

Dissertation

CREATION OF ALGORITHMS AND TOOLS FOR THE STUDY OF BRAIN DEVELOPMENT WITH THE IDENTIFICATION OF CONFOUNDS

Anton Orlichenko
Advisor: Yu-Ping Wang

November 2024



Outline

- 1 Background on fMRI and Neurodevelopment
- 2 Problem Statement and Goals
- 3 Age-Related Functional Connectivity Changes in the UKB Longitudinal Cohort
- 4 Latent Similarity Model Identifies Important Functional Connectivity for Phenotype Prediction
- 5 ImageNomer: Description of a Functional Connectivity Analysis Tool and Identification of Demographic Confounds
- 6 Summary and Acknowledgements

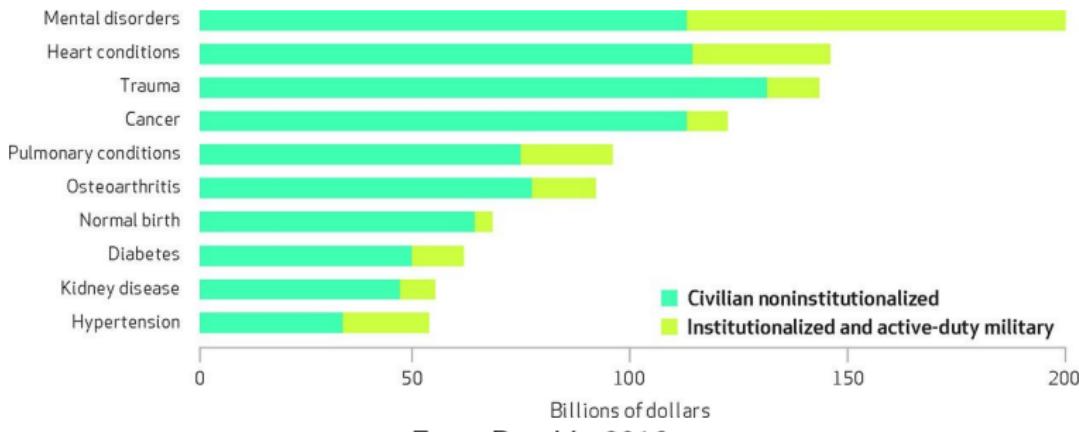


Background on fMRI and Neurodevelopment



Clinical Problem

- Schizophrenia, ADHD, depression, and other mental illnesses cost the U.S. \$201+ billion annually¹
- Dementia and Alzheimer's cost the U.S. \$157+ billion annually²
- Diagnosis of these diseases may be unreliable until symptoms become severe, when treatment options are more limited



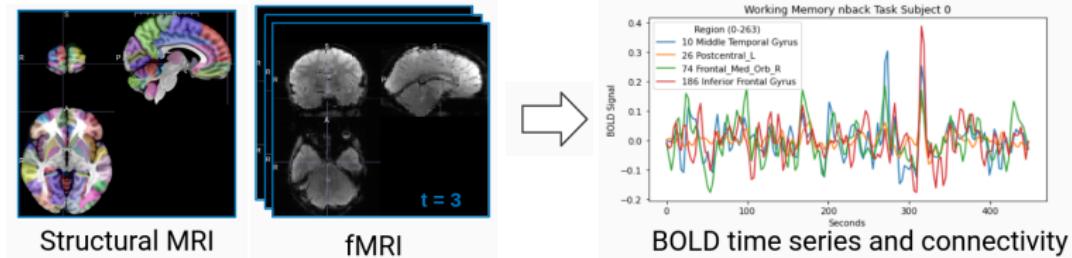
From Roerhig 2016.

¹ Roerhig 2016 <https://doi.org/10.1377/hlthaff.2015.1659>

² Hurd et al. 2013 doi:10.1056/NEJMsa1204629

fMRI

- Functional magnetic resonance imaging (fMRI) provides a non-invasive estimate of brain activity by exploiting the blood oxygen level-dependent (BOLD) signal

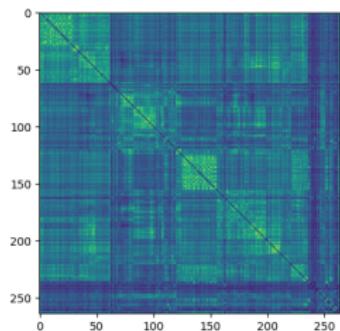


Example: extraction of BOLD timeseries from fMRI.

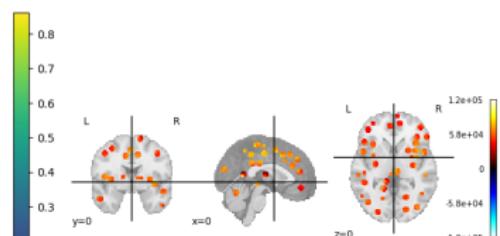


Functional Connectivity (FC)

- Functional connectivity (FC) is calculated as the Pearson correlation of BOLD signal between brain regions
- These brain regions are typically found using either ICA³ or an atlas⁴
- FC is the most common starting point for predictive fMRI studies and it is considered a brain fingerprint



mean FC of subjects in dataset



using 264 region Power template

³Calhoun et al. 10.1016/j.neuroimage.2008.10.057

⁴Power et al. 10.1016/j.neuron.2011.09.006

Power Atlas Functional Networks

- Example of Power atlas⁵ partition of brain into functional networks

Functional Networks

ROIs	Name	ROIs	Name
0-29	Somatomotor Hand	156-180	Frontoparietal
30-34	Somatomotor Mouth	181-198	Salience
35-48	Cinguloopercular	199-211	Subcortical
49-61	Auditory	212-220	Ventral Attention
62-119	Default Mode	221-231	Dorsal Attention
120-124	Memory	232-235	Cerebellar
125-155	Visual	236-263	Uncertain

⁵ Power et al. 2011 10.1016/j.neuron.2011.09.006

fMRI and Mental Health

- fMRI has been used to predict disease status and phenotypes such as age, sex, and general fluid intelligence⁶
- Machine learning based predictions of brain age have been correlated with future Alzheimer's diagnosis⁷
- Most diagnoses of mental disorders are still made by psychiatrists or physicians based on cognitive tests via Diagnostic and Statistical Manual of Mental Disorders (DSM)⁸

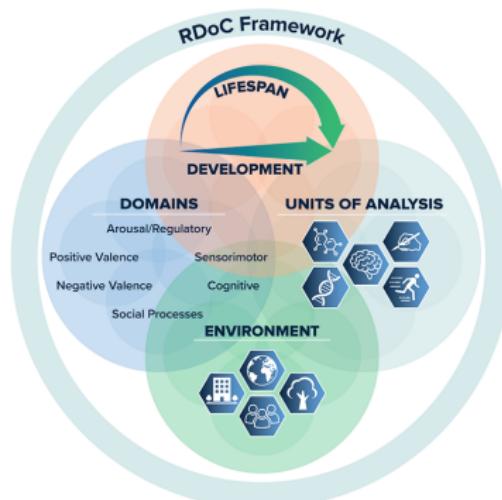
⁶ Qu et al. 2021 10.1109/TBME.2021.3077875

⁷ Millar et al. 2022 10.1016/j.neuroimage.2022.119228

⁸ <https://www.ndcn.ox.ac.uk/divisions/fmrib/what-is-fmri/how-is-fmri-used>

Research Domain Criteria (RDoC)

