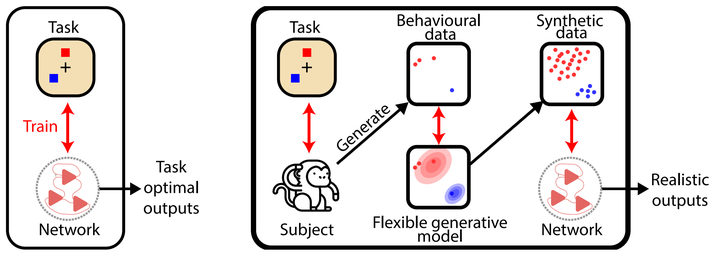
Setting up for failure: automatic discovery of the neural mechanisms of cognitive errors

Puria Radmard[1 2], Paul M. Bays[1], Máté Lengyel[1 2 3]
pr450@cam.ac.uk

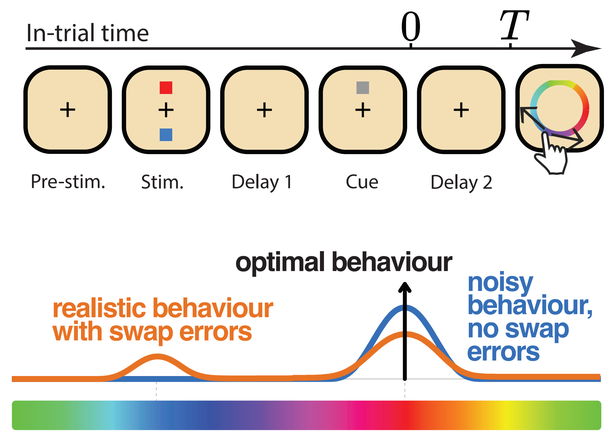
Data-driven automated mechanism discovery

- Typically, recurrent neural network (RNN) models of cortex are trained to achieve task-optimal behaviour, not realistic behaviour with suboptimalities and errors
- Acheiving realistic errors with these networks requires normatively inspired, but often heuristic and piecemeal ablations to the network design.
- Instead, we train RNNs to reproduce realistic errors with a **Denoising Diffusion Probabilistic Model (DDPM)-style objective**, and show that the resulting neural dynamics are more realistic than RNNs trained with mean-squared error (MSE)
- To allow DDPM-RNN training, we generate synthetic behaviour with a statistical model which captures errors (Radmard et al., 2025)
- Our approach allows **automatic mechanism discovery without human-in-the-loop design**



Test bed: swap errors in visual working memory

- As a test bed for our method, we use a **multi-item delayed estimation visual working memory task**
- Behaviour can be suboptimal not only due to response noise, but also due to so-called **swap errors to uncued items**
- Swap errors make the **response distribution multimodal**
- Typical RNN training uses **mean-squared error, which cannot capture this multimodality**
- **We instead train the RNN like a conditional DDPM, with synthetic behavioural data as the target dataset**



Network and training implementation

- Two-item working memory task displays **swap errors** which results in a **multimodal response distribution**, which RNNs trained with MSE cannot capture
- Dendritic tree RNN (Lyo & Savin, 2024): deep nonlinearity per neuron, biologically plausible pyramidal connectivity

Network dynamics

$$\boldsymbol{r}_{t+1} = \left[\left(1 - \frac{\mathrm{d}t}{\lambda}\right)\boldsymbol{r}_t + \frac{\mathrm{d}t}{\lambda}\mathcal{F}(\boldsymbol{r}_t, \boldsymbol{s}_t; \theta_r)\right] + \sigma_t \boldsymbol{\epsilon}_t$$

- DDPM objective applied only to behavioural subspace $\boldsymbol{x}_t$ during post-cue delay; nullspace m carries memory of task variables

DDPM denoising mean

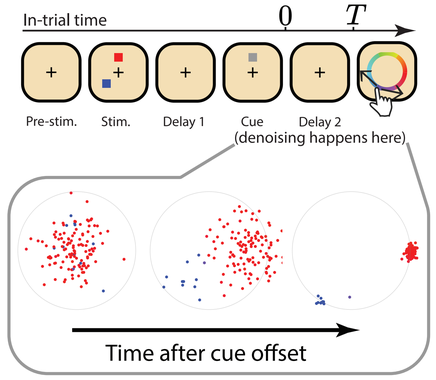
$$q(\boldsymbol{x}_{\tau-1}|\boldsymbol{x}_\tau) = \mathcal{N}(\boldsymbol{x}_\tau; \sqrt{1-\beta_\tau}\boldsymbol{x}_\tau, \beta_\tau I)$$

