Supplementary Results for "Automated Eloquent Cortex Localization in Brain Tumor Patients Using Multi-task Graph Neural Networks"

1 Introduction

In this supplementary section, we include experiments that separately address potential confounders to our analysis, model optimization, and performance of our MT-GNN method on healthy HCP data. Our first experiment is to assess model performance while accounting for certain potential confounders such as language laterality, tumor size, age, and gender. We include this experiment to observe any correlation between the confounders and model performance. Regarding model optimization, we explore with different architectures to maximize overfitting to our training data. We then explore the effect of adding dropout to this overfit model and observe validation accuracy. We include this experiment to show the robustness and generalization capabilities of the main model presented in the manuscript. Regarding the healthy HCP data, we observe the performance of our MT-GNN method on healthy HCP data without synthetic tumors. Our motivation for including this experiment is to observe how well our model does on the uncontaminated HCP data, to act as a realistic "ground truth" for model performance before adding the simulated tumor.

2 Confounder analysis

We assess model performance (both AUC and TPR for all four tasks) against four separate confounders, language laterality, tumor size, age, and gender to observe if there is a strong correlation between performance and the confounding variables. Here, laterality refers to a quantitative measure between -1 and 1 that describes handedness of the subject, and the same 10 fold-CV evaluation was used. The performance metrics are based on the same repeated 10-fold CV splits used in the paper. Likewise, we separated the testing performance based on gender and used a t-test to determine significance of model performance on men vs. women. We include the gender analysis table in and correlation plots with associated lines of best fit and p-values for the quantitative confounders in Section 2. STable 1 shows the gender analysis performance, where each task has a p-value greater than 0.05, indicating no significant change in performance.

STable 1: Gender confounder analysis.

Task	Male TPR	Male AUC	Female TPR	Female AUC	P-value
Language	$.74 \pm .013$	$.75 \pm .031$	$.76 \pm .016$	$.76 \pm .026$	.52
Finger	$.85 \pm .017$	$.84 \pm .029$	$.84 \pm .017$	$.83 \pm .019$	.42
Foot	$.82 \pm .031$	$.79 \pm .032$	$.81 \pm .039$	$.83 \pm .019$	.37
Tongue	$.80 \pm .019$	$.81 \pm .027$	$.81 \pm .018$	$.79 \pm .021$	.33

SFigs. 1-2 shows 8 separate plots of model AUC and TPR performance across all four tasks using tumor size and age as the respective confounding variables. SFig. 3 shows the AUC and TPR for just the language task against language laterality. The p-values were calculated from the correlation coefficient between the confounder and model performance, where the line of best fit is shown in red. As shown by $p > 0.05$, there is no significant correlation between any of the confounders and model performance for any of the four tasks.

