NETHOBENCH: FORECASTING DESIGN AND CONDITIONING SHAPE THE STRUCTURAL REALISM OF GENERATED NEURAL DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting models are increasingly used to model neural population activity, yet existing benchmarks typically evaluate them by pointwise predictive fidelity to held-out observations. In stochastic, partially observed systems, agreement with one sampled realization is distinct from preserving the statistical and dynamical structure of the process. We introduce Nethobench, a benchmark that evaluates structural agreement across five metric families: distributional, temporal–spectral, relational, geometric, and state-dynamical. We validate Nethobench using targeted perturbations and stochastic systems with known dynamics, and apply it to mouse calcium imaging, mouse Neuropixels recordings, and human intracranial recordings, using empirical references to calibrate interpretation. Nethobench detects targeted structural failures and assigns high scores to independent samples from the correct synthetic process despite trajectory mismatch. It tracks structural changes during training and under domain shift. Model comparisons reveal how forecasting formulation, architecture, training regime, and loss shape structural profiles, while OmniMouse input ablations improve all five family means when behavior is added to video in one held-out session. In Brain Treebank, analog retrieval passes more structural checks while a multistep GRU better predicts the observed realization. These results motivate reporting calibrated family-level structural metrics alongside pointwise accuracy, proper probabilistic scores, and task-specific endpoints when evaluating neural forecasts across species and recording modalities.

1 INTRODUCTION

Forecasting models are increasingly used as generative surrogates of neural population activity. (Lueckmann et al. (2025); Duan et al. (2025); Azabou et al. (2023); Calcagno et al. (2024)). Predicting future activity from past observations provides a fundamental test of our understanding of complex systems, including the brain, by testing whether learned patterns generalize to unseen activity. Such a test requires distinguishing three evaluation questions. Does a forecast match the observed realization? Does a predictive distribution assign appropriate probabilities under the conditioning context? And which population-level properties does generated activity preserve? Pointwise error addresses the first question ; proper scoring rules and calibration the second (Gneiting & Raftery (2007)). The third remains less systematically addressed by existing benchmarks and requires explicit structural diagnostics to identify which aspects of population activity a model reproduces or distorts.

Assessing *structural agreement* - defined as agreement on specified summaries under a fixed observation transform, support, and horizon - is essential in stochastic, partially observed systems. A same process can indeed produce realizations that differ point by point, while possessing similar marginal statistics, temporal dependence, inter-region relationships, latent geometry, and state-transition organization. Conversely, conditional-mean forecasters produce low pointwise errors, while suppressing the variability of the original realizations. We thus argue that structural evaluation should complement predictive-distribution evaluation, when using generative models to emulate dynamical processes.

Neuro-ethological recordings make these questions especially relevant: freely moving animals engage in spontaneous behavior, where neural dynamics are non-stationary, the behavioral state space is high dimensional and incompletely observed, and trajectory distributions are non-ergotic on short

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

timescales (Musall et al. (2019); Stringer et al. (2019); Dennis et al. (2021); Mazor & Laurent (2005); Sussillo & Barak (2013); Cunningham & Yu (2014)). Building generative models of these processes introduces choices that can favor different aspects of the recorded structure. One-step and multistep objectives impose different long-horizons constraints (Ben Taieb & Hyndman (2012); Benechehab et al. (2024)) ; teacher forcing and self-feedback change the histories encountered during training (Bengio et al. (2015)); and losses prioritize different statistics (Jiang et al. (2024); Le Guen & Thome (2019)). Evaluating these choices motivates the need for evaluation benchmarks that reveal which structural properties are preserved.

Existing benchmarks standardize stimulus-response prediction, held-out neuron prediction, and representational similarity (Schrimpf et al. (2020); Turishcheva et al. (2024a;b); Pei et al. (2022); Lueckmann et al. (2025); Duan et al. (2025)). Related methods compare neural representations and dynamics (Versteeg et al. (2025); Nejatbakhsh et al. (2024); Ostrow et al. (2023); Williams et al. (2022)). Nethobench complements these approaches with a family-resolved framework for diagnosing structural agreement in generated neural activity.

Our contributions are threefold. **First**, we define five metric families - distributional, temporal, relational, geometric, and state-dynamical (Figure 1) - and an evaluation contract for comparing generated activity with reference recordings, using empirical references to contextualize agreement. **Second**, controlled oracles and perturbations test sensitivity known successes and failure modes (Figure 2); evaluation on human intracranial forecasts expose a pointwise fidelity-structural fidelity ranking reversal and evaluations of behavior- and stimulus-conditioned forecasts show that conditioning can shape the structure of generated neural activity (Figure 3). **Third**, a common widefield task compares forecasting formulations, architectures, training regimes, losses, and horizons, identifying the structural gains and failures associated with these design choices (Figure 4). Together, these experiments support reporting structural properties a forecast preserves, together with its prediction error and predictive uncertainty.

2 NETHOBENCH: EVALUATING STRUCTURAL REALISM IN NEURAL FORECASTS

We now formalize Nethobench as an evaluation framework for comparing reference and generated neural activity. Given aligned tensors $X, Y \in \mathbb{R}^{M \times T \times R}$, where M, T and R denote sequences, time points and recorded regions or units, the benchmark asks which statistical and dynamical properties are preserved. Family profiles complement pointwise fidelity by distinguishing marginal, temporal, relational, geometric and state-transition agreement.

2.1 METRIC FAMILIES

Nethobench reports structural agreement in five metric families (Figure 1). Each family reflects established approaches to characterizing neural population activity in systems and computational neuroscience Cohen & Kohn (2011); Murray et al. (2014); Buzsáki & Mizuseki (2014); Cunningham & Yu (2014); Mazzucato et al. (2015). Reporting these families separately makes failures interpretable: a model can preserve marginal activity statistics while losing temporal structure, or reproduce the dominant latent subspace while distorting inter-region dependencies.

Distribution. This family tests whether generated activity preserves reference marginal statistics, including value distributions, tails, moments, and mean activity. It detects amplitude compression, mean shifts, excessive smoothing, and loss of rare events that may be missed by short-horizon error metrics.

Temporal-spectral. This family tests whether activity evolves with reference-like temporal structure. It combines trajectory summaries in a reference-defined latent space with autocorrelation of activity and its increments, and power-sensitive spectral agreement. Together these expose changes in trajectory organization, temporal dependence, frequency content, and amplitude; no single component identifies the dynamical law.

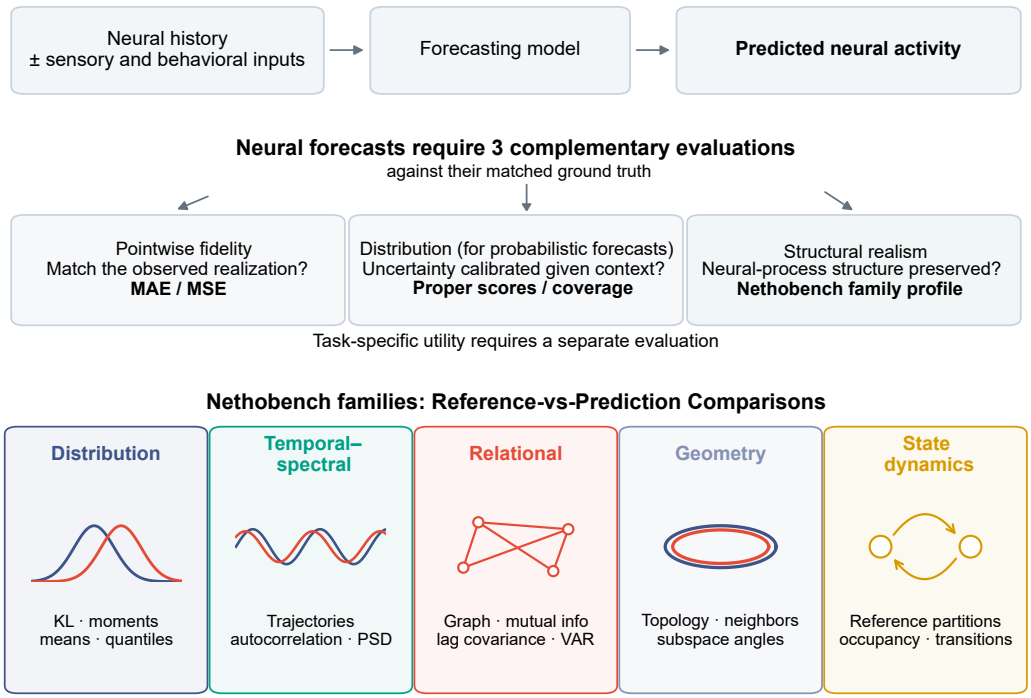


Figure 1: **Overview of neural forecasting challenges and benchmark tasks** Five metric families assess distributional, temporal-spectral, relational, geometric, and state-dynamical agreement with reference activity. These structural profiles complement pointwise accuracy and predictive-distribution evaluation.

Relational. This family tests whether inter-region dependencies are preserved. Nethobench compares graph topology, cross-region mutual information, lagged covariance, and a fitted one-step VAR-response proxy to summarize instantaneous and delayed population coordination. These summaries do not identify causal interactions. The pinned lag/VAR implementation pools sequences and one matrix triangle; a within-sequence, directed variant is tested separately in Appendix C.6.

Geometry. This family tests whether generated activity preserves reference-like latent geometry. It combines persistent topology, local neighborhood preservation, and principal-subspace alignment, separating dominant-subspace agreement from within-manifold structure.

State dynamics. This family tests whether generated activity visits and transitions between latent states like the reference. Reference embeddings and cluster centers define occupancy and multi-lag transition metrics, measuring both where activity lies and how it flows through latent space. These states are coarse-grained partitions of continuous activity, not inferred biological or cognitive states.

Aggregation and reporting. The 18 retained components are mapped to agreement scores in $[0, 1]$ and averaged within each family. With F_D, F_T, F_R, F_G, F_S denoting the five family scores, the fixed composite is

$$C = 0.22F_D + 0.18F_T + 0.24F_R + 0.18F_G + 0.18F_S. \quad (1)$$

Unavailable components are excluded with within-family weight renormalization, and their absence is reported. Family profiles remain the primary object of interpretation. Real-real and shuffled comparisons contextualize scores under a particular dataset and window length, providing empirical references. Detailed transforms, hyperparameters, and missingness rules appear in Appendix C.1.

Metric selection and robustness. A six-criterion audit covers 16 of 18 retained components (Figure S4); ACF and PSD have dedicated temporal controls (Figures S4b-S2). Matched-support

tests assess sensitivity to family weights, PCA dimensionality, state counts, and lags (Figures S5-S7; Appendix C.3). Family scores covary (Figure S6a-c), but their differing responses to controlled corruptions support failure localization (Figures S4d-S6d). These analyses assess diagnostic coverage and robustness, but we do not claim statistical independence among families or uniquely optimal weights.

3 Controlled and empirical evaluation

3.1 Controlled validation with structured perturbations and stochastic oracle dynamics

We test Nethobench using known generating processes and specified corruptions (Appendix C.2).

Independent realizations and targeted failures. Independent samples from the same generating process receive high structural agreement despite trajectory mismatch, while targeted perturbations produce distinct family-level deficits. Using a latent AR-based calcium generator (Figure 2a), we obtain mean structural agreement of 0.883 and pointwise fidelity (Appendix C.1) of 0.321 across 20 oracle realizations (b). Although both scores range from 0 to 1, they measure different properties and behave as expected: shared process statistics yield high structural agreement despite low pointwise fidelity. We then test sensitivity to thirteen perturbations, which produce distinct, overlapping family deficits (c), with class-averaged scores declining across normalized doses (d). These averages do not imply that every individual corruption is monotonic or affects only one family.

Structural agreement is not determined by pointwise error. Five transformations of widefield activity yield distinct family-score profiles despite identical MSE relative to the original recordings (Figure 2e; Appendix C.2). Complementary controls show that regional time shuffling preserves distributional agreement while disrupting temporal and relational structure (f), and that mean-preserving variance contraction affects several families (g), demonstrating sensitivity to reduced temporal variability. Dedicated ACF/PSD controls and component ablations further characterize sensitivity to temporal distortions (Figures S1-S2).

Beyond a single synthetic process. Evaluations across nonlinear, switching, chaotic, recurrent and spiking-calcium systems show that structural scores depend on which process properties a candidate preserves (Figure S3). In spiking-calcium, an independent oracle exceeds a pure shuffle in all five seeds, whereas a mixture retaining approximately 72% of the reference exceeds the oracle, consistent with its retained structure and pointwise proximity. Synthetic BOLD controls further reveal dependence on the observation transform and short-window component availability. These controls characterize behavior in the tested synthetic systems; without establishing validity on real fMRI.

3.2 External validation across models and recording modalities

OmniMouse. We test whether Nethobench detects structural differences among conditional neural predictions generated under different modality-conditioning configurations. We apply the public OmniMouse-80M checkpoint (Willeke et al. (2026)) to one held-out naturalistic-video session, keeping targets, the evaluated neuron panel, preprocessing, and four sequence-level splits fixed across configurations. We generate two successive 30-sample neural predictions, yielding 2 s of prediction-only scoring support at 30 Hz. These forecasts enable us to evaluate conditional prediction with observed target-interval covariates, on top of prediction from exclusively past neural inputs (Figure S10; Appendix B.2).

In this session, adding behavior to video raises mean scores in all five families, with composites of 0.401 versus 0.351. Adding video to neural history and behavior similarly raises the five means, with composites 0.400 versus 0.377 (Figure 3a; Appendix Table 5). Video plus behavior and full conditioning have nearly identical composites but have different family profiles: the former has higher temporal-spectral and relational means, the latter higher geometry and state dynamics. These are descriptive contrasts and limited to one session; real-real and shuffled references contextualize the scores. A separate native-horizon analysis with shared context yields a composite ordering consistent

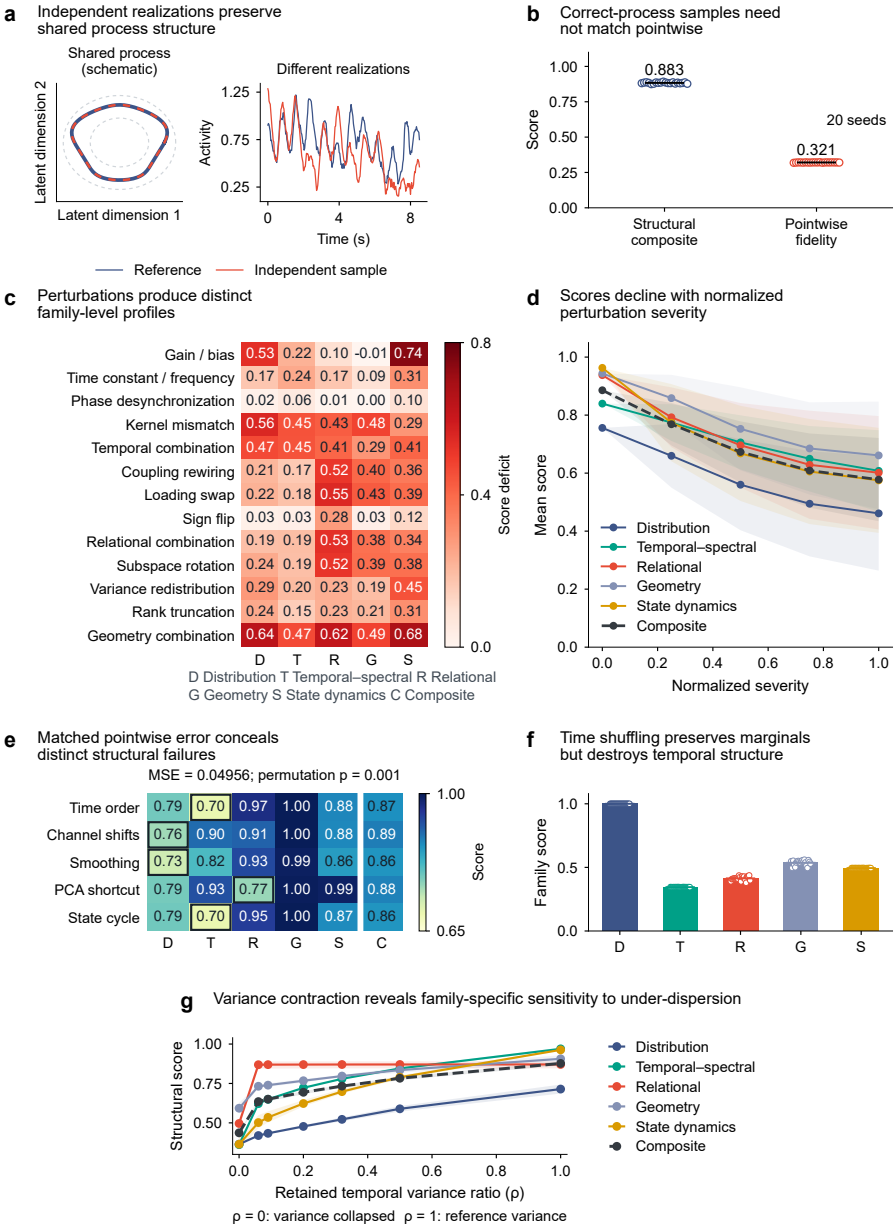


Figure 2: **Controlled validation of structural diagnostics.** (a) Latent AR-based calcium generator AR: process schematic and independent simulated traces. (b) Twenty oracle realizations: structural agreement and the non-proper fidelity sidecar measure different targets (mean ± SEM); 0.883 and 0.321 are not comparable scales. (c) Maximum-dose deficits for 13 corruptions; signed annotations retain improvements relative to the oracle. (d) Mean dose responses across the 13 classes; bands show SD across corruption classes. (e) Five widefield controls with matched MSE have different family profiles ; blocked permutation $p = 0.001$, 999 permutations. (f) Regional time shuffling on Widefield imaging data preserves marginals while disrupting other properties (20 shuffles). (g) Mean-preserving temporal-variance contraction. Full definitions and uncertainty conventions are given in Appendix C.2.

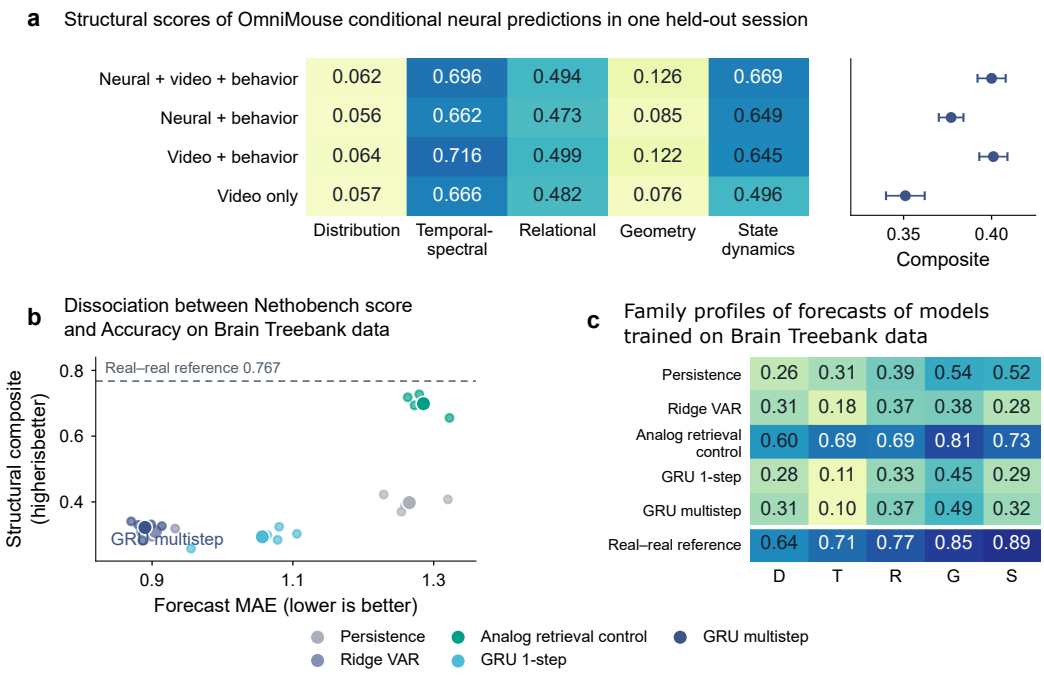


Figure 3: **Structural profiles and realization accuracy answer different questions.** (a) OmniMouse family means and composite mean ± SEM across four sequence-level splits in one held-out session, scoring 60 predicted neural samples (2 s) with initial recorded context excluded (Figure S10b). Enabled video and behavioral inputs include observations from the predicted interval. Neural-history configurations encode a 25-sample prefix, consisting of recorded activity for the first call and model predictions for the second. These within-session contrasts are descriptive. (b) MAE and Nethobench composite score of 5 Brain Treebank participants (small points) and equally weighted method means (large points): the analog retrieval control has higher structural agreement, while GRU multistep has lower native MAE. (c) Family scores averaged across Brain Treebank participants, averaging seeds within movies and movies within five participants. The dashed line in (b) and last row in (c) show the empirical real-real reference (composite 0.767): symmetric comparisons of alternating 25-window halves, averaged within participants. Common-component sensitivity appears in Figure S21. Scored forecasts are 20s long. Human preprocessing is offline; training-only contact-selection and retrieval checks are detailed in Appendix B.6.

with OmniMouse's reported response correlation and loss, while revealing family-specific differences (Appendix Figure S10-Table 4)

Brain Treebank. We test whether Nethobench extends beyond animal calcium imaging to human intracranial electrophysiology, capturing structural differences between forecasts across this change in species and recording modality. To this end, we evaluate offline high-gamma activity from 64 contacts in each of five Brain Treebank participants (Wang et al. (2024)), with three movie runs per participant. Models use 5 s of neural history to forecast 20 s, holding out each movie and fitting forecasts to the other two. Comparators are persistence, ridge VAR, an analog retrieval control, and GRUs trained with one-step or multistep objectives (Appendix B.3).

GRU multistep achieves the lowest pointwise error (MAE: 0.890), whereas analog retrieval achieves the highest structural score (Nethobench composite: 0.699) despite larger MAE (1.285; Figure 3b). Retrieval's advantage extends across all five family means (c), demonstrating that better pointwise accuracy need not imply better structural preservation. Retrieval uses only training data and matches on observed context, but its structural advantage does not establish that the retrieved continuations are appropriate futures or that it is the better forecaster overall.

The GRU multistep-minus-one-step structural contrast averages 0.0285 and is positive in all five participants (exact two-sided $p = 0.0625$). Participant, seed, and component diagnostics appear in Figures S20-S21; human perturbation controls assess sensitivity to noise, channel shifts, low-variance attenuation, and temporal block permutation (Figure S22). Training-only contact selection reproduces 14 of 15 folds from the original all-movie QC; omitting the remaining fold preserves the accuracy-structure ordering. The original contact selection and noncausal high-gamma preprocessing limit this experiment to an offline structural comparison rather than a fully prospective forecasting validation (Appendix B.6).

Additional recording modalities. We further applied Nethobench to single-neuron two-photon calcium traces (Figure S16) and IBL repeated-site Neuropixels population firing rates (Figure S17 and S18). For two-photon forecasts, a population-conditioned MLP shows higher geometric than distributional or temporal agreement. On Neuropixels data, structural scores decline under cross-session and cross-laboratory transfer. Target-only probabilistic evaluation shows that stochastic draws improve energy scores over their mean predictions but retain calibration and temporal-dependence errors. A separate ten-seed matched-capacity conditioning control gives no clear structural benefit for the tested task covariates (Appendix B.4). The modalities expose distinct limitations in transfer and predictive variability.

3.3 Sensitivity to forecasting design and training choices

We next ask whether Nethobench can distinguish models trained on the same neural forecasting task but differing in prediction formulation, training regime, architecture, and loss design (Ben Taieb & Hyndman (2012); Benechehab et al. (2024); Bengio et al. (2015); Jiang et al. (2024); Le Guen & Thome (2019)). We compare eight forecasters: a linear autoregressive baseline (**VAR**); a recurrent neural network (**RNN**); a stochastic latent state-space model (**SSM**); two next-token Transformers (**1-step** and **Sequifier**); matched sequential Transformers trained with teacher forcing (**TF**) or self-feedback (**AR**); and a teacher-forced variant with additional tail-quantile and symmetric-KL losses (**TF_QL_KL**). A comparison of Transformers' forecasting regimes is given in Figure S12. After a 90-step observed context (3s), each model generates a 720-step forecast (24s) without access to recorded future activity, advancing the rollout using its own predictions or latent states. Each model is trained with three independent seeds and evaluated on the same 222 held-out sequences, partitioned into four evaluation groups. Scores are averaged across groups within each seed; we report means and standard deviations across 3 training seeds in Tables 6-7. Implementation details are provided in Appendix D.

Which properties distinguish architectures? SSM has the highest mean composite (0.602) and the highest distributional, temporal-spectral, geometric, and state dynamics means; TF has the highest relational mean (0.696; Figure 4a). RNN and 1-step have similar composites (0.544 and 0.541) but different state dynamics (0.613 versus 0.464) and relational means (0.496 versus 0.672), illustrating distinctions obscured by the composite alone. All model composites lie between the time-and-region shuffle (0.439) and real-real reference (0.830), although individual families do not necessarily follow this ordering. Despite similar pointwise-fidelity scores on the same forecasts and evaluation splits, models exhibit distinct structural profiles, revealing differences in population organization and dynamics that pointwise fidelity alone does not resolve.

Which properties change with formulation and loss? The five planned contrasts reveal family-specific differences across forecasting formulations and show that additional marginal-distribution losses need not improve the structural composite (Figure 4b). TF has higher temporal-spectral, relational, and state dynamics means than 1-step, with a slightly lower geometry score. Sequifier changes the next-token implementation, with a higher temporal-spectral mean but lower relational and geometric means than 1-step. Among the matched sequential models, TF has a higher composite than AR (0.576 versus 0.532), principally through temporal-spectral and relational differences. Additional tail-quantile and symmetric-KL losses do not yield a higher composite: TF_QL_KL scores 0.564 versus TF's 0.576. Thus, alignment of a training loss with selected marginal statistics does not ensure a higher structural composite. These patterns do not establish general superiority of an architecture, training regime, or loss.

a Model design produces distinct structural profiles

	Distribution	Temporal-spectral	Relational	Geometry	State dynamics	Composite
VAR	0.345	0.342	0.614	0.757	0.340	0.482
RNN	0.425	0.466	0.496	0.760	0.613	0.544
SSM	0.502	0.502	0.562	0.811	0.668	0.602
1-step	0.399	0.421	0.672	0.735	0.464	0.541
Sequifier	0.389	0.493	0.578	0.663	0.495	0.522
AR	0.437	0.403	0.604	0.714	0.500	0.532
TF	0.439	0.480	0.696	0.714	0.541	0.576
TFQLKL	0.438	0.473	0.679	0.678	0.544	0.564
Real–real reference	0.721	0.843	0.830	0.893	0.888	0.830
Time + region shuffle	0.562	0.278	0.396	0.440	0.503	0.439

b Planned contrasts: descriptive differences of means

	Distribution	Temporal–spectral	Relational	Geometry	State dynamics	Composite
TF − 1-step	+0.040	+0.059	+0.024	-0.021	+0.077	+0.035
TF − AR	+0.002	+0.077	+0.092	+0.000	+0.041	+0.044
Sequifier − 1-step	-0.010	+0.072	-0.094	-0.072	+0.031	-0.019
TFQLKL − TF	-0.001	-0.007	-0.017	-0.036	+0.003	-0.012
SSM − RNN	+0.077	+0.036	+0.066	+0.051	+0.055	+0.058

c Horizon-resolved profiles

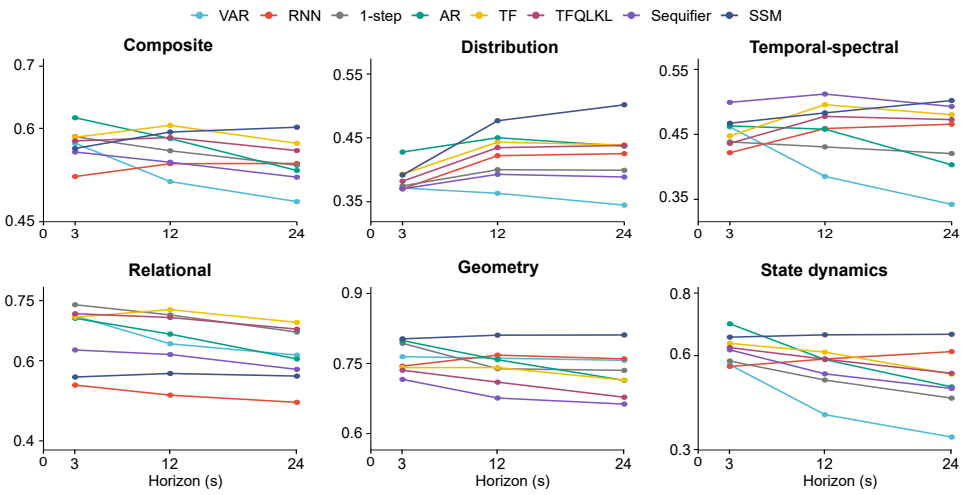


Figure 4: **Forecasting-design contrasts expose family-specific differences.** (a) Model family and composite means on the common 24-s forecast-only support. Real-real and time-and-region shuffle references use the same 222 targets and evaluation groups; reference means average four group scores within each of 20 random draws, then average across draws. (b) Five planned model contrasts, calculated from the rounded means in (a); positive differences favor the first model only on the indicated property. (c) Mean profiles at three forecast horizons. The 24-s means agree with the tables within displayed rounding. The complete 18-component comparison is given in Figure S23 and Tables 6-7 with SD across the 3 training seeds.

