A Caveat Regarding the Unfolding Argument: Implications of Plasticity

Authors
Vikas N. O'Reilly-Shah, MD, PhD 1#
Alessandro Maria 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

Corresponding Author
Professor of Anesthesiology & Pain Medicine
University of Washington School of Medicine, RR450, 1959 NE Pacific St, Seattle, WA 98195
voreill@uw.edu

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 boundary condition on 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. We show that plasticity is a concrete instance of lenient dependency between behavioral and internal observables, restoring empirical testability to theories that incorporate plasticity on perception-relevant timescales. Our results do not establish that recurrence, plasticity, or process are necessary for consciousness; they establish that the unfolding argument does not preclude empirical investigation of whether these properties matter.

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 (Seth & Bayne 2022). Modern physicalist theories of consciousness, in particular integrated information theory (IIT), have focused on the causal structure of neural network systems (Albantakis et al. 2023). 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 (2021; 2019) 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. 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 (Doerig et al. 2019):

- **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. (2019) 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 (Usher 2021). Others have questioned its philosophical basis, asserting that its restrictive, behaviorist-like methodology is inappropriate for consciousness science (Usher et al. 2023). Kleiner and Hoel (2021) 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 (Kleiner & Hoel 2021; Ganesh 2020). Notably, in an article responding to various critiques, Herzog, Schurger, and Doerig (2022) claim an extension of the unfolding argument to all functions that Turing Complete (TC) systems can calculate, and therefore "any function, and even every process can be implemented with many different causal structures." This is a significant generalization. Understood in the broader context, the unfolding argument and its Turing Completeness generalization can be seen as technical instantiations of the classical multiple realizability (MR) argument from philosophy of mind (Shapiro 2000; Kim 1992). MR, as originally formulated by

Putnam (1967), holds that the same mental state type can be realized by systems with different physical or causal structures. The UA makes this claim precise for neural networks; the TC generalization extends it to arbitrary computational substrates.

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. Plasticity helps to address the Kleiner-Hoel challenge. Our technically grounded argument provides a critical caveat to the unfolding argument and establishes a new set of theoretical bounds on the testability of 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 (Rabinovich et al. 2006). 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 (Chambers & Rumpel 2017; Magee & Grienberger 2020; Fauth & Tetzlaff 2016; Zenke & Gerstner 2017).

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 $R_{\square}$, 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 $F_{\square}$ such that $F_{\square}$ reflects the unfolded functional equivalent of the RNN $R_{\square}$ 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 $R_{\square}$ results in divergence between $R_{\square}$ and the originally unfolded FNN $F_{\square}$. A key distinction here is that the trajectory divergence we are identifying is a divergence in the *function space* - the mapping itself between inputs and outputs is evolving. This is as compared to the *state space* trajectory; a fixed RNN may describe a state space trajectory *within* its fixed weights in a way that a fixed FNN cannot (as pointed out by Usher (2021)). Concretely, a fixed-weight RNN processing a sequence traverses its hidden state attractor but always implements the same function φ, say, recognizing faces. A plastic RNN

might start implementing φ_0 (recognizing faces) but end, after plasticity implementing φ_t (recognizing faces and associating learned names). The function itself has changed - movement in function space, not just state space.

Claims

To ensure that the claims made in this manuscript are clear, we enumerate them below. These claims are the subject of the analysis presented in and taken up by the remainder of the manuscript.

1. **We do claim that the UA's premise does not apply to plastic systems, a caveat to UA premise P2.** The UA compares systems with fixed input-output functions *at a single time step*. Plastic systems do not have fixed functions; they traverse function space over time. The question "can a system with different internal organizational structure compute the same function?" presupposes there is a single function to compare. For plastic systems, this presupposition is false, because the internal organization is undergoing temporal evolution. The argument remains relevant for theories that can be reduced to static organizational structures. The Plasticity and the Unfolding Argument section outlines these arguments; S1 Supplementary Appendix Sections 3-7 details them.
2. **We do claim that perturbational differences outlined directly challenge UA premise P3**. This difference is distinct from Usher's original perturbation argument (Usher 2021); both challenge the unfolding argument, but Usher's argument operates in the hidden state space trajectory of a fixed RNN, whereas our argument operates in the function space trajectory of a plastic RNN, a more powerful result. Information-theoretic differences also provide a challenge to P3.
3. **We do <u>not</u> claim hypercomputation, or that substrate matters for consciousness.** We agree that the Turing Complete generalization implies that any plastic RNN could be implemented in an alternative substrate. A plastic RNN implemented in silicon with the same dynamical internal

organization as one in biological tissue would be functionally indistinguishable on our analysis. However, these two systems might be distinguishable on perturbation. Additionally, in order for the alternative substrate to causally track the behavior of the original, it would need access to the input data stream of the original. Finally, a Turing Machine cannot *a priori* know the trajectory of the plastic RNN without precomputing it. These points are causally relevant to multiple realizability; neither operates outside the computability parameters of the standard Church-Turing thesis.

4. **We do <u>not</u> claim, or advocate for, any specific theory of consciousness.** We present boundary conditions on the UA that may or may not apply to specific theories. Implications for particular formulations are discussed, but this manuscript should not be viewed as advocating for a specific theory.

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: (Usher 2021; Usher et al. 2023; Negro 2020; Ganesh 2020)), 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. The summaries of these arguments, presented below, serve as plain language explanations of these rigorous proofs. A note on terminology as it relates to this section: we use the terms 'learning' and 'plasticity' in this section interchangeably and only in the purely mathematical sense. Both terms, for the purposes of this section, refer to a nontrivial change to the network parameters (weights) between timesteps. *Learning*, here, is not meant in the more biological, task-directed sense.

Let us continue to examine an RNN $R_{\square}$ with an evolving input-output map as a function of time. Let us also continue to consider the unfolded equivalent FNN $F_{\square}$. 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 plasticity 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 $\square^{R}_{t} : X^t \rightarrow Y$. Since the FNN is static, this functional equivalence breaks for all $t > t_0$ (Figure 1).

The key insight is that plasticity, 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 time step through nontrivial weight changes (plasticity). 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. We also note that the folding-unfolding relationship is asymmetric: any feedforward network can be trivially folded into a recurrent network, but the converse does not hold in general, establishing strict containment.

Figure 1: **Visual representation of the functional divergence argument.** A recurrent neural network R with plastic weights evolves through plasticity rule P, 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 updates and its parameters evolve, its implemented function changes from f_0 to $f_{\square\square}$. Since F cannot modify its function f_F, functional equivalence necessarily breaks at some time $t_{\square} > 0$ (shown in red). This divergence persists for all subsequent times,

demonstrating that no single static network can maintain functional equivalence with a plastic 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. (2019), because plastic RNNs cannot be reduced to static input-output mappings without losing these essential properties. Usher (2021) informally makes the case for this for standard RNNs in his initial work on the unfolding argument. 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 (Albantakis et al. 2023; Northoff & Zilio 2022; Hirschhorn et al. 2020; O’Reilly-Shah 2025b).

