Learning interpretable positional encodings in transformers depends on initialization

arXiv

Takuya Ito[1], Luca Cocchi, Tim Klinger, Parikshit Ram, Murray Campbell, Luke J. Hearne

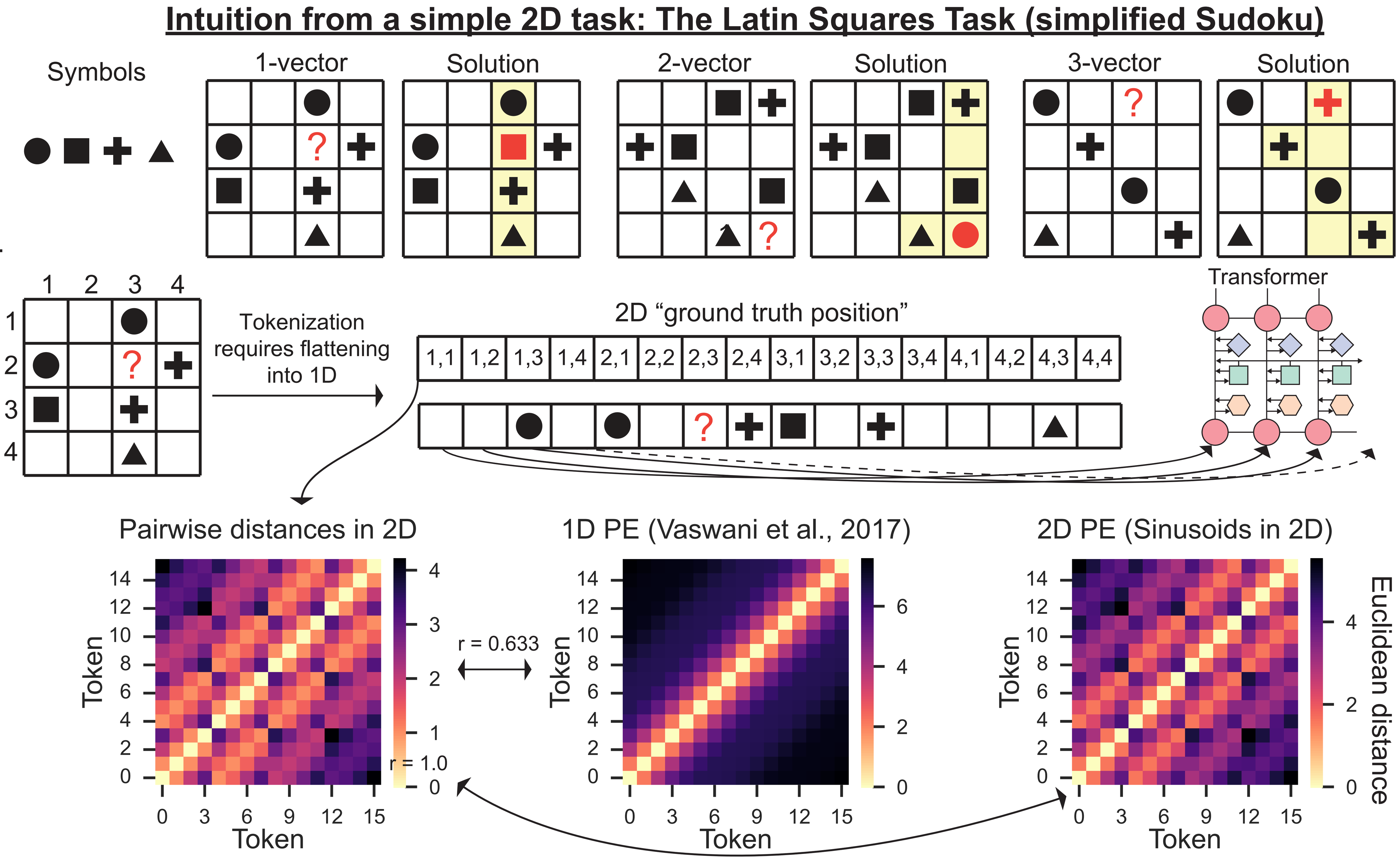
[1]Mathematics of Computation, IBM Research, [2]QIMR Berghofer Medical Research Institute, QLD, Australia

