LEARNING POSITIONAL ENCODINGS IN TRANSFORMERS DEPENDS ON INITIALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The attention mechanism is central to the transformer's ability to capture complex dependencies between tokens of an input sequence. Key to the successful application of the attention mechanism in transformers is its choice of positional encoding (PE). The PE provides essential information that distinguishes the position and order amongst tokens in a sequence. Most prior investigations of PE effects on generalization were tailored to 1D input sequences, such as those presented in natural language, where adjacent tokens (e.g., words) are highly related. In contrast, many real world tasks involve datasets with highly non-trivial positional arrangements, such as datasets organized in multiple spatial dimensions, or datasets for which ground truth positions are not known, such as in biological data. Here we study the importance of learning accurate PE for problems which rely on a non-trivial arrangement of input tokens. Critically, we find that the choice of initialization of a learnable PE greatly influences its ability to learn accurate PEs that lead to enhanced generalization. We empirically demonstrate our findings in a 2D relational reasoning task and a real world 3D neuroscience dataset, applying interpretability analyses to verify the learning of accurate PEs. Overall, we find that a learned PE initialized from a small-norm distribution can 1) uncover interpretable PEs that mirror ground truth positions, 2) learn non-trivial and modular PEs in a real-world neuroscience dataset, and 3) lead to improved downstream generalization in both datasets. Importantly, choosing an ill-suited PE can be detrimental to both model interpretability and generalization. Together, our results illustrate the feasibility of learning identifiable and interpretable PEs for enhanced generalization.

1 INTRODUCTION

Transformers commonly use ordered sequences of data, like words in a sentence. The position and order of these words are crucial to their correct interpretation. In transformers, sequences of tokens (e.g., words) are processed in parallel – not sequentially. Thus, to process tokens correctly in their intended sequence, the transformer must encode a notion of position and/or ordering of tokens. This information is encoded in its positional encoding (PE) layer – a model parameter that tags each input token with a unique location. For many common forms of data, such as natural language, text, and audio, the labeling of ground truth positional information is straightforward, since tokens are ordered sequences in 1D. This led to the original design of 1D sinusoidal PEs, which were successfully applied to natural language data, and provided general spatial information about language tokens (rather than data-specific information) (Vaswani et al., 2017). More recent investigations into the role of PE in transformers has led to a proliferation of PE schemes, each specially designed for 1D text with different properties (Su et al., 2022; Shaw et al., 2018; Vaswani et al., 2017; Raffel et al., 2020; Li et al., 2024; Kazemnejad et al., 2023; Shen et al., 2023; Golovneva et al., 2024; Press et al., 2022). However, many interesting problems require input sequences that are not in 1D (e.g., image datasets), or where position information is non-trivial or not known (e.g., biological data). The choice of PE significantly affects the performance of transformer models, even in simple string-based tasks (Kazemnejad et al., 2023; Ruoss et al., 2023; McLeish et al., 2024). Thus, understanding how to disambiguate and learn ideal position information from data directly would likely provide improved performance while affording increased model flexibility.

If learning the optimal PE for a task can enhance downstream generalization performance, what strategies can we use to achieve this? Recent work in deep learning theory suggests various parame-

terizations of simple neural network models, such as weight initializations, can greatly influence their learned internal representations (Woodworth et al., 2020; Chizat et al., 2020). In particular, those studies found that weight initializations in neural networks from large-norm distributions (e.g., a normal distribution with a large standard deviation) learned random, high-dimensional representations that would "memorize" input-output relations. This learning regime is commonly-referred to as the *lazy* learning or *neural tangent kernel* (NTK) regime, as it fails to learn a structured representation of the task or input. In contrast, neural networks that were initialized from a small-norm distribution (e.g., $\mathcal{N}(0, \sigma)$ for small σ) tended to learn structured representations that accurately reflected the organization of input features and were robust to noise. This is referred to as the *rich* or *feature* learning regime (Woodworth et al., 2020; Chizat et al., 2020). Although this theoretical framework was initially developed for simple neural networks (e.g., feed-forward networks with few hidden layers), the insights drawn from it should apply to various model architectures, including transformers (Zhang et al., 2024; Kunin et al., 2024). Given the recent interest in studying the impact of PEs on generalization, we aimed to evaluate whether the norm of PE initialization would influence the ability to properly learn a structured and accurate PE that would enhance generalization.

Here we studied how the initialization of learnable PEs in transformers influence representation learning and downstream generalization. We focused on problems containing sequences with nontrivial positioning and ordering, comparing the generalization performance of models with learned PEs to other common PE schemes (such as absolute and relative PEs). We also examined the interpretability of the learned PEs. *Our primary aim was to evaluate the hypothesis that learnable PEs in the feature rich learning regime would produce interpretable position information that mirrored ground truth knowledge and improve generalization performance.* We tested this hypothesis in a synthetic task and a real world biological dataset: 1) The Latin Squares Task (LST), a relational reasoning task in 2D that is analogous to simplified Sudoku, and 2) a real-world neuroscience dataset, where the task is to predict masked 3D brain activity from spatially and heterogeneously distributed brain regions. Overall, we found that when the PE is appropriately initialized with a small norm, learnable PEs can uncover ground truth PEs in the LST task, 2) learn non-trivial and modular positional representations of brain regions in a neuroscience dataset, and 3) lead to improved downstream generalization in both datasets. These results indicate the importance of PE choice for generalization performance, and provide insights into how to optimally discover PEs for a variety of tasks in which the ground truth PE is nontrivial or not known.

1.1 Contributions

We highlight three principal conclusions of this study.

1. Using a 2D relational reasoning task with known ground truth positions, we demonstrate that the ability to approximate the ground truth PE depends on initializing a PE parameter with a small norm.
2. We demonstrate the generality of this approach to learning nontrivial PEs in a real world 3D neuroscience dataset, where only small-norm initialized PEs learn a representation of the functional position of brain regions that reflect brain network modularity.
3. We demonstrate that in both sets of experiments, learning an accurate PE enhances downstream generalization relative to alternative and commonly-used PEs.

2 Datasets and Model Evaluation

2.1 The Latin Square Task (LST)

We first evaluate the effect of initialization on learning accurate PEs on the LST, a 2D relational reasoning task that is analogous to the game Sudoku (Fig. 1A-D). The LST is a nonverbal relational reasoning task developed in line with the psychological theory of *Relational Complexity* (Birney et al., 2006; Halford et al., 1998). Prior work in humans has demonstrated the reliability of the LST and its relationship to fluid intelligence ((Hearne et al., 2020; Birney et al., 2012; Hartung et al., 2022)). Each puzzle in the LST involves the presentation of a four-by-four grid populated with stimuli (e.g., shapes, numbers, etc.,), blank spaces, and a single target probe location, noted with a question mark. The fundamental rule of the LST is that in a complete puzzle (i.e., there remain no empty squares),

each shape can only appear once in each row and column. In our setup, the agent's aim is to infer the unknown target stimuli based on the organization of the elements within the LST grid.

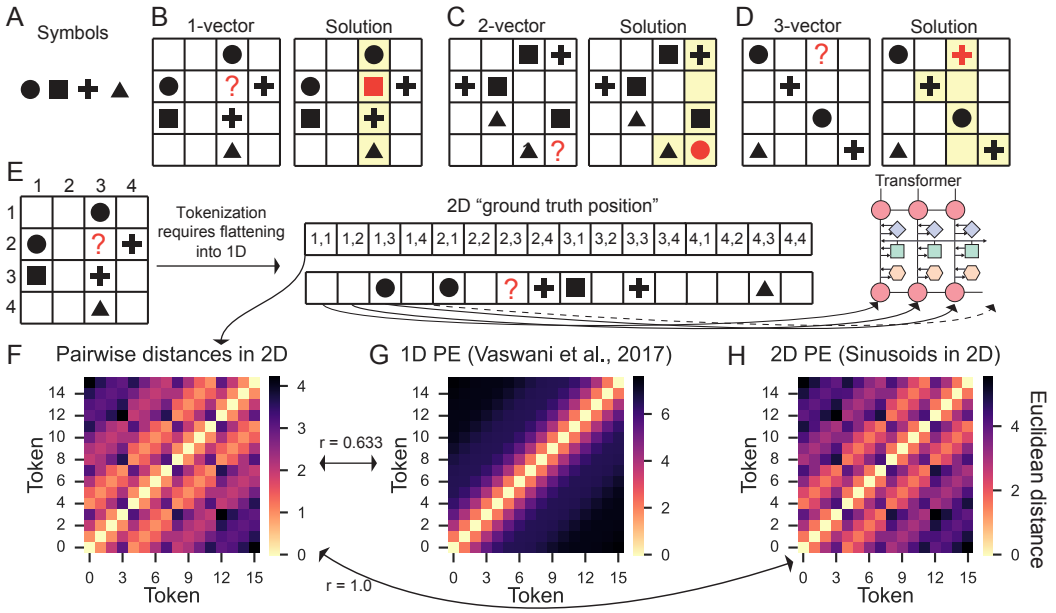


Figure 1: A-D) The Latin Square Task (LST). The LST involves the presentation of a 2-dimensional 4-by-4 grid populated with A) four possible symbols, blank spaces and a single target probe ("?"). The agent's aim is to solve for the target square with the rule that *each shape can only appear once in every row and column*. The reasoning complexity required for an LST puzzle can be manipulated by varying the number of distinct vectors that must be integrated to solve the problem. Examples of B) 1-, C) 2-, and D) 3-vector LST puzzles (left) and their solutions (right). E) Performing the LST is intuitive in 2D. However, when tokenizing the task for neural networks, the input must be flattened into 1D. The LST is significantly more challenging when row and column information is lost. F) The pairwise distances between token positions according to rows and columns provides the "ground truth" of how tokens relate to each other in 2D space. G) Naively using the 1D PE with sines and cosines (Vaswani et al., 2017) would provide incorrect token-wise position information, since it only considers the closeness of tokens in 1D. H) In contrast, recomputing absolute positions in 2D (with sines and cosines in the embedding dimensions) would preserve the position information of a 2D grid, even after flattening the sequence into 1D. (Note that any choice of PE that preserves relative distances in 2D space would be appropriate.)

The number of relations needed to solve a given LST puzzle can be manipulated by changing the organization of the elements within the grid (Fig 1A-D). For instance, one-vector puzzles require integration of information across a single row *or* column, while two-vector problems involve integration across a single row *and* column. Three-vector puzzles require information integration across three rows and/or columns. Importantly, performing the LST task requires positional informational information of the rows and columns. Flattening the grid to 1D without preserving the position information would significantly increase the difficulty of the LST task (Fig. 1E).

For our LST experiments, we generated eight thousand training puzzles, and assigned to one-, two-, or three-vector conditions. We ensured that the similarity between any generated training set puzzle was distinct from the generalization (validation) set of puzzles (i.e., the *Jaccard dissimilarity* > 0.8 for a test puzzle to any individual training puzzle).