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 (Schurger et al. 2015; Eisen et al. 2024; Eisen et al. 2025) and other control-theoretic measures. Indeed, the clinical use of perturbational complexity index (PCI) (Casarotto et al. 2016) to assess for signatures of consciousness is reliant on a long-window response to perturbation (Casali et al. 2013), a feature present in RNNs with specific intrinsic organizational structure and absent in FNNs (Maschke et al. 2024; Casali et al. 2013). 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.

This function space perturbation is more powerful than Usher's (2021) state space perturbation point because it operates at the level of the input-output mapping itself rather than the internal dynamics. Usher (2021) correctly observes that a fixed RNN and its unfolded FNN counterpart traverse different state space trajectories - the hidden states diverge even when both systems compute the same function. However, a proponent of the UA could respond that these are precisely the kind of internal differences that P4 declares inadmissible: so long as the input-output function remains identical, the differing state trajectories are not experimentally accessible without examining the system's internal workings. Our perturbation argument sidesteps this objection. When plasticity is present, input perturbations do not merely alter the hidden state trajectory while leaving the computed function intact; they alter the function itself. A plastic RNN modifies its weights during processing, changing what the system computes for all subsequent inputs. This change is observable at the input-output level without any appeal to internal measurement. The FNN, lacking this mechanism, returns to its fixed baseline after the input perturbation exits its input window. This divergence in observable input perturbation response is therefore immune to the P4 objection that constrains Usher's original argument.

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) (Song et al. 2023). 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 $\{F\square, F\square, ..., F\square\}$ 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.

The Turing Completeness Generalization

These plasticity arguments address the original formulation of the unfolding argument. However, Herzog et al. (2022) have proposed a stronger version, generalizing the UA beyond the RNN/FNN comparison to all Turing Complete (TC) systems. They argue that "any function, and even every process can be implemented with many different causal structures," concluding that "empirical results about consciousness are not provably due to causal structure."

On its face, this generalization would render our plasticity-based boundary conditions moot: one could concede that no single static FNN tracks a plastic RNN while maintaining that some TC system with different internal organizational structure could serve as a behaviorally identical substitute. We argue that this inference, while intuitive, is actually comprised of two distinct claims: that any

computable function can be realized on diverse substrates (which Turing Completeness does guarantee) and that any specific realization of a computation can be reproduced with different internal organization in a manner that is indistinguishable under experimental probing (which it does not). Computational class membership guarantees the existence of some program computing a given function; it does not guarantee the existence of a realization that is process-equivalent to an arbitrary implementation in a different system.

This distinction is consistent with the multiple realizability literature, which likewise distinguishes between computational equivalence and causally relevant computational differences (Shapiro 2000; Kim 1992). Perturbation provides an empirical lever. As we demonstrate for the plasticity case (S1 Supplementary Appendix, Section 6), systems with different internal organization respond differently to mid-course interventions, and these differences manifest at the input-output level, not solely in internal observables. A system that reuses the same parameters at every step (an RNN) and one that instantiates independent parameter copies at each step (an unfolded FNN, or an alternative TC realization) will generically exhibit different perturbation profiles. This extends to any proposed TC substitute: matching the input-output function of a plastic RNN at a fixed time horizon does not entail matching its response to perturbation, and perturbation response is experimentally accessible without violating P4. We do not address the question of whether such perturbational differences are relevant to consciousness. Differences under perturbation establish that TC systems implementing the same computation are empirically distinguishable. A fuller formal treatment of this non-entailment is beyond the scope of the present manuscript and will be developed in forthcoming work.

Additionally, for systems engaged in ongoing environmental interaction (where future inputs depend on prior outputs and on unpredictable environmental dynamics) the input stream to the plastic RNN or any TC implementation of it cannot be pre-specified on a static tape. Any TC system that causally tracks the behavior of a plastic RNN in such a setting must itself receive the same

environmental inputs and update its internal state accordingly. Such a system would, by construction, share the original's dynamical internal organization and would not constitute the kind of "different causal structure, same function" substitute that the UA or TC generalization requires.

The Interactive Turing Machine (ITM) literature formalizes this distinction between function-computation and process-execution, referring to *a priori* unpredictable environmental inputs as ω-streams (Goldin & Wegner 2008; van Leeuwen & Wiedermann 2001; Cabessa & Siegelmann 2012; Cabessa & Siegelmann 2014). While some in the ITM community were interested in so-called super-Turing computation, modern formulations of ITMs do not appeal to hypercomputation (Fu 2022; Martin et al. 2022; Martin et al. 2023; Jaeger et al. 2023). The implications of environmental coupling on computability and computational multiple realizability are important but, again, beyond the scope of this manuscript. A fuller discussion of ITMs and these implications will be developed in forthcoming work. We note here only that this environmental-coupling process-based consideration may provide an obstruction to unfolding and the TC generalization that is independent of, and orthogonal to, the plasticity arguments we have presented.

Lenient Dependency and the Restoration of Empirical Testability

Our results have bearing on a critical result in the consciousness literature: the Kleiner and Hoel (2021) substitutability argument. This argument demonstrates that if behavioral reports and internal observables are independent, then any minimally informative theory is already falsified (K&H Theorem 3.10) and if strictly dependent, then the theories are unfalsifiable (K&H Theorem 4.3). Only lenient dependency, where reports constrain but do not determine internal observables, permits empirical testability. Kleiner and Hoel identify no current theory meeting this criterion.

However, plastic systems exhibit this intermediate relationship. The intuition is straightforward: the longer you observe a plastic system's behavior, the more you can 'rule out' possible kinds of systems

that might be producing it. A static surrogate might match the plastic system's outputs for a short window, but it cannot match the persistent perturbation traces and growing information capacity that our formal results establish. As the observation horizon lengthens, systems that lack plasticity produce observably different behavioral profiles, and the space of candidate systems consistent with the full behavioral record narrows.

More formally, our perturbation and information-theoretic results provide the concrete exclusion mechanism. A localized perturbation at time k induces output differences at all subsequent times in a plastic RNN, whereas a static FNN forgets the perturbation after a fixed number of steps determined by its window size m (S1 Supplementary Appendix, Section 6). For any observation horizon longer than this window, the static surrogate's behavior diverges from the plastic system's: it can no longer account for the observed behavior. Similarly, the mutual information between target and RNN output grows with observation horizon, while a window-limited FNN hits a ceiling imposed by its fixed window (S1 Supplementary Appendix, Section 3). Together, these results progressively exclude static surrogates as the observation horizon lengthens.

This rules out the first horn of the Kleiner-Hoel dilemma: extended behavioral observation of a plastic system excludes static surrogates, because no static unfolding or finite family of feedforward networks can reproduce the behavioral profile over time (S1 Supplementary Appendix, Section 3). But the second horn, that behavior uniquely determines internal organization, is not implied either. Even systems that obey the same plasticity laws and produce identical behavioral traces may differ internally in ways that leave no observable signature, such as how individual units are labeled or how internal states are coordinated. Behavior constrains what the underlying system could be, without pinning it down to exactly one possibility. This is the intermediate relationship Kleiner and Hoel termed lenient dependency: the only regime in which theories of consciousness remain empirically testable. Hoel

(2025), citing a preprint version of this paper, recently arrived at a convergent conclusion via a different mathematical approach.

