DECODING DYNAMIC VISUAL EXPERIENCE FROM CALCIUM IMAGING VIA CELL-PATTERN-AWARE SSL

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised learning (SSL) holds a great deal of promise for applications in neuroscience, due to the lack of large-scale, consistently labeled neural datasets. However, most neural datasets contain heterogeneous populations that mix stable, predictable cells with highly stochastic, stimulus-contingent ones, which has made it hard to identify consistent activity patterns during SSL. As a result, self-supervised pretraining has yet to show clear signs of benefits from scale on neural data. Here, we present a novel approach to self-supervised pretraining, POYO-SSL that exploits the heterogeneity of neural data to improve pretraining and achieve benefits of scale. Specifically, in POYO-SSL we pretrain only on predictable (statistically regular) neurons—identified on the pretraining split via simple higher-order statistics (skewness and kurtosis)—then we fine-tune on the unpredictable population for downstream tasks. On the Allen Brain Observatory dataset, this strategy yields approximately 12–13% relative gains over from-scratch training and exhibits smooth, monotonic scaling with model size. In contrast, existing state-of-the-art baselines plateau or destabilize as model size increases. By making predictability an explicit metric for crafting the data diet, POYO-SSL turns heterogeneity from a liability into an asset, providing a robust, biologically grounded recipe for scalable neural decoding and a path toward foundation models of neural dynamics.

1 INTRODUCTION

Learning useful representations from neural data poses a fundamental challenge for machine learning, as datasets from varied lab settings are not only small-scale but the signals themselves are complex, highly-dimensional, and only-partially observed (limitation of recording technology), while available labels are typically too scarce and weak for effective supervision. Self-supervised learning (SSL) offers a powerful paradigm to address this data scarcity, as it provides a way to learn from large amounts of data with limited access to labels, thereby allowing many datasets to be combined. This could be particularly useful for reconstructing perceptions or intentions directly from neural activity, e.g. for Brain-Computer Interfaces (BCIs).

However, successful self-supervised learning (SSL) fundamentally relies on exploiting statistical regularities within the data. For instance, objectives like masked modeling and sequence prediction are effective in the language domain precisely because language is inherently predictable, governed by robust statistical patterns and structural regularities (Harris, 1954; Tenney et al., 2019; Sinha et al., 2021; Yu et al., 2024; Lan et al., 2019; Li & Jurafsky, 2017). Neural decoding, in contrast, poses a unique challenge to this prerequisite of predictability. We only record a small, biased subset of neurons from the full circuit, creating a heterogeneous sample where predictability is not uniform. This unpredictability often correlates with cell type: inhibitory and corticothalamic neurons tend to exhibit more regular dynamics, while excitatory pyramidal cells appear sparser and more stochastic in isolation, partly because we lack access to the broader network signals that drive them. Training SSL models indiscriminately on this mixed-signal data is therefore counterproductive, as the loss becomes dominated by the unpredictable neurons, pulling the model's focus from the relevant and regular patterns it should be learning.

We test the **Statistical Regularity Hypothesis**: that self-supervised learning (SSL) efficiency scales with the statistical regularity of the selected neural subset. This principle is motivated by the ob-

servation that different neural populations, such as inhibitory interneurons and modulatory neurons exhibit fundamentally distinct statistical dynamics. Our hypothesis leads to a "data diet" approach for neuroscience SSL, where, unlike conventional methods that rely on task difficulty, we propose that the intrinsic statistical properties of neurons should guide the learning curriculum.

To validate this, we introduce **POYO-SSL**, a framework that uses higher-order statistics (skewness and kurtosis) as proxies for regularity to first pre-train on the most stable neural populations, overcoming prior methods' homogeneous treatment of heterogeneous populations. Our results confirm the hypothesis: by transforming neural heterogeneity from a challenge into an asset, this approach improves data efficiency by 1.98x and enables high-fidelity movie reconstruction directly from neural recordings, offering a principled, biologically-grounded recipe for scalable neural decoding.

Our contributions are threefold:

- We introduce a biologically-grounded pretraining paradigm that uses statistical regularity (rather than task-based difficulty) to guide data selection, selectively learning from neurons with highly regular responses first before training on more stochastic neurons.
- We present an end-to-end decoder architecture that transforms neural population activity into high-fidelity visual reconstructions, operating independently of external stimulus information.
- We demonstrate that functional heterogeneity, when properly leveraged through our regularity-based data diet, enables robust model scaling unlike conventional approaches that plateau with increased capacity.

Terminology 1. We refer to our setup as a *hybrid objective*, a simple form of curriculum learning (Bengio et al. (2009)). The primary objective is masked reconstruction on neural dynamics, while a supervised auxiliary cross-entropy on primitive stimuli serves as an "easy" initial step to stabilize training and prevent representational collapse. Importantly, no downstream labels are used during this pretraining phase.

Terminology 2. We define a neuron population as *predictable* from a self-supervised learning (SSL) perspective: its activity must contain sufficient statistical regularity for a model to successfully reconstruct masked portions of its signal. We empirically link this SSL-defined predictability to low skewness and kurtosis in calcium traces. Thus, while our definition aligns with the neuroscientific concept of stable firing patterns, it remains a fundamentally operational one, tied to the success of the masked reconstruction task.

2 Related Work

Decoding Models for Neuroscience Recent neural decoding models span diverse architectures and learning paradigms. Transformer-based approaches such as POYO (Azabou et al., 2023) and POYO+ (Azabou et al., 2024) enable multi-session learning but depend on full supervision, limiting scalability to unlabeled data. Self-supervised methods like CEBRA (Schneider et al., 2023) relax label requirements for single-session training but require labels for multi-session training. In visual reconstruction, fMRI-based frameworks have reached high fidelity (Chen et al., 2023; Joo et al., 2024) through masked modeling and large generative models, but rely on indirect stimulus-to-brain mappings from fMRI's slow hemodynamic signal. While these approaches set benchmarks for fMRI, direct comparison is challenging due to modality differences. In contrast, our method learns directly from neural recordings using the intrinsic structure of population dynamics without auxiliary labels or stimulus information.

SSL in Neuroscience Most self-supervised approaches to neural data assume population homogeneity and ignore functional specialization. Models such as Neuro-BERT Wu et al. (2022) treat all neurons equally, while contrastive or task-aware methods Song et al. (2023); Zhao et al. (2024) depend on external supervision rather than intrinsic circuit structure. These frameworks overlook that predictable neurons (inhibitory interneurons and modulatory pathways) differ fundamentally from stimulus-encoding neurons in computational role and temporal dynamics. Recent work by Johnson et al. (2022) characterized such heterogeneity through in vivo imaging, while our predictability-based selection offers distinct computational advantages by identifying and pretraining on regulatory neurons, enabling SSL to capture circuit-level dynamics and improving scalability beyond uniform population models.

Data-Centric SSL and Neural Heterogeneity Our approach aligns with the emerging "data diet" perspective in machine learning, which posits that the quality of pre-training data is as critical as its quantity (Paul et al., 2021; Zhuang et al., 2025). However, we distinguish our framework from these methods in a fundamental way: while standard approaches prune training *samples* (e.g., specific images or text), our strategy selects *neurons* (feature sources). In neural recordings, heterogeneity is intrinsic to the sensor array itself, not just the examples. We demonstrate that adding more neurons can paradoxically lead to a "scaling collapse"—a failure mode unique to heterogeneous neural populations. By selecting neurons based on statistical regularity, we resolve this collapse and transform heterogeneity from a liability into an asset for scaling.

3 METHODS

3.1 DATASET AND PARTITION

We use the Allen Brain Observatory (BO) calcium imaging dataset, featuring recordings from 13 Cre driver lines, which we partition into pretraining and finetuning sets (de Vries et al., 2020). To form the pretraining set, we identified a "predictable" subset by applying a knee-detection algorithm (Algorithm S1) to the per-line skewness and kurtosis distributions. This a priori process selected four lines (SST, VIP, PVALB, and NTSR1) that fell below the statistical knee—corresponding to major inhibitory interneuron classes and one modulatory excitatory line. To prevent data leakage, animals, sessions, and neuron IDs were kept strictly disjoint across all splits. Crucially, this design ensures that our model is evaluated on **novel biological subjects**. While the visual stimulus (movie clip) is shared across experiments, the neural population responses are animal-specific and unique to each session. Therefore, high performance on the test set reflects the model's ability to decode the generalized neural code rather than memorizing stimulus-response pairs. This statistical partitioning is empirically validated by its correspondence to neurons with regular firing patterns, aligning our data-driven approach with established neuroscience principles. Finally, to guarantee a fair comparison, we explicitly verified that all models (including baselines and ablations) were evaluated on this identical held-out test split.

3.2 CELL-PATTERN-AWARE SSL

3.2.1 DATA-EFFICIENT SELECTION CRITERIA

Pretraining	
Data Size	134 sessions, 80,146 samples
Selection Criteria	$skewness \leq 3.51$, $kurtosis \leq 22.62$
Hardware	4×V100 (KISTI `cas_v100nv_4`)
Fine-tuning (Movie decoding, Drifting Gratings)	
Selected Data	299 sessions, 1,170,931 samples
Selection Criteria	$skewness > 3.51$, $kurtosis > 22.62$
Frames (movie decoding)	900
Hardware	4×V100 (KISTI `cas_v100nv_4`)

Table 1: **Computationally-Efficient Pretraining** Summary of dataset scale (sessions and samples), predictable-neuron selection criteria (skewness and kurtosis computed on per-neuron $\Delta F/F$ traces over the full recording), and computational setup for pretraining and fine-tuning.
Notes. (1) Selected Data = number of predictable (pretraining) / unpredictable (finetuning) sesions / samples after skewness/kurtosis filtering. (2) For movie decoding, training batches preserved temporal order, whereas validation and test batches were randomly shuffled to evaluate generalization beyond temporal continuity.

