Evaluating Foundation Models for the Brain: A Dynamical Systems Perspective

Anonymous Author(s)
Affiliation
Address
email

Abstract

Foundation models promise to transform neuroscience and brain–computer interfaces (BCIs), but their evaluation remains fragmented and often misleading. Standard benchmarks that emphasize in-distribution accuracy fail to capture what truly matters in dynamical domains: the ability to generalize across conditions that differ from those seen during training. In this perspective, we propose a unified framework for evaluating brain foundation models through the lens of dynamical systems theory. We introduce a generalization spectrum—a hierarchy of distribution shifts spanning system identity, parameter regimes, attractor structure, initial conditions, and observation noise—that clarifies what kinds of robustness should be expected from models claiming to be "foundational." We then map this spectrum onto a brain-specific taxonomy of distribution shifts—surface (hardware and noise), functional (state and task), and structural (subject and species)—to ground the framework in neuroscientific practice. Building on this foundation, we outline success criteria for brain foundation models: strong out-of-distribution generalization with minimal data, benchmark-validated performance across diverse tasks, and scalable power-law improvements with model and dataset size. This framework provides a common language for machine learning, neuroscience, and clinical research, and offers a roadmap for building evaluation cultures that distinguish narrow task solvers from truly foundational models.

1 Introduction

Foundation models [1] have begun to reshape research in neuroscience and brain–computer interfaces (BCIs). By training at scale across diverse datasets [3], they promise reusable representations that can accelerate clinical translation, enable more robust neural decoding, and provide a scientific lens onto brain dynamics. Yet, despite rapid progress, the evaluation of brain foundation models (BFMs) remains fragmented and inconsistent. Most current studies rely on narrow downstream benchmarks—such as classification accuracy on specific EEG datasets—that reward interpolation within the training distribution. Such metrics obscure a central scientific and practical question: *can these models generalize across the distribution shifts that inevitably arise in neural systems and their measurement?*

Traditional evaluation frameworks in machine learning, which emphasize performance on held-out i.i.d. test sets, are ill-suited for dynamical systems [8]. Brains, like other complex dynamical systems, exhibit multiple operating regimes, attractors, and sensitivities to initial conditions. Perturbations in hardware, subject state, or physiology generate test conditions that differ qualitatively from training data. Evaluating BFMs therefore requires moving beyond binary distinctions between "in-distribution" and "out-of-distribution." Instead, it demands a structured understanding of the *spectrum* of generalization challenges that models must confront.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

In this paper, we propose a framework for evaluating BFMs that unifies theory, neuroscience practice, and benchmarking protocols. At its core is a **generalization spectrum** derived from dynamical systems theory, which organizes distribution shifts into a principled hierarchy. From highest-to-lowest level of abstraction, they are: system identity, parameter regime, attractor structure, initial conditions, and observational noise. We map this spectrum onto a brain-specific taxonomy of shifts—**surface** (hardware and noise), **functional** (state and task), and **structural** (subject and species)—to clarify how evaluation should reflect neuroscientific realities. Finally, we outline concrete success criteria and benchmarking platforms that can instantiate these ideas, providing a roadmap for community standards. Our aim is to move evaluation from an ad hoc collection of tasks toward a rigorous scientific probe that distinguishes narrow solutions from truly foundational models.

2 The Generalization Spectrum for Dynamical Systems

At the heart of evaluating brain foundation models lies the question of how well a model generalizes to conditions beyond those represented in its training data. In static machine learning settings, this is often framed as a binary distinction: a model either performs well on a held-out i.i.d. test set ("in-distribution") or fails on qualitatively different data ("out-of-distribution"). However, this dichotomy collapses in the context of dynamical systems such as the brain. Here, distribution shifts occur along a continuum, ranging from mild perturbations to fundamental changes in the underlying system. Capturing this continuum requires a framework that recognizes multiple, nested levels of generalization.

We propose a **generalization spectrum**, which organizes distribution shifts in dynamical systems into five levels of increasing granularity (see Appendix A for a more rigorous mathematical formulation):

1. **System-level generalization.** Models are trained on one class of dynamical system but tested on a fundamentally different class. For example, a model trained on Lorenz dynamics but tested on Rössler dynamics, or in neuroscience, a model trained on human EEG but applied to rodent electrophysiology. This represents the most extreme form of generalization [27].