NB: This escape route from the unfolding and substitutability arguments does not yield any additional *a priori* credibility for any theory of consciousness. Theories incorporating plasticity simply avoid the UA's critique and remain viable candidates for empirical investigation. This is a restoration of testability, not a claim of correctness.

Discussion

Our formal analysis maps the scope conditions of the unfolding argument, identifying where its premises hold and where they require qualification. We build upon Doerig et al.'s (2019) important contribution by identifying specific boundary conditions; specifically, the analyses presented establish a natural division in the theoretical landscape. It remains true that any theory claiming special computational significance of recurrence *per se* fails to meet the challenge of the unfolding argument or the Turing Completeness generalization. No theory gains credibility on our analysis; we metaphorically open the door, but do not walk through it. Our analysis says simply that empirical testability is restored for theories that explicitly incorporate plasticity.

Specific Implications for Major Theories of Consciousness

This finding has significant implications for physicalist theories of consciousness. It restores to testability theoretical frameworks where the *how* of information processing - involving time, dynamics, and fast plasticity - is as important as the *what* of the input-output function. Our findings do not challenge the view that, computationally, the plastic RNN structure could be realized in different causal forms. The arguments we have made here have no implications for the specific *implementation* of the computations performed by a plastic RNN.

The arguments related to resource constraints echo Dennett (1993): "sometimes an *impossibility in fact* is theoretically more interesting than a *possibility in principle*." 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 open-ended, plastic processes. Each static FNN would represent an instant, but not the evolving state of the brain. Static FNNs lack the capacity for this continual state integration and adaptation.

For Integrated Information Theory (IIT) (Albantakis et al. 2023), the original challenge from Doerig and colleagues (2019) 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. A relevant distinction is the *capacity* for consciousness versus the *actual instantiation* of conscious states (Mudrik et al. 2023). Measures like Φ in IIT are computed from the system's causal structure - specifically, the transition probability matrix (TPM) - 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 (Casarotto et al. 2016; Casali et al. 2013). .

Computationally, the underlying causal structure of a candidate system and its TPM can still be expressed in any unfolded FNN or Turing Complete system. Consequently, IIT remains challenged and our findings do not alter the conclusions of Doerig et al. (2021; 2019) (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. IIT proponents themselves have used the unfolding equivalence didactically, emphasizing that 'being is not doing.' Our analysis suggests that incorporating plasticity into IIT models could resolve this tension, though IIT proponents would need a principled rather than ad hoc reason for doing so.)

Global Neuronal Workspace Theory (GNWT) (Mashour et al. 2020; Dehaene & Naccache 2001) proposes that consciousness involves information being "globally broadcast" within a dynamic workspace involving recurrent processing. Our analysis adds nuance to the claim in Doerig et al. that it is protected by being a functional, rather than structural, theory. Thinking mechanistically, a static implementation of a global workspace would be vulnerable to the unfolding argument. However, a dynamic workspace that incorporates plasticity, allowing for learning-dependent changes in information access patterns, is testable in our view. This framing makes the *evolution* of the workspace during perceptual information processing episodes the key to developing testable predictions.

Recurrent Processing Theory (RPT), developed by Lamme (2018), was also one of the original targets of the unfolding argument. Our results do not rescue RPT from the unfolding argument. RPT posits that consciousness arises when recurrent interactions between visual areas integrate features into coherent perceptual wholes, distinguishing it from unconscious feedforward processing (Lamme 2018). In fairness, Lamme (2018) explicitly identifies plasticity as a potential "missing ingredient" that distinguishes conscious from unconscious recurrent processing, but we leave it to the defenders of RPT to develop the point.

The state space theory (SST) of consciousness (O’Reilly-Shah 2025b; O’Reilly-Shah 2026) posits that consciousness arises from delay coordinate embedding (DCE) in plastic, recurrent neural networks. Recent mathematical (Duan et al. 2023; Hart 2025) and empirical (Uribarri & Mindlin 2022; Ostrow et al. 2024) work demonstrate that RNNs (LSTMs, reservoir computers) trained on dynamical

time series data make use of embedding structure for prediction. These findings potentially link the behavior of neural RNNs to perception (O'Reilly-Shah & Selvitella 2026). DCE inherently requires temporal depth because it is incoherent to speak of DCE without time series data. Plasticity ensures the embedding evolves to match internal and environmental dynamics monitored and reported on via neuronal spike trains, creating an exploitable relation (Shea 2018). These are fundamental commitments and arise directly from the core hypothesis (O'Reilly-Shah 2025a). This caveat paper formalizes obstructions mentioned in the original article where SST was developed (O'Reilly-Shah 2025b).

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 by He (2023), there may be multiple different sorts of neurobiological mechanisms working in concert or separately in support of consciousness. 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 (Motanis et al. 2018; Regehr 2012; Barri & Mongillo 2022; Li et al. 2024).

Our mathematical proofs establish boundaries in principle, but biological realities such as metabolic energy costs and stability-plasticity tradeoffs may impose additional constraints. As stated, 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.

One might object that even if plastic RNNs cannot be tracked by finite FNN libraries, one could instantiate the same computation by continuously replacing the RNN with its FNN, or just implementing it in another Turing Complete system. This alternative implementation would need to, at each moment, track precisely the dynamical trajectory and plasticity-mediated changes of the original system. Such a hypothetical alternative-implementation/FNN zombie with the dynamical and plasticity properties of the original RNN-based system would be identical in terms of empirical evidence for consciousness. However, a few observations:

1. Such a system is no longer a static FNN in any sense relevant to the unfolding argument. It reflects the computational function space trajectory of the original plastic RNN implemented in an alternative substrate. Standard Universal Approximation Theorems guarantee that a static FNN can approximate a specific target function to arbitrary accuracy (see e.g. Augustine 2024). They do not, however, guarantee that a single static architecture can approximate an evolving family of functions (a trajectory through function space) without itself undergoing parameter modification.
2. A static FNN could not change structure/weight in response to input without an external agent acting on the FNN (or would require instant-to-instant replacement.) A plastic RNN, or its implementation on a Turing Complete system, has the power to modify itself.

Whether consciousness requires dynamical properties (temporal thickness, real-time state-space and function-space traversal, history-dependent trajectories) or whether these are merely contingent features of biological consciousness remains an open empirical question. Our arguments support *only* the conclusion that plasticity provides a way to make the recurrence hypothesis testable and within the realm of science.

Our formal arguments are based on discrete-time RNNs. We remain agnostic on whether *continuous*-time dynamics are essential to consciousness. Some have argued that this induces a requirement for super Turing computation (Lucas 1961; Penrose 1989) or relates to the importance of biological substrate in consciousness (e.g., Seth 2025; Milinkovic & Aru 2025). Unlike substrate dependence arguments, our result does not privilege biological implementation; it identifies a formal property, plasticity, that is substrate-neutral. We are also agnostic on the question of instantiation versus simulation: whether a perfect simulation of a plastic RNN *is* conscious or *acts as if* conscious. This is a question our formal arguments cannot resolve. What we have established is more modest but, we believe, significant: theories with specific computational properties are empirically testable.

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 should consider whether the boundary conditions identified here bear on their formulations. 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 (Cleeremans 2007).

