# PREDICTING TISSUE-SPECIFIC GENE EXPRESSION FROM BLOOD USING AI

Presented by
Kartika Waluyo, 1000555
Vrinda Rajendar Rajanahally, 1129446

22 October 2021

CSL™ WEHI brighter together 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

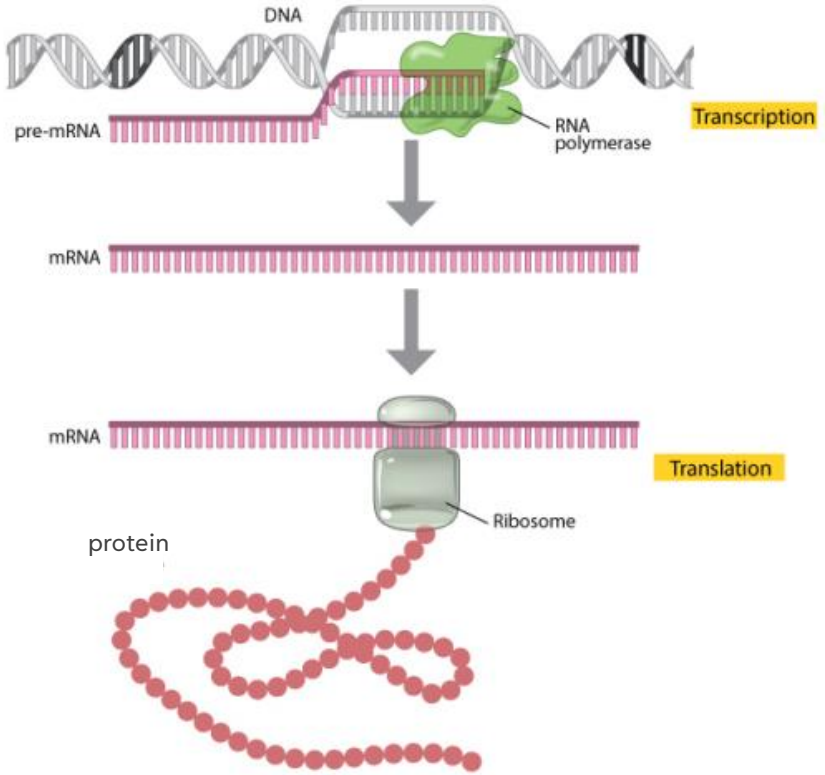
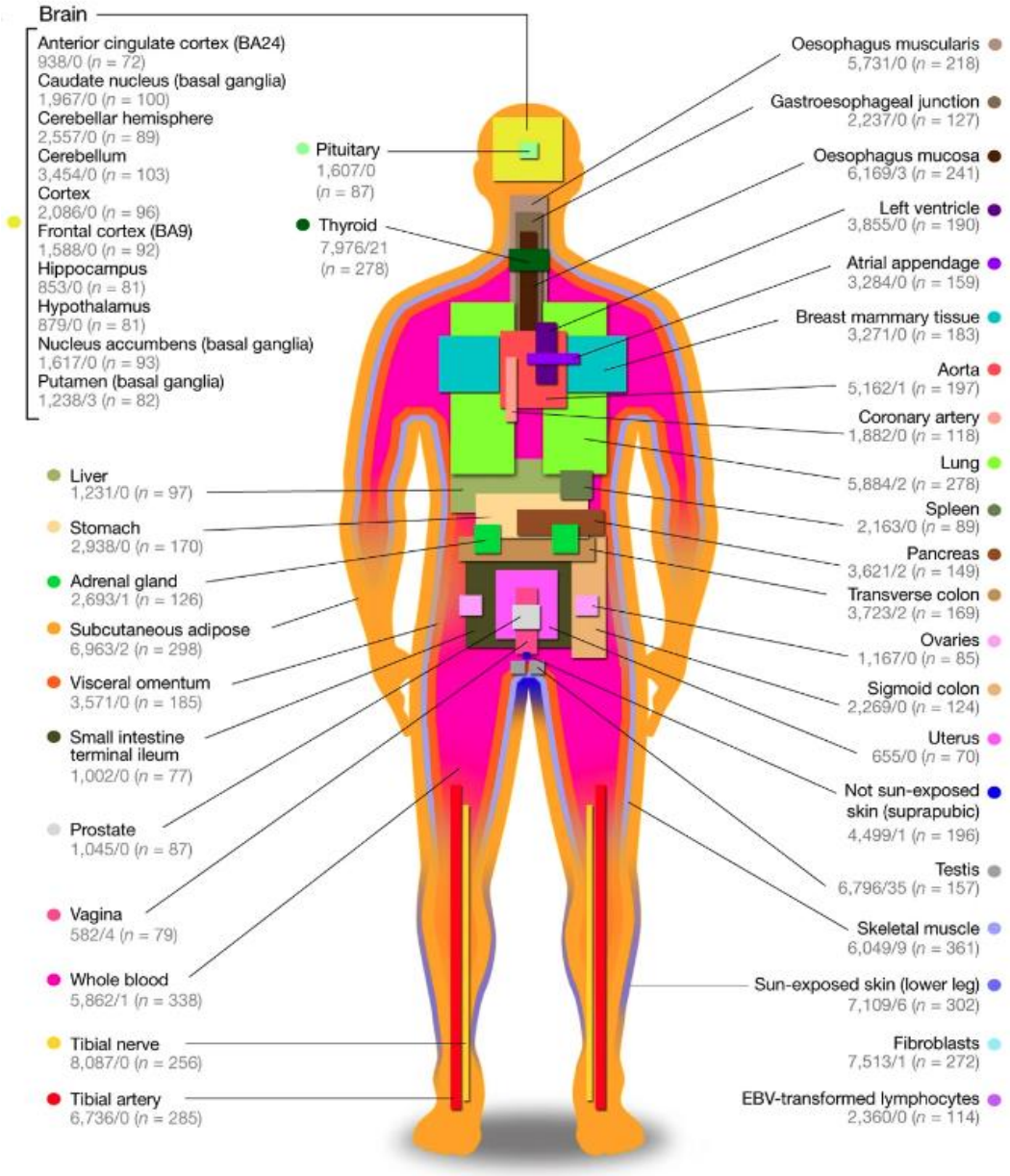



Image source: https://www.nature.com/




# 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/

Image source: https://www.researchgate.net/figure/Sample-and-data-types-in-the-GTEx-v8-study-A-Illustration-of-the-54-tissue-types_fig1_336246636