We hypothesize that neurons showing **statistical regularity** are ideal for effective SSL pretraining. Within our framework, we operationally define this as *predictability*—the inherent structure enabling effective masked reconstruction. To identify these neurons without labels, we leverage per-neuron **skewness** and **kurtosis**. We refer to the selected subset as exhibiting **near-Gaussian activity**

(mean skewness 1.87, kurtosis 7.32), characterized by symmetric, **thin-tailed** distributions suitable for learning general features. In stark contrast, excluded neurons exhibit **heavy-tailed**, sparse bursting (mean kurtosis 148.51), better reserved for task-specific fine-tuning. For rigorous empirical validation of these metrics, see Appendix B.

To objectively partition the data, we applied a **knee-detection algorithm** (Satopaa et al. (2011)) to find a data-driven threshold across the 13 discrete CRE lines. Specifically, we identified the knee point on the sorted distribution of per-line mean statistics, establishing a cutoff based on cell-type categories rather than individual neuron scores. While this approach failed for lower-order statistics like event rate and Fano factor, it revealed a clear breakpoint for both skewness and kurtosis, providing a principled basis for our split. The resulting data-driven thresholds (skewness ≤ 3.51, kurtosis ≤ 22.62) identified a "predictable" subset comprising four CRE lines: **SST, VIP, PVALB, and NTSR1**. This statistically derived group is also biologically coherent, consisting of three major inhibitory interneuron classes and one regulatory corticothalamic excitatory line (NTSR1), all of which are crucial for stabilizing neural circuits. This convergence of statistical and biological criteria validates that our method effectively captures neurons showing statistically regular firing pattern. Crucially, these thresholds were determined *a priori* as a single, fixed criterion to partition the dataset, not as a tunable hyperparameter, which is why a sensitivity analysis was not performed.

3.2.2 Model Framework

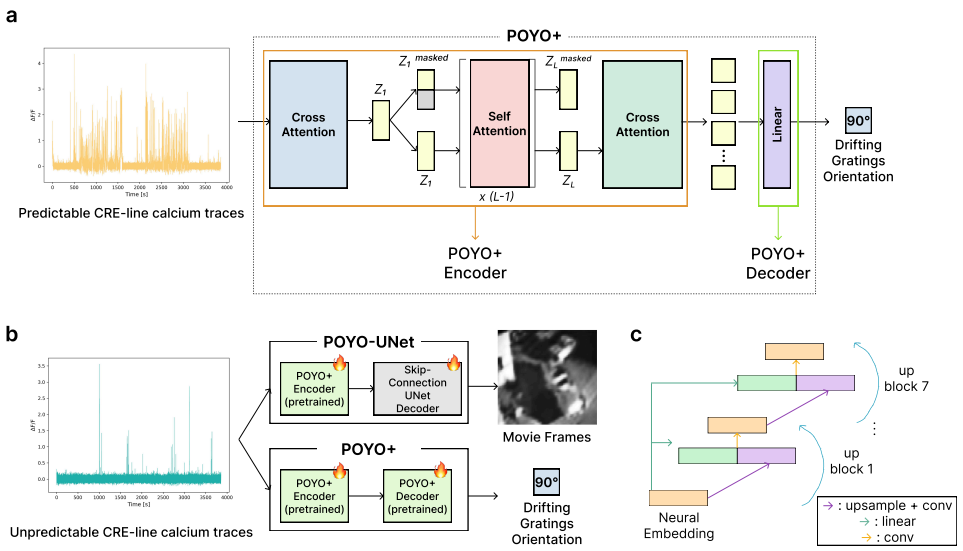

Figure 1: **Overall Framework of POYO-SSL. (a)** Pretraining strategy using predictable calcium traces with masked reconstruction learning (50% masking on temporal dimension). **(b)** Task-specific finetuning with unpredictable traces using either skip-connection UNet decoder (complex tasks) or original POYO+ decoder (simple tasks). **(c)** Skip-Connection UNet Decoder architecture replacing traditional encoder skip connections with neural embedding projections.

Predictable Neuron Pretraining with Auxiliary Classification We introduce a latent masked modeling approach to train our model: masked and an unmasked views of the same sample are fed independently through the encoder, the latent representation of the unmasked view is then used as target for the latent representation of the masked variant. To avoid representational collapse Grill et al. (2020); Chen et al. (2020), we use a supervised auxiliary loss. This auxiliary loss *bootstraps* early selectivity while masking-based reconstruction *shapes* representations for downstream decoding. The primitive labels also serve as guidance to stabilize early optimization.

Our architecture is based on the POYO+ Azabou et al. (2024) architecture: calcium traces are tokenized into a sequence of input tokens that are then compressed, using a cross-attention block, into a sequence of latent tokens, which we note $Z_1 = \{\mathbf{z_1^{(1)}}, \cdots, \mathbf{z_1^{(L)}}\}$, where L is the number of latent tokens and $\mathbf{z_1^{(i)}} \in \mathbb{R}^d$ is the latent embedding. Each latent token $\mathbf{z_1^{(i)}}$

has an associated timestamp relative to the context window. We introduce the following temporal masking scheme: we causally mask a percentage of the latent tokens to form $Z_1^{\text{masked}} = \{\mathbf{z_1^{(1)}}, \cdots, \mathbf{z_1^{(L-M)}}, <\text{MASKED}> \cdots, <\text{MASKED}>\}$. We selected a masking ratio of 50% empirically, i.e. the second half of the context window is masked. We use a siamese network (see Figure 1) to feed both Z_1 and Z_1^{masked} through the same self-attention blocks which yields Z_L and Z_L^{masked} respectively. Finally, we use Z_L as the target for Z_L^{masked}.

During pre-training, the model is trained on a joint objective, consisting of self-supervised masked reconstruction and fully-supervised classification of drifting grating orientations. This auxiliary classification task stabilizes the early training dynamics before the model focuses on the complex downstream movie decoding task.

The pre-training loss is as follows:

$$\text{Loss}_{\text{pretrain}} = \text{Loss}_{\text{L1}}(Z_L^{\text{masked}}, Z_L) + \lambda \cdot \text{Loss}_{\text{CrossEntropy}}(\text{DG}_{\text{predicted}}, \text{DG}_{\text{true}}) \tag{1}$$

where λ is a loss weight that we empirically found $\lambda = 0.01$ to be optimal through grid search ($\lambda \in 0.001, 0.01, 0.1$, with performance degrading by 7-11% for $\lambda < 0.001$ or $\lambda > 0.1$. We keep the cross-entropy weight small so CE accelerates convergence while masking drives representation formation. This hybrid objective operationalizes a curriculum learning strategy, where the simple auxiliary task provides a stable foundation for the more demanding masked reconstruction objective. Details are provided in Appendix D.

Task-Specific Fine-tuning on Unpredictable Neurons Finetuning uses unpredictable CRE-line traces with task-specific decoders. For classification and simple regression tasks such as drifting-grating orientation prediction, we use the POYO+ multi-task decoder, and for complex movie frame reconstruction we employ a dedicated vision-specialized Skip-Connection U-Net decoder.

The finetuning loss is as follows:

$$Loss_{movie} = 50\, Loss_{focal} + 50\, Loss_{L1} + 50\, Loss_{FFT} + Loss_{perceptual} + 0.1\, Loss_{SSIM} \tag{2}$$

$$Loss_{DG} = Loss_{CrossEntropy}(DG_{\text{predicted}}, DG_{\text{true}}) \tag{3}$$

Loss weights in Eq. 2 were determined through a systematic grid search over [0.1-100] using SSIM validation score. The different loss terms in the movie reconstruction loss corresponds to specialized components (Focal (Lin et al. (2017)), FFT (Fast Fourier Transform, (Zhao et al. (2016))), Perceptual (Johnson et al. (2016)), and SSIM (Wang et al. (2004))) that ensure high-fidelity image reconstruction. See Appendix H for details on each loss term.

Skip-Connection U-Net Decoder To address the challenge of reconstructing high-resolution movie frames, we designed a specializaed decoder, as this dense prediction task requires custom vision modules that were not designed in the POYO+ decoder. Our new U-Net-inspired decoder generates frames from a single neural embedding. In each upsampling stage, a direct projection of the latent vector (e.g., to $128 \times 2 \times 2$, $64 \times 4 \times 4$) is concatenated with the upsampled feature map and fused with a 1×1 convolution. These repeated latent injections are crucial for maintaining semantic information across all scales, enabling the faithful reconstruction of fine visual details from a compact neural representation. See Appendix G for more details.

3.3 Numerical Analysis

3.3.1 Loss Landscape Analysis

To understand the challenge of optimizing representation learning models on neural data, we projected neural activity onto its first two principal components (PCs) and approximated the reconstruction loss landscape. The loss at each grid point in the PC space was estimated using a k-nearest neighbor approach ($k = 5$), which considered the local variance of nearby data points and a distance penalty term. Landscapes were smoothed for visualization via a Gaussian filter ($\sigma = 1.0$) (Li et al. (2018)).

3.3.2 INFORMATION THEORY ANALYSIS

We used Fisher Information as a metric for data quality, where $I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log p(x|\theta)\right)^2\right]$ quantifies the amount of information each data point provides about underlying model parameters Amari (1998), with higher values indicating better parameter estimation and convergence. For a quasi-Gaussian signal, this can be approximated as the inverse of the signal variance ($I \approx 1/\sigma^2$). Based on this, we defined the **Effective Dataset Size** (D_{eff}) as the raw data size weighted by its quality, where a higher Fisher Information value corresponds to a larger effective size. This allows for a more accurate comparison of dataset utility beyond simple data point counts (Kaplan et al. (2020)).

3.4 REPRESENTATION ANALYSIS

We quantified the properties of the learned latent spaces using several metrics, including t-SNE for visualization, Intrinsic Dimension (ID) for efficiency (Levina & Bickel (2004)), and metrics to assess geometric dissimilarity and structural integrity. To assess dissimilarity between latent spaces learned by different models, we used Procrustes disparity (Dryden & Mardia (2016)) and Centered Kernel Alignment (CKA) (Kornblith et al. (2019)). To evaluate local structure, we used a Temporal Neighborhood Preservation score (Venna & Kaski (2001)) (see Appendix I for all definitions).

