Review

Learning Neural Evolution Operators: From Decoding to Identifiable Causal State-Space Models

Armin Hakkak Moghadam Torbati *

Posted Date: 5 March 2026

doi: 10.20944/preprints202603.0480.v1

Keywords: neural decoding; neural population dynamics; latent state-space models; evolution operators; low-dimensional manifolds; causal neural dynamics; perturbation

 doi:10.20944/preprints202603.0480.v1

Review

Learning Neural Evolution Operators: From Decoding to Identifiable Causal State-Space Models

Armin Hakkak Moghadam Torbati [1,2]

[1] Laboratory of Functional Anatomy, Faculty of Human Motor Sciences, Universite Libre de Bruxelles (ULB), B-1050 Brussels, Belgium; armin.hakkak.moghadamtorbati@ulb.be

[2] Laboratoire de Neuroanatomie et Neuroimagerie Translationnelles, UNI—ULB Neuroscience Institute, ULB, B-1050 Brussels, Belgium

Abstract

Neural decoding has demonstrated that population activity contains behaviorally relevant information, yet predictive accuracy alone does not constitute mechanistic explanation. Decoding models establish statistical mappings between neural responses and task variables but leave the underlying computational processes underdetermined. We argue that neural computation is more appropriately framed within a dynamical state-space perspective, in which population activity reflects the evolution of latent states governed by structured transition operators. Across empirical and theoretical work, neural trajectories increasingly appear as low-dimensional, nonlinear flows shaped by recurrent circuit structure and contextual inputs. This shift reframes the central scientific objective: not merely extracting representations, but learning the evolution operator that governs state transitions. However, even accurate reconstruction of latent dynamics does not guarantee mechanistic validity. Observational data typically constrain only an equivalence class of admissible operators, rendering the inferred dynamics structurally non-identifiable. We therefore propose that causal neural dynamics must be defined through perturbation and experimental design. By introducing directional constraints on state transitions, targeted interventions collapse equivalence classes and enable identification of operators that remain valid under manipulation. In this framework, evolution operators are treated as falsifiable hypotheses whose mechanistic status depends on predictive stability under perturbation. This perspective recasts neural modeling as the search for perturbation-validated dynamical laws governing population activity, moving the field from decoding-based description toward causal dynamical explanation.

Keywords: neural decoding; neural population dynamics; latent state-space models; evolution operators; low-dimensional manifolds; causal neural dynamics; perturbation

1. The Decoding Paradigm Has Reached Its Limits

Neural decoding has become a central tool for linking brain activity to stimuli, actions, and cognitive variables, and its empirical successes are undeniable. However, even recent high-profile syntheses emphasize that decoding, particularly when framed as “reading out” information from neural activity, does not necessarily link to neural mechanisms and therefore cannot be equated with an explanation of brain computation [1]. Decoding can demonstrate that information is present in a neural population in a format that a classifier can exploit, yet still fail to reveal the process by which circuits generate, transform, and use that information. This limitation is not merely semantic. As Kriegeskorte and Douglas argue, decoding primarily reveals the “products” rather than the "process" of neural computation, and statistically significant predictive performance places only weak constraints on computational theory [2]. High decoding accuracy does not imply that the fitted model captures the brain’s transformation. It may instead exploit correlated structure or redundancies in the data. Interpreting decoder weights or feature relevance as mechanistic evidence, therefore, risks conflating predictive sufficiency with explanatory adequacy [2].

 doi:10.20944/preprints202603.0480.v1

Importantly, this interpretational gap can be formalized within a causal framework. Weichwald and colleagues demonstrate that the meaning of feature relevance depends critically on the direction of the data-generating process. In common decoding settings, such as predicting stimuli from neural responses, the inference is anti-causal. In such cases, variables identified as "relevant" by a decoder need not correspond to true causes or effects in the underlying system, and genuine causal variables may remain undetected [3]. Ambiguity in decoding is thus structurally embedded in the inference direction rather than being a mere statistical artifact. One might argue that Bayesian formulations restore mechanistic grounding. Yet Lange et al. clarify that Bayesian "decoding" typically infers stimuli from neural responses using stimulus–response statistics, whereas Bayesian "encoding" concerns inference within an internal generative model [4]. Successful Bayesian decoding therefore does not, by itself, identify the brain's internal model or the dynamical circuit mechanisms that implement inference. The explanatory gap persists even under normative probabilistic framing [4].

These concerns extend beyond conceptual critique and can be demonstrated empirically. Jonas and Kording demonstrate that when standard neuroscientific analysis tools are applied to a classical microprocessor, a system whose computational architecture is fully known, they recover statistically compelling patterns yet fail to reconstruct the processor's hierarchical computational organization [5]. Even with complete access to activity and connectivity, data abundance and sophisticated analysis do not guarantee a mechanistic understanding if the analytical framework is misaligned with the system's computational structure.

A related fragility appears in applied brain modeling. Statistical models trained under independent and identically distributed assumptions can achieve impressive performance in controlled laboratory settings yet fail under distribution shifts, demographic variation, or altered recording conditions. A causal analysis of brainwave modeling makes explicit how anti-causal prediction, hidden confounders, and unmodeled generative factors undermine generalization and interpretability [6]. Modeling $P(Y \mid X)$ alone does not ensure robustness or causal insight even with high predictive accuracy.

