Unsupervised Discovery of Individual Differences in Neural Network Models of Behavior

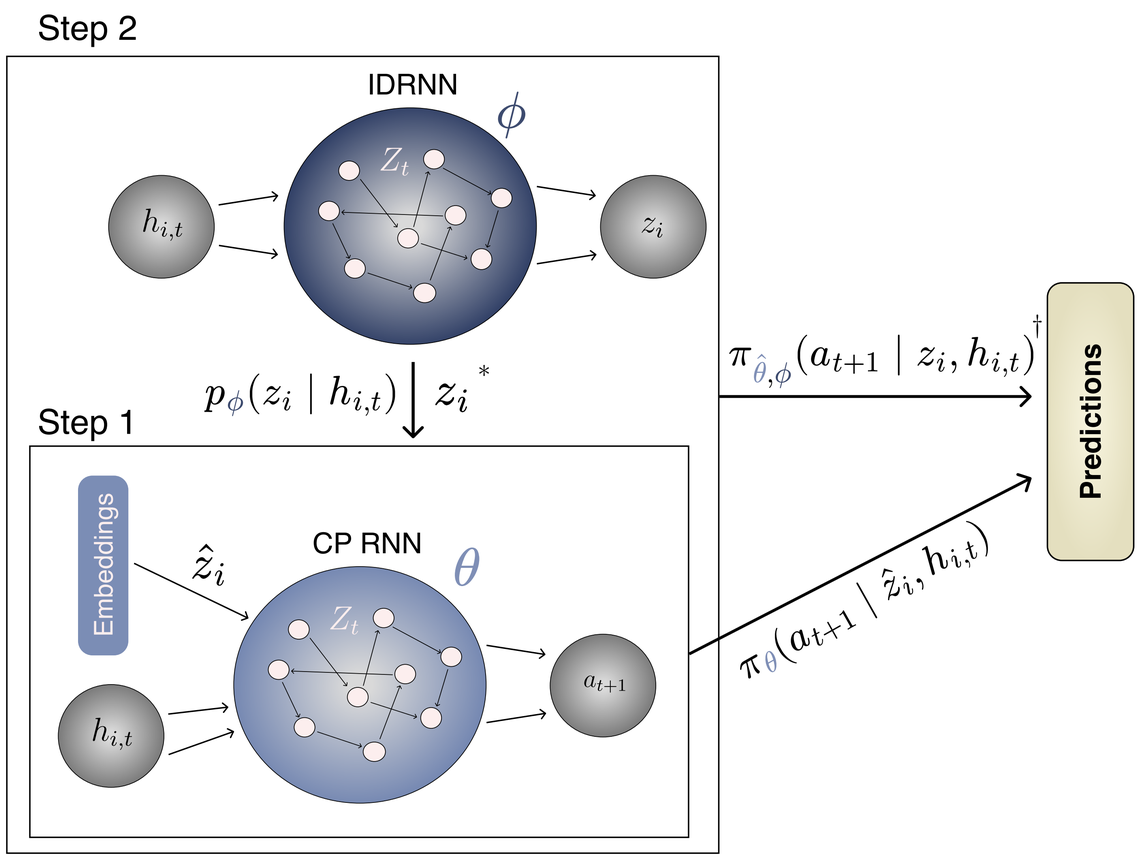
Marvin Mathony [1] Alireza Modirshanechi [*1,2] Eric Schulz [*1]

[1]Helmholtz Munich [2]Max Planck Institute For Biological Cybernetics [*]Equal contribution