E-mail: taku.ito1@gmail.com

Motivation: The importance of positional encoding choice for transformer generalization

* Choice of positional encodings (PE) in transformers have been shown to be critical for learning and generalization.
* Most investigations into PE have been tailored towards 1D string-based tasks, such as arithmetic or context-free grammars using pre-specified PEs (e.g., ROPE, or absolute PEs)
* Here we investigate the importance on a suite of tasks with sequence data organized in higher dimensions (>greater than 1D sequences)
* Specifically, we study the conditions by which we can ***learn interpretable positional encodings***, and study how they impact generalization
* Inspired by recent work on rich and lazy representation learning, we explored how initialization of a learnable PE parameter influences interpretability and generalization in transformers

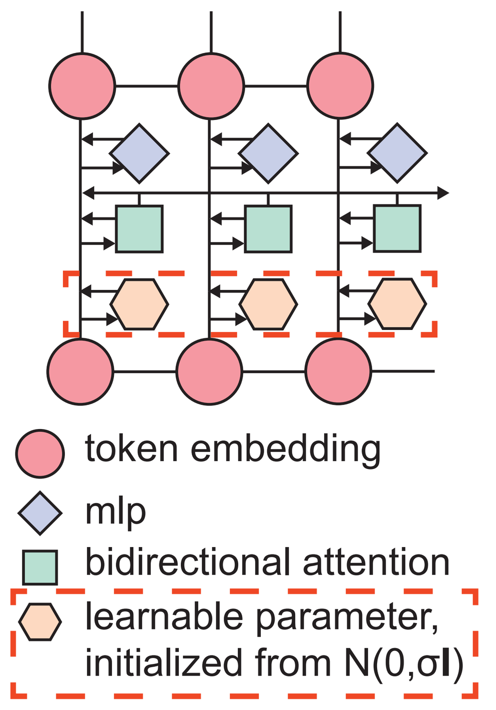
Intuition from a simple 2D task: The Latin Squares Task (simplified Sudoku)



Experiment 1: Learning interpretable positional encodings in the Latin Squares Task

