A Caveat Regarding the Unfolding Argument: Implications of Plasticity for Computational Theories of Consciousness

Authors
Vikas N. O'Reilly-Shah, MD, PhD 1#
Alessandro Selvitella Ph.D. 2,3
Aaron Schurger Ph.D. 4,5

1 Department of Anesthesiology & Pain Medicine, University of Washington School of Medicine, Seattle, WA, USA
2 Department of Mathematical Sciences & Laboratory of Data Science, Purdue University Fort Wayne, Fort Wayne, IN, USA
3 eScience Institute, University of Washington, Seattle, WA, USA
4 Department of Psychology, Crean School of Health and Behavioral Sciences, Chapman University, Orange, CA, USA
5 INSERM U992, Cognitive Neuroimaging Unit, NeuroSpin Center, Gif sur Yvette 91191, France

XXX

Disclosure of funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of interest: All authors declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Author contributions: All authors meet ICMJE criteria for authorship. Conceptualization: V-OS, A-Sch; Formal Analysis: A-Sel; Writing - Original Draft: : All authors; Writing - Review & Editing: All authors.

Acknowledgements: The authors would like to thank their respective universities for their generous support of the authors' time.

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work the author(s) used OpenAI ChatGPT, Anthropic Claude, and Google Gemini models in order to polish written content, assist with the formatting of proofs in LaTeX, and assist with the coding of TikZ figures in LaTeX. After using this tool/service, the authors reviewed and edited the content. The authors take full responsibility for the manuscript in its entirety. Citations were manually retrieved from traditional sources (e.g., PubMed).

Data Availability Statement: There is no new data associated with the article.

Abstract

The unfolding argument in the neuroscience of consciousness posits that causal structure cannot account for consciousness because any recurrent neural network (RNN) can be “unfolded” into a functionally equivalent feedforward neural network (FNN) with identical input-output behavior. Subsequent debate has focused on dynamical properties and philosophy of science critiques. We examine a novel caveat to the unfolding argument for RNN systems with rapid plasticity in their connection weights. We demonstrate through rigorous mathematical proofs that rapid plasticity negates the functional equivalence between RNN and FNN. Our proofs address history-dependent plasticity, dynamical systems analysis, information-theoretic considerations, perturbational stability, complexity growth, and resource limitations. We demonstrate that neuronal systems that possess properties such as plasticity, history-dependence, and complex temporal information encoding have features that cannot be captured by a static FNN. Our findings delineate limitations of the unfolding argument that apply if consciousness arises from temporally extended dynamic neural processes rather than static input-output mappings. This work has implications for theories of consciousness, and for the fields of computational neuroscience and philosophy of mind more generally, providing new constraints for theories of consciousness.

Keywords

Consciousness, Unfolding Argument, Recurrent Neural Networks, Feedforward Neural Networks, Takens’ Theorem, Dynamical Systems, Neural Plasticity, Information Theory

Introduction

Despite significant advances in both neuroscience and computation, a scientific explanation of consciousness remains elusive [1]. Modern physicalist theories of consciousness, in particular integrated information theory (IIT), have focused on the causal structure of neural network systems [2]. Recurrent causal structures are necessary for a system to have cause-effect power upon itself, an axiomatic principle in IIT for the development of consciousness from such a system. Among the many challenges facing physicalist theories of consciousness, the so-called *unfolding argument* (UA) poses a significant obstacle to any theory that asserts recurrence as a fundamental requirement.

The Unfolding Argument

Doerig and colleagues developed the unfolding argument as a challenge to causal structure theories of consciousness - theories that identify consciousness with specific patterns of causal interaction rather than input-output behavior [3,4]. Their argument targets theories like Integrated Information Theory (IIT) and Recurrent Processing Theory (RPT), which propose that consciousness depends on how neural elements interact with each other, not merely on what function they compute.

The unfolding argument rests on four key premises, paraphrased from the original paper [4]:

- **P1**: Scientific investigation must rely on physical measurements, and in the investigation of consciousness these are all ultimately forms of subjective report
- **P2**: Any recurrent neural network (RNN) can be "unfolded" into a functionally equivalent feedforward neural network (FNN)

- **P3**: Systems with identical input-output functions cannot be distinguished by experiments that rely on physical measurements, except for measurements of the internal workings of the system - i.e. brain activity
- **P4**: We cannot use measures of brain activity as *a priori* indicators of consciousness since understanding the brain basis of consciousness is what we seek to explain in the first place.

From these premises, Doerig et al. conclude that causal structure theories face a fatal dilemma: they are either false (if they accept that unfolded feedforward networks can be conscious) or unscientific (if they maintain that functionally identical systems with different causal structures have different consciousness despite being empirically indistinguishable).

The unfolding argument has faced significant criticism since it was introduced. Critiques have challenged its second premise, arguing that the claimed functional equivalence is not robust to dynamic or structural perturbations [5]. Others have questioned its philosophical basis, asserting that its restrictive, behaviorist-like methodology is inappropriate for consciousness science [6]. Kleiner and Hoel generalized this challenge into a broader "substitution argument," demonstrating that the independence of a theory's predictive data (internal observables) from its inferential data (reports) leads to the *a priori* falsification of any minimally informative theory. This generalization has prompted further debate, with critics arguing that certain classes of functionalist theories are immune to such substitutions [7,8].

However, critical gaps in the literature exist. First, there has been a failure to rigorously define the systems under consideration (FNN and RNN) from the outset. Neither the Doerig paper nor any subsequent paper has made any attempt to mathematically formalize the argument, relying instead on

references to the computer science and neural network literature. Second, there has been a failure to consider how synaptic plasticity might alter the functional properties of a recurrent neural network.

Therefore, in this paper, we explore how the non-stationarity introduced by plasticity in a recurrent neural network results in a system that cannot be durably replicated by any static feedforward architecture. Our technically-grounded argument provides a critical caveat to the unfolding argument and provides a new set of theoretical bounds on theories of consciousness.

The Plasticity Caveat

It is well understood that biological neural networks do not merely process static inputs; the dynamical trajectory of neuronal activity is key to understanding how they function [9]. Biological neural networks are not static computational devices but continuously evolving systems where connectivity patterns change based on activity history. This plasticity operates across multiple timescales, including millisecond-scale synaptic facilitation that is relevant to the timescale on which perception unfolds [10–13].

To formally explore the implications of plasticity for the unfolding argument and for theories of consciousness, we begin by defining the systems under consideration (S1 Supplementary Appendix, Introduction). We recognize that the biophysics underlying actual neurobiological systems is continuous in nature; the biological systems under consideration are continuously evolving. However, for the purposes of the mathematical discussion, we treat these as discretely evolving systems (systems evolving in discrete time steps). The unfolding argument itself relies on a treatment of neural network systems as discrete (an RNN is unfolded into the equivalent FNN based on the timesteps of the input to the RNN). Beyond that, treatment of these systems as discretely evolving captures the essence of the key

considerations associated with the introduction of plasticity, while avoiding complexities necessitated in the treatment of continuously evolving systems.

We proceed to reiterate the unfolding argument using the rigorously defined mathematical systems under consideration (S1 Supplementary Appendix, Introduction). We then mathematically introduce the implications for the unfolding argument when the input-output mapping in the RNN is allowed to change over time (plasticity in the RNN). Formally, let us consider the RNN R_ϕ, which denotes an RNN with plasticity defined as such.

When such plasticity is present, the unfolding argument becomes substantively bounded. Suppose we have the FNN F_ϕ such that F_ϕ reflects the unfolded functional equivalent of the RNN R_ϕ at the time of unfolding (call it t_0). We demonstrate that for all $t > t_0$, the evolution of the input-output map in R_ϕ results in divergence between R_ϕ and the originally unfolded FNN F_ϕ. As described below, we formally delineate the functional, structural, dynamical, and information-theoretic implications of this divergence. Also, extending an argument initially put forth by Usher, we formally delineate the implications from a complexity growth and resource constraint perspective [5].

