Towards zero-shot adaptation of predictive models of neurons encoding posterior probability

Suhas Shrinivasan[1] **Ralf M. Haefner**[2] **Fabian H. Sinz**[1] **Edgar Y. Walker**[4],
[1] Institute for Computer Science and Campus Institute for Data Science,
University of Göttingen, Göttingen, Germany
[2] Brain and Cognitive Sciences, University of Rochester, Rochester, New York, USA
[4] Department of Physiology and Biophysics, and Computational Neuroscience Center,
University of Washington, Seattle, USA

Abstract

Understanding how the brain adapts to changing sensory environments is a key challenge in neuroscience, with implications for AI. Neural predictive models are trained to predict neuronal responses to stimuli from a given stimulus distribution. Therefore, they cannot account for possible neural adaptations to new sensory contexts with shifts in the stimulus distribution, thus requiring the models to be retrained on newly recorded datasets in order to adapt them. In this work, we propose a zero-shot adaptation approach by leveraging Bayesian theories of perception and neural representation that suggest that (1) sensory neurons encode posterior distributions over latent variables in an internal generative model of the world and (2) that the brain preserves the mapping from latent causes to observations in its generative model, while adapting the prior distribution to new contexts. By employing advances in machine learning and generative models, we validate our approach on synthetic data, demonstrating the performance of our zero-shot adapted models to models retrained with new neural data. Our work not only lays the foundation for a normative approach to adapting neural predictive models to domain shifts, but also paves the way for an empirical method for testing Bayesian theories of neural representations.

1 Introduction

Efficient adaptability to changing environments is a hallmark of intelligence, and understanding how neural systems adjust to shifts in sensory context remains a challenge. Previous work has shown that neuronal responses are influenced not only by the properties of the incident stimuli but also by the statistics of the broader stimulus distribution [1–4]. Typical machine learning-based neural predictive models, known as system identification (SI) models, are trained to predict neuronal responses to stimuli from a given stimulus distribution ("context") [5–13]. While highly effective, SI models are not typically capable of accounting for neural adaptations due to shifts in the context, requiring them to be retrained on neural data recorded under the new contexts. Specifically, once trained on a given context T_A that entails a distribution $p(\mathbf{r} \mid \mathbf{x}, T_A)$ of neuronal responses $\mathbf{r}$ conditioned on a given stimulus $\mathbf{x}$, their performance would drop when tested on a new context T_B that entails a new response distribution $p(\mathbf{r} \mid \mathbf{x}, T_B)$ adapted to T_B, requiring the SI model to be retrained on newly recorded neural data from T_B.

Here we address this challenge using normative, Bayesian theories on neural coding and advances in probabilistic machine learning that allows for a *zero-shot generalization* of neural predictive models to novel contexts. Our approach assumes that sensory neurons encode posterior distributions over latent variables in an internal generative model of the world [14–16]. Specifically we assume that

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

sensory neuronal responses represent samples from the posterior (a neural sampling code, NSC) [15, 17–24]. Accordingly, in the brain's generative model, sensory observations $\mathbf{x}$ are caused by latent variables $\mathbf{r}$, and sensory neurons compute the posterior $p(\mathbf{r} \mid \mathbf{x})$, where under NSC, the latent variable $\mathbf{r}$ corresponds to neuronal responses. Under these assumptions, recent work [24] demonstrated that the brain's generative model could be learned using maximum likelihood estimation (MLE) from recorded stimulus-response data from a given context, and the posterior of the learned model acts as a neural predictive model. Here, we extend the learning of the brain's generative model to account for neuronal adaptations to a new context that requires *no neural data from the new context*. Our approach rests on a normative hypothesis that the brain preserves the mapping from latent causes to the observed stimuli (likelihood, "physical mechanism") across contexts, and adapts the latent variable distribution (prior expectation) to fit new contexts (Fig 1**A**) [24–26], consequently requiring us to only learn the prior in the new context thereby also adapting the posterior.

Here, we demonstrate the feasibility of our zero-shot adaptation method on synthetic data generated from an existing NSC model of the brain [19] simulating two different stimulus distributions (contexts). Our results show that the adapted model achieves performance close to a fully retrained system identification model, without requiring any new neural data from the new context.

2 Theory

Background Recent work [24] showed that, given a dataset of recorded stimulus-response pairs $\mathcal{D} := \{\mathbf{x}^{(i)}, \mathbf{r}^{(i)}\}_{i=1}^{N}$—assuming that the neurons follow NSC—one can learn the brain's generative model, parameterized as $p(\mathbf{x} \mid \mathbf{r}; \theta_L)\, p(\mathbf{r}; \theta_P)$ via maximum likelihood estimation (MLE): $(\theta_L^*, \theta_P^*) = \arg\max_{\theta_L, \theta_P} \left[\sum_{i=1}^{N} \log p\left(\mathbf{x}^{(i)} \mid \mathbf{r}^{(i)}; \theta_L\right) + \log p\left(\mathbf{r}^{(i)}; \theta_P\right)\right]$, where θ_L and θ_P are the parameters for the likelihood and the prior respectively. From this model, one can obtain an approximate posterior $q(\mathbf{r} \mid \mathbf{x}; \phi_{\text{NSC}})$ with parameters ϕ_{NSC}, which serves as a predictive model of neuronal responses (NSC predictive model). In contrast, SI models—the de-facto neural predictive models—learn the mapping $q(\mathbf{r} \mid \mathbf{x}; \phi_{\text{SI}})$, with parameters ϕ_{SI}, directly on $\mathcal{D}$ [5–13]. Both the NSC predictive model and the SI model attempt to capture the underlying true distribution $p(\mathbf{r} \mid \mathbf{x})$. Given that $\mathcal{D}$ consists of a fixed stimulus distribution $p(\mathbf{x})$ ("context"), an NSC predictive model offers no advantage in predictive power over an SI model [24] [1].

