

# Predicting Tissue-Specific Gene Expression from Blood Using AI

Presented by

Kartika Waluyo, 1000555

Vrinda Rajendar Rajanahally, 1129446

Supervised by

Roberto Bonelli PhD, CSL Research

Brendan Ansell PhD, WEHI

Milica Ng

Prof Melanie Bahlo, WEHI

Monther Alhamdoosh PhD, CSL Research

Ziad Al Bkhetan PhD, The University of Melbourne

Prof Michael Kirley, The University of Melbourne



25 June 2021

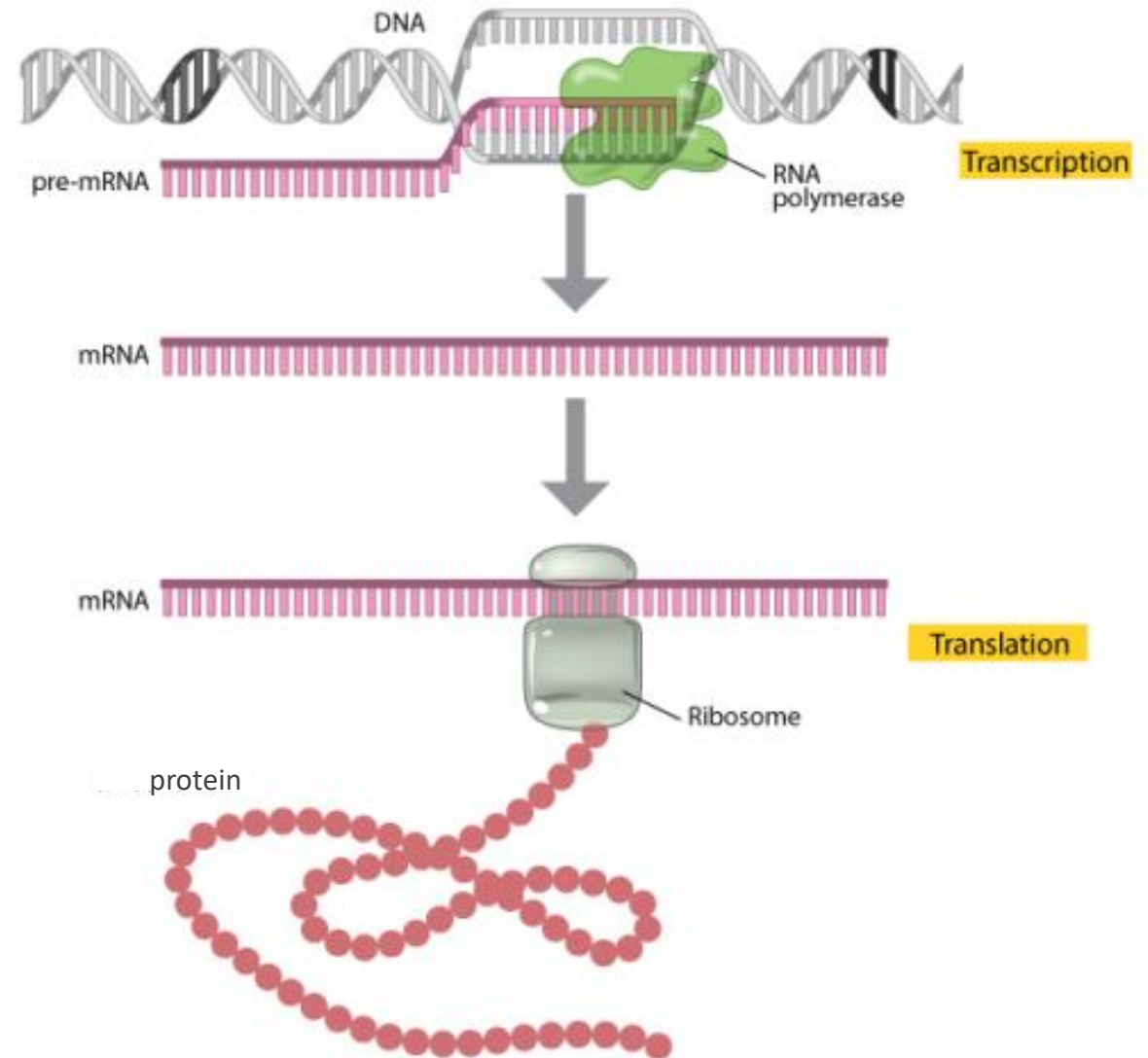


THE UNIVERSITY OF  
MELBOURNE

# Introduction

## Gene expression

- Measured through RNA sequencing
- Indicates the “activation level” of genes in a tissue
- Used to understand disease mechanisms and assess the effectiveness and safety of treatments



# Data

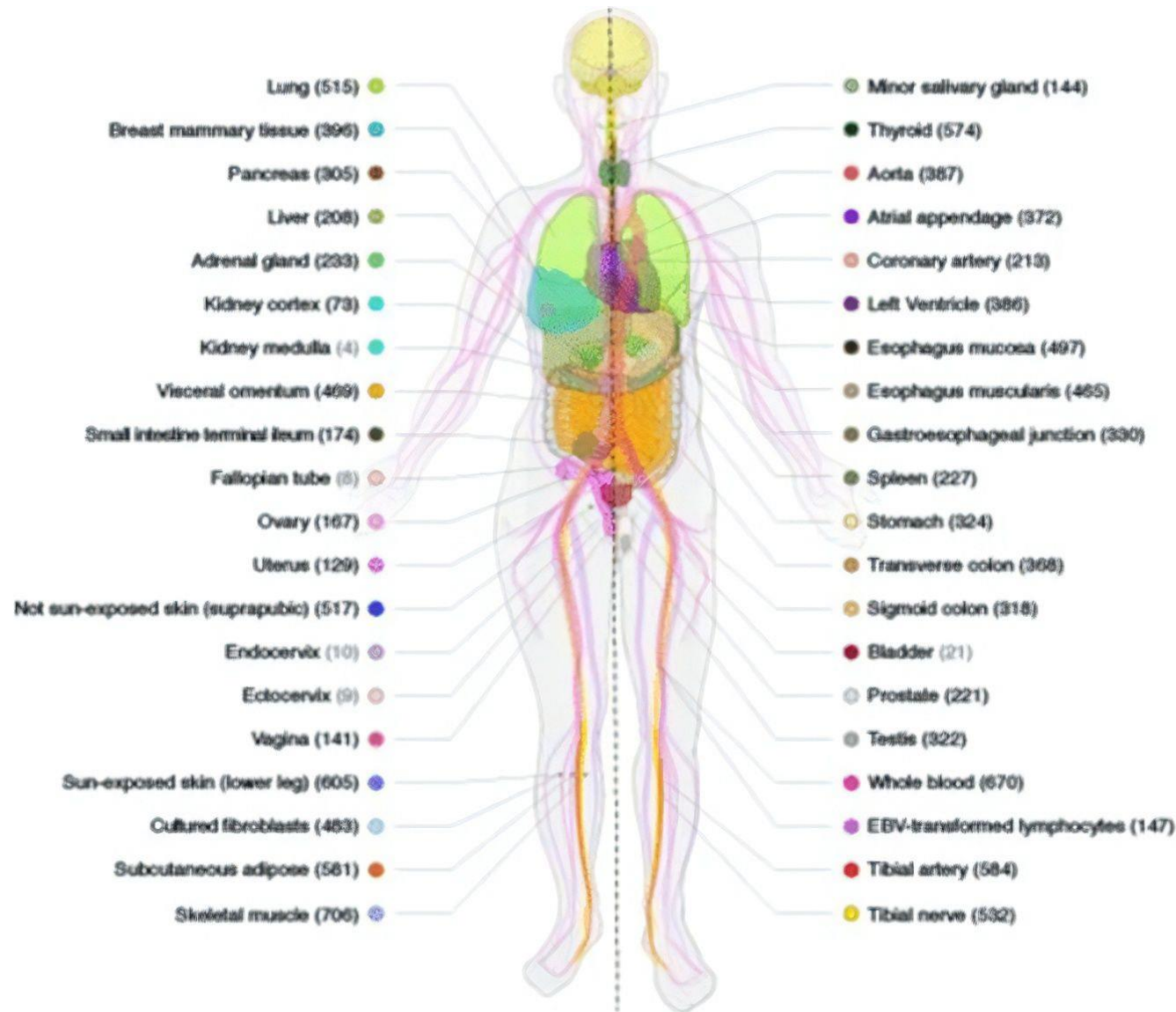


## The Genotype-Tissue Expression (GTEx)

Publicly available dataset containing gene expression of 54 post-mortem tissues from almost 1,000 individuals

