A Random Forests Analysis of Reading-Related Skills and White Matter Tractography

Julie A. Van Dyke[1], Kazunaga Matsuki[1,2], Hannah R. Jones[4], Peter J. Molfese[1,3], Andrew Jahn[1], Clinton L. Johns[1], Dave Kush[1], Morgan L. Bontrager[1]

[1]Haskins Laboratories, [2]McMaster University, [3]University of Connecticut, [4]University of Rochester

Haskins Laboratories

INTRODUCTION

Random Forests is a machine-learning based, non-parametric statistical technique for estimating *variable importance* (Breiman, 2001; Matsuki et al., in press).
- *Variable importance refers to a predictor's contribution to explaining response variability*
- *Especially effective when predictors are collinear, as with reading-related skills*
- *Built-in protection against model-overfitting makes it very effective when ratio of predictors to observations is high.*
- Data-driven method for identifying potentially interesting variables.
  - Especially suited to exploring DTI data, where functional and theoretical models are still nascent

METHOD

Subjects: 74 participants, aged 16-24 recruited from the New Haven, CT community. Many were not college-attending.
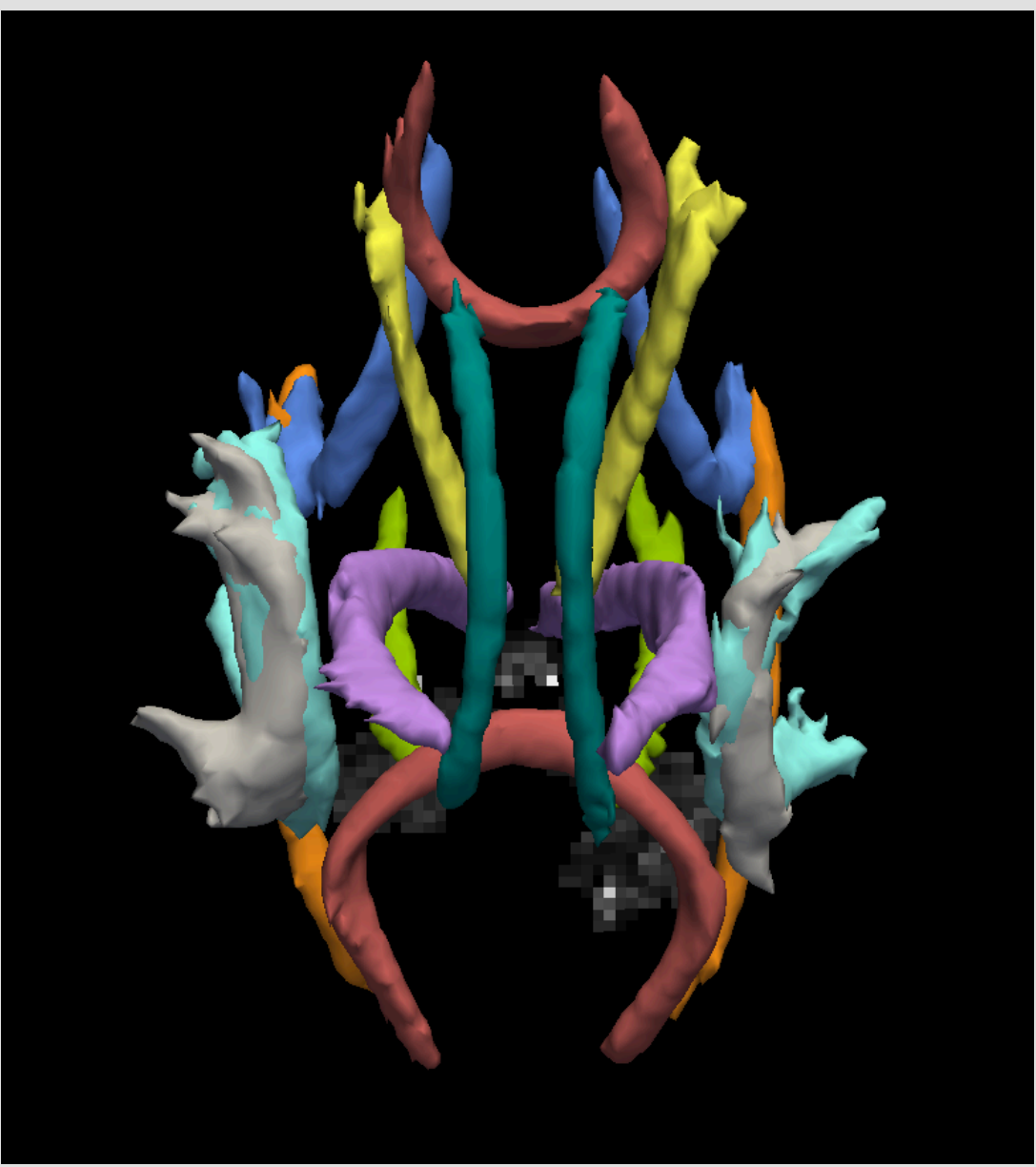
Skill measures: A 5+ hour battery of reading assessments delivered over 2 days, yielding 15 composite measures (see ASSESSMENTS)