Our formal results engage directly with the critical tradition of multiple realizability. On Shapiro's analysis, the question is whether the structural difference between the plastic RNN (whose weights and arity evolve) and static FNN (whose weights and arity are fixed) is relevant to the function at issue (Shapiro 2000). As we've established, it is indeed a relevant question for plastic systems. Dynamical internal organization constitutes a relevant property in precisely Shapiro's sense: it determines the system's future input-output behavior. The distinction between function-level and process-level MR (Shapiro 2000; Polger & Shapiro 2016) thus becomes a diagnostic criterion: theories

that can be reduced to static functional descriptions inherit all the vulnerabilities of classical MR, while theories that invoke genuine temporal dynamics escape them.

Our analysis identifies which theoretical frameworks remain viable candidates for empirical investigation in light of the unfolding argument's scope conditions. A theory constructed to be immune to the unfolding argument does not make it a stronger contender as a theory of consciousness. The unfolding argument and its caveats are a valuable guide. It helps identify which theoretical approaches can yield genuine insight into the nature of consciousness while remaining grounded in empirical science.

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 on the timescale of conscious neural processes. We emphasize that nothing in our analysis provides independent reason to believe plasticity is relevant to consciousness; our results identify boundary conditions on a logical argument, not evidence for any particular theory. The dynamic view is both scientifically respectable, grounded in rigorous mathematical proofs, and empirically productive, suggesting specific experimental approaches for advancing our understanding of consciousness.

References

Albantakis, L. et al., 2023. Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLoS computational biology*, 19(10), p.e1011465. Available at: http://dx.doi.org/10.1371/journal.pcbi.1011465.

Augustine, M.T., 2024. A Survey on Universal Approximation Theorems. *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/2407.12895.

Barri, A. & Mongillo, G., 2022. Short-term synaptic plasticity: Microscopic modelling and (some) computational implications. *Advances in experimental medicine and biology*, 1359, pp.105–121. Available at: http://dx.doi.org/10.1007/978-3-030-89439-9_5 [Accessed September 23, 2025].

Cabessa, J. & Siegelmann, H.T., 2012. The computational power of interactive recurrent neural networks. *Neural Computation*, 24(4), pp.996–1019. Available at: http://dx.doi.org/10.1162/NECO_a_00263 [Accessed January 27, 2026].

Cabessa, J. & Siegelmann, H.T., 2014. The super-Turing computational power of plastic recurrent neural networks. *International Journal of Neural Systems*, 24(8), p.1450029. Available at: http://dx.doi.org/10.1142/S0129065714500294.

Casali, A.G. et al., 2013. A theoretically based index of consciousness independent of sensory processing and behavior. *Science translational medicine*, 5(198), p.198ra105. Available at: https://pubmed.ncbi.nlm.nih.gov/23946194/.

Casarotto, S. et al., 2016. Stratification of unresponsive patients by an independently validated index of brain complexity: Complexity Index. *Annals of neurology*, 80(5), pp.718–729. Available at: https://pubmed.ncbi.nlm.nih.gov/27717082/.

Chambers, A.R. & Rumpel, S., 2017. A stable brain from unstable components: Emerging concepts and implications for neural computation. *Neuroscience*, 357, pp.172–184. Available at: http://dx.doi.org/10.1016/j.neuroscience.2017.06.005 [Accessed August 7, 2025].

Cleeremans, A., 2007. Consciousness: the radical plasticity thesis. In R. Banerjee & B. K. Chakrabarti, eds. *Progress in Brain Research*. Elsevier, pp. 19–33. Available at: https://www.sciencedirect.com/science/article/pii/S0079612307680030.

Dehaene, S. & Naccache, L., 2001. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79(1--2), pp.1–37.

Dennett, D.C., 1993. *Consciousness Explained*, Harlow, England: Penguin Books.

Doerig, A. et al., 2019. The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and cognition*, 72, pp.49–59. Available at: http://dx.doi.org/10.1016/j.concog.2019.04.002.

Doerig, A., Schurger, A. & Herzog, M.H., 2021. Hard criteria for empirical theories of consciousness. *Cognitive neuroscience*, 12(2), pp.41–62. Available at: http://dx.doi.org/10.1080/17588928.2020.1772214.

Duan, X.-Y. et al., 2023. Embedding theory of reservoir computing and reducing reservoir network using time delays. *Physical Review Research*, 5(2), p.L022041. Available at: http://dx.doi.org/10.1103/PhysRevResearch.5.L022041 [Accessed January 28, 2026].

Eisen, A.J. et al., 2024. Propofol anesthesia destabilizes neural dynamics across cortex. *Neuron*, 0(0). Available at: http://www.cell.com/article/S089662732400446X/abstract [Accessed July 25, 2024].

Eisen, A.J. et al., 2025. Similar destabilization of neural dynamics under different general anesthetics. *Neuroscience*. Available at: https://www.biorxiv.org/content/10.1101/2025.08.21.671540v1.full.pdf.

Fauth, M. & Tetzlaff, C., 2016. Opposing effects of neuronal activity on structural plasticity. *Frontiers in neuroanatomy*, 10, p.75. Available at: http://dx.doi.org/10.3389/fnana.2016.00075 [Accessed August 7, 2025].

Fu, Y., 2022. A thesis for interaction. *Theoretical Computer Science*, 897, pp.114–134. Available at: http://dx.doi.org/10.1016/j.tcs.2021.10.001 [Accessed January 22, 2026].

Ganesh, N., 2020. No substitute for functionalism -- A reply to "Falsification & consciousness." *arXiv [cs.OH]*. Available at: http://arxiv.org/abs/2006.13664 [Accessed August 18, 2025].

Goldin, D. & Wegner, P., 2008. The interactive nature of computing: Refuting the strong church–Turing thesis. *Minds and machines*, 18(1), pp.17–38. Available at: http://dx.doi.org/10.1007/s11023-007-9083-1.

Hart, A.G., 2025. Generic and isometric embeddings in reservoir computers. *Chaos (Woodbury, N.Y.)*, 35(11), p.111103. Available at: http://dx.doi.org/10.1063/5.0301957.

He, B.J., 2023. Towards a pluralistic neurobiological understanding of consciousness. *Trends in cognitive sciences*, 27(5), pp.420–432. Available at: http://dx.doi.org/10.1016/j.tics.2023.02.001.

Herzog, M.H., Schurger, A. & Doerig, A., 2022. First-person experience cannot rescue causal structure theories from the unfolding argument. *Consciousness and cognition*, 98, p.103261. Available at: http://dx.doi.org/10.1016/j.concog.2021.103261.

Hirschhorn, R. et al., 2020. Windows of integration hypothesis revisited. *Frontiers in human neuroscience*, 14, p.617187. Available at: http://dx.doi.org/10.3389/fnhum.2020.617187 [Accessed October 24, 2025].

Hoel, E., 2025. A disproof of large language model consciousness: The necessity of continual learning for consciousness. *arXiv [q-bio.NC]*. Available at: http://arxiv.org/abs/2512.12802 [Accessed December 19, 2025].

Jaeger, H., Noheda, B. & van der Wiel, W.G., 2023. Toward a formal theory for computing machines made out of whatever physics offers. *Nature Communications*, 14(1), p.4911. Available at: http://dx.doi.org/10.1038/s41467-023-40533-1.