4 RESULTS

4.1 EXPERIMENTAL SETUPS AND BASELINES

To isolate the benefits of our cell-pattern-aware pre-training, we compare our main model, **POYO-SSL**, against a crucial baseline:

- **Supervised Baseline (From-Scratch):** To rigorously quantify the performance gains from our SSL stage, we compare against a baseline sharing an identical encoder–decoder architecture but trained end-to-end on the downstream tasks without pre-training.
- **Architecture Ablation Studies:** To disentangle the contributions of encoder representations and decoder capacity, we include three capacity-matched variants. Capacity matched means total parameters are within $\pm 3\%$ of our model.: (i) *MLP Encoder* $\rightarrow$ *MLP Decoder*, which maps neurons directly to pixels through a deep fully-connected network with no spatial inductive bias; (ii) *POYO Encoder* $\rightarrow$ *MLP Decoder*, which retains our SSL encoder but replaces the U-Net decoder with a purely linear decoder to test whether learned representations alone can drive performance; (iii) *POYO Encoder* $\rightarrow$ *U-Net Decoder without skip connections*, which preserves the U-Net hierarchy but removes lateral skip pathways to assess the importance of multiscale feature fusion.

We compare to POYO+ (Azabou et al., 2024) which is a state-of-the-art model. To benchmark against external SSL methods, we evaluated an adapted CEBRA baseline (Schneider et al., 2023) by training its encoder and feeding representations to our vision decoder. This yielded an SSIM of $\sim$0.48, confirming that contrastive latent spaces optimized for behavioral alignment do not transfer effectively to high-fidelity pixel generation. For CEBRA, we report the best performance between training from scratch and fine-tuning strategies. Regarding Neuro-BERT (Wu et al., 2022), the lack of an official implementation prevented a reproducible adaptation, and thus it was excluded.

4.2 EFFECT OF CELL-PATTERN-AWARE SSL

4.2.1 CELL-PATTERN-AWARE SSL ENABLES SMOOTH LOSS LANDSCAPE

Our analysis of the masked reconstruction loss landscape elucidates a fundamental dichotomy in the nature of the optimization problems presented by the two neural populations. Predictable neurons induce a geometrically well-posed landscape characterized by a smooth, convex-like surface (roughness $\sigma_L = 14.8546$), which is highly amenable to gradient-based optimization methods. In stark contrast, unpredictable neurons give rise to a treacherous, non-convex landscape (roughness $\sigma_L = 2048.4712$) plagued by a multitude of spurious local minima. Crucially, the quantitative contrast remains striking even with the expanded FOV: despite the inclusion of steep basin walls, the 'unpredictable' landscape remains $\sim$138$\times$ rougher than the 'predictable' one. This confirms that our conclusion is robust to the choice of scale: the structural optimization gap between the two

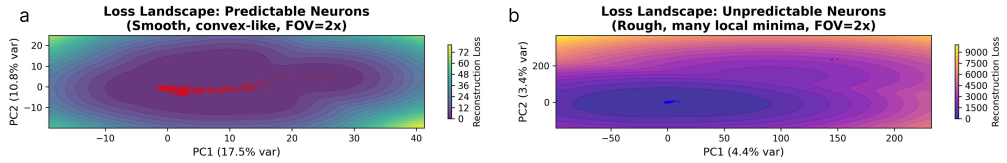

Figure 2: **Loss Landscape Topology Reveals a Dichotomy in Optimization Difficulty.** Masked reconstruction loss landscapes for predictable and unpredictable neurons, projected onto their first two principal components (PCs) with an expanded field of view (FOV=2x). **(a)** The landscape from predictable neurons is smooth and convex-like, clearly revealing high-loss boundaries that enclose data points (red dots) in a single basin, indicating a well-posed optimization problem. **(b)** In contrast, the landscape from unpredictable neurons is rugged and non-convex, characterized by numerous local minima, which presents a challenging, ill-posed problem.

populations is massive, regardless of the field of view. This topological difference explains why the pre-training task transforms from a simple optimization challenge to a complex, ill-posed problem, thereby providing a rigorous geometric basis for the superior performance of the predictable-first pre-training curriculum.

4.2.2 Predictable Neurons contains richer representation

Metric	Predictable	Unpredictable	Ratio (Pred./Unpred.)
Fisher Information (Data Quality)	64.51±0.55/-0.65	33.47±0.46/-0.35	**1.93x**
Data Quality Ratio (Efficiency)	34.41	17.39	**1.98x**
Effective Dataset Size	71.5 M	227.5 M	-

Table 2: **Information-Theoretic Analysis of Data Quality.** A quantitative comparison of predictable and unpredictable neural populations. The analysis reveals that predictable data is information-theoretically superior, providing a basis for its enhanced performance and scalability. Values are reported as mean ± 95% CI.

Our analysis revealed that predictable neural data is information-theoretically richer, which translates directly to greater data efficiency. We quantified this using **Fisher Information**, finding that the predictable dataset had a value of 64.5 compared to 33.5 for the unpredictable dataset, indicating that each predictable data point contains **1.93 times more information** for model training (Table 2). Consequently, while the raw dataset sizes were comparable, the quality-adjusted **Effective Dataset Size** (D_{eff}) was significantly larger for the predictable population, making each of its data points 1.98 times more efficient for training.

4.2.3 Cell-Pattern-Aware SSL achieves high performance

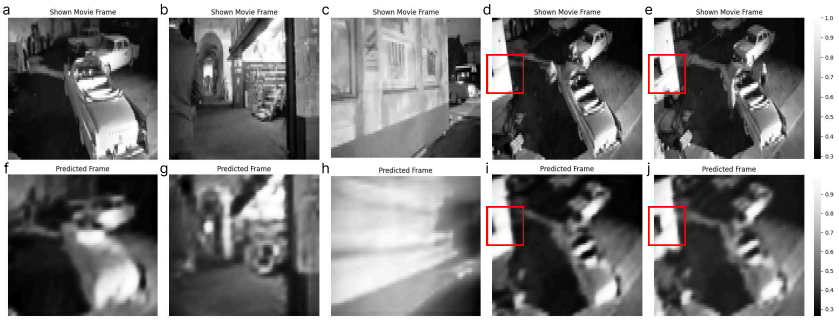

Figure 3: **End-to-end neural-to-vision decoding**. **(a-e)** depict movie frames presented to the mouse, **(f-j)** depict reconstructed frames. Our model captures subtle frame-to-frame variations (red boxes), demonstrating true reconstruction rather than frame memorization.

Method	Pretrain Data	Finetune Data	Movie SSIM↑	DG Accuracy↑
POYO-SSL (Ours)	Predictable	Unpredictable	**0.593±0.013**	**0.555±0.022**
Baseline: Train on All	N/A (From Scratch)	All (Pred. + Unpred.)	0.528±0.023	0.492±0.041
Architecture Ablation Studies				
MLP Enc.→MLP Dec.	Predictable	Unpredictable	0.449±0.022	–
POYO+ Enc.→MLP Dec.	Predictable	Unpredictable	0.503±0.019	–
POYO+ Enc.→UNet Dec. without skip connection	Predictable	Unpredictable	0.466±0.047	–
CEBRA Enc.→UNet Dec.	Predictable	Unpredictable	0.481±0.010	–
Data-Selection Ablation Studies				
Inhibitory-only SSL	Inhibitory	Excitatory	0.544±0.030	0.537±0.025
Reverse SSL	Unpredictable	Predictable	0.489±0.032	0.213±0.037
Mixed SSL	Unpred. + partial Pred.	Unpredictable	0.543±0.049	0.313±0.012
Random subset SSL	Random (Size-matched)	Remaining	0.532±0.044	0.254±0.011
Pretraining Objective Ablation Studies				
Random Masking Loss	Predictable	Unpredictable	0.540±0.017	0.548±0.028
Masking Loss only	Predictable	Unpredictable	0.496±0.050	0.099±0.019
Large CE weight (0.1)	Predictable	Unpredictable	0.552±0.052	0.482±0.033
Small CE weight (0.001)	Predictable	Unpredictable	0.532±0.042	0.469±0.015
Cross-Entropy Loss only	Predictable	Unpredictable	0.506±0.057	0.452±0.026

Table 3: **Performance comparison across multiple visual decoding tasks.** Our proposed framework, POYO-SSL, consistently outperforms baseline models, demonstrating the effectiveness and generalizability of cell-pattern-aware self-supervised learning. Best results are shown in bold. Movie decoding task is denoted as movie, drifting gratings decoding task is denoted as DG. Values are depicted as mean ± 95% CI across three seeds (with $p<0.05$ (paired t-test)). Dashes indicate tasks not applicable to image-only decoders.

As shown in Table 3, our cell-pattern-aware pretraining delivers significant performance gains across diverse downstream tasks, demonstrating the generalizability of the learned representations. On the complex movie decoding task, our approach achieves SSIM score of 0.593 for direct neural-to-visual reconstruction. This high fidelity reflects genuine reconstruction capabilities rather than simple pattern memorization, as the model successfully captures subtle frame-to-frame variations (Figure 3). Equally notably, on the drifting-gratings classification task, it reaches 55.5% accuracy, substantially outperforming the from-scratch baseline (49.2%). This dual success underscores that our pretraining strategy is effective for both high-fidelity generative tasks and classification challenges.

Ablation studies highlight the benefits of our approach, indicating that both the architecture and the learning objective tailored to the data's statistics are important factors. The superior performance of temporal masking over random masking underscores the value of the objective and lends functional support to our selection criteria. Temporal masking preserves local temporal dependencies critical for neural dynamics (typically 50-100ms receptive fields in V1 neurons), while random masking disrupts these patterns. This result suggests the curated neurons ("predictable" neurons) indeed possess the predictable temporal structure that a specialized task can effectively exploit. Furthermore, data-selection ablations indicate that data quality can outweigh quantity; reversing the curriculum to pretrain on unpredictable neurons leads to worse performance than training from scratch, suggesting that pretraining on highly stochastic data may establish a less effective inductive bias for downstream learning. Overall, our approach of selectively pretraining on neurons with regular firing patterns leverages population heterogeneity to enable stable and scalable representation learning.

