WiDS Datathon++ 2025: Predicting Brain Age and Exploring Sex Differences

Ruihang Han
Boston University
jasonhan@bu.edu

2025-05-04

1. Introduction

Understanding how brain networks develop differently between males and females during childhood and adolescence is critical to identifying sex-specific vulnerabilities to neuropsychiatric disorders. The WiDS Datathon++ 2025 challenges participants to model brain age using resting-state functional connectivity data and to explore sex-based differences in model behavior and neurodevelopmental patterns.

In this study, we aim to predict chronological age from high-dimensional fMRI-derived connectome features and associated metadata in a population of children and adolescents aged 5 to 21. Each subject's brain is represented as a 200×200 functional connectivity matrix, vectorized into ~20,000 features, and enriched with demographic, behavioral, and psychological information (e.g., sex, BMI, factor scores).

We employ multiple machine learning approaches, including Ridge regression and XGBoost, to build accurate age prediction models. To reduce dimensionality and improve interpretability, we apply Principal Component Analysis (PCA) and Lasso-based variable selection. We then conduct a detailed analysis of sex-specific differences in model performance, feature contributions, and brain network patterns, using tools such as residual diagnostics, SHAP values, partial dependence plots, and connectome-based network analysis.

Our goal is twofold: (1) to achieve robust brain-age prediction and (2) to uncover interpretable sex differences in functional brain development, contributing to a better understanding of normative and atypical trajectories during youth.

2. Data and Preprocessing

The training dataset consists of 1,104 adolescent participants between the ages of 5 and 21. Each participant is associated with two primary sources of data: a functional brain connectivity matrix derived from resting-state fMRI scans, and a set of metadata that includes demographic, physiological, and psychological information.

Each fMRI-derived connectome is represented as a 200×200 symmetric matrix, capturing pairwise correlations among 200 predefined brain regions. To render these matrices suitable for modeling, only the upper triangular portion was extracted and vectorized, resulting in approximately 19,900 connectivity features per subject. A parallelized pipeline was employed to expedite the processing of over one thousand individual .tsv files, each corresponding to a single participant.

Following feature extraction, the connectivity vectors were merged with metadata obtained from the accompanying training_metadata.csv file. Key metadata variables include the target outcome (chronological age),

as well as sex, race, ethnicity, body mass index (BMI), parental education, handedness, study site, and four standardized psychological factors (p-factor, internalizing, externalizing, and attention scores).

A comprehensive quality assessment of the merged dataset was conducted to verify structural consistency and evaluate data completeness. Fewer than 0.01% of values were missing, predominantly in categorical fields such as race, ethnicity, and parental education. To address missingness, continuous variables were imputed using the median of observed values, while categorical variables were encoded with an explicit "Missing" level to preserve potential patterns associated with nonresponse.

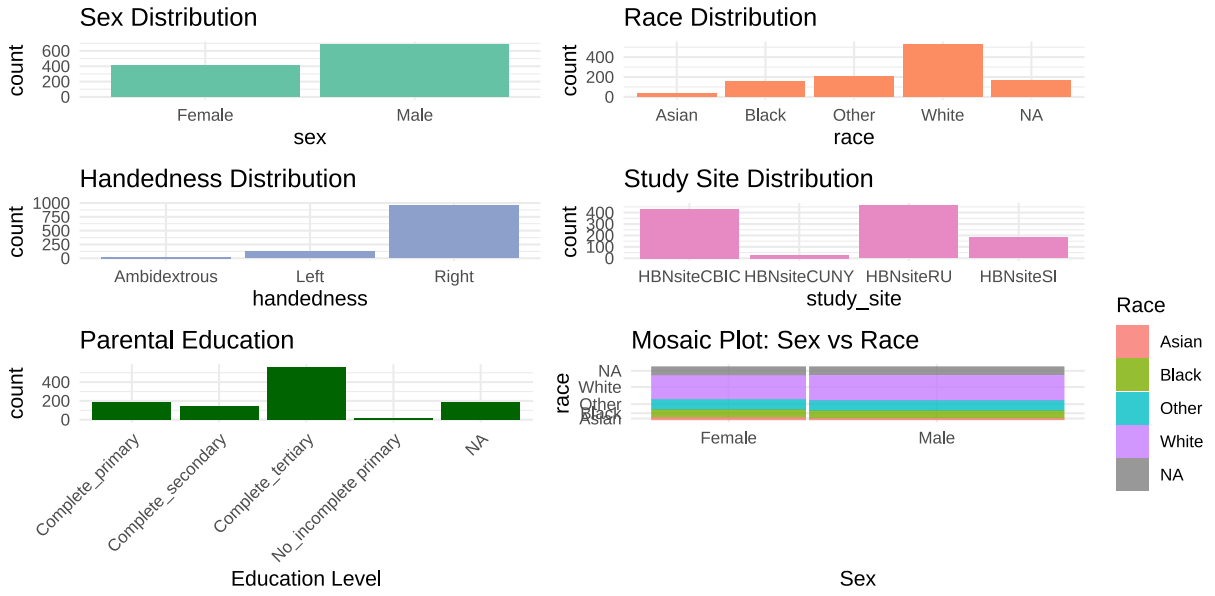
The final dataset comprises 1,104 observations and over 20,000 features, and was stored in both .csv and .rds formats to facilitate efficient access for subsequent modeling tasks.

3. Exploratory Data Analysis (EDA)

Note: The exploratory data analysis (EDA) presented in this section was conducted by my teammate and integrated here with light modifications.

3.1 Demographic Variable Distributions

We began by examining categorical metadata to understand the demographic makeup of the dataset. Frequency counts and bar plots reveal the distribution across variables such as sex, race, handedness, study site, and parental education.



The dataset reveals a moderate gender imbalance (688 males vs. 416 females), and a predominantly right-handed population. White is the most common race, followed by Other, Black, and Asian groups. Most participants were recruited from HBNsiteRU and HBNsiteCBIC, while parental education levels skew toward tertiary completion. The dataset shows a gender imbalance, with more male participants (688) than female (416). In terms of race, the majority identify as White (530), followed by Other (210), Black (157), and Asian (39), with some missing values. The handedness distribution is heavily skewed toward right-handed individuals (954), with fewer left-handed (121) and ambidextrous (29) participants. Most data were collected from two major study sites, HBNsiteRU and HBNsiteCBIC, while HBNsiteCUNY contributed the fewest. Regarding parental education, a large portion of Parent 1 had completed tertiary education (562), and

similarly high levels were observed for Parent 2. Ethnicity data indicates most participants are not Hispanic or Latino (712), while 283 are identified as Hispanic or Latino.

Race distribution is largely consistent across sex categories. White participants are the majority in both groups, followed by Other and Black. This consistency suggests minimal interaction between race and sex in this sample.

3.2 Summary Statistics of Continuous Variables

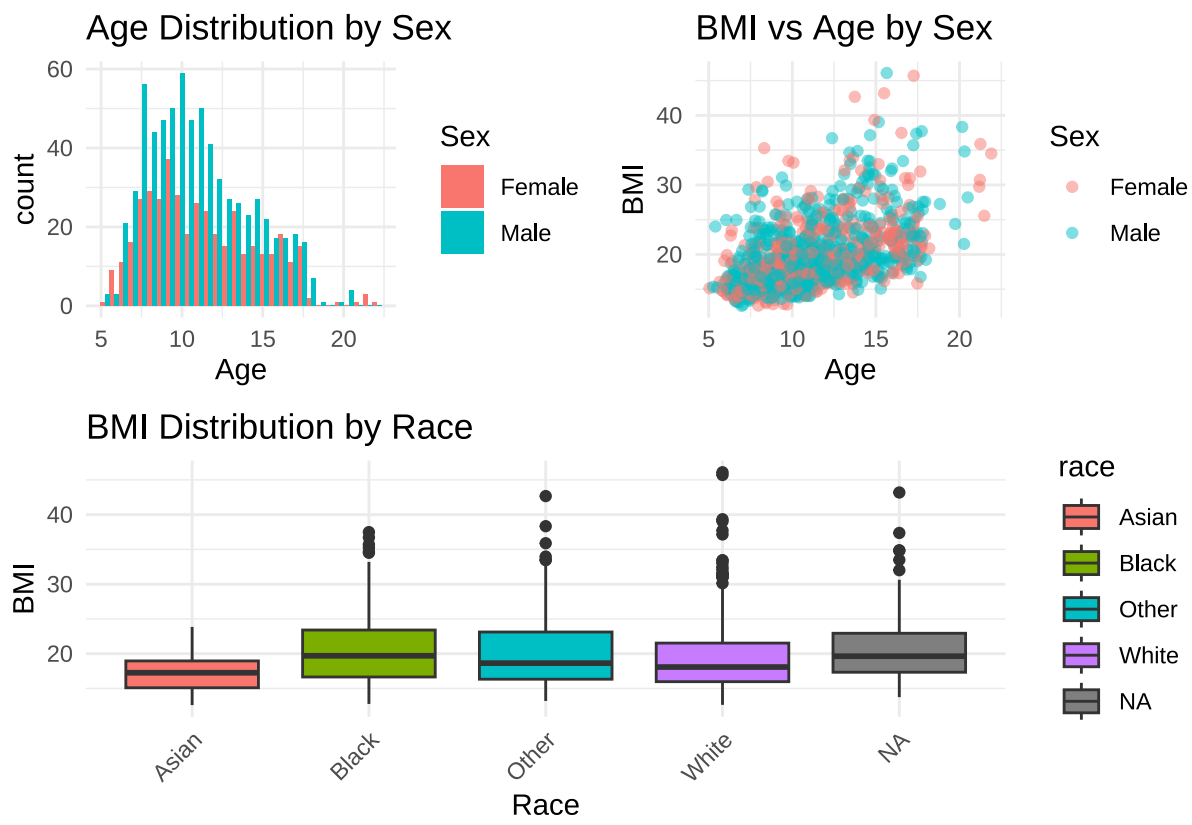
We summarized key continuous variables—age, BMI, and standardized psychological factors—using descriptive statistics.

Participants range in age from approximately 5 to 22 years (mean 11.2), with BMI spanning from 12.6 to 46.1. Mental health scores are z-scored and exhibit moderate variability. The distributions provide a foundation for modeling age while accounting for psychological and developmental diversity.

All psychological factor scores are z-scored, with mean values close to 0, as expected. The p-factor, a general psychopathology indicator, shows a range from -1.61 to 2.98. Internalizing and externalizing scores (e.g., anxiety/depression vs. behavioral dysregulation) also vary broadly, from -2.26 to 2.82 and -2.15 to 4.24, respectively. The attention factor spans from -3.18 to 2.48, with slightly more negative skew. These distributions reflect moderate psychological variability across participants and support the use of these variables as covariates or predictors in modeling brain age. The presence of 18 missing BMI values will need to be addressed via imputation or exclusion strategies in downstream analysis.

3.3 Age and BMI Trends

To explore growth and health patterns, we examined age and BMI trends across sex and race groups using histograms, scatterplots, and boxplots.