Background and aims

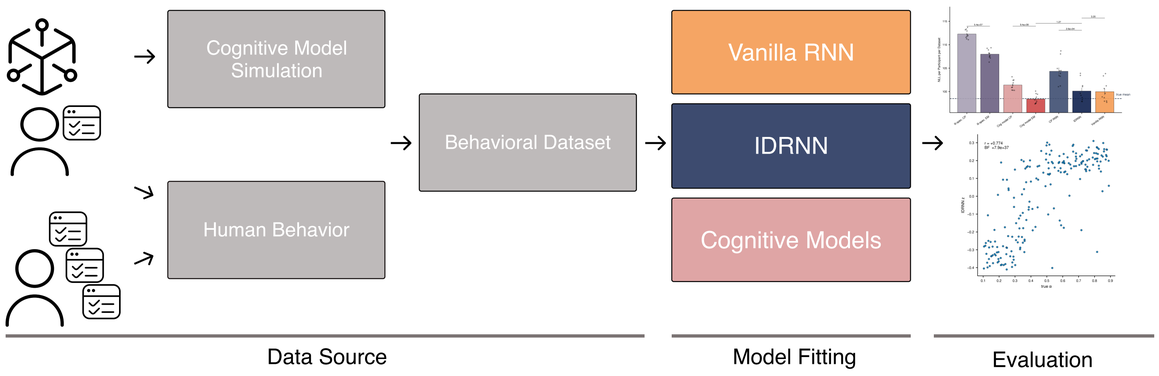
- Neural networks are the most predictive models of human behavior, yet it remains unclear what their internal representations encode.
- Because individual differences (ID) are learned implicitly, participant-specific information becomes entangled with task-related variation.
- **We develop an architecture that explicitly models participant representations and makes them accessible for analysis.**
- This allows us to ask a fundamental question: when neural networks explain behavioral variation, do their latent representations capture meaningful properties of individuals?
- We test this question in synthetic environments with known ground truth and in human reinforcement learning datasets, containing personality and clinical measures.