- The National Institutes of Mental Health (NIMH) have created the **Research Domain Criteria (RDoC)** to put mental health diagnoses on a more rigorous scientific basis⁹
- Using multidimensional biomarkers including genomics and brain imaging



10

⁹ Insel et al. 10.1176/appi.ajp.2010.09091379

¹⁰ <https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/about-rdoc>

Usefulness of Functional Connectivity

Study	Dataset	# of Subjects	Predictive Task	Accuracy
Hu et al. 2019	PNC	857	Age (child vs. young adult)	99%
Zhang et al. 2018	HCP	820	Sex	87%
Orlichenko et al. 2023	PNC, BSNIP	830	Race	85%
Du et al. 2023	COBRE + study-specific	688	Schizophrenia	73-76%
Zou et al. 2022	study-specific	80	Bipolar disorder	68-72%
Suo et al. 2020	study-specific	122	PTSD	$\rho = 0.298$
Ingalhalikar et al. 2022	ABIDE	988	Autism	60-90%
Hua et al. 2023	ADNI	235	Preclinical Alzheimer's disease	75-90%
Shi et al. 2022	FCP/INDI	100	Parkinson's disease	67%
Park et al. 2020	UKB	1,497	Obesity	$\rho = (0.308, 0.556)$

Table: List of representative studies using fMRI-based functional connectivity to predict patient demographics or illness status. In some cases, a correlation-based methodology was employed instead of a training and test set.



Problem Statement and Goals



Problem Statement

We seek to better understand the structure and function of the human brain, in order to

- ① characterize neurological disease for diagnosis and treatment
- ② understand normal development

but come up against **challenges** and **opportunities**.



Challenge 1

Small Sample Size but High Dimensionality

The median cohort size for fMRI studies in 2017 and 2018 was 23 subjects,^a while

- number of voxel-level features can be $> 10^6$
- number of connectivity-level features can be $> 10^4$

^aSzucs and Ioannidis 2020 10.1016/j.neuroimage.2020.117164

Challenge 2

Site-Specific Effects and Demographic Confounds

- Variation in scanner configuration, imaging protocol, and imaging parameters can lead to differences in the characteristics of resulting images^a
- Correlations among features of interest and demographics can result in unintentional inclusion of bias and masking of signal of interest^b

^aMali et al. 2021 10.3390/jpm11090842

^bSnoek et al. 2019 10.1016/j.neuroimage.2018.09.074

- This may lead to incorrect conclusions about biological basis of identified features



Challenge 3

Lack of Convenient Software for Correlational Analysis of FC and SNPs

- Most software for analysis of medical imaging data requires in-depth programming knowledge (e.g., scikit-learn^a or nilearn)
- GUI-based software such as SPM12 does not offer way of analyzing FC or genomic data such as SNPs^b

^aPedregosa et al. 2011 10.5555/1953048.2078195

^b<https://www.fil.ion.ucl.ac.uk/spm/software/spm12/>

- We would like to create a GUI program to explore FC and omics datasets



Opportunity

Large Public Datasets

Repositories exist for large amounts of sMRI, fMRI, and genomic data

- UK Biobank (UKB) has 40,000+ subjects with fMRI scans from a largely homogenous population^a
- OpenNeuro has 800+ small open access datasets with varied demographics^b

^aSudlow et al. 2015 10.1371/journal.pmed.1001779

^bMarkiewicz et al. 2021 10.7554/eLife.71774

Since fMRI data is high dimensional, use of large public datasets can remedy the curse of dimensionality by having large sample size



Age-Related Functional Connectivity Changes in the UKB Longitudinal Cohort



Old Age-Related Changes in FC Have Been Uncertain

- Studies have shown that changes occur in the default mode network (DMN) during late middle and old age¹¹
- The exact direction of change in FC does not always appear constant¹²
- Given the recent interest in using fMRI to predict pre-clinical Alzheimer's disease, we believe a knowledge of normative changes in FC during old age is essential¹³
- One problem is that most studies have employed cross-sectional cohorts
- The UKB presents a unique opportunity with a 2,800-subject longitudinal cohort

¹¹ Fjell et al. 2017 10.1002/hbm.23403

¹² Staffaroni et al. 2018 10.1523/JNEUROSCI.3067-17.2018

¹³ Millar et al. 2023 10.7554/eLife.81869



FC Increases Significantly in the Somatomotor-Visual Connection

- We find that, in the UKB longitudinal cohort, FC increases significantly in the Somatomotor-Visual system connection

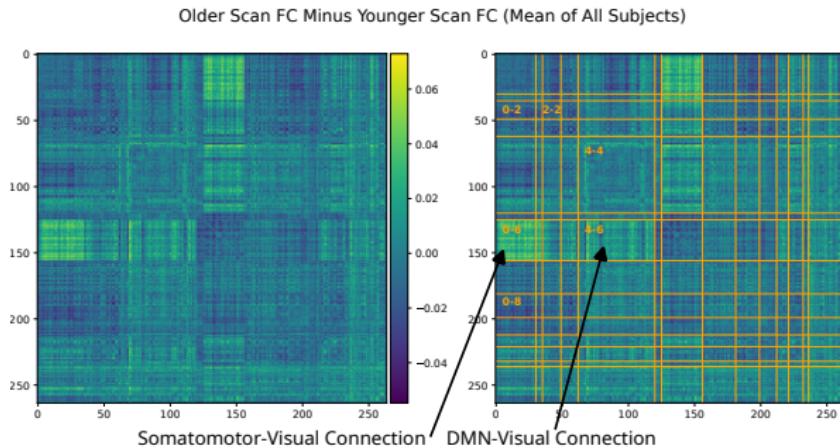
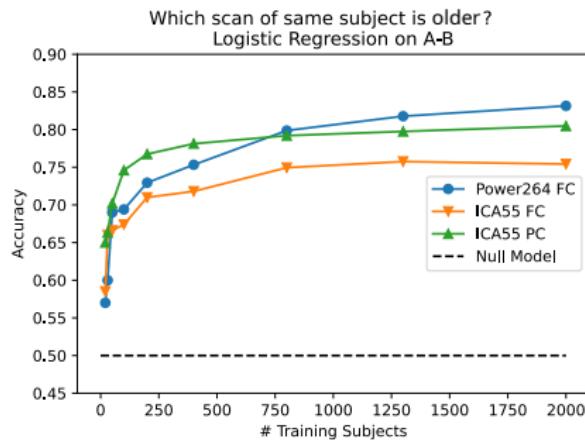


Figure: Difference of older scan minus younger scan in the UKB longitudinal cohort. Delineations are Power atlas functional networks.



FC Can Discriminate Between Younger and Older Scan

- The full FC can predict scan age at a level of 85% in the full longitudinal cohort using linear classifier
- Figure shows capability of predicting older scan of a pair as a function of the number of training subjects.