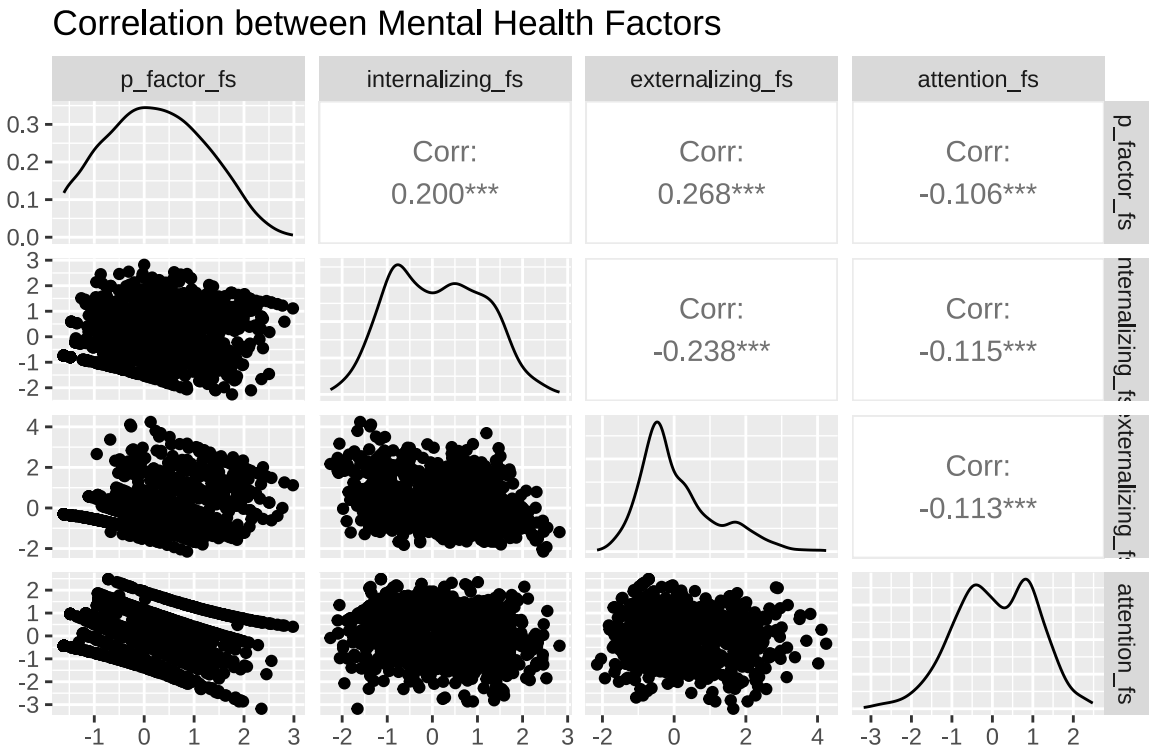
The age distribution is slightly right-skewed, with most participants concentrated between ages 8 and 13. Males outnumber females at nearly every age, consistent with the overall sex imbalance observed earlier.

The BMI scatterplot reveals a general upward trend with age, reflecting natural growth and weight gain. Both sexes show a similar trajectory, though male participants are slightly more represented in the higher BMI range at older ages.

The boxplot of BMI by race shows meaningful variation across groups. Asian participants tend to have the lowest median BMI and narrower spread, whereas Black participants display higher median BMI and greater variability, including more high-end outliers. These patterns may reflect differences in body composition, socioeconomic factors, or access to healthcare and nutrition.

3.4 Psychological Factors and Group Differences

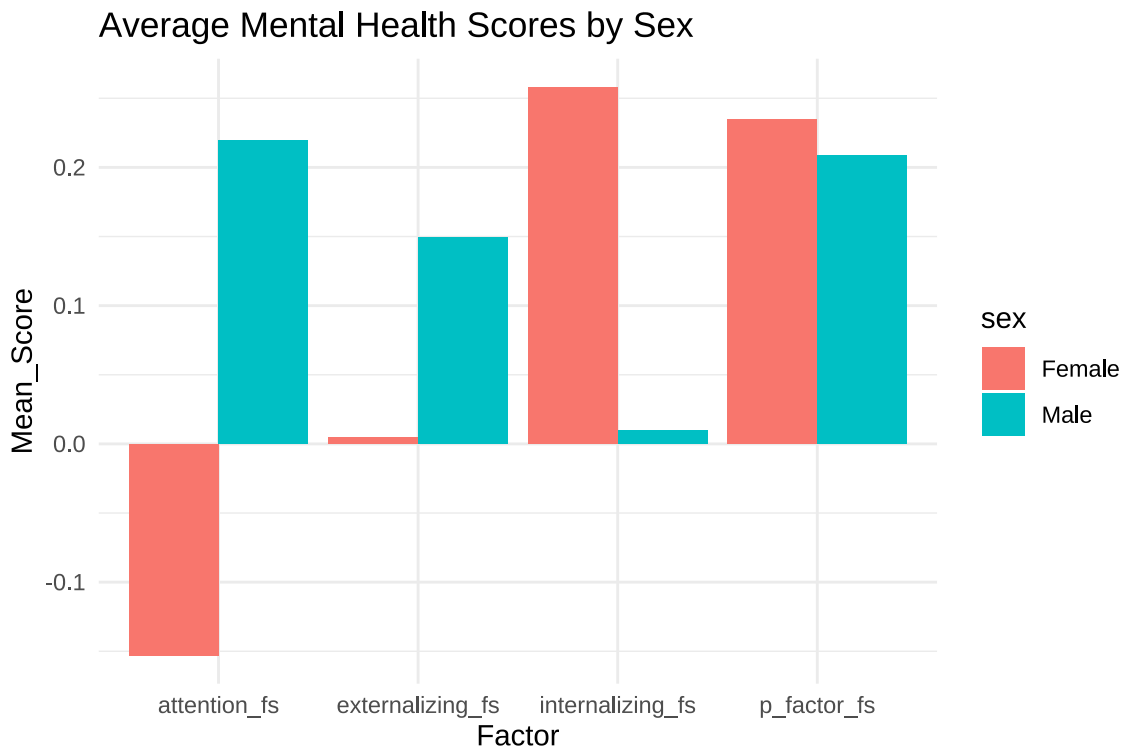
We examined the structure and sex-based distribution of psychological scores, including the general psychopathology factor (`p_factor_fs`) and three specific domains: internalizing, externalizing, and attention. These scores are standardized (z-scored) and reflect latent dimensions derived from psychological assessments.



Moderate positive correlation between `p_factor_fs` and both `internalizing_fs` (r 0.20) and `externalizing_fs` (r 0.27), indicating that higher general psychopathology is associated with greater emotional and behavioral problems.

Negative correlations between `attention_fs` and the other factors, especially with `externalizing_fs` (r –0.24), suggest that attentional difficulties may constitute a separate dimension of mental health burden, distinct from mood- and behavior-related symptoms.

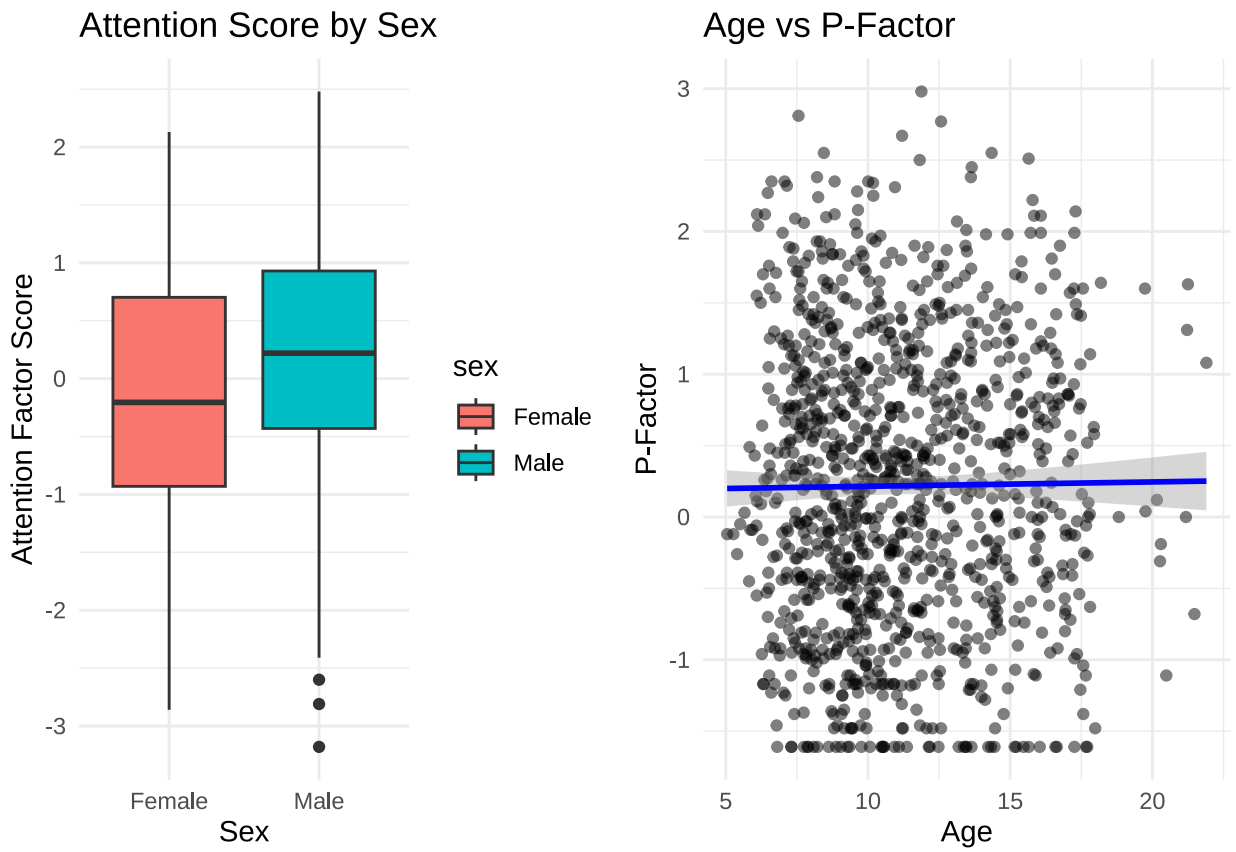
The density plots along the diagonal show that all variables are approximately normally distributed with mild deviations.



The bar chart displays the average scores of four standardized psychological factors by sex: general psychopathology (`p_factor_fs`), internalizing symptoms (e.g., anxiety and depression), externalizing behaviors (e.g., aggression and rule-breaking), and attention-related difficulties (e.g., inattention and hyperactivity). Clear differences emerge across sexes:

- **Females** exhibit higher mean values in both the `p_factor_fs` and `internalizing_fs` domains, suggesting a greater burden of generalized emotional symptoms and internal distress.
- **Males** score higher in `externalizing_fs` and, most notably, in `attention_fs`, indicating that behavioral dysregulation and attentional difficulties are more prevalent among male participants.

The negative mean `attention_fs` score among females contrasts sharply with the positive score among males, which is consistent with developmental psychology literature. Studies have shown that males are more likely to exhibit overt behavioral symptoms such as hyperactivity and impulsivity, while females tend to present with more internalized symptoms. These sex-linked patterns provide strong justification for stratifying modeling outcomes by sex in subsequent analyses and underscore the potential for differential neurodevelopmental trajectories between males and females.

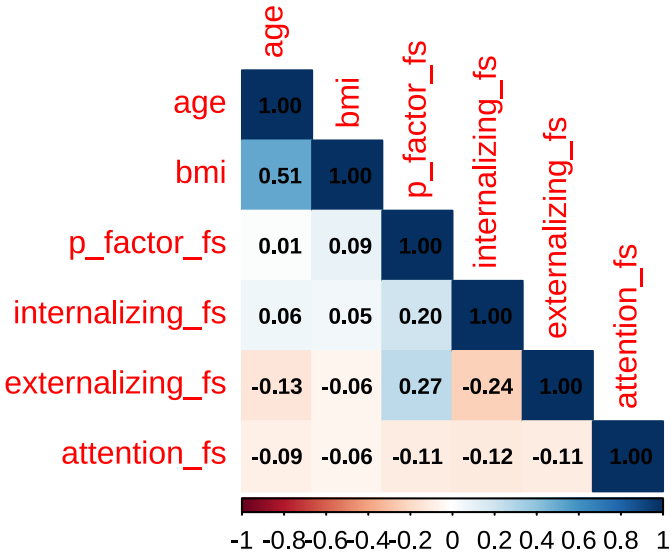


The boxplot shows that males have higher attention factor scores than females, both in terms of median and overall range. Male participants display more variability and a greater number of high outliers, indicating more pronounced attention-related difficulties. In contrast, female scores are generally lower and more tightly distributed, suggesting fewer symptoms of inattention or hyperactivity.

