**1 Methods and Materials**

**1.1 Participants**

Structural MRI was collected from ### children recruited to participate in a three-wave longitudinal study. (Demographics, covariate distributions)

**1.2 Procedure**

(recruitment)

**1.3 Questionnaires**

*Child Behavior Checklist*.

We administered the Childhood Behavior Checklist (CBCL) to assess participants’ problem behaviors…

**1.4 Imaging**

*Structural MRI Acquisition*.

As described previously in [NAME of SB AUTHORS] et al. (YYYY), subjects were scanned with a XXXX…

*Preprocessing and Subcortical Segmentation*.

We obtained measurements of subcortical grey matter volume (SGMV) for subcortical regions using FreeSurfer’s automated parcellation algorithm and labels based on the Desikan-Killiany Atlas.

**1.5 Statistical Analysis**

*Gradient Boosting Machines*

Gradient boosting refers to the process of iteratively modeling a response variable by sequentially adding simple models (e.g., shallow decision trees) to an ensemble that cumulatively generate a final prediction. Each new model predicts the residuals of the existing in order to further minimize a loss function (e.g., misclassification rate or mean-squared error). At every iteration, a new model is added that maximally decreases the prediction error from the previous iteration. As a result, the observations (e.g., the subjects) associated with the largest residuals will be most influential in the fitting of each subsequent model. This amounts to most efficiently minimizing the loss function by iteratively progressing along its gradient. Gradient Boosted Machines (GBMs; Friedman, 2001) can be therefore generalized to predict a variety of categorical or continuous response types, minimize any differentiable loss function, and often perform well in prediction and classification problems despite noisy and heterogeneous data.

*Classification of Previously Institutionalized Group Status*

To identify robust differences between the previously institutionalized (PI) and comparison (COMP) groups within the domain of SGMV, we trained, tuned, and validated 100 GBM classifiers on 100 novel resamples of the data. Since our goal is to build a model that is sensitive and specific in its identification of PI group status, each resample began by randomly assigning 70% of the PI group to the training dataset (train). To control for the effect that age and group sample size imbalance might have on model performance, we added an equal number of randomly selected, age-matched COMPs to train (Snoek, Miletić, & Scholte, 2019). The remaining 30% of PI subjects and approximately 45% of COMP subjects were assigned to the testing dataset (test). Finally, we generated a tuning dataset (tune) by randomly extracting 40% of the subjects in train. To maintain independence between train, tune, and test, all resamples were generated by subject/participant ID (Poldrack, Huckin, & Varoquaux, 2020), so subjects with repeat scans would either have both scans in the same dataset, or one scan in a single dataset and the other scan dropped. Finally, to control for the confounding effect of total Intra-Cranial Volume (ICV), we built a regression model to predict all SGMV features on ICV using train (Snoek, Miletić, & Scholte, 2019). The regression model parameters were used to adjust SGMV features in tune and test to avoid data leakage (Poldrack, Huckins, & Varoquaux, 2020).

For each of the 100 resamples, we trained a GBM to classify group membership using the ‘gbm’ function in the R package ‘gbm’ (Greenwell et al., 2019). The models contained an ensemble of gradient boosted decision trees and predicted the probability of membership in the PI group using a logit link function. Optimal hyperparameter values were assessed and selected using a grid search. We investigated the hyperparameters that determine the minimum number of subjects allowed in each tree’s terminal nodes, the weight applied to the residuals from each new tree, and the number of splits permitted by each tree, denoted *n.minobsinnode*, *shrinkage,* and *interaction.depth* by the ‘gbm’ authors respectively. For each combination of hyperparameter values, we fit a candidate model with train data and calculated the area under the receiver operating characteristic curve (AUC) for its classifications on tune. We retained the final model, and hyperparameters, that maximized this metric. Following model selection, we obtained the predictions of each final model for its respective test set. We computed this final vector of predicted probabilities by averaging the predictions made for each subject, when each subject occupied a given test set. Therefore, the cross-validation estimate of final model performance is the AUC computed from these aggregated predictions with respect to the true group labels.