All these arguments suggest that the limitations of decoding are not primarily technical but conceptual. Improving classifier architectures or increasing dataset size is unlikely to resolve the explanatory gap. What is missing is not predictive power, but an explicit account of how neural population states evolve according to structured dynamical laws, interact across time, and implement computation through state transitions. This motivates a shift from static representational mappings toward dynamical, state-space formulations in which computation is expressed through the causal evolution of latent neural states (see Figure 1) [7,8]. In what follows, we formalize this shift through four foundational principles that together redefine the scientific objective from decoding accuracy to perturbation-validated dynamical law.

 doi:10.20944/preprints202603.0480.v1

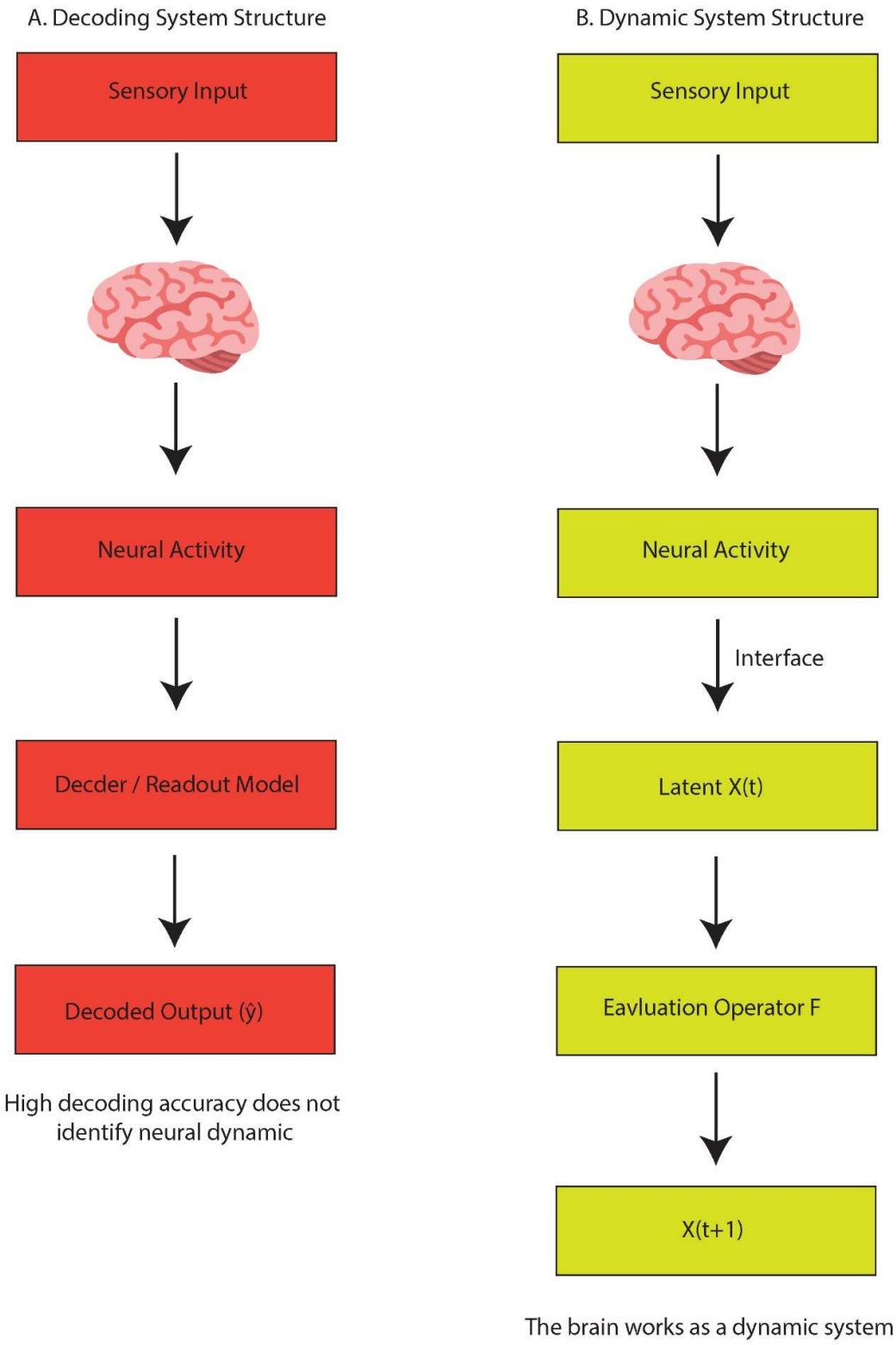

Figure 1. From decoding to dynamical state-space computation. (**A**) In the classical decoding framework, sensory input gives rise to neural activity that is mapped through a readout model to produce a decoded variable (ŷ). High predictive accuracy in this setting demonstrates statistical sufficiency but does not identify the underlying neural transformation or computational mechanism. (**B**) In a dynamical systems framework, neural activity is interpreted as observations of an evolving latent state $x(t)$. Computation is implemented through

 doi:10.20944/preprints202603.0480.v1

the structured evolution operator F, which governs transitions $x(t) \rightarrow x(t+1)$. In this view, the explanatory target shifts from decoding outputs to identifying the causal evolution law underlying population state trajectories. **Principle 1:** Neural computation is expressed through latent state evolution, not static readout decoding.

2. Brain Activity as Nonlinear State-Space Dynamics

2.1. From Representation to Dynamics

The classical representational framework treats neural activity as encoding task-relevant variables: stimuli, movements, decisions, or internal estimates. In this view, neural firing rates are functions of behavioral parameters. A different interpretation emerges when neural populations are considered not as static encoders, but as evolving dynamical systems.

Churchland et al. [9] demonstrated that motor cortical population activity during reaching exhibits transient rotational dynamics revealed through jPCA. These quasi oscillatory trajectories were present even though the behavior itself was not rhythmic. The population state r(t), rather than representing kinematic variables, evolved according to a dynamical rule of the form:

$$dr/dt = f(r(t)) + u(t) \quad (1)$$

The preparatory state set the initial condition of the trajectory, and subsequent evolution unfolded largely autonomously. What appeared as heterogeneous, multiphasic single-neuron responses became coherent when viewed as structured rotations in a low-dimensional subspace. Representation, in this framing, is incidental to dynamics. Marques et al. [10] extended this perspective by showing that zebrafish foraging behavior is governed by persistent internal states forming a stochastically activated nonlinear dynamical system. Whole-brain recordings revealed metastable states whose transitions were not reducible to stimulus encoding. Internal state functioned as a latent dynamical variable shaping perception and action over extended time scales. These findings suggest that neural computation may be more accurately described as structured evolution in state space rather than as a static mapping from inputs to outputs.

