Review article

Neural and phenotypic representation under the free-energy principle

Maxwell J.D. Ramstead [a,b,c,*], Casper Hesp [c,d,e,f], Alexander Tschantz [i,j], Ryan Smith [g], Axel Constant [b,c,h], Karl Friston [c]

[a] Division of Social and Transcultural Psychiatry, Department of Psychiatry, McGill University, Montreal, Quebec, Canada
[b] Culture, Mind, and Brain Program, McGill University, Montreal, Quebec, Canada
[c] Wellcome Centre for Human Neuroimaging, University College London, London, WC1N3BG, UK
[d] Department of Psychology, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, the Netherlands
[e] Amsterdam Brain and Cognition Centre, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, the Netherlands
[f] Institute for Advanced Study, University of Amsterdam, Oude Turfmarkt 147, 1012 GC Amsterdam, the Netherlands
[g] Laureate Institute for Brain Research, Tulsa, OK, USA
[h] Theory and Method in Biosciences, Level 6, Charles Perkins Centre D17, Johns Hopkins Drive, University of Sydney, NSW, 2006, Australia
[i] Sackler Centre for Consciousness Science, University of Sussex, Brighton, UK
[j] Department of Informatics, University of Sussex, Brighton, UK

ARTICLE INFO

ABSTRACT

The aim of this paper is to leverage the free-energy principle and its corollary process theory, active inference, to develop a generic, generalizable model of the representational capacities of living creatures; that is, a theory of phenotypic representation. Given their ubiquity, we are concerned with distributed forms of representation (e.g., population codes), whereby patterns of ensemble activity in living tissue come to represent the causes of sensory input or data. The active inference framework rests on the Markov blanket formalism, which allows us to partition systems of interest, such as biological systems, into internal states, external states, and the blanket (active and sensory) states that render internal and external states conditionally independent of each other. In this framework, the representational capacity of living creatures emerges as a consequence of their Markovian structure and nonequilibrium dynamics, which together entail a dual-aspect information geometry. This entails a modest representational capacity: internal states have an intrinsic information geometry that describes their trajectory over time in state space, as well as an extrinsic information geometry that allows internal states to encode (the parameters of) probabilistic beliefs about (fictive) external states. Building on this, we describe here how, in an automatic and emergent manner, information about stimuli can come to be encoded by groups of neurons bound by a Markov blanket; what is known as the *neuronal packet hypothesis*. As a concrete demonstration of this type of emergent representation, we present numerical simulations showing that self-organizing ensembles of active inference agents sharing the right kind of probabilistic generative model are able to encode recoverable information about a stimulus array.

1. Introduction

Living creatures are endowed with a representational capacity: they behave as if states of their bodies encode information or probabilistic beliefs about salient features of their environment, which they seem to leverage to generate contextually appropriate, adaptive behaviour (Egan, 2018; Kiefer and Hohwy, 2018; Ramstead et al., 2020, 2019a, 2019b). Most multicellular creatures have evolved specialised cell populations – nervous systems, comprising neural and nonneural (e.g., glial) cells – that are dedicated to realizing this representational capacity. Through continuous tuning of network connections, groups of neurons self-organize to form functionally integrated ensembles that look as if they are recapitulating the statistical structure of the environment, in a manner that organisms can leverage to generate adaptive patterns of behaviour.

Population codes, where neural ensembles look as if they encode

probabilistic beliefs about the statistical structure of the environment (i. e., probability densities over the latent causes of sensory data), are as ubiquitous as they are well documented (Pouget et al., 2000). Single neurons are known to respond preferentially to specific kinds of stimuli; e.g., neurons in primary visual cortex respond preferentially to specific orientations and luminance contrasts of a stimulus presented in their receptive field. Neurons also look as if they work together to infer the most likely causes of sensory input in distal sensory epithelia, such that stimulus information can only be fully recovered by examining activation patterns across many cells. This has been shown extensively in the visual system (Usrey and Reid, 1999). It is well known, for instance, that neurons in primary visual cortex look as if they infer the causes of stimuli (i.e., the patterns of photons) that impinge upon retinal neurons (Maunsell and Van Essen, 1983); that is, the average activity pooled over these neural populations looks as if they track features of the stimulus presented to retinal neurons. Something similar can be said of many other neural populations, e.g., place cells in the hippocampus (Bair, 1999; Borst and Theunissen, 1999) and topographically organized neurons in motor cortex (Tolhurst et al., 1983).

How are these coordinated feats of representation possible? How does specialised neural tissue, and indeed any kind of phenotypic state, come to encode information about the world in which the organism is embedded? Is it possible to formulate a general theory of how the confluence of brains, bodies, and their behaviours can represent the environment? Can we construct a general theory of phenotypic representation?

In this paper, we leverage the apparatus of the free-energy principle and its corollary process theory, known as active inference, to propose a general theory of neural representation. This framework opens, more broadly, onto a theory of phenotypic representation, where the brains, bodies, and behaviours of organisms – and indeed, the coordinated behaviour of groups of organisms – are cast as engaging in inference about the causes of their sensory states at several nested spatial and temporal scales.

The free-energy principle is a novel multiscale framework for studying belief-driven adaptive action (Friston, 2020). The free-energy principle provides us with a theory derived from first principles regarding what it means to exist as a system with a phenotype (i.e., to exist in a nonequilibrium steady state regime). On this account, to exist as a living system entails continually revisiting the neighbourhood of the same set of states (e.g., remaining within a certain range of body temperatures or ecological environments). Technically, such systems are endowed with a random dynamical or pullback attractor. This engenders a nonequilibrium steady state density that we can associate with the phenotype of a living system. Active inference, in turn, provides us with a (Bayesian) mechanics that explains how organisms remain within their phenotypic bounds (Ramstead et al., 2018). The framework rests on the Markov blanket formalism, which allows us to partition systems of interest into internal states, external states, and the blanket (active and sensory) states that render internal states conditionally independent of external states. This fundamental nonequilibrium Markovian structure (Friston et al., 2020a, 2020b) will guide our investigations here. We argue that representational capacities of living creatures can be formalized in a biologically realistic way via the active inference framework. We develop a generic and generalizable model of the representational capacities of living creatures or, indeed, particles with *particular* states (namely, internal states and their Markov blanket).

We draw on one prescient operationalization of this distributed representational capacity – the construct of a *neuronal packet* – and provide a proof of principle for the *neuronal packet hypothesis* (Yufik, 1998; Yufik and Friston, 2016). According to the neuronal packet hypothesis, cortical neurons perform inference about the causes of patterns in sensory neuron stimulation by forming neuronal packets, which are neural ensembles wrapped in a Markov blanket, such that neuronal

packets come to represent recoverable features of the stimulus. In one sense, the notion of a neuronal packet can be read as homologous to the notion of a neuronal group, ensemble or assembly that has a long pedigree in neurology and neuroscience (Buzsaki, 2010; Edelman, 1993, 1998; Gerstein and Kirkland, 2001; Hebb, 1949; Mountcastle, 1997; Nicolelis and Lebedev, 2009; Sherrington, 1911; von der Malsburg, 1981).

We argue that the representational capacity of creatures is a consequence of their Markovian structure. The Markovian partition of a system in a nonequilibrium steady state regime equips the internal states of a system with a dual-aspect information geometry. Internal states have an intrinsic information geometry that describes their trajectory over time in state space, as well as an extrinsic geometry based on probabilistic beliefs about fictive external states (that is, beliefs that may or may not correspond to true external states), encoded by internal states. Here, we show simulation results that concretely demonstrate in silico that – when suitably configured under the right probabilistic model – the self-organizing ensembles of active inference agents (ensembles of simulated neurons) that share the same model are able to encode recoverable information about a stimulus array. We argue that the configurations at which the system arrives through such self-organised coordination can be cast as a form of embodied inference (i.e., arriving at posterior state estimates over the causes of sensory stimulation).

Although we focus on neuronal ensembles, it is important to highlight that the account that we develop here, based as it is on the free-energy principle, can be applied to any group of living (sub-)systems that entertain a shared generative model. This means that the representational capacity that we discuss in this paper is general. It need not be restricted to cells, or even to animals, and has the resources to account for potential collective representational capacities across all kinds of self-organising systems; e.g., plant dynamics (Calvo and Friston, 2017).

Extant work on this topic has focused on the representational capacities of individual systems (Friston, 2013; Kuchling et al., 2019), sometimes composed of other systems (Palacios et al., 2020). In this paper, we extend this line of reasoning to *patterns of group activity* that enable a group of agents (be they neurons or other elements of larger systems that realise the appropriate dynamics) to *collectively represent* features of an external environment. This extension is crucial for understanding the intentional capacities of living systems, where hierarchical and distributed dynamics are ubiquitous.

2. Background: multiscale active inference

2.1. Active inference and Markov blankets

The free-energy principle begins from the observation that living creatures tend to revisit the neighbourhood of the same states of being (Ramstead et al., 2018). This is formalized as the claim that the dynamics of living creatures are underwritten by a nonequilibrium steady state (Friston, 2020). Living systems exist far from thermodynamic equilibrium, in a regime of characteristic states that we associate with a nonequilibrium steady-state density over the system's states, with high probability states corresponding to typical states given 'the kind of creature that I am', i.e., to those that are consistent with an organism's phenotype. The free-energy principle explains how living creatures remain alive (i.e., away from equilibrium), by continuously generating patterns of adaptive action that bring about preferred sensory data or observations compatible with their survival (Friston, 2020). Thus, according to active inference, to exist as a living thing means to continually produce sensory evidence of one's own existence (i.e., evidence that 'I am in or near a nonequilibrium steady state); what is sometimes called self-evidencing (Hohwy, 2016).

Active inference rests on a few fundamental, interlocking

mathematical constructs: these comprise the nonequilibrium steady state densities just discussed, as well Markov blankets, generative models, and variational free-energy (Ramstead et al., 2019a, 2019b; Ramstead et al., 2020, 2019a, 2019b). A growing literature appeals to Markov blankets to account for biological self-organization and the ability of living systems to remain alive over time (Constant et al., 2018; Friston, 2013, 2020; Hipólito, 2019; Kirchhoff et al., 2018; Kuchling et al., 2019; Palacios et al., 2020; Ramstead et al., 2018, 2019a, 2019b). This scheme gives us a principled definition of what it means to exist as a system (Kirchhoff et al., 2018; Ramstead et al., 2018). Minimally, to exist as a system implies a form of conditional independence with respect to the rest of the world; and Markov blankets operationalize this independence.

A Markov blanket defines a system of interest by identifying the states that mediate its probabilistic relations with the events and regularities of the environment in which it is embedded, and which cause or generate those sensory states. By inducing conditional independence (Pearl, 1988), Markov blankets enable us to define the boundaries between a system and its environment, and thereby delimit the system as such (Friston, 2013, 2020; Friston et al., 2015). We define a system of interest, delimiting it from the system in which it is embedded (e.g., a living creature in its ecological niche, or a cell in the intercellular milieu), by defining a third set of states – the Markov blanket itself – that mediate the causal influences between them. The Markov blanket is further partitioned into sensory and active states: active states affect, but are not affected by, external states; whereas sensory states affect, but are not affected by, internal states. Given this partition, internal states are independent of external states, given blanket states. See Fig. 1.