To assess the statistical significance of the out-of-sample predictions, we conducted permutation testing. To do so, we independently randomized the response vectors of the train, tune, and test sets 1000 times. We then fit models on train to classify these random group labels from the corresponding set of SGMV features. As in the case of the true model, we aggregated the predictions from these null models across the 100 resamples to obtain an average null distribution of 1000 prediction vectors. With these vectors we computed a distribution of 1000 null AUCs. Finally we compute and report the permutation p-value of the cross-validated AUC of the true model as the proportion of null AUCs greater than the true value.

*Prediction of CBCL Scores Using Gradient Boosting Machines*.

To assess the relationship between SGMV and our CBCL subscales of interest, we modified our GBM pipeline, described above, to separately regress each of the CBCL responses on the SGMV features. GBMs predicting CBCL scores contained boosted regression trees, and minimized mean-squared error. We generated a balanced training dataset (train) by stratifying over 4 quantiles of the CBCL response vector within each generated our initial train, tune, and test sets by

Replicating the resampling procedure above three times (once for each CBCL response), we generated 100 resamples with which to train, tune, and test gradient boosting regressors of the CBCL *t*-scores of interest. Within each resample, all three models’ hyperparameters and prediction errors were evaluated using the root mean squared error (RMSE. For each outcome, we generated a final vector of predictions by aggregating the test set predictions to each subject when said subjects occupied the test set. With these predictions, we could then estimate the RMSE in prediction of each CBCL outcome by the SGMV predictors.

To assess statistical significance of the three estimates of predictive performance, we drew 10,000 bootstrap resamples of the data and recomputed the RMSE from the vector of aggregated test set predictions, for the subjects sampled into the bootstrap each time. We then generated another vector of aggregate test set predictions by replacing the test set predictions generated by each model with the mean response for the subjects in that given train set. Using this vector, we could estimate the cross-validated prediction RMSE from an ensemble of intercept-only models. We similarly formulated a bootstrap confidence interval around this prediction error, and compared its limits to that of the modeled error to determine if the models significantly reduced prediction error relative to what would be expected from a model without any SGMV predictors.

*Computation of Cross-Validated Permutation Variable Importance*

We obtained each SGMV feature’s cross-validated permutation variable importance (CVPVI) and their p-values using the R package ‘vip’ (**CITATION**). We compute the variable importance by measuring its contribution to the AUC in prediction of the test data (Janitza et al., 2015); specifically, the variable importance is the percentage that the model improves prediction, in terms of AUC on test data, that results from adding the given variable to the model. Furthermore, we compute the statistical significance of each variable’s CVPVI by comparing them to their null distribution, obtained by repeating the above measurement of each variable’s CVPVI for each of the null model (Altmann, 2013).

For each variable, to obtain its CVPVI in the classification model we permute each variable at a time, regenerate the predictions of the trained model for the test subjects, and recalculate the prediction AUC. The true AUC from the true test set is divided by that resulting from the perturbed feature set. To avoid introducing noise into our estimates of each CVPVI due to the randomness of permutation, we repeated the permutation step 100 times before deriving the null prediction AUCs. This is conducted within each of the 100 cross-validation iterations, as well as for each of the corresponding distribution of null models. The final vector of CVPVI for the *p* variables given variable is the average of the 100 CVPVI vectors obtained from each true model, and the null distribution for each variable is obtained by averaging the 100 *p* x 1000 matrices of null CVPVI containing the 1000 null CVPVIs for each variable. Therefore, the CVPVI for each variable *i* in the set of SGMV predictors is defined,

where *R* represents the number of resamples, *p* refers to the number of permutations of each variable, *AUCR* is the prediction AUC for the model trained and validated with the *R*-th (cross-validation or null) resample, and represents the prediction AUC following the *p*-th permutation of the *i*-th variable. The resultant null CVPVI is expected to be 1 and the percent that a CVPVI is greater than 1 reflects the percent improvement in AUC attributable to that variable when added to the null model.