2.2 Human functional MRI dataset

In the second experiment, we evaluated the effect of learning accurate PEs in a real world 3D neuroscience dataset. In this context, we conceptualized 'tokens' as distinct brain regions across the cortical mantle with the goal of predicting masked brain activity with a self-supervised objective. Naively, the PE of brain regions might manifest as their physical location in 3D space. However, decades of neuroscience research has revealed a modular brain organization, whereby different brain regions belong to distinct functional networks (or communities) (Power et al., 2011; Yeo et al., 2011; Ji et al., 2019; Schaefer et al., 2018). Thus, the goal of using this real-world dataset was to

assess the degree to which models could recover this "modular" network organization in terms of PE. We used publicly available human functional magnetic resonance imaging (fMRI) data from the Human Connectome Project (HCP) dataset (`www.humanconnectomeproject.org`). We used the resting-state fMRI data from a subset ($n = 100$; $n_{train} = 70$; $n_{test} = 30$) of the HCP 1200 participant pool (Van Essen et al., 2013). The subset of 100 participants was selected based on quality control assessments that were previously described in Ito et al. (2020). Each subject contained 4800 timepoints of brain activity (sampled at 720ms). We used the Glasser et al. (2016) parcellation that identified 360 distinct cortical regions (tokens), and the CAB-NP network partition, which partitioned 360 cortical regions into 12 distinct functional networks based on resting-state functional connectivity profiles (Ji et al., 2019). Details on the preprocessing of fMRI data can be found in Appendix A.2.

2.3 Model architecture and training

For all experiments, we used a standard encoder-only transformer architecture with four layers and embedding dimension of 160 for the LST, and 64 for the fMRI data (Vaswani et al., 2017). The primary transformer manipulation was the choice of PE, and includes a mix of absolute, relative, and learnable PEs (described in later sections). The formulations and definitions of common PEs are detailed in Appendix A.1. For each model variation, we trained on 15 seeds. For simplicity of analyzing attention maps, we trained models with only a single attention head (fully-connected, bidirectional attention unless specified otherwise). However, we have included results for models with multiheaded attention (2 and 4 heads) on the LST task in Fig. A14, which overall reduce generalization performance. The context window for the model was either 16 tokens long for the LST (given the 4×4 structure of the LST paradigm), or 360 tokens long (for the number of brain regions in the Glasser et al. (2016) brain atlas). We used the Adam optimizer with a learning rate of 0.0001. For comparable analysis, all models were trained for a fixed number of training steps (4000 epochs for the LST; 8000 puzzles per epoch; 50k training steps for human brain data). Results reported in the main text were trained without regularization, given that we were interested in understanding the role of PE initialization in isolation (no dropout, no weight decay). However, for completeness, we include results using weight decay (AdamW, with weight decay=0.1) in the Appendix for the LST task, which yielded qualitatively similar results (Fig. A7, Table A5, Table A6). A single model/seed could be trained (4000 epochs of LST) on one NVIDIA V100 GPU in under 45 minutes. All model training was performed on an internal cluster.

The objective function for the LST task paradigm was to choose the correct symbol in the location with the red question mark (Fig. 1A-D) using a cross entropy loss. Generalization in the LST experiment was performed on new puzzles that had low dissimilarity to the training set (Jaccard dissimilarity > 0.8). For the human brain data, the objective function was to predict the activity of the masked brain regions in the input by minimizing the mean squared error (MSE) of masked brain activity. Generalization in the human dataset was to evaluate MSE prediction of masked brain activity data from a separate human participant.

3 Experiments and results

3.1 PE initialization influences downstream generalization in the LST

Prior work in deep learning theory suggests that choice of weight initialization can influence learned representations and downstream generalization (Jacot et al., 2020; Woodworth et al., 2020; Chizat et al., 2020). In particular, the smaller the norm of the distribution from which the neural network is initialized, the more structured the learned representations will be. Thus, we first assessed whether these intuitions would generalize to learnable PEs initialized from different Normal distributions, controlling for the standard deviation (i.e., norm). For each token embedding, we initialized a learnable PE parameter from a multivariate Normal distribution, denoted $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \sigma\mathbf{I}$, and $\mathbf{I}$ denotes the identity matrix scaled by σ (Fig. 2A).

We initialized PEs from distributions with $\sigma \in \{0.1, 0.2, 0.3, ..., 2.0\}$. (For each PE initialization, we trained on 15 seeds.) We trained all models for 4000 epochs, and found that all models converged (Fig. 2B). Remarkably, though the choice of σ was the only source of variation across model parameterizations, models exhibited a wide range of generalization performance (Fig. 2C). Compared to the default initialization choice of $\sigma = 1$, which had a generalization performance of 0.89 (Table

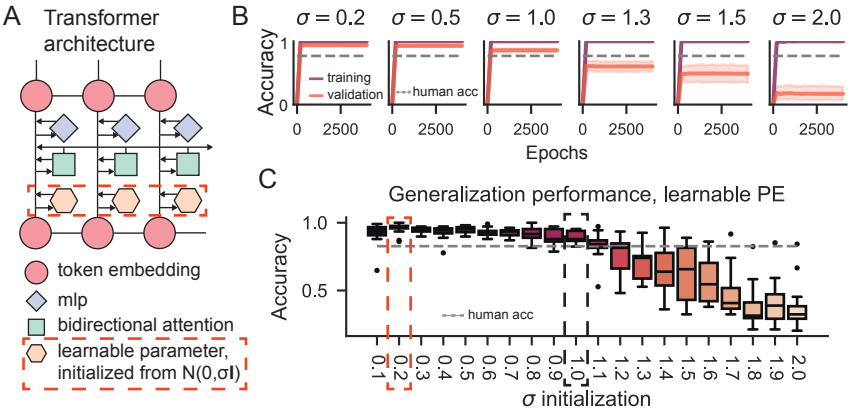


Figure 2: A) Transformer parameterization. We parameterize a learnable PE initialized from different distributions (i.e., $\mathcal{N}(0, \sigma\mathbf{I})$, varying σ), and study the effect initialization on downstream generalization. B) The training and validation performance across 4000 epochs for examplar initializations. C) Despite the learnability of all initialized models, the choice of σ strongly influenced downstream generalization. Boxplots reflect variability across 15 random seeds. (See also Appendix Table A2.)

A2), we found that the optimal generalization performance was produced with PEs initialized with small norms (e.g., $\sigma = 0.2$; Acc=0.96; Table A2). Moreover, consistent with the NTK regime, PE's initialized from a distribution with large σ converged, but generalized poorly (e.g., $\sigma = 2.0$, Acc=0.38; see also Table A2).

While our results are consistent with the hypothesis that small norm initialized models tend to learn the most generalizable representations, we did observe a slight reduction in generalization performance for PEs initialized at $\sigma = 0.1$ relative to $\sigma = 0.2$. In practice, we found that the poor generalization for very small values of σ is due to the behavior of the Adam optimizer. Specifically, vanilla SGD outperformed the generalization ability of models trained with Adam for small values of $\sigma \in \{0.01, 0.05, 0.1\}$ (see Fig. A6). Thus, it is important for practitioners to consider optimizers when encouraging the rich/feature training regime for very small initializations.

3.2 Learnable PEs outperform commonly-used PE schemes in the LST

We next sought to benchmark the optimal learnable PE ($\sigma = 0.2$, `learn-0.2`) relative to other standard PE schemes. These schemes included absolute 1D PE (`1d-fixed`, using sines and cosines; Vaswani et al. (2017)), relative PE (`relative`; Shaw et al. (2018)), and rotary PE (`rope`; Su et al. (2022)). In addition, we also evaluated performance on PE schemes that have been shown to be beneficial for algorithmic and compositional generalization tasks, including no PE with a causal attention mask (`c-nope`) (Kazemnejad et al., 2023) and random PE without a learnable parameter (`random`) (Ruoss et al., 2023). (Note that the `c-nope` model can implicitly learn position information due to the nature of the causal attention mask.) We also included a baseline control model without any specified PE (`nope`), which should not be able to learn the task in a systematic way, due to the permutation invariance of bidirectional attention in the absence of PE. Finally, we included a "ground truth" PE – an absolute 2D PE based on sines and cosines (`2d-fixed`) – to compare how similarly the various PEs produced attention mechanisms to this ground truth model (see Appendix A.1). (Note, that the term "ground truth" applies to any rotation of the absolute 2D PE, or any PE that preserves the original 2D row and column LST information, as depicted in Fig. 1E,F).)

All models, except for the `nope` and `c-nope` models converged (accuracy after 4000 epochs, `c-nope`=0.56, `nope`=0.51). Poor performance was expected for the `nope` model due to the lack of any explicit or implicit PE information. However, it was surprising that `c-nope` models neither learned nor generalized, given prior research suggesting that this PE scheme had above average performance on compositional and length generalization tasks (Kazemnejad et al., 2023). Nevertheless, as anticipated, we found the ground truth `2d-fixed` model to exhibit the highest generalization performance (Table 1; Appendix Fig. A9A,D). Remarkably, the next highest performing model was the `learn-0.2` model ($\sigma = 0.2$), followed by the `relative`, `random`, `rope`, and `1d-fixed`

Table 1: Training and validation performance of common PEs and the `learn-0.2` PE on the LST.

PE	Validation acc	Validation SD	Training acc	Training SD
2d-fixed (ground truth)	**0.977**	**0.073**	1.000	0.000
learn-0.2	**0.956**	**0.039**	1.000	0.000
relative	0.920	0.042	1.000	0.000
random	0.888	0.046	1.000	0.000
rope	0.805	0.115	1.000	0.000
1d-fixed	0.781	0.185	0.999	0.000
nope	0.334	0.020	0.509	0.002
c-nope	0.314	0.042	0.559	0.102

PE models, respectively. While we did not find a robust statistical difference between the ground truth `2d-fixed` model (97.7%) and the `learn-0.2` model (95.6%) (`2d-fixed` vs. `learn-0.2`, t-test, $t(13) = 0.96$, $p = 0.35$), there was a significant difference between the `learn-0.2` model with the next highest-performing model (`relative`) ($t(13) = 2.51$, $p = 0.03$). Further, when applying L2 regularization (via weight decay), the generalization performance of learnable PEs was virtually indistinguishable from the ground truth `2d-fixed` PE (both models generalized with 99% accuracy; Table A6; Fig. A7). Finally when performing a perturbation analysis, where we systematically injected noise to the token embeddings and evaluated downstream performance, we found that models that were most robust to noise were those initialized from a low-norm distribution (Fig. A10). Overall, these findings corroborate prior theoretical intuitions in deep learning theory, and suggests their applicability to learning rich transformer PEs in structured reasoning tasks.

3.3 Learnable PE models discover ground truth attention maps and positions

Next, we sought to understand how different PE initializations influenced the learned representations (e.g., attention maps and learned PEs) within the transformer. Since the underlying structure of the LST paradigm is a 2D grid, we used the `2d-fixed` model as the ground truth model. This allowed us to assess how learned attention maps would deviate from optimal attention maps. First, we extracted the attention weights for each model, and computed the cosine similarity of attention weights between the ground truth and learnable PE models (Fig. 3A; Table A3). We then correlated the similarity between learned attention maps and the ground truth with generalization performance. We found that the degree of agreement of learned attention maps with the ground truth (`2d-fixed`) predicted improved generalization ($\rho = 0.96$, $p < 0.0001$; Fig. 3B). Importantly, learned PEs with small σ tended to learn attention representations that were most aligned with the ground truth. (We also computed the Jensen-Shannon Divergence as a complementary distance measure, finding similar patterns; Appendix Fig. A8.) In the Appendix, we also include comparative analyses of the attention maps of models with common PEs (`1d-fixed`, `relative`, etc.) and compare those to the ground truth model (Fig. A9B-D).

