Transformers and Cortical Waves: Encoders for Pulling In Context Across Time

Lyle Muller[1], Patricia S. Churchland[2], Terrence J. Sejnowski[3,4]
[1]Department of Mathematics, Western University
[2]Department of Philosophy, University of California at San Diego
[3]Computational Neurobiology Laboratory, Salk Institute for Biological Studies
[4]Department of Neurobiology, University of California at San Diego

Abstract: The capabilities of transformer networks such as ChatGPT and other Large Language Models (LLMs) have captured the world's attention. The crucial computational mechanism underlying their performance relies on transforming a complete input sequence – for example, all the words in a sentence – into a long "encoding vector" that allows transformers to learn long-range temporal dependencies in naturalistic sequences. Specifically, "self-attention" applied to this encoding vector enhances temporal context in transformers by computing associations between pairs of words in the input sequence. We suggest that waves of neural activity, traveling across single cortical regions or across multiple regions at the whole-brain scale, could implement a similar encoding principle. By encapsulating recent input history into a single spatial pattern at each moment in time, cortical waves may enable temporal context to be extracted from sequences of sensory inputs, the same computational principle used in transformers.

Highlights

- Transformer networks learn to predict long-range dependencies by concatenating input sequences into a long "encoding vector".

- Sensory inputs, however, arrive at the periphery one word and one saccade at a time, raising the question of how sensory cortex could implement a similar computational principle while processing incoming inputs in real time.

- We suggest that a computational role we have previously identified for cortical waves in sensory cortex may subserve the same underlying computational principle as the transformers' "encoding vector" to provide temporal context.

- Self-attention in transformers assigns association strengths between pairs of words that can be far apart in a sequence. Self-attention could be implemented in brains by interacting waves in the cortex and basal ganglia over a wide range of time scales.

Main Text

Transformer networks use encoding vectors to capture long-range dependencies

Cortical mechanisms for spatial context are well established, mediated by the long-range horizontal connections that give rise to non-classical receptive fields[1–4]. The contextual modulations of non-classical receptive fields allow spatial contrasts between inputs to neurons in neighboring columns to be directly encoded into the responses of their classical receptive fields. Equally important, however, is temporal context, which occurs when reading sequences of words in sentences. Consider these sentences in French and English:

“Le chat à traversé la rue parce qu'il faisait chaud.”

and

“The cat crossed the street because it was hot.”

Contextual information is indispensable in translating one sentence to the other. In the English sentence, “it” may refer to the cat, the street, or the weather more generally. In the French sentence, by contrast, “il” could not refer to the street but only to either *chat* or the weather. Decisions regarding the referent of “it” are determined by context – whether within the sentence, in the context of a neighboring sentence, or in the context of a whole paragraph. Experienced readers parse this effortlessly, based on their experience. Language is chock full of these context-dependencies, which can make for surprises, as when Groucho Marx announced “This morning I shot an elephant in my pajamas. How it got in my pajamas I don’t know.”

Over the past twenty years, interest in predicting sequences of words steadily increased in the field of natural language processing (NLP). Vast amounts of text had become available online, and there was great interest in learning models to predict and generate naturalistic language from this text. Neural networks that could generate sequences were an obvious choice for this task. Specifically, networks where model neurons are connected together in a dense web, called *recurrent neural networks* (RNNs), had long been known to be useful for generating sequences[5]. RNNs are distinct from feedforward models, such as convolutional neural networks[6,7], where neurons are organized into successive processing layers with no internal, intra-layer connections. Inputs to an RNN affect neurons within the network, which in turn propagate their activity to other neurons through a dense and loopy web of interconnections (Fig. 1, top left). An RNN receiving words in a sentence as inputs, one by one, can build up an internal state that can, in turn, capture dependencies within a natural language sequence[8]. Various techniques had been developed to

train connections between neurons in RNNs[9–11], and specialized model neurons formed the basis of networks such as long-short term memory (LSTM)[12] that provided important early advances in sequence prediction tasks. In applying RNNs to sequence prediction for natural language tasks, however, researchers began to realize the difficulties in training RNNs to pick up on long-range dependencies[13–15], which are critical for language prediction, such as with the context-dependent gist a human reader picks up with ease.

To address the problem of long-range dependencies, a mechanism was introduced by Bahdanau, Cho, and Bengio in 2014 to allow an RNN to learn which parts of a source sentence were the most valuable for making correct predictions[16]. This mechanism was called “attention”, in a loose association with the process of human attention to different sensory items. This mechanism allowed the network to “focus” on those pieces of the input sequence that would most effectively drive its internal state to produce the correct prediction. This mechanism for identifying predictively valuable segments proved very effective in helping RNNs learn natural language prediction tasks. Equipped with attention mechanisms, RNNs gained proficiency on sentence-level translation tasks, in part tackling the problem of long-range dependencies. A breakthrough in utilizing attention for long-range dependencies, however, was made in 2017 with the introduction of transformer networks[17]. The main innovation behind transformers was surprising: use only the attention mechanism itself, combined with relatively simple feedforward layers, to predict the next word. This simplified architecture provided the foundation for advances exhibited in the Generative Pre-Trained transformer (GPT) architecture[18], which led to the current large-language model (LLM) chat agents such as ChatGPT, LLaMa and PaLM 2[19].

In a language translation task, the transformer architecture is divided into an Encoder (which, in the above example, would process the sentence in French) and a Decoder (which would output the sentence in English) (Fig. 1, bottom left). The key step is the self-attention module, where a set of features learned for each word item interacts with features for the other items in the input sequence (Fig. 1, right). The size of this feature vector is called the “embedding dimension”. If the feature vector for one word matches with another, then the two words will have a strong link in the self-attention process. For example, in a given input sequence, “popcorn” and “ribosome” will be less strongly linked than “popcorn” and “movie”. Once this process is computed in parallel for all the words in the input sequence, the array of numbers storing the embedding vectors for all words in the input passes into a simple feedforward network. This is the basic function of one Encoder module in the transformer architecture. In practice, the self-attention mechanism is repeated many times within a single Encoder, a process that is called “multi-headed attention”. After several encoding layers with self-attention (Fig. 1), the resulting encoding vector then passes to a multi-layer Decoder, where the relationships the encoding vector has captured with self-attention aid in the correct prediction of the next output. During the training process of the

transformer, the connections that make up the self-attention and feedforward modules in the Encoder learn how to create a very high-dimensional encoding vector that can effectively drive the decoder to predict correct output sentences. The long vectors used to encode the inputs, together with the self-attention across the components of the vector, provide very broad context for making predictions. In essence, the temporal context of spoken words is represented in a transformer by the spatial context within the encoding vector.

The transformer architecture introduced the idea that the attention mechanism was all that is needed for language prediction tasks. As transformer networks scaled up, the encoding vectors they produced became surprisingly proficient at capturing the long-range dependencies in language that were previously difficult to capture with standard RNNs, which received words sequentially. The breakthrough represented by the transformer is that *the computation itself is simple* - the self-attention mechanism, iterated with feedforward networks, dramatically increased computational efficiency, meaning these networks could be scaled to larger and larger problems. As has become clear in GPT models, these networks can successfully produce coherent pages of text and, in some cases, display impressive generalization and reasoning[20]. The utility of this encoding vector, and its focus on capturing the relationships between words in an input sequence, is central to these advances in language prediction. Grasping what this encoding vector can teach us about computation more generally could advance our understanding of neural networks, both artificial and biological.