Model manipulation

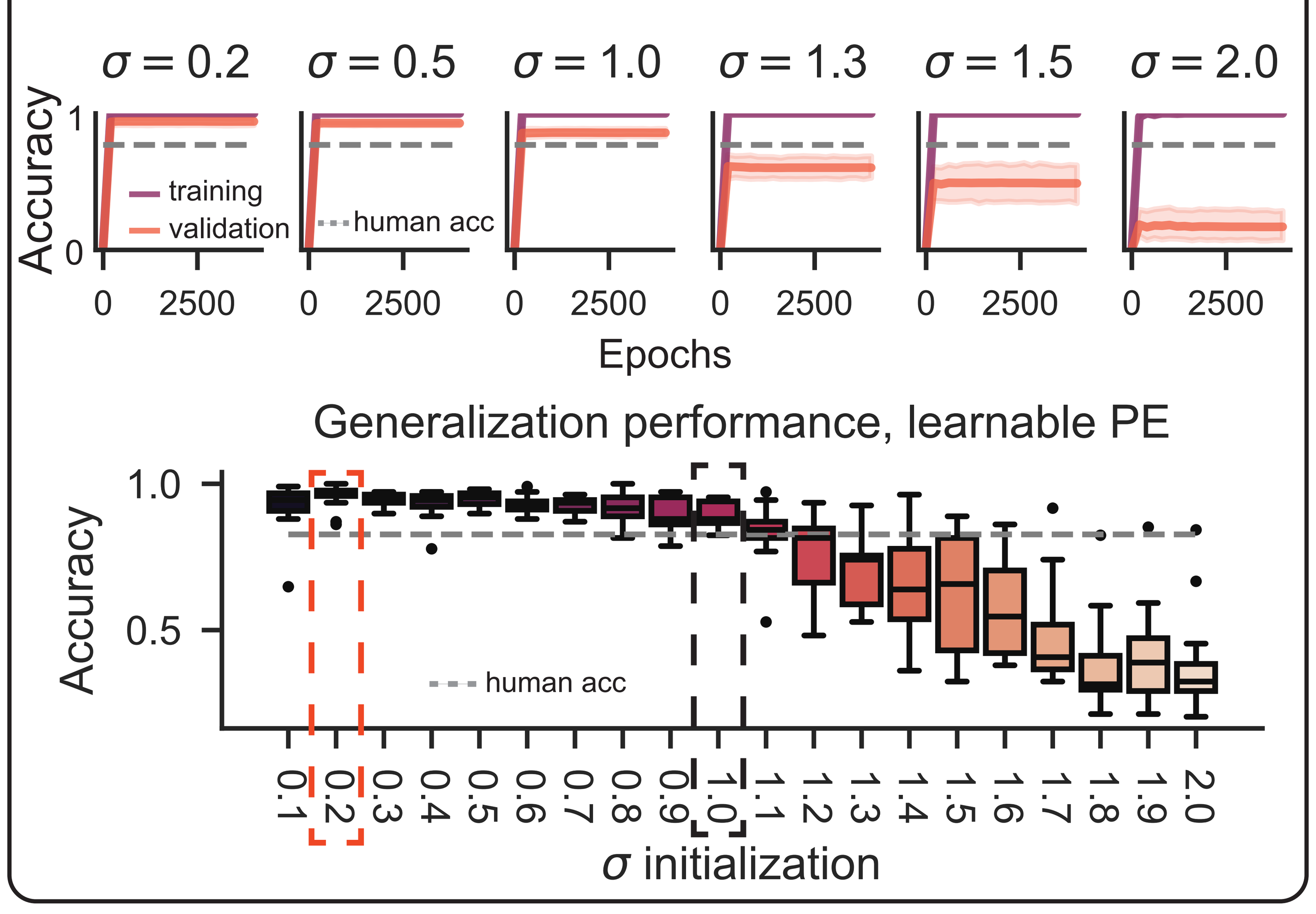
Replace traditional PEs with a learnable parameter initialized from **N(0,σI)**, and manipulate σ



Generalization

Enhanced test set generalization when training a learnable PE embedding initialized from small σ

