retinal ganglion cells”. In: *Annual review of vision science* 8 (2022), pp. 135–169 (cit. on p. 73).

[323] JL Schnapf et al. “Spectral sensitivity of human cone photoreceptors”. In: *Nature* 325.6103 (1987), pp. 439–441 (cit. on p. 74).

[324] Kristian Donner. “Temporal vision: measures, mechanisms and meaning”. In: *Journal of Experimental Biology* 224.15 (2021), jeb222679 (cit. on pp. 74, 77).

[325] Robert M Shapley et al. “The effect of contrast on the transfer properties of cat retinal ganglion cells.” In: *The Journal of physiology* 285.1 (1978), pp. 275–298 (cit. on p. 74).

[326] Kerry J Kim et al. “Temporal contrast adaptation in the input and output signals of salamander retinal ganglion cells”. In: *Journal of Neuroscience* 21.1 (2001), pp. 287–299 (cit. on p. 74).

[327] Helene Marianne Schreyer et al. “Nonlinear spatial integration in retinal bipolar cells shapes the encoding of artificial and natural stimuli”. In: *Neuron* 109.10 (2021), pp. 1692–1706 (cit. on p. 74).

[328] Dennis M Dacey et al. “Colour coding in the primate retina: diverse cell types and cone-specific circuitry”. In: *Current opinion in neurobiology* 13.4 (2003), pp. 421–427 (cit. on p. 74).

[329] Tom Baden et al. “Understanding the retinal basis of vision across species”. In: *Nature Reviews Neuroscience* 21.1 (2020), pp. 5–20 (cit. on pp. 74, 123).

[330] Maria Neimark Geffen et al. “Retinal ganglion cells can rapidly change polarity from Off to On”. In: *PLoS biology* 5.3 (2007), e65 (cit. on p. 75).

[331] Greg Schwartz et al. “Synchronized firing among retinal ganglion cells signals motion reversal”. In: *Neuron* 55.6 (2007), pp. 958–969 (cit. on p. 75).

[332] EJ Chichilnisky et al. “Functional asymmetries in ON and OFF ganglion cells of primate retina”. In: *Journal of Neuroscience* 22.7 (2002), pp. 2737–2747 (cit. on p. 75).

[333] Qing Shi et al. “Functional characterization of retinal ganglion cells using tailored nonlinear modeling”. In: *Scientific reports* 9.1 (2019), p. 8713 (cit. on p. 75).

[334] Neha Parikh et al. “Saliency-based image processing for retinal prostheses”. In: *Journal of neural engineering* 7.1 (2010), p. 016006 (cit. on p. 75).

[335] Ce Feng et al. “Edge-preserving image decomposition based on saliency map”. In: *2014 7th International Congress on Image and Signal Processing*. IEEE. 2014, pp. 159–163 (cit. on p. 75).

[336] C Erickson-Davis et al. “What do blind people “see” with retinal prostheses”. In: *Observations and qualitative reports of epiretinal implant users (Neuroscience)* (2020) (cit. on p. 75).

[337] Tom Baden et al. “Spikes in retinal bipolar cells phase-lock to visual stimuli with millisecond precision”. In: *Current Biology* 21.22 (2011), pp. 1859–1869 (cit. on p. 77).

[338] Alma Ganczer et al. “Transiency of retinal ganglion cell action potential responses determined by PSTH time constant”. In: *Plos one* 12.9 (2017), e0183436 (cit. on p. 77).

[339] Simon B Laughlin et al. “The metabolic cost of neural information”. In: *Nature neuroscience* 1.1 (1998), pp. 36–41 (cit. on p. 80).

[340] Jeremy E Niven et al. “Energy limitation as a selective pressure on the evolution of sensory systems”. In: *Journal of Experimental Biology* 211.11 (2008), pp. 1792–1804 (cit. on p. 80).

[341] Kiersten Ruda et al. “The functional organization of retinal ganglion cell receptive fields across light levels”. In: *bioRxiv* (2022), pp. 2022–09 (cit. on p. 81).

[342] Adriana Olmos et al. “A biologically inspired algorithm for the recovery of shading and reflectance images”. In: *Perception* 33.12 (2004), pp. 1463–1473 (cit. on p. 83).