# PROBLEM

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

# GOAL

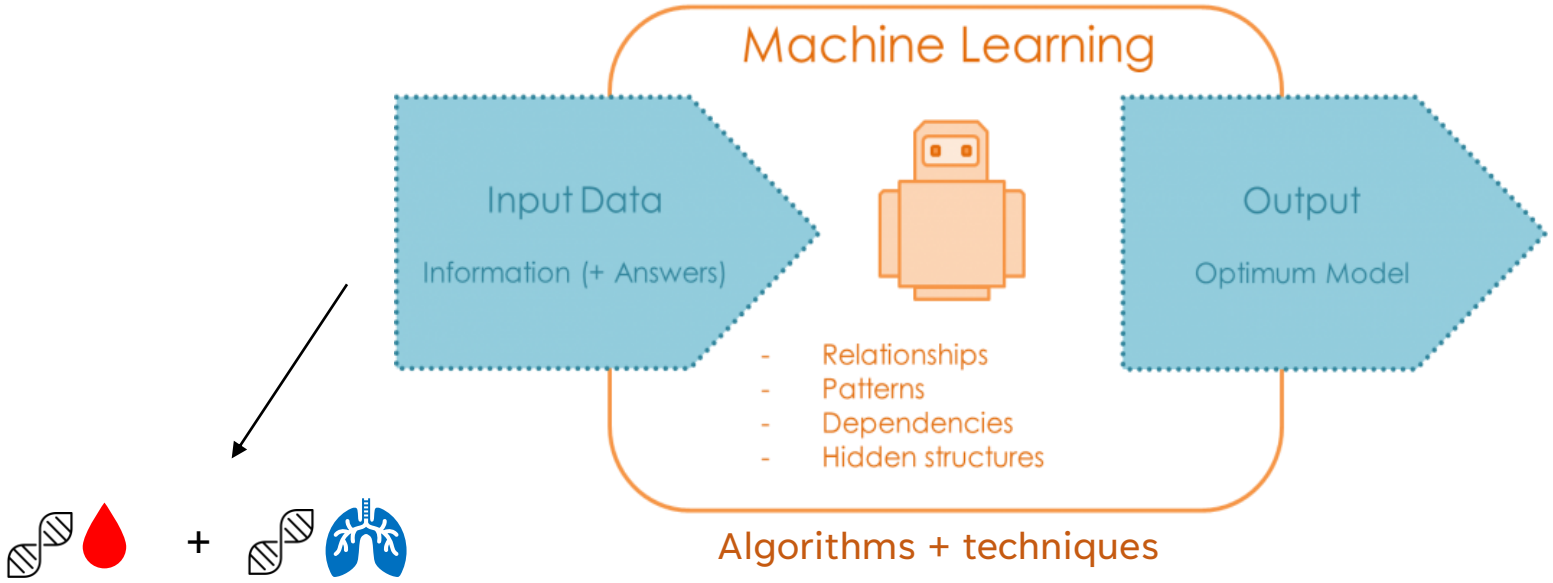
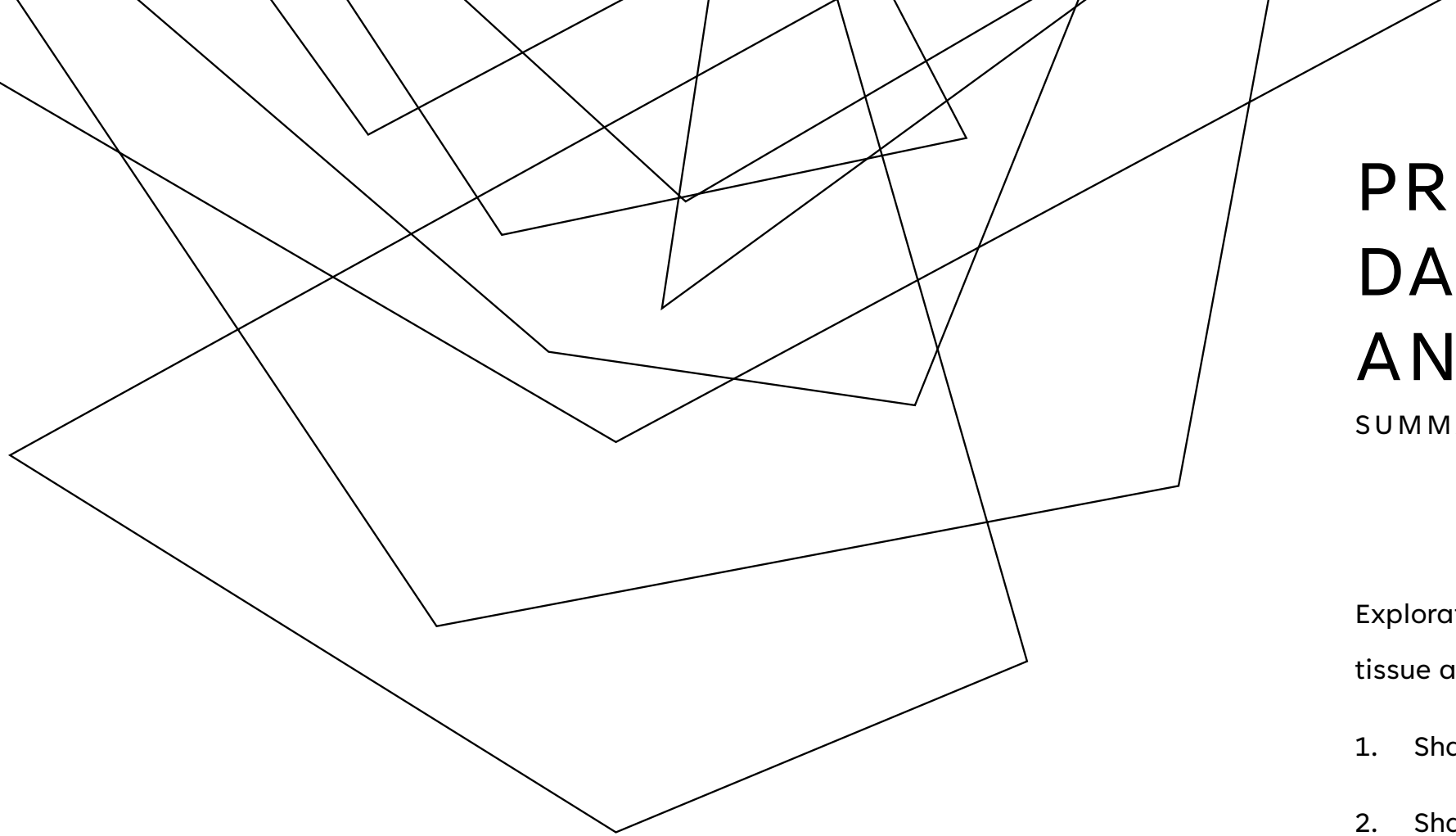



Image source: https://quantdare.com/

# DATA PRE-PROCESSING

| | NOC2L | KLHL17 | PLEKHN1 | HES4 | ISG15 | AGRN | C1orf159 |
|---|---|---|---|---|---|---|---|
| GTEX-111YS | 4.623645 | 1.9305234 | -0.757455296 | -2.04196059 | 3.202670 | 0.34305222 | 2.644519 |
| GTEX-1122O | 3.782539 | 1.2749997 | -2.178112437 | -0.37572051 | 3.300398 | 1.14820160 | 1.781592 |
| GTEX-1128S | 4.513360 | 3.5022390 | 0.886702570 | -0.07257490 | 2.585954 | 2.74835104 | 3.438855 |
| GTEX-117YW | 5.452794 | 3.5097032 | -0.361040851 | 1.20675685 | 4.961396 | 2.66612637 | 3.895245 |
| GTEX-11DXX | 3.375196 | 1.8264574 | -2.001951642 | -3.55957777 | 2.441123 | 0.46472915 | 1.706664 |
| GTEX-11DXZ | 4.473020 | 2.6155735 | -0.020073254 | 0.27319390 | 3.814766 | 2.14063671 | 2.751952 |
| GTEX-11EI6 | 4.503837 | 3.7731937 | -0.089162153 | 0.96350723 | 3.581990 | 1.66687554 | 3.257955 |
| GTEX-11EMC | 4.689939 | 2.7730465 | 1.233797566 | 2.48327357 | 6.018670 | 2.15817703 | 3.950064 |
| GTEX-11EQ9 | 3.966465 | 1.7293461 | -1.366626237 | -2.34489017 | 2.402989 | 1.00785226 | 2.256939 |
| GTEX-11GSP | 4.713622 | 3.0192610 | -0.137642769 | 1.52595816 | 4.664902 | 1.89930115 | 3.458261 |

| | NOC2L | KLHL17 | PLEKHN1 | HES4 | ISG15 | AGRN | C1orf159 |
|---|---|---|---|---|---|---|---|
| GTEX-111YS | 5.564605 | 3.605865 | -0.26752148 | 3.184108 | 3.877916 | 6.992831 | 3.997613 |
| GTEX-1122O | 5.874426 | 3.966150 | 0.51738596 | 3.422275 | 4.272583 | 7.030410 | 4.070818 |
| GTEX-1128S | 5.970547 | 4.392573 | 1.41139215 | 3.491767 | 4.466593 | 6.707747 | 4.035884 |
| GTEX-117YW | 5.166227 | 4.029355 | 1.74797744 | 5.045113 | 5.187642 | 7.266217 | 4.812552 |
| GTEX-11DXX | 5.958369 | 4.006143 | 0.23257945 | 2.753910 | 3.901730 | 7.207189 | 4.116435 |
| GTEX-11DXZ | 5.771633 | 4.627956 | 2.85381134 | 3.622749 | 5.236977 | 8.041496 | 4.574765 |
| GTEX-11EI6 | 5.211195 | 4.826346 | 1.46932243 | 4.520421 | 5.145422 | 7.703033 | 4.195147 |
| GTEX-11EMC | 5.748173 | 3.427566 | 1.95369092 | 4.087649 | 3.857217 | 6.868381 | 4.153474 |
| GTEX-11EQ9 | 6.000552 | 3.625458 | 0.79048291 | 2.635259 | 3.707483 | 6.842565 | 3.919237 |
| GTEX-11GSP | 6.049516 | 4.244247 | 1.64588696 | 5.739212 | 5.385897 | 7.577261 | 4.412140 |