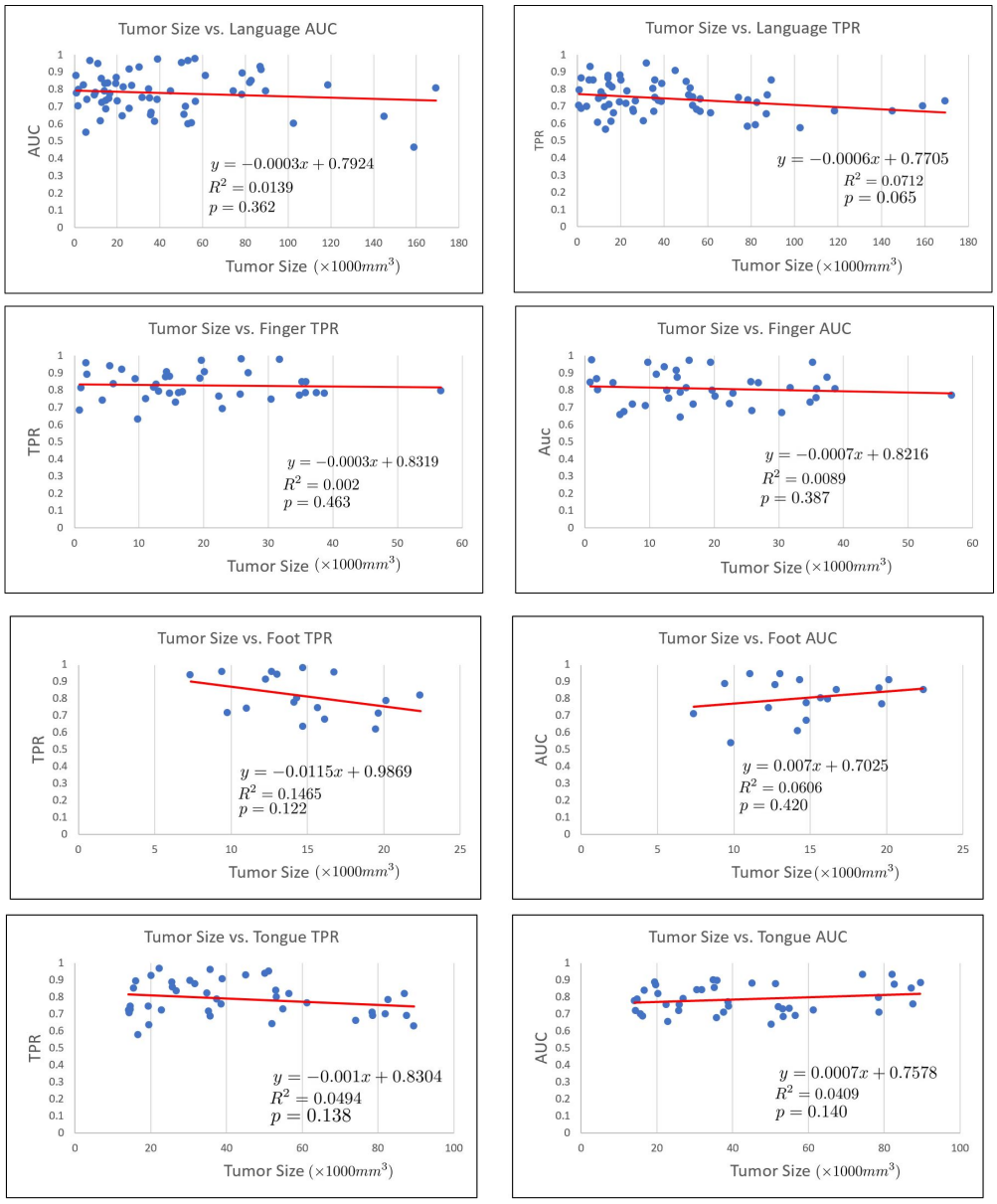
3 Model optimization experiment

3.1 Model augmentation

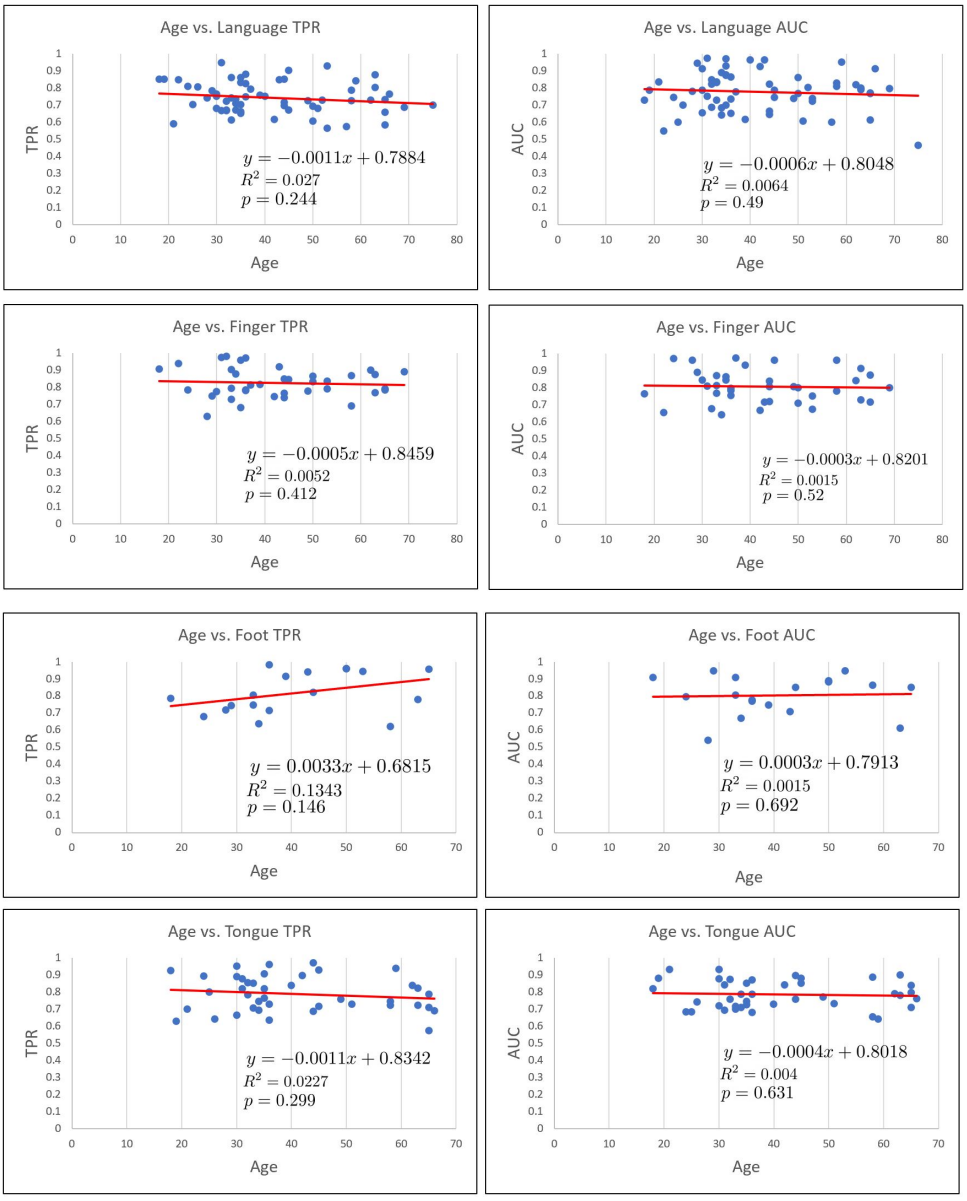
One common strategy in deep learning is to construct an architecture that overfits to the training data and then use regularization tricks, such as dropout, to close the generalization gap on a separate validation set. SFigs. 4-6 show training (blue) and validation (orange) curves for the task-specific TPR with the original model across three different scales of the Craddocks atlas ($N = 318, N = 384, N = 432$). The dotted black line represents the main JHH results from the manuscript using the $N = 384$ atlas. We can observe that the original model does not fully overfit to the training data, as all four blue curves do not saturate at 1. To arrive at a model that will overfit the training data, we increased