University of Plymouth

PEARL **https://pearl.plymouth.ac.uk**

Faculty of Science and Engineering School of Engineering, Computing and Mathematics

2021-08

Biological constraints on neural network models of cognitive function

Pulvermuller, F

http://hdl.handle.net/10026.1/17554

10.1038/s41583-021-00473-5
Nature Reviews Neuroscience
Springer Science and Business Media LLC

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.

Perspective

Biological constraints on neural network models of cognitive function

Friedemann Pulvermüller[1-4]*, Rosario Tomasello[1,4], Malte R. Henningsen-Schomers[1,4] and Thomas Wennekers[5]

[1] *Brain Language Laboratory, Department of Philosophy and Humanities, WE4, Freie Universität Berlin, 14195 Berlin, Germany.*

[2] *Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, 10099 Berlin, Germany*

[3] *Einstein Center for Neurosciences Berlin, 10117 Berlin, Germany.*

[4] *Cluster of Excellence 'Matters of Activity', Humboldt-Universität zu Berlin, 10099 Berlin, Germany.*

[5] *School of Engineering, Computing and Mathematics, University of Plymouth, Plymouth PL48AA, UK.*

* e-mail: friedemann.pulvermuller@fu-berlin.de

Paper accepted for publication in *Nature Reviews Neuroscience* 5/2021

Abstract | Neural network models are potential tools for improving our understanding of complex brain functions. To address this goal, these models need to be neurobiologically realistic. However, although neural networks have advanced dramatically in recent years and even achieve human-like performance on complex perceptual and cognitive tasks, their similarity to aspects of brain anatomy and physiology is imperfect. Here, we discuss different types of neural models, including localist, auto-associative and hetero-associative, deep and whole-brain networks, and identify aspects under which their biological plausibility can be improved. These aspects range from the choice of model neurons and of mechanisms of synaptic plasticity and learning, to implementation of inhibition and control, along with neuroanatomical properties including area structure and local and long-range connectivity. We highlight recent advances in developing biologically grounded cognitive theories and in mechanistically explaining, based on these brain-constrained neural models, hitherto unaddressed issues regarding the nature, localization and ontogenetic and phylogenetic development of higher brain functions. In closing, we point to possible future clinical applications of brain-constrained modelling.

Introduction

Cognition calls for mechanistic explanation. The superb and specific cognitive capacities of humans and higher mammals may depend on specific structural and functional features of their brains. If so, these neurobiological features must play a decisive role in explanations of cognitive capacities. Despite substantial progress in understanding brain function in general, explaining how structural and functional features of neural tissue bring about cognition, language and thought has remained a challenge. We propose that such explanation can only be achieved if neural networks for modelling cognition incorporate a broad range of features that make them similar at different levels to real neurobiological networks. This article will discuss recent attempts to progress towards this goal and highlight seven brain constraints that may help to make neural networks more neural (**Fig. 1**). We discuss established types of cognitive neural models (**Fig. 2**) and give examples of neural simulations that integrate several brain constraints in novel ways. These networks offer perspectives on modelling neurocognitive mechanisms in and across multiple brain areas using huge numbers of realistic neurons and their local and global interactions through short- and long-range neuroanatomical connections, and hence provide opportunities for better understanding the mechanisms of cognition, including perception, categorization, attention, memory, language and semantic-conceptual processing.

Neural models of cognition

Localist networks. A standard way to present theories about cognition has been in the form of abstract box-and-arrow diagrams, with each 'box' indicating a module specialized for one cognitive sub-process (for example, sound processing) and arrows indicating information transmission between modules. These models were usually based on behavioural experiments involving healthy participants — for example, in tasks processing differences between speech and other sounds, which motivated the postulate of different modules for speech and non-linguistic acoustic stimuli[1,2]. Although some of these models were refined in light of studies in individuals with neurological impairments[3,4], they were generally formulated without explicit reference to neuronal circuits, despite researchers' demands for biological foundations of models of mental processing[5-7].

An important step towards addressing the neural substrate was taken by so-called localist models of cognition and language[8-12], which filled the boxes of modular models with single artificial 'neurons' thought to locally represent cognitive elements[13] such as perceptual features and percepts, phonemes, word forms, meaning features, concepts and so on (**Fig. 1a**). The 1:1 relationship between the artificial neuron-like computational–algorithmic implementations and the entities postulated by cognitive theories made it easy to connect the two types of models. However, the notion that individual neurons each carry major cognitive functions is controversial today and difficult to reconcile with evidence from neuroscience research[14,15]. This is not to dispute the great specificity of some neurons' responses[16], but rather to highlight the now dominant view that even these very specific cells "do not act in isolation but are part of cell assemblies representing familiar concepts", objects or other entities[17,18]. A further limitation of the localist models was that they did not systematically address the mechanisms underlying the formation of new representations and their connections.

Auto-associative networks. Neuroanatomical observations suggest that the cortex is characterized by ample intrinsic and recurrent connectivity between its neurons and, therefore, it can be seen as an associative memory[19,20]. This position inspired a family of artificial neural networks, called 'auto-associative networks' or 'attractor networks'[21-32].

Auto-associative network models implement neurons with connections between their neuron members, so that each neuron interlinks with several or even all of the other neurons included in the set. This contrasts with the hetero-associative networks discussed below, where connections run between sub-populations of network neurons without any connections within each neuron pool. To simulate the effect of learning in auto-associative networks, so-called learning rules are included that change the connection weights between neurons as a consequence of their prior activity. For example, biologically founded unsupervised Hebbian learning, which strengthens connections between co-activated neurons[5], is frequently applied and leads to the formation of strongly connected cell assemblies within a weakly connected auto-associative neuron pool (**Fig. 2b**). These cell assemblies can function as distributed network correlates or representations of perceptual, cognitive or 'mixed' context-dependent perceptual–cognitive states[6,30,32-34]. Therefore, the observations that cortical neurons work together in groups and that representations are distributed across such groups [14,18] can both be accommodated by this artificial network type, along with learning mechanisms, thus overcoming major shortcomings of localist networks.

Additional cognitively relevant features of auto-associative networks include the ability of a cell assembly to fully activate after only partial stimulation — a possible mechanism for Gestalt completion; that is, the recognition of an object (such as a cat) given only partial input (tail and paws). The mechanism is illustrated in **Fig. 2b**, where stimulation of neurons α and β is sufficient for activating the cell assembly formed by neurons α-to-γ.

Furthermore, auto-associative networks integrate the established observations that: cortical neural codes can be sparse (that is, only a small fraction of available neurons respond to a given (complex) stimulus)[15,18,22,35,36]; and that some (other) neurons respond to elementary and frequently occurring features of several stimuli (thus behaving in a less-sparse manner)[37]. The reason for this lies in cell assembly overlap; that is, the possibility that two or more such circuits can share neurons while remaining functionally separate. This is illustrated in **Fig. 2b**, by the 'overlap neuron' of cell assemblies α-to-γ and γ-to-ε.

Auto-associative networks can model a wide spectrum of cognitive processes, ranging from object, word and concept recognition to navigation, syntax processing, memory, planning and decision making[21,22,27,28,30,32,36,38-40]. Some models use several interlinked auto-associative

network components to model the interaction between multiple cortical areas in cognitive processing[36,41-50]. Furthermore, auto-associative components can be included in more complex networks, for example the deep recurrent networks discussed below.

Hetero-associative networks. Hetero-associative or feedforward networks include two or more neuron pools or 'layers' with connections between, but not (as in auto-associative networks) within, layers [51,52], and offer a different pathway towards distributed representations and learning. These models consist of different neuron pools connected in sequence, a connection scheme inspired by neurobiological structures; for example, by next-neighbour feedforward connections between layers of the retina or between cortical areas[53].

A frequently used class of models called multiple-layer perceptrons, which are frequently used in parallel-distributed processing, includes three neuron layers, with one layer receiving the input, one generating the output and one 'hidden' layer in between (**Fig. 2c**)[52]. In such networks, representations of cognitive entities are dense and nonsparse; that is, they are distributed across all neurons of the hidden layer, such that an activation vector across the entire layer is the network correlate of an object, word, meaning or thought[54]. Note that the dense and 'fully distributed' nature of these representations contrasts with the sparse (but also distributed) representations characterizing many auto-associative networks.

To implement a form of learning in these networks, input and output 'teacher' signals are fed to the input and output layers, respectively[52], and synaptic weights are modulated to make the network learn the relationships between inputs and teacher outputs. To adjust the synaptic weights, supervised learning algorithms are applied that adjust weights according to the error gradient that is calculated backwards from the output layer to the synapse in question. Initially, an algorithm called 'error back-propagation'[55] was applied; however, more recently, a range of learning rules based on error gradients across the network have become available, thus allowing for different variants of 'gradient-dependent' learning[56].

In order to implement memory processes, one additional layer was added to the three-layer architecture and connected reciprocally to the 'hidden' layer[57]; this 'simple recurrent network' architecture allows for reverberant activity. Similar to auto-associative networks, feedforward networks including three or four layers have successfully addressed a broad range of cognitive functions[58-60].

Deep neural networks. To further enhance their computational power, more layers were added to hetero-associative networks, thus resulting in deep neural networks (DNNs; **Fig. 2d**)[56,61-64]. A neurobiological motivation for increasing the number of layers from three to six and more comes from the neuroanatomical structure of the ventral occipitotemporal stream for object processing, in which neuroanatomical connections lead from the primary visual cortex (V1) to adjacent areas and, via multiple levels, to anterior temporal areas [65].

DNNs have been varied in several important ways. For example, convolutional DNNs include topographic projections between (some of) their layers to facilitate the joint processing of adjacent inputs, and recurrent DNNs include reciprocal connections between layers and/or recurrent links within layers that make it possible to maintain information over time. These modifications have produced extremely powerful devices that have reached near-human performance levels in several cognitive domains, including, for example, object classification[66,67] and speech recognition[68-70]. DNNs have been proposed to perform an over-parameterized blind process of directly fitting complex data sets with immanent regularization,

and to allow for generalization based on local interpolation, thus taking on functions previously attributed to discrete rules[71].

Despite their broad success, specific limitations of DNNs have recently been pointed out. They sometimes show inappropriate generalization behaviour[72-75], such as classifying, with high confidence, images completely unrecognizable to humans as instances of specific objects[73]. Similarly, an extensive body of research on so-called 'adversarial examples' has shown that minimal perturbations to images can, despite being imperceptible to humans, nonetheless cause gross misclassifications by DNNs[72,76]. It remains to be investigated whether these partial failures point to relevant differences between DNN-immanent and brain-internal perception mechanisms that can be remediated by algorithmic improvements, or rather reflect limitations of the stimulus sets applied during training, which differ from the realistic stimuli available to living beings during their ontogeny. A recent study found that introducing a hidden layer that better matches primate V1 improved robustness to adversarial examples in a neural network for image classification, suggesting that a higher degree of biological realism might be key[77].

Today, there are various neural network architectures that use different types of processing components, coding and learning rules. Supervised and unsupervised learning, sparse and dense distributed coding, and hetero-associative and auto-associative network components co-exist across approaches, and some models even mix these features — for example, by integrating both hetero-associative and auto-associative layers, as in the case of deep-recurrent and reservoir-computing networks[48,69,78,79]. The choice of network features is driven by processing efficiency, including error minimization, learning speed, effective use of computing resources and so on. Neurobiological plausibility is frequently used as a source of inspiration, as mentioned, but artificial neural networks are usually not designed to structurally and functionally resemble specific parts of the brain.

Whole-brain networks. The connections between the layers of neural networks are typically simple, with neighbouring areas being linked; however, this differs from the connectivity structure of the cortex, in which numerous areas are connected in intricate ways. Approximating the complex subdivision of brains into areas and nuclei, and implementing the neuroanatomical connectivity between these components, is a main goal of 'whole-brain' modelling. Rather than modelling all parts of the brain, most 'whole-brain' models focus on cortical areas and forebrain nuclei, or selections thereof.

Neuroanatomists have mapped the areal structure of the cortex and brain at coarse and fine-grained levels[80-82]. Information about structural brain connectivity is available from invasive tracer studies in animals[83,84], which have revealed fibre tracts and their directionality, and from non-invasive brain-imaging applicable in both animals and humans. The latter use diffusion tensor or diffusion weighted imaging (DTI/DWI) along with deterministic or probabilistic tractography [85-88]. By parcellating the cortex into a set of areas and translating its structural connectivity into between-area links, a whole-brain or multiple-area network model can be obtained in which each cortical area and anatomical connection has its corresponding network node and link, respectively (**Fig. 2e**).

Whole-brain models can be applied to investigate the functional consequences of neuroanatomical structure and connectivity. Multiple studies have simulated the spatial and temporal dynamics of brain activity that spontaneously emerges in resting conditions to explain the interplay between and functional coupling of areas[89-93]. It transpires that the structural connectivity imposed by cortical anatomy, along with dynamic parameters, determines and explains the emergence of functional communities of areas in the resting state[94-99].

Whole-brain networks provide an important tool for understanding aspects of brain activity. However, it is not uniformly functioning whole areas, but rather neurons and their interaction in neuronal circuits that are the functional units that carry brain function and cognition. The crucial neuronal circuits and their interconnecting links are in part determined by neuroanatomical structure, but an equally important contribution to circuit formation comes from synaptic plasticity. Therefore, to move towards improved biological plausibility, it is necessary to incorporate neurons and plasticity into whole-brain network simulations, thus taking advantage of the constraints realized by the aforementioned associative neural networks (for example, see refs. [57,62,100,101]). In essence, whole-brain network models can be improved by adding more detail at the sub-area and neuronal levels; correspondingly, DNNs can be made more realistic by adding neuroanatomical information about within-area and between-area connectivity.

