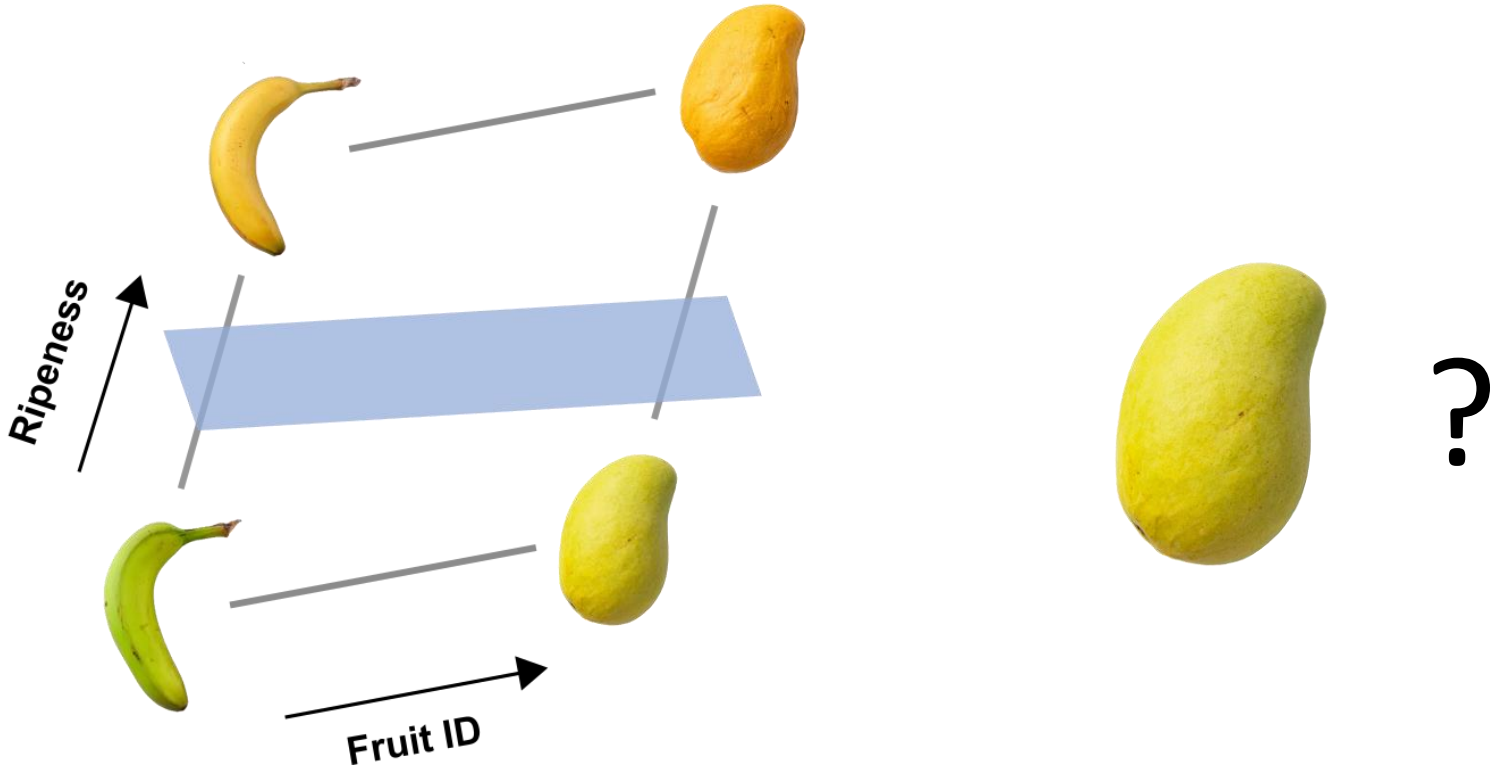
Disentangling Representations through Multi-task Learning