If neural computation unfolds in time, what kind of dynamics does it follow? Linear dynamical systems offer tractability but are fundamentally limited in expressivity. Empirical observations increasingly point to an intrinsically nonlinear structure.

Duncker and Sahani [11] emphasized that population trajectories often reside on low-dimensional manifolds embedded in high-dimensional neural space. However, identifying such manifolds is only a first step, while the central challenge lies in estimating the dynamical law governing transitions on them. Computational processes such as integration, attractor formation, and flexible trajectory generation require nonlinear flow fields.

Runfola et al. [12] proposed a mechanistic account for the emergence of low-dimensional structure based on time-scale separation. Fast oscillatory fluctuations average out, leading high-dimensional dynamics to collapse onto slow invariant manifolds. In this account, low-dimensional structure arises from nonlinear self-organization rather than from simple linear projection.

Predictive learning frameworks further reinforce this view. Recanatesi et al. [13] showed that recurrent networks trained to predict future observations spontaneously develop nonlinear manifold representations of latent variables. Greco et al. [14] demonstrated in humans that predictive learning reshapes representational geometry to align with environmental statistical structure, linking prediction error encoding to geometric reorganization of neural representations.

Nonlinear dynamics thus appear not merely as a modeling preference, but as a biological necessity arising from recurrent computation, time-scale separation, and predictive adaptation.

 doi:10.20944/preprints202603.0480.v1

2.2. Latent Manifolds and State-Space Formalization

Beiran et al. [15] investigated low-rank recurrent neural network (RNN) trained on flexible timing tasks and showed that constraining connectivity rank restricts dynamics to a low-dimensional nonlinear manifold. Tonic contextual inputs modulate trajectory speed along this manifold without altering its geometry, enabling generalization beyond training conditions. Neural data from the frontal cortex confirmed signatures of this geometry-preserving control.

Importantly, the relationship between observed activity and causal contribution is not straightforward. Fakhar et al. [16] demonstrated that recorded activity patterns may not reliably indicate causal influence on behavior. Through multi-site perturbation and Shapley-value analysis, they showed dissociations between decodability and causal contribution. This finding underscores the need to distinguish between descriptive or observed state trajectories and genuine causal structure within the dynamical system.

Schneider et al. [17] introduced CEBRA, a nonlinear contrastive embedding framework that jointly leverages neural and behavioral data to produce consistent and identifiable latent spaces. This methodological advance improves reliability and interpretability of neural state-space estimation across sessions and subjects.

All these findings collectively show that neural activity evolves on structured low-dimensional manifolds shaped by nonlinear recurrent dynamics and modulated by contextual inputs, while also requiring causal validation. This geometric perspective is illustrated schematically in Figure 2, where population trajectories are shown as structured flows evolving on a low-dimensional manifold embedded within high-dimensional neural space.

These empirical and theoretical developments converge on a common formal description of neural computation within a latent state-space framework. The cumulative evidence motivates a formal shift from decoding variables to modeling evolution operators. Neural population activity can be expressed as:

$$x(t+1) = F(x(t), u(t)) + \varepsilon(t) \quad (2)$$

$$y(t) = G(x(t)) + \eta(t)$$

Here, x(t) is a latent neural state evolving on a nonlinear manifold, F is the evolution operator, u(t) contextual input, and G the observation mapping. The central scientific objective is to infer F, the structured dynamical law governing state transitions.

To sum up, empirical demonstrations of rotational dynamics [9], metastable internal states [10], manifold-based computational flows [11], geometry-preserving parametric control [15], mechanistic emergence of invariant manifolds [12], predictive reshaping of representational geometry [14], and identifiable latent embeddings [17] collectively argue that brain computation is best understood as nonlinear state-space evolution.

At the same time, causal dissociations between activity and contribution [16] warn that state-space models must be validated through perturbation and intervention.

The representational question: what is encoded?. Is therefore subsumed by a deeper dynamical question: What invariant manifolds, flow fields, and control inputs govern the causal evolution of neural population states?

This shift prepares the ground for the next step: learning evolution operators rather than merely extracting representations.