2. **Regime-level generalization.** The governing equations remain the same, but system parameters change, potentially crossing bifurcation points that alter qualitative behavior. In neuroscience, this corresponds to parameter shifts induced by pharmacological manipulations or long-term plasticity [26].

3. **Attractor-level generalization.** Within a fixed system and parameter setting, the dynamics may admit multiple attractors or operating modes. Generalization here requires predicting trajectories that lie in a different basin of attraction than those seen during training. For neural data, this is analogous to capturing transitions between resting state, task-engaged state, or pathological rhythms [9, 12].

4. **Initial-condition generalization.** Even within a single attractor, trajectories initialized at different states can evolve in ways unseen during training, especially in chaotic systems where small perturbations can amplify over time. For brain data, this includes variability across trials within the same subject and task [18, 12].

5. **Noise-level generalization.** The system dynamics remain fixed, but the observation channel is perturbed—for example, changes in sensor noise, hardware differences, or preprocessing pipelines. This is the minimal but most practically pervasive form of distribution shift in EEG and other neural recording modalities [13, 10].

This spectrum reframes generalization as a structured hierarchy rather than a binary label. It highlights that failures at different levels carry different implications: a model that fails at the noise level may be unsuitable for deployment, while one that succeeds up to attractor-level generalization but not at system-level still has strong scientific value. By explicitly situating evaluation within this hierarchy, we can diagnose the strengths and limits of brain foundation models with greater precision, and align expectations with both scientific and practical goals.

3 Mapping the Spectrum to Brain-Specific Shifts

While the generalization spectrum provides a theoretical backbone grounded in dynamical systems, its value lies in how it maps onto the practical distribution shifts encountered in neuroscience and brain–computer interface (BCI) research. To bridge theory and practice, we introduce a complementary taxonomy of shifts that arise in neural data: **surface**, **functional**, and **structural**. Each corresponds to a subset of levels in the generalization spectrum, anchoring abstract categories of dynamical change in concrete neuroscientific settings.

3.1 Surface Shifts: Noise and Hardware Variability

Surface shifts are the most immediate and practically pervasive. They occur when the observation channel changes while the underlying neural dynamics remain fixed. Examples include variation across EEG caps, differences in amplifier quality, electrode impedance changes, or the presence of movement artifacts. Such shifts map primarily onto the *noise-level* and, in some cases, the *initial-condition* levels of the spectrum. Success at this level indicates robustness to the day-to-day realities of neural data collection, and is a minimal requirement for BFMs intended for deployment [17, 10, 4].

3.2 Functional Shifts: Brain State and Task Dynamics

Functional shifts reflect changes in the operating mode of the brain without altering its structural identity. They include transitions between alertness and drowsiness, resting state and task-engaged state, or healthy and pathological rhythms. These correspond to *attractor-level* and *regime-level* generalization: the system identity is preserved, but parameters or initial conditions shift the brain into qualitatively different modes of operation. Evaluating generalization at this level probes whether a model captures the global structure of neural dynamics beyond the specific states represented in training data [19, 20].

3.3 Structural Shifts: Subject and Species Differences

Structural shifts are the most demanding, arising when the identity of the system itself changes. This includes variation across subjects, recording modalities, or species. Such shifts map directly to *system-level* generalization in the spectrum. For example, a model trained on human EEG that transfers to nonhuman primate recordings must abstract beyond superficial statistics to capture dynamical principles shared across brains. Success at this level is critical for building BFMs that support cross-subject BCIs or comparative neuroscience [25].

3.4 Unifying View

Together, the surface–functional–structural taxonomy provides an interpretable bridge between the abstract dynamical systems spectrum and the practical realities of brain research. It clarifies which generalization challenges a given benchmark actually tests, and highlights where progress is most urgently needed. Surface robustness ensures reliability, functional generalization supports adaptability, and structural generalization enables scalability across individuals and contexts. By situating evaluation within this taxonomy, the neuroscience and ML communities can converge on a common language for diagnosing and comparing the capabilities of brain foundation models.