Pantelis Vafidis, Aman Bhargava & Antonio Rangel

California Institute of Technology

April 2025

Disentangled representations are interpretable, generalizable world models

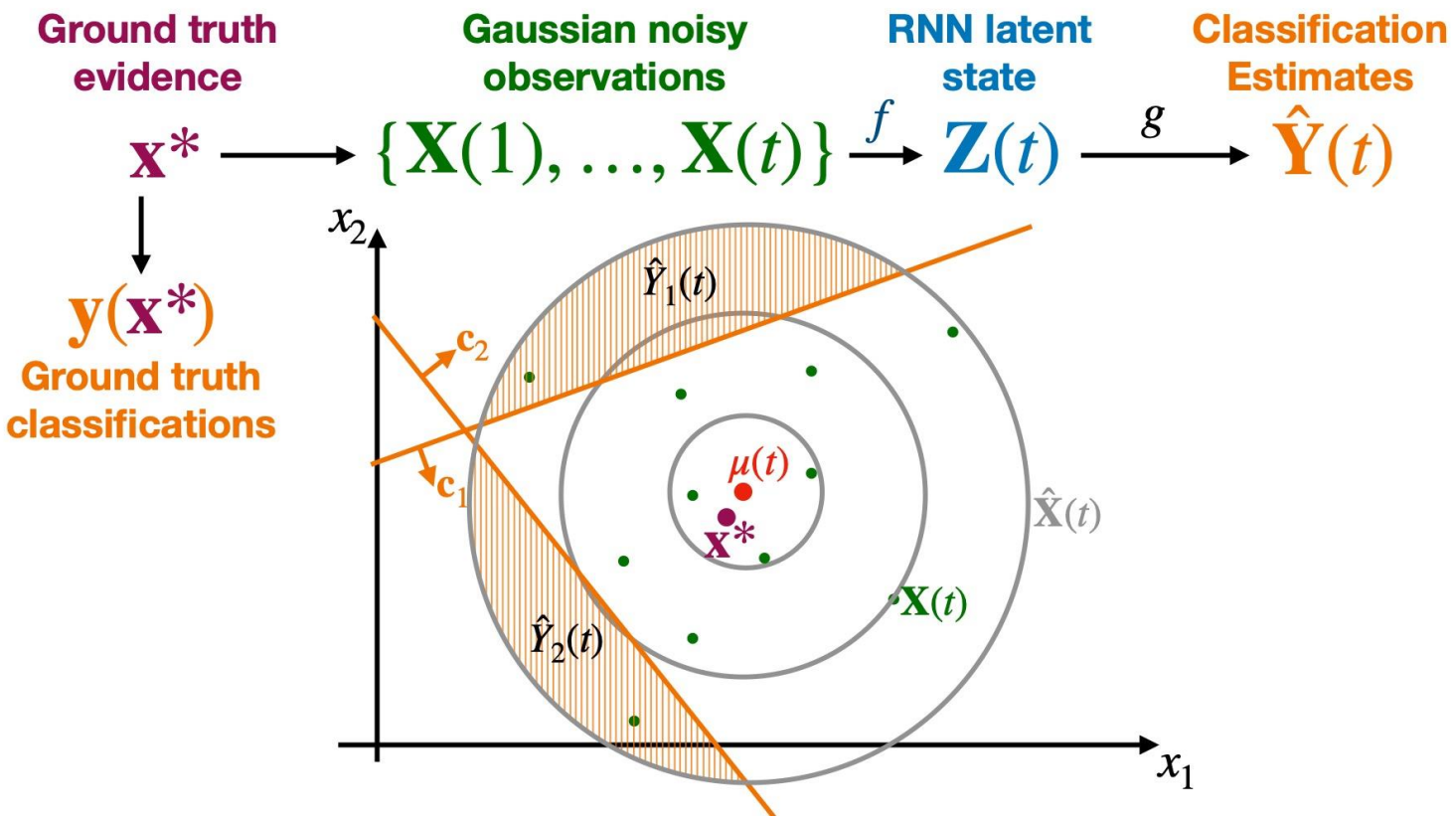


Definition: Disentangled reps: latent factors are orthogonal
Abstract reps: factors are linearly decodable and approx. orthogonal

Disentangled representations support generalization in various brain areas (Saez et al. 2015, Boyle et al. 2022, Bongioanni et al. 2021). Can we guarantee their emergence in neural networks?

How to learn disentangled representations? Multi-task learning

Idea: if we place pressure from many tasks at the same time, representations will not collapse but maintain a generalizable structure



Not a new idea! Caruana (1997). Recent empirical evidence by Johnston & Fusi (2023) and Maziarka et al. (2021) in FFNN

We: prove that MTL leads to disentanglement (and abstractness)

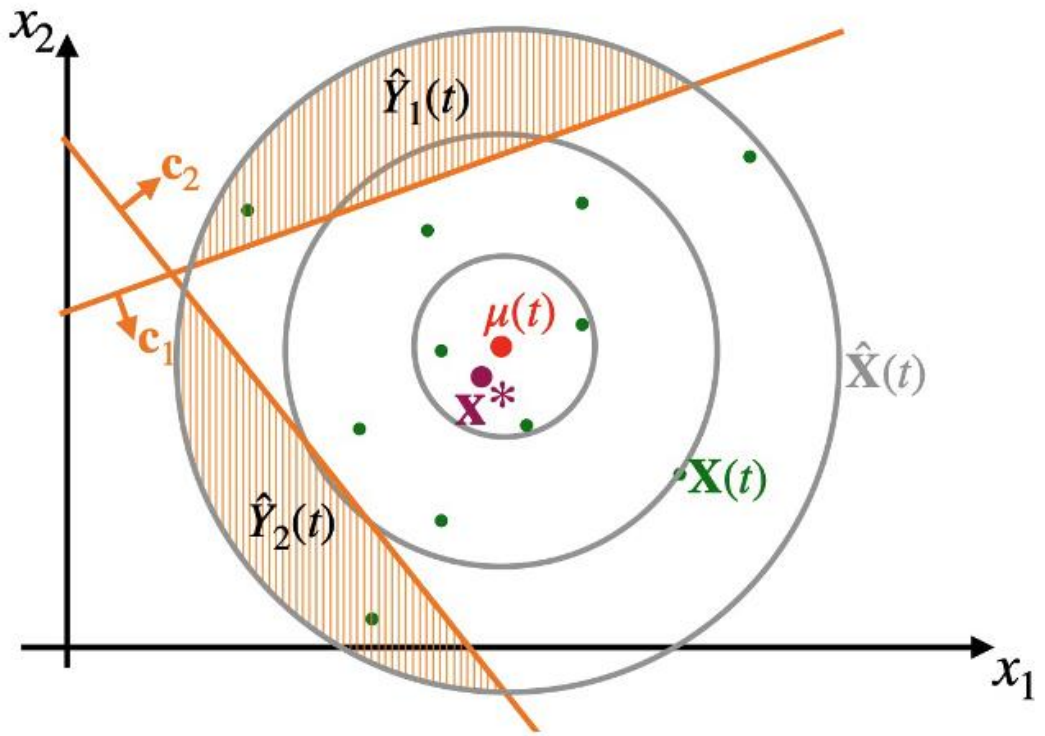
Key theoretical insight: classification estimates represent distances