4.2.4 The Representational Advantage of Cell-Pattern-Aware SSL

Analysis of the learned representations reveals a stark contrast between the strategies (Figure 4). Qualitatively, t-SNE visualizations show that our POYO-SSL model learns a well-structured manifold that captures the data's temporal continuity, while baseline approaches like reverse SSL and from-scratch training yield disorganized or collapsed representations. This visual observation is

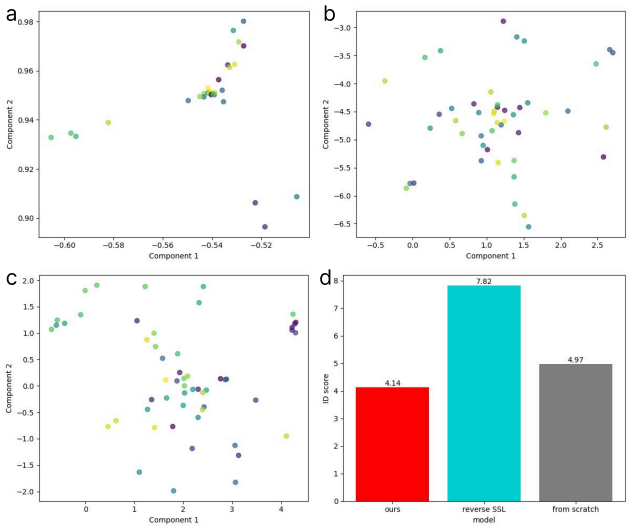

Figure 4: **POYO-SSL learns a more efficient and structured latent manifold. (a-c)** t-SNE visualization of latent spaces from our POYO-SSL model (a), a reverse SSL model (b), and a from-scratch model (c). Point color reflects the temporal progression of frames. **(d)** Quantitative comparison of the Intrinsic Dimension (ID) for each model.

validated by multiple quantitative metrics. Our model's latent space is more efficient, with a significantly lower intrinsic dimension (ID) of 4.14 compared to the from-scratch (4.97) and reverse SSL (7.82) models. It also better preserves local temporal structure, evidenced by a higher Temporal Neighborhood Preservation score (0.2355 vs. 0.1584 and 0.0960). Furthermore, high Procrustes disparity (>0.98) and low Centered Kernel Alignment (CKA,≈0.13) confirm that the methods learn fundamentally different feature spaces. Taken together, these results demonstrate that our selective pre-training is crucial for learning a concise and structured representation of the neural code.

4.3 POYO-SSL Enables Stable Model Scaling

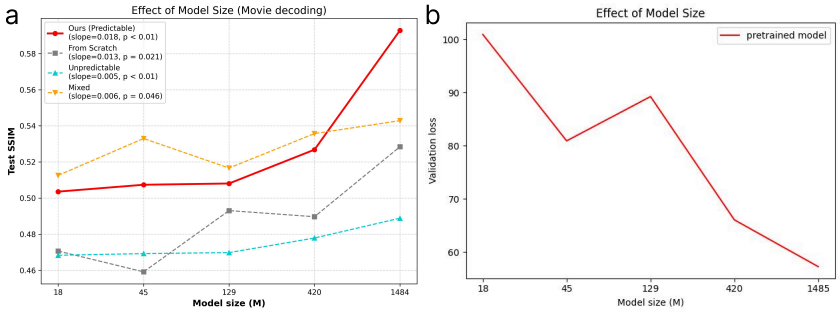

Figure 5: **Pre-training with predictable neurons is crucial for effective model scaling. (a)** Test SSIM performance versus model size for different pre-training strategies. Only the model pretrained exclusively on predictable neurons (red) demonstrates robust, positive scaling with model capacity (slope=0.018, $p < 0.01$ under bootstrap analysis). In contrast, training from scratch (gray) or including unpredictable neurons in pre-training (cyan, yellow) leads to flat or erratic scaling (slopes ≈ 0.005–0.013). **(b)** Corresponding validation loss during pre-training on the predictable set, showing a general downward trend that indicates successful learning.

A key advantage of our framework is its ability to enable stable model scaling, a critical property for building more powerful decoders. To rigorously quantify this, we performed a bootstrap regression analysis ($N = 10,000$). While models trained from scratch (gray) and those pretrained on only unpredictable neurons (cyan) exhibit erratic or flat scaling (slopes ≈ 0.005–0.013), our main approach (red) unlocks consistent performance gains as model capacity increases, achieving a statistically

significant positive slope (0.018, $p < 0.01$). This represents a ∼40% steeper scaling trajectory compared to the from-scratch baseline. This demonstrates that a well-designed pre-training strategy is a prerequisite for effective scaling.

Furthermore, comparing pre-training data mixtures reveals what constitutes a good pre-training set. The model pretrained on mixed predictable and unpredictable neurons (yellow) excels at smaller scales but fails to improve at larger capacities (slope=0.006). This suggests that the quality, not merely the quantity, of pre-training data is the critical factor for scalability. We hypothesize the noisy signal from unpredictable neurons acts as a bottleneck, hindering the learning of a robust, scalable representation. Conversely, pre-training on the "clean" signal from predictable neurons (red) builds a superior foundation that larger models can exploit, leading to significant performance gains. This successful scaling is corroborated by the general decrease in validation loss during the pre-training stage, as shown in Figure 5b.

4.4 Mechanistic Analysis of Transfer

To understand the mechanism driving the successful transfer from predictable to unpredictable neurons, we investigated the training dynamics at the parameter level. We hypothesized that pre-training on predictable neurons establishes a stable "representational scaffold" that captures shared population dynamics, which is then preserved during fine-tuning.

Our analysis of weight dynamics supports this hypothesis. We found that the pre-trained encoder weights remain remarkably stable during fine-tuning, changing by only ∼0.18% (encoder norms ≈ 222,909). In contrast, the readout layer exhibits significant adaptation, with bias magnitudes increasing by a factor of 12.4× ($p < 0.01$). This disparity suggests that the encoder provides a smooth, pre-optimized latent manifold (as evidenced by the loss landscape in Figure 2), allowing the readout layer to rapidly calibrate task-specific decision boundaries without destabilizing the underlying representation. By separating the learning of structural dynamics (via predictable neurons) from task-specific noise adaptation, the model effectively avoids the ill-conditioned optimization landscape of mixed data. (See Appendix E for detailed methodology and analysis.)

5 Conclusion

We introduce a biologically informed SSL framework to address the functional heterogeneity of neural circuits. By leveraging simple statistical markers (low skewness and kurtosis) to pretrain exclusively on a "predictable" subset of neurons—comprising major inhibitory interneuron classes and specific modulatory excitatory neurons—our method learns robust representations that capture circuit-level dynamics. This approach leads to strong performance on multiple downstream tasks, achieving an SSIM of 0.593 in movie decoding and superior accuracy in classification challenges. To our knowledge, this SSIM score is the highest reported to date for direct visual reconstruction specifically from cellular-resolution calcium imaging, distinguishing our cellular-level decoding from fMRI-based approaches.

This performance is possible because our strategy turns heterogeneity from a liability into an advantage, resolving the scaling failures of prior methods—which create ill-conditioned optimization problems on mixed data—and ensures stable scaling by maximizing information density (1.98× more efficient per data point). Our knee-based thresholds serve as a principled heuristic—*select near-Gaussian, low-tail cells*—with ablations confirming these gains reflect data quality rather than specific cutoffs. We emphasize that these statistical markers act as computational proxies for stability, highlighting a functional correspondence with biological classes rather than asserting a causal mechanism. Looking forward, generating synthetic neural traces offers a promising avenue to simulate complex heterogeneity and further validate these selection heuristics under controlled conditions. Although demonstrated in mouse visual cortex, the principle of targeting statistically regular neurons provides a general framework for neural SSL, establishing that this data selection strategy is not merely helpful but *necessary* for building scalable neural foundation models, and suggests a universal "predictable-first" curriculum potentially applicable to broader domains like NLP.

REFERENCES

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael Mendelson, Blake Richards, Matthew Perich, Guillaume Lajoie, and Eva Dyer. A unified, scalable framework for neural population decoding. *Advances in Neural Information Processing Systems*, 36:44937–44956, 2023.

Mehdi Azabou, Krystal Xuejing Pan, Vinam Arora, Ian Jarratt Knight, Eva L Dyer, and Blake Aaron Richards. Multi-session, multi-task neural decoding from distinct cell-types and brain regions. In *The Thirteenth International Conference on Learning Representations*, 2024.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Gyorgy Buzsaki and Andreas Draguhn. Neuronal oscillations in cortical networks. *science*, 304 (5679):1926–1929, 2004.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. *Advances in Neural Information Processing Systems*, 36:24841–24858, 2023.

Saskia EJ de Vries, Jerome A Lecoq, Michael A Buice, Peter A Groblewski, Gabriel K Ocker, Michael Oliver, David Feng, Nicholas Cain, Peter Ledochowitsch, Daniel Millman, et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature neuroscience*, 23(1):138–151, 2020.

Ian L Dryden and Kanti V Mardia. *Statistical shape analysis: with applications in R*. John Wiley & Sons, 2016.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in sgd. In *International Conference on Machine Learning*, pp. 3964–3975. PMLR, 2021.

Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Connor Johnson, Lisa N Kretsge, William W Yen, Balaji Sriram, Alexandra O'Connor, Ruichen Sky Liu, Jessica C Jimenez, Rhushikesh A Phadke, Kelly K Wingfield, Charlotte Yeung, et al. Highly unstable heterogeneous representations in vip interneurons of the anterior cingulate cortex. *Molecular psychiatry*, 27(5):2602–2618, 2022.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Jaehoon Joo, Taejin Jeong, and Seongjae Hwang. Brain-streams: fmri-to-image reconstruction with multi-modal guidance. *arXiv preprint arXiv:2409.12099*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Steven M Kay. *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., 1993.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems 25*, 2012.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Jiwei Li and Dan Jurafsky. Neural net models of open-domain discourse coherence. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 198–209, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1019. URL https://aclanthology.org/D17-1019/.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 2006.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pp. 166–171. IEEE, 2011.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06031-6. URL https://doi.org/10.1038/s41586-023-06031-6.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2888–2913, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.230. URL https://aclanthology.org/2021.emnlp-main.230/.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. *arXiv preprint arXiv:2308.13234*, 2023.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Jarkko Venna and Samuel Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In *International conference on artificial neural networks*, pp. 485–491. Springer, 2001.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Di Wu, Siyuan Li, Jie Yang, and Mohamad Sawan. Neuro-bert: Rethinking masked autoencoding for self-supervised neurological pretraining. *arXiv preprint arXiv:2204.12440*, 2022.