4 Success Criteria for Brain/EEG Foundation Models

If foundation models are to fulfill their promise for neuroscience and BCIs, their evaluation must go beyond narrow task accuracy. We propose three ranked criteria that define what it means for a brain foundation model (BFM) to be truly *foundational*. Each criterion isolates a distinct property: *transferability*, *breadth*, and *scaling potential*. Together, they offer a principled way to assess whether models capture reusable structure in neural dynamics.

1. Out-of-Distribution Generalization with Data Efficiency (Transferability)

The first and most fundamental requirement is that BFMs enable rapid transfer to new tasks with minimal labeled data. This criterion probes whether a model's internal representations encode reusable neural structure, rather than task-specific correlations. Success means that, when faced with an unfamiliar decoding or forecasting problem, the BFM can be adapted with orders-of-magnitude less supervision than a model trained from scratch [27]. Transferability directly reduces the cost of deploying models across subjects, tasks, and clinical settings.

2. Benchmark-Validated Performance (Breadth)

The second requirement is breadth: a BFM must demonstrate consistent superiority across diverse and standardized evaluation suites. While transferability measures how well the model adapts to new situations, benchmark breadth measures how reliably it performs across a spectrum of existing tasks, brain states, modalities, and noise conditions. Strong results on established baselines [21, 15, 20] indicate that the benefits of large-scale pretraining are realized not just in isolated cases, but across the field. Breadth ensures that a model is not a niche tool, but a robust platform others can trust and build upon.

3. Scalable Power-Law Behavior (Scaling Potential)

The third requirement concerns the trajectory of progress: foundational models must improve predictably as more data and parameters are added. The criterion of power-law scaling is the hallmark of foundation models in other domains [14, 11]. Scaling potential provides the strategic justification for long-term investment: if gains follow a reliable scaling curve, then expanding datasets and architectures will continue to pay off. It also offers a common yardstick to compare very different modeling approaches, since scaling laws expose whether progress is driven by structure learning or by brittle overfitting.

4.1 Interpreting Scaling Failures

Consider the case of building an EEG foundation model. When scaling behavior departs from power-law trends, the implications differ depending on context:

- **Plateau below threshold:** Performance saturates before reaching levels required for practical applications, suggesting that EEG alone may be insufficient unless paired with other modalities.
- **Plateau above threshold:** Performance saturates at levels well above application requirements, implying that the task may be too simple to justify large-scale foundation modeling.
- **Model underperforming, data valuable:** If power-law scaling exists but is weak (small exponent), it's possible that the data contain reusable structure, but current architectures fail to unlock it. Investment and efforts should shift toward advancing ML techniques. Signals of architectural underfitting can be corroborated by intrinsic-dimension estimates of the objective landscape [16].
- **No power law (noise-dominated regime):** If model/data scaling yields no performance improvements, this suggests that the signal-to-noise ratio of EEG may be too low to support general-purpose foundation models, restricting utility to narrow, task-specific applications.

Together, these success criteria establish a rigorous checklist for BFMs. A model that demonstrates strong OOD transfer, validates across diverse benchmarks, and scales predictably with resources can legitimately be called foundational. Conversely, models that fail to meet these criteria should be recognized as narrow tools, valuable in context but not general platforms for neuroscience.

5 Benchmarking Protocols and Platforms

Establishing rigorous success criteria requires equally rigorous protocols for measurement. In natural language processing and computer vision, shared benchmarks such as GLUE [24, 23] and ImageNet [6] accelerated progress by standardizing evaluation and making comparisons transparent. Analogous

platforms in control and RL (e.g., OpenAI Gym [2], DeepMind Control Suite [22], D4RL [7]) demonstrate how shared tasks and standard metrics shape progress. For brain foundation models (BFMs), we argue that the community must similarly converge on shared evaluation protocols that probe robustness, efficiency, and scaling across the generalization spectrum. Again, taking EEG foundation models as a case study, we offer EEG-specific diagnostics.