- **Optimal $\hat{\mathbf{Y}}(t)$:** Compute $\hat{\mathbf{X}}(t) = P\left(\mathbf{x}^* \,|\, \mathbf{X}(1)\ldots\mathbf{X}(t)\right)$. Integrate probability mass on each side of boundaries $(\mathbf{c}_i, b_i)$, $i \in [N_{task}]$.

$$\hat{Y}(t) \triangleq \Pr\{\mathbf{c}^\top \mathbf{x}^* > b \,|\, \mathbf{X}(1)\ldots\mathbf{X}(t)\}$$

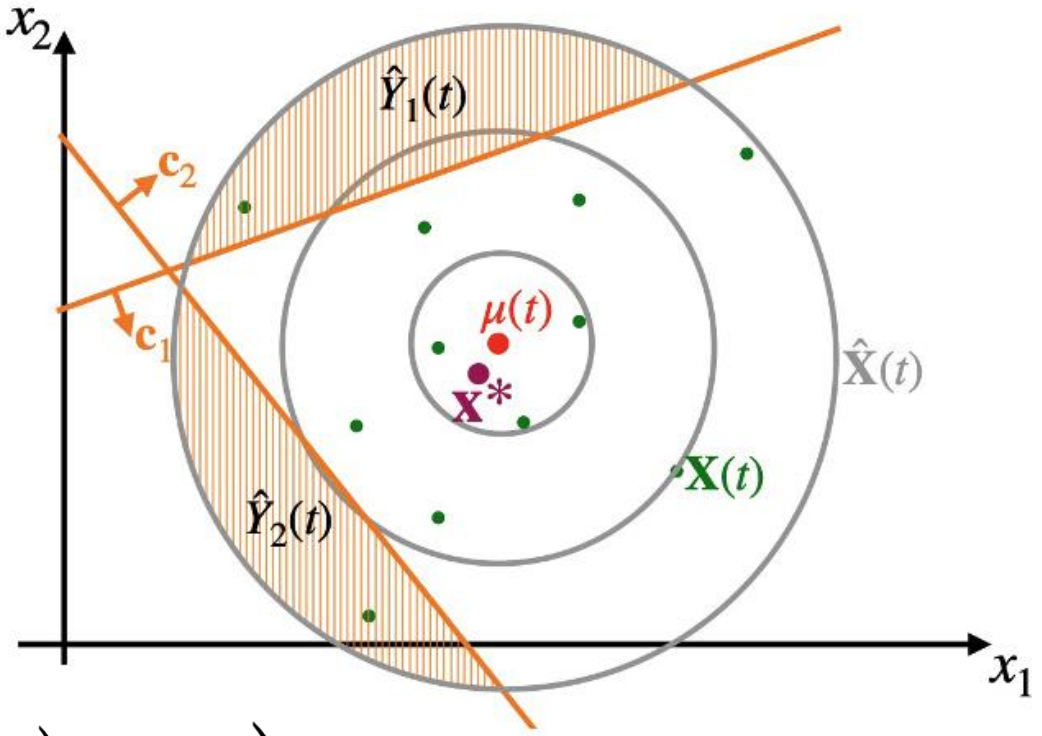
$$= \Phi(k\sqrt{t}/\sigma)$$

$$\Longrightarrow \quad k = \frac{\sigma}{\sqrt{t}}\Phi^{-1}(\hat{Y}(t))$$



Given enough distances, ground truth is uniquely identifiable

- $\hat{\mathbf{X}}(t) = P\left(\mathbf{x}^* \mid \mathbf{X}(1)\ldots\mathbf{X}(t)\right)$
  - $= \mathcal{N}\left(\mu(t), \Sigma(t)\right)$ where
  - $\mu(t) = \text{mean}\left(\mathbf{X}(1)\ldots\mathbf{X}(t)\right)$
- **Theorem**: If decision boundary matrix $\mathbf{C}$ is full-rank and $N_{task} \geq D$, then then $\left(\hat{\mathbf{Y}}(t), t, \mathbf{b}, \mathbf{C}, \sigma\right)$ are sufficient to reconstruct the exact value of $\mu(t)$, the optimal estimator for $\mathbf{x}^*$.