Next, to directly interpret what embeddings the learnable PEs converged to during training, we evaluated whether learnable PEs approximated the actual ground truth `2d-fixed` PE (Fig. 3D). This involved estimating the distance (i.e., L2 norm) between the `2d-fixed` PE and the learned PE embedding *after* an orthogonal Procrustes transform was applied. An orthogonal Procrustes transform was applied to rotate and match embedding dimensions according to maximal similarity since embedding dimensions were arbitrary in the learnable PE models. We found that small-norm initialized PEs could better approximate the PEs from the `2d-fixed` PE scheme (Fig. 3E). Critically, more similar ground truth PE approximation (measured by L2 norm) near perfectly predicted downstream generalization ($\rho = -0.98$, $p < 0.0001$; Fig. 3F). These results indicate that 1) low-norm initializations can discover ground truth PEs and their subsequent attention maps, and 2) these discovered PEs predicted downstream generalization.

3.4 Learning interpretable PEs in human brain data for improved generalization

We next investigated the feasibility of learning interpretable PEs in a real-world dataset in which it is difficult to specify a ground truth PE. In neuroscience, a central goal is to be able to predict

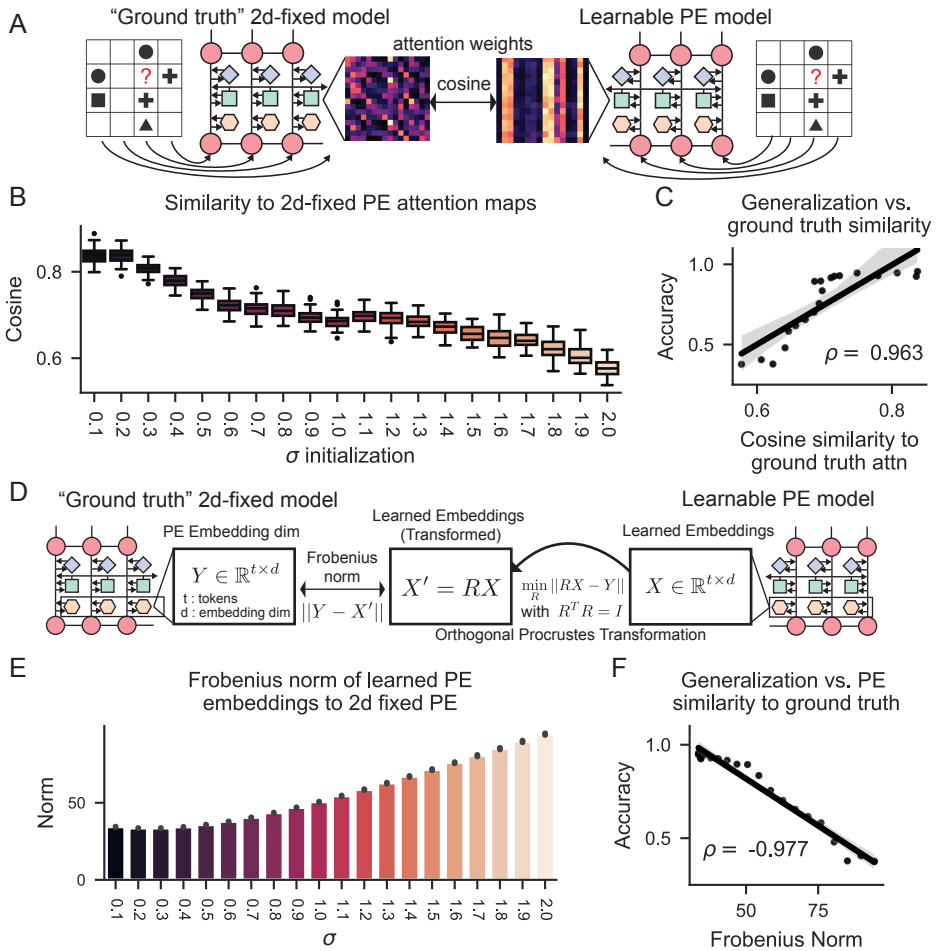


Figure 3: A,B) We compared the attention maps of each learnable PE model to those derived from the "ground truth" `2d-fixed` model, and computed the cosine similarity of the attention maps. C) A strong rank correlation between generalization performance and the agreement of attention maps to the `2d-fixed` model. D) We directly compared the learned PE embeddings to the ground truth `2d-fixed` PE. Since the embedding dimensions of the learned PE models are random and not indexed in the same way to the `2d-fixed` embedding, we matched their embedding dimensions using an orthogonal Procrustes transform. After matching the dimensions of the PE embeddings, we computed the Frobenius norm to calculate the distance between the learned PE and the ground truth `2d-fixed` PE. E) The distance (Frobenius norm) between the learned PEs and `2d-fixed` PE, for every σ. (Note that bar plot colors correspond to differences in x-axis values.) F) We found a strong relationship between the PE agreement with the ground truth PEs and generalization performance ($\rho = -0.977$). (See also Appendix Table A3.)

distributed brain activity using the activity of other brain regions (Bassett and Sporns, 2017). To achieve this, we sought to build a generalizable transformer model that would predict the brain activity of target regions using the brain activity of other regions. This task can be formalized as masked pretraining, where the input tokens to the model are contemporaneous brain activity across different brain regions with masked (or missing) activity values (Fig 4B). This requires the model to predict missing brain activity from its surrounding context, i.e., the brain activity values of other brain regions. This approach is analogous to masked pretraining in the BERT model (Devlin et al., 2019). We trained transformer models with a mix of fixed and learnable PEs (fixed: `1d-fixed`, `relative`, `rope`, `random`; learnable PEs initialized with $\sigma \in \{0.1, 0.2, 1.0, 2.0\}$), minimizing the MSE of masked brain activity. Figure 4 shows results where we performed masked pretraining with 50% masking (Fig. 4C), and tested with 90% masking (Fig. 4D). (We show results with 15%, 75%, and 90% mask pretraining in Appendix Figs. A11,A12,A13) Successful generalization of the trained model would involve predicting the masked brain activity of a separate subject's data (i.e., test subjects). As expected based on the first set of results, we found that small-norm initialized PEs

(`learn-0.1` and `learn-0.2`) achieved the best performance in both their training and validation sets after a fixed number of training steps (50k). Interestingly, relative PEs fared the worst, followed by absolute `1d-fixed` and `random` PEs.

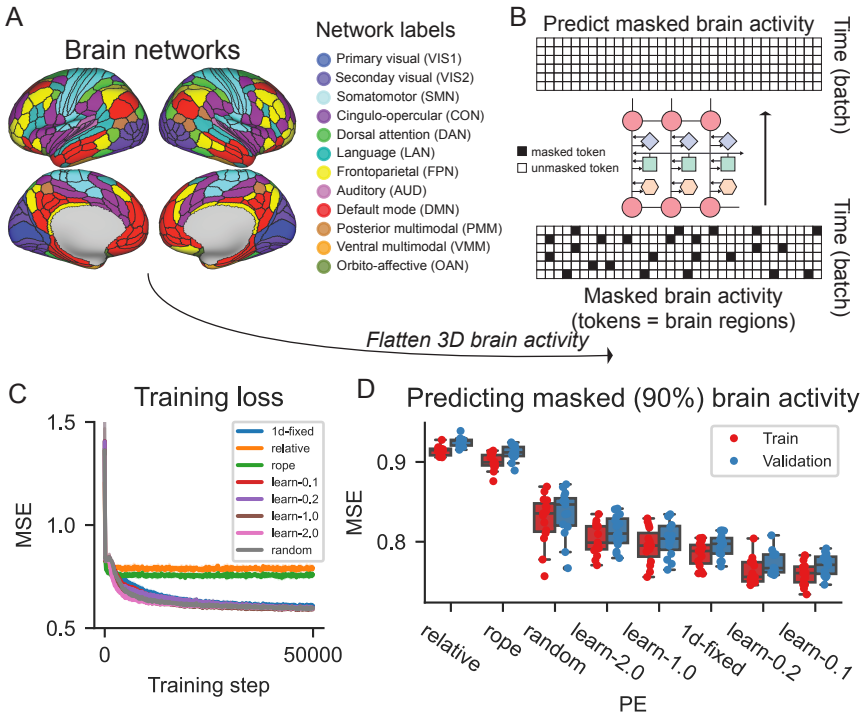


Figure 4: A) The brain is organized into functional networks that are heterogeneously distributed throughout the brain in 3D space. B) A central goal within neuroscience is to be able to predict the activity of brain region(s) using the activity of other brain regions (i.e., tokens). We sought to understand the influence of PE on a model's ability to predict contemporaneous brain activity. This can be formalized as a self-supervised masked prediction task, where a transformer is trained to predict contemporaneous, spatially masked brain activity. To test for generalization, we evaluated the MSE on a validation dataset, which involved predicting the masked activity of data collected from different participants with an increased masking threshold (90%). (Models were trained with a 50% masking. See Figures A11,A12,A13 for results with 15%, 75%, and 90% masked pretraining objectives.) C) We found that transformers endowed with different PE schemes converged to different MSE loss values. D) Notably, learnable PEs initialized with small norms (`learn-0.1` and `learn-0.2`) converged to the lowest MSE for both training and validation datasets. (X-axis is sorted from highest to lowest MSE.)

Having established the superior predictive performance of richly learned PEs (small σ), we next evaluated whether these models learned meaningful yet nontrivial position information from masked pretraining of brain activity. In the masked pretraining setup, we conceptualized 'tokens' as distinct brain regions across the cerebral cortex. Unlike the first set of analyses investigating the LST, the ground-truth PE of brain regions is not known. Naively, the PE of brain regions could be encoded as their physical location in 3D space, which would be trivial to map. However, decades of neuroscience research has revealed a modular brain network organization, whereby different brain regions that are heterogeneously distributed throughout the cortex belong to distinct functional networks (or communities) (Fig. 4A) (Power et al., 2011; Yeo et al., 2011; Ji et al., 2019; Schaefer et al., 2018). In other words, two brain regions that are distant in 3D space may actually be "functionally" close (e.g., yellow regions in Fig. 4A). (Prior work has indicated this functional closeness is determined by anatomical connectivity; Vázquez-Rodríguez et al. (2019).) We therefore sought to address whether richly learned PEs could recover this modular functional organization by measuring the distance of learned PE parameters.