How do structural profiles change across rollout durations? Evaluation over cumulative 3-, 12-, and 24,s windows reveals model-specific changes in structural agreement (Figure 4c). VAR's temporal-spectral and state-dynamics means decline as the evaluated rollout expands, while AR begins among the stronger composites but falls below TF over the 24s window. SSM's distributional agreement rises, with comparatively stable geometry and state-dynamics means. Because the windows

are cumulative, these profiles characterize progressively longer generated trajectories; rising scores may partly reflect more stable estimation rather than improved late-horizon predictions.

Replicated training and model-quality controls. Structural agreement changes nonmonotonically across the sampled training checkpoints: epoch 100 exceeds epoch 10 in all 15 paired seeds, whereas epoch 30 exceeds epoch 100 in 14 of 15 (Figure S19; Appendix C.4).

4 DISCUSSION

Understanding what a neural forecasting model has learned requires examining the activity it generates at several levels. Nethobench complements existing neural prediction benchmarks by adding explicit diagnostics of population structure alongside conventional measures of predictive accuracy. Across our evaluations, forecasting design, rollout horizon and conditioning shape which structural properties are preserved, supporting evaluation against the requirements of a model's intended scientific use.

These profiles can inform model development. Matched-MSE controls distinguish structural failures that pointwise error cannot separate (Figure 2e); the Brain Treebank comparison shows that structural and pointwise criteria can favor different models (Figure3b-c); and nonmonotonic training trajectories show that additional training need not improve structural agreement (Figure S19e). Researchers could use these diagnostics to reject excessively smoothed forecasts when preserving variability is essential, select checkpoints on validation data for prespecified structural properties, or identify poorly preserved interregional dependencies. Such decisions should reflect the intended scientific use, since the highest composite need not identify the most suitable model for every purpose.

Structural agreement concerns selected summaries under a specified observation transform and evaluation window. It does not establish recovery of a generating process, appropriate uncertainty or causal mechanisms. Brain Treebank retrieval illustrates this distinction: recorded continuations can preserve population structure without being appropriate futures for a given context. Pooled agreement can likewise conceal incorrect stimulus- or behavior-dependent responses. Proper probabilistic scores, conditional calibration and task-specific validation therefore remain necessary.

Methodological choices further delimit sensitivity. PCA-based components can miss weak but behaviorally relevant signals, and clustered states should not be equated with cognitive states. Comparisons across neural populations require additional alignment assumptions, while dependence on preprocessing and window length calls for experiment-specific real-real and shuffled references. Families are partly correlated, but their differing perturbation responses support diagnostic complementarity. Robustness within the tested settings does not establish uniquely optimal metrics or weights, or stable fine-grained model rankings.

For neural foundation models, matched evaluations across model sizes, pretraining datasets and adaptation strategies could reveal whether scaling and transfer preserve temporal dependence, population coordination and long-horizon variability on unseen animals or tasks. Evaluation within stimulus and behavioral conditions would distinguish conditional agreement from agreement driven by condition prevalence. Persistent family-specific deficits could also motivate targeted training objectives.

In systems neuroscience, Nethobench could help evaluate models intended as experimental surrogates by testing whether they reproduce measured changes in timescales, population coordination or transition structure under held-out behavioral or intervention conditions. Generated datasets could also support controlled tests of neural analysis methods, with independent checks that known effects are recovered. A key next step is to establish whether specific structural improvements predict better behavioral decoding or recovery of experimentally induced changes. This would connect the benchmark's diagnostic value to the question that motivates neural modeling: which aspects of brain activity must a model preserve to support reliable scientific inference?

AI USE STATEMENT

Generative AI tools assisted with analysis and plotting code, numerical consistency checks, figure preparation, and manuscript editing. Quantitative figures are based on recorded or simulated data; conceptual schematics are illustrative. AI assistance did not replace verification against the experimental records, and the authors are responsible for the final text, methods, and claims.

ETHICS STATEMENT

The benchmark evaluates processed neural recordings and secondary analyses of public datasets. Human analyses use the de-identified Brain Treebank release and do not attempt participant identification. Structural scores are not measures of clinical validity, cognition, or causal mechanism, and the evaluated models are not validated for clinical decision-making. Dataset access and reuse remain subject to the originating datasets' terms and governance. Reporting empirical references, component availability, and failure cases reduces the risk of treating a single benchmark score as evidence of biological equivalence.

REPRODUCIBILITY STATEMENT

Appendices B–E specify dataset splits, preprocessing, model objectives, score transforms, component availability, and uncertainty units. The figures retain model- and seed-level controls alongside aggregate scores. We identify the evaluation implementation and distinguish fixed-checkpoint comparisons, training replications, corruption realizations, and participant-level inference; these sources of variation are not pooled into a single uncertainty estimate.

REFERENCES

Alexandre Andre, Shivashriganesh P. Mahato, Vinam Arora, Keshav Balaji, Divyansha Lachi, Nanda H. Krishna, Jingyun Xiao, Yizi Zhang, Ximeng Mao, Wenrui Ma, Han Yu, International Brain Laboratory, Daniel Birman, Niccolò Bonacchi, Gaelle A. Chapuis, Joana A. Catarino, Felicia Davatolhagh, Mayo Faulkner, Laura Freitas-Silva, Fei Hu, Julia M. Huntenburg, Anup Khanal, Inês Laranjeira, Petrina Lau, Guido T. Meijer, Nathaniel J. Miska, Jean-Paul Noel, Alejandro Pan-Vazquez, Georg Raiser, Cyrille Rossant, Karolina Z. Socha, Anne E. Urai, Miles J. Wells, Steven J. West, Olivier Winter, Blake Richards, Guillaume Lajoie, Cole Hurwitz, Mehdi Azabou, Matthew R. Whiteway, Liam Paninski, and Eva L. Dyer. Brainwidebench: Benchmarking large-scale pretraining and across-animal transfer in multi-region neural recordings, 2026. URL https://arxiv.org/abs/2609.22064.

Antonis Antoniades, Yiyi Yu, Joseph Canzano, William Wang, and Spencer LaVere Smith. Neuroformer: Multimodal and multitask generative pretraining for brain data, 2024. URL https://arxiv.org/abs/2311.00136.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael J. Mendelson, Blake Richards, Matthew G. Perich, Guillaume Lajoie, and Eva L. Dyer. A unified, scalable framework for neural population decoding, 2023. URL https://arxiv.org/abs/2310.16046.

Hubert Banville, Stéphane d'Ascoli, Simon Dahan, Jérémy Rapin, Marlène Careil, Yohann Benchetrit, Jarod Lévy, Saarang Panchavati, Antoine Ratouchniak, Mingfang, Zhang, Elisa Cascardi, Katelyn Begany, Teon Brooks, and Jean-Rémi King. Neuralbench: A unifying framework to benchmark neuroai models, 2026. URL https://arxiv.org/abs/2605.08495.

Souhaib Ben Taieb and Rob J. Hyndman. Recursive and direct multi-step forecasting: the best of both worlds. Working Paper 19/12, Department of Econometrics and Business Statistics, Monash University, 2012. URL https://robjhyndman.com/papers/rectify.pdf.

A. Benechehab, G. Paolo, A. Thomas, M. Filippone, and B. Kégl. Multi-timestep models for model-based reinforcement learning, 2024. URL https://openreview.net/forum?id=Rh4DmXaf8R. Preprint.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *CoRR*, abs/1506.03099, 2015. URL http://arxiv.org/abs/1506.03099.

György Buzsáki and Kenji Mizuseki. The log-dynamic brain: how skewed distributions affect network operations. *Nature Reviews Neuroscience*, 2014. doi: 10.1038/nrn3687. URL https://doi.org/10.1038/nrn3687.

Salvatore Calcagno, Isaak Kavasidis, Simone Palazzo, Marco Brondi, Luca Sità, Giacomo Turri, Daniela Giordano, Vladimir R. Kostic, Tommaso Fellin, Massimiliano Pontil, and Concetto Spampinato. Quantformer: Learning to quantize for neural activity forecasting in mouse visual cortex, 2024. URL https://arxiv.org/abs/2412.07264.

Marlene R Cohen and Adam Kohn. Measuring and interpreting neuronal correlations. *Nature Neuroscience*, 2011. doi: 10.1038/nn.2842. URL https://doi.org/10.1038/nn.2842.

John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17:1500–1509, 2014.

Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Teon Brooks, Katelyn Begany, Joséphine Raugel, Hubert Banville, and Jean-Rémi King. A foundation model of vision, audition, and language for in-silico neuroscience, 2026. URL https://arxiv.org/abs/2605.04326.

Emily Jane Dennis, Ahmed El Hady, Angie Michaiel, Ann Clemens, Dougal R. Gowan Tervo, Jakob Voigts, and Sandeep Robert Datta. Systems neuroscience of natural behaviors in rodents. *Journal of Neuroscience*, 41(5):911–919, 2021. doi: 10.1523/JNEUROSCI.1877-20.2020.

Yu Duan, Hamza Tahir Chaudhry, Misha B. Ahrens, Christopher D Harvey, Matthew G Perich, Karl Deisseroth, and Kanaka Rajan. Poco: Scalable neural forecasting through population conditioning, 2025. URL https://arxiv.org/abs/2506.14957.

Andrea Giovannucci, Johannes Friedrich, Pat Gunn, Jérémie Kalfon, Brandon L Brown, Sue Ann Koay, Jiannis Taxidis, Farzaneh Najafi, Jeffrey L Gauthier, Pengcheng Zhou, Baljit S Khakh, David W Tank, Dmitri B Chklovskii, and Eftychios A Pnevmatikakis. Caiman an open source tool for scalable calcium imaging data analysis. *eLife*, 8:e38173, jan 2019. ISSN 2050-084X. doi: 10.7554/eLife.38173. URL https://doi.org/10.7554/eLife.38173.

Joshua I. Glaser, Ari S. Benjamin, Raeed H. Chowdhury, Matthew G. Perich, Lee E. Miller, and Konrad P. Kording. Machine learning for neural decoding, 2020. URL https://arxiv.org/abs/1708.00909.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

International Brain Laboratory. Reproducibility of in vivo electrophysiological measurements in mice. *eLife*, 13:RP100840, 2024. doi: 10.7554/eLife.100840.

Ruoxi Jiang, Peter Y. Lu, Elena Orlova, and Rebecca Willett. Training neural operators to preserve invariant measures of chaotic attractors, 2024. URL https://arxiv.org/abs/2306.01187.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Nikolaus Kriegeskorte and Pamela K. Douglas. Interpreting encoding and decoding models, 2019. URL https://arxiv.org/abs/1812.00278.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Vincent Le Guen and Nicolas Thome. Shape and time distortion loss for training deep time series forecasting models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/466accbac9a66b805ba50e42ad715740-Paper.pdf.

Bryan M. Li, Theoklitos Amvrosiadis, Nathalie Rochefort, and Arno Onken. Synthesising realistic calcium traces of neuronal populations using GAN, 2020. URL https://arxiv.org/abs/2009.02707.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Jan-Matthis Lueckmann, Alexander Immer, Alex Bo-Yuan Chen, Peter H. Li, Mariela D. Petkova, Nirmala A. Iyer, Luuk Willem Hesselink, Aparna Dev, Gudrun Ihrke, Woohyun Park, Alyson Petruncio, Aubrey Weigel, Wyatt Korff, Florian Engert, Jeff W. Lichtman, Misha B. Ahrens, Michał Januszewski, and Viren Jain. Zapbench: A benchmark for whole-brain activity prediction in zebrafish, 2025. URL https://arxiv.org/abs/2503.02618.

Leon Luithlen. sequifier - causal transformer models for multivariate sequence modelling, 2025. URL https://github.com/0xideas/sequifier.

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, Berlin, Heidelberg, 2005. doi: 10.1007/978-3-540-27752-1.

Jason Manley, Sihao Lu, Kevin Barber, Jeffrey Demas, Hyewon Kim, David Meyer, Francisca Martínez Traub, and Alipasha Vaziri. Simultaneous, cortex-wide dynamics of up to 1 million neurons reveal unbounded scaling of dimensionality with neuron number. *Neuron*, 112(10): 1694–1709.e5, 2024. doi: 10.1016/j.neuron.2024.02.011.

Mackenzie Weygandt Mathis, Adriana Perez Rotondo, Edward F. Chang, Andreas S. Tolias, and Alexander Mathis. Decoding the brain: From neural representations to mechanistic models. *Cell*, 187(21):5814–5832, 2024. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2024.08.051. URL https://www.sciencedirect.com/science/article/pii/S0092867424009802.

Ofer Mazor and Gilles Laurent. Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. *Neuron*, 48(4):661–673, 2005. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2005.09.032. URL https://www.sciencedirect.com/science/article/pii/S0896627305008925.

Luca Mazzucato, Alfredo Fontanini, and Giancarlo La Camera. Dynamics of multistable states during ongoing and evoked cortical activity. *Journal of Neuroscience*, 35(21):8214–8231, 2015. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.4819-14.2015. URL https://www.jneurosci.org/content/35/21/8214.

John D Murray, Alberto Bernacchia, David J Freedman, Ranulfo Romo, Jonathan D Wallis, Xinying Cai, Camillo Padoa-Schioppa, Tatiana Pasternak, Hyojung Seo, Daeyeol Lee, and Xiao-Jing Wang. A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience*, 17:1661–1663, 2014. doi: 10.1038/nn.3862. URL https://doi.org/10.1038/nn.3862.

Simon Musall, Matthew T Kaufman, Ashley L Juavinett, Steven Gluf, and Anne K Churchland. Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, 22: 1677–1686, 2019.

Amin Nejatbakhsh, Victor Geadah, Alex H. Williams, and David Lipshutz. Comparing noisy neural population dynamics using optimal transport distances, 2024. URL https://arxiv.org/abs/2412.14421.

Mitchell Ostrow, Adam Eisen, Leo Kozachkov, and Ila Fiete. Beyond geometry: Comparing the temporal structure of computation in neural circuits with dynamical similarity analysis, 2023. URL https://arxiv.org/abs/2306.10168.

Felix Pei, Joel Ye, David Zoltowski, Anqi Wu, Raeed H. Chowdhury, Hansem Sohn, Joseph E. O'Doherty, Krishna V. Shenoy, Matthew T. Kaufman, Mark Churchland, Mehrdad Jazayeri, Lee E. Miller, Jonathan Pillow, Il Memming Park, Eva L. Dyer, and Chethan Pandarinath. Neural latents benchmark '21: Evaluating latent variable models of neural population activity, 2022. URL https://arxiv.org/abs/2109.04463.

Michael Scheuerer and Thomas M. Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015. doi: 10.1175/MWR-D-14-00269.1.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 2020. doi: 10.1101/407007. URL https://www.biorxiv.org/content/early/2020/01/02/407007.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):eaav7893, 2019. doi: 10.1126/science.aav7893.

David Sussillo and Omri Barak. Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Computation*, 25(3):626–649, 03 2013. ISSN 0899-7667. doi: 10.1162/NECO_a_00409. URL https://doi.org/10.1162/NECO_a_00409.

Polina Turishcheva, Paul G. Fahey, Laura Hansel, Rachel Froebe, Kayla Ponder, Michaela Vystrčilová, Konstantin F. Willeke, Mohammad Bashiri, Eric Wang, Zhiwei Ding, Andreas S. Tolias, Fabian H. Sinz, and Alexander S. Ecker. The dynamic sensorium competition for predicting large-scale mouse visual cortex activity from videos, 2024a. URL https://arxiv.org/abs/2305.19654.

Polina Turishcheva, Paul G. Fahey, Michaela Vystrčilová, Laura Hansel, Rachel E Froebe, Kayla Ponder, Yongrong Qiu, Konstantin Friedrich Willeke, Mohammad Bashiri, Ruslan Baikulov, Yu Zhu, Lei Ma, Shan Yu, Tiejun Huang, Bryan M. Li, Wolf De Wulf, Nina Kudryashova, Matthias H. Hennig, Nathalie Rochefort, Arno Onken, Eric Wang, Zhiwei Ding, Andreas S. Tolias, Fabian H. Sinz, and Alexander S Ecker. Retrospective for the dynamic sensorium competition for predicting large-scale mouse primary visual cortex activity from videos. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b. URL https://openreview.net/forum?id=gViJjwRUlM.

Christopher Versteeg, Jonathan D. McCart, Mitchell Ostrow, David M. Zoltowski, Clayton B. Washington, Laura Driscoll, Olivier Codol, Jonathan A. Michaels, Scott W. Linderman, David Sussillo, and Chethan Pandarinath. Computation-through-dynamics benchmark: Simulated datasets and quality metrics for dynamical models of neural activity. *bioRxiv*, 2025. doi: 10.1101/2025.02.07.637062. URL https://www.biorxiv.org/content/early/2025/02/20/2025.02.07.637062.

Christopher Wang, Adam Yaari, Aaditya K Singh, Vighnesh Subramaniam, Dana Rosenfarb, Jan DeWitt, Pranav Misra, Joseph R Madsen, Scellig Stone, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brain treebank: Large-scale intracranial recordings from naturalistic language stimuli. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 96505–96540. Curran Associates, Inc., 2024. doi: 10.52202/079017-3060.

Eric Y. Wang, Paul G. Fahey, Zhuokun Ding, Stelios Papadopoulos, Kayla Ponder, Marissa A. Weis, Andersen Chang, Taliah Muhammad, Saumil Patel, Zhiwei Ding, Dat Tran, Jiakun Fu, Casey M. Schneider-Mizell, MICrONS Consortium, R. Clay Reid, Forrest Collman, Nuno Maçarico da Costa, Katrin Franke, Alexander S. Ecker, Jacob Reimer, Xaq Pitkow, Fabian H. Sinz, and Andreas S. Tolias. Foundation model of neural activity predicts response to new stimulus types. *Nature*, 640: 470–477, 04 2025. doi: 10.1038/s41586-025-08829-y.

Konstantin F. Willeke, Polina Turishcheva, Alex Gilbert, Goirik Chakrabarty, Hasan A. Bedel, Paul G. Fahey, Yongrong Qiu, Marissa A. Weis, Michaela Vystrčilová, Taliah Muhammad, Lydia Ntanavara, Rachel E. Froebe, Kayla Ponder, Zheng Huan Tan, Emin Orhan, Erick Cobos, Sophia

Sanborn, Katrin Franke, Fabian H. Sinz, Alexander S. Ecker, and Andreas S. Tolias. Omnimouse: Scaling properties of multi-modal, multi-task brain models on 150b neural tokens, 2026. URL https://arxiv.org/abs/2604.18827.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations, 2022. URL https://arxiv.org/abs/2110.14739.

APPENDIX

- Appendix A: Related Work
- Appendix B: Experimental Set Up and Datasets
- Appendix C: Validation and robustness analyses
- Appendix D: Models
- Appendix E: Code, Data and Reproducibility
- Appendix F: Supplementary Figures

A RELATED WORK

Predictive modeling and neural forecasting. Predictive modeling is central to systems and computational neuroscience. In encoding models, neural responses are predicted from external variables such as stimuli, actions, or task variables; in decoding models, stimuli, behavior, or latent variables are predicted from neural population activity (Kriegeskorte & Douglas (2019); Glaser et al. (2020); Mathis et al. (2024)). These approaches test what information neural populations carry, but predictive success does not imply recovery of the mechanisms or dynamics that generated the activity. Recent work has extended this paradigm in scale and scope, from large-scale neural population decoders to foundation-style models that predict brain responses under naturalistic multimodal stimulation (Azabou et al. (2023); Wang et al. (2025); d'Ascoli et al. (2026)).

Neural forecasting is a narrower predictive setting in which the target is future neural activity itself, generated from recent neural history, and potentially, stimulus and behavioral variables (Antoniades et al. (2024); Calcagno et al. (2024); Duan et al. (2025)). When recordings are viewed as samples from a stochastic dynamical process, forecasting models can therefore serve as generative surrogates of neural dynamics.

Neural benchmarks and the scope of Nethobench. Existing neural benchmarks - including Brain-Score, Sensorium, Dynamic Sensorium, the Neural Latents Benchmark, and NeuralBench - standardize the evaluation of models predicting held-out neural, behavioral, or stimulus variables (Schrimpf et al. (2020); Turishcheva et al. (2024a;b); Pei et al. (2022); Banville et al. (2026)), but they are not designed to assess the long-horizon structure of generated neural activity. Recent forecasting benchmarks - such as ZAPBench, PoCo, and BrainWideBench (Lueckmann et al. (2025); Duan et al. (2025); Andre et al. (2026)) - move closer to this setting, while emphasizing pointwise measures such as correlation and mean squared error, or probabilistic measures such as likelihood. Pointwise metrics require trajectory alignment, whereas probabilistic measures evaluate the predicted distribution. But neither provides a family-resolved account of which structural properties are preserved.

Complementary methods treat neural activity as a representation or dynamical system. RSA, CKA, and generalized shape metrics compare representational geometry under different invariances ((Kriegeskorte et al., 2008; Kornblith et al., 2019; Williams et al., 2022)). Dynamical Similarity Analysis more directly compares the temporal organization of computation across neural systems ((Ostrow et al., 2023)), while optimal-transport approaches compare noisy, stochastic neural population dynamics without reducing them to deterministic trajectories ((Nejatbakhsh et al., 2024)). This system-level perspective has also begun to enter benchmarking. The Computation-through-Dynamics Benchmark similarly shows in controlled settings that reconstruction accuracy need not imply recovery of the correct dynamics (Versteeg et al. (2025)).

Nethobench is designed for this complementary evaluation problem. Its scope is not all neural modeling, and not all predictive modeling. It specifically targets generative neural forecasting models, evaluating them over long rollout horizons. Within this scope, Nethobench asks which structural properties of the reference neural process are preserved by the generated activity.

B Experimental Set Up and Datasets

B.1 Widefield calcium forecasting

The processed widefield dataset contains 287 sessions of approximately 21 minutes from six GCaMP6 transgenic male mice (Ai93 and Ai94), totaling approximately 101 hours. During the recordings, the head-fixed mice performed an uncued, self-initiated lever-pull task for a water reward, with a 3-second unrewarded lockout period imposed after each successful pull. Neural activity was captured via widefield calcium imaging through a chronic transcranial window at 30 Hz with a spatial resolution of 68 μm/pixel.

Following rigid alignment to a 2D projection of the Allen Common Coordinate Framework v3, locaNMF was applied to decompose the imaging data into anatomical components. Consequently, activity is summarized over 16 cortical regions at 30 Hz. We use these regional calcium time series rather than raw images; the benchmark does not test image segmentation, anatomical registration, or spike inference. The processed release is identified as `Nethobench/nethobench-widefield-v1` and distributed under CC-BY-4.0. Recording identifiers define the data partitions.

We split the data at the original recording-sequence level: all subsequences derived from a given recording are assigned to exactly one split. Formally, letting $s(i)$ denote the parent recording sequence of example i, we enforce

$$s(i) \neq s(j) \quad \text{whenever } i \in \text{train},\ j \in \text{val/test}.$$

This prevents temporally adjacent subsequences from the same recording from appearing in different splits.

Training examples pair 90 context samples with a 90-sample target block. Non-overlapping blocks are constructed within recordings, and longer evaluation targets concatenate consecutive blocks without crossing recording boundaries.Channel normalization uses training-context mean and standard deviation, with numerical stabilization 10^{-6}, and is applied unchanged to evaluation data. The principal comparison scores only the 720 generated samples following 90 observed samples: 24 s of forecast from 3 s of context. For each model and forecast horizon, scores are computed separately on four disjoint groups of evaluation sequences and averaged within each training seed. We report the mean and sample standard deviation across three training-seed means. This SD describes training variability on the fixed evaluation set, rather than variability across animals or evaluation groups. Horizon analyses score the first 3, 12, and 24 seconds of generated activity, excluding the observed context.

Widefield references use the same 222 evaluation targets as the model comparison, excluding context and retaining the full 720-bin forecast window. Each of 20 random draws is scored jointly within four fixed evaluation groups of 56/56/55/55 sequences, using the same evaluator as the model comparisons. Real-real companions are finite 720-bin windows sampled from the same recording without overlapping their paired target window. Temporal shuffling independently permutes time within each channel, whereas pooled shuffling permutes all time-channel entries within each sequence. Scores are averaged equally across the four groups and then across draws (Figure S11). Their values calibrate this observation scale and support; they are not attainable upper and lower bounds on all forecasts. A separate five-checkpoint robustness panel uses 100 sequences of 630 bins and includes context in its evaluation tensor (Appendix C.3).

Evaluation records and provenance. The evaluation pipeline saves component, family, composite, and pointwise-fidelity scores for each model, training seed, and evaluation split, with separate records for each forecast horizon. Structural and pointwise scores use the same prediction–reference arrays and evaluation groups. JSON caches and CSV exports retain the individual scores, selected sequence indices, and aggregation summaries. Reported variability is the sample SD across training-seed means after averaging the four evaluation groups within each seed; within-seed split variability is exported separately.

B.2 OmniMouse conditional neural prediction.

We evaluate the public OmniMouse-80M checkpoint (Willeke et al., 2026) on one held-out naturalistic-video session containing 128 evaluated neurons. Four configurations retain neural history, video

Table 1: **OmniMouse modality-conditioning configurations.** Inputs refer to the first model call.

Manuscript configuration	Saved strategy	Neural input	Video processing	Behavior input
Neural history + video + behavior	`causal_response_video_behavior`	Samples 0–24	Full 60-frame clip encoded; subsequent feature selection configured as frames 30–59	All 40 samples
Neural history + behavior	`causal_response_behavior`	Samples 0–24	None	All 40 samples
Video + behavior	`video_behavior`	None	Full 60-frame clip encoded; full-clip features retained	All 40 samples
Video only	`stimulus_driven`	None	Full 60-frame clip encoded; full-clip features retained	None

and behavior; neural history and behavior; video and behavior; or video alone. Behavioral variables condition predicted neural activity; behavioral outputs are not scored. All configurations use the same targets, neuron panel, preprocessing and four non-overlapping sequence-level evaluation splits.

Input timing and prediction task. We reconstruct the inference protocol from the published repository. Let $t = 0$ denote the beginning of the neural target interval. Each model call use a two-second input window, nominally spanning $[-1, 1)$ s, with neural responses and stimulus video sampled at 30 Hz and behavioral covariates at 20 Hz. Neural targets are input-window samples 30-59, corresponding to $[0, 1)$ s. For the two configurations using neural history, only samples 0-24 are encoded, with sample times from -1 to -0.2 s. The remaining configurations receive no neural response values as context. These timings describe the configured sampling grid.

When behavior conditioning is enabled, all 40 behavioral samples from the two-second window are encoded, including observations during the neural target interval. When video conditioning is enabled, the complete 60-frame clip enters the video encoder. The `causal_response_video_behavior` strategy subsequently selects video features using the configured frame range 30-59, whereas `video_behavior` and `stimulus_driven` retains features for the full clip (Table 1). Because selection occurs after video encoding, the configured range does not restrict the encoder's raw input to target-interval frames. Accordingly, "causal' refers to the neural response prefix and does not imply past-only availability of every conditioning modality.