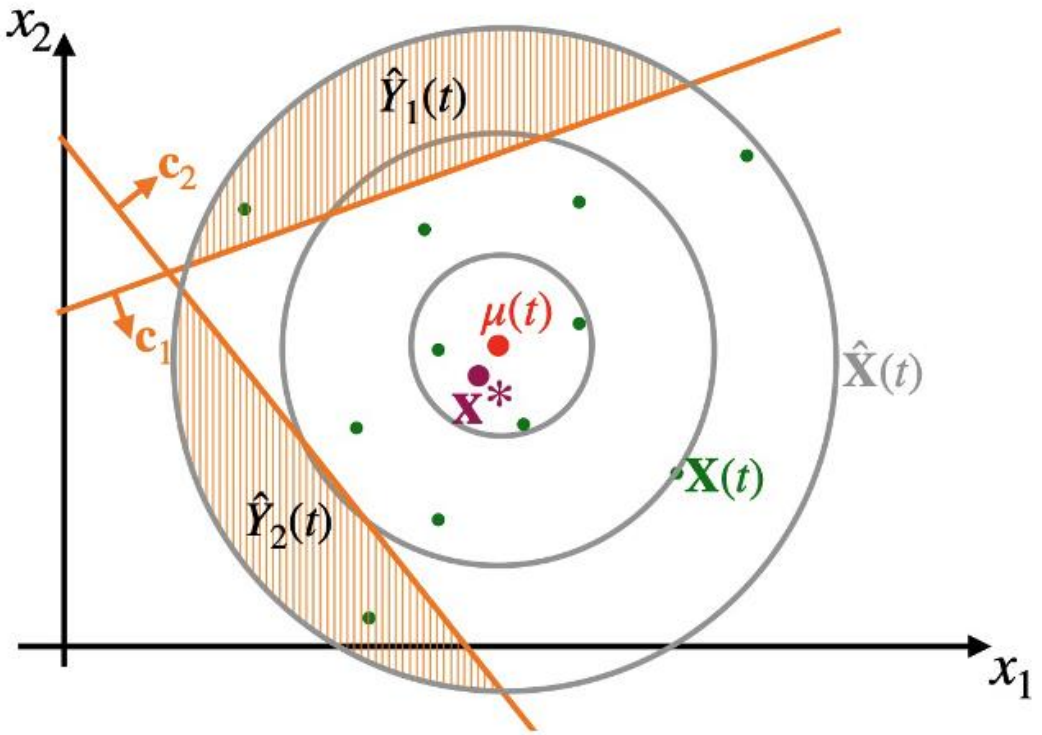


$$\mu(t) = (\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{C}^\top\left(\frac{\sigma}{\sqrt{t}}\Phi^{-1}\left(g(\mathbf{Z}(t))\right) + \mathbf{b}\right)$$

...and linearly decodable for sigmoid activation function *g*

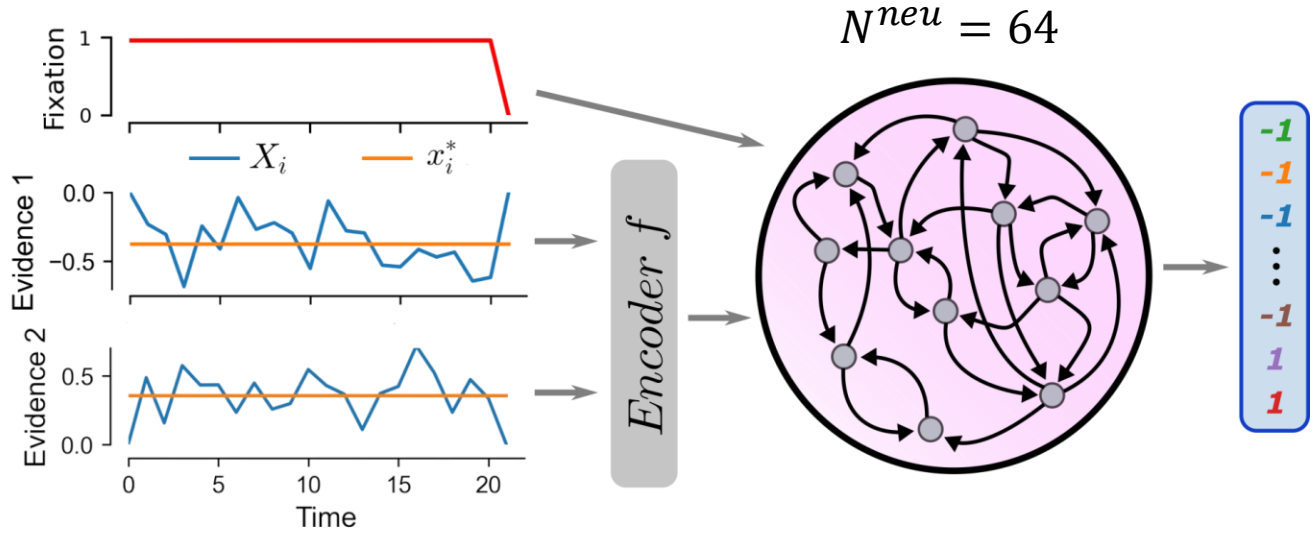
$$\mu(t) = (\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{C}^\top\left(\frac{\sigma}{\sqrt{t}}\Phi^{-1}\big(g(\mathbf{Z}(t))\big) + \mathbf{b}\right)$$