[343] Ilya Buldyrev et al. “Inhibitory mechanisms that generate centre and surround properties in ON and OFF brisk-sustained ganglion cells in the rabbit retina”. In: *The Journal of physiology* 591.1 (2013), pp. 303–325 (cit. on p. 85).

[344] Ian Van Der Linde et al. “DOVES: a database of visual eye movements”. In: *Spatial vision* 22.2 (2009), p. 161 (cit. on p. 86).

[345] Anne Hsu et al. “Quantifying variability in neural responses and its application for the validation of model predictions”. In: *Network: Computation in Neural Systems* 15.2 (2004), pp. 91–109 (cit. on p. 88).

[346] Nicol S Harper et al. “Network receptive field modeling reveals extensive integration and multi-feature selectivity in auditory cortical neurons”. In: *PLoS Computational Biology* 12.11 (2016), e1005113 (cit. on p. 91).

[347] Andrew Francl et al. “Deep neural network models of sound localization reveal how perception is adapted to real-world environments”. In: *Nature Human Behaviour* 6.1 (2022), pp. 111–133 (cit. on p. 91).

[348] Blake A Richards et al. “A deep learning framework for neuroscience”. In: *Nature Neuroscience* 22.11 (2019), pp. 1761–1770 (cit. on p. 91).

[349] Ryota Kobayashi et al. “Made-to-order spiking neuron model equipped with a multi-timescale adaptive threshold”. In: *Frontiers in Computational Neuroscience* (2009), p. 9 (cit. on p. 91).

[350] Cyrille Rossant et al. “Fitting neuron models to spike trains”. In: *Frontiers in Neuroscience* 5 (2011), p. 9 (cit. on p. 91).

[351] Skander Mensi et al. “Parameter extraction and classification of three cortical neuron types reveals two distinct adaptation mechanisms”. In: *Journal of Neurophysiology* 107.6 (2012), pp. 1756–1775 (cit. on p. 91).

[352] Christian Pozzorini et al. “Temporal whitening by power-law adaptation in neocortical neurons”. In: *Nature Neuroscience* 16.7 (2013), pp. 942–948 (cit. on p. 91).

[353] Sophie Denève et al. “Efficient codes and balanced networks”. In: *Nature Neuroscience* 19.3 (2016), pp. 375–382 (cit. on pp. 91, 104).

[354] Basile Confavreux et al. “A meta-learning approach to (re) discover plasticity rules that carve a desired function into a neural network”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 16398–16408 (cit. on p. 91).

[355] Lukas Braun et al. “Online learning of neural computations from sparse temporal feedback”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 16437–16450 (cit. on p. 91).

[356] Timo Wunderlich et al. “Demonstrating advantages of neuromorphic computation: a pilot study”. In: *Frontiers in Neuroscience* 13 (2019), p. 260 (cit. on pp. 91, 104, 125).

[357] Louis Lapicque. “Recherches quantitatives sur l’excitation electrique des nerfs traitee comme une polarization”. In: *Journal de physiologie et de pathologie générale* 9 (1907), pp. 620–635 (cit. on p. 91).

[358] Marie Levakova et al. “Adaptive integrate-and-fire model reproduces the dynamics of olfactory receptor neuron responses in a moth”. In: *Journal of the Royal Society Interface* 16.157 (2019), p. 20190246 (cit. on p. 91).

[359] Yuan Zeng et al. “Temporal learning with biologically fitted SNN models”. In: *International Conference on Neuromorphic Systems 2021*. 2021, pp. 1–8 (cit. on p. 91).

[360] Guillaume Bellec et al. “A solution to the learning dilemma for recurrent networks of spiking neurons”. In: *Nature Communications* 11.1 (2020), pp. 1–15 (cit. on pp. 91, 94, 95).

[361] Bojian Yin et al. “Effective and efficient computation with multiple-timescale spiking recurrent neural networks”. In: *International Conference on Neuromorphic Systems 2020*. 2020, pp. 1–8 (cit. on pp. 91, 94).