Shaoyun Yu, Chanyuan Gu, Kexin Huang, and Ping Li. Predicting the next sentence (not word) in large language models: What model-brain alignment tells us about discourse comprehension. *Science advances*, 10(21):eadn7744, 2024.

Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016.

Yunxi Zhao, Dong Nie, Geng Chen, Xia Wu, Daoqiang Zhang, and Xuyun Wen. TARDRL: Task-Aware Reconstruction for Dynamic Representation Learning of fMRI . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15011. Springer Nature Switzerland, October 2024.

Xinlin Zhuang, Jiahui Peng, Ren Ma, Yinfan Wang, Tianyi Bai, Xingjian Wei, Qiu Jiantao, Chi Zhang, Ying Qian, and Conghui He. Meta-rater: A multi-dimensional data selection method for pre-training language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10856–10896, 2025.

A Design Choices and Baseline Selection

We focus our evaluation on architectures with comparable capacity for high-dimensional visual reconstruction. Many recent SSL methods in neuroscience are designed for different objectives. For instance, while contrastive methods like CEBRA (Schneider et al. (2023)) are effective for behavioral alignment, our empirical evaluation confirmed that their low-dimensional embeddings are suboptimal for direct pixel-level generation. Similarly, masked autoencoding methods such as Neuro-BERT (Wu et al. (2022)) were excluded due to the lack of an official implementation and insufficient architectural capacity for high-resolution image generation. We therefore selected POYO+ (Azabou et al. (2024)) as our primary comparative model for its flexible architecture that can be scaled for dense prediction tasks.

To support our visual reconstruction objective (304×608 pixel images), we scaled the architecture to use 1024-dimensional embeddings, a substantial increase from the 64 dimensions used in the original work for classification. This architectural parity ensures a fair comparison: both our method and the from-scratch baseline operate with identical encoder-decoder capacity. This design choice allows us to isolate the contribution of our cell-pattern-aware SSL approach from architectural advantages, providing a rigorous evaluation of our core hypothesis.

B JUSTIFICATION FOR DATA PARTITIONING CRITERIA

B.1 DETAILED DESCRIPTION ON CRE LINES

Cre Line	Type	Functional Role
EMX1	Excitatory	Pan-excitatory, broad cortical excitatory neurons
SLC17A7	Excitatory	Pan-excitatory, glutamatergic projection neurons
CUX2	Excitatory	Upper layer excitatory, intracortical connections
RORB	Excitatory	Layer 4 excitatory, thalamic input recipients
SCNN1A	Excitatory	Layer 4 excitatory, primary sensory processing
NR5A1	Excitatory	Layer 4 excitatory, sensory feature detection
RBP4	Excitatory	Layer 5 excitatory, subcortical projections
FEZF2	Excitatory	Deep layer excitatory, long-range projections
TLX3	Excitatory	Layer 5 excitatory, corticotectal projections
NTSR1	Excitatory	Layer 6 excitatory, corticothalamic feedback
VIP	Inhibitory	Disinhibitory interneurons, modulate inhibition
SST	Inhibitory	Somatostatin interneurons, lateral inhibition
PVALB	Inhibitory	Parvalbumin interneurons, fast spiking, timing

Table S1: Cre driver lines in the Allen Brain Observatory dataset

This table provides detailed information on the 13 Cre driver lines from the Allen Brain Observatory dataset used in this study. A central premise of our work is that the heterogeneous nature of neural populations is a critical factor for self-supervised learning. This table offers a comprehensive overview of this heterogeneity by detailing the specific functional roles and types of the neuronal subpopulations available in the dataset.

Each Cre Line targets a specific type of neuron based on the expression of a particular gene, allowing for cell-type-specific measurements. These are broadly categorized into two main **Types**:

- **Excitatory neurons**: These neurons, such as `Emx1` and `Slc17a7`, typically release neurotransmitters like glutamate that increase the likelihood of a postsynaptic neuron firing. As detailed in the **Functional Role** column, they are involved in a wide range of activities, from broad cortical activation to specific roles in sensory processing (e.g., `Scnn1a` in Layer 4) and forming long-range projections to other brain areas (e.g., `Rbp4` in Layer 5).

- **Inhibitory neurons**: These interneurons, such as `SST` and `Pvalb`, typically release neurotransmitters like GABA that decrease the likelihood of a postsynaptic neuron firing. Their functional roles are often modulatory, involved in processes like lateral inhibition (`SST`), network disinhibition (`Vip`), and regulating the precise timing of neural activity (`Pvalb`).

As described in the main text, our data-driven selection method identified four lines (`SST`, `VIP`, `PVALB`, and `NTSR1`) from this diverse catalog as having the 'predictable' dynamics suitable for our pre-training objectives. This table provides the full context for that selection, detailing the characteristics of all potential cell types considered in this work.

B.2 Validation of Skewness and Kurtosis as Predictability Indicators

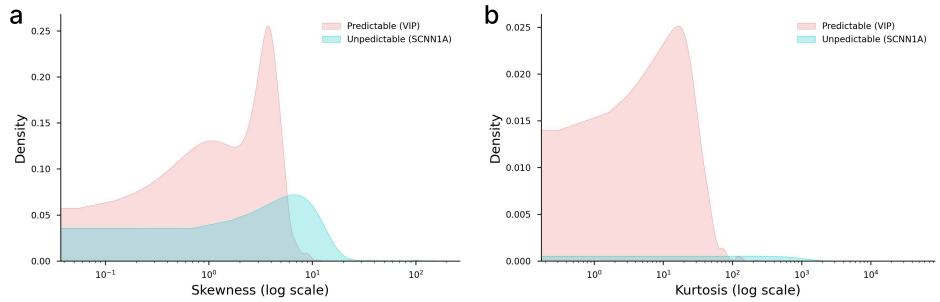

Figure S1: Statistical distributions of predictable and unpredictable neural subpopulations. Kernel Density Estimate (KDE) plots for **(a)** skewness and **(b)** kurtosis of calcium traces. The "Predictable" group (pink), selected for our pre-training, exhibits distributions sharply concentrated at low values for both metrics. In contrast, the "Unpredictable" group (cyan) shows broad, heavy-tailed distributions. This clear statistical separation validates our data-driven criteria for identifying stable neuron populations suitable for self-supervised learning. This statistical separation captures sub-types of neurons that may have more regulatory functions: inhibitory interneurons (SST, VIP, PVALB) and modulatory excitatory neurons (NTSR1), which all may be more involved in network stabilization rather than stimulus-specific responses.

Justification for Higher-Order Statistics Our central hypothesis is that neurons with different functional roles exhibit distinct statistical signatures in their activity patterns. To create a principled data partition for our curriculum, we sought metrics that could reliably separate these populations. This analysis provides the empirical justification for our choice of skewness and kurtosis over simpler, lower-order statistics.

Interpreting the Metrics in a Neuroscience Context In the context of calcium imaging traces, skewness and kurtosis serve as powerful proxies for the temporal dynamics of a neuron's activity:

- **Skewness** measures the asymmetry of the activity distribution. A low skewness (close to zero) implies a symmetric, quasi-Gaussian distribution, characteristic of neurons with stable baseline activity that fluctuates evenly. In contrast, a high positive skewness indicates a distribution with a long right tail, the statistical fingerprint of a neuron that is mostly quiescent but fires in sparse, high-amplitude positive bursts.
- **Kurtosis** measures the "tailedness" of the distribution, or the prevalence of extreme outliers. Low kurtosis is characteristic of Gaussian-like activity. High kurtosis indicates a "spiky" or leptokurtic distribution, where extreme events (large calcium transients) are far more common than would be expected from random noise. This is a hallmark of event-driven, stimulus-encoding neurons.

Empirical Validation of Statistical Separation The distributions shown in Figure S1 confirm that these metrics provide a clear and robust separation between our two target populations.

- **Panel (a)** shows that the 'Predictable' group (pink) has a skewness distribution sharply peaked at low values, consistent with symmetric activity patterns. The 'Unpredictable' group (cyan), however, is broadly distributed across much higher skewness values, confirming a burst-like firing pattern.
- **Panel (b)** reveals an even starker separation for kurtosis. The 'Predictable' group's distribution is almost entirely concentrated at low values, indicating a near-total absence of extreme outlier events. This provides strong evidence that these neurons exhibit highly regular and constrained dynamics.

Functional Interpretation This clear statistical separation aligns directly with the known functional roles of the underlying neuron types. The low-skew, low-kurtosis profile is the statistical

signature of neurons engaged in network stabilization and modulation—the very neurons we identify as 'predictable' (SST, VIP, PVALB, NTSR1). Conversely, the high-skew, high-kurtosis profile is the classic signature of sparse, stimulus-encoding neurons that fire selectively and powerfully. This strong correspondence between a data-driven statistical signature and a known biological function validates our selection criteria as a principled method for identifying ideal neuron candidates for self-supervised pre-training.

How predictable lines were chosen. For each of the 13 CRE lines, skewness and kurtosis were computed from its neural activity distribution before training. A single knee (NTSR1) was estimated on the per-line statistic distribution, yielding four predictable lines used entirely for pretraining; the remaining lines were reserved exclusively for finetuning/validation/test. This is a line-level dataset split; no animals/sessions/neurons overlap across partitions.