Both SI and NSC predictive models face performance degradation when there are shifts in the underlying $p(\mathbf{r} \mid \mathbf{x})$. In this work, we focus on adaptations in the neuronal response distribution to stimuli driven by shifts in the stimulus distribution (new contexts) [1–4]. In order to capture these neuronal adaptations, SI models would have to be retrained on a newly recorded dataset of adapted neuronal responses and stimuli under the new context. However, as we show next, under normative assumptions, the brain's adaptation to context changes can be attributed to adapting the prior distribution $p(\mathbf{r})$ *only* (Fig 1**A**), yielding a corresponding shift in the posterior $p(\mathbf{r} \mid \mathbf{x})$. Importantly, this implies that the NSC predictive model—which is the posterior of the learned generative model of the brain—can be adapted by only relearning the shifted prior distribution while leaving the likelihood $p(\mathbf{x} \mid \mathbf{r})$ unchanged (Fig 1**B**).

The brain's model adaptation and zero-shot generalization of NSC models If the brain maintains the generative model of the world $\mathbf{r} \longrightarrow \mathbf{x}$, changes in the stimulus distribution $p(\mathbf{x})$ can be accounted for either from shifts in the likelihood $p(\mathbf{x} \mid \mathbf{r})$ or the prior $p(\mathbf{r})$. Under the assumption that the latents $\mathbf{r}$ are the result of causal representation learning [25, 26], the likelihood $p(\mathbf{x} \mid \mathbf{r})$ represents invariant physical mechanisms (i.e., how latent variables like animal identity give rise to observations), while $p(\mathbf{r})$ reflects how the latents are distributed in a given context (e.g., relative frequency of different animals). Contextual changes will therefore change the brain's $p(\mathbf{r})$, while keeping $p(\mathbf{x} \mid \mathbf{r})$ constant (Fig 1**A**).[24]). Indeed, this normative assumption is implicitly present in the NSC literature [19]. For example, Haefner, Berkes, and Fiser [19] model tasks with varying contexts to affect the prior over sensory neural responses in the brain's generative model $p(\mathbf{r})$ rather than the stimulus likelihood $p(\mathbf{x} \mid \mathbf{r})$. Consequently neuronal responses under NSC, considering two contexts T_A and T_B can be expressed as:

$$\mathbf{r}^A \sim p(\mathbf{r} \mid \mathbf{x}, T_A) \propto p(\mathbf{x} \mid \mathbf{r})\, p(\mathbf{r} \mid T_A)$$
$$\mathbf{r}^B \sim p(\mathbf{r} \mid \mathbf{x}, T_B) \propto p(\mathbf{x} \mid \mathbf{r})\, p(\mathbf{r} \mid T_B)\,.$$

[1] Apart from possible advantages in terms of inductive biases.

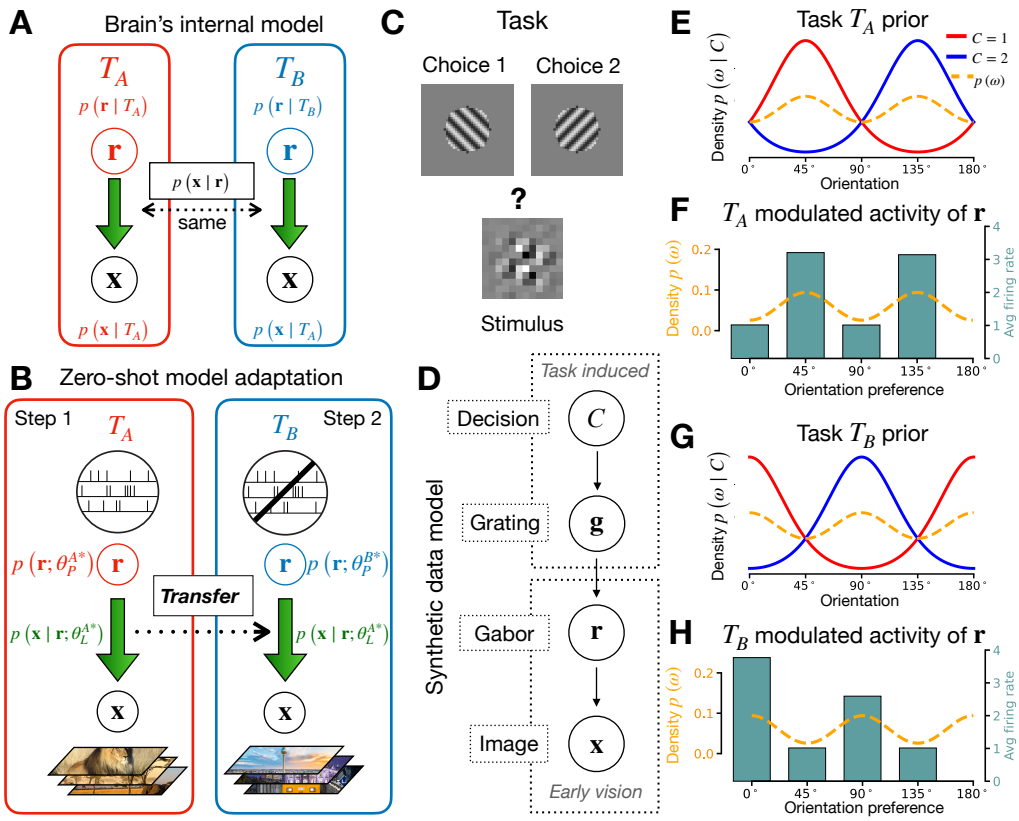

Figure 1: **Context dependence of the brain**. **A** Brain's internal generative model under contexts T_A and T_B. The hypothesis is that the brain retains the likelihood, but adapts prior to fit different contexts. **B** Our zero-shot model adaptation procedure. First learn the generative model using neural responses and images from T_A. Second transfer the learned likelihood to T_B and learn prior under T_B using *only images and no neural data* in T_B. **C** Tasks as contexts: "Categorize a noisy image stimulus into one of two orientation categories." Each task defines a stimulus distribution (here: via orientation distribution) (A.4) **D** synthetic data model, used to simulate neural responses and images (A.2). The model simulates neural activity $\mathbf{r}$ modulated by context (tasks). **E** Probability distributions over orientation in T_A: $45°$ (red) vs $135°$ (blue) task with uncertainty; overall task prior $p(\omega)$ that modulates neural activity. **E** Probability distributions over orientation ω for classes 1 and 2. The distributions vary in mean for a given class. **F** Activity of the four-$\mathbf{r}$ neurons from synthetic data model, modulated by T_A. Each neuron has an orientation preference (x-axis). **G & H** Same as E & F but for T_B.