Two-step prediction. The primary protocol (Figure S10b) uses a three-second source window nominally spanning $[-1, 2)$ s, through two successive model calls. The first call predicts neural responses over $[0, 1)$ s. The second shifts the input window forward by one second and predicts $[1, 2)$ s. For neural-history configurations, the selected population's neural context is replaced with first-call predictions, of which the first 25 samples form the encoded prefix. Video and behavioral inputs remain recorded observations in each shifted window. Thus, prediction is recursive in neural responses only for the neural-history configurations, while enabled covariates remain observed throughout the predicted interval.

Scoring support. The primary protocol scores the concatenated 60 predicted samples against the corresponding recorded activity, excluding the initial recorded context. The separate native-horizon protocol (Figure S10a) concatenates 30 recorded context samples with one 30-sample prediction and compares this with the same recorded context followed by the recorded target. Both protocols therefore score 60-sample sequences. This support does not meet the default tail-quantile component's 80-sample requirement, so available-component aggregation must be interpreted with this limitation.

Tables 4 and 5 retain the supplied aggregate means and SEMs across the four splits. The real-real reference uses distinct recorded sequences; temporal, time-and-region and factorized controls retain different amounts of structure. They cannot be interpreted as interchangeable null hypotheses. Generalization across sessions, mice, behavioral variables or conditioning protocols is not established by this single-session comparison.

B.3 BRAIN TREEBANK HUMAN INTRACRANIAL RECORDINGS

Cohort and observation transform. We use Brain Treebank (Wang et al., 2024), NEMAR `nm000253`, version 1.0.0 (DOI: `10.82901/nemar.nm000253`; metadata commit `7fe9b07131937b08ab07c211ec9ed08948571a84`). The selection includes participants with at least three available runs, taking the first three in acquisition order: participants 01–04, runs 01/02/03, and participant 06, runs 01/02/05. These are 15 participant–movie runs totaling 32.1 hours, not a claim of 15 distinct movies.

Contacts must have BIDS-good status in all three runs. Before common-average referencing, gross quality control excludes contacts whose pooled first-difference median absolute deviation is above 20 times or below 1/20 of the within-participant candidate median. The calculation samples differences at a stride of 2048 original samples. We retain 64 contacts by evenly spaced acquisition-order indices. This contact panel and gross QC use all three runs, including the held-out run; they are not a completely held-out-blind preprocessing pipeline. They do not use forecasting errors or retrieval outcomes.

The 2048 Hz signals undergo common-average referencing, fourth-order zero-phase Butterworth filtering in 70–115 and 125–150 Hz bands, and Hilbert-power extraction. Band powers are averaged, resampled to 16 Hz, and log-transformed. Each contact is normalized by its training-movie median and 1.4826 times its training-movie MAD, and clipped to $[-8, 8]$. A sample with log power below -20 in more than 25% of selected contacts marks a dropout; any context-plus-future window containing it is excluded. Normalization and training PCA use unflagged training samples. Zero-phase filtering is an offline observation transform, so the forecasts are not a prospective causal signal-processing benchmark.

Splits, retrieval and prediction. Within each participant, one movie is held out and the other two provide training data. Each held-out run contributes 50 non-overlapping windows after excluding 60 s at each edge. A window contains 80 context samples (5 s) and 320 future samples (20 s), giving 750 evaluation windows overall. Persistence repeats the last observed vector. Ridge VAR uses 16 lags, ridge coefficients from $\{0.1, 1, 10\}$, and a maximum fitted spectral radius of 0.995. Analog retrieval projects observed contexts into a 16-component whitened PCA representation fitted on training movies, retrieves the nearest training context, and outputs its recorded continuation. Both the 5-s context and the 20-s continuation in the library come exclusively from the two training movies. The held-out movie and future do not enter the library or nearest-context selection. This control tests whether a training continuation passes structural checks; it is not evidence of conditional correctness.

The GRUs have two layers, 128 hidden units and dropout 0.1. Training uses seeds 17, 29 and 43, batch size 64, learning rate 5×10^{-4}, weight decay 10^{-4}, gradient clipping at 1, a maximum of 60 epochs, and early-stopping patience 8. Candidate windows advance by 16 samples. Each epoch uses 4096 training windows; validation uses 512 windows from a 10% portion of each training movie. Training/validation window boundaries are not purged of all potential overlap; this affects the independence of checkpoint selection within training movies, while the third movie remains excluded from fitting. The multistep objective supervises horizons 1, 4, 16 and 64; the one-step model uses the next observation. Both generate autonomously at evaluation.

Scoring and inference. Temporal metrics use 16 Hz. Each deterministic method contributes one evaluation per fold, while each GRU contributes three training-seed evaluations: 135 model/fold/seed rows. Scores average seeds within movie, movies within participant, and then the five participants equally. Persistence produces an undefined ACF score in 15 rows; removing ACF from every method preserves model ordering and gives a multistep-minus-one-step composite difference of approximately 0.0280. The real–real reference averages the two directions of comparing alternating halves of the 50 evaluation windows (25 versus 25); its mean composite is 0.767.

For the primary GRU contrast, the participant is the independent unit. The paired hierarchical bootstrap resamples participants and then their movies, after averaging seeds within movie and keeping methods paired (10,000 replicates, seed 20260917). It yields $\Delta C = 0.0285$, 95% interval $[0.0180, 0.0415]$. Enumerating all $2^5 = 32$ participant sign assignments gives two-sided $p = 0.0625$. These summaries are both reported: a positive bootstrap interval does not override the resolution of the five-participant randomization test. Leave-one-participant-out analyses test dependence on individual participants, and the 45 movie/seed contrasts display within-participant variation.

Human corruption controls. For each held-out fold, four transformations are evaluated at $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$, giving $15 \times 4 \times 5 = 300$ evaluations. Additive noise uses $Y = X + \alpha\epsilon$, with standard-normal ϵ. Temporal block permutation divides each 320-bin future into ten 32-bin (2 s) blocks, selects a nested random-priority prefix of $\mathrm{round}(10\alpha)$ blocks, and cyclically remaps it jointly across contacts. Channel-wise circular shifts use fixed contact-specific base shifts scaled by α. Lower-variance attenuation uses training-fitted PCA and scales components 33–64 by $1 - \alpha$, retaining the other components; PCA is fitted on at most 100,000 valid training samples. The removed half accounts for 21.3–40.3% of training variance. It is not infinitesimal variance or independently identified task-relevant information. Severity zero is identity. One nested corruption realization per fold and type is held fixed across doses. Participant-mean dose correlations are descriptive associations across five doses, not independent tests at each point.

B.4 IBL Neuropixels population forecasts

Data and support. The repeated-site IBL cohort (International Brain Laboratory, 2024) comprises 17 sessions, four laboratories and 2176 stimulus-aligned trials. Twelve regional population-rate channels are binned at 20 ms, transformed as $\log(1 + \text{rate in Hz})$, and normalized using training-context statistics. Each sequence contains 25 context bins and 75 future bins (0.5 and 1.5 s). A region requires at least five quality-filtered units for the anatomical-coverage count. Missing regional channels are zero-filled before normalization in the stored tensors, so a common 12-channel shape does not imply equal regional coverage. The within-session, held-out-session and held-out-laboratory splits contain 436, 512 and 128 evaluated target contexts, respectively. They are fixed partitions, not independent draws of laboratory splits.

Forecast models and temporal controls. The population examples compare PSTH, VAR, latent-dynamical VAR and a regularized Transformer. The latter adds mean, 90th-quantile and lag-product penalties to MSE; the stored model identifier does not denote a negative-binomial likelihood. The GRU and latent SSM-VAE transfer comparisons retain training seeds 7, 11 and 19. Connected summary points score their ensemble-averaged predictions, rather than averaging the seed-level scores. These structural transfer panels score all 100 bins, including copied context. Pearson horizon comparisons and all probabilistic analyses instead exclude context. Their absolute scores are not directly comparable.

Order controls operate on the 436 within-session target tensors of shape 75×12 at 50 Hz. A partial shuffle selects $\mathrm{round}(75\alpha)$ bins per trial from a nested random order and permutes them jointly across channels. Selected positions may remain fixed by chance. Jitter applies contact-specific circular shifts bounded by $M(\alpha) = \mathrm{round}(\alpha\lfloor 75/8 \rfloor)$; shifts are $\mathrm{round}(UM)$ for fixed $U \sim \mathcal{U}[-1, 1]$ across doses. Maximum jitter is nine bins (180 ms). Error bars summarize three corruption RNG seeds (7, 11, 19), not three independently sampled animals or model fits.

Probabilistic evaluation. We evaluate $K = 18$ predictive draws per context on the target only. For target vector x and draws y_k, the empirical energy score is

$$\widehat{\mathrm{ES}} = \frac{1}{\sqrt{TR}} \left[K^{-1} \sum_k \|y_k - x\|_2 - (2K^2)^{-1} \sum_{k,l} \|y_k - y_l\|_2 \right].$$

The temporal variogram score (Scheuerer & Hamill, 2015) compares target increment magnitudes with ensemble-expected increment magnitudes over the specified within-channel lags (Appendix C.5). Lower values are better. An empirical marginal reference consists of 18 other real target sequences selected by distinct nonzero offsets, equally spaced as integer indices from 1 to $N - 1$, and added modulo N to the target index. A target never supplies its own reference, but reference sequences are reused between contexts. This is neither a deployable conditional predictor nor a mathematical lower bound.

Central-interval coverage and ranks among the 18 draws (19 possible ranks) pool dependent context/time/channel entries. No independent-sample confidence intervals are attached. The variance ratio divides ensemble variance by pooled target variance, which includes between-context differences; unity is not a conditional-calibration target. Structural composites use 18 Monte Carlo replicates, each taking one draw per corresponding context; points and sample SD quantify draw

variability, not training-seed uncertainty. The within-session spread control multiplies centered deviations by $0.5, 1, 2, 4$; energy and variogram values retain their own units and should only be compared within each metric.

Conditioning control. A separate ten-seed matched-capacity GRU experiment compares correct versus shuffled covariates: signed contrast, stimulus contrast, probability-left, previous choice and previous feedback. Current-trial choice, movement, feedback and reward are excluded. Complete covariate rows are shuffled within session and training/validation split; true held-out covariates are used for evaluation. The mean condition-stratified composite contrast is -0.000358 (paired 95% bootstrap interval $[-0.001922, 0.001317]$; exact two-sided sign-flip $p = 0.6914$, Holm-adjusted $p = 1$; four of ten seeds positive). This does not establish a benefit for this structural endpoint, nor a general absence of behavioral or functional information.

B.5 Two-photon calcium example

A separate single-mouse example uses two-photon calcium imaging through an implanted GRIN lens in a head-fixed mouse during exposure to appetitive rewards. The recording comprises calcium $\Delta F/F$ traces from 88 basolateral-amygdala principal neurons sampled at 10Hz for approximately 66 minutes, following CaImAn motion correction and source extraction (Giovannucci et al., 2019). These traces are not inferred spike counts. The population-conditioned MLP combines per-neuron temporal features and a population context embedding, with a residual output and learned persistence skip. The illustrated inference protocol applies one-step predictions autoregressively within chunks, then resets context to recorded activity (Figure S16). It therefore receives intermittent observations and is not comparable to a continuously autonomous rollout of the same total duration. The example illustrates a metric profile in another observation regime; it is not an independent multi-animal replication or a controlled architecture comparison with the widefield models.

B.6 Brain Treebank contact-selection and retrieval audit

Training-movie-only contact selection. We repeated the original gross-QC algorithm using only the two training movies in each held-out fold. The BIDS-good intersection, pooled first-difference MAD thresholds, and evenly spaced selection were recomputed independently for each fold. Sparse raw-signal differences used exactly the original stride of 2048 samples, resolutions, and acquisition-order indices; they were retained for reproduction. As a check on the implementation, pooling all three movies exactly reproduced the original 64 selected contacts for every participant.

Training-only QC selected the same 64 contacts in 14 of 15 folds. For participant P01 with movie 02 held out, the retained pool changed from 154 to 153 contacts, and the evenly spaced 64-contact panels shared 46 contacts. The other 14 selections were identical. The main analysis therefore retains its original offline participant-level panel: these checks do not retroactively make its preprocessing test-blind. A complete refit with the changed contact panel is required to replace the affected fold. In particular, changing contacts also changes common-average referencing and the fitted normalization and models; deleting a few output channels is not an equivalent correction.

Excluding the changed-QC fold and reaveraging movies within participants preserves the main dissociation: the analog retrieval control has the highest composite (0.7034), and GRU multistep has the lowest MAE (0.8871). The GRU multistep-minus-one-step contrast is 0.0255, positive in all five participants. This is an omission sensitivity using existing forecasts, not a contact-panel refit.

Programmatic retrieval exclusions. For all 15 folds we verified that the held-out movie is absent from the two training-run identifiers, every candidate context/continuation lies wholly within a permitted training run, and every recorded selected candidate belongs to that pool. All 750 stored analog predictions exactly equal their selected 320-bin training continuation. The selector's inputs are training runs, training quality masks, training indices, and observed contexts; neither the held-out future nor a held-out-movie library enters its PCA fit or nearest-context search. This audit checks stored matches, candidate bounds and the executed selector's data contract; it does not substitute a new nearest-neighbor search or demonstrate prospective preprocessing.

Offline observation transform. The Butterworth filters are applied in both temporal directions and the Hilbert transform has nonlocal support. Consequently, a fixed finite onset exclusion cannot certify causal signal construction. We retain this limitation even when examining forecasts away from their boundary. Prospective claims require a one-sided transform, fold-specific contact selection and model refitting under that transform. The present experiment supports an offline comparison between realization accuracy and structural agreement, not a validated online forecasting system.

Forecast-onset sensitivity on fixed predictions. We rescored all 135 saved model/seed/fold arrays after excluding the first 32 forecast bins (2 s), leaving 50 sequences of 288 bins and 64 contacts per array. The same pinned component definitions and 16-Hz sampling rate were used. Native errors were recomputed on this identical 18-s support, then training seeds were averaged within movie, movies within participant, and the five participants equally. Analog retrieval remains the structural-composite leader (0.6936; MAE 1.2870), while GRU multistep retains the lowest MAE (0.8923; composite 0.3125). These are descriptive within-support comparisons; the values are not directly comparable to full 20-s scores. Removing onset samples does not remove nonlocal preprocessing dependence and is not a substitute for a causal-transform refit.

C Validation and robustness analyses

C.1 Metric definitions and implementation

Inputs and aggregation. Reference and prediction tensors must be aligned to identical sequence, time and channel keys, with shape $M \times T \times R$. Unaligned samples are not silently truncated. Different components compare different summaries: some pool time and sequences, others first score individual sequences or channels. Let $s_j \in [0, 1]$ be a finite component similarity. Each family is the arithmetic mean of its available components, and the five families receive the fixed weights in Equation 1. There are four distributional, three temporal-spectral, four relational, two geometric and five state dynamics components. If a component is undefined, its within-family weight is redistributed over the finite components. If an entire family is absent, outer weights are likewise renormalized. Missing values are not zeros and component availability must accompany a result.

The implementation distinguishes a similarity from its associated raw discrepancy. Frequently used transforms are $q(d) = (1 + d)^{-1}$ for $d \geq 0$ and $c(a, b) = \mathrm{clip}((1 + \mathrm{corr}(a, b))/2, 0, 1)$. The latter is an affine, clipped Pearson score, not raw signed correlation or absolute correlation. For vectors, $r(a, b) = q(\mathrm{RMSE}(a, b)/\sigma_a)$, falling back from reference SD to mean absolute reference magnitude and then 1 for degenerate scales. A generic matrix comparison applies $0.6c + 0.4r$ to strictly upper-triangular entries, with available-term renormalization. These choices define numerical agreement rather than an identified generative likelihood.

Distributional components. *Symmetric KL:* shared histogram support is derived from pooled reference and prediction 0.001/0.999 quantiles, with 60 interior bins and two tail bins. Histograms use additive 10^{-12} stabilization. For each sequence/channel pair with at least ten finite samples,

$$d_{mr} = \tfrac{1}{2}\{D_{\mathrm{KL}}(p_{mr} \| q_{mr}) + D_{\mathrm{KL}}(q_{mr} \| p_{mr})\}.$$

Within a sequence, $q(d_{mr})$ values are geometrically averaged over channels. The final score is half their sequence mean plus half their tenth percentile. The retained divergence is symmetric KL; the generic implementation label "KL/JSD" does not mean that JSD was used.

Mean shift: absolute reference–prediction temporal mean differences are divided by the pooled reference channel IQR. The largest $\lceil 0.1R \rceil$ channel discrepancies are averaged per sequence, then across sequences, before applying q. *Tail quantiles:* 99 quantiles at levels 0.01–0.99 are computed, retaining levels at or below 0.10 and at or above 0.90. IQR-scaled absolute differences are averaged over retained levels, then over the largest $\lceil 0.25R \rceil$ channel discrepancies and sequences. The score is q of this distance. At least 80 samples per sequence/channel are required; long sequences can be subsampled to 1200 samples with fixed RNG seed 0.

Moments: pooled per-channel variance, skewness and excess kurtosis yield

$$d_r = \left| \log \frac{v_{Y,r} + \epsilon}{v_{X,r} + \epsilon} \right| + 0.5|\gamma_{Y,r} - \gamma_{X,r}| + 0.25|\kappa_{Y,r} - \kappa_{X,r}|.$$

The component averages $q(d_r)$ over finite channels, requiring at least 24 samples. Skewness and kurtosis use finite-sample bias correction. The retained pooled-moment score should not be confused with auxiliary weighted-moment visualizations or differentiable training losses.

Temporal-spectral components. The temporal family is the equal mean of trajectory, ACF and power-sensitive PSD agreement. *Trajectory* combines occupancy/velocity comparisons (weight 0.6) and path-feature comparisons (0.4) in reference-fitted PCA coordinates. Occupancy uses 12-bin axis histograms; speed and turning are compared through quantiles at 0.1, 0.3, 0.5, 0.7 and 0.9. Path summaries include path length, displacement, radius, displacement-to-path ratio and lag-one speed dependence. The PCA variance target is 85%, with component caps of three for occupancy and four for path features. These are selected trajectory statistics, not an exact path-distribution distance.

ACF compares both raw activity and first increments using relative ℓ_2 discrepancies of autocorrelation vectors. Physical delays correspond to $\{1, 2, 4, 8, 16, 32, 48\}/30$ seconds, converted to rounded, unique valid sample lags using each dataset's sampling frequency. Constant signals can make this component undefined. *PSD* uses within-sequence Welch density estimates with constant detrending and at most 256 samples per segment, excluding DC. With normalized non-DC spectra p, q and total powers P_X, P_Y, its agreement combines spectral overlap and power:

$$s_{\mathrm{PSD}} = \left(\sum_f \sqrt{p_f q_f} \right) \sqrt{\frac{\min(P_X, P_Y)}{\max(P_X, P_Y)}}.$$

Two zero-power spectra score 1 and a single zero-power spectrum scores 0. ACF and PSD aggregate finite sequence/channel scores with equal weights on their mean and tenth percentile. The separately plotted PSD-shape and directionality diagnostics are not additional retained components. Sampling rates are 30 Hz for widefield/OmniMouse, 16 Hz for Brain Treebank and 50 Hz for IBL; synthetic BOLD controls account for their fourfold downsampling.

Relational components. *Correlation graph* uses pooled Pearson correlations with zero diagonal. It combines Jaccard overlap of the strongest 20% of absolute edges, agreement in absolute-weight degree, and binary clustering-coefficient agreement by a geometric mean of the finite terms, clipped to $[10^{-6}, 1]$. Thus this score is not simply elementwise correlation-matrix error. *Cross-region mutual information* uses a five-nearest-neighbor estimator, averages the two regression directions and clips negative estimates to zero, with up to 1500 pooled samples and RNG seed 0. Its matrix comparison is $0.6c + 0.4r$.

Lagged covariance averages matrix agreements at lags 1, 2 and 4. *VAR-response proxy* compares ridge VAR(1) coefficient matrices with ridge 0.01. The latter is the operational "impulse response" component: it summarizes a fitted one-step linear response, not a multi-horizon impulse-response identification or causal influence estimate. For these additional matrix metrics, reference and prediction are each median/IQR scaled separately, with SD/unit-scale fallbacks. The implementation pools finite rows and applies lagged operations to the flattened data; transitions across sequence boundaries are therefore included in these particular covariance/VAR estimates. Upper-triangular extraction also retains only one orientation of their asymmetric matrices. These components must not be interpreted as complete directed coupling recovery. ACF, PSD and discrete-state transitions use their own sequence-aware computations.

Geometric components. *Manifold agreement* combines stratified persistence-lifetime agreement (0.75) with a local contribution (0.25), the latter combining pooled persistence and five-nearest-neighbor distance-profile agreement. Persistence is computed for H_0 and H_1 with a reference PCA cap of three dimensions and deterministic point subsampling; stratified sampling takes two points from up to 48 sequences. The persistence comparison uses lifetime-distribution summaries, not a full optimal matching of diagrams. *Subspace agreement* is the mean $\cos^2 \theta_i$ of principal angles between covariance eigenspaces. Each subspace aims to explain 85% of variance, capped at six dimensions, and both are truncated to their smaller rank. Covariances use shrinkage after separate robust scaling. High subspace agreement does not guarantee that all low-variance dimensions, manifold topology or task-relevant information are preserved. This limitation matters because cortical population dimensionality can continue to grow with the number of recorded neurons (Manley et al., 2024). Native-space temporal checks and the Brain Treebank lower-variance perturbation test provide

complementary evidence, without establishing that the reduced representations preserve all neural information.

State-dynamical components. A reference median/IQR scaler and three-dimensional reference PCA define a common representation. Mini-batch k-means centers are fitted only to reference points (at most 10,000 fit points, ten initializations, RNG seed equal to K); occupancy evaluation uses at most 30,000 points per dataset. The five components comprise occupancies at $K = 11, 12$ and transition histograms at $K = 11$, lags 1, 2 and 3. Transitions are counted within sequences. Each flattened count matrix is normalized by its total count; it is a joint transition histogram, not a row-normalized conditional transition matrix. For normalized histograms p, q,

$$s_{\text{hist}} = 0.40 \sum_i \min(p_i, q_i) + 0.25c(p, q) + 0.35r(p, q),$$

with available-term renormalization. The states are computational partitions, not annotated behavioral or latent biological states. For smooth activity, a transition may simply cross a cluster boundary; its interpretation depends on the embedding, state count and lag. The parameter controls quantify sensitivity to these choices without validating discrete biological states.

Sidecars and numerical scope. The displayed pointwise-fidelity sidecar combines the implementation's error and paired mutual-information scores with weights 0.65 and 0.35. It is excluded from the structural composite and is not a proper scoring rule. The main comparisons additionally use native MAE/MSE where available. The scorer is deterministic conditional on tensors, configuration and fixed subsampling seeds; this does not remove finite-window uncertainty or data-selection effects. Different components have distinct normalization and pooling conventions, so figures illustrating intermediate moments, correlations or spectra are diagnostic explanations rather than new component definitions.

C.2 Controlled synthetic and empirical validation

AR-based oracle and perturbations. The controlled generator uses a latent first-order autoregressive process and a calcium observation transform, with six latent dimensions, 16 observed regions, 100 sequences and 768 retained time bins after a 128-bin burn-in. Generator parameters are fixed by seed 11; the reference uses realization seed 101 and the 20 independent oracle realizations use seeds 501–520. Shared generator parameters do not imply identical latent paths. The overlapping curves in Figure 2a are a conceptual sketch, whereas the adjacent first-channel traces are measured simulations from reference 101 and oracle 501.

The perturbation inventory changes gain/bias; time constants/frequency; phase synchronization; observation kernels; combinations of temporal parameters; coupling; channel loadings; signs; relational combinations; subspace orientation; variance allocation; rank; and geometric combinations. Each of 13 classes is evaluated at five normalized doses using realization seed 701. Dose is normalized within a manipulation and is not a common physical unit across classes. The heatmap shows the mean oracle score minus the maximum-dose perturbed score. Negative differences are retained numerically; an independent oracle is not a strict upper bound. Dose-response bands are SD across the 13 classes, not confidence intervals over seeds. Mean monotonic degradation does not imply that each class decreases monotonically or affects a unique family.

Metric-level corruption screen. Figure S1a retains 34 specified corruptions and 16 component readouts, with the reference fixed while transforming a fixed prediction. Rows cover mean/bias, scale/tails, additive noise, temporal rearrangement, region remapping, spectral contamination and smoothing. For component m, corruption c and severity α, the displayed sensitivity is

$$\max_{\alpha} \frac{\max(S_m^{\text{base}} - S_{m,c,\alpha}, 0)}{S_m^{\text{base}} + \epsilon}.$$

This maximum-over-dose summary describes sensitivity, not specificity, a seed-averaged effect or an inference test. Its temporal column is trajectory agreement. ACF and PSD coverage is evaluated in the separate temporal-control panel, rather than inferred from the 16-column matrix.

Matched-error and temporal controls. Five widefield transformations—time-order change, channel shifts, smoothing, a PCA shortcut and state cycling—are amplitude-matched to MSE 0.04956. Family-profile differences are tested with 999 blocked permutations, yielding $p = 0.001$; paired evaluation-sequence bootstraps (200 replicates per corruption) characterize profile uncertainty. The separate 20-shuffle comparison permutes time independently within each region. The contraction control defines

$$Y_\rho = \bar{Y} + \sqrt{\rho}(Y - \bar{Y}), \qquad 0 \le \rho \le 1,$$

where means are taken over time within region/window. This preserves means while retaining fraction ρ of temporal variance. The plotted means and supplied 95% interval boundaries are preserved; they should not be interpreted as training-seed or animal-level uncertainty.

Dedicated temporal controls use 100 held-out widefield sequences of 630 bins and 16 regions, with transformations confined within sequences. Block shuffling, lag jitter, Gaussian smoothing and phase scrambling are evaluated over the full displayed dose grids, with 20 corruption realizations per setting; a fixed-dose time reversal is also retained. Larger blocks preserve more temporal order, so block length is not an increasing severity scale. A signed lagged-covariance directionality diagnostic remains outside the composite. Bands show the 10th–90th percentile over corruption realizations. Across 360 evaluations, removing PSD gives rank correlation 0.998 and mean absolute composite change 0.0046; removing ACF gives 0.983 and 0.0070. On 500 paired evaluation-sequence bootstrap rows from five fixed models, both rank correlations are 0.997, with mean absolute changes 0.0209 and 0.0206. Ablations average the two remaining temporal components equally and retain the temporal family weight of 0.18. Rank stability does not establish component redundancy or unchanged absolute scores.

Richer generators and the spiking–calcium control. Figure S3 evaluates nonlinear latent dynamics, switching linear dynamics, Lorenz dynamics, a low-rank recurrent network, and spiking–calcium observations. Five generator seeds (7, 11, 19, 23, 31) each yield eight sequences of 160 bins and 12 channels. The independent oracle is a new realization from the same generator. A noisy-reference control adds Gaussian noise with SD 0.05; it is not an empirical split-half comparison. All 18 retained components are finite, including trajectory, ACF and power-sensitive PSD.

The spiking–calcium comparison distinguishes an independent oracle, a pure shuffle, and an MSE-matched reference–shuffle mixture. For a channel-wise shuffle $\pi(X)$ across pooled sequence/time samples,

$$X_{\mathrm{mix}} = (1 - \alpha)X + \alpha\pi(X), \qquad \alpha = \sqrt{0.16/\mathrm{MSE}(\pi(X), X)}.$$

The implementation's clipping limits are inactive here. Across seeds, $\alpha = 0.2821$–0.2846, retaining 71.54–71.79% reference coefficient. Mean composites are 0.73005 for the mixture, 0.66640 for the oracle and 0.50360 for the pure shuffle, with MSE 0.160, 1.770 and 1.993. The mixture exceeds the oracle in distributional, temporal, relational and geometric agreement; the oracle exceeds it in state dynamics. Relational and geometric gains contribute 0.04900 to the total mixture-minus-oracle difference of 0.06365. The oracle exceeds the pure shuffle in all five seeds. This comparison explains the reference-retaining control's advantage; it does not identify a finite-sample mechanism or establish that a pointwise-close mixture is a viable forecast.

Synthetic BOLD observations. Nonlinear, switching and low-rank recurrent systems undergo hemodynamic-response convolution and fourfold downsampling. Five seeds, three systems and five controls yield 75 scores. Controls comprise repeated observations of the same neural realization with independent observation noise/drift, a noisy neural reference, an independent oracle, an MSE-matched transition mixture before the observation transform, and a pure BOLD time shuffle. Each seed is averaged equally over the three systems, and figures show the five seed means. The 40-bin sequences fail the quantile component's 80-sample threshold, so these scores use 17/18 components, redistributing only the within-distribution weight. All three temporal components remain included and the sampling-frequency parameter changes from 30 to 7.5. Mean composites are respectively 0.848, 0.848, 0.745, 0.778 and 0.500. Neither ordering nor its differences from the neural-domain panel isolate a causal effect of hemodynamic smoothing. These controls are synthetic observation-model tests, not validation on real fMRI.