 doi:10.20944/preprints202603.0480.v1

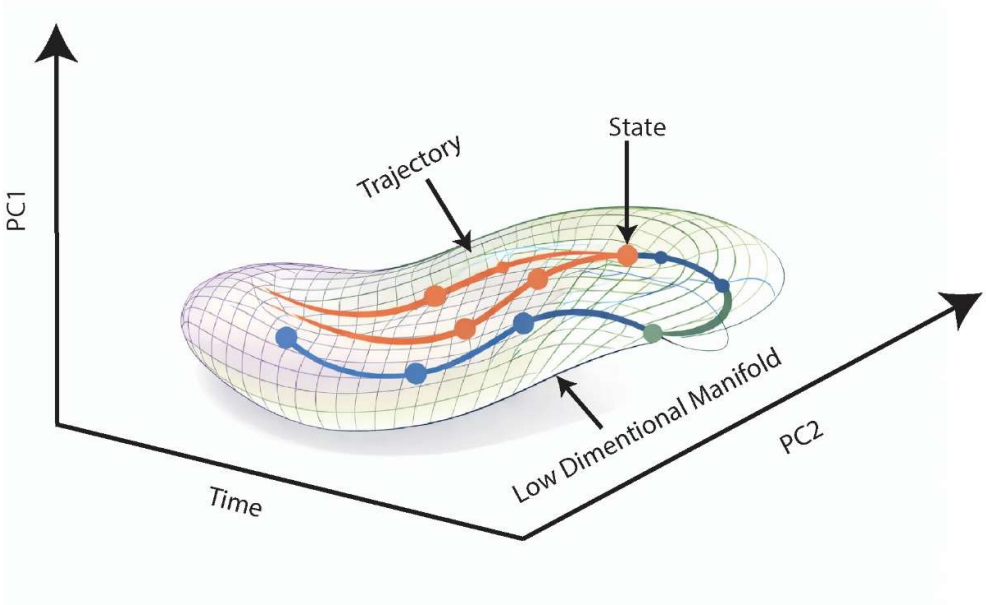

Figure 2. Neural dynamics as structured flow on low-dimensional manifolds. Population activity trajectories (colored paths) unfold over time on a low-dimensional nonlinear manifold embedded in high-dimensional neural space. Each point corresponds to a latent neural state, and arrows indicate directed evolution governed by a structured flow field. Task context or control inputs modulate the geometry and velocity of trajectories without necessarily altering manifold topology. This perspective reframes neural computation as constrained evolution within a dynamical landscape rather than static representation of task variables. **Principle2:** Neural population activity evolves on low-dimensional manifolds governed by structured, task-dependent flow fields.

3. Learning Evolution Operators, Not Just Representations

The key challenge in understanding brain activity goes beyond just representation learning. It involves recovering the evolution law that governs how latent neural states evolve over time. As Durstewitz et al. [18] highlight, the goal of reconstructing neural dynamics is not simply dimensionality reduction or pattern matching but creating a surrogate dynamical system that captures the flow structure, attractors, and topological features of the system's evolution. In the state-space framework, a system is not defined solely by its latent states but by the transition operator, which determines how the system evolves. Paninski et al. [19] formalize this with a model that includes both transition and observation dynamics. To make meaningful inferences, the focus must shift from learning just a latent representation to learning the evolution operator itself, which is central to understanding neural computation. This shift is critical because without recovering the dynamics that drive neural state transitions, any learned representation risks being just an incomplete snapshot, not an explanation of the underlying processes.

One of the most challenging aspects of learning hidden dynamics from neural data lies in the need for appropriate regularization and inference methods. While it is clear that directly modeling the evolution operator offers a more accurate and mechanistic understanding of neural dynamics, this task is fraught with difficulties such as overfitting and sensitivity to hyperparameters. The Latent Factor Analysis via Dynamical Systems (LFADS) model introduced by Sussillo et al. [20] demonstrates a powerful method for inferring single-trial latent dynamics from neural spike trains by using a RNN as the evolution operator. However, despite its successes in modeling the latent states of neural populations, LFADS is highly sensitive to the quality of hyperparameter tuning, which can lead to issues with overfitting and model instability if not properly regularized . Keshtkaran et al. [21] further emphasize these challenges, revealing that the performance of LFADS and other similar models is highly dependent on careful tuning and regularization of the underlying

 doi:10.20944/preprints202603.0480.v1

neural networks, particularly in the context of small datasets and when behavioral task information is sparse. Their work highlights the necessity of robust training procedures to ensure that the inferred dynamics are reliable and generalizable across conditions. Without proper regularization, even the most sophisticated models risk learning trivial or non-mechanistic solutions that fail to capture the true nature of neural computation. Thus, as the field progresses, it becomes clear that improving the robustness of these models through better regularization and tuning strategies is essential for learning trustworthy evolution operators.

While LFADS and related models learn a single nonlinear transition function, empirical neural data often violate the assumption of a globally homogeneous flow. Neural dynamics can change abruptly across behavioral epochs or physiological states, making stationary operators inadequate. Switching state-space models explicitly address this limitation by introducing regime-dependent transition dynamics, allowing different linear systems to govern evolution at different times [21]. However, discrete switching introduces its own artifacts, including boundary discontinuities and sensitivity to regime assignments. The Gaussian Process Switching Linear Dynamical System resolves this by replacing hard switching with smoothly interpolated locally linear regimes, combining interpretability with continuous uncertainty-aware inference over the flow field [22]. Yet even regime-based models assume that at any time point the system belongs to a single dynamical mode. Decomposed Linear Dynamical Systems extend this idea further by representing evolution as a sparse combination of elementary linear operators, effectively learning a dictionary of dynamical primitives whose time-varying coefficients define the flow [23]. This compositional view avoids combinatorial explosion in switching models and allows overlapping or concurrent subprocesses to be expressed without sacrificing structure. These developments suggest that learning neural evolution requires structured operators, either regime-conditioned or sparsely composed, rather than unconstrained nonlinear functions, if interpretability and mechanistic insight are to be preserved.

The evolution operator must not only be nonlinear but also structured to preserve interpretability. Simply learning a single, continuous-time evolution function without consideration for temporal regimes or state-specific dynamics is insufficient. Abbaspourazad et al. introduced the DFINE model, which enables flexible inference of neural dynamics by learning both a nonlinear latent manifold and a linear dynamic system on top of it, enabling a more accurate description of complex neural processes while preserving interpretability. This combination of manifold and linear dynamics allows the model to adapt to different neural contexts and behaviors, accounting for nonlinearity without losing the power of tractable, interpretable dynamics [24]. Moreover, the work of ElGazzar and van Gerven emphasizes the utility of Stochastic Differential Equation (SDE) in continuous time, which offer a more nuanced understanding of latent dynamics by incorporating both drift and diffusion terms into the evolution law. This hybrid approach of integrating neural networks with mechanistic SDE models allows for the explicit handling of uncertainty in the learned dynamics, crucial for robust and interpretable neural modeling [25]. However, as Sedler et al. [26] point out, the architecture of the evolution operator is pivotal for interpretability. They argue that RNN, commonly used for dynamical modeling, fail to provide meaningful insights into the underlying neural processes due to their limited expressiveness when it comes to capturing complex flows. In contrast, Neural ODEs offer a superior alternative by allowing the dynamics to be modeled more flexibly, facilitating better generalization to real-world data [24]. Moreover, Gosztolai et al. advocate for Geometric Deep Learning to enhance the interpretability of neural dynamics. Their MARBLE model decomposes neural dynamics into local flow fields, capturing dynamical changes across multiple conditions and allowing for a structured, interpretable view of the evolution operator. By using this geometric approach, MARBLE offers a new way of linking the latent dynamics with task-specific behavior, overcoming the limitations of traditional methods [27]. These approaches highlight the importance of combining flexible, non-linear dynamics with structured representations, emphasizing the need for evolution operators that are not only powerful but also interpretable.