Direct Response to the Unfolding Argument

Our analysis does not refute the original unfolding argument within its stated constraints - namely, functional input-output equivalence over finite time horizons with fixed parameters. Rather, we formally delineate the boundary conditions within which this equivalence holds. This has significant implications for computational theories of consciousness. Specifically, we show that for theories that explicitly incorporate plastic recurrent dynamics, the challenge presented by the unfolding argument is avoided. The argument remains relevant for theories that can be reduced to static causal structures (arguably [5,6,8,14]), but loses force when applied to genuinely dynamic, adaptive systems that change

their own structure through experience on the same short time scales as the neuronal interactions underlying perceptual decision making.

First, for premise P2, which asserts that any recurrent system can be unfolded into a functionally equivalent feedforward one, we demonstrate that this premise is invalidated under conditions of plasticity. We demonstrate that this claimed equivalence breaks down due to learning-induced functional divergence, the exponential growth of a plastic system's functional repertoire, and the prohibitive resource scaling required for unfolding.

Similarly, premise P3 is also undermined. This premise claims that functionally identical systems are empirically indistinguishable. This statement is not directly challenged, but rather, we show that the antecedent condition to P3 is not met because we demonstrate that plastic recurrent systems and their static feedforward counterparts are not functionally identical. Our arguments do not involve looking 'inside' the network, but rather demonstrate that the experimenter can ascertain based on the behavior of the system if it is feedforward vs plastic recurrent. This is because their learning trajectories may diverge over relevant timescales (depending on the learning rate), because they possess different information-theoretic capacities for memory and prediction, and because they exhibit distinguishable responses to perturbation. To distinguish systems empirically, one could conduct controlled experiments where multiple systems are tested under the same perturbation protocol. The key is that the different behavioral patterns emerge from the input-output function itself, not from examining internal architecture.

Proofs

We present here a description of the formal arguments that establish precise mathematical boundaries on the applicability of the unfolding argument to RNN systems with plasticity. The formal proofs are detailed in the S1 Supplementary Appendix, which are available for review and appraisal by the technically and mathematically savvy reader. The summaries of these arguments, presented below, serve as plain language explanations rather than as substitutions for the mathematically rigorous arguments made in S1 Supplement.

Let us continue to examine an RNN R_ϕ with an evolving input-output map as a function of time. Let us also continue to consider the unfolded equivalent FNN F_ϕ. As stated, we concede that those networks would reflect an equivalent input-output function at the time of unfolding (call it t_0). While an RNN and an FNN might be functionally equivalent at that initial time point, any non-trivial learning in the RNN will alter its parameters and, due to the evolving nature of neural network functions, necessarily change its input-output mapping over time. This evolution in the input-output function is mathematically represented as $\phi^R{}_t : X^t \rightarrow Y$. Since the FNN is static, this functional equivalence breaks for all $t > t_0$ (Figure 1). The key insight is that learning, a ubiquitous feature of biological neural systems, inherently creates time-varying functions that are incompatible with any static representation (S1 Supplementary Appendix, Learning Limits Unfolding). In fact, even more fundamentally, a static FNN processes a fixed number of variables as input (corresponding to the matching time t_0), while the input space of the RNN expands to calculate the output with information gathered at each learning time. This increases the dimensionality of the input space. Biologically plausible mechanisms like Hebbian learning or spike-timing-dependent plasticity continuously modify synaptic strengths based on neural

activity. This directly challenges premise P2 of the unfolding argument; if consciousness depends on systems that learn and adapt, no single static FNN can replicate their behavior over extended periods.

Figure 1: **Visual representation of the functional divergence argument.** A recurrent neural network R with plastic weights evolves through learning rule L, while a static feedforward network F maintains fixed parameters W_F. Initially at t=0, both networks can be functionally equivalent, implementing the same input-output mapping f_0=f_F. However, as R learns and its parameters evolve, its implemented function changes from f_0 to f_{t*}. Since F cannot modify its function f_F, functional equivalence necessarily breaks at some time $t_* > 0$ (shown in red). This divergence persists for all subsequent times, demonstrating that no single static network can maintain functional equivalence with a learning system over arbitrary time intervals.

Next, we present an information-theoretic analysis, targeting the claim that FNNs can match the information processing capabilities of RNNs (S1 Supplementary Appendix, Learning Implies More Information). The core insight is that the fixed-window architecture of an FNN imposes a hard bound on its memory, while the internal state of the equivalent RNN has a much more robust informational capacity. This is particularly relevant for biological functions like working memory and episodic memory, which require integrating information over extended and variable time periods. We formally prove that, due to the monotonicity of mutual information, a fixed-window FNN is necessarily less informative for prediction tasks than an RNN that can capture this historical information (Figure 2). This challenges premise P2 of the unfolding argument by demonstrating that RNNs and static FNNs can be distinguished by their information processing capacities, especially in tasks requiring long-term memory.

Figure 2: **Visual representation of the argument from information-theoretic limitations.** The complete history of the process up to time t is represented by H_t (large gray box), which includes both recent past and distant past information (hatched area, $k < t\text{-}m$). A static Feedforward Neural Network (FNN) predicts the next state x_{t+1} (green circle) using only information from a fixed recent window W_t (blue box), capturing predictive information I_F (blue arrow). A Recurrent Neural Network (RNN), through its internal state h_t, can potentially summarize the entire history H_t and uses this to predict x_{t+1}, capturing predictive information I_R (red arrow). There is non-zero information from the distant past (hatched area) relevant for predicting x_{t+1} (represented by the dashed gray arrow indicating a long-range dependency). This information is available to the RNN but not to the FNN, which is limited to the window W_t. Consequently, the FNN's predictive capacity is strictly limited compared to the RNN's ($I_F < I_R$).

We then examine the fundamental non-equivalence from the perspective of dynamical systems theory (S1 Supplementary Appendix, Learning at Discrete Times Needs Countable Unfolding). This examination targets the assumption that an FNN can capture the essential dynamical properties of an RNN. The critical insight here is that RNNs are orbit generators capable of long-term temporal evolution, whereas FNNs compute finite-horizon maps (Figure 3). When the NNs are considered from the perspective of the function space they represent, a static FNN is a single point in the function space, while an RNN changes at each learning time (discrete in our definition) and so represents an orbit in the function space rather than a single point. Using the language of dynamical systems, we show that RNNs and their unfolded FNN counterparts belong to different mathematical categories. An FNN represents a fixed, time-T map of the RNN's dynamics but cannot generate the full orbit that characterizes the behavior of the discrete-time RNN. This challenges the sufficiency of functional equivalence assumed by Doerig et al., because plastic RNNs cannot be reduced to static input-output mappings without losing these essential properties. Usher informally makes the case for this for standard RNNs in his initial work on the unfolding argument [5]. This proof adds rigor and definitively demonstrates it to be the case for plastic RNNs.

Figure 3: **Visual representation of the argument from dynamical systems perspective.** A static FNN represents a fixed point in function space (red), while a learning RNN generates an orbit (blue trajectory). The unfolding theorem can create an FNN equivalent to any single point on the RNN's orbit (purple), but cannot capture the orbit-generating dynamics in function space.

Next, we leverage principles of feedback control and stability to target the assumption that an open-loop FNN can replicate closed-loop RNN dynamics for the case of discrete time networks (S1 Supplementary Appendix, Learning Implies Perturbation Response). We demonstrate that plastic RNNs

and static FNNs respond fundamentally differently to temporal perturbations in their input-output behavior. While FNNs forget the effects of any input change after a fixed number of steps (determined by their input window size), RNNs maintain persistent traces of perturbations in their hidden states indefinitely (Figure 4). This temporal persistence property of RNNs - their ability to integrate and carry forward information from the distant past - cannot be replicated by static FNNs regardless of their size. This has direct relevance to consciousness, which arguably requires sustained integration of information across time rather than the bounded memory of windowed processing [2,15–17]. This challenges premise P3 by showing that evolving RNNs and static FNNs respond differently to perturbations in their input-output function behavior, potentially making them experimentally distinguishable through stability tests [18–20] and other control-theoretic measures. Indeed, the clinical use of perturbational complexity index (PCI) [21] to assess for signatures of consciousness is reliant on a long-window response to perturbation [22], a feature present in RNNs with specific intrinsic causal structure and absent in FNNs [22,23]. It should be noted that this does not violate P4 as we are specifically discussing differences in the input-output behavior, not differences in the internal workings of the networks.