C.3 METRIC-SELECTION AND ROBUSTNESS ANALYSES

Candidate inventory. The candidate audit contains 45 metric candidates and six criteria: oracle agreement, targeted-corruption selectivity, model utility, nonredundancy, coverage and interpretability. Eighteen candidates belong to the retained structural specification; complete criterion records are available for 16, while ACF and PSD have dedicated controls rather than complete screening records. Twenty-five candidates have an unmeasured nonredundancy entry that used a neutral 0.5 fallback in the rubric. These entries are masked, not treated as measured evidence, in Figure S4. Group summaries use available entries with criterion-specific denominators. Interpretability is a design judgment, and utility on a selection panel is not independent validation on that panel. Predictive MAE/correlation proxies address different targets and are not declared inferior because of a structural-screening rubric.

Parameter and family-weight sensitivity. Twelve alternative PCA, state-count and lag settings are compared on exactly 45 matched seed/checkpoint observations (15 training seeds at epochs 10, 30 and 100). Rank correlations always compare the same condition set. Separately, 5000 shared family-weight draws use RNG seed 20260918. Their mixture is 70% Dirichlet$(80w)$, 10% equal family weights, 10% uniformly selected single-family omission, and 10% uniformly selected single-family weight doubling, with renormalization; $w = (0.22, 0.18, 0.24, 0.18, 0.18)$. Each draw applies the same weights to every condition. Components, within-family averaging and score transforms are fixed.

The violin distributions in Figure S5 show, for each draw, the mean of 15 paired seed contrasts. Their central 95% ranges describe specification sensitivity, not sampling confidence intervals. Mean contrasts 100–10, 30–10 and 30–100 are positive for all draws. The first two are positive in every seed for all draws, but 30–100 is positive in every seed for only 0.0088 of draws. Thus stable mean ordering coexists with individual reversals and nonmonotonic training trajectories.

Dependence and omission. Family correlations are computed separately on the 45 training observations and 360 temporal-control observations. Both contain dependent repeated evaluations and support descriptive correlations rather than independent-replicate inference. Average-linkage clustering uses distance $1 - |\rho_{\mathrm{Spearman}}|$ among the 18 components. For a full composite C and a composite C_{-f} omitting one family with renormalized weights, the loss of corruption sensitivity is

$$(1 - C) - (1 - C_{-f}) = C_{-f} - C.$$

A positive value means omission makes the corruption appear less severe. This measures an aggregation change, not causal unique information, semantic decoding or incremental downstream utility.

Five-checkpoint panel and PCA choices. The fixed-checkpoint analysis aligns five Transformer forecasts to a common 100-sequence, 630-bin, 16-region tensor. This panel is distinct from the principal eight-model, forecast-only experiment. At state-representation PCA dimensions 2, 3, 5 and 8, rank correlations with the default are 0.90, 1.00, 0.90 and 1.00. The TF distributional-loss variant remains first, while 1-step and Sequifier exchange order at dimensions 2 and 5. Retained reference variance after median/IQR scaling is 81.0%, 88.4%, 94.9% and 98.4%. These values do not quantify retained task information. Replacing the temporal family by native-space $(\mathrm{ACF} + \mathrm{PSD})/2$ yields pooled $\rho = 0.9761$ over 500 rows from 100 paired sequence bootstraps and five models. These are evaluation resamples, not independent fits.

Empirical references and subsets. Twenty random 50/50 partitions of the same 100-sequence pool provide real–real references. Twenty transformations each provide independent within-region time permutation, joint time–region permutation, and regional marginal resampling with replacement. Reference scores compare 50 versus 50 sequences, whereas the five forecast markers score all 100; these are descriptive reference points rather than sample-size-matched normalized performances. Repeated partitions are not independent subjects.

All 31 nonempty family subsets retain relative default weights and renormalize. Across 100 paired sequence bootstraps, subset and full scores are compared on the same five models, producing 3100 subset/bootstrap rows. Pairwise agreement averages over all ten model pairs; a tie in only one

ordering counts as disagreement. The temporal family alone has median rank correlation 1.00, central bootstrap range $[0.90, 1.00]$, and mean pairwise agreement 0.971. One family can therefore recover coarse ordering on this panel. Multiple family profiles still distinguish structural failure types; the subset experiment establishes neither a universal minimum number of families nor individual indispensability. Best subsets are selected descriptively within this panel, without an independent selection/test split.

C.4 Replicated endpoints and training trajectories

The source-domain generator combines latent network dynamics and state switching, regional spiking with refractory effects and inter-region coupling, AR(2) calcium rise/decay filtering, and neuropil, observation-noise and baseline-drift terms. "Biophysical" names this artificial data-generating process, not the forecasting architecture. The forecasters use the Sequifier Transformer architecture. CalciumGAN is a WaveGAN-based model with a Wasserstein training objective for synthesizing population calcium traces (Li et al., 2020); here its generated traces supply the transfer domain. The endpoint comparison uses five paired runs (seeds 2001–2005) of a low-capacity model (approximately 11,000 parameters, 16-bin context, one layer, one epoch) and a higher-capacity model (approximately 1.58 million parameters, 128-bin context, six layers, validation-selected weights and last logged epoch labels 65–69). Capacity, context, depth and training co-vary. Both forecasts are rescored on exactly matching targets within domain: 96 source sequences at bins 128–1039 (912 generated bins) and 128 CalciumGAN-transfer sequences at bins 128–1023 (896 generated bins), retaining 16 regions. Context is excluded, although forecast origins differ with context length. CalciumGAN is a generated target domain, not another animal cohort.

Mean higher-minus-lower composite differences are 0.1317 in the source domain and 0.1206 in transfer. Unadjusted paired Student-t 95% intervals with four degrees of freedom are $[-0.0074, 0.2707]$ and $[0.0212, 0.2199]$; both exact two-sided sign-flip tests give $p = 0.125$. We do not infer uniform family improvement or isolate the effect of an individual design variable.

A separate 15-seed training experiment evaluates epochs 10, 30 and 100 on 45 validation sequences of 810 generated bins (Figure S19e). Mean composites are 0.4742, 0.6163 and 0.5680, respectively. Epoch 100 exceeds epoch 10 in all seeds, while epoch 30 exceeds epoch 100 in 14/15. The latter exact paired sign-flip $p = 0.000122$ becomes $p = 0.000183$ after Holm adjustment over the three checkpoint contrasts. Thus, structural agreement changes nonmonotonically during training; "peak" refers only to the sampled checkpoints.

A separate fixed-epoch comparison evaluates six forecasting configurations - TF, TF_QL_KL, AR, 1-step, RNN and SSM - with 3 training seeds per configuration (Figure S19f-g). All configurations are evaluated at epoch 150 on the same validation sequences. Panel f reports the Nethobench composite and panel g the normalized forecast MSE. Points represent individual training seeds, lines connect matching seed identifiers across configurations, and bars show the seed mean ± SEM. These fixed-epoch comparisons are distinct from the validation-selected checkpoints used in the principal widefield comparison. Model capacity and compute are not matched, and seed SEM describes training variability on the fixed evaluation set rather than animal or session uncertainty.

C.5 Proper scores and statistical conventions

Before probabilistic scoring, targets and draws are transformed by the same per-region target median and $1.4826\,\mathrm{MAD}$ (SD/unit-scale fallbacks). This evaluation normalization is shared across methods and does not fit a forecasting model. Energy distances are divided by $\sqrt{TR}$, with the pairwise draw term including the zero diagonal. The lag-restricted variogram uses $p = 0.5$, lags $L = \{1, 2, 4, 8, 16\}$ and normalized weights $w_\ell \propto 1/\ell$:

$$\widehat{\mathrm{VS}} = \sum_{\ell \in L} w_\ell \operatorname*{mean}_{t,r} \left[|x_{t+\ell,r} - x_{t,r}|^{1/2} - \frac{1}{K} \sum_k |y_{k,t+\ell,r} - y_{k,t,r}|^{1/2} \right]^2 .$$

Scores are averaged over contexts. We report the empirical finite-ensemble implementation rather than claiming an unbiased estimator of an infinite-ensemble score. Neither energy nor a lag-restricted variogram alone localizes every structural failure.

All uncertainty units are stated locally: training-seed SD after averaging evaluation groups for the principal widefield comparison; split SEM for OmniMouse; seed variation for replicated fits; corruption-realization ranges for perturbations; specification ranges for weight draws; and participant-level inference for Brain Treebank. Independent synthetic realizations are not biological replicates. Exact paired sign-flip tests enumerate the available seed/participant signs when feasible, and adjusted results specify the contrast family. We avoid interpreting raw differences between different metric scales as effect sizes or pooling the distinct experiments into a global significance claim.

C.6 WITHIN-SEQUENCE RELATIONAL AND CONDITIONAL-TRANSITION SENSITIVITY

The primary figures and Tables 6-7 use the implementation described in Appendix C.1. We separately evaluate within-sequence, directed relational comparisons on fixed synthetic controls (Figure S24). These variants were evaluated in supplementary sensitivity analyses; the principal empirical figures and tables retain the original implementation.

Sequence-safe directed relational comparisons. Reference and forecast are median/IQR scaled separately as in the pinned implementation. For lag ℓ, define valid within-sequence pairs

$$\mathcal{P}_\ell = \{(m,t) : 1 \le m \le M,\ 1 \le t \le T-\ell,\ Z_{m,t}, Z_{m,t+\ell} \text{ are finite}\}.$$

Let A_ℓ and B_ℓ contain the corresponding earlier and later row vectors. The corrected lagged covariance is

$$\widehat{\Gamma}_\ell = (A_\ell - \bar{A}_\ell)^\top (B_\ell - \bar{B}_\ell)/(|\mathcal{P}_\ell| - 1),$$

using lags $1, 2, 4$. The one-step VAR-response proxy is

$$\widehat{V} = \left[(A_1^\top A_1 + 0.01I)^{-1} A_1^\top B_1\right]^\top.$$

Both comparisons use every directed off-diagonal entry, $\mathrm{vec}_{i\neq j}(\cdot)$, with the same $0.6c + 0.4r$ similarity as the pinned component. No last-to-first transition connects different sequences. Tests verify pair counts, sequence-permutation invariance, channel-permutation invariance, and sensitivity to a change confined to the previously omitted triangle. The two old components were first reproduced from every saved tensor before their differences were computed.

The corrected relational variant changes the composite by at most 0.00485 in the neural controls and 0.01059 in simulated BOLD. It preserves candidate ordering in all 40 system-by-seed groups (Figure S24). These are finite-panel sensitivity results; they do not establish unchanged rankings for the principal widefield models or other datasets.

Rescoring frozen forecasts further preserves the mean composite ordering of the three tested widefield models and all Brain Treebank models, including the latter's structural-accuracy reversal (Figure S25).

Conditional state transitions. For each of the same reference-fitted $K = 11$ partitions and lags $1, 2, 3$, let N_{ij} count transitions within sequences. The pinned score normalizes all entries by $\sum_{ij} N_{ij}$ and therefore combines occupancy and transition information. The alternative compares conditional rows, $P_{ij} = N_{ij}/\sum_j N_{ij}$, on origin states with positive reference counts. It averages the existing histogram similarity uniformly over these states, retaining its overlap/correlation/RMSE terms. An origin state visited in the reference but never left in the prediction receives zero agreement; absent reference origin states are excluded. The two occupancy components are unchanged. Thus, the alternative reduces weighting by how often an origin state occurs, while still penalizing unsupported predicted states.

Conditional-row scoring changes neural composites by at most 0.01327 and BOLD composites by at most 0.01134, preserving ordering in 38 of 40 system-by-seed groups. Applying both variants yields maximum absolute changes 0.01375 and 0.00972, respectively. These checks expose a modeling choice rather than establish uniquely correct state partitions or independent families. All variant component scores, composites, support masks and ordering comparisons are retained with the analysis.

D MODELS

We compare linear, recurrent, Transformer, and stochastic latent state-space forecasters. We first define the shared widefield protocol, then specify the models and their training objectives. The two-photon extension uses a separate model and observation protocol.

D.1 SHARED WIDEFIELD PROTOCOL

Let $x_t \in \mathbb{R}^N$ denote neural activity, $X = x_{t-T_{\text{in}}+1:t}$ the context, and $Y = x_{t+1:t+H}$ the training target. The principal comparison uses $N = 16$ regions sampled at 30 Hz and $T_{\text{in}} = 90$ (3 s). The training horizon is $H = 90$, except for the next-step Transformers, which use $H = 1$ (Table 2). Data preparation and partitions are specified in the experimental-setup appendix.

Optimization and model selection. On the same train-val-test partition, models were trained with seeds 101, 102, and 103. Shared AdamW (Loshchilov & Hutter, 2019) settings are collected in Table 3. For the reported widefield comparison, each run is represented by the checkpoint with the lowest validation mean-forecast MAE, independently of the training objective. For deterministic models this is the MAE of the model output; for the SSM it is the MAE after averaging decoder means over 16 sampled latent trajectories, with fixed validation seeds across epochs. Training-progress analyses instead use the specified intermediate checkpoints.

Autonomous evaluation. Only the initial 90-step context is observed. Each subsequent prediction is conditioned on generated activity. Multi-step Transformers, RNN, and SSM generate 90-step blocks and re-encode the latest 90 generated observations at each boundary; Transformer positions and attention caches, recurrent states, and SSM initial-state distributions are reset there. Within a block, states are retained. VAR and next-step Transformers update a sliding 90-step context after every prediction. The observed context is excluded from scoring. The 3-, 12-, and 24-s analyses score the first 90, 360, and 720 forecast steps: these are cumulative windows, so changes in scores also reflect changes in the amount of data available for metric estimation.

Table 2: Principal widefield model configurations. H is the number of supervised future steps per training example. Teacher forcing supplies only preceding target observations; all evaluation rollouts use self-generated history. MLP widths list consecutive layer dimensions.

Model	Architecture	Training inputs	H	Objective
VAR	Affine map, $1440 \to 16$	Teacher forcing	90	MAE
RNN	GRU: 8 layers, width 64; layer-normalized readout	Teacher forcing	90	MAE
1_step	Causal Transformer: width 64, 8 layers, 8 heads, FFN 128	Context only	1	MAE
Sequifier	Causal Transformer: width 64, 8 layers, 8 heads, FFN 128	Context only	1	MAE
TF	Same backbone as 1_step	Teacher forcing	90	MAE
AR	Same backbone as TF	Self-feedback	90	MAE
$\text{TF}_{\text{QL_KL}}$	Same backbone as TF	Teacher forcing	90	MAE + histogram and quantile losses
SSM	Latent width 12; context/posterior GRUs 96; transition MLP $12 \to 64 \to 12$; decoder $12 \to 32 \to 32 \to 16$	Target-informed posterior (training only)	90	Regularized variational objective

D.2 DETERMINISTIC FORECASTERS

Linear autoregression (VAR). The VAR baseline predicts $\hat{x}_{t+1} = W \operatorname{vec}(X) + b$, allowing every output region to depend linearly on every region and lag in the context (Lütkepohl, 2005). Training applies this map to ground-truth sliding windows across the target block; inference recursively substitutes predictions for unavailable observations.

Table 3: Optimization settings for the principal widefield comparison. The neural-model column covers the Transformers, RNN, and SSM. Accumulation gives 2048 examples per full optimizer update. SSM-specific regularization is defined in Section D.4.

Setting	Neural models	VAR
Optimizer	AdamW	AdamW
Learning rate	OneCycle, maximum 3×10^{-4}	Constant 10^{-3}
Schedule	Cosine; warm-up fraction 0.3; division factors 25 and 100	None
Weight decay	10^{-2}	10^{-5}
Batch size / accumulation	1024 / 2	2048 / 1
Maximum epochs	150	150
Minimum validation improvement	10^{-6}	10^{-6}
Dropout	0.05 (Transformers and RNN); none (SSM)	None
Gradient clipping / precision	Global norm 1; mixed precision on CUDA	No clipping; float32

Recurrent model (RNN). A unidirectional GRU encodes X from a zero hidden state. The first prediction is an affine readout of the layer-normalized top-layer state. The GRU then consumes the preceding target observation during training, or its own preceding prediction during evaluation, before applying the same readout. Dropout acts between recurrent layers. Training minimizes MAE over all H future steps.

Shared Transformer backbone. The 1_step, TF, AR, and $\mathrm{TF}_{\mathrm{QL_KL}}$ variants use an affine observation embedding ϕ, learned positional embeddings, and a learned future-start token s. Each causal decoder block contains pre-layer-normalized self-attention and a GELU feed-forward network with residual connections. A final layer normalization and affine readout produce the observations. Positions are relative to the current context–forecast block, with $T_{\mathrm{in}} + H$ learned positions.

Training regimes. For teacher forcing (TF), the input tokens are $[\phi(X), s, \phi(Y_1), \ldots, \phi(Y_{H-1})]$; a causal mask prevents each prediction from accessing its own or later targets. All future outputs are computed in one pass. AR instead recursively feeds back its predictions during training. Its attention history grows throughout the block, and gradients propagate through predictions and cached states without detachment. The 1_step baseline uses the same backbone with $H = 1$, so only the context and start token are supplied. TF and AR minimize MAE; $\mathrm{TF}_{\mathrm{QL_KL}}$ uses TF with the augmented objective below.

Sequifier. Sequifier (Luithlen, 2025) is an additional causal Transformer implementation trained on the next-step ℓ_1 objective ($H = 1$) and applied recursively at evaluation. Its dimensions and optimization settings are given in Tables 2 and 3. Operationally, Sequifier encodes input regions separately and predicts the next activity vector from the final observed position; our 1_step model jointly embeds the activity vector and predicts from an appended learned future-start token. Both feed predictions back one step at a time during evaluation.

D.3 Distribution-augmented Transformer objective

For a batch of B examples, let $\mathcal{L}_{\mathrm{MAE}} = (BHN)^{-1} \sum_{b,i,n} |\hat{Y}_{bin} - Y_{bin}|$. The augmented objective is

$$\mathcal{L}_{\mathrm{aug}} = \frac{\mathcal{L}_{\mathrm{MAE}}}{r_{\mathrm{MAE}}} + \lambda_{\mathrm{KL}} \frac{\mathcal{L}_{\mathrm{KL}}}{r_{\mathrm{KL}}} + \lambda_{\mathrm{QTL}} \frac{\mathcal{L}_{\mathrm{QTL}}}{r_{\mathrm{QTL}}}. \tag{2}$$

Each r_j is the detached component value on the first training batch plus 10^{-8}, held fixed thereafter. The MAE-only Transformer objective uses zero distributional weights. For $\mathrm{TF}_{\mathrm{QL_KL}}$, the weights are $\lambda_{\mathrm{KL}} = 0.02$ and $\lambda_{\mathrm{QTL}} = 0.08$. Both added losses compare within-window marginals; neither directly constrains temporal order or cross-region dependence.

Soft-histogram loss. For each region, 33 Gaussian bins span the 0.001–0.999 quantiles of pooled target and detached prediction values across batch and time. The Gaussian width is half the bin width.

Soft counts are summed over time, stabilized by 10^{-8}, and normalized separately for each sequence and region. If p_{bn} and q_{bn} are the resulting target and prediction histograms, respectively, define

$$D_{bn} = \tfrac{1}{2}\{D_{\mathrm{KL}}(p_{bn}\|q_{bn}) + D_{\mathrm{KL}}(q_{bn}\|p_{bn})\}, \qquad s_b = \exp\left[-\frac{1}{N}\sum_n \log(1 + D_{bn})\right], \quad (3)$$

$$\mathcal{L}_{\mathrm{KL}} = 1 - \tfrac{1}{2}\,\mathrm{mean}_b(s_b) - \tfrac{1}{2}Q_{0.10}(\{s_b\}). \quad (4)$$

Here Q_q is the empirical quantile. Histogram support is detached, but soft assignments remain differentiable. The divergence is symmetric KL.

Tail-quantile loss. Temporal quantiles are computed at 99 levels from 0.01 to 0.99, retaining $q \le 0.10$ or $q \ge 0.90$. For each sequence and region, the mean absolute prediction–target difference over these tail quantiles is divided by the target region's interquartile range pooled across batch and time (minimum 10^{-8}). Average the largest $\lceil N/4 \rceil$ regional discrepancies per sequence and then across sequences to obtain D_{QTL}. The loss is $\mathcal{L}_{\mathrm{QTL}} = 1 - (1 + D_{\mathrm{QTL}})^{-1}$.

D.4 Conditional latent state-space model (SSM)

Generative model. A context GRU parameterizes the diagonal Gaussian initial state $p(z_0 \mid X)$. The conditional model is

$$p(Y, z_{0:H} \mid X) = p(z_0 \mid X)\prod_{i=1}^{H} p(z_i \mid z_{i-1})p(Y_i \mid z_i), \quad (5)$$

with $p(z_i \mid z_{i-1}) = \mathcal{N}(m(z_{i-1}), \mathrm{diag}(\sigma_z^2))$ and $p(Y_i \mid z_i) = \mathcal{N}(d(z_i), \mathrm{diag}(\sigma_x^2))$. The transition combines linear and nonlinear dynamics,

$$m(z) = [1 - g(z)] \odot Az + g(z) \odot f(z), \qquad g(z) = \mathrm{sigmoid}(W_g z + b_g). \quad (6)$$

The linear matrix is normalized to spectral norm at most 0.99; this bound does not constrain the full nonlinear transition. The transition MLP f has tanh activations after both affine layers, while decoder d has tanh hidden layers and a linear output. Transition and observation variances are learned, diagonal, and state-independent. Log variances are clipped to $[-12, 8]$; transition scales use softplus plus 10^{-4}. Table 2 gives the layer dimensions. Initial-state and transition standard deviations start near 0.1; the posterior mean correction starts at zero and the gate bias at -4, initially favoring the linear transition.

Variational training. A second GRU, initialized from the context encoder, processes the target prefix $Y_{1:i}$. Its state and z_{i-1} enter a $108 \to 96 \to 24$ MLP with a tanh hidden layer to parameterize

$$q(z_i \mid z_{i-1}, X, Y_{1:i}) = \mathcal{N}\big(m(z_{i-1}) + \Delta\mu_i, \mathrm{diag}(\sigma_{q,i}^2)\big)\,.$$

The objective uses reparameterized latent samples:

$$\mathcal{L}_{\mathrm{SSM}} = \mathcal{L}_{\mathrm{NLL}} + \beta_e\big(\mathcal{K}_1 + 0.25\mathcal{K}_{\mathrm{over}}\big). \quad (7)$$

The Gaussian negative log-likelihood sums over time and channels and averages over examples. $\mathcal{K}_1$ likewise sums latent KL terms over time and dimensions, applying a free-bits floor of 0.05 nats to each example–time–dimension term before averaging across examples. β_e increases linearly from zero at epoch 1 to 0.2 at epoch 45 and then remains fixed. The initial state is sampled from $p(z_0 \mid X)$; there is no separate initial-state KL.

Overshooting compares target posterior distributions with endpoint transition Gaussians after 5, 15, and 30 autonomous steps from four evenly spaced posterior anchors per horizon. Intermediate latents are sampled; anchors and target posterior parameters are detached. For each horizon, $D_{\mathrm{KL}}(q_{\mathrm{target}}\|p_{\mathrm{endpoint}})$ uses the same per-dimension floor, sums over latent dimensions, averages over anchors and examples, and is multiplied by H; $\mathcal{K}_{\mathrm{over}}$ averages these values over horizons.

Forecasts used for evaluation. Evaluation uses only the context-conditioned generative model, never the target-informed posterior. Initial states and transitions are sampled, but the output is the decoder mean $d(z_i)$ without observation noise. The configured evaluation draws one stochastic latent trajectory per context; this is the primary prediction. It represents latent-process variability, not a full observation sample. Each trajectory supplies its own context at subsequent block boundaries. The separately exported deterministic trajectory recursively propagates latent means; because the dynamics are nonlinear, it generally differs from the Monte Carlo mean used for checkpoint selection.

D.5 TWO-PHOTON EXTENSION: POPULATION-CONDITIONED MLP

The separate 88-neuron calcium example uses a shared temporal MLP to encode each neuron's 90-step history, adds learned neuron embeddings, and conditions these features on a population summary. The summary concatenates each neuron's last value, temporal mean, maximum, and change over the last five steps. A population encoder produces feature-wise scale and shift terms. The prediction head emits H_{MLP} residuals r_{in} per neuron and adds a learned persistence skip:

$$\hat{Y}_{in} = \max\{0, \mathrm{softplus}(a_n^i x_{tn} + r_{in}) - \log 2\}, \qquad a_n = \mathrm{sigmoid}(\alpha_n).$$

The local and population widths are 64 and 128, with GELU activations and dropout 0.05. The training configuration uses $H_{\mathrm{MLP}} = 16$, $a_n = 0.965$ at initialization, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, batch size 1024, 30 epochs, and gradient clipping at norm 1. Training and checkpoint selection use weighted smooth-ℓ_1 loss, with unit smooth-ℓ_1 transition threshold and weights $1+2\min\{10, \max(0, Y_{in}/s)\}$. Here s is the 95th percentile of strictly positive training-target values, bounded below by 10^{-6} (or 1 if none are positive).

The inference implementation supports direct block prediction followed by resets to recorded context. These forecasts receive intermittent observations and are not equivalent to the autonomous widefield rollouts above.

E CODE, DATA AND REPRODUCIBILITY

Overview. We release the processed widefield-calcium forecasting dataset (Appendix B.1), the anonymized code used to train models, generate rollouts, compute Nethobench scores, and reproduce the main figures, and the widefield model suite checkpoints. The two-photon dataset (Appendix B.5) will be released on Zenodo upon publication. Dataset construction, splitting, normalization, and long-rollout generation are described in Appendix B. Model architectures and training details are described in Appendix D, and metric definitions are provided in Appendix C.

Dataset. The processed benchmark dataset is available at:

`https://huggingface.co/datasets/Nethobench/nethobench-widefield-v1`

Code. The code is provided in two anonymized repositories for review, released under MIT License. The Nethobench repository contains the final metric implementations, family-level and composite scoring code, documentation, and examples intended for public reuse.

`https://anonymous.4open.science/r/nethobench-1D12`

The forecasting repository contains the model training and evaluation pipeline for widefield and two-photon calcium imaging, including rollout-generation scripts and Nethobench-evaluation workflows:

`https://anonymous.4open.science/r/sequential_neural_pred-3CFB`

Checkpoints. The widefield models suite checkpoints are available at:

`https://anonymous.4open.science/r/submission_checkpoints-86D5`

Reproducibility workflow. To reproduce the main benchmark analyses, users should download the processed dataset, train or load a forecasting model, generate closed-loop rollouts from the provided context windows, and evaluate the generated trajectories against held-out target windows using the Nethobench scoring code. The reported outputs are the five family-level structural-realism scores, the final composite score, and pointwise-fidelity sidecar metrics.

E.1 EVIDENCE ROADMAP

The roadmap separates support, availability and units of variation. Sequences, model seeds, corruption draws and participants are not interchangeable independent units. Absolute scores should only be compared on matched supports and definitions.

Evaluation questions and evidence.

Question	Evidence and interpretation boundary
Temporal order	Figures 2, S2, S17: shuffles, blocks, jitter, ACF/PSD ablations. Sensitivity is not complete dynamical identification.
Under-dispersion	Figure 2g and Figure S11: mean-preserving variance contraction; Figure S18: proper scores and coverage. Structural sensitivity and conditional calibration are separate.
Empirical anchors	Figures S8 and S11: real–real references and structured shuffle or resampling controls; no universal ceiling or floor.
Parameter sensitivity	Figures S5, S7, S24: weights, PCA, states, lags and explicit implementation variants. Fixed-panel stability does not certify new supports.
Family dependence	Figures S6 and S9: correlated scores, omissions and all subsets. Diagnostic complementarity does not require all families for ranking.
Stochastic validity	Figures 2b and S18: correct-process samples versus predictive-distribution scoring. Structural agreement does not imply conditional calibration.
Cross-modality scope	Figures 3, S3, S16, S17: distinct recordings and observation transforms; scores are not directly comparable across them.
Human protocol	Appendix B.6: training-only contact selection and exact analog-library exclusions; offline transforms and the changed contact fold remain limitations.

