Prospectus

Analysis of Functional Connectivity and its Fusion with Multi-Omics Using Contrastive Learning

Anton Orlichenko
Advisor: Yu-Ping Wang

June 2023

Outline

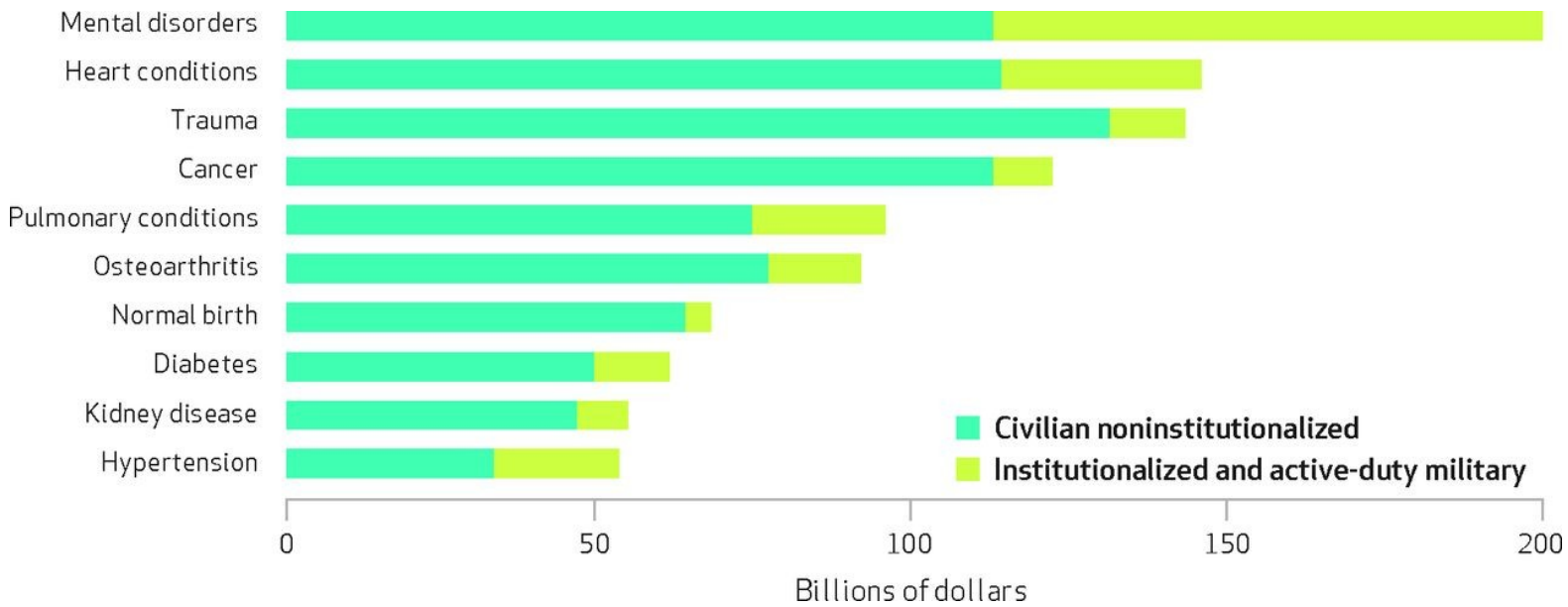
TULANE UNIVERSITY

Background on fMRI and Cognitive Science

TULANE UNIVERSITY

Clinical Problem

- Schizophrenia, ADHD, depression, and other mental illnesses cost the U.S. \$201+ billion annually[1]
- Dementia and Alzheimer's cost the U.S. \$157+ billion annually[2]
- Diagnosis of these diseases may be unreliable until symptoms become severe, when treatment options are more limited



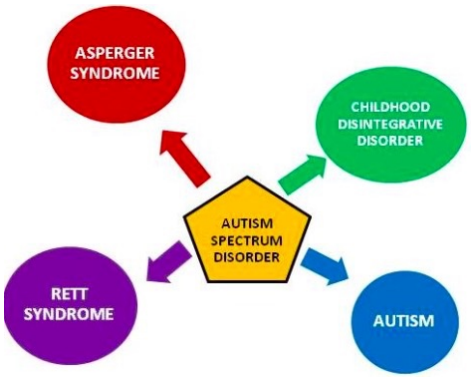
From Roerhig 2016.

[1] Roerhig 2016 https://doi.org/10.1377/hlthaff.2015.1659

[2] Hurd et al. 2013 doi:10.1056/NEJMsa1204629

Statistics

- 1 in 300 people are living with schizophrenia[3]
- 1 in 10 children may be affected by ADHD[4]
- 1 in 10 people may have had a major depressive episode in the past several years[5]
- 1 in 36 children may be diagnosed with autism spectrum disorder[6]
- 1 in 9 Americans over 65 years old are living with Alzheimer's[7]



[8]

[3] Desai et al. 2013 10.1111/jphs.12027

[4] Dincer et al. 2022 10.1177/10870547221099963

[5] Brody et al. 2018 Prevalence of depression among adults aged 20 and over

[6] Maenner et al. 2020 10.15585/mmwr.ss7011a1

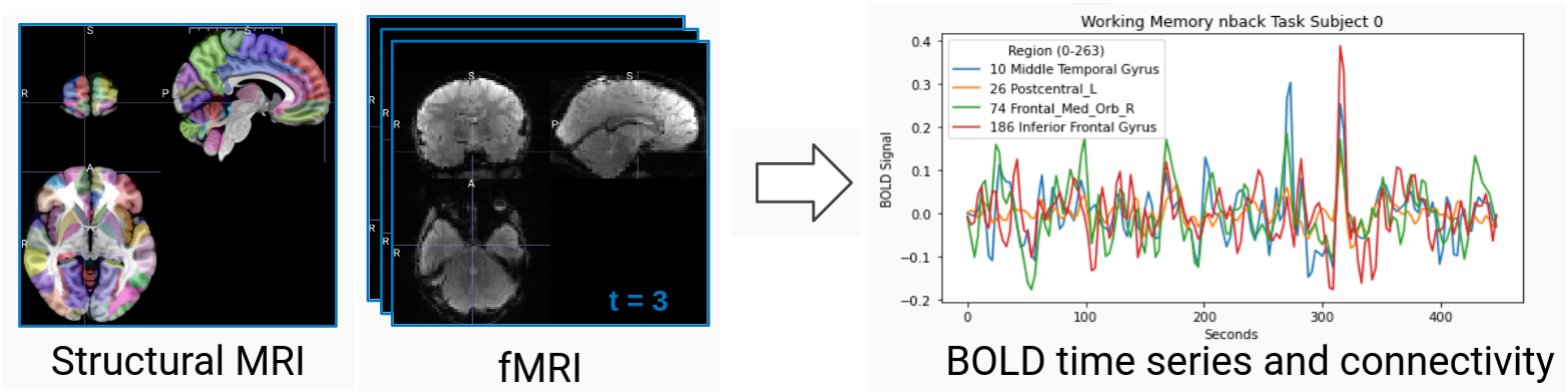
[7] 2023 Alzheimer's disease facts and figures

[8] https://speechandot.com/what-are-the-types-of-autism-spectrum-disorder/

TULANE UNIVERSITY

fMRI

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

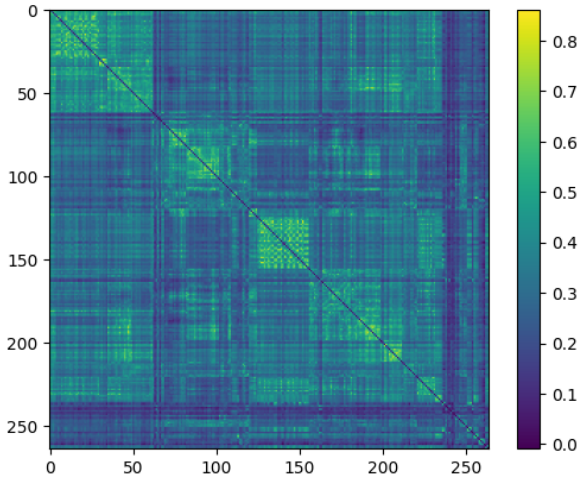


Example extraction of BOLD timeseries from fMRI volumes.

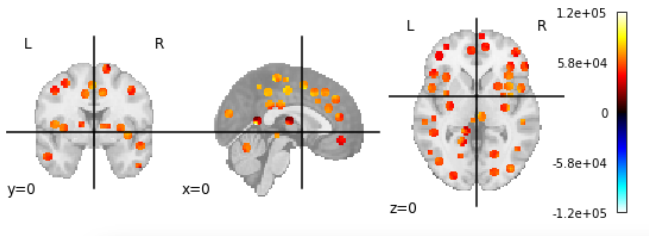
TULANE UNIVERSITY

Functional Connectivity (FC)

- Functional connectivity (FC) is the temporal Pearson correlation of BOLD signal between brain regions
- These brain regions are typically found using either ICA[9] or an atlas[10]
- FC or a variant is the most common starting point for predictive fMRI studies



mean FC of subjects in dataset



using 264 region Power template

[9] Calhoun et al. 10.1016/j.neuroimage.2008.10.057

[10] Power et al. 10.1016/j.neuron.2011.09.006

TULANE UNIVERSITY

Power Atlas Functional Networks

- Example of Power atlas partition of brain into functional networks

Functional Networks

ROIs		ROIs	
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

TULANE UNIVERSITY

fMRI and Mental Health

- fMRI has been used to predict disease status and (endo)phenotypes such as age, sex, and general fluid intelligence[11]
- Machine learning predictions of brain age have been correlated with future Alzheimer's diagnosis[12]
- Most diagnoses of mental disorders are still made by psychiatrists or physicians based on cognitive tests[13]

[11] Qu et al. 2021 10.1109/TBME.2021.3077875

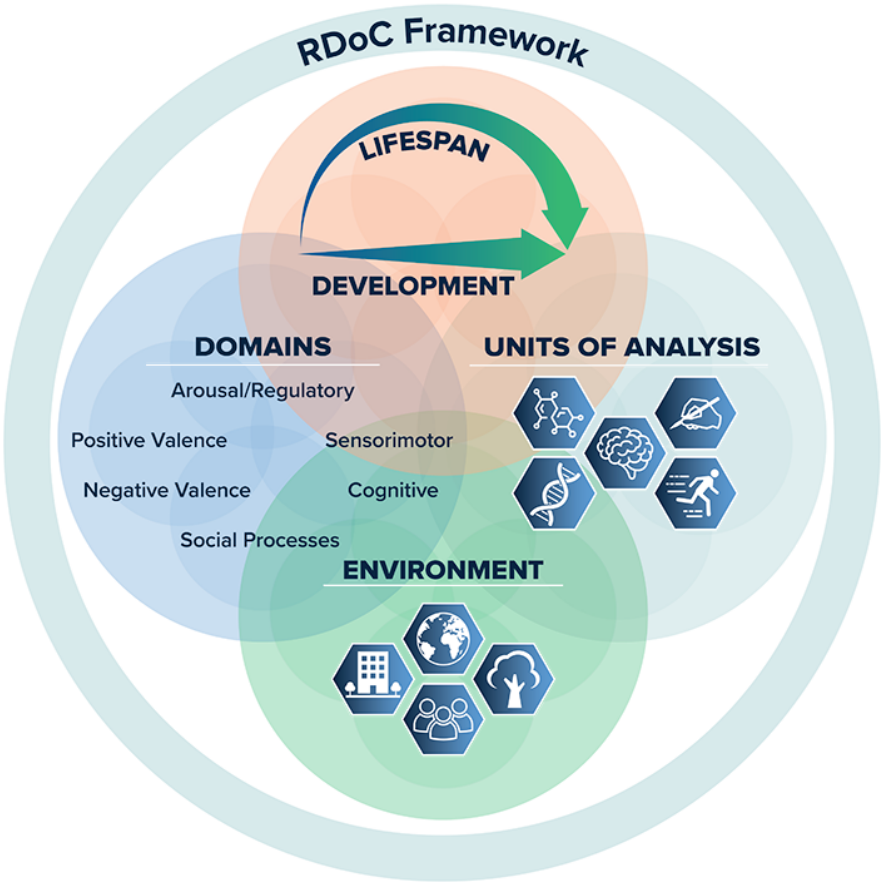
[12] Millar et al. 2022 10.1016/j.neuroimage.2022.119228