Figure 4: **Visual representation of the argument from perturbational stability.** A localized input perturbation δ is introduced at time k (red spike). The static Feedforward Neural Network (FNN, orange) shows a transient response that completely disappears after m steps ($t \geq k+m$), due to its bounded sliding window memory. The Recurrent Neural Network (RNN, blue) maintains persistent traces of the perturbation indefinitely through its hidden state dynamics.

Our last area of formal investigation addresses the growth in complexity inherent to plastic systems and the implications of this for realistic, physically constrained biological systems. We provide a constructive proof demonstrating that plasticity within an RNN results in the ability of the network to approximate an arbitrarily large family of functions, whereas any single, static FNN is only able to approximate itself (S1 Supplementary Appendix, Learning Implies Approximation Capacity) [24]. This creates an unbridgeable gap between the functional repertoire of a plastic RNN and the capacity of any

single, static FNN. A related result returns to the first argument, Corollary 3.5 (S1 Supplementary Appendix, Learning Limits Unfolding, No Finite Family of FNNs Can Track a Plastic RNN). When paired, these results form a powerful demonstration of the limits of FNNs in terms of their capability to reconstitute the functionality inherent to plastic RNNs (Figure 5). These results imply that even with perfect knowledge of the future behavior of a plastic RNN, it is impossible to switch to a finite library of FNNs to recapitulate that functionality. (NB: due to stochasticity and to chaotic behavior [yielding a deterministic yet non-predictable system] inherent to biological neural network systems, such perfect knowledge itself is impossible to attain.)

Figure 5: **Visual representation of the argument from complexity growth and resource constraints.** A plastic RNN (blue) achieves expanding functional capacity over time while operating with fixed biological resources (gold bar, constant width). The RNN's approximation capacity grows vertically through learning, eventually capable of approximating arbitrary functions within the continuous function space $C(X, Y)$. In contrast, static FNNs (red points) form a finite library $\{F_1, F_2, ..., F_{\square}\}$ where each network can only approximate itself, creating a bounded functional repertoire. The hatched region represents the mathematical proof that no finite collection of static networks can access the functional space achievable by a plastic RNN with fixed resources. While the RNN maintains constant resource requirements, equivalent FNN implementations would require unbounded resource growth (approaching infinity).

While the concept of resource capacity is not well-defined in the neural network literature, these arguments also have implications for the recapitulation of biological RNNs in FNN systems. By 'resource capacity,' we refer to the physical and computational constraints of a network, such as its total number of neurons and connections, as well as the metabolic or computational costs associated with its operation. A biological RNN operates using a physically constrained set of resources that, while dynamic and plastic, remains bounded over time. The functionality of this biological RNN cannot be replicated by a finite library of FNNs and, through learning, can approximate any arbitrary set of input-output functions. Thus, real, physically constrained, plastic biological RNNs cannot be

re-instantiated in a FNN. The brain has a well-recognized capacity for lifelong (but probably bounded) learning within what might be considered quite strict metabolic, anatomical, and developmental constraints. These formal findings challenge the practical applicability and relevance of premise P2, as maintaining functional equivalence over time would require an unbounded library of FNNs.

Discussion

Our formal analysis reveals that the unfolding argument's challenge to causal structure theories only holds if the connection weights in the network are unchanging on the time scale of conscious perceptual events. We build upon Doerig et al.'s important contribution by identifying specific boundary conditions; specifically, the analyses presented establish a natural division in the theoretical landscape, transforming the unfolding challenge from a universal critique into a diagnostic tool. This tool distinguishes theories based on whether they depend on static or genuinely dynamic properties. On one side of this divide are static causal structure theories, which remain vulnerable to the original unfolding argument. This category includes theories that identify consciousness with fixed connectivity patterns, reduce it to instantaneous causal relationships, or treat neural architecture as unchanging hardware. On the other side are dynamic causal structure theories, which are protected by the boundary conditions we have established (depending on the time scale). These are theories that incorporate learning and adaptation, emphasize temporal evolution and history-dependence, and treat plasticity as a fundamental component of conscious processing. This creates a natural taxonomy of consciousness theories based on their vulnerability to the unfolding challenge, with dynamic and plastic theories being offered a counter-argument based on the mathematical boundaries we have established.

Specific Implications for Major Theories of Consciousness

This finding has significant implications for physicalist theories of consciousness. It supports theoretical frameworks where the *how* of information processing – involving time, dynamics, plasticity, and specific architectures – is as important as the *what* of the input-output function. This is underscored by neuroanatomical evidence. Recurrent connections, for instance, often terminate near the axon hillock, allowing them to exert powerful influence over neuronal output, whereas feedforward inputs tend to land more distally on the dendrites. Such architectural details, which directly impact computational dynamics, are abstracted away and lost in a static, structurally agnostic FFN.

The arguments related to resource constraints echo Dennett: "sometimes an *impossibility in fact* is theoretically more interesting than a *possibility in principle*." [25] Imagine the brain's recurrent neural networks, which may be represented as a vast set (in the mathematical sense) of recurrent neural networks. (It is due to plasticity that we should represent it as a set of RNNs rather than as a single RNN; plasticity means that the underlying weights, connections, and even node count of the network will change over time.) While it may be possible to unfold each RNN in this set into an equivalent finite set of feedforward architectures, this would still fail to capture the open-ended, plastic processes occurring in the brain. Each static FNN would represent an instant, but not the evolving state of the brain. It would perhaps be too strong to contend that consciousness *requires* flexible adaptation and learning. However, in the case that this contention is conceded, we have demonstrated that static FNNs lack the capacity for this continual state integration and adaptation.

For Integrated Information Theory (IIT), [2] the original challenge from Doerig and colleagues contended that a system with a high degree of integrated information (Φ) could be functionally

mimicked by a feedforward system with negligible Φ, making IIT either false or unscientific [4]. However, our boundary conditions substantially constrain this critique. When plasticity is introduced, both Φ and the so-called transition probability matrix (TPM) are no longer static properties but become temporally dynamic. Learning-induced changes to connectivity will necessarily alter both the system's integrated information over time and the temporal evolution dynamics. This temporal evolution of Φ and TPM provides novel, testable hypotheses. For instance, one could measure whether Φ increases during learning tasks when subjects acquire new conscious representations, or test whether disrupting plasticity (e.g., through synaptic blockers) affects both learning and consciousness in parallel [26]. These types of experiments could be done *in silico* or in simple model systems (e.g. *Caenorhabditis elegans*) to provide systems with tractable measurement of Φ. Additionally, the TPM should show measurable changes across timescales that correspond to conscious processing episodes rather than remaining static. (NB: this is much more difficult to provide a tractable measurement of, due to the complexity of TPM measurement.) Because plastic systems with different causal structures can diverge functionally over time, the premise of sustained functional equivalence breaks down. Consequently, while static formulations of IIT remain challenged, a dynamical version that incorporates the evolution of Φ and TPM during learning may evade the unfolding critique. The finding in the original paper that arbitrary functions could be built with arbitrarily high Φ is of no consequence on this view; we concede that plastic RNNs, and conscious systems, can indeed be trained on arbitrary input-output functions. (Note, the foregoing is based on a strong interpretation of IIT; see e.g. Mediano et al. 2022 for a more nuanced discussion of the point [27]. IIT proponents themselves have used the unfolding equivalence didactically, emphasizing that 'being is not doing.' Our analysis additionally suggests that incorporating plasticity into IIT models could resolve this tension.)

Global Neuronal Workspace Theory (GNWT), which proposes that consciousness involves information being "globally broadcast" within a dynamic workspace involving recurrent processing, also finds support [1]. Our analysis adds nuance to the claim that it is protected by being a functional, rather than structural, theory. A static implementation of a global workspace might be vulnerable to the unfolding argument. However, a dynamic workspace that incorporates plasticity, allowing for learning-dependent changes in information access patterns, avoids it. This reframing makes the evolution of the workspace during perceptual information processing episodes a key testable prediction, meaning GNWT's immunity to the unfolding argument depends on whether it incorporates these dynamic properties.