What this implies from a model learning perspective is that once the generative model $p\left(\mathbf{x} \mid \mathbf{r}; \theta_L^{A*}\right) p\left(\mathbf{r}; \theta_P^{A*}\right)$ has been learned on a given task T_A, the generative model for task B only requires learning a new prior, while the likelihood can be *transferred*: $p\left(\mathbf{x} \mid \mathbf{r}; \theta_L^{A*}\right) p\left(\mathbf{r}; \theta_P^{B*}\right)$.

Crucially, the new prior distribution $p\left(\mathbf{r}; \theta_P^B\right)$ can be learned solely from stimuli sampled from the stimulus distribution under the new context $p\left(\mathbf{x} \mid T_B\right)$ (Fig 1**B**), by maximizing the log-likelihood of the sampled stimuli under the generative model:

$$\underset{\theta_P^B}{\operatorname{argmax}}\, p(\{\mathbf{x}\}_{i=1}^M; \theta_P^B) = \underset{\theta_P^B}{\operatorname{argmax}} \sum_{i=1}^{M} \log \int_{\mathbf{r}} p\left(\mathbf{x}^{(i)} \mid \mathbf{r}; \theta_L^{A*}\right) p\left(\mathbf{r}; \theta_P^B\right) \mathrm{d}\mathbf{r}$$

In order to evaluate the integral above in practice, we employ Monte-Carlo sampling (find method summary in Algorithm 1, and full objective derivation in Appendix A.1). Note that the likelihood function retains learned context T_A parameters θ_L^{A*}. Learning θ_P^B completes the adaptation of the generative model $p\left(\mathbf{x}, \mathbf{r}; \theta_P^{B*}, \theta_L^{A*}\right)$ under T_B. The posterior under T_B can then be approximated via

variational inference [24]. Because only samples from the stimulus distribution, but not neuronal responses, from context T_B are necessary, this constitutes a zero-shot adaptation of the model to T_B.

3 Experiments

The goal of this work is to show that this form of zero-shot adaption is feasible. We therefore focus on synthetic data from classic NSC models.

Synthetic data We generated two datasets (10k image-neuronal response pairs each: $\{\mathbf{x}^{(i)}, \mathbf{r}^{(i)}\}_{i=1}^{10k}$) using an existing model of the brain (synthetic data model) [19] (Fig 1**D**), implementing NSC in a classic sparse-coding model [27, 28] in two task contexts, T_A and T_B (Fig 1**C**). Both (orientation discrimination) tasks differ in the to-be-discriminated orientations, with each task T defining a different stimulus distribution $p(\mathbf{x} \mid T)$. The different stimulus distributions imply different priors which influence the firing rate of sensory neurons (features) through feedback signals. For example, neurons preferring $45°$ and $135°$ orientations exhibit higher *a priori* activity under T_A where stimuli with oblique orientations are overrepresented, while neurons preferring $0°$ and $90°$ orientations show increased activity under T_B where stimuli with cardinal orienations are overrepresented. The synthetic data model specifies a true generative model $p(\mathbf{x}, \mathbf{r} \mid T)$ per context T with likelihood $p(\mathbf{x} \mid \mathbf{r})$ and prior $p(\mathbf{x} \mid T)$ (more details under A.2 and A.4).

Test of our method The key models that we train in order to test our approach are: an NSC predictive model for T_A (NSC T_A), the zero-shot adapted NSC predictive model for T_B (NSC Zero-Shot), and a SI model for T_B (SI T_B). The NSC predictive models necessitate the training of generative models, one for T_A and one for T_B. The T_A generative model—consisting of a likelihood and a prior—is learned using image-response pairs from T_A, and the generative model is adapted for T_B by (1) transferring the learned T_A likelihood and learning only a new prior using *only images and no responses* from T_B (following Algorithm 1). We model the likelihood as an isotropic Gaussian with mean and variance defined by nonlinear functions of the response. The prior is a normalizing flow with a multivariate normal base distribution and a sequence of invertible transformations. The approximate posterior distribution is modeled by a factorized Gamma distribution, with parameters modeled as MLPs. For the system identification models, we utilize the same density function and model architecture as described for the posterior, differing only in the training objective. Refer to the appendix (A.3) for detailed descriptions.

Algorithm 1 Zero-Shot Adaptation of NSC Predictive Model to Novel Context T_B

Require: T_B-stimuli $\{\mathbf{x}\}_{i=1}^{M} \sim p(\mathbf{x} \mid T_B)$; learned T_A-likelihood $p\left(\mathbf{x} \mid \mathbf{r}; \theta_L^{A*}\right)$

Learn T_B **Prior**: $p\left(\mathbf{r}; \theta_P^B\right)$

1: Initialize prior parameters θ_P^B
2: Repeat until convergence:
$\theta_P^B \leftarrow \arg\max_{\theta_P^B} \sum_{i=1}^{M} \mathrm{LME}_{j=1}^{K} \left\{\log p\left(\mathbf{x}^{(i)} \mid \mathbf{r}^{(j)}; \theta_L^{A*}\right)\right\}$ where $\{\mathbf{r}^{(j)}\}_{j=1}^{K} \sim p\left(\mathbf{r}; \theta_P^B\right)$
3: Save converged prior parameters θ_P^{B*}

