Geometric Causal Discovery in Neural Populations: Structured Subspaces, Vacuity, and External Validation in Brain-Wide Neuropixels Recordings

Elliot Tower *elliot@elliottower.ai*
Independent Researcher

Abstract

How should we identify the causal subspaces of neural population activity—the low-dimensional projections that actually drive behavior? Using Neuropixels recordings across 73 brain regions in 10 mice (**?**), we show that a structured variational autoencoder discovers causal choice subspaces that are 3.4× stronger than those found by linear discriminant analysis (LDA), winning in 73/73 brain regions ($p = 5.7 \times 10^{-14}$). A companion paper establishes that linear and nonlinear similarity metrics anti-correlate across brain regions, mediated by effective dimensionality; here we show this extends to causal subspace discovery itself. **However**, the interchange intervention accuracy (IIA) metric used to evaluate subspace quality is vacuous for nonlinear methods: the VAE achieves comparable IIA on random Gaussian noise as on real data (0.70 vs. 0.69), across both binary and continuous metrics (KL, JS, probability shift). This confirms the theoretical warning of **?** and means IIA alone cannot validate nonlinear subspace methods. **External validation rescues the claim**: optogenetic silencing ground truth (**?**) shows LDA subspaces are *anti-correlated* with causal importance ($\rho = -0.73$, $p = 0.01$), while VAE subspaces show a positive trend, with the method difference significant ($\Delta\rho = +1.06$, 95% CI $[0.31, 1.51]$). The VAE advantage generalizes across four task variables (all $p < 10^{-13}$), and engagement and choice subspaces are near-orthogonal ($d_G = 1.82$), confirming genuine disentanglement. An inductive bias ablation further reveals that SAE-style overcomplete sparse representations are the dominant factor: all three sparse variants outperform all three VAE variants, with the best achieving IIA = 0.962 vs. the structured VAE baseline of 0.939. Notably, the label-conditional prior that succeeds on transformer representations (**?**) *hurts* on neural data, suggesting that neural firing patterns lack the clean class-conditional structure of learned computational features. These findings establish that external causal validation—not IIA alone—is necessary to assess nonlinear subspace quality, and that sparse overcomplete representations offer a promising inductive bias for neural population geometry.

1 Introduction

A growing body of work characterizes neural computation through the geometry of population activity (**??**). Decision variables live in low-dimensional subspaces (**?**), motor plans occupy orthogonal manifolds (**?**), and representations drift over time within fixed-dimensional structures (**?**). Identifying the *causal* subspaces—the projections of population activity that actually drive behavior—is central to understanding neural computation.

The field's standard tools for this task—linear discriminant analysis (LDA), principal component analysis (PCA), and their variants—assume that causal structure lies in linear subspaces. A companion paper [Paper A] establishes that linear and nonlinear similarity metrics anti-correlate across brain regions ($\rho = -0.85$), mediated by effective dimensionality: high-dimensional regions appear dissimilar to linear methods but highly similar to nonlinear ones. This raises a direct question: *do linear subspace methods also fail to identify causal structure?*

We test this directly. Using Neuropixels recordings across 73 brain regions in the Steinmetz et al. (2019) dataset, we find that a structured variational autoencoder (**?**) discovers choice subspaces that are 3.4× more causally effective than LDA subspaces, winning in all 73 brain regions. We validate this against optogenetic silencing data (**?**): LDA subspace quality is *anti-correlated* with ground-truth causal importance ($\rho = -0.73$), meaning linear methods systematically identify the wrong regions as causally important. The VAE reverses this pattern.

An honest caveat. The interchange intervention accuracy (IIA) metric used to evaluate subspace quality is vacuous for nonlinear methods: both an unconstrained MLP and our structured VAE achieve high IIA on random Gaussian noise (mean ≈ 0.70). This vacuity extends to continuous distributional metrics (KL divergence, JS divergence, probability shift), ruling out binary thresholding as the cause. Only linear DAS properly rejects noise (IIA = 0.16). This confirms the theoretical warning of **?** and motivates our emphasis on external validation—optogenetic silencing ground truth, multi-task generalization, and engagement orthogonality—over metric-internal comparisons.