Website:

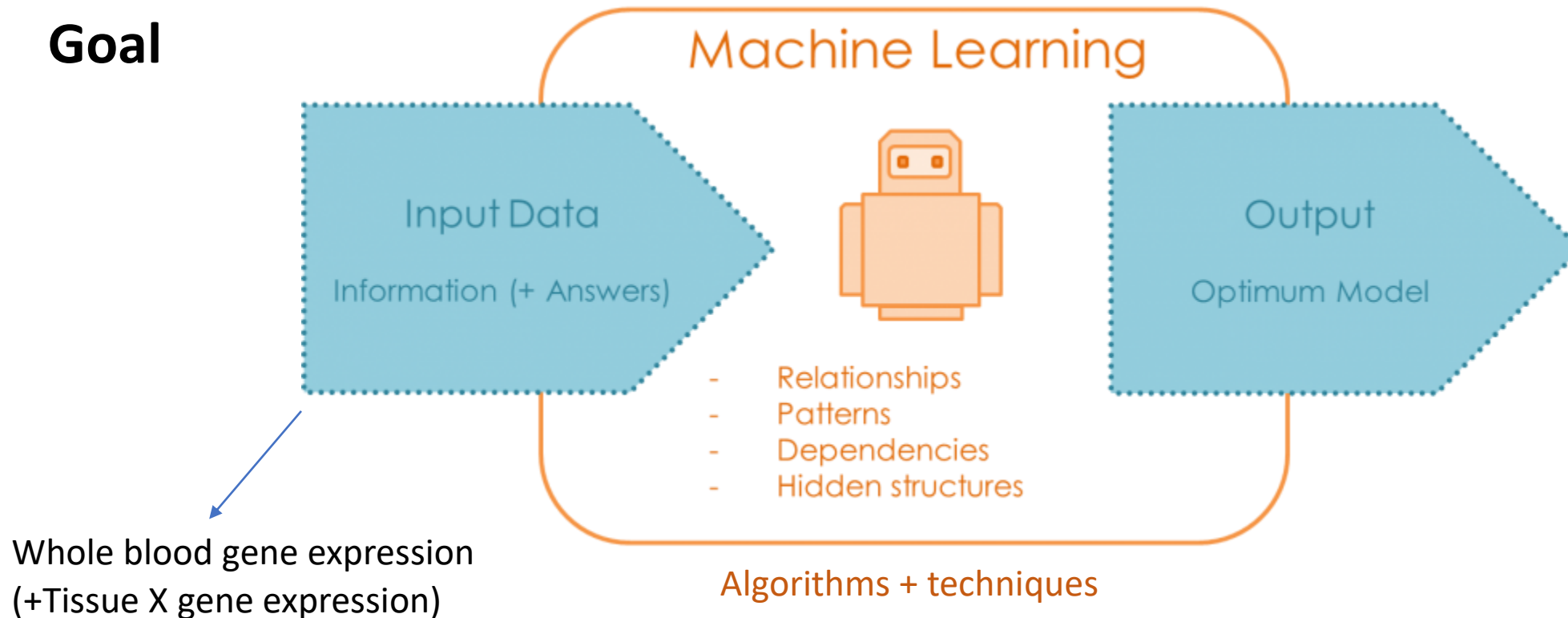
<https://www.gtexportal.org/home/>



# Problem

- Gene expression varies across tissues
- Obtaining samples from heart, lung, brain and other organs is often not possible

# Goal



# Data Preprocessing



	GTEX-111YS	GTEX-1122O	GTEX-1128S	GTEX-117YW	GTEX-11DXX
ENSG00000188976	4.62364473	3.78253909	4.5133603	5.4527936	3.37519628
ENSG00000187961	1.93052342	1.27499967	3.5022390	3.5097032	1.82645735
ENSG00000187583	-0.75745530	-2.17811244	0.8867026	-0.3610409	-2.00195164
ENSG00000188290	-2.04196059	-0.37572051	-0.0725749	1.2067569	-3.55957777
ENSG00000187608	3.20267015	3.30039771	2.5859542	4.9613959	2.44112346



	GTEX-111YS	GTEX-1122O	GTEX-1128S	GTEX-117YW	GTEX-11DXX
ENSG00000188976	5.56460543	5.8744263	5.970547249	5.16622735	5.95836923
ENSG00000187961	3.60586511	3.9661503	4.392572536	4.02935543	4.00614281
ENSG00000187583	-0.26752148	0.5173860	1.411392148	1.74797744	0.23257945
ENSG00000188290	3.18410789	3.4222746	3.491766740	5.04511280	2.75391016
ENSG00000187608	3.87791649	4.2725832	4.466592815	5.18764207	3.90172955



	GTEX-1128S	GTEX-117XS	GTEX-1192X	GTEX-11DXW	GTEX-11DXY
ENSG00000188976	4.5133603	5.41962562	4.40573764	5.62804043	4.97935857
ENSG00000187961	3.5022390	2.86169747	3.54214773	3.29359001	3.82458092
ENSG00000188290	-0.0725749	0.70906184	0.88278386	1.70362024	-0.56079737
ENSG00000187608	2.5859542	5.66514689	5.17640485	4.86172548	4.24214853
ENSG00000188157	2.7483510	1.73139723	1.64804437	2.11621022	2.45405401