capacity of the original model by increasing the number of feature maps in the convolutional layer and adding two fully-connected (FC) layers. Specifically, we increased M from 8 to 16, increased H_1 from 27 to 50 and added two FC layers of sizes $H_3 = 25$ and $H_4 = 20$. The overfit model is shown in SFig. 7. As shown in SFigs. 8-10, the model presented in SFig. 7 overfits to the training data, as each training curve saturates at around 1.

3.2 Adding dropout to overfit model

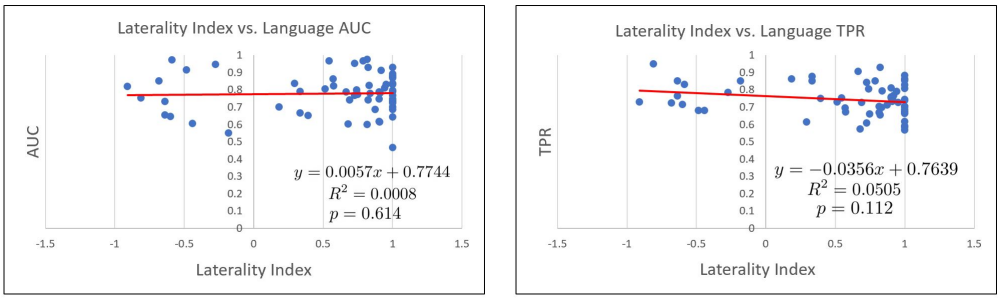
Though the model in SFig. 7 fits the training data well, it performs poorly when applied to unseen test data. Now that we have identified a model with enough capacity to fit the training data, our next goal is to decrease the generalization gap. To do this, we employ dropout with $p = 0.5$ in between each hidden layer