Capturing relationships by encoding a full input sequence in parallel

In learning and using context, do brains rely on anything like a transformer? The fundamental insight of the encoding vector is to capture *in parallel* predictively valuable relationships between all the items in an input sequence, rather than handling inputs one by one as with standard RNNs. Sensory inputs, however, arrive at our brains one word and one saccade at a time. This appears to pose a fundamental difficulty for brain systems to make use of transformer-style contextual information; that is, to capture relationships *in parallel* and to operate on sensory input, which is not only astronomically high-dimensional but also continuously arriving at the periphery. Yet somehow, brains do seem to have solved some version of the broad context problem. Could brain circuits implement an encoding strategy that is similar to that of transformers?

The tactic of encoding in parallel many elements in a temporal sequence may at first appear at odds with our current understanding of sensory processing in the brain. Regions in sensory cortex contain neurons that respond selectively to the onset of sensory inputs[21]. For example, the orientation of a bar of light may be encoded by the spike rate of an orientation-selective unit in primary visual cortex (V1), the tone of a sound may be encoded by the rate of a frequency-

selective neuron in primary auditory cortex (A1), or (to take an example from a cognitive system) the position of a rodent during navigation may be encoded by the spike rate of “place cells” in the hippocampus. Hubel and Wiesel, in their pioneering work on neuronal selectivity in visual cortex[22,23], established a model in which the sensory encoding is fully driven by the input, and where an individual neuron’s response properties are determined by just-now events in its receptive field. In this model of sensory encoding, trial-to-trial fluctuations that deviate from the average response expected from the receptive field are thought to be a product of noise[24], or to represent uncertainty[25,26]. True enough, feedback projections[27] from higher cortical areas may provide minimal context for incoming sensory inputs, or impose minimal bias on this input[28], and much of the feedback is not sensory but efference copy from motor commands[29,30]. These refinements are, however, exceedingly modest relative to the contextual capacities enabled by transformers. On the other hand, the powerful transformer strategy of encoding entire input sequences in parallel, along with their predictive relationships by virtue of a large encoding vector, appears ill-suited to real neural networks as characterized by the classical framework of sensory function. At first blush, the transformer strategy seems beyond the brain’s reach, assuming that neuronal encoding remains a fixed function of single input features, such as visual orientation or auditory pitch at one moment of time. That assumption, we suggest, may benefit from another look, in the light of new recordings from arrays of electrodes.

Recent research has demonstrated that rather than being based solely on just-now features of sensory input that are currently present, the selectivity of single neurons may take future, predicted features as well as past features into account. One recent study of place cells in bats noticed that, by shifting the present position of the animal forward in time in the data analysis, hippocampal place fields became sharper and new, well-formed place fields emerged[31]. This result suggests that, especially at the high speeds flown by bats relative to place field size in the hippocampus, selectivity may be enhanced by future, anticipated inputs, rather than restricted to present input stimuli. Anticipatory responses to moving stimuli have also been observed in the visual system, in the peripheral circuits of the salamander and rabbit retina[32] and, more recently, in monkey V1[33]. These results indicate that, in contrast to simply providing a reflection of present sensory input, the maps of the sensory and cognitive systems may be playing more dynamic roles in neural computation. If that is the case, the question is this: what dynamical structure in neurobiological networks could enable the type of highly structured encoding that occurs in transformers to handle contextual features across time?

Waves in single regions of visual cortex: parallel encoding of the recent past

The key computational insight of the transformer architecture is to encode the words in an input sequence in parallel, in a highly structured encoding that allows extracting meaningful

relationships. In terms of the visual system, we might consider a simple input sequence to be a series of points of light presented briefly at successive times. The key circuit element would be a way to link the activity patterns evoked by each stimulus, even after the initial activity pattern has subsided, to generate predictively valuable signals. How might neural circuitry be organized to achieve transformer-like richness?

Recent work has demonstrated, using large-scale optical imaging techniques and multi-electrode arrays, that small visual stimuli drive waves that propagate out from the point of input across the visual cortex[34,35]. These waves propagate at the same speed as the unmyelinated long-range horizontal fibers that connect neurons across cortical areas[36], traveling over a substantial portion of the map of visual space in tens of milliseconds. These unmyelinated horizontal fibers, which project many millimeters to connect neurons across an individual cortical region[37], are thus a candidate network mechanism underlying waves in single cortical regions, such as V1. Computer network models of spiking neural networks in the visual cortex create waves that match those observed in experimental recordings, especially when known distance-dependent axonal conduction delays are added[38]. In experimental recordings, these waves modulate neural excitability and thus the responses to incoming input[39,40]. Concordantly, large-scale spiking network models also reveal that the local balance of excitation and inhibition of neurons is modulated as the waves pass through local circuits, providing a mechanism for the modulation of neural excitability[38]. Further, experimental observations indicate that waves evoked in the awake state do not cross the boundaries between different cortical areas[34], in contrast to waves that occur in anesthetized animals[41,42]. This restriction suggests that waves occurring during normal, waking visual processing respect the retinotopic maps in individual regions of the visual system. With neurons thus organized, the waves could yield structured spatiotemporal patterns in response to a series of brief input stimuli. Another important feature of these waves, in both recordings and models, is that they are sparse: when a wave passes over a cortical region, only a small fraction of the neurons spike, which contrasts with the dense waves that occur, for example, during epileptic seizures. Unlike dense waves, sparse waves can propagate across single cortical regions, along long-range horizontal fibers, modulating but not interfering with the feedforward output.

Taken together, the experimental and modeling results raise the possibility that stimulus-evoked waves are not pointless doodads, but may play a significant computational role in sensory processing. In this case, however, what computation could this be? Waves of neural activity traveling over the retinotopic map seem at first to be inconsistent with the standard framework for sensory processing. Within this canonical framework, models of the visual system generally consider feedforward inputs from the retina, with precise retinotopic projections from one layer of neurons to the next, to process incoming visual inputs through successively elaborated receptive

field selectivity[43]. If a single point stimulus can evoke a wave that travels over a large part of an individual visual area, it appears this wave could potentially disrupt the processing of other sensory stimuli as it propagates. In this way, stimulus-evoked waves may appear incompatible with the classical conception of precise retinotopic maps and retinotopic projections.

We previously suggested that, in contrast to this initial impression, waves may be able to have a clear computational role in processing visual input by providing temporal context[35]. The key property underlying this role is that waves traveling radially outward from a point of input could encode both *where* (in retinotopic space) and *when* a stimulus occurred. For example, with a small, punctate input that evokes a wave (Fig. 2, top), a decoder could tell *where* the input occurred, by using the center point of the wave on the retinotopic map, and *when* it occurred, by using the distance from the center and the fact that these waves travel at a specific range of speeds. In a case with multiple stimuli, such as multiple inputs presented in a sequence, the spatiotemporal pattern of waves traveling along the horizontal fiber network evoked by the sequence could enable decoding of both the sequence of stimulus positions and their onset times (Fig. 2, bottom). In this way, waves could provide a mechanism for the sensory cortex to encode stimuli in the recent past, in a highly structured manner that enables extracting meaningful relationships across space and time.

As waves of activity spread laterally within the cortex, they influence the spiking activity of neighboring neurons after a delay caused by conduction through unmyelinated long-range horizontal axons (Fig. 3a). As the wave progresses through the tissue, it influences the spiking activity of more distant neurons after longer time delays, as visualized in Fig. 3b as an expanding spacetime cone. In a natural scene, many neurons will be activated, creating interference patterns between all the expanding waves. This is reminiscent of a hologram formed by spatial interference fringes, which contain all the information needed to recreate a 3D object when illuminated by a coherent light source. In the cortex, spatial input is mixed with temporal delays to create a spacetime representation, containing information needed to recover the spatial and temporal history of the sensory inputs.