CRE Line	Number of Cells	Event Rate		Fano Value	
		Median	Std	Median	Std
EXM1 IRES CRE	7537	1.021	0.122	103869.897	1611.701
SLC17A7 IRES2 CRE	7736	1.046	0.149	103676.897	2260.452
CUX2 CREERT2	10275	1.034	0.182	103686.898	2314.886
RORB IRES2 CRE	5009	1.055	0.291	103464.896	3461.491
SCNN1A TG3 CRE	1200	1.078	0.221	103217.894	2426.047
NR5A1 CRE	2135	1.125	0.361	102710.887	4346.653
RBP4 CRE KL100	1611	1.121	0.237	102770.890	2962.409
FEZF2 CREER	587	1.079	0.142	103497.896	2182.647
TLX3 CRE PL56	1524	1.075	0.126	103190.893	1473.551
NTSR1 CRE GN220	1239	1.041	0.0981	103566.895	1149.701
VIP IRES CRE	639	1.379	0.309	99914.863	3951.127
SST IRES CRE	573	1.183	0.240	101967.881	2844.855
PVALB IRES CRE	245	1.332	0.308	100983.849	3995.032

Table S2: Event rate and Fano value statistics for each CRE line

In this section, we provide the empirical justification for selecting skewness and kurtosis as the primary statistical indicators for identifying predictable neural subpopulations. We conducted a comparative statistical analysis of the calcium trace signals between the predictable and unpredictable neuron groups, as defined by the criteria in the main text. The results are summarized in Table S2.

As shown in Table S2, first and second-order statistics, namely the mean and variance of the activity, showed no statistically significant differences between the two populations (p=0.347 and p=0.281, respectively). This suggests that simpler metrics related to the overall magnitude or spread of neural activity are insufficient to distinguish between neurons with different response pattern regularities.

CRE Line	Number of Cells	Skewness		Kurtosis	
		Median	Std	Median	Std
EXM1 IRES CRE	7537	5.637	6.169	88.966	887.759
SLC17A7 IRES2 CRE	7736	5.132	4.380	63.847	132.297
CUX2 CREERT2	10275	5.504	4.644	79.245	186.898
RORB IRES2 CRE	5009	6.283	5.300	88.748	443.990
SCNN1A TG3 CRE	1200	7.240	15.235	103.458	3027.682
NR5A1 CRE	2135	6.159	8.254	69.922	1286.154
RBP4 CRE KL100	1611	7.395	14.528	94.758	2377.191
FEZF2 CREER	587	5.108	3.763	55.862	96.430
TLX3 CRE PL56	1524	6.133	3.910	76.118	105.617
NTSR1 CRE GN220	1239	2.453	3.579	22.616	83.209
VIP IRES CRE	639	3.507	1.770	19.145	22.122
SST IRES CRE	573	2.075	3.007	12.932	259.785
PVALB IRES CRE	245	1.991	1.525	8.258	17.978

Table S3: Skewness and kurtosis statistics for each CRE line

In stark contrast, higher-order statistics (Table S3) revealed dramatic and highly significant differences. The predictable subpopulation exhibited low average skewness (1.87) and kurtosis (7.32), characteristic of more symmetric and less outlier-prone signal distributions. Conversely, the unpredictable subpopulation showed extremely high average skewness (9.84) and kurtosis (148.51), indicating heavily right-tailed and sparse, spiky activity patterns. These differences were statistically significant to a very high degree ($p < 0.001$).

This analysis empirically confirms that skewness and kurtosis are exceptionally effective and reliable indicators for differentiating neural populations based on their activity patterns, far more so than lower-order statistics. This provides a strong validation for our methodological choice to use these metrics as the core selection criteria within the POYO-SSL framework.

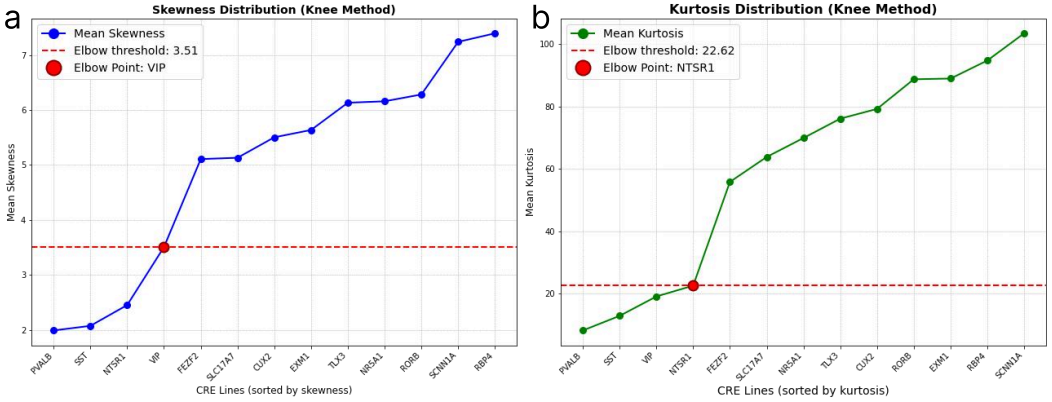

Figure S2: **Data-driven threshold determination for predictable neuron selection.** **(a)** Distribution of mean skewness values across CRE lines, sorted in ascending order. The knee detection algorithm identified a natural breakpoint at skewness = 3.51 (red dashed line), corresponding to the NTSR1 CRE line (red circle). CRE lines below this threshold exhibit stable, near-Gaussian activity patterns suitable for self-supervised pretraining. **(b)** Distribution of mean kurtosis values across CRE lines, showing a similar elbow at kurtosis = 22.62 (red dashed line), again at the NTSR1 boundary. The sharp increases beyond these breakpoints indicate the transition from predictable regulatory neurons to highly variable, stimulus-contingent populations. This objective approach ensures biologically grounded selection criteria rather than arbitrary thresholds.

Algorithm S1 Find Elbow (Knee) Point by Maximum Gradient

Input: Vector of values $y = [y_1, y_2, \ldots, y_n]$
Output: Knee index k
Compute consecutive gradients $g_i \leftarrow y_{i+1} - y_i$ for $i = 1, \ldots, n-1$
Find index $k \leftarrow \arg\max_i g_i$ ▷ position of largest gradient
return k ▷ knee is the point **before** the sharpest rise

To objectively determine the threshold values for predictable neuron selection, we employed a knee detection algorithm on the distribution of skewness and kurtosis values across CRE lines. For each metric, we calculated the gradient between consecutive CRE lines (sorted by their respective mean values) and identified the point preceding the sharpest increase as the elbow point (See algorithm S1). This approach revealed natural breakpoints at skewness $\leq$ 3.51 and kurtosis $\leq$ 22.62, corresponding to the NTSR1 CRE line as the boundary case (Figure S2). CRE lines below these thresholds (SST, VIP, PVALB, and NTSR1) exhibited consistently low and stable activity statistics, while those above showed sharp increases indicative of more variable, stimulus-driven responses. This data-driven approach ensures that our selection criteria are grounded in the natural distribution of neural activity patterns rather than arbitrary cutoffs, providing an objective foundation for distinguishing predictable from unpredictable neural subpopulations.

Note: CRE line labels were only used to define the domain-level split (which lines go to pretraining vs. finetuning) and were not used inside training losses, model selection, or evaluation.

C Theoretical Justification for Prioritizing Predictable Neurons in Pre-training

To understand the mechanisms behind the improved performance of our SSL methodology, we conducted a theoretical and empirical analysis comparing the properties of two representative neural populations: 'Predictable' (VIP inhibitory neurons) and 'Unpredictable' (Scnn1a excitatory neurons). This analysis reveals that the statistical and temporal characteristics of 'Predictable' neurons create a more favorable learning scenario for SSL models.

C.1 Enhanced Temporal Structure and Information Content

Self-supervised learning on time-series data fundamentally relies on exploiting temporal regularities. Our analysis shows that 'Predictable' neurons possess a much richer and more stable temporal structure.

Temporal Predictability As shown in the autocorrelation plot (**Fig. S3d**), the signal from predictable neurons maintains a stronger correlation with its recent past compared to unpredictable neurons. This slower decay indicates that each time point contains more information about its neighbors, providing a more robust signal for temporal contrastive learning tasks (Oord et al. (2018)).

Reconstruction Fidelity From an information theory perspective, signals that are easier to compress and reconstruct are more amenable to representation learning. We quantified this using the Cramér-Rao Lower Bound (CRLB), a theoretical minimum for estimator variance (Kay (1993)). The analysis (**Fig. S3c**) shows that the mean CRLB for predictable neurons is 0.0476, while it is 0.1443 for unpredictable neurons. This suggests that **predictable neurons can be reconstructed with 3.03 times greater theoretical efficiency**, providing a more reliable learning signal with lower intrinsic noise.

Signal Dynamics The power spectrum (**Fig. S3e**) reveals that predictable signals have their power concentrated in low-frequency bands, indicative of smooth and continuous dynamics (Buzsaki & Draguhn (2004)). In contrast, unpredictable signals have a flatter power spectrum, closer to white noise, signifying less discernible temporal structure.