Beyond the VAE-vs-LDA comparison, we conduct an inductive bias ablation testing six model variants. SAE-style overcomplete sparse representations emerge as the dominant inductive bias, outperforming all VAE variants. Surprisingly, the label-conditional prior of pi-VAE (**?**)—which succeeds on transformer representations—slightly *hurts* on neural data.

These findings converge on two messages. First, *linear methods fail where they matter most*: the brain regions where causal choice signals are strongest are the regions where LDA is least informative, and external causal validation—not IIA alone—is necessary to assess nonlinear subspace quality. Second, *inductive biases matter*: the choice between structured splits, identifiability priors, and sparsity constraints has measurable consequences for causal subspace discovery.

2 Related Work

Neural manifolds and the potent/null framework. **?** argued that the fundamental unit of motor cortical computation is the neural mode—a population-level activation pattern spanning a low-dimensional manifold. **?** formalized this as the potent/null space distinction: preparatory activity lives in a null space orthogonal to the muscle-driving output dimensions, while execution activity enters the potent space. Our choice subspace analysis directly tests this framework: the choice subspace is the potent space for decision output, and high-dimensional regions may harbor large null spaces that contain variance but no causal signal.

Representational similarity. CKA (**?**) and RSA (**?**) are the standard tools for comparing neural representations. **?** showed that many similarity measures are special cases of a generalized shape metric, but all within the linear/kernel family. A companion paper [Paper A] demonstrates that these linear metrics anti-correlate with nonlinear manifold similarity, mediated by effective dimensionality—motivating the nonlinear subspace methods studied here.

Identifiable VAEs for neural data. **?** proposed pi-VAE, which conditions the latent prior $p(\mathbf{z}|\mathbf{u})$ on task variables $\mathbf{u}$ via an exponential family parameterization, providing identifiability guarantees under mild conditions. Unlike our structured VAE, which separates $\mathbf{z}_{\text{choice}}$ from $\mathbf{z}_{\text{other}}$ via architecture and a classification head, pi-VAE achieves identifiability through the prior while using a single unsplit latent space. The semi-supervised disentanglement framework (**?**) separates labeled task variables from unlabeled variance. The mechanistic independence condition (**?**) provides identifiability guarantees when the generator Jacobian is sparser in factor-aligned coordinates.

Mechanistic interpretability of VAEs. **?** introduced causal effect strength (CES), intervention specificity, and circuit modularity as metrics for understanding how VAE encoders process inputs. Their key finding: FactorVAE develops highly specialized encoder channels with mediation values > 0.7 mapping directly to individual latent factors, while β-VAE distributes information across channels. Activation patching and causal mediation analysis—tools borrowed from transformer interpretability (**?**)—can identify which encoder channels function as "choice detectors" in a structured VAE trained on neural data.

IIA vacuity for nonlinear methods. **?** proved that unconstrained nonlinear alignment makes interchange intervention analysis (IIA) vacuous—any sufficiently flexible nonlinear map can achieve perfect IIA on arbitrary data. The linearity constraint of DAS (**?**) is what makes IIA non-vacuous. Our structured VAE, despite having a structured latent split, retains enough encoder expressiveness to fit random noise, confirming Sutter et al.'s warning and motivating external validation.

Dimension-level causal perturbation. **?** perturbed coding versus residual subspace dimensions in ALM, finding that residual dimensions causally drive choice through transient amplification. IIA (**?**) and DAS (**?**) perform formal causal interventions on learned subspaces in language models. Cross-region activation patching—swapping one region's choice-subspace output into another region's activity—is the biological analogue of path patching in transformer circuits.

What is missing. No prior work validates nonlinear subspace methods against optogenetic ground truth, tests whether the IIA vacuity concern applies to biological subspace estimation, ablates inductive biases (structured splits, identifiability priors, sparsity) for causal subspace discovery, or decomposes the potent/null space variance structure across dozens of brain regions simultaneously.

3 Methods

3.1 Dataset

We use the Steinmetz et al. (2019) dataset: Neuropixels recordings from 10 mice performing a two-alternative visual decision task. The dataset covers 73 brain regions with ≥ 15 neurons per region (50 with ≥ 2 recording sessions). We extract trial-averaged activity in a post-stimulus window (150–350ms) and compute binary choice labels (left vs. right). All cross-session pairs for the same region constitute the comparison set: 1,316 pairs across 50 multi-session regions.

