High-level information integration in the brain via large-scale attractor dynamics

Tamas Spisak (tamas.spisak@uk-essen.de)
Center for Translational Neuro- and Behavioral Sciences (C-TNBS), University Medicine Essen, Germany

Abstract

Understanding how high-level information integration arises from large-scale brain activity requires bridging computational principles with neural dynamics. We propose a theoretical framework where large-scale brain dynamics emerge as trajectories around attractors in course-grained recurrent networks whose dynamics precisely map to computations. The core network model in our framework combines principles of self-organization with attractor network theory and Bayesian inference, offering a recursive, multi-level description, applicable to large-scale empirical data. A key feature of these networks is the emergent orthogonality of the attractors, which maximizes storage capacity and computational efficiency. Crucially, this orthogonality allows mapping complex attractor dynamics onto simpler, interpretable bipartite architectures, revealing how a wide variety of computations can be implemented implicitly by network-wide stochastic attractor dynamics. We propose this framework as a model for large-scale brain dynamics. Our approach aligns with previous literature and is supported by emerging evidence, such as observations of orthogonal brain attractors, akin to canonical resting state networks. The framework yields testable predictions and offers a principled yet simple approach to understanding, explaining, and predicting large-scale brain dynamics and corresponding behavior.

Keywords: Large-Scale Brain Dynamics; Attractor Networks; Computational Model; Self-Organization; Free Energy Principle

Introduction

Large-scale brain dynamics are organized around stable, self-sustaining patterns, known as *attractor states* (Deco & Rolls, 2003; Khona & Fiete, 2022; Englert et al., 2024). The brain not only operates as an attractor network; it is also capable of becoming and remaining one through a self-directed process of development and learning. While the mathematical apparatus of attractor networks is well established (Amit, 1989), understanding the concrete computations carried out by these collective self-organized dynamics still poses a key challenge.

Here we propose a formal framework that links the principles of adaptive self-organizing dynamics to the concrete computations implemented by attractors in the brain at arbitrary scales.

Our framework is rooted in the free energy principle (FEP) (Friston, 2010). The FEP is ideally suited to establish this link, as it formally equates the process of maintaining statistical separation from the environment (a form of self-organization) with performing Bayesian inference (a form of computation) through variational free energy minimization.

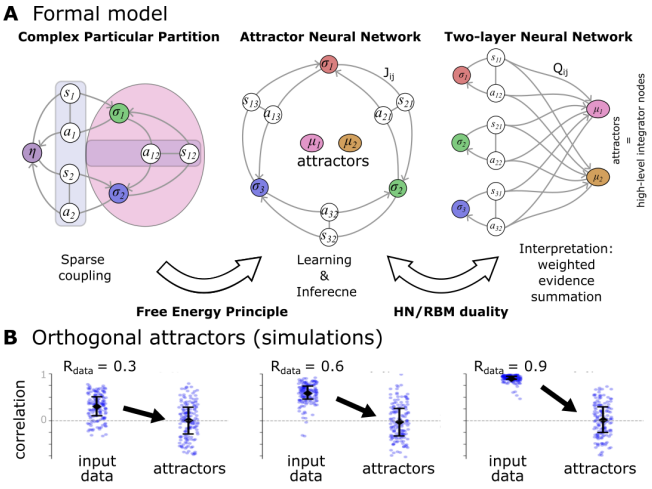

Figure 1: Derivation of the formal model. **A** A deep particular partition, with interacting subparticles σ_i (left), which gives rise to an attractor network (middle). As this network establishes approximately orthogonal attractors (panel **B**), the network is equivalent to an interpretable bipartite architecture (**A** right).

Here we give a high-level overview of the resulting formal model and describe some of the main properties, including its potential to link attractor dynamics to computational primitives through a mathematical duality between two different network topologies. We conclude by discussing the framework in light of existing literature and initial results.

Self-Organizing Attractor Networks from FEP

The derivation of our framework (Spisak & Friston, 2025) starts from a general formulation of random dynamical systems and a universal partitioning of them, known as a *particular partition*. This separates internal (μ) from external (η) states via the "Markov blanket" consisting of sensory (s) and active (a) states, so that: $\eta \perp \mu \mid s, a$. Maintaining this partition for extended time periods drives internal states to infer external causes via free energy minimization: $\dot{\mu} \propto -\nabla_\mu F$.