Somatotmotor-Visual Connectivity Best Predicts Scan Age

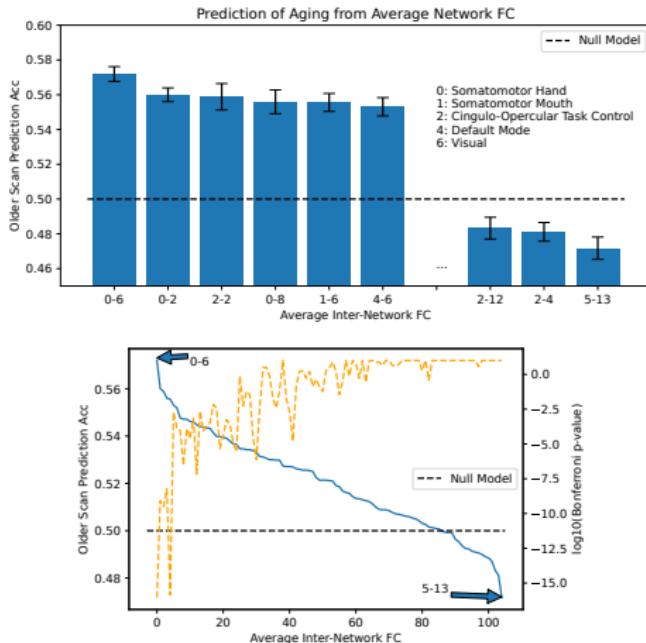


Figure: Top: best and worst inter-network connectivities for prediction. Bottom: Prediction accuracy for all 105 inter-network connectivities.



Somatotmotor and Visual Network ROIs

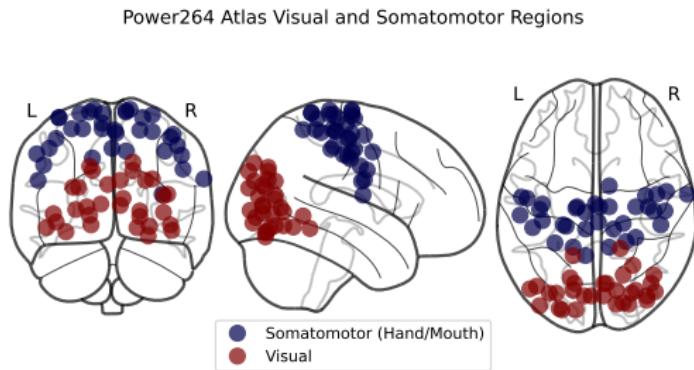
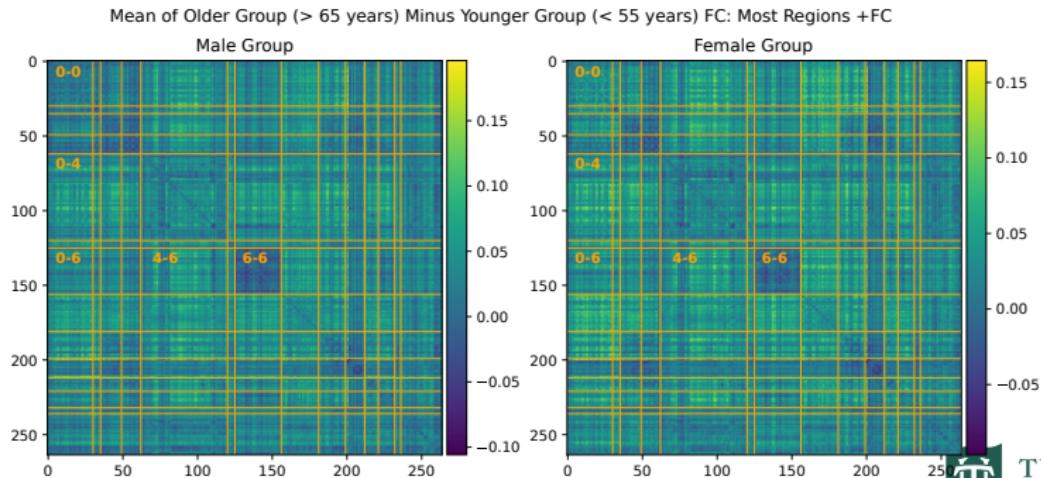


Figure: SMT and VIS regions in the Power264 atlas.



Connectivity Increases Throughout Brain in Cross-Sectional Cohort

- Average resting state FC has a significant increase in almost all inter-network connections in the UKB cross-sectional cohort (40,000+ subjects)



TULANE
UNIVERSITY

Connectivity Changes are Highly Significant

Regions	Male (Young to Old)			p-value
	Mean FC Increase	Std Dev of Avg FC	Avg FC	
SMT-VIS (0-6)	0.031	0.13		$< 10^{-23}$
SMT-DMN (0-4)	0.045	0.11		$< 10^{-78}$
DMN-VIS (4-6)	0.042	0.11		$< 10^{-60}$
VIS-VIS (6-6)	-0.014	0.10		$< 10^{-7}$
Total FC	0.035	0.09		$< 10^{-64}$

Table: Average FC changes with age in the UKB cross-sectional cohort from young subjects (< 55 years old) to old subjects (> 65 years old).

SMT=Somatomotor Network, VIS=Visual Network, DMN=Default Mode Network, FC=Functional Connectivity

- Similar results for female group



Latent Similarity Model Identifies Important Functional Connectivity for Phenotype Prediction



Problem Statement

- fMRI has been shown to predict age, sex, psychiatric disease status, and brain age associated with Alzheimer's disease
- Prediction is hindered, however, by the small sample size but high feature number.
- We have thus developed a **kernel-based deep learning model** that is robust at small sample sizes
- We also create a **greedy tree-based algorithm** for feature selection
- Seeking to improve prediction performance, we also use the model to **integrate omics (SNPs) with FC data**



Latent Similarity Model Overview

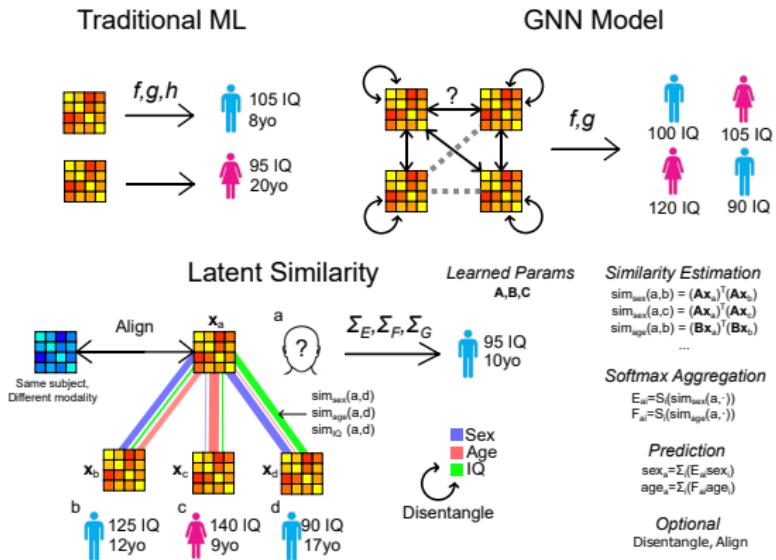


Figure: In traditional ML, estimation of response variables is decoupled from inter-subject similarity, whereas graph neural network (GNN) models require additional degrees of freedom to estimate edges between subjects.