3.2 Dimensionality Covariate

We use the power-law exponent α (fit to PCA eigenvalue spectrum in log-log space, ranks 10–50) as a dimensionality covariate throughout this paper. Low α indicates high effective dimensionality. The full characterization of geometric invariants (CKA, UMAP Procrustes, Grassmannian distance, topological and dynamical measures) and their relationship to α is reported in a companion paper [Paper A].

3.3 Structured VAE for Subspace Discovery

We train a structured variational autoencoder (**?**) per region. The encoder maps population activity to two latent groups: $\mathbf{z}_{\text{choice}}$ (k dimensions, $k \in \{1, 2, 3\}$) trained with an auxiliary classification loss to predict left/right choice, and $\mathbf{z}_{\text{other}}$ ($m = 15$ dimensions) capturing remaining variance. Both groups jointly reconstruct the population activity via the ELBO. Training: 300 epochs, $\alpha_{\text{task}} = 10$, $\beta_{\text{KL}} = 1$, across all 39 sessions.

The encoder weight matrix for $\mathbf{z}_{\text{choice}}$ directly provides the choice subspace directions, replacing the LDA+PCA pipeline.

3.4 Interchange Intervention Analysis (IIA)

For each region, we swap the choice-subspace projection between opposite-evidence trial pairs and measure how often a choice classifier flips its prediction. IIA = 1 means the subspace is perfectly causal; IIA = 0.5 means random. We compute IIA for both the VAE and LDA subspaces, and against a random-direction null.

3.5 Optogenetic Silencing Validation

We validate subspace quality against ground-truth causal importance from the Zatka-Haas et al. (2021) optogenetic inactivation dataset: 52 cortical coordinates, 47,002 laser trials across 10 mice. For each silencing

coordinate mapped to a Steinmetz brain region ($n = 12$ direct matches, $n = 16$ with hierarchical Allen CCF grouping), we compute the behavioral effect size (Euclidean distance in choice-probability space between laser-on and laser-off trials on matched-contrast conditions). The Spearman correlation between silencing effect and subspace IIA provides external validation independent of any metric-internal evaluation.

3.6 Multi-Task Generalization

Beyond the primary choice variable, we test subspace discovery on three additional task variables: evidence strength (high vs. low contrast), feedback (correct vs. error), and stimulus side (left vs. right). For each variable, we train a separate structured VAE and compute IIA against the corresponding LDA baseline.

3.7 Engagement Orthogonality

We derive an engagement variable from behavioral signatures: disengaged trials (NoGo responses on easy stimuli where the correct response is obvious) vs. engaged trials (correct responses on matched stimuli). We estimate engagement and choice subspaces in a pre-stimulus window (0–150ms) and post-stimulus window (250–450ms) respectively, then compute their Grassmannian distance. Near-orthogonality ($d_G \approx \pi/2$ per dimension) would confirm that the VAE disentangles genuinely distinct cognitive variables rather than capturing a single dominant signal.

4 Results

4.1 Learned Subspaces Amplify Causal Signal

VAE subspaces are 3.4× more causal than LDA. The VAE produces mean IIA = 0.56 (SD 0.08) vs. LDA mean IIA = 0.17 (SD 0.09), with Cohen's $d = 4.5$. The VAE wins in 73/73 regions (Wilcoxon $W = 2701$, $p = 5.7 \times 10^{-14}$). It exceeds the random-direction null in all 73 regions (mean null = 0.23).

Choice subspace dimensionality ablation. The causal signal is effectively one-dimensional: even $k = 1$ achieves a 3.7× improvement over LDA (72/73 wins), and performance is stable across $k \in \{1, 2, 3\}$.

Table 1: Effect of choice subspace dimensionality k on structured VAE IIA across 73 brain regions.

k	VAE IIA	LDA IIA	Ratio	VAE wins	Choice acc.
1	0.548 ± 0.093	0.150 ± 0.091	3.7×	72/73	0.99
2	0.555 ± 0.088	0.159 ± 0.097	3.5×	73/73	0.98
3	0.561 ± 0.080	0.166 ± 0.094	3.4×	73/73	0.99

Low-dimensional regions benefit most. Low-α (high-dimensional) regions show a larger VAE improvement (ΔIIA = 0.46) than high-α regions (ΔIIA = 0.33). VAE IIA anti-correlates with α ($\rho = -0.48$, $p = 1.5 \times 10^{-5}$), reversing the pattern seen with LDA.