2.2. Generative models, variational free-energy, and the mechanics of inference

Active inference is a formulation of *Bayesian mechanics*, which tells us about the flow of internal and active states and how this keeps the Markov blanket in play (i.e., keeps the boundaries that separate an organism from its environment intact). For action and perception to be adaptive and knowledge-driven, some structures must exist that harness this knowledge in a form amenable for use in guiding adaptive action. These structures are the internal states of creatures, shielded behind their Markov blankets. Active inference allows us to cast perception and action (i.e., the flow of internal and active states, respectively, in response to perturbations from sensory states) as maximising the evidence for an implicit statistical model, which living creatures embody and which guides their production of adaptive action through the selection of (beliefs about) sequences of actions, or policies (Friston, 2020).

Formally, generative model is a joint probability density over the variables of a system; in the case of active inference, this a joint probability density over organismic states (so-called particular states, i.e.,

internal, sensory, and active states) and fictive external states (Ramstead et al., 2019a, 2019b). Under the free-energy principle, the generative model is just an interpretation of the nonequilibrium steady state density that we associate with the organism's phenotype (Friston, 2020). Heuristically, we can either look at this density from the point of view of the physics of flow, as the underlying manifold that structures the flow of states; or as a joint probability density over all systemic states (Ramstead et al., 2020). In brief, the presence of a Markov blanket in a system at nonequilibrium steady state means that we can look at the system's internal states as if they parameterize beliefs about (probability distributions over) external states. Because the external states now play the role of random variables they are often referred to as the latent or hidden states (i.e., hidden behind the Markov blanket). These hidden states can be read as explanatory fictions that the organism entertains to explain how its data were caused (Ramstead et al., 2020).

A modest representational capacity follows from this Markovian, nonequilibrium setup (Friston, 2020; Friston, Wiese, et al., 2020; Parr et al., 2020a, 2020b). Indeed, the Markovian structure just described leads to a dual-aspect information geometry that licences a description of internal states where they encode a conditional (Bayesian) belief about the external world. Information geometry is essentially concerned with measurement of spaces that represent (the sufficient statistics of) probability distributions. Simply put, the existence of the Markovian partition implies that, for any blanket states of a system, one can identify (an average over) internal states and a corresponding conditional belief (i.e., probability density) about external states. Thus, in addition to the 'intrinsic' information geometry of the system – which captures the probability of internal states over time – the Markovian structure endows the system with an 'extrinsic' information geometry, which captures beliefs about concurrent external states (Friston et al., 2020a, 2020b).

According to the free-energy principle, living systems enact dynamics (patterns of action and perception) that entail a generative model (Friston, 2012; Ramstead et al., 2019a, 2019b). What this means is that, mathematically, the behaviour looks as if it was produced by a specific causal process (a specific generative process), the structure of which is recapitulated in the generative model. In producing adaptive behaviour, living systems will look as if they are performing inference about the causal structure of their ecological niche, and maximising the evidence for a model of the world that they implicitly embody. More specifically, they do so by optimizing a bound on the (negative log) of this evidence known as the variational free-energy – which the organism can compute from the dynamics of its sensory states. By minimizing variational free-energy (a.k.a. maximizing model evidence), internal states of an organism come to encode a probabilistic 'best guess' about the causes of sensory input (Kiefer and Hohwy, 2018; Ramstead et al., 2019a, 2019b). This is known as the (variational) free-energy principle (Friston, 2010, 2020).

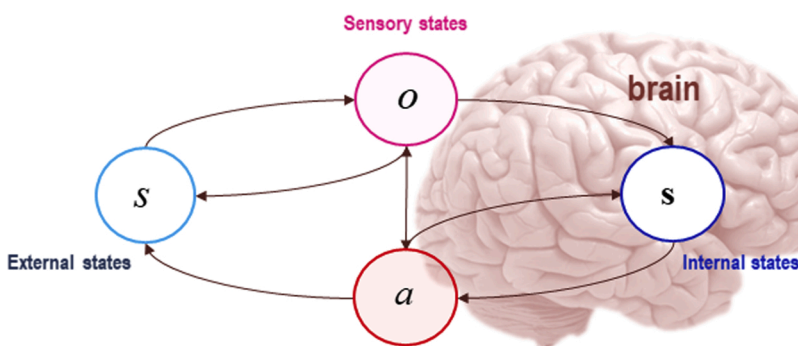To summarize, the entailment of a generative model by adaptive

Fig. 1. Markov blankets under active inference.
This figure depicts the statistical relations that define the Markov blanket. The existence of a Markov blanket induces certain conditional independences: the presence of the blanket partitions the system into internal states (denoted **s**), external states (denoted *s*), and the blanket states themselves – which render internal and external states independent of each other, when conditioned on blanket states. The Markov blanket itself comprises active states (*a*) and sensory states (*o*). From Ramstead et al. (2019a), 2019b.

behaviour, and the concomitant partition of the system into internal and external states separated by blanket states (and thereby the maintenance of the Markov blanket), leads to the induction of beliefs or knowledge about external states, parameterized by internal states. The physical states of the system then come to encode (the sufficient statistics) of distinct posterior beliefs about latent states. These are explanatory fictions or hypotheses about what is causing the sensory states. This means that the physical quantities that organisms embody encode the (parameters of) posterior or conditional density (i.e., Bayesian beliefs) over hidden or latent states that might best explain the manner in which sensory data were generated (Ramstead et al., 2019a, 2019b). The existence of a partition into internal, blanket, and external states thus also induces beliefs or knowledge about external states encoded by internal states and leads to the emergence of representational capacity in living creatures (Friston, 2012, 2020; Friston et al., 2020a, 2020b; Ramstead et al., 2020). The ensuing Bayesian mechanics suggests a deep if somewhat deflationary connection between existence and knowledge; namely, to exist is to know – in the sense of holding Bayesian beliefs (John Campbell, personal communication).

2.3. Nested Markov blankets and multiscale self-organization

This Markovian structure of living systems is iterated recursively across all levels of biological organization (Friston, 2020; Kirchhoff et al., 2018; Palacios et al., 2020). Since most living systems are composed of parts that are themselves living systems (e.g., organisms are made of organs, which are made of cells), the Markovian structure iterates recursively, allowing us to explain the dynamics of any system that exists at nonequilibrium steady state across levels of description – from cells to societies (Friston, 2020; Kirchhoff et al., 2018; Ramstead

et al., 2018, 2019a, 2019b). It is thus blankets of blankets, as it were, all the way up, and all the way down. See Fig. 2.

The strategy deployed by proponents of active inference to model self-organization at several scales is to show that some higher-level phenomenon (e.g., pattern formation) emerges from the coordinated inference dynamics of component phenomena at the subordinate scale, given that they share a generative model (Friston, 2013; Kirchhoff et al., 2018). In this work, each individual agent is equipped the same generative model (which is a metaphor for a shared genetic code). The parameters of this generative model encode knowledge about the kinds of signals that different cell types receive and send, as a function of the kind of cell they are. (We discuss this in detail in the Methods and Results sections below). Based on this shared knowledge, each cell in the ensemble is able to infer its place and its role (as an internal, active, or sensory state) in the higher-order system, as a function of the evidence that each cell accumulates that is indicative of such a role, in terms of probabilistic relations between roles as defined by the Markovian partition – e.g., internal states do not communicate directly with other internal states (which are shrouded behind their blankets), but can communicate with blanket states, etc.

In other words, an ensemble self-organizes to some target configuration because its components share the same expectations about the typical observations or data that are generated by the behaviour of it and other members of the ensemble. For a system of units sharing a generative model, the free-energy minimum for each cell coincides with the free-energy minimum for the ensemble (Friston, 2013; Friston et al., 2015; Kuchling et al., 2019; Palacios et al., 2020). Target morphologies emerge as the result of group inference operating on a shared generative model; where each unit comes to 'know its place' by inferring what role it must be playing, given the signals it is receiving from other units and
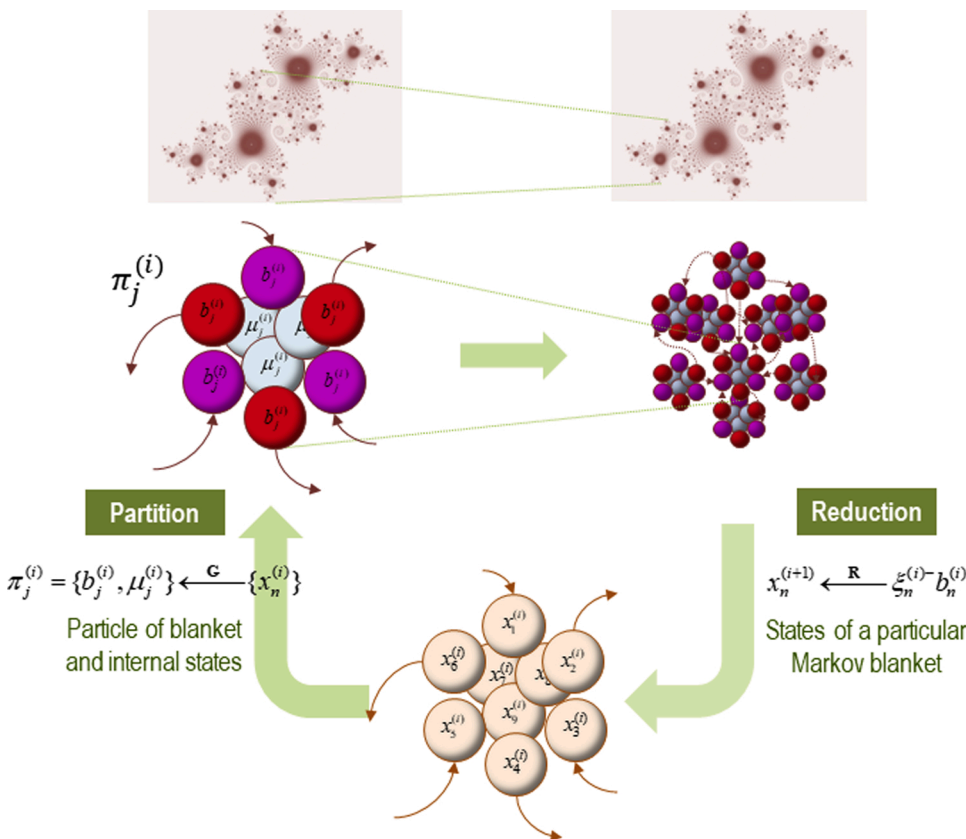

Fig. 2. Nested Markov blankets.
This figure depicts the recursive structure of nested Markov blankets. At each successive level of the nested hierarchy, superordinate dynamics at successively larger and slower scales emerge from, and constrain, dynamics arising at subordinate, smaller and faster scales. We can read the figure starting with the bottom panel. This panel depicts an ensemble of nine states – denoted x – which have not yet been differentiated into the partition prescribed by a Markov blanket. The formalism deploys renormalization group theory to derive a scale-free dynamics (Friston, 2020). In brief, we start at the bottom panel with a group of vector states (here, nine are depicted). We use a grouping operator **G** to group these states according to their conditional independences, by computing the Jacobian of their adjacency matrix. This leaves us with a particular partition of the same states, now grouped into blanket (sensory and active) and internal states. We then use a dimension reduction operator **R** to derive dynamics at the next scale. Here, the fast and stable modes of the blanket states and the internal states are dropped. The principal eigenstates of each blanket are taken to define new vector states at the scale above; and the process repeats until all states have been accounted for. Depicted in the upper panels are particles, which are defined in terms of the conditional dependences between these states and that define the Markov blanket of each particle. Here, a particle π is a mixture of blanket states (b) and internal states (μ). From Friston (2020).

those it sends.