In the case of regression, the variable importance reflects the degree to which a variable contributes to reducing prediction error. Since the prediction error, in this case RMSE, exists on an arbitrary scale, we computed each variable’s importance to reflect the percent decrease in prediction error attributable to the variable. Similarly to the PVI computed for the classification models, a given variable’s PVI is derived from aggregating across the 100 resamples. We therefore define the PVI in the regression model for variable *i*,

where *R* represents the number of resamples, *p* refers to the number of permutations of each variable, *RMSER* is the prediction RMSE for the model trained and validated with the *R*-th resample, and represents the prediction RMSE following the *p*-th permutation of the *i*-th variable.

**2 Results**

**2.1 Prediction of Groups and Model Assessment**

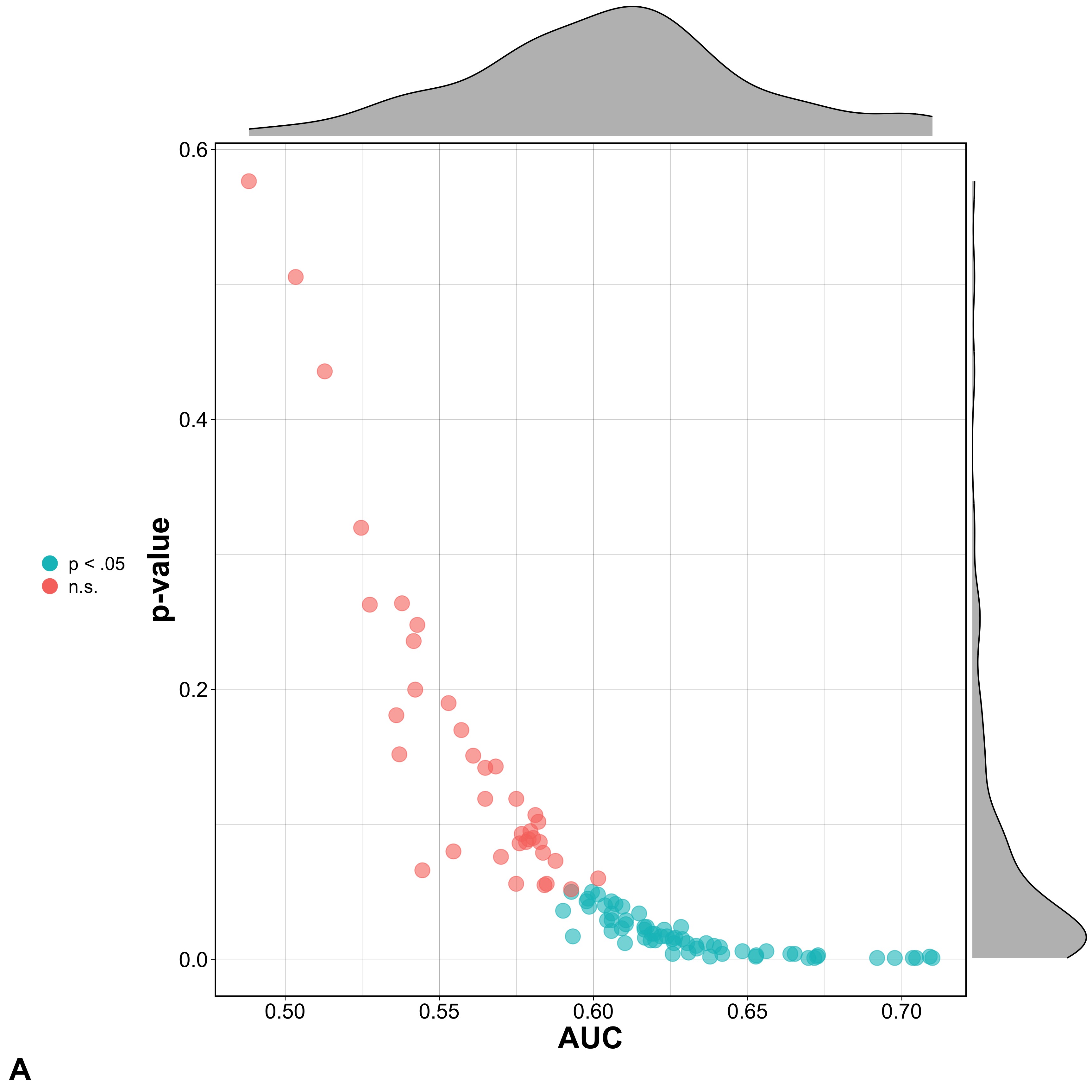
Figure 1a shows the distribution of prediction AUCs and their permutation *p*-values within each of the 100 cross-validation splits. The resultant split-specific Receiver-Operating Characteristic (ROC) curves are plotted together in Figure 1b. On average, the AUC in prediction of PI group membership for a final model to its respective test set was 0.605 (SD=0.050), while the largest was AUC 0.710 and the lowest 0.488. At an alpha of .05, 63% of these models performed significantly better than their permuted null, while the average *p* value was .070 (SD=0.102).