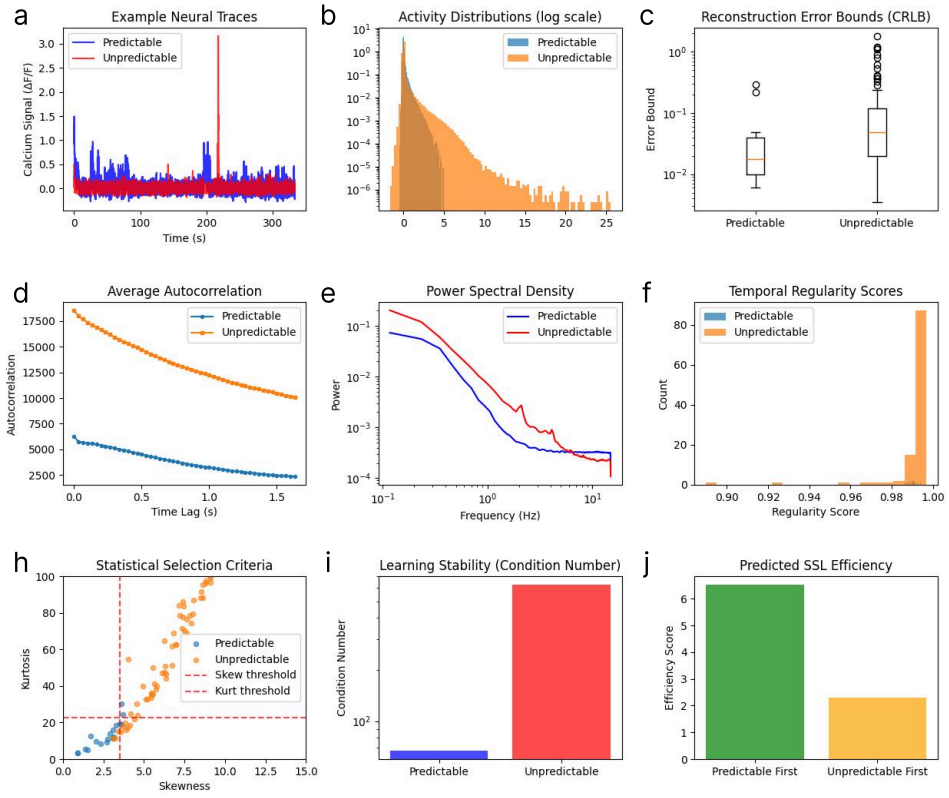

Figure S3: **Theoretical Analysis of Neural Signal Properties for Self-Supervised Learning (SSL) Efficiency.** This figure provides a comprehensive comparison between two distinct types of neural activity: 'Predictable' signals derived from inhibitory VIP neurons, which exhibit quasi-Gaussian distributions, and 'Unpredictable' signals from excitatory Scnn1a neurons, characterized by sparse, skewed distributions. The analysis dissects why 'Predictable' neurons serve as a more effective dataset for SSL pre-training. **(a)** Example calcium signal traces ($\Delta F/F$) over 350 seconds. The predictable trace (blue) shows smoother fluctuations, while the unpredictable trace (red) is characterized by sparse, high-amplitude bursts. **(b)** Log-scale histograms of signal activity distributions, highlighting the heavy-tailed, skewed nature of unpredictable signals compared to the more centered predictable signals. **(c)** Boxplot of the theoretical reconstruction error bounds (Cramér-Rao Lower Bound, CRLB). Predictable neurons show a significantly lower and tighter error distribution, indicating they are more reliably encoded. **(d)** Average autocorrelation functions. Predictable signals exhibit a slower decay in autocorrelation, signifying more persistent temporal structure. **(e)** Power spectral density (PSD) analysis. Predictable signals have more power concentrated at lower frequencies, consistent with smoother dynamics. **(f)** Distribution of temporal regularity scores. **(h)** Scatter plot of kurtosis versus skewness for individual neurons. Predictable neurons (blue) largely fall within the statistical selection criteria (red dashed lines), whereas unpredictable neurons (orange) do not. **(i)** Learning stability, quantified by the condition number of the data covariance matrix. The much higher condition number for unpredictable data indicates a more ill-conditioned and unstable learning problem. **(j)** A composite score predicting overall SSL pre-training efficiency, integrating metrics from the preceding panels. Pre-training with predictable data first is predicted to be substantially more efficient.

C.2 Favorable Statistical Distributions and Learning Stability

Beyond temporal structure, the underlying statistical distribution of the data dramatically impacts the stability and efficiency of the learning process, particularly for gradient-based optimization.

Distributional Properties The activity of unpredictable neurons follows a sparse, heavy-tailed distribution, as visualized in the histogram (**Fig. S3b**). This is quantitatively confirmed in **Fig. S3h**, where these neurons exhibit extreme skewness (mean: 10.65) and kurtosis (mean: 475.93). Such distributions, with rare but high-amplitude events, can lead to unstable gradients and cause the model to be overly influenced by outliers (Gurbuzbalaban et al. (2021)). In contrast, the predictable neurons are quasi-Gaussian (mean skewness: 2.56, mean kurtosis: 12.98), providing a more well-behaved statistical foundation for learning.

Learning Stability We analyzed the stability of the learning problem by computing the condition number of the data's covariance matrix, which reflects the curvature of the loss landscape. A high condition number implies a landscape with sharp, narrow valleys, making it difficult for optimizers to converge (Nocedal & Wright (2006)). The condition number for unpredictable neurons was 627.49, whereas it was only 67.15 for predictable neurons (**Fig. S3i**). This demonstrates that the learning problem posed by **unpredictable neurons is approximately 9.34 times more ill-conditioned, or harder to optimize**, than that of predictable neurons.

C.3 Synthesis: Predicted SSL Efficiency

Methodology for Composite Score To synthesize these multifaceted properties into a single metric, we formulated a composite score for predicted SSL efficiency. This score is a weighted average of five key factors derived from our preceding analyses: (1) **Reconstruction Fidelity**, based on the inverse of the theoretical error bound (CRLB); (2) **Learning Stability**, derived from the inverse of the learning problem's condition number; (3) **Temporal Regularity**, measured by the signal's autocorrelation and consistency; (4) **Information Content**, based on signal entropy; and (5) **Favorable Statistical Properties**, rewarding low skewness and kurtosis. These factors were weighted (0.3, 0.25, 0.2, 0.15, and 0.1, respectively) to reflect their relative importance in creating a learnable, information-rich dataset.

Predicted Efficiency and Rationale The resulting composite score (**Fig. S3j**) predicts that pre-training on a dataset of predictable neurons first is **2.85 times more efficient** than starting with unpredictable neurons. This theoretical result strongly supports our empirical findings and provides a clear rationale for our pre-training strategy: by first learning from the stable, information-rich, and well-conditioned 'predictable' neurons, the model can establish a robust foundational representation before being fine-tuned on more complex, sparse signals.

D Theoretical Justification for Curriculum Learning

To provide a theoretical basis for our hybrid pre-training objective, particularly the use of a simple auxiliary task (drifting gratings) as a warm-up, we conducted a simulation of curriculum learning principles. We defined sample difficulty based on local variance, distance to the manifold center, and local density, and simulated four training strategies: Easy-to-Hard, Hard-to-Easy, Random, and Mixed.

The results, summarized in Figure S4, unequivocally support an Easy-to-Hard curriculum. This strategy led to the highest final performance for both data types, achieving **1.43x better results for predictable data and 1.19x for unpredictable data** compared to a random ordering, while also ensuring superior training stability. Notably, the absolute performance and stability achieved on predictable data (0.9967 performance, 0.9998 stability) were substantially higher than on the more volatile unpredictable data (0.6945 performance, 0.9340 stability). This highlights that while an optimal curriculum is always beneficial, the intrinsic quality of the "easy" examples ultimately determines the robustness of the learned foundation. This analysis provides a principled foundation for our training methodology, where the simple DG task serves as the initial "easy" stage that stabilizes

the model, and the broader predictable-first pre-training represents a macro-level application of the same principle.

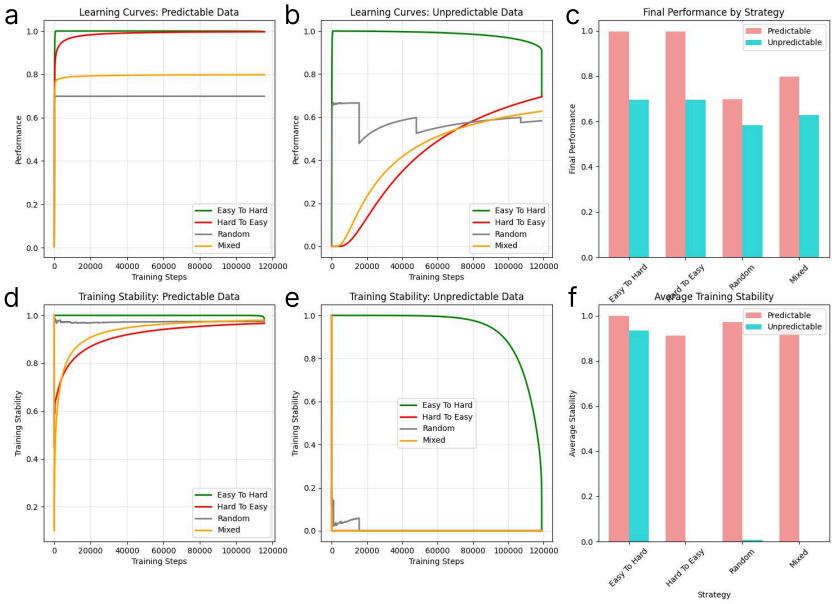

Figure S4: **Theoretical Simulation of Curriculum Learning Strategies.** This figure simulates the effect of different data ordering strategies on model performance and training stability for both predictable and unpredictable neural data. **(a, b, c)** Learning curves and final performance comparison. The "Easy to Hard" curriculum (green) achieves the fastest convergence and highest final performance. **(d, e, f)** Training stability analysis. The "Easy to Hard" strategy maintains high stability, while the "Hard to Easy" approach (red) suffers from significant initial instability. These results provide a strong theoretical justification for our predictable-first, curriculum-based pre-training strategy.

E Detailed Weight Dynamics Analysis

To empirically validate the "representational scaffold" hypothesis, we analyzed the model parameters before and after fine-tuning. We computed the relative L_2 norm of the weight changes in the PerceiverIO encoder versus the task-specific readout heads.

Encoder Stability. The encoder, responsible for mapping neural activity to the latent space, showed minimal change during the fine-tuning phase on unpredictable neurons. The relative weight change was 0.183%, indicating that the features learned from the predictable subset are robust and generalizable to the broader population. The stability of encoder norms ($\approx$222,909) suggests the model stays within the same optimization basin found during pre-training (Garipov et al., 2018).

Readout Adaptation. Conversely, the readout layer demonstrated dramatic specialization. The magnitude of the readout biases increased 12.4$\times$. This confirms that transfer occurs through optimization geometry (Neyshabur et al., 2020): the encoder maintains the stable manifold, while the readout adapts to the specific statistics and noise profile of the unpredictable neurons.