[362] Bojian Yin et al. “Accurate online training of dynamical spiking neural networks through Forward Propagation Through Time”. In: *Nature Machine Intelligence* (2023), pp. 1–10 (cit. on pp. 91, 95, 101).

[363] Nicolas Perez-Nieves et al. “Spiking Network Initialisation and Firing Rate Collapse”. In: *arXiv:2305.08879* (2023) (cit. on pp. 92, 94, 95).

[364] Bojian Yin et al. “Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks”. In: *Nature Machine Intelligence* 3.10 (2021), pp. 905–913 (cit. on pp. 93, 111).

[365] Jean-Sébastien Jouhanneau et al. “In vivo monosynaptic excitatory transmission between layer 2 cortical pyramidal neurons”. In: *Cell reports* 13.10 (2015), pp. 2098–2106 (cit. on p. 94).

[366] Marc-Oliver Gewaltig et al. “NEST (NEural Simulation Tool)”. In: *Scholarpedia* 2.4 (2007), p. 1430 (cit. on p. 95).

[367] Abigail Morrison et al. “Advancing the boundaries of high-connectivity network simulation with distributed computing”. In: *Neural computation* 17.8 (2005), pp. 1776–1801 (cit. on p. 95).

[368] Abigail Morrison et al. “Exact subthreshold integration with continuous spike times in discrete-time neural network simulations”. In: *Neural Computation* 19.1 (2007), pp. 47–79 (cit. on p. 95).

[369] Ronald J Williams et al. “A learning algorithm for continually running fully recurrent neural networks”. In: *Neural Computation* 1.2 (1989), pp. 270–280 (cit. on p. 95).

[370] Anil Kag et al. “Training recurrent neural networks via forward propagation through time”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 5189–5200 (cit. on p. 95).

[371] James M Murray. “Local online learning in recurrent networks with random feedback”. In: *Elife* 8 (2019), e43299 (cit. on p. 95).

[372] Malu Zhang et al. “Rectified linear postsynaptic potential function for backpropagation in deep spiking neural networks”. In: *IEEE Transactions on Neural Networks and Learning Systems* 33.5 (2021), pp. 1947–1958 (cit. on p. 95).

[373] Shibo Zhou et al. “Spiking Neural Networks with Single-Spike Temporal-Coded Neurons for Network Intrusion Detection”. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 8148–8155 (cit. on p. 95).

[374] Shibo Zhou et al. “Temporal-coded deep spiking neural network with easy training and robust performance”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 12. 2021, pp. 11143–11151 (cit. on p. 95).

[375] Christian Pehle et al. *Norse - A deep learning library for spiking neural networks*. Version 0.0.7. Documentation: https://norse.ai/docs/. Jan. 2021 (cit. on pp. 100, 112, 115).

[376] Wei Fang et al. “SpikingJelly: An open-source machine learning infrastructure platform for spike-based intelligence”. In: *Science Advances* 9.40 (2023), eadi1480 (cit. on pp. 100, 112, 115).

[377] Garrick Orchard et al. “Converting static image datasets to spiking neuromorphic datasets using saccades”. In: *Frontiers in Neuroscience* 9 (2015), p. 437 (cit. on pp. 101, 110).

[378] Benjamin Cramer et al. “The heidelberg spiking data sets for the systematic evaluation of spiking neural networks”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2020) (cit. on pp. 101, 110).

[379] Ed S Lein et al. “Genome-wide atlas of gene expression in the adult mouse brain”. In: *Nature* 445.7124 (2007), pp. 168–176 (cit. on pp. 102, 112).

[380] Michael J Hawrylycz et al. “An anatomically comprehensive atlas of the adult human brain transcriptome”. In: *Nature* 489.7416 (2012), pp. 391–399 (cit. on pp. 102, 112).

[381] Mark CW van Rossum. “A novel spike distance”. In: *Neural Computation* 13.4 (2001), pp. 751–763 (cit. on pp. 103, 113, 130).

[382] Jason K Eshraghian et al. “Training spiking neural networks using lessons from deep learning”. In: *Proceedings of the IEEE* (2023) (cit. on p. 103).

[383] Christian Pozzorini et al. “Automated high-throughput characterization of single neurons by means of simplified spiking models”. In: *PLoS Computational Biology* 11.6 (2015), e1004275 (cit. on p. 104).