# PRELIMINARY DATA ANALYSIS

SUMMARY OF PART 1

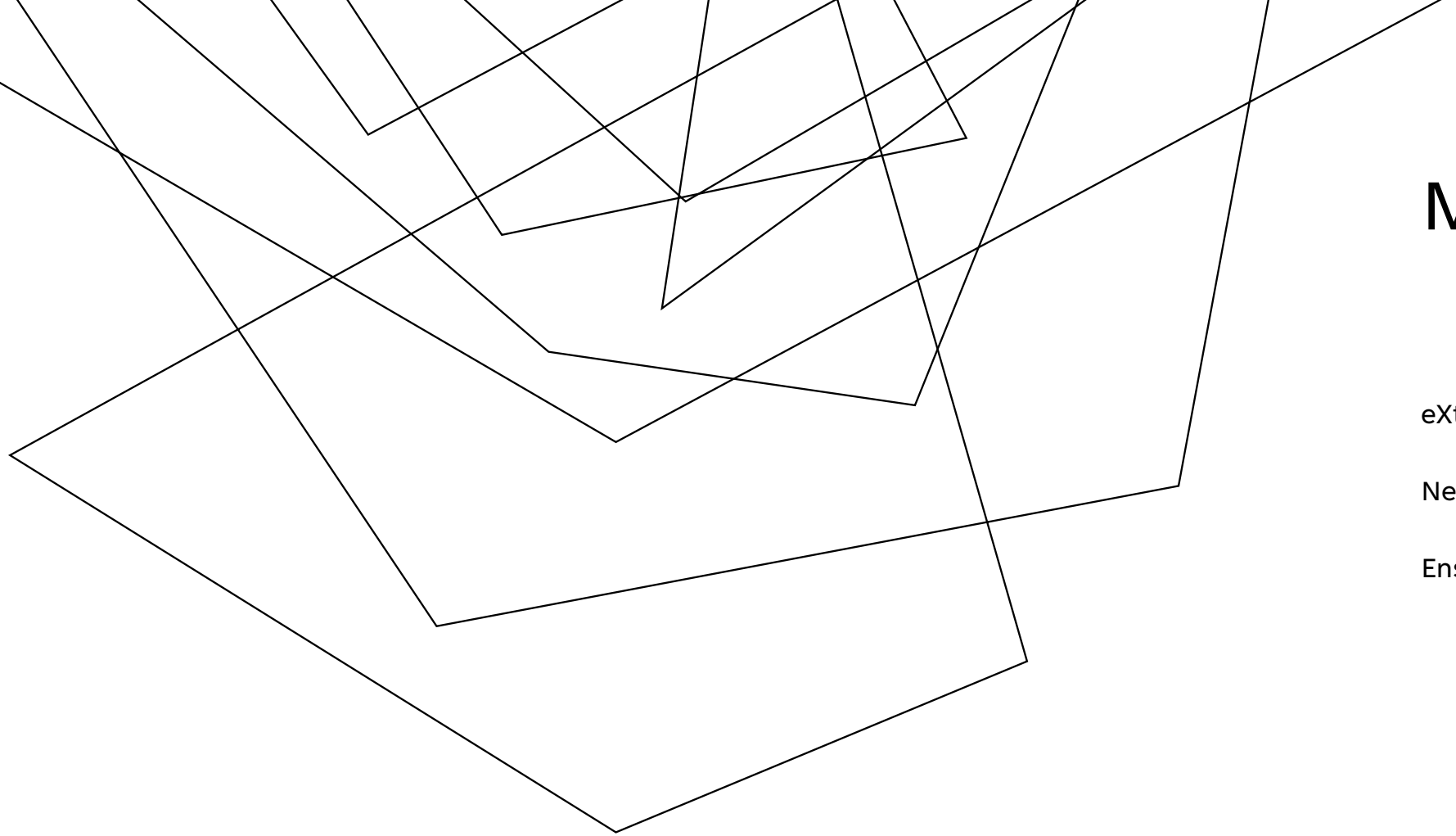
Exploration of **similarities** between each tissue and whole blood