The scatterplot of age and p-factor scores shows little to no relationship between the two. The trend line is flat, indicating that general psychological distress levels remain stable across the sampled age range.

3.5 Correlation Matrix of Continuous Predictors

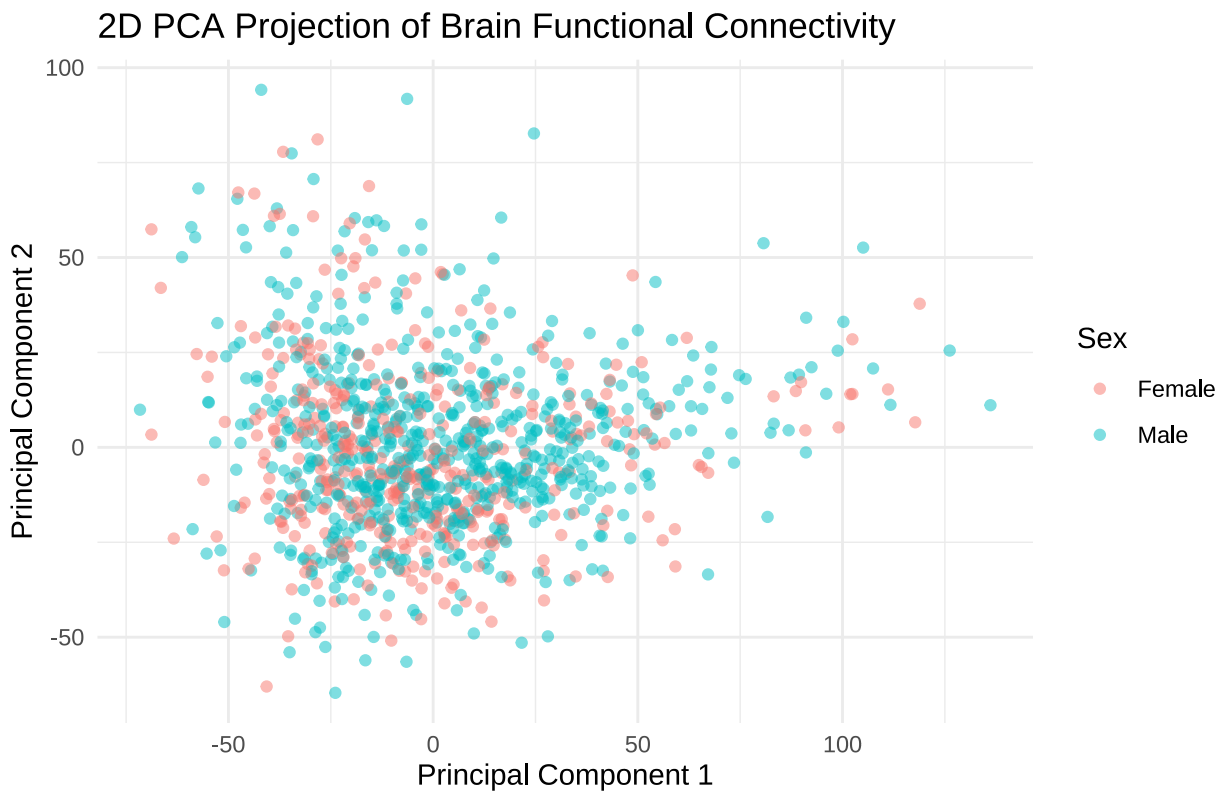
We constructed a correlation heatmap across key continuous predictors to identify linear dependencies and potential multicollinearity:



Age and BMI show a moderate positive correlation (r = 0.51), which is consistent with expected growth patterns. The psychological scores—p-factor, internalizing, externalizing, and attention—are only weakly correlated with each other and with age or BMI. This indicates that they represent distinct constructs and can be modeled independently without strong concern for multicollinearity.

3.6 PCA of Brain Connectivity Features

To explore the structure of high-dimensional functional connectivity features, we applied Principal Component Analysis (PCA) to the set of connectome variables. PCA provides a reduced-dimensional representation that captures the most salient variance in the data, making it useful for visualizing broad patterns such as potential grouping by sex.



The PCA projection onto the first two components reveals substantial overlap between sexes, with no clear clustering or boundary that separates male and female participants. This suggests that sex does not account for large-scale variance in the dominant principal components. Instead, brain connectivity patterns appear to be more individually distributed or may vary along more subtle dimensions not captured in PC1 and PC2. This visual result supports the need for more nuanced modeling approaches when analyzing sex-related brain connectivity differences.

4. Modeling and Prediction

4.1 Data Import and Splitting

The dataset was first loaded from a preprocessed `.rds` file. To evaluate model performance, we split the data into training (80%) and validation (20%) sets using stratified sampling based on the `sex` variable to preserve group proportions in both subsets.

4.2 PCA for Dimensionality Reduction (90% Variance)

To reduce dimensionality and mitigate noise in the connectome features, we applied Principal Component Analysis (PCA) to the subset of variables starting with `V`, which represent the functional connectivity structure. Components were standardized before decomposition. We retained the minimum number of principal components required to explain at least 90% of the total variance, balancing representational fidelity and computational efficiency. The same transformation was applied to both training and validation sets using the loadings obtained from the training data.

```
## Number of principal components to retain: 559
```

The PCA procedure indicated that 559 components were required to retain 90% of the total variance in the connectome features. This relatively large number reflects the high dimensionality and complexity of the original connectivity matrix. While PCA helped reduce noise and multicollinearity, the slow decay in variance suggests that information is broadly distributed across many dimensions, limiting the potential for aggressive dimensionality reduction.

4.3 Metadata Processing and Encoding

We extracted and preprocessed relevant metadata variables including demographic factors (e.g., sex, race, handedness), cognitive assessments (e.g., p-factor, internalizing, externalizing, attention), and study site information. Missing values in continuous variables such as BMI were imputed using the median. For categorical variables, missing levels were explicitly retained as a separate category labeled "Missing" to preserve potentially informative patterns. All categorical variables were then converted to dummy (one-hot) encoded format to ensure compatibility with downstream machine learning models.

4.4 Merging PCA Features and Metadata

To prepare the final modeling datasets, we combined the PCA-transformed connectome features with the encoded metadata using `participant_id` as the join key. This ensured that each observation retained both its neuroimaging and contextual information. We then defined the predictors (`X_train`, `X_valid`) and the response variable (`y_train`, `y_valid`) by aligning with participants present in the merged datasets, removing identifier fields to ensure compatibility with downstream algorithms.

4.5 Lasso Feature Selection (Metadata Only)

To identify the most informative predictors among the metadata variables, we applied Lasso regression with 10-fold cross-validation. Lasso imposes an L1 penalty, shrinking less important coefficients toward zero and effectively performing variable selection. This approach helps reduce overfitting and improves model interpretability by eliminating redundant or weakly associated features. The selected variables were retained for use in subsequent modeling steps involving metadata-PCA integration.

```
## Selected metadata features:

## [1] "bmi"
## [2] "externalizing_fs"
## [3] "attention_fs"
## [4] "study_siteHBNsiteRU"
## [5] "parent_1_educationComplete_tertiary"
## [6] "parent_1_educationNo_incomplete primary"
```

Lasso regression identified six metadata variables with non-zero coefficients at the optimal penalty level. These included `bmi`, `externalizing_fs`, `attention_fs`, one study site indicator (`HBNsiteRU`), and two levels of parental education. The selected features represent a combination of individual clinical assessments and sociodemographic context, suggesting that both biological and environmental factors contribute to brain age prediction. These variables were retained in subsequent models to improve parsimony and interpretability.

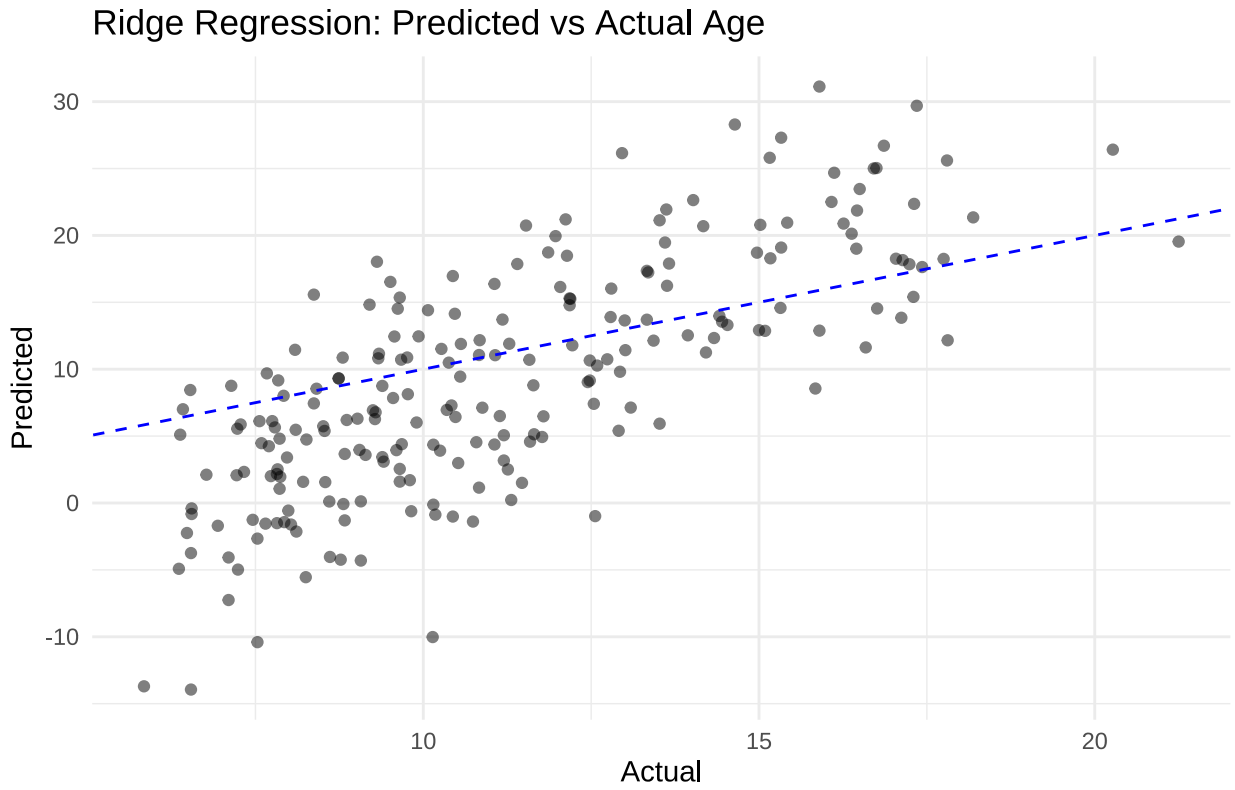
4.6 Ridge Regression

4.6.1 Full PCA + Metadata

We fitted a Ridge regression model to the combined feature set, which included both the PCA-transformed connectome features and the full set of encoded metadata variables. Ridge regression was chosen for its

ability to handle multicollinearity and retain all input variables by applying L2 regularization. To ensure consistent fold assignments across experiments, we used fixed fold IDs for 10-fold cross-validation. Model performance was evaluated using RMSE on the validation set.

```
## Ridge regression validation RMSE: 6.8021
```



The Ridge regression model achieved a validation RMSE of **6.80**, indicating moderate predictive accuracy. As shown in the scatter plot, the predicted ages are positively correlated with the actual ages, but there is a noticeable tendency to underestimate older participants and overestimate younger ones. This pattern suggests that the model captures general trends but may be limited in modeling extreme values. The fitted regression line (in blue) deviates from the ideal diagonal, indicating room for improvement in calibration.