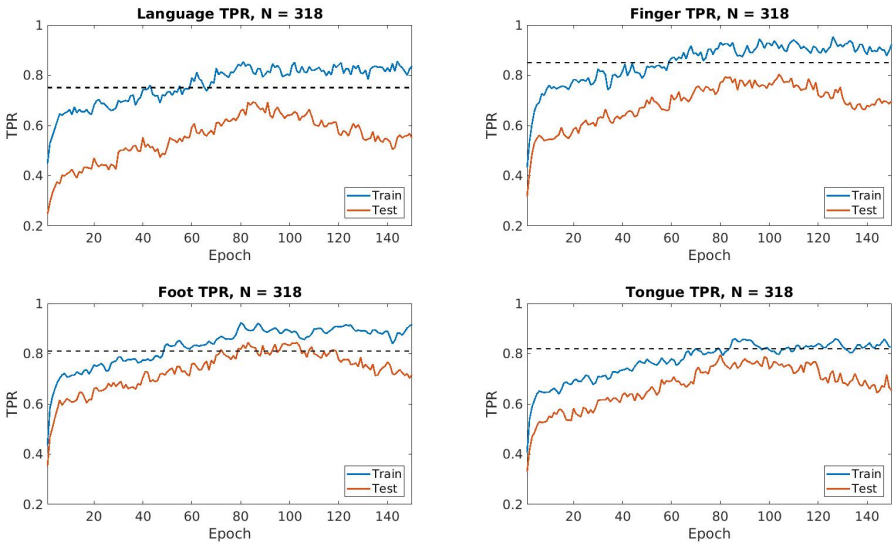
SFigure 1: Tumor size vs. AUC and TPR for each task



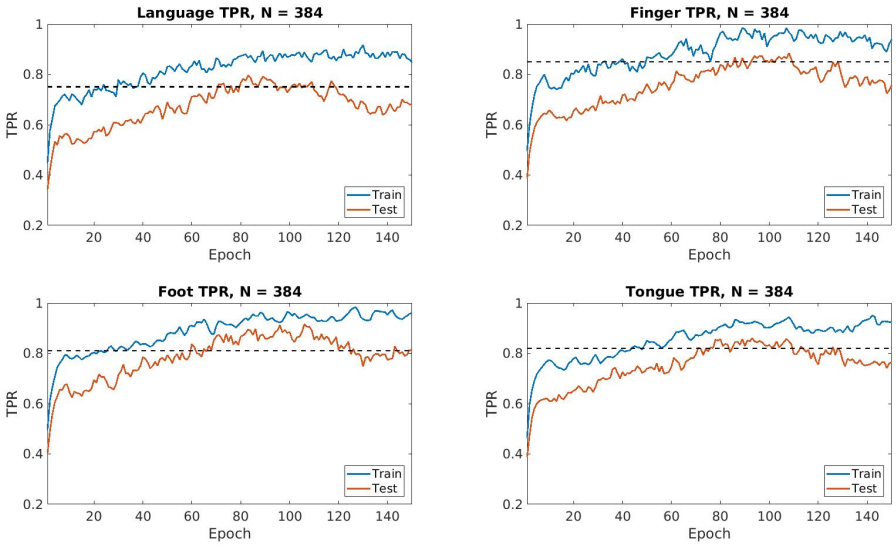
SFigure 2: Age vs. AUC and TPR for each task



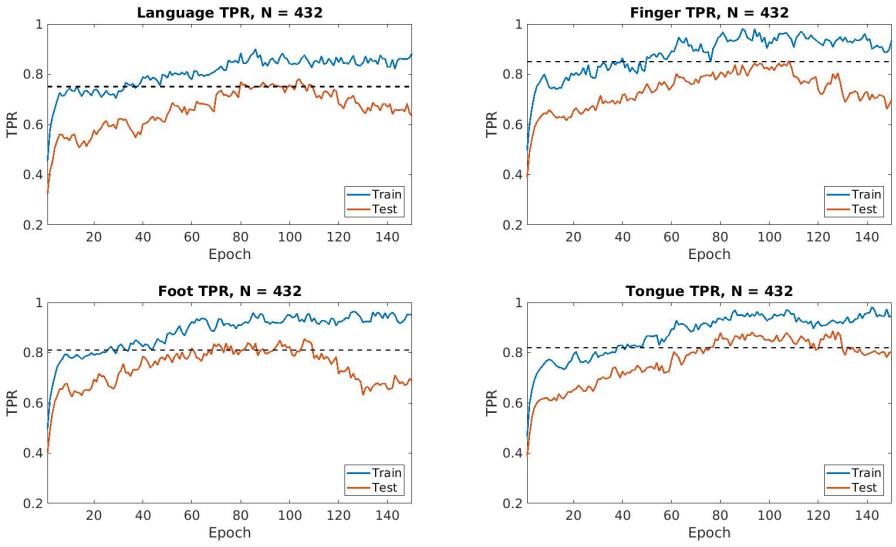
SFigure 3: Language laterality index vs. language AUC and TPR.