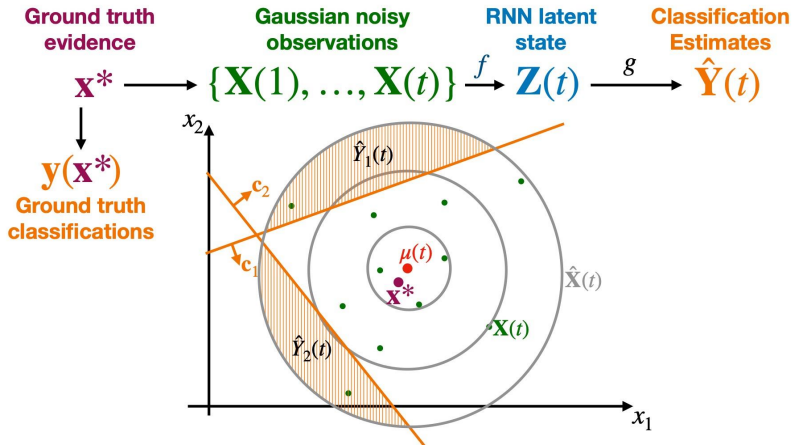
$$\mu(t) \approx \underbrace{\frac{2\sqrt{3}\sigma}{\pi\sqrt{t}}(\mathbf{C}^\top\mathbf{C})^{-1}\mathbf{C}^\top}_{\text{linear}}\underbrace{(\mathbf{Z}(t) + \mathbf{b})}_{\text{affine}}$$



Trilateration Theorem! Furthermore, rep is orthogonal for $N_{task} \gg D$

Experimental setup

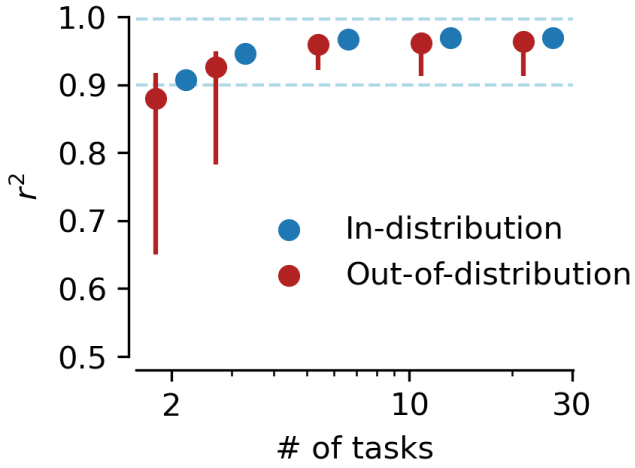


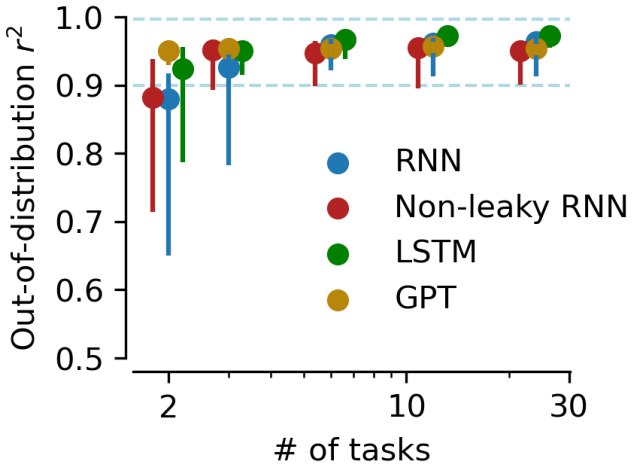


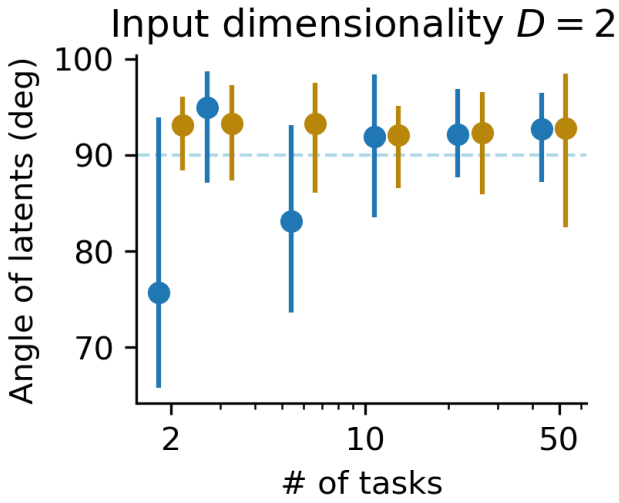
f is a randomly initialized, fixed MLP

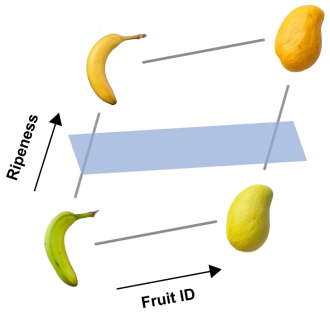
$x^* \in \mathbb{R}^D, D = 2$ for now but we are going to increase it later

Trained networks zero-shot generalize out-of-distribution









Regression generalization metric (Johnston & Fusi 2023)

Transformers orthogonalize at theoretical minimum N_{task}!