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

# 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 |



# MODELS

eXtreme Gradient Boosting

Neural Network

Ensemble Model

# CROSS-VALIDATION

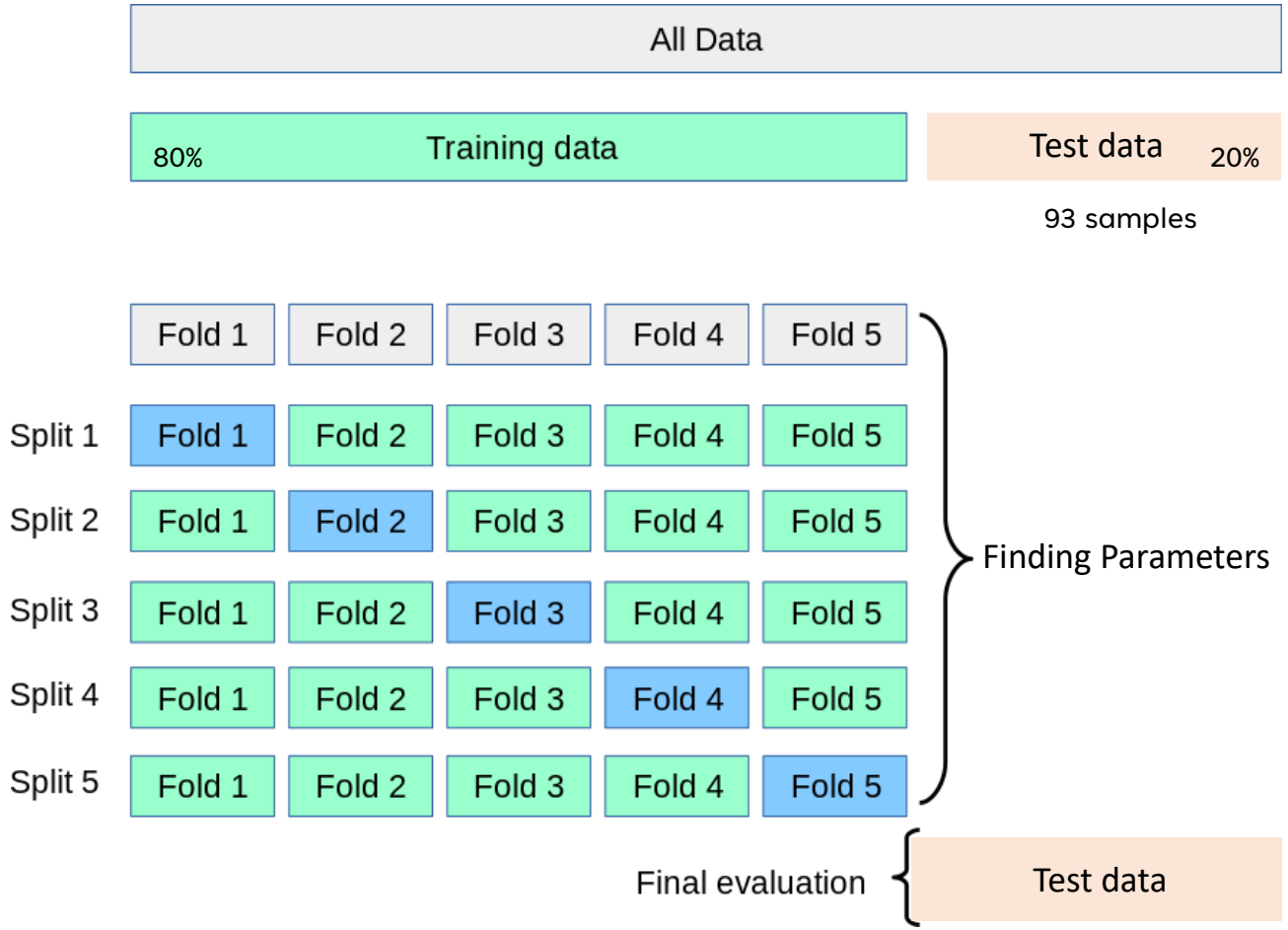





Image source: https://scikit-learn.org/stable/modules/cross_validation.html

# EXTREME GRADIENT BOOSTING (XGBOOST)

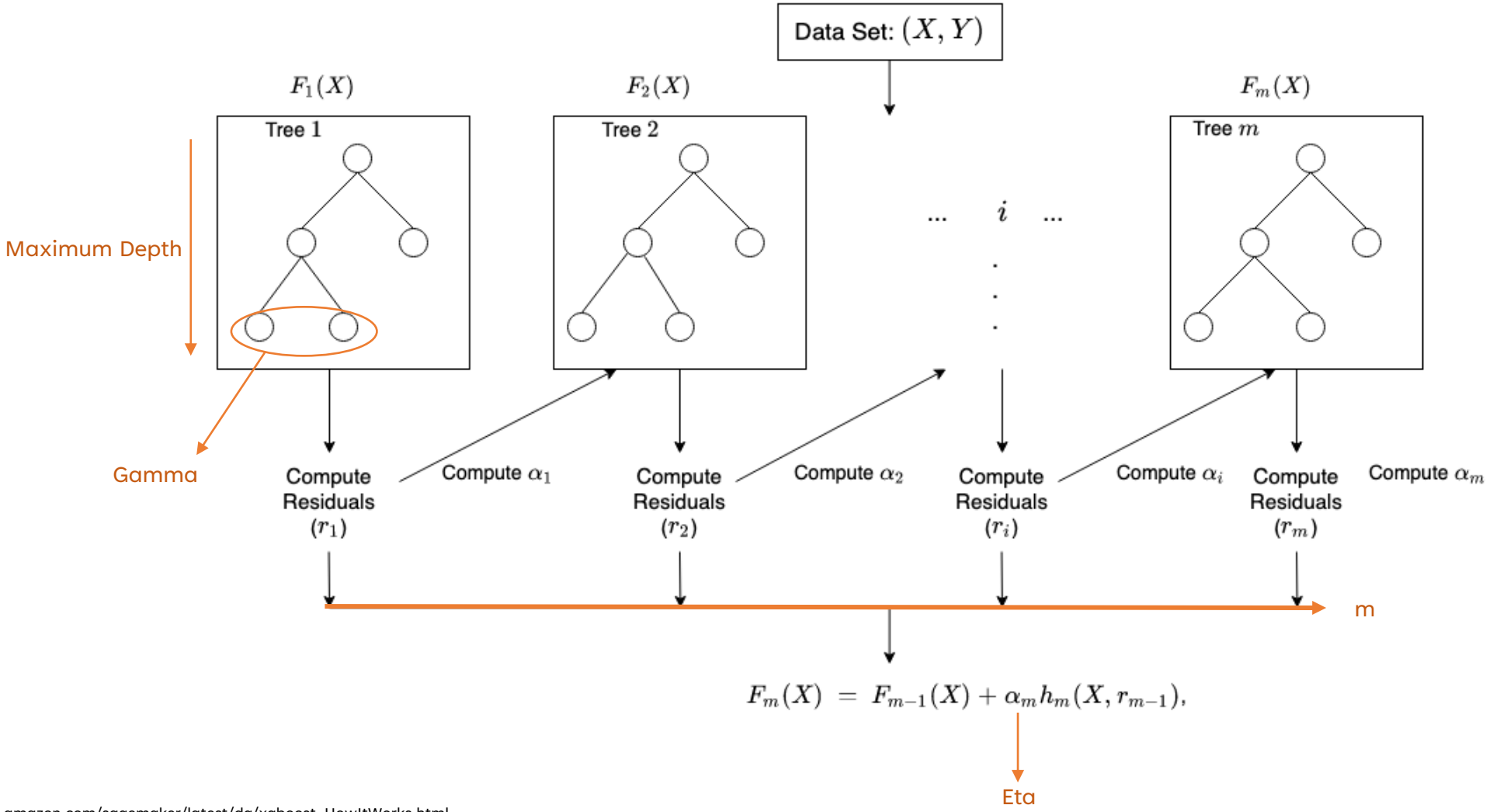
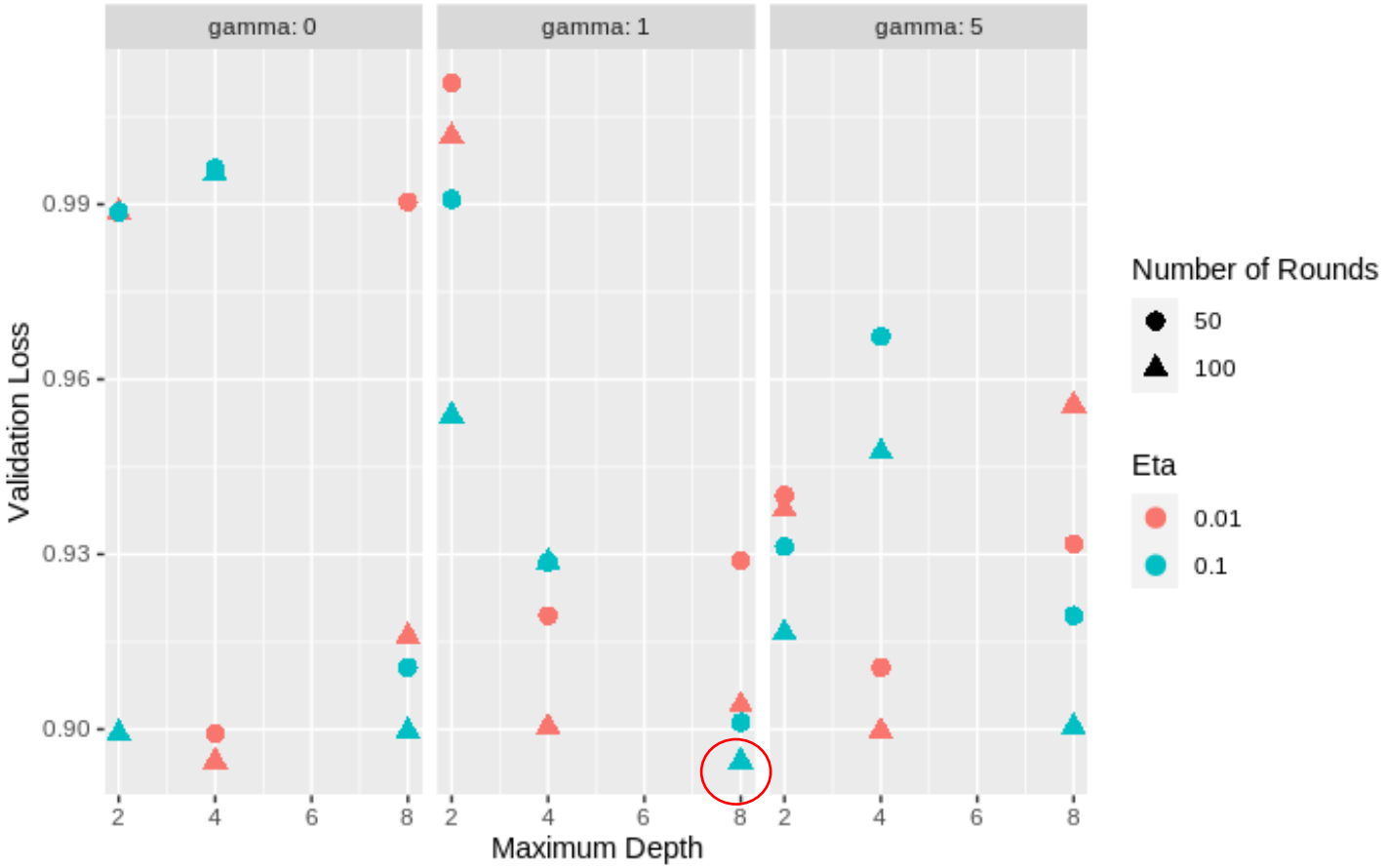



