Ring Attractor dynamics underlies the learning of stable working memory representations in a dual task

A. Mahrach[1], X. Zhang[2,3], D. Li[2,3], C.T. Li[2,3], A. Compte[1]

[1] IDIBAPS, Barcelona, Spain
[2] Institute of Neuroscience, State Key Laboratory of Neuroscience, Chinese Academy of Sciences, CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, China
[3] School of Future Technology, University of Chinese Academy of Sciences, Beijing, China

Working memory (WM) is a cognitive function that allows the short-term maintenance and manipulation of information when no longer accessible to the senses. It relies on temporarily storing stimulus features in the activity of neuronal populations. Recent studies have addressed how WM is protected from distraction. It is proposed that pre and post-distraction population activity decomposes into orthogonal subspaces to preserve WM. However, whether WM orthogonalization is innate or acquired through learning is unknown, and the network mechanisms supporting it are unclear. Here, we probe WM orthogonalization using calcium imaging data from the mouse prelimbic (PrL) and anterior cingulate (ACC) cortices as they learn to perform an olfactory dual task. The dual task combines an outer Delayed Paired-Association task (DPA) with an inner Go-NoGo task. We examined how neuronal activity reflected the process of protecting the DPA sample information against Go/NoGo distractors. As mice learned the task, we measured the overlap between the neural activity onto the low-dimensional subspaces that encode sample or distractor odors. Early in the training, pre-distraction activity overlapped with both sample and distractor subspaces. Later in the training, pre-distraction activity was strictly confined to the sample subspace, resulting in a more robust sample code. We present a mechanistic insight into how these low-dimensional WM representations evolve with learning in a recurrent neural network model of excitatory and inhibitory neurons with low-rank connections. The model links learning to (1) the orthogonalization of sample and distractor WM subspaces and (2) the orthogonalization of each subspace with irrelevant inputs. We validated (1) by measuring the angular distance between the sample and distractor subspaces through learning in the data. Prediction (2) was validated in PrL through the photoinhibition of ACC to PrL inputs, which turned back late training into early training dynamics. Moreover, our model imposes a double well attractor network dynamics on a one-dimensional ring and suggests that learning optimally adjusts the location of the attractors on this ring. We validated this theoretical prediction 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 distracting tasks.