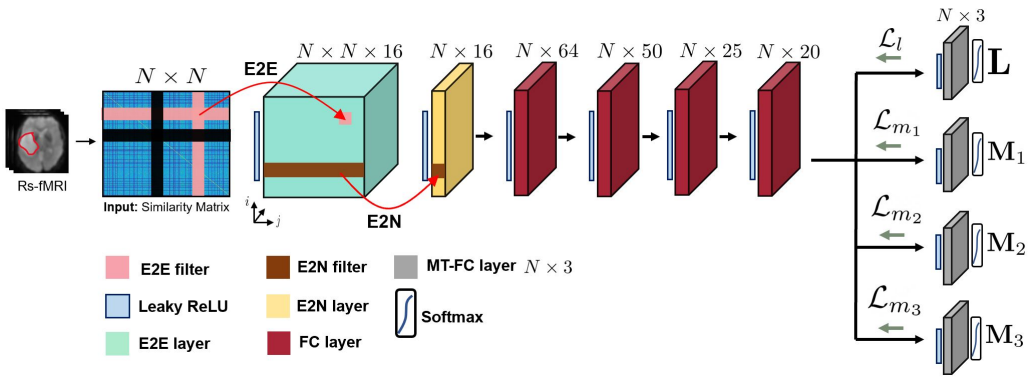
SFigure 4: Task-specific TPR for $N = 318$ atlas with original model. We observe that the training TPR (blue) does not saturate at 1.



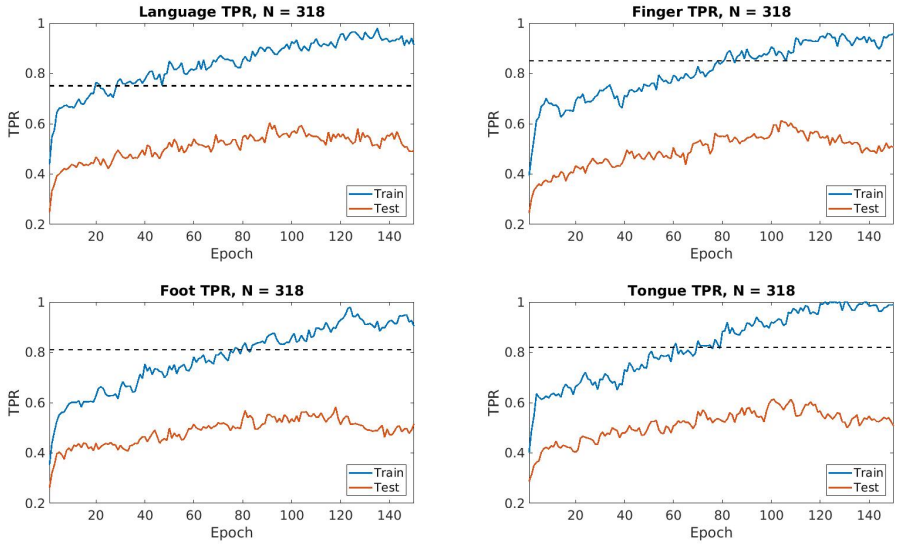
SFigure 5: Task-specific TPR for $N = 384$ atlas with original model. We observe that the training TPR (blue) does not saturate at 1.