Image source: https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html

# HYPERPARAMETER TUNING

On 50 random genes

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



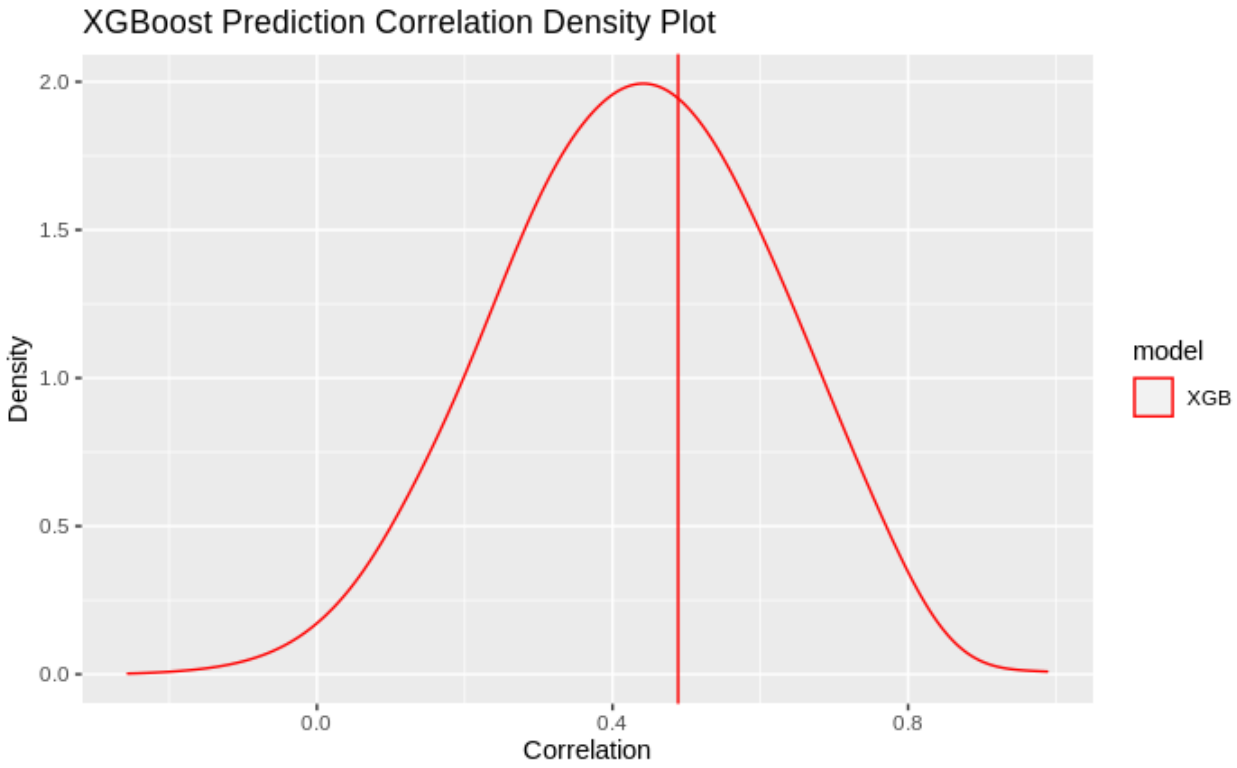

Loss Function

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

Best hyperparameter combination:
Gamma = 1
Maximum depth = 8
Eta = 0.1
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 | | |

# PERCEPTRON

SINGLE-LAYER NEURAL NETWORK

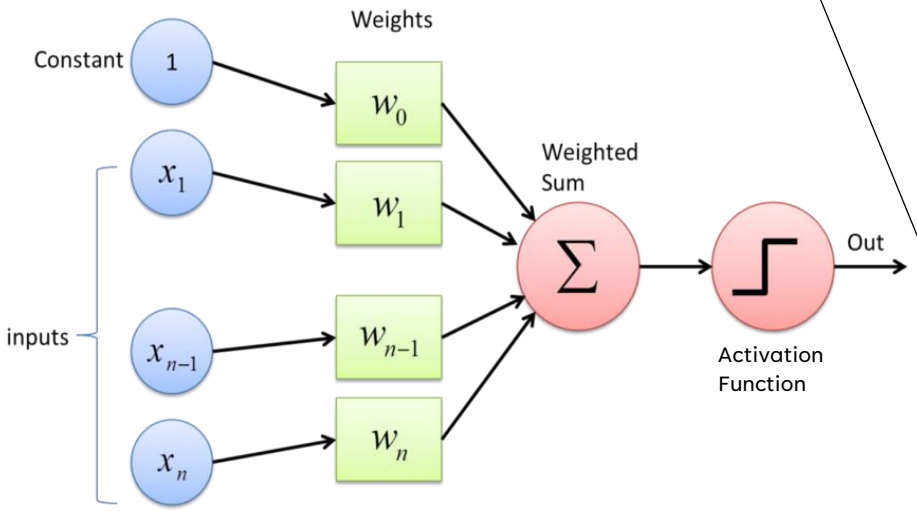



Image source: https://www.norwegiancreations.com/2019/04/introduction-to-neural-network/