5.1 EEG-Specific Diagnostics

Electroencephalography provides a practical testbed for BFM evaluation due to its accessibility, prevalence, and inherent challenges of noise and variability. We propose a suite of diagnostic tasks that capture distinct dimensions of robustness and generalization:

1. **Task-specific benchmarking:** Compare foundation models to baselines trained from scratch with matched data.
2. **OOD task generalization:** Evaluate zero- or few-shot transfer to tasks not included in pretraining.
3. **Scaling law validation:** Plot loss against dataset and model sizes to test for power-law behavior.
4. **Embedding richness:** Compare learned embeddings to raw time-series and handcrafted EEG features using a fixed decoder.
5. **Decoder vs. embedding contribution:** Vary decoder complexity while freezing embeddings to separate representational power from decoding capacity.
6. **Channel degradation:** Measure performance as EEG channels are progressively removed to simulate hardware constraints.
7. **Real-world signal robustness:** Quantify performance under common noise sources such as blinks, muscle artifacts, and impedance variability [13, 5].
8. **Cross-hardware generalization:** Evaluate transfer across medical-grade and consumer EEG systems.

This diagnostic suite provides a multi-dimensional profile of a BFM, making it possible to identify strengths, weaknesses, and suitability for practical use cases.

6 Outlook and Call to Action

Brain foundation models promise to accelerate discovery, translation, and application across neuroscience and brain–computer interfaces. Yet without principled evaluation, it is impossible to know whether these models capture reusable structure in brain dynamics or merely interpolate within narrow datasets. We have outlined a framework for rigorous evaluation built on three pillars: a dynamical-systems-inspired *generalization spectrum*, a neuroscience-grounded *surface–functional–structural taxonomy*, and concrete *benchmarking protocols*.

Progress in this area requires not only better models but also better evaluation cultures. Community benchmark suites—analogous to ImageNet for vision [6] and GLUE for language [24, 23]—are needed to anchor claims of generality, ensure reproducibility, and accelerate progress through shared goals. By organizing evaluation around structured distribution shifts, researchers can identify both the strengths and limitations of a model and draw clearer implications for scientific and clinical use.

Looking ahead, we call for collaboration across machine learning, neuroscience, and clinical communities to adopt such frameworks. Surface robustness ensures reliability, functional generalization supports adaptability, and structural generalization enables cross-subject and cross-species transfer. Together, these capabilities would make brain foundation models genuinely foundational.

In short, foundation models will only be as transformative as the standards by which they are judged. By uniting theory, taxonomy, and practice, evaluation itself can become a scientific probe that distinguishes narrow tools from models capturing the deeper principles of brain dynamics.

References

[1] Bommasani, R. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

[2] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym.

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[4] Casson, A. J., Yates, D. C., Smith, S. J., Duncan, J. S., and Rodriguez-Villegas, E. (2010). Wearable electroencephalography. *IEEE engineering in medicine and biology magazine*, 29(3):44–56.

[5] Delorme, A., Sejnowski, T., and Makeig, S. (2007). Enhanced detection of artifacts in eeg data using higher-order statistics and independent component analysis. *Neuroimage*, 34(4):1443–1449.

[6] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

[7] Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. (2021). D4rl: Datasets for deep data-driven reinforcement learning.

[8] Gilpin, W. (2021). Chaos as an interpretable benchmark for forecasting and data-driven modelling. *arXiv preprint arXiv:2110.05266*.

[9] Göring, N., Hess, F., Brenner, M., Monfared, Z., and Durstewitz, D. (2024). Out-of-domain generalization in dynamical systems reconstruction. *arXiv preprint arXiv:2402.18377*.

[10] Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., and Hämäläinen, M. S. (2014). Mne software for processing meg and eeg data. *neuroimage*, 86:446–460.

[11] Hestness, J., Narang, S., Ardalani, N., Diamos, G. F., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. (2017). Deep learning scaling is predictable, empirically. *ArXiv*, abs/1712.00409.

[12] Hu, J., Hu, Y., Chen, W., Jin, M., Pan, S., Wen, Q., and Liang, Y. (2024). Attractor memory for long-term time series forecasting: A chaos perspective. *Advances in Neural Information Processing Systems*, 37:20786–20818.