The modular network organization of the human brain is spatially distributed across the cortex. When flattening brain regions across the cortex into a 1D tensor, each brain region's (i.e., token's) assignments are heterogeneously distributed across that tensor (Fig. 5A). (In this context, a PE scheme that places adjacent tokens closer to each other, such as in the original `1d-fixed` PE with sines and cosines from Vaswani et al. (2017), would be clearly ill-suited.) To evaluate whether learned PEs

learned a modular organization that reflected the known functional network organization of the brain, we measured the distance between every pair of tokens' PE. This involved computing an orthogonal Procrustes transform to rotate and match the embedding dimensions of each token's PE prior to computing their distance (Frobenius norm) (Figure 5B). The reason this is necessary is because since learned PEs are first randomly initialized, the embedding dimension of each token's PE are not necessarily aligned (e.g., position 1's embedding dimension i does not necessarily correspond to embedding dimension i of position 2). Thus, after aligning the embedding dimensions across tokens, we computed the distance for every pair of tokens, yielding a 2D distance matrix. (We then scaled distance in this 2D matrix to range from 0 and 1, and computed the complement (i.e., $1-d_{scaled}$), such that closer PEs would have higher values. We computed both the network modularity and network clustering (i.e., segregation) with respect to the known network partitions (see A.3 for mathematical definitions; Rubinov and Sporns (2010)). In brief, modularity is a statistic that quantifies the degree to which the distance matrix can be cleanly subdivided into the brain's network partitions. Network clustering is a statistic that quantifies the ratio between within-module distances and across-module distances, where modules are defined using a network partitioning from Ji et al. (2019). We found that the modularity of small-norm initialized PEs (`learn-0.1` and `learn-0.2`) had the highest overall network modularity and segregation relative to other learnable PEs. This implies that the small-norm initialized PEs learned interpretable PEs with respect to the brain's known biological networks. These findings support the hypothesis that learnable PEs (as opposed to off-the-shelf PEs) in the rich training regime can improve generalization, while successfully learning nontrivial and interpretable position information (Figure 5C,D).

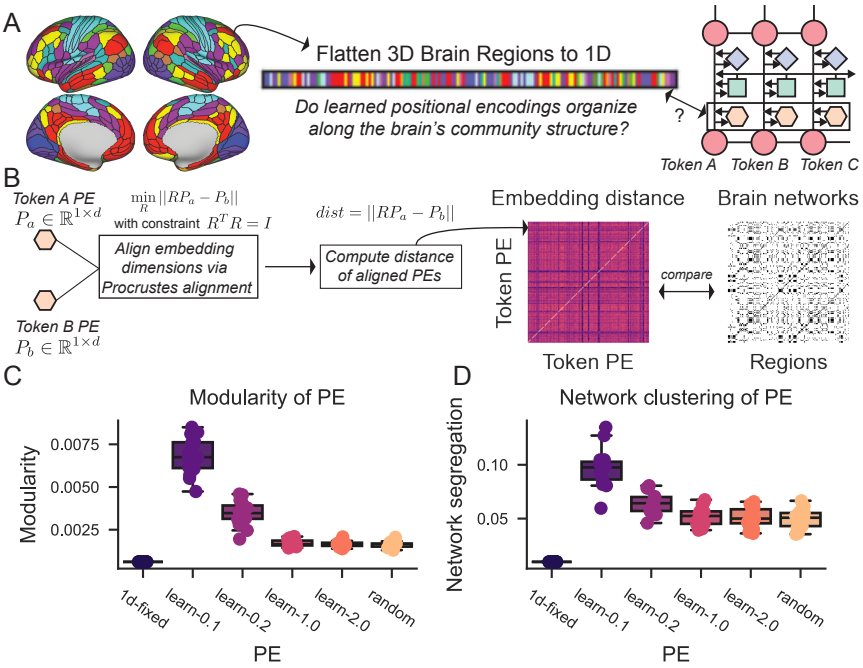


Figure 5: A) We evaluated whether learnable PEs could discover the brain's modular functional network organization. When flattening the brain from a 3D to 1D tensor, the network affiliation of each brain region is heterogenously distributed across a 1D tensor in a disorganized manner. B) To assess whether learnable PEs learned accurate positional brain (network) information, we computed the distance between positional embeddings between every pair of token positions after aligning their embedding dimensions (d) through an orthogonal Procrustes transform. This allowed us to construct a token-by-token embedding distance matrix, which we then compared to brain network organization. C) The modularity of PEs with respect to the brain's network organization. Models that had learnable PEs initialized from small-norm distributions learned a modular PE organization that was consistent with prior neuroscience studies (Ji et al., 2019) D) The network clustering of PEs, which assessed whether PEs of tokens that belong to the same network are closer in space than PEs that do not belong to the same network.

4 DISCUSSION

4.1 RELATED WORK

Recent studies have revealed that the choice of PE can strongly influence transformer generalization (Li et al., 2024; Kazemnejad et al., 2023; Golovneva et al., 2024; Ruoss et al., 2023; McLeish et al., 2024; Shen et al., 2023; Csordás et al., 2022; Ontañón et al., 2022; Zhou et al., 2024; Zhang et al., 2024). However, most of these investigations have been limited to evaluating the generalizability of various PE schemes on 1D string-based tasks (e.g., sequence learning tasks for arithmetic, context-free grammars, or compositional tasks). In contrast, many important problems require the encoding of sequences that are not in 1D, and where position information is non-trivial or not known, which we investigate here. Additionally, work in deep learning theory has provided insight into the impact on model initialization and representation learning, yet focus primarily on simple neural networks rather than transformers (Chizat et al., 2020; Woodworth et al., 2020; Jacot et al., 2020; Kunin et al., 2024; Lippl and Stachenfeld, 2024). In this study, we apply insights from deep learning theory to transformer models to effectively learn (and improve generalization) to nontrivial sequence tasks, such as tasks requiring reasoning in 2D, or tasks in which ground truth position information is organized non-trivially (e.g., 3D neuroscience data).

4.2 LIMITATIONS AND FUTURE DIRECTIONS

We have demonstrated that learnable PEs initialized from small-norm distributions can 1) approximate the ground truth PE, and 2) outperform many commonly-used PEs. However, there remain several limitations of the present study which future studies can explore. First, though we consider the use of the LST (with a 2D organization) a strength of this study due to the visual interpretability of the paradigm's positional information, it is unclear how well this approach will generalize to tasks with an arbitrary number of elements, tasks in which there are dynamic changes in the number of elements (e.g., length generalization problems), or tasks in which there are specific distribution shifts. Second, the current learned PE is limited insofar that the embedding is linearly superimposed on a given token (i.e., $token + pe$). While this makes it difficult to potentially generalize to more complex tasks, a natural future research direction would be to learn nonlinear PE embeddings that allow for PEs to be flexibly generated as a *function* of the token embedding, thereby learning token abstractions (i.e., $pe(token)$). This nonlinear formulation of PE as a function of the token embedding would, in theory, have significantly greater expressive ability, and potentially endow transformers with the ability to recognize more complex formal languages (Delétang et al., 2023; Merrill et al., 2024). Finally, while we have empirically demonstrated the applicability of prior learning theory on neural network initialization to the choice of PE initialization (Chizat et al., 2020; Woodworth et al., 2020), in practice, it remains unclear how the precise choice of small σ will impact generalization performance in practice (e.g., dependency on task and architecural choices, such as number of layers, embedding dimension, etc.). It will be important for future theoretical work to more carefully characterize how the choice of σ may be influenced by transformer architecture and optimizer.

4.3 CONCLUSION

There are many tasks and problems in which it is difficult to know the ground truth position information of input sequences. Examples of such problems include reasoning on parse trees and directed graphs (where node distances and relations are not preserved when flattening into a 1D sequence for parallel transformer processing), or inference on real-world biological datasets in which the ground truth structure is important for prediction yet difficult to know (e.g., 3D neuroscience data explored here, or co-expression of genes in a DNA sequence based on 3D chromatin conformation) (Szabo et al., 2019; Ji et al., 2021). In this study, we sought to understand how to learn position information directly from data using insights from deep learning theory. In particular, we found that an optimally-learned PE 1) outperformed commonly-used PEs, 2) learned attention maps and PE embeddings that were closely aligned to a ground truth PE scheme, and 3) enhanced generalization performance. *Critically, learning an optimal and interpretable PE depended on its initialization in both a synthetic reasoning task and a real-world biological dataset.* We anticipate these results will spur future investigations into the importance and utility of learnable PEs for structured learning and generalization.

5 ETHICS STATEMENT

All human brain data used in this study is publicly available archived data from the Human Connectome Project (www.humanconnectomeproject.org). All participants gave signed, informed consent in accordance with the protocol approved by the local institutional review board. All other methods and results reported in this study adhere to the ICLR Code of Ethics.

6 REPRODUCIBILITY STATEMENT

Code, LST data, and environments associated with reproducing models, results, and figures in this study are provided in the supplementary materials. Human fMRI data is publicly available, though is not included in the supplementary material due to its size.

REFERENCES

Danielle S Bassett and Olaf Sporns. Network neuroscience. *Nature Neuroscience*, 20(3):353, 2017. ISSN 1097-6256. doi: 10.1038/nn.4502.

Yashar Behzadi, Khaled Restom, Joy Liau, and Thomas T Liu. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, 37(1):90–101, 2007. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2007.04.042. URL http://www.sciencedirect.com/science/article/pii/S1053811907003837.

Damian P. Birney, Graeme S. Halford, and Glenda Andrews. Measuring the Influence of Complexity on Relational Reasoning: The Development of the Latin Square Task. *Educational and Psychological Measurement*, 66(1):146–171, February 2006. ISSN 0013-1644. doi: 10.1177/0013164405278570. URL https://doi.org/10.1177/0013164405278570. Publisher: SAGE Publications Inc.

Damian P. Birney, David B. Bowman, Jens F. Beckmann, and Yuan Zhi Seah. Assessment of Processing Capacity: Reasoning in Latin Square Tasks in a Population of Managers. *European Journal of Psychological Assessment*, 28(3):216–226, January 2012. ISSN 1015-5759, 2151-2426. doi: 10.1027/1015-5759/a000146. URL https://econtent.hogrefe.com/doi/10.1027/1015-5759/a000146.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On Lazy Training in Differentiable Programming. *arXiv:1812.07956 [cs, math]*, January 2020. URL http://arxiv.org/abs/1812.07956. arXiv: 1812.07956.

Róbert Csordás, Kazuki Irie, and Jürgen Schmidhuber. The Devil is in the Detail: Simple Tricks Improve Systematic Generalization of Transformers, February 2022. URL http://arxiv.org/abs/2108.12284. arXiv:2108.12284 [cs].

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. Neural Networks and the Chomsky Hierarchy, February 2023. URL http://arxiv.org/abs/2207.02098. arXiv:2207.02098 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

Matthew F Glasser, Stephen M Smith, Daniel S Marcus, Jesper L R Andersson, Edward J Auerbach, Timothy E J Behrens, Timothy S Coalson, Michael P Harms, Mark Jenkinson, Steen Moeller, Emma C Robinson, Stamatios N Sotiropoulos, Junqian Xu, Essa Yacoub, Kamil Ugurbil, and David C Van Essen. The Human Connectome Project's neuroimaging approach. *Nature neuroscience*, 19(9):1175–87, 2016. ISSN 1546-1726. doi: 10.1038/nn.4361. URL http://www.nature.com/neuro/journal/v19/n9/pdf/nn.4361.pdf%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/27571196.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Contextual Position Encoding: Learning to Count What's Important, May 2024. URL http://arxiv.org/abs/2405.18719. arXiv:2405.18719 [cs].

Graeme S. Halford, William H. Wilson, and Steven Phillips. Relational complexity metric is effective when assessments are based on actual cognitive processes. *Behavioral and Brain Sciences*, 21(6):848–860, December 1998. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X98401766. URL https://www.cambridge.org/core/product/identifier/S0140525X98401766/type/journal_article.

Johanna Hartung, Benjamin Goecke, Ulrich Schroeders, Florian Schmitz, and Oliver Wilhelm. Latin square tasks: A multi-study evaluation. *Intelligence*, 94:101683, September 2022. ISSN 0160-2896. doi: 10.1016/j.intell.2022.101683. URL https://www.sciencedirect.com/science/article/pii/S0160289622000642.