Kim, J., 1992. Multiple Realization and the Metaphysics of Reduction. *Philosophy and Phenomenological Research*, 52(1), pp.1–26. Available at: http://dx.doi.org/10.2307/2107741.

Kleiner, J. & Hoel, E., 2021. Falsification and consciousness. *Neuroscience of consciousness*, 2021(1), p.niab001. Available at: https://dx.doi.org/10.1093/nc/niab001 [Accessed August 18, 2025].

Lamme, V.A.F., 2018. Challenges for theories of consciousness: seeing or knowing, the missing ingredient and how to deal with panpsychism. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 373(1755). Available at: http://dx.doi.org/10.1098/rstb.2017.0344 [Accessed November 24, 2025].

van Leeuwen, J. & Wiedermann, J., 2001. The Turing machine paradigm in contemporary computing. In *Mathematics Unlimited — 2001 and Beyond*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1139–1155. Available at: http://dx.doi.org/10.1007/978-3-642-56478-9_59.

Li, G., McLaughlin, D.W. & Peskin, C.S., 2024. A biochemical description of postsynaptic plasticity-with timescales ranging from milliseconds to seconds. *Proceedings of the National Academy of Sciences of the United States of America*, 121(7), p.e2311709121. Available at: http://dx.doi.org/10.1073/pnas.2311709121 [Accessed September 23, 2025].

Lucas, J.R., 1961. Minds, machines and gödel. *Philosophy (London, England)*, 36(137), pp.112–127. Available at: http://dx.doi.org/10.1017/s0031819100057983.

Magee, J.C. & Grienberger, C., 2020. Synaptic plasticity forms and functions. *Annual review of neuroscience*, 43(1), pp.95–117. Available at: http://dx.doi.org/10.1146/annurev-neuro-090919-022842 [Accessed August 7, 2025].

Martin, A., Magnaudet, M. & Conversy, S., 2022. Modelling interactive computing systems: Do we have a good theory of what computers are? *Zagadnienia Filozoficzne w Nauce*, (73), pp.77–119. Available at: http://dx.doi.org/10.59203/zfn.73.597 [Accessed January 22, 2026].

Martin, Magnaudet, M. & Conversy, S., 2023. Computers Interactive Machines: Can We Build Explanatory Abstraction? *Minds Machines*, 33, pp.77–106.

Maschke, C. et al., 2024. Critical dynamics in spontaneous EEG predict anesthetic-induced loss of consciousness and perturbational complexity. *Communications biology*, 7(1), p.946. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11300875/.

Mashour, G.A. et al., 2020. Conscious processing and the global neuronal workspace hypothesis. *Neuron*, 105(5), pp.776–798. Available at: http://dx.doi.org/10.1016/j.neuron.2020.01.026.

Mediano, P.A.M. et al., 2022. The strength of weak integrated information theory. *Trends in cognitive sciences*, 26(8), pp.646–655. Available at: http://dx.doi.org/10.1016/j.tics.2022.04.008.

Milinkovic, B. & Aru, J., 2025. On biological and artificial consciousness: A case for biological computationalism. *Neuroscience and Biobehavioral Reviews*, 181(106524), p.106524. Available at: http://dx.doi.org/10.1016/j.neubiorev.2025.106524.

Motanis, H., Seay, M.J. & Buonomano, D.V., 2018. Short-term synaptic plasticity as a mechanism for sensory timing. *Trends in neurosciences*, 41(10), pp.701–711. Available at: http://dx.doi.org/10.1016/j.tins.2018.08.001 [Accessed September 23, 2025].

Mudrik, L. et al., 2023. Theories of consciousness and a life worth living. *Current Opinion in Behavioral Sciences*, 53, p.101299. Available at: https://www.sciencedirect.com/science/article/pii/S2352154623000530.

Negro, N., 2020. Phenomenology-first versus third-person approaches in the science of consciousness: the case of the integrated information theory and the unfolding argument. *Phenomenology and the cognitive sciences*, 19(5), pp.979–996. Available at: http://dx.doi.org/10.1007/s11097-020-09681-3.

Northoff, G. & Zilio, F., 2022. From shorter to longer timescales: Converging Integrated Information Theory (IIT) with the temporo-spatial theory of consciousness (TTC). *Entropy (Basel, Switzerland)*, 24(2), p.270. Available at: http://dx.doi.org/10.3390/e24020270 [Accessed October 24, 2025].

O'Reilly-Shah, V.N., 2025a. Computational dynamic monism: Process metaphysics for the state space theory of consciousness. *SSRN Electron. J.* Available at: https://papers.ssrn.com/abstract=6042716 [Accessed January 22, 2026].

O'Reilly-Shah, V.N., 2025b. Delay coordinate embedding as neuronally implemented information processing: The state space theory of consciousness. *Journal of consciousness studies: controversies in science & the humanities*, 32(1), pp.127–159. Available at: https://www.ingentaconnect.com/content/imp/jcs/2025/00000032/f0020001/art00006.

O'Reilly-Shah, V.N., 2026. State Space Theory as a Unifying Framework for Consciousness. *Nonlinear Dynamics, Psychology, and Life Sciences*.

O'Reilly-Shah, V.N. & Selvitella, A.M., 2026. Smooth embeddings in contracting recurrent networks driven by regular dynamics: A synthesis for neural representation. *arXiv [q-bio.NC]*. Available at: http://dx.doi.org/10.48550/arXiv.2601.19019 [Accessed January 28, 2026].

Ostrow, M., Eisen, A. & Fiete, I., 2024. Delay embedding theory of neural sequence models. *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/2406.11993 [Accessed January 14, 2025].

Penrose, R., 1989. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, OUP Oxford. Available at: https://play.google.com/store/books/details?id=0mVEBJ34v9EC.

Polger, T.W. & Shapiro, L.A., 2016. *The Multiple Realization Book* L. A. Shapiro, ed., Oxford: Oxford University Press UK.

Putnam, H., 1967. Psychological Predicates. In W. H. Capitan & D. D. Merrill, eds. *Art, mind, and religion*. University of Pittsburgh Press, pp. 37–48.

Rabinovich, M.I. et al., 2006. Dynamical principles in neuroscience. *Reviews of modern physics*, 78(4), pp.1213–1265. Available at: http://link.aps.org/pdf/10.1103/RevModPhys.78.1213 [Accessed April 7, 2025].

Regehr, W.G., 2012. Short-term presynaptic plasticity. *Cold Spring Harbor perspectives in biology*, 4(7), p.a005702. Available at: http://dx.doi.org/10.1101/cshperspect.a005702 [Accessed September 23, 2025].

Schurger, A. et al., 2015. Cortical activity is more stable when sensory stimuli are consciously perceived. *Proceedings of the National Academy of Sciences of the United States of America*, 112(16), pp.E2083–92. Available at: http://dx.doi.org/10.1073/pnas.1418730112.

Seth, A.K., 2025. Conscious artificial intelligence and biological naturalism. *The Behavioral and Brain Sciences*, pp.1–42. Available at: http://dx.doi.org/10.1017/S0140525X25000032 [Accessed February 2, 2026].