Mathematical Description of Similarity

$$\text{similarity}(a, b) = \langle \phi(\mathbf{x}_a), \phi(\mathbf{x}_b) \rangle$$

$$= \mathbf{x}_a \mathbf{A} \mathbf{A}^T \mathbf{x}_b^T,$$

$$\mathbf{M} = \text{diag}(\infty),$$

$$\mathbf{E} = \text{Softmax}_{Row}((\mathbf{1} - \mathbf{M}) \odot \mathbf{X} \mathbf{A} \mathbf{A}^T \mathbf{X}^T), \quad (1)$$

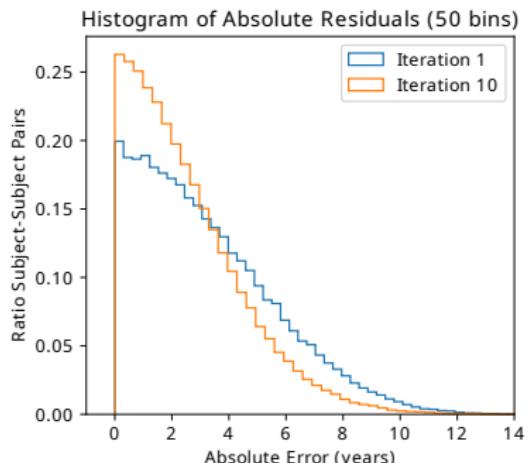
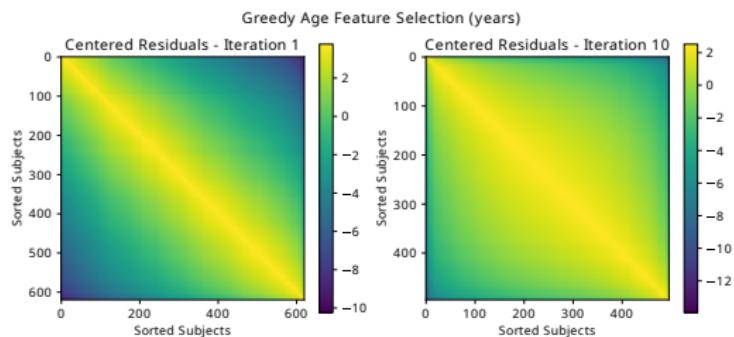
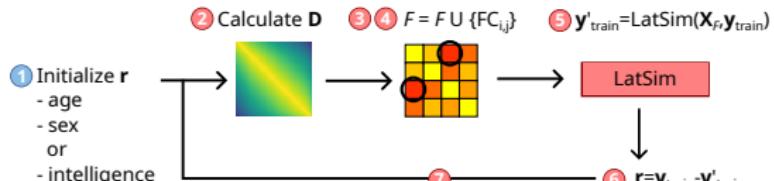
$$\hat{\mathbf{y}} = \mathbf{E} \mathbf{y}_{train}$$

$$\underset{\mathbf{A}}{\text{minimize}} \quad \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 \quad (\text{Regression})$$

$$\underset{\mathbf{A}}{\text{minimize}} \quad \text{CE}(\mathbf{y}, \mathbf{p}_{\hat{\mathbf{y}}}) \quad (\text{Classification})$$

- $\phi(\mathbf{x}_a)$ is a kernel function acting on input features \mathbf{x}_a for subject a
- $\mathbf{E} \in \mathbb{R}^{N \times N}$ is the final similarity matrix
- $\mathbf{M} \in \mathbb{R}^{N \times N}$ is a mask to remove self-loops in predictions
- $\hat{\mathbf{y}}$ is the prediction on the training or test set

Greedy Feature Selection



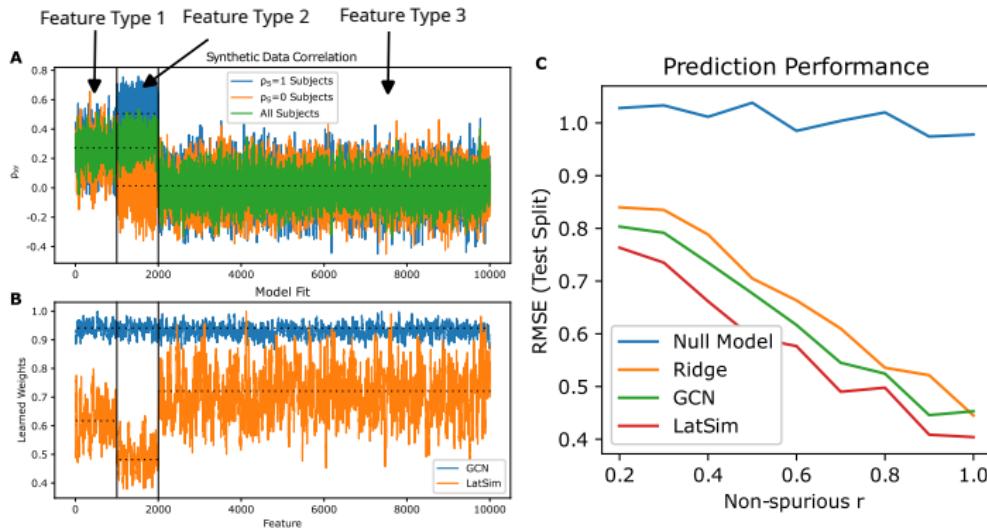
Simulation Experiment

- We performed a simulation experiment to test LatSim in the presence of a ground truth dataset.
- A set of $N_{train} = 40$, $N_{val} = 120$, and $N_{test} = 120$ subjects with 10,000 normally-distributed features x_{ni} was created
- There were three types of features, those correlated with response variable for all subjects, for $\frac{1}{2}$ of subjects, and those uncorrelated and just noisy



Simulation Results

- LatSim performs better than linear model and graph convolutional network (GCN) model and clearly distinguishes between the three types of features



Brain Development Study

- We utilize a 620-subject subset of the Philadelphia Neurodevelopmental Cohort (PNC) dataset of normal children 8-22 years old

	Number of Subjects
Males	286
Females	334
Total	620

	Min	Mean	Max
Age (months)	103	180±39	271
Age (years)	8.6	15±3.3	22.6
WRAT score	70	102±15.7	145

Table: Demographic information for the subset of the PNC dataset used in our experiments.



Prediction Accuracy

- To test LatSim, we predict subject age, sex, and wide range achievement test score (WRAT or fluid intelligence) based on functional connectivity
- LatSim is significantly better than other methods at predicting age and WRAT score

Model	Age (RMSE, years)		Sex (Accuracy)		Intelligence (RMSE, WRAT score)	
	N=30	N=496	N=30	N=496	N=30	N=496
Null	3.3		0.54		15.7	
M-GCN	4.47	2.37	0.51	0.75	23.27	15.59
MLP	4.52	2.43	0.53	0.8	21.17	15.64
GCN	3.89	2.16	0.49	0.8	16.29	14.38
Linear	4.36	2.34	0.54	0.77	19.8	14.97
LatSim	2.86	2.05	0.55	0.82	15.59	14.26
p-value	2.2e-6	5e-3	0.32	0.11	0.02	0.30

Table: Results of PNC experiments.