Experiments confirm theory

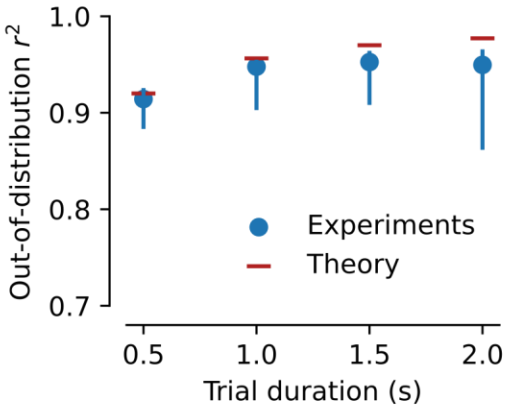
Theorem 3.1 (Disentangled Representation Theorem). *If* $\mathbf{C} \in \mathbb{R}^{N_{task} \times D}$ *is a full-rank matrix and* $N_{task} \geq D$ *and noise* $\sigma > 0$, *then*

1. *any optimal estimator of* $\mathbf{y}(\mathbf{x}^*)$ ***must encode a finite-sample, maximum likelihood estimate*** $\mu(t)$ *of the ground truth evidence variable* $\mathbf{x}^*$ *in its latent state* $\mathbf{Z}(t)$, *and*

2. *if the activation function is sigmoid-like,* $\mu(t)$ *will be* ***linearly decodable from*** $\mathbf{Z}(t)$, *thus implying that* $\mathbf{Z}(t)$ *contains an abstract representation of* $\mu(t)$ *(Ostojic & Fusi, 2024).*

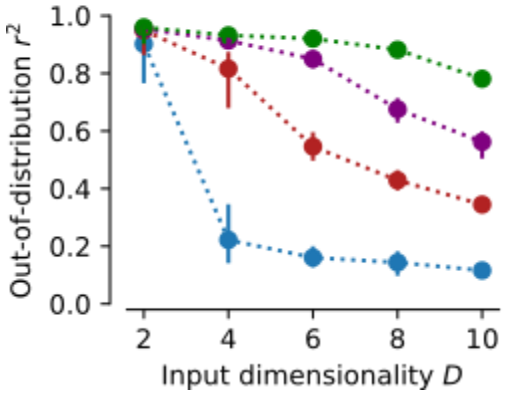
3. *Furthermore, the representation is guaranteed to be disentangled (orthogonal) as* $N_{task} \gg D$ *for random decision boundaries.*