Luke J. Hearne, Damian P. Birney, Luca Cocchi, and Jason B. Mattingley. The Latin Square Task as a measure of relational reasoning: A replication and assessment of reliability. *European Journal of Psychological Assessment*, 36(2):296, 2020. Publisher: Hogrefe Publishing.

Takuya Ito, Scott L. Brincat, Markus Siegel, Ravi D. Mill, Biyu J. He, Earl K. Miller, Horacio G. Rotstein, and Michael W. Cole. Task-evoked activity quenches neural correlations and variability across cortical areas. *PLOS Computational Biology*, 16(8):e1007983, August 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007983. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007983. Publisher: Public Library of Science.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *arXiv:1806.07572 [cs, math, stat]*, February 2020. URL http://arxiv.org/abs/1806.07572. arXiv: 1806.07572.

Jie Lisa Ji, Marjolein Spronk, Kaustubh Kulkarni, Grega Repovš, Alan Anticevic, and Michael W Cole. Mapping the human brain's cortical-subcortical functional network organization. *NeuroImage*, 185:35–57, 2019. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2018.10.006. URL http://www.sciencedirect.com/science/article/pii/S1053811918319657.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, August 2021. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btab083. URL https://academic.oup.com/bioinformatics/article/37/15/2112/6128680.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The Impact of Positional Encoding on Length Generalization in Transformers, November 2023. URL http://arxiv.org/abs/2305.19466. arXiv:2305.19466 [cs].

Daniel Kunin, Allan Raventós, Clémentine Dominé, Feng Chen, David Klindt, Andrew Saxe, and Surya Ganguli. Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning, June 2024. URL http://arxiv.org/abs/2406.06158. arXiv:2406.06158 [cs, stat].

Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, Sanjiv Kumar, and Srinadh Bhojanapalli. Functional Interpolation for Relative Positions Improves Long Context Transformers, March 2024. URL http://arxiv.org/abs/2310.04418. arXiv:2310.04418 [cs].

Samuel Lippl and Kim Stachenfeld. When does compositional structure yield compositional generalization? A kernel theory, May 2024. URL http://arxiv.org/abs/2405.16391. arXiv:2405.16391 [cs, q-bio].

Sean McLeish, Arpit Bansal, Alex Stein, Neel Jain, John Kirchenbauer, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, Jonas Geiping, Avi Schwarzschild, and Tom Goldstein. Transformers Can Do Arithmetic with the Right Embeddings, May 2024. URL http://arxiv.org/abs/2405.17399. arXiv:2405.17399 [cs].

William Merrill, Jackson Petty, and Ashish Sabharwal. The Illusion of State in State-Space Models, April 2024. URL http://arxiv.org/abs/2404.08819. arXiv:2404.08819 [cs].

Santiago Ontañón, Joshua Ainslie, Vaclav Cvicek, and Zachary Fisher. Making Transformers Solve Compositional Tasks, March 2022. URL http://arxiv.org/abs/2108.04378. arXiv:2108.04378 [cs].

Jonathan D. Power, Alexander L. Cohen, Steven M. Nelson, Gagan S. Wig, Kelly Anne Barnes, Jessica a. Church, Alecia C. Vogel, Timothy O. Laumann, Fran M. Miezin, Bradley L. Schlaggar, and Steven E. Petersen. Functional Network Organization of the Human Brain. *Neuron*, 72(4): 665–678, 2011. ISSN 08966273. doi: 10.1016/j.neuron.2011.09.006. URL http://dx.doi.org/10.1016/j.neuron.2011.09.006.

Ofir Press, Noah A Smith, and Mike Lewis. TRAIN SHORT, TEST LONG: ATTENTION WITH LINEAR BIASES ENABLES INPUT LENGTH EXTRAPOLATION. 2022.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, July 2020. URL http://arxiv.org/abs/1910.10683. arXiv:1910.10683 [cs, stat].

Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069, 2010. ISSN 10538119. doi: 10.1016/j.neuroimage.2009.10.003. URL http://dx.doi.org/10.1016/j.neuroimage.2009.10.003.

Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. Randomized Positional Encodings Boost Length Generalization of Transformers, May 2023. URL http://arxiv.org/abs/2305.16843. arXiv:2305.16843 [cs, stat].

Alexander Schaefer, Ru Kong, Evan M. Gordon, Timothy O. Laumann, Xi-Nian Zuo, Avram J. Holmes, Simon B. Eickhoff, and BT Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, 2018. Publisher: Oxford University Press.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-Attention with Relative Position Representations, April 2018. URL http://arxiv.org/abs/1803.02155. arXiv:1803.02155 [cs].

Ruoqi Shen, Sébastien Bubeck, Ronen Eldan, Yin Tat Lee, Yuanzhi Li, and Yi Zhang. Positional Description Matters for Transformers Arithmetic, November 2023. URL http://arxiv.org/abs/2311.14737. arXiv:2311.14737 [cs].

Stephen M. Smith, Christian F. Beckmann, Jesper Andersson, Edward J. Auerbach, Janine Bijsterbosch, Gwenaëlle Douaud, Eugene Duff, David A. Feinberg, Ludovica Griffanti, Michael P. Harms, Michael Kelly, Timothy Laumann, Karla L. Miller, Steen Moeller, Steve Petersen, Jonathan Power, Gholamreza Salimi-Khorshidi, Abraham Z. Snyder, An T. Vu, Mark W. Woolrich, Junqian Xu, Essa Yacoub, Kamil Uğurbil, David C. Van Essen, and Matthew F. Glasser. Resting-state fMRI in the Human Connectome Project. *NeuroImage*, 80:144–168, October 2013. ISSN 10538119. doi: 10.1016/j.neuroimage.2013.05.039.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding, August 2022. URL http://arxiv.org/abs/2104.09864. arXiv:2104.09864 [cs].

Quentin Szabo, Frédéric Bantignies, and Giacomo Cavalli. Principles of genome folding into topologically associating domains. *Science Advances*, 5(4):eaaw1668, April 2019. doi: 10.1126/sciadv.aaw1668. URL https://www.science.org/doi/10.1126/sciadv.aaw1668. Publisher: American Association for the Advancement of Science.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy E J Behrens, Essa Yacoub, Kamil Ugurbil, and WU-Minn HCP Consortium. The WU-Minn Human Connectome Project: an overview. *NeuroImage*, 80:62–79, October 2013. ISSN 1095-9572. doi: 10.1016/j.neuroimage.

2013.05.041. URL http://www.ncbi.nlm.nih.gov/pubmed/23684880http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3724347.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Bertha Vázquez-Rodríguez, Laura E. Suárez, Ross D. Markello, Golia Shafiei, Casey Paquola, Patric Hagmann, Martijn P. van den Heuvel, Boris C. Bernhardt, R. Nathan Spreng, and Bratislav Misic. Gradients of structure–function tethering across neocortex. *Proceedings of the National Academy of Sciences*, 116(42):21219–21227, October 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1903403116. URL https://www.pnas.org/content/116/42/21219. Publisher: National Academy of Sciences Section: PNAS Plus.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and Rich Regimes in Overparametrized Models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, July 2020. URL http://proceedings.mlr.press/v125/woodworth20a.html. ISSN: 2640-3498.

B. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zollei, J. R. Polimeni, B. Fischl, H. Liu, and R. L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, September 2011. ISSN 0022-3077, 1522-1598. doi: 10.1152/jn.00338.2011.

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, and Zhi-Qin John Xu. Initialization is Critical to Whether Transformers Fit Composite Functions by Inference or Memorizing, May 2024. URL http://arxiv.org/abs/2405.05409. arXiv:2405.05409 [cs].

Yongchao Zhou, Uri Alon, Xinyun Chen, Xuezhi Wang, Rishabh Agarwal, and Denny Zhou. Transformers Can Achieve Length Generalization But Not Robustly, February 2024. URL http://arxiv.org/abs/2402.09371. arXiv:2402.09371 [cs].

A APPENDIX / SUPPLEMENTAL MATERIAL

A.1 POSITIONAL ENCODING DEFINITIONS

Below, we provide the formal definitions for the common PEs that we evaluated learnable PEs against: `1d-fixed`, `2d-fixed`, `relative`, and `rope`.

1d-fixed (Vaswani et al., 2017). 1D absolute PEs were defined from Vaswani et al. (2017) (Vaswani et al., 2017). For a given position pos (here $1 \leq pos \leq 16$, for elements in the LST grid), we define

$$PE_{(pos,2i)} = sin(\frac{pos}{10000^{2i/d_{model}}})$$

$$PE_{(pos,2i+1)} = cos(\frac{pos}{10000^{2i/d_{model}}})$$

where i is the embedding dimension, and d_{model} is the dimensionality of the embedding vector. Note that $PE_{(pos,2i)}$ (sines) is reserved for even embedding dimensions, and $PE_{(pos,2i)}$ (cosines) is reserved for odd embedding dimensions.

2d-fixed. 2D absolute PEs were a 2D generalization of `1d-fixed` (Vaswani et al., 2017). The primary distinction is that rather than $1 < pos < 16$, there are two position variables, $1 \leq pos_w \leq 4$ and $1 \leq pos_h \leq 4$ (for width and height of grid). Positional encoding for a row w is defined by

$$PE_{(pos_w,2i)} = sin(\frac{pos_w}{10000^{2i/d_{model}}})$$

$$PE_{(pos_w,2i+1)} = cos(\frac{pos_w}{10000^{2i/d_{model}}})$$

Positional encoding for a column h is defined by

$$PE_{(pos_h,2i)} = sin(\frac{pos_h}{10000^{2i/d_{model}}})$$

$$PE_{(pos_h,2i+1)} = cos(\frac{pos_h}{10000^{2i/d_{model}}})$$

For a 2D PE encoding, half the embedding dimensionality is reserved for encoding rows; the other half of the embedding dimensionality is reserved for encoding columns. Thus, for $d_{model} = 160$, embedding dimensions 0-79 are reserved for encoding rows. Embedding dimensions 80-159 are reserved for encoding columns.

relative (Shaw et al., 2018). `relative` PE modifies standard self-attention to incorporate the relative positions of tokens. This implies that calculation of PE is wrapped within the self-attention module. The relative position embedding parameter between a token at position i and j is a_{j-1}. In brief, self attention is then modified to include relative position information by modifying attention between tokens i and j as

$$e_{ij} = \frac{x_i W^Q (x_j W^K)^T + x_i W^Q (a_{ij}^K)^T}{\sqrt{d_z}}$$

where x_i and x_j are the embeddings for tokens i and j, and W^Q andd W^K are the query and key matrices, respectively. Additional details can be found in (Shaw et al., 2018).

rope (Su et al., 2022). `rope` applies a rotation to the token embeddings based on their positions in a higher dimensional space. For a token at position p with an embedding x, let $x = [x_1, x_2, ..., x_d]$, where d is even. Then, for each pair of dimensions, apply the rotation

$$\begin{pmatrix} \hat{x}_{2k} \\ \hat{x}_{2k+1} \end{pmatrix} = \begin{pmatrix} cos(\theta_p) & -sin(\theta_p) \\ sin(\theta_p) & cos(\theta_p) \end{pmatrix} \begin{pmatrix} x_{2k} \\ x_{2k+1} \end{pmatrix}$$

with $\theta_p = \frac{p}{10000^{2k/d}}$. The resulting embedding is the concatenation of rotated pairs. Additional details can be found in the original paper (Su et al., 2022).

