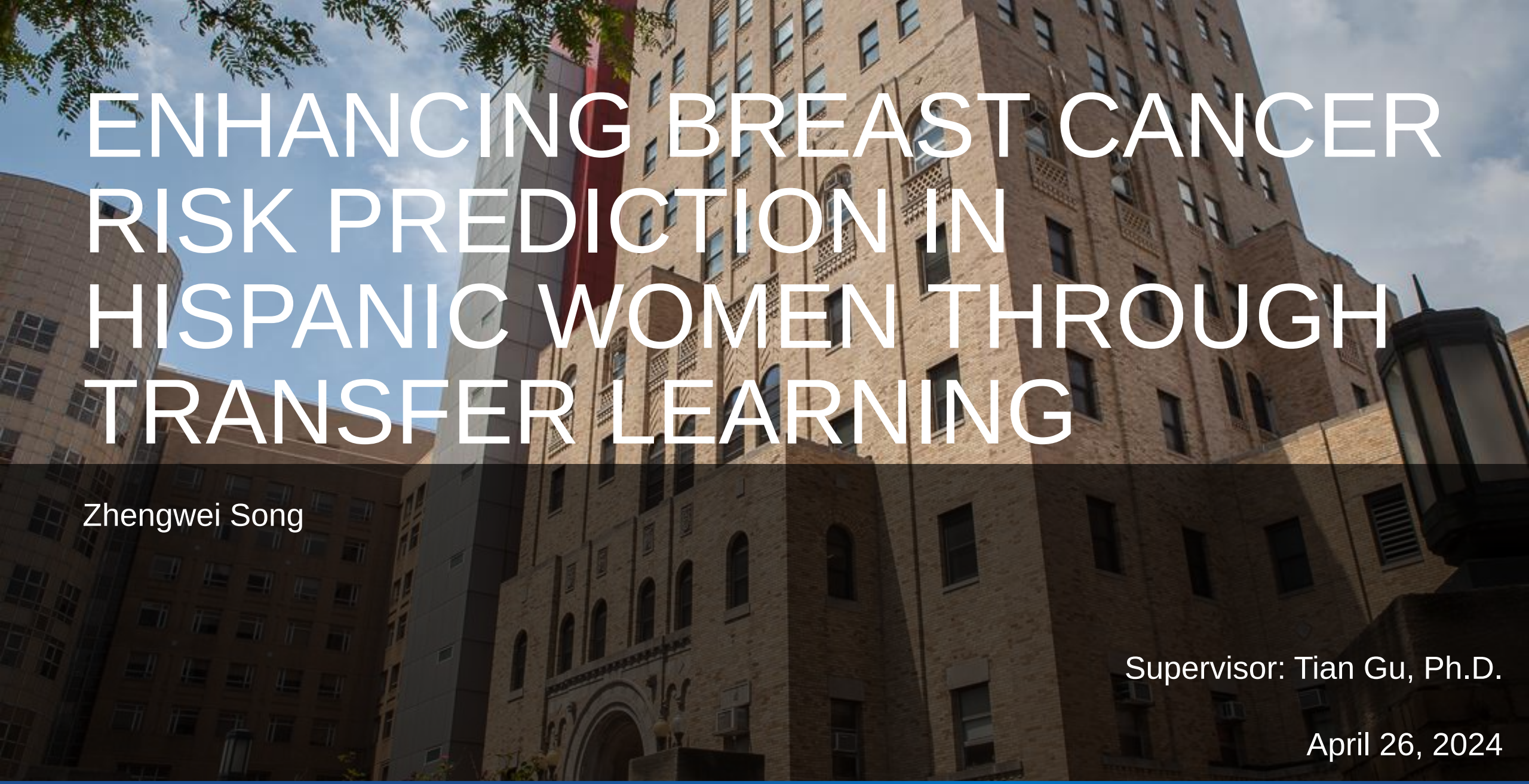


# ENHANCING BREAST CANCER RISK PREDICTION IN HISPANIC WOMEN THROUGH TRANSFER LEARNING

Zhengwei Song

Supervisor: Tian Gu, Ph.D.