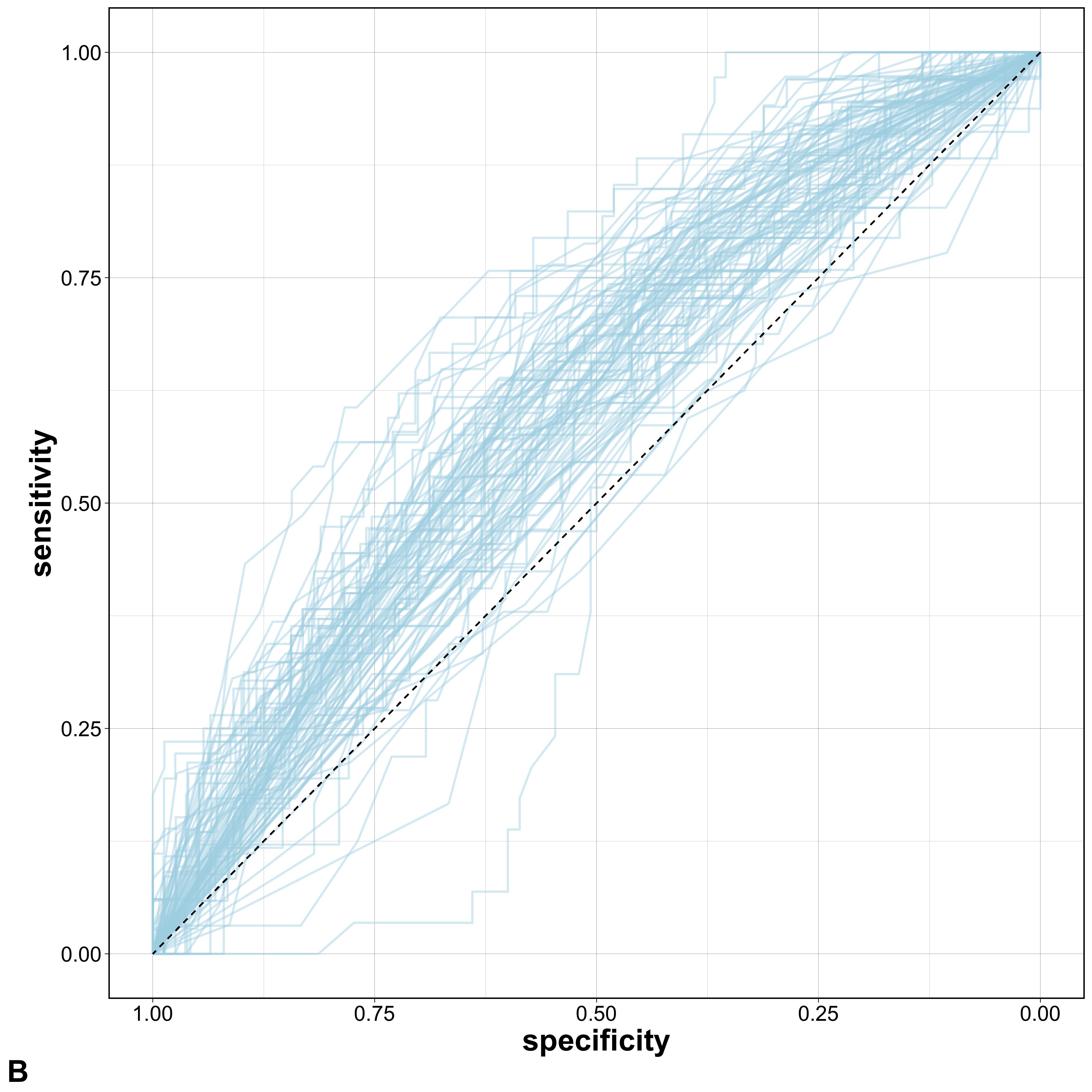
To test if SGMV could robustly and significantly predict the probability of PI group membership, we conducted a permutation test for the AUC derived from the vector of test-set probability predictions averaged across the 100 resamples. As depicted in figure 2, these probabilities proved a highly accurate predictor of true group membership, AUC=0.680 [0.595, 0.766]. The permutation test showed that this observed value was greater than all 10,000 null values, *p* < .001, providing conclusive evidence that our model decodes robust differences between the PI and COMP groups in terms of SGMV.