Seth, A.K. & Bayne, T., 2022. Theories of consciousness. *Nature reviews. Neuroscience*, 23(7), pp.439–452. Available at: http://dx.doi.org/10.1038/s41583-022-00587-4.

Shapiro, L.A., 2000. Multiple Realizations. *Journal of Philosophy*, 97(12), pp.635–654.

Shea, N., 2018. *Representation in Cognitive Science*, London, England: Oxford University Press. Available at: http://dx.doi.org/10.1093/oso/9780198812883.001.0001 [Accessed January 27, 2026].

Song, C. et al., 2023. Minimal width for universal property of deep RNN. *Journal of machine learning research: JMLR*, 24(121), pp.121:1–121:41. Available at: http://jmlr.org/papers/v24/22-1191.html [Accessed October 24, 2025].

Uribarri, G. & Mindlin, G.B., 2022. Dynamical time series embeddings in recurrent neural networks. *Chaos, solitons, and fractals*, 154(111612), p.111612. Available at: http://dx.doi.org/10.1016/j.chaos.2021.111612 [Accessed November 20, 2025].

Usher, M., 2021. Refuting the unfolding-argument on the irrelevance of causal structure to consciousness. *Consciousness and cognition*, 95(103212), p.103212. Available at: http://dx.doi.org/10.1016/j.concog.2021.103212 [Accessed May 19, 2025].

Usher, M. et al., 2023. When philosophical nuance matters: safeguarding consciousness research from restrictive assumptions. *Frontiers in psychology*, 14, p.1306023. Available at: http://dx.doi.org/10.3389/fpsyg.2023.1306023.

Zenke, F. & Gerstner, W., 2017. Hebbian plasticity requires compensatory processes on multiple timescales. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 372(1715). Available at: http://dx.doi.org/10.1098/rstb.2016.0259 [Accessed August 7, 2025].

Figure 1

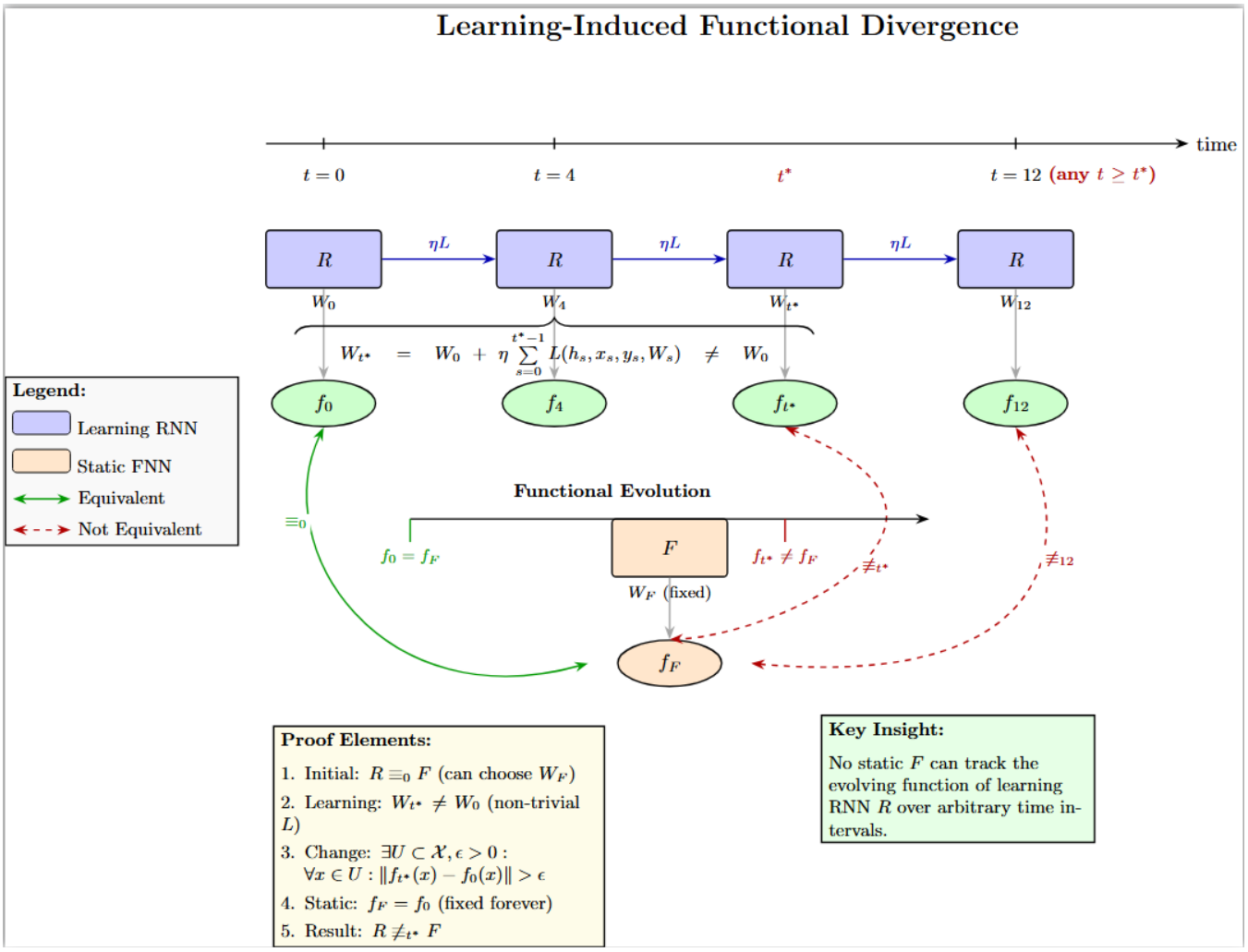

Learning-Induced Functional Divergence
time
$t = 0$
$t = 4$
t^*
$t = 12$ (any $t \geq t^*$)
R
ηL
W_0
W_4
W_{t^*}
W_{12}
$W_{t^*} = W_0 + \eta \sum_{s=0}^{t^*-1} L(h_s, x_s, y_s, W_s) \neq W_0$
f_0
f_4
f_{t^*}
f_{12}
Legend:
Learning RNN
Static FNN
Equivalent
Not Equivalent
Functional Evolution
$\equiv_0$
$f_0 = f_F$
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