April 26, 2024

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH

# Introduction

# Breast Cancer in Hispanic Women

- Second leading cause of cancer death.
- 30% lower diagnosis rate compared to non-Hispanic Whites.
- Increased risk of developing aggressive tumors: HER2+ and triple-negative breast cancer.
- Lack of effective prediction tools for early prevention

Current Studies:

- Genetic and clinical factors integration has improved risk stratification in White women, while underrepresentation in studies leads to suboptimal risk prediction in Hispanic women.

# Objective

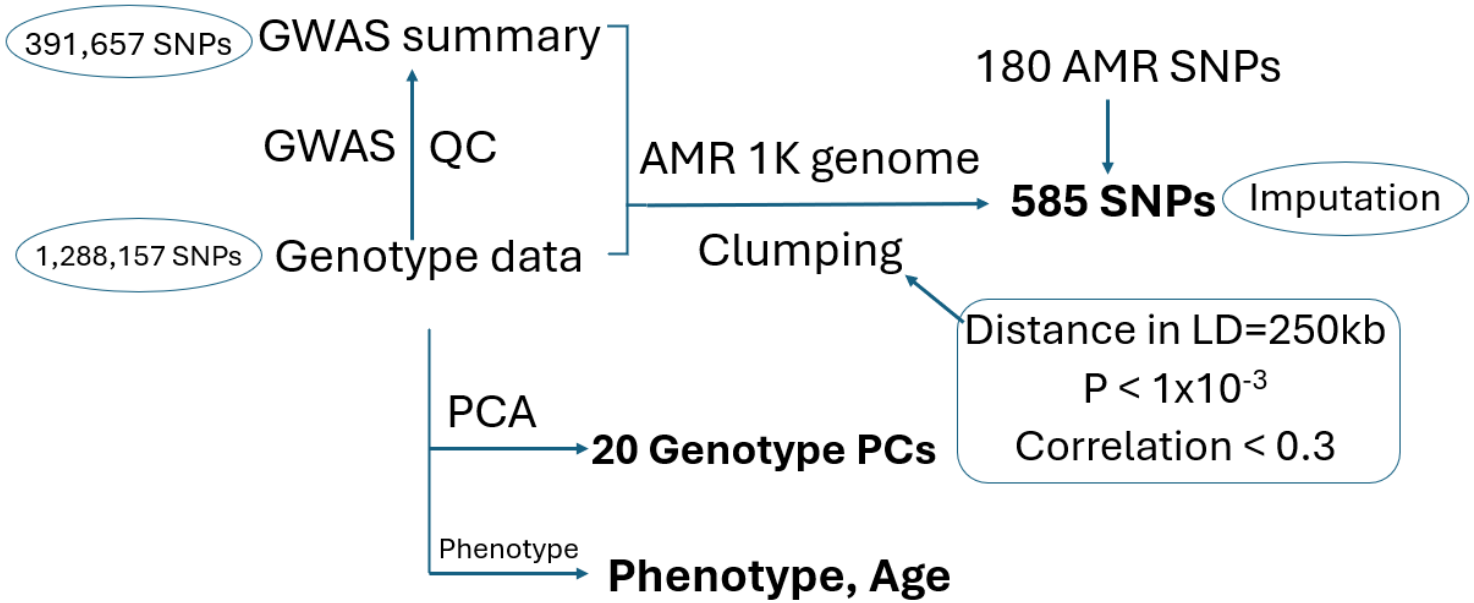
Implement novel **transfer learning** models to improve risk prediction by utilizing estimates from other populations (White and Asian) to benefit Hispanic women, that is, to **enhance the accuracy of breast cancer risk prediction for Hispanic women**.