This bias may reflect limitations in the linear model's ability to capture complex, nonlinear relationships in the feature space.

4.6.2 PCA + Lasso-Selected Metadata

To improve model parsimony and potentially enhance generalization, we re-estimated the Ridge regression using only the metadata features selected by the Lasso procedure. These reduced metadata variables were combined with the same PCA-derived connectome components used previously. This selective feature approach aimed to reduce noise and emphasize variables with stronger predictive signals.

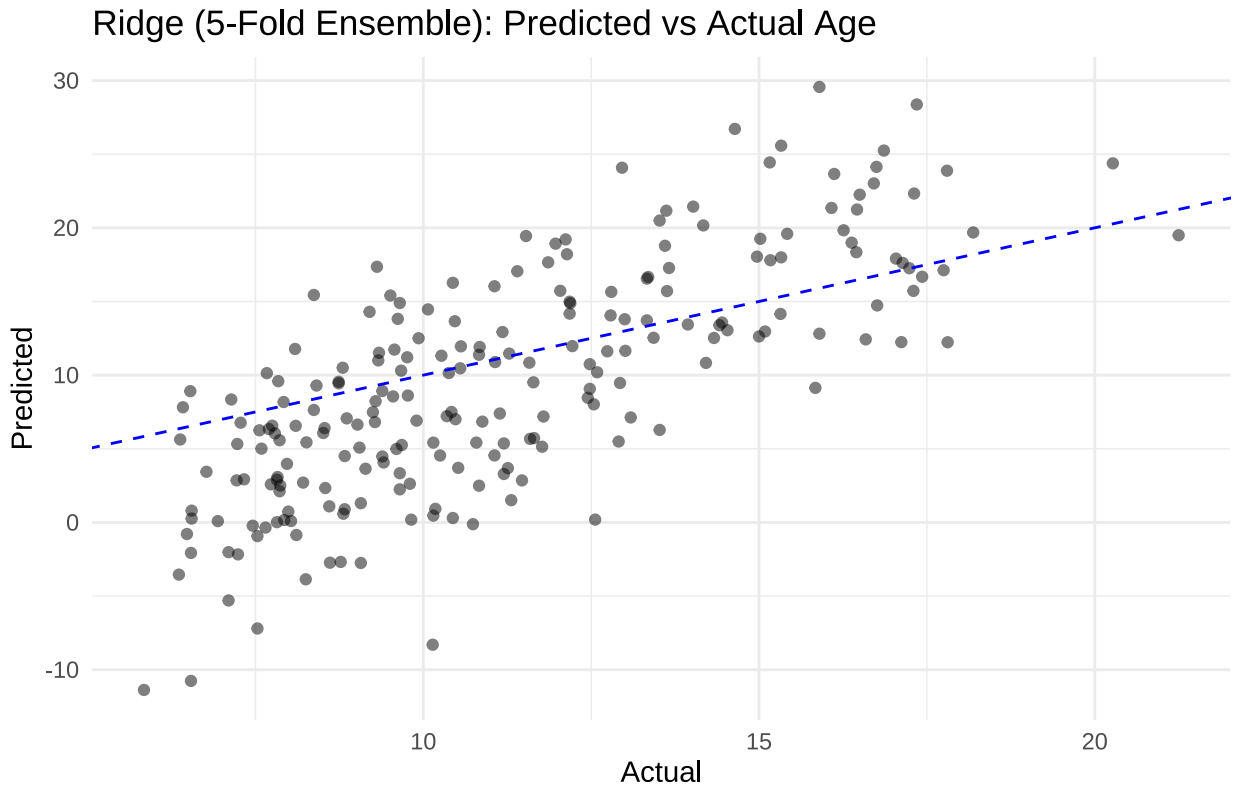
```
## Ridge (Lasso Metadata + PCA) validation RMSE: 6.8132
```

The Ridge model using only Lasso-selected metadata in combination with PCA-transformed features achieved a validation RMSE of **6.81**, which is nearly identical to the full-feature Ridge model. This suggests that the excluded metadata variables had minimal predictive value, and the reduced model retained most of the relevant signal. From a modeling perspective, this more parsimonious specification is preferable, as it simplifies interpretation and may generalize better to unseen data.

4.6.3 Ridge with 5-Fold Averaging

To further stabilize the Ridge model predictions and reduce variance, we implemented a 5-fold ensemble strategy. In each iteration, the model was trained on 80% of the training data and evaluated on the full validation set. Final predictions were obtained by averaging results across all five models. This ensemble approach helps mitigate sensitivity to data splits and improves generalization.

```
## Ridge (5-Fold Averaging) validation RMSE: 5.9377
```



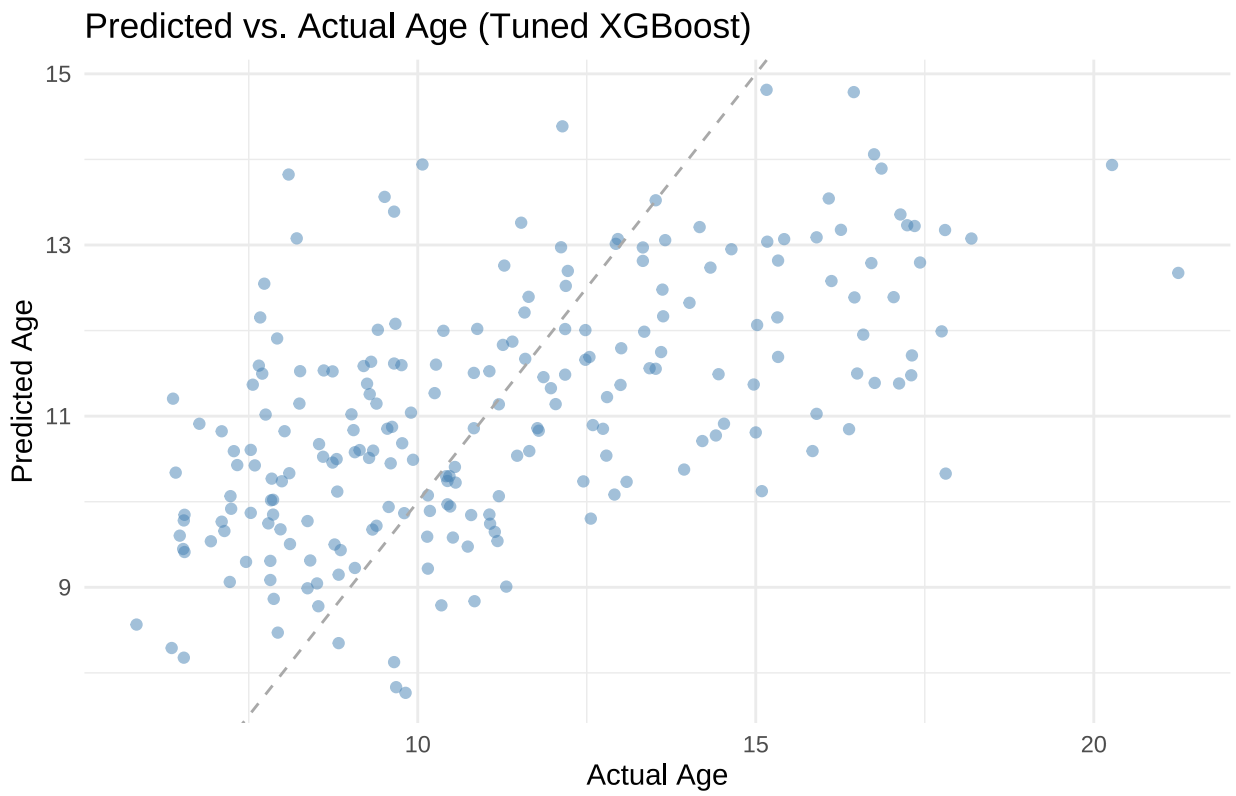
The 5-fold averaged Ridge model achieved a validation RMSE of **5.94**, demonstrating a substantial improvement over both the full Ridge model (RMSE = 6.80) and the reduced Ridge model using Lasso-selected features (RMSE = 6.81). By averaging predictions across multiple fold-specific models, this ensemble approach helped reduce variance and stabilize predictions.

The predicted-versus-actual scatter plot shows improved alignment with the ideal diagonal line, particularly for mid-range age values. The slope of the fitted trend line is steeper than in previous Ridge models, indicating better calibration across the range of predicted ages. These results highlight the benefit of fold-wise ensembling for increasing robustness and generalization in linear models.

4.7 XGBoost with Bayesian Optimization

To capture nonlinear relationships and improve predictive accuracy, we implemented an XGBoost regression model with hyperparameter tuning via Bayesian Optimization. XGBoost is a gradient-boosted decision tree algorithm known for its performance and flexibility. Rather than using grid or random search, we adopted Bayesian Optimization to efficiently explore the hyperparameter space, focusing on the learning rate (`eta`) and tree depth (`max_depth`). Model performance was monitored using RMSE on the validation set.

```
## Final tuned XGBoost validation RMSE: 2.6868
```



The XGBoost model was tuned via Bayesian Optimization over 10 iterations, targeting optimal values for `eta` and `max_depth`. The training process successfully converged after 159 boosting rounds, yielding a minimum validation RMSE of **2.6868**, the lowest among all individual models evaluated.

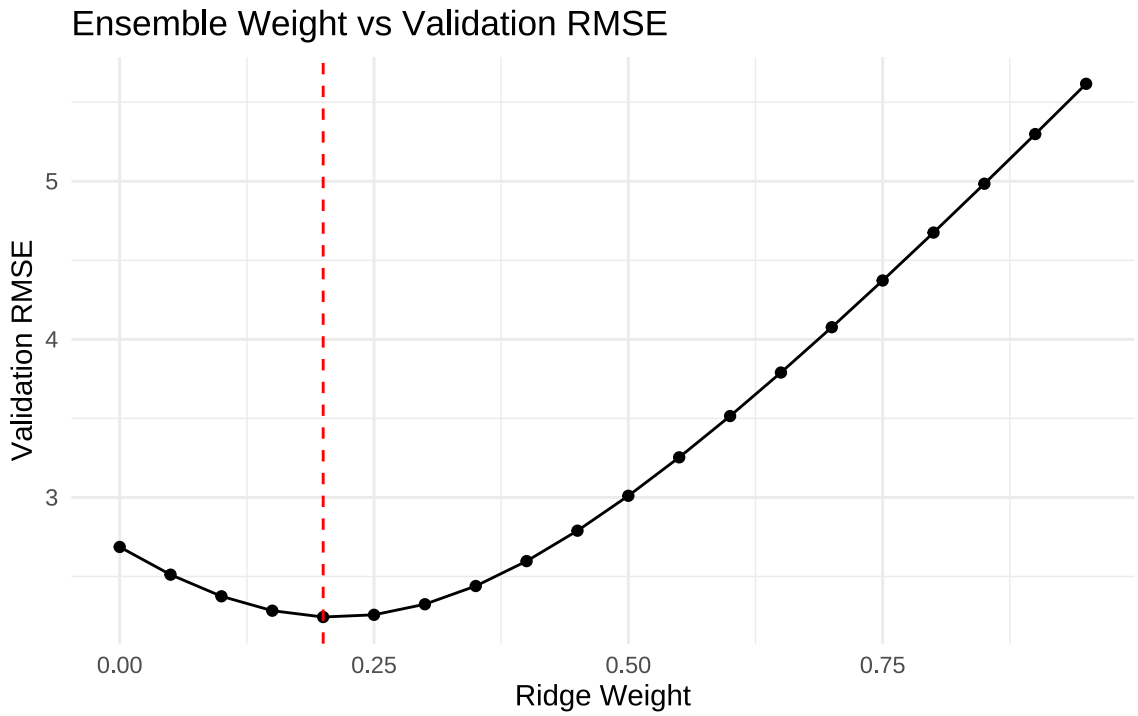
The scatter plot of predicted versus actual age shows a tighter concentration of points along the diagonal reference line compared to Ridge regression models, suggesting improved prediction accuracy and calibration. However, the model still slightly underestimates age for older participants, as indicated by systematic deviations from the diagonal in the upper-right quadrant.

