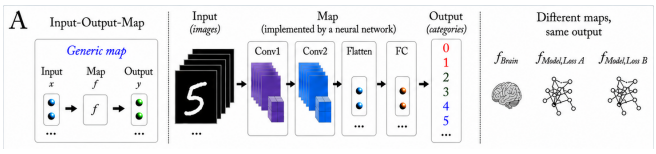
Decoding Alignment without Encoding Alignment: A Missing Component for Interpretability

Johannes Bertram, Luciano Dyballa, T. Anderson Keller, Savik Kinger, Steven W. Zucker

How should networks be compared? We show they can look identical under RSA & CKA while their internal circuits and functionality differ. Encoding manifolds organize units; Gromov-Wasserstein measures distance.

Motivation

Input-Output-Maps are **underspecified:** different maps can lead to same behavior over fixed-dataset classification tasks.



Decoding Manifold: Stimuli are represented in **neural activity coordinates.** Specifies **behavior** of a neural population.

Decoding Metrics: Metrics like **RSA, CKA,** etc. quantify similarity of stimulus representation based on the **decoding manifold.**

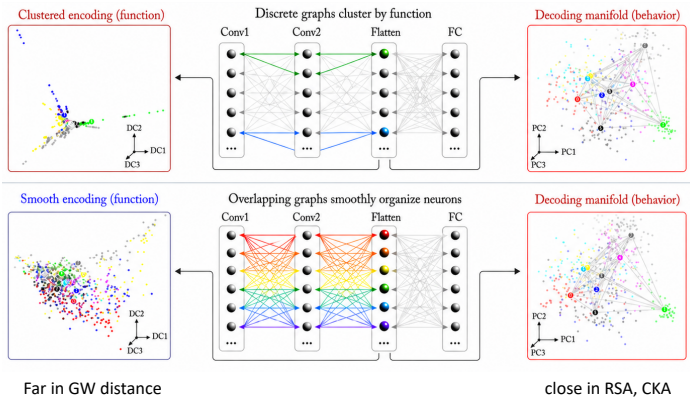
Encoding Manifold: Neurons in **stimulus-response coordinates.** Neurons with similar response properties are close. Specifies the **functional organization** of a neural population.

Gromov-Wasserstein (GW): Measures how close two **encoding manifolds** are by finding the optimal transport plan between neurons.

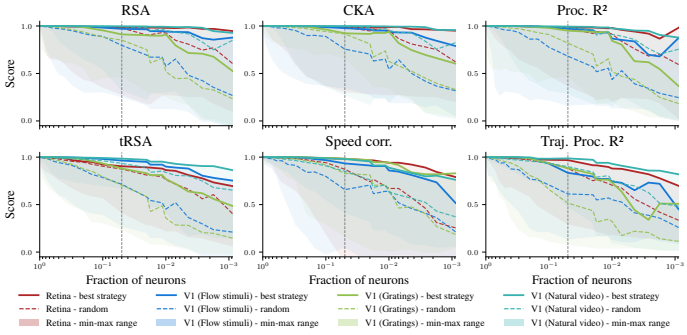
$$\mathrm{GW}(\mathcal{X},\mathcal{Y}) = \sqrt{\inf_{\pi\in\Pi(\mu,\nu)} \sum_{i,i',j,j'} |\mathbf{C}^{ii'}_{\mathcal{X}} - \mathbf{C}^{jj'}_{\mathcal{Y}}|^2 \pi_{ij}\pi_{i'j'}}$$

$\mathbf{C}_{\mathcal{X}}, \mathbf{C}_{\mathcal{Y}}$: Pairwise Euclidean distance matrices
$\Pi(\mu,\nu)$: Joint probability measures with marginals μ,ν
π_{ij} : Mass transported from I to j

Example: Two networks (rows) trained on **MNIST can show similar classification accuracy and I/O maps** (stimulus organization in the decoding manifold, measured by RSA, CKA, etc.) but with **different functional implementations** (clustered (top, test) vs smooth (bottom, baseline) encoding manifold).



Observation: Decoding Metrics Saturate



Decoding metrics can be recovered by small subpopulations (bold bars). However, all neurons are not the same: shaded regions indicate that not all subpopulations achieve full recovery of decoding.

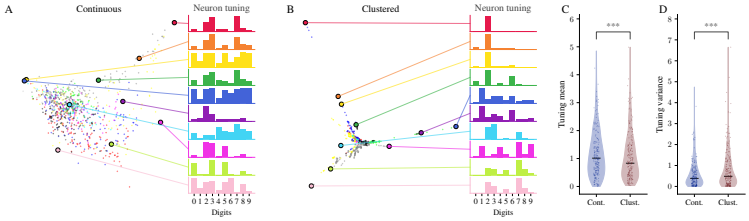
To determine whether neurons are relevant for a particular computation, we investigate neural population behavior from a functional perspective.