Diffusion data acquisition and analysis: TORTOISE 2.0.1 (Pierpaoli et al., 2010) was used for co-registration and non-linear tensor fitting. TRACULA (Yendiki et al., 2011) within Freesurfer 5.3 was used to extract Fractional Anisotropy (FA), Mean Diffusivity (MD), tract Volume, and Average Path length for 18 white-matter tracts. Volume and Path Length were corrected for total brain volume.

ASSESSMENTS

Skill Assessments and Reference Labels

CONSTRUCT	ABBREVIATION	ASSESSMENTS
COMPLEX MEMORY	COMPMEM	Operation Span
		Sentence Span
SPATIAL MEMORY	SPACEMEM	Corsi Blocks
INHIBITION	INHIB	Stop Signal
		Stroop
EXPERIENCE	EXP	Author Recognition
		Magazine Recognition
		Title Recognition
READING COMPREHENSION	RCOMP	Gates MacGinitie
		Nelson Denny
		WJ3 Reading Comp
IQ	IQ	WASI IQ
VOCABULARY	VOCAB	PPVT
WORD READING	WREAD	TOWRE
		WJ3 Word ID
NON-WORD READING	NWREAD	TOWRE
		WJ3 Word Attack
FLUENCY	FLU	WJ3 Reading Fluency
ORAL COMPREHENSION	ORAL	WJ3 Oral Comprehension
SYNTAX	SYN	CELF Recalling Sentences
		CELF Formulating Sentences
		CELF Sentence Assembly
PHONOLOGICAL AWARENESS	PA	CTOPP2 PA Composite
PHONOLOGICAL MEMORY	PM	CTOPP2 Phono Memory Composite
RAPID NAMING	RAN	CTOPP2 RAN Composite