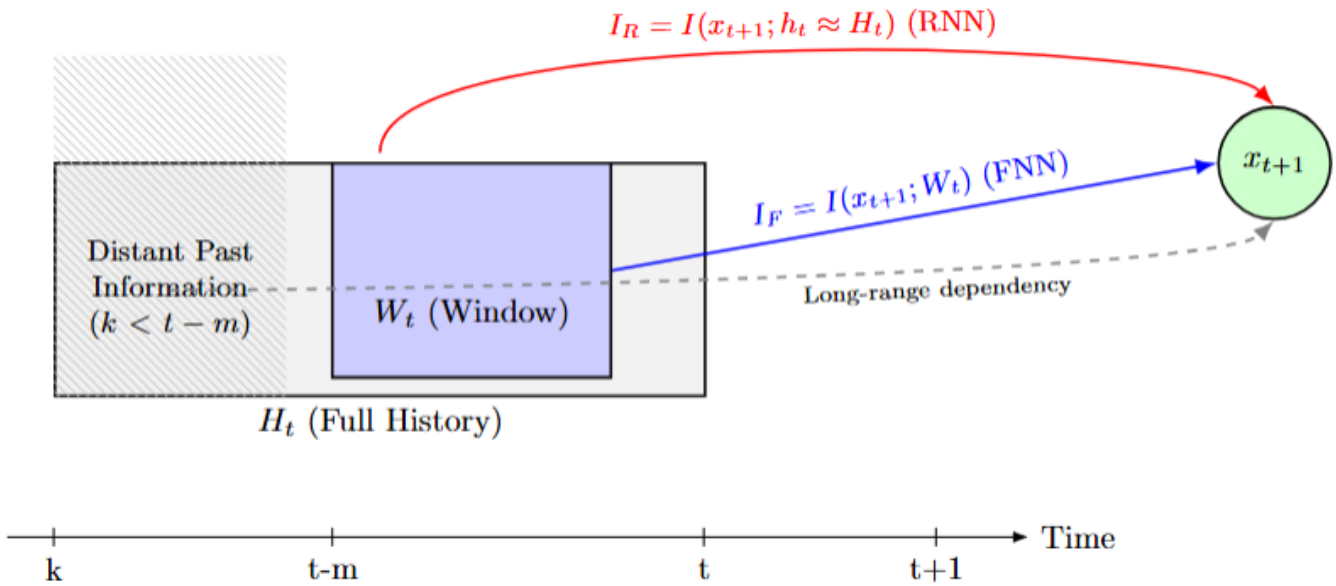

$I_R = I(x_{t+1}; h_t \approx H_t)$ (RNN)
x_{t+1}
$I_F = I(x_{t+1}; W_t)$ (FNN)
Distant Past
Information
$(k < t - m)$
W_t (Window)
Long-range dependency
H_t (Full History)
Time
k
t-m
t
t+1

Figure 3

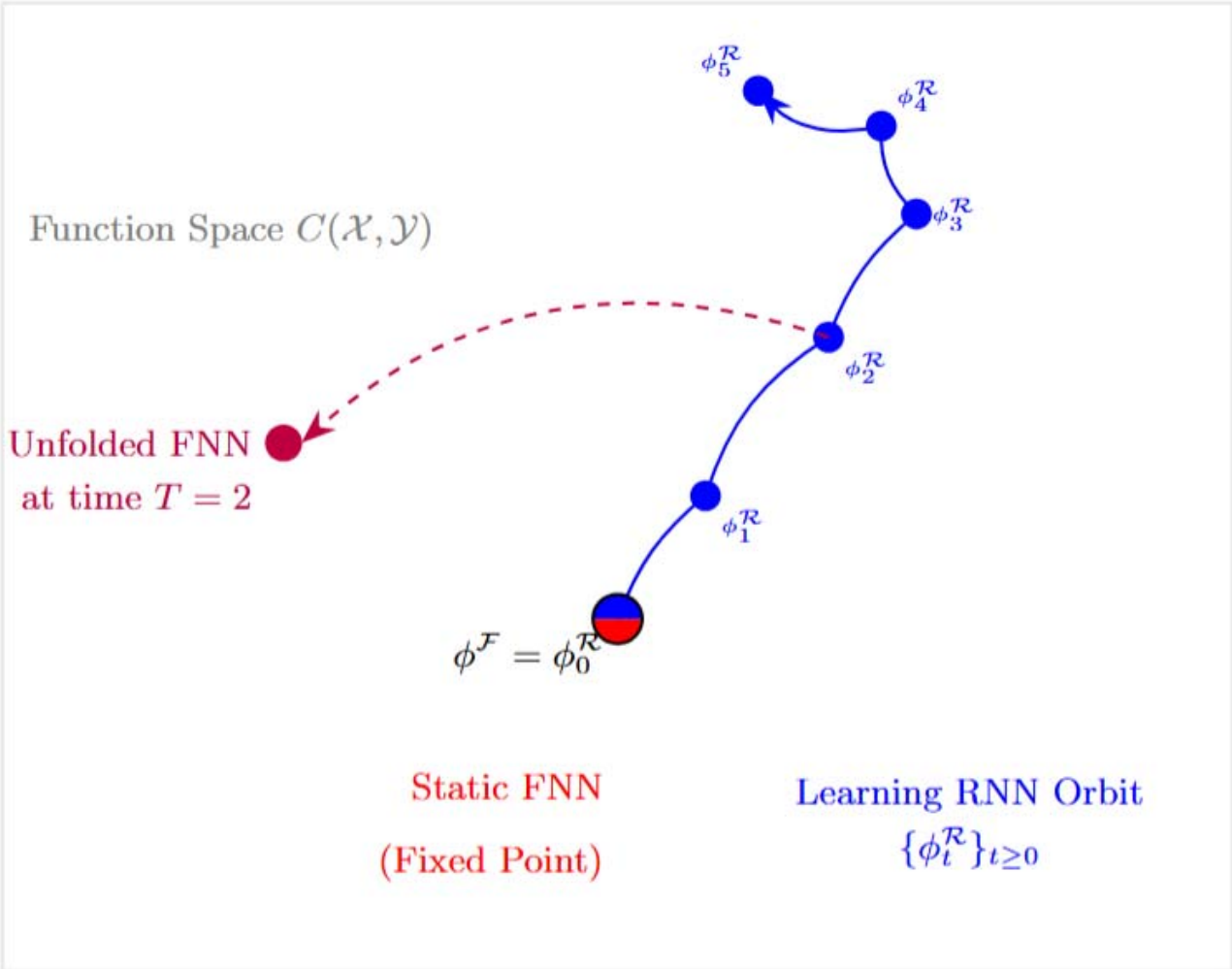

$\phi_5^{\mathcal{R}}$
$\phi_4^{\mathcal{R}}$
$\phi_3^{\mathcal{R}}$
Function Space $C(\mathcal{X}, \mathcal{Y})$
$\phi_2^{\mathcal{R}}$
Unfolded FNN
at time $T = 2$
$\phi_1^{\mathcal{R}}$
$\phi^{\mathcal{F}} = \phi_0^{\mathcal{R}}$
Static FNN
(Fixed Point)
Learning RNN Orbit
$\{\phi_t^{\mathcal{R}}\}_{t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