# Data (Genotype & Phenotype)

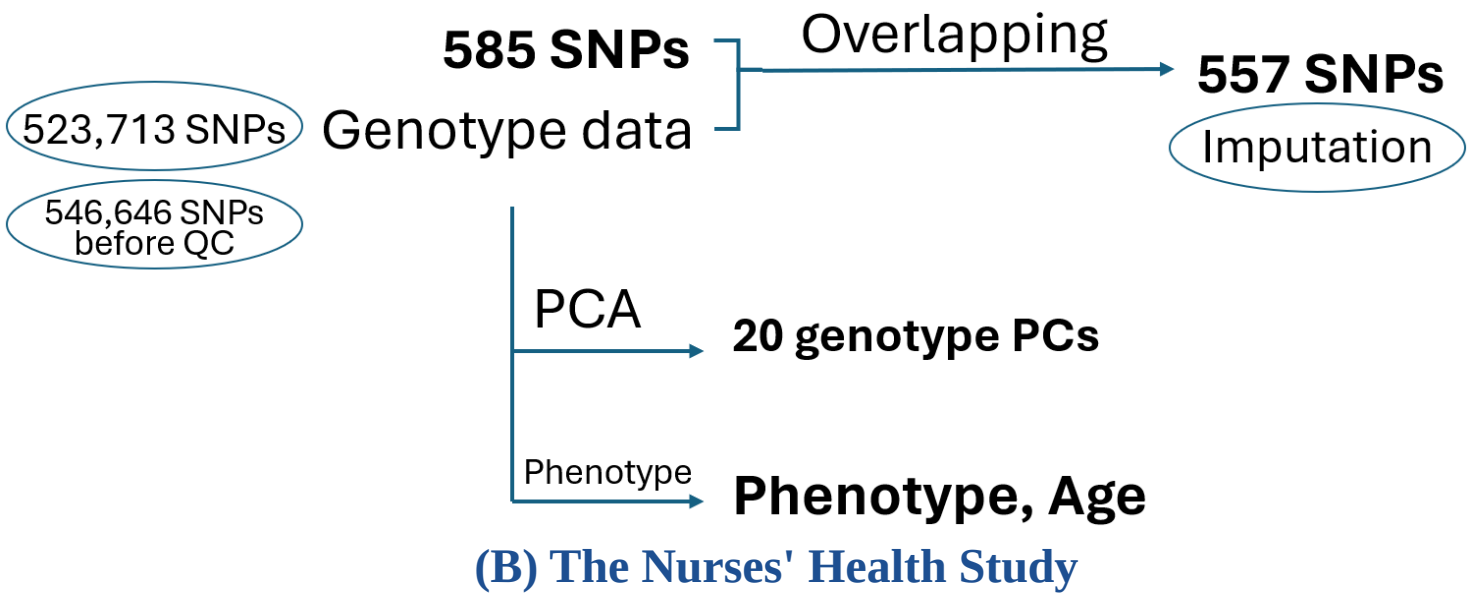
- **The Multiethnic Cohort Study**: a prospective cohort study that was initiated between 1993 and 1996 and includes subjects from various ethnic groups including **Hispanics** and **Japanese-Americans** primarily from greater Los Angeles, CA
- **The Cancer Genetic Markers of Susceptibility (CGEMS) study**: a prospective cohort study by National Cancer Institute (NCI) in **European** cohort that was provided a blood sample between 1989 and 1990 and were free of diagnosed breast cancer at blood collection and followed for incident disease until May 2004


- Target Data: **n=1064 Hispanics (AMR, 520 cases and 544 controls)** descent populations from The Multiethnic Cohort Study
- Two source data: **$n_1$=2287 European (EUR, 1145 cases & 1142 controls)** and **$n_2$=1707 Japanese (JPN, 885 cases & 822 controls)** descent populations, from CGEMS and the Multiethnic Cohort Study, respectively
- Both datasets were conducted another preprocessing steps, including outcome (i.e. Breast Cancer) normalization, missing data imputation, and Linkage Disequilibrium (LD) Clumping