Greedy Selection Algorithm Age Results

- The greedy tree-based selection algorithm is shown to be superior to saliency or random features for prediction using only a few features

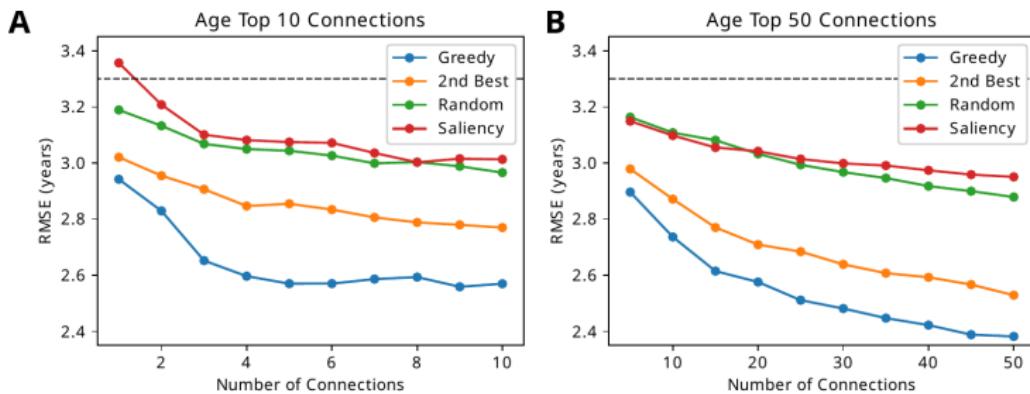


Figure: Greedy feature selection for age prediction.

Greedy Selection Algorithm Sex Results

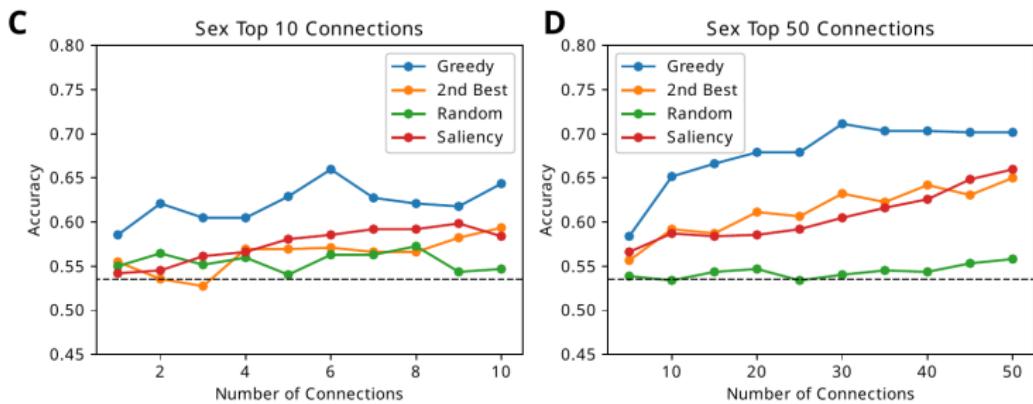


Figure: Greedy feature selection for sex prediction.

Greedy Selection Algorithm WRAT Score Results

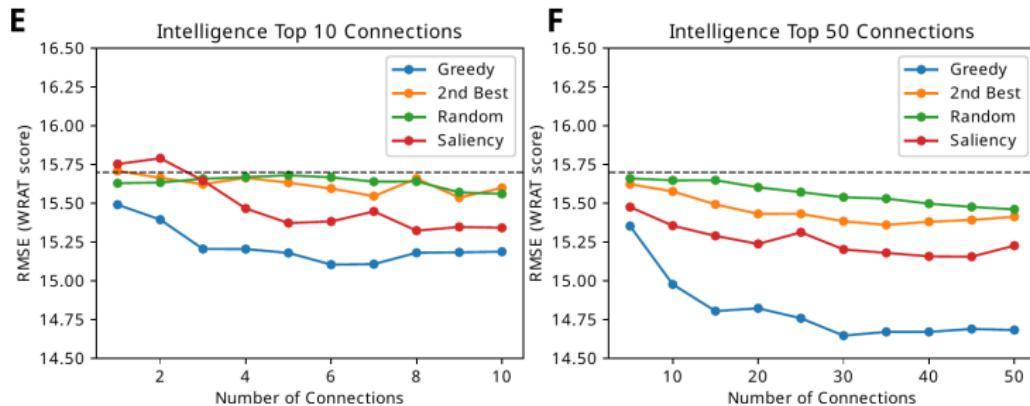
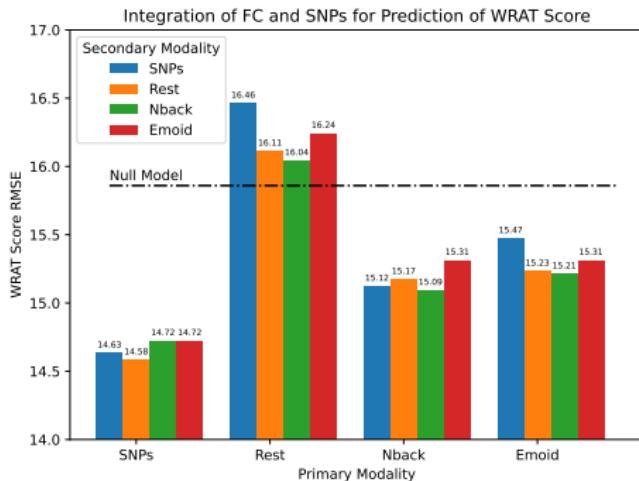


Figure: Greedy feature selection for WRAT score prediction.

Integration of SNPs and FC for Prediction of WRAT Score

- We used single nucleotide polymorphism data along with FC to predict WRAT score
- Original set of a million+ SNPs reduced to 35,621 common SNPs
- SNPs integration with FC yield superior prediction of WRAT score



Significant SNPs for WRAT Prediction

- Surprisingly, most SNPs significant for WRAT prediction had a large disequilibrium between different ethnic groups

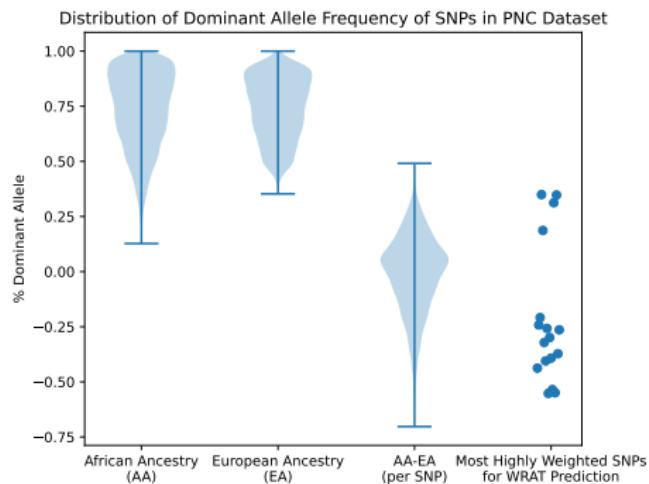


Figure: Plot of the distribution of dominant allele frequency among the two main ethnic groups in the PNC dataset, as well as difference between the two groups for the 20 most highly weighted SNPs.