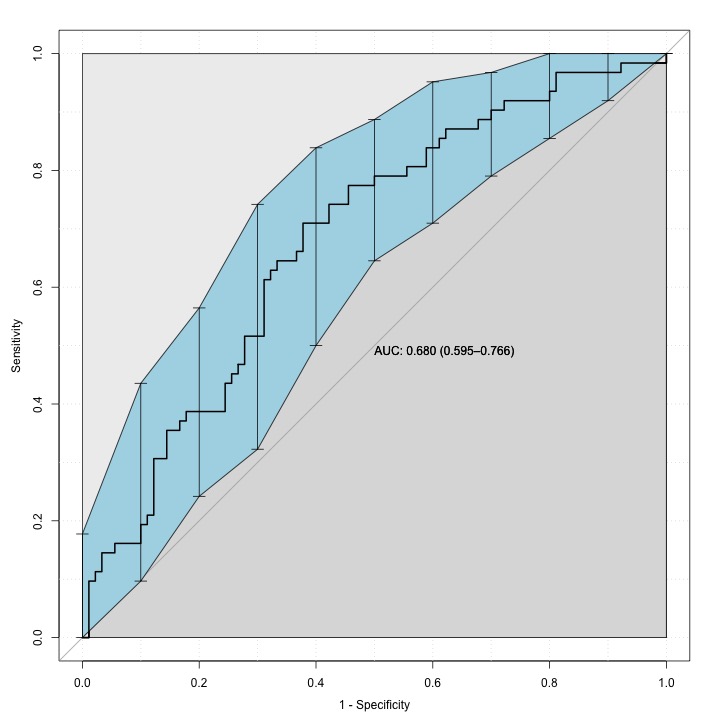
The results of the GBM regression showed that SGMV was not a reliably strong predictor of any of the CBCL outcomes. The RMSE and 95% bootstrap confidence intervals pertaining to the averaged test-set prediction of the internalizing, externalizing, and total CBCL *t*-score outcomes were 11.71 [10.76, 12.65], 11.96 [10.83, 13.09], and 12.92 [11.84, 14.00], respectively. The corresponding average intercept-only predictions showed slightly, though not significantly, larger prediction errors for the three scales: RMSEinternalizing=11.96 (95% CI [11.02, 12.90]), RMSEexternalizing=12.15 (95% CI [11.01, 13.28]), and RMSEtotal=13.20 (95% CI [12.10, 14.29]).