Learn T_B **Posterior**: $q\left(\mathbf{r} \mid \mathbf{x}; \phi_{\text{NSC}}^B\right)$

4: $\{\mathbf{x}'^{(i)}, \mathbf{r}'^{(i)}\}_i^V \sim p\left(\mathbf{x} \mid \mathbf{r}; \theta_L^{A*}\right) p\left(\mathbf{r}; \theta_P^{B*}\right)$
5: $\phi_{\text{NSC}}^{B*} \leftarrow \arg\max_{\phi_{\text{NSC}}^B} \sum_{i=1}^{V} \log q\left(\mathbf{r}'^{(i)} \mid \mathbf{x}'^{(i)}; \phi_{\text{NSC}}^B\right)$
6: Return learned posterior parameters ϕ_{NSC}^{B*}

3.1 Results

Predictive models under T_A We first learned the generative model under T_A (A.5), and as a sanity check, computed the NSC posterior (NSC T_A), and compared its log-likelihood performance to a gold-standard system identification model trained under T_A (SI T_A). We found that the performance of the two match in neural predictive performance, supporting that the generative model was captured well (Fig 2**D**).

Generalization to T_B Under T_B, only the prior was relearned while keeping the likelihood learned from T_A fixed (Fig 2**B**). The new prior is learned without using any neural responses but only

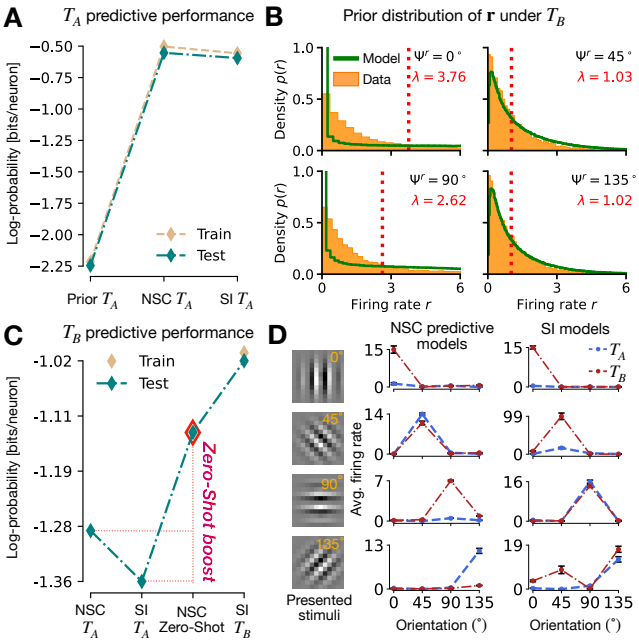

Figure 2: **Zero-shot adaptation to novel context T_B** **A** Performance of predictive models on T_A, trained on T_A data: Prior model serves as the baseline; NSC T_A is the posterior model learned via variational inference on the T_A generative model; SI T_A is the system identification model trained to predict T_A responses to T_A images. **B** Subplots show response distributions for single neurons r_j under T_B, with orange representing response distribution, green line the density learned by the flow prior *using only T_B images, without response data*, ψ^r the orientation preference of the r_j and λ its mean. **C** Performance of predictive models on T_B, demonstrating zero-shot generalization of the NSC posterior. NSC T_A and SI T_A are baselines trained on data under T_A; NSC Zero-Shot shows the posterior learned from the generative model *adapted* to T_B using our proposed method *using only T_B images, and no neural data* (Algorithm 1); SI T_B is the system identification model trained on T_B. **D** Left column shows the presented stimulus; right shows mean predictions from the NSC predictive and SI models under T_A and T_B for neurons with their orientation preference on the x-axis. Labels T_A and T_B here refer to the context of the predictive model. Note that task T_A places higher prior on $45°$ and $135°$ and context T_B places higher prior on $0°$ and $90°$ (see Fig 1**E—H**).

on stimuli from T_B (Algorithm 1). We then compute the posterior model under T_B using the newly learned prior, and evaluate its log-likelihood performance on the adapted neuronal responses conditioned on images under T_B. Importantly, we also train a system identification model directly on image-response pairs on T_B as the gold-standard performance. As baselines, we also evaluate the performance of the predictive models fit only on data from T_A. We find that our adapted model (NSC-Zero-Shot) yields promising results (Fig 2**C**): it significantly outperforms the baseline models (NSC-T_A, SI-T_A), moving closer to the performance of the system identification model (SI-TB) which was trained explicitly on neural data from T_B, demonstrating the capabilities of our zero-shot generalization to novel context.

Lastly, we tested whether the predictive models provide predictions that match our intuition, such as reflecting the prior of the tasks, and the tuning properties of neurons. We find the predictive models indeed: (1) predict a relatively higher activation for a neuron whose preferred orientation is present in the stimulus, and (2) reflect the task specific prior (e.g. firing rate of $0°$ neuron in T_B is higher than in T_A) (Fig 2**D**).

4 Discussion

In this work, we have proposed a novel approach based on NSC and Bayesian inference to achieve zero-shot generalization of neural predictive models across stimulus distributional shifts.

While our method demonstrates promising results on zero-shot adaptation of neural predictive models, it opens up a few key areas for further investigation. First, we observed that as we increase the dimensionality of our datasets, the Monte-Carlo approximation for prior learning suffered from high variance, making training challenging. Second, we observed a notable gap still present in the performance between NSC Zero-Shot and the system identification model which can be attributed to the approximate nature of the zero-shot prior. These issues may all be potentially addressed by employing other approximation methods such as importance-weighted variational inference [29] to evaluate the integral for prior learning. Furthermore, once the prior is learned, computing the posterior

of generative models that differ in the priors falls under Bayesian model reduction (BMR) [30–32]. Given that we parameterize our priors and posteriors flexibly using normalizing flows and deep neural networks, we used generic scalable variational inference relying on gradient descent and backpropagation [33, 34] to learn the posterior. However, following advances from BMR would likely improve the performance of our adapted posteriors.