Architecture



* sample N times at test time
† average over N at test time

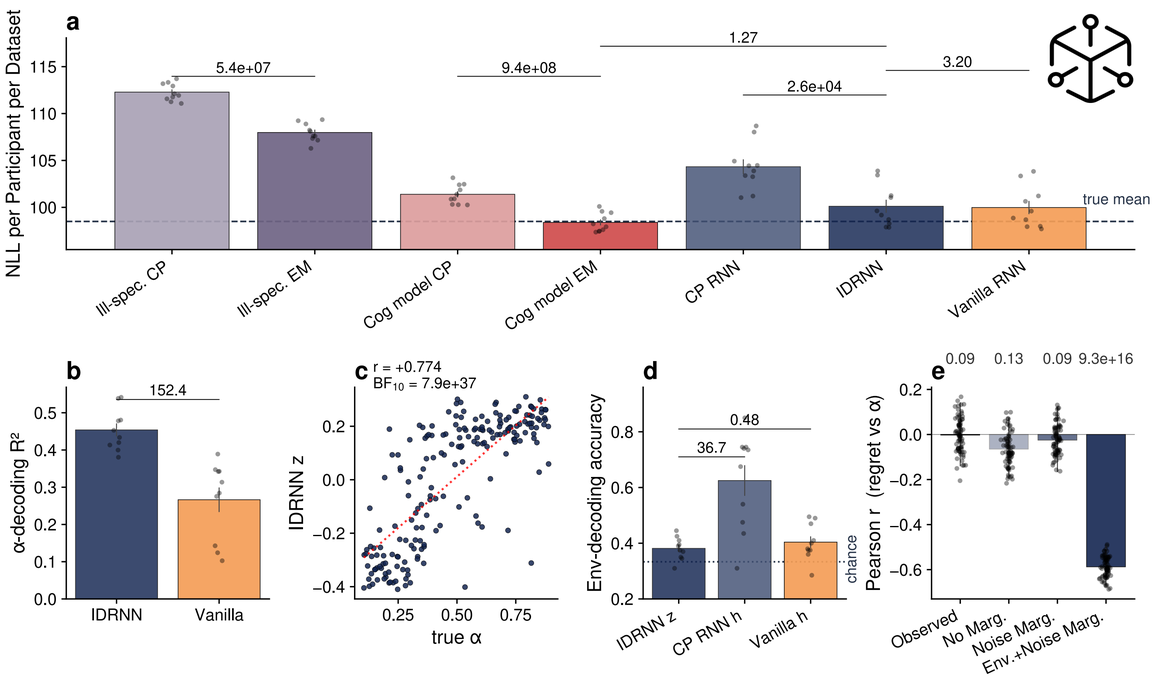
Modeling Pipeline



Synthetic Data

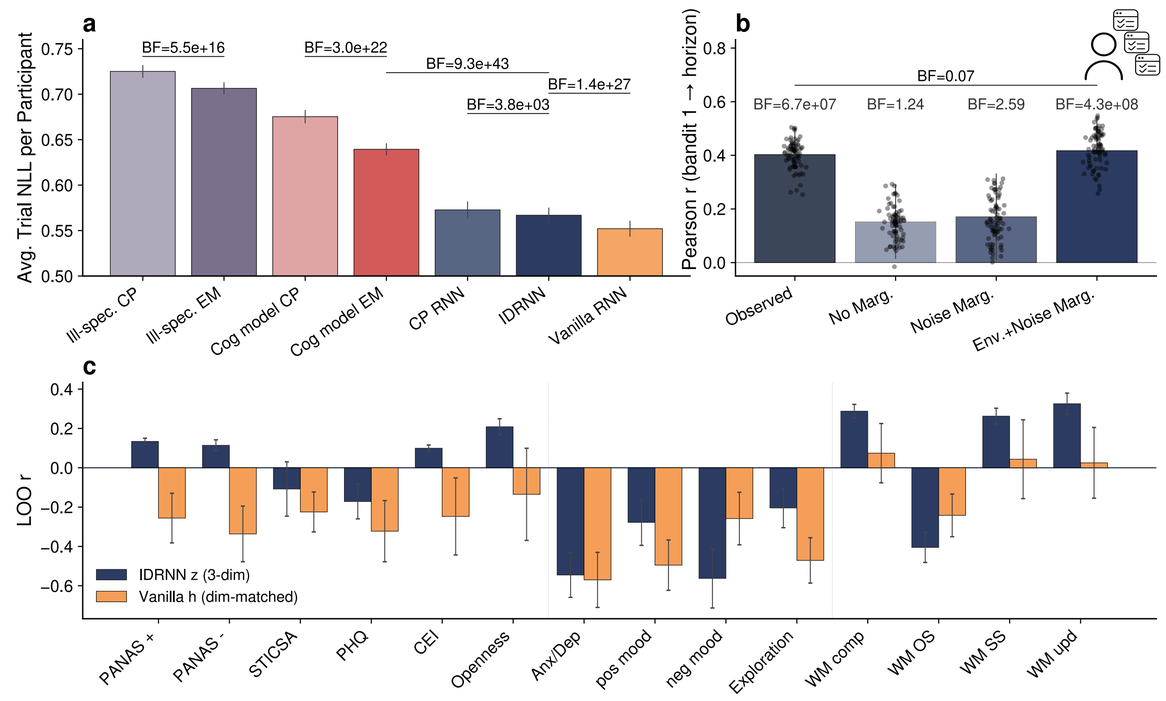
We achieve comparable predictive fit to vanilla models, recover ground truth parameter structure and disentangle between-subject and environmental variance.

Generative recovery of individual differences (panel e here, panel b Human Data): By marginalizing over environment and action stochasticity, we obtain a measure of behavior driven primarily by participant-specific representations.



Human Data: learning subject representations across 3 reinforcement learning tasks

