Network structure stabilizes neural manifolds under representational drift

Representational drift, the ongoing change in neuronal activity even under stable behavioral conditions, has been observed across many cortical and subcortical brain regions (e.g. Rokni et al., 2007; Ziv et al., 2013; Aschauer et al., 2022). Paradoxically, representational similarity (i.e. the similarity between population activity patterns evoked by different stimuli) remains stable over time despite this drift (e.g. Gallego et al., 2020; Noda et al., 2025). How can this stable representational similarity co-exist with ongoing activity changes on the level of neuronal populations?

We address this using binary feedforward networks with random connectivity, combining simulations and analytical results. We show that random projections generically preserve the similarity structure of inputs in output space: output similarity is a monotonically increasing function of input similarity, independent of the specific network connectivity. This holds for any sufficiently random or high-dimensional network with a large enough output space to faithfully represent the input similarity structure. Modeling drift as random synaptic rewiring, this input-similarity preservation directly implies that representational similarities remain stable, even though population activity vectors change substantially. Furthermore, this also holds for Hebbian rewiring and this framework can be extended to recurrent network architectures.

If connectivity shaped by learning from high-dimensional inputs behaves statistically like random projections, similar similarity-preserving properties should also emerge in trained deep neural networks. Importantly, these networks are not random but acquire structured connectivity through learning from high-dimensional inputs. Here, we model drift by continuing training beyond performance saturation. Within each layer, this leads to changing activity while representational similarity is largely preserved. Interestingly, drift magnitude increases with layer depth, mirroring experimental observations of greater drift in higher order regions, e.g. hippocampus, compared to early sensory areas, positioning DNNs as a powerful model system for studying representational drift.

Together, our results suggest that similarity-preserving properties characteristic of random projections provide a useful effective description for the experimentally observed coexistence of activity drift and representational stability.

References:

Aschauer et al., Learning-induced biases in the ongoing dynamics of sensory representations predict stimulus generalization, 2022, Cell Rep.

Gallego et al., Long-term stability of cortical population dynamics underlying consistent behavior, 2020, Nat. Neurosci.

Noda et al., Homeostasis of a representational map in the neocortex, 2025, Nat. Neurosci.

Rokni et al., Motor learning with unstable neural representations, 2007, Neuron

Ziv et al., Long-term dynamics of CA1 hippocampal place codes, 2013, Nat. Neurosci.