**2.2 Assessment of Variable Importance**

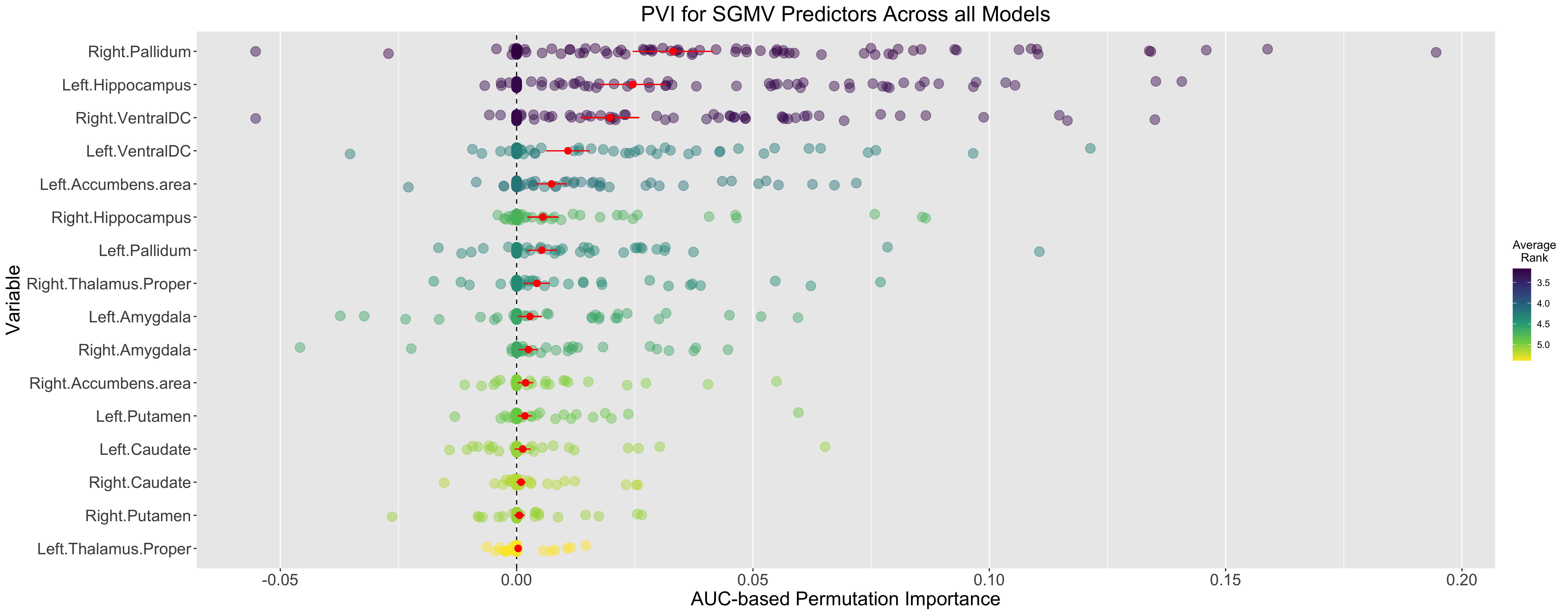
The variables most important for decoding group status are displayed in Figure 3. The three top-ranked variables in terms of average magnitude of their PVIs include the right pallidum, left hippocampus, and right ventral diencephalon. The average PVI attributable to these variables is .033, .025, and .020, indicating that for a new dataset. Compared to the PVI associated with each other variable in a given forest, these three PVIs ranked, on average, 3rd highest. The left accumbens, left ventral diencephalon, left pallidum, and right thalamus all exhibited an average rank between 4.2 and 4.4, with PVIs of .011, .007, .006, and .004 respectively.

Figures

****

**Figure 1.** Plots of the performance of the model built within each cross-validation iteration. Graph **A** shows the AUC (x-axis) and p-values (y-axis) for each optimal model’s test-set predictions. *p*-values are computed by comparing the observed AUC to 1000 null values after permuting the test set’s group labels 1000 times. Graph B shows the 100 ROC curves for the test set predictions of each model; one blue line is plotted for each cross-validation iteration.

**Figure 2.** The ROC curve generated from the vector of each subject’s predicted probability of PI group membership, averaged across the model that excluded the subjects from the training or tuning subsets. The blue region shows the bootstrap 95% confidence interval from 2000 resamples, while the dark grey region quantifies the AUC. The 45% degree line bisecting the graph represents an AUC of .5, corresponding to random predictions.



**Figure 3.** Each variable’s 100 variable importance scores, computed at each cross-validation iteration on a given model’s respective test set. The red dot indicates the average PVI across the 100 cross-validation repetitions, accompanied by the 95% confidence interval of the standard error for the mean PVI estimate. Colors indicate the average rank for the variables across the 100 models. Within a given model, ranks can range from 1 to *i*, where *i* is the number of SGMV predictors and a rank of 1 corresponds to the largest PVI*i*.