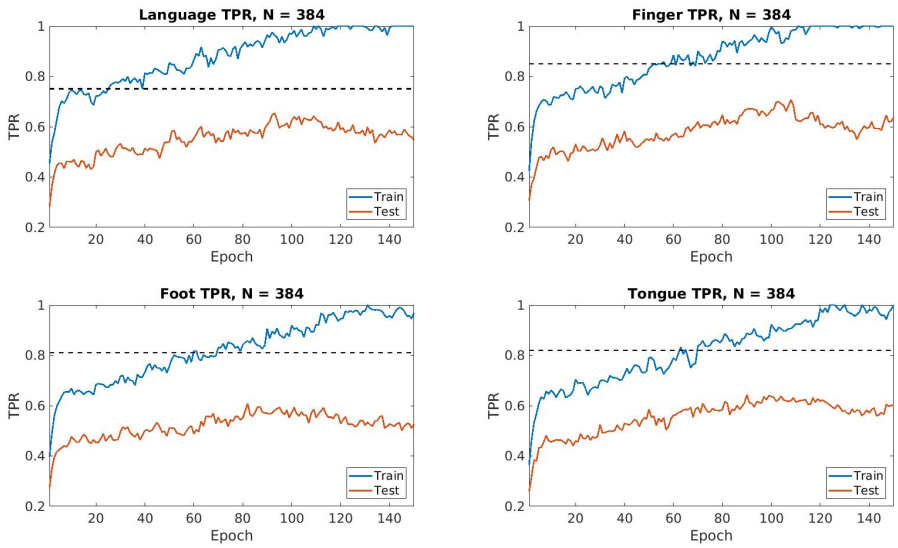
SFigure 6: Task-specific TPR for $N = 432$ atlas with original model. We observe that the training TPR (blue) does not saturate at 1.



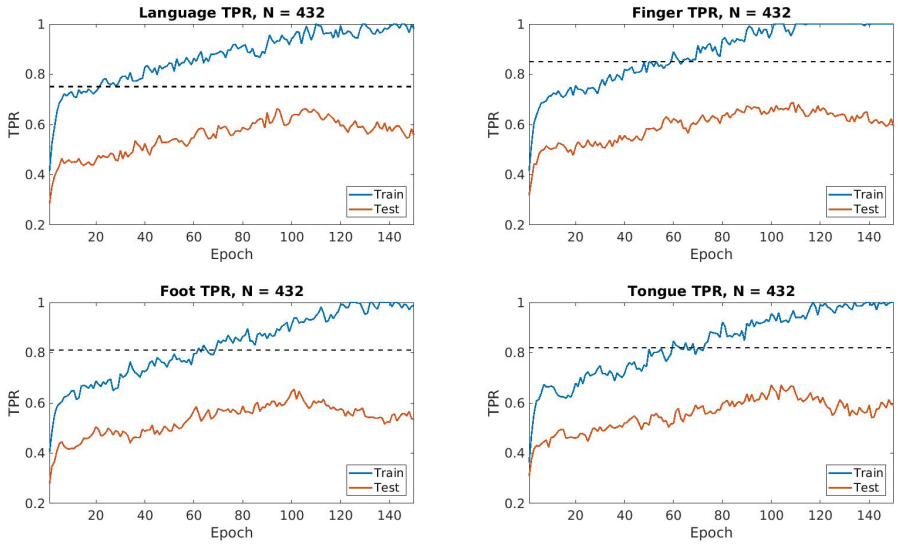
SFigure 7: Higher capacity model used for model optimization experiment.



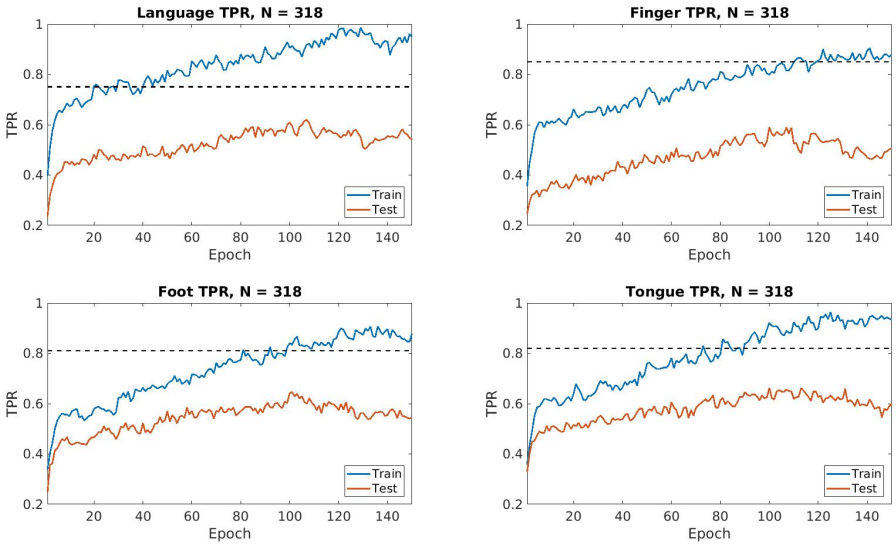
SFigure 8: Task-specific TPR for $N = 318$ atlas with higher capacity model, where training (blue) saturates at 1 but validation (orange) decreases.