It can be shown that, if the internal state μ comprises interacting *subparticles* (where one subparticle's internal state σ_i can be another's external state, Fig. 1A left), the system's dynamics can give rise to arbitrarily complex attractor network structures. Under plausible parametrizations (e.g., continuous Bernoulli states for subparticles with bias b_i representing evidence), the joint distribution of this system (referred to as a *deep particular partition*) takes the functional form of a continuous-state stochastic Hopfield network (a type of Boltzmann Machine): $P(\sigma) \propto \exp(\sum_i b_i \sigma_i + \sum_{ij} J_{ij} \sigma_i \sigma_j)$ (Fig. 1A middle), where J_{ij} is the coupling weights implemented by the boundary states of the particle.

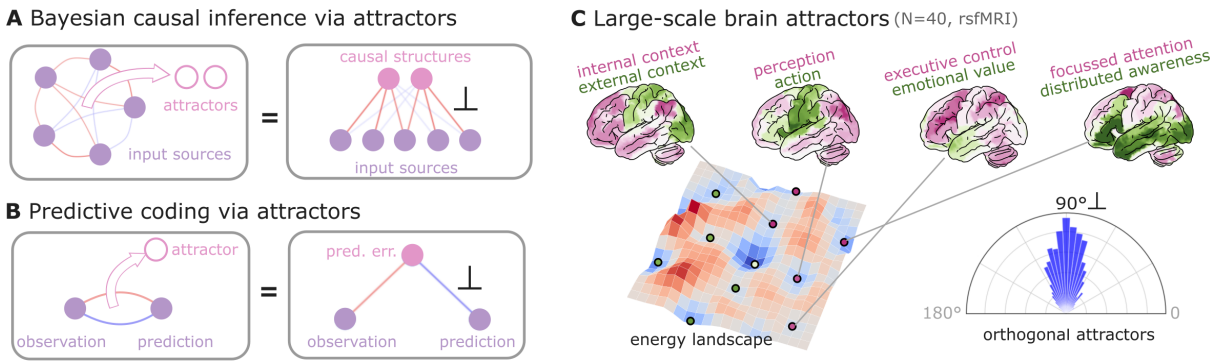

Figure 2: In the proposed framework, attractor states implement computational primitives, like weighted evidence integration (Bayesian causal inference, **A**) or predictive coding (**B**). Brain attractor states reconstructed from resting state fMRI data (**C**) resemble canonical resting state networks and exhibit approximate orthogonality. Adapted from (Englert et al., 2024).

Local Inference Dynamics: Minimizing variational free energy (VFE) from the point of view of a single node of the attractor network, σ_i, given observations from all other nodes, $\sigma_{\setminus i}$ yields a stochastic variant of the standard Hopfield update rule $\mathbb{E}_q[\sigma_i] = L(b_i + \sum_{j \neq i} J_{ij}\sigma_j)$, where L is the Langevin function and J_{ij} is the coupling strength, equivalent to local approximate Bayesian inference (Spisak & Friston, 2025).

Macro-scale Bayesian Inference: Being a special case of Boltzmann machines, the network implements Markov Chain Monte Carlo sampling (Neal, 1992) from the global posterior distribution $P(\sigma|\mathbf{b},\mathbf{J})$, given the prior represented by attractors and external input (likelihood) $\mathbf{b}$. This reveals a recursive structure: local Bayesian inference in the subparticles gives rise to collective macro-scale Bayesian inference at the network level (which itself is also a VFE-minimizing particle). This allows **multiple valid levels of description** - an important requirement for modeling large-scale brain dynamics with coarse-graining and renormalization groups (Binder, 1981).

Learning & Orthogonalization: Minimizing VFE with respect to couplings J_{ij} yields a learning rule that contrasts observed vs. predicted correlations: $\Delta J_{ij} \propto \sigma_i\sigma_j - L(b_i + \sum_{k \neq i} J_{ik}\sigma_k)\sigma_j$. It can be shown both mathematically and with simulations (Fig. 1B, (Spisak & Friston, 2025)) that this learning rule – akin to Sanger's rule (Sanger, 1989) – intrinsically drives attractor states towards *orthogonality*, which optimizes memory capacity and computational robustness (Personnaz, Guyon, & Dreyfus, 1985; Kanter & Sompolinsky, 1987).

Duality: attractors = information integration: Emergent attractor orthogonality enables a mathematical duality between our attractor network and two-layer bipartite architectures, akin to Restricted Boltzmann Machines, with an additional layer of 'integrator' and inter-layer weights Q (Fig. 1A right), so that $J \approx QQ^T$ (Barra, Bernacchia, Santucci, & Contucci, 2012). This maps attractors directly onto the integrator nodes, translating their dynamics to a language of interpretable computational primitives (Fig. 2A-B), like predictive coding (Rao & Ballard, 1999) and weighted evidence integration (Bayesian Causal Inference (Körding et al., 2007)).