Brain constraints

To make neural networks of neurocognitive function more realistic, it is necessary first to develop models of neuronal mechanisms across different levels, ranging from the micro-levels of neurons and local cortical circuits to the macro-level of cortical areas and global connectivity, and second to apply constraints from neuroscience at these different levels (**Fig. 1**). Previous proposals have already discussed various different constraints on neural networks[7,38,100,102,103], but have frequently listed neurobiological constraints side-by-side with technical or practical criteria (for example 'computational efficiency' or 'scalability'[100]), or addressed biological plausibility very generally (for example as 'biological realism' in ref. [7]).

Below, we spell out point-by-point what 'biological realism' implies and discuss a list of specific neurobiological constraints applicable to network models of cognitive functions — some of which are also suitable for neural networks more generally. As cognitive processes are in focus, mechanisms in the forebrain and especially in the cortex receive special attention. The proposed constraints are not thought to represent categorical features that must necessarily be met by models in a binary fashion. Rather, we conceive them as dimensions along which neural models can be adjusted gradually, in view of their specific purposes.

Integration at different levels. Most previous modelling has focused on one specific grain size, aiming to approximate neuronal function at the level of either single neurons[104,105], neuronal interaction in local cortical circuits[106-109] or global interplay between cortical areas (see Whole-brain networks). To simultaneously apply constraints at different levels of brain structure and function, these different levels must be addressed and integrated into a single model. Furthermore, multi-level modelling is required to exploit multiple sources of experimental data for model validation[101,110-112]. These different sources include behavioural performance (for example, accuracy and response times in cognitive experiments), and neurophysiological activity, where, once again, different spatial scales come into play, ranging from the micro-level of single-cell and multi-cell recordings to that of local field potentials and macroscopic local area activations as revealed by non-invasive neuroimaging techniques, such as electroencephalography (EEG), magnetoencephalography (MEG) and functional MRI. These physiological data also come at different temporal scales, ranging from millisecond delays between neurons and high-frequency oscillations, to slow potential shifts and neurometabolic changes. The activity of recorded neurons or imaged areas, and even the similarities between these activation patterns[113,114] or the transient functional interactions they reveal[115,116], can be related to activation patterns of artificial neurons, local neuron clusters or 'area' components of a model. Likewise, spatio-temporal patterns of activity within a local cortical neighbourhood or across different cortical areas can be compared with the patterns produced and integrated

by networks. Such comparison of activity patterns across different scales requires that components of the brains and networks resemble each other structurally, and that it is possible to identify brain parts with network parts. Only in this case can the physiological measures be used to validate the neural models. The constraints below address different aspects of such similarity.

Neuron models. The functional units of the cortex and brain are neurons. These neurons receive inputs that are translated into postsynaptic potentials and finally into an output of discrete action potentials, whereby action potential frequency can vary continuously[117,118]. All neural networks are composed of artificial correlates of neurons, but the level of detail with which neuronal function is simulated varies considerably [104,105,119].

Mean field models use neurons with continuous inputs and outputs (thus ignoring the spiking of most real neurons), together with a transfer function that transduces the former into the latter[120,121]. They can be interpreted to simulate the firing probability of single neurons or the cumulative activity of local neuron circuits. The more sophisticated spiking 'integrate-and-fire' neurons model the summation of postsynaptic potentials and resultant neuronal firing, and can be extended to integrate dendritic compartments and non-linear interactions between their inputs[104,122-125]. Multi-compartment and biophysical neuron models can include even more detail, including subdivisions of the dendritic tree, postsynaptic ion-channel dynamics and dendritic action potentials[104,126-129]. Thus, a dimension of progression addressing the degree of realism of the neuron model may advance from mean field to integrate-and-fire to biophysical neuron models. Furthermore, some neuronal activity is difficult to explain by the input and can thus be seen as noise[130]. The addition of noise, along with adaptation (that is, the reduction of activity with prolonged activation), can thus further increase the degree of realism of neuron models.

However, please note that the most detailed neuron model is not always the best choice for a given research question. Not only can relatively basic neuron models yield excellent descriptions of neuronal activity[104]; the greater computational resources required by sophisticated neuron models also currently limit their applicability to large-scale simulations of within-area and across-area interactions relevant for cognition.

Synaptic plasticity and learning. Evidently, the inclusion of learning mechanisms is a crucial biological ingredient of biologically plausible networks. As mentioned, localist and whole-brain models typically lack this feature. To model multiple learning systems in the brain, the implementation of both major forms of learning, supervised and unsupervised, is necessary.

Learning based on biologically plausible Hebbian principles is relevant for all cognitive domains. It is approximated by various different learning rules[131]. Some of these rules are elementary; for example, implementations of the 'fire-together-wire-together' principle lead to long-term potentiation (LTP) of connections between co-activated neurons. Other rules include more neurophysiological detail, in particular ones that can implement both LTP and, as a consequence of uncorrelated or anticorrelated activation, long-term depression (LTD)[132,133]. Synaptic plasticity dependent on the timing of action potentials, known as spike-timing dependent plasticity (STDP), is a consequence of Hebbian learning realized in sophisticated implementations[134-137]. Thus, a progression can be seen from the absence of learning, to LTP-based Hebbian learning and, ultimately, Hebbian LTP-plus-LTD learning rules and the addition of STDP.

Supervised learning involves the use of a feedback signal that informs the individual or network) whether performance was appropriate, or wrong or erroneous. However, the choice of algorithms used in supervised-learning simulations has been guided not only by biological plausibility[7,138]; the computational efficacy of gradient-dependent learning has played a major role [55,56,62]. It is controversial whether these latter algorithms are biologically realistic and applicable to sophisticated learning in specific cognitive domains. Some researchers have criticized the lack of strong biological support for mechanisms that compute and feed error gradients back through feedforward neuron networks[7,139], whereas others point to recent emerging evidence for neurobiological mechanisms that could, in part, support aspects thereof[56,140-142]. The mechanisms by which error gradients might gradually propagate backwards through a population of biological neurons remain a target of ongoing research.

A further putative problem concerns the type of feedback required for gradient-dependent learning. In DNN simulations, thousands of learning trials are available, whereby, for example, object pictures are classified with regard to category membership (for example, whether they show cats or cups). However, in many cases of classification and language learning in humans, such feedback does not play a major role — and is, if at all, only rarely available[143]. It rather seems that, in some cases of cognitive learning, the absence of certain inputs is sufficient. For example, to 'preempt' the use of specific linguistic constructions — that is, to block them from the learner's repertoire — it is sufficient to hear other constructions in contexts where the pre-empted constructions could in principle be used[144,145]. This type of cognitive learning is not explained by feedback-driven supervised learning but invites modelling based on Hebbian plasticity[146]. Nevertheless, explicit feedback is important in some types of learning (such as reinforcement learning) and its biologically realistic implementation is crucial [7,48,138].

Inhibition and regulation. Brains are regulated systems, and cortical activity is regulated by control mechanisms at different levels. These include the microscopic local circuit level, at which excitatory cortical neurons interact with local inhibitory cells, and the macroscopic, more global level of interacting brain parts, at which cortical activity is regulated through information exchange with the thalamus, basal ganglia and other subcortical structures[6,147,148]. Many distributed neural networks used for simulating cognition are composed of excitatory units only and lack inhibition mechanisms. Other modelling frameworks implement regulation at an abstract level, for example by predefining a maximum number of neurons in an area that are allowed to become active at one time[36,119]. More realistic neural network models include excitatory and inhibitory neuronal elements, such that activity regulation and control is achieved by their interplay, a mechanism crucial for arousal and attention (see Perspectives section). Few models have realized both local and global regulatory mechanisms[42-44,109,110,149-154]. Inclusion of inhibition and regulation mechanisms at both the local circuit level and more global levels (such as the area level) is an important feature of making neurocognitive networks biologically plausible.

Area structure. The cortex is structured into a set of areas, whereby area definition is primarily based on anatomical criteria and sometimes refined using functional information (see, for example, the Brainnetome Atlas)[80-82,155]. Depending on the question to be addressed by a simulation, a network model may implement one, a specific selection of, or all cortical areas along with subcortical nuclei. Each area or nucleus can be realized as a separate 'layer' or model area including a predefined number of artificial neurons. Dimensions of progressing towards biological realism include the range of brain parts and regions covered by the model

(from one, to several, to whole-brain) and the granularity of the modelled areas, moving from coarse to more fine-grained area subdivisions.

Within-area local connectivity. Pyramidal cells, the most common excitatory neurons in cortex, each carry 10,000–40,000 dendritic spines[20,156], each of which in turn typically contain one synapse[20]. Therefore, one of these cells may make contact with a few tens of thousands of other cortical cells — within a pool of 15–32 billions of neurons in human cortex overall[157]. The number of connections (more than 10^{14}) is too huge to be determined, item-by-item, by the genetic code alone, and therefore stochastic principles must also co-determine whether a specific pair of neurons is connected. The probability that two adjacent pyramidal neurons are connected decreases with the distance between them[20,101,158,159]. In sum, neuroanatomical studies indicate that local excitatory connections within a cortical area are sparse and show a neighbourhood bias towards links between adjacent neurons [20,160].

In view of these features, the lack of within-layer connections of hetero-associative networks does not seem biologically realistic. Many networks that include auto-associative layers or areas[21,22,28,161] include full connectivity between all neurons within these areas and, similarly, the memory layers of simple recurrent networks[57] have all-to-all recurrent connections. Such full auto-associativity likewise contrasts with the sparseness of intrinsic local cortical connections. The brain constraint of sparse, local and partly random connections with a neighbourhood bias has been realized in neural networks that connect auto-associativity with between-area hetero-associative connections[24,149,150,162]. Most advanced with regard to the local connectivity constraint are microcircuit models that realize different cell types and their location in layers of the neocortex[109,110,152,154,163]. Nonetheless, for most neural networks available today, the implementation of within-area connectivity constraints still leads to an increase in biological realism[101].

Between-area global connectivity. The connections between areas of cortex follow some general rules. Most links are reciprocal. Adjacent areas are almost always interlinked, and second-next neighbours are connected in many cases[20,164]. However, longer-distance links are sparser, and much effort has been spent to map them precisely using invasive techniques (for example, with tracers) and non-invasive techniques (such as DTI/DWI)[101,165-172].

If two areas are interlinked, their connections are in most cases reciprocal and show topographic projections, such that local neighbouring relationships are preserved. Between-area connections are carried by long axon branches of cortical pyramidal cells. These axon branches pass through the white matter and can reach neurons in distant areas, where they branch and make contact with a local neighbourhood of neurons. Most hetero-associative artificial networks implement between-area links as all-to-all connections. An advance in biological realism can be achieved by introducing random connectivity with biases towards topographical projections and neighbourhood links, as, for example, in multi-area auto-associative and convolutional deep networks[43,44,173]. At the macroscopic level, it is biologically motivated to replace the typical linear lineup of next-neighbour-connected areas (**Fig. 2c,d**) with realistic connectivity schemes (see, for example, refs. [65,168,174] and section on 'Whole-brain models'), ideally taking into account any connection asymmetries and even the number of axons per fibre bundle along with axonal conduction delays. Therefore, essential brain constraints on artificial neural networks come from the connectivity structure of between-area links as documented by neuroanatomical research.

Additional constraints on global connectivity may be taken from measures of functional interaction, including correlation-based undirected functional connectivity and directed effective connectivity[175-178], although we suggest applying such data as neurophysiological evidence for validating anatomically constrained models rather than as a-priori constraints (see 'Integration of modelling at different levels').

Recent progress and trends

We propose that models of brain function should attempt to integrate several and, ideally, all of the seven brain constraints mentioned above. The integration of microscopic and macroscopic levels is crucial to this endeavor (see constraint 'Integration of modelling at different levels'). Microcircuit networks provide a detailed picture of the functional interplay between connected excitatory and inhibitory neurons located in different cortical layers within one local cluster of neurons (constraints 'neuron models', 'inhibition and regulation' and 'within-area local connectivity')[109,110,152,153]. Several microcircuit models have been integrated into models of multiple cortical areas by taking into account long-distance connectivity as well (constraints 'area structure' and 'between-area global connectivity')[154,174,179-181].

The model by Schmidt et al.[154] convincingly integrates local-circuit with area-structure and global-connectivity constraints and simulates resting-state activity of the human brain at a hitherto unprecedented level of detail (**Fig. 3**). However, this model is computationally demanding and touches the limits of simulation capacities of current cutting-edge computing equipment, even though it restricts itself to modelling 32 areas, each of which simulates neurons below just 1 mm^2 of the cortical surface. These restrictions indicate that, in view of keeping computational efforts manageable, it is necessary to carefully select the degree to which constraints can gradually be met. Furthermore, applying all constraints in their most extreme form could result in computationally unrealizable models, which may, even if realized, be too complex to be helpful as tools for better understanding brain function.

Multiple-area networks including microcircuits open novel perspectives for neural modelling of cognition. After inclusion of biological learning mechanisms (constraint 'synaptic plasticity and learning'), they may be applied to simulate and potentially explain higher cognitive functions. Below, we address putative benefits of adding the aforementioned neurobiological constraints to neural models of cognition.