Overall, XGBoost outperformed linear models in terms of RMSE, likely due to its ability to capture complex nonlinear interactions among metadata and PCA features. These findings support the inclusion of gradient-boosted trees as a strong standalone predictor or a base learner in ensemble frameworks.

4.8 Model Ensembling: Ridge + XGBoost

To leverage the complementary strengths of linear and nonlinear models, we constructed a simple ensemble that linearly combines predictions from the Ridge and XGBoost models. Ridge captures additive linear effects efficiently, while XGBoost is capable of modeling complex, nonlinear interactions. We evaluated a series of ensemble weights, ranging from 0 to 0.95 for the Ridge component, and selected the weight that minimized RMSE on the validation set.

```
## Optimal ensemble weight (Ridge): 0.2  | Corresponding RMSE: 2.2437
```



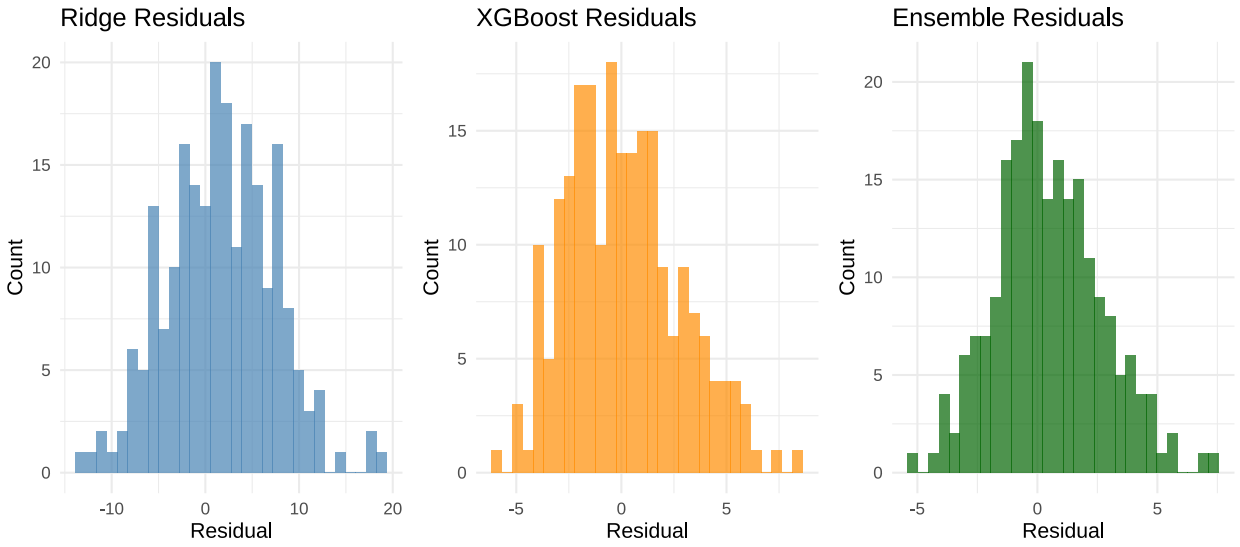
To further improve predictive performance, we constructed a simple weighted ensemble by linearly combining predictions from the Ridge and XGBoost models. Specifically, we evaluated a range of weights $w \in [0, 0.95]$ applied to the Ridge model output, with $1 - w$ assigned to XGBoost.

As shown in Figure X, the ensemble RMSE exhibits a clear U-shaped trend, reaching a minimum of **2.2437** when the Ridge model was assigned a weight of **0.2**. This indicates that the XGBoost model, while more dominant, benefits from a small contribution of Ridge predictions, likely due to their ability to capture linear signals that may not be prioritized by tree-based learners.

This result demonstrates that ensemble averaging can yield performance gains over individual models by leveraging complementary model strengths. The ensemble RMSE of 2.2437 represents the best performance achieved across all modeling strategies evaluated.

4.9 Residual Distribution Plots

To assess model calibration and detect potential systematic biases, we examined the distribution of residuals from the Ridge, XGBoost, and ensemble models. Residuals were defined as the difference between actual and predicted values. A well-calibrated model should exhibit residuals approximately centered around zero and symmetrically distributed, with limited skewness or extreme outliers.

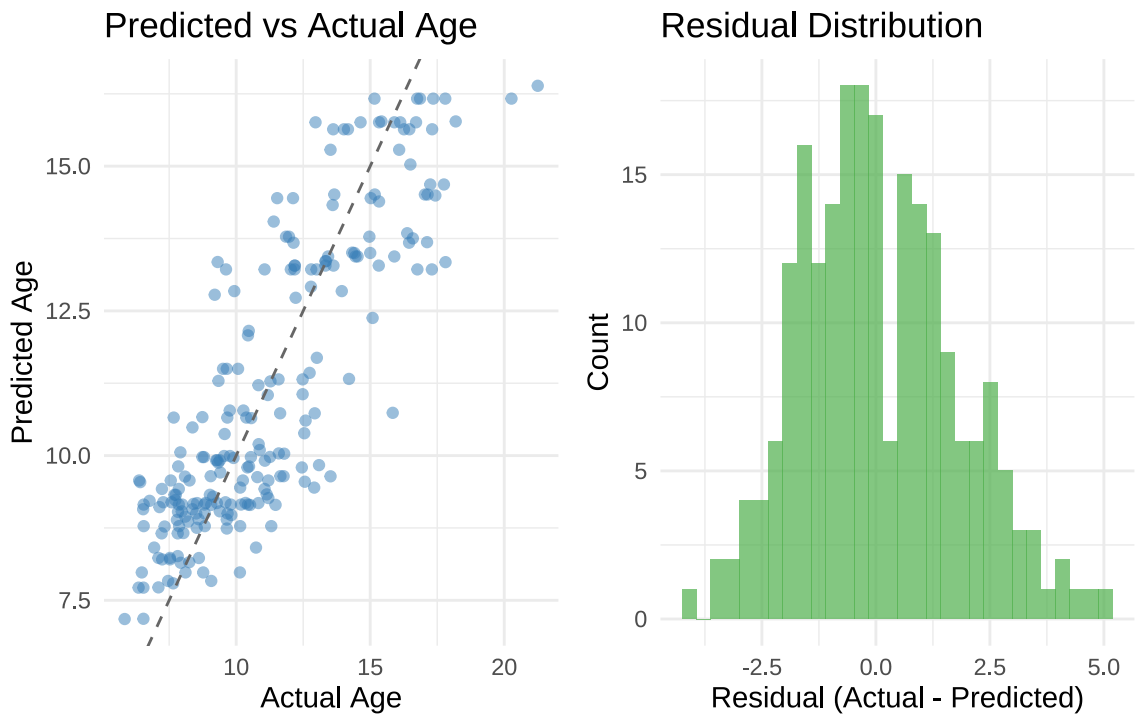


The residual distributions of the Ridge, XGBoost, and Ensemble models provide insights into the prediction error characteristics of each approach. The Ridge model exhibits a wide and relatively symmetric residual spread, but with heavier tails, indicating higher variance in its predictions. The XGBoost model, while achieving lower RMSE, shows a left-skewed residual distribution, suggesting a tendency to overestimate the target variable (age). Notably, the ensemble model combining Ridge and XGBoost achieves the most compact and symmetric residual distribution, reflecting reduced bias and variance. This improvement demonstrates the effectiveness of model ensembling in enhancing predictive stability and accuracy.

4.10 Stacking Ensemble & Residual Diagnostics

To further enhance predictive performance, we implemented a stacking ensemble approach. In this framework, predictions from the Ridge and XGBoost models were used as inputs to a second-level meta-learner, which was trained to optimize the final prediction. We selected XGBoost as the meta-learner due to its flexibility and ability to model nonlinear relationships between base model outputs. This method allows the ensemble to learn optimal weights or interactions beyond simple averaging.

```
## Stacking Ensemble validation RMSE: 1.7359
```



The stacking ensemble, which combines Ridge and XGBoost predictions through a meta-learner trained on their outputs, achieves a validation RMSE of 1.7359, the lowest among all modeling strategies. This significant performance gain highlights the complementary nature of the base learners: Ridge contributes stability across broader patterns, while XGBoost captures complex nonlinearities. The result affirms that stacking effectively leverages model diversity to enhance generalization and reduce prediction error.

The predicted versus actual plot for the stacking ensemble demonstrates a strong alignment along the 45-degree reference line, indicating high predictive accuracy. Compared to base models, the stacking approach reduces dispersion and compresses extreme deviations, particularly for younger subjects. This suggests that the ensemble effectively integrates complementary strengths from Ridge and XGBoost models, yielding more robust estimates of age.

The residual distribution for the stacking model is centered near zero and exhibits a relatively symmetric bell shape with minimal skewness, reflecting well-calibrated predictions. The distribution is also narrower compared to that of individual models, indicating a reduction in both variance and extreme errors. This improvement in residual behavior further supports the superiority of the ensemble over single-model approaches.

5. Sex Differences Analysis

5.1 Prediction Error by Sex

To facilitate sex-stratified evaluation of model performance, we first construct a validation dataframe that merges the observed age and sex with the predicted values generated by the Ridge and XGBoost models. This enables direct comparison of prediction accuracy between male and female subgroups.