Table R1: Evaluation supports, units of variation and interpretation limits. T1–T4 denote supplementary tables.

Figure/ table	Dataset	Scientific question	Evaluation support	Context included?	Component availability	Uncertainty unit	Independent unit	Principal limitation
1	All	Evaluation contract	Experiment-specific aligned tensors	Varies	Five families /18 retained components	Schematic; no interval	None	Not a validity guarantee
2	Synthetic / widefield	Oracle; controlled failures	100×768×16 oracle; fixed widefield controls	No / separate windows	18	Oracle and shuffle draws; classes; sequence bootstrap	Generator realization / sequence	Different targets and score scales; no strict oracle ceiling
3; T2	OmniMouse / human	Conditioning; accuracy vs structure	60×128 /50×320×64 per fold	No	Short-window mask / persistence ACF absent	Four-split SEM / participant summaries	One session /five participants	One Omnimouse session; observed target-interval covariates ; offline human transform and QC
4; T3–4	Widefield	Design contrasts	720 future bins×16 at 30 Hz	No	18 supplied	training-seed SD ; figure means only	Sequence-level split	-
S1	Fixed prediction / widefield	Metric sensitivity	34 corruptions across doses; fixed 100 × 630 × 16 temporal panel	No	16-screen inventory; ACF/PSD tested separately	Maximum-over-dose summaries /20 corruption draws	Fixed prediction / tensor	Sensitivity is not exclusive selectivity
S2	Widefield / fixed models	ACF; PSD coverage	360 controls; 500 bootstrap rows	Yes (fixed 630-bin tensor)	18 with ACF/PSD ablations	Corruption-realization ranges / paired sequence resamples	Fixed tensor / sequence	Reused samples; resamples are not independent fits
S3	Synthetic / simulated BOLD	Observation-regime validity	8×160×12 per seed; BOLD 40 bins	No	18 /17 (no QNT)	Five seed means	Generator seed	Mixture retains reference; simulated BOLD is not real fMRI
S4	Candidate audit	Metric retention	45 candidates, six criteria	N/A	Complete records for 16/18; ACF/PSD separate	Descriptive available-entry means	Candidate metric	Missing criteria and retrospective selection evidence

Figure/ table	Dataset	Scientific question	Evaluation support	Context included?	Component availability	Uncertainty unit	Independent unit	Principal limitation
S5	Synthetic training	Definition robustness	45 seed/checkpoint rows; 5000 weight draws	No (810 future bins after 90-bin context)	18 with parameter and weight variants	Specification ranges over shared weight draws	15 training seeds	Specification ranges are not sampling confidence
S6	Training / controls	Dependence; localization	45 training evaluations; 360 controls; 200 bootstraps/corruption	No / yes	18	Descriptive correlations / paired sequence bootstrap	Training seed / fixed evaluation tensor	Repeated scores; families are not independent
S7	Training /five models	Parameter stability	45 sequences$\times$810 future bins; $100 \times 630 \times 16$	No / yes	18 with variants	Paired seed contrasts /100 sequence bootstraps per model	Training seed / sequence	Not the principal eight-model panel
S8	Fixed widefield	Empirical anchors	$100 \times 630 \times 16$; 50/50 references; five forecasts use all 100	Yes	18	20 reused partitions /20 realizations per null	Reused sequence pool	Reference and forecast sample sizes differ

Figure/ table	Dataset	Scientific question	Evaluation support	Context included?	Component availability	Uncertainty unit	Independent unit	Principal limitation
S9	Fixed widefield	Family subsets	31 subsets×100 resamples×5 models	Yes	Five families /31 nonempty subsets	Paired sequence bootstrap	Sequence	Best subsets selected on the same panel
S10; T1	OmniMouse	Scoring support	30 recorded +30 predicted /60 forecast-only samples	Yes / no	QNT unavailable	Four-split SEM	One session	No session replication; observed target-interval covariates
S11	Widefield	Variance contraction	Paired 720-bin windows; four joint batches	No	18 plus PSD sidecars	Paired-bootstrap intervals / joint-batch SEM	Window pair / joint batch	-
S12	Widefield models	Supervision / rollout	90-context/90-target training blocks; rollout validation	Varies by panel	Auxiliary losses	Supplied curves	Model run	Losses are not retained components
S13–14	Widefield models	Submetric diagnostics	Matched GT/prediction diagnostics from the principal export	No	18 plus diagnostics	Supplied aggregate diagnostics	Sequence / split as supplied	-
S15	Widefield models	Scale sensitivity	Context 30/90/300; regions 4/8/16; epochs 10/30/100	Protocol-specific	Family summaries	Supplied variation	Model run	No biological replication
S16	Two-photon	Single-neuron applicability	88 neurons; 10 Hz; chunked rollout	Context resets between chunks	As supplied	Four-split SEM	One mouse	Not a continuous autonomous rollout
S17	IBL	Population transfer	100-bin full window /75-bin target; 12 regions; 436/512/128 contexts	Yes / no	Recorded component-availability mask	Three training seeds /three corruption RNG seeds	Session / fixed laboratory partition	Supports differ; missing channels are zero-filled
S18	IBL	Probability; conditioning	75×12 targets; 18 draws; 436/512/128 contexts	No	Proper scores and calibration separate from structural components	18 structural Monte Carlo replicates /10 training seeds	Context / training seed	Pooled dependent calibration entries; no general conditioning inference
S19	Synthetic / transfer	Training; endpoints	96×912 /128×896 endpoints; 45×810 training panel	No	18	Five paired /15 training /10 architecture seeds	Training seed	Capacity, context, depth and training co-vary

Figure/ table	Dataset	Scientific question	Evaluation support	Context included?	Component availability	Uncertainty unit	Independent unit	Principal limitation
S20–21	Human iEEG	Participant robustness	$50 \times 320 \times 64$ per held-out movie	No	18; persistence ACF absent	Hierarchical participant/movie bootstrap; sign flips; movie/seed contrasts	Five participants	Offline transform and QC; sign test resolves only to $p = 0.0625$
S22	Human iEEG	Corruption calibration	15 folds×4 types×5 doses	No	18	Participant SEM; one nested draw per fold/type	Five participants	Perturbations do not establish task-specific utility
S23	Widefield	Component audit	720 future bins×16; reported model means	No	18 supplied	Figure means; training-seed SD	Sequence-level split	-
S24	Synthetic / simulated BOLD	Implementation sensitivity	80 neural +75 BOLD comparisons	No	18 /17; two variants	All five seed values	Generator seed	-
S25	Widefield / human iEEG	Composite robustness to relational rescoring	58×720×16 /50×320×64 per movie	No	18; lagged covariance and VAR rescored	Paired seed / participant values; mean bars	Training seed / participant	Tested-panel stability, not invariance of every metric or comparison

F SUPPLEMENTARY FIGURES

Table 4: **OmniMouse evaluation at the native horizon with shared context.** Nethobench compares 30 recorded context samples followed by 30 predicted samples against the same context followed by the recorded future. Nethobench values are mean ± SEM across four sequence-level splits from one held-out session. The two grey columns reproduce response correlation (higher is better) and response loss (lower is better) reported by the OmniMouse authors under their evaluation protocol. Bold indicates the best mean or reported value among the four conditioning configurations in each column, excluding controls and references. The shared context contributes to absolute agreement. Temporal shuffling is reported as a control, not a universal score floor.

Conditioning / reference	Distribution	Temporal-spectral	Relational	Geometry	State dynamics	Composite	Response correlation ↑	Response loss ↓
Neural history + video + behavior	**0.111** ± 0.002	**0.855** ± 0.006	**0.589** ± 0.007	0.260 ± 0.041	**0.780** ± 0.026	**0.507** ± 0.012	**0.398**	**0.429**
Neural history + behavior	0.106 ± 0.002	0.839 ± 0.006	0.579 ± 0.008	**0.266** ± 0.068	0.740 ± 0.018	0.494 ± 0.018	0.340	0.463
Video + behavior	0.110 ± 0.001	0.850 ± 0.006	0.585 ± 0.008	0.233 ± 0.069	0.766 ± 0.028	0.497 ± 0.020	0.365	0.461
Video only	0.099 ± 0.001	0.827 ± 0.004	0.577 ± 0.008	0.191 ± 0.074	0.679 ± 0.007	0.466 ± 0.016	0.282	0.513
Persistence control	0.148 ± 0.002	0.739 ± 0.010	0.542 ± 0.009	0.500 ± 0.066	0.881 ± 0.010	0.544 ± 0.015	—	—
Real–real reference	0.202 ± 0.005	0.910 ± 0.009	0.756 ± 0.013	0.793 ± 0.047	0.956 ± 0.007	0.704 ± 0.010	—	—
Temporal shuffle	1.000 ± 0.000	0.745 ± 0.003	0.721 ± 0.024	0.927 ± 0.037	0.924 ± 0.005	0.860 ± 0.010	—	—
Time-and-region shuffle	0.155 ± 0.003	0.687 ± 0.006	0.599 ± 0.009	0.474 ± 0.094	0.822 ± 0.009	0.534 ± 0.020	—	—
Fully factorized null	0.156 ± 0.003	0.670 ± 0.003	0.600 ± 0.015	0.477 ± 0.051	0.818 ± 0.005	0.532 ± 0.011	—	—

Table 5: **OmniMouse evaluation on two-block forecast-only sequences.** Nethobench compares two successively generated 30-sample blocks (2 s at 30 Hz) against the corresponding recorded future, excluding the initial context from scoring. Enabled video and behavioral covariates remain recorded observations spanning both predicted intervals. Values are mean ± SEM across four sequence-level splits from one held-out session. Bold indicates the highest mean among the four conditioning configurations in each column, excluding references and controls. Temporal shuffling is reported as a control, not a universal score floor.

Conditioning / reference	Distribution	Temporal-spectral	Relational	Geometry	State dynamics	Composite
Neural history + video + behavior	0.062 ± 0.002	0.696 ± 0.007	0.494 ± 0.006	**0.126** ± 0.063	**0.669** ± 0.055	0.400 ± 0.008
Neural history + behavior	0.056 ± 0.002	0.662 ± 0.006	0.473 ± 0.002	0.085 ± 0.031	0.649 ± 0.058	0.377 ± 0.007
Video + behavior	**0.064** ± 0.001	**0.716** ± 0.003	**0.499** ± 0.008	0.122 ± 0.061	0.645 ± 0.059	**0.401** ± 0.008
Video only	0.057 ± 0.001	0.666 ± 0.012	0.482 ± 0.003	0.076 ± 0.018	0.496 ± 0.058	0.351 ± 0.011
Real–real reference	0.144 ± 0.005	0.904 ± 0.014	0.643 ± 0.029	0.591 ± 0.034	0.917 ± 0.019	0.620 ± 0.017
Temporal shuffle	1.000 ± 0.000	0.630 ± 0.005	0.619 ± 0.016	0.873 ± 0.044	0.861 ± 0.010	0.794 ± 0.009
Time-and-region shuffle	0.107 ± 0.002	0.513 ± 0.014	0.441 ± 0.002	0.041 ± 0.037	0.649 ± 0.017	0.346 ± 0.008
Fully factorized null	0.108 ± 0.002	0.504 ± 0.018	0.440 ± 0.001	0.029 ± 0.022	0.652 ± 0.012	0.343 ± 0.007

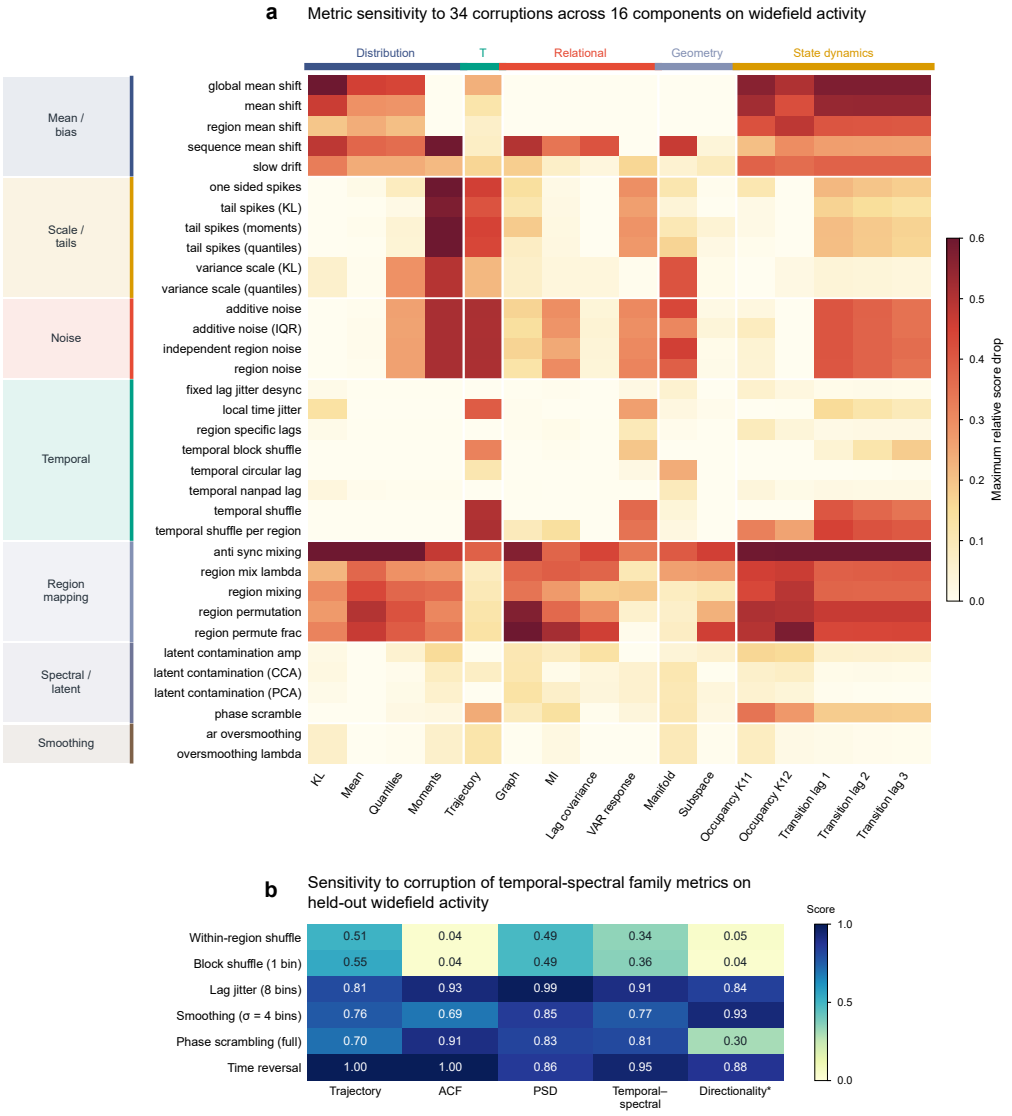


Figure S1: **Metric-level responses to specified corruptions.** (a) Maximum relative score drops over tested severities for 34 corruptions and 16 components; the reference is fixed. Labeled row blocks distinguish corruption classes. (b) Dedicated trajectory, ACF, PSD and directionality controls on held-out widefield activity; cells average 20 corruption realizations at the displayed dose. Directionality is a signed lagged-covariance diagnostic excluded from the composite. Replicates are transformations of a fixed $100 \times 630 \times 16$ tensor. Overlapping responses do not imply exclusive family selectivity (Appendix C.2).

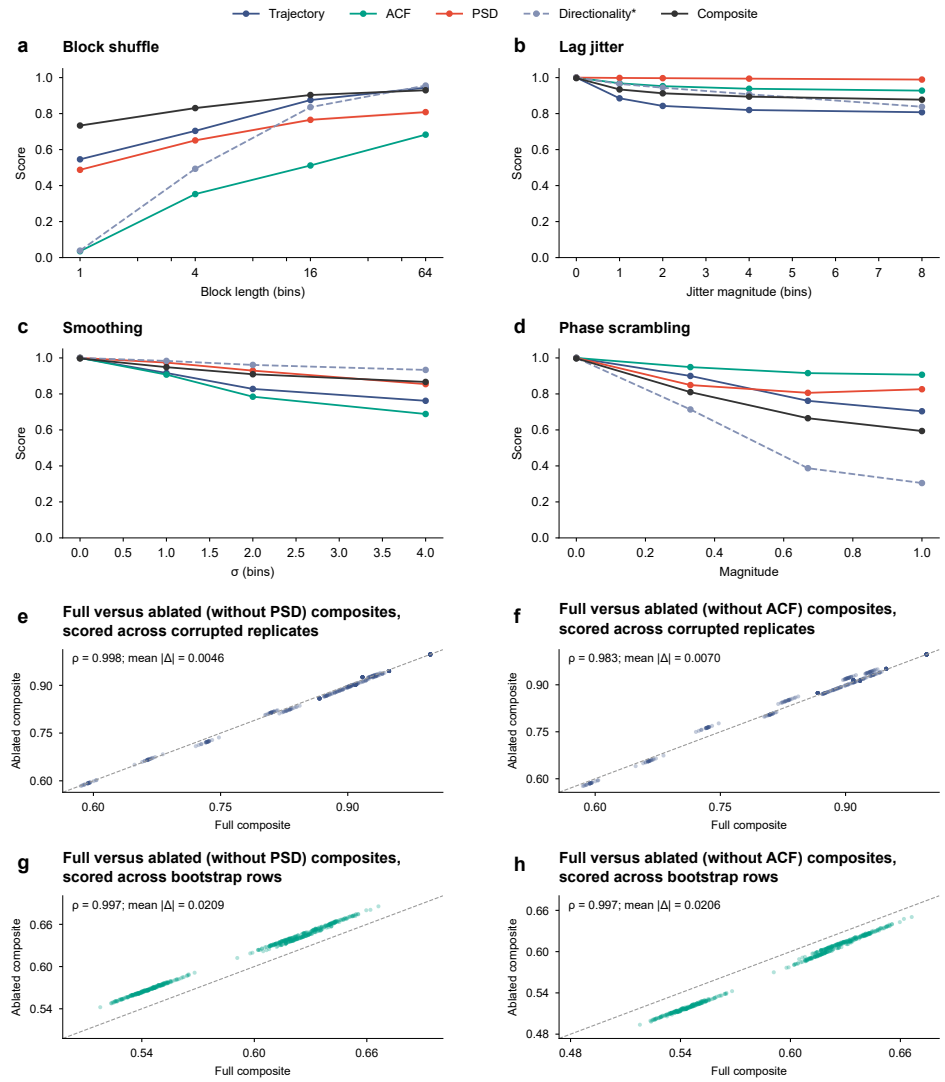


Figure S2: **Temporal sensitivity and component ablations.** (a–d) All doses of block shuffling, jitter, smoothing and phase scrambling. Lines are means and bands span the 10th-90th percentile over 20 corruption realizations. Longer blocks preserve more temporal order. (e–f) Full versus ablated composites, scored across 360 temporal controls, including time reversal. (g–h) Analogous comparisons, scored across 500 evaluation-bootstrap rows (100 paired sequence resamples, five fixed models). Removing ACF or PSD averages the remaining two temporal components equally while retaining family weight 0.18. Directionality remains outside the composite.

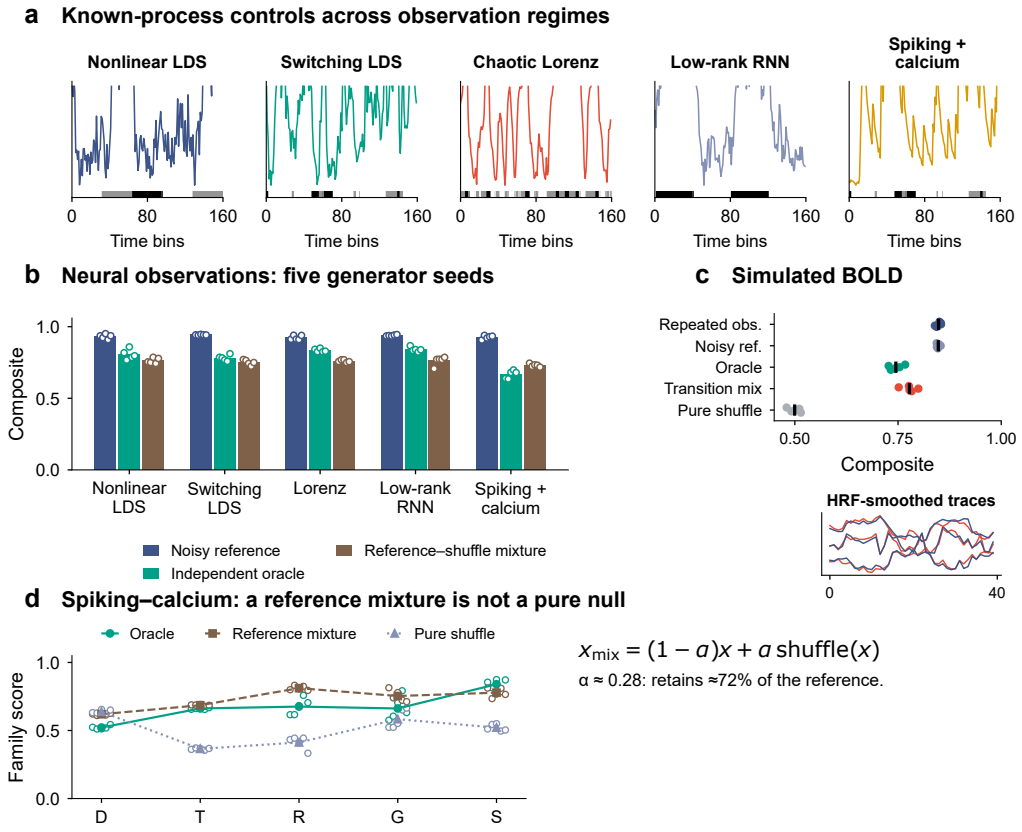


Figure S3: **Structural validation across richer synthetic systems and simulated BOLD.** (a) Display-scaled available trace segments and latent/task-state strips from five known-process systems. (b) Five generator seeds, all 18 retained components: noisy reference, independent oracle, and reference–shuffle mixture. Bars are means; points retain individual seeds. The noisy reference is not an independent real–real half-split. (c) Simulated BOLD controls average three systems within each seed; 40-bin support provides 17/18 components because quantiles are unavailable. The inset retains the reference-like (blue) and corrupted (red) HRF-smoothed trace examples. (d) Spiking–calcium profiles distinguish a mixture retaining approximately 72% of the reference from a pure shuffle. The independent oracle exceeds the pure shuffle in all five seeds. D/T/R/G/S denote distributional, temporal-spectral, relational, geometric and state dynamics families. The mixture's closeness to the reference and its family profile explain its higher composite. Controls do not establish mechanistic validity or real-fMRI performance. Human calibration is reported separately in Figure S22.

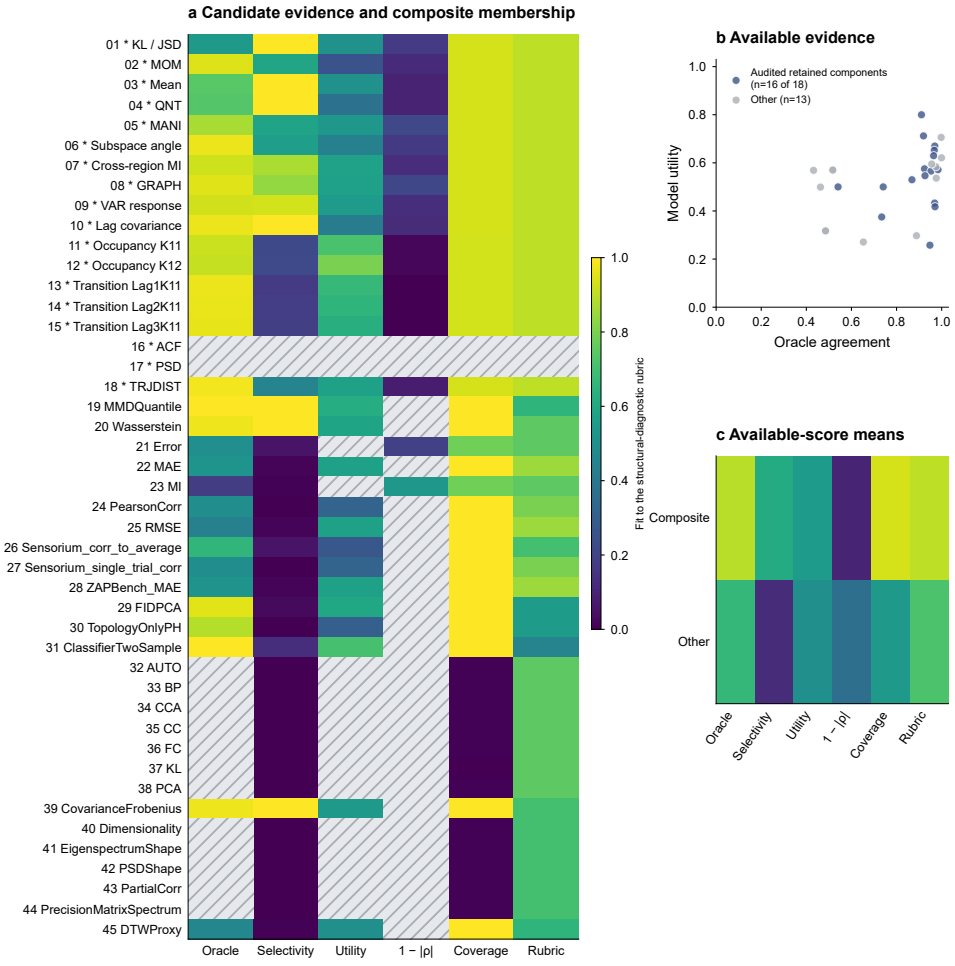


Figure S4: **Metric construction and retrospective design evidence.** (a) Six screening criteria for 45 candidates; asterisks mark 18 retained components. Hatched cells indicate unavailable measurements. (b) Oracle agreement versus utility for components with both available (16 of 18 retained). (c) Available-entry criterion means. Utility on a selection panel is not independent validation. Predictive sidecars address different targets. Dedicated temporal controls support ACF/PSD coverage without supplying missing screening records (Appendix C.3).

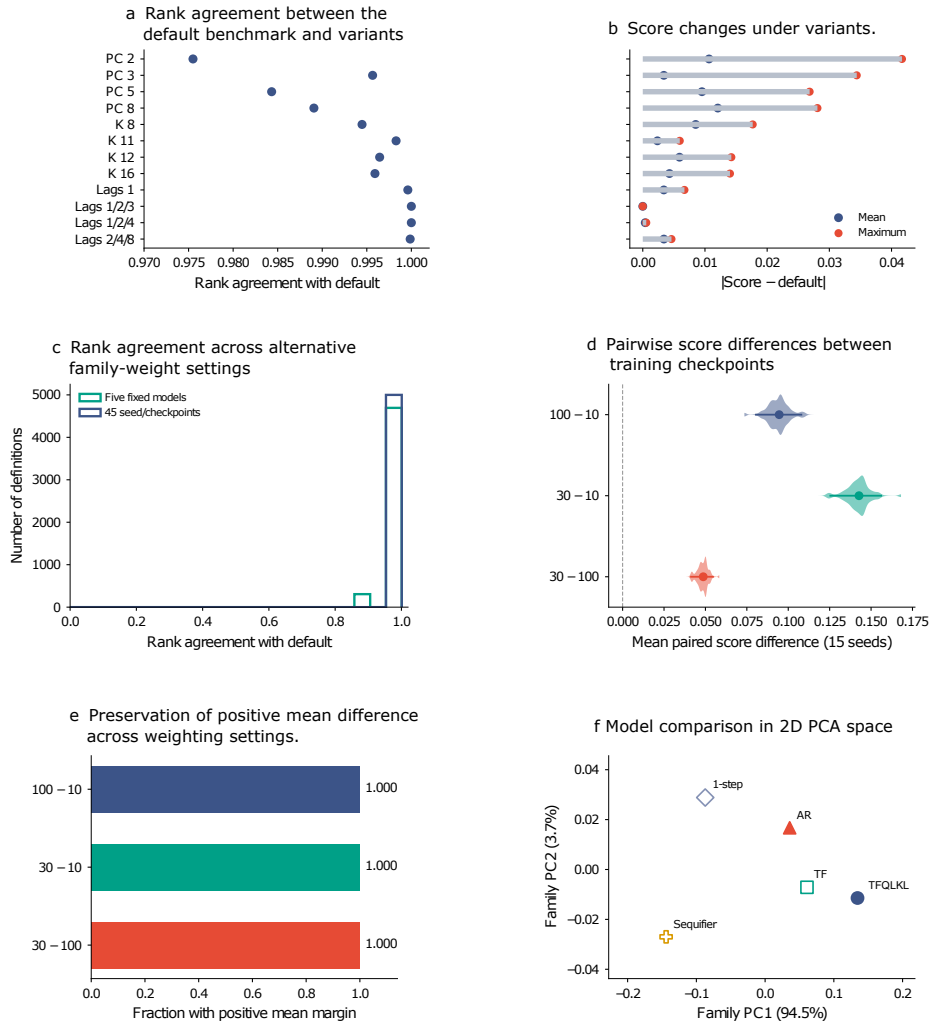


Figure S5: **Robustness to benchmark definition and family weighting.** (a) Rank agreement between the default benchmark and variants using different PCA dimensions, latent-state counts (K), or transition lags across 45 matched seed-checkpoint conditions. (b) Mean and maximum absolute score changes under the same variants. Rankings remain stable although absolute scores shift slightly. (a-b) 15 seeds, 3 checkpoints. (c) Rank agreement across 5,000 alternative family-weight settings for five models and 45 seed-checkpoint conditions. The ensemble mixes Dirichlet$(80w)$ draws (70%), equal weights (10%), single-family omission (10%) and single-weight doubling (10%), with renormalization. (d) Mean paired score differences between training checkpoints across 15 seeds under each weighting. Intervals describe specification sensitivity. (e) Fraction of weighting settings preserving a positive mean difference. All three contrasts remain positive across all settings. (f) PCA visualization of the five-dimensional family-score profiles. Filled markers indicate Pareto-nondominated models. The PCA axes are used only for visualization and do not represent directions of improvement. Definitions and RNG are specified in Appendix C.3.