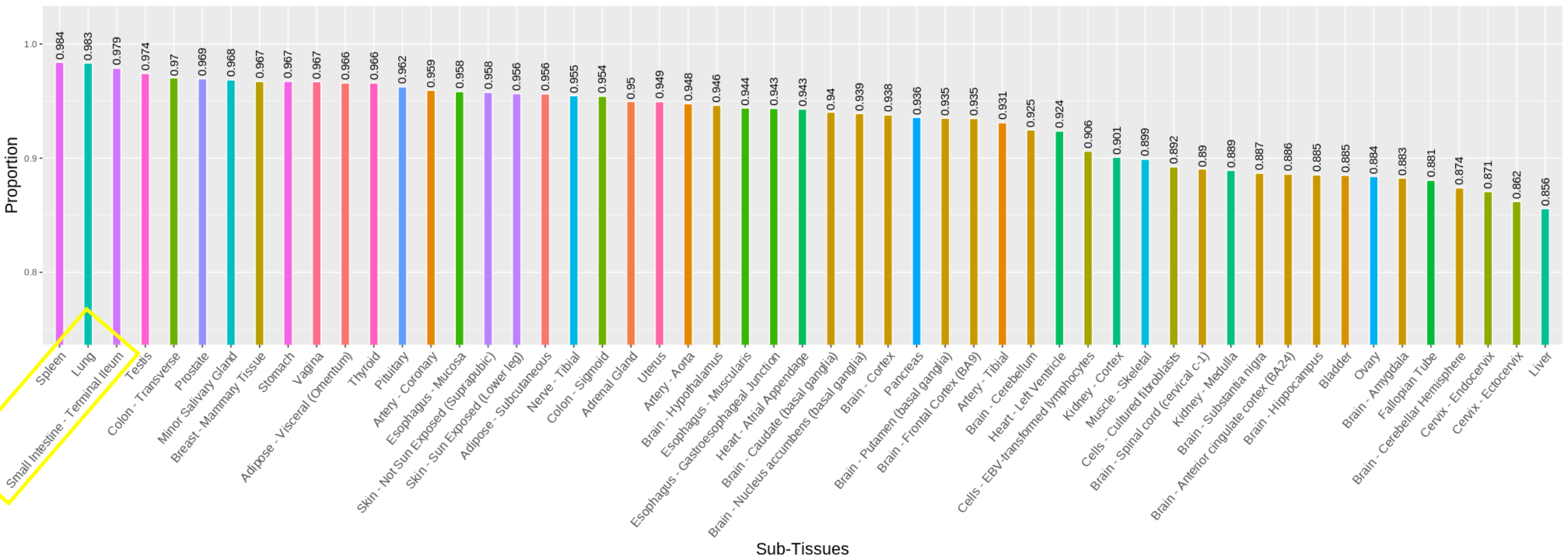
	GTEX-1128S	GTEX-117XS	GTEX-1192X	GTEX-11DXW	GTEX-11DXY
ENSG00000188976	6.36783476	6.2316949	6.46433195	6.3887099	6.2711988467
ENSG00000187961	5.56040664	4.9513539	6.16913252	5.3212697	5.0969189121
ENSG00000188290	1.61186294	1.4614192	2.71896260	2.4761530	1.4588005060
ENSG00000187608	0.70806198	1.5332473	1.69497304	0.6651525	1.6285311222
ENSG00000188157	4.52051422	4.8785337	5.12529845	4.8907246	5.1851027306

# Preliminary Data Analysis

Exploration of **similarities** between each tissue and whole blood

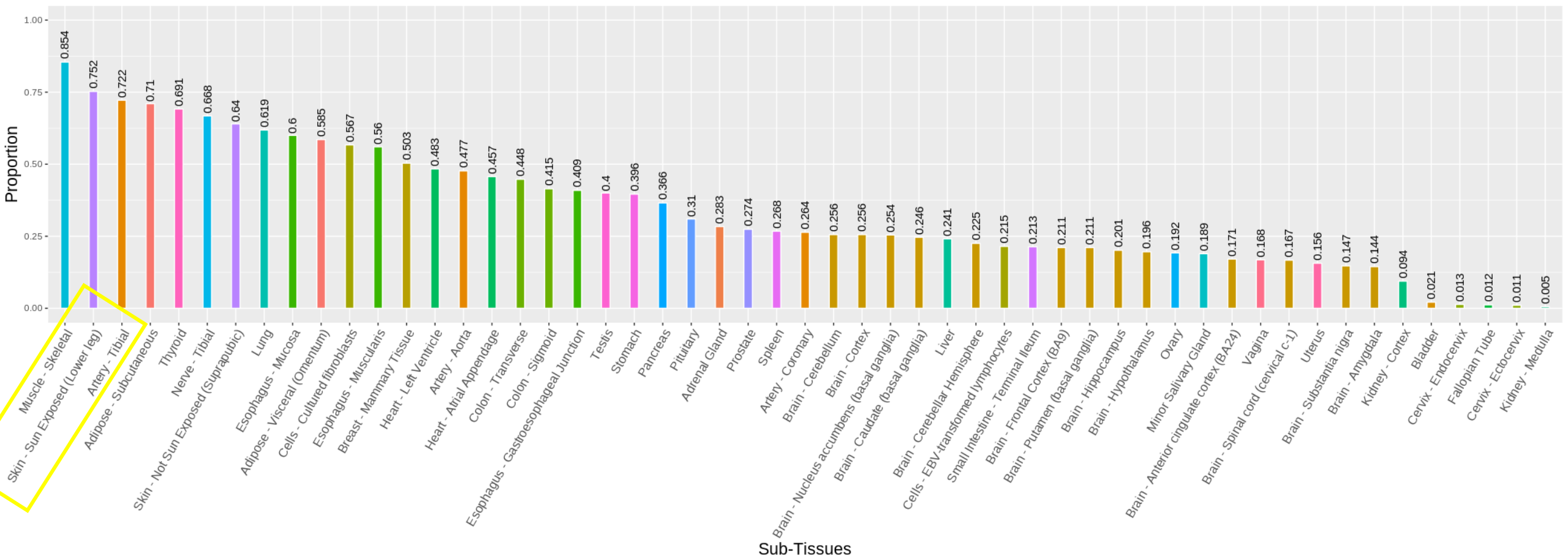
1. Shared expressed **genes**
2. Shared **donors**
3. Gene expression **correlation**