[13] Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., Mckeown, M. J., Iragui, V., and Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(2):163–178.

[14] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

[15] Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013.

[16] Li, C., Farkhoor, H., Liu, R., and Yosinski, J. (2018). Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*.

[17] Lopez-Calderon, J. and Luck, S. J. (2014). Erplab: an open-source toolbox for the analysis of event-related potentials. *Frontiers in human neuroscience*, 8:213.

[18] Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. *Tellus*, 21(3):289–307.

[19] Makeig, S., Debener, S., Onton, J., and Delorme, A. (2004). Mining event-related brain dynamics. *Trends in cognitive sciences*, 8(5):204–210.

[20] Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., and Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001.

[21] Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., and Ball, T. (2017). Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420.

[22] Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., de Las Casas, D., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., Lillicrap, T., and Riedmiller, M. (2018). Deepmind control suite.

[23] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2020). Superglue: A stickier benchmark for general-purpose language understanding systems.

[24] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). Glue: A multi-task benchmark and analysis platform for natural language understanding.

[25] Wu, D., Xu, Y., and Lu, B.-L. (2020). Transfer learning for eeg-based brain–computer interfaces: A review of progress made since 2016. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):4–19.

[26] Yin, Y., Ayed, I., de Bézenac, E., Baskiotis, N., and Gallinari, P. (2021). Leads: Learning dynamical systems that generalize across environments. *Advances in Neural Information Processing Systems*, 34:7561–7573.

[27] Zhang, Y. and Gilpin, W. (2024). Zero-shot forecasting of chaotic systems. *arXiv preprint arXiv:2409.15771*.

A Rigorous Definitions of the Generalization Spectrum

Setup and notation. Let $(\mathcal{X}, \|\cdot\|)$ be a smooth n-dimensional state space and let a (possibly parameterized) dynamical system be given by a smooth vector field $f_{\Sigma,\theta} : \mathcal{X} \to \mathbb{R}^n$, where $\Sigma \in \mathcal{S}$ indexes the *system identity* (e.g., family of governing equations) and $\theta \in \Theta \subset \mathbb{R}^p$ collects continuous parameters. Denote by $\phi^t_{\Sigma,\theta} : \mathcal{X} \to \mathcal{X}$ the associated flow (or time-t map for discrete time). Let $\mathcal{A}(f_{\Sigma,\theta})$ be the set of attractors of $f_{\Sigma,\theta}$, and for $A \in \mathcal{A}(f_{\Sigma,\theta})$ let μ_A be an invariant probability measure supported on A (e.g., an SRB measure in chaotic regimes). Let $\mathsf{B}(A) \subset \mathcal{X}$ denote the basin of attraction of A.

We assume an *observation channel* given by a measurable map $h : \mathcal{X} \to \mathcal{Y}$ (deterministic sensing) and a family of noise laws $\{\mathsf{K}_\eta(\cdot \mid y)\}_{\eta \in \mathcal{H}}$ on $\mathcal{Y}$ (stochastic sensing), so that observed data Y_t is generated from

$$X_t = \phi^t_{\Sigma,\theta}(X_0), \qquad Y_t \sim \mathsf{K}_\eta(\cdot \mid h(X_t)).$$

A training configuration is a tuple

$$\mathbf{M}_{\mathrm{train}} = (\Sigma, \theta, A, \rho, \eta, h),$$

where ρ is a distribution of initial conditions supported on $\mathsf{B}(A)$ (often $\rho = \mu_A$ for on-attractor sampling). A test configuration $\mathbf{M}_{\mathrm{test}}$ is defined analogously, possibly with altered components.

A learning algorithm produces a predictor g (e.g., a forecaster, decoder, or controller). Let $\ell : \mathcal{Y}^{\mathbb{N}} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss comparing a prediction (which may depend on a context window) to the next observation. For a configuration $\mathbf{M}$, define the *asymptotic pathwise loss*

$$\mathcal{L}(g; \mathbf{M}) \;=\; \limsup_{T \to \infty} \; \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \ell\big(g(Y_{0:t-1}),\, Y_t\big) \right], \tag{1}$$

where the expectation is over $X_0 \sim \rho$ and the observation noise. Under ergodicity of μ_A and stationarity of the observation channel, (1) coincides with the μ_A-expectation of a one-step loss by Birkhoff's theorem.