[13] https://www.ndcn.ox.ac.uk/divisions/fmrib/what-is-fmri/how-is-fmri-used

TULANE UNIVERSITY

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 basis[14]



[15]

[14] Insel et al. 10.1176/appi.ajp.2010.09091379

[15] https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/about-rdoc

fMRI in the Clinic

- Clinically, fMRI is used for pre-surgical planning and to follow the progression of concussion[16]

Example of a clinical lab setup for monitoring concussion[17]

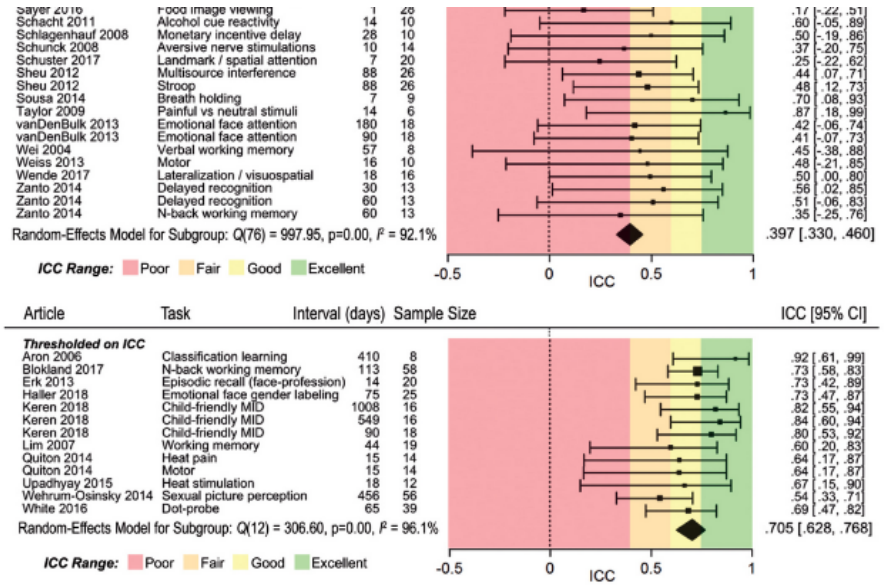
[16] Kaushal et al. 2019 10.1002/hbm.24440

[17] https://www.cognitivefxusa.com/blog/fmri-brain-scans-duke-study-implications

TULANE UNIVERSITY

Potential Pitfalls

- Voxel-based fMRI analysis has been criticized for having poor test-retest replicability[18]
- This may explain why FC has become popular as a predictive tool



[18] Elliott et al. 2020 10.1177/0956797620916786

Problem Statement and Goals

TULANE UNIVERSITY

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 **two challenges** and **one opportunity**.

TULANE UNIVERSITY

Challenges

High Cost of Data Acquisition

The cost of acquiring an fMRI scan for a single subject is \$500-\$1000.[a]

[a] Szucs and Ioannidis 2020 10.1016/j.neuroimage.2020.117164

- Same magnitude of cost for acquiring genomic data.

Small Sample Size, 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 $> 10^4$

[a] Szucs and Ioannidis 2020 10.1016/j.neuroimage.2020.117164

TULANE UNIVERSITY

Opportunity

Large Amount of Unlabeled Data

Repositories exist for large amounts of unlabeled or out-of-distribution data

- UKB has 50,000 subjects with fMRI scans from a largely homogenous population[a]
- OpenNeuro has 800+ small open access datasets with varied demographics[b]

---

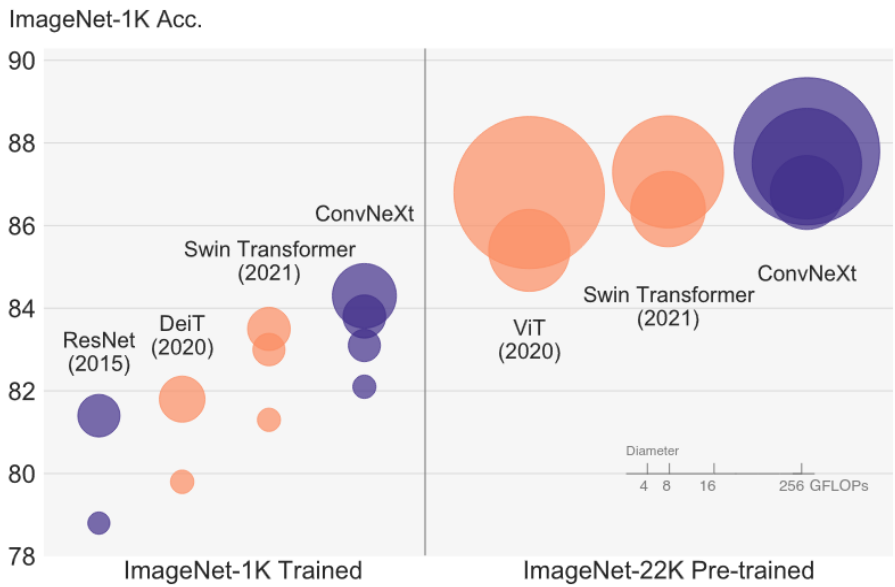
[a] Sudlow et al. 2015 10.1371/journal.pmed.1001779

[b] Markiewicz et al. 2021 10.7554/eLife.71774

TULANE UNIVERSITY

Idea: Pre-training Improves Performance

- It has been shown in the image classification and natural language processing AI literature that pre-training models on large unlabeled datasets improves performance

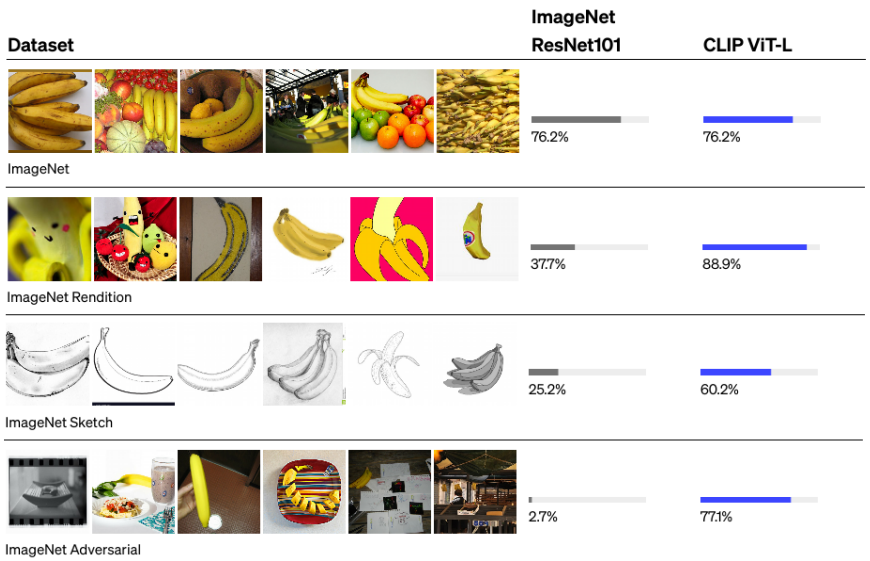


Pre-training yields a 4-6% improvement in image classification.[19]

[19] Liu et al. 2022 10.48550/arXiv.2201.03545

Contrastive Learning Fuses Modalities

- Pre-training learns important features and often falls under the umbrella of **contrastive learning**
- CLIP fuses features of natural language and images, allowing, e.g., AI-generate art



Contrastive Language Image Pretraining.[20]

[20] https://openai.com/research/clip

Goals

The goals of this dissertation research are the following:

1. **Aim 1: Improve model performance** in the small sample size, high dimensionality regime
2. **Aim 2: Create tools** for data exploration, quality control, and detection of unreported covariates/confounders
3. **Aim 3: Construct a generative model** for the decomposition of FC to create augmentations for use in pre-training
4. **Aim 4: Find associations** between brain networks and genotype or gene expression using contrastive learning

TULANE UNIVERSITY

Specific Aims

Aim 1: Latent Similarity for Small Sample Size, High Dimensionality Datasets

TULANE UNIVERSITY

Problem of High Dimensionality

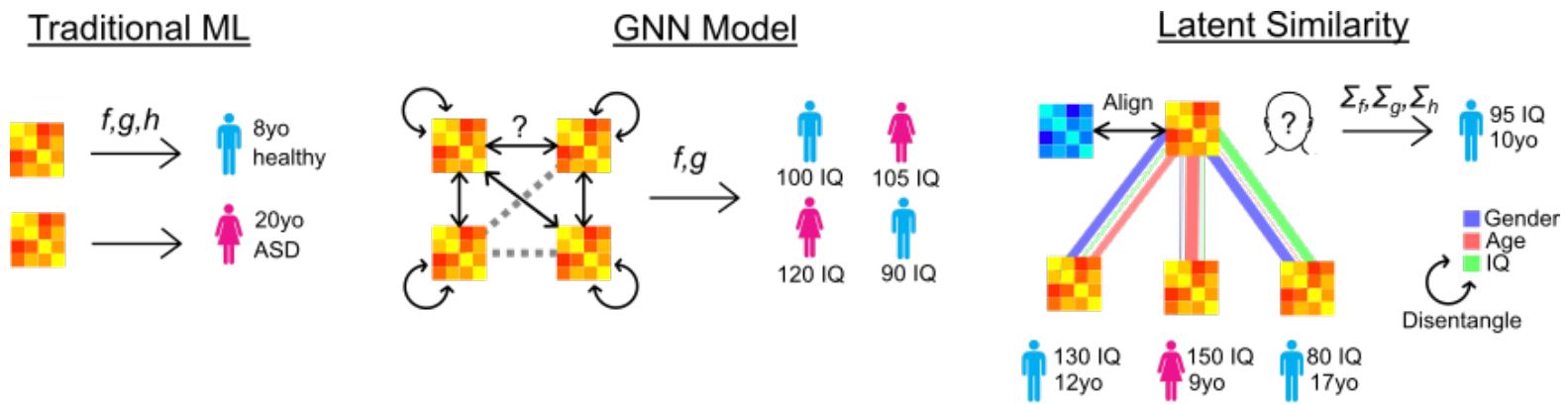
Question

Can we create a model that is robust with very low sample sizes and high dimensional features?