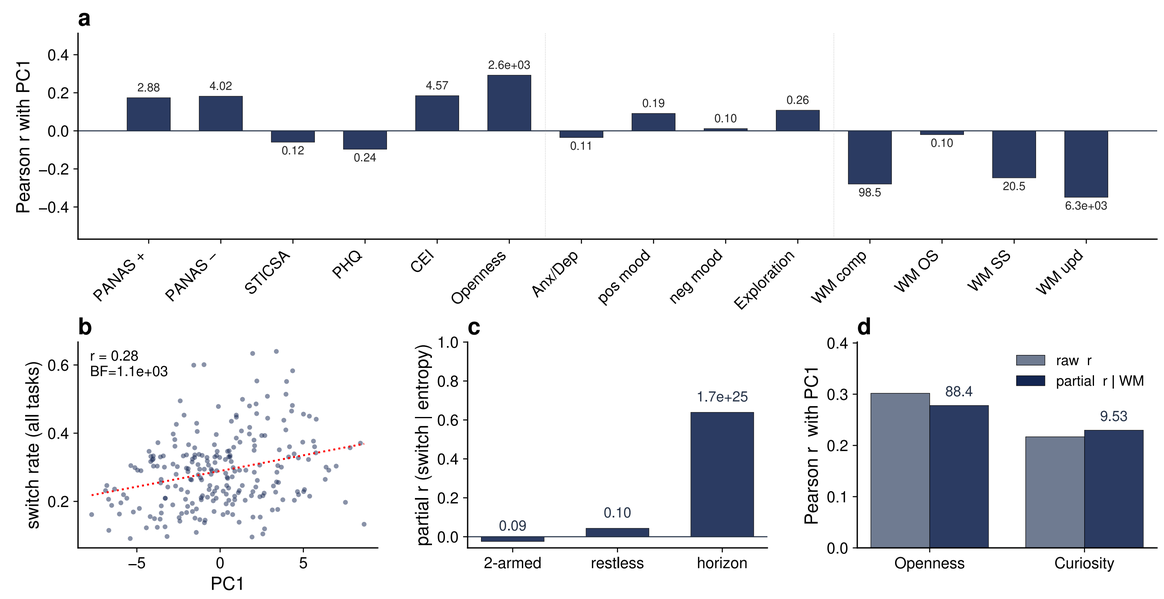
Subject-level representations learned across 3 tasks reveal relationships to psychological traits that could not be uncovered with previous methods.



PC1 of subject-specific representations encodes trait related exploration/exploitation axis

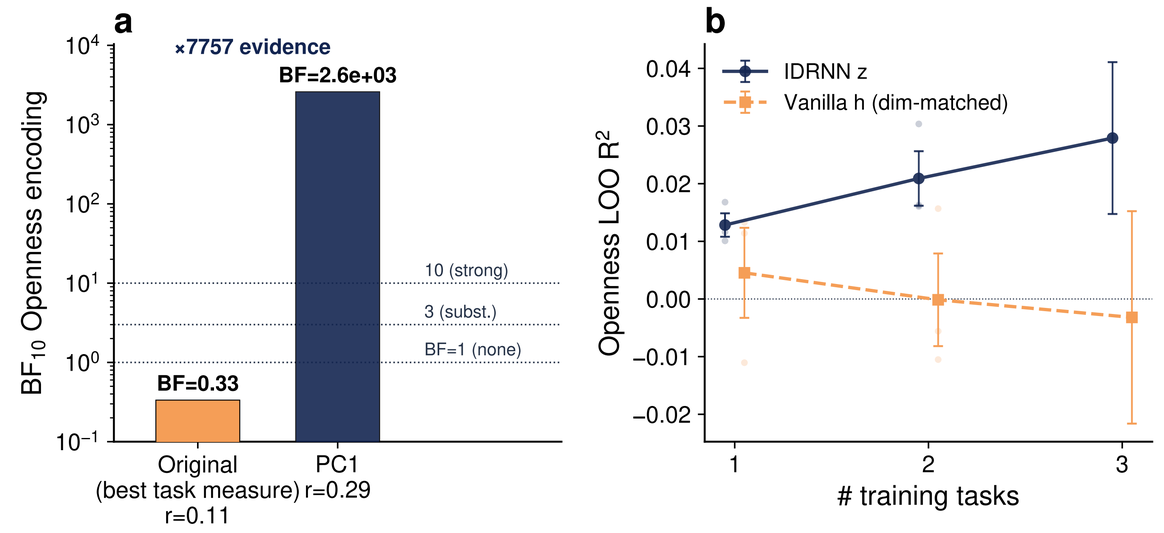
First principal component across seeds and participants encodes an exploration/exploitation axis - with WM capacity on the exploitation end and curiosity related personality traits on the exploration end.

WM capacity and personality traits are separate sources of individual differences in exploration behavior.



IDRNN allows to accumulate evidence across contexts

Because IDRNN learns static subject-specific representations, it can integrate behavioral information of an individual across contexts.



Conclusion

IDRNN learns a **static**, **per-subject latent** that isolates what it means—behaviorally—to be a given participant.

We demonstrate the benefits of this approach through decoding analyses and generative simulation.

This recovers an **exploration/exploitation–personality link** that conventional behavioral or model-based measures miss.

@marvinmathony marvin.mathony@helmholtz-munich.de @marvinmathony.bsky.social