Recurrent Processing Theory (RPT), developed by Lamme and colleagues, was one of the original targets of the unfolding argument and provides particularly strong support for our plasticity-based analysis. RPT posits that consciousness arises when recurrent interactions between visual areas integrate features into coherent perceptual wholes, distinguishing it from unconscious feedforward processing [28]. Critically, Lamme explicitly identifies plasticity as a potential "missing ingredient" that distinguishes conscious from unconscious recurrent processing. He proposes that not all recurrent connections are constitutive of consciousness, but specifically those recurrent processes that can induce network plasticity through mechanisms like NMDA receptor activation and synaptic modifications at various timescales, see e.g. Section 5 and Figure 3 from Lamme 2018. This aligns precisely with our formal arguments: RPT incorporating rapid plasticity would avoid the unfolding argument by maintaining the kind of temporally extended, history-dependent dynamics that our proofs demonstrate cannot be captured by static feedforward networks. Under this plasticity-enriched version of RPT, only recurrent processes capable of modifying network connectivity on perception-relevant

timescales would generate conscious experience, providing both protection from the unfolding argument and a principled boundary for consciousness.

The recently introduced state space theory (SST) of consciousness [17] arguably gets around the UA, as it explicitly incorporates multiple dynamic elements. SST posits that consciousness arises from plastic, recurrent "engines" performing delay coordinate embedding. This process of dynamic state reconstruction cannot be captured by static networks. Also, its reliance on history-dependent plasticity prevents lasting functional equivalence, and its hierarchical structure requires the kind of unbounded temporal integration that static FNNs cannot provide. An orbit-based perspective on the (lack of) evolution of a static FNN versus the function orbit of a plastic RNN lends the strongest support for SST. Within SST, conscious experience is cast as the temporal evolution of a high-dimensional state vector within a dynamic, layered set of structured state spaces. On this view, consciousness arises not merely from static states, but from the trajectory, or orbit, that the system traces through its state space in response to sensory inputs, memory, and internal dynamics. If we model the associated neural substrate as a dynamical system, akin to a recurrent neural network (RNN), then conscious processes are naturally understood as the real-time evolution of hidden states, continually integrating perceptual history and ongoing experience. SST therefore exemplifies how theories that integrate multiple dynamic and plastic mechanisms gain cumulative protection from the unfolding challenge.

In broader terms, these formal arguments shift the landscape by explicitly acknowledging the impact of the original unfolding argument on static theories of consciousness and providing a mechanism by which these theories can adjust to avoid those impacts. Nota bene: theories that ascribe functional importance to plasticity, dynamics, and architectural properties of neural systems do not, on

this view, have any additional *a priori* credibility; they simply avoid failing due to the impacts of the unfolding argument.

By formally demonstrating the limitations of unfolding under plausible conditions, this work helps delineate the computational regimes where recurrence is not merely an implementation choice but a fundamental requirement, thereby refining the search space for the neural correlates and computational principles of consciousness.

Restoring Empirical Testability

Our argument establishes that plastic recurrent architectures cannot be reduced to static feedforward ones, allowing certain theories to escape the unfolding argument. This mathematical distinction between plastic and static systems suggests certain empirical differences that could be investigated.

If one accepts that consciousness requires the kind of irreducible recurrence that plasticity preserves (a theoretical commitment beyond our argument), then several testable predictions might follow. First, learning capacity could correlate with consciousness because plastic systems inherently combine computation with adaptation, while static systems only compute [29]. Second, temporal integration spans could matter because plastic RNNs maintain information across arbitrary timescales through evolving internal states, unlike static networks with fixed temporal windows. Third, disrupting plastic feedback loops would impair consciousness differently than disrupting static ones, since the former carry historically-accumulated information while the latter only transmit current signals. Fourth, developmental plasticity trajectories might predict conscious capabilities if consciousness depends on the accumulation of adaptive changes over time.

However, these predictions require additional theoretical assumptions about consciousness that our mathematical argument alone does not establish. What we can definitively say is only that theories incorporating short-time-scale plasticity avoid the unfolding argument's critique and therefore remain viable candidates for empirical testing, not that they are correct or that these specific predictions necessarily follow.

Limitations, Extensions, and Open Questions

A significant limitation concerns the relationship between the timescales of consciousness and plasticity. Many conscious processes occur rapidly, while significant structural plasticity can take much longer. This raises the empirical questions of whether rapid forms of synaptic plasticity are sufficient to provide protection from unfolding and whether conscious processes depend more on ongoing plastic changes or their accumulated historical effects. As advocated for in He 2023, there may be multiple different sorts of neurobiological mechanisms working in concert or separately in support of consciousness [30]. In support of our argument, short-term synaptic plasticity and spike timing dependent plasticity have been well documented to occur on relevant timescales for perception and consciousness, including one review identifying these as relevant for sensory timing [31–34].

Another limitation is that our mathematical proofs establish boundaries in principle, but biological realities such as metabolic energy costs and stability-plasticity tradeoffs may impose additional constraints. Finally, our proofs are theory-neutral; they demonstrate which dynamic properties are lost in unfolding but do not, by themselves, establish which of these properties are necessary for consciousness. To clarify how these results bear on specific theories, we must distinguish between two ways a theory might link architecture to consciousness.

The first concerns the *capacity* for consciousness versus the *actual instantiation* of conscious states.[35] Measures like Φ in IIT are computed from the system's causal structure - specifically, the transition probability matrix - and purport to quantify the degree of consciousness a system possesses by virtue of its architecture. According to IIT, Φ is meant to measure both dispositional and occurrent properties of conscious systems: the amount of integrated information is the amount of consciousness actually present at that moment. So if a network happens to be in a highly integrated state, IIT would say it is experiencing a correspondingly high level of consciousness. By contrast, measures like PCI capture only an occurrent property: what the system is currently doing, not what it is capable of [21,22].

The unfolding argument primarily targets dispositional claims: it demonstrates that any causal structure can be replicated (in terms of input-output function) by alternative architectures with arbitrarily different Φ values. Our plasticity caveat, however, concerns instantiation. A plastic RNN and its static FNN "equivalent" may share the same input-output function at time t_0, but they diverge in their ongoing dynamics, in what they are actually *doing* as they process information over time. The FNN computes a fixed mapping; the RNN undergoes a trajectory through function space in modifying its own parameters through learning.

This distinction sharpens the theoretical landscape. Theories identifying consciousness with structural capacity remain vulnerable to the unfolding argument regardless of our caveat. However, theories identifying consciousness with instantiated processes (the actual temporal unfolding of dynamics, including real-time parameter modification through plasticity) escape both the original argument and its implications. Notably, IIT's proponents claim that Φ measures actual consciousness, not merely capacity for consciousness. If this claim is to be maintained, our analysis suggests that incorporating the dynamic evolution of the TPM through plasticity may be necessary to preserve IIT against unfolding-style objections. What matters for consciousness may

not be the structural capacity to integrate information, but the ongoing process of actually doing so.

One might object that even if plastic RNNs cannot be tracked by finite FNN libraries, at each moment some alternative FNN implementation exists or one could continually 'bolt on' to the existing FNN architecture at the next time step. Even if granted, the alternative implementation at each moment would need to track precisely the dynamical trajectory and plasticity-mediated changes of the original system. We grant that this magic alternative-implementation FNN zombie with the dynamical and plasticity properties of the original RNN-based system could, in fact, be conscious (depending on the theory). Note: such a system is no longer an FNN in any sense relevant to the unfolding argument. It is an RNN implemented in alternative substrate. Whether consciousness requires dynamical properties (temporal thickness, real-time state-space traversal, history-dependent trajectories) or whether these are merely contingent features of biological consciousness remains an open empirical question that our formal arguments help to sharpen but do not resolve. We welcome further debate on this idea. The question of what constitutes a 'genuine' implementation raises a related but distinct issue: whether consciousness might require properties that exceed discrete computation entirely.

Our formal arguments are based on discrete-time RNNs and remain agnostic on whether biologically relevant continuous dynamics or non-halting environmental interaction are essential to consciousness, as some have argued (Lucas 1961; Penrose 1989). We are not claiming that consciousness involves super-Turing computation, nor that digital systems are in principle incapable of consciousness. Rather, our analysis suggests that the unfolding argument may rest on assumptions about discretizability and functional characterizability that deserve scrutiny.