SFigure 9: Task-specific TPR for $N = 384$ atlas with higher capacity model, where training (blue) saturates at 1 but validation (orange) decreases.



SFigure 10: Task-specific TPR for $N = 432$ atlas with higher capacity model, where training (blue) saturates at 1 but validation (orange) decreases.



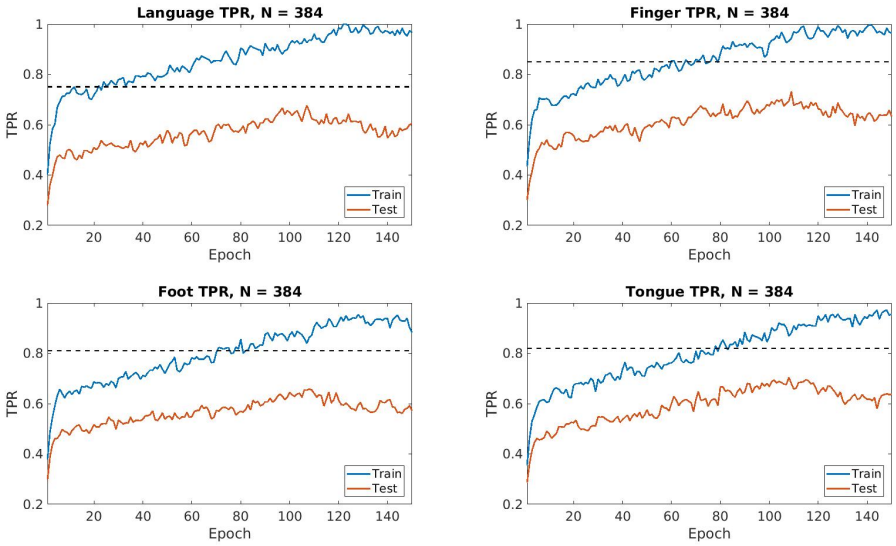
SFigure 11: Task-specific TPR for $N = 318$ atlas with higher capacity model and dropout. The validation accuracy is lower than that of the original model.

of the model except for after the E2E layer. Across three different scales of the Craddocks atlas, SFigs. 9-11 show the training (blue) and validation (orange) curves for the overfit model with dropout. The black dashed line indicates the original model performance on the $N = 384$ atlas. We observe that the validation is consistently lower than the black dashed line, indicating that the original model in the manuscript has the best testing performance to unseen data.