2.4. Markov blankets in the brain and the neuronal packets hypothesis

As discussed above, Markov blankets have been used to define the boundaries and dynamics of systems interacting across scales (Friston, 2020; Hipólito, 2019; Kirchhoff et al., 2018; Palacios et al., 2020; Ramstead et al., 2018, 2019a, 2019b). This applies to the brain as well (Friston et al., 2020a, 2020b; Hipólito et al., 2020; Palacios et al., 2019; Parr and Friston, 2018). To start with, single neurons have their Markov blankets (Kiebel and Friston, 2011). The sensory states of the neuron are the synapses through which it receives signals from other cells. The active states of neurons are their axons and synaptic terminals through which they convey action potentials and patterns of neurotransmitter release onto other cells. Each neuron has a preferred receptive field profile or response vector that represents its preferred stimulus type, which we can associate with the afferent synaptic connections of the neurons that encode its likelihood function (i.e., that encode what inputs are those that provide evidence for the hypothesis it represents). Active inference can be used to explain neural behaviour at three different scales: the fast timescale of inference (changes in activation or depolarization), the slower timescale of associate learning (changes in synaptic efficacy or gain), and the even slower timescale of structure learning (structural change in dendritic architecture) (Kiebel and Friston, 2011). Empirically, there is emerging evidence for the self-evidencing behaviour of neuronal ensembles in terms of dendritic architectures; even in cell cultures (Isomura and Friston, 2018).

On the active inference account, the architecture of (nested) Markov blankets discussed above might form the basis of brain organization generally (Friston et al., 2020a, 2020b; Hipólito et al., 2020). The causal factors that figure in generative models are thought to correspond to anatomically segregated brain areas (Friston and Buzsáki, 2016; Parr and Friston, 2018). The evidence once thought to suggest modular processing in the brain may be better interpreted as evidence for structures enabling this factorization. In other words, functionally specialised brain regions may encode factors of a factorized probability density, which are conditionally independent but related to each other through their mean fields (Friston et al., 2017a, 2017b; Parr et al., 2020a, 2020b). This means that neuronal populations may encode distinct factors, each segregated by their Markov blanket (e.g., the identity of an object vs. its spatial location).

This leaves open the question of how *neural ensembles* come to encode information. The *neuronal packet hypothesis* (Yufik and Friston, 2016) proposes that neurons are able to encode information about external causes by forming coherent, bounded ensembles, enshrouded behind a Markov blanket. The self-organization of a Markovian ensemble of neurons in response to perturbations (action potentials) coming from sensory neurons can thus be interpreted as a form of posterior inference, with each packet encoding a different hypothesis (Yufik, 2019; Yufik and Friston, 2016). This is broadly consistent with theories of structural representation, according to which neurons encode exploitable information about a target domain by mirroring the (mathematical or statistical) structure of that target domain (Kiefer and Hohwy, 2018, 2019; Kiefer, 2020). The neuronal packet hypothesis goes further and proposes that the capacity of an organism to infer and interpret the causes of its sensory states results from the binding together of neuronal populations through the formation of a boundary (i.e., a Markov blanket) around the neuronal packet. To 'mobilise' such a packet – in response to sensory perturbation – would be to understand a world composed of interacting causes (Yufik, 2019; Yufik and Friston, 2016).

Empirical evidence supports the neuronal packet hypothesis, suggesting that neurons process information by forming task-dependent, transient coalitions on the fly; for an extensive review of the empirical evidence for this type of 'neural reuse' (Anderson, 2014). Thus, brain activity arises from coalitions of neurons functioning together as a cohesive unit, in a context- and goal-dependent manner. This allows for the combination of information about more or less independent causal factors encoded by the physical states of brain networks (Parr et al., 2020a, 2020b). Neurons are thus trying to form functional coalitions dynamically, to bring to bear the factors that they encode in a multifactorial coordinated inference; see also Park and Friston (2013). Strikingly, evidence for the Markovian structure of brain dynamics comes from recent work in neuroimaging, which demonstrates that brain regions form Markovian 'parcels' (Friston et al., 2020a, 2020b).

In recent active inference models, neurons are endowed with the prior belief (encoded in their shared generative model) that, in the absence of external perturbations, their signalling profiles will align or synchronise (Palacios et al., 2019). This might provide the initial impetus for packet formation. Each neuron is equipped with a generative model that allows it to infer its functional identity (or role in the Markovian partition) in the superordinate ensemble. It thus follows that perceptual inference arises as the neuronal system self-organizes into coherent packets in response to sensory perturbations.

One further phenomenon of interest that may be worth considering in this context is structural self-organization – and associated functional segregation (Zeki, 2005) – during early development and neuronal differentiation. Instead of neurons being genetically hard-wired to come to represent specific stimuli, some work suggests that common patterns of functional segregation may be more weakly specified by the inherited constraints on large-scale axonal structure (e.g., the path of optic nerve signals to visual cortices via the thalamus), where the functional role of a neuron (or of a locally connected population of neurons) instead emerges from self-organization in response to the statistics of the signals it receives early in development (Bienenstock et al., 1982; Peters and Yilmaz, 1993; Sin et al., 2002; Sur et al., 1988; Sur and Rubenstein, 2005; von der Malsburg et al., 2010; Von Melchner et al., 2000; Zeki and Shipp, 1989). This has been most starkly demonstrated in studies of experimentally manipulated large-scale neuronal pathways, e.g., studies that redirect visual signals from the optic nerve to the medial geniculate nucleus and auditory cortices. Remarkably, in these cases, populations of auditory cortex neurons become sensitive to specific visual properties similar to those normally seen in visual cortex. Other studies of developmental windows, such as work showing selective blindness to particular visual features (e.g., vertical contours) not present in the early environment, and associated pruning of the necessary synaptic connections to implement the inference of such properties, further illustrates this point (Blakemore and Cooper, 1970; Buonomano and Merzenich, 1998; Huang and Reichardt, 2001; Muir and Mitchell, 1973; Rice and Barone, 2000; Singer et al., 1981; Thoenen, 1995). Thus, there is a strong sense in which (under broad genetic constraints on large-scale axonal wiring patterns) neural populations infer what specific functional role they play based on the early inputs they receive.

2.5. Neural representation in the active inference framework

This paper builds on the technical developments of active inference models of hierarchical self-assembly and self-organization (Friston et al., 2015; Kirchhoff et al., 2018; Palacios et al., 2019, 2020). This work has shown that a Markov blanket emerges naturally around a group of active inference agents that share the right kind of generative model (Friston et al., 2015; Kirchhoff et al., 2018; Kuchling et al., 2019). The strategy deployed in these papers is to recast the problem of superordinate boundary formation – the emergence of a Markov blanket –as the consequence of inference conditioned on a shared generative model.

Extending this work, our strategy in this paper is to explain the self-organization of neural tissue into neuronal packets at a superordinate level by modelling coordinated inferences of single neurons sharing a generative model. The technical novelty in our setup is that each

individual neuron must now infer to which higher-order Markov blanketed system it belongs, in addition to the role that it plays in that superordinate system. This role in the superordinate ensemble is implemented by leveraging the partition of the system into different kinds of states, which follows from the existence of a Markov blanket at the superordinate scale (and beliefs about membership in that blanketed system): individual neurons must infer whether they play the role of a sensory state, an active state, or an internal state, where each of these roles is defined in terms of the conditional independencies that define the Markov blanket.

We model neuronal packet formation as the result of inferences about membership to a higher-order Markov blanketed ensemble, which can be cast as a form of inference. Each neuron has an individual response vector; that is, each neuron responds preferentially to some features being present in their receptive field. When sensory neurons detect their preferred stimulus in their receptive field, they send signals (i.e., action potentials) to cortical neurons, e.g., neurons in the primary visual context receiving signals from sensory neurons in the retina (through the lateral geniculate nucleus). We model the later type of neuron in this work. We hypothesize that cortical neurons use these signals as evidence for a specific inference: *that they are the internal states of a neuronal packet at the superordinate level*. Neurons exchange signals as well, which express their beliefs about blanket membership at the superordinate scale. This group inference and communication leads to the formation of a neuronal packet, which encodes a hypothesis about what might have caused the sensory impressions on visual receptors. Neuronal packet membership can be seen as a measure of evidence for that hypothesis. Thus, neuronal packet formation corresponds to the self-organization of neural ensembles enabled by inferences about membership to a higher-order Markov blanketed system. This provides a generalizable mechanism for distributed neural representations, premised on a nonequilibrium, Markovian, and recursively nested architecture.

In short, we hypothesize a specific architecture that links neurons in the sensory epithelia to neurons in the brain. The response vector of sensory neurons encodes the kinds of target features in their receptive fields to which they are most sensitive (e.g., orientation and contrast). This sensitivity means that, in the presence of their preferred stimulus, the rate at which action potentials are generated increases. Our hypothesis is that, when they detect increases in particular signalling patterns from sensory neurons, cortical neurons take this as evidence to *infer that they are internal states of a higher-order ensemble*. Essentially, on this account, the action potentials generated by each cortical neuron broadcast to other cortical neurons a signal that conveys its beliefs about whether it is an internal state of a superordinate ensemble.

3. Methods

To concretely demonstrate our proposed formulation, we simulate $N = 256$ 'particles' operating in a two-dimensional phase space, where the location of a particle is denoted $x \in \mathbb{R}^2$ and the equations of motion are largely based on those used by Palacios and colleagues (2020). Each particle – representing an individual neuron – is modelled as an active Bayesian agent, partitioned into sensory s, active a and internal μ states. The internal states parameterize beliefs about (i.e., a probability density over) hidden external causes of sensory input and are optimised using Gaussian filtering (Palacios et al., 2020). In other words, they are updated in order to minimise variational free-energy. In the current context, the internal states parameterize beliefs about the identity of the particle $\psi \in \mathbb{R}^4$, where each identity corresponds to membership in a superordinate ensemble.