Time

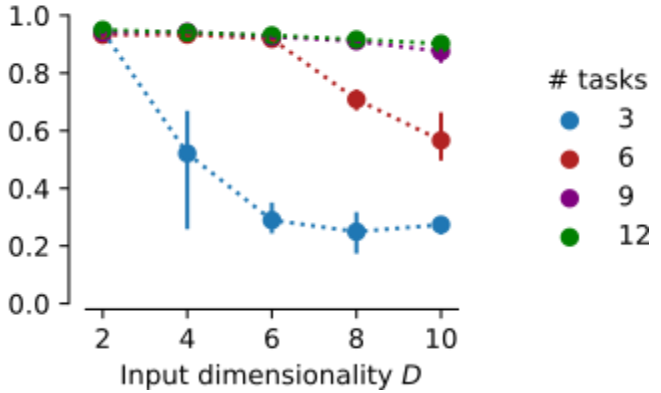


Dimensionality

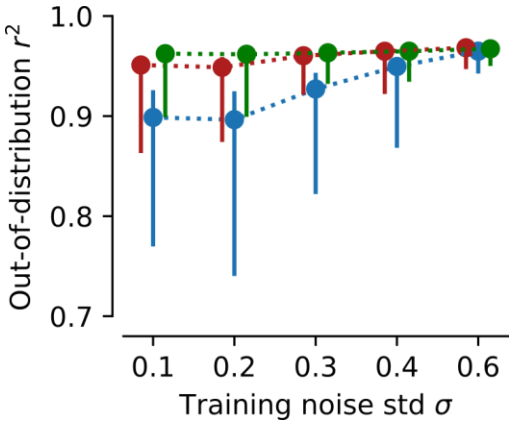
RNN



GPT



Noise



Discussion

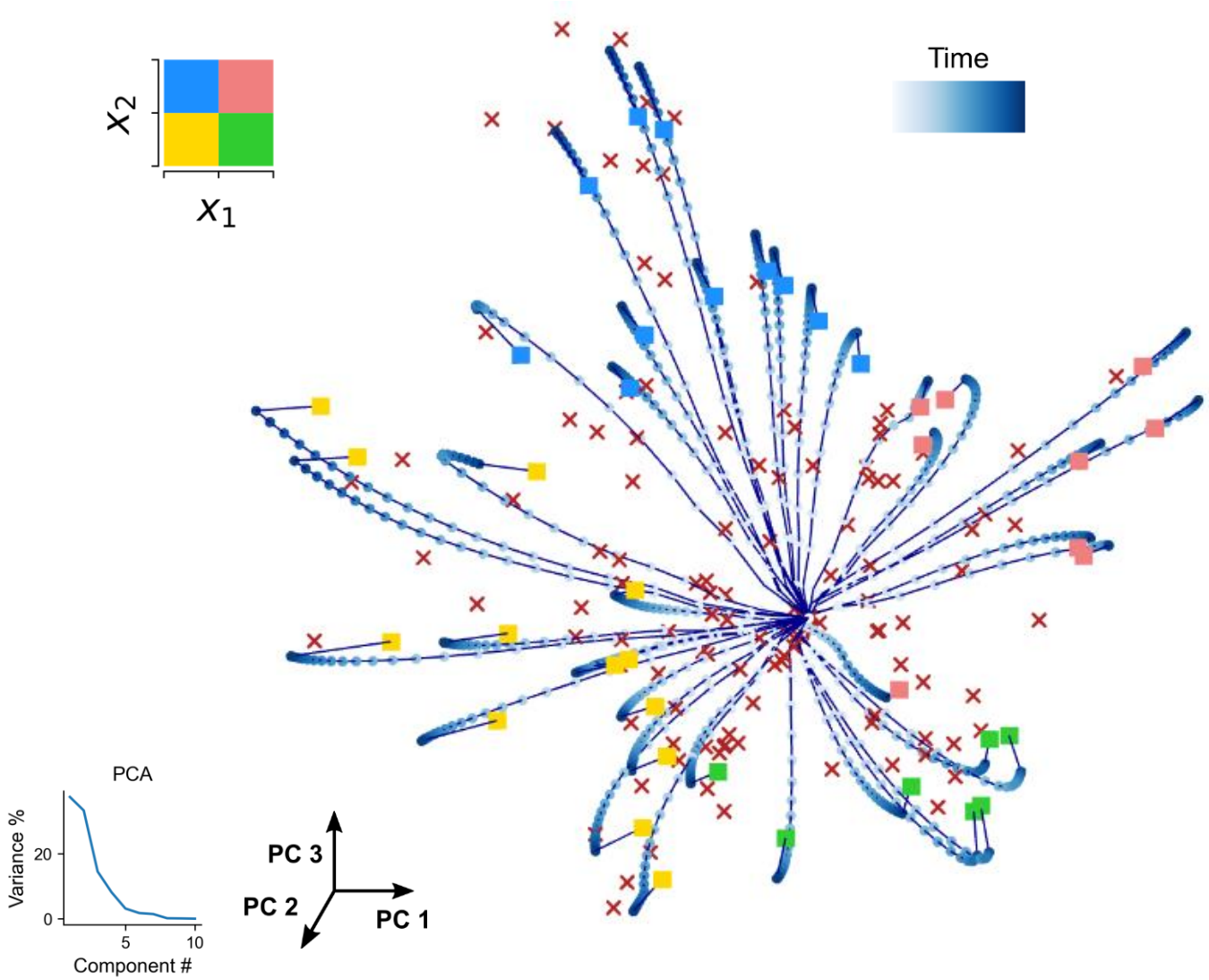
Our work:

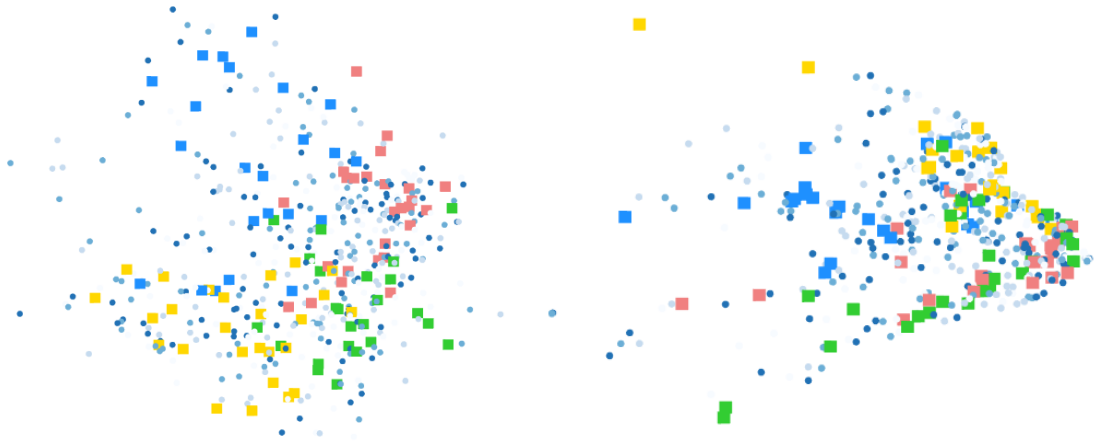
- Proves mathematically and demonstrates experimentally why and when disentangled representations emerge; *competence at multiple tasks*
- Offers an explanation for representation alignment across artificial and biological systems (Templeton et al 2024)
- Suggests that the cortex, with its massively parallel processing architecture (cortical columns) might be the locus for world model construction in brains!

For more, come find us on Saturday April 26 3:00-5:30 pm @ Poster Session 6!

Thank you!