Across discrete, switching, decomposed, manifold-based, and continuous-time formulations, the common objective is to recover the structured dynamical law that generates neural population

 doi:10.20944/preprints202603.0480.v1

trajectories. However, the central problem is no longer whether neural representations can be decoded, but whether the evolution operator governing latent state transitions can be reliably inferred as a structurally valid mechanism. Reconstruction accuracy alone does not guarantee that the learned operator reflects the true underlying mechanism. Highly expressive models may accurately reproduce observed trajectories while differing significantly in their internal dynamics. This raises a deeper question: under what conditions is the inferred evolution law identifiable, and when does it reflect genuine structure rather than a statistical approximation? These questions move beyond mere reconstruction and into the realm of causal and structural validity, which we address in the next section.

4. Identifiability and Causality: The Missing Pieces

Learning an evolution operator from neural data, even when reconstruction is excellent, does not mean achieving a unique and interpretable mechanism. The issue is not just data scarcity or model weakness. The problem is that observational data (passive recording) often lead to more than one explanation that is consistent with the data. In other words, we often arrive at a model that fits the data, but not a causal operator. This gap between "statistical fit" and "mechanistic explanation" was highlighted in previous sections as the shift from decoding to dynamics, and here it must be framed more precisely in terms of identifiability and causality [3,5].

A specific source of this ambiguity is "observational equivalence": multiple causal/dynamical structures can generate similar temporal distributions, meaning that observational data alone usually do not lead to a unique solution. Even in the causal discovery literature of time series, it is emphasized that recovering the causal structure depends heavily on strong assumptions, and the output is often defined as an equivalence class (not a unique graph/mechanism) [28].

In the context of neural data, causal tests based on information criteria (such as directed information) show that the results are highly sensitive to conditioning, assumptions about hidden variables, and how confounding is controlled. In other words, the statistical tools alone do not produce "causal certainty" [29]. Therefore, without additional constraints (often design or intervention-related), observing the trajectories only identifies a family of operators consistent with the data, not a unique causal operator.

More importantly, non-identifiability does not just create "uncertainty", it can lead to systematic error. Sachdeva et al. directly challenge this common intuition. They show that in models where tuning, coupling, and latent sources are jointly modeled, non-identifiability can cause persistent bias in parameter estimates instead of increasing variance. This means we can end up with answers that have low variance across initializations, but are systematically deviating from the ground truth [30]. This shows even if a latent-dynamics model achieves excellent reconstruction, it may consistently report the "wrong" operator, because the observational data may not have been sufficient to disentangle the contributions of coupling, tuning, and latent drive.

Another source of ambiguity comes from the parametrization of the observation map or readout. That means how we map the latent state to the neural activity space. The Versteeg et al. demonstrates that if the readout is non-injective, changes in the latent state do not necessarily appear in the observed data. As a result, the model can "invent" artificial latent dynamics during training to improve reconstruction that does not correspond to the actual mechanistic process [31]. They argue that this issue makes reconstruction an unreliable indicator for interpretability of dynamics, and the solution is to enforce (approximately) injectivity so that all latent variance must manifest in the neural space, and the model does not get rewarded for inventing ineffective states. Therefore, operator learning without proper constraints on the observation/architecture becomes structurally non-identifiable or misleading. Moreover, identifiability should be seen as a property of the joint "model + data + data collection design," not something that will automatically resolve by increasing model capacity alone. Wang and Hao provide a computational framework for practical identifiability, specifically showing that practical identifiability is tied to the invertibility (or full-rank condition) of the Fisher Information Matrix, and can be improved with optimal measurement point designs and

 doi:10.20944/preprints202603.0480.v1

data-collection algorithms [32]. In more classical latent-state models, Liu and Culpepper show that identifiability in families of HMMs/RHMMs depends on conditions and design, explicitly linking their results to "study/intervention design". That is, the type of data and structure you collect determines whether parameters/transitions are identifiable [33].

Overall, without design-centered constraints (which should include task design, perturbation, diverse conditions, or complementary measurements), the latent dynamical operator is often not uniquely identifiable or interpretable mechanistically. Thus, the logical next step is to examine how perturbation and intervention can break the equivalence class and validate the learned operator as a stable and manipulable mechanism, rather than just a good statistical fit. This inferential structure is summarized schematically in Figure 3. We will discuss that in the next section.