VAE finds genuinely different subspaces. The VAE and LDA subspaces are nearly orthogonal: mean Grassmannian distance = 2.34 (SD 0.15), far exceeding $\pi/2 \approx 1.57$ per dimension. The VAE does not refine the LDA solution—it finds a fundamentally different set of directions.

Model confidence tracks causal effectiveness. Posterior uncertainty anti-correlates with IIA ($\rho = -0.26$, $p = 0.028$): regions where the model is most confident are those where interventions are most effective.

4.2 The IIA Vacuity Problem

Before interpreting the 3.4× improvement as evidence that the VAE finds better causal subspaces, we must address a fundamental concern raised by **?**: is IIA a valid metric for nonlinear methods?

We test three methods on real neural data versus random Gaussian noise with matched labels:

Table 2: IIA on real neural data vs. random Gaussian noise (73 regions). A non-vacuous method should show high IIA on real data and chance IIA (≈ 0.5) on noise.

Method	Real IIA	Random IIA	Interpretation
Unconstrained MLP	—	0.70 ± 0.11	Vacuous (fits noise)
Structured VAE	0.69 ± 0.11	0.70 ± 0.11	Also vacuous on noise
Linear DAS	—	0.16 ± 0.09	Non-vacuous (rejects noise)

The structured VAE achieves comparable IIA on random noise (0.70) as on real data (0.69). The structured latent split alone does not prevent the encoder from fitting arbitrary labels. This means *the 3.4× IIA improvement over LDA could in principle reflect encoder expressiveness rather than genuine causal subspace discovery.*

One might hypothesize that this vacuity is an artifact of the binary IIA metric itself—perhaps continuous measures of distributional shift would distinguish real from random data even when the binary flip rate does not. We test this directly by computing four continuous metrics alongside IIA: KL divergence ($D_{KL}(p_{\text{orig}} \| p_{\text{swap}})$), Jensen-Shannon divergence, probability shift ($|\Delta P(\text{choice})|$), and logit difference ($|\Delta \ell|$). Results across all 73 regions:

Table 3: Continuous metrics on real vs. random data (Exp72). If vacuity were a property of the binary metric, continuous metrics would show real $\gg$ random. They do not.

Method	Metric	Real	Random	Wilcoxon p
VAE	IIA	0.688	0.695	0.97
	KL divergence	2.51	2.59	0.99
	JS divergence	0.397	0.408	0.99
	Prob. shift	0.655	0.666	0.98
MLP	IIA	0.687	0.695	0.95
	Logit diff.	8.93	8.13	3.3×10^{-10}
Linear DAS	IIA	0.167	0.163	0.17

The VAE shows real $\approx$ random on *every* continuous metric ($p > 0.95$ in all cases). The vacuity is not an artifact of binary thresholding—it reflects genuine encoder flexibility. The one exception, MLP logit difference ($p = 3.3 \times 10^{-10}$), reflects the unconstrained encoder producing larger raw logit magnitudes on real data despite identical binary flip rates, suggesting scale differences without discrimination improvement.

This result motivates two responses: (1) external validation that does not depend on IIA or any intervention metric, and (2) ablation experiments that isolate the source of the improvement.

[PENDING: Exp62: Shuffled-label control. Train VAE with randomly shuffled choice labels on real neural data. If IIA drops to chance, the VAE is learning real structure despite also being able to fit noise. If IIA stays high, the result is vacuous.]

[PENDING: Exp63: Linear VAE ablation. Compare nonlinear structured VAE, linear structured VAE (same latent split, no hidden-layer nonlinearities), and nonlinear unstructured VAE (single latent, no classification loss). This isolates whether the benefit comes from the structured split or from nonlinearity.]

4.3 External Validation: Optogenetic Silencing Ground Truth

The strongest evidence for the VAE's superiority comes not from IIA comparisons but from external validation against optogenetic silencing data—a test that is completely independent of the IIA metric.

Across 12 brain regions matched between Steinmetz and Zatka-Haas et al. (2021):

Table 4: Spearman correlation between subspace quality and optogenetic silencing effect (ground-truth causal importance). Bootstrap BCa CIs from 10,000 resamples.