- Makes very conservative predictions with a small training set
- Increases in predictive power with a large training set

TULANE UNIVERSITY

Solution: Latent Similarity



- As in contrastive learning, we explicitly model subject-subject latent similarities[21][22]
- Traditional ML models (Ridge, MLP, CNN) do not explicitly consider inter-subject relationships
- GNN models require many degrees of freedom to estimate edges

[21] Orlichenko et al. 2022 10.1109/TBME.2022.3232964

[22] Tool available at `https://github.com/aorliche/LatentSimilarity`

TULANE UNIVERSITY

Similarity Kernel

- We model a linear similarity kernel to project data to a very low dimensional subspace
- Followed by softmax aggregation
- Can perform cross-modality alignment and disentanglement
- A similar approach was recently used in the image domain[23]

$$\begin{aligned} \text{sim}(a, b) &= \langle \phi(\mathbf{x}_a), \phi(\mathbf{x}_b) \rangle \\ &= \mathbf{x}_a \mathbf{A}\mathbf{A}^\mathsf{T} \mathbf{x}_b^\mathsf{T}, \\ \mathbf{E} &= S_{Row}((\mathbf{1} - \text{diag}(\infty)) \odot \mathbf{X}\mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{X}^\mathsf{T}), \end{aligned} \tag{1}$$

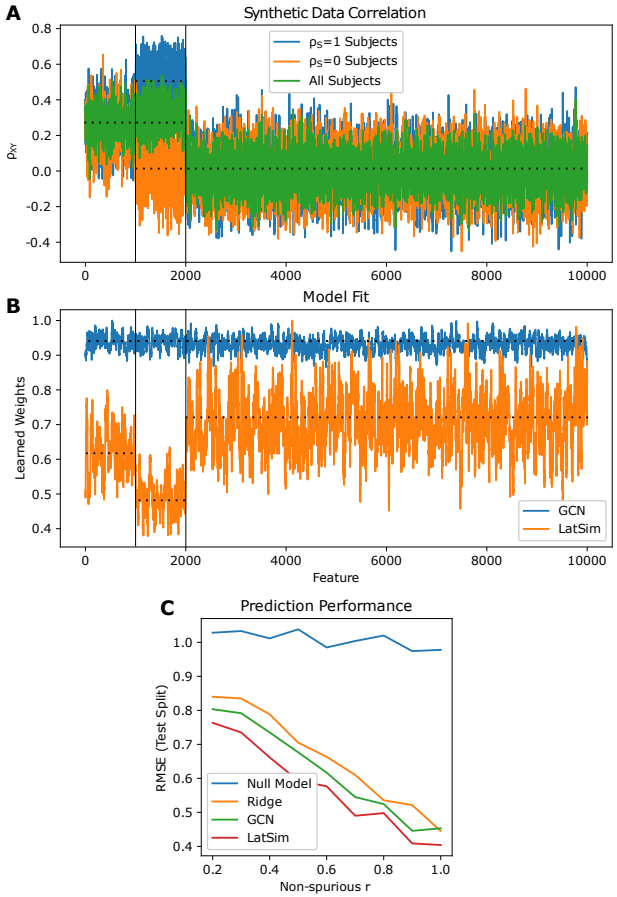
$$S(\mathbf{z})_i = \frac{e^{z_i/\tau}}{\sum_{j=0}^{N} e^{z_j/\tau}},$$

TULANE UNIVERSITY

[23] Zheng et al. CVPR 2022 10.48550/arXiv.2203.06915

Simulation Results

- We find that LatSim performs better in simulation studies on synthetic data



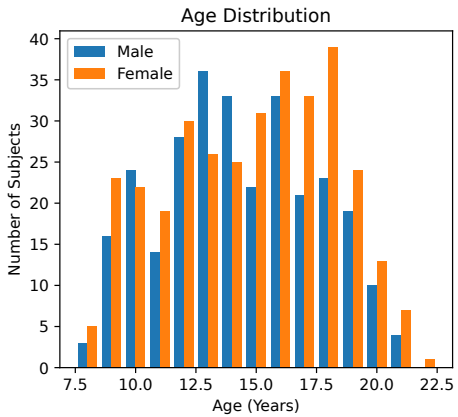
TULANE UNIVERSITY

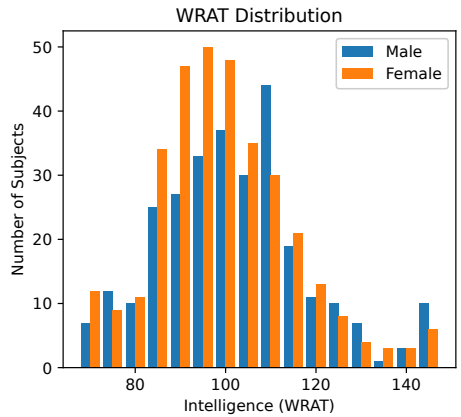
PNC Dataset

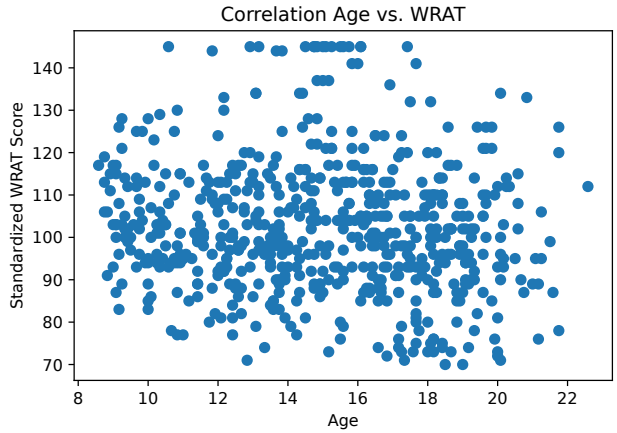
- We test on Philadelphia Neurodevelopmental Cohort (PNC), using reduced number of subjects

	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

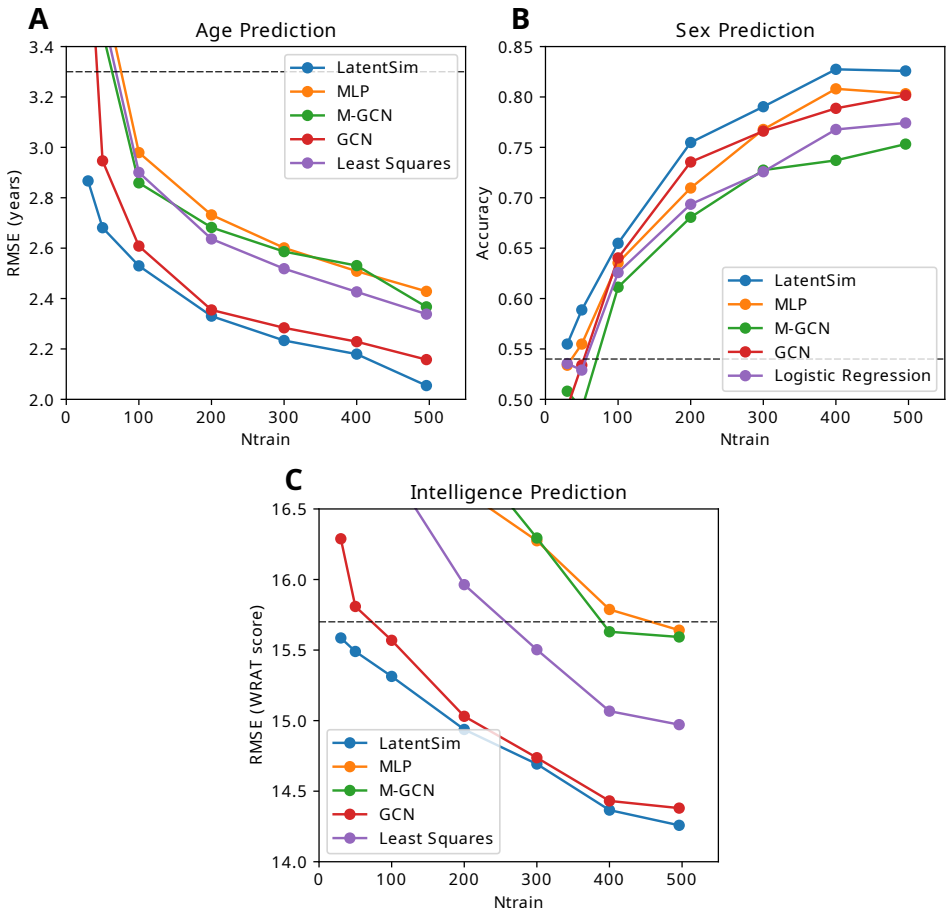






Latent Similarity FC Prediction

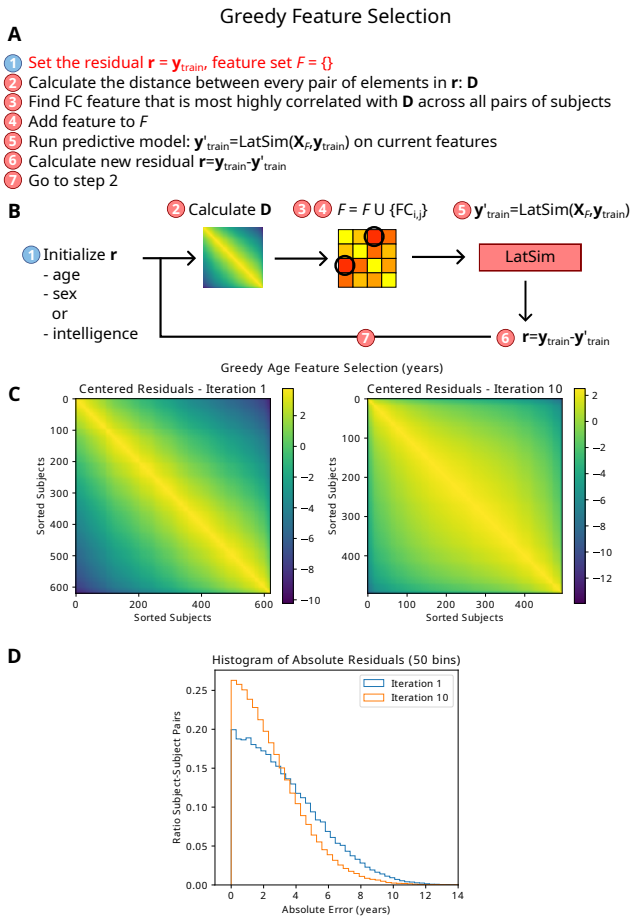
- We achieve superior predictive accuracy, especially at small sample sizes (predict age, sex, and WRAT score)



TULANE UNIVERSITY

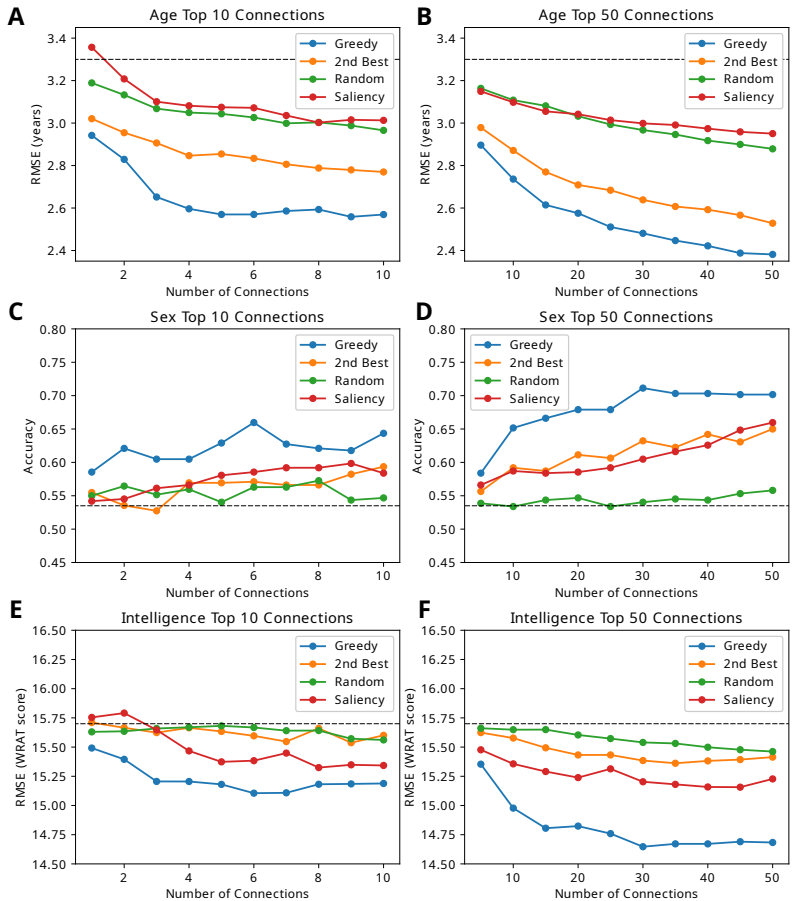
Latent Similarity Greedy Selection

- We include a greedy feature selection algorithm which uses LatSim in its inner loop