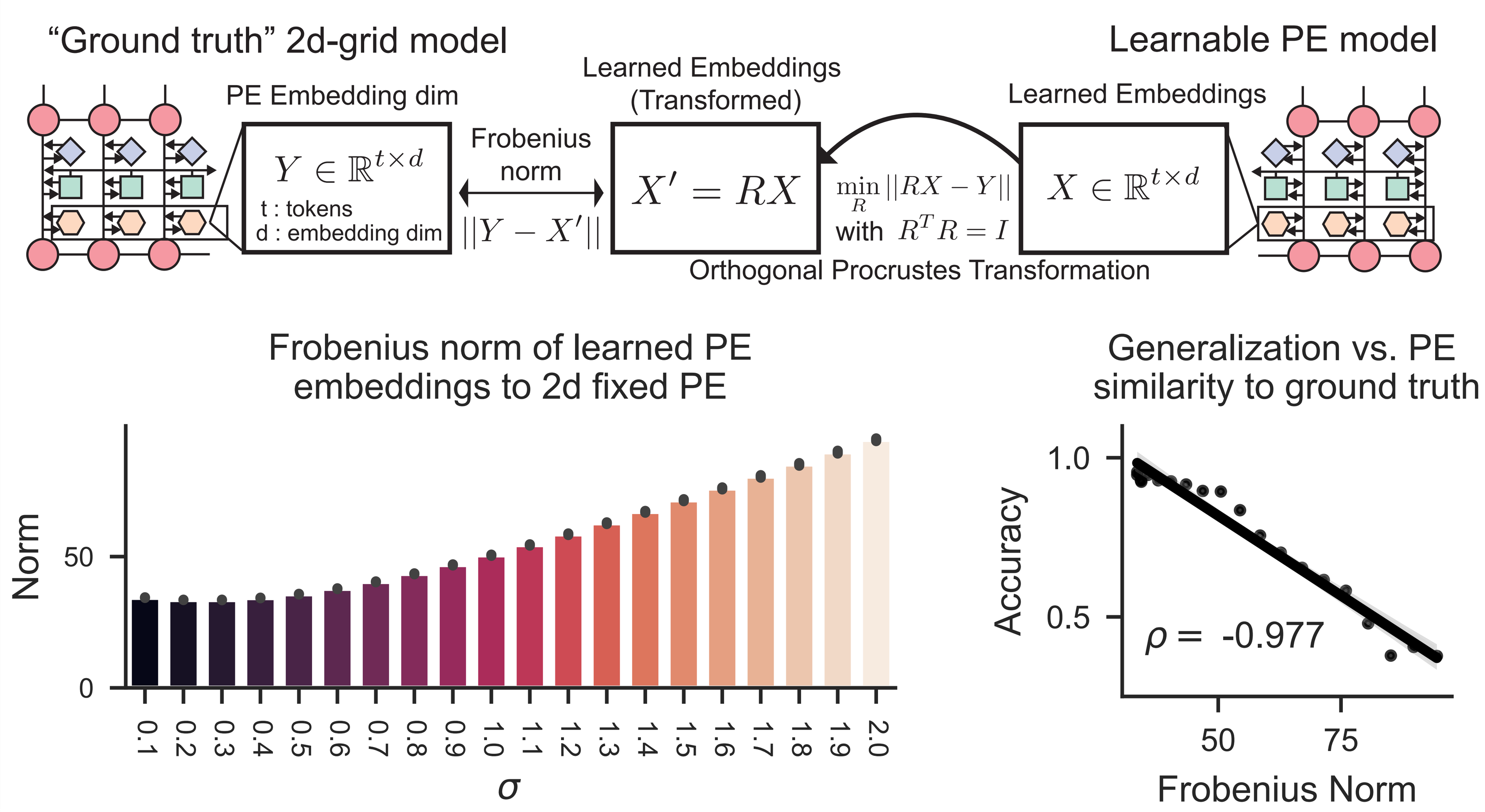
Large σ: Lazy learning | Small σ: Rich learning



Interpretability analyses

Enhanced interpretability of PE embeddings when using a learnable PE embedding initialized with small σ

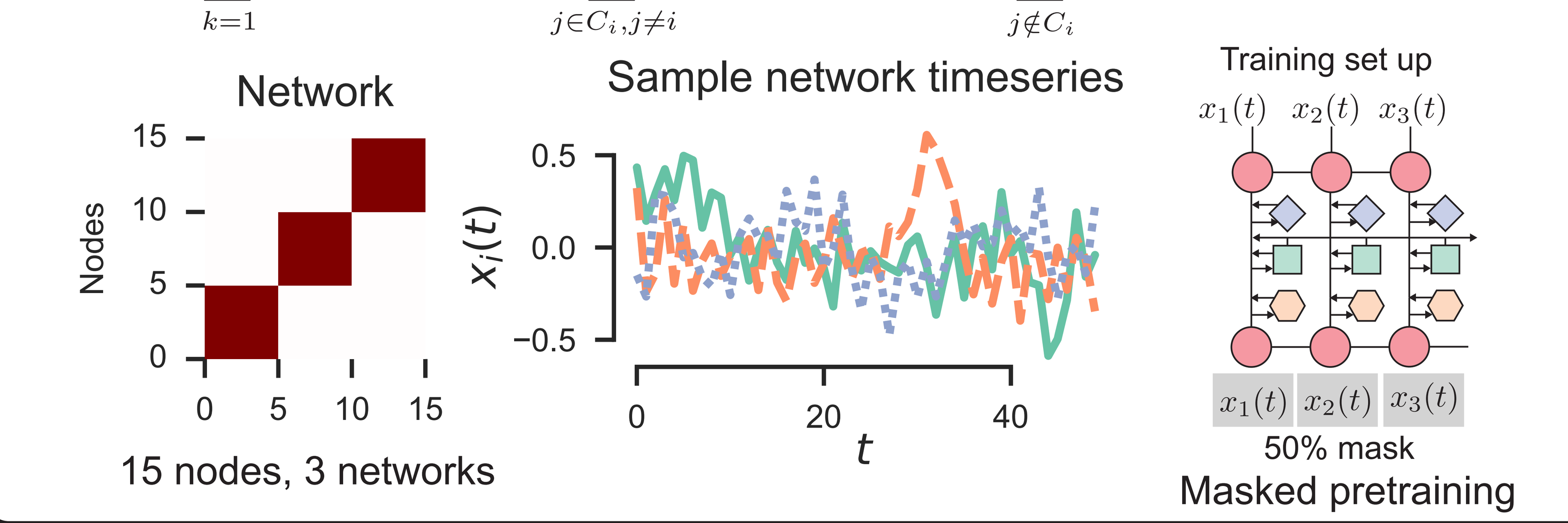
Learned PE embedding mimics a 2D grid structure, consistent with the LST input