# MULTILAYER PERCEPTRON

NEURAL NETWORK

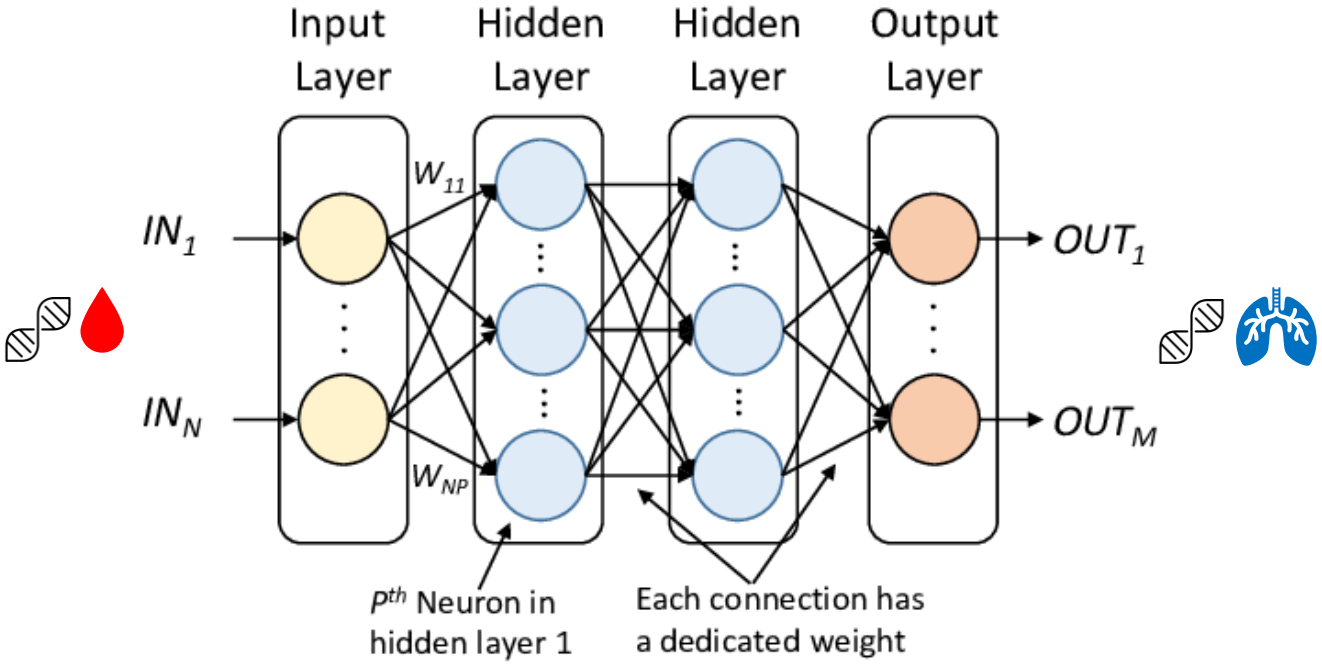
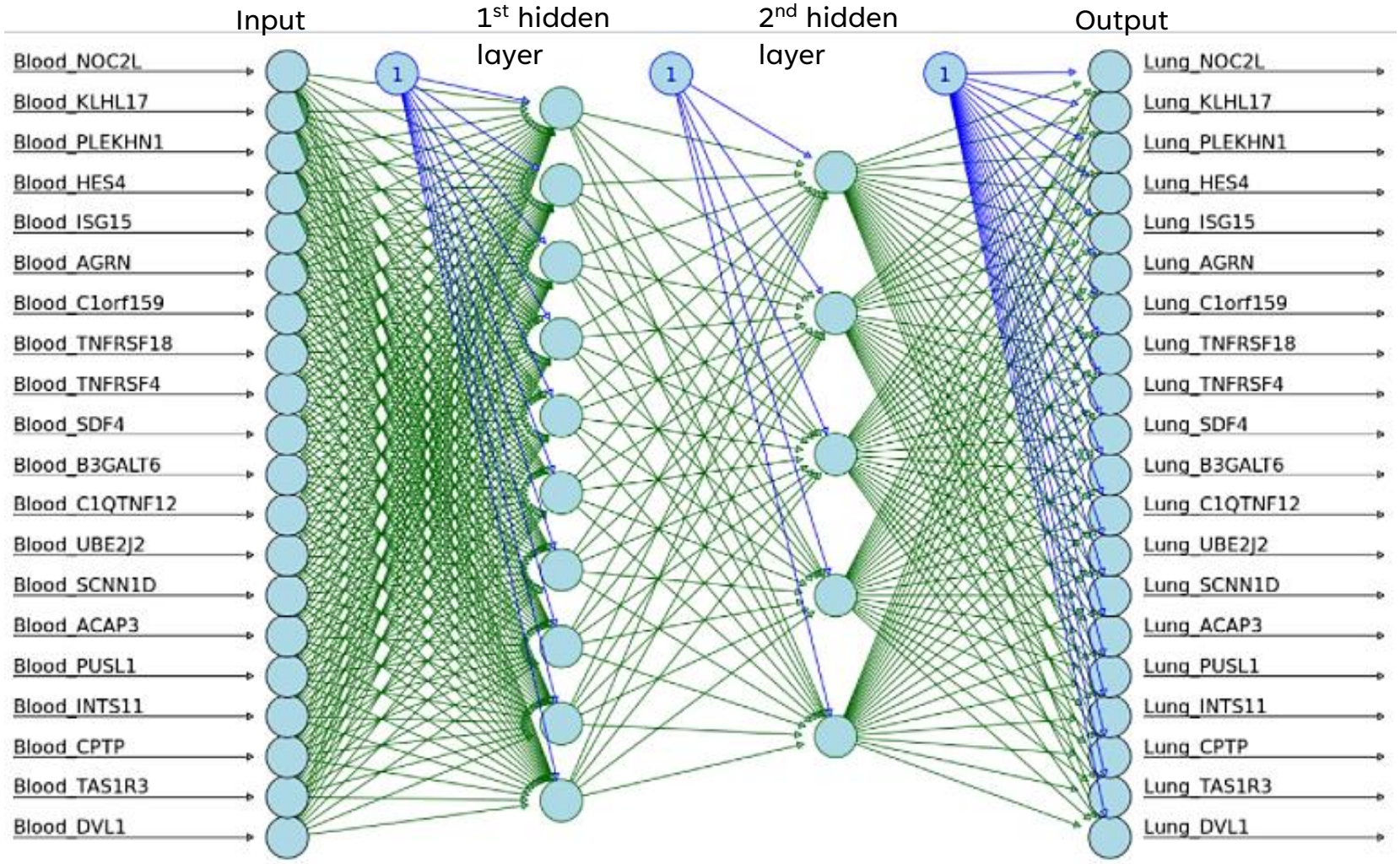


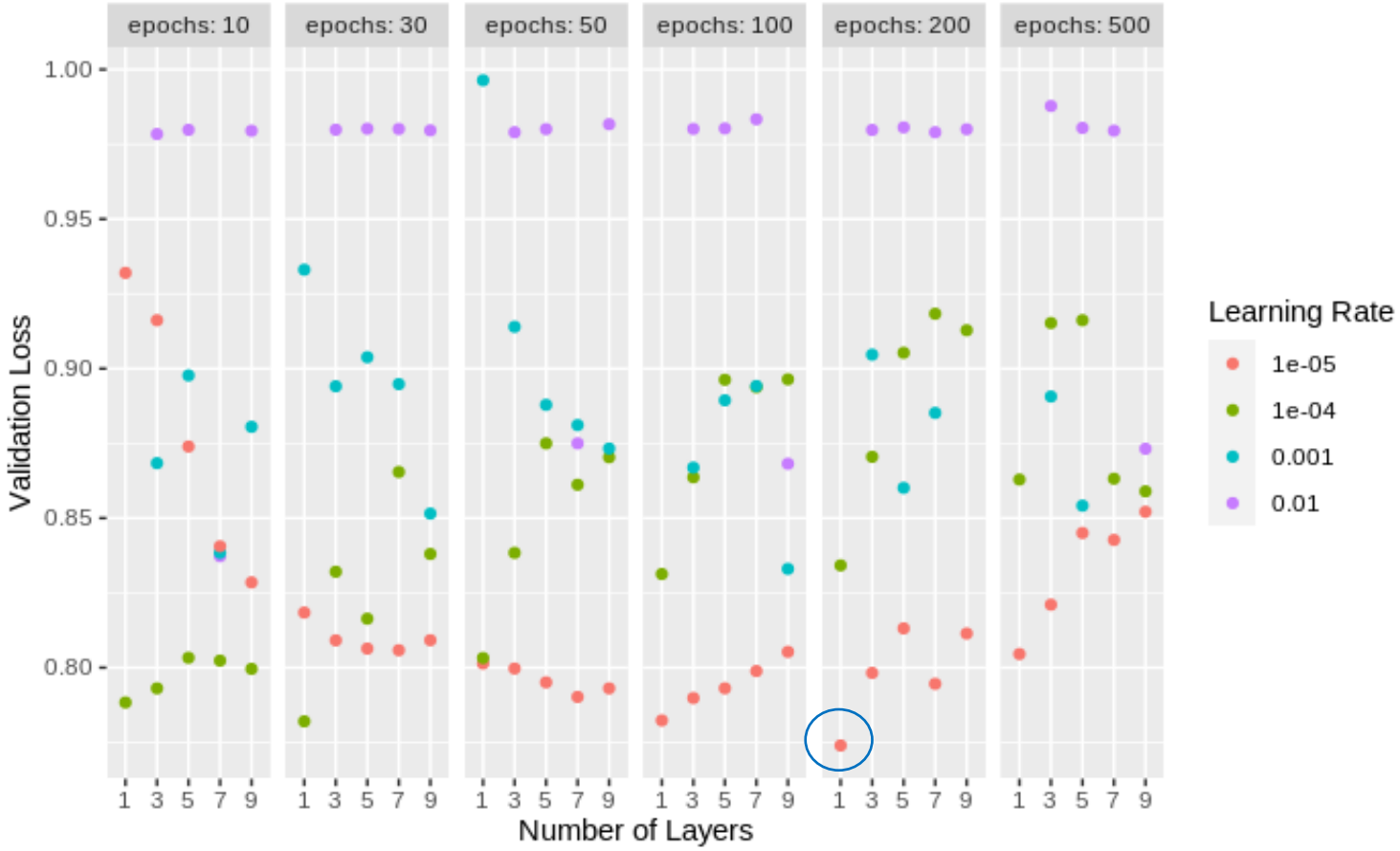

Image source: https://www.simplilearn.com/tutorials/deep-learning-tutorial/multilayer-perceptron