# Data




**(A) The Multiethnic Cohort Study**




**(B) The Nurses' Health Study**

# Method

# Ridge

For each sample in each ethnic group, the linear model takes the form as

- Residualized Breast Cancer (y) ~ $\beta_0 + \beta_1*SNP_1 + \beta_2*SNP_2 + \ldots + \beta_{557}*SNP_{557} + \beta_{558}*PC_1 + \beta_{559}*PC_2 + \ldots + \beta_{577}*PC_{20} + \varepsilon$

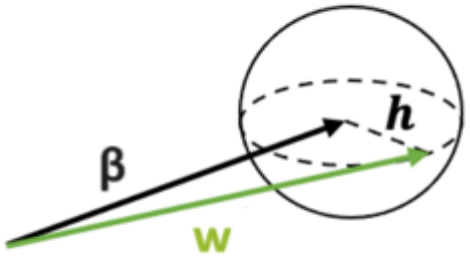
Additionally, with Ridge, the final parameter takes the form as

- $\widehat{\boldsymbol{\beta}}(\lambda) = \ argmin_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left( y_i - \ \beta_0 - \ \sum_{j=1}^{577} \beta_j x_{ij} \right)^2 + \ \lambda \ \sum_{j=1}^{577} \beta_j^2 \right\}$, where n=1707 for JPN, n=2287 for EUR, n=1064 for AMR

- Source Estimates: $\widehat{\boldsymbol{\beta}}_{JPN}, \widehat{\boldsymbol{\beta}}_{EUR}$

# Transfer Learning Models

A

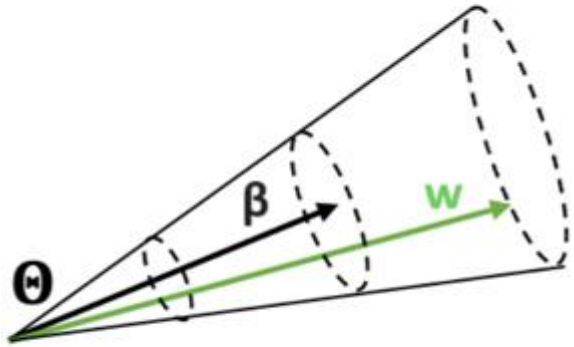



Distance-based similarity measure: $||\boldsymbol{\beta} - \boldsymbol{w}||_2 \leq \boldsymbol{h}$

$\widehat{\boldsymbol{w}}_{distTL} = argmin_w \frac{1}{n}\|Y_k - X_k w\|_2^2 + \lambda_d \left\|w - \hat{\beta}_k\right\|^2$, where $\lambda_d$ is the tuning parameters, k=1 or 2; $\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}}_{JPN}$, $\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}}_{EUR}$, or $\widehat{\boldsymbol{\beta}}_2 = (\widehat{\boldsymbol{\beta}}_{JPN}, \widehat{\boldsymbol{\beta}}_{EUR})$

**Geometric illustration of the distance-based similarity characterization (A); the angle-based characterization (B).**

B




Angle-based similarity measure: $\sin\Theta\,(\boldsymbol{\beta}, \boldsymbol{w}) \leq \boldsymbol{d}$

$\widehat{\boldsymbol{w}}_{angleTL} = argmin_w \frac{1}{n}\|Y_k - X_k w\|_2^2 + \lambda_a \|w\|^2 - 2\eta\hat{\beta}_k^T w$, where $\lambda_a$ and η are the tuning parameters, k=1 or 2; $\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}}_{JPN}$, $\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}}_{EUR}$, or $\widehat{\boldsymbol{\beta}}_2 = (\widehat{\boldsymbol{\beta}}_{JPN}, \widehat{\boldsymbol{\beta}}_{EUR})$

# Transfer Learning Models

- **Target-Only** Model (baseline model using AMR Only): $\widehat{\boldsymbol{w}}_{\boldsymbol{target}} = argmin_{\boldsymbol{w}} \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|^2$;