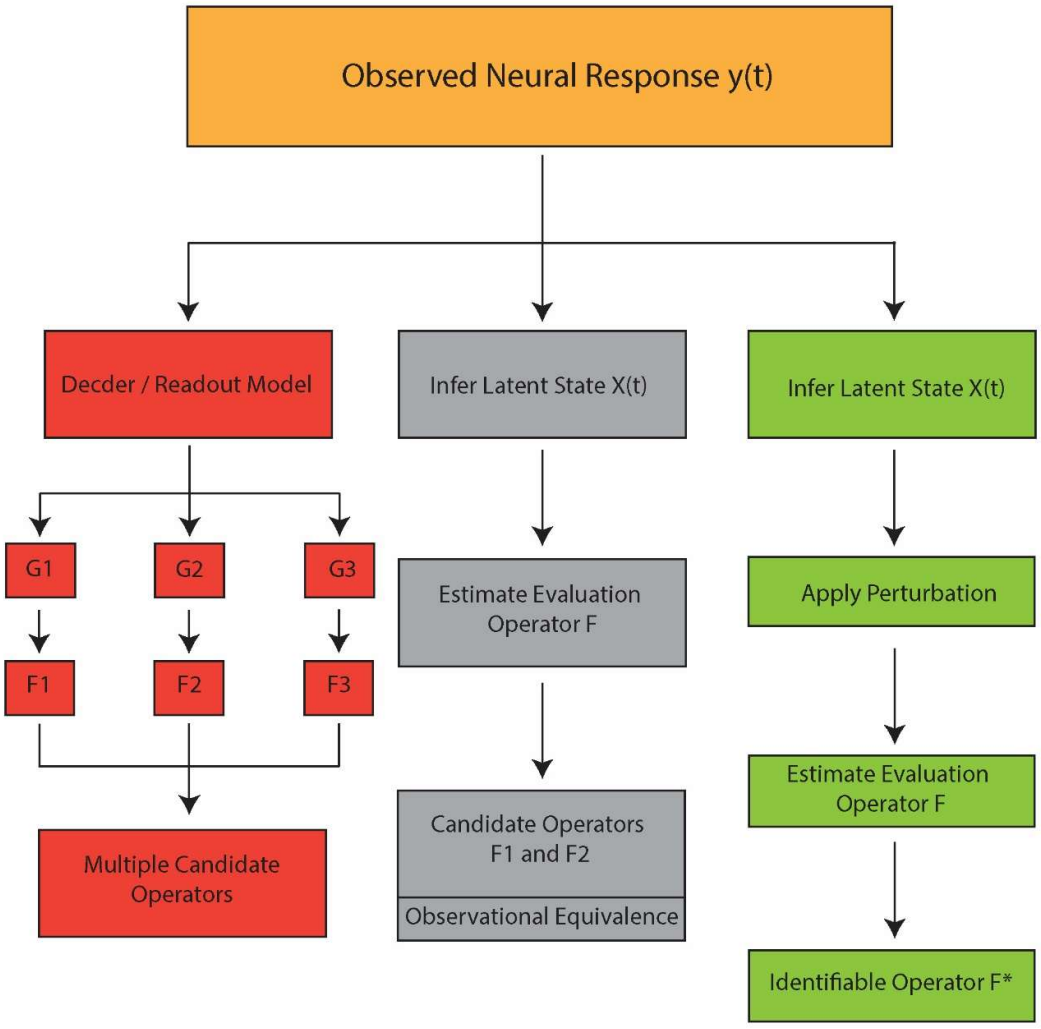

Figure 3. Starting from observed neural responses $y(t)$, three inferential pathways yield distinct epistemic outcomes. **Decoding pathway (left).** Readout models G_1, G_2, G_3 map activity to task variables, each compatible with different implied dynamical operators. High predictive accuracy constrains correlations but leaves the underlying transition structure underdetermined, resulting in multiple candidate operators. **Passive dynamical inference (center).** A latent state $x(t)$ and evolution operator F are inferred from observational trajectories. However, distinct operators (e.g., F_1, F_2) can generate statistically indistinguishable temporal distributions. Reconstruction fidelity therefore identifies an equivalence class of admissible operators rather than a unique

 doi:10.20944/preprints202603.0480.v1

causal law. **Perturbation-augmented inference (right).** Introducing structured perturbations or controlled experimental manipulations imposes additional constraints on state transitions. Operators inconsistent with perturbed responses are eliminated, collapsing the equivalence class and yielding an identifiable operator F^*. **Principle 3:** Observational neural recordings constrain only an equivalence class of evolution operators, not a unique causal mechanism. **Principle 4:** Perturbation and experimental design collapse this equivalence class, enabling identifiability of the causal evolution operator.

5. Toward Perturbation-Validated Neural Dynamics

A consistent theme across recent dynamical work is that the latent geometry inferred from population activity becomes mechanistically meaningful only when it constrains how the system responds to targeted perturbations. In spiking network models explicitly constructed to exhibit low-dimensional activity, Wärnberg and Kumar show that the "intrinsic manifold" can be a direct consequence of circuit connectivity, and that forcing activity outside this manifold effectively requires substantially larger changes in synaptic weights than producing patterns within it, providing a concrete mechanistic account for why within-manifold behaviors are easier to learn and express than outside-manifold behaviors [34]. Complementing this connectivity-based argument, Vinograd et al. provide causal evidence for a continuous (line) attractor encoding an internal affective state by using model-guided perturbations targeted to neurons contributing to the attractor. On-manifold perturbations integrate stimulation pulses and drive persistent displacement along the attractor, whereas transient off-manifold perturbations rapidly relax back into the attractor [35]. Finally, O'Shea et al. explicitly frame dynamical hypotheses as being distinguishable by their perturbation predictions, and use optogenetic and electrical microstimulation perturbations during reaching to constrain the dynamical class consistent with motor-cortical population activity, illustrating how perturbation responses can rule out otherwise observationally plausible dynamical explanations [36]. Therefore, these works motivate perturbation not as an "add-on" to descriptive state-space analysis, but as a primary route to exposing which latent dimensions are dynamically potent, which are dynamically silent, and which geometrical structures (e.g., attractors/manifolds) are implemented by circuit mechanisms rather than fitting artifacts.