Results

MNIST Experiment: Unit Tuning and Topology

Network trained to have a smooth encoding manifold has multi-tuned units.

Network trained to have clustered encoding has sharply-tuned units.



A: Baseline model encoding manifold shows **smooth** organization with neurons being activated by different stimuli.

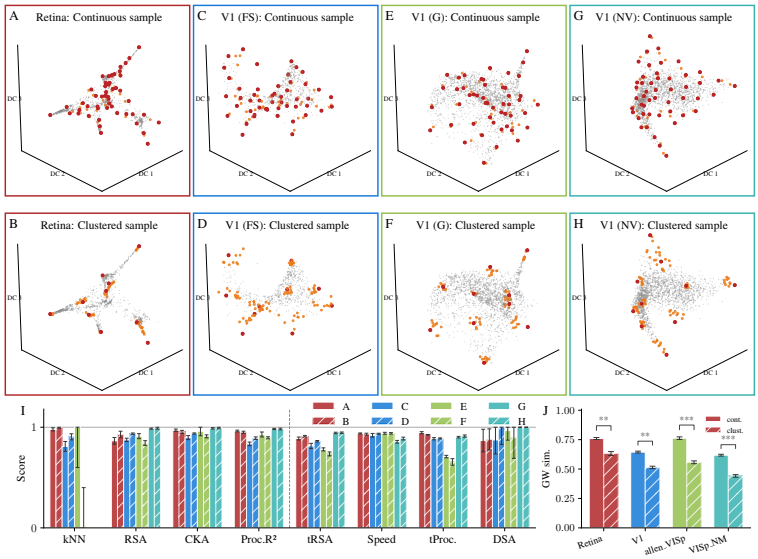
B: Test model trained to yield more selective neurons produces an encoding manifold **clustered** by preferred digits and sharp tuning properties.

C, D: Tuning quantification: test model shows lower mean tuning with higher variance, indicating strongly selective neurons.

Table: Decoding metrics are blind to the model differences, but **GW** captures the distinct topologies. Functional differences in terms of connectivity align with this, showing more sparse functional graphs in the test model.

Metric	Baseline (λ=0)	Clustered (λ=10)	Sig.
RSA	0.958 ± 0.001	0.961 ± 0.002	ns
CKA	0.976 ± 0.001	0.976 ± 0.001	ns
Proc. R^2	0.950 ± 0.002	0.941 ± 0.008	ns
GW Sim (ours)	**0.958** ± 0.001	**0.502** ± 0.022	***
Functional Differences			
Blur robustness, mean acc. ($\sigma_{\text{blur}} \geq 4$)	0.318 ± 0.010	0.282 ± 0.013	*
FC Attribution Entropy	5.919 ± 0.008	5.756 ± 0.007	***
FC Attribution Overlap	0.463 ± 0.002	0.407 ± 0.005	***
FC Assortativity	0.077 ± 0.001	0.086 ± 0.002	*
Conv2 Attribution Entropy	3.232 ± 0.008	3.097 ± 0.019	***
Conv2 Attribution Overlap	0.902 ± 0.003	0.889 ± 0.006	ns
Conv2 Assortativity	0.277 ± 0.035	0.501 ± 0.003	***

Biology: Decoding is Blind to Functional Topology



Smooth or clustered subsamples of the encoding manifold produce **indistinguishable decoding** results across biological datasets.

These differences are, however, captured by our functional **GW** metric.

Conclusion

Decoding (behavior) and **Encoding** (function) are **complementary** properties in neural computation.

Decoding metrics are **blind** to the **functional** perspective.

Thus, to claim similarity between two systems, one has to **additionally** attend to the **functional perspective**.

References

Luciano Dyballa, Andra M. Rudzite, Mahmood S. Hoseini, Mishek Thapa, Michael P. Stryker, Greg D. Field, and Steven W. Zucker. Population encoding of stimulus features along the visual hierarchy. Proceedings of the National Academy of Sciences, 121(4), January 2024a. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2317773121.

Johannes Bertram, Luciano Dyballa, Anderson Keller, Savik Kinger, and Steven W. Zucker. How 'Neural' is a Neural Foundation Model?, 2026

Facundo Mémoli. Gromov–Wasserstein Distances and the Metric Approach to Object Matching. 11(4):417–487, 2011. ISSN 1615-3383. doi: 10.1007/s10208-011-9093-5

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis connecting the branches of systems neuroscience. 2, 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited, 2019

This article has received funding from the European Commission's Marie Sk\l{}odowska-Curie Action under grant agreement no. 101207931 (LD), and by the MICIU /AEI /10.13039/501100011033 / FEDER, UE Grant no. PID2024-155187OB-I00 (LD). SWZ was supported by NIH grant 1R01EY031059 and NSF Grant 1822598.

Paper

Code