Particles emit signals $\psi \in \mathbb{R}^4$, where each signal corresponds to a particular identity. In particular, each particle emits signals ψ as a function of their internal states μ:

$$\psi = \sigma(\mu) \tag{1}$$

Where σ is the softmax function. The internal states $\mu \in \mathbb{R}^4$ of a particle encode model evidence (negative free-energy) for their identity, such that the distribution ψ_i represents the beliefs of a particle about its own identity. Particles thus transmit beliefs about their respective identity. These transmitted signals are sensed by neighbouring particles. Signals decay exponentially as a function of distance in their shared phase space (e.g., as in local vs. long-range synaptic connections within a cortical region or diffuse neuromodulator release), such that the local signal intensity $m \in \mathbb{R}^4$ for particle is given by:

$$m = \sum_{j}^{N} \psi_j \exp\left(-\frac{1}{2}\Delta \overrightarrow{x}_j^2\right) \tag{2}$$

where the vector between particle locations is given by $\Delta \overrightarrow{x}_j = \overrightarrow{x}_j - \overrightarrow{x}$.

We introduce two numerical innovations with respect to the simulations reported by Palacios and colleagues (2020). Rather than particles sensing local signal intensities directly, they sense the (2D) spatial flow vector of the four signals, $\overrightarrow{R} \in \mathbb{R}^2 \times \mathbb{R}^4$, the relative rate of change in the signal as a function of location, or its proportional intensity gradient:

$$\overrightarrow{R} = \frac{\nabla_x m}{m} \tag{3}$$

By combining Eqs. (1–3), we can show that the flow vector of any exponentially decaying variable satisfies:

$$\overrightarrow{R} = \sum_{j \neq i}^{N} \Delta \overrightarrow{x}_j \, \sigma\left(\mu_j - \frac{1}{2}\Delta \overrightarrow{x}_j^2\right) \tag{4}$$

This equation shows that, for any particle, the flow vector $\overrightarrow{R}$ points directly to the (intensity-weighted) signal source. Moreover, the magnitude $R = |\overrightarrow{R}|$ of this vector is proportional to the distance from the signal source. These are remarkable properties for a signal that can always be computed locally, given the reasonable assumption of an exponential decrease of signal intensity with distance.

As well as sensing the flow vector magnitude s_R, each particle internally senses the signal it transmits s_ψ, such that the complete sensory state s is:

$$s = \begin{pmatrix} s_\psi \\ s_R \end{pmatrix} = \begin{pmatrix} \psi + \Omega_\psi \\ R + \Omega_R \end{pmatrix} \tag{5}$$

where Ω_ψ and Ω_R are Gaussian fluctuations.

To enable self-organising behaviour, we equip the particles with prior beliefs about the magnitude of flow vectors R. As Bayesian agents act to realise prior beliefs (i.e., maximise model evidence), expectations about R will promote behaviour (i.e., movement in state space) that realises the corresponding belief about distance from the intensity-weighted signal source.

We specify a unique prior $P(R|\psi)$ for each signal identity ψ, meaning that the signals (flow vector magnitudes) that are expected by each particle depend on their inferred identity:

$$P(R|\psi) = \mathcal{N}(\mathbf{R}_\psi, \Pi_R) \tag{6}$$

where $\mathbf{R}_\psi \in \mathbb{R}^4$ is a set of four conditional means for each signal, given a particular identity ψ and given the corresponding set of precisions Π_R (inverse variances). Because R is directly proportional to the (intensity-weighted) distance to the signal source, the Markov blanket structure of the superordinate ensemble can be encoded in these identity-specific expectations $\mathbf{R}_\psi$:

$$\mathbf{R}_\psi = \begin{bmatrix} D & D & 2D & 2D \\ D & D & D & 2D \\ 2D & D & D & D \\ 2D & 2D & D & D \end{bmatrix} \qquad (7)$$

Where D is a scaling parameter. Effectively, this matrix encodes the compatibility of spatial proximity between the four identities. Specifically, particles can (infer they) belong to one of two higher-order ensembles, where these ensembles are themselves superordinate particles consisting of internal and blanket states. Particles can form beliefs about whether they are an internal state ψ_1 or a blanket state ψ_2 of one superordinate Markov blanket, or alternatively, a blanket state ψ_3 or an internal state ψ_4 of the alternative superordinate Markov blanket. These identities differ only in terms of their associated expectations $P(R|\psi)$, where inferring that one is an internal state ψ_1 of one Markov blanket involves the expectation of being far away (distance of 2D) from particles that are part of the other superordinate particle, $\psi_{3,4}$, and vice versa. Particles which identify as ψ_2 or ψ_3 correspond to the blanket states that separate the internal states of the superordinate ensembles. Accordingly, blanket states ψ_2 are compatible with proximity (distance of 1D) to the corresponding internal states ψ_1 as well as the blanket states ψ_3 of the other ensemble. Reciprocally, blanket states ψ_3 are compatible with proximity to internal states ψ_4 and blanket states of the other ensemble ψ_2.

The agents have a categorical belief distribution over identities and their actual expectations are a Bayesian model average of the expectations corresponding with each identity:

$$\mathbf{R} = \mathrm{E}_\psi[R] = \psi\mathbf{R}_\psi \qquad (8)$$

We now derive updates for the model evidence for internal states μ. The flow of the internal states is given by gradient flows on variational free-energy F:

$$f_\mu(\tilde{s}, \ \tilde{a}, \tilde{u}) = (Q_\mu - \Gamma_\mu) \, \nabla_\mu F \qquad (9)$$

Here, the tilde indicates generalised coordinates, containing both the variable and all its temporal derivatives. We refer readers to Palacios and colleagues (2020) for a full description. We consider a first-order approximation for small Gaussian fluctuations; in which case we obtain the rate of change of model evidence μ:

$$\dot{\mu} = \nabla_\mu \mathbf{R} \cdot \Pi_R \varepsilon_R + \nabla_\mu \psi \cdot \Pi_\psi \varepsilon_\psi - \Pi_\mu \mu \qquad (10)$$

Here, Π_R, Π_ψ, Π_μ are sensory and internal precisions (inverse variances) that normalise model evidence μ and the prediction errors ε_R, ε_ψ:

$$\varepsilon_\psi = s_\psi - \psi$$

$$\varepsilon_R = s_R - \mathbf{R} \qquad (11)$$

In other words, the model evidence for each particle identity varies with the normalised prediction error, where the prediction error quantifies the difference between the (magnitude of the) sensed flow vector and the prior beliefs about the magnitudes of flow vectors.

Finally, we move on to consider active states, which correspond to a chemical emitter ($a_\psi \in \mathbb{R}^4$) and motor actuators ($a_x \in \mathbb{R}^2$) that move the particle through phase space x. The flow of active states mirrors that of internal states:

$$f_{\tilde{a}}(\tilde{s}, \ \tilde{a}, \tilde{u}) = (Q_a - \Gamma_a) \, \nabla_{\tilde{a}} F \qquad (12)$$

Intracellular self-communication serves as a minimal type of memory that is regulated by a chemical emitter that communicates identity beliefs with normalised Gaussian fluctuations (mentioned above):

$$a_\psi = \psi$$

$$\dot{a}_\psi = -\Pi_\psi \varepsilon_\psi \qquad (13)$$

As discussed, motor actuators regulate the spatial location of cells. We still neglect friction for now ($\Pi_x = 0$), but we increase the realism of cell movement by choosing for a second-order approximation: $\tilde{a}_x = \begin{pmatrix} x \\ \dot{x} \end{pmatrix}$, where $\dot{x}$ is the velocity. Since we are calculating the flow of both position and velocity as defined by the free-energy functional, this model also incorporates acceleration:

$$\dot{x} = -\nabla_x \overrightarrow{R} \cdot \Pi_R \varepsilon_R$$

$$\ddot{x} = \nabla_x \overrightarrow{R} \cdot \lambda \Pi_R \dot{\mathbf{R}} \qquad (14)$$

where λ parametrises the autocorrelation of fluctuations over time, which sets the smoothness of fluctuations and regulates the magnitude of second-order effects (see Friston et al., 2010 for its derivation) and $\dot{\mathbf{R}}$ is the temporal derivative of expectations about R:

$$\dot{\mathbf{R}} = \mathbf{R}\psi(1 - \psi)\dot{\mu} \qquad (15)$$

Collecting all results, we can now integrate over time and use a Taylor series expansion to simulate trajectories of particles over time in terms of x, a_ψ and μ:

$$x_{t+1} = x_t + \dot{x}_t \Delta t + \frac{1}{2}\ddot{x}_t \Delta t^2$$

$$a_{\psi,t+1} = \psi + \Pi_\psi \Omega_\psi \Delta t$$

$$\mu_{t+1} = \mu_t + \dot{\mu}_t \Delta t \qquad (16)$$

In summary, all we have done here is write down a relatively simple generative model of what any neuron or particle expects to sense, depending upon its identity (e.g., functional selectivity) and spatial relationship to other particles that share the same generative model. This generative model means that one can compute variational free-energy gradients and the ensuing flow of particular states, which can be read as neuronal dynamics or movement. This dynamics will necessarily show self-organising, self-assembling, and self-evidencing aspects because it minimises a variational bound on the evidence for a (shared) model of the external milieu. Crucially, the external milieu is made up of particles that are all trying to do the same thing and thereby fulfil each other's expectations. Technically, because we are dealing with continuous states and Gaussian fluctuations, the variational free-energy gradients become prediction errors. This can be seen in the above equations where the flow of internal and active states is a function of various prediction errors. Biophysically, these prediction errors would correspond to the setpoints of electrochemical variables, such that the particles keep moving when prediction errors are all zero and each particle has found its variational free-energy minimum (Fig. 3)

In order to illustrate the concepts outlined in the previous sections, we constructed two sets of numerical experiments. These will illustrate, respectively:

- The emergence of a two Markov blanketed systems, which results from the application of an external field that provides evidence of ensemble membership ($\psi_{1,2}$) to a small subset of particles.
- The transient and dynamic nature of Markov blanket formation in response to changing external fields.