We measure *generalization* by comparing the same learned g across configurations via the *generalization gap*

$$\Gamma(g; \mathbf{M}_{\mathrm{train}} \to \mathbf{M}_{\mathrm{test}}) \;=\; \mathcal{L}(g; \mathbf{M}_{\mathrm{test}}) - \mathcal{L}(g; \mathbf{M}_{\mathrm{train}}).$$

Below we define five nested families of allowable test configurations; moving outward increases dynamical novelty.

Level 1: System generalization (new governing equations). Here the test system may belong to a *different* identity class. Formally, let $\Sigma^\star \in \mathcal{S}$ satisfy $\Sigma^\star \notin \mathrm{supp}(\mathcal{D}^{\Sigma}_{\mathrm{train}})$, where $\mathcal{D}^{\Sigma}_{\mathrm{train}}$ is the distribution over system identities used to generate training data. Allow arbitrary $\theta^\star \in \Theta$, $A^\star \in \mathcal{A}(f_{\Sigma^\star,\theta^\star})$, $\rho^\star \ll \mathsf{B}(A^\star)$, while keeping the observation family (h, η) fixed (unless otherwise specified). A *Level 1 test* admits any $\mathbf{M}_{\mathrm{test}} = (\Sigma^\star, \theta^\star, A^\star, \rho^\star, \eta, h)$ with $\Sigma^\star$ disjoint from the training support. Success at Level 1 requires small Γ uniformly over such $\mathbf{M}_{\mathrm{test}}$ (e.g., in expectation over a held-out $\mathcal{D}^{\Sigma}_{\mathrm{test}}$).

Level 2: Regime generalization (parameter/bifurcation shift). Fix the identity Σ but shift parameters beyond the training support. Let $\Theta_{\mathrm{train}}, \Theta_{\mathrm{test}} \subset \Theta$ be disjoint (up to negligible overlap). A *Level 2 test* fixes Σ and admits any $\mathbf{M}_{\mathrm{test}} = (\Sigma, \theta^\star, A^\star, \rho^\star, \eta, h)$ with $\theta^\star \in \Theta_{\mathrm{test}}$, $A^\star \in \mathcal{A}(f_{\Sigma,\theta^\star})$, $\rho^\star \ll \mathsf{B}(A^\star)$. Note this explicitly includes *bifurcation* events: qualitative changes in $\mathcal{A}(f_{\Sigma,\theta})$ as θ crosses critical sets.

Level 3: Attractor generalization (new invariant set, same system/regime). Fix (Σ, θ). Let $A \neq A^\star$ be distinct attractors of $f_{\Sigma,\theta}$ with basins $\mathsf{B}(A)$ and $\mathsf{B}(A^\star)$. A *Level 3 test* keeps (Σ, θ) fixed and admits any $\mathbf{M}_{\mathrm{test}} = (\Sigma, \theta, A^\star, \rho^\star, \eta, h)$ with $\rho^\star \ll \mathsf{B}(A^\star)$. This probes whether g captures global structure beyond the specific invariant set seen in training.

Level 4: Initial-condition generalization (local within-basin shifts). Fix (Σ, θ, A) and its basin $\mathsf{B}(A)$. Let ρ be the training initial distribution (often μ_A or a neighborhood thereof). A *Level 4 test* perturbs initial conditions within $\mathsf{B}(A)$ while keeping all other components fixed: $\mathbf{M}_{\mathrm{test}} =$

$(\Sigma, \theta, A, \rho^\star, \eta, h)$ with $\mathrm{supp}(\rho^\star) \subset \mathsf{B}(A)$ and $\rho^\star$ *not* absolutely continuous w.r.t. ρ in general (to allow targeted shifts).