To be clear about the scope of this suggestion: the process-based view outlined above does not require hypercomputation; it requires only that consciousness supervene on dynamics rather than function. Indeed, the hidden assumption that the brain is a Turing machine dissolves when one considers that it is a dynamical, plastic, interacting entity. Work looking at interaction machines in the Turing sense has shown that these machines are strictly more powerful than a Turing machine, because they represent a different kind of computation that can't be reduced to input-output functions [37,38]. However, the deeper question of whether continuous physical processes can ever be fully captured by discrete computational simulation (as opposed to approximation) may prove relevant. The distinction between simulating dynamics and instantiating them, between a numerical approximation of a trajectory and the trajectory itself, echoes the distinction between computing a function and undergoing a process. We leave this as an open question that future work may help to resolve.

Meta-Theoretical Implications and the Path Forward

By delineating these boundaries, our analysis shifts the burden of proof in consciousness science. Proponents of static theories must now explain why consciousness does not appear to require the dynamic, adaptive, and plastic properties that are so central to biological systems. Conversely, proponents of dynamic theories must specify precisely which of these properties they deem necessary and provide empirical evidence to support their claims. This suggests new research directions focusing on the relationship between consciousness and learning, development, and pathological plasticity [29].

Our analysis suggests that progress in consciousness science may benefit from embracing dynamic theoretical frameworks that cannot be reduced to static input-output mappings. We acknowledge that just because a theory has been constructed to be immune to the unfolding argument does not necessarily make it a stronger contender as a theory of consciousness. Based on our arguments,

the unfolding argument and its caveats become a valuable guide. It helps identify which theoretical approaches can yield genuine insight into the nature of consciousness while remaining grounded in empirical science. Our arguments suggest that consciousness science may benefit from embracing the difficulties and complexities inherent to a view of the brain as an intrinsically dynamic, adaptive system.

Conclusions

In conclusion, our formal analysis of the unfolding argument's boundaries tells us that theories under which consciousness depends on recurrence might not be invalidated by the unfolding argument if they incorporate, or at least allow for, fast plasticity in recurrent networks. This dynamic view is both scientifically respectable, grounded in rigorous mathematical proofs, and empirically productive, suggesting specific experimental approaches for advancing our understanding of consciousness.

The unfolding argument need not eliminate theories that depend on recurrence, provided they incorporate plasticity on the timescale of conscious neural processes. Specifically, theories that rely on recurrent dynamics (like IIT and GWT) might avoid the unfolding argument's critique by adopting plastic rather than static architectures. However, this revision is not universally applicable. Theories that do not depend on recurrence in the first place remain unaffected by our analysis, particularly those that abstract away from implementation details. We identify a specific modification (incorporating plasticity) that could allow certain recurrence-based theories to remain viable despite the unfolding argument's constraints. This caveat to the unfolding argument restores viability to theories of consciousness that focus on neural wiring, if they allow for plasticity on the timescale of conscious neural processes.

References

1. Seth AK, Bayne T. Theories of consciousness. Nat Rev Neurosci. 2022;23: 439–452. doi:10.1038/s41583-022-00587-4

2. Albantakis L, Barbosa L, Findlay G, Grasso M, Haun AM, Marshall W, et al. Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. PLoS Comput Biol. 2023;19: e1011465. doi:10.1371/journal.pcbi.1011465

3. Doerig A, Schurger A, Herzog MH. Hard criteria for empirical theories of consciousness. Cogn Neurosci. 2021;12: 41–62. doi:10.1080/17588928.2020.1772214

4. Doerig A, Schurger A, Hess K, Herzog MH. The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. Conscious Cogn. 2019;72: 49–59. doi:10.1016/j.concog.2019.04.002

5. Usher M. Refuting the unfolding-argument on the irrelevance of causal structure to consciousness. Conscious Cogn. 2021;95: 103212. doi:10.1016/j.concog.2021.103212

6. Usher M, Negro N, Jacobson H, Tsuchiya N. When philosophical nuance matters: safeguarding consciousness research from restrictive assumptions. Front Psychol. 2023;14: 1306023. doi:10.3389/fpsyg.2023.1306023

7. Kleiner J, Hoel E. Falsification and consciousness. Neurosci Conscious. 2021;2021: niab001. doi:10.1093/nc/niab001

8. Ganesh N. No substitute for functionalism -- A reply to “Falsification & consciousness.” arXiv [cs.OH]. 2020. Available: http://arxiv.org/abs/2006.13664

9. Rabinovich MI, Varona P, Selverston AI, Abarbanel HDI. Dynamical principles in neuroscience. Rev Mod Phys. 2006;78: 1213–1265. doi:10.1103/revmodphys.78.1213

10. Chambers AR, Rumpel S. A stable brain from unstable components: Emerging concepts and implications for neural computation. Neuroscience. 2017;357: 172–184. doi:10.1016/j.neuroscience.2017.06.005

11. Magee JC, Grienberger C. Synaptic plasticity forms and functions. Annu Rev Neurosci. 2020;43: 95–117. doi:10.1146/annurev-neuro-090919-022842

12. Fauth M, Tetzlaff C. Opposing effects of neuronal activity on structural plasticity. Front Neuroanat. 2016;10: 75. doi:10.3389/fnana.2016.00075

13. Zenke F, Gerstner W. Hebbian plasticity requires compensatory processes on multiple timescales. Philos Trans R Soc Lond B Biol Sci. 2017;372. doi:10.1098/rstb.2016.0259

14. Negro N. Phenomenology-first versus third-person approaches in the science of consciousness: the case of the integrated information theory and the unfolding argument. Phenomenol Cogn Sci. 2020;19: 979–996. doi:10.1007/s11097-020-09681-3

15. Northoff G, Zilio F. From shorter to longer timescales: Converging Integrated Information Theory (IIT) with the temporo-spatial theory of consciousness (TTC). Entropy (Basel). 2022;24: 270. doi:10.3390/e24020270

16. Hirschhorn R, Kahane O, Gur-Arie I, Faivre N, Mudrik L. Windows of integration hypothesis revisited. Front Hum Neurosci. 2020;14: 617187. doi:10.3389/fnhum.2020.617187

17. O'Reilly-Shah VN. Delay coordinate embedding as neuronally implemented information processing: The state space theory of consciousness. J Conscious Stud. 2025;32: 127–159. doi:10.53765/20512201.32.1.132

18. Schurger A, Sarigiannidis I, Naccache L, Sitt JD, Dehaene S. Cortical activity is more stable when sensory stimuli are consciously perceived. Proc Natl Acad Sci U S A. 2015;112: E2083–92. doi:10.1073/pnas.1418730112

19. Eisen AJ, Kozachkov L, Bastos AM, Donoghue JA, Mahnke MK, Brincat SL, et al. Propofol anesthesia destabilizes neural dynamics across cortex. Neuron. 2024;0. doi:10.1016/j.neuron.2024.06.011

20. Eisen AJ, Bardon AG, Ballesteros JJ, Bastos AM, Donoghue JA, Mahnke MK, et al. Similar destabilization of neural dynamics under different general anesthetics. Neuroscience. bioRxiv; 2025. Available: https://www.biorxiv.org/content/10.1101/2025.08.21.671540v1.full.pdf

21. Casarotto S, Comanducci A, Rosanova M, Sarasso S, Fecchio M, Napolitani M, et al. Stratification of unresponsive patients by an independently validated index of brain complexity: Complexity Index. Ann Neurol. 2016;80: 718–729. doi:10.1002/ana.24779

22. Casali AG, Gosseries O, Rosanova M, Boly M, Sarasso S, Casali KR, et al. A theoretically based index of consciousness independent of sensory processing and behavior. Sci Transl Med. 2013;5: 198ra105. doi:10.1126/scitranslmed.3006294

23. Maschke C, O'Byrne J, Colombo MA, Boly M, Gosseries O, Laureys S, et al. Critical dynamics in spontaneous EEG predict anesthetic-induced loss of consciousness and perturbational complexity. Commun Biol. 2024;7: 946. doi:10.1038/s42003-024-06613-8