Most Significant SNPs

SNP	Weight	AA %	EA %	Allele	Type	Gene	Role
rs919927	1.513	0.696	0.565	A>G	intron	CTNNA2	Actin Binding in Brain
rs7854368	1.053	0.325	0.7628	A>G,T	intron	SUSD1	Calcium Binding Integral Membrane
rs7603462	1.015	0.634	0.877	A>C,G,T	non-coding		
rs2476773	0.899	0.507	0.716	C>A,G,T	intron	NALF1	Stretch activated calcium channel
rs7816586	0.866	0.891	0.541	A>C,G	intron	EPHX2	Epoxide hydrolase
rs898647	0.725	0.523	0.609	A>C,G,T	non-coding		
rs1956533	0.705	0.570	0.509	T>A,C	intron	RAD51B	DNA Repair by homologous repli
rs12457893	0.703	0.740	0.554	A>C,T	intron	BCL2	Cell death, blocks apoptosis
rs4692540	0.636	0.513	0.835	G>A,C,T	non-coding		
rs1036542	0.630	0.328	0.733	T>C,G	intron	ANKRD44	Recognition of phosphoproteins, Neurodevelopmental disorder-assoc
rs2135354	0.531	0.384	0.642	G>A,C,T	intron	ANK2	Membrane-cytoskeleton linkage
rs920249	0.529	0.323	0.875	A>C,G,T	intron	SPATS2L	RNA binding
rs7630982	0.501	0.590	0.889	G>A	nc-transcript	LINC02054	Glioma susceptibility
rs2855711	0.490	0.484	0.877	A>G,T	intron	ETV6	Vascular transcription factor
rs2252268	0.455	0.242	0.777	A>C,G,T	intron	CDH4	Cadherin cell-cell adhesion glycop
rs223498	0.448	0.809	0.497	A>C,G	intron	MANBA	Lysosome enzyme
rs9285480	0.419	0.406	0.955	C>A,T	intron	AHI1	Cerebellar and cortical developme
rs9518729	0.415	0.714	0.367	C>T	non-coding		
rs2211479	0.355	0.338	0.711	T>C,G	non-coding		
rs7824311	0.345	0.457	0.721	G>A	intron	KCNQ3	Neuronal potassium channel

Table: Description of the twenty most highly weighted SNPs for WRAT score prediction in the PNC dataset. All highly weighted SNPs are non-coding.

AA=African Ancestry, EA=European Ancestry



LatSim Resilience to Spurious Correlations

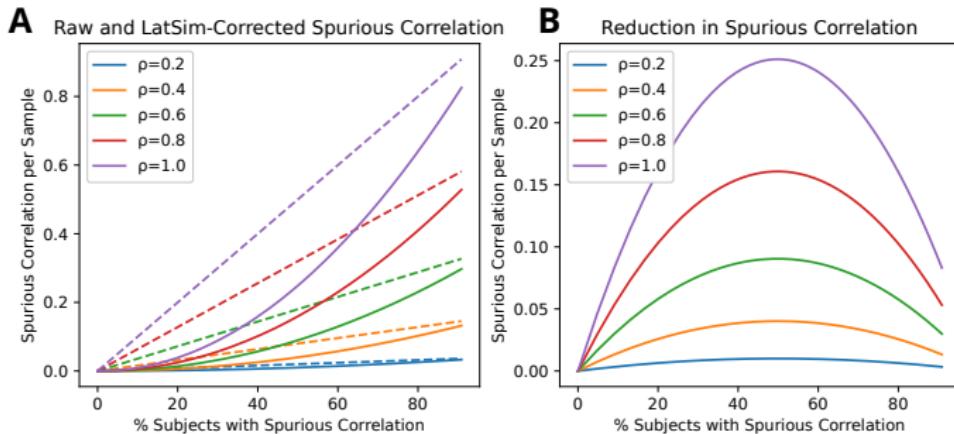


Figure: **A.** Spurious correlation per sample in a traditional ML model (dashed lines) versus LatSim (solid lines). **B.** The absolute reduction in spurious correlation as a function of frequency in the sample.

ImageNomer: Description of a Functional Connectivity Analysis Tool and Identification of Demographic Confounds



Problem Statement

- Existing software packages for analysis of fMRI, FC, and FC-like measures such as partial correlation connectivity are either mostly text-based (programmatic interface)
- Or they have incomplete feature sets for identifying correlations with phenotypes
- We present a new software tool called ImageNomer for visualization and analysis of FC or SNPs associated with phenotypes

	correlation analysis	GUI	no Matlab depend.	demographics	FC/PC/ imaging	SNPs	custom groups	easy to use
numpy/PyTorch	[+]	[-]	[+]	[+]	[-]	[-]	[+]	[-]
sklearn/nilearn	[+]	[-]	[+]	[+]	[+]	[-]	[+]	[-]
GIFT	[-]	[+]	[-]	[-]	[+]	[-]	[+]	[+]
BrainNet viewer	[-]	[+]	[-]	[-]	[+]	[-]	[-]	[+]
COINSTAC	[+]	[+]	[+]	[+]	[+]	[-]	[-]	[-]
Infinitome	[-]	[+]	[+]	[-]	[+]	[-]	[-]	[+]
ImageNomer	[+]	[+]	[+]	[+]	[+]	[+]	[+]	[+]

Figure: Comparison of existing tools for analysis of FC



TULANE
UNIVERSITY

ImageNomer Architecture

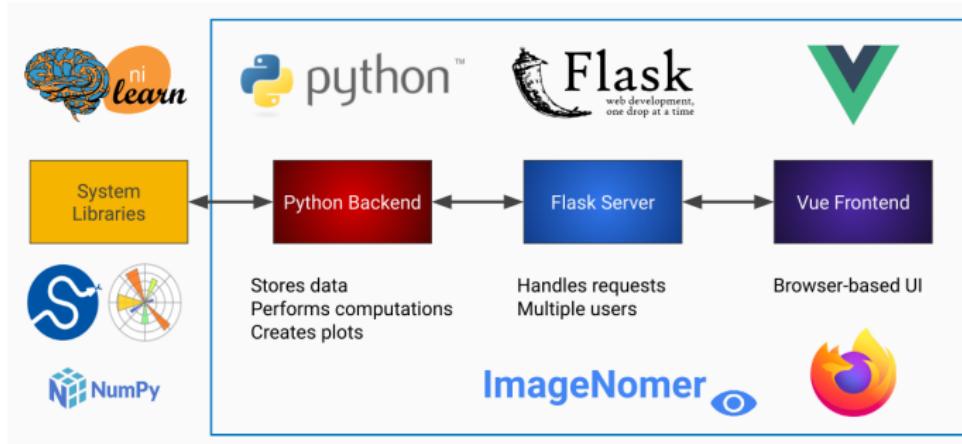
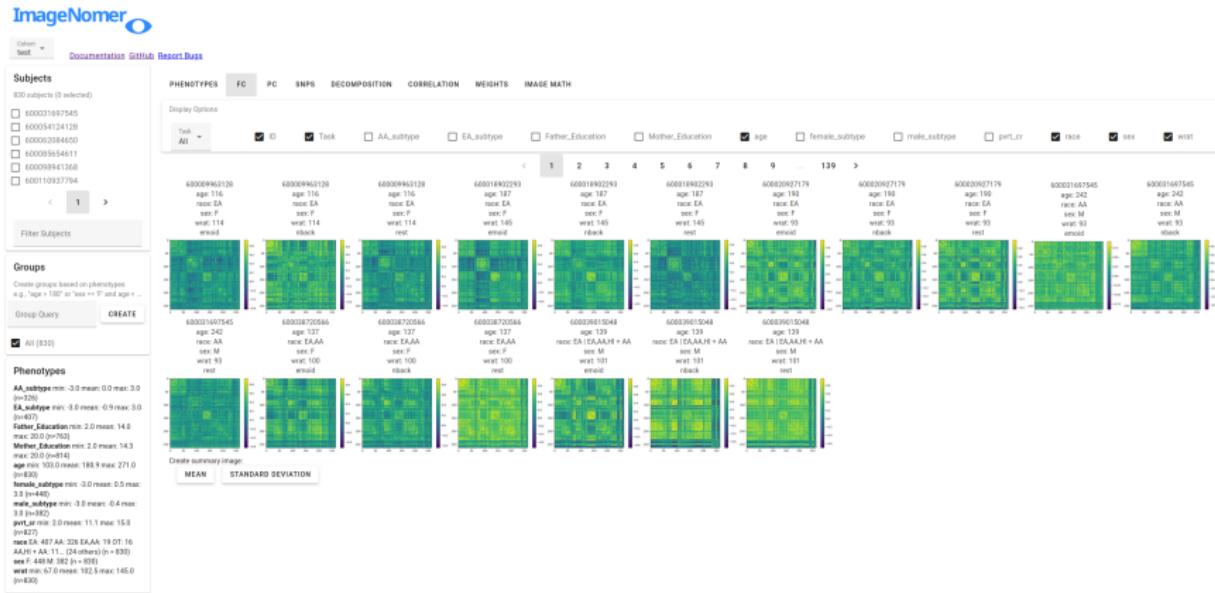