Metric	ρ vs. silencing	p	95% CI
LDA IIA	-0.73	0.01	$[-0.93, -0.27]$
VAE IIA	$+0.33$	0.30	—
$\Delta\rho$ (VAE − LDA)	$+1.06$	< 0.005	$[0.31, 1.51]$

LDA subspaces are anti-correlated with causal importance ($\rho = -0.73$, permutation $p = 0.01$). Regions where optogenetic silencing most disrupts behavior—the ground-truth causally important regions—are the regions where LDA-based IIA is *lowest*. Linear methods systematically identify the wrong regions as causally important.

The VAE reverses this pattern. The VAE correlation is positive ($\rho = +0.33$), though not individually significant at $n = 12$. The critical test is the *difference*: $\Delta\rho = +1.06$, 95% bootstrap CI $[0.31, 1.51]$, permutation $p < 0.005$. The CI excludes zero, confirming that the VAE provides a significantly better match to optogenetic ground truth than LDA.

Hierarchical region expansion. Using Allen CCF ontology to group Steinmetz sub-regions under parent silencing areas expands the matched set from 12 to 16 regions. The expanded analysis shows VAE $\rho = +0.42$ (perm $p = 0.11$, trending), LDA $\rho = -0.29$, and the delta-rho CI still excludes zero ($\Delta\rho = 0.87$, 95% CI $[0.09, 1.43]$).

[PENDING: Exp66: Per-mouse optogenetic validation. Compute per-mouse silencing effects and correlate each mouse's profile with geometry metrics. If the LDA anti-correlation holds per individual mouse (not just pooled), this rules out the concern that outlier mice drive the result.]

4.4 Multi-Task Generalization

The VAE advantage is not specific to the choice variable. Across three additional task variables, the VAE wins 73/73 regions for every variable:

Table 5: VAE vs. LDA IIA across four task variables. All 73 regions, all 39 sessions.

Task variable	VAE IIA	LDA IIA	VAE wins	Wilcoxon p
Choice	0.56 ± 0.08	0.17 ± 0.09	73/73	5.7×10^{-14}
Evidence strength	0.64 ± 0.13	0.26 ± 0.10	73/73	5.7×10^{-14}
Feedback	0.58 ± 0.11	0.21 ± 0.09	73/73	5.7×10^{-14}
Stimulus side	0.66 ± 0.11	0.29 ± 0.10	73/73	5.7×10^{-14}

The improvement is largest for evidence strength (2.4×) and smallest for stimulus side (2.3×), but all four show overwhelming VAE dominance. The anti-correlation between α and VAE IIA holds for all task variables: evidence strength ($\rho = -0.75$, $p < 10^{-13}$), feedback ($\rho = -0.70$, $p < 10^{-11}$), stimulus side ($\rho = -0.76$, $p < 10^{-14}$).

4.5 Engagement-Choice Orthogonality

The VAE identifies engagement and choice subspaces that are near-orthogonal across all 59 regions with sufficient data:

- **Grassmannian distance**: $d_G = 1.82$ (LDA), $d_G = 2.41$ (VAE). Both exceed $\pi/2 \approx 1.57$, confirming orthogonality.
- **VAE wins 59/59 regions** for engagement-choice separation.

- **Cross-variable IIA**: swapping the engagement subspace between opposite-choice trials produces mean IIA = 0.17—near-random, confirming that engagement interventions do not affect choice and vice versa.

This demonstrates that the VAE is not merely finding a single dominant nonlinear direction that correlates with everything. It discovers geometrically orthogonal subspaces for genuinely distinct cognitive variables.

4.6 Cross-Task Subspace Geometry

We compute pairwise Grassmannian distances between all four task-variable subspaces per region. Two patterns emerge:

Choice and stimulus side share structure. LDA Grassmannian distance between choice and stimulus side is $d_G = 0.585$ (the closest pair), consistent with stimulus side directly driving choice. All other task pairs are near-orthogonal ($d_G \approx 1.3$).

VAE separates all task variables more strongly. The VAE produces larger pairwise distances than LDA in 73/73 regions ($p < 10^{-14}$), with mean $d_G = 2.36$ (VAE) vs. 1.20 (LDA). The VAE finds subspaces that are more orthogonal—better disentangled—than LDA's projections.