24. Song C, Hwang G, Lee J ho, Kang M-J. Minimal width for universal property of deep RNN. J Mach Learn Res. 2023;24: 121:1–121:41. doi:10.48550/arXiv.2211.13866

25. Dennett DC. Consciousness Explained. Harlow, England: Penguin Books; 1993.

26. Albantakis L, Hintze A, Koch C, Adami C, Tononi G. Evolution of integrated causal structures in animats exposed to environments of increasing complexity. PLoS Comput Biol. 2014;10: e1003966. doi:10.1371/journal.pcbi.1003966

27. Mediano PAM, Rosas FE, Bor D, Seth AK, Barrett AB. The strength of weak integrated information theory. Trends Cogn Sci. 2022;26: 646–655. doi:10.1016/j.tics.2022.04.008

28. Lamme VAF. Challenges for theories of consciousness: seeing or knowing, the missing ingredient and how to deal with panpsychism. Philos Trans R Soc Lond B Biol Sci. 2018;373. doi:10.1098/rstb.2017.0344

29. Cleeremans A. Consciousness: the radical plasticity thesis. In: Banerjee R, Chakrabarti BK, editors. Progress in Brain Research. Elsevier; 2007. pp. 19–33. doi:10.1016/S0079-6123(07)68003-0

30. He BJ. Towards a pluralistic neurobiological understanding of consciousness. Trends Cogn Sci. 2023;27: 420–432. doi:10.1016/j.tics.2023.02.001

31. Motanis H, Seay MJ, Buonomano DV. Short-term synaptic plasticity as a mechanism for sensory timing. Trends Neurosci. 2018;41: 701–711. doi:10.1016/j.tins.2018.08.001

32. Regehr WG. Short-term presynaptic plasticity. Cold Spring Harb Perspect Biol. 2012;4: a005702. doi:10.1101/cshperspect.a005702

33. Barri A, Mongillo G. Short-term synaptic plasticity: Microscopic modelling and (some) computational

implications. Adv Exp Med Biol. 2022;1359: 105–121. doi:10.1007/978-3-030-89439-9_5

34. Li G, McLaughlin DW, Peskin CS. A biochemical description of postsynaptic plasticity-with timescales ranging from milliseconds to seconds. Proc Natl Acad Sci U S A. 2024;121: e2311709121. doi:10.1073/pnas.2311709121

35. Mudrik L, Mylopoulos M, Negro N, Schurger A. Theories of consciousness and a life worth living. Current Opinion in Behavioral Sciences. 2023;53: 101299. doi:10.1016/j.cobeha.2023.101299

36. Butlin P, Long R, Bayne T, Bengio Y, Birch J, Chalmers D, et al. Identifying indicators of consciousness in AI systems. Trends Cogn Sci. 2025;0. doi:10.1016/j.tics.2025.10.011

37. Wegner P, Goldin D. Computation beyond turing machines. Commun ACM. 2003;46: 100–102. doi:10.1145/641205.641235

38. Goldin D, Wegner P. The interactive nature of computing: Refuting the strong church–Turing thesis. Minds Mach (Dordr). 2008;18: 17–38. doi:10.1007/s11023-007-9083-1

Figure 1

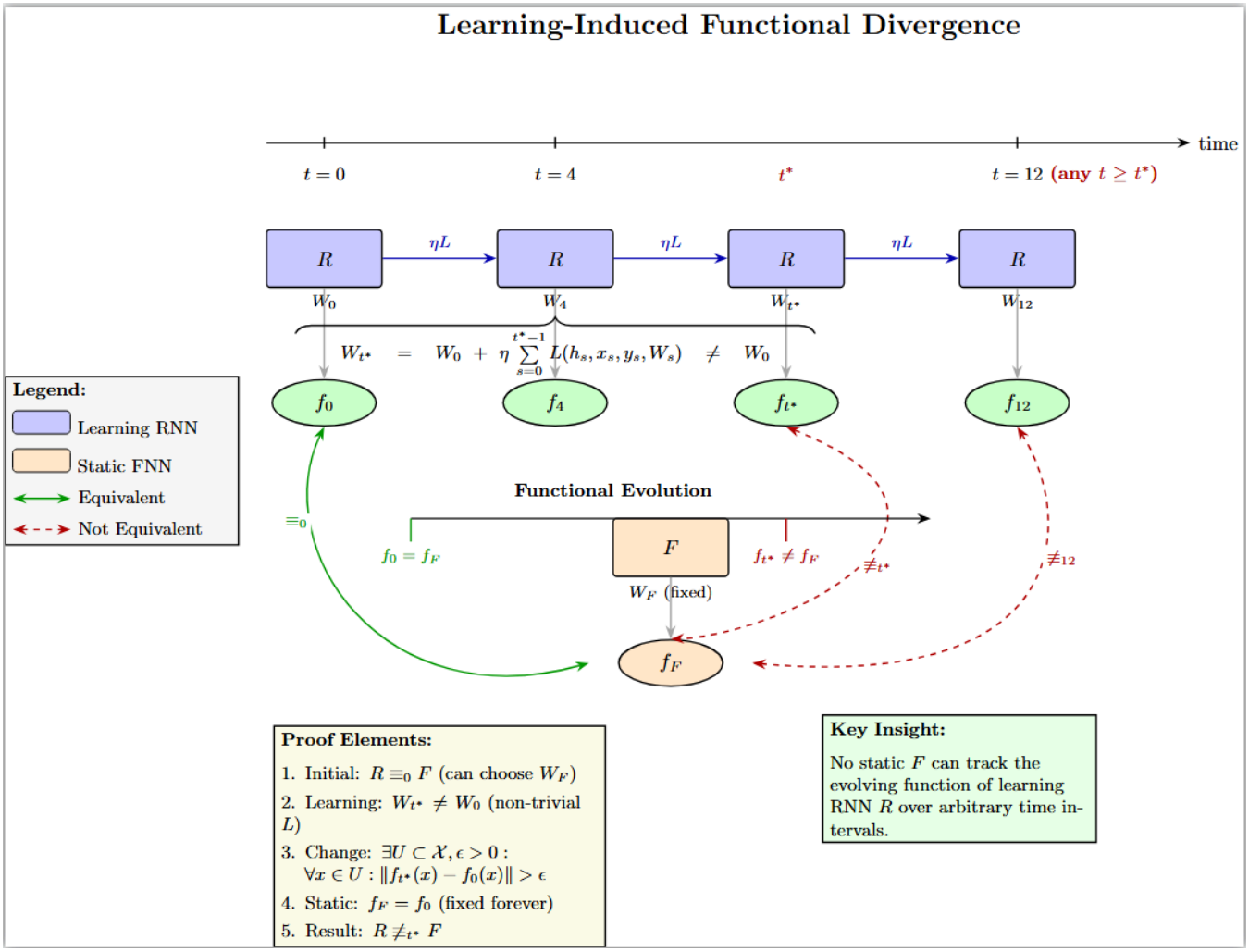

Learning-Induced Functional Divergence
time
$t = 0$
$t = 4$
t^*
$t = 12$ (any $t \geq t^*$)
R
ηL
W_0
W_4
W_{t^*}
W_{12}
$W_{t^*} = W_0 + \eta \sum_{s=0}^{t^*-1} L(h_s, x_s, y_s, W_s) \neq W_0$
f_0
f_4
f_{t^*}
f_{12}
Legend:
Learning RNN
Static FNN
Equivalent
Not Equivalent
Functional Evolution
$\equiv_0$
$f_0 = f_F$
F
W_F (fixed)
$f_{t^*} \neq f_F$
$\not\equiv_{t^*}$
$\not\equiv_{12}$
f_F
Proof Elements:
1. Initial: $R \equiv_0 F$ (can choose W_F)
2. Learning: $W_{t^*} \neq W_0$ (non-trivial L)
3. Change: $\exists U \subset \mathcal{X}, \epsilon > 0 :$ $\forall x \in U : \|f_{t^*}(x) - f_0(x)\| > \epsilon$
4. Static: $f_F = f_0$ (fixed forever)
5. Result: $R \not\equiv_{t^*} F$
Key Insight:
No static F can track the evolving function of learning RNN R over arbitrary time intervals.