Linking cognitive theory to the brain. Cognitive models may perfectly fit the data obtained in cognitive tasks, independent of whether they are implemented algorithmically or in a neural network. However, there is typically more than one theory, and hence algorithm or network, that can model a given data set. An advantage of neural models is that they allow the introduction of functional neuroscience constraints that can be used to decide between these alternatives[9].

Object recognition and classification: As mentioned, the evidence for cognitive neural model evaluation can come from single-cell and multiple-cell recordings[111,182], fMRI activation patterns of areas, or spatio-temporal activation patterns revealing the orchestration of neural activity across different neuron populations or areas[162,183-185]. For example, a large range of models of object recognition are able to describe human object classification performance, but only neural models with some similarily to brain structure can be tested using neurophysiological data recorded at different levels of human and monkey inferior temporal cortex[186,187]. For such testing, it is particularly useful to apply not only measures that reveal the activated loci during object perception and classification, but also the more sophisticated method of

representational similarity analysis (RSA). RSA allows one to relate the degrees of similarity between brain activation patterns elicited by different stimuli to the similarity structure of the stimuli themselves, which can be measured at different levels (for example, perceptual or semantic)[113]. Notably, RSA has revealed that neurometabolic activity in inferior temporal cortex reflects semantic similarities of visually perceived objects[186]. Application and comparison of alternative neural (and non-neural) models have shown that the best fit of area-specific activation patterns and pattern similarities in temporal cortex signifying object perception was achieved with a feedforward convolutional DNN trained by applying supervised learning[184]. Furthermore, recent research indicates that inclusion of recurrent connections in DNNs is crucial for capturing aspects of the fast neurophysiological dynamics observed in temporal cortex[188]. Over and above the structural and functional constraints already addressed by these models, further structural constraints on within- and between-area connectivity may be considered in future (constraints 'inhibition and regulation', 'within-area local connectivity' and 'between-area global connectivity').

Attention. Research on attention offers another example for the application of neuroscience constraints to cognitive theory. Psychologists had long theorized about the fact that visual selective attention biases object and feature perception and that different perceptual elements compete with each other for attention, an interaction described by the biased competition model of attention[189]. This cognitive theory was translated into a neurocomputational model of early and higher visual cortex (including V2 and V4). Pools of excitatory and inhibitory neurons with within-area and between-area connections inspired by neuroscience data were used to model the interaction between representations, attentional bias and resource-limiting inhibition and to help to explain results of behavioural and neurophysiological recording experiments in non-human primates[182,190]. In a separate line of work using a convolutional DNN, the effects of attention on visual object classification were recently shown to depend on the level at which biases are applied, with modulation at higher layers being relatively more effective than modulation at lower layers in biasing attention toward specific object categories[191,192]. This observation may help to clarify the role of anterior temporal and prefrontal areas in visual attention processing[193]. Future biological modelling may marry within-area inhibition with deep architectures, constraining the latter using the neuroanatomical connectivity of the visual 'what' and 'where' streams, and/or may apply Hebbian learning instead of backpropagation for weight tuning (constraints 'synaptic plasticity and learning', 'area structure', 'within-area local connectivity' and 'between-area global connectivity').

Language processing and reasoning. A relatively straightforward way of modelling cognitive brain functions in a neurobiological format is to select a set of areas and nuclei, model each using a set of neurons, assign cognitive functions to model areas and determine the interactions among them using predefined or learned mappings between area-specific activation vectors. This strategy has given rise to biologically inspired models of cognitive functions, including concept, language and number processing[47,194-198]. One such model[47] is particularly impressive because of its size (2.5 million neurons) and implementation of the neuron-model constraint (integrate-and-fire neurons) in various cognitive tasks, addressing, for example, number recognition, question-answering and even fluid reasoning[199]. Twenty brain regions (areas and nuclei) are represented in the model, each by an ensemble of neurons, although global connections are sometimes idealized (for example, showing a linear line-up of areas from primary visual to anterior-temporal). Within areas, there is either no or full connectivity, and inhibitory neurons are missing. Therefore, three of the seven constraints have been addressed ('Neuron models', 'Area structure' and 'between-area global connectivity'). Other brain-constrained models using smaller vocabularies and task ranges than those used in this large model have successfully integrated conceptual and language processing with action and perception mechanisms and have been applied in robots[194,197,198,200].

Explaining neuroscience findings. Beyond putting cognitive theories into a neurobiological environment, brain constrained modelling can also provide answers to long-standing questions in neuroscience about how specific capacities of the brain are mechanistically implemented, how they emerge in ontogeny, how they came about in phylogeny and why they are situated in the specific brain parts where they are observed to be situated. We now turn specifically to this explanatory perspective.

Memory. Some models discussed above have provided a mechanistic explanation for the competition aspect of attention, which emerged from local inhibitory connections and between-area projections[182]; however, the other component of the biased-competition model of attention, the bias (such as that towards an object or part of visual space), is typically presented to the model from the outside and thus remains without model-immanent explanation. A mechanism for biases may come from network models of memory that account for memory dynamics based on neuronal function and structural connectivity. Here, memory mechanisms at the subcellular level (such as synaptic plasticity) are important, but in these models the questions of why certain types of memory develop in specific species, brain parts and circuits are still partly open. With appropriately adjusted activation thresholds and synaptic weights, a fully connected auto-associative memory is characterized by persistent attractor states, a possible correlate of working memory[22,201,202]. In models using spiking model neurons, auto-associative networks with all-to-all connectivity within their 'areas' gave rise to firing patterns resembling those recorded from neurons in inferior temporal lobe during working memory experiments[39,46,203].

Further work has addressed the question of why different cortical areas typically show different predominant memory activity patterns[46,204]. Hebbian learning principles suggest that memory circuits are built from neurons firing together during sensory stimulation and thus located in areas where the to-be-stored information arrives; that is, in primary areas. However, most neurons with memory characteristics develop in multimodal areas (in prefrontal, anterior-temporal and posterior-parietal cortex) distant from primary areas. A deep neural model implementing six frontal and temporal areas and their neuroanatomical connectivity structure showed that unsupervised Hebbian learning applied to this network gives rise to firing patterns and realistic distributions of memory cells across primary, secondary and multimodal areas, thus providing an explanation of cortical memory topographies and dynamics based on neuron function and corticocortical connectivity[185,204]. On a related thread, a recent study asked why a specific form of memory, verbal working memory (that is, memory for spoken words and language-like acoustic stimuli), developed in the context of specific evolutionary changes of the connectivity structure of fronto-temporal areas and is specific to humans. It was shown that the specific increase in fronto-temporal connectivity observed in primate evolution led to the emergence of distributed neuronal circuits for articulatory-acoustic units[205], thus offering an explanation of a specifically human trait based on phylogenetic structural change (**Fig. 4**). This model and the previously described one realize the seven constraints discussed in this article (with special emphasis on 'between-area global connectivity'), although the number of areas and the level of detail of local microcircuit simulation could still be increased.

Concepts. Similar explanatory advances could be achieved in the domain of language and conceptual processing. It is well-known that multimodal areas in frontal, temporal and parietal lobe are important for conceptual and semantic processing generally, whereas modality-preferential sensory and motor areas contribute to the processing of specific semantic categories[206-209]. An explanation of why these particular sites are relevant for concepts generally or specifically had been missing. Brain theory suggests that the convergence of multimodal information is essential for conceptual mechanisms[210]. To get from the visual

perception of a cat to the related concept and, for example, the knowledge of what kind of sound a cat typically produces, information from different modalities needs to be integrated. Such integration of multimodal information requires connections bridging between modality-specific neural systems in different cortical areas (which are distant from each other), which implies a role of large-scale connectivity structure in conceptual processing. Modelling of word learning processes in frontal and temporal cortex that applied realistic early language learning scenarios, unsupervised learning and anatomically constrained local and global connectivity, revealed a distribution of neuronal circuits for conceptual and semantic processing consistent with the data. Specifically, most neurons of conceptual circuits, which formed as a consequence of co-processing information about signs and their related objects or actions **(Fig. 5**), were housed in the model's multimodal convergence hubs, thus explaining the general 'pull' away from the areas where sensorimotor information reaches the cortex towards the most strongly connected connector hub areas. By contrast, the stimulation of specific modality-specific areas during conceptual learning and the resultant correlated neural activation topographies helped to explain the category-specific contribution of modality-preferential areas (**Fig. 5**)[185,211-213]. Again, realistic between-area connectivity was a crucial factor of these explanations; all seven constraints were addressed.

Future applications

Besides the aforementioned theoretical and explanatory advances offered by brain-constrained modelling, this novel strategy offers very practical application strategies for the future. One such application addresses neuroplasticity, aiming at predicting and explaining reorganization of cognitive functions after lesion or deprivation. In this context, recent modelling attempts have targeted, for example, altered language processing in patients with focal lesions of language-relevant cortical areas[214] and sensory deprivation in blind people[50]. These modelling experiments have led to accounts of well-documented neuroplasticity phenomena; for example, the takeover of visual areas during language and cognitive processing in congenitally blind individuals[50].

In future, it may become possible to perform neurocomputational modelling constrained by specific features of individual brains. Results obtained with individually constrained neural networks may open new perspectives on predicting future neuroplastic dynamics[215] and may be used for planning personalised therapy or surgery, for example for individuals with brain tumours[216-218]. Brain-constrained modelling applied to particular populations and even individual cases may thus open fruitful future perspectives.

Conclusion

Based on a brief overview of neural network simulations of cognition and their successes, we here suggest a move towards more biologically oriented modelling. According to the position put forward in this Perspective, neuroscience constraints have priority over other aims, such as processing efficacy and big data processing. Beyond the imperative to build models that bridge between the macro-scale and micro-scales, in order to be testable against physiological data at different levels, the proposed constraints address neuron models, learning algorithms and regulation mechanisms along with neuroanatomical constraints on model architecture that address area subdivisions and short-range and long-range connectivity. As concrete examples of brain-constrained modelling, we have reviewed biologically motivated neurocomputational implementations of neurocognitive theories and model-based explanations of the brain mechanisms underlying object perception attention, memory, concepts and language along

with their ontogenetic and phylogenetic development. The future outlook for brain-constrained modelling includes the generation of networks that realize features of individual brains, which may, for example, be useful in assessing which parts of cortex are crucial for retaining specific cognitive functions in a given individual.

References

1 Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. Perception of the speech code. *Psychological Review* **74**, 431-461 (1967).
2 Fodor, J. A. *The modulatity of mind*. (MIT Press, 1983).
3 Shallice, T. *From neuropsychology to mental structure*. (Cambridge University Press, 1988).
4 Ellis, A. W. & Young, A. W. *Human cognitive neuropsychology*. (Lawrence Erlbaum Associates Ltd., 1988).
5 Hebb, D. O. *The organization of behavior. A neuropsychological theory*. (John Wiley, 1949).
6 Braitenberg, V. in *Theoretical approaches to complex systems. (Lecture notes in biomathematics, vol. 21)* (eds R. Heim & G. Palm) 171-188 (Springer, 1978).
7 O'Reilly, R. C. Six principles for biologically based computational models of cortical cognition. *Trends Cogn Sci* **2**, 455-562 (1998).
8 Dell, G. S. A spreading-activation theory of retrieval in sentence production. *Psychological Review* **93**, 283-321 (1986).
9 MacKay, D. G. *The organization of perception and action. A theory of language and other cognitive skills*. (Springer-Verlag, 1987).
10 Grainger, J. & Jacobs, A. M. Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review* **103**, 518-565 (1996).
11 Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M. & Gagnon, D. A. Lexical access in aphasic and nonaphasic speakers. *Psychological Review* **104**, 801-838 (1997).
12 Dijkstra, T. *et al.* Multilink: a computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition* **22**, 657-679 (2019).
13 Barlow, H. Single units and cognition: a neurone doctrine for perceptual psychology. *Perception* **1**, 371-394 (1972).
14 Abeles, M. *Corticonics - Neural circuits of the cerebral cortex*. (Cambridge University Press, 1991).
15 Quiroga, R. Q., Kreiman, G., Koch, C. & Fried, I. Sparse but not 'grandmother-cell'coding in the medial temporal lobe. *Trends in cognitive sciences* **12**, 87-91 (2008).
16 Perrett, D. J., Mistlin, A. J. & Chitty, A. J. Visual neurons responsive to faces. *Trends in Neurosciences* **10**, 358-364 (1987).
17 Quiroga, R. Q. Concept cells: the building blocks of declarative memory functions. *Nat Rev Neurosci* **13**, 587-597, doi:10.1038/nrn3251 (2012).
18 Quiroga, R. Q. Plugging in to human memory: advantages, challenges, and insights from human single-neuron recordings. *Cell* **179**, 1015-1032 (2019).
19 Braitenberg, V. in *Architectonics of the cerebral cortex* (eds M.A.B. Brazier & H. Petsche) 443-465 (Raven Press, 1978).
20 Braitenberg, V. & Schüz, A. *Cortex: statistics and geometry of neuronal connectivity*. 2 edn, (Springer, 1998).
21 Willshaw, D. J., Buneman, O. P. & Longuet-Higgins, H. C. Non-holographic associative memory. *Nature* **222**, 960-962. (1969).
22 Palm, G. *Neural assemblies*. (Springer, 1982).
23 Palm, G. Cell assemblies as a guideline for brain research. *Concepts in Neuroscience* **1**, 133-147 (1990).
24 Palm, G., Knoblauch, A., Hauser, F. & Schüz, A. Cell assemblies in the cerebral cortex. *Biol Cybern* **108**, 559–572, doi:10.1007/s00422-014-0596-4 (2014).
25 Lundqvist, M., Rehn, M., Djurfeldt, M. & Lansner, A. Attractor dynamics in a modular network model of neocortex. *Network* **17**, 253-276, doi:10.1080/09548980600774619 (2006).