# NEURAL NETWORK ILLUSTRATION



# HYPERPARAMETER TUNING

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



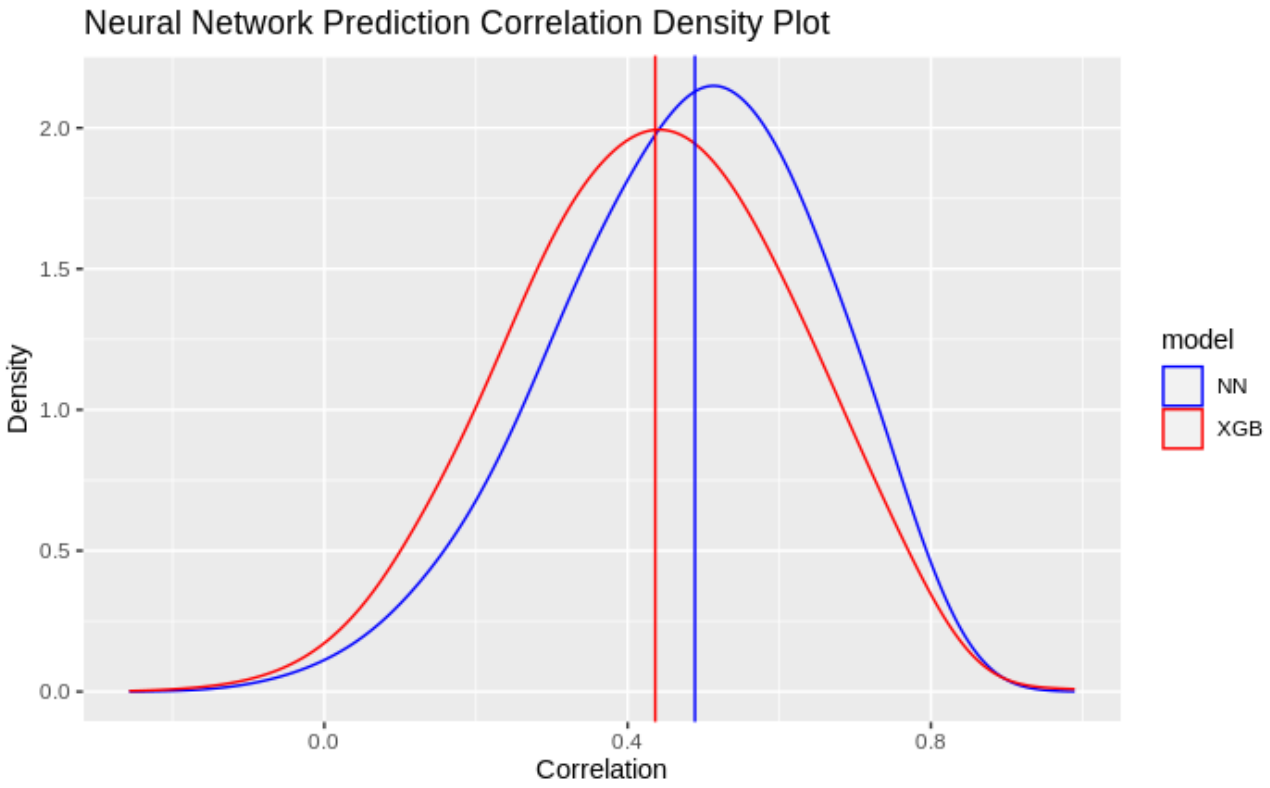

Loss Function

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

Best hyperparameter combination:
1 hidden layer of 100 units
Learning rate = $e^{-5}$
Epochs = 200

# EVALUATION

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



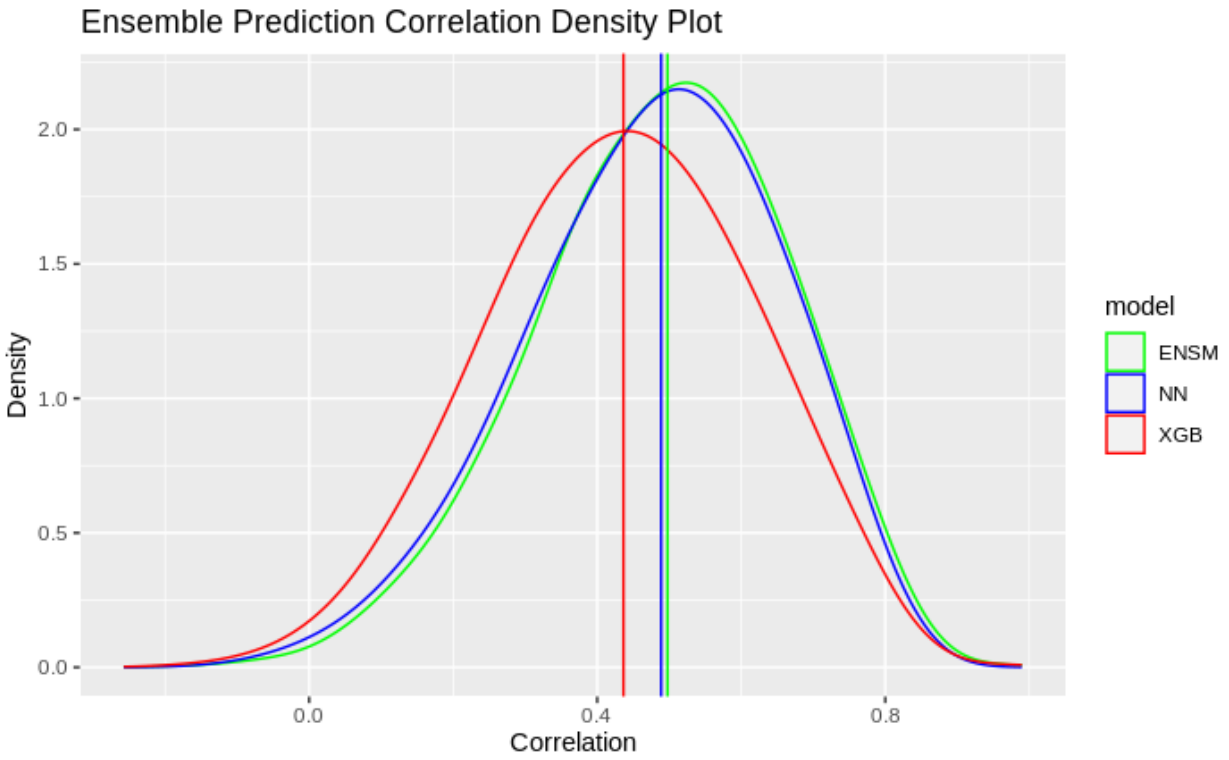

| Model | Median Correlation | |
|---|---|---|
| | Training | Test |
| XGBoost | 0.990975 | 0.4364047 |
| Neural Network | 0.6278658 | 0.4887672 |
| Ensemble | | |

# ENSEMBLE MODEL

$$ENSpred = \frac{1}{2}NNpred + \frac{1}{2}XGBpred = \frac{1}{2}\begin{bmatrix} \hat{y}_{11} & \cdots & \hat{y}_{1p} \\ \vdots & \ddots & \\ \hat{y}_{n1} & \cdots & \hat{y}_{np} \end{bmatrix} + \frac{1}{2}\begin{bmatrix} \hat{z}_{11} & \cdots & \hat{z}_{1p} \\ \vdots & \ddots & \\ \hat{z}_{n1} & \cdots & \hat{z}_{np} \end{bmatrix}$$