Figure: Overview of the ImageNomer architecture.

Screenshot of ImageNomer

- Main view of the ImageNomer program showing resting state FC for all subjects along with demographic data.¹⁴



¹⁴ <https://aorliche.github.io/ImageNomer/live/>

ImageNomer Capabilities

ImageNomer has the following capabilities:

- Examine individual subject FC and partial correlation-based (PC) connectivity
- Display distributions of phenotypes
- Carry out correlation analysis between phenotypes, FC, and SNPs
- Display p-value maps for correlations
- Perform calculations on single or average FC maps
- Display components for FC decompositions (such as PCA)
- Correlate decomposition components with phenotypes or SNPs
- Display weights from linear models
- Summarize and average weights from multiple models



PNC Dataset Demographics Visualized in ImageNomer

A view of ImageNomer plots displaying demographics in the PNC dataset.

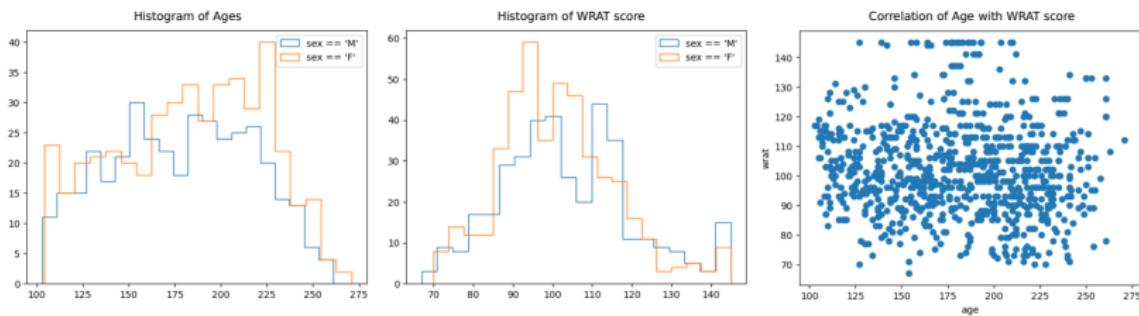


Figure: Demographics of our subset of the PNC dataset. Plots of age vs sex, WRAT vs sex, and WRAT vs age are shown. All plots created using the GUI of ImageNomer, without programming input.

Confounding Effect of Race on WRAT Score

Using ImageNomer we quickly identify a confound of race on WRAT score.

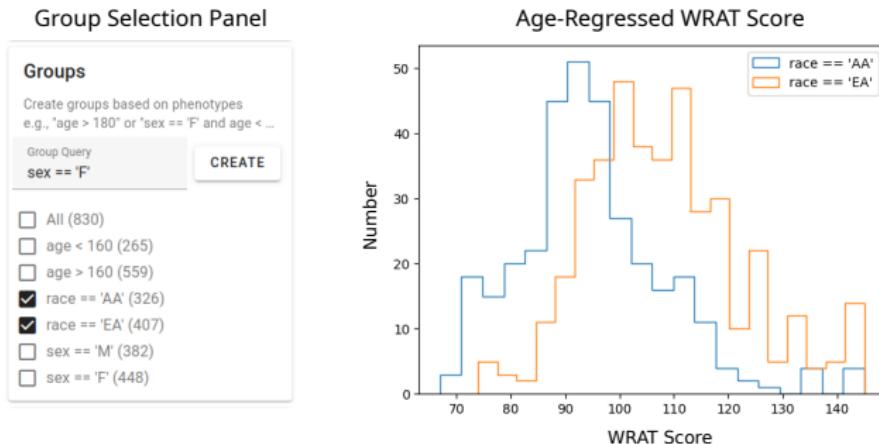
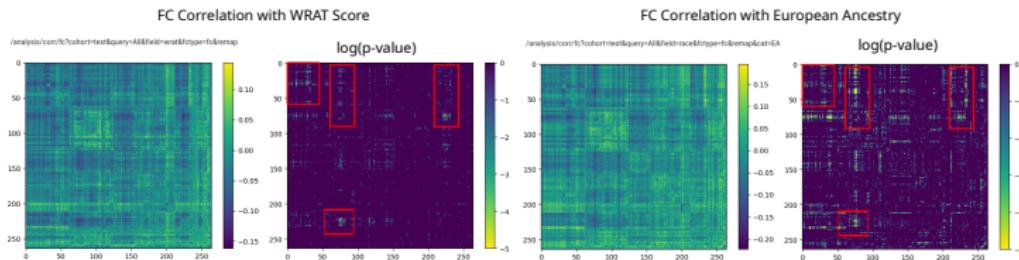


Figure: Examining race bias on WRAT score in the PNC dataset using ImageNomer. We find whereas age has been regressed from WRAT score, there is still a large racial bias.



FC Regions and SNPs Associated with WRAT Score and Race

- We find that while FC and SNPs can predict WRAT score, prediction ability disappears when controlling for race
- Meanwhile, FC and SNPs associated with WRAT score are also associated with race
- Out of the top 20 SNPs correlated with WRAT score, 6 are highly correlated with race.



DemoVAE Model to Confirm Demographic Confounds

- We created a generative model to recapitulate the FC distribution in a dataset conditioned on user-specified demographics (age, sex, race)