Proportion of Overlapping Genes of Whole Blood with all other Tissues



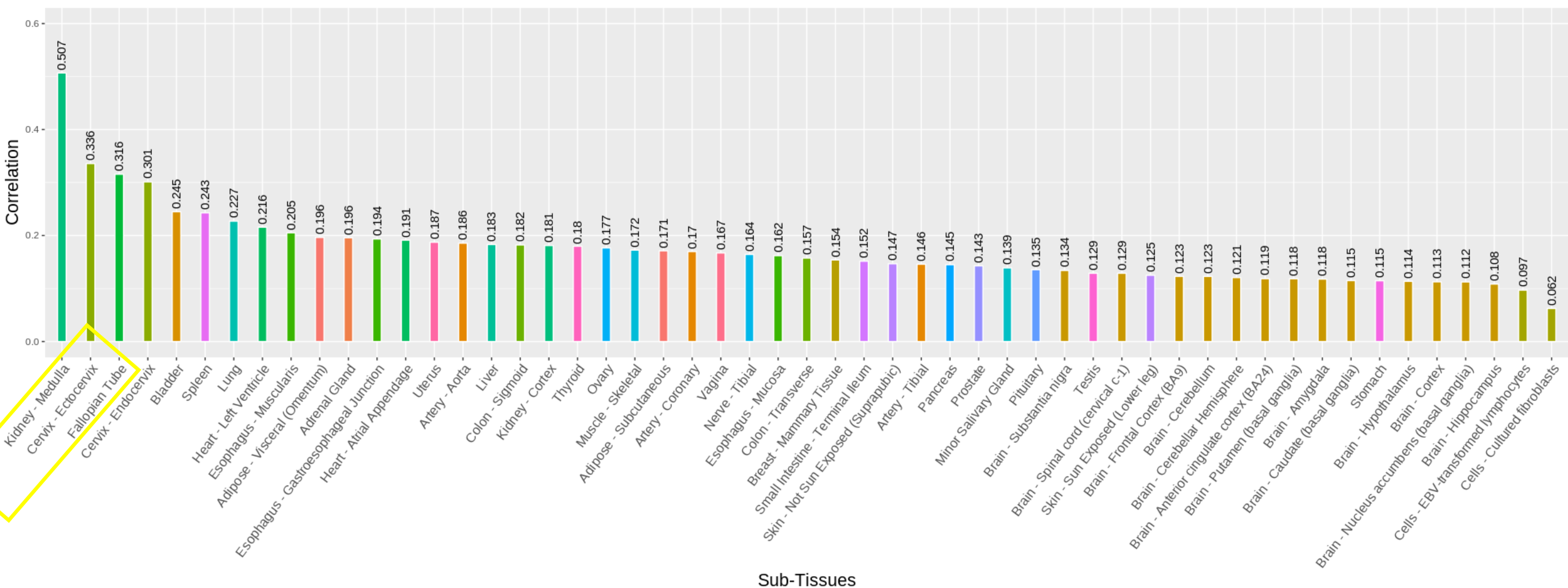


Proportion of Overlapping Donors of Whole Blood with all other Tissues





# Mean of Absolute Correlation of a Tissue and Whole Blood, based on their Shared Genes



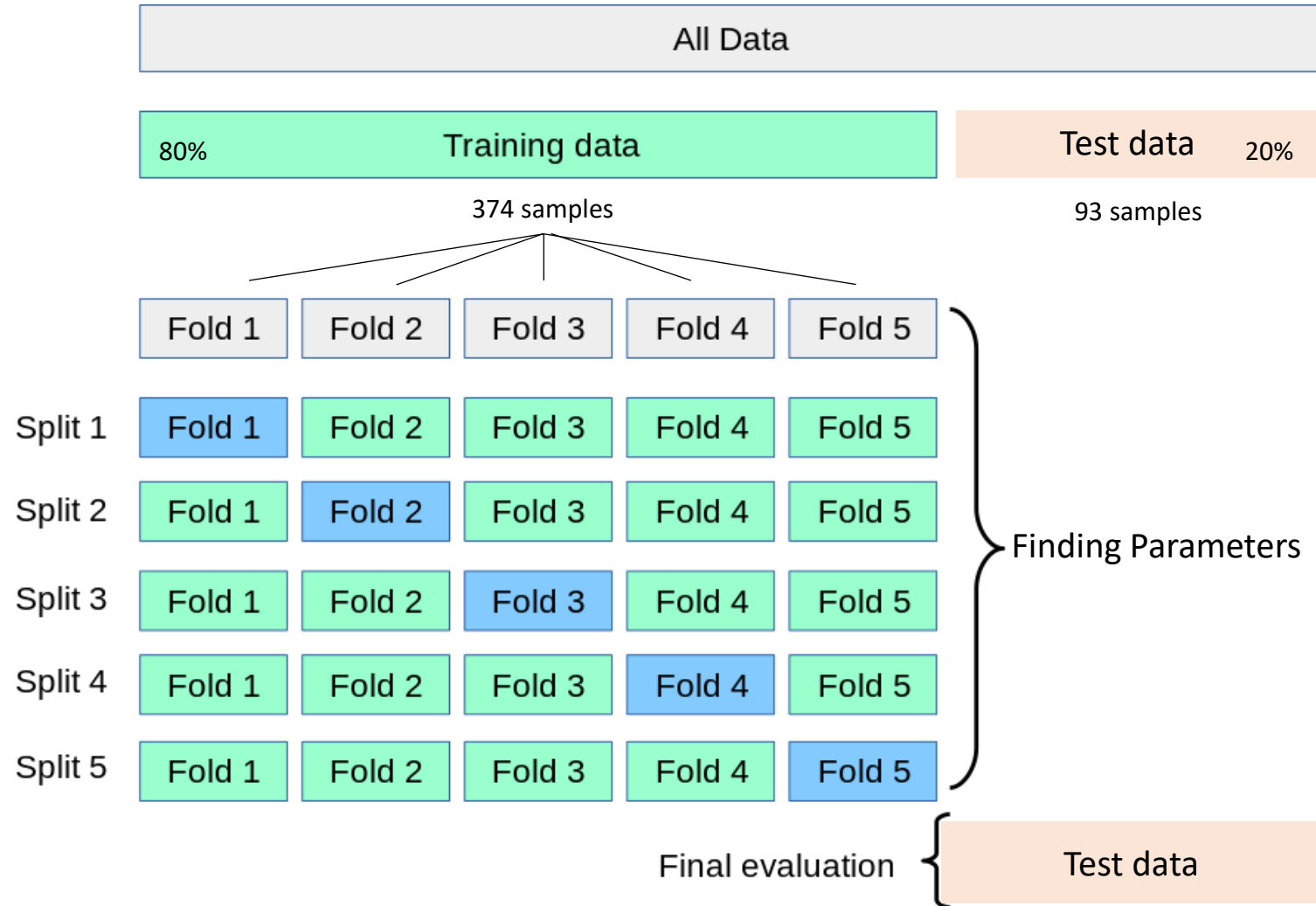
Pairwise correlation between each gene in whole blood's gene expressions with the same gene in other tissues' gene expressions.

# Tissue Selection

Rank tissues based on the previously observed measures

Subtissue	Donor Proportion	Donor Rank	Gene Proportion	Gene Rank	Mean Absolute Correlation	Correlation Rank	Average Rank
Lung	0.619	8	0.983	2	0.227	7	5.67
Adipose - Visceral (Omentum)	0.585	10	0.966	11	0.196	10	10.33
Spleen	0.268	26	0.984	1	0.243	6	11
Thyroid	0.691	5	0.966	12	0.18	19	12
Adipose - Subcutaneous	0.71	4	0.956	18	0.171	22	14.67
Esophagus - Muscularis	0.56	12	0.944	25	0.205	9	15.33
Breast - Mammary Tissue	0.503	13	0.967	8	0.154	28	16.33
Colon - Transverse	0.448	17	0.97	5	0.157	27	16.33
Esophagus - Mucosa	0.6	9	0.958	15	0.162	26	16.67
Nerve - Tibial	0.668	6	0.955	19	0.164	25	16.67
Artery - Aorta	0.477	15	0.948	23	0.186	15	17.67
Skin - Not Sun Exposed (Suprapubic)	0.64	7	0.958	16	0.147	30	17.67
Colon - Sigmoid	0.415	18	0.954	20	0.182	17	18.33
Adrenal Gland	0.283	24	0.95	21	0.196	11	18.67
Heart - Atrial Appendage	0.457	16	0.943	27	0.191	13	18.67
Esophagus - Gastroesophageal Junction	0.409	19	0.943	26	0.194	12	19