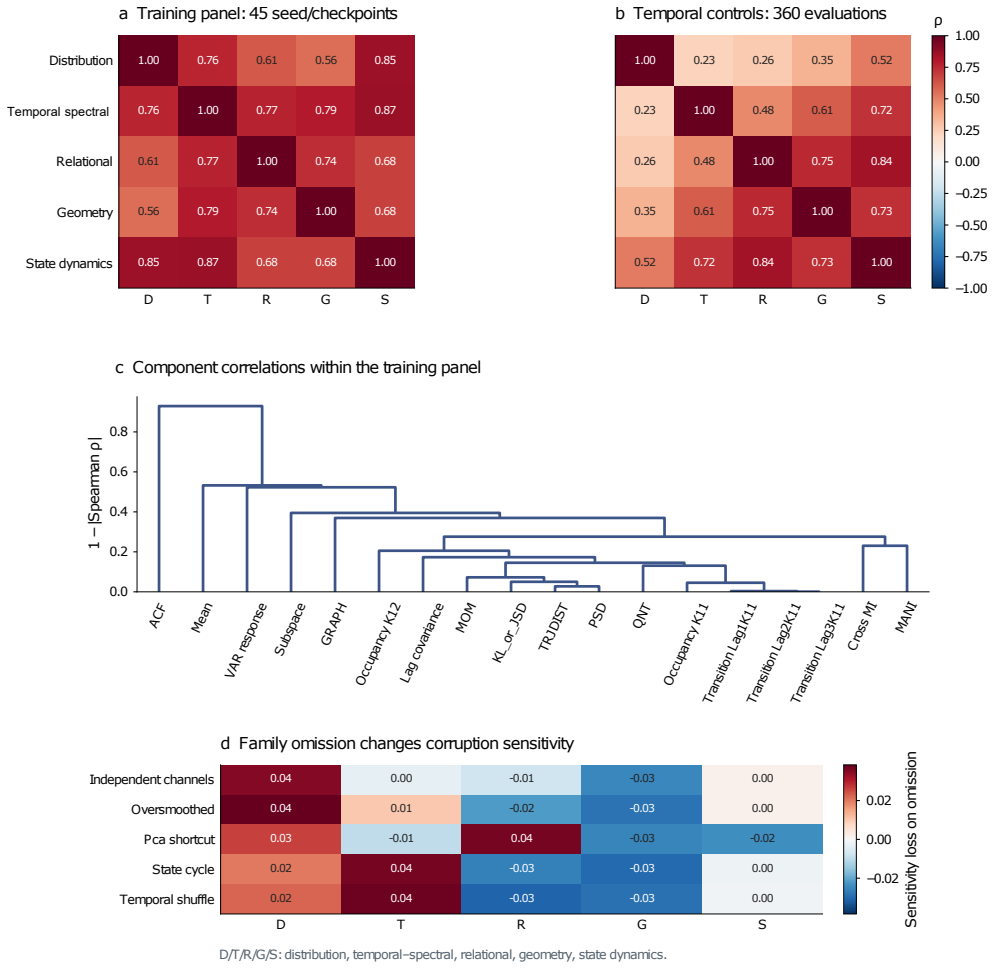


Figure S6: **Family dependence and diagnostic complementarity.** (a–b) Descriptive family correlations across 45 training evaluations and, separately, 360 temporal controls. Repeated evaluations. (c) Average-linkage clustering of 18 components using $1 - |\rho|$. (d) Sensitivity loss on omission, $C_{-f} - C$, with remaining family weights renormalized: positive values mean that omission reduces sensitivity to the specified corruption. The heatmap shows point estimates; 200 paired sequence bootstrap records per corruption accompany the analysis. Correlation and omission effects do not establish causal unique information or downstream utility. Multiple profiles localize distinct failures even when one family reproduces a coarse model ordering.

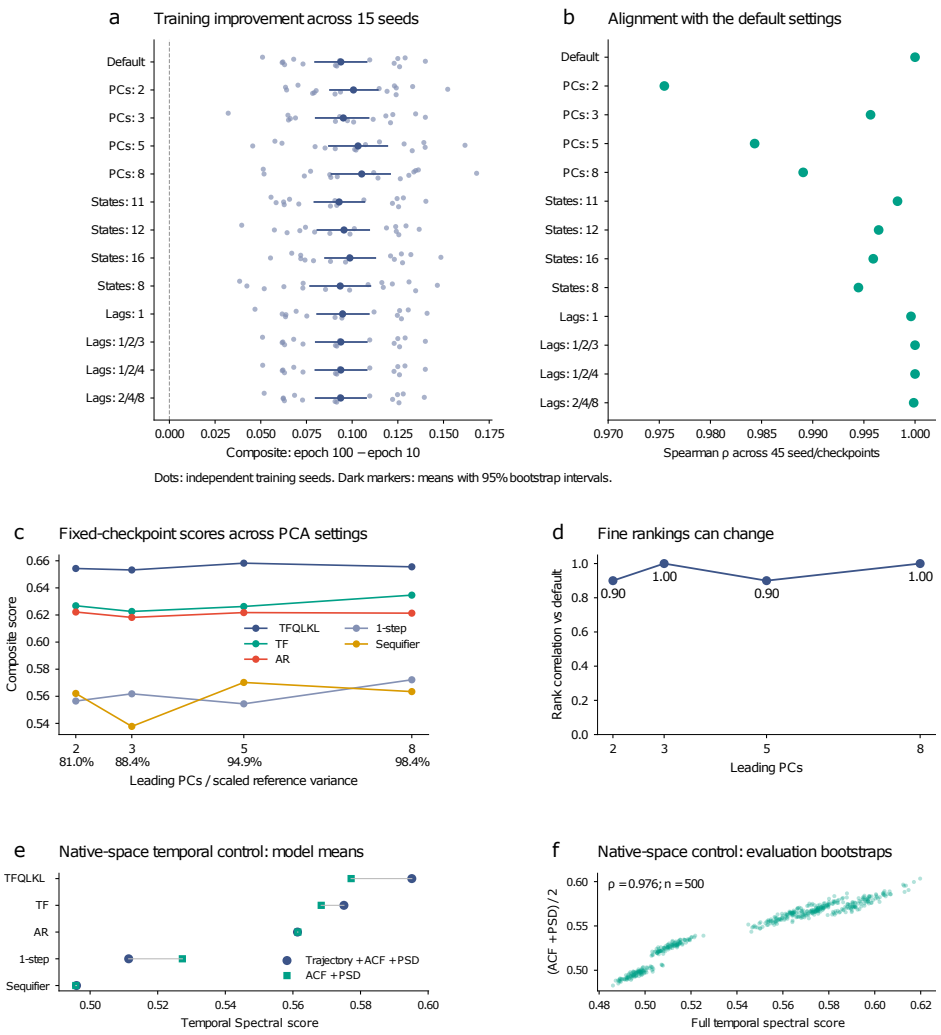


Figure S7: **Sensitivity to benchmark parameters and temporal components.** (a–b) 15 independently initialized teacher-forced, one-step Transformers (8 layers, 8 heads, width 64), trained with MAE on 16-region widefield activity and evaluated on a fixed split of 45 held-out sequences, scoring 810 forecast bins after 90 context bins. Across the default and 12 alternative PCA, state-count, and lag settings, all seeds improve from epoch 10 to 100; intervals are seed-bootstrap 95% intervals. Rank agreement with default settings is assessed across 45 seed/checkpoint evaluations (15 seeds, epochs 10, 30, and 100). (c–d) Five fixed Transformer checkpoints evaluated on one aligned $100 \times 630 \times 16$ tensor. Varying PCA dimensionality preserves the top-ranked model but swaps 1-step and Sequifier at 2 and 5 PCs (Spearman $\rho = 0.90$–1.00). (e–f) Removing the trajectory component preserves the five-model temporal–spectral ordering. Full and ACF/PSD-only scores correlate strongly ($\rho = 0.976$) across 100 paired sequence bootstraps per model (500 evaluations).

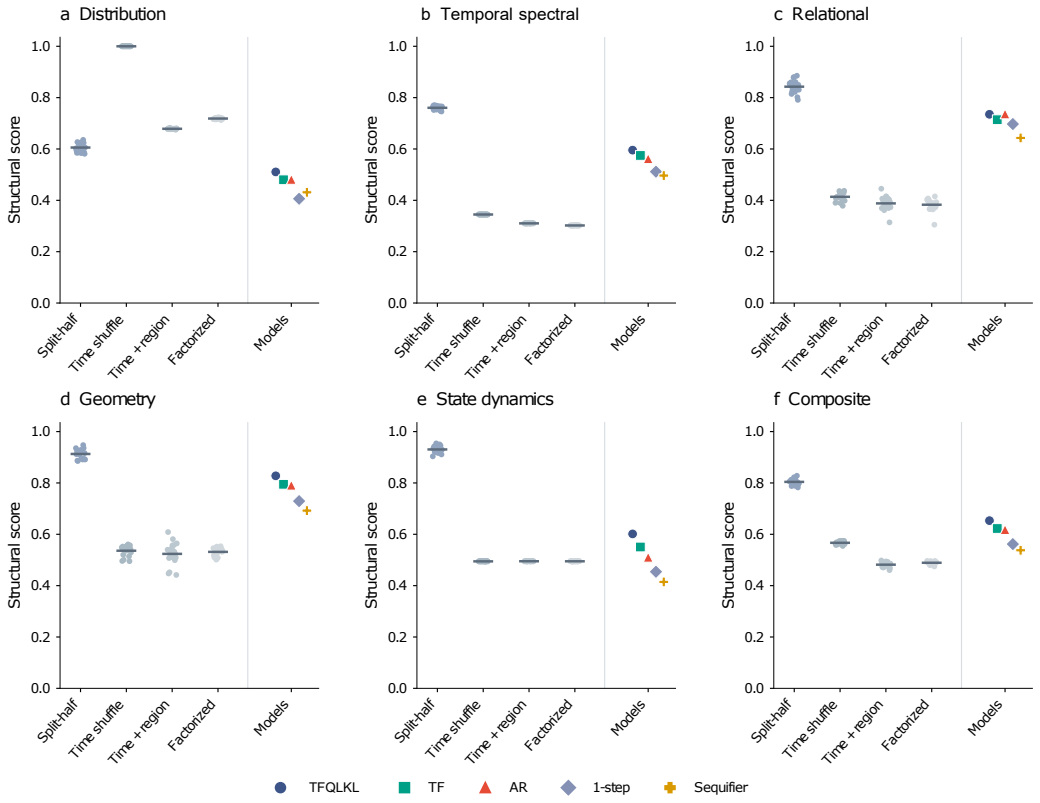
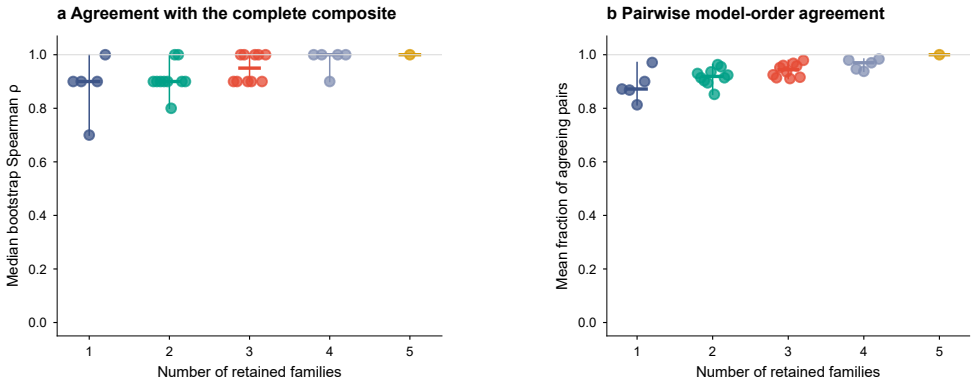


Figure S8: **Empirical references, structured nulls and forecast scores.** All values use one $100 \times 630 \times 16$ widefield tensor at 30 Hz. Real–real points represent 20 random disjoint 50/50 partitions that reuse the sequence pool. Each null has 20 realizations: within-region time permutation, joint time–region permutation, or regional marginal resampling. Five forecast markers score all 100 sequences. Thus reference/model comparisons are descriptive and not sample-size matched. Horizontal marks show reference/null means. Temporal shuffling preserves marginals and need not fall below a forecast or real–real reference. These empirical anchors are not universal bounds; copied-context support differs from the principal experiment.



c Best observed subset(s) at each size

Size	Best observed subset(s)	Median ρ	95% bootstrap range	Pair agreement
1	T	1.00	[0.90, 1.00]	0.97
2	R+S	1.00	[0.80, 1.00]	0.96
3	T+R+S	1.00	[0.90, 1.00]	0.98
4	T+R+G+S	1.00	[0.90, 1.00]	0.98
5	D+T+R+G+S	1.00	[1.00, 1.00]	1.00

Figure S9: **Family-subset agreement on five fixed forecasting models.** All 31 nonempty subsets retain relative default weights and renormalize. Each is compared with the full composite on 100 paired evaluation-sequence bootstraps. (a) Median bootstrap rank agreement per subset. (b) Mean agreement over all ten model pairs; ties in only one ordering count as disagreement. (c) Best observed subsets selected descriptively; brackets give central 95% bootstrap ranges. The temporal family alone attains median $\rho = 1.00$ (range $[0.90, 1.00]$), with pairwise agreement 0.971. One family can recover coarse ordering here, while multiple family profiles remain useful for diagnostic localization. No universal minimum family count follows.

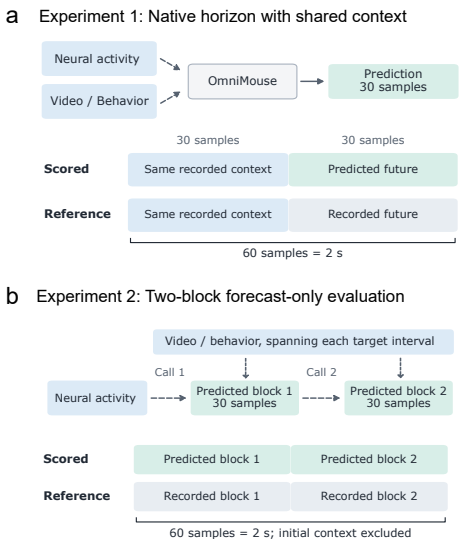


Figure S10: **OmniMouse input and scoring protocols.** (a) Native-horizon evaluation concatenates 30 recorded context samples with 30 predicted neural samples and compares this sequence with the same recorded context followed by the recorded target. This shared context contributes to absolute agreement. Configurations without neural history receive no neural response values as model input. (b) Prediction-only evaluation scores two successive 30-sample predicted blocks against the corresponding recorded activity, excluding initial recorded context. For neural-history configurations, the second call encodes the first 25 predictions from the preceding block. In both protocols, enabled video and behavioral inputs include observations from the target interval. The complete video clip enters the encoder before any strategy-specific feature selection (Appendix B.2).

Table 6: **Family-level Nethobench and pointwise-fidelity scores.** Model values are mean ± sample standard deviation across three training seeds, after averaging four evaluation-group scores within each seed. Empirical reference values are mean ± Monte Carlo sample SD across 20 random draws, after averaging four fixed evaluation groups of 56/56/55/55 sequences within each draw; all draws use the same 222 targets and 720-bin forecast window. Pointwise fidelity is computed on the same forecasts and evaluation groups as the structural scores and is reported separately from the structural composite.

	Model	Nethobench scores						
		D	T	R	G	S	Composite	Pointwise fidelity
Models	VAR	0.345 ± 0.005	0.342 ± 0.008	0.614 ± 0.014	0.757 ± 0.020	0.340 ± 0.006	0.482 ± 0.006	0.354 ± 0.004
	RNN	0.425 ± 0.010	0.466 ± 0.011	0.496 ± 0.012	0.760 ± 0.012	0.613 ± 0.018	0.544 ± 0.006	0.357 ± 0.003
	SSM	0.502 ± 0.020	0.502 ± 0.013	0.562 ± 0.014	0.811 ± 0.024	0.668 ± 0.028	0.602 ± 0.018	0.309 ± 0.003
	1-step	0.399 ± 0.011	0.421 ± 0.010	0.672 ± 0.011	0.735 ± 0.014	0.464 ± 0.029	0.541 ± 0.007	0.370 ± 0.003
	Sequifier	0.389 ± 0.010	0.493 ± 0.010	0.578 ± 0.031	0.663 ± 0.073	0.495 ± 0.064	0.522 ± 0.030	0.324 ± 0.023
	AR	0.437 ± 0.007	0.403 ± 0.014	0.604 ± 0.024	0.714 ± 0.016	0.500 ± 0.005	0.532 ± 0.009	0.363 ± 0.003
	TF	0.439 ± 0.013	0.480 ± 0.009	0.696 ± 0.006	0.714 ± 0.009	0.541 ± 0.012	0.576 ± 0.006	0.364 ± 0.001
	TFQLKL	0.438 ± 0.010	0.473 ± 0.007	0.679 ± 0.009	0.678 ± 0.028	0.544 ± 0.018	0.564 ± 0.009	0.364 ± 0.002
Refs.	Real–real reference	0.721 ± 0.038	0.843 ± 0.013	0.830 ± 0.012	0.893 ± 0.041	0.888 ± 0.017	0.830 ± 0.012	–
	Time+region shuffle	0.562 ± 0.021	0.278 ± 0.008	0.396 ± 0.013	0.440 ± 0.037	0.503 ± 0.039	0.439 ± 0.007	–

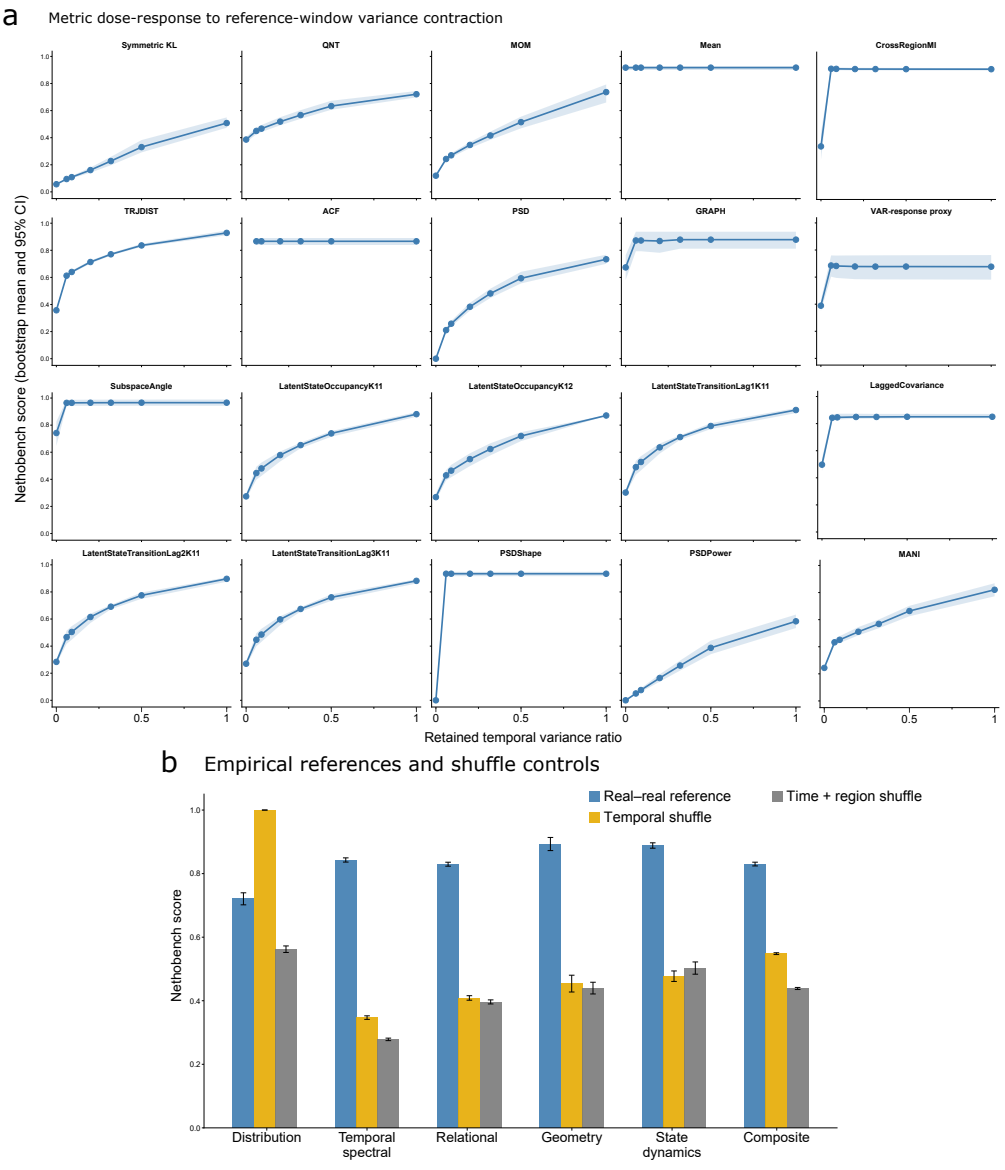


Figure S11: **Empirical references and sensitivity to temporal-variance contraction in widefield neural activity. (a)** Metric-level responses to progressively contracting the temporal variance of the second window in each split-half pair. For each region, the contracted activity was defined as $Y_\rho = \bar{Y} + \sqrt{\rho}(Y - \bar{Y})$, thereby retaining a fraction ρ of the original temporal variance while preserving the temporal mean. Thus, $\rho = 1$ corresponds to the unmodified real window, whereas $\rho = 0$ produces a temporally constant signal at each region's mean. Lines show bootstrap-mean Nethobench scores, and shaded bands indicate 95% confidence intervals obtained by paired resampling of joint batch scores. **(b)** Family-level and composite scores for matched empirical references, using the same 222 evaluation targets and 720-timestep forecast-only windows as the principal model comparison. For each target, the real-real reference samples a non-overlapping 720-timestep companion window from the same recording. Temporal shuffling independently permutes time within each region, preserving regional marginals while disrupting temporal structure and inter-region coordination. Time-and-region shuffling permutes all entries within each target sequence, additionally disrupting regional identity and regional marginals. Each of 20 random draws is scored within four fixed evaluation groups of 56/56/55/55 sequences. Bars show means across draws after equally averaging the four group scores within each draw; error bars show the Monte Carlo sample SD of these draw means.

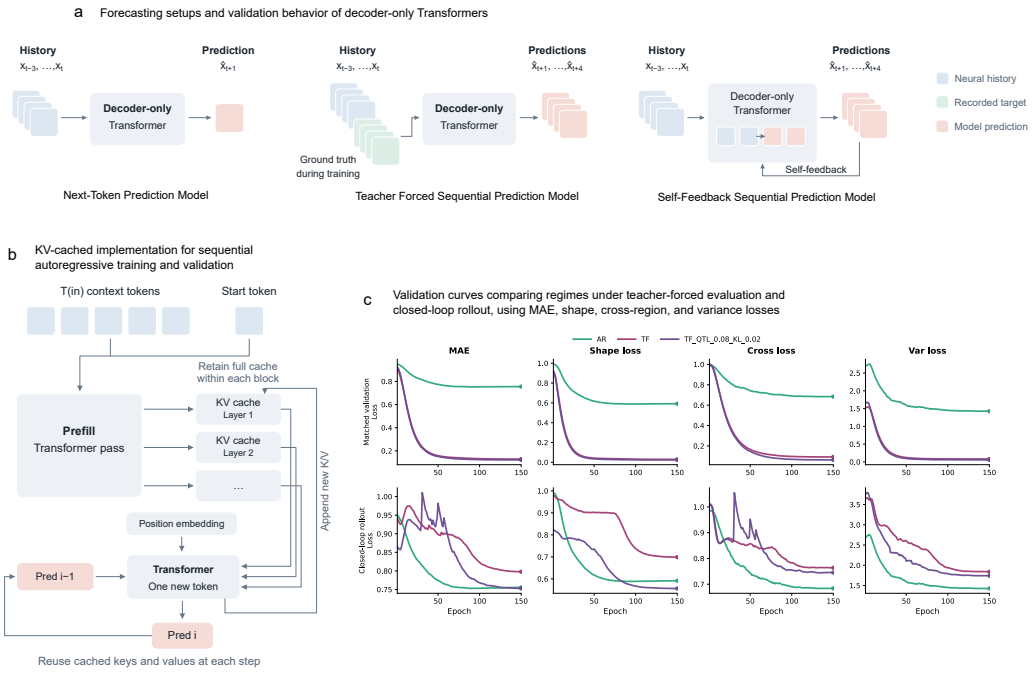


Figure S12: **Forecasting formulations, autoregressive implementation, and validation behavior of decoder-only Transformers.** **(a)** Comparison of next-token, teacher-forced sequential, and self-feedback sequential prediction. The next-token model predicts a single future timestep from the observed history; the teacher-forced model predicts a future sequence conditioned on shifted ground-truth targets; and the self-feedback model generates the sequence autoregressively by conditioning on its own previous predictions. **(b)** KV-cached implementation of self-feedback prediction. A causal prefill pass over the context and start token initializes the layer-wise key–value caches, after which only the newly generated token is processed at each step. The principal implementation retains the growing cache throughout a block (Appendix D.2). The cache resets at block boundaries. **(c)** Validation losses across training for self-feedback (AR), teacher-forced (TF), and distribution-augmented teacher-forced (TF-QL-KL) models. Rows show training-regime-matched validation (top) and common closed-loop rollout (bottom); columns report mean absolute error (MAE), temporal-shape, cross-region, and variance losses. These auxiliary diagnostics differ from the retained benchmark components.