26 Lansner, A. Associative memory models: from the cell-assembly theory to biophysically detailed cortex simulations. *Trends Neurosci* **32**, 178-186, doi:10.1016/j.tins.2008.12.002 (2009).
27 Hopfield, J. J. & Tank, D. W. Computing with neural circuits: A model. *Science* **233**, 625-633 (1986).
28 Hinton, G. E. & Shallice, T. Lesioning an attractor network: investigation of acquired dyslexia. *Psychological Review* **98**, 74-95 (1991).
29 Sommer, F. T. & Wennekers, T. Models of distributed associative memory networks in the brain. *Theory in Biosciences* **122**, 55-69 (2003).
30 Rigotti, M., Ben Dayan Rubin, D., Wang, X. J. & Fusi, S. Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. *Front Comput Neurosci* **4**, 24, doi:10.3389/fncom.2010.00024 (2010).
31 Huyck, C. R. & Passmore, P. J. A review of cell assemblies. *Biol Cybern* **107**, 263-288, doi:10.1007/s00422-013-0555-5 (2013).
32 Lindsay, G. W., Rigotti, M., Warden, M. R., Miller, E. K. & Fusi, S. Hebbian Learning in a Random Network Captures Selectivity Properties of the Prefrontal Cortex. *J Neurosci* **37**, 11021-11036, doi:10.1523/JNEUROSCI.1222-17.2017 (2017).
33 Ballintyn, B., Shlaer, B. & Miller, P. Spatiotemporal discrimination in attractor networks with short-term synaptic plasticity. *Journal of computational neuroscience* **46**, 279-297 (2019).
34 Seeholzer, A., Deger, M. & Gerstner, W. Stability of working memory in continuous attractor networks under the control of short-term plasticity. *PLoS Comput Biol* **15**, e1006928, doi:10.1371/journal.pcbi.1006928 (2019).
35 Olshausen, B. A. & Field, D. J. Sparse coding of sensory inputs. *Curr Opin Neurobiol* **14**, 481-487, doi:10.1016/j.conb.2004.07.007 (2004).
36 Papadimitriou, C. H., Vempala, S. S., Mitropolsky, D., Collins, M. & Maass, W. Brain computation by assemblies of neurons. *Proc Natl Acad Sci U S A* **117**, 14464-14472, doi:10.1073/pnas.2001893117 (2020).
37 Hubel, D. *Eye, brain, and vision*. 2 edn, (Scientific American Library, 1995).
38 Wennekers, T., Garagnani, M. & Pulvermüller, F. Language models based on Hebbian cell assemblies. *J Physiol Paris* **100**, 16-30 (2006).
39 Zipser, D., Kehoe, B., Littlewort, G. & Fuster, J. M. A spiking network model of short-term active memory. *Journal of Neuroscience* **13**, 3406-3420. (1993).
40 Pulvermüller, F., Garagnani, M. & Wennekers, T. Thinking in circuits: Towards neurobiological explanation in cognitive neuroscience. *Biol Cybern* **108**, 573-593, doi:10.1007/s00422-014-0603-9 (2014).
41 Dominey, P. F. Complex sensory-motor sequence learning based on recurrent state representation and reinforcement learning. *Biol Cybern* **73**, 265-274 (1995).
42 Bibbig, A., Wennekers, T. & Palm, G. A neural network model of the cortico-hippocampal interplay and the representation of contexts. *Behav Brain Res* **66**, 169-175. (1995).
43 Knoblauch, A. & Palm, G. Scene segmentation by spike synchronization in reciprocally connected visual areas. I. Local effects of cortical feedback. *Biol Cybern* **87**, 151-167 (2002).
44 Knoblauch, A. & Palm, G. Scene segmentation by spike synchronization in reciprocally connected visual areas. II. Global assemblies and synchronization on larger space and time scales. *Biol Cybern* **87**, 168-184 (2002).
45 Dominey, P. F. & Inui, T. Cortico-striatal function in sentence comprehension: Insights from neurophysiology and modeling. *Cortex* **45**, 1012-1018 (2009).
46 Verduzco-Flores, S., Bodner, M., Ermentrout, B., Fuster, J. M. & Zhou, Y. Working memory cells' behavior may be explained by cross-regional networks with synaptic facilitation. *PLoS One* **4**, e6399 (2009).
47 Eliasmith, C. *et al.* A large-scale model of the functioning brain. *Science* **338**, 1202-1205 (2012).
48 Cazin, N. *et al.* Reservoir computing model of prefrontal cortex creates novel combinations of previous navigation sequences from hippocampal place-cell replay with spatial reward propagation. *PLoS Comput Biol* **15**, e1006624, doi:10.1371/journal.pcbi.1006624 (2019).
49 Drude, L., von Neumann, T. & Haeb-Umbach, R. in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 11-15 (IEEE).

50 Tomasello, R., Wennekers, T., Garagnani, M. & Pulvermüller, F. Visual cortex recruitment during language processing in blind individuals is explained by Hebbian learning. *Sci Rep* **9**, 3579 (2019).
51 Minsky, M. & Papert, S. *Perceptrons*. (MIT Press, 1969).
52 McClelland, J. L. & Rumelhart, D. E. *Parallel distributed processing: explorations in the microstructure of cognition*. (MIT Press, 1986).
53 Hubel, D. *Eye, brain, and vision*. (Freeman, 1988).
54 McClelland, J. L. & Rumelhart, D. E. Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General* **114**, 159-188 (1985).
55 Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **323**, 533-536 (1986).
56 Richards, B. A. *et al.* A deep learning framework for neuroscience. *Nat Neurosci* **22**, 1761-1770, doi:10.1038/s41593-019-0520-2 (2019).
57 Elman, J. L. *et al. Rethinking innateness. A connectionist perspective on development*. (MIT Press, 1996).
58 Rumelhart, D. E. & McClelland, J. L. in *Parallel distributed processing: explorations in the microstructure of cognition* (eds J.L. McClelland & D.E. Rumelhart) (MIT Press, 1986).
59 Elman, J. L. Finding structure in time. *Cognitive Science* **14**, 179-211 (1990).
60 Rogers, T. T. & McClelland, J. L. *Semantic cognition. A parallel distributed processing approach*. (MIT Press, 2004).
61 Hinton, G. E. Learning multiple layers of representation. *Trends Cogn Sci* **11**, 428-434, doi:10.1016/j.tics.2007.09.004 (2007).
62 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444 (2015).
63 Kriegeskorte, N. & Golan, T. Neural network models and deep learning. *Current Biology* **29**, R231-R236 (2019).
64 Yu, Y., Si, X., Hu, C. & Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput* **31**, 1235-1270, doi:10.1162/neco_a_01199 (2019).
65 Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex* **1**, 1-47, doi:10.1093/cercor/1.1.1 (1991).
66 Krizhevsky, A., Sutskever, I. & Hinton, G. E. in *Advances in neural information processing systems.* 1097-1105.
67 Zhou, H.-Y., Liu, A.-A., Nie, W.-Z. & Nie, J. Multi-View Saliency Guided Deep Neural Network for 3-D Object Retrieval and Classification. *IEEE Transactions on Multimedia* **22**, 1496-1506 (2019).
68 Dahl, G. E., Yu, D., Deng, L. & Acero, A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing* **20**, 30-42 (2012).
69 Graves, A., Mohamed, A.-R. & Hinton, G. in *2013 IEEE international conference on acoustics, speech and signal processing.* 6645-6649 (IEEE).
70 Smit, P., Virpioja, S. & Kurimo, M. Advances in subword-based HMM-DNN speech recognition across languages. *Computer Speech and Language* **66**, 101-158 (2021).
71 Hasson, U., Nastase, S. A. & Goldstein, A. Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron* **105**, 416-434, doi:10.1016/j.neuron.2019.12.002 (2020).
72 Szegedy, C. *et al.* Intriguing properties of neural networks. *arXiv* **1312.6199**, 1-10 (2014).
73 Nguyen, A., Yosinski, J. & Clune, J. in *Computer Vision and Pattern Recognition (CVPR 2015).* (ed IEEE) 427-436 (IEEE).
74 Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires re-thinking generalization. *arXiv* **1611**, 03530v03532 (2017).
75 Alcorn, M. A. *et al.* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4845-4854 (2019).
76 Carlini, N. & Wagner, D. Towards Evaluating the Robustness of Neural Networks. *arXiv* **1608.04644v2** (2017).
77 Dapello, J. *et al.* Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. *BioRxiv* **2020.06.16.154542**, doi:https://doi.org/10.1101/2020.06.16.154542 (2020).

78 Devereux, B. J., Clarke, A. & Tyler, L. K. Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Scientific reports* **8**, 1-12 (2018).
79 Tanaka, G. *et al.* Recent advances in physical reservoir computing: A review. *Neural Netw* **115**, 100-123, doi:10.1016/j.neunet.2019.03.005 (2019).
80 Brodmann, K. *Vergleichende Lokalisationslehre der Grobhirnrinde*. (Barth, 1909).
81 Fan, L. *et al.* The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cereb Cortex* **26**, 3508-3526, doi:10.1093/cercor/bhw157 (2016).
82 Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171-178, doi:10.1038/nature18933 (2016).
83 Pandya, D. N. & Yeterian, E. H. in *Cerebral cortex. Vol. 4. Association and auditory cortices* (eds A. Peters & E.G. Jones) 3-61 (Plenum Press, 1985).
84 Yeterian, E. H., Pandya, D. N., Tomaiuolo, F. & Petrides, M. The cortical connectivity of the prefrontal cortex in the monkey brain. *Cortex* **48**, 58-81, doi:10.1016/j.cortex.2011.03.004 (2012).
85 Waugh, J. L. *et al.* A registration method for improving quantitative assessment in probabilistic diffusion tractography. *NeuroImage* **189**, 288-306 (2019).
86 Sarwar, T., Ramamohanarao, K. & Zalesky, A. Mapping connectomes with diffusion mri: deterministic or probabilistic tractography? *Magnetic Resonance In Medicine* **81**, 1368-1384 (2019).
87 Descoteaux, M., Deriche, R., Knosche, T. R. & Anwander, A. Deterministic and probabilistic tractography based on complex fibre orientation distributions. *IEEE Transactions On Medical Imaging* **28**, 269-286 (2008).
88 Behrens, T. E. J., Berg, H. J., Jbabdi, S., Rushworth, M. F. S. & Woolrich, M. W. Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *Neuroimage* **34**, 144-155 (2007).
89 Kötter, R. Neuroscience databases: tools for exploring brain structure-function relationships. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **356**, 1111-1120 (2001).
90 Bressler, S. L. & Menon, V. Large-scale brain networks in cognition: emerging methods and principles. *Trends Cogn Sci* **14**, 277-290, doi:10.1016/j.tics.2010.04.004 (2010).
91 Hagmann, P. *et al.* Mapping the structural core of human cerebral cortex. *PLoS Biol* **6**, e159, doi:10.1371/journal.pbio.0060159 (2008).
92 Honey, C. J., Kotter, R., Breakspear, M. & Sporns, O. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc Natl Acad Sci U S A* **104**, 10240-10245, doi:10.1073/pnas.0701519104 (2007).
93 Honey, C. J. *et al.* Predicting human resting-state functional connectivity from structural connectivity. *Proc Natl Acad Sci U S A* **106**, 2035-2040, doi:10.1073/pnas.0811168106 (2009).
94 Deco, G. & Jirsa, V. K. Ongoing cortical activity at rest: criticality, multistability, and ghost attractors. *J Neurosci* **32**, 3366-3375, doi:10.1523/JNEUROSCI.2523-11.2012 (2012).
95 Deco, G., Jirsa, V. K. & McIntosh, A. R. Resting brains never rest: computational insights into potential cognitive architectures. *Trends Neurosci* **36**, 268-274, doi:10.1016/j.tins.2013.03.001 (2013).
96 Deco, G., Tononi, G., Boly, M. & Kringelbach, M. L. Rethinking segregation and integration: contributions of whole-brain modelling. *Nat Rev Neurosci* **16**, 430-439, doi:10.1038/nrn3963 (2015).
97 Petersen, S. E. & Sporns, O. Brain Networks and Cognitive Architectures. *Neuron* **88**, 207-219, doi:10.1016/j.neuron.2015.09.027 (2015).
98 Nakagawa, T. T., Adhikari, M. H. & Deco, G. Large-scale Computational Models of Ongoing Brain Activity. *Computational Models of Brain and Behavior*, 425 (2017).
99 Avena-Koenigsberger, A., Misic, B. & Sporns, O. Communication dynamics in complex brain networks. *Nat Rev Neurosci* **19**, 17-33, doi:10.1038/nrn.2017.149 (2017).
100 Palm, G. Neural Information Processing in Cognition: We Start to Understand the Orchestra, but Where is the Conductor? *Front Comput Neurosci* **10**, 3, doi:10.3389/fncom.2016.00003 (2016).