Furthermore, while our approach assumes that neurons follow NSC, we note that NSC is only one—albeit prominent—implementation of neurons encoding the posterior. Future work can explore how to generalize the adaptation methodology to assume neurons to encode an aspect or function of the posterior—such as encoding its sufficient statistics—and our work paves the way for this direction.

Finally, a crucial avenue for future work is applying our method on empirical data. This would meaningfully extend the learning of generative models from cortical networks [24, 35, 36] to predicting responses in new sensory contexts.

Overall, our method not only proposes a novel normative approach to achieving zero-shot generalization of predictive models of neural activity in response to stimulus distribution shifts, but also presents a rigorous test—once fitted and tested on experimentally recorded neuronal response data—of whether the empirically observed neural adaptations align with the Bayesian Brain hypothesis or reflect alternative mechanisms like efficient [37] or predictive coding [38].

Acknowledgements

We thank all the reviewers for their valuable and constructive feedback. We additionally thank Xaq Pitkow, Andreas Tolias, and members of Sinz- and Walker-lab for helpful and stimulating discussions. SS and FHS are supported by the German Research Foundation (DFG): SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms, TP 06, project number: 276693517. RMH was supported by NIH/R01 EY028811 and NSF/CAREER IIS-2143440.

References

[1] Barry Wark, Brian Nils Lundstrom, and Adrienne Fairhall. “Sensory adaptation”. In: *Current opinion in neurobiology* 17.4 (2007), pp. 423–429.

[2] Samuel G Solomon and Adam Kohn. “Moving sensory adaptation beyond suppressive effects in single neurons”. In: *Current biology* 24.20 (2014), R1012–R1022.

[3] Eero P Simoncelli and Bruno A Olshausen. “Natural image statistics and neural representation”. In: *Annual review of neuroscience* 24.1 (2001), pp. 1193–1216.

[4] Marlene R Cohen and William T Newsome. “Context-dependent changes in functional circuitry in visual area MT”. In: *Neuron* 60.1 (2008), pp. 162–173.

[5] Daniel LK Yamins et al. “Performance-optimized hierarchical models predict neural responses in higher visual cortex”. In: *Proceedings of the national academy of sciences* 111.23 (2014), pp. 8619–8624.

[6] William F Kindel, Elijah D Christensen, and Joel Zylberberg. “Using deep learning to reveal the neural code for images in primary visual cortex”. In: *arXiv preprint arXiv:1706.06208* (2017).

[7] David Klindt et al. “Neural system identification for large populations separating “what” and “where””. In: *Advances in Neural Information Processing Systems* 30 (2017).

[8] Martin Schrimpf et al. “Brain-score: Which artificial neural network for object recognition is most brain-like?” In: *BioRxiv* (2018), p. 407007.

[9] Santiago A Cadena et al. “Deep convolutional models improve predictions of macaque V1 responses to natural images”. In: *PLoS computational biology* 15.4 (2019), e1006897.

[10] Konstantin-Klemens Lurz et al. “Generalization in data-driven models of primary visual cortex”. In: *BioRxiv* (2020), pp. 2020–10.

[11] Mohammad Bashiri et al. “A flow-based latent state generative model of neural population responses to natural images”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 15801–15815.

[12] Santiago A Cadena et al. "Diverse task-driven modeling of macaque V4 reveals functional specialization towards semantic tasks". In: *bioRxiv* (2022), pp. 2022–05.

[13] Konstantin F Willeke et al. "Deep learning-driven characterization of single cell tuning in primate visual area V4 unveils topological organization". In: *bioRxiv* (2023), pp. 2023–05.

[14] Tai Sing Lee and David Mumford. "Hierarchical Bayesian inference in the visual cortex". In: *JOSA a* 20.7 (2003), pp. 1434–1448.

[15] József Fiser et al. "Statistically optimal perception and learning: from behavior to neural representations". In: *Trends in cognitive sciences* 14.3 (2010), pp. 119–130.

[16] Ralf M Haefner et al. "How does the brain compute with probabilities?" In: *arXiv preprint arXiv:2409.02709* (2024).

[17] Patrik Hoyer and Aapo Hyvärinen. "Interpreting neural response variability as Monte Carlo sampling of the posterior". In: *Advances in neural information processing systems* 15 (2002).

[18] Pietro Berkes et al. "Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment". In: *Science* 331.6013 (2011), pp. 83–87.

[19] Ralf M Haefner, Pietro Berkes, and József Fiser. "Perceptual decision-making as probabilistic inference by neural sampling". In: *Neuron* 90.3 (2016), pp. 649–660.

[20] Gergő Orbán et al. "Neural variability and sampling-based probabilistic representations in the visual cortex". In: *Neuron* 92.2 (2016), pp. 530–543.

[21] Rodrigo Echeveste et al. "Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference". In: *Nature neuroscience* 23.9 (2020), pp. 1138–1149.

[22] Camille Rullán Buxó and Cristina Savin. "A sampling-based circuit for optimal decision making". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 14163–14175.

[23] Dylan Festa et al. "Neuronal variability reflects probabilistic inference tuned to natural image statistics". In: *Nature communications* 12.1 (2021), p. 3635.

[24] Suhas Shrinivasan et al. "Taking the neural sampling code very seriously: A data-driven approach for evaluating generative models of the visual system". In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023. URL: `https://openreview.net/forum?id=Pl416tPkNv`.

[25] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[26] Bernhard Schölkopf et al. "Toward causal representation learning". In: *Proceedings of the IEEE* 109.5 (2021), pp. 612–634.

[27] Bruno A Olshausen and David J Field. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images". In: *Nature* 381.6583 (1996), pp. 607–609.

[28] Bruno A Olshausen and David J Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision research* 37.23 (1997), pp. 3311–3325.