# Cross-validation



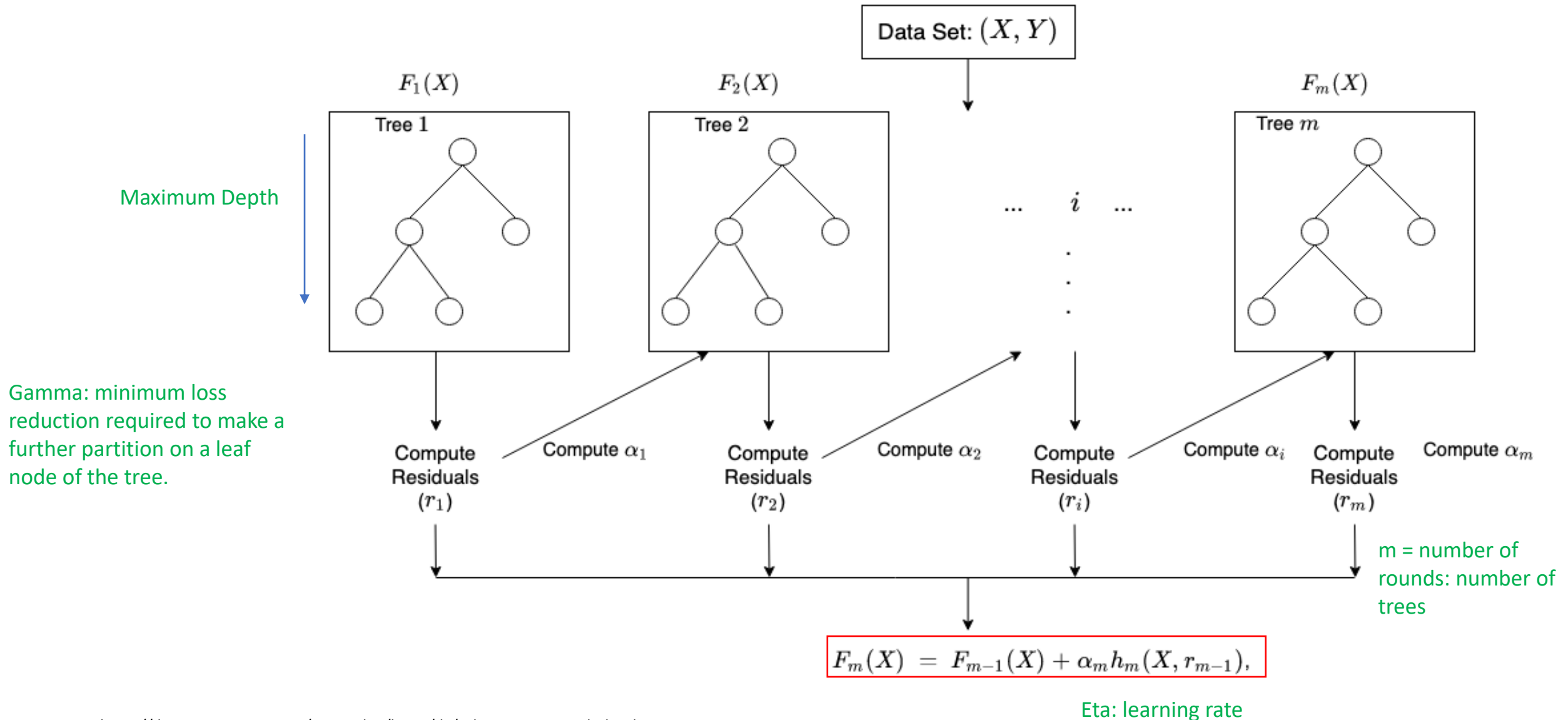
# Models

eXtreme Gradient Boosting

Neural Network

Ensemble Model

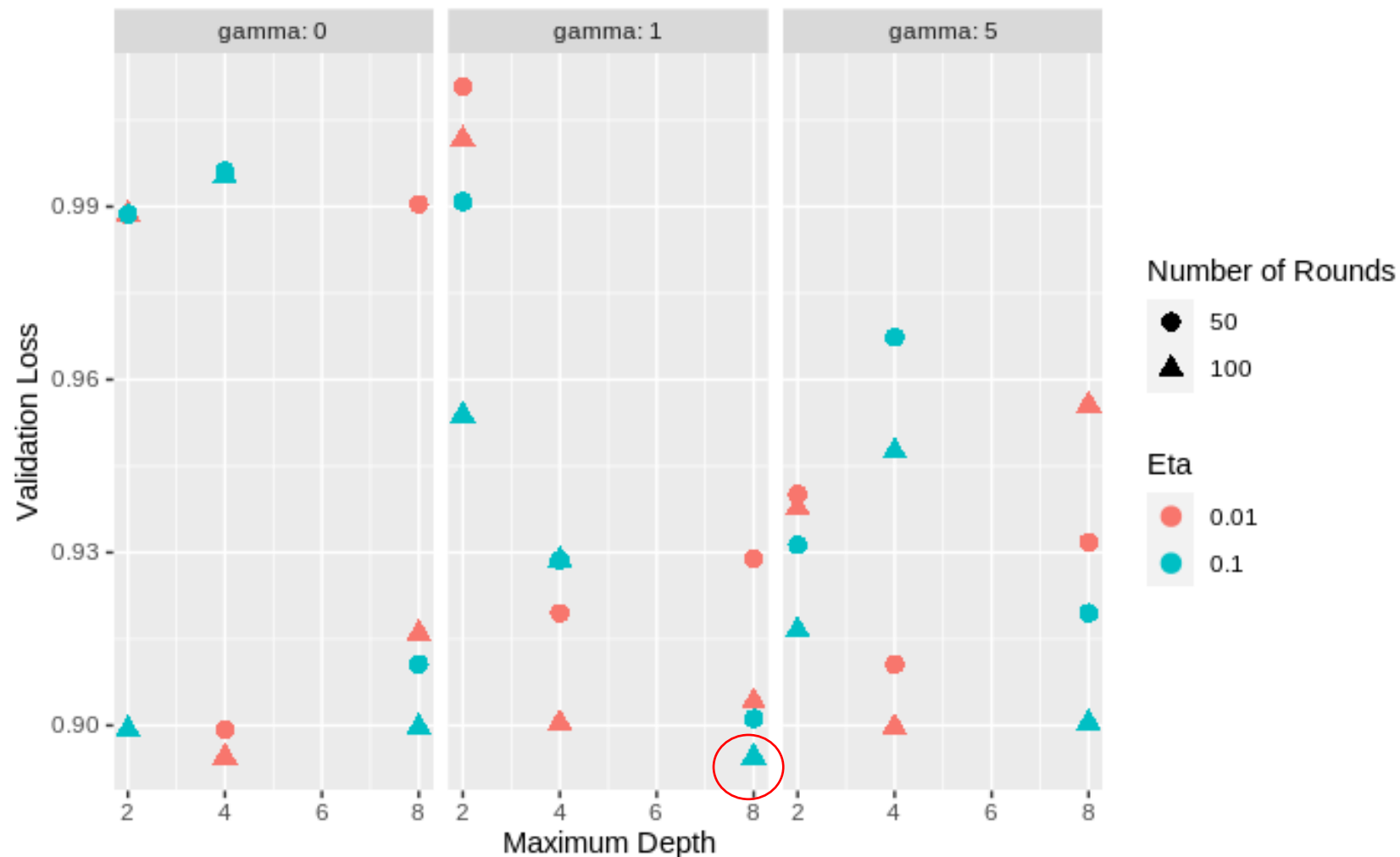
# eXtreme Gradient Boosting (XGBoost)