Following our introduction of this framework for computations with waves in sensory cortex, it has been increasingly appreciated that waves can provide spatial structure for neural population activity that could be useful for computation[44], and recent theoretical work has demonstrated that waves can indeed enable this conjunctive encoding of where and when a stimulus occurred[45]. Further, networks generating waves can learn to make simple predictions of naturalistic movie inputs[45], and waves in recurrent neural networks can drive elementary computer vision tasks such as image segmentation[46]. Adding to the computational profile, traveling waves in recurrent neural network models can be rapidly trained to remember sequences and extract information from them

over long time intervals[47]. These results demonstrate that waves can play computational roles in the visual system, making short-term predictions of dynamic sensory input possible by linking space and time directly on the retinotopic map. Moreover, and this was a surprise, with waves as an organizational principle, encoding precision was not sacrificed, but augmented.

There are key reasons to suggest that this connection between waves and transformer networks may be more than a superficial resemblance. One intriguing point is that the specific way transformers keep track of the order in an input sequence - called *positional encoding* - bears similarity to a discrete Fourier basis, a systematic ordering of sinusoids of different frequencies[17]. In recent work, we have found a specific mathematical link between this discrete Fourier basis and spatiotemporal dynamics, from generation of waves to complex spatiotemporal patterns, in networks of nonlinear oscillators that can model neural populations[48]. Capitalizing on this result, future theoretical work may be able to identify significant links between the spatiotemporal dynamics of neural populations and the computational principles underlying the transformer architectures that have proven highly successful in predicting natural language input. Additionally, new RNN architectures may be inspired that both agree with the main architectural features of single cortical regions and can utilize principles learned from transformers.

Waves and transformers: bringing the encodings together

The potential similarity between transformers and waves is that they may be tapping into the same computational principle, albeit with somewhat different physical mechanisms: by processing inputs in parallel, using a highly structured encoding, transformer networks and cortical waves may enable extracting meaningful relationships from these sequences. In the case of the transformer, the long encoding vector contains the attention mechanism that enables capturing the long-range dependencies critical for natural language processing. In the case of waves in visual cortex, the highly structured spatiotemporal patterns, earlier tagged as sparse, may enable encoding temporal relationships directly onto populations of neurons over the retinotopic map, facilitating flexible storage of the recent past in a way that enables extracting the temporal relationships from the spatial map.

This potential similarity between the computational principle underlying both waves and transformers may be able to shed light on the function of waves in single regions of visual cortex. Since the introduction of the feedforward model of the visual system by Hubel and Wiesel[22,23], and its refinement through successive network implementations[43,49], we have implicitly thought that the visual cortex contains some veridical image of sensory input, just filtered in some way by the receptive field selectivity in each area. *Waves in single regions of visual cortex, however, indicate that input encoding in the visual system may be much more sophisticated, as a local*

population of neurons can influence networks far across the retinotopic map in a highly structured manner. Encoding long input sequences in parallel provides transformers with an advanced capacity to extract meaningful relationships in natural sensory input. This stunning but conceptually simple achievement suggests that roughly comparable encoding can be conveniently implemented by nervous systems to extract relationships across input sequences. Although the complex and fluctuating activity patterns of the visual system -- from spontaneous activity in the absence of visual input, to neural responses to simple stimuli, and finally to dynamics in response to naturalistic movie input -- may at first appear to be noisy, it remains possible that these fluctuations are not mere noise. Instead, they may be a reflection of computations that extract meaningful relationships from the continuous stream of visual input and create short-term predictions of incoming stimuli. Although we have focused on neural data, the psychological experiments regarding expectation effects, such as priming, show that there are clear speed and accuracy benefits in making accurate predictions of upcoming sensory inputs, for example, in auditory anticipations driven by experience[50,51].

Self-attention

As shown in Fig. 1, the input to each layer projects to self-attention, which then is combined with the feedforward projection. So-called self-attention is a novel addition to deep feedforward networks. The foundational paper for LLMs was entitled “Attention is All You Need,” emphasizing its importance[17]. Without self-attention, a transformer would be a conventional feedforward network with limited capabilities. Here is how GPT-4 described self-attention: “Imagine you're reading a book and come across a sentence that refers to something mentioned a few pages back. You might flip back to remind yourself. Self-attention allows the model to look at other words in the sentence to better understand the current word.” This is different from how “attention” is used in neuroscience, which typically is focused on single sensory items, but self-attention could be considered a generalization of attention that links items across time.

This raises the question of how “self-attention” could be implemented in brains. State space models have recently replaced the matrix self-attention mechanism, and are much more computationally efficient[52]. The state space models that generate traveling waves are in the same class and can also implement efficient self-attention. We therefore propose that cortical waves on different spatial and time scales are responsible for self-attention. Within cortical areas, self-attention could be learned with synaptic plasticity on the long-range horizontal connections, which can extend many millimeters across the cortex with time delays in the 100 ms range.

How could events originating from separate cortical areas be linked together? Two major brain regions have reciprocal loops with the cortex that are candidates, the basal ganglia and the

cerebellum. The reciprocal loops between the cortex and the basal ganglia are topographically organized (Fig. 4). However, it takes 100 ms to go around the loop, long after the feedforward cortical inputs that generated the waves have passed. This may be mitigated by the essential feature of the spacetime population code, which is extended in time, perhaps for many seconds, at the higher levels of processing in the prefrontal cortex. Though speculative, this is an interesting new way to view the basal ganglia,[53] which is known to be involved in learning and generating sequences of actions to achieve goals. Regarding the cerebellum, temporal context is essential for coarticulation in speech. The motor cortex also supports waves[34,35], and transformer-style self-attention could facilitate coordinating muscular contractions by extending motor representations over time as spatial patterns of activity.

Scaling up spacetime cortical codes more broadly requires distant cortical regions to interact on a larger scale. When meeting someone for the first time, the next day you can recall their face, the sound of their voice, their handshake, where you met, and, perhaps, their name, represented in neurons distributed throughout the cortex. Although the basal ganglia receives converging inputs from nearby cortical areas, it cannot learn self-attention globally across the entire cortex. Cortical waves occur at the macroscopic, whole-brain scale during sleep. Sleep spindles are bursts of spikes at 10-14 Hertz lasting 1-2 seconds[54], which occur thousands of times during the night and are involved in memory consolidation[55–57]. Spindles accompany replay of temporal sequences stored in the hippocampus from previous experiences[58–60]. Spindles can be preceded by bursts of theta activity in the hippocampus[61] and aid in transfer of newly formed engrams from the hippocampus to neocortex for long-term storage[62,63]. In recordings from human cortex, spindle waves encompass all of cortex in a rotating pattern[64]. Rotating waves could implement global self-attention by linking together aspects of experiences that have broad representations throughout the cortex. The long-range association fibers in the white matter link all these brain areas, with time delays in the 10s of millisecond range. Following sleep, distant parts of the cortex could be linked by global spacetime codes.

This strategy may be particularly important in infants, who sleep 12-16 hours daily and need to organize complex sensorimotor experiences[65]. Waves occur in the retina before the eyes open and are needed to sharpen the topographic projections in the thalamus[66]. Interestingly, these retinal waves are also important for the development of the long-range horizontal connections in the cortex[67]. Cortical waves occur spontaneously early in brain development and may be equally important in preparing the cortex for the flood of sensorimotor signals that babies experience.