101 van Albada, S. J. *et al.* Bringing anatomical information into neuronal network models. *arXiv* **arXiv:1312.6026** (2020).
102 Arbib, M. A., Billard, A., Iacoboni, M. & Oztop, E. Synthetic brain imaging: grasping, mirror neurons and imitation. *Neural Netw* **13**, 975-997 (2000).
103 Kell, A. J. & McDermott, J. H. Deep neural network models of sensory systems: windows onto the role of task constraints. *Curr Opin Neurobiol* **55**, 121-132, doi:10.1016/j.conb.2019.02.003 (2019).
104 Gerstner, W. & Naud, R. Neuroscience. How good are neuron models? *Science* **326**, 379-380, doi:10.1126/science.1181936 (2009).
105 Teeter, C. *et al.* Generalized leaky integrate-and-fire models classify multiple neuron types. *Nature communications* **9**, 1-15 (2018).
106 Schwalger, T., Deger, M. & Gerstner, W. Towards a theory of cortical columns: From spiking neurons to interacting neural populations of finite size. *PLoS Comput Biol* **13**, e1005507, doi:10.1371/journal.pcbi.1005507 (2017).
107 Malagarriga, D., Pons, A. J. & Villa, A. E. Complex temporal patterns processing by a neural mass model of a cortical column. *Cognitive neurodynamics* **13**, 379-392 (2019).
108 Jansen, B. H. & Rit, V. G. Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biological cybernetics* **73**, 357-366 (1995).
109 Potjans, T. C. & Diesmann, M. The cell-type specific cortical microcircuit: relating structure and activity in a full-scale spiking network model. *Cereb Cortex* **24**, 785-806, doi:10.1093/cercor/bhs358 (2014).
110 Einevoll, G. T. *et al.* The Scientific Case for Brain Simulations. *Neuron* **102**, 735-744, doi:10.1016/j.neuron.2019.03.027 (2019).
111 O'Connell, R. G., Shadlen, M. N., Wong-Lin, K. & Kelly, S. P. Bridging Neural and Computational Viewpoints on Perceptual Decision-Making. *Trends Neurosci* **41**, 838-852, doi:10.1016/j.tins.2018.06.005 (2018).
112 Hahn, G., Ponce-Alvarez, A., Deco, G., Aertsen, A. & Kumar, A. Portraits of communication in neuronal networks. *Nat Rev Neurosci* **20**, 117-127, doi:10.1038/s41583-018-0094-0 (2019).
113 Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis - connecting the branches of systems neuroscience. *Front Syst Neurosci* **2**, 4 (2008).
114 Carota, F., Nili, H., Pulvermüller, F. & Kriegeskorte, N. Distinct fronto-temporal substrates of distributional and taxonomic similarity among words: evidence from RSA of BOLD signals. *Neuroimage* **224**, 117408, doi:10.1016/j.neuroimage.2020.117408 (2021).
115 Papadopoulou, M., Friston, K. & Marinazzo, D. Estimating directed connectivity from cortical recordings and reconstructed sources. *Brain topography* **32**, 741-752 (2019).
116 Shen, K. *et al.* Exploring the limits of network topology estimation using diffusion-based tractography and tracer studies in the macaque cortex. *NeuroImage* **191**, 81-92 (2019).
117 Kandel, E. R., Schwartz, J. H. & Jessell, T. M. *Principles of neural sciences*. 4 edn, (McGraw-Hill, Health Professions Division, 2000).
118 Matthews, G. G. *Cellular physiology of nerve and muscle*. (John Wiley & Sons, 2009).
119 O'Reilly, R. C. & Munakata, Y. *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. (MIT Press, 2000).
120 Deco, G., Jirsa, V. K., Robinson, P. A., Breakspear, M. & Friston, K. The dynamic brain: from spiking neurons to neural masses and cortical fields. *PLoS Comput Biol* **4**, e1000092 (2008).
121 Breakspear, M. Dynamic models of large-scale brain activity. *Nat Neurosci* **20**, 340-352, doi:10.1038/nn.4497 (2017).
122 Burkitt, A. N. A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input. *Biol Cybern* **95**, 1-19, doi:10.1007/s00422-006-0068-6 (2006).
123 Brette, R. *et al.* Simulation of networks of spiking neurons: a review of tools and strategies. *J Comput Neurosci* **23**, 349-398, doi:10.1007/s10827-007-0038-6 (2007).
124 Gerstner, W. & Kistler, W. M. *Spiking neuron models: Single neurons, populations, plasticity*. (Cambridge university press, 2002).
125 Li, S. *et al.* Dendritic computations captured by an effective point neuron model. *Proc Natl Acad Sci U S A* **116**, 15244-15252, doi:10.1073/pnas.1904463116 (2019).
126 London, M. & Häusser, M. Dendritic computation. *Annu Rev Neurosci* **28**, 503-532, doi:10.1146/annurev.neuro.28.061604.135703 (2005).

127 Bono, J. & Clopath, C. Modeling somatic and dendritic spike mediated plasticity at the single neuron and network level. *Nat Commun* **8**, 706, doi:10.1038/s41467-017-00740-z (2017).
128 Venkadesh, S., Komendantov, A. O., Wheeler, D. W., Hamilton, D. J. & Ascoli, G. A. Simple models of quantitative firing phenotypes in hippocampal neurons: Comprehensive coverage of intrinsic diversity. *PLoS Comput Biol* **15**, e1007462, doi:10.1371/journal.pcbi.1007462 (2019).
129 Gidon, A. *et al.* Dendritic action potentials and computation in human layer 2/3 cortical neurons. *Science* **367**, 83-87, doi:10.1126/science.aax6239 (2020).
130 Faisal, A. A., Selen, L. P. & Wolpert, D. M. Noise in the nervous system. *Nat Rev Neurosci* **9**, 292-303, doi:10.1038/nrn2258 (2008).
131 Gerstner, W. & Kistler, W. M. Mathematical formulations of Hebbian learning. *Biol Cybern* **87**, 404-415 (2002).
132 Tsumoto, T. Long-term potentiation and long-term depression in the neocortex. *Progress in Neurobiology* **39**, 209-228 (1992).
133 Artola, A. & Singer, W. Long-term depression of excitatory synaptic transmission and its relationship to long-term potentiation. *Trends in Neurosciences* **16**, 480-487 (1993).
134 Gerstner, W., Kempter, R., van Hemmen, J. L. & Wagner, H. A neuronal learning rule for sub-millisecond temporal coding. *Nature* **383**, 76-81, doi:10.1038/383076a0 (1996).
135 Kempter, R., Gerstner, W. & Van Hemmen, J. L. Hebbian learning and spiking neurons. *Physical Review E* **59**, 4498 (1999).
136 Caporale, N. & Dan, Y. Spike Timing-Dependent Plasticity: A Hebbian Learning Rule. *Annu Rev Neurosci* **31**, 25-46 (2008).
137 Rumbell, T., Denham, S. L. & Wennekers, T. A spiking self-organizing map combining STDP, oscillations, and continuous learning. *IEEE Trans Neural Netw Learn Syst* **25**, 894-907, doi:10.1109/TNNLS.2013.2283140 (2014).
138 Mollick, J. A. *et al.* A systems-neuroscience model of phasic dopamine. *Psychol Rev* **127**, 972-1021, doi:10.1037/rev0000199 (2020).
139 Thorpe, S. J. & Imbert, M. in *Connectionism in perspective* (eds R. Pfeifer, Z. Schreter, F. Fogelman-Soulie, & L. Steels) 63–92 (North-Holland/Elsevier Science, 1989).
140 Marblestone, A. H., Wayne, G. & Kording, K. P. Toward an Integration of Deep Learning and Neuroscience. *Front Comput Neurosci* **10**, 94, doi:10.3389/fncom.2016.00094 (2016).
141 Pozzi, I., Bohté, S. & Roelfsema, P. A biologically plausible learning rule for deep learning in the brain. *arXiv preprint arXiv:1811.01768* (2018).
142 Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J. & Hinton, G. Backpropagation and the brain. *Nat Rev Neurosci* **21**, 335-346, doi:10.1038/s41583-020-0277-3 (2020).
143 Marcus, G. F. Negative evidence in language acquisition. *Cognition* **46**, 53-85 (1993).
144 Goldberg, A. E. *Constructions at work: The nature of generalisation in language*. (Oxford University Press, 2006).
145 Goldberg, A. E. *Explain me this: Creativity, competition and the partial preductivity of constructions*. (Princeton University Press, 2019).
146 Pulvermüller, F. Neural reuse of action perception circuits for language, concepts and communication. *Prog Neurobiol* **160**, 1-44, doi:10.1016/j.pneurobio.2017.07.001 (2018).
147 Yuille, A. L. & Geiger, D. in *The handbook of brain theory and neural networks* (ed M. A. Arbib) 1228-1231 (A Bradford Book/MIT Press, 2003).
148 Gurney, K., Prescott, T. J., Wickens, J. R. & Redgrave, P. Computational models of the basal ganglia: from robots to membranes. *Trends Neurosci* **27**, 453-459 (2004).
149 Knoblauch, A., Markert, H. & Palm, G. in *International work-conference on the interplay between natural and artificial computation 2005* Vol. 3562 *Lecture Notes In Computer Science* (eds J. Mira & J.R. Alvarez) 405-414 (Springer, 2005).
150 Sommer, F. T. & Wennekers, T. Associative memory in networks of spiking neurons. *Neural Netw* **14**, 825-834 (2001).
151 Garagnani, M., Wennekers, T. & Pulvermüller, F. A neuroanatomically-grounded Hebbian learning model of attention-language interactions in the human brain. *European Journal of Neuroscience* **27**, 492-513 (2008).
152 Binzegger, T., Douglas, R. J. & Martin, K. A. A quantitative map of the circuit of cat primary visual cortex. *J Neurosci* **24**, 8441-8453, doi:10.1523/JNEUROSCI.1400-04.2004 (2004).

153 Thomson, A. M. & Lamy, C. Functional maps of neocortical local circuitry. *Front Neurosci* **1**, 19-42, doi:10.3389/neuro.01.1.1.002.2007 (2007).
154 Schmidt, M. *et al.* A multi-scale layer-resolved spiking network model of resting-state dynamics in macaque visual cortical areas. *PLoS Comput Biol* **14**, e1006359, doi:10.1371/journal.pcbi.1006359 (2018).
155 Van Essen, D. C., Glasser, M. F., Dierker, D. L., Harwell, J. & Coalson, T. Parcellations and hemispheric asymmetries of human cerebral cortex analyzed on surface-based atlases. *Cereb Cortex* **22**, 2241-2262, doi:10.1093/cercor/bhr291 (2012).
156 Elston, G. N., Benavides-Piccione, R. & DeFelipe, J. The pyramidal cell in cognition: a comparative study in human and monkey. *J Neurosci* **21**, RC163 (2001).
157 Haug, H. Brain sizes, surfaces, and neuronal sizes of the cortex cerebri: a stereological investigation of man and his variability and a comparison with some mammals (primates, whales, marsupials, insectivores, and one elephant). *American Journal of Anatomy* **180**, 126-142 (1987).
158 Hellwig, B. A quantitative analysis of the local connectivity between pyramidal neurons in layers 2/3 of the rat visual cortex. *Biological Cybernetics* **82**, 111-121. (2000).
159 Perin, R., Berger, T. K. & Markram, H. A synaptic organizing principle for cortical neuronal groups. *Proceedings of the National Academy of Sciences* **108**, 5419-5424 (2011).
160 Kaas, J. H. Topographic maps are fundamental to sensory processing. *Brain Res Bull* **44**, 107-112 (1997).
161 Hopfield, J. J. & Tank, D. W. "Neural" computation of decisions in optimization problems. *Biol Cybern* **52**, 141-152 (1985).
162 Garagnani, M., Lucchese, G., Tomasello, R., Wennekers, T. & Pulvermüller, F. A spiking neurocomputational model of high-frequency oscillatory brain responses to words and pseudowords. *Frontiers in Computational Neuroscience* **10**, 145, doi:doi: 10.3389/fncom.2016.00145 (2017).
163 Douglas, R. J., Martin, K. A. & Whitteridge, D. A canonical microcircuit for neocortex. *Neural computation* **1**, 480-488 (1989).
164 Young, M. P., Scannell, J. W. & Burns, G. *The analysis of cortical connectivity*. (Springer, 1995).
165 Eichert, N. *et al.* What is special about the human arcuate fasciculus? Lateralization, projections, and expansion. *Cortex* **118**, 107-115, doi:10.1016/j.cortex.2018.05.005 (2019).
166 Rojkova, K. *et al.* Atlasing the frontal lobe connections and their variability due to age and education: a spherical deconvolution tractography study. *Brain Struct Funct* **221**, 1751-1766, doi:10.1007/s00429-015-1001-3 (2016).
167 Fernandez-Miranda, J. C. *et al.* Asymmetry, connectivity, and segmentation of the arcuate fascicle in the human brain. *Brain Struct Funct* **220**, 1665-1680, doi:10.1007/s00429-014-0751-7 (2015).
168 Rilling, J. K. Comparative primate neuroimaging: insights into human brain evolution. *Trends in Cognitive Sciences* **18**, 46-55 (2014).
169 Petrides, M., Tomaiuolo, F., Yeterian, E. H. & Pandya, D. N. The prefrontal cortex: comparative architectonic organization in the human and the macaque monkey brains. *Cortex* **48**, 46-57, doi:10.1016/j.cortex.2011.07.002 (2012).
170 Thiebaut de Schotten, M., Dell'Acqua, F., Valabregue, R. & Catani, M. Monkey to human comparative anatomy of the frontal lobe association tracts. *Cortex* **48**, 82-96, doi:10.1016/j.cortex.2011.10.001 (2012).
171 Ardesch, D. J. *et al.* Evolutionary expansion of connectivity between multimodal association areas in the human brain compared with chimpanzees. *Proc Natl Acad Sci U S A* **116**, 7101-7106, doi:10.1073/pnas.1818512116 (2019).
172 Barbeau, E. B., Descoteaux, M. & Petrides, M. Dissociating the white matter tracts connecting the temporo-parietal cortical region with frontal cortex using diffusion tractography. *Sci Rep* **10**, 8186, doi:10.1038/s41598-020-64124-y (2020).
173 Kietzmann, T., McClure, P. & Kriegeskorte, N. in *Oxford Research Encyclopedia, Neuroscience* (Oxford University Press, 2019).