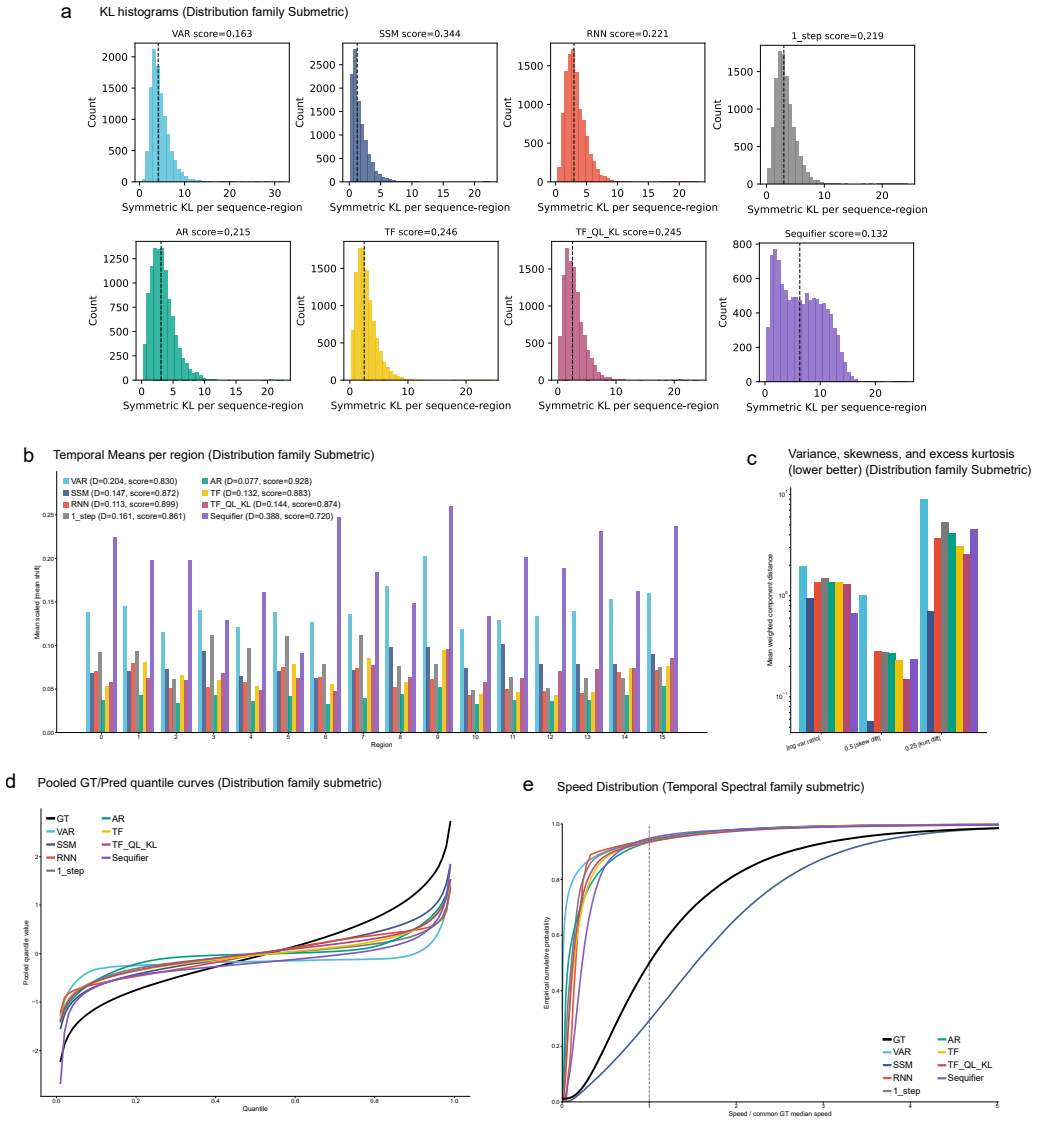


Figure S13: **Model-specific diagnostics underlying selected Nethobench submetrics (Distribution and Temporal-spectral families).** Each panel compares model predictions with matched ground-truth (GT) activity and displays component-related diagnostics. Panel annotations are diagnostic summaries and need not equal the retained aggregations in Table 7. **(a)** Distributions of symmetric KL divergence across sequences and regions. **(b)** IQR-scaled absolute temporal-mean shifts across regions. **(c)** Weighted diagnostic discrepancies in variance, skewness, and excess kurtosis; the retained pooled-moment score is defined in Appendix C.1. **(d)** Pooled GT and predicted quantile functions. **(e)** Empirical cumulative distributions of speed in the GT-defined latent space, normalized by the common GT median. Together, these diagnostics localize the statistical and dynamical differences summarized by the Nethobench Distribution and Temporal-spectral families.

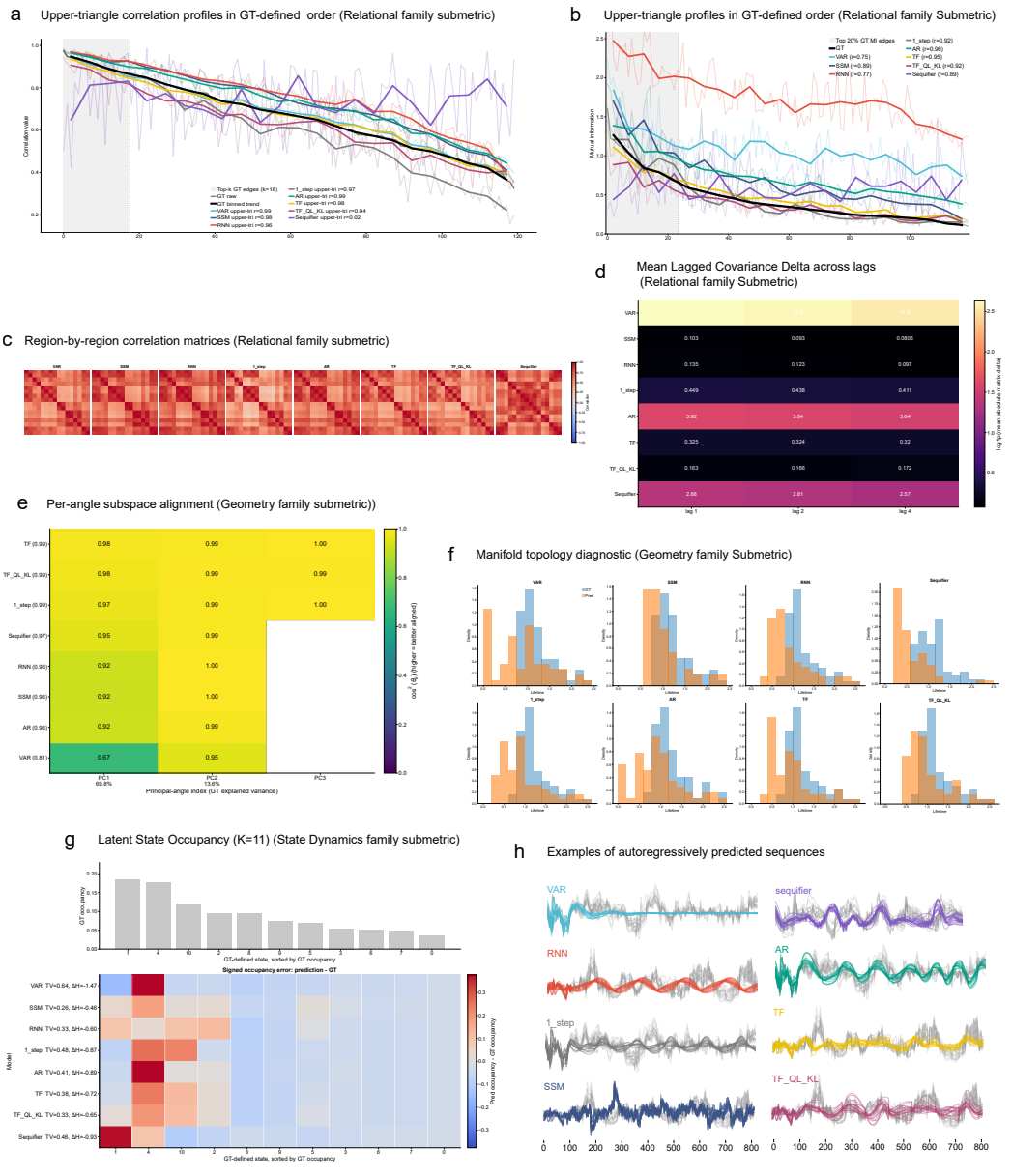


Figure S14: **Model-specific diagnostics underlying selected Nethobench submetrics (Relational, Geometry and State dynamics families).** Each panel compares model predictions with matched ground-truth (GT) activity and displays component-related diagnostics. Panel annotations are diagnostic summaries and need not equal the retained aggregations in Table 7. **(a)** Region-pair correlations ordered by GT edge strength. **(b)** Mutual-information profiles for the top 20% of GT edges, shown in GT-defined order. **(c)** Region-by-region correlation matrices. **(d)** Mean absolute lagged-covariance discrepancies at lags 1, 2, and 4. **(e)** Per-dimension subspace alignment, measured by the squared cosine of each principal angle. **(f)** Persistence-lifetime distributions for GT and predicted manifolds. **(g)** GT occupancy across $K = 11$ latent states and signed model occupancy errors in the same GT-defined state space. **(h)** Representative autoregressive rollouts. Together, these diagnostics localize the statistical and dynamical differences summarized by the Nethobench Relational, Geometry and State dynamics families.

Table 7: **Metric-level Nethobench scores by family and model.** Model values are mean ± sample standard deviation across three training seeds, after averaging four evaluation-group scores within each seed. Empirical reference values are mean ± Monte Carlo sample SD across 20 random draws, after averaging four fixed evaluation groups of 56/56/55/55 sequences within each draw; all draws use the same 222 targets and 720-bin forecast window. Pointwise fidelity is computed on the same forecasts and evaluation groups as the structural scores and is reported separately from the structural composite.

		Models								Empirical references	
Family	**Metric**	**VAR**	**RNN**	**SSM**	**1-step**	**Sequifier**	**AR**	**TF**	**TFQLKL**	**Real-real reference**	**Time+region shuffle reference**
Distribution	Symmetric KL	0.091 ± 0.003	0.117 ± 0.007	0.274 ± 0.026	0.120 ± 0.005	0.135 ± 0.008	0.116 ± 0.003	0.147 ± 0.010	0.143 ± 0.007	0.508 ± 0.043	0.402 ± 0.031
	Quantiles	0.396 ± 0.009	0.436 ± 0.016	0.499 ± 0.021	0.426 ± 0.009	0.461 ± 0.055	0.439 ± 0.008	0.454 ± 0.016	0.454 ± 0.014	0.721 ± 0.028	0.446 ± 0.020
	Moments	0.076 ± 0.008	0.260 ± 0.012	0.376 ± 0.032	0.204 ± 0.032	0.227 ± 0.063	0.272 ± 0.019	0.285 ± 0.024	0.295 ± 0.020	0.737 ± 0.083	0.478 ± 0.045
	Mean activity	0.816 ± 0.008	0.888 ± 0.012	0.858 ± 0.009	0.847 ± 0.007	0.732 ± 0.020	0.921 ± 0.011	0.871 ± 0.009	0.861 ± 0.009	0.917 ± 0.017	0.923 ± 0.007
	Family score	0.345 ± 0.005	0.425 ± 0.010	0.502 ± 0.020	0.399 ± 0.011	0.389 ± 0.010	0.437 ± 0.007	0.439 ± 0.013	0.438 ± 0.010	0.721 ± 0.038	0.562 ± 0.021
Temporal spectral	Trajectory distance	0.337 ± 0.022	0.500 ± 0.027	0.643 ± 0.014	0.462 ± 0.025	0.607 ± 0.031	0.463 ± 0.032	0.529 ± 0.019	0.542 ± 0.018	0.929 ± 0.021	0.472 ± 0.012
	Autocorrelation	0.624 ± 0.007	0.566 ± 0.006	0.434 ± 0.002	0.599 ± 0.006	0.650 ± 0.015	0.566 ± 0.004	0.616 ± 0.004	0.579 ± 0.005	0.866 ± 0.027	0.034 ± 0.001
	Power spectral density	0.066 ± 0.002	0.331 ± 0.022	0.429 ± 0.025	0.200 ± 0.008	0.222 ± 0.015	0.180 ± 0.010	0.296 ± 0.020	0.297 ± 0.016	0.734 ± 0.039	0.329 ± 0.013
	Family score	0.342 ± 0.008	0.466 ± 0.011	0.502 ± 0.013	0.421 ± 0.010	0.493 ± 0.010	0.403 ± 0.014	0.480 ± 0.009	0.473 ± 0.007	0.843 ± 0.013	0.278 ± 0.008
Relational	Graph structure	0.768 ± 0.036	0.460 ± 0.022	0.554 ± 0.056	0.751 ± 0.031	0.667 ± 0.047	0.783 ± 0.072	0.794 ± 0.025	0.757 ± 0.029	0.879 ± 0.078	0.307 ± 0.060
	Cross-region MI	0.662 ± 0.026	0.511 ± 0.009	0.642 ± 0.012	0.791 ± 0.009	0.734 ± 0.020	0.750 ± 0.015	0.809 ± 0.003	0.777 ± 0.012	0.912 ± 0.011	0.427 ± 0.037
	Lagged covariance	0.512 ± 0.019	0.671 ± 0.004	0.714 ± 0.023	0.644 ± 0.024	0.438 ± 0.079	0.405 ± 0.042	0.694 ± 0.014	0.616 ± 0.043	0.851 ± 0.021	0.381 ± 0.027
	VAR-response proxy	0.515 ± 0.061	0.343 ± 0.031	0.337 ± 0.039	0.499 ± 0.026	0.474 ± 0.029	0.480 ± 0.021	0.485 ± 0.041	0.565 ± 0.029	0.677 ± 0.107	0.470 ± 0.033
	Family score	0.614 ± 0.014	0.496 ± 0.012	0.562 ± 0.014	0.672 ± 0.011	0.578 ± 0.031	0.604 ± 0.024	0.696 ± 0.006	0.679 ± 0.009	0.830 ± 0.012	0.396 ± 0.013
Geometry	Manifold geometry	0.604 ± 0.041	0.532 ± 0.020	0.674 ± 0.027	0.637 ± 0.029	0.583 ± 0.032	0.569 ± 0.011	0.520 ± 0.024	0.519 ± 0.021	0.821 ± 0.057	0.746 ± 0.080
	Subspace angle	0.910 ± 0.023	0.988 ± 0.009	0.948 ± 0.023	0.833 ± 0.038	0.743 ± 0.125	0.859 ± 0.024	0.909 ± 0.031	0.836 ± 0.041	0.966 ± 0.028	0.134 ± 0.026
	Family score	0.757 ± 0.020	0.760 ± 0.012	0.811 ± 0.024	0.735 ± 0.014	0.663 ± 0.073	0.714 ± 0.016	0.714 ± 0.009	0.678 ± 0.028	0.893 ± 0.041	0.440 ± 0.037
State dynamics	State occupancy ($K = 11$)	0.346 ± 0.008	0.652 ± 0.033	0.658 ± 0.031	0.473 ± 0.032	0.490 ± 0.072	0.511 ± 0.012	0.549 ± 0.009	0.552 ± 0.022	0.881 ± 0.020	0.529 ± 0.050
	State occupancy ($K = 12$)	0.328 ± 0.029	0.588 ± 0.018	0.638 ± 0.017	0.465 ± 0.040	0.496 ± 0.046	0.472 ± 0.034	0.535 ± 0.038	0.538 ± 0.031	0.871 ± 0.003	0.554 ± 0.069
	State transition, lag 1 ($K = 11$)	0.361 ± 0.009	0.652 ± 0.029	0.696 ± 0.031	0.490 ± 0.030	0.520 ± 0.070	0.533 ± 0.011	0.568 ± 0.008	0.572 ± 0.022	0.912 ± 0.023	0.462 ± 0.024
	State transition, lag 2 ($K = 11$)	0.339 ± 0.010	0.602 ± 0.023	0.682 ± 0.031	0.457 ± 0.030	0.494 ± 0.071	0.502 ± 0.006	0.536 ± 0.007	0.539 ± 0.020	0.897 ± 0.022	0.481 ± 0.026
	State transition, lag 3 ($K = 11$)	0.325 ± 0.011	0.570 ± 0.021	0.668 ± 0.031	0.436 ± 0.029	0.476 ± 0.071	0.481 ± 0.006	0.516 ± 0.007	0.517 ± 0.018	0.882 ± 0.020	0.488 ± 0.031
	Family score	0.340 ± 0.006	0.613 ± 0.018	0.668 ± 0.028	0.464 ± 0.029	0.495 ± 0.064	0.500 ± 0.005	0.541 ± 0.012	0.544 ± 0.018	0.888 ± 0.017	0.503 ± 0.039
Overall	**Composite**	0.482 ± 0.006	0.544 ± 0.006	0.602 ± 0.018	0.541 ± 0.007	0.522 ± 0.030	0.532 ± 0.009	0.576 ± 0.006	0.564 ± 0.009	0.830 ± 0.012	0.439 ± 0.007
	Pointwise fidelity	0.354 ± 0.004	0.357 ± 0.003	0.309 ± 0.003	0.370 ± 0.003	0.324 ± 0.023	0.363 ± 0.003	0.364 ± 0.001	0.364 ± 0.002	–	–

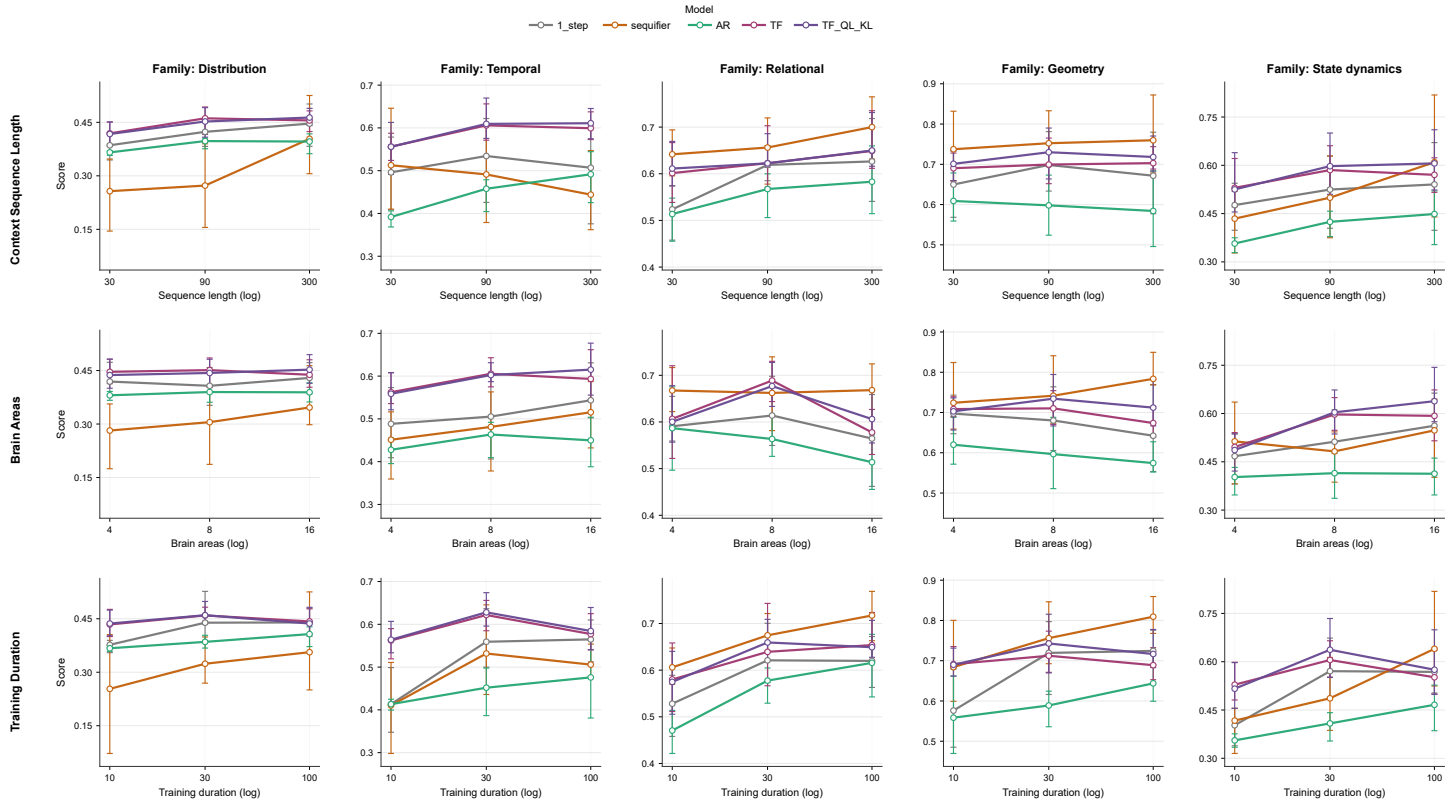


Figure S15: **Family-level sensitivity to experimental scale.** Nethobench family scores for five Transformer configurations as functions of context length (30, 90 and 300 samples), number of areas (4, 8 and 16), and training duration (10, 30 and 100 epochs). These are descriptive trends within the evaluated configurations. Changes with increased scale are family-specific and model-dependent: some patterns are stable across axes, while others sharpen or weaken with larger temporal context, system dimensionality, or training budget.

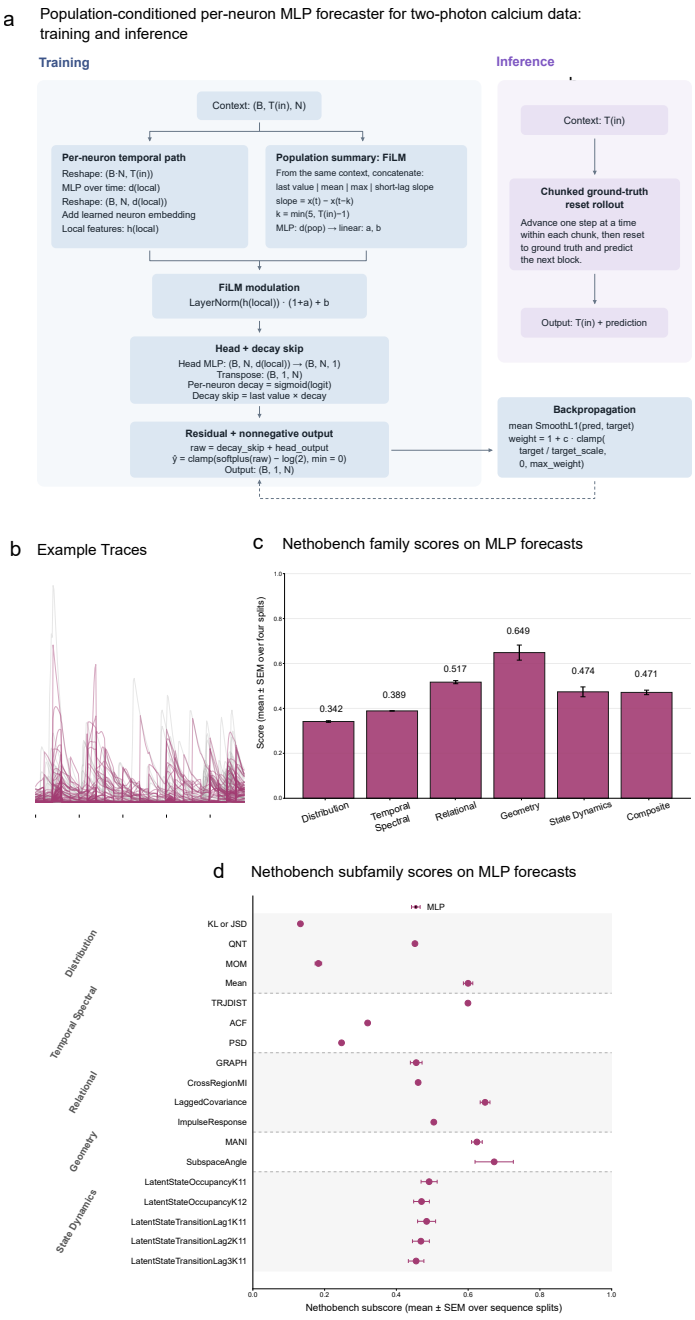


Figure S16: **Two-photon MLP baseline and Nethobench evaluation.** **(a)** Schematic of the MLP forecaster, which combines per-neuron temporal features with a population-level context embedding and predicts the next calcium-activity sample through a residual head plus learned persistence skip. Chunked inference: the one-step predictor advances autoregressively within each chunk, then the context is reset to recorded activity before the next chunk. This is not a continuously autonomous long-horizon forecast. **(b)** Example calcium-activity traces. **(c)** Family-level Nethobench scores and composite (0.471), with mean ± SEM across four sequence splits. Geometry (0.649) exceeds distributional (0.342) and temporal-spectral (0.389) agreement. **(d)** Nethobench submetric profile across all five families. Split uncertainty is within one mouse; it does not estimate between-animal variation (Appendix B.5).

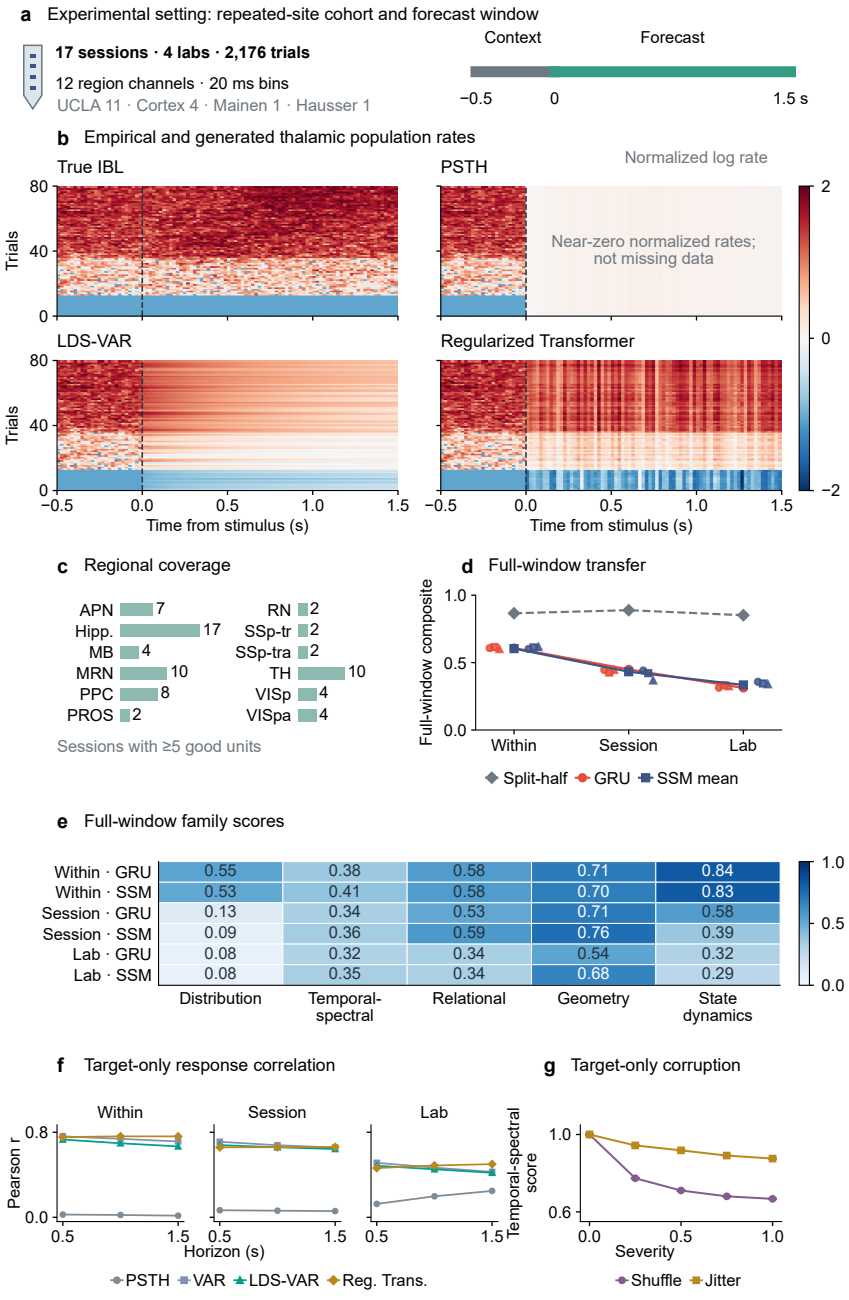


Figure S17: **IBL Neuropixels population forecasts and transfer: structural agreement declines under transfer.** (a) Seventeen sessions, four laboratories and 2176 trials. (b) Identically ordered regional-rate maps; white PSTH entries are finite near-zero normalized rates. (c) Regional coverage requires five quality-filtered units; missing channels are zero-filled before normalization. (d–e) Full-window transfer composites and profiles, including context. Small markers show three training seeds; connected points score ensemble-mean predictions. (f) Target-only Pearson correlations by horizon. (g) Target-only order controls on 436 trials at 50 Hz; bars show SD over three corruption realizations. Shuffle severity controls selected positions, and maximum channel jitter is nine bins. Sampling supports, controls and model identities are detailed in Appendix B.4.