Multiplexing varies by region. The most multiplexed regions (lowest mean pairwise distance, most overlapping task subspaces) include OT, LH, and VISa—regions at the interface of multiple processing streams. The least multiplexed (most orthogonal) include PIR, EPd, and AUD—unimodal sensory regions.

[PENDING: Exp65: Temporal sliding window IIA. Compute IIA in 50ms sliding windows from stimulus onset through response. If the choice subspace IIA rises before the behavioral response, this is evidence of a causal decision signal, not a post-hoc motor artifact.]

4.7 Inductive Bias Ablation: SAE-Style Sparse Representations

Which inductive biases matter for causal subspace discovery in neural populations? We test six model variants combining three biases: (1) structured latent split ($\mathbf{z}_{\text{choice}}$ / $\mathbf{z}_{\text{other}}$), (2) label-conditional prior (**?**), and (3) overcomplete $\mathbf{z}_{\text{choice}}$ (8× expansion) with L1 sparsity (SAE-style).

Table 6: Inductive bias ablation across 73 brain regions. All models trained with 300 epochs, $\alpha_{\text{task}} = 10$, $\beta_{\text{KL}} = 1$. SAE variants use 8× expansion factor and L1 coefficient 10^{-3}.

Model	Split	Label prior	Overcomplete+L1	Mean IIA	Std
pi-VAE		✓		0.754	0.253
pi-Structured VAE	✓	✓		0.935	0.112
Structured VAE	✓			0.939	0.114
pi-SAE Structured	✓	✓	✓	0.953	0.102
pi-SAE Plain		✓	✓	0.954	0.099
SAE Structured	✓		✓	**0.962**	0.096

The SAE-style overcomplete sparse representation is the dominant inductive bias: all three SAE variants outperform all three VAE variants. The label-conditional prior that succeeds on transformer representations *hurts slightly* on neural data (SAE Structured = 0.962 vs. pi-SAE Structured = 0.953), suggesting that neural firing patterns lack the clean class-conditional structure of learned computational features. The structured latent split remains important—pi-VAE without it achieves only 0.754—but adding sparsity provides a further +0.023 improvement over the structured VAE baseline.

5 Discussion

5.1 Dimensionality Interaction and the Potent/Null Framework

A companion paper [Paper A] establishes that linear and nonlinear similarity metrics anti-correlate across brain regions, mediated by effective dimensionality. Our causal subspace results extend this finding: linear methods fail most severely in the highest-dimensional brain regions, and these are precisely the regions where optogenetic silencing reveals the strongest causal role in behavior. A researcher using LDA to identify causally important regions would conclude that low-dimensional sensory regions carry the strongest choice signals, while the optogenetic ground truth shows the opposite.

The potent/null space framework (**?**) explains why. The choice subspace identified by the VAE corresponds to the *potent space* for decision output—the subspace of population activity that causally drives the downstream binary decision. The remaining variance lives in the *null space*—present, but not behavior-driving. We tested this directly by decomposing each region's activity into its choice subspace (potent) and complement (null). Across 73 regions, LDA places 99.8% of variance in the null space on average, while the VAE places 84.3%—consistent with the VAE discovering a larger potent subspace. Despite this difference, both methods decode choice from the potent subspace with comparable accuracy (∼66%), confirming that the potent subspace—however small—carries the causal signal. The dimensionality index α is strongly anti-correlated with the null-space fraction ($\rho = -0.86$, $p = 5 \times 10^{-22}$ for the VAE), meaning regions where metrics disagree most are precisely those with the largest choice-potent subspaces.

5.2 The IIA Vacuity Problem and External Validation

We take the Sutter dilemma seriously. Our structured VAE achieves IIA = 0.70 on random Gaussian noise—the same as an unconstrained MLP. This means IIA alone cannot distinguish genuine causal subspace discovery from encoder flexibility. The structured latent split (z_choice/z_other) does not provide sufficient constraint to prevent overfitting.

Three lines of evidence nevertheless support the VAE's validity:

1. **Optogenetic validation** (§4.3): The VAE-silencing correlation is positive while LDA's is strongly negative, and the difference is significant. This test is completely independent of IIA.