  → Testing $\widehat{\boldsymbol{y}}'_{\boldsymbol{target}} = \boldsymbol{X}'\widehat{\boldsymbol{w}}_{target} + \boldsymbol{\epsilon}$;

- **Source-Only** Model (directly using JPN or EUR): $\widehat{\boldsymbol{w}}_{\boldsymbol{src}} = argmin_{\boldsymbol{w}} \frac{1}{N_k}\|\boldsymbol{Y}_{\boldsymbol{k}} - \boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{w}\|_2^2 + \lambda_k\|\boldsymbol{w}\|^2$,

  → Testing $\widehat{\boldsymbol{y}}'_{\boldsymbol{src}} = \boldsymbol{X}'\widehat{\boldsymbol{w}}_{src} + \boldsymbol{\epsilon}$;

- Distance-based Transfer Learning (**DistTL**): $\widehat{\boldsymbol{w}}_{\boldsymbol{DistTL}} = argmin_{\boldsymbol{w}} \frac{1}{n}\|\boldsymbol{Y}_{\boldsymbol{k}} - \boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{w}\|_2^2 + \lambda_d\left\|\boldsymbol{w} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}\right\|^2$,

  → Testing $\widehat{\boldsymbol{y}'}_{\boldsymbol{DistTL}} = \boldsymbol{X}'\widehat{\boldsymbol{w}}_{\boldsymbol{DistTL}} + \boldsymbol{\epsilon}$;

- Angle-based Transfer Learning (**AngleTL**): $\widehat{\boldsymbol{w}}_{\boldsymbol{AngleTL}} = argmin_{\boldsymbol{w}} \frac{1}{n}\|\boldsymbol{Y}_{\boldsymbol{k}} - \boldsymbol{X}_{\boldsymbol{k}}\boldsymbol{w}\|_2^2 + \lambda_a\|\boldsymbol{w}\|^2 - 2\eta\hat{\beta}_k^T\boldsymbol{w}$

  → Testing $\widehat{\boldsymbol{y}}'_{\boldsymbol{AngleTL}} = \boldsymbol{X}'\widehat{\boldsymbol{w}}_{\boldsymbol{AngleTL}} + \boldsymbol{\epsilon}$;

- **Aggregated**: Testing $\widehat{\boldsymbol{y}'}_{\boldsymbol{weighted}} = w_1*\widehat{\boldsymbol{y}'}_{\boldsymbol{target}} + w_2*\widehat{\boldsymbol{y}'}_{\boldsymbol{src-JPN}} + w_3*\widehat{\boldsymbol{y}'}_{\boldsymbol{src-EUR}} + w_4*\widehat{\boldsymbol{y}'}_{\boldsymbol{DistTL-JPN}} + w_5*\widehat{\boldsymbol{y}'}_{\boldsymbol{DistTL-EUR}} + w_6*\widehat{\boldsymbol{y}'}_{\boldsymbol{DistTL-J\&E}} + w_7*\widehat{\boldsymbol{y}'}_{\boldsymbol{AngleTL-JPN}} + w_8*\widehat{\boldsymbol{y}'}_{\boldsymbol{AngleTL-EUR}} + w_9*\widehat{\boldsymbol{y}'}_{\boldsymbol{AngleTL-J\&E}}$

  where weight w is obtained from independent **training data** from

$\boldsymbol{y}_{\mathbf{AMR}} = w_1*\widehat{\boldsymbol{y}}_{\boldsymbol{target}} + w_2*\widehat{\boldsymbol{y}}_{\boldsymbol{src-JPN}} + w_3*\widehat{\boldsymbol{y}}_{\boldsymbol{src-EUR}} + w_4*\widehat{\boldsymbol{y}}_{\boldsymbol{DistTL-JPN}} + w_5*\widehat{\boldsymbol{y}}_{\boldsymbol{DistTL-EUR}} + w_6*\widehat{\boldsymbol{y}}_{\boldsymbol{DistTL-J\&E}} + w_7*\widehat{\boldsymbol{y}}_{\boldsymbol{AngleTL-JPN}} + w_8*\widehat{\boldsymbol{y}}_{\boldsymbol{AngleTL-EUR}} + w_9*\widehat{\boldsymbol{y}}_{\boldsymbol{AngleTL-J\&E}}$