A.2 FMRI DATA AND PREPROCESSING

Data were previously collected as part of the Human Connectome Project and made publicly available (Van Essen et al., 2013). All participants gave signed, informed consent in accordance with the protocol approved by the Washington University institutional review board. Whole-brain multiband echo-planar imaging acquisitions were collected on a 32-channel head coil on a modified 3T Siemens Skyra with TR = 720 ms, TE = 33.1 ms, flip angle = 52°, Bandwidth = 2,290 Hz/Px, in-plane FOV = 208x180 mm, 72 slices, 2.0 mm isotropic voxels, with a multiband acceleration factor of 8. Data were collected across two days, with the first two resting-state fMRI sessions collected on the first day, and another two sessions collected on the second day. Each resting-state fMRI session lasted 14.4 minutes. Additional details on imaging sessions can be found in Smith et al. (2013).

Data were provided in a minimally preprocessed format. We performed additional preprocessing steps in accordance with Ito et al. (2020), which we paraphrase below. We first parcellated minimally preprocessed data into 360 brain regions using the Glasser et al. (2016). In addition, we removed the first five frames of each run, de-meaning and de-trending the timeseries, and performing nuisance regression on the minimally preprocessed data. Nuisance regression included removing motion signals and physiological noise. Six primary motion parameters were included, along with their derivatives and quadratic timeseries. Physiological noise was modeled using aCompCor on the timeseries extracted from the white matter and ventricles (Behzadi et al., 2007). For aCompCor, the first 5 principal components from the white matter and ventricles were extracted separately and included in the nuisance regression. We also included the derivatives of each of those components, and the quadratics of all noise regressors. In total, the nuisance regression modeled contained 64 nuisance regressors.

A.3 MODULARITY AND CLUSTERING IN FMRI BRAIN DATA

Modularity and network clustering (segregation) are common measures in the neuroscience literature, particularly when applied to fMRI data (Rubinov and Sporns, 2010). We adopted these measures to calculate the modularity and network clustering of learned PE parameters. The distance d between PEs was scaled between 0 and 1, and we calculated the complement $(1 - d)$ such that higher values indicated two PEs were closer. Both modularity and network clustering were calculated with respect to the predefined network partition (Fig. 4; Ji et al. (2019)). Modularity Q^W of the learned, PE distance matrix W was calculated as

$$Q^W = \frac{1}{l^W} \sum_{i,j \in N} \left[W_{ij} - \frac{k_i^W k_j^W}{l^W} \right] \delta_{m_i, m_j}$$

where l^W is the sum of all weights in W, N are tokens (brain regions), W_{ij} is the distance between token i and token j, k_i^W is the weighted degree of token i, m_i is the module containing node i, and $\delta_{m_i,m_j} = 1$ if $m_i = m_j$ and 0 otherwise (as determined by the network partition from Ji et al. (2019)).

Network clustering C is measured as the difference in mean within module and across module distances, as a proportion of the within-module distance

$$C = \frac{1}{|M|} \sum_{m \in M} \left[\frac{\bar{W}_{m_{in}} - \bar{W}_{m_{out}}}{\bar{W}_{m_{in}}} \right]$$

where M is the full set of modules (networks), $\bar{W}_{m_{in}}$ is the mean within-module distance, and $\bar{W}_{m_{out}}$ is the across-module distance.

Table A2: Training and generalization performance for all learnable PE models initialized with different $\mathcal{N}(0, \sigma)$ (no regularization). This is the corresponding data table for Fig. **??**.

PE	Training acc	Training SD	Validation acc	Validation SD
0.100	1.000	0.000	0.925	0.083
0.200	**1.000**	**0.000**	**0.956**	**0.039**
0.300	1.000	0.000	0.946	0.022
0.400	1.000	0.000	0.929	0.048
0.500	1.000	0.000	0.946	0.025
0.600	1.000	0.000	0.929	0.027
0.700	1.000	0.000	0.926	0.027
0.800	1.000	0.000	0.916	0.048
0.900	1.000	0.000	0.896	0.060
1.000	1.000	0.000	0.894	0.043
1.100	1.000	0.000	0.835	0.101
1.200	1.000	0.000	0.756	0.141
1.300	1.000	0.000	0.702	0.126
1.400	1.000	0.000	0.654	0.161
1.500	1.000	0.000	0.616	0.208
1.600	1.000	0.000	0.582	0.168
1.700	1.000	0.000	0.480	0.167
1.800	1.000	0.000	0.378	0.153
1.900	1.000	0.000	0.406	0.161
2.000	1.000	0.000	0.377	0.171

Table A3: Attention map similarity between learnable PEs and the ground truth `2d-fixed`. Corresponding data table for Fig. 3.

σ	Cosine	SD
0.100	0.836	0.018
0.200	0.838	0.016
0.300	0.808	0.013
0.400	0.779	0.013
0.500	0.748	0.014
0.600	0.722	0.016
0.700	0.715	0.016
0.800	0.710	0.016
0.900	0.694	0.015
1.000	0.685	0.015
1.100	0.697	0.015
1.200	0.692	0.017
1.300	0.685	0.016
1.400	0.672	0.016
1.500	0.657	0.017
1.600	0.647	0.022
1.700	0.642	0.016
1.800	0.624	0.021
1.900	0.606	0.023
2.000	0.577	0.019

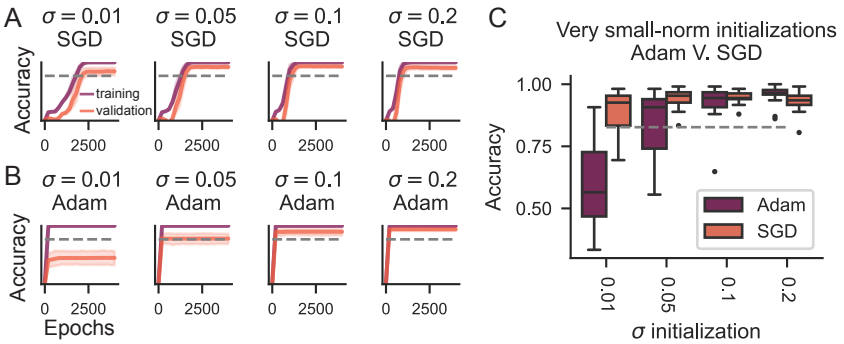


Figure A6: Different optimizers (Adam, SGD) can lead to different generalization performances in very small-norm PE initializations. We noticed that for models that were initialized with very small-norm PEs, generalization performance was reduced, counter to the theoretical claims in the NTK theory. However, we realized that this could be attributed to variability in the adaptive learning rates inherent to the Adam optimizer. Thus, we evaluated whether using vanilla SGD (learning rate=0.001) would ameliorate these reduced accuracies. Indeed, we found that for very small-norm initializations, SGD tended to ameliorate the reduced generalization effects observed with Adam. A) Training trajectories for $\sigma \in 0.01, 0.05, 0.1, 0.2$ using vanilla SGD (learning rate=0.001). Note that the slow training is significantly more obvious for small-norm initializations with SGD, consistent with theory. B) Corresponding training trajectories using Adam. C) A clear generalization discrepancy emerges when using Adam vs. SGD.

Table A4: Performance differences across different optimizers (Adam, SGD) in very small-norm PE initializations (without regularization). Corresponding data table for Fig. A6C.

σ	Adam acc	Adam sd	SGD acc	SGD sd
0.010	0.588	0.186	0.885	0.101
0.050	0.836	0.140	0.941	0.042
0.100	0.925	0.083	0.947	0.027
0.200	0.956	0.039	0.930	0.044

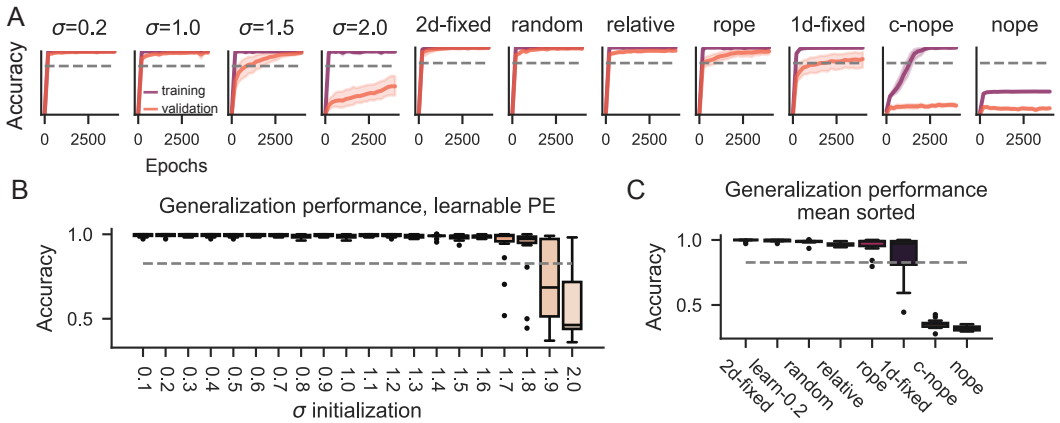


Figure A7: Generalization performance of learnable and common PEs with weight decay (0.1). To assess the impact of learnable PEs in a more realistic setting, we measured the effect of L2 regularization with a weight decay parameter of 0.1. In general, weight decay improves the generalization of learnable PE models across initializations, except for particularly high σ. Importantly, learnable PEs with weight decay become nearly indistinguishable to the ground-truth PE. Nevertheless, the general pattern remains for learnable PE models: the smaller the σ, the greater the generalization. A) Training and validation trajectories across training. B) Generalization performance for learnable PE models with weight decay = 0.1. C) Generalization performance for common PE models with weight decay = 0.1.

Table A5: Training and validation performance of learnable PEs with weight decay (0.1). Corresponding data table for Fig. A7B.

σ	Validation acc	Validation SD	Training acc	Training SD
0.100	0.996	0.006	1.000	0.000
0.200	0.995	0.007	1.000	0.000
0.300	0.994	0.008	1.000	0.000
0.400	0.994	0.007	1.000	0.001
0.500	0.993	0.007	0.999	0.002
0.600	0.993	0.007	1.000	0.001
0.700	0.993	0.009	0.999	0.003
0.800	0.993	0.009	0.999	0.003
0.900	0.993	0.008	1.000	0.001
1.000	0.993	0.008	1.000	0.001
1.100	0.990	0.008	0.999	0.002
1.200	0.989	0.010	1.000	0.001
1.300	0.989	0.012	1.000	0.000
1.400	0.987	0.010	0.999	0.002
1.500	0.985	0.013	0.999	0.002
1.600	0.983	0.018	0.999	0.002
1.700	0.929	0.138	1.000	0.001
1.800	0.900	0.180	1.000	0.001
1.900	0.707	0.238	0.992	0.024
2.000	0.586	0.238	0.996	0.012

Table A6: Training and validation performance of common PEs with weight decay (0.1). Corresponding data table for Fig. A7C.