[29] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. "Importance weighted autoencoders". In: *arXiv preprint arXiv:1509.00519* (2015).

[30] Karl Friston, Thomas Parr, and Peter Zeidman. "Bayesian model reduction". In: *arXiv preprint arXiv:1805.07092* (2018).

[31] Lancelot Da Costa et al. "Active inference on discrete state-spaces: A synthesis". In: *Journal of Mathematical Psychology* 99 (2020), p. 102447.

[32] Lancelot Da Costa et al. "Possible principles for aligned structure learning agents". In: *arXiv preprint arXiv:2410.00258* (2024).

[33] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[34] Danilo Rezende and Shakir Mohamed. "Variational inference with normalizing flows". In: *International conference on machine learning*. PMLR. 2015, pp. 1530–1538.

[35] Takuya Isomura, Hideaki Shimazaki, and Karl J Friston. "Canonical neural networks perform active inference". In: *Communications Biology* 5.1 (2022), p. 55.

[36] Takuya Isomura et al. “Experimental validation of the free-energy principle with in vitro neural networks”. In: *Nature Communications* 14.1 (2023), p. 4547.

[37] Horace B Barlow et al. “Possible principles underlying the transformation of sensory messages”. In: *Sensory communication* 1.01 (1961), pp. 217–233.

[38] Rajesh PN Rao and Dana H Ballard. “Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects”. In: *Nature neuroscience* 2.1 (1999), pp. 79–87.

[39] Konstantin-Klemens Lurz et al. “Bayesian Oracle for bounding information gain in neural encoding models”. In: *The Eleventh International Conference on Learning Representations*. 2023. URL: `https://openreview.net/forum?id=iYC5hOMqUg`.

[40] Adam Paszke et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32 (2019).

A Appendix

A.1 Novel context prior learning

In this work, we learn a new prior distribution $p\left(\mathbf{r};\theta_P^B\right)$ under context T_B solely from stimuli sampled from the stimulus distribution from the new context $p\left(\mathbf{x} \mid T_B\right)$ (Fig 1**B**, Algorithm 1). Here we derive the corresponding log-likelihood maximization objective:

$$
\begin{aligned}
&\underset{\theta_P^B}{\operatorname{argmax}}\, p(\{\mathbf{x}\}_{i=1}^M;\theta_P^B) \\
&= \underset{\theta_P^B}{\operatorname{argmax}} \log p(\{\mathbf{x}\}_{i=1}^M;\theta_P^B) \\
&= \underset{\theta_P^B}{\operatorname{argmax}} \sum_{i=1}^M \log p\left(\mathbf{x}^{(i)};\theta_P^B\right) \\
&= \underset{\theta_P^B}{\operatorname{argmax}} \sum_{i=1}^M \log \int p\left(\mathbf{x}^{(i)} \mid \mathbf{z};\theta_L^{A*}\right) p\left(\mathbf{z};\theta_P^B\right) \mathrm{d}\mathbf{z} \\
&\approx \underset{\theta_P^B}{\operatorname{argmax}} \sum_{i=1}^M \log \frac{1}{S} \sum_{j=1}^K p\left(\mathbf{x}^{(i)} \mid \mathbf{z}^{(j)};\theta_L^{A*}\right), \text{ where } \left\{\mathbf{z}^{(j)}\right\}_{j=1}^K \sim p\left(\mathbf{z};\theta_P^B\right) \\
&\approx \underset{\theta_P^B}{\operatorname{argmax}} \sum_{i=1}^M \left\{\log \sum_{j=1}^K \exp\left(\log p\left(\mathbf{x}^{(i)} \mid \mathbf{z}^{(j)};\theta_L^{A*}\right)\right) - \log K\right\} \\
&\approx \underset{\theta_P^B}{\operatorname{argmax}} \sum_{i=1}^M \left\{\underset{j=1,\ldots,K}{\mathrm{LSE}} \log p\left(\mathbf{x}^{(i)} \mid \mathbf{z}^{(j)};\theta_L^{A*}\right) - \log K\right\} \\
&\approx \underset{\theta_P^B}{\operatorname{argmax}} \sum_{i=1}^M \left\{\underset{j=1,\ldots,K}{\mathrm{LME}} \log p\left(\mathbf{x}^{(i)} \mid \mathbf{z}^{(j)};\theta_L^{A*}\right)\right\},
\end{aligned}
$$

where in the integral in step 4 has been converted into a Monte-Carlo sum in step 5; "LSE" in step 7 stands for the Log-Sum-Exp operator that evaluates expression in step 6 with better numerical stability; "LME" in step 8 stands equivalently for Log-Mean-Exp operator, a short hand for expression in step 7.

A.2 Synthetic data model

Here we describe the NSC model we use to generate synthetic data. The model was introduced by Haefner, Berkes, and Fiser [19].

The model assumes that sensory neurons $\mathbf{r}$ represent the presence of oriented Gabor features, and higher-level cortical areas encode the task-relevant grating variables ($\mathbf{g}$) and decision variable D associated with the orientation discrimination task. The brain's goal is to infer both the local sensory features ($\mathbf{r}$), oriented "objects" ($\mathbf{g}$) and the decision variable D using posterior inference given an incident stimulus $\mathbf{x}$, i.e., compute $p\left(D, \mathbf{g}, \mathbf{r} \mid \mathbf{x}\right)$.

The model assumes the brain has learned two choices in an orientation discrimination task: $D = 1$ and $D = 2$, corresponding to different stimulus orientations. The choices are equally probable:

$$p_D(D = 1) = p_D(D = 2) = 0.5.$$

The brain learns task-relevant orientations ψ_1 and ψ_2 with uncertainty, modeled by a circular Gaussian (von Mises) distribution:

$$g_i \mid D \sim \text{Bernoulli}\left\{\frac{1}{n_g l_0(\kappa)} \exp\left[\kappa \cos 2\left(\psi_i^{(g)} - \psi_D\right)\right]\right\},$$

where g_i is a binary variable indicating the presence of a grating of orientation $\psi_i^{(g)}$, ψ_D corresponds to the target orientation for decision D, κ represents the concentration or sharpness of the orientation tuning in the von Mises (circular Gaussian) distribution ($\kappa = 0$ reflects no knowledge about task-relevant orientations, while $\kappa \to \infty$ indicates perfect knowledge of these orientations).