If perturbation reveals latent structure, closed-loop perturbation operationalizes it. Stimulation is no longer delivered at pre-set times to pre-set cells but is guided by ongoing population activity and behavioral context. Zhang et al. describe an all-optical closed-loop system that integrates real-time calcium imaging analysis with holographic optogenetic stimulation, enabling automated recruitment of neurons into stimulation ensembles, rapid functional mapping followed by immediate closed-loop photoinhibition, and activity-guided targeting of ensembles based on ongoing population patterns during behavior [37]. This kind of closed-loop "read–write" pipeline creates the experimental substrate for testing dynamical models in the regime where trial-to-trial variability and state-dependence matter, rather than averaging them away. In parallel, Marin Vargas et al. present a distinct closed-loop instantiation in which a learned policy operating at the muscle level is aligned to primate M1/S1 activity, and its latent trajectory representation can be decoded from M1 to achieve direct neural control of the controller in real time, yielding coherent grasp trajectories with improved robustness relative to joint-angle decoding [38]. Finally, Sourmpis et al. make the validation logic explicit for data-constrained circuit models. They evaluate reconstructed RNN by testing whether they predict responses to unseen optogenetic interventions, and show that biologically informed inductive biases can substantially improve generalization to perturbed trials even when a generic RNN matches unperturbed activity well [39]. Therefore, it is defined a pragmatic progression from perturbation-as-probe to perturbation-as-control: real-time targeting enables state-conditional interventions, and model-based perturbation tests provide an objective metric for whether an inferred dynamical model remains valid under the very manipulations required to establish a mechanism.

The central methodological claim of this section is that learned evolution operators should be treated as falsifiable hypotheses. They must not merely reconstruct observed trajectories, but

 doi:10.20944/preprints202603.0480.v1

correctly predict how trajectories change under intervention. O'Shea et al. articulate this logic directly by contrasting competing dynamical hypotheses that are all compatible with low-dimensional observational trajectories, yet make distinct predictions about how perturbations engage circuit dynamics. Their perturbation experiments are then used to constrain the hypothesis space by demonstrating qualitative mismatches with specific classes of dynamics [36]. Sourmpis et al. sharpen the same point in the context of modern reconstruction. They propose a perturbation test for reconstructed networks and show an instructive contradiction. Their models can exhibit higher similarity on held-out unperturbed trials yet generalize poorly on perturbed trials, implying that conventional goodness-of-fit on passive recordings is not sufficient to establish mechanistic adequacy of the learned operator [39]. Vinograd et al. provide an additional, stringent form of falsification for latent geometric claims by using model-guided on- versus off-manifold perturbations. They demonstrate differential integration versus relaxation responses that would not be expected if the inferred attractor structure were merely epiphenomenal, thereby turning an inferred state-space object into a perturbation-validated dynamical mechanism [35]. Finally, the closed-loop systems described by Zhang et al. [37] and Marin Vargas et al. [38] expand what is falsifiable in practice by enabling interventions contingent on ongoing activity and by embedding neural activity into real-time control loops, conditions under which incorrect operators are expected to fail in identifiable, behaviorally relevant ways. In this framing, perturbation does not merely "support" a dynamical model, it defines the criterion by which an inferred evolution operator earns mechanistic status by surviving targeted attempts to break its predictions.

6. Conclusions

Neural decoding has demonstrated that population activity contains behaviorally relevant information, yet predictive sufficiency alone does not constitute mechanistic explanation. We argued that neural computation is more accurately framed as structured evolution in a latent state space, governed by an evolution operator whose form, not merely its outputs, defines the computational mechanism. Learning such operators moves the field beyond representation toward dynamical law, but observational data constrain only an equivalence class of admissible operators. Reconstruction accuracy, even when high, does not guarantee causal validity. Perturbation and experimental design transform this inferential landscape by introducing directional constraints on state transitions, intervention breaks dynamical symmetries and collapses the equivalence class. In this framework, a neural evolution operator earns mechanistic status only insofar as it survives targeted manipulation and correctly predicts counterfactual trajectories. Causal neural dynamics are therefore not defined by reconstruction fidelity, but by invariance and falsifiability under perturbation. This shift from decoding to dynamics, from reconstruction to intervention, recasts neural modeling as the search for perturbation-validated evolution laws governing population activity.

References

1. Mathis MW, Rotondo AP, Chang EF, Tolias AS, Mathis A. Decoding the brain: From neural representations to mechanistic models. Cell. 2024;187(21):5814-32.
2. Kriegeskorte N, Douglas PK. Interpreting encoding and decoding models. Current opinion in neurobiology. 2019;55:167-79.
3. Weichwald S, Meyer T, Özdenizci O, Schölkopf B, Ball T, Grosse-Wentrup M. Causal interpretation rules for encoding and decoding models in neuroimaging. Neuroimage. 2015;110:48-59.
4. Lange RD, Shivkumar S, Chattoraj A, Haefner RM. Bayesian encoding and decoding as distinct perspectives on neural coding. Nature neuroscience. 2023;26(12):2063-72.
5. Jonas E, Kording KP. Could a neuroscientist understand a microprocessor? PLoS computational biology. 2017;13(1):e1005268.
6. Barmpas K, Panagakis Y, Zoumpourlis G, Adamos DA, Laskaris N, Zafeiriou S. A causal perspective on brainwave modeling for brain–computer interfaces. Journal of Neural Engineering. 2024;21(3):036001.