Superior Performance of Greedy Selection

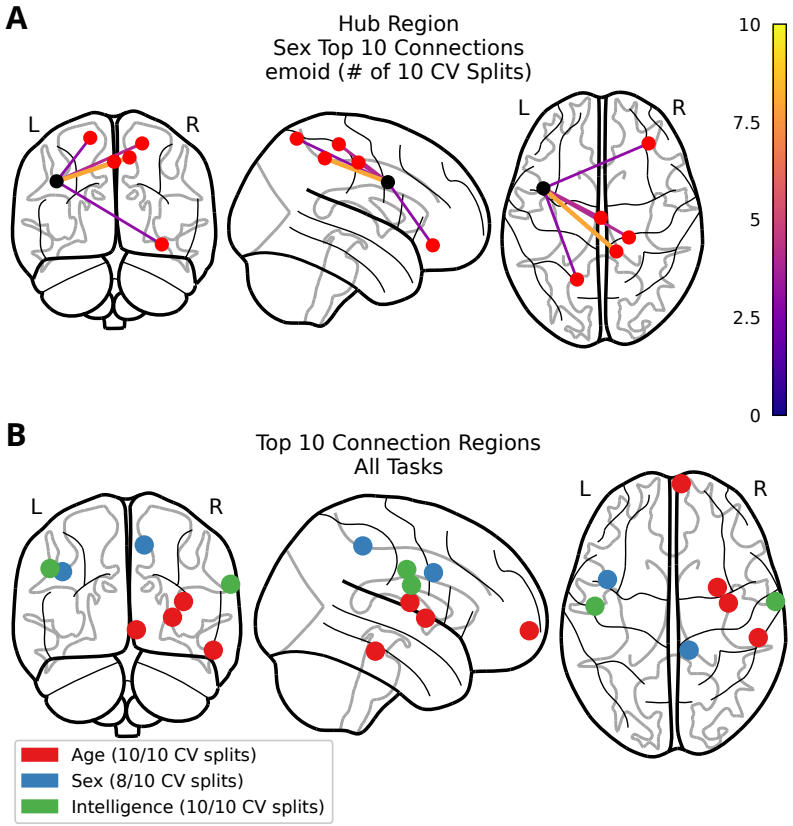
- We found bootstrap greedy selection features to be superior to 2nd choice, random features, and saliency (gradient-based method)



TULANE UNIVERSITY

Key Connections

- We identified several key connections useful for each predictive task
- However, these connections may be replaced by others with small to medium predictive penalty



TULANE UNIVERSITY

Computational Benefits

- LatSim is much faster than all but Ridge or Least Squares models
- Why choose a slow model to get similar or worse results?

Table: Training time for all 10 folds of 10-fold cross validation.

Model	LatSim	Lstsq	Logistic	GCN	MLP	M-GCN
Epochs	200	-	100	1e4	1e4	5e3
Training Time	**4.3s**	**<1s**	63.4s	406s	364s	5912s

Latent Similarity Takeaways

Proof of Concept for Similarity Method

- We show that similarity-based methods, at the core of contrastive learning and data fusion, can achieve superior results at small sample sizes
- LatSim also competitive with other models at larger sample sizes
- Much faster than, e.g., scikit-learn implementation of logistic regression

TULANE UNIVERSITY

Aim 2: ImageNomer, A Data Exploration Tool Reveals Racial Confound in fMRI Data

TULANE UNIVERSITY

Problem of Dataset Exploration and Quality Control

Question

- How to quickly browse through and validate FC and genomic data?
- How to find correlations in phenotypes and visualize importance of features for prediction?

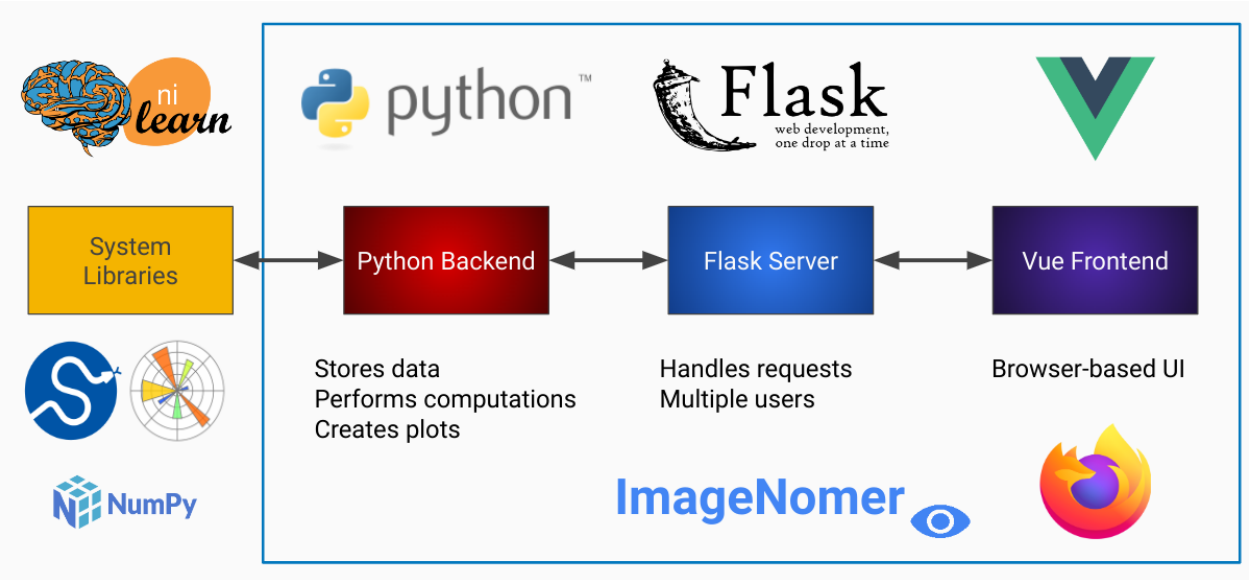
Preview of Findings

- We create an easy-to-use tool for exploring and visualizing correlations in an fMRI dataset
- We find a previously ignored confound in fMRI data affecting achievement score prediction

TULANE UNIVERSITY

Our Solution

- We create a browser-based tool called ImageNomer to visualize subjects and get statistics about a dataset

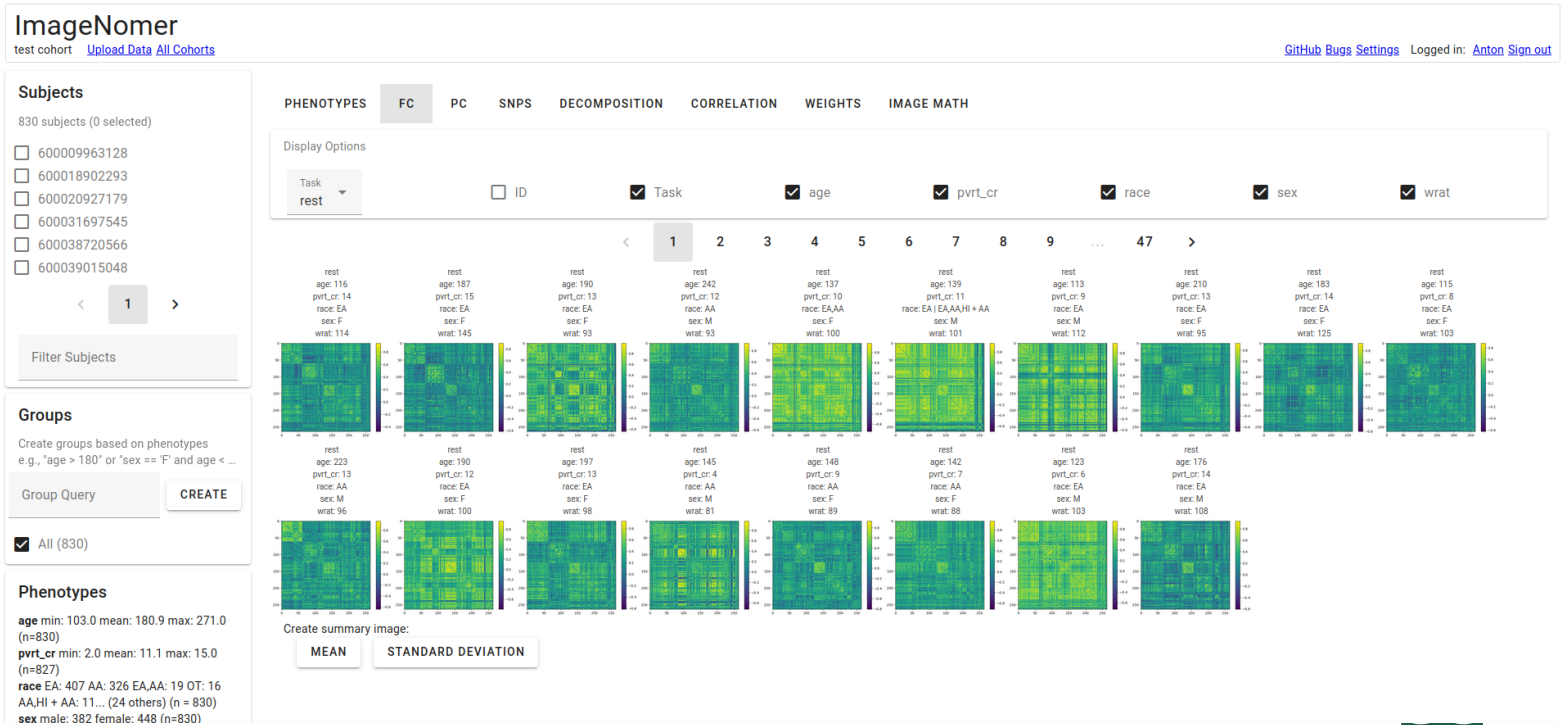


Architecture of the tool.[24]

[24] Orlichenko et al. 2023 10.1117/12.2654311

ImageNomer Main View

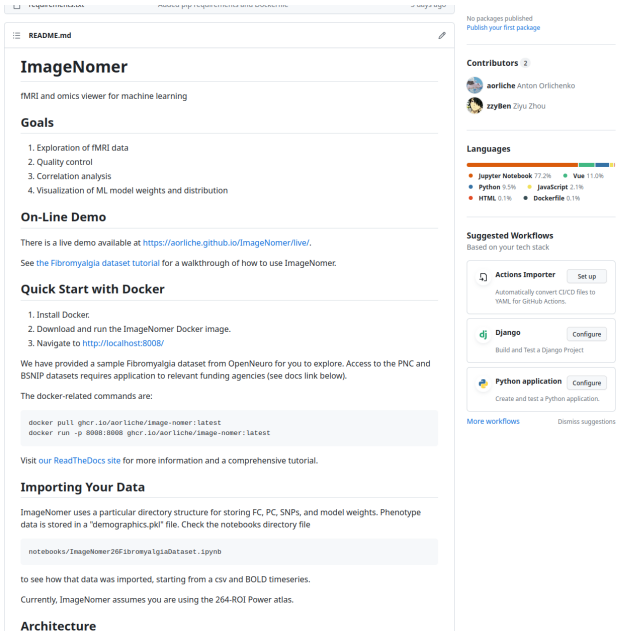
- A screenshot of the main FC view in ImageNomer.
- Tool is available at [25]. See links for tutorial.
- `https://aorliche.github.io/ImageNomer/live/`

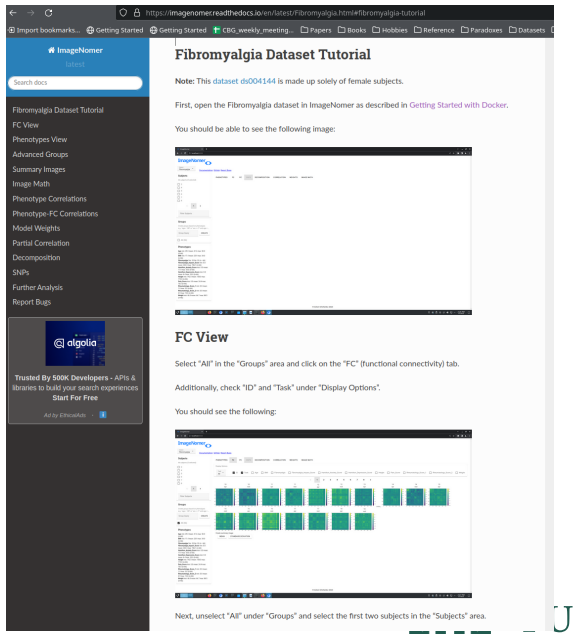


[25] `https://github.com/TulaneMBB/ImageNomer`

ImageNomer Documentation

- On-line documentation available at ReadTheDocs[26]
- Tutorial walks through how to use features of ImageNomer
  - Phenotype distribution
  - Phenotype-phenotype or phenotype-FC correlation
  - Visualization of model weights
  - Image math