Sensory responses, represented by $\mathbf{r}$, are modeled as:

$$\tau_i = E[r_i \mid g] = 1 + \delta \sum_{k=1}^{n_g} g_k \exp\left[\lambda \cos 2\left(\psi_i^{(x)} - \psi_k^{(g)}\right)\right],$$

where ψ_i^x is the orientation preference of r_i, λ controls the strength of the relationship between a grating variable g_k (representing a task-relevant orientation) and a Gabor-shaped feature x_i (it adjusts how much the similarity between the grating's orientation and the sensory feature influences the expected response of the sensory neuron), and δ modulates the overall strength of the task-related influence on the sensory responses in the model.

The probability of the sensory input r_i given the grating g is:

$$p(r_i \mid g) = \frac{1}{\tau_i} \exp\left(-\frac{r_i}{\tau_i}\right),$$

where τ_i is the expected value of r_i. Finally, the likelihood of the image $\mathbf{x}$ given $\mathbf{r}$ is:

$$p(\mathbf{x} \mid \mathbf{r}) = \mathcal{N}\left(\mathbf{x} \mid \sum_{i=1}^{n_r} \mathrm{PF}_i r_i, \sigma_x^2 \mathbb{I}\right).$$

where PF_i contains the projective fields for each r_i, which is a Gabor filter with orientation $\phi_i^{(x)}$, and n_r denotes the number of sensory neurons.

The image presented to the subject/used here as the observation is a noisy version of the linear combination of these projective fields, weighted by the sensory responses (latents) $\mathbf{r}$.

A.2.1 Parameter values used for data generation

In our synthetic data model, the prior probability for each decision, p_c, is set to 0.5. The task-relevant orientations, ψ_1 and ψ_2, are defined as $\frac{\pi}{4}$ and $\frac{3\pi}{4}$ for task T_A, and 0 and $\frac{\pi}{2}$ for task T_B, respectively. The concentration parameter for the von Mises distribution, κ, is set to 10.0, indicating strong orientation tuning. We set $\psi^{(g)}$t to 9 equally spaced orientation angles between 0 and π. The strength of the interaction between the sensory responses and the grating orientations, λ, is set to 10.0, and task modulation feedback strength, δ, is set to 5. The sensory neurons are assumed have PFs with orientations $\psi^{(x)} = \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$. The observation noise, $\sigma_{\mathbf{x}}$, is 0.1, and the image size is set to 12×12 pixels. We use $10,000$ samples for each task, with a fixed random seed of 42 to ensure reproducibility, with a 0.7:0.2:0.1 train:validation:test split.

A.3 Detailed description of models

Our models mostly mirror those used in [24], and describe them in detail here.

We model the likelihood as an isotropic Gaussian distribution $p\left(\mathbf{x} \mid \mathbf{r}^{(i)}\right) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{2(i)} \cdot \mathbf{I}\right)$, where the parameters mean, $\boldsymbol{\mu}^{(i)} \in \mathbb{R}^{|\mathbf{x}|}$ and variance, $\boldsymbol{\sigma}^{(i)} \in \mathbb{R}_{>0}^{|\mathbf{x}|}$ are functions of response, $\mathbf{r}^{(i)}$, and $|\mathbf{x}|$ is the number of dimensions of $\mathbf{x}$. We consider a nonlinear function $\boldsymbol{\mu} = w_{\boldsymbol{\mu}}\mathrm{MLP}(\mathbf{r}^{(i)}) + b_{\boldsymbol{\mu}}$ and $\boldsymbol{\sigma} = \exp^{w_{\boldsymbol{\sigma}}\mathrm{MLP}(\mathbf{r}^{(i)}) + b_{\boldsymbol{\sigma}}}$, where $\mathrm{MLP}(\cdot)$ stands for multi-layer perceptron, where choosing a linear mapping was left as a hyperparameter.

We model the prior as a normalizing flow: $p(\mathbf{r}; \theta_P) = p_{\text{base}}(T^{-1}(\mathbf{r}; \theta_P)) \cdot \mid \frac{\partial T^{-1}(\mathbf{r}; \theta_P)}{\partial \mathbf{r}} \mid$, where we choose p_{base} to be a full-covariance multivariate normal distribution (not factorized as in [24], and hence our flow learns dependencies among neurons), and T^{-1} represents the following series of invertible mappings with learnable parameters θ_P: $[\text{affine}, \tanh, \text{affine}, \tanh, \text{affine}, \text{softplus}^{-1}]$,

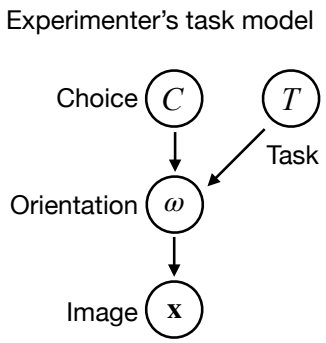

Figure 3: Experimenter's task model that creates the two different stimulus distributions (contexts).

where $\text{softplus}^{-1}(y) = \log(e^y - 1)$, $\text{affine}(y) = ay + b$ with learnable parameters a and b. softplus^{-1} ensures that the support of $\mathbf{r}$ is non-negative, since the prior is modeling the distribution of (non-negative) firing rates. Each dimension (i.e., neuron $\mathbf{r}$) after the correlated base distribution is treated independently, with the affine layers applied per dimension and restricted to be diagonal.