4. Results

4.1. The emergence of a neuronal packet

The results of the first series of numerical simulations are presented
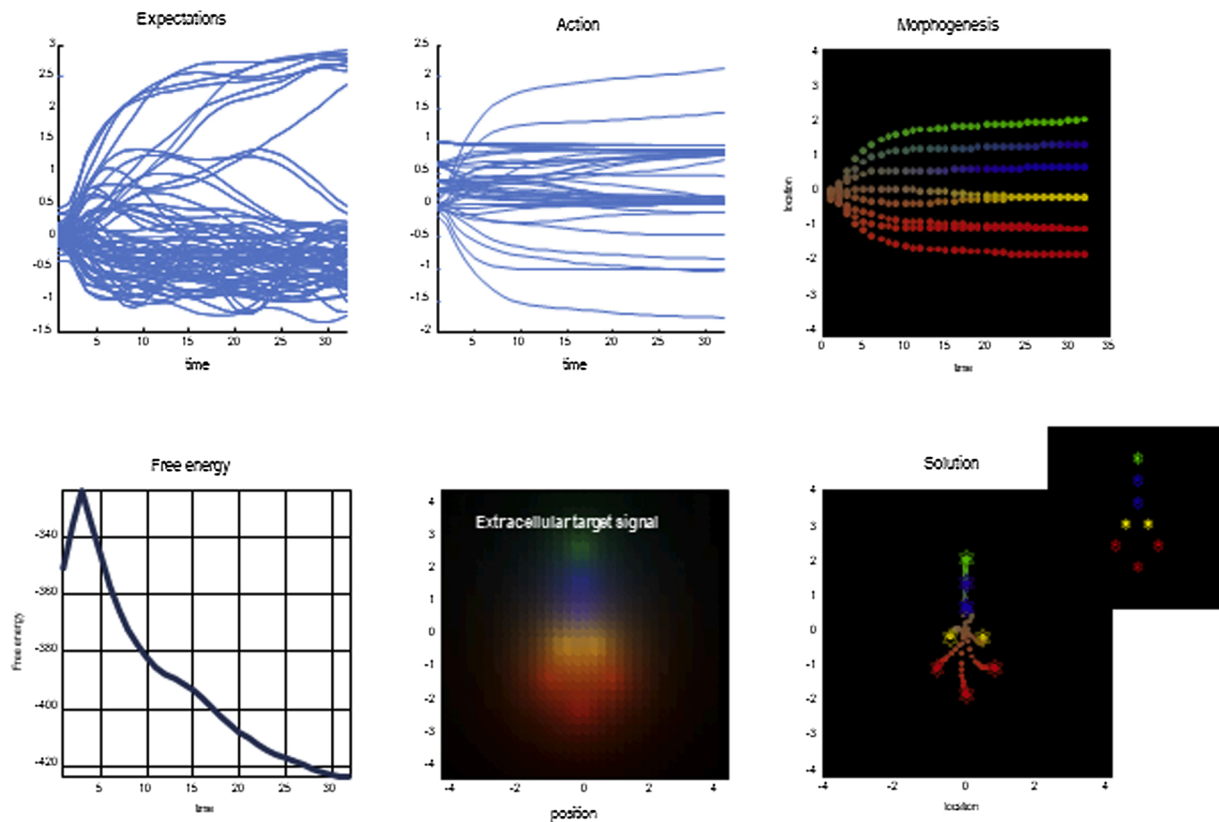
Fig. 3. Morphogenesis under the free-energy principle.

This figure depicts the numerical results reported in Friston et al. (2015), which were expanded upon by Palacios et al. (2020) and by Kuchling et al. (2019), and which are described further in the text. The bottom right panel depicts the target configuration and the movement of cells towards this configuration, and the central bottom panel depicts the signalling profile associated with each target position, which each cell uses to find its place. The top panels display the expectations of each cell as it moves towards the final position. The bottom left panel depicts the free-energy of the ensemble. Free-energy increases initially as each cell figures out what kind of cell it is, and then drops to the global minimum, which coincides with each cell finding its place in the target configuration.

in Fig. 4. In these simulations, a subpopulation of the individual neurons was stimulated to induce the precise posterior belief that they were states of one Markov blanketed system at the superordinate level ($\psi_{1,2}$). This was achieved by coupling the selected neurons to an external field, which, as described as above, is a metaphor for signals received from neurons in the sensory epithelia of the organism.

We introduce a stimulus function $f_{stim}(x) \in \langle 0, 1 \rangle$ that specifies the stimulus intensity (between 0 and 1) as a function of phase space position. This function determines the evidence a neuron receives for being a member of the first ensemble $\psi_{1,2}$, while the complement of this function $1 - f_{stim}(x)$ determines the evidence a neuron receives for the alternative ensemble $\psi_{3,4}$.

At the start of the simulation, neurons were initialised with uniform beliefs about their identity. The stimulus field was then introduced at time step $T = 10$, causing the neurons within this field to quickly form beliefs that were members of the first superordinate ensemble ($\psi_{1,2}$). Conversely, the remaining neurons formed beliefs that they are members of the second superordinate ensemble ($\psi_{3,4}$). Over the course of the simulation, the shared expectations about proximity relations $\mathbf{R}_\psi$ caused neurons to converge to either internal ($\psi_{1,4}$) or blanket ($\psi_{2,3}$) states. Specifically, neurons which were not in spatial proximity to states of the opposing higher-level ensemble formed beliefs that they were internal states, whereas neurons which were in spatial proximity to both ensembles formed beliefs that they were blanket states. This caused the first Markov blanket to form around the field introduced by the stimulus function, and the second Markov blanket to envelope the first. The

emergence of this structure meant that when conditioned on their blanket states ψ_2, the internal states ψ_1 of the first Markov blanket were independent of the states of the second Markov blanket $\psi_{3,4}$, and vice versa.

These results may also speak to the role of local field potentials in neural synchronization. The prior belief that they will synchronise with each other in the absence of stimulation lends to the individual neurons that make up a neuronal system a kind of coherence, which gives it the look and feel of a fictitious 'centripetal' force. Conversely, when neurons are stimulated by contradictory driving inputs, they will self-organize into distinct and segregated groups, which will look like a 'centrifugal' force. The group inference that realizes population coding effectively balances these two opposing tendencies in self-organization.

4.2. Splitting ensembles and multiple ensembles: towards an account of dynamic, transient Markov boundary formation

We move on to consider dynamic nature of Markov blanket formation in response to changing external stimulation. Following the first set of simulations, we initialise neurons with uniform beliefs about their identity and introduce an external stimulus at $T = 10$. However, we now transform the external stimulus at $T = 200$. This change in stimulation causes a subpopulation of neurons to form incompatible beliefs and promotes the reorganisation of the superordinate ensembles.

The results of these simulations are shown in Fig. 5 and 6, which differ only in the stimulus function used. As with the first set of
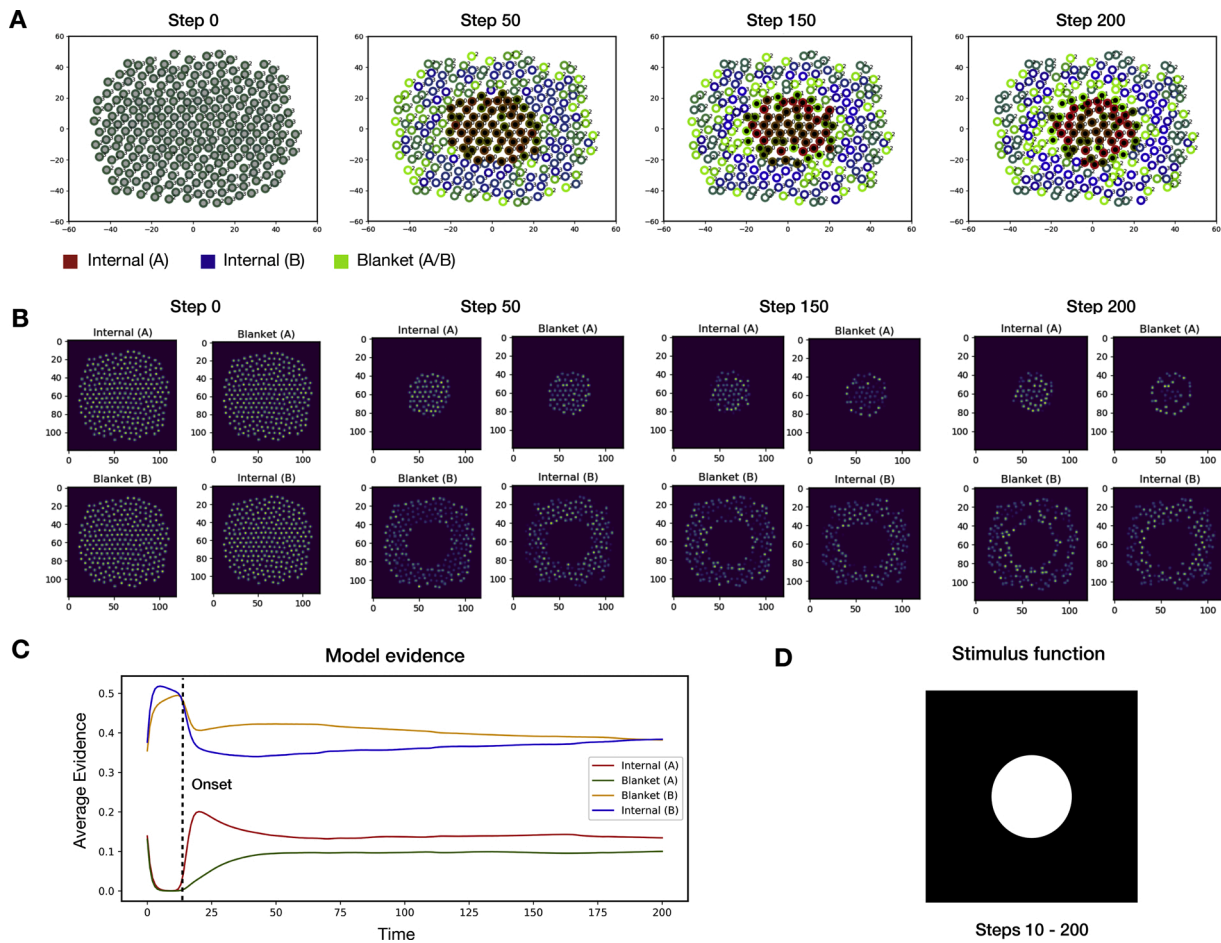
Fig. 4. The emergence of a neuronal packet.

This figure depicts the first series of numerical results, which demonstrate the emergence of a Markov blanket in response to external stimulation – or coupling with an external field. **(A)** A visualisation of individual neurons at different times during the stimulation. The colour of a neuron represents its beliefs about membership in superordinate Markov blankets, while its location denotes its position in phase space. Black and white represent distinct Markov blankets; while line colour represents role in the Markovian partition (as internal or blanket states of the respective blanketed ensemble). At the start of the stimulation, neurons are initialised with uniform beliefs about their identity. At time step $T = 10$, an external stimulus is introduced (described in the main text). The external stimulation causes neurons to divide into two sets of internal and blanket states, organised around the stimulus. Over time, shared expectations about what each neuron should sense cause the neurons to divide into internal and blanket states. The internal states of the first ensemble (A) organise around the stimulus centre, while the blanket states of this ensemble organise around the edges of the stimulus. Of the remaining neurons, those which are in spatial proximity to the first ensemble (A), or are at the external boundary of the ensemble, form beliefs that they are blanket states of the second ensemble (B), while the other neurons form beliefs that they are internal states of that ensemble. The external stimulation thus induces the formation of two Markov blankets, the internal states of which are conditionally independent of one another, conditioned on the blanket states. **(B)** A visualisation of the signals emitted by neurons, averaged over the ensemble, for each location in phase space. In panels (A) and (B), the two axes represent the two abstract states of a two-dimensional state space. For simplicity, here, we assume this is a physical space where the agents move, with real values $\mathbb{R}^2$. **(C)** The ensemble-average evidence for each type of identity as a function of time. **D)** The stimulus function used in this simulation, where the white denotes the stimulus intensity.

simulations, these results show the emergence of two Markov blanketed systems around the initial stimulus. However, the results additionally demonstrate that once the location of the stimulus changes, the entire ensemble self-organizes anew. Specifically, a new Markov blanket organises around the location of the new stimulus, which is a direct result of the evidence provided by the stimulus field. The new blanket forms on a slower time scale to that of the original, due to the fact that members of this new ensemble originally believed they were members of the alternative ensemble ($\psi_{3,4}$). Moreover, since the original external field has been removed, the corresponding neurons no longer receive evidence that they are states of the first ensemble ($\psi_{1,2}$). Instead, they receive evidence that they are members of the opposing ensemble ($\psi_{3,4}$), leading to the gradual dissolution of the first Markov blanket. Taken together,

these results demonstrate the dynamic nature of Markov blanket formation when coupled to a dynamic external field.