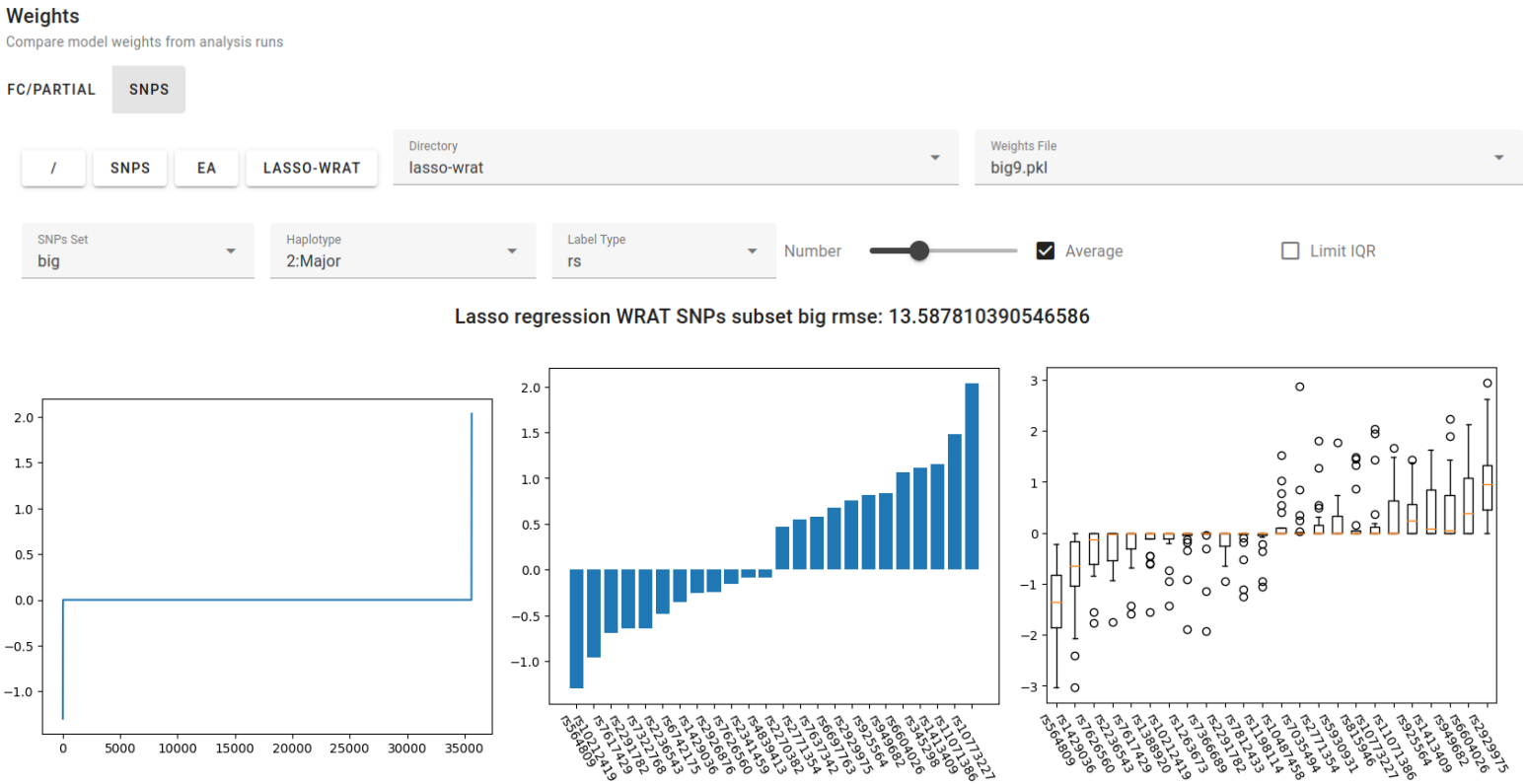




[26] https://imagenomer.readthedocs.io/en/latest/index.html

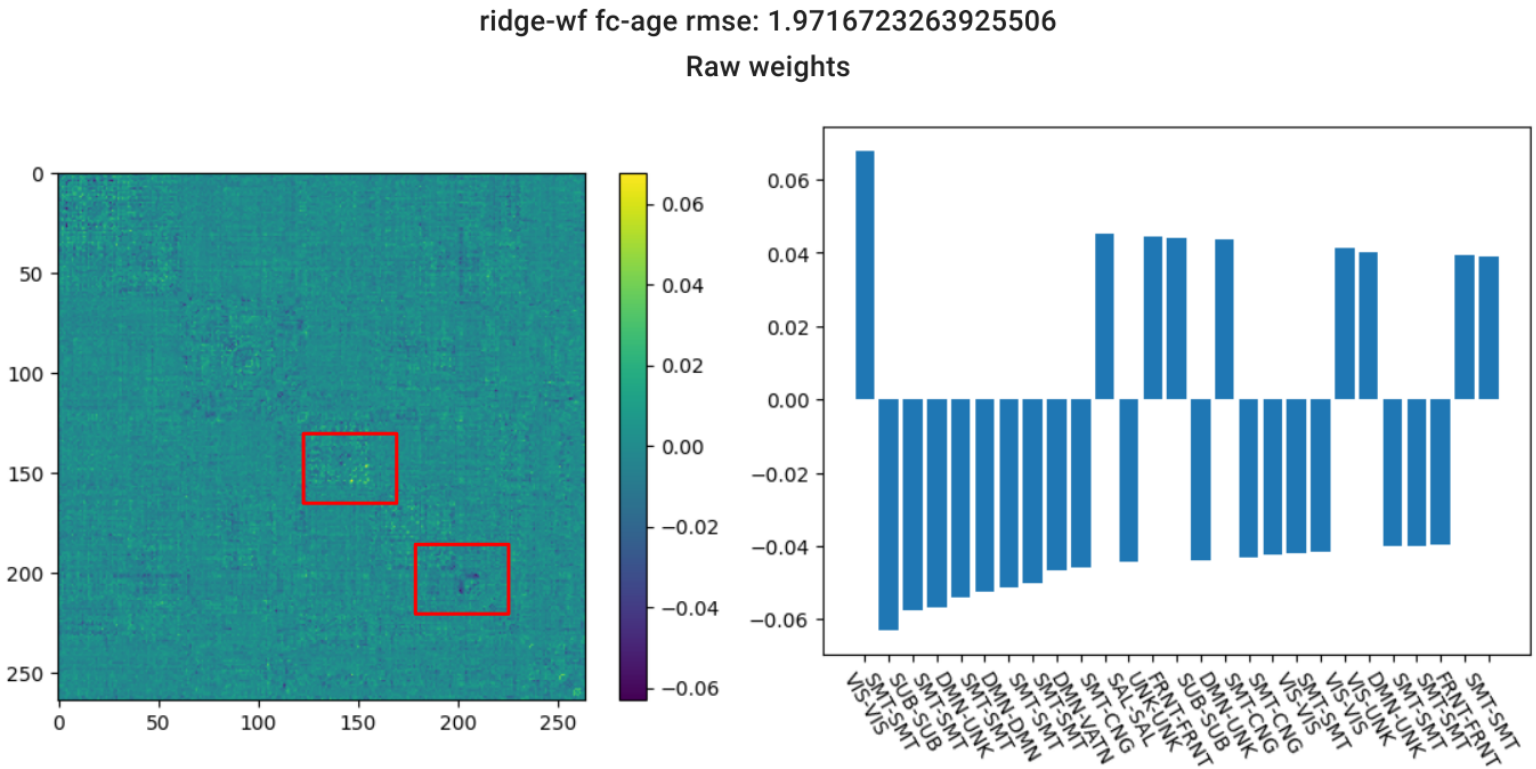
ImageNomer SNP Visualization

- Besides fMRI data, we can process SNP information



ImageNomer FC Model Weights Visualization

- Using the ImageNomer GUI, one can identify which regions machine learning models flag as important for prediction

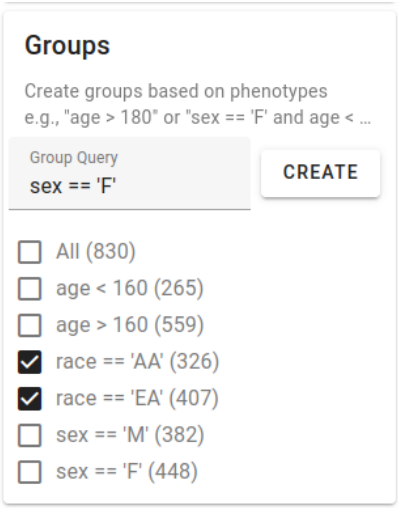


TULANE UNIVERSITY

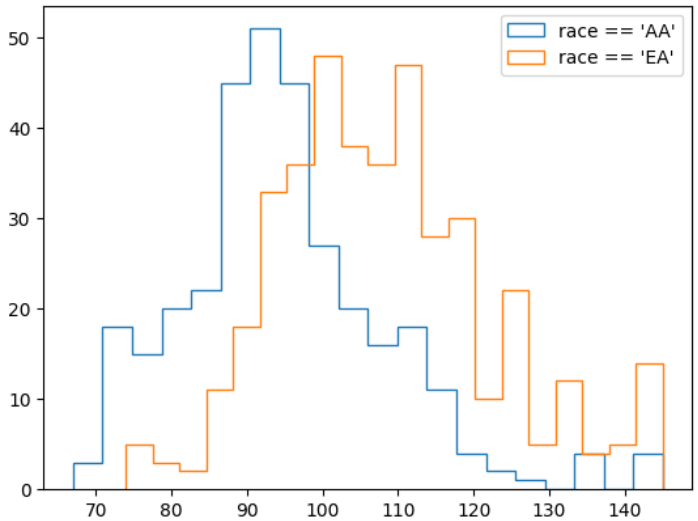
Using Phenotype Explorer, Potential Race Confound is Apparent

- By creating plots of phenotype correlations, we see race or SES is a potential confound in WRAT score[27] prediction
- ...as long as FC data can be used to predict race

Group Selection Panel



Age-Regressed WRAT Score

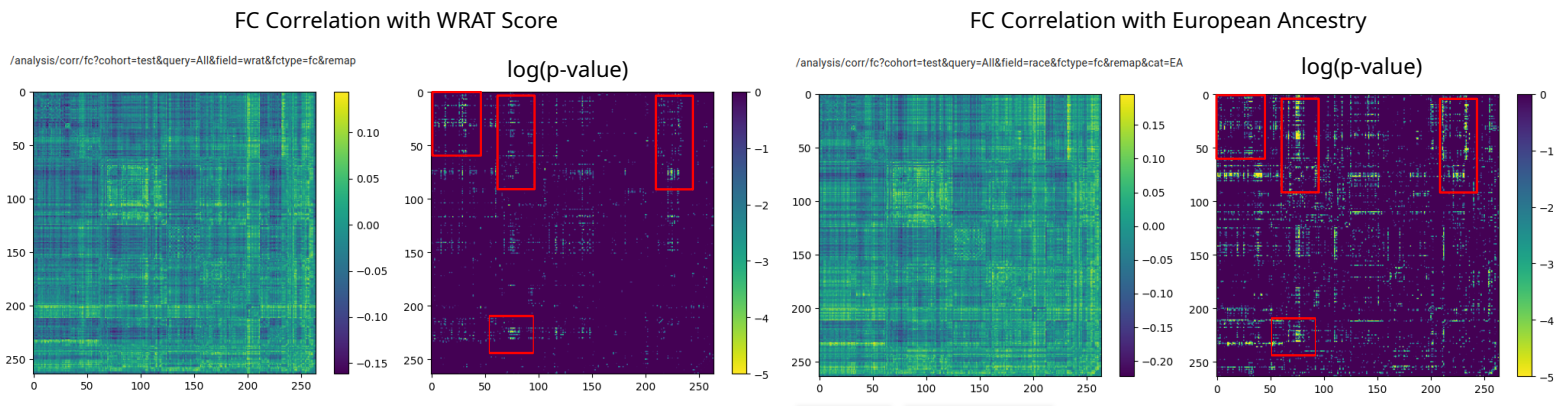


TULANE UNIVERSITY

[27] Sayegh et al. 2014 10.1093/arclin/acu059

Achievement-FC Correlation is Subset of Race-FC Correlation

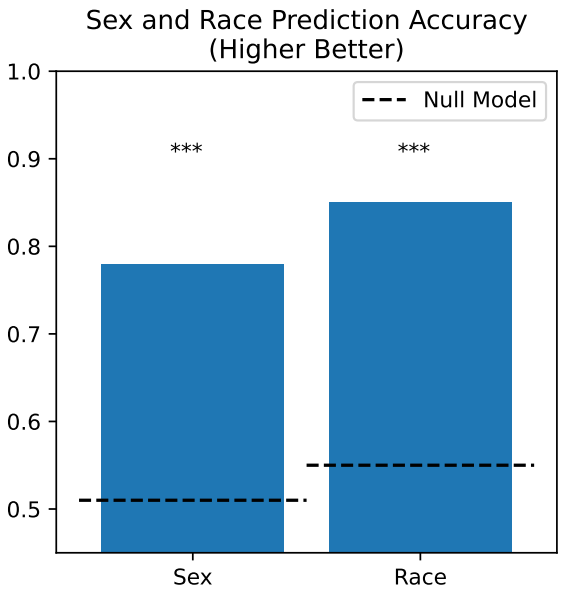
- We perform correlation analysis between FC and WRAT score and apply Bonferroni-corrected t-test
- We find achievement-FC correlation is a subset of race-FC correlation