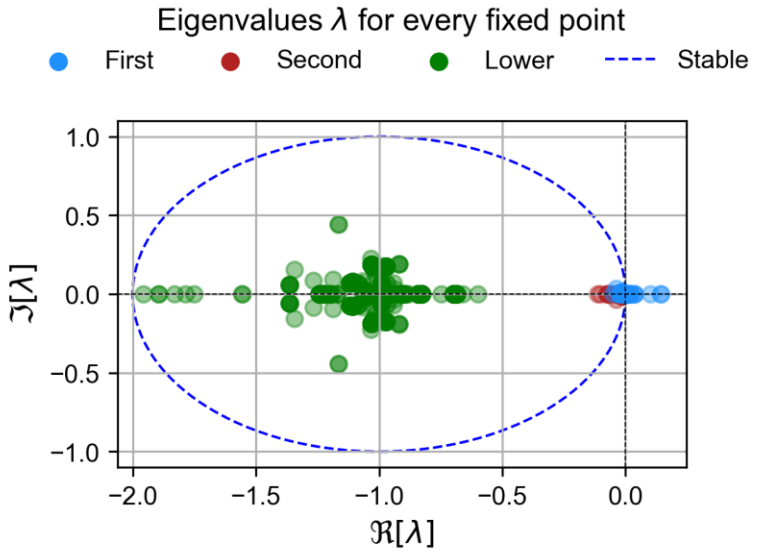
RNNs learn continuous attractors



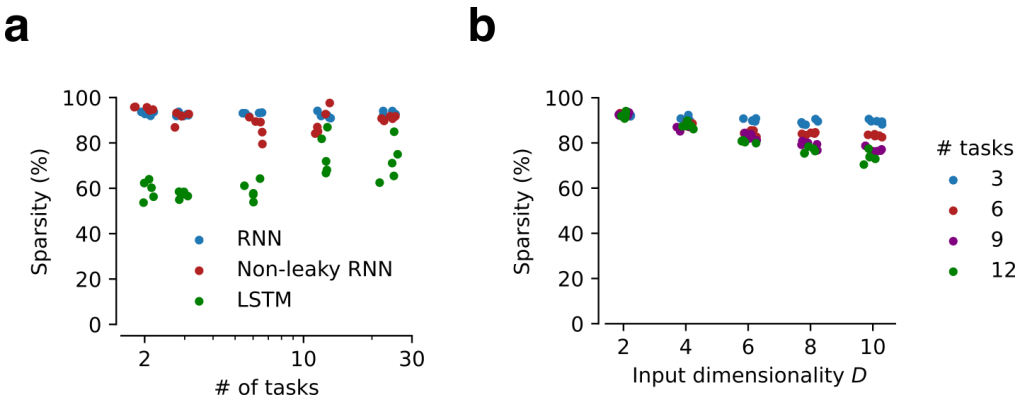


After noising ($X(t)$)

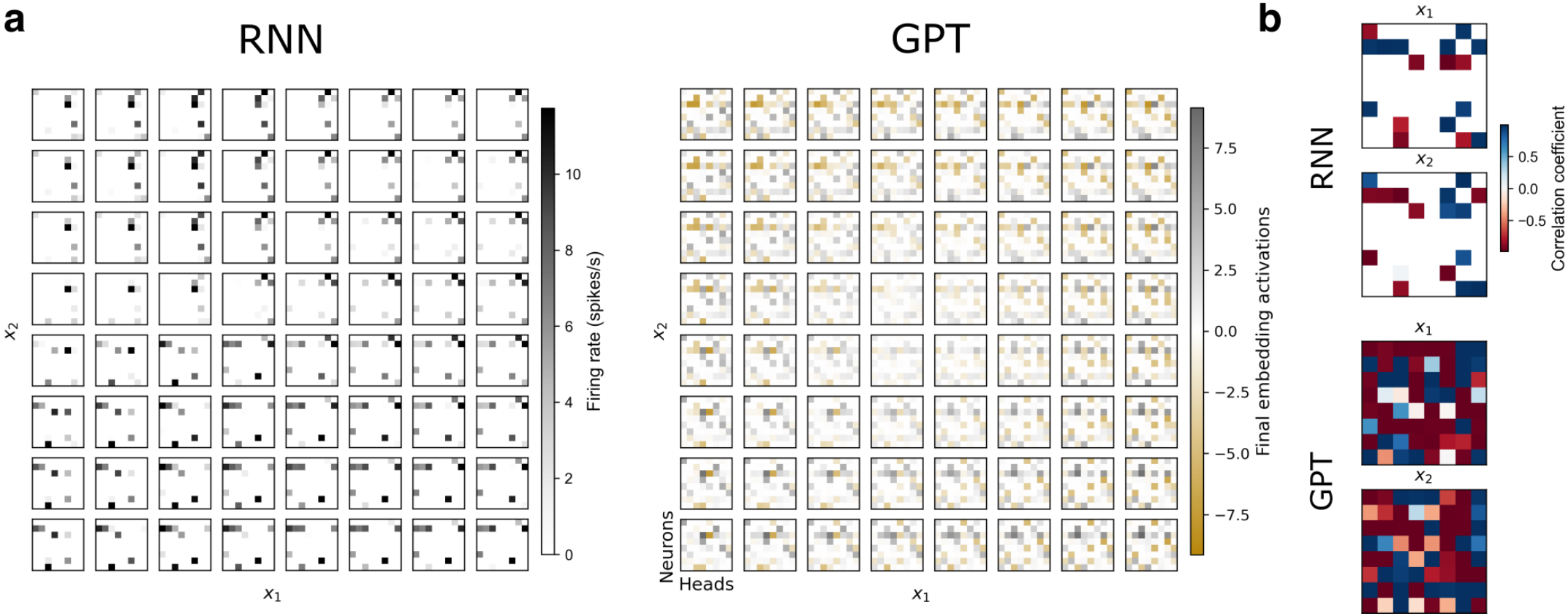
After mixing (f($X(t)$))



Representations learned



Vinje & Gallant (2000)



Disentanglement in high dimensions