We model the posterior distribution of responses conditioned on images as a factorized Gamma distribution, following state-of-the-art (SOTA) work in system identification [39]: $p\left(\mathbf{r} \mid \mathbf{x}^{(i)}\right) = \prod_{j=1}^{S} p_{\Gamma}\left(\mathbf{r}_j \mid \boldsymbol{\alpha}^{(i)}, \boldsymbol{\beta}^{(i)}\right)$, where $\mathbf{x}^{(i)}$ is the ith image, $\mathbf{r}_j$ is the jth neuron out of $\mid \mathbf{r} \mid = S$ total neurons, and the parameters concentration, $\boldsymbol{\alpha}^{(i)}$ and rate, $\boldsymbol{\beta}^{(i)}$ are functions of the image, $\mathbf{x}^{(i)}$, modeled as an MLP.

We split the 10k datapoints in our simulated datasets into 7k for train, 2k for validation and 1k for testing. We train all models using backpropagation and gradient descent, implemented using the PyTorch library [40].

A.4 Orientation discrimination task

We identified a change in the stimulus distributions as part of specific tasks. We assume a two-alternative forced choice (2AFC) task design, where subjects are trained on two separate orientation classification tasks, Tasks 1 and 2 (Fig 1**C**). In each trial, the subject is shown an oriented Gabor patch (angle ω), drawn from one of two classes—$C = 1$ or $C = 2$—each described by a circular Gaussian probability distribution (von-Mises distribution). The two tasks differ in mean orientations (see the two task distributions in Fig 1**E** and **G**) but share the same variability ($\kappa = 1$) [4]. For both tasks, the class was randomly chosen with equal probability, meaning $p\left(C = 1\right) = p\left(C = 2\right) = 0.5$ for every trial. The stimulus orientation was then drawn from the task-specific distribution $p\left(\omega \mid C, T\right)$.

The exact generative model used to define the relevant probability distributions in the task in presented in Fig 3 as the experimenter's task generative model.

For our simulated data, we had two contexts or tasks $T = T_A$ and $T = T_B$. For T_A, we chose mean orientations to be 45° and 135° (Fig 1**E**), and for T_B we chose 0° and 90° (Fig 1**G**).

This framework is conceptualized with monkeys as subjects and recordings from monkey V1, but the general approach and underlying principles, including the experimental design, are more broadly applicable.

A.5 Learning generative model under T_A

Evaluating the generative model $p\left(\mathbf{x}, \mathbf{r}; \theta_L^A, \theta_P^A\right)$ requires evaluating the learned prior $p\left(\mathbf{r}; \theta_P^A\right)$ and the likelihood $p\left(\mathbf{x} \mid \mathbf{r}; \theta_L^A\right)$. For the synthetic data model, computing $p\left(\mathbf{r}\right)$ is computationally expensive since it involves computing the sum $p(\mathbf{r}) = \sum_C \sum_g \; p(\mathbf{r} \mid \mathbf{g}) p(\mathbf{g} \mid C, T) p(C)$, and consequently we evaluate the model qualitatively by examining the learned densities (Fig 4**A**). Our flow prior model shows a promising fit to the data distribution. For neurons with 0° and 90° orientation preferences, where the average firing rate is relatively low ($\lambda \approx 1$), the model achieves a near-perfect fit, whereas for other cases ($\lambda > 3$), we observe a slight mismatch, likely due to

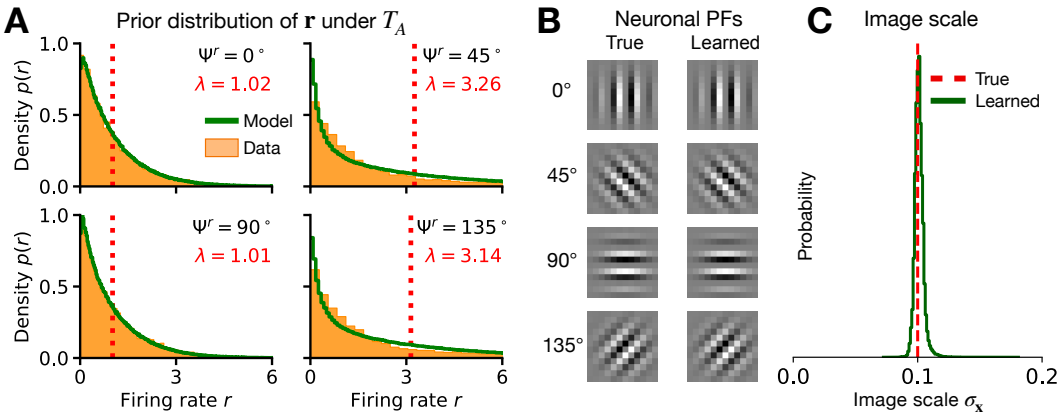

Figure 4: **A** Subplots show response distributions for single neurons r_j under T_A, with the green line representing the density learned by the flow prior via MLE on T_A responses, ψ^r denotes orientation preference and λ the avg firing rate. **B** PFs learned by the MLP likelihood model compared to the true PFs. **C** Image scale parameter learned by the likelihood model.

fitting via MLE on observed data with high variance. The likelihood $p\left(\mathbf{x} \mid \mathbf{r}; \theta_L^A\right)$ models the true distribution $p(\mathbf{x} \mid \mathbf{r}) = \mathcal{N}\left(\mathbf{x} \mid \mu_{\mathbf{x}} = \sum_{i=1}^{n_r} \mathrm{PF}_i r_i, \sigma_{\mathbf{x}}^2 \mathbb{I}\right)$, where n_r is the number of neurons, and PF_i represents the oriented Gabor-based projective field of neuron r_j. The MLP-based likelihood effectively captures this distribution, achieving accurate PF reconstruction ($\mu_{\mathbf{x}}$) (Fig 4**B**) as well the scale ($\sigma_{\mathbf{x}}^2$) (Fig 4**C**). This strong performance is critical for generalizing the model to adapted responses under the shifted stimulus distribution (T_B), as we transfer the likelihood model without re-learning it for T_B.