PE	Validation acc	Validation SD	Training acc	Training SD
2d-fixed	0.997	0.008	0.999	0.001
learn-0.2	0.994	0.008	1.000	0.000
random	0.987	0.016	0.999	0.003
relative	0.965	0.012	0.999	0.004
rope	0.959	0.060	1.000	0.001
1d-fixed	0.872	0.166	0.998	0.001
c-nope	0.352	0.036	0.999	0.003
nope	0.322	0.017	0.508	0.002

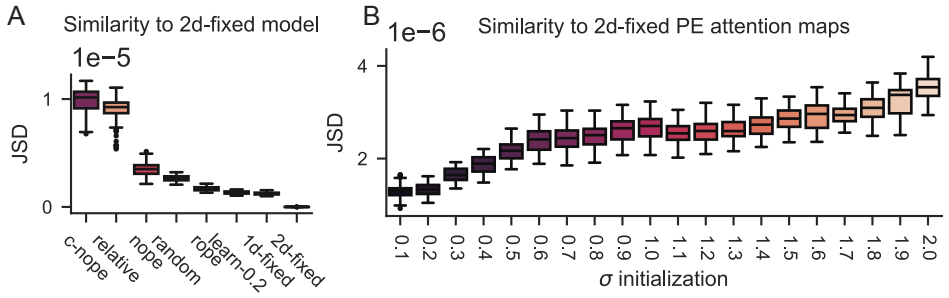
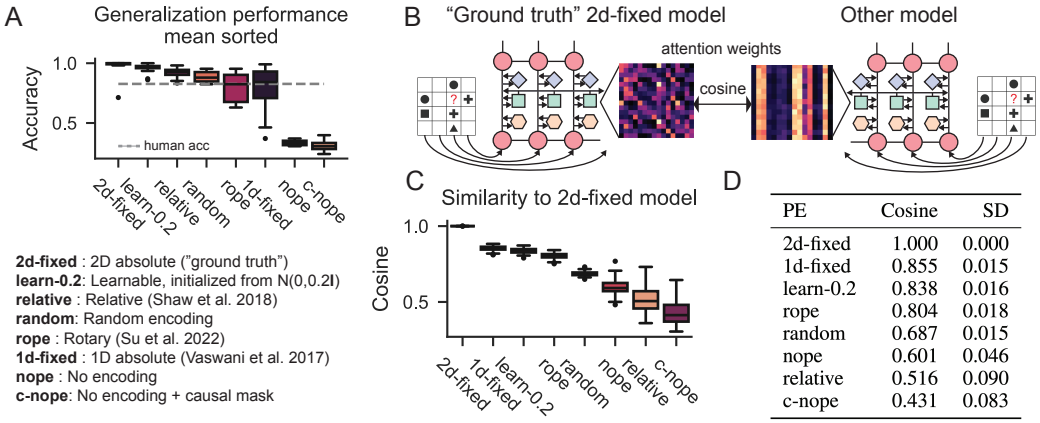


Figure A8: We computed the Jensen-Shannon Divergence of the attention weights for every learnable PE model and the ground truth `2d-fixed` model. This Figure is a comparable analysis to Fig. 3, but using the Jensen-Shannon Divergence distance metric applied to attention maps (instead of cosine similarity).



PE	Cosine	SD
2d-fixed	1.000	0.000
1d-fixed	0.855	0.015
learn-0.2	0.838	0.016
rope	0.804	0.018
random	0.687	0.015
nope	0.601	0.046
relative	0.516	0.090
c-nope	0.431	0.083

Figure A9: Comparing top-performing learnable PE models (`learn-0.2`) with commonly-used PE schemes. A) We compare the generalization performance of `learn-0.2` model with models using common PEs. Notably, we included a `2d-fixed` PE as the "ground truth", since it obeyed the 2D organization of the LST task. The `learn-0.2` model outperformed all other PE models. B) We compared the attention maps of each model to those derived from the "ground truth" `2d-fixed` model. C,D) We found that aside from the `1d-fixed` model (which does not generalize well), `learn-0.2` learned the closest attention map to the `2d-fixed` model. This was expected, since the `2d-fixed` and `1d-fixed` PEs are highly similar by design (the baseline cosine similarity between the two schemes is 0.72). In contrast, the `learn-0.2` learned an attention map that was highly similar to the `2d-fixed` model, despite having no similarity to the `2d-fixed` PE scheme at initialization (cosine at initialization = 0.00). (We also note that the `rope` model had high baseline similarity to the `2d-fixed`, since by construction, a component of the `rope` encoding is highly similar to the `1d-fixed` PE scheme.) Thus, we found that despite having no prior bias towards the `2d-fixed` PE scheme, a small-norm initialized learnable PE is capable of learning an attention map that approximates an attention map derived from the ground truth PE.

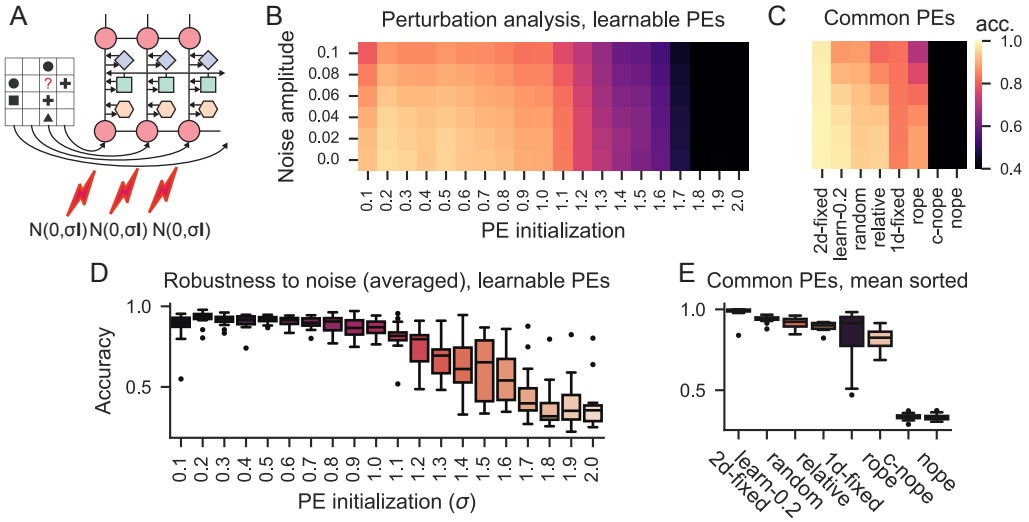


Figure A10: A) We injected noise with different amplitudes into each token embedding, and assessed downstream generalization performance. B) Generalization performance with noisy inputs across all initializations for learnable PEs, and C) common PEs. D) Average performance across noisy inputs (collapsing across rows in B). E) Besides the ground truth model (`2d-fixed`), the `learn-0.2` model was most robust to perturbation compared to common PEs.

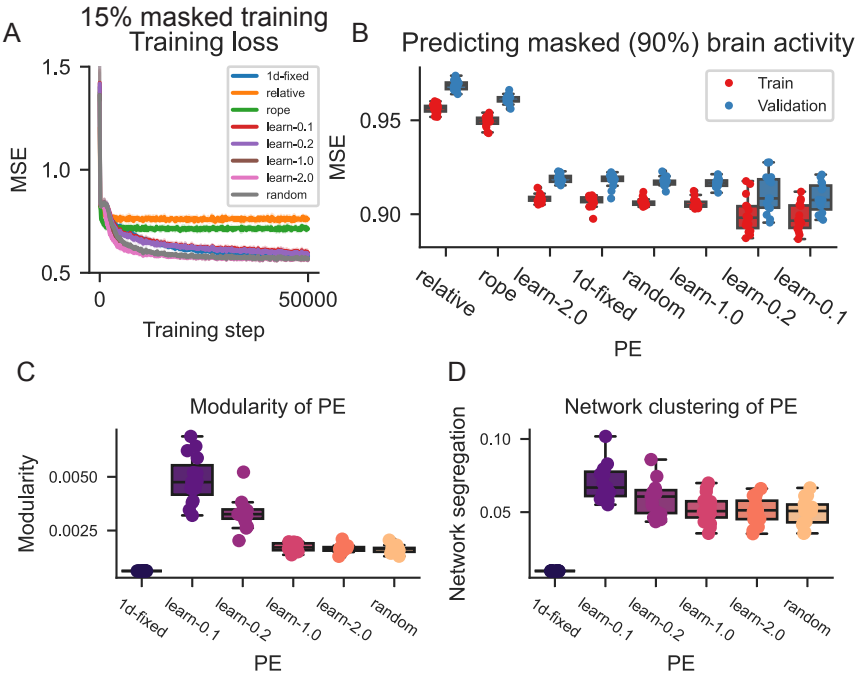


Figure A11: Training and evaluating models on fMRI data with 15% masked pretraining. A) Training trajectory for each model with 15% masked pretraining. B) MSE of each model on training and testing datasets at the end of training (after 50k training steps) predicting on 90% masked input. (X-axis is sorted by highest to lowest MSE.) Consistent with results in the main text, models with a learnable PE parameter (initialized from a small-norm distribution) achieved the lowest generalization MSE. C) The modularity of PEs with respect to the brain's network organization (analogous analysis to Fig. 5C). D) The network clustering of PEs within a model, which assessed whether PEs of tokens that belong to the same network are closer in space than PEs that do not belong to the same network (analogous analysis to Fig. 5D). Consistent with results in the main text, models with a learnable PE parameter initialized from a small-norm distribution learned a similar network modularity/clustering consistent with the brain's known network organization.

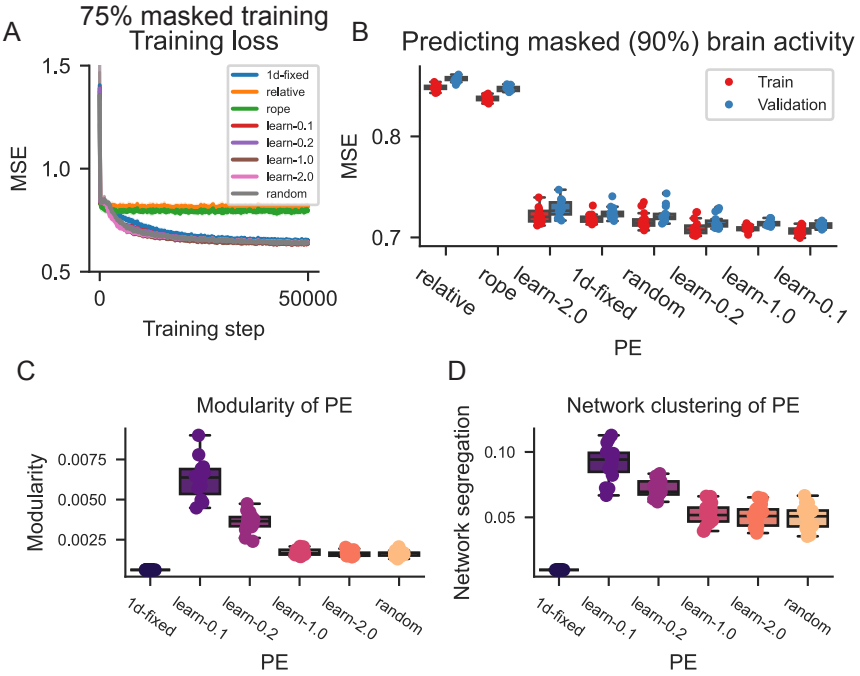


Figure A12: Training and evaluating models on fMRI data with 75% masked pretraining. A) Training trajectory for each model with 75% masked pretraining. B) MSE of each model on training and testing datasets at the end of training (after 50k training steps) predicting on 90% masked input. (X-axis is sorted by highest to lowest MSE.) Consistent with results in the main text, models with a learnable PE parameter (initialized from a small-norm distribution) achieved the lowest generalization MSE. C) The modularity of PEs with respect to the brain's network organization (analogous analysis to Fig. 5C). D) The network clustering of PEs within a model, which assessed whether PEs of tokens that belong to the same network are closer in space than PEs that do not belong to the same network (analogous analysis to Fig. 5D). Consistent with results in the main text, models with a learnable PE parameter initialized from a small-norm distribution learned a similar network modularity/clustering consistent with the brain's known network organization.