Experiment 2: Learnable PEs recover interpretable network clusters in network simulation

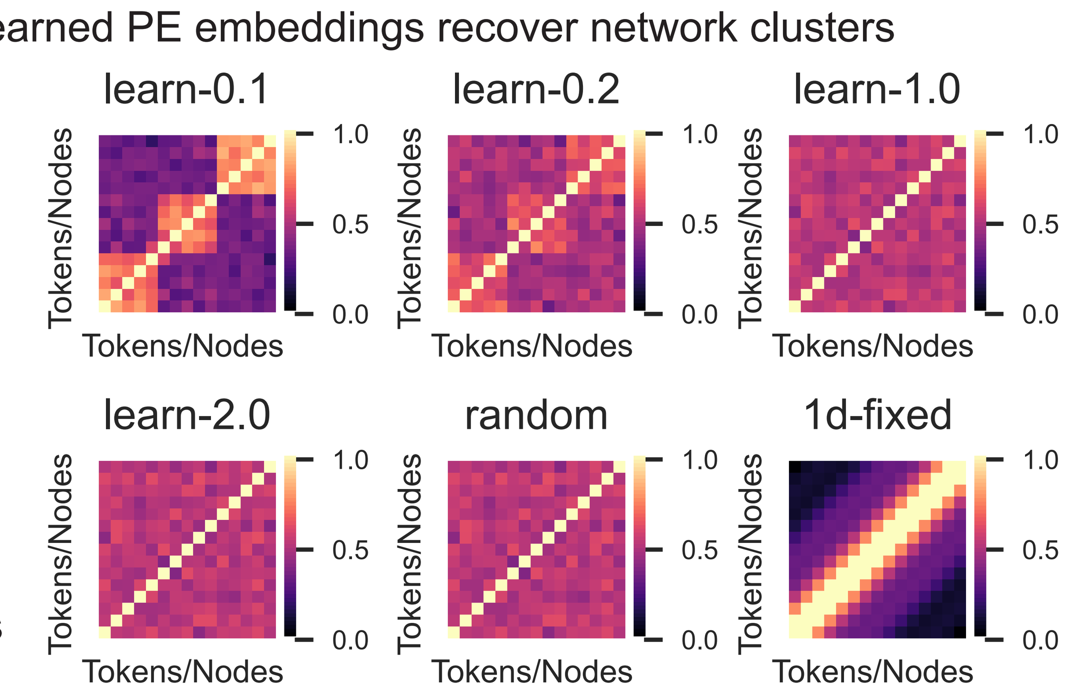
Experimental setup: Simulate nonlinear autoregressive network simulation

$$x_i(t) = \sum_{k=1}^{p} w_{i,k} \cdot x_i(t-k) + \sum_{j \in C_i, j \neq i} \lambda_{ij} \cdot f(x_j(t-1)) + \sum_{j \notin C_i} \eta_{ij} \cdot f(x_j(t-1)) + \epsilon_i(t)$$