Experimental Predictions

We predict that cortical waves propagate information about the local history of previous inputs through the long-range horizontal projections in the cortex. These unmyelinated fibers are relatively slow compared with the fast feedforward myelinated fibers[68]. Evolution, ever wily, may have used these slower conduction speeds to advantage in sensory cortex, exploiting the larger time window to preserve longer traces of past sensory stimuli. These inputs modulate, after a delay, the membrane potentials of distant neurons and, in so doing, modulate their spiking. This is the essence of what we mean by *temporal context.*

As strong feedforward sensory input arrives at a cortical neuron, its response will be mixed with information about inputs from distant spatial locations at previous times (Fig. 3b). This can be experimentally tested by reconstructing previous sensory inputs from current activity in large populations of neurons. If information about the past is being encoded along with current information in the same spike trains, methods from machine learning should be able to reconstruct past inputs. Some evidence indicates that multiple sequential inputs may be decoded from EEG recordings in humans in a one-second window[69]. Similar experiments could be performed leveraging multielectrode arrays or optical recordings.

More work needs to be done to understand whether cortical waves can provide fundamental insight into the visual system, whose well-known trial-to-trial variability in single neuron responses to identical visual stimuli[70–73] may in part be due to spontaneous waves[74]. Studies in anesthetized animals have provided evidence that spontaneous activity shapes stimulus-evoked responses[75,76], and recent work in awake animals has found that spontaneous activity modulates both the excitability of local neural populations in visual cortex and perceptual sensitivity in a challenging visual detection task[40]. Previous work has found, in addition, that stimulus-evoked activity can actively shape spontaneous activity: optical imaging experiments in the visual cortex of the anesthetized rat found that, when waves were evoked by a repeated visual stimulus, spontaneous waves following stimulation became more similar to the evoked[77]. Future computational and mathematical modeling work may be able to illuminate the complex interaction between spontaneous and stimulus-evoked waves in the visual cortex during normal perceptual processing in awake animals, and provide specific, testable hypotheses for future experiments with large-scale recordings.

Converging information from thousands of lateral inputs onto a single cortical neuron should, by the central limit theorem in probability theory, produce fluctuations that are Gaussian distributed, regardless of the statistical distribution of the inputs. Consequently, the correlations in the fluctuations of the membrane potentials driven by these weaker inputs are governed by linear dynamics, which preserves information[78]. Large-scale optical recordings of membrane potentials

from populations of neurons using voltage-sensitive dyes could probe the information content of these fluctuations.

Discussion

Population coding by waves traveling over topographic maps may not be as intuitive at first as the traditional conceptual framework for coding with receptive fields in sensory maps. Both, however, might be relevant to understanding neuronal function, specifically as different levels of description of sensory systems. Receptive fields are measured under carefully controlled conditions, controlling sensory stimuli to be as nearly identical as possible and then obtained by averaging neural responses to tens or hundreds of stimulus presentations. Receptive fields thus capture information about a neuron's responses *on average* to features of sensory stimuli, with trial-to-trial fluctuations about this average thought to be a product of noise. This framework for neural coding has been highly successful in understanding responses to repeated visual stimuli and in understanding the elaboration of neuronal selectivity across the visual system. However, when faced with an incoming stream of complex whole-field visual inputs, extracting meaningful relationships across time, as occurs in transformer networks, may provide the visual system an important advantage in predicting upcoming inputs and preparing behavioral responses.

An important difference between the cortical architecture and the transformer architecture is that transformers are feedforward, but cortical circuits are recurrent. The activity dynamics in recurrent networks is an essential mechanism for generating traveling waves[47]. Multiple sequences can be learned and replayed in the large, sparsely connected recurrent networks found in the upper layers of the cortex and the hippocampus[79]. One consequence is that fewer layers of recurrent networks may be needed to achieve the same levels of performance in transformers. Another difference is that cortical layers receive feedback projections from higher layers, along with lateral feedback within layers. The synaptic plasticity in these pathways could be responsible for developing dynamic internal representations of sequences rather than the static feedforward representations learned in transformers. The units in transformers can have mixed excitatory and inhibitory outputs. In the cortex, neurons are primarily either excitatory or inhibitory, some of which, like inhibitory parvalbumin basket cells, are specialized for generating oscillations, and inhibitory somatostatin interneurons that may coordinate working memory[80]. Evolution could have found alternative architectures to those in transformers for solving the same computational problems.

Finally, beyond the “mesoscopic” scale of individual cortical regions, there may be possible roles for the transformer architecture at the “macroscopic”, whole-brain scale. Transformers are autoregressive: the feedforward network generates one word at a time, which is then looped back and added to the previous prediction (cf. “output decoder prediction for next step” (Fig. 1, bottom left panel). Similarly, the loop between the cortex and basal ganglia generates sequences of actions, including sequences of words in humans. The output from the cortico-basal ganglia loop also controls the sequence of saccades you make as you read this sentence[53]. The self-attention mechanism in transformers could be implemented within the basal ganglia (Fig. 4), which integrates information from all areas of the cortex, helping to stitch together relationships between words, actions, and intentions across time. This raises the possibility that inputs to the cortex from the basal ganglia may also generate waves of neural activity at the whole-brain scale, involving structured sequences of activation between multiple cortical areas orchestrating complex actions. Finally, sleep spindles are waves across the entire cortex, creating a possible basis for the global workspace by stitching together disparate elements of sensory experience into a unified engram through known mechanisms of synaptic plasticity.

Donald Hebb is well known for his prediction that the coincidence of presynaptic and postsynaptic activity triggers synaptic plasticity[81]. Less well-known is why Hebb proposed this learning algorithm. He was concerned with the problem of how a transient sensory input could continue to circulate in the brain[82]. He proposed this learning algorithm as a mechanism for creating chains of neurons to maintain a trace of the sensory input, which he called a “phase sequence”. The modern version of Hebbian synaptic plasticity, spike-timing dependent plasticity (STDP), in which pre before post leads to LTP and post before pre produces LTD, enhances traveling waves[83]. Our proposal for the function of cortical waves, which could be called phase sequences, shows how Hebb’s theories might be realized.

Conclusion

Traveling waves are ubiquitous in the cerebral cortex, having now been observed in sensory, motor, and cognitive regions and propagating at many different time and spatial scales[34,35]. We offer a possible function for waves in single cortical regions in providing temporal context for sequences of sensory inputs, such as words in the auditory cortex and saccadic fixations during reading in the visual cortex. Waves mix old information with new information to provide a new type of spacetime population code. This form of encoding has computational advantages similar to those found in the transformer architecture of LLMs, which maps temporal sequences into a long input vector. Evolution may have found an alternative method to achieve the same functionality, taking advantage of cortical dynamics in recurrent networks. At any moment of time,

the dynamic spacetime population code, the equivalent of the long input vector in a transformer, is projected to the next layer in the cortical hierarchy, engendering even longer spacetime codes.

Throughout biology, evolution has repeatedly exploited the physics of oscillators to make extensive use of waves in microscopic systems on a wide range of time scales, from the rotation of flagella, to whisking, digesting, egg-laying, and swimming[84]. Our hypothesis is that yet another evolutionary adaptation deploys waves of neural activity, specially suited to sparse spiking dynamics in cortex, in mammalian brains to support spacetime coding. Brains rule the waves.

Acknowledgements

The authors thank Andrew Keller and Max Welling for inspiring discussions on waves in trained recurrent neural networks, and Arjun Karuvally and Hava Siegelmann for discussing their models showing that traveling waves in recurrent networks can implement variable binding, an important computational primitive. April Benasich helped us analyze sleep spindle waves in infants and how these waves could influence brain development. Luisa Liboni provided helpful comments on the manuscript. This research was supported by grants from ONR N000014-23-1-2069, NIH DA055668-01, MH132644-01, NSF NeuroNex DBI-2014862, the Swartz Foundation, BrainsCAN at Western University through the Canada First Research Excellence Fund (CFREF), NSF/CIHR NeuroNex DBI-2015276, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Western Academy for Advanced Research, NIH U01-NS131914, and NIH R01-EY028723.