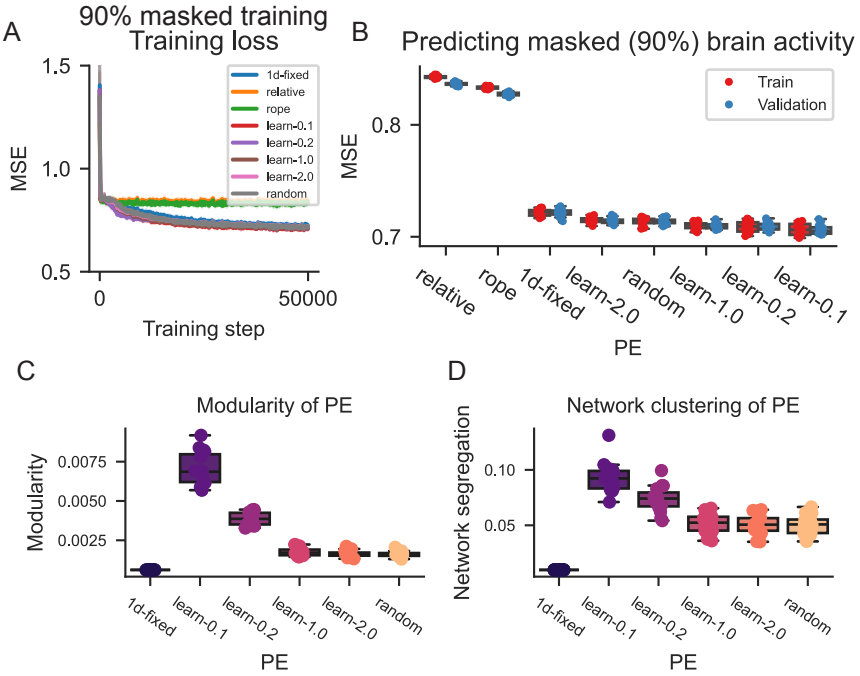
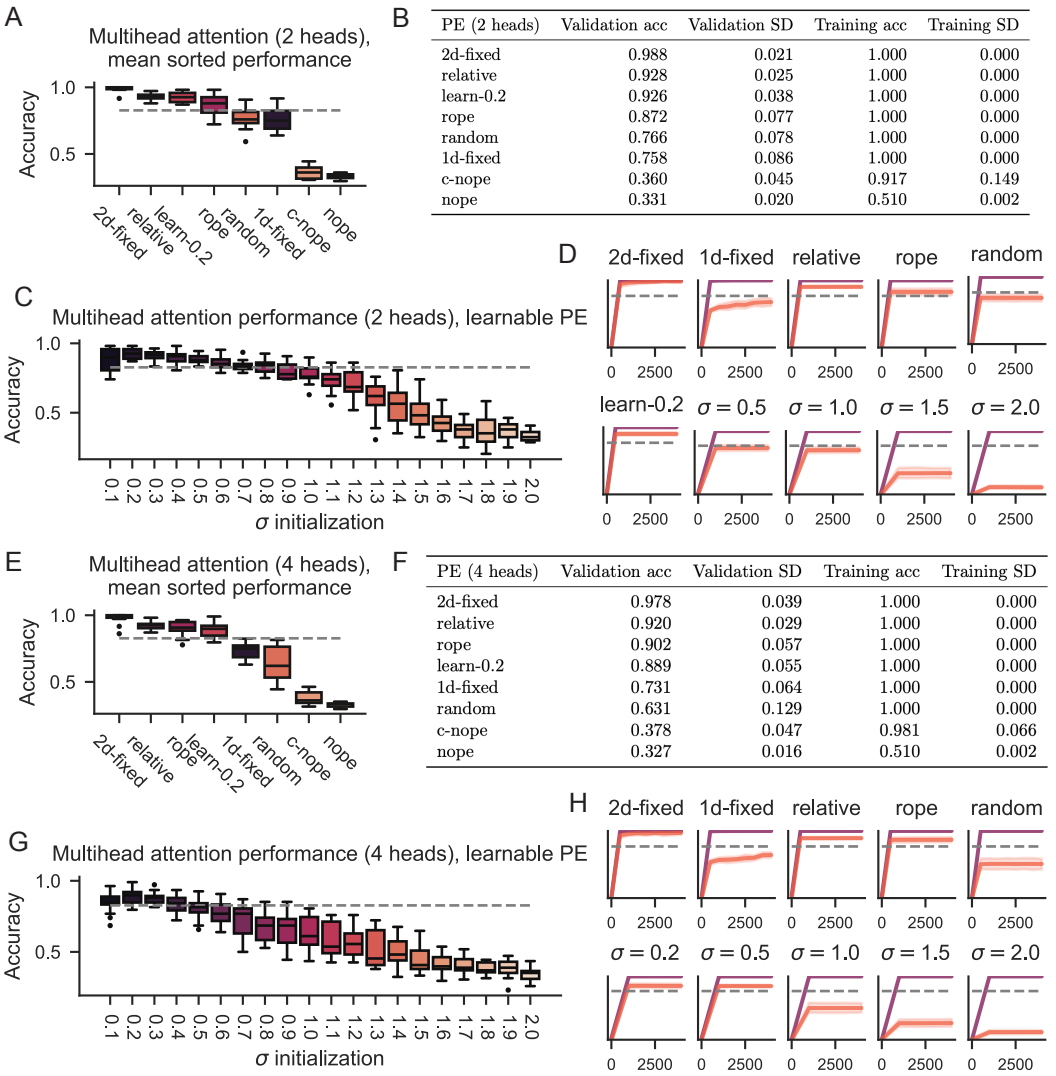


Figure A13: Training and evaluating models on fMRI data with 90% masked pretraining. A) Training trajectory for each model with 90% masked pretraining. B) MSE of each model on training and testing datasets at the end of training (after 50k training steps) predicting on 90% masked input. (X-axis is sorted by highest to lowest MSE.) Consistent with results in the main text, models with a learnable PE parameter (initialized from a small-norm distribution) achieved the lowest generalization MSE. C) The modularity of PEs with respect to the brain's network organization (analogous analysis to Fig. 5C). D) The network clustering of PEs within a model, which assessed whether PEs of tokens that belong to the same network are closer in space than PEs that do not belong to the same network (analogous analysis to Fig. 5D). Consistent with results in the main text, models with a learnable PE parameter initialized from a small-norm distribution learned a similar network modularity/clustering consistent with the brain's known network organization.



PE (2 heads)	Validation acc	Validation SD	Training acc	Training SD
2d-fixed	0.988	0.021	1.000	0.000
relative	0.928	0.025	1.000	0.000
learn-0.2	0.926	0.038	1.000	0.000
rope	0.872	0.077	1.000	0.000
random	0.766	0.078	1.000	0.000
1d-fixed	0.758	0.086	1.000	0.000
c-nope	0.360	0.045	0.917	0.149
nope	0.331	0.020	0.510	0.002

PE (4 heads)	Validation acc	Validation SD	Training acc	Training SD
2d-fixed	0.978	0.039	1.000	0.000
relative	0.920	0.029	1.000	0.000
rope	0.902	0.057	1.000	0.000
learn-0.2	0.889	0.055	1.000	0.000
1d-fixed	0.731	0.064	1.000	0.000
random	0.631	0.129	1.000	0.000
c-nope	0.378	0.047	0.981	0.066
nope	0.327	0.016	0.510	0.002

Figure A14: Model performances when incorporating multihead (2 and 4) attention mechanisms. Overall, we find that adding attention heads tends to reduce generalization performance across the board. The only architecturein which we see improvements are models with `rope` PE. Nevertheless, despite their improvements, `rope` models still do not outperform models with `relative` PEs, nor do they outperform `learn-0.2` models with a single attention head (single head `learn-0.2` performance = 95.6%; Table 1). We also notably see that `learn-0.2` models tend to degrade in their performance, likely due to the increase in free parameters. A,B) Performance of models with 2 attention heads. C) Performance and of models with learnable PEs (2 attention heads). D) Training trajectories for example model architectures (2 attention heads). E,F) Performance of models with 4 attention heads. G) Performance of models with learnable PEs (4 attention heads). H) Training trajectories for example model architectures (4 attention heads).

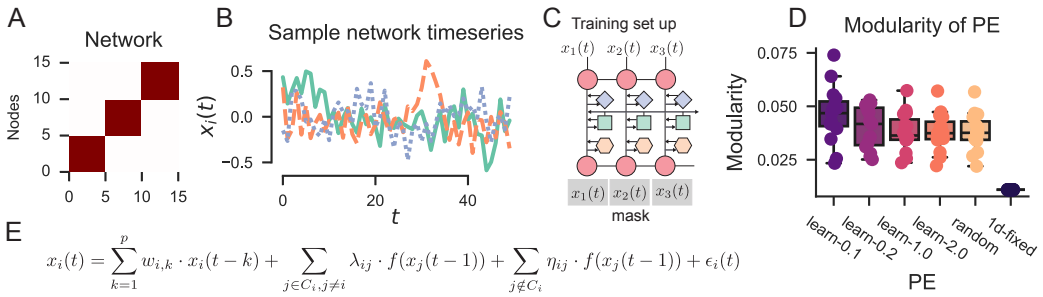


Figure A15: Recovering community structure in PEs using a nonlinear multivariate autoregressive model (NMAR). We sought to verify that recovery of PEs could be achieved in nonlinear stochastic systems. We thus experimented with a toy, NMAR system with 15 nodes divided into 3 modules (which determine module / ground truth PE). Importantly, intra-module nodes interacted more strongly with each other than inter-module nodes. Overall, we found that even in models with explicit nonlinear relationships between tokens/nodes, we can recover PE structure using learnable PEs initialized with small-norms. A) Network structure with 15 nodes divided into 3 modules. B) Example time series from 3 randomly selected nodes generated from the NMAR. Note that when generating the training dataset, we generated data with 20000 time points. C) The training set up was analogous to our empirical fMRI experiments, where contemporaneous timepoints were masked, and the model was trained to predict masked data. D) The network modularity of PEs, which was measured in relation to the "ground truth" modules, was highest in learnable PE models initialized with small norm (`learn-0.1`). This implied that PE embeddings of small-norm learnable PEs learned an accurate representation of the model's network structure. E) The equation governing the timeseries generation for node $x_i(t)$. The number of time lags p was chosen to be 3, and $w_{i,k} \sim \mathcal{U}(0.2, 0.5)$. C refers to the module, and the other remaining parameters were chosen as follows: $\lambda_{ij} \sim \mathcal{U}(0.02, 0.2)$, $\eta_{ij} \sim \mathcal{U}(0.005, 0.01)$, $\epsilon_i(t) \sim \mathcal{N}(0, 0.2)$, and $f(x) = \sin(x)$.