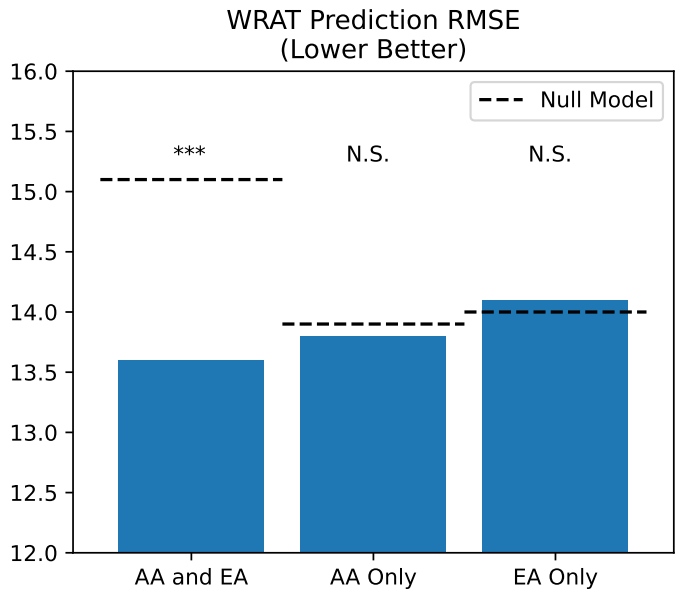


TULANE UNIVERSITY

Race Can Be Predicted, Unbiased Achievement is Harder

- We test the effect of the confound by predicting WRAT score from FC on whole cohort and within race groups
- When controlling for race, WRAT score prediction not significantly better than guess





TULANE UNIVERSITY

ImageNomer Takeaways

Proof of Concept for Large Dataset and Multi-Omics Work

- We create an omics and fMRI visualization tool that requires minimal programming experience to use
- We show that we are able to analyze datasets with
  - thousands of subjects
  - hundreds of demographic features
  - very high FC and SNP dimensionality
- In the process, we find a confound that was unaddressed in and potentially contaminated previous work

TULANE UNIVERSITY

Aim 3: Angle Basis: A Generative Model and Decomposition of Functional Connectivity

TULANE UNIVERSITY

Problem of fMRI Data for Contrastive Learning

Question

Contrastive learning typically uses augmentations or different views of images/text to create positive/negative pairs

- How to create data-driven augmentations for fMRI data?

Preview of Findings

- We create a theoretically-motivated generative model for FC
- Model has several uses besides augmentations

TULANE UNIVERSITY

Angle Basis Overview

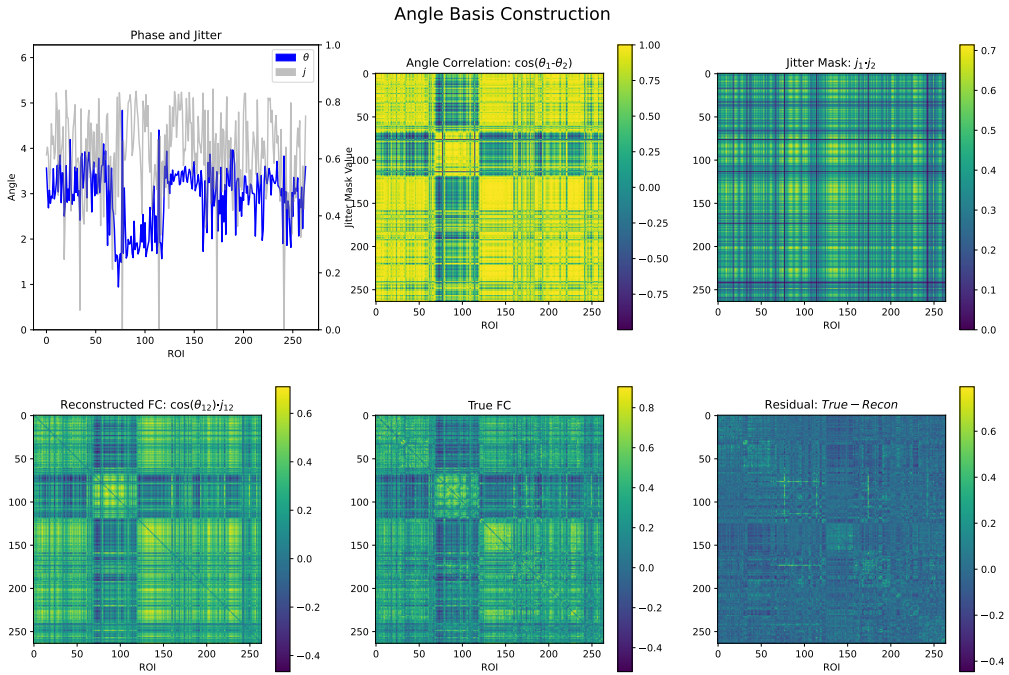
Key Features

1. 10x data compression
2. 97.3% fingerprinting identification (residual) vs 62.5% for FC
3. Modest 5% AUC prediction improvement over FC
4. Data augmentation for contrastive learning
5. Generation of synthetic FC from user-input patient characteristics (network-level explainability/interpretability)
6. Does not rely on knowledge of a population; a single subject is enough

TULANE UNIVERSITY

Example of Angle Basis Decomposition

- In searching for network level explanations for FC-based prediction, we created a simple yet powerful decomposition and compression of FC[28]



Tool available at [29]

[28] Orlichenko et al. 2023 10.48550/arXiv.2305.10541

[29] https://github.com/aorliche/AngleBasis

Angle Basis Decomposition

- Based on the well-known phase lock value (PLV)

$$
\begin{aligned}
\mathrm{H}[x](t) &= \frac{1}{\pi}\mathrm{p.v.}\int_{-\infty}^{\infty}\frac{x(\tau)}{t-\tau}d\tau \\
x_a(t) &= x(t) + i\mathrm{H}[x(t)] \\
x_a(t) &= a(t)e^{i\theta(t)} \\
\theta_{cd}(t) &= \theta_c(t) - \theta_d(t) \\
\mathrm{PLV}_{cd} &= \frac{1}{T}\left|\sum_{t=1}^{T} e^{i\theta_{cd}(t)}\right|
\end{aligned}
\tag{2}
$$

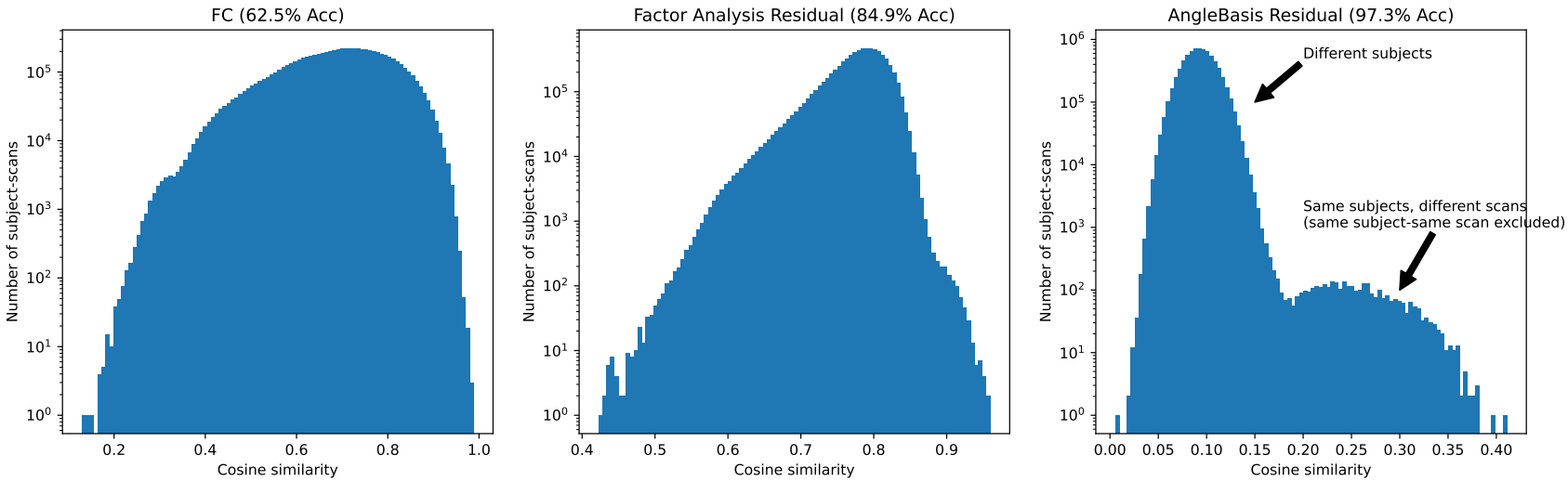
- We use a basis of orthogonal sine waves and jitter mask to compress and reconstruct FC

$$
\begin{aligned}
\hat{\rho}_{cd}^{(n)} &= j_c^{(n)} \cdot j_d^{(n)} \cdot \cos(\theta_c^{(n)} - \theta_d^{(n)}) \\
\tilde{\rho}_{cd} &= \frac{1}{N}\sum_{n=1}^{N}\hat{\rho}_{cd}^{(n)}
\end{aligned}
\tag{3}
$$

TULANE UNIVERSITY

Fingerprinting Accuracy

- We find that the residual of angle basis has 97.3% accuracy in identifying the same subject from different scans
- Compared to 62.5% for FC



TULANE UNIVERSITY

Classification Improvement

- On their own, angle basis and residual have equal predictive ability compared to FC
- Ensemble of the two is significantly better than FC