# 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 |

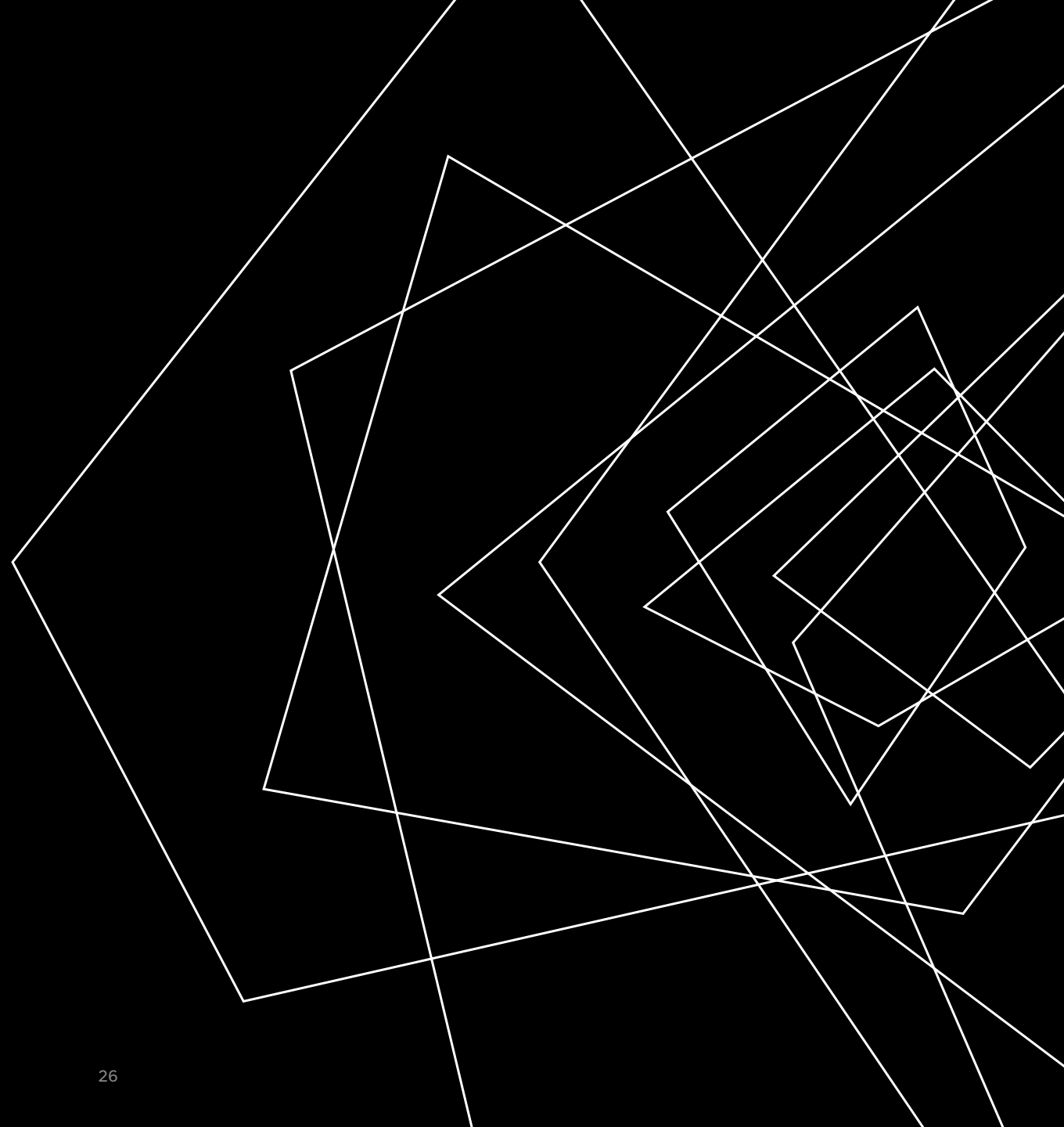
# FUTURE DIRECTIONS

Training the models on all other tissues, on the entire dataset

Stacking ensemble for improved prediction

Biological validation

Implementation of the web application

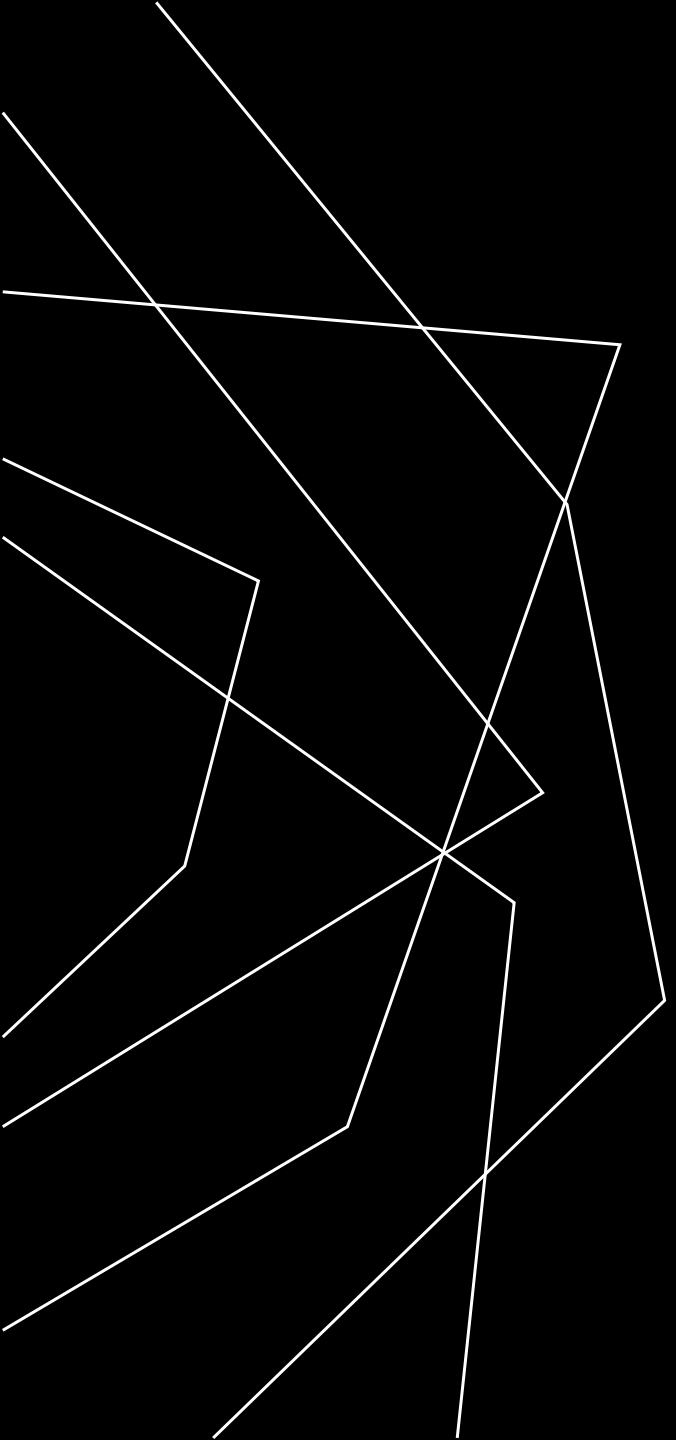


# WE WOULD LIKE TO EXTEND A HEARTFELT THANK YOU TO ALL OUR MENTORS INVOLVED IN THIS PROJECT!

Dr. Roberto Bonelli, CSL Research
Dr. Brendan Ansell, WEHI

Dr. Milica Ng, CSL Research
Dr. Monther Alhamdoosh, CSL Research
Dr. Melanie Bahlo, WEHI

Dr. Ziad Al Bkhetan, The University of Melbourne
Dr. Michael Kirley, The University of Melbourne



# THANK YOU