[384] Jan Stuijt et al. “μBrain: An event-driven and fully synthesizable architecture for spiking neural networks”. In: *Frontiers in Neuroscience* (2021), p. 538 (cit. on p. 104).

[385] Yuming He et al. “A 28.2 μC neuromorphic sensing system featuring SNN-based near-sensor computation and event-driven body-channel communication for insertable cardiac monitoring”. In: *2021 IEEE Asian Solid-State Circuits Conference (A-SSCC)*. IEEE. 2021, pp. 1–3 (cit. on p. 104).

[386] Priyadarshini Panda et al. “Toward scalable, efficient, and accurate deep spiking neural networks with backward residual connections, stochastic softmax, and hybridization”. In: *Frontiers in Neuroscience* 14 (2020), p. 653 (cit. on p. 104).

[387] Sharan Chetlur et al. “cudnn: Efficient primitives for deep learning”. In: *arXiv:1410.0759* (2014) (cit. on p. 104).

[388]Adam Paszke et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on p. 109).

[389]James Bradbury et al. *JAX: composable transformations of Python+NumPy programs*. Version 0.2.5. 2018 (cit. on p. 109).

[390]Jacques Kaiser et al. “Synaptic plasticity dynamics for deep continuous local learning (DECOLLE)”. In: *Frontiers in Neuroscience* 14 (2020), p. 424 (cit. on p. 111).

[391]Alexandre Bittar et al. “A surrogate gradient spiking baseline for speech command recognition”. In: *Frontiers in Neuroscience* 16 (2022) (cit. on p. 111).

[392]David Harker. *Creating scientific controversies: uncertainty and bias in science and society*. Cambridge University Press, 2015 (cit. on p. 122).

[393]Mehdi Mirza et al. “Conditional generative adversarial nets”. In: *arXiv preprint arXiv:1411.1784* (2014) (cit. on p. 122).

[394]Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems* 27 (2014) (cit. on p. 122).

[395]Bowen Zhang et al. “Styleswin: Transformer-based gan for high-resolution image generation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11304–11314 (cit. on p. 122).

[396]Antonia Creswell et al. “Generative adversarial networks: An overview”. In: *IEEE signal processing magazine* 35.1 (2018), pp. 53–65 (cit. on p. 122).

[397]Prafulla Dhariwal et al. “Diffusion models beat gans on image synthesis”. In: *Advances in neural information processing systems* 34 (2021), pp. 8780–8794 (cit. on p. 122).

[398]Jascha Sohl-Dickstein et al. “Deep unsupervised learning using nonequilibrium thermodynamics”. In: *International conference on machine learning*. PMLR. 2015, pp. 2256–2265 (cit. on p. 122).

[399]Robin Rombach et al. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695 (cit. on p. 122).

[400]Andreas Blattmann et al. “Stable video diffusion: Scaling latent video diffusion models to large datasets”. In: *arXiv preprint arXiv:2311.15127* (2023) (cit. on p. 122).

[401]Martin Šálek et al. “Changes in home range sizes and population densities of carnivore species along the natural to urban habitat gradient”. In: *Mammal Review* 45.1 (2015), pp. 1–14 (cit. on p. 123).

[402]Michael Pecka et al. “Experience-dependent specialization of receptive field surround for selective coding of natural scenes”. In: *Neuron* 84.2 (2014), pp. 457–469 (cit. on p. 123).

[403]Dawn Finzi et al. “A single computational objective drives specialization of streams in visual cortex”. In: *bioRxiv* (2023), pp. 2023–12 (cit. on p. 124).

[404] Freddy Trinh et al. “Cochlear tuning characteristics arise from temporal prediction of natural sounds”. In: *bioRxiv* (2023), pp. 2023–10 (cit. on p. 124).

[405] Jennifer K Bizley et al. “Visual–auditory spatial processing in auditory cortical neurons”. In: *Brain research* 1242 (2008), pp. 24–36 (cit. on p. 124).

[406] Giuliano Iurilli et al. “Sound-driven synaptic inhibition in primary visual cortex”. In: *Neuron* 73.4 (2012), pp. 814–828 (cit. on p. 124).