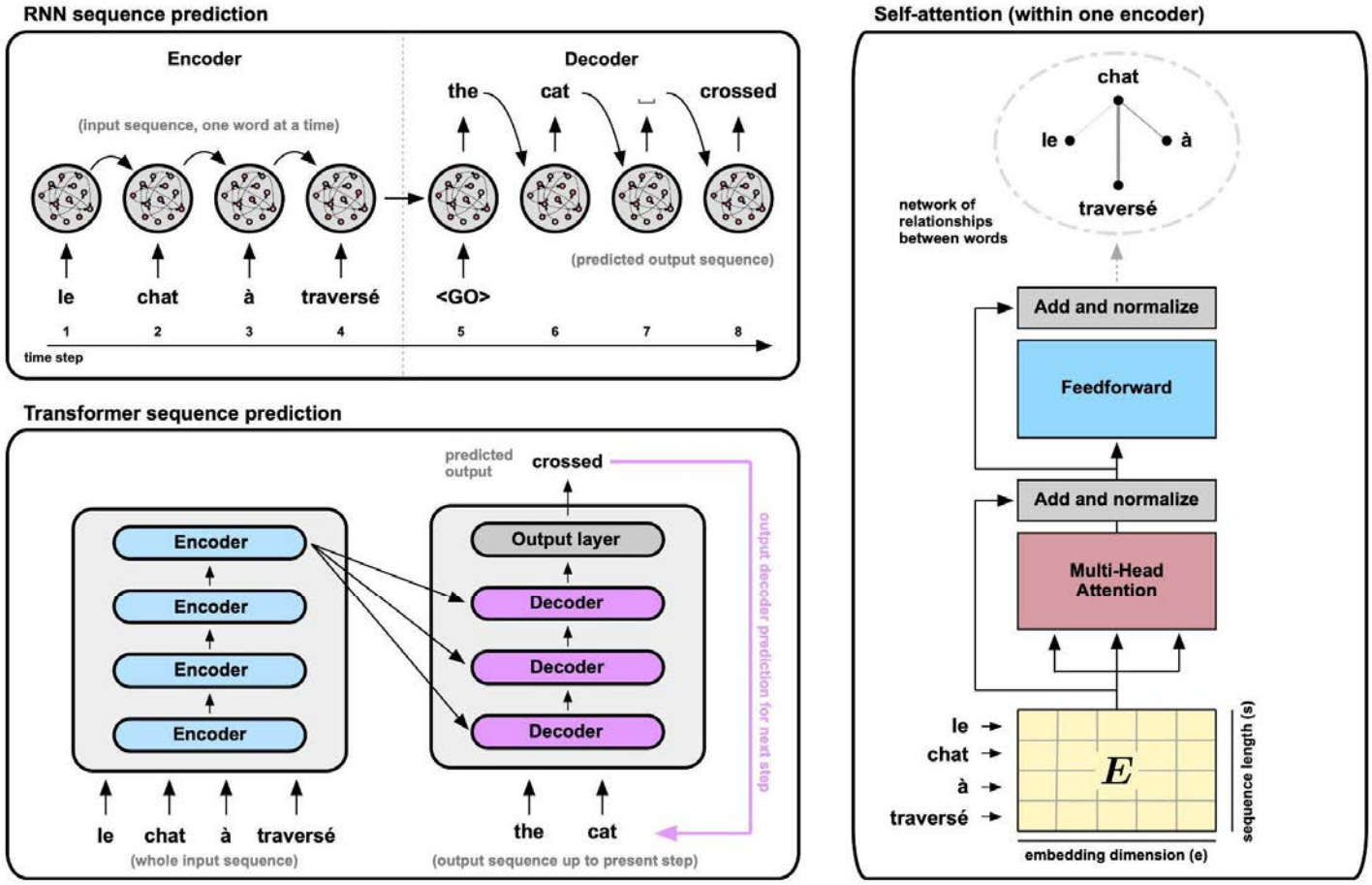

Figure 1. Recurrent neural network architecture, transformer networks, and self-attention in language prediction tasks. (Top left) RNNs for language prediction tasks take each word in a sequence as input, one at a time. The inputs are processed by the RNN, whose state passes from time step to time step (horizontal arrows) in order to build up a representation of the sequence. After the entire input sequence is fed into the RNN, a "go" signal cues the network to generate the output sequence, again one word at a time. Each generated output word is fed back into the RNN to recursively generate the output sequence. **(Bottom left)** Instead of taking each input one at a time, transformers take in the whole input sequence, which is processed through a series of Encoders. GPT-4 has a context length of 128k tokens (about 240 pages at 400 words per page). The output of the last Encoder is then an input to the Attention mechanism in the Decoder modules. The output of the complete Encoder-Decoder is the predicted next word in the sequence. This prediction is then appended to the input to the decoder to start the prediction for the next step. **(Right)** Within a single layer of the Encoder and Decoder, the sequence encoding (E) is passed to a multi-head attention module. The result of this calculation is the self-attention score, which is added to its input and passed on to a traditional feedforward layer. This self-attention mechanism enables the data-driven discovery of the network of relationships between words in the input sequence (top).