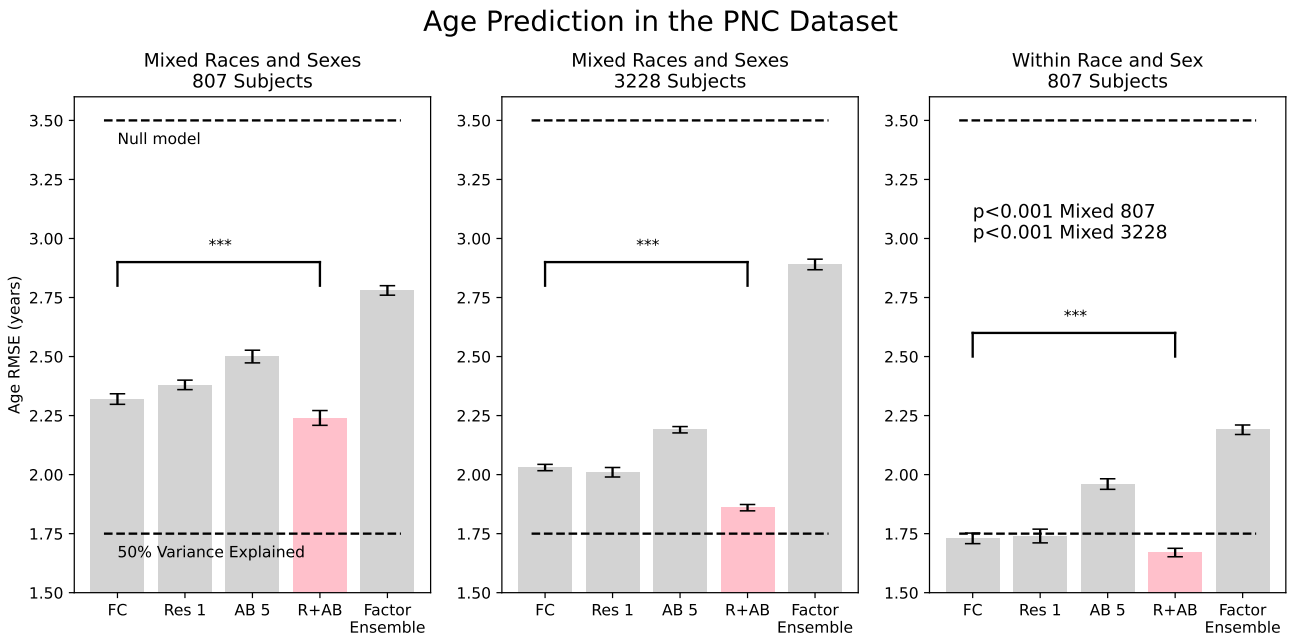
Classification Accuracy (AUC)

Dataset	Predictive Task	FC	Deep AE Ens	AB+Res Ens	p-value
BSNIP	SZ/NC	0.785 ± 0.048	0.701 ± 0.062	$\mathbf{0.804 \pm 0.040}$	0.240
BSNIP	Sex	0.755 ± 0.023	0.645 ± 0.085	$\mathbf{0.791 \pm 0.022}$	0.002
BSNIP	Race	0.845 ± 0.022	0.739 ± 0.051	$\mathbf{0.866 \pm 0.023}$	0.108
PNC	Sex	0.886 ± 0.006	0.744 ± 0.065	$\mathbf{0.923 \pm 0.010}$	< 0.001
PNC	Race	0.946 ± 0.007	0.812 ± 0.040	$\mathbf{0.973 \pm 0.003}$	< 0.001
BSNIP→PNC	Sex	0.667 ± 0.017	0.601 ± 0.032	$\mathbf{0.700 \pm 0.013}$	< 0.001
BSNIP→PNC	Race	0.807 ± 0.018	0.710 ± 0.010	$\mathbf{0.847 \pm 0.012}$	< 0.001
PNC→BSNIP	Sex	0.629 ± 0.019	0.572 ± 0.019	$\mathbf{0.667 \pm 0.013}$	< 0.001
PNC→BSNIP	Race	0.800 ± 0.010	0.702 ± 0.022	$\mathbf{0.832 \pm 0.009}$	< 0.001

TULANE UNIVERSITY

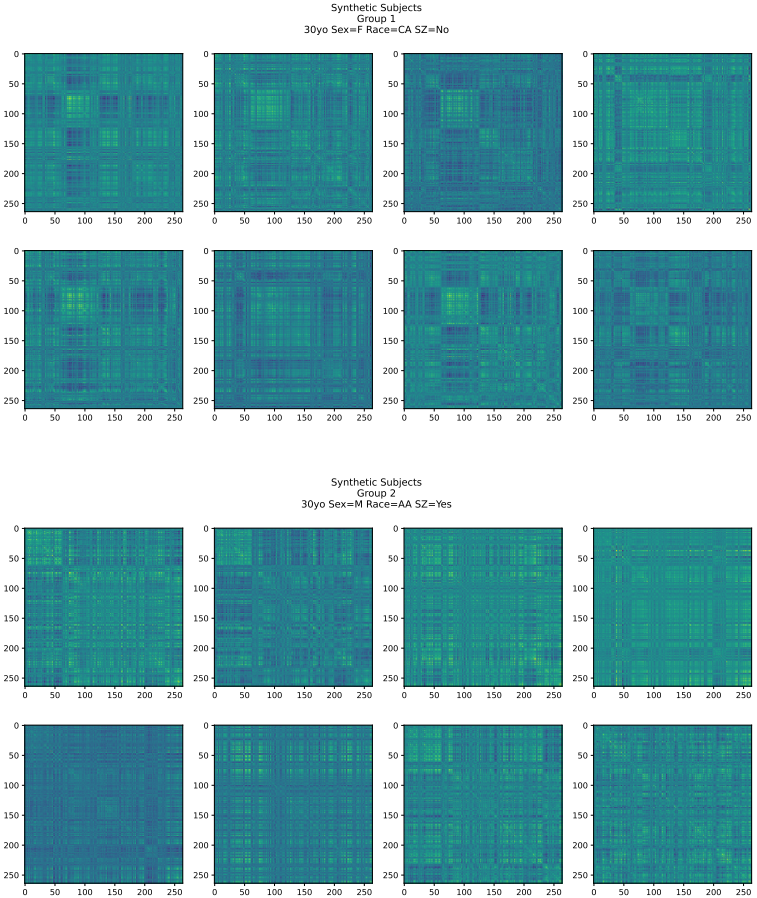
Importance of Within-Group Predictions

- Age regression results also show better prediction error with ensemble of angle basis and residual compared to just FC
- Moreover, we see the importance of making prediction within sex and race-matched groups



TULANE UNIVERSITY

FC Generation for Explainability

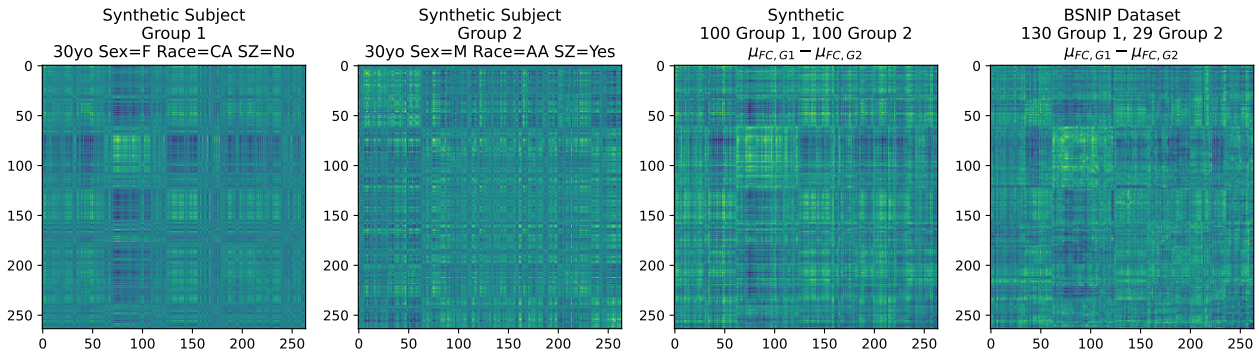


First 8 synthetic subjects from 30yo normal CA female group and 30yo schizophrenic AA male group.

TULANE UNIVERSITY

Validation on Group Differences

- Synthetic subjects have similar group differences compared to demographic-matched real subject FCs
- Synthetic generation allows for extrapolation as well as visualization

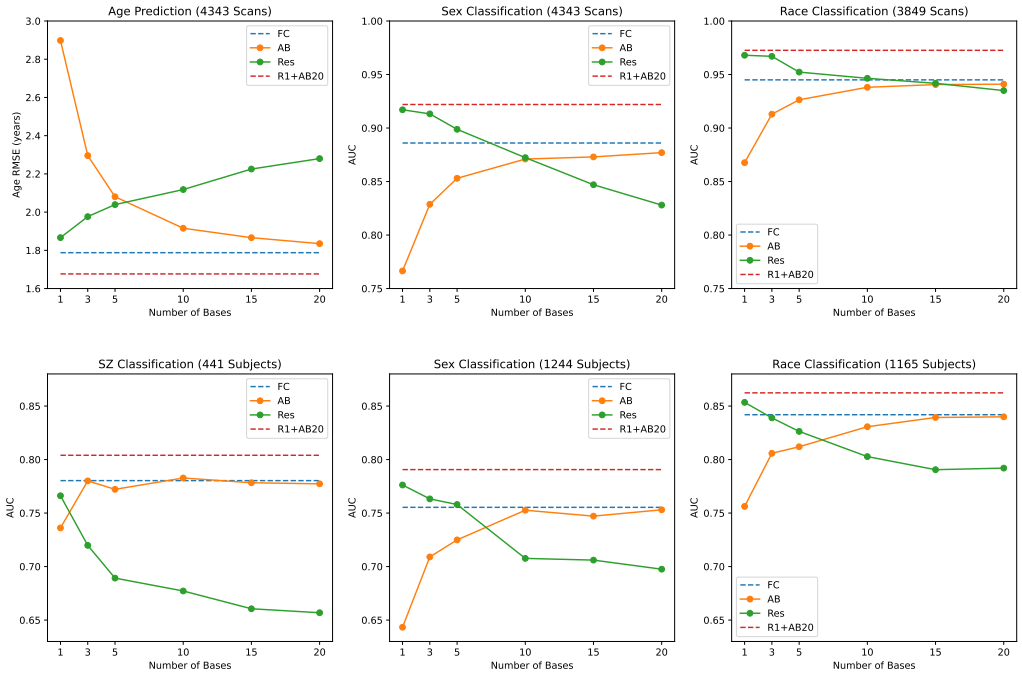


- Females are known to have increased intra-DMN connectivity[30]

[30] Ficek-Tani et al. 2022 10.1093/cercor/bhac491

Prediction Accuracy as a Function of Number of Bases

- Both angle basis and residual have predictive value
- Varies inversely as the number of sinusoidal bases and jitter masks is increased



AngleBasis Takeaways

Proof of Concept for Augmentation and Manipulation of FC

- We are able to decompose, compress, and generate FC
  - to identify subjects with 97.3% accuracy
  - to achieve a modest 5% classification AUC gain
  - to generate synthetic FC based on patient demographics, increasing intuition and explainability
- In the process, we are now able to perform theoretically-motivated augmentations, as in the mixup model[a]

[a] Zhang et al. 2017 10.48550/arXiv.1710.09412

TULANE UNIVERSITY

Aim 4: Contrastive Learning for Fusing Omics with Brain Imaging

TULANE UNIVERSITY

Problem of Leveraging Unlabeled Datasets for Discovering Causal Associations with Omics

Question

Can we use contrastive learning

- to leverage the large amounts of out-of-distribution data (UKB, OpenNeuro) available to create better models?
- to fuse brain imaging with multi-omics to discover links between omics and fMRI endophenotype?

Important

Work is only beginning on this Aim, thus many of the ideas in this section are speculative.