5. Discussion

This paper has presented a model of neural and phenotypic representation developed using the active inference framework. One potential advantage that our model offers is that it builds on the current neural process theory associated with active inference. The current process theory models single-neuron activity – which can also simply be equated with activity in populations of neurons – as more or less uniformly encoding posterior beliefs over states; moreover, it assumes that the synaptic connection patterns giving rise to the selective stimulus
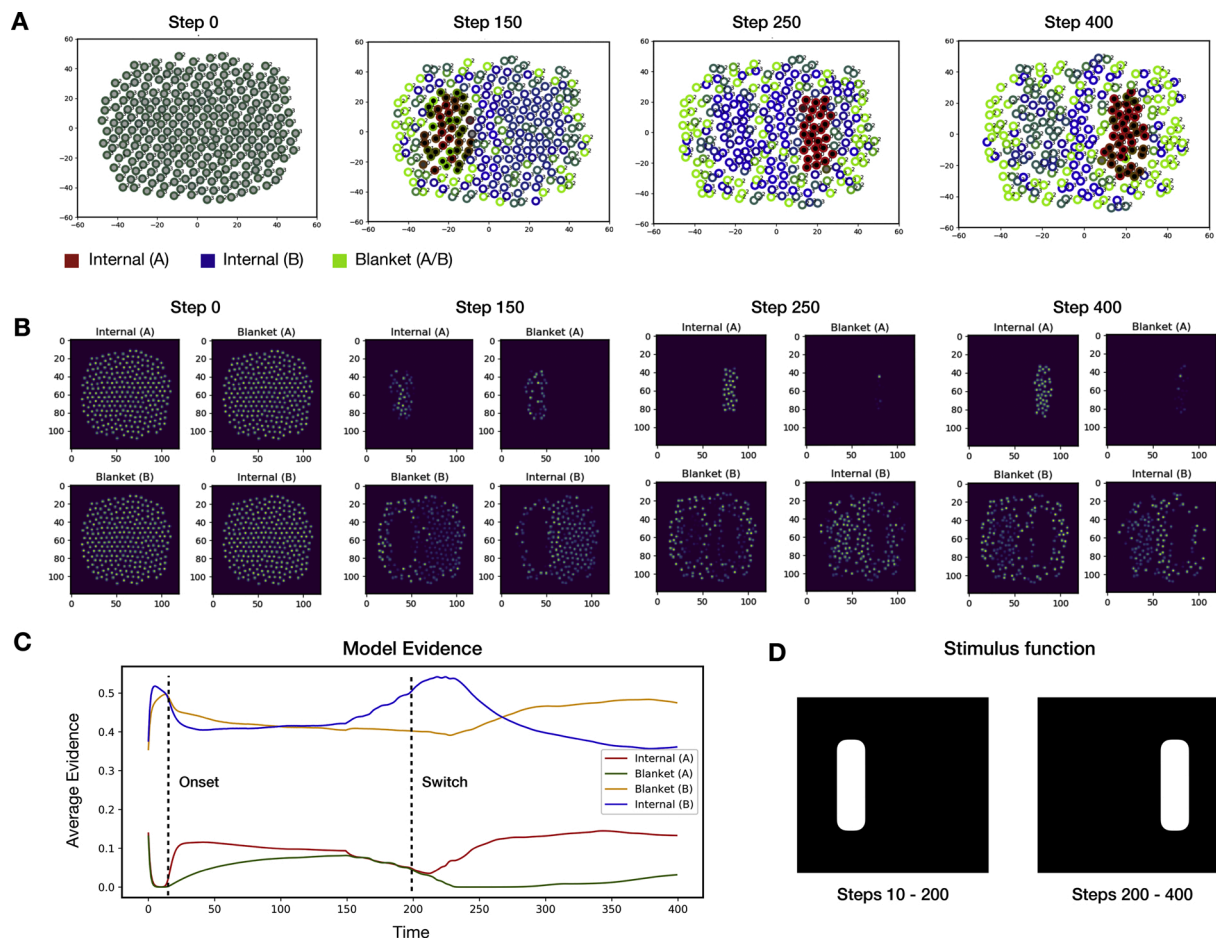
Fig. 5. Dynamic and transient packet formulation.

This figure depicts the second series of numerical results, which demonstrate the dynamic nature of Markov blanket formulation in response to transient external stimulation. **(A)** A visualisation of individual neurons at different times during the stimulation. As in the first set of simulations, the neurons are initialised with uniform beliefs and at time step $T = 10$ an external stimulus is introduced (described in the main text). Following the results of the first simulation, the external field induces the formation of two Markov blanketed systems organised around the stimulus. At time step $T = 200$, the location at which the stimulus is applied is changed. This leads a subpopulation of neurons to receive evidence that is incompatible with their Bayesian beliefs, causing the dissolution of the original Markov blanket and the emergence of a new Markov blanket around the changed site of stimulation. These results demonstrate the dynamic and transient nature of Markov blanket formation when coupled to an external field. **(B)** A visualisation of the signals emitted by neurons at different times during the stimulation. Again, in panels (A) and (B), the two axes represent the two abstract states of a two-dimensional state space. For simplicity, here, we assume this is a physical space where the agents move, with real values $\mathbb{R}^2$. **(C)** The ensemble-average evidence for each type identity as a function of time. **D)** The stimulus function used in this simulation. As above, white denotes the stimulus intensity. Here, the stimulus changes location at 200 time steps.

sensitivity of a given neuron (or population) are formally specified as single entries in a likelihood matrix connecting sensory epithelia to those neurons. Heuristically, learning in these models is Hebbian, in that synaptic connections in the likelihood matrix are strengthened based on the strength of the activity of a neuron (or population), coincident with current observations.

The key points are that – in extant models – no distinction is made between the function of a single neuron versus a population of neurons encoding posterior beliefs over states, or learning stimulus sensitivity, and that the activation levels that are modelled are single trajectories of values. Our results in this paper demonstrate that coherent activity in neural populations can arise in response to environmental perturbation and allow for population-wide tuning of synaptic connections during development – such that different populations of neurons could learn differential sensitivities to different stimulus features. This avoids possible concerns about modelling 'grandmother cells' – and may offer a means of redundant coding that allows a system to be robust to noise or injury.

On a related note, the formalism used to model the emergence of a selective responses and specialisation (e.g., receptive fields) – as a free energy minimising process – may resolve arguments about whether probabilistic representations are represented by the sample density of population responses; i.e., a population code in the sense of Beck et al. (2008); Pouget et al. (2013); Zemel et al. (1998) – or whether the population response encodes the parameters or sufficient statistics of a probabilistic representation; i.e., an ensemble code. On the present account, the two are the same thing. In other words, the quantity that matters – in terms of a representational or extrinsic information geometry – is the ensemble average of a population of neurons that constitute the internal states of a larger Markov blanket.

A subtle, technical point here is that the extrinsic information geometry that undergirds neuronal representation or encoding rests upon the conditional expectation or average of internal states, e.g., neurons that are internal to a larger Markov blanket. However, mathematically, this expectation is over time, not over internal states (Friston, 2020). To endow ensemble averages with a representational attribute, it then
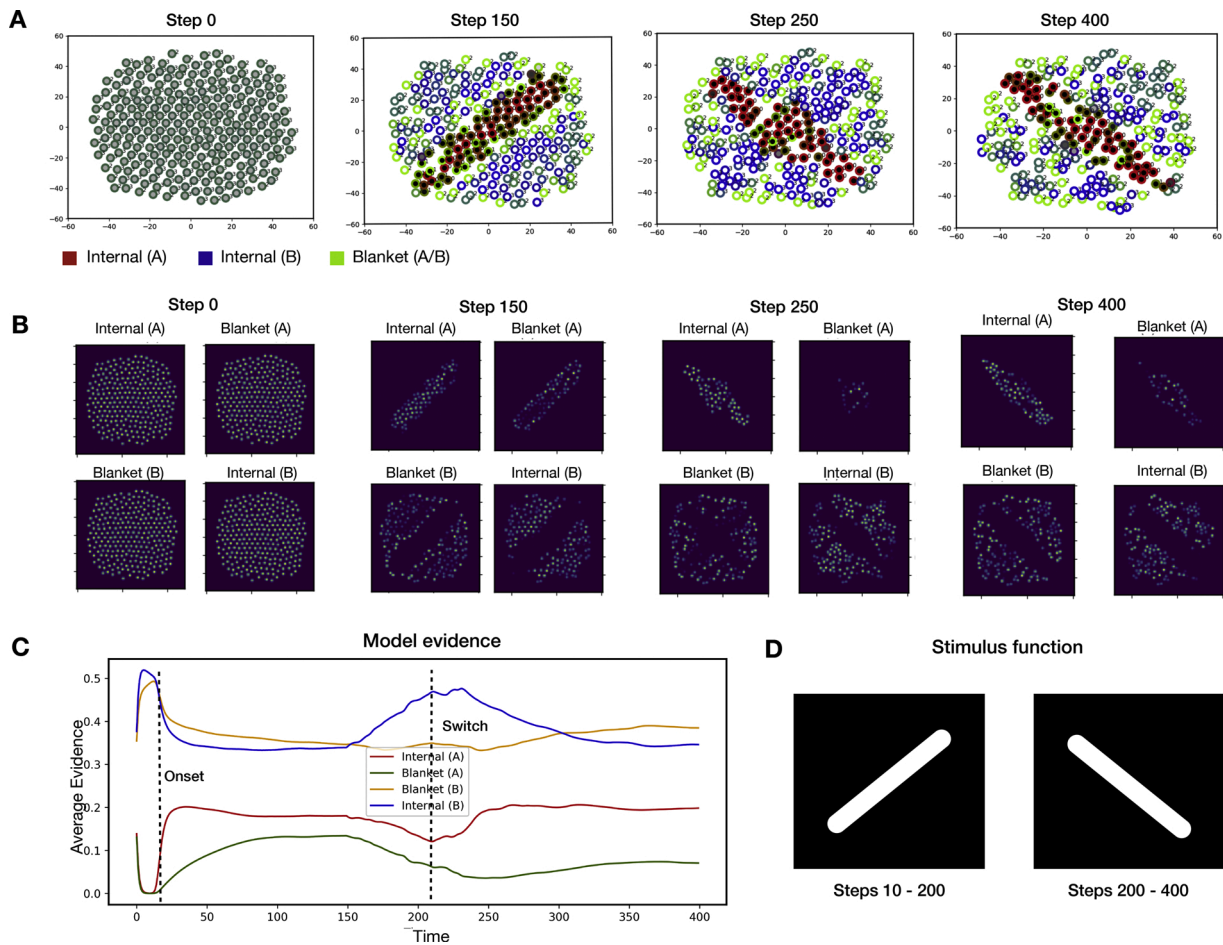
Fig. 6. Dynamic and transient packet formulation (B).