# Hyperparameter Tuning

On 50 random genes

On gamma, maximum depth, eta, and number of rounds



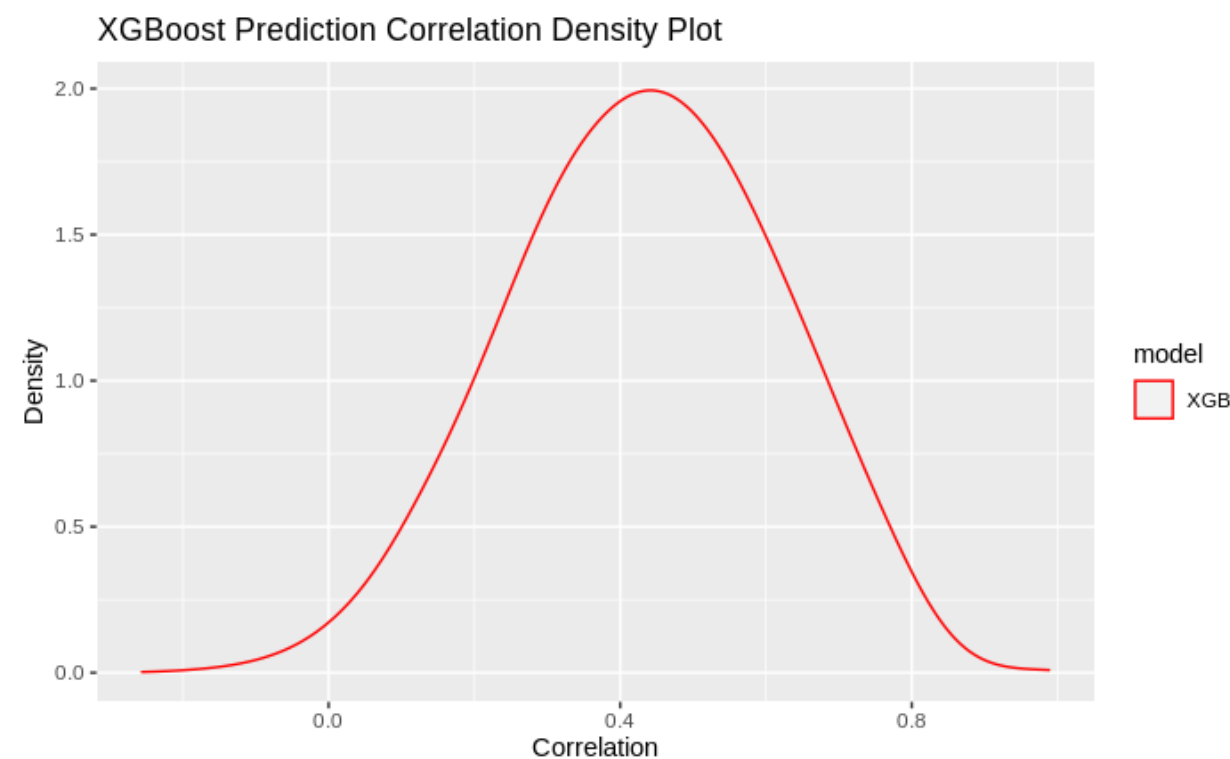
Loss Function

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2$$

Best hyperparameter combination: gamma=1, maximum depth=8, eta=0.1, and number of rounds=100

# Evaluation

Compute the correlation between each gene in the prediction and its true value

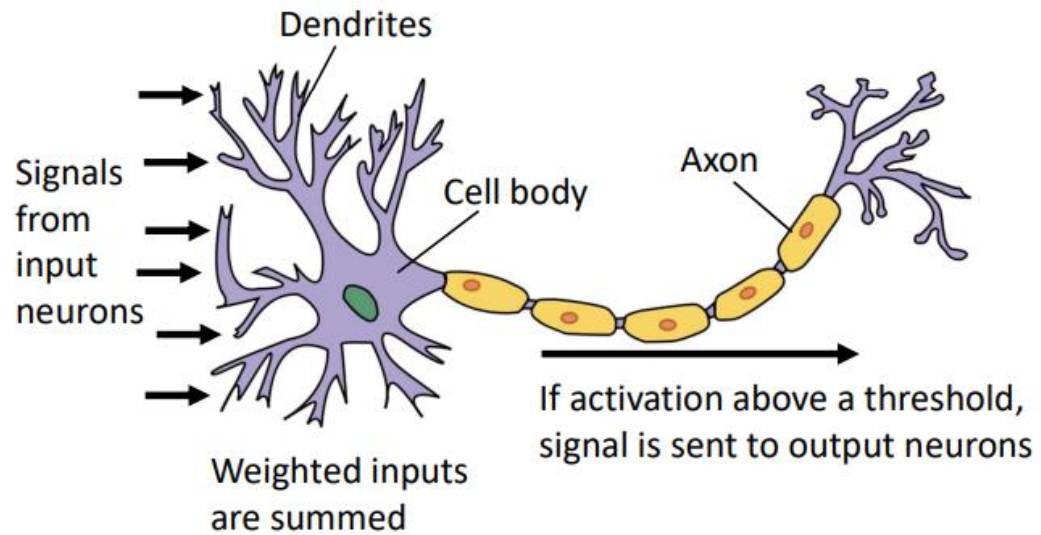


Model	Median Correlation	
	Training	Test
XGBoost	0.990975	0.4364047
Neural Network		
Ensemble		