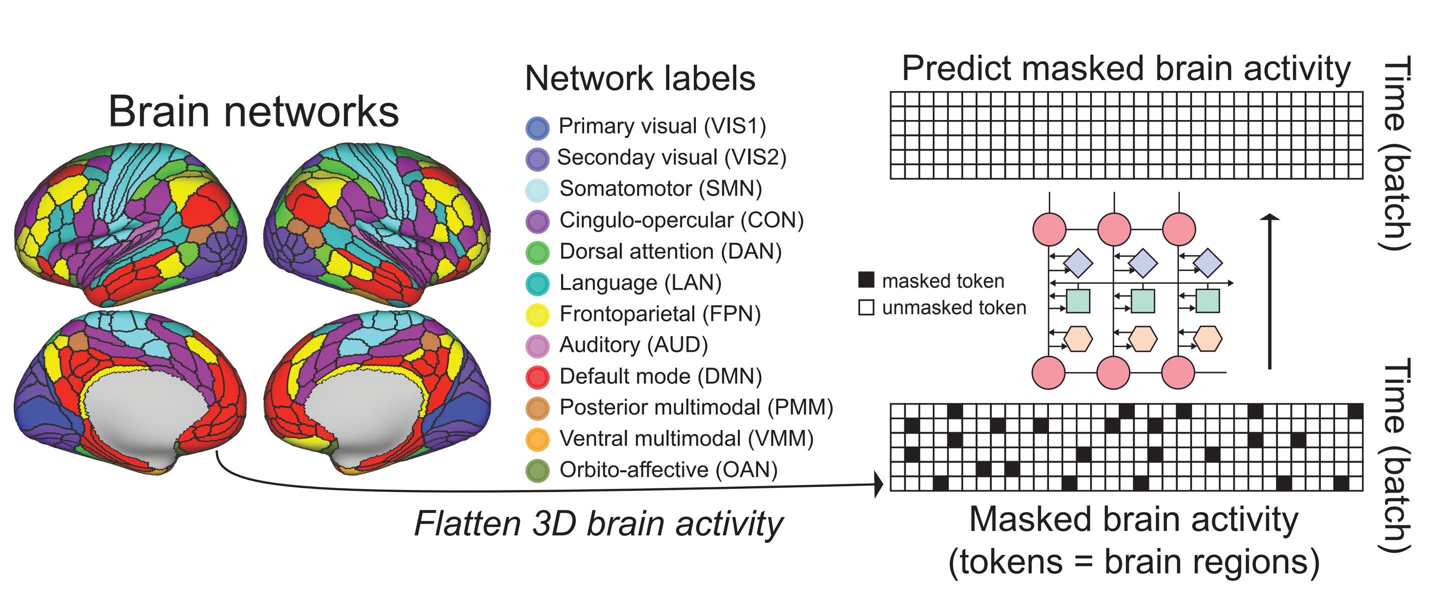
Interpretability analyses: Richly learned PE embeddings recover network clusters

* We measured the cosine distance of learned PE embeddings across a range of PE models
* Small σ models (e.g., learn-0.1) accurately recovered the distance between PE embeddings
* Large σ models (e.g., learn-1.0) did not
* Models with random and 1d-fixed PEs also did not exhibit network structure