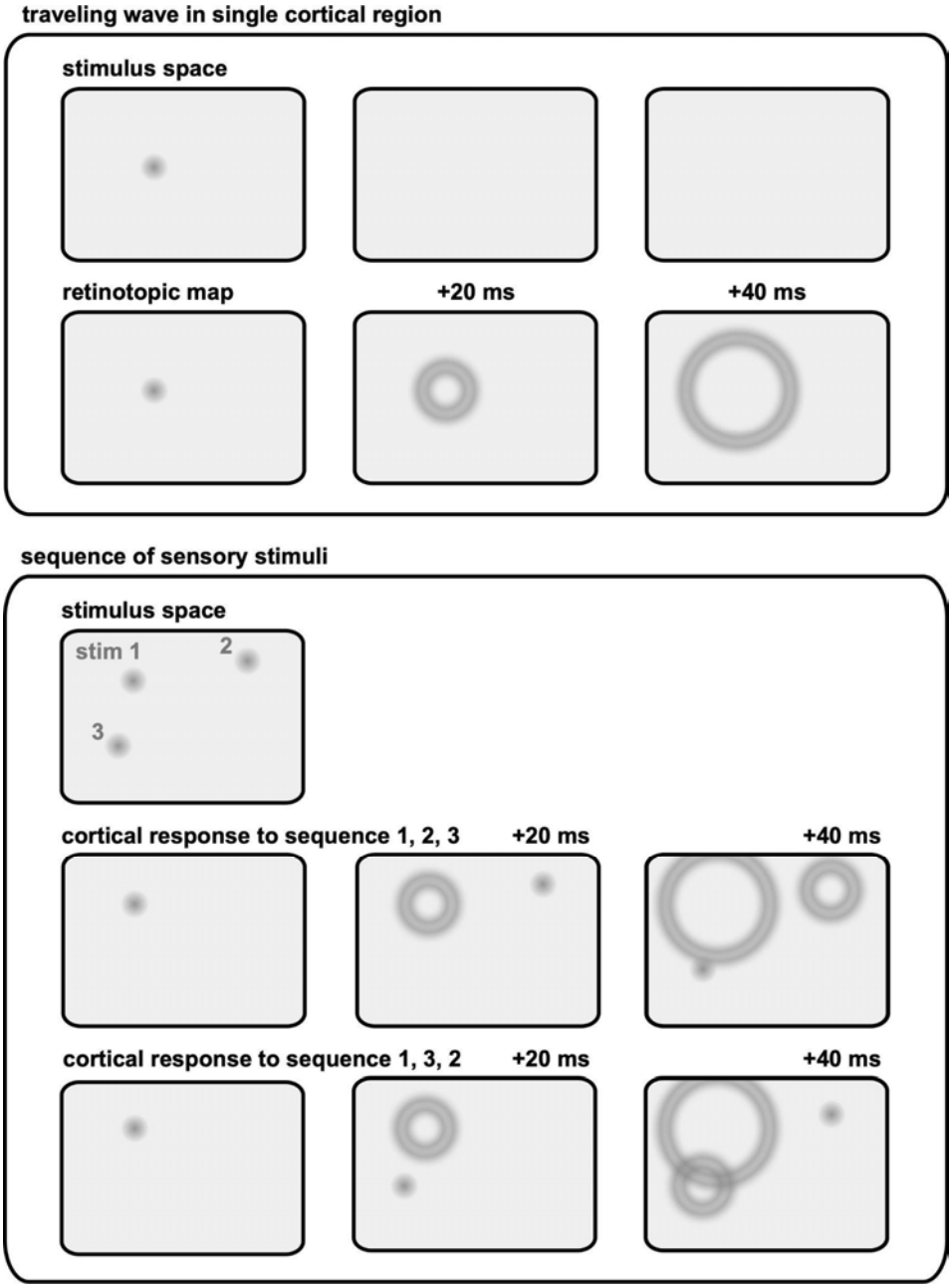

Figure 2. Waves occur in single cortical regions and responses to multiple stimuli. (top box) Recent studies in awake, behaving animals have found that small, punctate visual stimuli (stimulus space, top row) can create waves of activity that propagate outward from the point of feedforward input (retinotopic map), similar to ripples in a pond created by dropping a pebble. **(bottom box)** In the case of three visual stimuli (stimulus space, top row), a specific temporal order of presentation (stimulus 1, then 2, then 3) can create one pattern of waves (cortical response to sequence 1, 2, 3), while another order of presentation (stimulus 1, then 3, then 2) can create a different spatiotemporal pattern (cortical response to sequence 1, 3, 2).

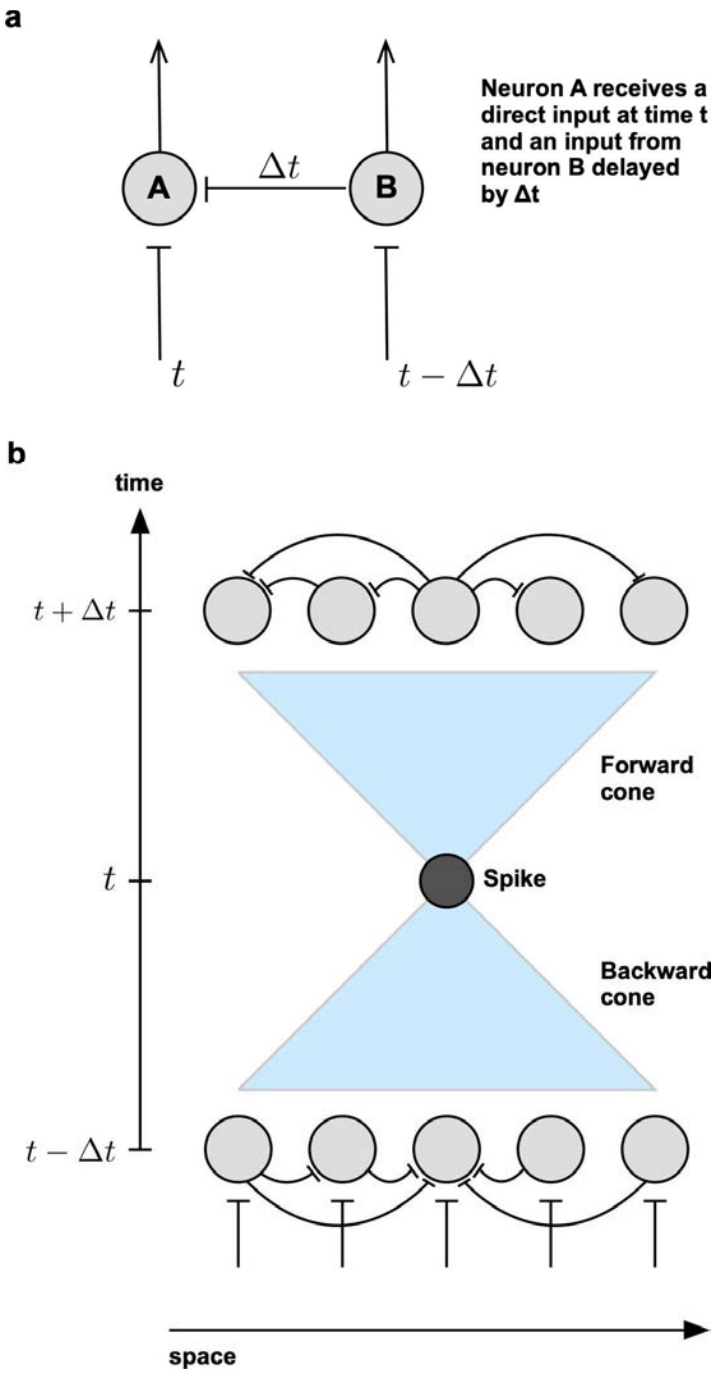

Figure 3. Time delays between laterally interacting neurons create a spacetime population code. (a) Neuron A receives a direct input at time *t* and an input from neuron B delayed by *Δt*. **(b)** The response of a spiking neuron (dark gray circle at the intersection of the two blue triangles) is influenced by the activity of all the interacting neurons in the backward spacetime cone (blue triangle from *t* - *Δt* to *t*), whose shape is structured by the temporal delays in the network. The spike of the neuron at time *t* influences, in turn, a population of interacting neurons within the forward spacetime cone (blue triangle from *t* to *t* + *Δt*). The cone extends back in time to include all inputs to the central neuron.