TULANE UNIVERSITY

Foundations of Contrastive Learning

- The basic foundation of contrastive learning is an objective function based on positive and negative pairs
- The InfoNCE (noise contrastive estimation) loss is shown below:

$$\mathcal{L}_{CL} = -\frac{1}{N}\sum_{i}^{N}\log\frac{e^{\mathbf{q}_i^T\mathbf{k}_i^+/\tau}}{e^{\mathbf{q}_i^T\mathbf{k}_i^+/\tau} + \sum_{j=1}^{K} e^{\mathbf{q}_i^T\mathbf{k}_{i,j}^-/\tau}} \tag{4}$$

- $\mathbf{q}_i$ is an image $\mathbf{x}_i$ mapped into a query latent via a query encoder $\mathbf{q}_i = f(\mathbf{x}_i)$ The goal is to learn $f(\cdot)$
- $\mathbf{k}_i$ is an image mapped into a key latent via a key encoder $g(\mathbf{x}_i)$ ($+$ and $-$ are positive and negative samples)
- N is the batch size
- K is the number of negatives per positive sample
- τ is a temperature for the Softmax function

LORAC

- fMRI data is very high dimensional and many studies have low sample sizes
- In the image domain, Wang et al. 2022[31] have proposed that salient features should lie on a low-dimensional manifold
- A low rank prior may be useful for our domain

$$\mathcal{L}_i = -\frac{1}{M-1}\sum_{m}^{M-1}\log\frac{e^{\mathbf{q}_{i,m}^{\mathsf{T}}\mathbf{k}_i^{+}/\tau}\cdot h(\mathbf{Q})}{e^{\mathbf{q}_{i,m}^{\mathsf{T}}\mathbf{k}_i^{+}/\tau}\cdot h(\mathbf{Q}) + \sum_{j=1}^{K} e^{\mathbf{q}_{i,m}^{\mathsf{T}}\mathbf{k}_{i,j}^{-}/\tau}} \tag{5}$$

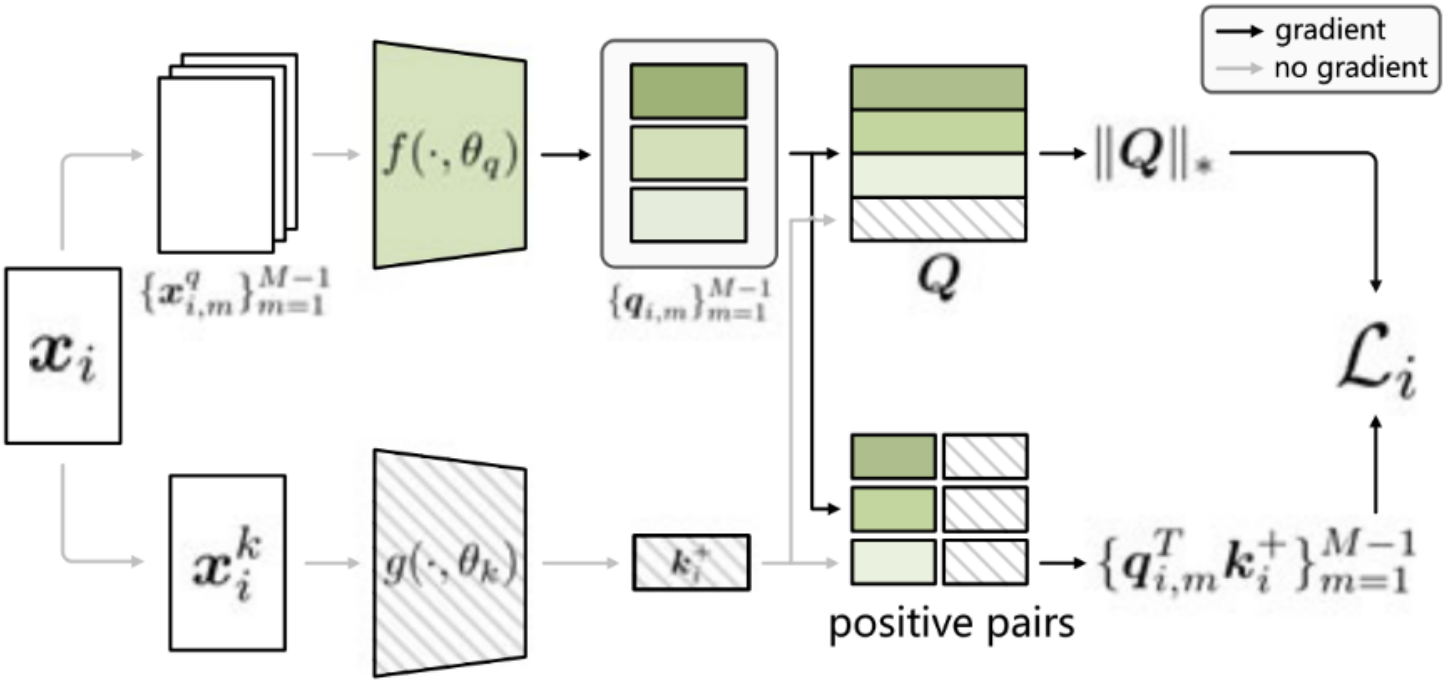
$$h(\mathbf{Q}) = e^{-\frac{||\mathbf{Q}||_*}{M\cdot\beta\cdot\tau}}$$

- The Low Rank Promoting Prior for Contrastive Learning (LORAC) modification to the InfoNCE loss

TULANE UNIVERSITY

[31] Wang et al. 2022 10.1109/TPAMI.2022.3180995

LORAC Training Schematic Diagram

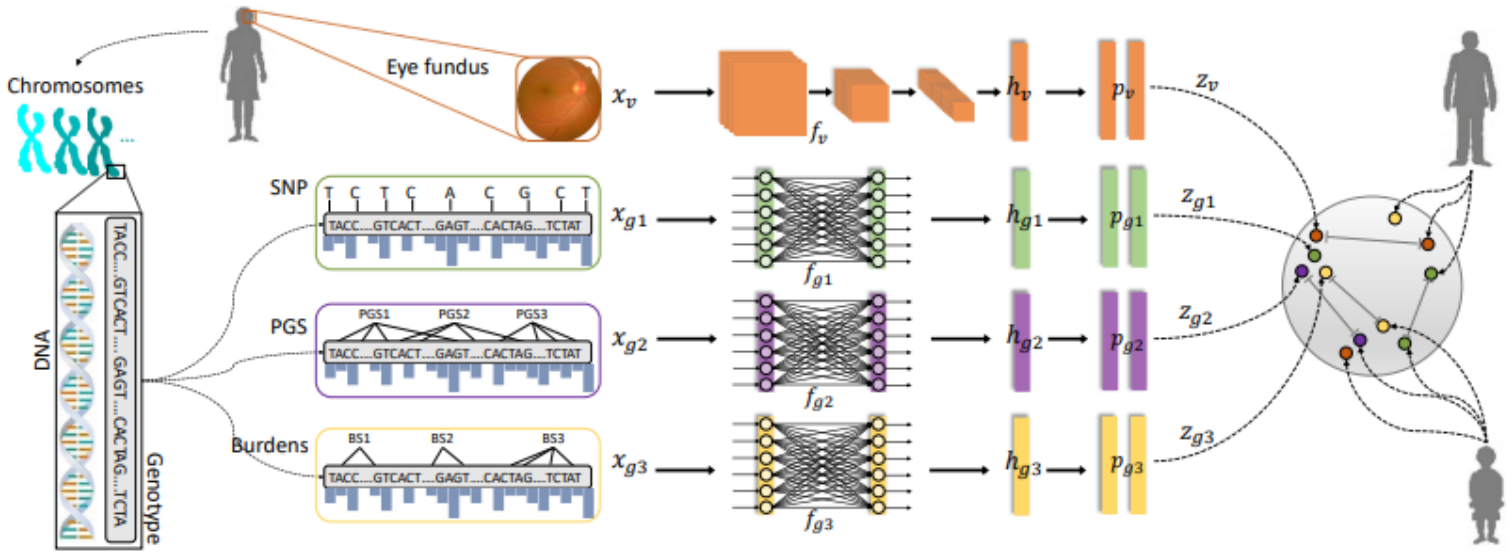


The training for LORAC consists of constructing the matrix **Q** out of augmented views of image $\mathbf{x_i}$ and is based on another framework using stop gradients and augmentations

TULANE UNIVERSITY

ContIG: Contrastive Learning for Medical Imaging and Genetics

- Contrastive learning has recently been used to integrate medical imaging and genomics[32]
- In this study, the medical imaging was of retinal images for diabetic retinopathy



TULANE UNIVERSITY

[32] Taleb et al. 2022 ContIG CVPR

Contrastive Learning and Multi-Omics Fusion Takeaways

Previous Aims Provide Foundation

- We can leverage work done in previous aims, especially the similarity kernel and augmentations, as a starting point
- Goal is to both
  - use large amounts of unlabeled data
  - move multi-omics and brain imaging to a common latent space
- ImageNomer software will be useful for joining and exploring many large or small datasets

TULANE UNIVERSITY

Summary and Acknowledgements

TULANE UNIVERSITY

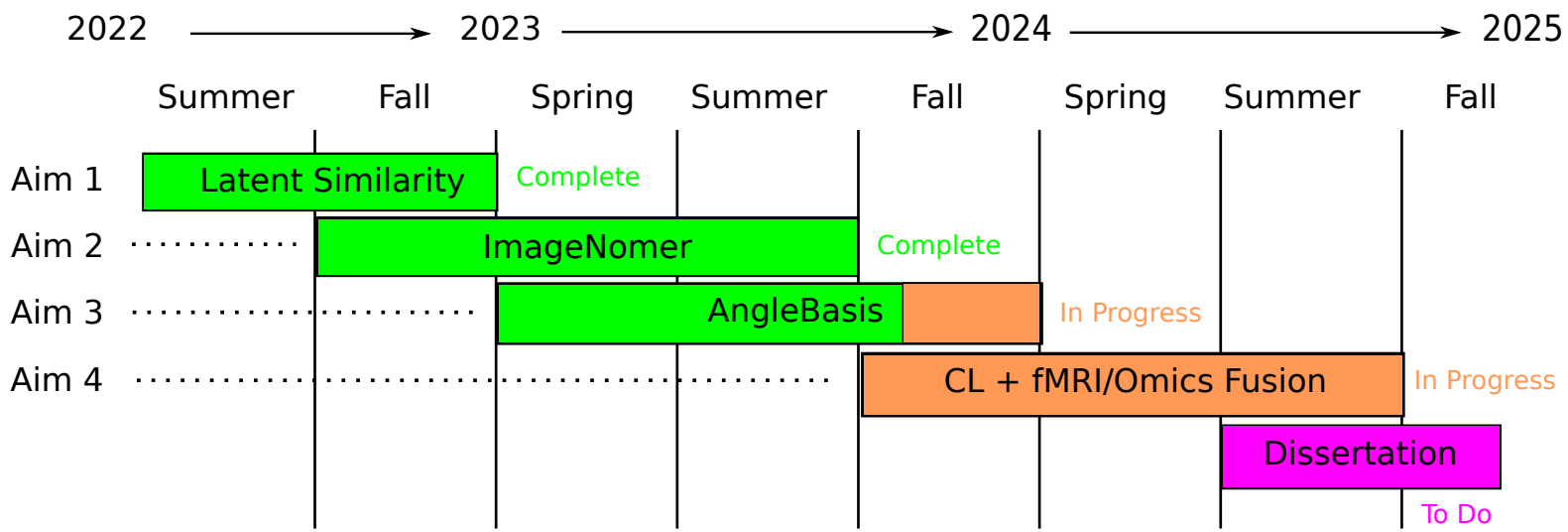
Summary

- The **Latent Similarity** framework works well for small sample size, high dimensionality data
- We apply a greedy feature selection algorithm for better feature selection
- **ImageNomer** allows for quick data exploration, feature visualization, and quality control
- We use ImageNomer to identify an unexpected race confound in fMRI data
- **AngleBasis** provides a generative model for the decomposition and compression of FC
- Our future goal is to leverage **contrastive learning** for fMRI-omics fusion and large unlabeled pre-training

TULANE UNIVERSITY

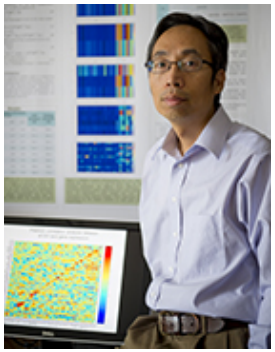
Timeline

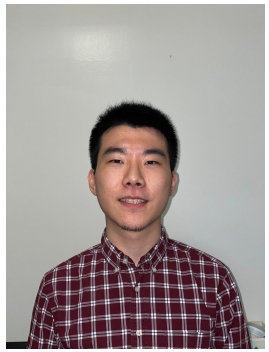
- Many of the specific aims have already been started or completed
- In addition, a large amount of data has been acquired from original sources and pre-processed (PNC, BSNIP, ADNI, UKB, etc.)



TULANE UNIVERSITY

Acknowledgements

- Lab members: Yu-Ping Wang (PI), Anton Orlichenko, Gang Qu, Binish Patel, Ziyu Zhou (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

TULANE UNIVERSITY