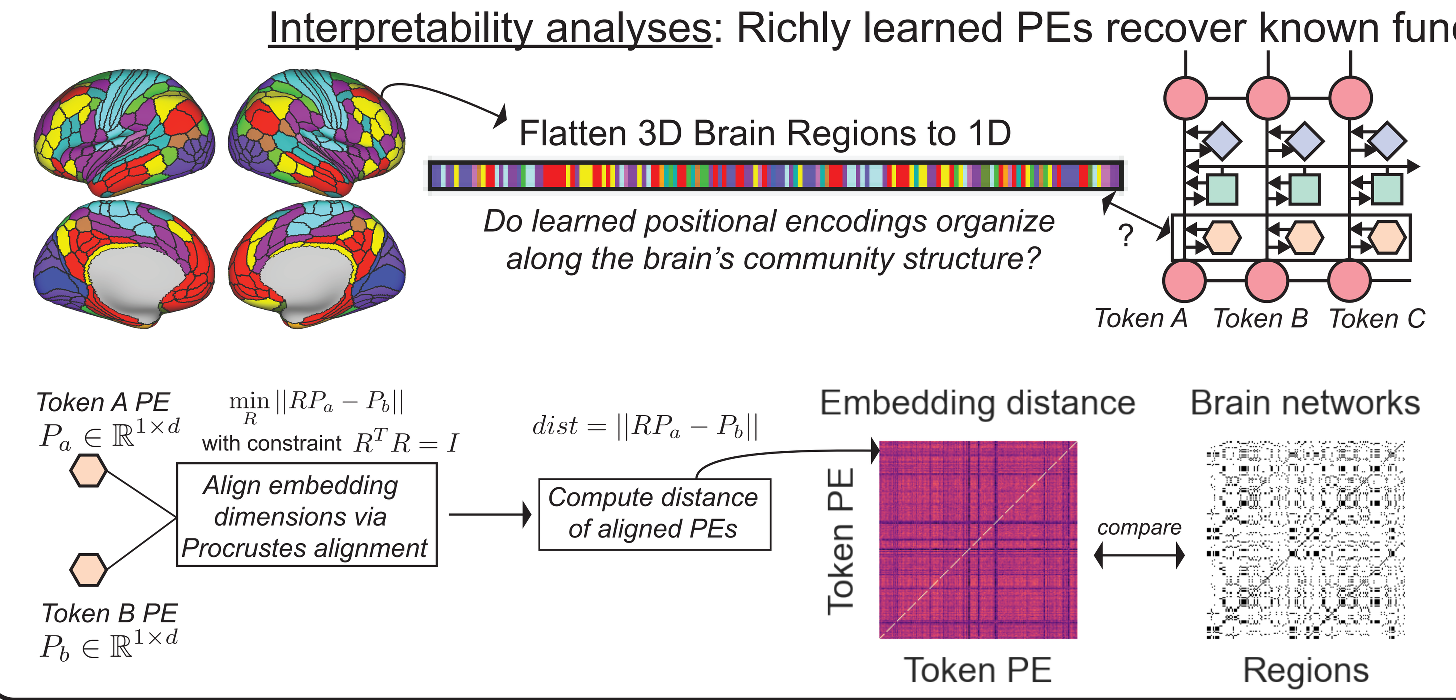


Experiment 3: Learnable PEs recover known network clusters in human brain fMRI data

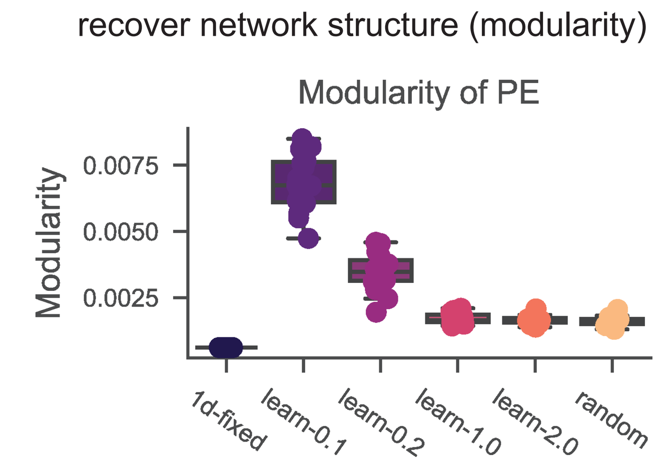
Experimental setup: Predicting whole-brain activity from masked inputs



Interpretability analyses: Richly learned PEs recover known functional network clusters



* We compute the distances between learned PEs in transformers
* PEs learned in the rich regime (small σ) recover network structure (modularity)



Conclusion

* We extend prior transformer generalization studies from 1D sequences to n-dimensional sequences, which requires positional encoding schemes for higher dimensions.
* We demonstrate that ***rich representation learning of positional encodings*** – which is induced by initializing parameters with a small norm – ***learns interpretable embeddings that also enhance generalization***