174 Schuecker, J., Schmidt, M., van Albada, S. J., Diesmann, M. & Helias, M. Fundamental Activity Constraints Lead to Specific Interpretations of the Connectome. *PLoS Comput Biol* **13**, e1005179, doi:10.1371/journal.pcbi.1005179 (2017).

175 Friston, K. J. Functional and effective connectivity: a review. *Brain Connect* **1**, 13-36, doi:10.1089/brain.2011.0008 (2011).

176 Friston, K., Moran, R. & Seth, A. K. Analysing connectivity with Granger causality and dynamic causal modelling. *Curr Opin Neurobiol* **23**, 172-178, doi:10.1016/j.conb.2012.11.010 (2013).

177 Sokolov, A. A. *et al.* Asymmetric high-order anatomical brain connectivity sculpts effective connectivity. *Netw Neurosci* **4**, 871-890, doi:10.1162/netn_a_00150 (2020).

178 Zarghami, T. S. & Friston, K. J. Dynamic effective connectivity. *Neuroimage* **207**, 116453, doi:10.1016/j.neuroimage.2019.116453 (2020).

179 Markov, N. T. *et al.* A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cereb Cortex* **24**, 17-36, doi:10.1093/cercor/bhs270 (2014).

180 Schmidt, M., Bakker, R., Hilgetag, C. C., Diesmann, M. & van Albada, S. J. Multi-scale account of the network structure of macaque visual cortex. *Brain Struct Funct* **223**, 1409-1435, doi:10.1007/s00429-017-1554-4 (2018).

181 Schmidt, M., Bakker, R., Hilgetag, C. C., Diesmann, M. & van Albada, S. J. Correction to: Multi-scale account of the network structure of macaque visual cortex. *Brain Struct Funct*, doi:10.1007/s00429-019-02020-6 (2020).

182 Deco, G. & Rolls, E. T. Neurodynamics of biased competition and cooperation for attention: a model with spiking neurons. *J Neurophysiol* **94**, 295-313 (2005).

183 Bojak, I., Oostendorp, T. F., Reid, A. T. & Kötter, R. Towards a model-based integration of co-registered electroencephalography/functional magnetic resonance imaging data with realistic neural population meshes. *Philos Trans A Math Phys Eng Sci* **369**, 3785-3801, doi:10.1098/rsta.2011.0080 (2011).

184 Khaligh-Razavi, S. M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput Biol* **10**, e1003915, doi:10.1371/journal.pcbi.1003915 (2014).

185 Tomasello, R., Garagnani, M., Wennekers, T. & Pulvermüller, F. A neurobiologically constrained cortex model of semantic grounding with spiking neurons and brain-like connectivity. *Front Comput Neurosci* **12**, 88, doi:10.3389/fncom.2018.00088 (2018).

186 Carlson, T. A., Simmons, R. A., Kriegeskorte, N. & Slevc, L. R. The emergence of semantic meaning in the ventral temporal pathway. *J Cogn Neurosci*, doi:10.1162/jocn_a_00458 (2013).

187 Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci Rep* **6**, 27755, doi:10.1038/srep27755 (2016).

188 Kietzmann, T. C. *et al.* Recurrence is required to capture the representational dynamics of the human visual system. *Proc Natl Acad Sci U S A* **116**, 21854-21863, doi:10.1073/pnas.1905544116 (2019).

189 Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annu Rev Neurosci* **18**, 193-222 (1995).

190 Buehlmann, A. & Deco, G. The neuronal basis of attention: rate versus synchronization modulation. *J Neurosci* **28**, 7679-7686 (2008).

191 Lindsay, G. W. & Miller, K. D. How biological attention mechanisms improve task performance in a large-scale visual system model. *Elife* **7**, doi:10.7554/eLife.38105 (2018).

192 Lindsay, G. W. Attention in Psychology, Neuroscience, and Machine Learning. *Front Comput Neurosci* **14**, 29, doi:10.3389/fncom.2020.00029 (2020).

193 Duncan, J., Assem, M. & Shashidhara, S. Integrated Intelligence from Distributed Brain Activity. *Trends Cogn Sci* **24**, 838-852, doi:10.1016/j.tics.2020.06.012 (2020).

194 Markert, H., Kaufmann, U., Kara Kayikci, Z. & Palm, G. Neural associative memories for the integration of language, vision and action in an autonomous agent. *Neural Netw* **22**, 134-143 (2009).

195 Ueno, T., Saito, S., Rogers, T. T. & Lambon Ralph, M. A. Lichtheim 2: synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron* **72**, 385-396 (2011).

196 Zhong, J., Cangelosi, A. & Wermter, S. Toward a self-organizing pre-symbolic neural model representing sensorimotor primitives. *Front Behav Neurosci* **8**, 22, doi:10.3389/fnbeh.2014.00022 (2014).
197 Cangelosi, A., Schlesinger, M. & Smith, L. B. *Developmental robotics: From babies to robots*. (MIT Press, 2015).
198 Heinrich, S. & Wermter, S. Interactive natural language acquisition in a multi-modal recurrent neural architecture. *Connection Science* **30**, 99-133 (2018).
199 Raven, J. & Court, J. *Manual for Raven's Progressive Matrices and Vocabulary Scales*. (Harcourt Assessment, 2004).
200 Rast, A. D. *et al.* Behavioral Learning in a Cognitive Neuromorphic Robot: An Integrative Approach. *IEEE Trans Neural Netw Learn Syst* **29**, 6132-6144, doi:10.1109/TNNLS.2018.2816518 (2018).
201 Rolls, E. T. & Deco, G. Networks for memory, perception, and decision-making, and beyond to how the syntax for language might be implemented in the brain. *Brain Res*, doi:10.1016/j.brainres.2014.09.021 (2014).
202 Fuster, J. M. & Bressler, S. L. Cognit activation: a mechanism enabling temporal integration in working memory. *Trends Cogn Sci* **16**, 207-218, doi:10.1016/j.tics.2012.03.005 (2012).
203 Fiebig, F. & Lansner, A. A Spiking Working Memory Model Based on Hebbian Short-Term Potentiation. *J Neurosci* **37**, 83-96, doi:10.1523/JNEUROSCI.1989-16.2017 (2017).
204 Pulvermüller, F. & Garagnani, M. From sensorimotor learning to memory cells in prefrontal and temporal association cortex: A neurocomputational study of disembodiment *Cortex* **57**, 1-21 (2014).
205 Schomers, M. R., Garagnani, M. & Pulvermüller, F. Neurocomputational consequences of evolutionary connectivity changes in perisylvian language cortex. *J Neurosci* **37**, 3045-3055, doi:10.1523/JNEUROSCI.2693-16.2017 (2017).
206 Binder, J. R. & Desai, R. H. The neurobiology of semantic memory. *Trends Cogn Sci* **15**, 527-536, doi:10.1016/j.tics.2011.10.001 (2011).
207 Kiefer, M. & Pulvermüller, F. Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex* **48**, 805-825 (2012).
208 Ralph, M. A., Jefferies, E., Patterson, K. & Rogers, T. T. The neural and computational bases of semantic cognition. *Nat Rev Neurosci* **18**, 42-55, doi:10.1038/nrn.2016.150 (2017).
209 Harpaintner, M., Sim, E. J., Trumpp, N. M., Ulrich, M. & Kiefer, M. The grounding of abstract concepts in the motor and visual system: An fMRI study. *Cortex* **124**, 1-22, doi:10.1016/j.cortex.2019.10.014 (2020).
210 Damasio, A. R. The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation* **1**, 123-132 (1989).
211 Garagnani, M. & Pulvermüller, F. Conceptual grounding of language in action and perception: a neurocomputational model of the emergence of category specificity and semantic hubs. *Eur J Neurosci* **43**, 721-737, doi:10.1111/ejn.13145 (2016).
212 Chen, L., Ralph, M. A. L. & Rogers, T. T. A unified model of human semantic knowledge and its disorders. *Nature Human Behaviour* **1**, 0039 (2017).
213 Tomasello, R., Garagnani, M., Wennekers, T. & Pulvermüller, F. Brain connections of words, perceptions and actions: A neurobiological model of spatio-temporal semantic activation in the human cortex. *Neuropsychologia* **98**, 111–129, doi:10.1016/j.neuropsychologia.2016.07.004 (2017).
214 Chang, Y.-N. & Ralph, M. A. L. A unified neurocomputational bilateral pathway model of spoken language production in healthy participants and recovery in post-stroke aphasia. *bioRxiv, doi: https://doi.org/10.1101/2020.02.21.959239* (2020).
215 Seghier, M. L. & Price, C. J. Interpreting and utilising intersubject variability in brain function. *Trends in Cognitive Sciences* **22**, 517-530 (2018).
216 Picht, T., Frey, D., Thieme, S., Kliesch, S. & Vajkoczy, P. Presurgical navigated TMS motor cortex mapping improves outcome in glioblastoma surgery: a controlled observational study. *J Neurooncol* **126**, 535-543, doi:10.1007/s11060-015-1993-9 (2016).
217 Cha, Y. J. *et al.* Prediction of response to stereotactic radiosurgery for brain metastases using convolutional neural networks. *Anticancer Research* **38**, 5437-5445 (2018).

218 Tuncer, M. S. *et al.* Towards a tractography-based risk stratification model for language area associated gliomas. *Neuroimage Clin* **29**, 102541, doi:10.1016/j.nicl.2020.102541 (2021).
219 Schaefer, A. *et al.* Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb Cortex* **28**, 3095-3114, doi:10.1093/cercor/bhx179 (2018).
220 Sotiropoulos, S. N. & Zalesky, A. Building connectomes using diffusion MRI: why, how and but. *NMR Biomed* **32**, e3752, doi:10.1002/nbm.3752 (2017).
221 Rilling, J. K. *et al.* The evolution of the arcuate fasciculus revealed with comparative DTI. *Nat Neurosci* **11**, 426-428 (2008).
222 Rilling, J. K., Glasser, M. F., Jbabdi, S., Andersson, J. & Preuss, T. M. Continuity, divergence, and the evolution of brain language pathways. *Front Evol Neurosci* **3**, 11 (2011).
223 Martin, A. The representation of object concepts in the brain. *Annu Rev Psychol* **58**, 25-45 (2007).
224 Barsalou, L. W. Grounded cognition. *Annu Rev Psychol* **59**, 617-645 (2008).
225 Borghi, A. M. *et al.* Words as social tools: Language, sociality and inner grounding in abstract concepts. *Physics of Life Reviews* **29**, 120-153, doi:10.1016/j.plrev.2018.12.001 (2019).
226 Grisoni, L., Tomasello, R. & Pulvermüller, F. Correlated Brain Indexes of Semantic Prediction and Prediction Error: Brain Localization and Category Specificity. *Cereb Cortex* **31**, 1553-1568, doi:10.1093/cercor/bhaa308 (2021).

Author contributions

F.P., T.W., R.T. and M.R.H.-S. researched data for the article. F.P. and T.W. contributed substantially to discussion of the content. F.P. wrote the article. All authors reviewed and/or edited the manuscript before submission.

Competing interests

The authors declare no competing interests.

Peer review information

Nature Reviews Neuroscience thanks [Referee name], [Referee name] and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements

The authors wish to thank A. Aertsen, N. Bray, A. Cangelosi, L. Fekonja, M. Garagnani, A. Glenberg, L. Grisoni, S. Harnad, A. Knoblauch, G. Palm, T. Picht, S. Rotter and W. Schäffner for comments and suggestions on earlier versions of this manuscript and related talks. Research funding was provided through the following organizations and research grants: European Research Council, Advanced Grant “Material constraints enabling human cognition, MatCo” (ERC-2019-ADG 883811); Deutsche Forschungsgemeinschaft (German Research Foundation) Excellence Strategy cluster “Matters of Activity, MoA” (DFG EXC 2025/1).

Related links

Brainnetome Atlas: http://atlas.brainnetome.org

Figure Captions

Fig. 1 | Seven constraints for making neural networks models more biologically plausible. Constraints address the integration of modelling across the levels of cortical neurons, local cortical circuits and macroscopic brain structures (**a**; left to right panels) and specifically highlight the nature of the neuron model (**b**), the implementation of synaptic plasticity and learning (**c**), regulation and control by way of interplay between excitatory and inhibitory neurons (**d**), gross anatomical structure and area subdivision (**e**) and local within-area connectivity (**f**) and global between-area connectivity (**g**). Most current network models used for modelling cognition focus on only one or a few of these aspects, whereas brain-constrained modelling works towards networks integrating all of them. Part **e** redrawn after ref. [109]. Part **g** adapted from ref. [219], Figure 4, panel in row 5, column 3, and ref. [220], Figure 1, panel in row 2 on the right.