For fine-grained control, decompose local perturbations using the Oseledets splitting (or, locally, the eigenspaces of the Jacobian $J_f(x)$): for $x \in A$, let $\mathcal{E}_s(x)$ and $\mathcal{E}_u(x)$ denote stable and unstable subspaces. Define *stable-direction tests* by pushing ρ via maps $x \mapsto x + \delta$ with $\delta \in \mathcal{E}_s(x)$ and small $\|\delta\|$; and *unstable-direction tests* analogously with $\delta \in \mathcal{E}_u(x)$. Unstable-direction tests typically induce rapid divergence in trajectory prefixes while remaining within $\mathsf{B}(A)$.

Level 5: Observation/noise generalization (surface/channel shift). Fix $(\Sigma, \theta, A, \rho)$. Let the observation channel vary within a specified family. Write the stochastic sensing as a Markov kernel $\mathsf{K}_\eta(\cdot \mid h(x))$ on $\mathcal{Y}$. A *Level 5 test* admits any $\mathbf{M}_{\mathrm{test}} = (\Sigma, \theta, A, \rho, \eta^\star, h^\star)$ with $(h^\star, \eta^\star)$ in a prescribed perturbation class. For example, one may constrain observation shifts by

$$\sup_{x \in \mathcal{X}} D\big(\mathsf{K}_{\eta^\star}(\cdot \mid h^\star(x)) \,\|\, \mathsf{K}_\eta(\cdot \mid h(x))\big) \;\le\; \varepsilon,$$

for a divergence D (e.g., total variation, Wasserstein, or KL where defined), or by structural constraints (e.g., channel dropouts, downsampling, additive noise with altered covariance). This level formalizes hardware changes, artifacts, and measurement noise mismatches.

Partial order and severity. The spectrum induces a natural partial order of novelty:

$$\text{Level 5} \;\preceq\; \text{Level 4} \;\preceq\; \text{Level 3} \;\preceq\; \text{Level 2} \;\preceq\; \text{Level 1},$$

since each outer level relaxes constraints of the inner ones. A *severity index* can be attached to a test by metrizing each coordinate: (i) a discrete metric on identities Σ; (ii) a parameter metric $d_\Theta(\theta, \theta^\star)$; (iii) an invariant-set distance $d_{\mathcal{P}}(\mu_A, \mu_{A^\star})$ (e.g., Wasserstein on invariant measures); (iv) a local shift size $\|\delta\|$ and alignment with $\mathcal{E}_{u/s}$ (e.g., via local Lyapunov exponents); (v) an observation-channel distance $\sup_x D(\mathsf{K}_{\eta^\star}(\cdot|h^\star(x))\|\mathsf{K}_\eta(\cdot|h(x)))$. These compose into a vector-valued difficulty label for each $\mathbf{M}_{\mathrm{test}}$.

Evaluation primitives. Given a trained predictor g from $\mathbf{M}_{\mathrm{train}}$, we report: (i) the asymptotic loss $\mathcal{L}(g; \mathbf{M}_{\mathrm{test}})$ at each level; (ii) the generalization gap $\Gamma(g; \mathbf{M}_{\mathrm{train}} \to \mathbf{M}_{\mathrm{test}})$; and (iii) *valid prediction time* τ_ϵ (the largest horizon for which mean error stays below ϵ) to separate short-horizon tracking from long-horizon invariant adherence. For Level 3, invariant-set fidelity can be scored by comparing empirical measures of forecasts to $\mu_{A^\star}$ (e.g., via $d_{\mathcal{P}}$) to capture attractor-shape preservation even when pointwise errors grow.

Remarks. (1) The framework extends to random or controlled dynamical systems by letting $f_{\Sigma,\theta}$ define a cocycle over a driving process and by augmenting the configuration with control policies; definitions are unchanged with flows replaced by skew-product flows. (2) In neuroscience applications, the Surface–Functional–Structural taxonomy is captured by: Level 5 (surface: sensing/noise), Levels 4–3 (functional: state/mode within fixed (Σ, θ)), and Levels 2–1 (structural: parameter or identity changes across subjects/species). (3) The dynamical viewpoint avoids i.i.d. assumptions: losses are time averages under invariant measures, OOD is expressed as explicit changes in $(\Sigma, \theta, A, \rho, h, \eta)$, and "difficulty" is measured by dynamical distances rather than solely by sample-distribution divergences.