This figure depicts the final series of numerical results, which further demonstrates the dynamic nature of Markov blanket formation in response to external stimuli. These numerical results utilise a different set of set stimuli relative to Fig. 5. **(A)** A visualisation of individual neurons at different times during the stimulation. As in the previous simulations, neurons are initialised with uniform beliefs and at time step $T = 10$ an external stimulus is introduced. This again induces the formation of two Markov blanketed systems organised around the induced stimulus. At time step $T = 200$, the location at which the stimulus is applied is changed, and this again causes the location of Markov blankets to dynamically reorganise. Unlike in Fig. 5, the two stimuli overlap, meaning a subset of the population receives consistent evidence after the stimulus is changed. These beliefs of these neurons do not change, leading to faster blanket formation, relative to Fig. 5. **(B)** A visualisation of the signals emitted by neurons at different times during the stimulation. As above, in panels (A) and (B), the two axes represent the two abstract states of a two-dimensional state space. For simplicity, here, we assume this is a physical space where the agents move, with real values $\mathbb{R}^2$. **(C)** The ensemble-average evidence for each type identity as a function of time. **D)** The stimulus function used in this simulation. As above, white denotes the stimulus intensity. Here, the stimulus changes location at 200 time steps.

becomes necessary to make the ergodic assumption that the average over a large number of internal states is the same as the time average. This only works if the internal states are exchangeable or can be treated as equivalent to each other. This simple observation means that there may be something quintessential about population codes; in the sense that they can only be attributes of ensembles, e.g., populations of neurons that play the role of internal states at a superordinate scale.

This work presented here also demonstrates novel Markov blanket dynamics, opening the door to the modelling of transient boundary formation and reconfiguration – and dynamics in functional and effective connectivity – consistent with a growing body of evidence supporting neural reuse (Anderson, 2014). Some systems have permanent boundaries that endure over time. This is the case, for instance, when the statistical boundaries of a system coincide with its physical ones, as in the case of a cell, where the statistical boundary is also hardwired into the physical structure of the system, as a literal barrier (a cell membrane). Other living systems, however, have more malleable boundaries, which dissolve and reform over time. These are systems like social systems (Veissière et al., 2020) and – arguably – neuronal packets (Yufik and Friston, 2016). When we wake up in our home, we are part of the intimate living unit; when we go to the office, we are part of another functional unit; and belonging to these groups constrains possible courses of action. This is also the case for neural ensembles, which seem to form and reform on the fly to match task requirements (Anderson, 2014). For systems such as these, the relevant boundaries are functional precisely because they can change, as when one individual member moves to another group, when a group splits into two, or when two groups merge. The boundaries of such neural (or social) systems are intrinsically malleable, which allows for adaptive reconfigurations of groups and group membership. We propose that one can associate the changing architecture and membership of superordinate ensembles with inference dynamics that track different perceptual contents.

Note that fluctuating and itinerant Markov blankets are consistent with non-equilibrium steady-state, due to a separation of temporal scales that emerges with Markov blankets of Markov blankets (Friston, 2020). In other words, any Markov blanket will change slowly from the point of view of Markov blankets at the scale below. For example, the Markov blanket of a city does not change hour by hour, as you commute from work to home. Similarly, the Markov blanket of your brain does not change second by second, as neuronal ensembles (i.e., packets or assemblies) split and merge. The key thing – from the perspective of ensemble dynamics – is that various (attracting) states are revisited after a sufficient period of time. For example, you return home every evening. And your functionally specialised neuronal ensembles are activated every few seconds by characteristic patterns of sensory afferents.

The framework that we develop here has implications that transcend neuronal dynamics. Indeed, the scheme offers a formal, generalizable account of phenomena that involve the formation of transient boundaries in nested systems. The ensuing formalism provides a generic and general mechanism by which coordinated movement in living creatures might act as an effective means to model or represent the causal situations, environments, and predicaments in which living creatures find themselves.

Group foraging behaviours, for instance, obey similar principles, where the coordinated patterns of movement in a direction count as accumulated evidence for specific hypotheses about the structure of the environment. This is the case, for instance, when ants leave and follow pheromone trails. In this setup, the more ants that follow a given path, the more evidence each provides to its conspecifics for the belief that there is something salient (a food item or danger) at the end of the trail (Jackson et al., 2004).

We suggest that it may be possible to extend the modelling strategy developed above to account for the ways in which social and cultural group membership functions as a form of inference about relevant causal structure in the human environment. For instance, political affiliation seems to track one's beliefs and attitudes regarding some domains, such as the reality of human-made climate change. One might speculate that for many human agents, the individual inference problem (namely, to answer for oneself the question, 'Do I believe that human-made climate change is real?') is solved by outsourcing the process of generating a working guess (computing the posterior) to one's close social circle, often by figuring out and learning those things believed and valued in one's social ingroup (Kirmayer and Ramstead, 2017; Ramstead et al., 2016; Veissière et al., 2020).

Another, less politically contentious example is the optimal foraging behaviour in ant colonies discussed above. Indeed, ant foraging behaviour might be modelled as a kind of group inference, where each foraging 'arm' of the ant population (enshrouded behind its own Markov blanket, here determined by chemical pheromone gradients) can be cast as representing a distinct hypothesis that 'there is food here', or that 'there is a dangerous threat here', and so on. One advantage of applying this formalism to such animal models – instead of human systems – is that they are comparatively easier to study, since they are simpler to extensively (and ethically) measure and since their (relatively less complex) behaviour is easier to model. Work on ant foraging and on other simple foraging systems, such as the multiple search fronts of foraging slime mould populations, routinely employs modelling strategies amenable to active inference (Chandrasekhar et al., 2018; Reid et al., 2012). Behavioural data regarding the decisions made by those simple systems is equally reproducible in silico.

Finally, one might be able to extend the above population coding strategy to *hierarchically nested* representational capacities. In such a scheme, neuronal packets at the superordinate level would engage in exactly the same kind of coordinated inferential dynamics, inferring their membership identity in yet a higher level of self-organization. This

could lead from a theory of neural representation to a theory of conceptual- or semantic-level processing (Yufik, 2019; Yufik and Friston, 2016) – as opposed to the perceptual processing formally modelled here – where neuronal packet formation at several nested scales in the brain explains the perception and multimodal integration of the perception of objects. Indeed, current formulations of understanding and abstraction rest upon deep generative models with relational structure that necessarily call upon Markov blankets to separate hierarchical levels; e.g., Friston et al. (2017a), 2017b. The time and context sensitive nature of Markov blankets – considered in this work – may reflect the context sensitive, state dependent, nonlinear generative models that are necessary for self-evidencing.

6. Conclusion

The aim of this paper was to develop a generic and generalizable model of the representational capacities of living creatures by using the resources of the free-energy principle and its corollary process theory, active inference. We developed an account of neuronal and, more broadly, phenotypic representation. We argued that the representational capacities of living creatures are a consequence of their Markovian structure and action in nonequilibrium regimes to counter the dissipating effects of random fluctuations and environmental itinerancy. We demonstrated *in silico* that information about external stimuli can be encoded by groups of neurons bound by a Markov blanket, thereby providing evidence for the *neuronal packet hypothesis* that inherits from theories of neuronal groups and assemblies (Buzsaki, 2010; Edelman, 1993, 1998; Hebb, 1949; von der Malsburg, 1981). Our numerical proof-of-principle simulations showed that self-organizing ensembles of particles that share the right kind of probabilistic generative model are able to encode recoverable information about a changing sensorium – and that they can dynamically evince patterns of functional self-organization that adapt to the demands of the situation. This model of neuronal packet formation thus entails a modest form of representational capacity and opens up important directions for future work.

7. Software note

The Python code for this paper is available at https://github.com/alec-tschantz/packets. To recreate the figures four, five and six, please run packets_figure_01.py, packets_figure_02.py and packets_figure_03.py, respectively. Further information on recreating the figures and data are available in the linked repository. This code is based on a generalised filtering scheme written in the SPM software for the MATLAB environment; namely, the DEM_cells.m and DEM_cells_cells.m scripts, which implement the self-organisation of a single ensemble Markov blanketed ensemble and an ensemble of such ensembles, respectively. The original routines are available at http://www.fil.ion.ucl.ac.uk/spm.

Acknowledgements

References

Anderson, M.L., 2014. After Phrenology: Neural Reuse and the Interactive Brain. The MIT Press, Cambridge, MA; London, England.

Bair, W., 1999. Spike timing in the mammalian visual system. Curr. Opin. Neurobiol. 9 (4), 447–453.

Beck, J.M., Ma, W.J., Kiani, R., Hanks, T., Churchland, A.K., Roitman, J., Pouget, A., 2008. Probabilistic population codes for Bayesian decision making. Neuron 60 (6), 1142–1152. https://doi.org/10.1016/j.neuron.2008.09.021.

Bienenstock, E.L., Cooper, L.N., Munro, P.W., 1982. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. J. Neurosci. 2 (1), 32–48.

Blakemore, C., Cooper, G.F., 1970. Development of the brain depends on the visual environment. Nature 228 (5270), 477–478. https://doi.org/10.1038/228477a0.

Borst, A., Theunissen, F.E., 1999. Information theory and neural coding. Nat. Neurosci. 2 (11), 947–957.

Buonomano, D.V., Merzenich, M.M., 1998. Cortical plasticity: from synapses to maps. Annu. Rev. Neurosci. 21, 149–186. https://doi.org/10.1146/annurev.neuro.21.1.149.

Buzsaki, G., 2010. Neural syntax: cell assemblies, synapsembles, and readers. Neuron 68 (3), 362–385. https://doi.org/10.1016/j.neuron.2010.09.023.

Calvo, P., Friston, K., 2017. Predicting green: really radical (plant) predictive processing. J. R. Soc. Interface 14 (131). https://doi.org/10.1098/rsif.2017.0096.

Chandrasekhar, A., Gordon, D.M., Navlakha, S., 2018. A distributed algorithm to maintain and repair the trail networks of arboreal ants. Sci. Rep. 8 (1), 1–19.