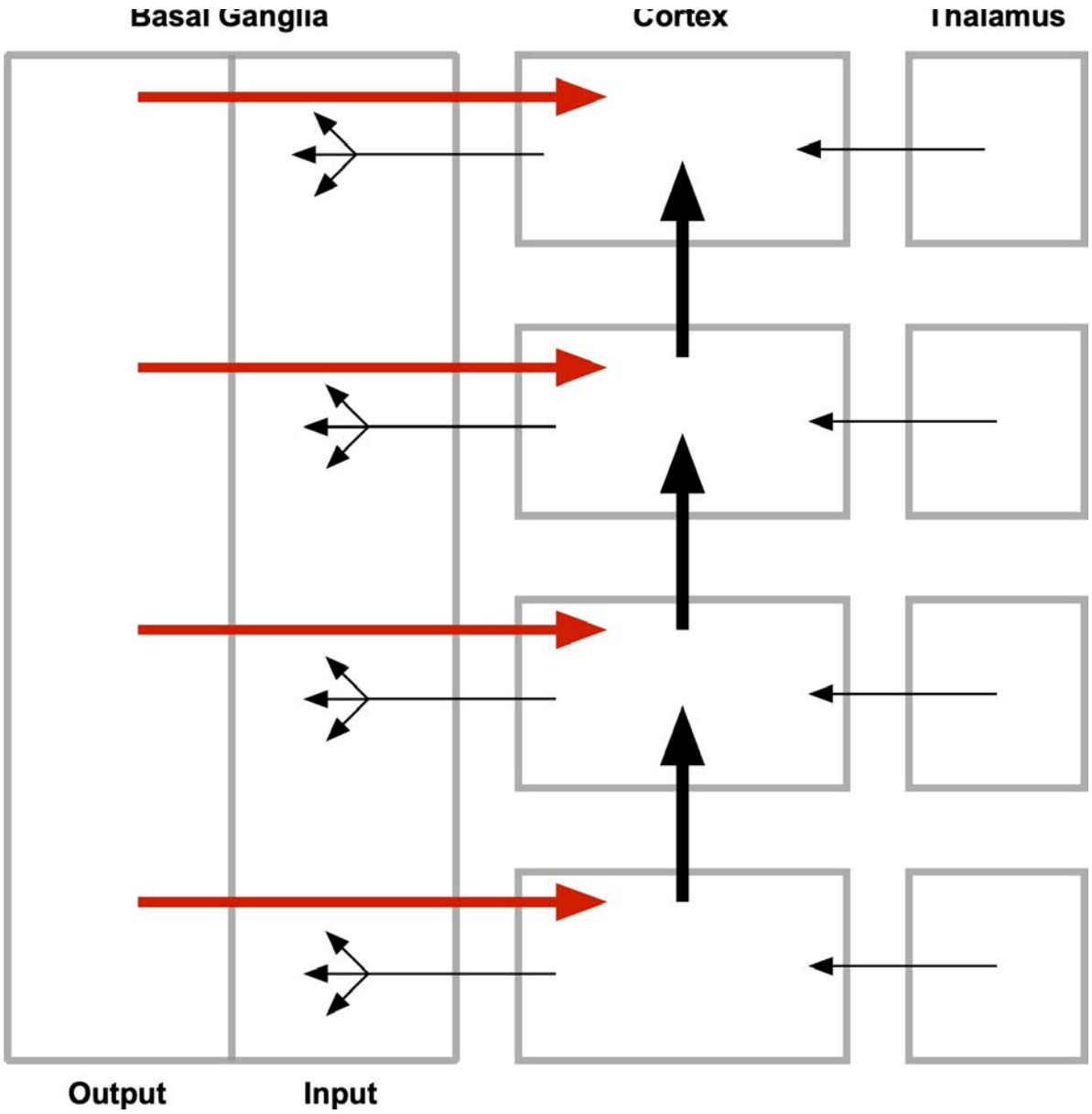

Figure 4. Schematic diagram of the loops between the basal ganglia and the cortex. Cortical areas project topographically to the basal ganglia, which then feedback topographically to the cortex. Compare this with the self-attention boxes (orange) in Fig. 2. Cortical hierarchies are found in sensory cortex, motor cortex and the prefrontal cortex. Associations between input to the basal ganglia can be learned through dopamine neurons, which carry reward prediction signals. The cortex receives inputs from the thalamus, similar to the encoder inputs that the decoder receives in a transformer.

References

1. Blakemore, C. & Tobin, E. A. Lateral inhibition between orientation detectors in the cat's visual cortex. *Exp. Brain Res.* **15**, 439–440 (1972).
2. Allman, J., Miezin, F. & McGuinness, E. Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annu. Rev. Neurosci.* **8**, 407–430 (1985).
3. Gilbert, C. D. Adult cortical dynamics. *Physiol. Rev.* **78**, 467–485 (1998).
4. Albright, T. D. & Stoner, G. R. Contextual influences on visual processing. *Annu. Rev. Neurosci.* **25**, 339–379 (2002).
5. Amari, S.-I. Learning Patterns and Pattern Sequences by Self-Organizing Nets of Threshold Elements. *IEEE Trans. Comput.* **C-21**, 1197–1206 (1972).
6. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, (2012).
7. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
8. Graves, A. Generating Sequences With Recurrent Neural Networks. *arXiv [cs.NE]* (2013).
9. Bryson, A. E. A gradient method for optimizing multi-stage allocation processes. *Symposium on digital computers and their applications*.
10. Werbos, P. J. Generalization of backpropagation with application to a recurrent gas market model. *Neural Netw.* **1**, 339–356 (1988).
11. Werbos, P. J. Backpropagation through time: what it does and how to do it. *Proc. IEEE* **78**, 1550–1560 (1990).
12. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
13. Bengio, Y., Frasconi, P. & Simard, P. The problem of learning long-term dependencies in recurrent networks. in *IEEE International Conference on Neural Networks* 1183–1188 vol.3 (1993).

14. Bengio, Y., Simard, P. & Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **5**, 157–166 (1994).
15. Kolen, J. F. & Kremer, S. C. *A Field Guide to Dynamical Recurrent Networks*. (John Wiley & Sons, 2001).
16. Bahdanau, D., Cho, K. & Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv [cs.CL]* (2014).
17. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, (2017).
18. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training.
19. Li, H. Language models: past, present, and future. *Commun. ACM* **65**, 56–63 (2022).
20. Brown, T. *et al.* Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
21. Swindale, N. Visual map. *Scholarpedia J.* **3**, 4607 (2008).
22. Hubel, D. H. & Wiesel, T. N. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* **148**, 574–591 (1959).
23. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–154 (1962).
24. Averbeck, B. B., Latham, P. E. & Pouget, A. Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* **7**, 358–366 (2006).
25. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).
26. Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9**, 1432–1438 (2006).
27. Angelucci, A. *et al.* Circuits for local and global signal integration in primary visual cortex. *J. Neurosci.* **22**, 8633–8646 (2002).
28. Rao, R. P. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation

of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).

29. Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S. & Churchland, A. K. Single-trial neural dynamics are dominated by richly varied movements. *Nat. Neurosci.* **22**, 1677–1686 (2019).
30. Li, J. S., Sarma, A. A., Sejnowski, T. J. & Doyle, J. C. Internal feedback in the cortical perception–action loop enables fast and accurate behavior. *Proceedings of the National Academy of Sciences* **120**, e2300445120 (2023).
31. Dotson, N. M. & Yartsev, M. M. Nonlocal spatiotemporal representation in the hippocampus of freely flying bats. *Science* **373**, 242–247 (2021).
32. Berry, M. J., Brivanlou, I. H., Jordan, T. A. & Meister, M. Anticipation of moving stimuli by the retina. *Nature* vol. 398 334–338 Preprint at https://doi.org/10.1038/18678 (1999).
33. Benvenuti, G. *et al.* Anticipatory responses along motion trajectories in awake monkey area V1. 2020.03.26.010017 (2020) doi:10.1101/2020.03.26.010017.
34. Muller, L., Reynaud, A., Chavane, F. & Destexhe, A. The stimulus-evoked population response in visual cortex of awake monkey is a propagating wave. *Nat. Commun.* **5**, 3675 (2014).
35. Muller, L., Chavane, F., Reynolds, J. & Sejnowski, T. J. Cortical travelling waves: mechanisms and computational principles. *Nat. Rev. Neurosci.* **19**, 255–268 (2018).
36. Girard, P., Hupé, J. M. & Bullier, J. Feedforward and feedback connections between areas V1 and V2 of the monkey have similar rapid conduction velocities. *J. Neurophysiol.* **85**, 1328–1331 (2001).
37. Stettler, D. D., Das, A., Bennett, J. & Gilbert, C. D. Lateral connectivity and contextual interactions in macaque primary visual cortex. *Neuron* **36**, 739–750 (2002).
38. Davis, Z. *et al.* Spontaneous traveling waves naturally emerge from horizontal fiber time delays and travel through locally asynchronous-irregular states. *Nature Communications* (2021).