Figure 2

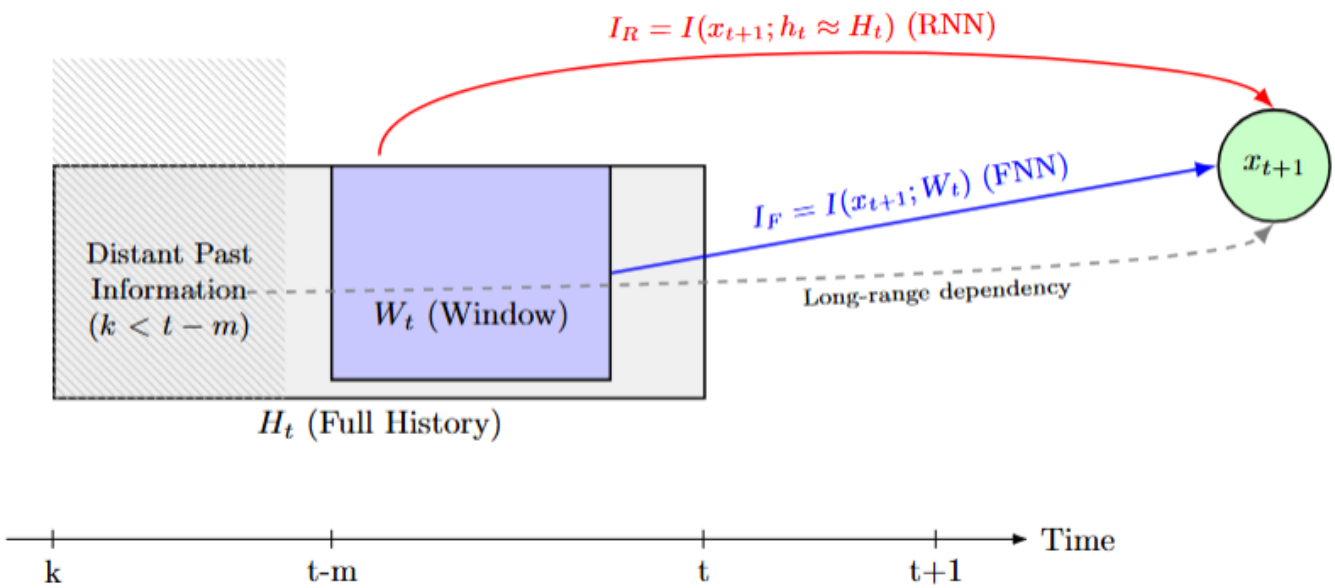

$I_R = I(x_{t+1}; h_t \approx H_t)$ (RNN)
x_{t+1}
$I_F = I(x_{t+1}; W_t)$ (FNN)
Distant Past
Information
$(k < t - m)$
W_t (Window)
Long-range dependency
H_t (Full History)
Time
k
t-m
t
t+1

Figure 3

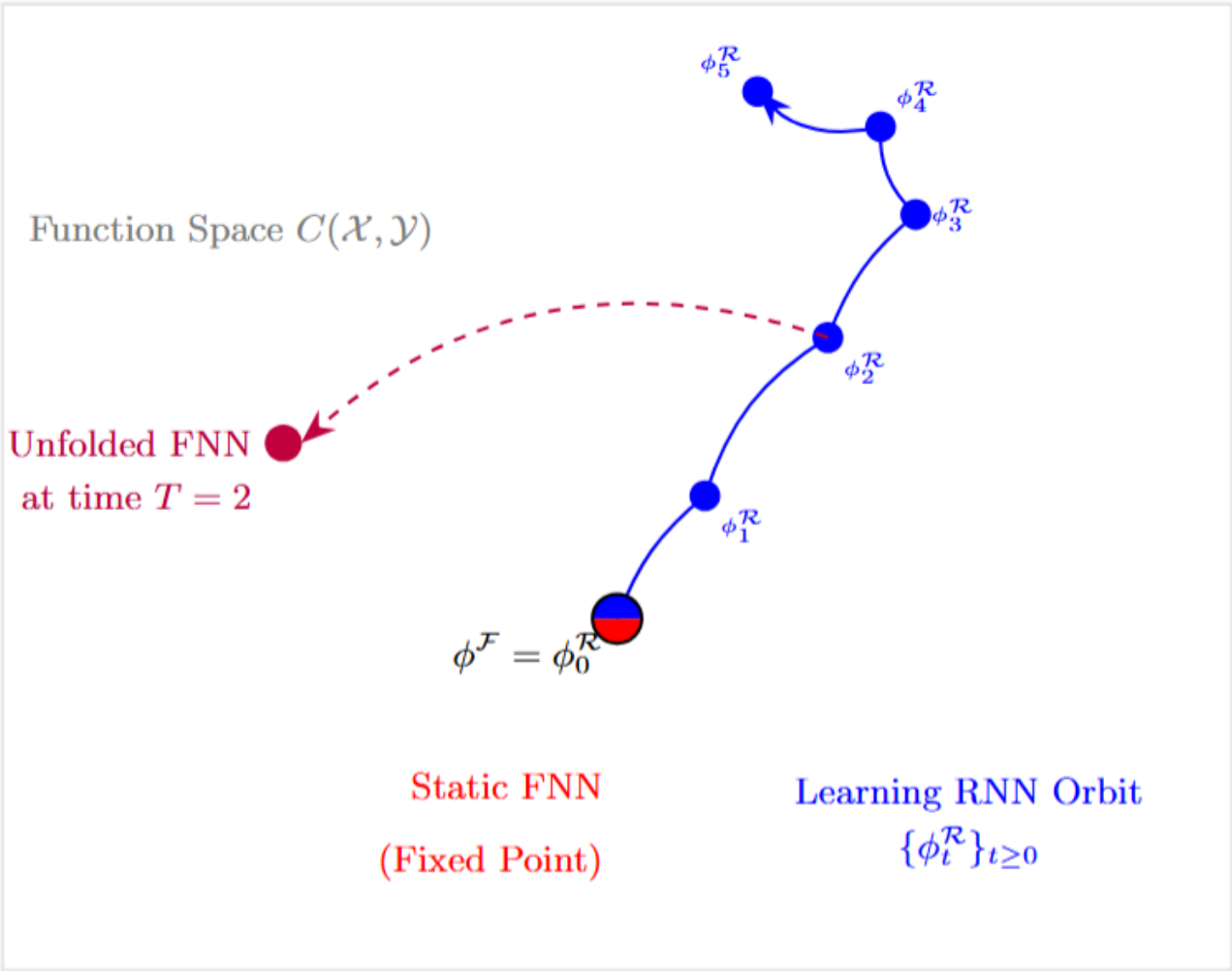

$\phi_5^{\mathcal{R}}$
$\phi_4^{\mathcal{R}}$
$\phi_3^{\mathcal{R}}$
Function Space $C(\mathcal{X}, \mathcal{Y})$
$\phi_2^{\mathcal{R}}$
Unfolded FNN
at time $T = 2$
$\phi_1^{\mathcal{R}}$
$\phi^{\mathcal{F}} = \phi_0^{\mathcal{R}}$
Static FNN
(Fixed Point)
Learning RNN Orbit
$\{\phi_t^{\mathcal{R}}\}_{t \geq 0}$