Constant, A., Ramstead, M.J., Veissière, S.P., Campbell, J., Friston, K.J., 2018. A variational approach to niche construction. J. R. Soc. Interface.

Edelman, G.M., 1993. Neural Darwinism: selection and reentrant signaling in higher brain function. Neuron 10 (2), 115–125.

Edelman, S., 1998. Representation is representation of similarities. Behav. Brain Sci. 21 (4), 449–467.

Egan, F., 2018. The nature and function of content in computational models. In: Sprevak, M., Colombo, M. (Eds.), The Routledge Handbook of the Computational Mind. Routledge, pp. 247–258.

Friston, K.J., 2010. The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. 11 (2), 127–138.

Friston, K., 2012. A free energy principle for biological systems. Entropy 14 (11), 2100–2121.

Friston, K.J., 2013. Life as we know it. J. R. Soc. Interface 10 (86). https://doi.org/10.1098/rsif.2013.0475.

Friston, K.J., 2020. A Free Energy Principle for a Particular Physics.

Friston, K., Buzsáki, G., 2016. The functional anatomy of time: what and when in the brain. Trends Cogn. Sci. 20 (7), 500–511.

Friston, K.J., Levin, M., Sengupta, B., Pezzulo, G., 2015. Knowing one's place: a free-energy approach to pattern regulation. J. R. Soc. Interface 12 (105).

Friston, K.J., Lin, M., Frith, C.D., Pezzulo, G., Hobson, J.A., Ondobaka, S., 2017a. Active inference, curiosity and insight. Neural Comput. 29 (10), 2633–2683. https://doi.org/10.1162/neco_a_00999.

Friston, K.J., Parr, T., de Vries, B., 2017b. The graphical brain: belief propagation and active inference. Netw. Neurosci. 1 (4), 381–414.

Friston, K.J., Fagerholm, E.D., Zarghami, T.S., Parr, T., Hipólito, I., Magrou, L., Razi, A., 2020a. Parcels and particles: markov blankets in the brain. arXiv preprint arXiv 2007, 09704.

Friston, K.J., Wiese, W., Hobson, J.A., 2020b. Sentience and the origins of consciousness: from Cartesian duality to Markovian monism. Entropy 22 (5), 516.

Gerstein, G.L., Kirkland, K.L., 2001. Neural assemblies: technical issues, analysis, and modeling. Neural Netw. 14 (6–7), 589–598. https://doi.org/10.1016/s0893-6080(01)00042-9.

Hebb, D.O., 1949. The Organization of Behavior. Wiley, New York.

Hipólito, I., 2019. A simple theory of every 'thing'. Phys. Life Rev. 31, 79–85.

Hipólito, I., Ramstead, M.J.D., Convertino, L., Bhat, A., Parr, T., Friston, K.J., 2020. Markov Blankets in the Brain.

Hohwy, J., 2016. The self-evidencing brain. Noûs 50 (2), 259–285.

Huang, E.J., Reichardt, L.F., 2001. Neurotrophins: roles in neuronal development and function. Annu. Rev. Neurosci. 24, 677–736. https://doi.org/10.1146/annurev.neuro.24.1.677.

Isomura, T., Friston, K., 2018. In vitro neural networks minimise variational free energy. bioRxiv. https://doi.org/10.1101/323550.

Jackson, D.E., Holcombe, M., Ratnieks, F.L., 2004. Trail geometry gives polarity to ant foraging networks. Nature 432 (7019), 907–909.

Kiebel, S.J., Friston, K.J., 2011. Free energy and dendritic self-organization. Front. Syst. Neurosci. 5.

Kiefer, A.B., 2020. Psychophysical identity and free energy. arXiv.

Kiefer, A., Hohwy, J., 2018. Content and misrepresentation in hierarchical generative models. Synthese 1–29.

Kiefer, A., Hohwy, J., 2019. Representation in the prediction error minimization framework. In: Symons, J., Calvo, P., Robins, S. (Eds.), Routledge Handbook to the Philosophy of Psychology.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., Kiverstein, J., 2018. The Markov blankets of life: autonomy, active inference and the free energy principle. J. R. Soc. Interface 15 (138), 20170792.

Kirmayer, L.J., Ramstead, M.J., 2017. Embodiment and enactment in cultural psychiatry. In: Durt, C., Fuchs, T., Tewes, C. (Eds.), Embodiment, Enaction, and Culture: Investigating the Constitution of the Shared World. MIT Press, pp. 397–422.

Kuchling, F., Friston, K., Georgiev, G., Levin, M., 2019. Morphogenesis as Bayesian inference: a variational approach to pattern formation and control in complex biological systems. Phys. Life Rev.

Maunsell, J.H., Van Essen, D.C., 1983. Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. J. Neurophysiol. 49 (5), 1127–1147.

Mountcastle, V.B., 1997. The columnar organization of the neocortex. Brain 120 (Pt 4), 701–722. https://doi.org/10.1093/brain/120.4.701.

Muir, D.W., Mitchell, D.E., 1973. Visual resolution and experience: acuity deficits in cats following early selective visual deprivation. Science 180 (4084), 420–422.

Nicolelis, M.A.L., Lebedev, M.A., 2009. Principles of neural ensemble physiology underlying the operation of brain–machine interfaces. Nat. Rev. Neurosci. 10 (7), 530–540. https://doi.org/10.1038/nrn2653.

Palacios, E.R., Isomura, T., Parr, T., Friston, K., 2019. The emergence of synchrony in networks of mutually inferring neurons. Sci. Rep. 9 (1), 1–14.

Palacios, E.R., Razi, A., Parr, T., Kirchhoff, M., Friston, K., 2020. On Markov blankets and hierarchical self-organisation. J. Theor. Biol. 486, 110089.

Park, H.-J., Friston, K., 2013. Structural and functional brain networks: from connections to cognition. Science 342 (6158).

Parr, T., Friston, K.J., 2018. The anatomy of inference: generative models and brain structure. Front. Comput. Neurosci. 12.

Parr, T., Da Costa, L., Friston, K., 2020a. Markov blankets, information geometry and stochastic thermodynamics. Philos. Trans. Math. Phys. Eng. Sci. 378 (2164), 20190159.

Parr, T., Sajid, N., Friston, K.J., 2020b. Modules or mean fields? Preprint.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, CA.

Peters, A., Yilmaz, E., 1993. Neuronal organization in area 17 of cat visual cortex. Cereb. Cortex 3 (1), 49–68. https://doi.org/10.1093/cercor/3.1.49.

Pouget, A., Dayan, P., Zemel, R., 2000. Information processing with population codes. Nat. Rev. Neurosci. 1 (2), 125–132.

Pouget, A., Beck, J.M., Ma, W.J., Latham, P.E., 2013. Probabilistic brains: knowns and unknowns. Nat. Neurosci. 16 (9), 1170–1178. https://doi.org/10.1038/nn.3495.

Ramstead, M.J., Veissière, S.P., Kirmayer, L.J., 2016. Cultural affordances: scaffolding local worlds through shared intentionality and regimes of attention. Front. Psychol. 7.

Ramstead, M.J., Badcock, P.B., Friston, K.J., 2018. Answering Schrödinger's question: a free-energy formulation. Phys. Life Rev. 24, 1–16.

Ramstead, M.J., Constant, A., Badcock, P.B., Friston, K.J., 2019a. Variational ecology and the physics of sentient systems. Phys. Life Rev.

Ramstead, M.J., Kirchhoff, M.D., Friston, K.J., 2019b. A tale of two densities: active inference is enactive inference. Adapt. Behav. 1059712319862774.

Ramstead, M.J., Friston, K.J., Hipólito, I., 2020. Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. Entropy.

Reid, C.R., Latty, T., Dussutour, A., Beekman, M., 2012. Slime mold uses an externalized spatial "memory" to navigate in complex environments. Proc. Natl. Acad. Sci. 109 (43), 17490–17494.

Rice, D., Barone, S., 2000. Critical periods of vulnerability for the developing nervous system: evidence from humans and animal models. Environ. Health Perspect. 108, 511–533. https://doi.org/10.2307/3454543.

Sherrington, C.S., 1911. The Integrative Action of the Nervous System. Retrieved from. http://content.apa.org/books/2009-00519-000.

Sin, W.C., Haas, K., Ruthazer, E.S., Cline, H.T., 2002. Dendrite growth increased by visual activity requires NMDA receptor and Rho GTPases. Nature 419 (6906), 475–480. https://doi.org/10.1038/nature00987.

Singer, W., Freeman, B., Rauschecker, J., 1981. Restriction of visual experience to a single orientation affects the organization of orientation columns in cat visual cortex. Exp. Brain Res. 41 (3–4), 199–215.

Sur, M., Rubenstein, J.L., 2005. Patterning and plasticity of the cerebral cortex. Science 310 (5749), 805–810.

Sur, M., Garraghty, P.E., Roe, A.W., 1988. Experimentally induced visual projections into auditory thalamus and cortex. Science 242 (4884), 1437–1441.

Thoenen, H., 1995. Neurotrophins and neuronal plasticity. Science 270 (5236), 593–598. https://doi.org/10.1126/science.270.5236.593.

Tolhurst, D.J., Movshon, J.A., Dean, A.F., 1983. The statistical reliability of signals in single neurons in cat and monkey visual cortex. Vision Res. 23 (8), 775–785.

Usrey, W.M., Reid, R.C., 1999. Synchronous activity in the visual system. Annu. Rev. Physiol. 61 (1), 435–456.

Veissière, S.P., Constant, A., Ramstead, M.J., Friston, K.J., Kirmayer, L.J., 2020. Thinking through other minds: a variational approach to cognition and culture. Behav. Brain Sci.

von der Malsburg, C., 1981. The Correlation Theory of Brain Function. Internal Report 81-82. Retrieved from Gottingen.

von der Malsburg, C., Phillips, W.A., Singer, W., 2010. Dynamic Coordination in the Brain : From Neurons to Mind.

Von Melchner, L., Pallas, S.L., Sur, M., 2000. Visual behaviour mediated by retinal projections directed to the auditory pathway. Nature 404 (6780), 871–876.

Yufik, Y.M., 1998. Virtual associative networks: a framework for cognitive modeling. In: Pribram, K.H. (Ed.), Brain and Values. Lawrence Erlbaum Associates, New York, pp. 109–177.

Yufik, Y.M., 2019. The understanding capacity and information dynamics in the human brain. Entropy 21 (3), 308.

Yufik, Y.M., Friston, K.J., 2016. Life and understanding: the origins of "understanding" in self-organizing nervous systems. Front. Syst. Neurosci. 10.

Zeki, S., 2005. The Ferrier Lecture 1995 behind the seen: the functional specialization of the brain in space and time. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 360 (1458), 1145–1183.

Zeki, S., Shipp, S., 1989. Modular connections between areas V2 and V4 of macaque monkey visual cortex. Eur. J. Neurosci. 1 (5), 494–506.

Zemel, R., Dayan, P., Pouget, A., 1998. Probabilistic interpretation of population code. Neural Computat. 10, 403–430.