# Neural Network

## Biological Neuron



## Artificial Neuron

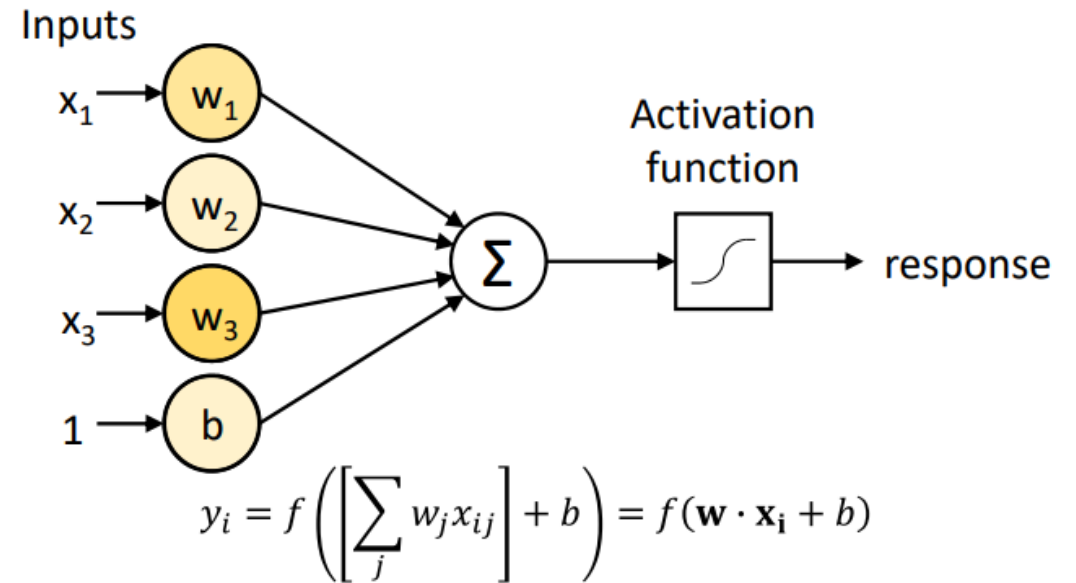
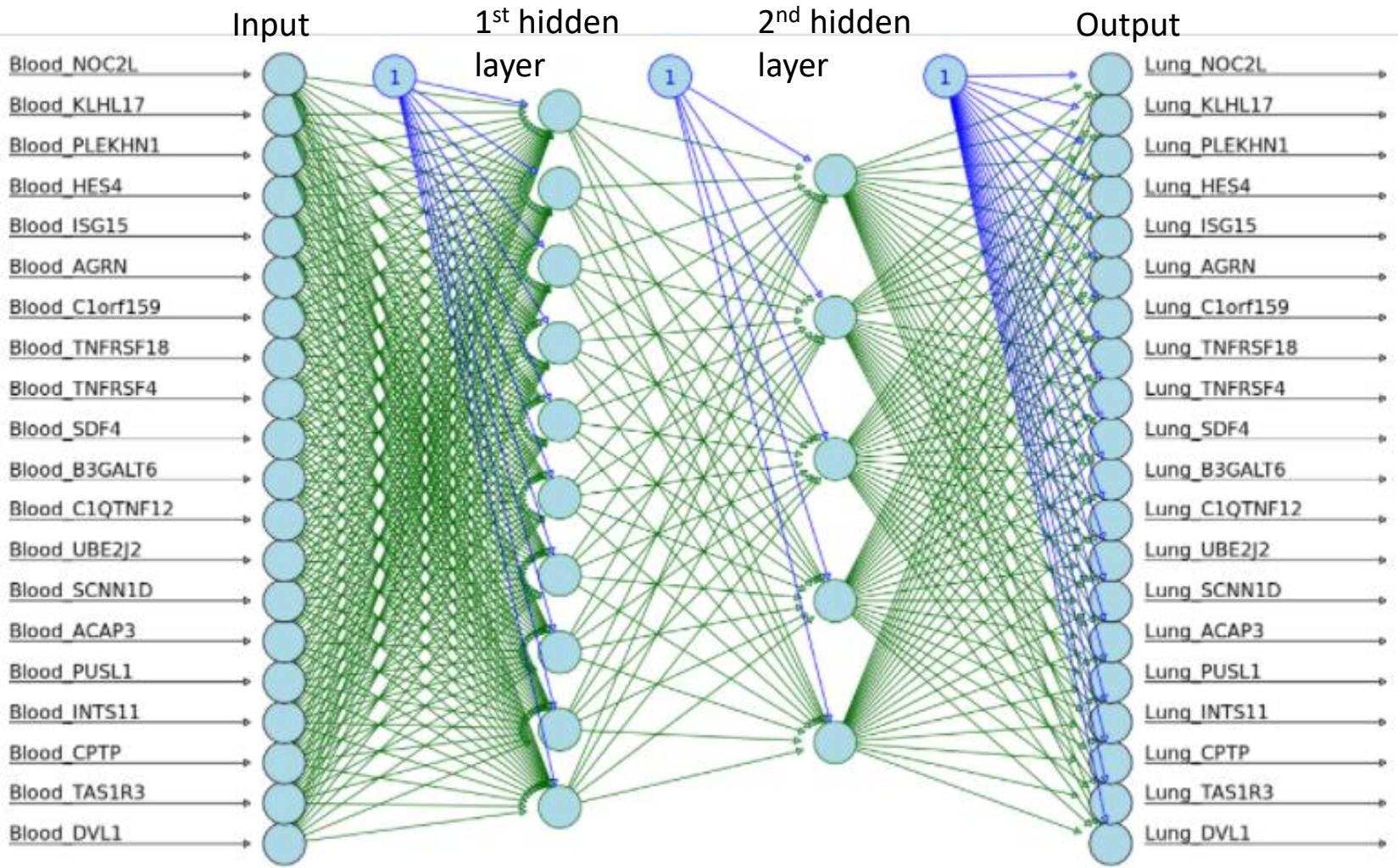
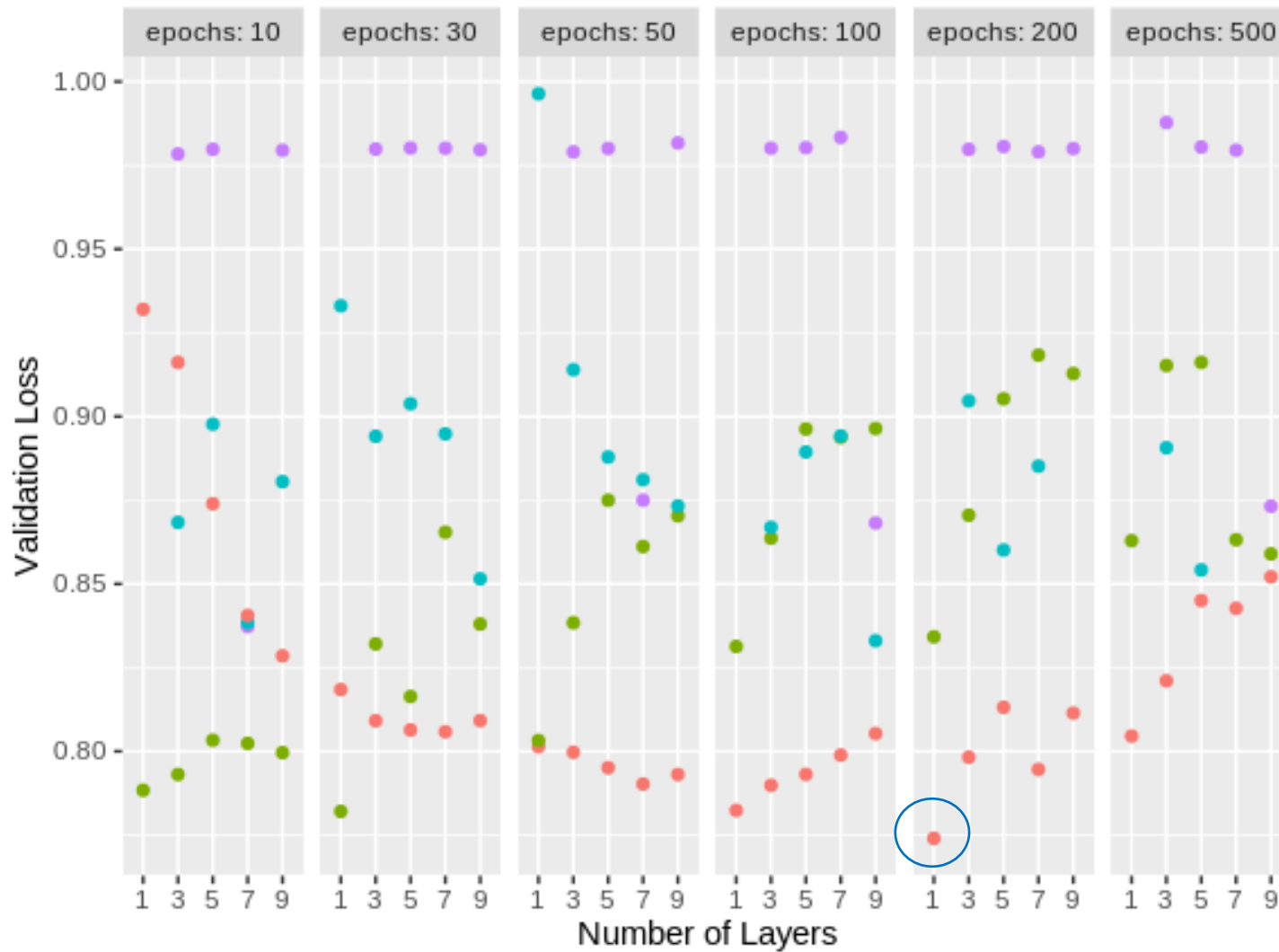