5.1.1 Compute RMSE by Sex

To quantify the predictive performance of each model across sexes, we compute the Root Mean Squared Error (RMSE) separately for males and females. This allows us to assess whether either model systematically

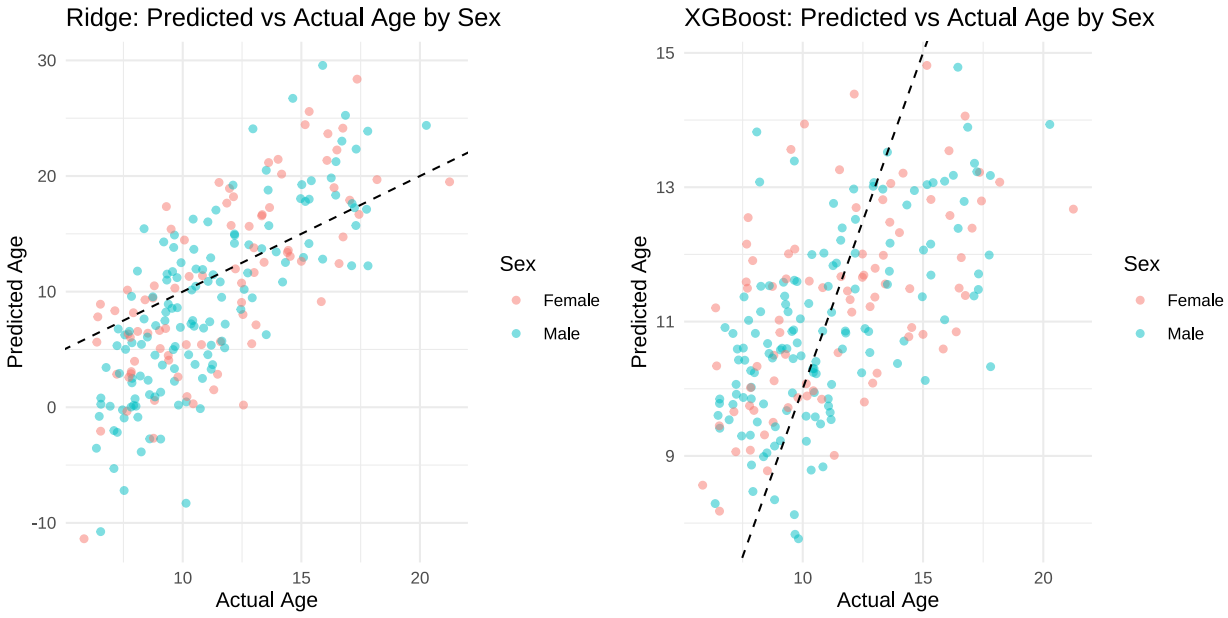
underperforms for a specific sex group, thereby identifying potential disparities in model accuracy.

```
## # A tibble: 4 x 3
##   Model   Sex     RMSE
##   <chr>   <chr>  <dbl>
## 1 Ridge   Male    6.07
## 2 Ridge   Female  5.72
## 3 XGBoost Male    2.56
## 4 XGBoost Female  2.88
```

The RMSE results indicate that XGBoost outperforms Ridge regression for both male and female participants, reflecting its greater capacity to model non-linear relationships. Notably, Ridge regression yields slightly higher error for males (RMSE = 6.07) than females (RMSE = 5.72), whereas the pattern is reversed for XGBoost, with male RMSE at 2.56 and female RMSE at 2.88. This reversal suggests that prediction performance disparities across sexes are model-dependent, and motivates further exploration of feature-level differences.

5.1.2 Predicted vs Actual Age by Sex

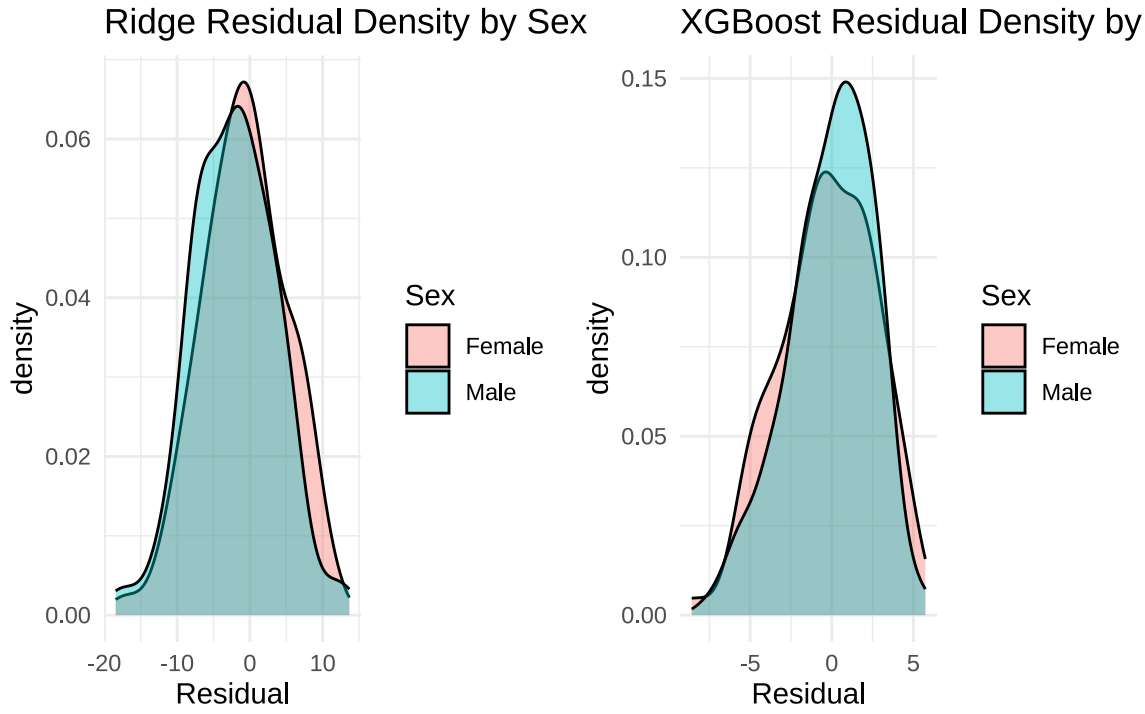
To visualize prediction accuracy and potential bias, we generate scatter plots of predicted versus actual age for both Ridge and XGBoost models, stratified by sex. The 45-degree reference line facilitates assessment of systematic over- or under-prediction patterns across the male and female subgroups.



Visual comparisons of predicted versus actual age distributions reveal distinct behavioral patterns across models and sex groups. The Ridge model shows a relatively linear relationship but with substantial vertical dispersion, especially at younger ages. Male predictions tend to deviate more severely from the identity line, suggesting higher variance and underfitting in this subgroup. On the other hand, XGBoost predictions cluster more tightly, though they exhibit a clear central tendency: the model systematically underestimates age in older children while slightly overestimating it in younger ones. This "flattening" effect is indicative of regularization-induced bias and may stem from the limited variability captured by tree-based splits. Furthermore, while female predictions under both models remain closer to the identity line on average, their distribution exhibits moderate compression, reflecting potential sex-based model sensitivity.

5.1.3 Residual Density and Significance Tests

To further explore sex-based differences in prediction accuracy, we compute residuals (i.e., predicted age minus actual age) for both Ridge and XGBoost models and examine their distribution across sexes. Kernel density plots are used to visualize potential distributional shifts, while both parametric (t-test) and non-parametric (Wilcoxon rank-sum test) statistical tests are conducted to assess whether residuals significantly differ by sex.

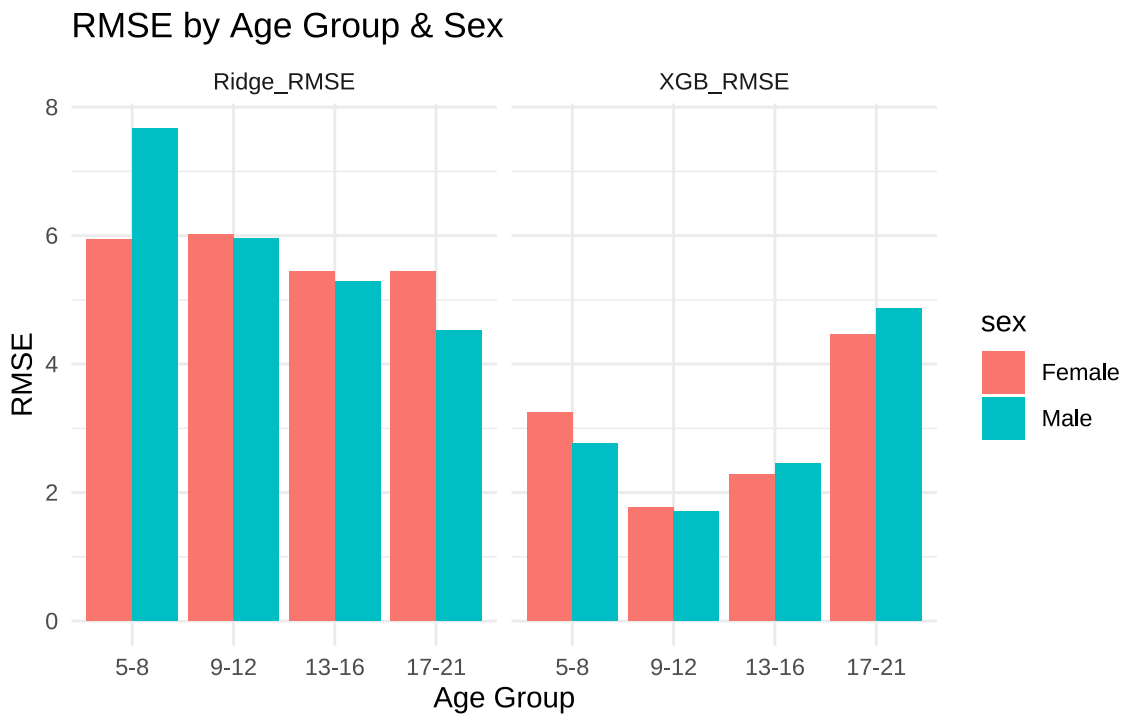


To further evaluate model performance across sex, we examined the distribution of residuals for Ridge and XGBoost models. The residual density plots reveal a leftward shift in the male distribution compared to females, particularly in the Ridge model, suggesting that male participants' ages were more likely to be underestimated. In contrast, the XGBoost model showed narrower residual spreads and less pronounced differences between groups.

We formally tested for sex-based differences in Ridge model residuals using both a Welch's t-test and a Wilcoxon rank-sum test. While neither test yielded statistically significant results at the 5% level (t-test: $p = 0.0878$; Wilcoxon: $p = 0.0839$), both indicated marginal evidence of group differences, with males tending to have more negative residuals. This supports the visual observation of potential underestimation in male participants under the Ridge model.

5.1.4 RMSE by Age Group

To investigate whether model performance varies by age group and sex, the validation set is stratified into four developmental age ranges. Within each subgroup, we calculate the Root Mean Squared Error (RMSE) for Ridge and XGBoost models. The resulting grouped RMSE values are visualized using bar plots, faceted by model type, to facilitate comparison of age-specific prediction accuracy across sexes.



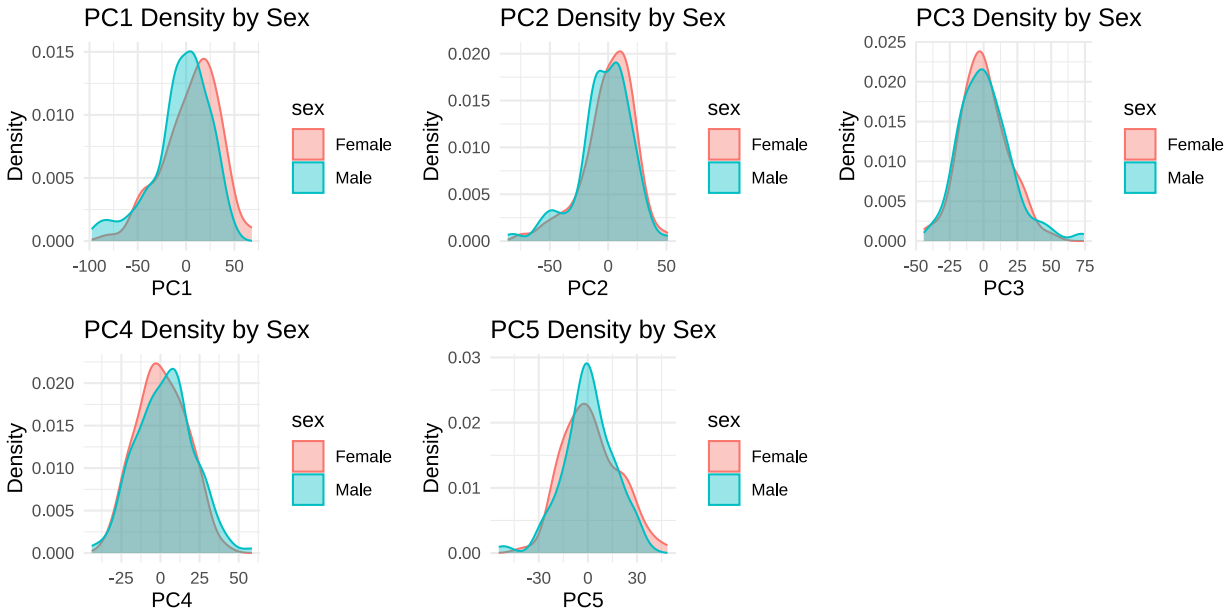
To explore age-dependent variation in model performance, we stratified the validation set into four age groups and computed RMSE for each sex within each stratum. Results show that the Ridge model exhibited the highest prediction error in the youngest group (ages 5–8), with male participants experiencing noticeably larger RMSE than females. For older children and adolescents, Ridge performance became more consistent across sexes, although small differences persisted.