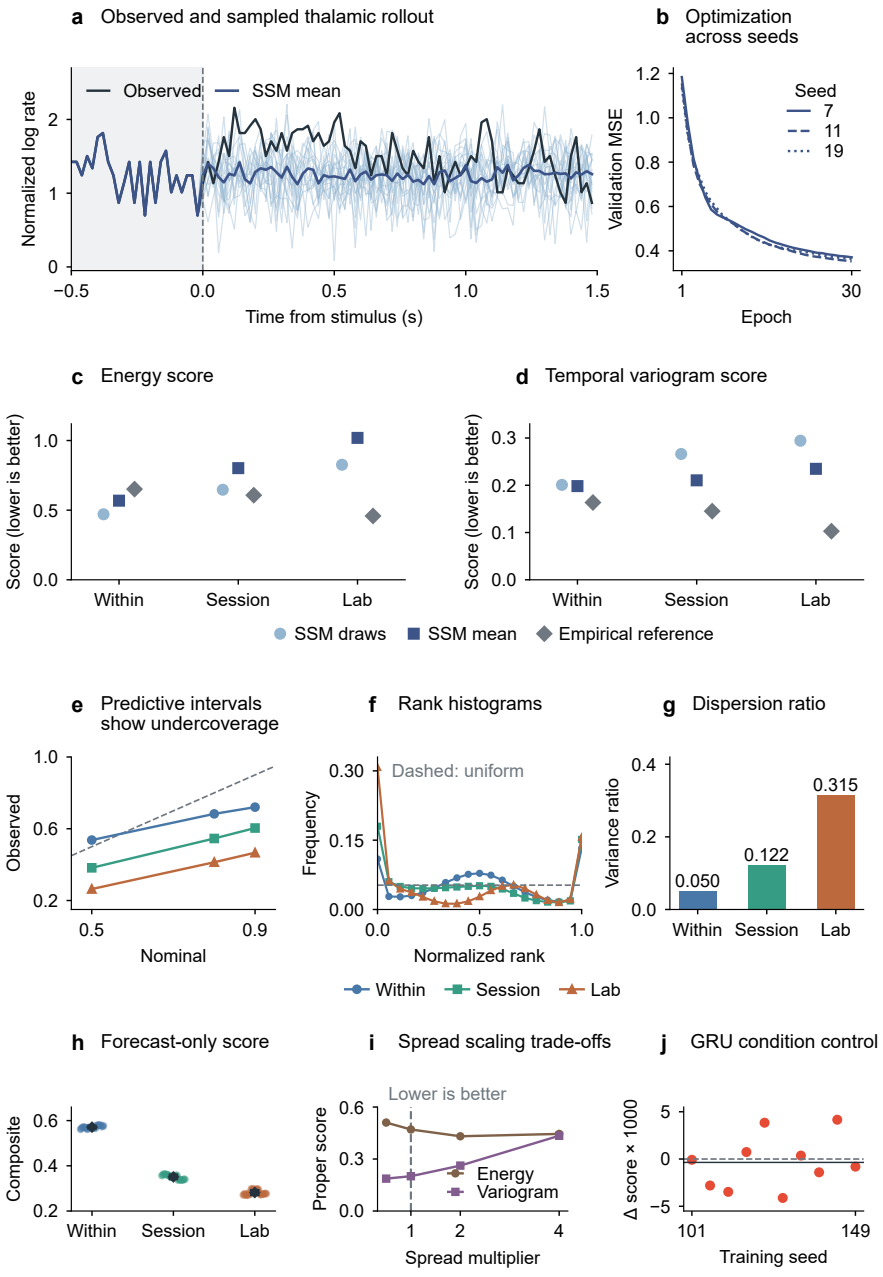


Figure S18: **Stochastic IBL Neuropixels forecasts, calibration and conditioning controls.** (a–b) First stored example and three SSM training histories (75 target bins × 12 regions). (c–d) Target-only energy and temporal variogram scores for 18 draws, their mean, and other-real-target reference ensembles. Draws improve energy versus the mean; variogram errors exceed the reference. (e–g) Coverage, ensemble ranks and variance ratios; pooled target variance includes between-context variation. (h) Eighteen structural Monte Carlo replicates (mean ± SD). (i) Spread sensitivity in each metric's own units. (j) Ten paired GRU conditioning seeds: correct-minus-shuffled mean composite -0.000358, exact $p = 0.6914$. These covariates and models do not establish a general absence of conditioning benefits (Appendix B.4).

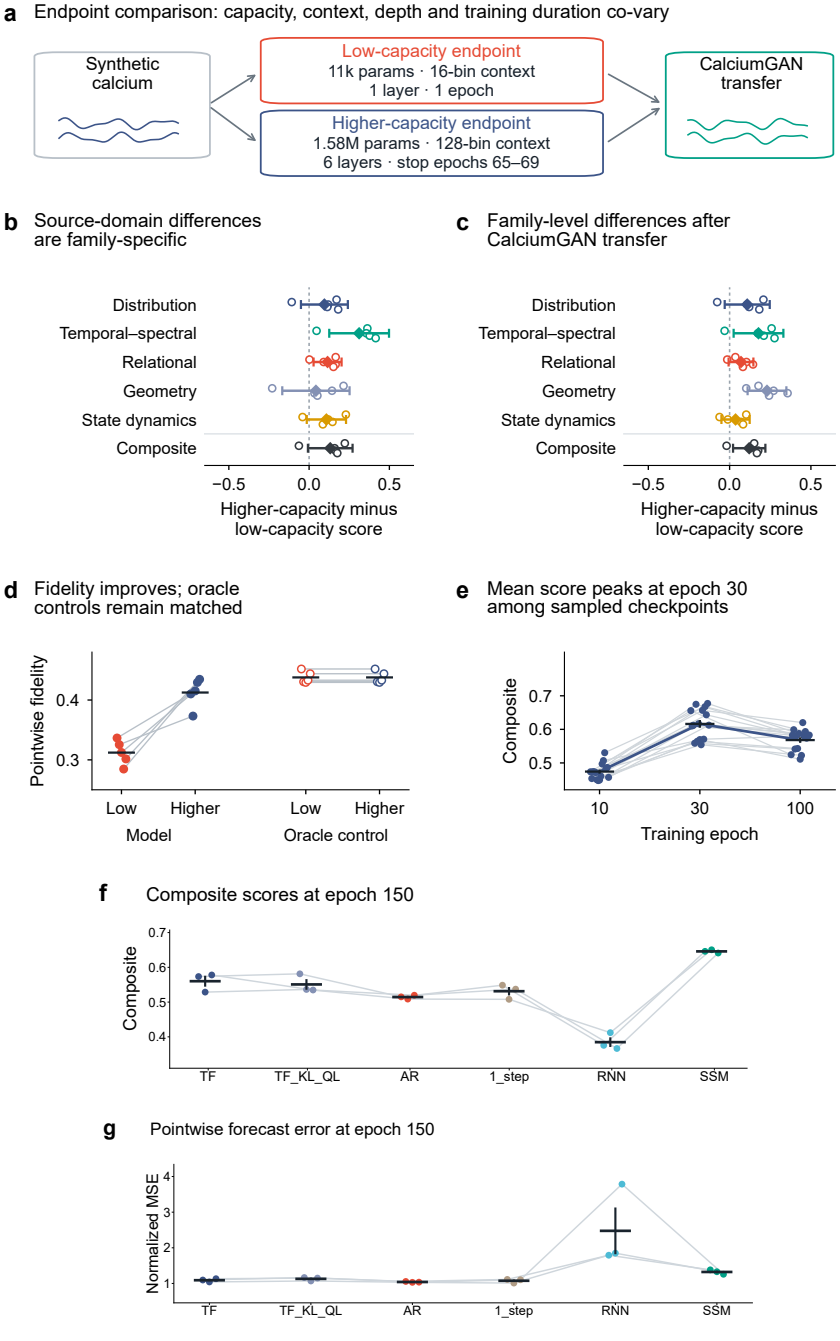


Figure S19: **Nethobench scores differ across model endpoints, training stages and model families.** (a) Endpoint settings co-vary in capacity, context, depth and training; traces are schematic. (b–c) Five paired source/CalciumGAN-transfer differences on matched generated targets. (d) Source-domain pointwise-fidelity controls (the same five paired seeds as b–c; filled/open markers show model/oracle scores). (e) Fifteen paired training seeds: epoch 100 exceeds 10 in all seeds, whereas 30 exceeds 100 in 14/15 (Holm-adjusted $p = 0.000183$). (f–g) Six model families evaluated at epoch 150 on the same validation sequences: (f) composite score and (g) normalized forecast MSE. Points show three training seeds, lines connect matched seeds, and bars show seed mean ± SEM. Model capacity and compute are not matched; CalciumGAN is a generated domain (Appendix C.4).

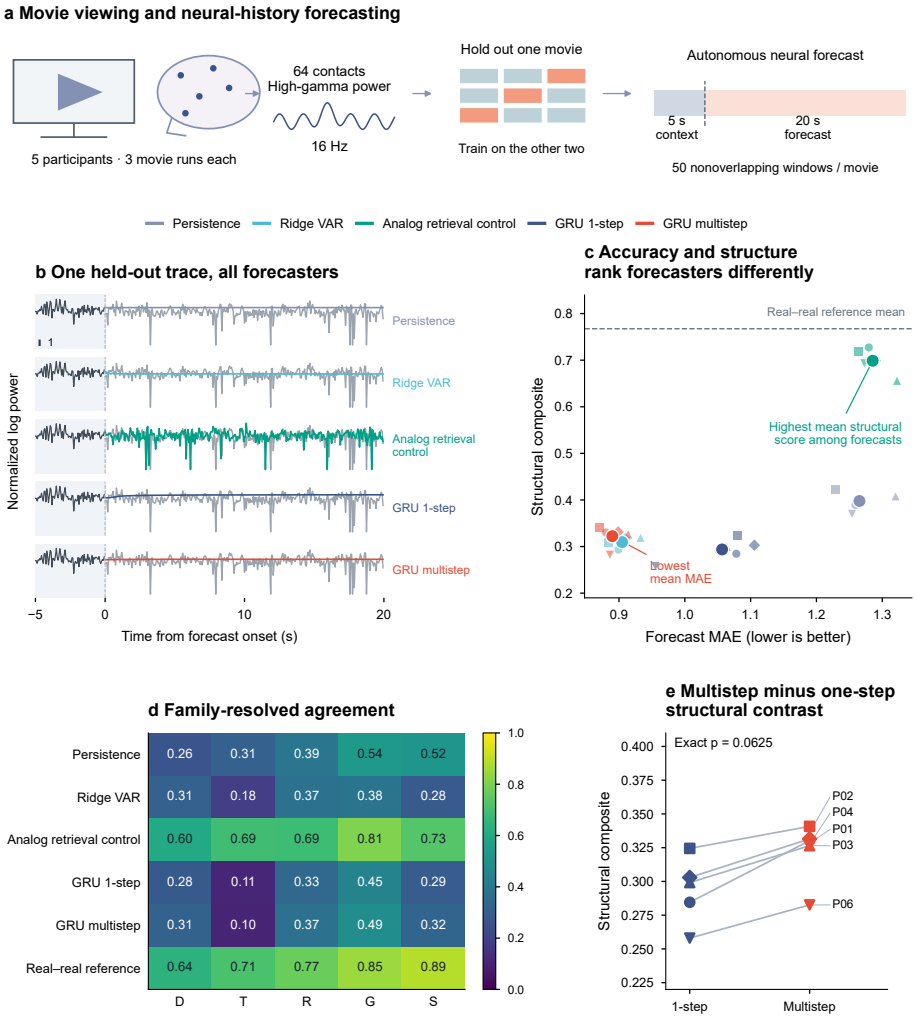


Figure S20: **Brain Treebank human neural forecasts separate structural agreement from pointwise accuracy.** (a) Five participants, high-gamma activity during naturalistic movie viewing (three movie runs per participant) and 64 contacts per participant, 5-s context and 20-s forecast. The task/contact drawings are illustrative. (b) A fixed first-participant/run/window/contact example, without smoothing. (c–d) Participant means and family profiles: training-only analog retrieval leads structural agreement, whereas multistep GRU has the lowest mean MAE. The real–real line is an empirical reference. (e) Paired GRU participant means: $\Delta C = 0.0285$, hierarchical 95% interval $[0.0180, 0.0415]$, exact two-sided $p = 0.0625$. Structural advantage does not establish conditional correctness, mechanisms or downstream utility. Retrieval exclusions and preprocessing boundaries are detailed in Appendices B.3 and B.6.

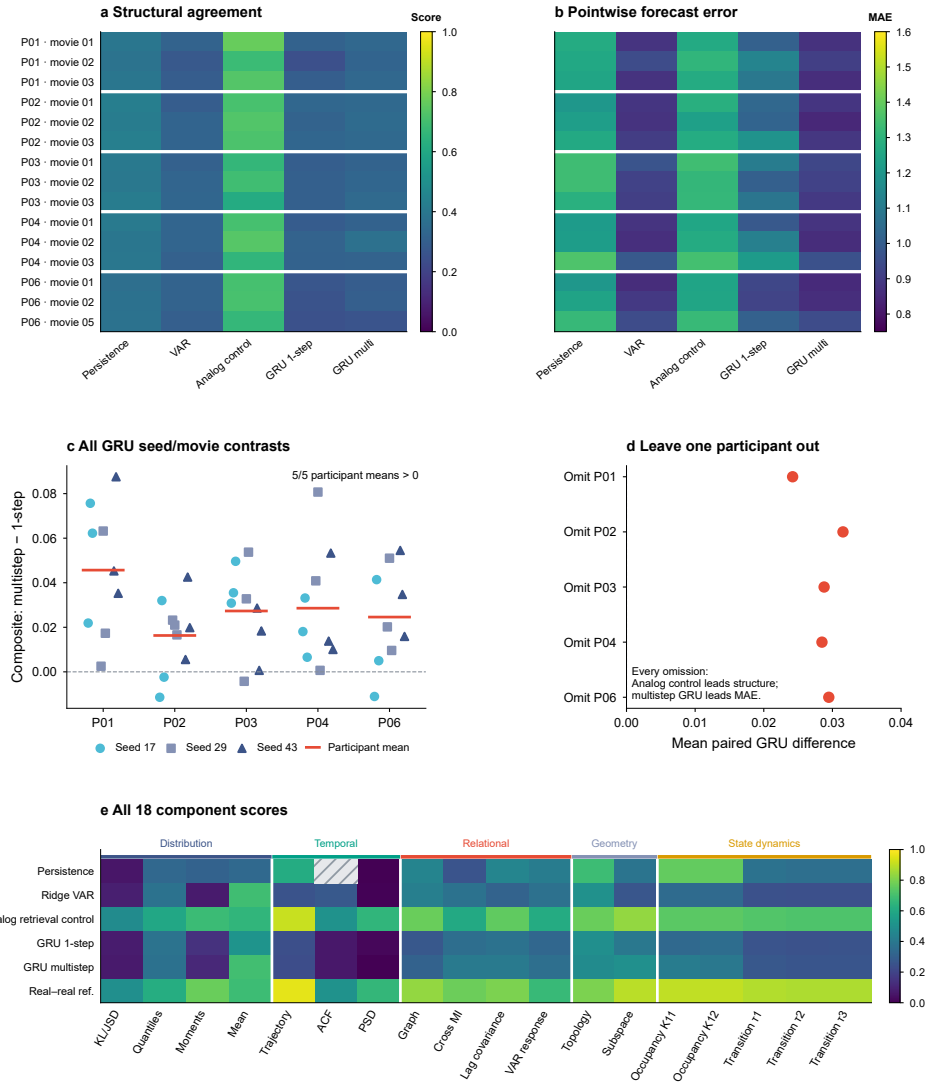


Figure S21: **Human forecast robustness by movie, seed, participant and component (Brain Treebank).** (a–b) All 15 held-out folds, averaging GRU seeds; 135 model/fold/seed evaluations underlie the summaries. Analog retrieval leads structure in every fold; multistep GRU leads MAE in 13 and VAR in two. (c) All 45 paired GRU seed/movie contrasts, including negative values. (d) Leave-one-participant-out contrasts. (e) All 18 components and the real-real reference. Persistence ACF is undefined; excluding ACF from every method preserves ordering and a positive mean GRU contrast (0.0280). The independent sample size is five participants, not 15 folds or 45 contrasts.

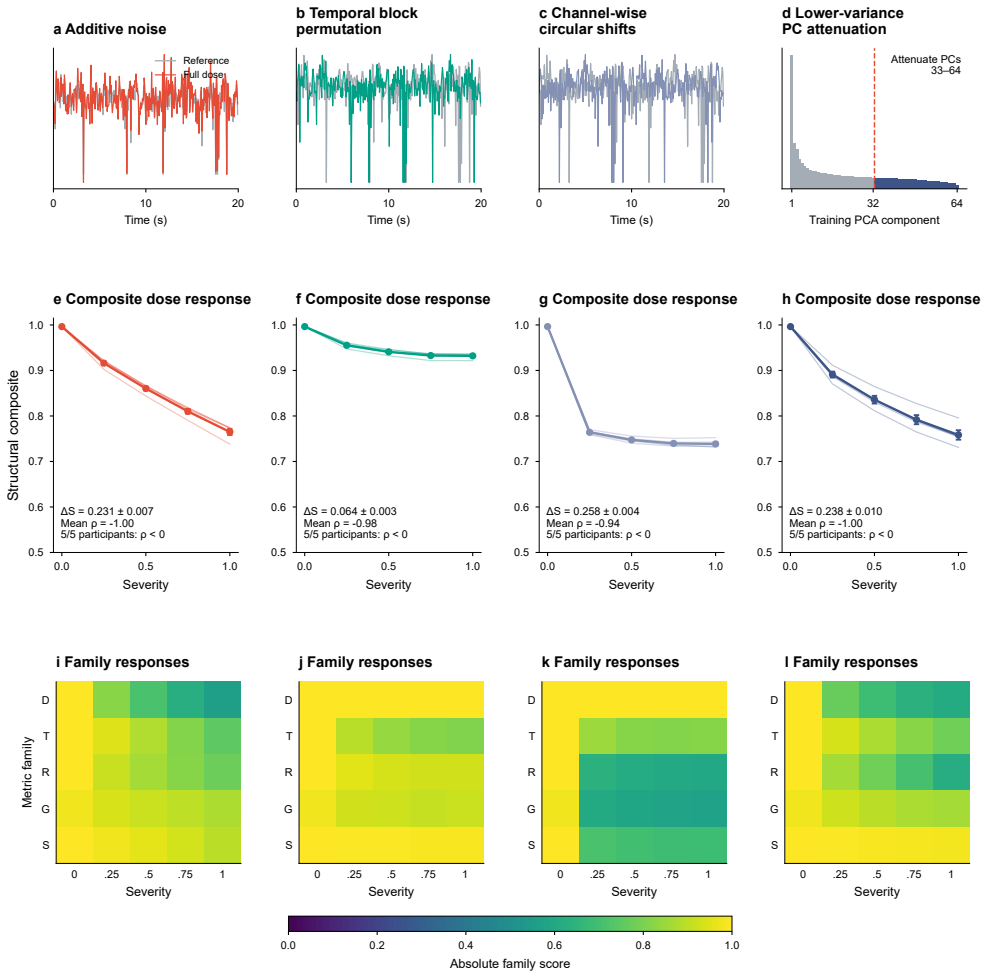


Figure S22: **Controlled perturbations of human neural activity from Brain Treebank.** (a–c) Full-dose noise, 2-s block permutation and channel shifts applied to a fixed held-out trace. (d) Training-PC spectrum; attenuation affects PCs 33–64, carrying 21.3–40.3% of training variance across folds. (e–h) Participant dose curves and mean ± participant SEM across all 300 evaluations. Full-dose drops are 0.231, 0.064, 0.258 and 0.238. All 20 participant/type dose correlations are negative, without implying strict monotonicity at every step. (i–l) Family profiles on common scales. Block permutation preserves samples and within-block paths. These controls do not establish preservation of arbitrarily small-variance or task-specific information.

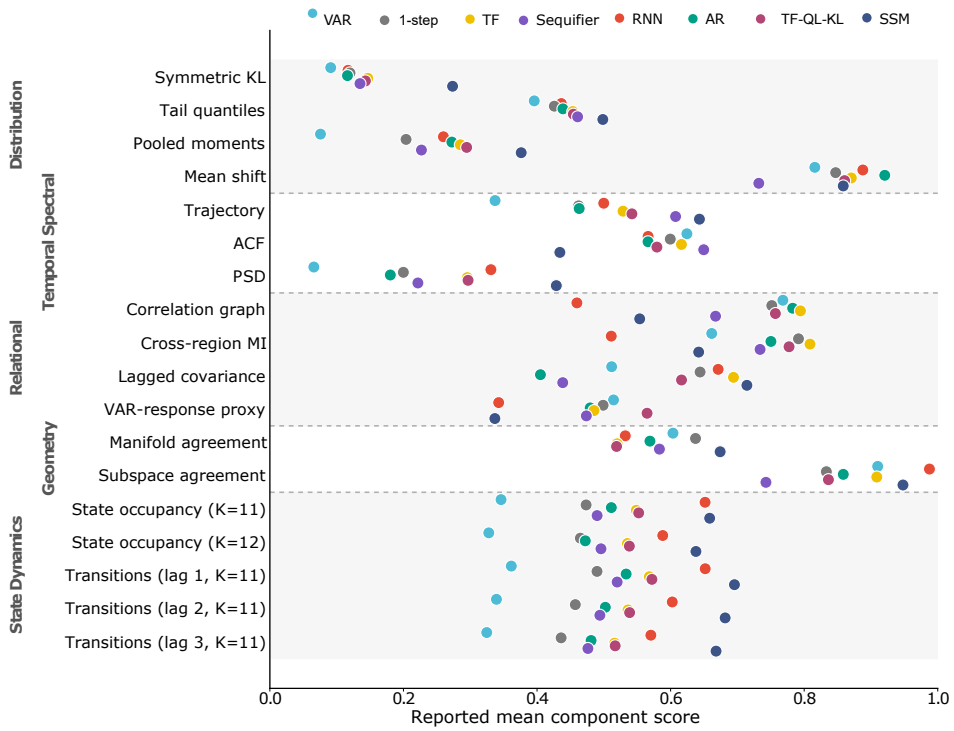


Figure S23: **Component-level audit of the principal widefield model panel.** Points are the reported model means; components are grouped by the five families.

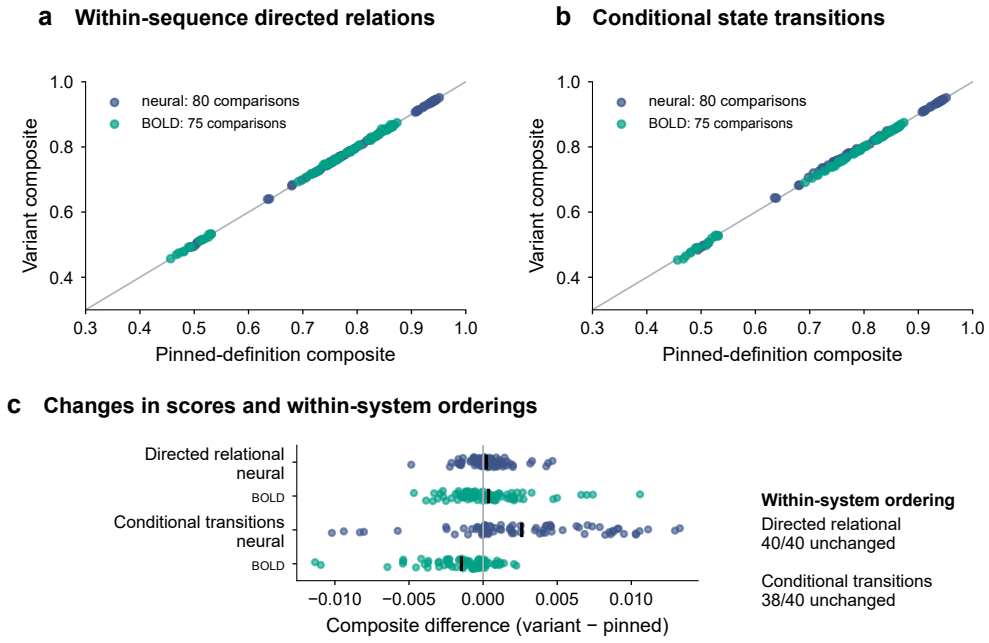


Figure S24: **Implementation sensitivity on fixed known-process controls.** (a) Sequence-safe lagged covariance/VAR comparisons using all directed off-diagonal entries. (b) Row-conditional state-transition comparisons with unchanged occupancy components. Each point is one fixed control comparison, not an independent biological replicate: 80 neural and 75 simulated-BOLD comparisons. (c) Composite changes for the two variants; vertical marks are means and points retain every comparison. The directed-relational variant preserves candidate ordering in all 40 system-by-seed groups; the conditional-transition variant preserves 38/40. These finite-panel checks do not establish unchanged principal-model rankings. Definitions, missing-origin handling and verification tests appear in Appendix C.6.

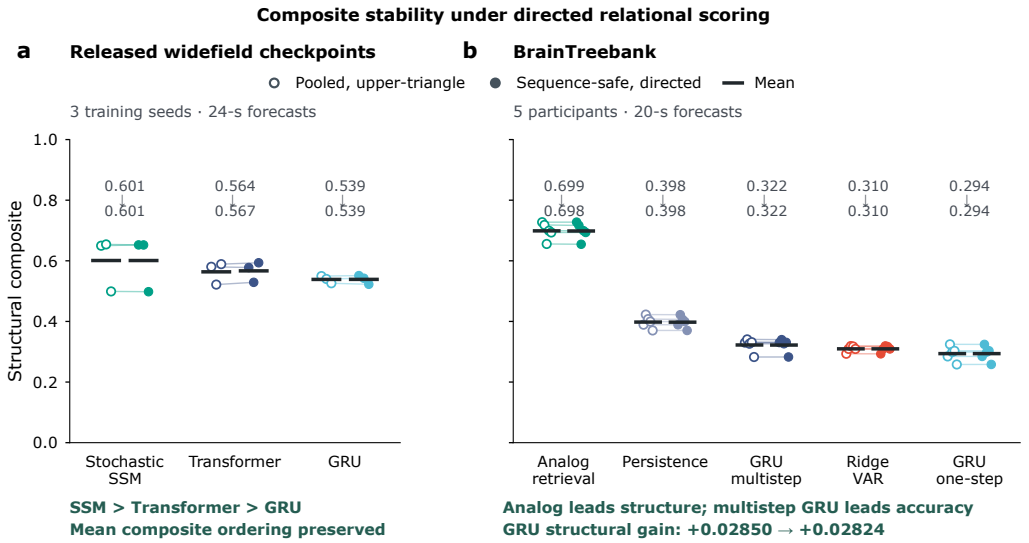


Figure S25: **Composite findings persist under sequence-safe, directed relational scoring.** Frozen forecasts were rescored using lag pairs restricted to their original sequences and both directed off-diagonal orientations in the lagged-covariance and ridge VAR(1) comparisons. Scaling, similarity transforms, all other components and aggregation weights were unchanged. Open circles show pooled, upper-triangle scoring, which flattens sequences and compares one off-diagonal matrix triangle; filled circles show sequence-safe, directed scoring, which retains within-sequence lag pairs and both off-diagonal orientations; lines pair the same evaluation unit, and black bars mark means. Numerical annotations give the pooled, upper-triangle mean followed by the sequence-safe, directed mean. ** (a) Released widefield checkpoints: three training seeds per model, evaluated on 58 validation sequences with 3-s context and 24-s forecasts. Each SSM checkpoint averages the scores of four fixed latent draws. Mean composite ordering remains SSM $>$ Transformer $>$ GRU: 0.60097 $\rightarrow$ 0.60094, 0.56362 $\rightarrow$ 0.56682 and 0.53856 $\rightarrow$ 0.53873, respectively. (b) BrainTreebank: five participant means, averaging training seeds within held-out movie and then three movies within participant (135 model/fold/seed evaluations; 5-s context and 20-s forecasts). The complete mean composite ordering is preserved. Analog retrieval remains highest in structural agreement (0.69878 $\rightarrow$ 0.69837); multistep GRU retains the lowest MAE (0.88990), which is unaffected by rescoring. Its composite advantage over one-step GRU changes from 0.02850 to 0.02824 and remains positive in all five participants. The structural and accuracy winners also persist in every leave-one-participant-out summary. Points represent training seeds in (a) and participants in (b); bars are means, not uncertainty intervals. These findings establish stability of the reported composite conclusions in the tested panels..