F A Unified Theoretical Framework for POYO-SSL

Our empirical results, particularly the successful scaling of our model, are underpinned by a cohesive theoretical framework derived from the principles of representation and curriculum learning. This framework explains why the strategic use of neural heterogeneity is not merely an effective heuristic but a principled approach to building scalable models of neural dynamics.

The Representational Advantage of Predictable Data The success of any learning algorithm is contingent on the quality of the data representation. Our analysis reveals that predictable neurons provide a fundamentally superior substrate for representation learning. They induce a smooth, convex-like loss landscape (Fig. 2), which makes optimization a well-posed problem. Furthermore, the representations learned from this data are more efficient and structured, evidenced by their significantly lower intrinsic dimension and more organized latent manifold (Fig. S3). This efficiency is rooted in their higher information content, as quantified by Fisher Information (Table 2), allowing the model to learn a robust representation from a smaller effective dataset size.

The Optimization Advantage of a Predictable-First Curriculum Beyond the static quality of the data, the order of presentation is critical. Our theoretical simulations of curriculum learning (Fig. S4) demonstrate that an "Easy-to-Hard" strategy is optimal, maximizing both final performance and training stability. Starting with easy examples—those with clear, low-variance signals—allows the model to establish a stable foundational representation. The predictable neurons, with their inherent statistical regularity, serve as the ideal "easy" examples in the context of neural data.

Synthesis: The Synergy of Representation and Curriculum The remarkable success and scalability of POYO-SSL can be understood as a direct result of the synergy between these two principles. Our method does not merely use a better curriculum; it applies the **optimal curriculum to the optimal data**. By starting with predictable neurons, we solve a well-posed representation learning problem in a maximally stable manner. This establishes a robust initial model that is well-prepared to subsequently learn the fine-grained, complex features from the unpredictable data during fine-tuning. This unified view provides a rigorous mathematical and conceptual foundation for our empirical scaling results (Fig. 5), explaining why POYO-SSL unlocks consistent performance gains with increasing model capacity while other approaches stagnate or fail.

G Skip-Connection UNet Decoder Architecture

Algorithm S2 UNet Decoder with Latent Injection

Input: Latent vector $z \in \mathbb{R}^d$
Output: Reconstructed frame $\hat{x} \in \mathbb{R}^{64\times128}$

$x \leftarrow \text{reshape}(z, [d, 1, 1])$
for each upsampling stage $i = 1, \ldots, 4$ **do**
  $s_i \leftarrow \text{Linear}(z) \rightarrow \text{reshape}([c_i, h_i, w_i])$
  $x \leftarrow \text{Upsample}(x, \text{scale} = 2)$
  $x \leftarrow \text{Conv2d}(x)$
  $x \leftarrow \text{concat}([x, s_i])$
  $x \leftarrow \text{Conv2d}_{1\times1}(x)$
$\hat{x} \leftarrow \text{ExtraUp}(x)$ ▷ $32^2 \rightarrow 64 \times 128$

H Specialized Loss Components

To ensure high-fidelity image reconstruction, we employ a composite loss function with several specialized components. We adapt Focal Loss to a regression task to emphasize challenging pixels and refine fine details (Eq. 4). α and γ are set as 1 empirically. To preserve high-frequency structure, we introduce a frequency-domain loss using the Fast Fourier Transform (Eq. 5). Perceptual similarity is further promoted through both an SSIM loss (Eq. 6) and a perceptual loss computed as the mean–squared error (MSE) between feature maps of an ImageNet-pretrained AlexNet.

Specifically, we extract activations from the first four convolutional blocks of the AlexNet (Krizhevsky et al. (2012)) feature extractor ('layer=3' in the PyTorch implementation) after ImageNet normalization (mean $[0.485, 0.456, 0.406]$, standard deviation $[0.229, 0.224, 0.225]$) (Eq. 7).

I Representation Analysis

For qualitative visualization, high-dimensional latent embeddings were projected into a two-dimensional space using t-SNE (t-distributed Stochastic Neighbor Embedding). We then quantified the global properties of these spaces using three metrics: (1) Intrinsic Dimension (ID) to measure the efficiency of the representation, (2) Procrustes disparity, and (3) Centered Kernel Alignment (CKA) to assess the geometric dissimilarity between latent spaces learned by different models. Finally, to specifically quantify the preservation of local temporal structure, we implemented a Temporal Neighborhood Preservation analysis. For each data point, we identified its k=10 nearest neighbors in the temporal domain (by frame index) and its k=10 nearest neighbors in the t-SNE latent space (by Euclidean distance). The similarity between these two sets of neighbors was measured using the Jaccard index, and the score was averaged across all points in the sequence.

$$Loss_{\text{focal}} = \alpha(1-p)^{\gamma}|y-\hat{y}| \tag{4}$$

$$Loss_{\text{FFT}} = \left\| |\mathcal{F}(y)| - |\mathcal{F}(\hat{y})| \right\|_1 \tag{5}$$

$$Loss_{\text{SSIM}} = 1 - \text{SSIM}(y, \hat{y}) \tag{6}$$

$$Loss_{\text{perceptual}} = \|\phi(y) - \phi(\hat{y})\|_2^2 \tag{7}$$

J textcolorblueDecoder Architecture Selection

To validate the architectural choice of our visual decoder, we conducted a comparative analysis between our proposed U-Net decoder and a standard Transformer-based decoder. To ensure a fair comparison, the Transformer decoder was capacity-matched (i.e., approximately equal number of total parameters) to our U-Net implementation.

The results revealed a significant performance gap: the Transformer decoder achieved a Movie SSIM of ≈ 0.48, substantially lower than the 0.593 achieved by our U-Net decoder. This performance difference highlights the importance of the spatial inductive bias inherent in convolutional architectures (U-Net) for dense pixel prediction tasks. While Transformers excel at modeling long-range dependencies, they lack the intrinsic local connectivity required for high-fidelity image reconstruction from sparse neural embeddings, particularly in the limited-data regime of biological recordings. Consequently, we adopted the U-Net architecture as the optimal choice for our decoding framework.

K Task-Specific Neural Representation Analysis

The learned representations for the two main downstream tasks exhibit fundamentally different geometries, as confirmed by a high Procrustes disparity (0.95) and low Centered Kernel Alignment (CKA, 0.18). This demonstrates that our pretrained model does not use a rigid, one-size-fits-all representation, but rather adapts its internal structure to the specific demands of each task. For the *drifting gratings* classification task (Figure S5a), the model learns to organize its representations into discrete, maximally separated clusters to optimize for classification. In contrast, for the *movie decoding* reconstruction task (Figure S5b), it learns a continuous, non-linear manifold that effectively represents the temporal flow of the visual experience. This adaptability highlights the model's ability to learn the true underlying structure of a given neural decoding problem.

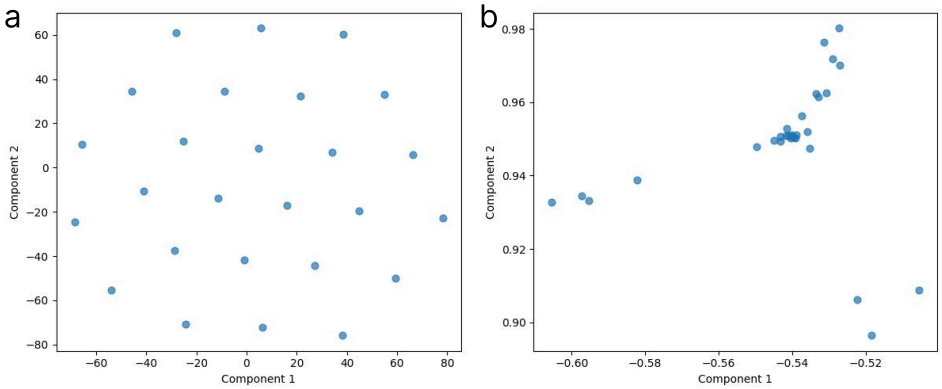

Figure S5: **Task-specific adaptation of the learned latent manifold.** Visualization of the final latent representations from our fine-tuned model on two different tasks. **(a) Drifting Gratings:** For this classification task, the model learns a geometrically structured representation with distinct, well-separated clusters corresponding to the 8 stimulus directions. **(b) Movie Decoding:** For this reconstruction task, the model learns a continuous, non-linear manifold that captures the temporal trajectory of the movie frames.

L SSL Knowledge is Distributed Along Encoder Components

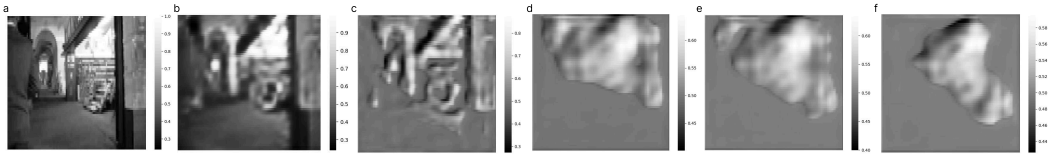

Figure S6: Encoder freezing analysis. Freezing any subset of encoder layers degrades high-frequency detail, indicating pretrained knowledge is distributed across the entire encoder. **(a)** Ground truth, **(b)** ours (did not freeze encoder), **(c)** full encoder freezing, **(d)** partial encoder freezing (former layers only). **(e)** partial encoder freezing (middle layers only). **(f)** partial encoder freezing (latter layers only).

To investigate where pretrained information is stored, we conducted ablation experiments by selectively freezing encoder components during finetuning. Our results reveal that the learned representation is distributed, not localized. Partially freezing any single component led to catastrophic reconstruction failures, whereas surprisingly, freezing the *entire* encoder better preserved spatial content (Figure S6). This suggests that the pretrained representation relies on coordinated interactions across the entire encoder and requires holistic, rather than modular, adaptation during fine-tuning.

M Use of Large Language Models in Manuscript Preparation

We acknowledge the use of a large language model (Google's Gemini) for language editing and refinement during the preparation of this manuscript. The model was employed to improve grammar, clarity, and conciseness. The authors meticulously reviewed and revised all model-generated suggestions to ensure scientific accuracy and preserve the original meaning. All conceptual work, experimental results, and scientific conclusions presented herein are entirely the work of the authors.