Figure 4

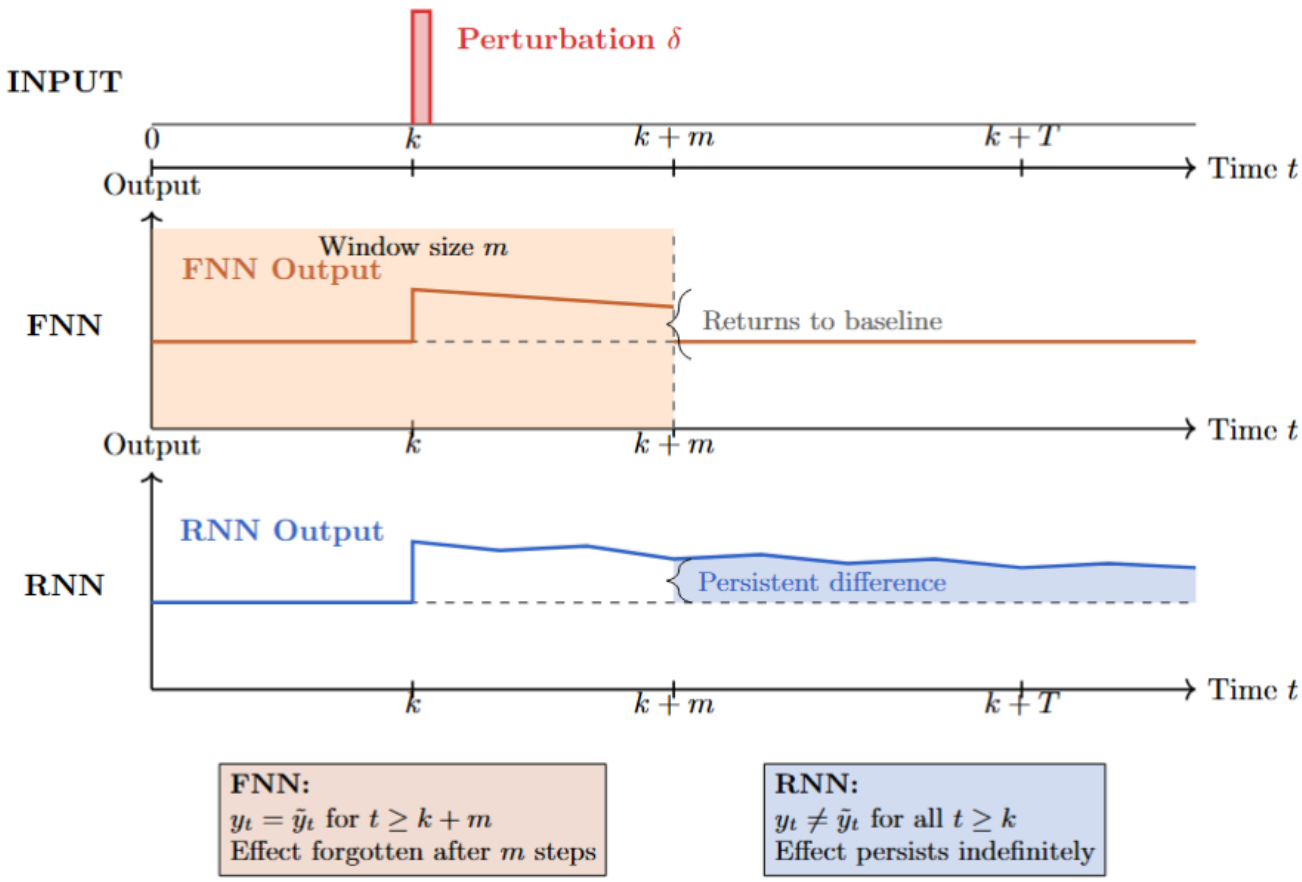
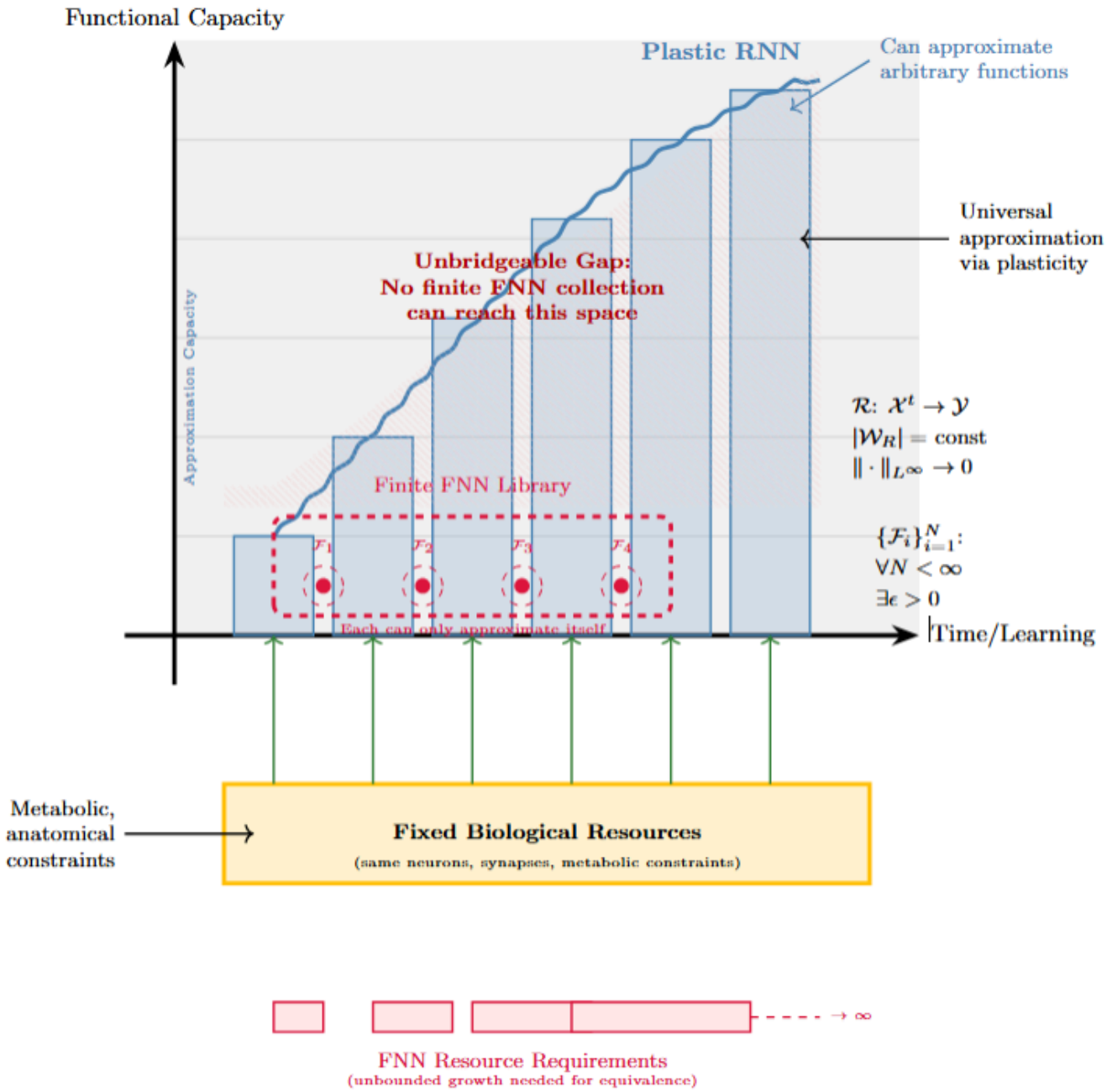

Perturbation δ
INPUT
0
k
$k+m$
$k+T$
Time t
Output
Window size m
FNN Output
FNN
Returns to baseline
Output
k
$k+m$
Time t
RNN Output
RNN
Persistent difference
k
$k+m$
$k+T$
Time t
FNN:
$y_t = \tilde{y}_t$ for $t \geq k+m$
Effect forgotten after m steps
RNN:
$y_t \neq \tilde{y}_t$ for all $t \geq k$
Effect persists indefinitely

Figure 5


Functional Capacity
Plastic RNN
Can approximate arbitrary functions
Universal approximation via plasticity
Unbridgeable Gap:
No finite FNN collection can reach this space
Approximation Capacity
$\mathcal{R}: \mathcal{X}^t \to \mathcal{Y}$
$|\mathcal{W}_R| = \text{const}$
$\|\cdot\|_{L^\infty} \to 0$
Finite FNN Library
$\mathcal{F}_1$
$\mathcal{F}_2$
$\mathcal{F}_3$
$\mathcal{F}_4$
$\{\mathcal{F}_i\}_{i=1}^N$:
$\forall N < \infty$
$\exists \epsilon > 0$
Each can only approximate itself
Time/Learning
Metabolic, anatomical constraints
Fixed Biological Resources
(same neurons, synapses, metabolic constraints)

$\to \infty$
FNN Resource Requirements
(unbounded growth needed for equivalence)