Illustration of how our neural network model might look



# Hyperparameter Tuning

On number of layers, learning rate, and number of epochs



Learning Rate

- 1e-05
- 1e-04
- 0.001
- 0.01

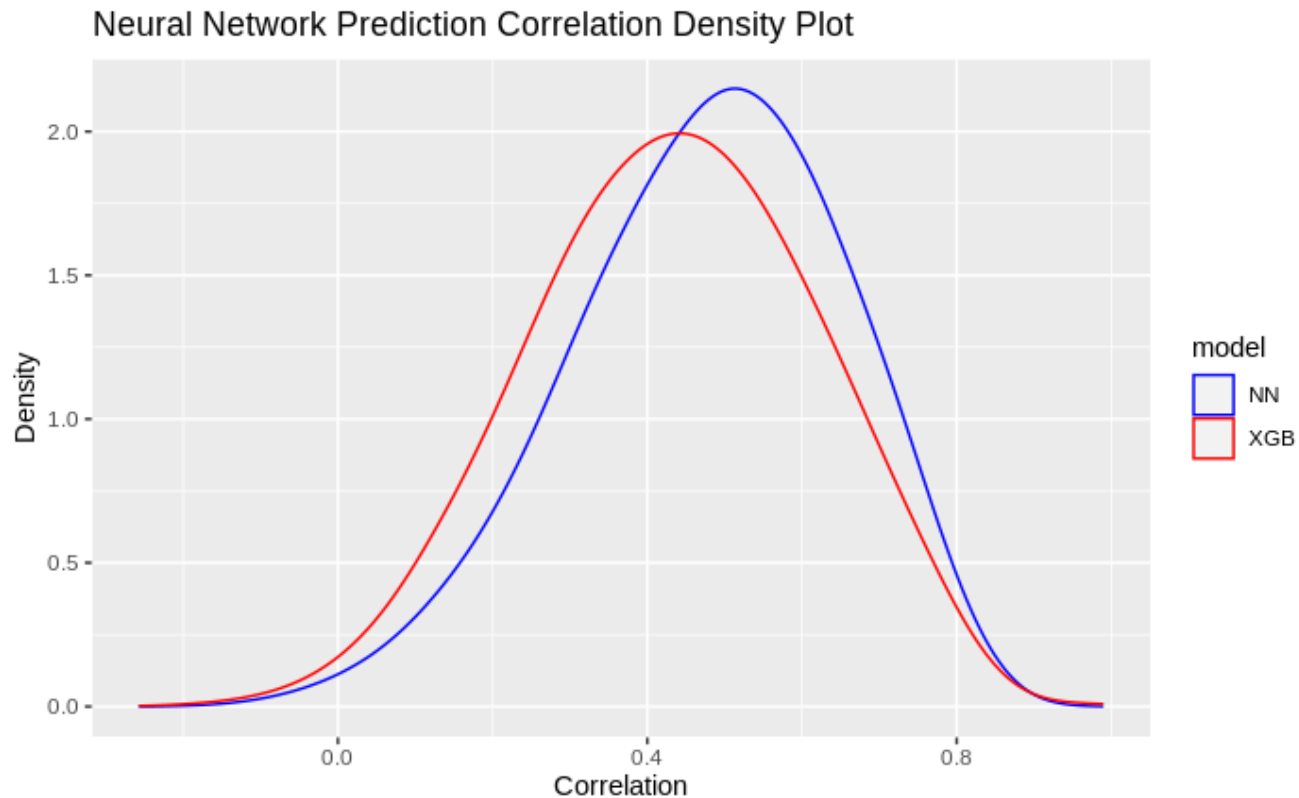
Loss Function

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Best parameter combination:  
1 hidden layer of 100 units,  $e^{-5}$  learning rate, and  
200 epochs

# Evaluation

Compute the correlation for each gene in the prediction and the true value



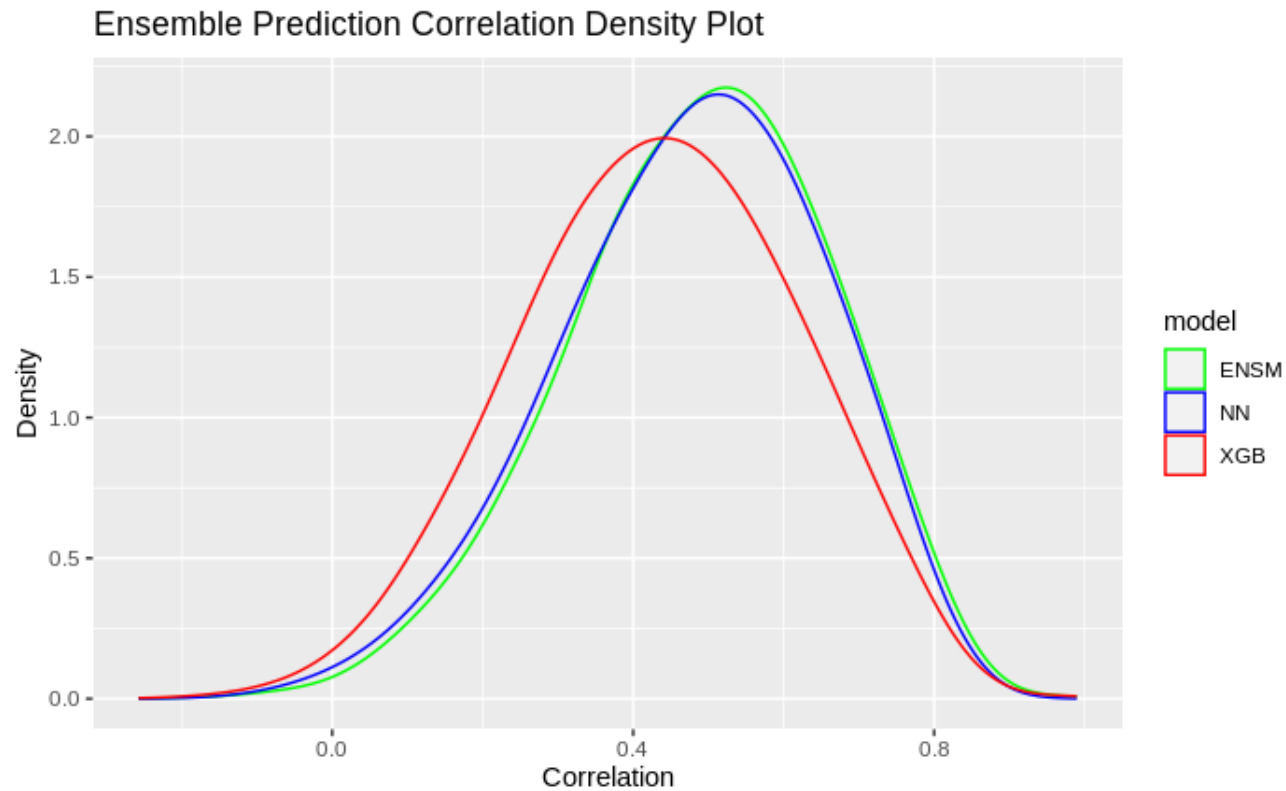
Model	Median Correlation	
	Training	Test
XGBoost	0.990975	0.4364047
Neural Network	0.6278658	0.4887672
Ensemble		

# Ensemble Prediction

$$ENS_{pred} = \frac{1}{2}NN_{pred} + \frac{1}{2}XGB_{pred} = \frac{1}{2} \begin{bmatrix} \hat{y}_{11} & \cdots & \hat{y}_{1p} \\ \vdots & \ddots & \\ \hat{y}_{n1} & \cdots & \hat{y}_{np} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \hat{z}_{11} & \cdots & \hat{z}_{1p} \\ \vdots & \ddots & \\ \hat{z}_{n1} & \cdots & \hat{z}_{np} \end{bmatrix}$$

# Evaluation

Compute the correlation for each gene in the prediction and the true value



Model	Median Correlation	
	Training	Test
XGBoost	0.990975	0.4364047
Neural Network	0.6278658	0.4887672
Ensemble	0.9205496	0.4976427

# Future Directions

- Training the models on all other tissues, on the entire dataset
- Stacking ensemble for improved prediction
- Biological validation
- Implementation of the web application



We would like to extend a heartfelt thank you to all  
our mentors involved in this project!

Roberto Bonelli PhD, CSL Research

Brendan Ansell PhD, WEHI

Milica Ng

Prof Melanie Bahlo, WEHI

Monther Alhamdoosh PhD, CSL Research

Ziad Al Bkhetan PhD, The University of Melbourne

Prof Michael Kirley, The University of Melbourne



THE UNIVERSITY OF  
MELBOURNE

# Questions?



THE UNIVERSITY OF  
MELBOURNE

# Thank You