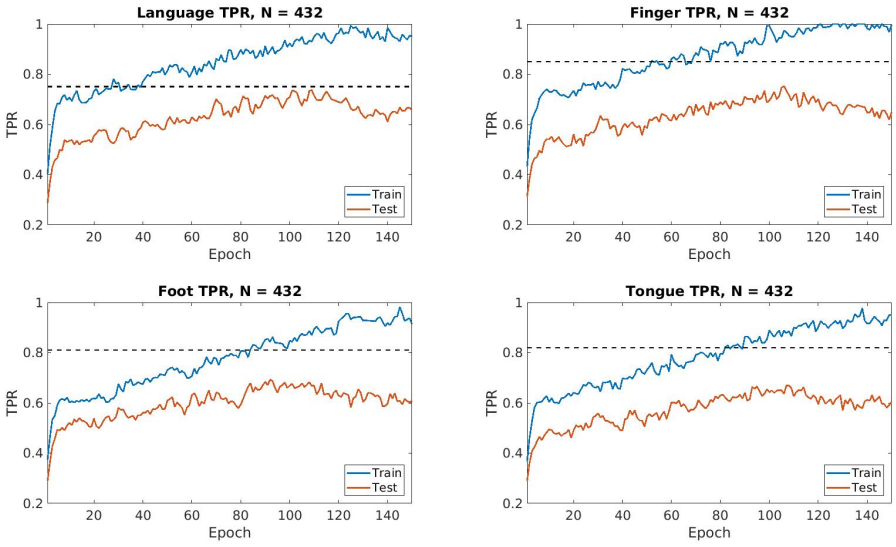
4 Healthy HCP experiment

4.1 Modelling changes

The work presented in the original manuscript treats eloquent cortex detection for tumor patients as a three-class classification problem, where tumor nodes are given their own class (i.e. not healthy and not belonging to the eloquent cortex). Removing the tumor class makes this a two-class classification problem, as each parcel is considered as either belonging to the eloquent cortex or not. Therefore, the MT-FC layers are now of size $N \times 2$ Other than the last MT-FC layers, we keep the layer dimensions consistent. To prevent biasing our hyperparameter selection, we once again use 10-fold CV on the separate healthy HCP2 dataset for hyperparameter selection, resulting in $\delta_l = (1.94, 0.54)$ and $\delta_m = (1.33, 0.54)$.



SFigure 12: Task-specific TPR for $N = 384$ atlas with higher capacity model and dropout. The validation accuracy is lower than that of the original model.



SFigure 13: Task-specific TPR for $N = 432$ atlas with higher capacity model and dropout. The validation accuracy is lower than that of the original model.

STable 2: Mean plus or minus standard deviation for eloquent class TPR and AUC for the HCP cohort (100 subjects).

Task	Method	Eloquent TPR	AUC	p-value
Language	MTGNN	$\mathbf{0.70 \pm 0.011}$	$\mathbf{0.72 \pm 0.009}$	
	FCNN	0.64 ± 0.02	0.66 ± 0.017	3.7**e-**51
	RF	0.35 ± 0.034	0.55 ± 0.032	5.1**e-**100
	SVM	0.40 ± 0.026	0.53 ± 0.019	≈ 0
Finger	MTGNN	$\mathbf{0.84 \pm 0.013}$	$\mathbf{0.84 \pm 0.007}$	
	FCNN	0.78 ± 0.013	0.75 ± 0.012	7.6**e-**106
	RF	0.44 ± 0.03	0.59 ± 0.026	1.6**e-**243
	SVM	0.42 ± 0.021	0.53 ± 0.014	≈ 0
Foot	MTGNN	$\mathbf{0.86 \pm 0.01}$	$\mathbf{0.82 \pm 0.012}$	
	FC-NN	0.75 ± 0.014	0.74 ± 0.013	2.7**e-**73
	RF	0.44 ± 0.027	0.59 ± 0.028	4.2**e-**100
	SVM	0.52 ± 0.022	0.52 ± 0.013	≈ 0
Tongue	MTGNN	$\mathbf{0.82 \pm 0.011}$	$\mathbf{0.80 \pm 0.008}$	
	FC-NN	0.77 ± 0.011	0.74 ± 0.011	3.7**e-**35
	RF	0.45 ± 0.027	0.58 ± 0.031	3.1**e-**155
	SVM	0.58 ± 0.021	0.55 ± 0.012	≈ 0

4.2 Healthy HCP ROI classification

We use a 10 repeated 10-fold CV evaluation strategy, where fold membership is different for each CV. Once again, we compare the MT-GNN with a multi-class linear SVM, a RF classifier, and a fully-connected neural network (FC-NN). The FC-NN hyperparameters were selected via 10-fold CV on the healthy HCP2 dataset as well and were set to be $\delta_l = (2.04, 0.44)$ and $\delta_m = (1.52, 0.44)$. Table 1 shows the eloquent class true positive rate (TPR), AUC, and FDR corrected p-value for the associated t-score comparing AUC's from the MT-GNN with the baseline methods. We observe that our model outperforms each baseline at each task. Compared to the results presented in Table 3, we observe that each method performs better, likely due to the absence of the simulated tumor, which disrupted healthy connections in these subjects. The performance gains from the MT-GNN to the baselines are slightly higher than those in Table 3 in the main manuscript, shown by even smaller p-values.

5 Discussion

In this supplementary document, we explored different potential confounders' effect on model performance, various model optimization strategies, and performance on the healthy HCP data. We observe no statistical significance in the correlation coefficients for each of the confounders. Regarding model optimization, we show confidence in the original model to generalize well to unseen testing data, as the overfit model or overfit model with dropout does not gen-

eralize as well. Finally, the healthy HCP result shows our method can identify localized functional subsystems of the eloquent cortex in healthy rs-fMRI scans. Therefore, the synthetic tumor experiment in the HCP Section 3.1 has a baseline comparison with the original data.