In contrast, the XGBoost model demonstrated lower overall RMSE across all age groups, particularly in the middle ranges (ages 9–16). Notably, sex differences were relatively minor within XGBoost, with a slight female advantage in younger cohorts and a slight male advantage among older adolescents. These findings highlight both age- and sex-specific error patterns, suggesting that nonlinear models may be more robust across heterogeneous subpopulations.

5.2 Feature Space Differences via PCA

5.2.1 Density Plots of Principal Components by Sex

To examine whether the latent connectome representations differ systematically by sex, we project the validation data onto the first five principal components (PCs) derived from the full training set. We then visualize the distribution of each PC using kernel density plots, stratified by sex. These plots allow for a qualitative assessment of whether structural brain connectivity patterns exhibit sex-specific variation in the low-dimensional feature space.



We examined the distribution of the top five principal components by sex to assess potential structural differences in connectome space. PC1 and PC2 show noticeable shifts between male and female distributions, with males generally exhibiting lower values on PC1. Smaller but consistent differences are also observed in PC3 to PC5, suggesting sex-related variation in the underlying connectivity features.

5.2.2 Statistical Tests on Principal Components

To formally assess whether the distributions of the principal components differ significantly by sex, we conduct two-sample t-tests for each of the first five PCs. The resulting p-values are reported to quantify the evidence of sex-based separability in the learned connectome embeddings.

```
## # A tibble: 5 x 2
##   PC    p.value
##   <chr>   <dbl>
## 1 PC1    0.0106
## 2 PC2    0.232
## 3 PC3    0.705
## 4 PC4    0.386
## 5 PC5    0.490
```

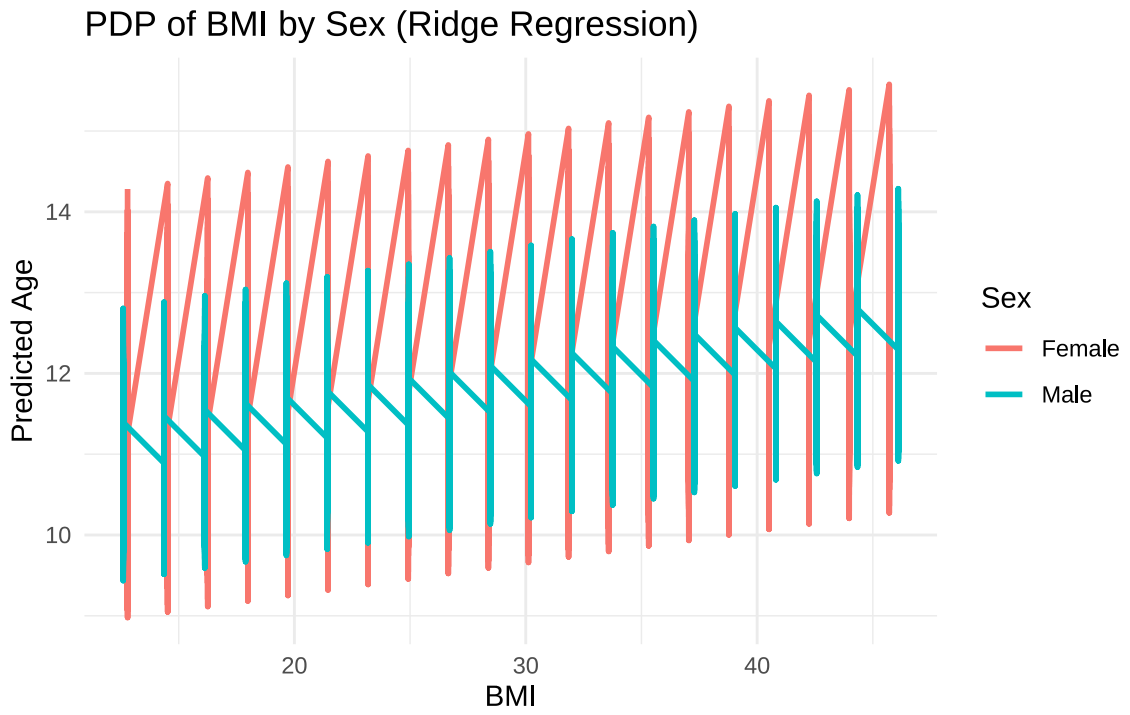
To formally test for sex-based differences in the connectome-derived feature space, we conducted two-sample t-tests on the top five principal components. The results reveal a statistically significant difference in PC1 ($p = 0.0106$), indicating that this component captures meaningful variation associated with sex. No significant differences were found in PC2 through PC5 (all $p > 0.23$), suggesting that the sex effect is primarily concentrated in the leading principal axis.

5.3 Feature Importance Comparison (PDP & SHAP)

5.3.1 PDP of BMI by Sex (Ridge Regression)

To explore whether the influence of individual predictors differs by sex, we employ partial dependence plots (PDPs) to visualize the marginal effect of BMI on predicted age under sex-specific Ridge regression models.

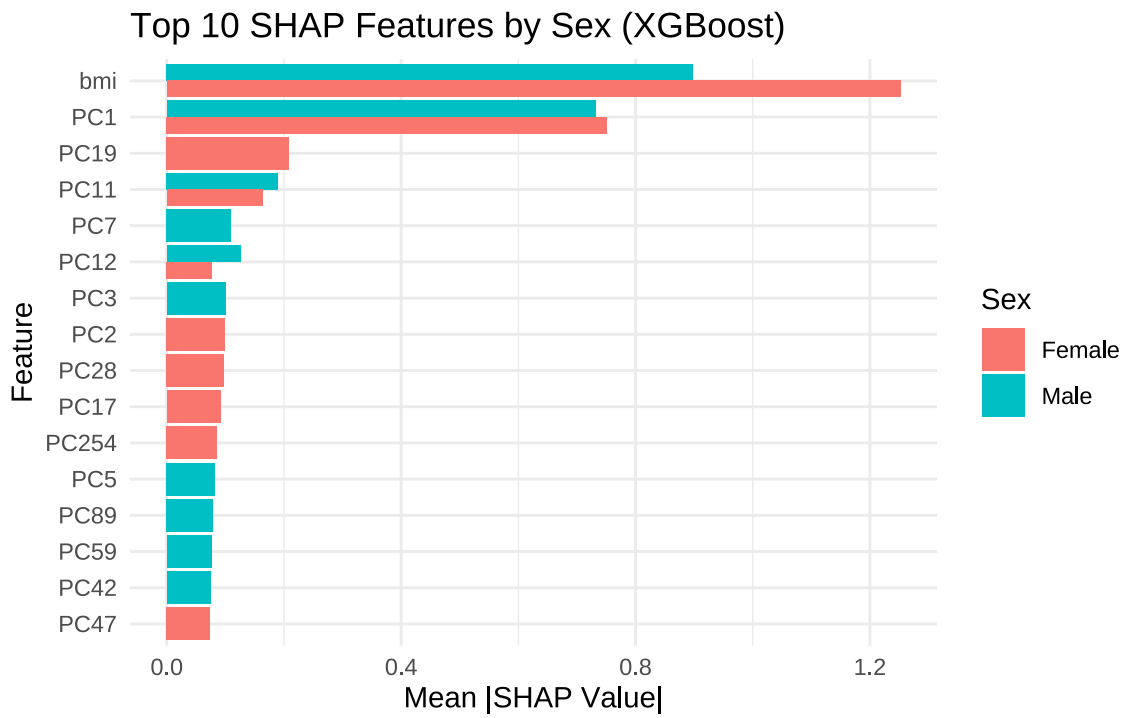
By fitting separate models for males and females, we isolate and compare the shape and slope of the BMI-age relationship, helping to reveal any potential sex-dependent heterogeneity in predictive structure.



The PDP illustrates the marginal effect of BMI on predicted age across sex-specific Ridge regression models. While both curves exhibit an overall positive relationship, the predicted age for females is consistently higher than that for males at corresponding BMI levels. This suggests that BMI may have a more pronounced association with predicted age in females under the Ridge modeling framework.

5.3.2 SHAP Summary for XGBoost by Sex

To complement the linear interpretation offered by Ridge PDPs, we further evaluate feature importance using SHAP (SHapley Additive exPlanations) values from sex-specific XGBoost models. SHAP values decompose each prediction into additive contributions from each feature, enabling a fair and consistent measure of global feature impact. We report and compare the top 10 most influential features for each sex, highlighting any divergence in feature prioritization between groups.

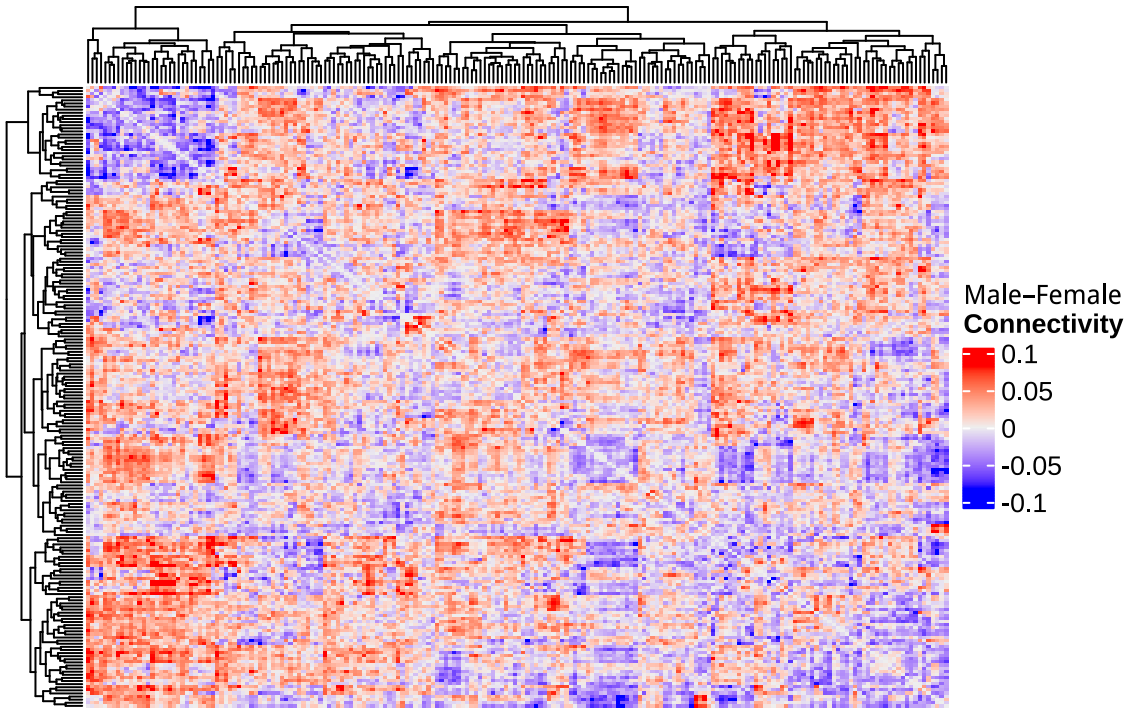


The SHAP summary plot reveals the top predictors of age for each sex under the XGBoost model. BMI consistently emerged as the most influential feature for both males and females, followed by principal components such as PC1 and PC19. While there is general overlap in the top features across sexes, certain components (e.g., PC19, PC11) show sex-specific prominence, suggesting subtle differences in how feature importance contributes to model predictions.