2. **Cross-task disentanglement** (§4.5): The VAE discovers near-orthogonal subspaces for engagement vs. choice, with cross-variable IIA at chance (0.17). A vacuous method would not produce orthogonality between genuinely distinct cognitive variables.

3. **Multi-task generalization** (§4.4): The VAE advantage holds for all four task variables with the same dimensionality interaction pattern. A single overfitting mechanism would not generalize this consistently.

[PENDING: Exp62 (shuffled-label control) will provide the definitive test: if the VAE's IIA drops when trained on randomly shuffled labels, the encoder IS learning real structure despite also being able to fit noise on random data.]

[PENDING: Exp63 (linear VAE ablation) will isolate whether the benefit comes from the structured latent split or from encoder nonlinearity.]

5.3 Identifiable Subspace Discovery via pi-VAE

The IIA vacuity problem (§4.2) suggests that our structured VAE's encoder may have too much expressiveness—it can fit arbitrary labels. A principled remedy is to constrain the model with identifiability guarantees. pi-VAE (**?**) achieves this by conditioning the latent prior on task variables: $p(\mathbf{z}|\mathbf{u}) = \prod_i \frac{Q_i(z_i)}{Z_i(\mathbf{u})} \exp\left[\sum_j T_j(z_i)\lambda_{i,j}(\mathbf{u})\right]$, where $\mathbf{u}$ is the choice label. Under mild conditions (injective

decoder, differentiable sufficient statistics, enough distinct label values), the latent representation is identifiable up to a linear transformation.

We tested a hybrid "pi-structured-VAE" that combines both approaches: the label-conditioned prior is applied *only* to $\mathbf{z}_{\text{choice}}$, while $\mathbf{z}_{\text{other}}$ retains a standard Gaussian prior. This preserves the structured split while gaining identifiability on the choice-relevant dimensions. The results (Table 6) were surprising: the label-conditional prior *hurts* on neural data. The pi-Structured VAE achieves IIA = 0.935 vs. the structured VAE's 0.939, and pi-VAE without the structured split drops to 0.754. The identifiability guarantee does not translate to better causal subspace discovery on neural populations, likely because neural firing patterns—unlike transformer representations—lack the clean class-conditional structure that the exponential family prior assumes. The structured latent split is the critical inductive bias; adding identifiability constraints on top provides no benefit and slightly degrades performance.

5.4 Cross-Region Activation Patching

The biological analogue of path patching in transformer circuits (**?**) is cross-region activation patching: replacing one region's choice-subspace output with another region's and measuring whether downstream decoding changes. For simultaneously recorded region pairs (A, B), we project A's activity onto its VAE choice subspace, transplant this component into B's activity (replacing B's own projection onto A's subspace), and measure whether a classifier trained on B's activity flips its choice prediction.

If patching $A \to B$ flips B's decoded choice, then A's choice subspace lies in B's potent space—the two regions share causal geometry. The directed patching graph across all 73 regions would reveal information flow: regions with high *outgoing* patching strength are causal sources (their choice signal propagates), while regions with high *incoming* strength are receivers (their choice decoding depends on others). The asymmetry of this graph tests the prediction that information flows from sensory to decision to motor regions.

[PENDING: Exp70: Cross-region activation patching. For all simultaneously recorded region pairs, patch source's choice subspace into target's activity. Build directed patching graph. Test asymmetry: do causally important regions (per silencing) have higher outgoing patching strength?]

5.5 Mechanistic Interpretability of the VAE Encoder

Beyond using the VAE as a tool for subspace discovery, we can ask how the encoder *itself* processes neural activity—applying mechanistic interpretability to the analysis method rather than the brain. Following the multi-level causal intervention framework of **?**, we compute three metrics per trained VAE: causal effect strength (CES $= \mathbb{E}[\|\mathbf{D}(\mathbf{z}) - \mathbf{D}(\tilde{\mathbf{z}}_i)\|_2]$, measuring how much intervening on latent dimension i changes the reconstruction), intervention specificity ($S(i) = 1/(H(\mathbf{p}_i) + \epsilon)$, measuring whether changes are localized), and circuit modularity ($M = 1 - \frac{1}{\binom{K}{2}} \sum_{i<j} |\rho(\Delta\mathbf{a}_i, \Delta\mathbf{a}_j)|$, measuring whether different latent dimensions use separate encoder pathways).