A Model for Large-Scale Brain Dynamics

We propose this framework as a model for large-scale brain dynamics as measured by fMRI, with the nodes σ_i representing activity in brain regions and coupling weights $\mathbf{J}$ reconstructed as the negative inverse covariance matrix of the activation timeseries. Theory suggests that fMRI may provide sufficient time resolution to capture the slow processes associated with macro-scale computations (Carr, 2012). Indeed, there is robust empirical evidence in the literature that large-scale brain dynamics ("activity flow") align with the derived inference rule (Cole, 2024). Further, there is initial evidence (Englert et al., 2024) indicating that: (i) functional connectome-based attractor networks exhibit efficient and rich attractor dynamics; (ii) brain attractors robustly map onto canonical resting state networks (Fig. 2C, N=40 rsfMRI); (iii) attractors extracted from fMRI data exhibit near-orthogonality (Fig. 2C) and; (iv) attractor dynamics can capture perceptual states, like pain modulation (Englert et al., 2024).

Discussion and Predictions

Our framework is based on self-organizing, and self-orthogonalizing, attractor networks that emerge from the FEP, and links them via a mathematical duality to interpretable computations. Our approach is grounded in plausible assumptions and borrows its construct validity from first principles of self-organization. Supported by initial fMRI evidence, the framework presents a principled yet simple theoretical model for large-scale brain dynamics. Testable predictions include that: (i) large-scale synaptic plasticity should reflect the derived orthogonalizing learning rule; (ii) network perturbations (task, TMS, pathology) should propagate consistent with model dynamics; and (iii) given that attractor dynamics represent an emergent level of computation (Rosas et al., 2024), cognitive function relying on high-level information integration should be predictable from *attractor timeseries* as much as from full neural data. Validating these and further predictions will open up novel opportunities in understanding, explaining and predicting large-scale brain dynamics and corresponding behavior.

Acknowledgment

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — Project-ID 422744262 - TRR 289; and Project-ID 316803389 – SFB 1280 “Extinction Learning”.

References

Amit, D. J. (1989). *Modeling brain function: The world of attractor neural networks*. Cambridge university press.

Barra, A., Bernacchia, A., Santucci, E., & Contucci, P. (2012). On the equivalence of hopfield networks and boltzmann machines. *Neural Networks*, *34*, 1–9.

Binder, K. (1981). Critical properties from monte carlo coarse graining and renormalization. *Physical Review Letters*, *47*(9), 693.

Carr, J. (2012). *Applications of centre manifold theory* (Vol. 35). Springer Science & Business Media.

Cole, M. W. (2024). The explanatory power of activity flow models of brain function. *arXiv preprint arXiv:2402.02191*.

Deco, G., & Rolls, E. T. (2003). Attention and working memory: a dynamical model of neuronal activity in the prefrontal cortex. *European Journal of Neuroscience*, *18*(8), 2374–2390.

Englert, R., Kincses, B., Kotikalapudi, R., Gallitto, G., Li, J., Hoffschlag, K., ... Spisak, T. (2024). Connectome-based attractor dynamics underlie brain activity in rest, task, and disease. *eLife*, *13*(RP98725). doi: 10.7554/eLife.98725.1

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138.

Kanter, I., & Sompolinsky, H. (1987). Associative recall of memory without errors. *Physical Review A*, *35*(1), 380.

Khona, M., & Fiete, I. R. (2022). Attractor and integrator networks in the brain. *Nature Reviews Neuroscience*, *23*(12), 744–766.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS one*, *2*(9), e943.

Neal, R. M. (1992). Connectionist learning of belief networks. *Artificial intelligence*, *56*(1), 71–113.

Personnaz, L., Guyon, I., & Dreyfus, G. (1985). Information storage and retrieval in spin-glass like neural networks. *Journal de Physique Lettres*, *46*(8), 359–365.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79–87.

Rosas, F. E., Geiger, B. C., Luppi, A. I., Seth, A. K., Polani, D., Gastpar, M., & Mediano, P. A. (2024). Software in the natural world: A computational approach to hierarchical emergence. *arXiv preprint arXiv:2402.09090*.

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks*, *2*(6), 459–473.

Spisak, T., & Friston, K. (2025). Self-orthogonalizing attractor neural networks emerging from the free energy principle. *arXiv preprint arXiv:2505.22749*.