 doi:10.20944/preprints202603.0480.v1

7. Vyas S, Golub MD, Sussillo D, Shenoy KV. Computation through neural population dynamics. Annual review of neuroscience. 2020;43(1):249-75.
8. Sani OG, Pesaran B, Shanechi MM. Dissociative and prioritized modeling of behaviorally relevant neural dynamics using recurrent neural networks. Nature neuroscience. 2024;27(10):2033-45.
9. Churchland MM, Cunningham JP, Kaufman MT, Foster JD, Nuyujukian P, Ryu SI, Shenoy KV. Neural population dynamics during reaching. Nature. 2012;487(7405):51-6.
10. Marques JC, Li M, Schaak D, Robson DN, Li JM. Internal state dynamics shape brainwide activity and foraging behaviour. Nature. 2020;577(7789):239-43.
11. Duncker L, Sahani M. Dynamics on the manifold: Identifying computational dynamical activity from neural population recordings. Current opinion in neurobiology. 2021;70:163-70.
12. Runfola C, Petkoski S, Sheheitli H, Bernard C, McIntosh AR, Jirsa V. A mechanism for the emergence of low-dimensional structures in brain dynamics. npj Systems Biology and Applications. 2025;11(1):32.
13. Recanatesi S, Farrell M, Lajoie G, Deneve S, Rigotti M, Shea-Brown E. Predictive learning as a network mechanism for extracting low-dimensional latent space representations. Nature communications. 2021;12(1):1417.
14. Greco A, Moser J, Preissl H, Siegel M. Predictive learning shapes the representational geometry of the human brain. Nature communications. 2024;15(1):9670.
15. Beiran M, Meirhaeghe N, Sohn H, Jazayeri M, Ostojic S. Parametric control of flexible timing through low-dimensional neural manifolds. Neuron. 2023;111(5):739-53. e8.
16. Fakhar K, Dixit S, Hadaeghi F, Kording KP, Hilgetag CC. When neural activity fails to reveal causal contributions. bioRxiv. 2023.
17. Schneider S, Lee JH, Mathis MW. Learnable latent embeddings for joint behavioural and neural analysis. Nature. 2023;617(7960):360-8.
18. Durstewitz D, Koppe G, Thurm MI. Reconstructing computational system dynamics from neural data with recurrent neural networks. Nature Reviews Neuroscience. 2023;24(11):693-710.
19. Paninski L, Ahmadian Y, Ferreira DG, Koyama S, Rahnama Rad K, Vidne M, Vogelstein J, Wu W. A new look at state-space models for neural data. Journal of computational neuroscience. 2010;29(1):107-26.
20. Sussillo D, Jozefowicz R, Abbott L, Pandarinath C. Lfads-latent factor analysis via dynamical systems. arXiv preprint arXiv:160806315. 2016.
21. Keshtkaran MR, Sedler AR, Chowdhury RH, Tandon R, Basrai D, Nguyen SL, Sohn H, Jazayeri M, Miller LE, Pandarinath C. A large-scale neural network training framework for generalized estimation of single-trial population dynamics. Nature methods. 2022;19(12):1572-7.
22. Hu A, Zoltowski D, Nair A, Anderson D, Duncker L, Linderman S. Modeling latent neural dynamics with gaussian process switching linear dynamical systems. Advances in Neural Information Processing Systems. 2024;37:33805-35.
23. Mudrik N, Chen Y, Yezerets E, Rozell CJ, Charles AS. Decomposed linear dynamical systems (dlds) for learning the latent components of neural dynamics. Journal of Machine Learning Research. 2024;25(59):1-44.
24. Abbaspourazad H, Erturk E, Pesaran B, Shanechi MM. Dynamical flexible inference of nonlinear latent factors and structures in neural population activity. Nature Biomedical Engineering. 2024;8(1):85-108.
25. ElGazzar A, van Gerven M. Generative modeling of neural dynamics via latent stochastic differential equations. arXiv preprint arXiv:241212112. 2024.
26. Sedler AR, Versteeg C, Pandarinath C. Expressive architectures enhance interpretability of dynamics-based neural population models. Neurons, behavior, data analysis, and theory. 2023;2023:10.51628/001c. 73987.
27. Gosztolai A, Peach RL, Arnaudon A, Barahona M, Vandergheynst P. MARBLE: interpretable representations of neural population dynamics using geometric deep learning. Nature Methods. 2025;22(3):612-20.
28. Gong C, Zhang C, Yao D, Bi J, Li W, Xu Y. Causal discovery from temporal data: An overview and new perspectives. ACM Computing Surveys. 2024;57(4):1-38.
29. Theocharous A, Gregoriou GG, Sapountzis P, Kontoyiannis I. Temporally causal discovery tests for discrete time series and neural spike trains. IEEE Transactions on Signal Processing. 2024;72:1333-47.

 doi:10.20944/preprints202603.0480.v1

30. Sachdeva P, Bak JH, Livezey J, Kirst C, Frank L, Bhattacharyya S, Bouchard KE. Resolving Non-identifiability Mitigates Bias in Models of Neural Tuning and Functional Coupling. bioRxiv. 2023.
31. Versteeg C, Sedler AR, McCart JD, Pandarinath C. Expressive dynamics models with nonlinear injective readouts enable reliable recovery of latent features from neural activity. ArXiv. 2023:arXiv: 2309.06402 v1.
32. Wang S, Hao W. A systematic computational framework for practical identifiability analysis in mathematical models arising from biology. Advanced Science. 2025;12(35):e04346.
33. Liu Y, Culpepper S. Designing learning intervention studies: Identifiability of heterogeneous hidden Markov models. Psychometrika. 2025;90(4):1258-83.
34. Wärnberg E, Kumar A. Perturbing low dimensional activity manifolds in spiking neuronal networks. PLOS computational biology. 2019;15(5):e1007074.
35. Vinograd A, Nair A, Kim JH, Linderman SW, Anderson DJ. Causal evidence of a line attractor encoding an affective state. Nature. 2024;634(8035):910-8.
36. O'Shea DJ, Duncker L, Goo W, Sun X, Vyas S, Trautmann EM, Diester I, Ramakrishnan C, Deisseroth K, Sahani M. Direct neural perturbations reveal a dynamical mechanism for robust computation. bioRxiv. 2022:2022.12. 16.520768.
37. Zhang Z, Dzialecka P, Russell LE, Ratto R, Buetfering C, Gauld OM, Selviah DR, Häusser M. A real-time all-optical interface for dynamic perturbation of neural activity during behavior. Cell Reports Methods. 2025;5(10).
38. Marin Vargas A, Chiappa AS, Perez Rotondo A, Mathis MW, Mathis A. Closed-loop imitation learning reveals muscle-centric and latent-goal codes in primate sensorimotor cortex. bioRxiv. 2026:2026.02. 01.703133.
39. Sourmpis C, Petersen CC, Gerstner W, Bellec G. Biologically informed cortical models predict optogenetic perturbations. eLife. 2026;14:RP106827.