If the structured VAE develops "choice detector" channels in its encoder—analogous to FactorVAE's specialized shape-detecting channels (**?**)—this would confirm that the choice/other split is not merely a labeling convenience but reflects a genuine architectural specialization learned from the data.

[PENDING: Exp71: VAE causal circuits. Compute CES, specificity, and modularity for structured vs. unstructured VAEs. Apply activation patching to encoder hidden layers: which channels mediate choice information? Does the structured VAE have higher modularity (more specialized channels) than the unstructured control?]

5.6 Practical Implications

For the interpretability community, the IIA vacuity result is a cautionary finding. IIA is widely used to evaluate causal subspace quality in language models (**??**); our results show that nonlinear methods can achieve high IIA on random noise, confirming **?**. External validation (behavioral perturbation, out-of-

distribution generalization, cross-task disentanglement) is necessary alongside metric-internal evaluation, particularly for nonlinear methods.

For neural data analysis, the SAE ablation results suggest that overcomplete sparse representations—a technique developed for transformer interpretability—transfer productively to neural population geometry. The structured latent split and sparsity constraint together provide the strongest causal subspaces, while identifiability priors designed for clean computational representations do not help on biological data.

5.7 Limitations

Silencing sample size. $n = 12$ direct matches ($n = 16$ with hierarchical grouping) for the optogenetic validation. **[PENDING: Exp66 (per-mouse analysis) will test whether the LDA anti-correlation holds per individual mouse.]**

IIA vacuity. The structured VAE's IIA is potentially vacuous (§4.2). We rely on external validation, but **[PENDING: exp62 (shuffled control) and exp63 (linear ablation) will provide the definitive internal controls.]**

No temporal resolution. All analyses use trial-averaged activity (150–350ms window). **[PENDING: Exp65 (temporal IIA) will test whether the choice subspace emerges before the behavioral response.]**

Task specificity. All results are for binary visual choice. Other task types (navigation, working memory) may show different patterns.

6 Conclusion

Linear subspace methods systematically underestimate causal structure in neural populations by 3.4× and are *anti-correlated* with optogenetic ground-truth causal importance ($\rho = -0.73$). A structured variational autoencoder recovers these signals, producing subspaces that generalize across four task variables, disentangle engagement from choice, and better predict optogenetic silencing effects.

However, the IIA metric used to evaluate these subspaces is vacuous for nonlinear methods—achieving comparable scores on random noise as on real data. This confirms the theoretical warning of **?** and means that external validation (optogenetic silencing, cross-task generalization, engagement orthogonality) is the gold standard for subspace quality assessment, not metric-internal evaluation.

An inductive bias ablation reveals that SAE-style overcomplete sparse representations are the dominant factor for causal subspace quality (IIA = 0.962 vs. 0.939 for the structured VAE baseline), while the label-conditional prior of pi-VAE—designed for identifiable representation learning—hurts on neural data. This suggests that techniques from transformer interpretability transfer productively to neural population geometry, but that the inductive biases must be adapted to the statistical structure of biological data.

Forward directions include cross-region activation patching (the biological analogue of path patching in transformer circuits) and mechanistic interpretability of the VAE encoder itself—applying causal mediation analysis to understand how the analysis tool processes neural activity.

A Sequential Discovery and Multiple Comparisons

Stage	Test	ρ	p	N_k	Survives
Stage 3: Causal subspace discovery					
1	VAE > LDA IIA (73/73 regions)	—	5.7×10^{-14}	1	✓
2	Silencing vs. LDA IIA	-0.73	0.01	2	✓
3	$\Delta\rho$ (VAE − LDA) vs. silencing	$+1.06$	< 0.005	3	✓
Stage 4: Generalization					
4	Multi-task VAE wins (4 variables)	—	$< 10^{-13}$	4	✓
5	Engagement-choice orthogonality	$d_G = 1.82$	—	5	✓
Stage 5: IIA vacuity and controls					
6	VAE random IIA ≈ MLP random IIA	0.70	$p = 0.38$ (no diff)	6	(diagnostic)

Table 7: Sequential discovery table. All five core claims (stages 3–4) survive Holm correction. Stage 5 is a diagnostic test, not a directional hypothesis. Stages 1–2 (metric comparison, dimensionality mediation) are reported in a companion paper [Paper A].