Figure 4

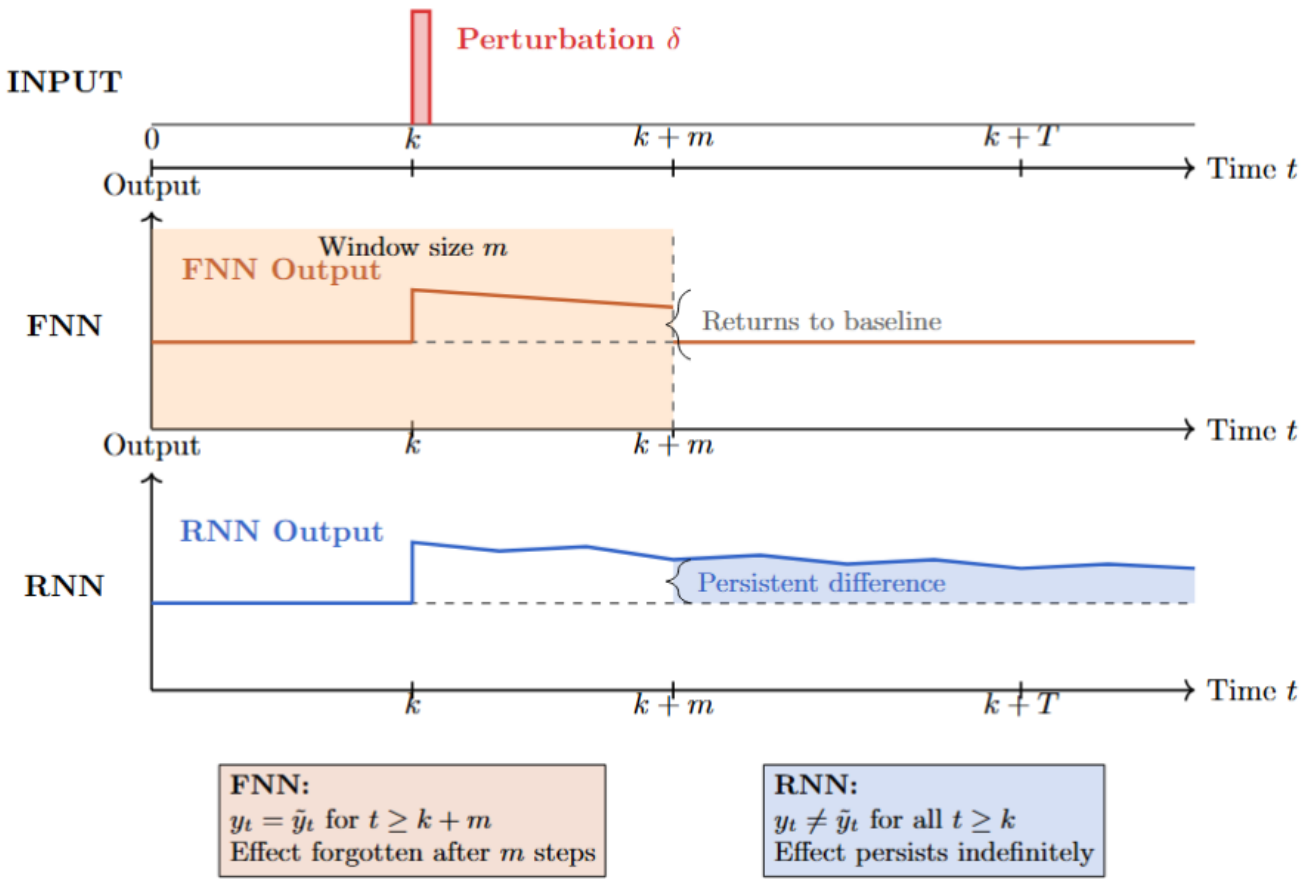
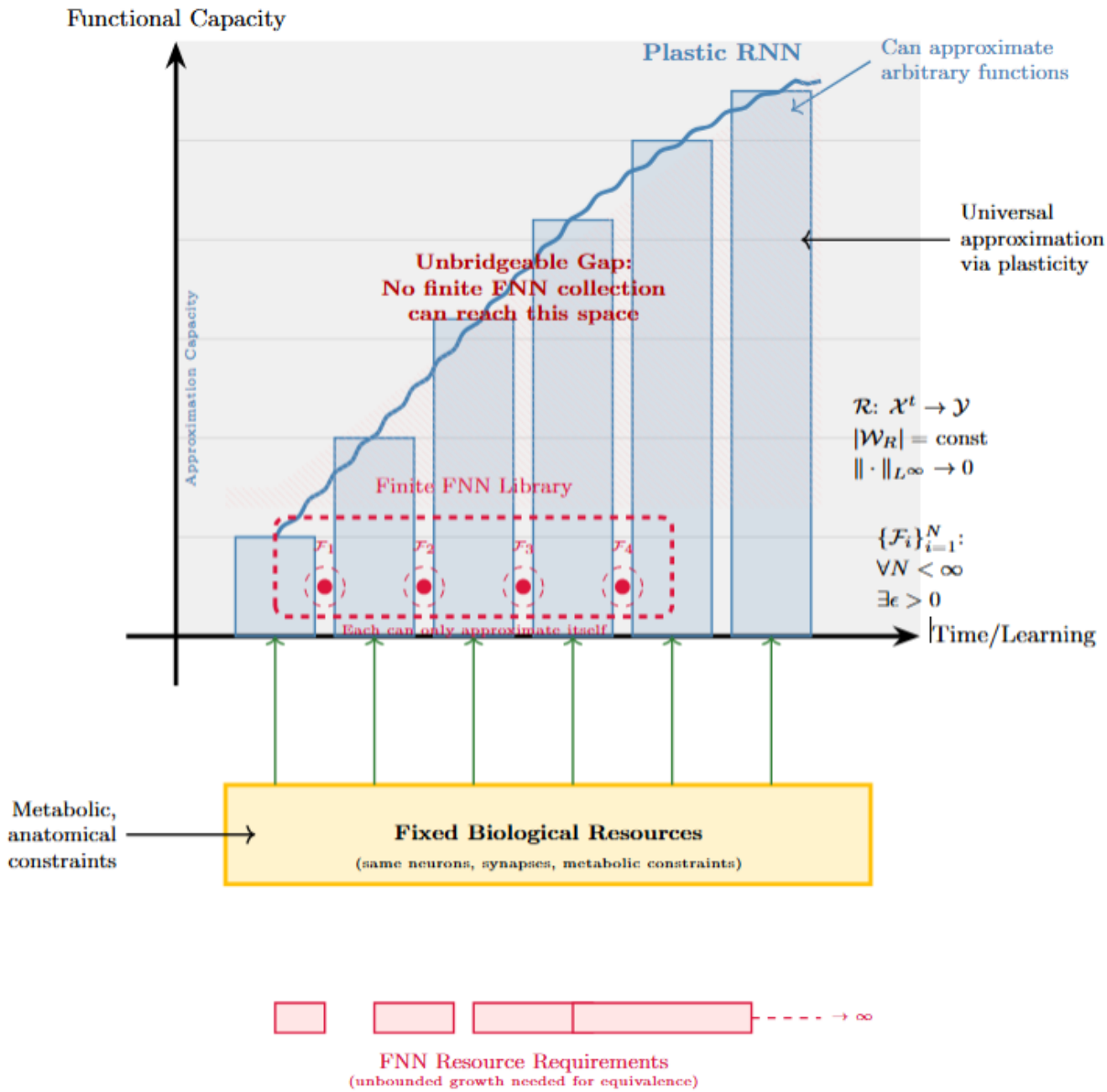

Perturbation δ
INPUT
0
k
$k+m$
$k+T$
Time t
Output
Window size m
FNN Output
FNN
Returns to baseline
Output
k
$k+m$
Time t
RNN Output
RNN
Persistent difference
k
$k+m$
$k+T$
Time t
FNN:
$y_t = \tilde{y}_t$ for $t \geq k+m$
Effect forgotten after m steps
RNN:
$y_t \neq \tilde{y}_t$ for all $t \geq k$
Effect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