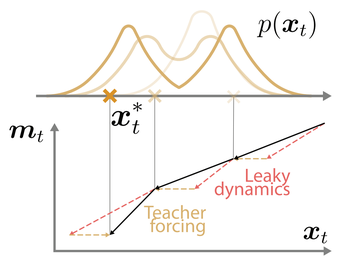
DDPM noising process

$$\hat{\boldsymbol{\mu}}_{\theta_r}(\boldsymbol{r}_t, t) = W_x\left[\left(1 - \frac{\mathrm{d}t}{\lambda}\right)\boldsymbol{r}_t + \frac{\mathrm{d}t}{\lambda}\mathcal{F}(\boldsymbol{r}_t, \boldsymbol{s}_t; \theta_r)\right]$$

DDPM objective (with regularisation)

$$\sum_{\text{i.i.d.trials } k}\left[\sum_{\substack{\text{denoising}\\ \text{timesteps } t}}\|\boldsymbol{\mu}_q(\boldsymbol{x}_t^{*,k}, \boldsymbol{x}_T^{*,k}) - \hat{\boldsymbol{\mu}}_{\theta_r}(\boldsymbol{r}_t^k, t)\|_2^2 \;+\; \gamma_2 \sum_{\substack{\text{all}\\ \text{timesteps } s}}\|\boldsymbol{r}_s^k\|_2^2\right] + \gamma_1\left(\|\bar{\boldsymbol{r}}_0\|_2^2 + \|\boldsymbol{\Sigma}_{\boldsymbol{r}_0}\|_{\mathrm{Fr}}^2\right)$$



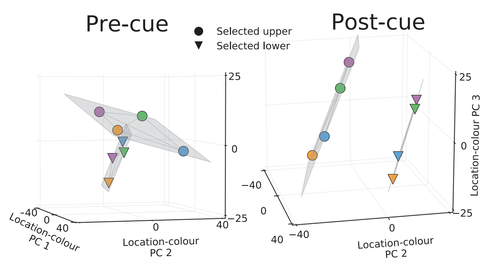


- **Teacher forcing** on $\boldsymbol{x}_t$ keeps trajectory in-distribution for valid DDPM training
- The first steps of denoising are also trained on moment matching to the unit normal

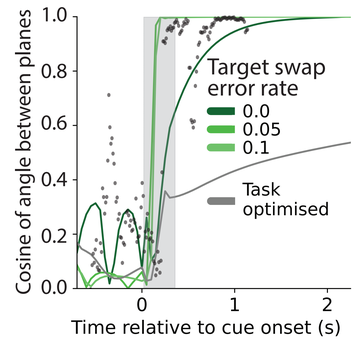
Only behaviour-matched networks recovered neural signatures

- Real macaque data with **fixed location items**: colour representations for **non-swap trials** start orthogonal, then align during cue
- Task-optimised networks (trained with MSE) and DDPM-RNNs with **unimodal behavioural data did not recover this**, but DDPM with swap errors did.
- With location reintroduced, this trend was **only recovered when swap errors depended on the stimuli distance**, as in other swap error datasets

Real data (Panichello and Buschman, 2021)



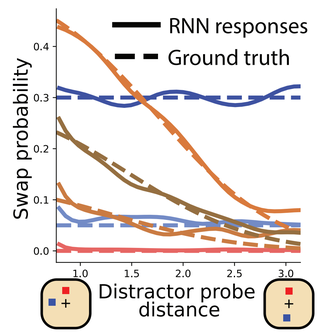
Fixed location



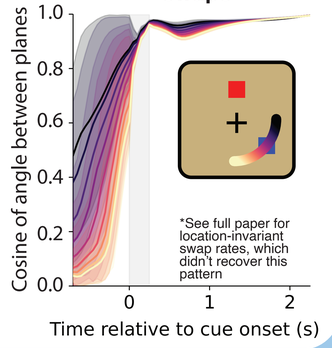
- The network further **predicts that closer, more swap-prone items are more aligned in pre-cue representations**

Behaviour-matched network predicts pre-cue geometry

- Extending to **variable location items**, we find success in using our method to fit to synthetic data with specific swap dependencies
- Cortex-like representations are only recovered when **swap errors diminish with distractor distance**, as in human data
- Plane representations of closer, more swap prone item pairs are more aligned before the cue, compared to further, less swapped pairs
- This is an **automatically discovered, experimentally falsifiable prediction**



Distance-dependent swaps*



This work was supported by the Wellcome Trust (Investigator Award 212262/Z/18/Z)