Fig. 2 | Networks for modelling cognitive functions. a | A localist network model includes nodes representing cognitive entities, for example word forms (middle layer), phonemes (bottom layer) and semantic features (top layer). Lines indicate links between nodes. Nodes sum up their inputs linearly, such that the activation of the phoneme nodes of /d/, /o/ and /g/ activate the word node for 'dog', which, in turn, activates semantic feature units (filled circles at the top) characterizing the related concept. **b** | Auto-associative networks include connections between their neurons, such that reverberating activity is possible; they are inspired by the local connectivity between adjacent cortical pyramidal cells[20,22]. This panel shows the connectivity matrix between five artificial neurons, α to ε. These neurons make up an auto-associative network that includes two discrete representations indicated in magenta (neurons α-to-γ) and cyan (neurons γ-to-ε). Numbers specify the presence (1) or absence (0) of a connection from the neuron listed on the left of the matrix to the neuron indicated at the top. Each neuron becomes active if and only if it receives at least two simultaneous inputs, thus resulting in the discrete representations maintaining activity over time. **c** | In hetero-associative networks, neuron populations ordered in 'layers' project onto each other serially, resembling connectivity in some neural structures. The typical three-layer networks used in many parallel-distributed processing (PDP) models include input and output layers plus a 'hidden' layer in-between. **d |** Deep neural networks include several hidden layers. The number of neurons per layer can vary substantially. Representations are activation vectors across all neurons of a layer. **e** | Whole-brain models implement global between-area connectivity. Here, between-area connectivity of the right hemisphere is shown for a single subject (top left panel) and a group of subjects both in anatomical topology (bottom left) and in matrix form (right). The matrix gives all connections between pairs of areas of the network model; colors indicate connection weights (fiber densities). Part **a** is adapted with permission from ref. [11]. Part **e** is adapted with permission from ref. [91].

Fig. 3 | Multi-level network for explaining neural dynamics based on neuroanatomical constraints. Neurophysiological activity in 32 cortical areas of the visual system of the macaque was modelled[154]. Probabilities of between-area connectivity (bottom left panel) were based on anatomical data and general connectivity principles. Applying further neuroanatomical constraints, each area was modelled as a 1 mm^2 patch of cortex with excitatory and inhibitory spiking cells (right panel) arranged in layers and specific connection probabilities within and between layer-specific cell populations (top left panel). The model was applied to reproduce and explain spiking neuronal activity from neurophysiological

recordings, activity propagation across areas and causal dynamic interactions between neuron populations and areas. The model unifies local and large-scale accounts of the cortex and clarifies how the detailed local and global connectivity of the cortex shapes its dynamics at multiple scales. Synaptic plasticity and learning were not modelled, although all of the other six constraints discussed in the main text were applied. Figure adapted, with permission, from ref. [180].

Fig. 4 | Model of evolutionary connectivity change in left fronto-temporal cortex and its functional consequences. Verbal working memory is a feature of humans that apes and monkeys apparently lack. **a |** 'Monkey' and 'human' models of six areas of fronto-temporal cortex involved in articulation and auditory perception (middle panel) were used to address this issue. Areas were modelled as sets of 625 mean-field excitatory neurons, each projecting randomly to local neighbourhoods of other excitatory units (coloured); each excitatory cell has a corresponding inhibitory 'cell' (in grey) projecting to a narrow local neighbourhood (left panel). Between-area connections implementing comparative DTI results[170,171,221,222] included next-neighbour connections between areas (green) and the second-next area connections specific to human perisylvian cortex (violet links in the 'human' model only). Correlation-based Hebbian plasticity was applied to imitate early sound and sign learning and to interlink articulatory and auditory information. **b |** After stimulation with learnt auditory–articulatory patterns, the monkey model showed weak and short-lived sequential activation of the model areas: A1, primary auditory cortex; AB auditory belt cortex; M1, primary articulatory motor cortex; PB, auditory parabelt cortex; PF, inferior prefrontal cortex; PM, premotor cortex. **c |** The same stimulation led to strong and long-lasting parallel activation in the areas of the human model. This prolonged activity can be interpreted as verbal working memory, a mechanism necessary for human language. The model applies all seven constraints discussed in the main text.. The left portion of part **a** is adapted with permission from ref. [211]. Parts **a-c** are adapted from ref. [205].

Fig. 5 | Brain-constrained model of semantic grounding. a | For simulating the infant's learning of the meaning of object-related and action-related words, a 12-area model was created including the six inferior-frontal and superior-temporal perisylvian areas of **Fig. 4** (A1, primary auditory cortex; AB auditory belt cortex; M1, primary articulatory motor cortex; PB, auditory parabelt cortex; PF, inferior prefrontal cortex; PM, premotor cortex), plus a ventral temporo-occipital visual stream (in green: AT, anterior-temporal cortex; TO, temporo-occipital cortex;V1, primary visual cortex) and a dorsolateral frontal action stream (in yellow-brown: lateral PF (PF_L), lateral PM (PM_L) and lateral M1 ($M1_L$)). Between-area connectivity is shown by arrows. Semantic learning and grounding of object and action words was modelled by co-presenting acoustic and articulatory information along with either semantic-referential object-related information or action-related information. This was done by co-activating specific patterns of spiking neurons in the different 'primary' areas of the model (M1 and $M1_L$, V1 and A1) and Hebbian correlation learning. After learning, 'auditory word comprehension' was simulated by presenting specific previously learned auditory patterns to area A1. As a result, specific circuits of neurons distributed across the network were activated, as indicated by the coloured dots in the insets (1 dot indicates 1 active model neuron; blue, object–word circuit; red, action–word circuit; yellow, both), shown in the black boxes representing areas. **b,c |** Distribution of circuit neurons across model areas. Bars give average numbers of neurons per area for object- (dark grey) and action-word circuits (light grey); whiskers give standard errors. Note the relatively stronger representation of object–word circuits in ventral-visual areas and that of action–word circuits in dorsolateral-frontal areas in part **a**[50,213], which offer

an explanation for well-known differences between the cortical mechanisms underlying action-related and object-related concepts[207,223-226]. All seven constraints discussed in the main text were implemented. Figure adapted, with permission, from ref. [185].

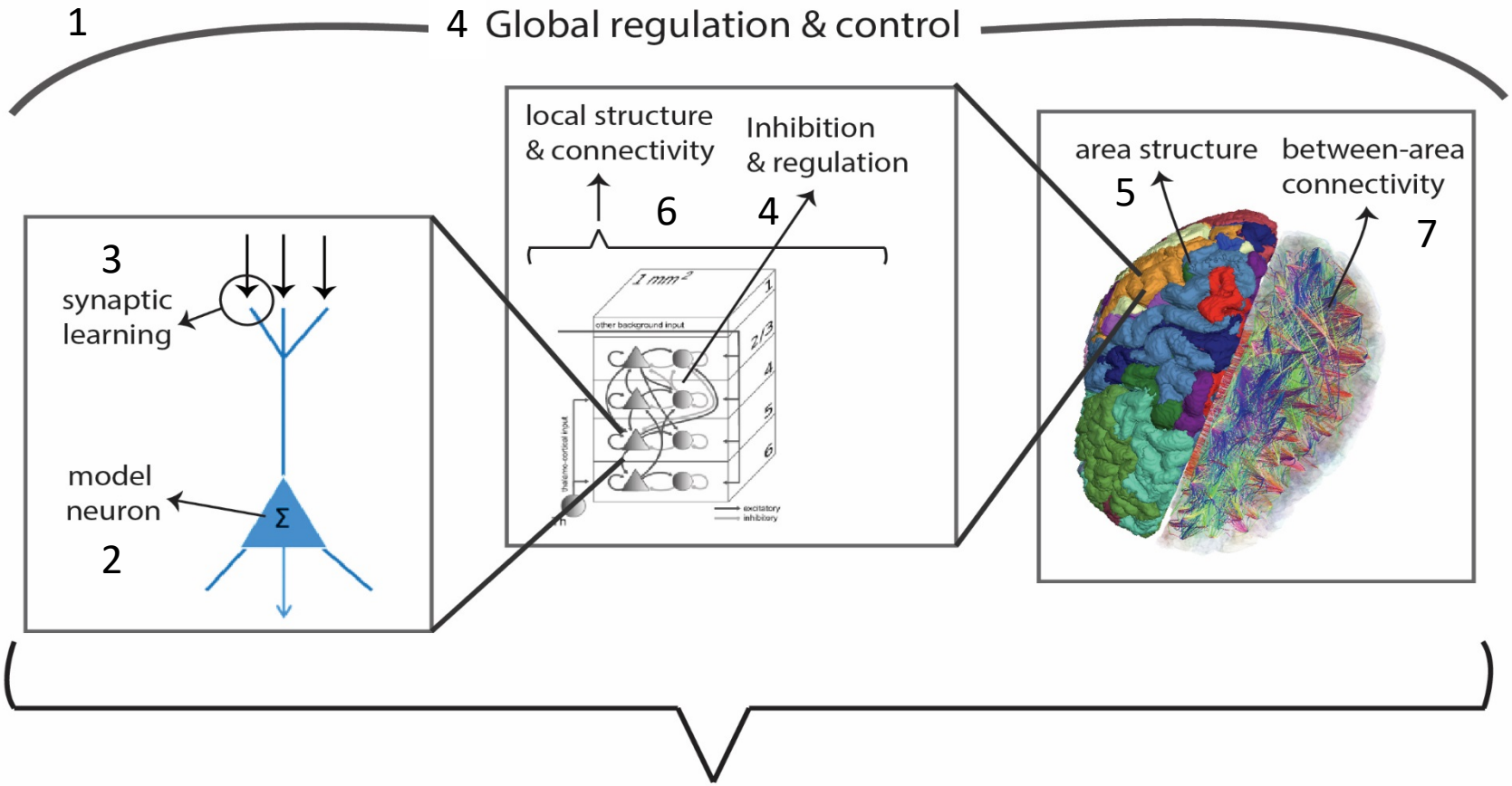
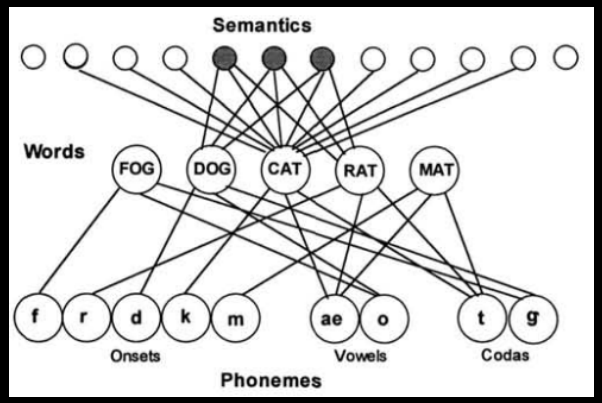

1
4 Global regulation & control
local structure
& connectivity
Inhibition
& regulation
6
4
3
synaptic
learning
model
neuron
2
Σ
area structure
between-area
connectivity
5
7
brain constrained model
simulates, explains and predicts brain activation
and behaviour in perceptual and cognitive tasks