# Transfer Learning Models

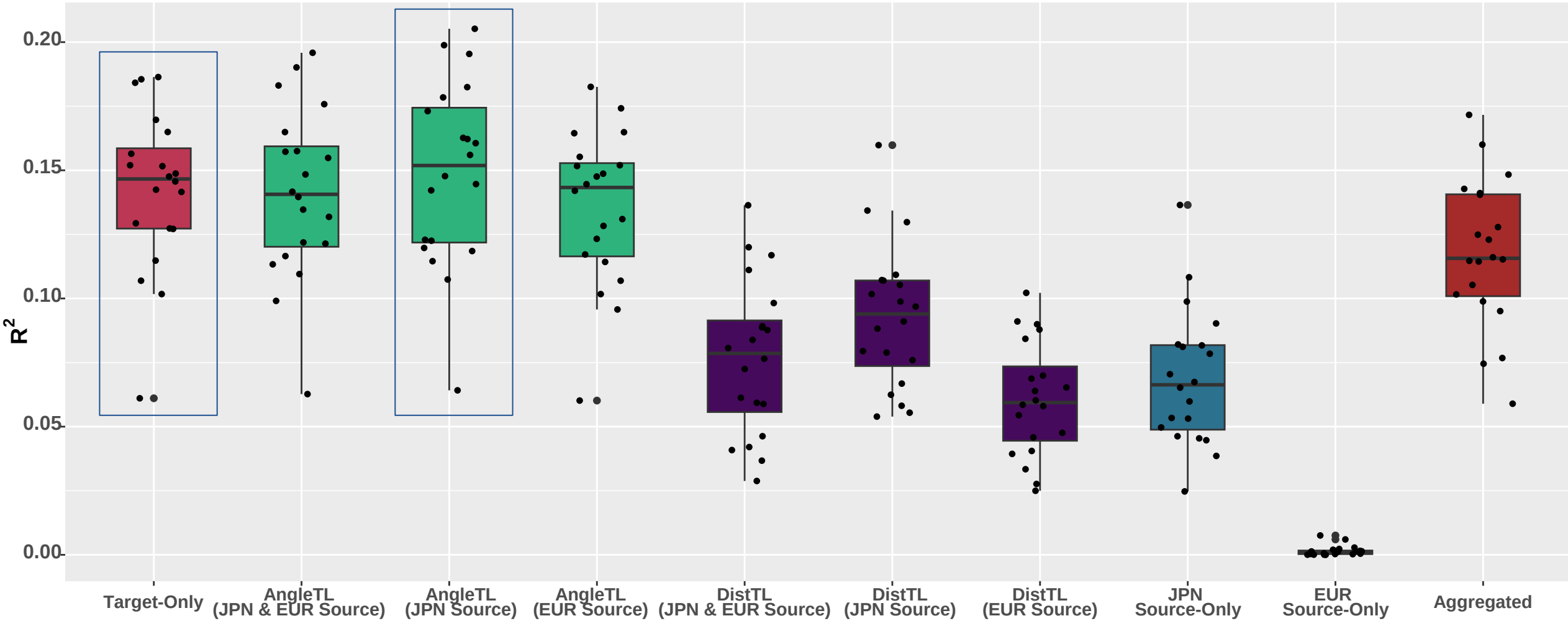
- **Comparison**

Model Performance Measure:

$$R^2 = (\text{Correlation between predicted } \widehat{y'} \text{ and observed } y'_{AMR})^2$$