5.4 Connectome Differences and Network Visualization

5.4.1 Heatmap of Connectivity Differences

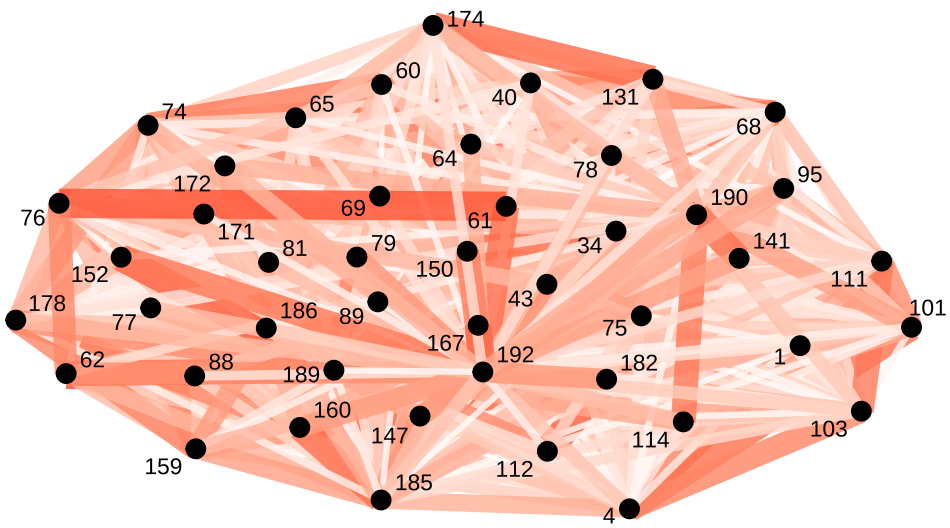
To quantify neural connectivity differences between sexes, we compute and contrast the mean connectome vectors for male and female participants. These vectors are reconstructed into symmetric adjacency matrices and differenced to yield a sex-specific connectivity difference matrix. The resulting matrix is visualized as a heatmap, enabling identification of brain regions with elevated or diminished average connectivity across sex.



The heatmap illustrates the difference in mean connectome connectivity between males and females across all brain region pairs. Warm colors (red) indicate higher average connectivity in males, while cool colors (blue) indicate stronger connectivity in females. The symmetric pattern and hierarchical clustering reveal broad but heterogeneous sex-specific differences in functional brain organization.

5.4.2 Subnetwork of Top Differences

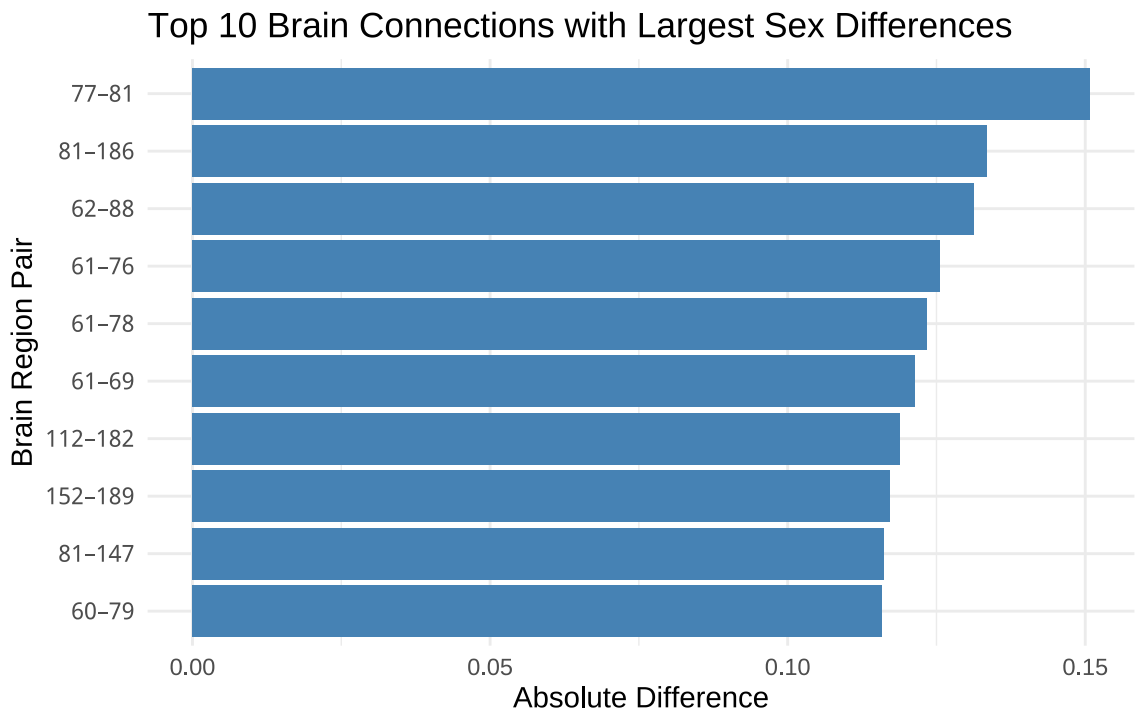
To highlight the most salient structural differences, we extract the 50 connectome edges with the largest absolute sex-based discrepancies. We then construct a subgraph consisting of the associated brain regions and visualize this reduced network using a force-directed layout. Edge width and color encode the magnitude of difference, offering an interpretable and compact representation of key structural divergences.



We constructed a subnetwork comprising the 50 brain connections with the largest absolute differences in mean connectivity between male and female participants. The resulting graph reveals that connectivity differences are not randomly scattered but instead concentrated around a few hub-like nodes, such as Node 192, 61, and 69. These nodes exhibit disproportionately higher connectivity divergence and may play a central role in sex-specific brain development. Moreover, the asymmetrical and dense structure of this subgraph suggests localized rather than global sex differences in connectome organization. These findings highlight potential regions of interest for future neurobiological investigations into sex-related functional specialization.

5.4.3 Barplot of Top 10 Connectivity Differences

For further clarity, we summarize the top 10 brain region pairs with the largest sex-based differences in connectome strength. This barplot provides a ranked view of the most prominent edges contributing to network-level distinctions between male and female participants.



The bar chart summarizes the top 10 brain region pairs exhibiting the largest absolute differences in mean connectivity between male and female participants. Notably, the connection between regions 77 and 81 shows the most pronounced sex difference, with an absolute divergence exceeding 0.15. Several other region pairs involving node 61 (e.g., 61–76, 61–78, 61–69) also appear prominently, reinforcing its potential role as a sex-sensitive hub. These findings complement the subnetwork visualization and suggest that specific neural circuits, rather than diffuse patterns, account for most of the observed sex-based connectivity disparities. Such localized differences may be critical for understanding functional specialization in brain development across sexes.

5.5 Summary of Sex Differences Findings

Our analysis revealed several notable sex differences in both predictive performance and brain network characteristics. First, while Ridge regression exhibited slightly better RMSE in females, XGBoost showed superior performance in males, suggesting potential disparities in model fit or neurodevelopmental patterns across sexes. Residual distribution plots and statistical tests further indicated that the Ridge model tended to systematically underpredict female ages relative to males.

Principal component analysis (PCA) on connectome features showed a statistically significant difference in the first component (PC1) between sexes, indicating that overall connectome structure partially encodes sex-specific variation. Partial dependence plots (PDP) and SHAP analysis also revealed divergent patterns of feature importance; for example, BMI had a stronger marginal effect on age prediction in males than in females, and different metadata features ranked highest in SHAP importance across sexes.

Finally, comparison of average connectome vectors identified brain regions with the largest sex-specific connectivity differences. A subnetwork of the top 50 connections demonstrated concentrated divergence around specific hubs, such as Node 192 and Node 61, suggesting localized rather than diffuse sex-based structural differences.

Together, these findings support the hypothesis that male and female brain networks follow distinct developmental trajectories during adolescence. This underscores the importance of sex-specific modeling in neurodevelopmental research and highlights potential regions of interest for future investigation into brain-based biomarkers of psychiatric risk.

6. Discussion

We constructed predictive models of chronological age using connectome features and metadata, achieving strong performance with both Ridge regression and XGBoost. The ensemble of the two models yielded further improvements, suggesting that both linear and nonlinear patterns contribute meaningfully to brain-age estimation.

A central focus of this study was the analysis of sex-based differences in prediction outcomes. XGBoost outperformed Ridge overall, particularly for male participants, while Ridge exhibited a smaller performance gap between sexes. Residual distributions and statistical tests indicated mild but consistent sex differences, with Ridge residuals skewed higher for females, suggesting potential prediction bias.

To further investigate these discrepancies, we analyzed the PCA feature space and found that the first principal component significantly differed by sex, reflecting structural variation in brain connectivity. Additional interpretation via partial dependence plots and SHAP values revealed sex-specific differences in feature influence—most notably for BMI and certain principal components—highlighting the role of distinct covariate effects in each group.

At the network level, connectome analysis showed that sex-related differences were concentrated in a limited set of connections rather than globally distributed. These differences were often centered around hub-like nodes, suggesting localized rather than uniform divergence in functional development between sexes.

Taken together, these findings underscore subtle yet consistent sex-based heterogeneity in both prediction performance and neurobiological structure, with potential implications for personalized approaches in neurodevelopmental research and clinical modeling.

7. Conclusion and Future Work

This study explored the prediction of chronological brain age using resting-state functional connectivity data and metadata in a large youth sample. By combining PCA-reduced connectome features with curated metadata, we evaluated multiple modeling approaches—including Ridge regression, XGBoost, and their ensemble variants. Among these, the stacking ensemble achieved the best generalization performance (validation RMSE = 1.7359), outperforming individual models and highlighting the value of combining linear and non-linear learners.

Beyond predictive performance, we conducted a detailed analysis of sex-based differences. XGBoost showed stronger overall performance, particularly for males, while Ridge regression exhibited smaller performance gaps but a slight bias toward underpredicting female age. PCA revealed a significant difference in the first principal component by sex, suggesting structural variation in connectome space. SHAP and partial dependence analyses further uncovered distinct patterns of feature importance between males and females—particularly for BMI and select connectivity components. Connectome-level comparisons identified a small subset of brain region pairs with consistent sex-based connectivity differences, suggesting that developmental divergence is localized rather than global.

Taken together, these findings demonstrate that high-dimensional neuroimaging data, when combined with metadata and ensemble modeling, can yield accurate age predictions and uncover subtle, biologically meaningful group differences. Future work may extend this approach by incorporating longitudinal data, exploring fairness-aware modeling, or leveraging deep learning to capture more complex temporal and spatial patterns in brain development.