39. Takahashi, K. *et al.* Large-scale spatiotemporal spike patterning consistent with wave propagation in motor cortex. *Nat. Commun.* **6**, 7169 (2015).
40. Davis, Z. W., Muller, L., Martinez-Trujillo, J., Sejnowski, T. & Reynolds, J. H. Spontaneous travelling cortical waves gate perception in behaving primates. *Nature* (2020) doi:10.1038/s41586-020-2802-y.
41. Roland, P. E. *et al.* Cortical feedback depolarization waves: a mechanism of top-down influence on early visual areas. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 12586–12591 (2006).
42. Xu, W., Huang, X., Takagaki, K. & Wu, J.-Y. Compression and reflection of visually evoked cortical waves. *Neuron* **55**, 119–129 (2007).
43. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025 (1999).
44. Singer, W. Recurrent dynamics in the cerebral cortex: Integration of sensory evidence with stored knowledge. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
45. Benigno, G. B., Budzinski, R. C., Davis, Z. W., Reynolds, J. H. & Muller, L. Waves traveling over a map of visual space can ignite short-term predictions of sensory input. *Nat. Commun.* **14**, 3409 (2023).
46. Liboni, L. H. B. *et al.* Image segmentation with traveling waves in an exactly solvable recurrent neural network. *arXiv [cs.CV]* (2023).
47. Anderson Keller, T., Muller, L., Sejnowski, T. & Welling, M. Traveling Waves Encode the Recent Past and Enhance Sequence Learning. *arXiv [cs.NE]* (2023).
48. Budzinski, R. C. *et al.* Geometry unites synchrony, chimeras, and waves in nonlinear oscillator networks. *Chaos* **32**, 031104 (2022).
49. Fukushima, K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36**, 193–202 (1980).
50. Otazu, G. H., Tai, L.-H., Yang, Y. & Zador, A. M. Engaging in an auditory task suppresses responses in auditory cortex. *Nat. Neurosci.* **12**, 646–654 (2009).

51. Manassi, M., Murai, Y. & Whitney, D. Serial dependence in visual perception: A meta-analysis and review. *J. Vis.* **23**, 18 (2023).
52. Gu, A. & Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv [cs.LG]* (2023).
53. Sejnowski, T. J. Large Language Models and the Reverse Turing Test. *Neural Comput.* **35**, 309–342 (2023).
54. Steriade, M., McCormick, D. A. & Sejnowski, T. J. Thalamocortical oscillations in the sleeping and aroused brain. *Science* **262**, 679–685 (1993).
55. Gais, S., Mölle, M., Helms, K. & Born, J. Learning-dependent increases in sleep spindle density. *J. Neurosci.* **22**, 6830–6834 (2002).
56. Clemens, Z., Fabó, D. & Halász, P. Overnight verbal memory retention correlates with the number of sleep spindles. *Neuroscience* **132**, 529–535 (2005).
57. Mednick, S. C. *et al.* The critical role of sleep spindles in hippocampal-dependent memory: a pharmacology study. *J. Neurosci.* **33**, 4494–4504 (2013).
58. Siapas, A. G. & Wilson, M. A. Coordinated interactions between hippocampal ripples and cortical spindles during slow-wave sleep. *Neuron* **21**, 1123–1128 (1998).
59. Peyrache, A., Battaglia, F. P. & Destexhe, A. Inhibition recruitment in prefrontal cortex during sleep spindles and gating of hippocampal inputs. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 17207–17212 (2011).
60. Buzsáki, G. Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus* **25**, 1073–1188 (2015).
61. Gonzalez, C. E. *et al.* Theta Bursts Precede, and Spindles Follow, Cortical and Thalamic Downstates in Human NREM Sleep. *J. Neurosci.* **38**, 9989–10001 (2018).
62. McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457

(1995).

63. Buzsaki, G. *Rhythms of the Brain*. (Oxford University Press, 2006).
64. Muller, L. *et al.* Rotating waves during human sleep spindles organize global patterns of activity that repeat precisely through the night. *Elife* **5**, (2016).
65. Peters, S. E. Dense array EEG measures of infant sleep. Preprint at https://doi.org/10.7282/T35142BT (2017).
66. Blankenship, A. G. & Feller, M. B. Mechanisms underlying spontaneous patterned activity in developing neural circuits. *Nat. Rev. Neurosci.* **11**, 18–29 (2010).
67. Kim, J., Song, M., Jang, J. & Paik, S.-B. Spontaneous Retinal Waves Can Generate Long-Range Horizontal Connectivity in Visual Cortex. *J. Neurosci.* **40**, 6584–6599 (2020).
68. Chklovskii, D. B., Schikorski, T. & Stevens, C. F. Wiring optimization in cortical circuits. *Neuron* **34**, 341–347 (2002).
69. King, J.-R. & Wyart, V. The Human Brain Encodes a Chronicle of Visual Events at Each Instant of Time Through the Multiplexing of Traveling Waves. *J. Neurosci.* **41**, 7224–7233 (2021).
70. Tomko, G. J. & Crapper, D. R. Neuronal variability: non-stationary responses to identical visual stimuli. *Brain Res.* **79**, 405–418 (1974).
71. Britten, K. H., Shadlen, M. N., Newsome, W. T. & Movshon, J. A. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J. Neurosci.* **12**, 4745–4765 (1992).
72. Shadlen, M. N., Britten, K. H., Newsome, W. T. & Movshon, J. A. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *J. Neurosci.* **16**, 1486–1510 (1996).
73. Nienborg, H., Cohen, M. R. & Cumming, B. G. Decision-related activity in sensory neurons: correlations among neurons and with behavior. *Annu. Rev. Neurosci.* **35**, 463–483 (2012).
74. Ringach, D. L. Spontaneous and driven cortical activity: implications for computation. *Curr.*

Opin. Neurobiol. **19**, 439–444 (2009).

75. Tsodyks, M., Kenet, T., Grinvald, A. & Arieli, A. Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science* **286**, 1943–1946 (1999).
76. Kenet, T., Bibitchkov, D., Tsodyks, M., Grinvald, A. & Arieli, A. Spontaneously emerging cortical representations of visual attributes. *Nature* **425**, 954–956 (2003).
77. Han, F., Caporale, N. & Dan, Y. Reverberation of recent visual experience in spontaneous cortical waves. *Neuron* **60**, 321–327 (2008).
78. Sejnowski, T. J. On the stochastic dynamics of neuronal interaction. *Biol. Cybern.* **22**, 203–211 (1976).
79. Chen, Y., Zhang, H. & Sejnowski, T. Predictive Sequence Learning in the Hippocampal Formation. *bioRxiv* 2022.05.19.492731 (2023) doi:10.1101/2022.05.19.492731.
80. Kim, R. & Sejnowski, T. J. Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks. *Nat. Neurosci.* **24**, 129–139 (2021).
81. Hebb, D. O. Organization of behavior: A neurophysiological theory. *(No Title)* (1949).
82. Sejnowski, T. J. The book of Hebb. *Neuron* **24**, 773–776 (1999).
83. Rao, R. P. N. & Sejnowski, T. J. Self-organizing neural systems based on predictive learning. *Philos. Trans. A Math. Phys. Eng. Sci.* **361**, 1149–1175 (2003).
84. Wan, K. Y. Active oscillations in microscale navigation. *Anim. Cogn.* (2023) doi:10.1007/s10071-023-01819-5.