# Fine-tuning of attractors on a ring underlies the learning of robust working memory in mice | Mahrach Alexandre

A. Mahrach1, X. Zhang2,3,4, D. Li2,3,4, C.T. Li2,3,4 A. Compte1
1 IDIBAPS, Barcelona, Spain
2 Institute of Neuroscience, State Key Laboratory of Neuroscience, CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China
3 Lingang Laboratory, Shanghai, China
4 Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, China

Working memory (WM) is a process that temporarily holds information when not accessible to the senses. It relies on storing stimulus features in neuronal activity. Despite its importance in cognition, WM is quite vulnerable due to its limited capacity and stability, especially when facing interferences from internal and external sources. Recent studies have explored the neural mechanisms that protect WM and proposed that pre-/post-distraction neural activity decomposes into orthogonal subspaces, thus protecting information. However, whether orthogonalization is acquired through learning is unknown, and the mechanisms supporting it are unclear. Here, we study the learning of orthogonalization using calcium imaging data from the mouse prelimbic (PrL) and anterior cingulate (ACC) cortices as they learn an olfactory dual task. The task combines an outer Delayed Paired-Association task (DPA) with an inner Go-NoGo task. We examined how neuronal activity reflected the process of shielding the DPA sample information against Go-NoGo. As mice learned the task, we measured the overlap between the activity and the low-dimensional subspaces that encode the sample or Go-NoGo odors. Early in the training, pre-Go-NoGo activity overlapped with both sample/Go-NoGo subspaces. Later in the training, pre-Go-NoGo activity was confined to the sample subspace, resulting in a more robust sample code. We present a mechanistic insight into how these low-dimensional representations evolve with learning in a recurrent network model of excitatory and inhibitory neurons with trained low-rank connections. The model links learning to (1) the orthogonalization of sample and Go-NoGo subspaces and (2) modifications of a

double-well attractor on a one-dimensional ring. We validated (1) by measuring the angular distance between the sample and Go-NoGo subspaces in the data and (2) by estimating an energy landscape for the recorded neural dynamics. In sum, our study underscores the crucial role attractor dynamics plays in shielding WM representations from concurrent tasks.