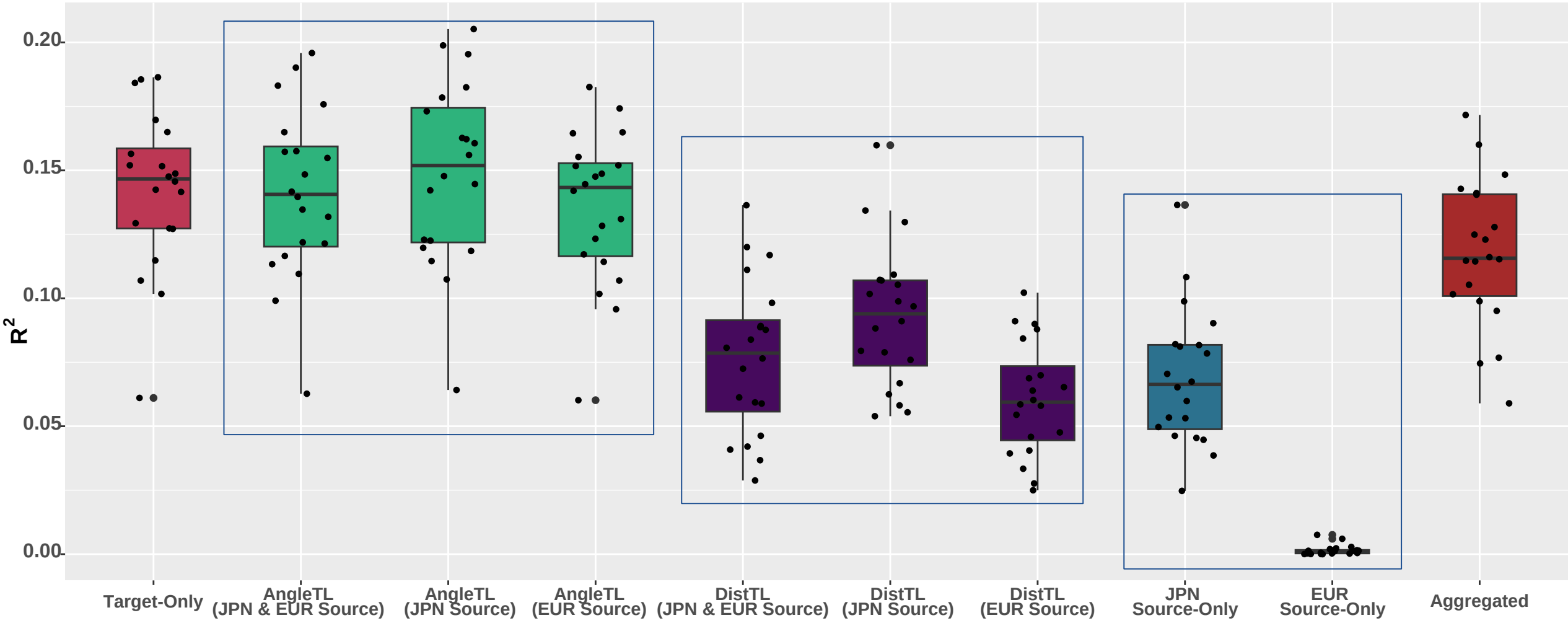
# Result

# $R^2$ across models



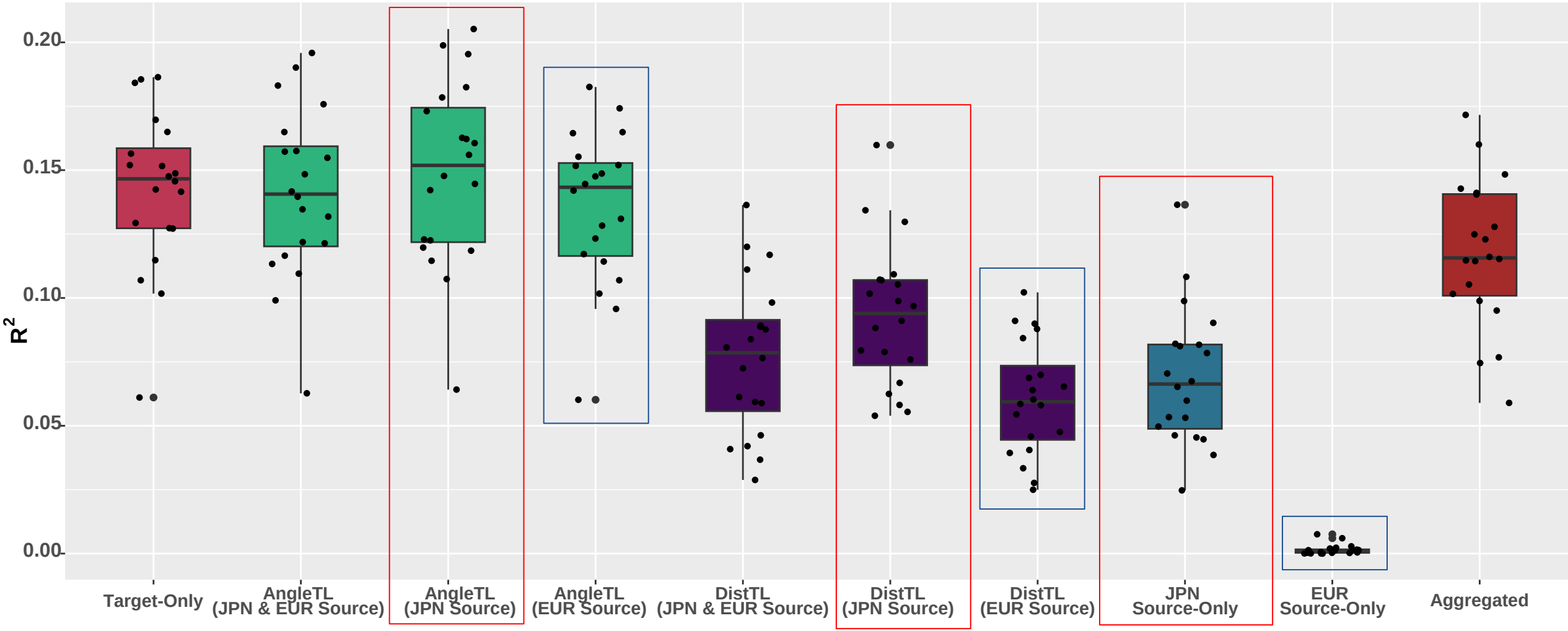

- AngleTL with JPN source showed a >10% improvement in $R^2$ values over target-only model

# $R^2$ across models




- AngleTL has superior predictive accuracy compared to source-only model and DistTL
- Under the same source, the performance of transfer learning (AngleTL or DistTL) is consistent with source-only.

# $R^2$ across models




- Models using EUR source all perform poorly, i.e. the similarity between AMR and EUR is low.

# Discussion

# Discussion

SNP Clumping & Data Quality:

- Reduced genetic marker redundancy
- Isolated most representative SNPs for analysis, which improved reliability of subsequent findings

Transfer Learning Performance:

- Potential to address risk prediction disparities in underrepresented populations, with utilization of data from well-researched populations to benefit underserved groups
- Demonstrated potential of transfer learning in genetic research

# Thank you

# Appendix

# Breast Cancer Genotype Data Quality Control (QC)

- Step1: **Genotype Missing Rate < 5%** (--geno 0.05), i.e. removing SNPs with NA accounting for more than 5%

- Step2: **Sample Missing Rate < 5%** (--mind 0.05), i.e. removing samples with NA accounting for more than 5%

- Step3: **Minor Allele Frequency (MAF) > 1%** (--maf 0.01), i.e. removing SNPs with MAF < 1%

- Step4: **Hardy-Weinberg Equilibrium (HWE) with p-value < $10^{-6}$** (--hwe 0.000001), i.e. removing SNPs with the p-value > $10^{-6}$