[407] George-Rafael Samantsidis et al. “‘What I cannot create, I do not understand’: functionally validated synergism of metabolic and target site insecticide resistance”. In: *Proceedings of the Royal Society B* 287.1927 (2020), p. 20200838 (cit. on p. 125).

[408] Henry Markram. “The blue brain project”. In: *Nature Reviews Neuroscience* 7.2 (2006), pp. 153–160 (cit. on p. 125).

[409] Emma Strubell et al. “Energy and policy considerations for deep learning in NLP”. In: *arXiv preprint arXiv:1906.02243* (2019) (cit. on p. 126).

[410] Roy Schwartz et al. “Green ai”. In: *Communications of the ACM* 63.12 (2020), pp. 54–63 (cit. on p. 126).

[411] Theodosius Dobzhansky. “Nothing in biology makes sense except in the light of evolution”. In: *The american biology teacher* 75.2 (2013), pp. 87–91 (cit. on p. 127).

[412] Alan M Turing. *Computing machinery and intelligence*. Springer, 2009 (cit. on p. 128).

[413] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901 (cit. on p. 128).

[414] Anthony Zador et al. “Catalyzing next-generation artificial intelligence through neuroai”. In: *Nature communications* 14.1 (2023), p. 1597 (cit. on p. 128).

[415] Gordon M Shepherd. *Neurobiology*. Oxford University Press, 1988 (cit. on p. 128).

[416] Nicholas Roy et al. “From machine learning to robotics: challenges and opportunities for embodied intelligence”. In: *arXiv preprint arXiv:2110.15245* (2021) (cit. on p. 128).

[417] Richard S Sutton et al. *Reinforcement learning: An introduction*. MIT press, 2018 (cit. on p. 128).

[418] Thomas Bäck et al. “An overview of evolutionary algorithms for parameter optimization”. In: *Evolutionary computation* 1.1 (1993), pp. 1–23 (cit. on p. 128).

[419] Carsen Stringer et al. “Spontaneous behaviors drive multidimensional, brainwide activity”. In: *Science* 364.6437 (2019), eaav7893 (cit. on p. 128).

[420] Nikolaus Kriegeskorte et al. “Representational similarity analysis-connecting the branches of systems neuroscience”. In: *Frontiers in systems neuroscience* (2008), p. 4 (cit. on p. 129).

[421] Ari Morcos et al. “Insights on representational similarity in neural networks with canonical correlation”. In: *Advances in neural information processing systems* 31 (2018) (cit. on p. 129).

[422] Simon Kornblith et al. “Similarity of neural network representations revisited”. In: *International conference on machine learning*. PMLR. 2019, pp. 3519–3529 (cit. on p. 129).

[423] Alex H Williams et al. “Generalized shape metrics on neural representations”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 4738–4750 (cit. on p. 129).

[424] SueYeon Chung et al. “Neural population geometry: An approach for understanding biological and artificial neural networks”. In: *Current opinion in neurobiology* 70 (2021), pp. 137–144 (cit. on p. 129).

[425] Gopal Santhanam et al. “Factor-analysis methods for higher-performance neural prostheses”. In: *Journal of neurophysiology* 102.2 (2009), pp. 1315–1330 (cit. on p. 129).

[426] Mark M Churchland et al. “Stimulus onset quenches neural variability: a widespread cortical phenomenon”. In: *Nature neuroscience* 13.3 (2010), pp. 369–378 (cit. on p. 129).

[427] Benjamin R Cowley et al. “Stimulus-driven population activity patterns in macaque primary visual cortex”. In: *PLoS computational biology* 12.12 (2016), e1005185 (cit. on p. 129).

[428] Yuan Zhao et al. “Variational latent gaussian process for recovering single-trial dynamics from population spike trains”. In: *Neural computation* 29.5 (2017), pp. 1293–1316 (cit. on p. 130).

[429] Jonathan D Victor et al. “Nature and precision of temporal coding in visual cortex: a metric-space analysis”. In: *Journal of neurophysiology* 76.2 (1996), pp. 1310–1326 (cit. on p. 130).