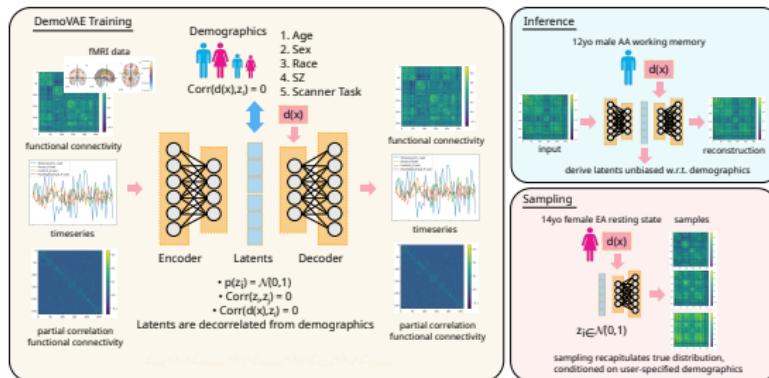
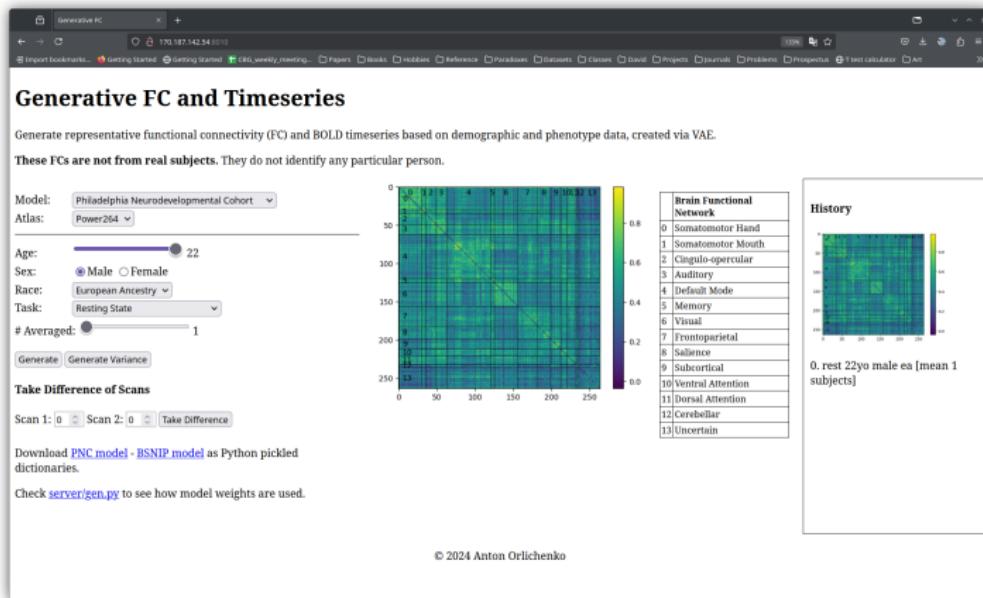


Figure: Overview of the **demographics-conditioned and decorrelated variational autoencoder (DemoVAE)** model. Instead of reconstruction based only on latent features $\mathbf{z} = E_\phi(\mathbf{x})$, the DemoVAE model uses demographics \mathbf{y} as input to the decoder $\hat{\mathbf{x}} = D_\theta(\mathbf{z}, \mathbf{y})$.

DemoVAE Demonstration

- DemoVAE can be used to create synthetic FC data based on user-input demographics that recapitulates real group differences¹⁵

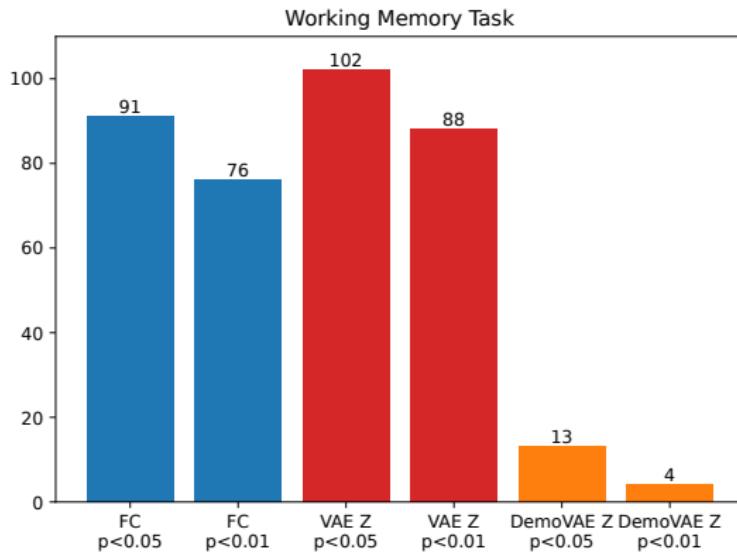


¹⁵ <https://aorliche.github.io/DemoVAE/>

Most FC-based Correlations Depend on Demographics

- Using the DemoVAE model to remove demographics from latent features, we find that most response variable-FC correlations depend on demographic information

Significant Correlations Between Tasks/Questionnaire/Demographic Fields
and FC/DemoVAE Latents in PNC Dataset



Summary and Acknowledgements



Conclusions

This thesis has:

- Identified robust changes of somatomotor-visual connectivity with age in the large longitudinal UK Biobank dataset
- Used a novel latent similarity model to predict subject phenotypes based on FC and/or SNPs
- Developed the ImageNomer software for analysis and exploration of FC and omic datasets
- Utilized ImageNomer to identify that race information in FC influences prediction of WRAT score
- Created a generative model (VAE) to remove the effects of demographics from FC latent features, showing most correlations depend upon demographics



Acknowledgements



- Lab members: Yu-Ping Wang (PI), Anton Orlichenko, Gang Qu, Binish Patel, Ziyu Zhou, Xiaoxuan Li (PhD students)
- We would like acknowledge the NIH (grants R01 GM109068, R01 MH104680, R01 MH107354, P20 GM103472, R01 EB020407, R01 EB006841, R56MH124925) and NSF (#1539067) for partial funding support



Committee Members and Collaborators

I would like to extend my sincere gratitude to the incredible help of my dissertation committee:

- Yu-Ping Wang (chair)
- Donald Gaver
- Hong-Wen Deng
- Zhengming Ding

And for the insightful contributions of our collaborators:

- Vince D. Calhoun
- Tony W. Wilson
- Julia M. Stephen



Publications

- Anton Orlichenko, Kuan-Jui Su, Hui Shen, Hong-Wen Deng, and Yu-Ping Wang. "Somatomotor-visual resting state functional connectivity increases after 2 years in the UK Biobank longitudinal cohort," *Journal of Medical Imaging* 11(2), 024010 (12 April 2024).
- Orlichenko A, Qu G, Zhang G, Patel B, Wilson TW, Stephen JM, Calhoun VD, Wang YP. Latent Similarity Identifies Important Functional Connections for Phenotype Prediction. *IEEE Trans Biomed Eng.* 2023 Jun;70(6):1979-1989. doi: 10.1109/TBME.2022.3232964. Epub 2023 May 19.
- Orlichenko A, Daly G, Zhou Z, Liu A, Shen H, Deng HW, Wang YP. ImageNomer: Description of a functional connectivity and omics analysis tool and case study identifying a race confound. *Neuroimage Rep.* 2023 Dec;3(4):100191. doi: 10.1016/j.ynirp.2023.100191. Epub 2023 Nov 7.
- Orlichenko A, Qu G, Zhou Z, Liu A, Deng HW, Ding Z, Stephen JM, Wilson TW, Calhoun VD, Wang YP. A Demographic-Conditioned Variational Autoencoder for fMRI Distribution Sampling and Removal of Confounds. *ArXiv [Preprint]*. 2024 May 13:arXiv:2405.07977v1.
- Qu G, Orlichenko A, Wang J, Zhang G, Xiao L, Zhang K, Wilson TW, Stephen JM, Calhoun VD, Wang YP. Interpretable Cognitive Ability Prediction: A Comprehensive Gated Graph Transformer Framework for Analyzing Functional Brain Networks. *IEEE Trans Med Imaging*. 2024 Apr;43(4):1568-1578. doi: 10.1109/TMI.2023.3343365. Epub 2024 Apr 3.