Figure 1

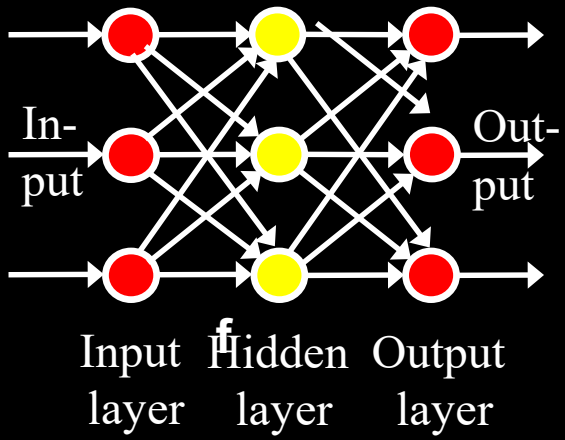
a Localist network model



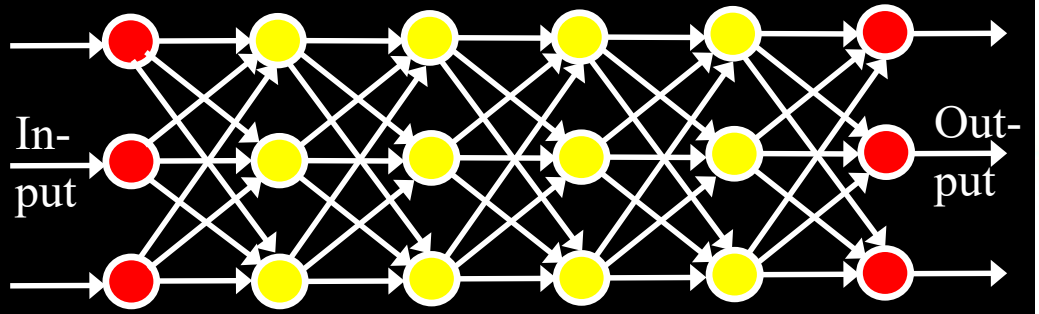
b Auto-associative matrix

	α	β	γ	δ	ε
α	1	1	1	0	0
β	1	1	1	0	0
γ	1	1	1	1	1
δ	0	0	1	1	1
ε	0	0	1	1	1

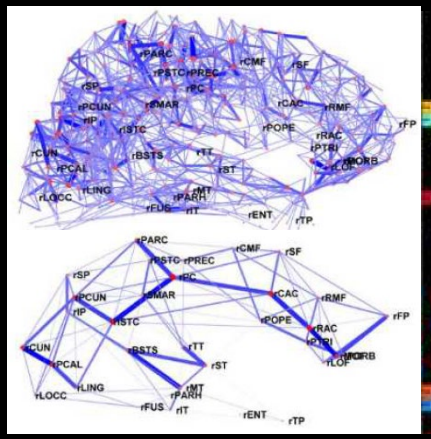
c PDP / 3 layer network



d Deep neural network



e Whole brain model



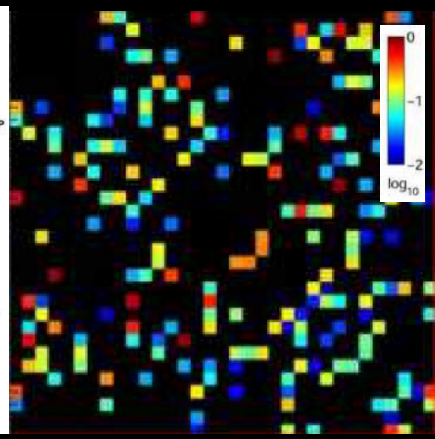


Figure 2

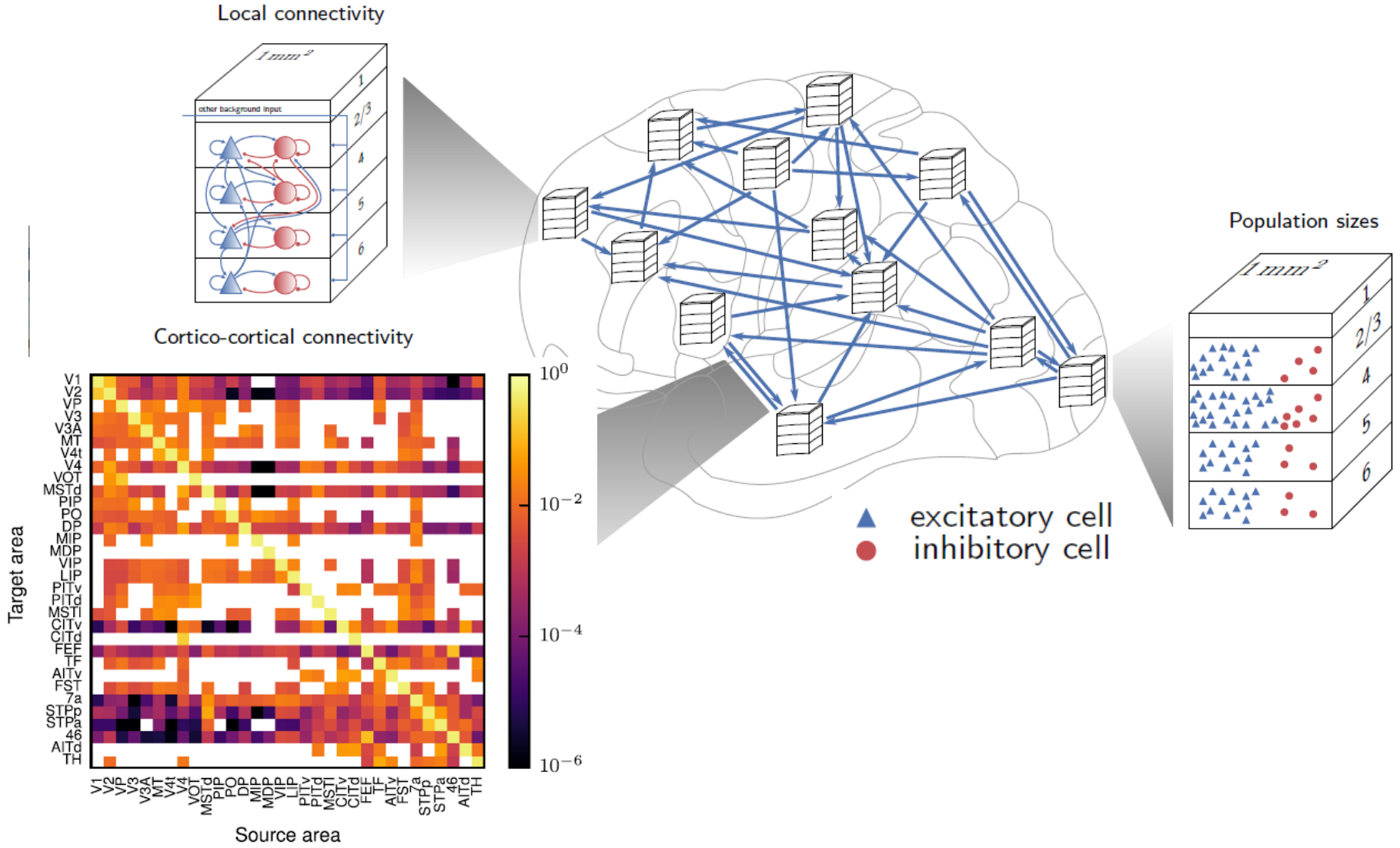

Local connectivity
1 mm²
other background input
1
2/3
4
5
6
Cortico-cortical connectivity
Target area
V1
V2
VP
V3
V3A
MT
V4t
V4
VOT
MSTd
PIP
PO
DP
MIP
MDP
VIP
LIP
PITv
PITd
MSTl
CITv
CITd
FEF
TF
AITv
FST
7a
STPp
STPa
46
AITd
TH
Source area
10^0
10^{-2}
10^{-4}
10^{-6}
excitatory cell
inhibitory cell
Population sizes
1 mm²
1
2/3
4
5
6

Figure 3

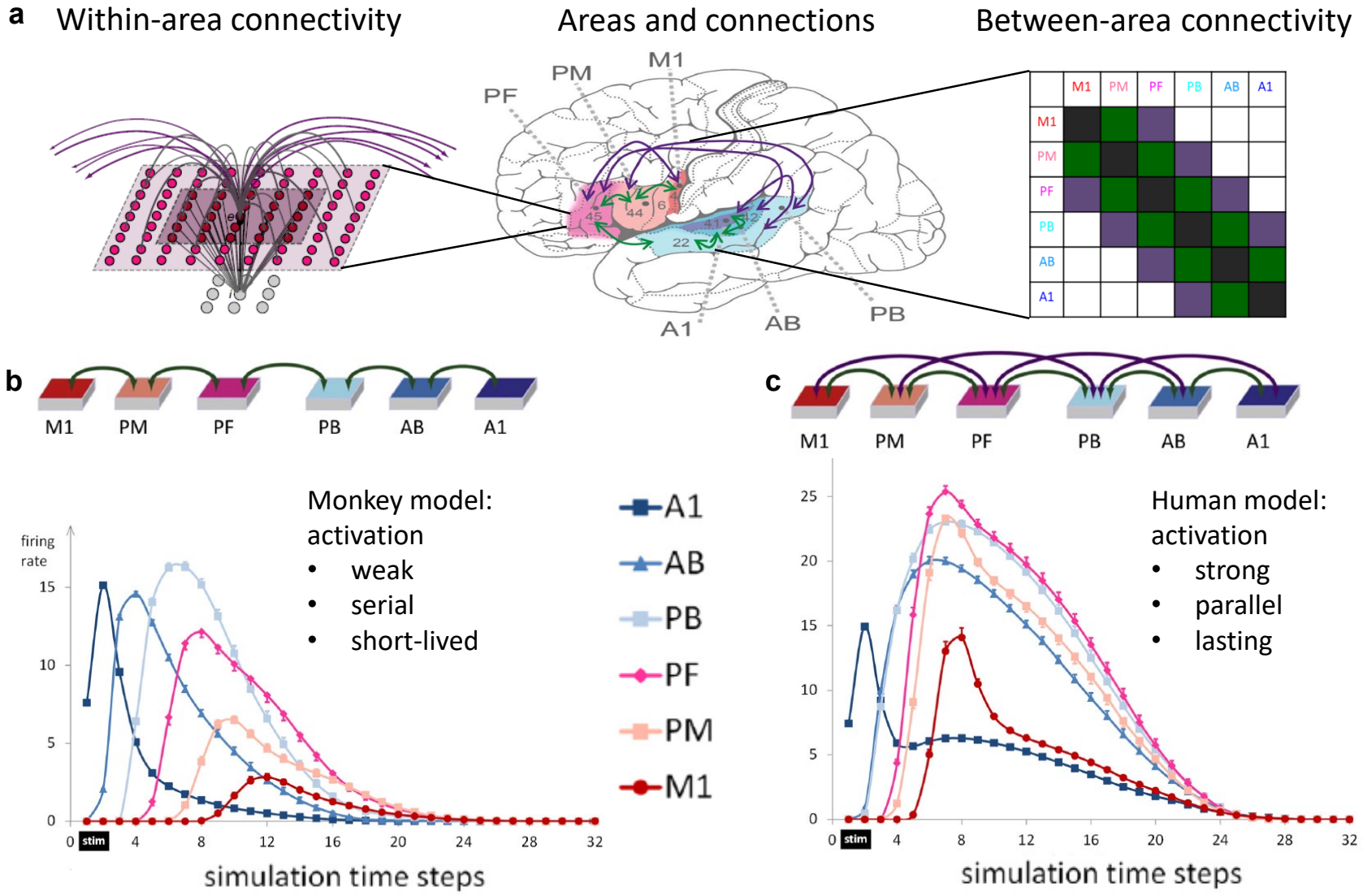

a
Within-area connectivity
Areas and connections
Between-area connectivity
M1
PM
PF
PB
AB
A1
b
c
Monkey model:
activation
• weak
• serial
• short-lived
Human model:
activation
• strong
• parallel
• lasting
firing rate
simulation time steps
stim

Figure 4

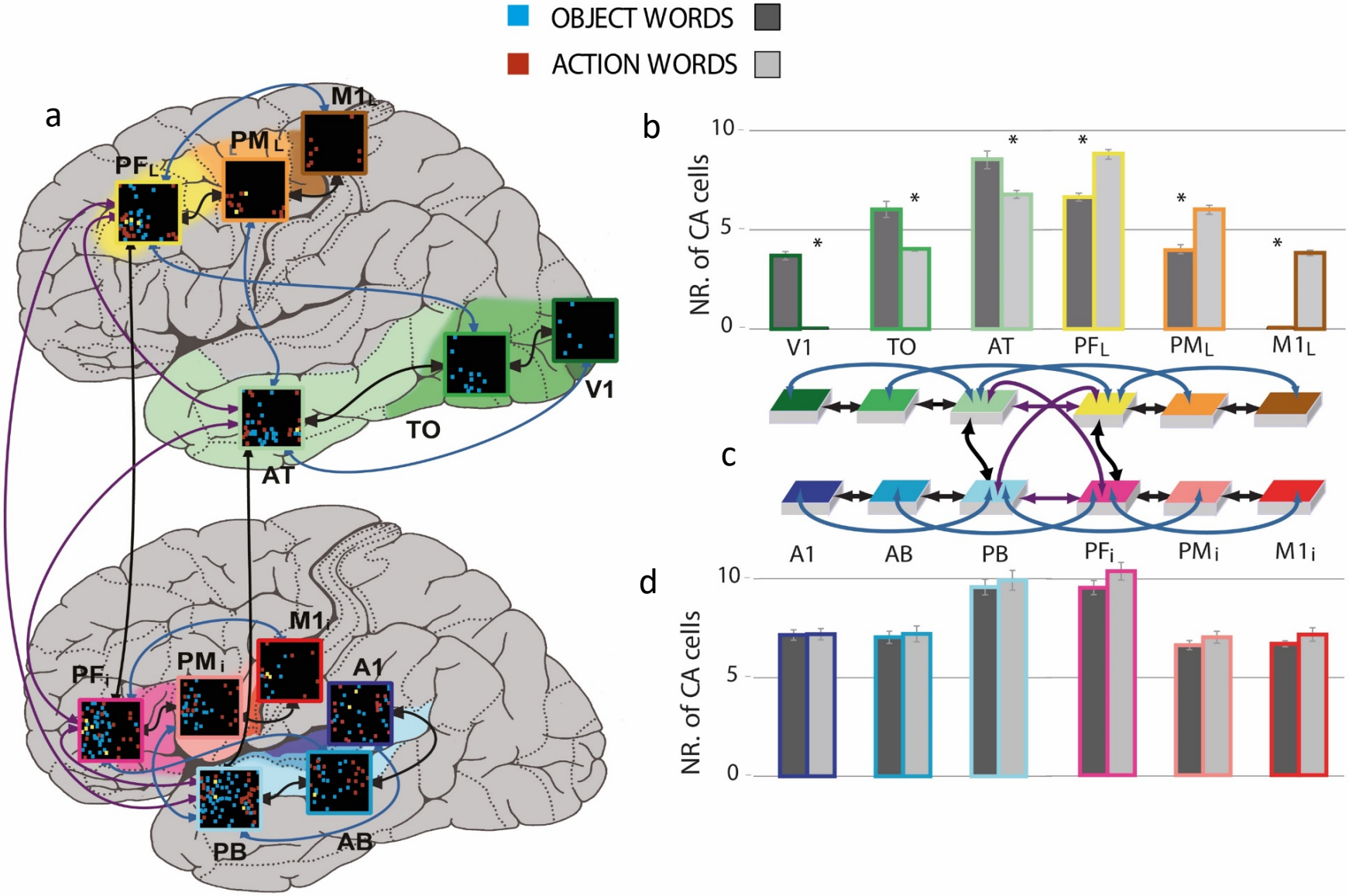

OBJECT WORDS
ACTION WORDS
a
M1L
PML
PFL
V1
TO
AT
M1i
PMi
PFi
A1
PB
AB
b
NR. of CA cells
10
5
0
V1
TO
AT
PFL
PML
M1L
c
A1
AB
PB
PFi
PMi
M1i
d
NR. of CA cells
10
5
0

Figure 5