TRACTOGRAPHY
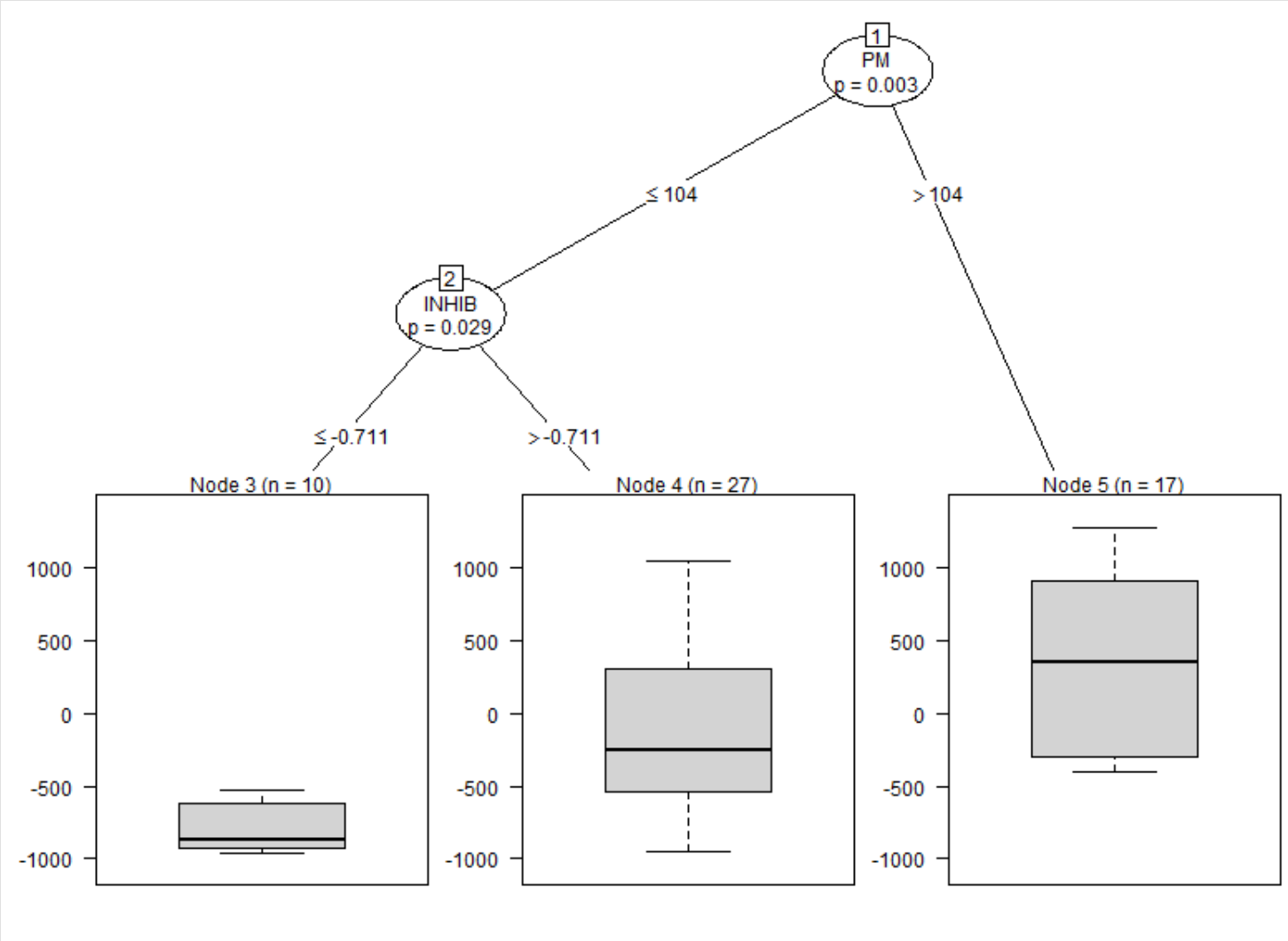
5100 Corpus Callosum Forceps Major
5101 Corpus Callosum Forceps Minor
5102 Left Anterior Thalamic Radiation
5103 Left Cingulum - Angular Bundle
5104 Left Cingulum - Cingulate Gyrus
5105 Left Corticospinal Tract
5106 Left Inferior Longitudinal Fasciculus
5107 Left Superior Longitudinal Fasciculus - Parietal
5108 Left Superior Longitudinal Fasciculus - Temporal
5109 Left Uncinate Fasciculus
5110 Right Anterior Thalamic Radiation
5111 Right Cingulum - Angular Bundle
5112 Right Cingulum - Cingulate Gyrus
5113 Right Corticospinal Tract
5114 Right Inferior Longitudinal Fasciculus
5115 Right Superior Longitudinal Fasciculus - Parietal
5116 Right Superior Longitudinal Fasciculus - Temporal
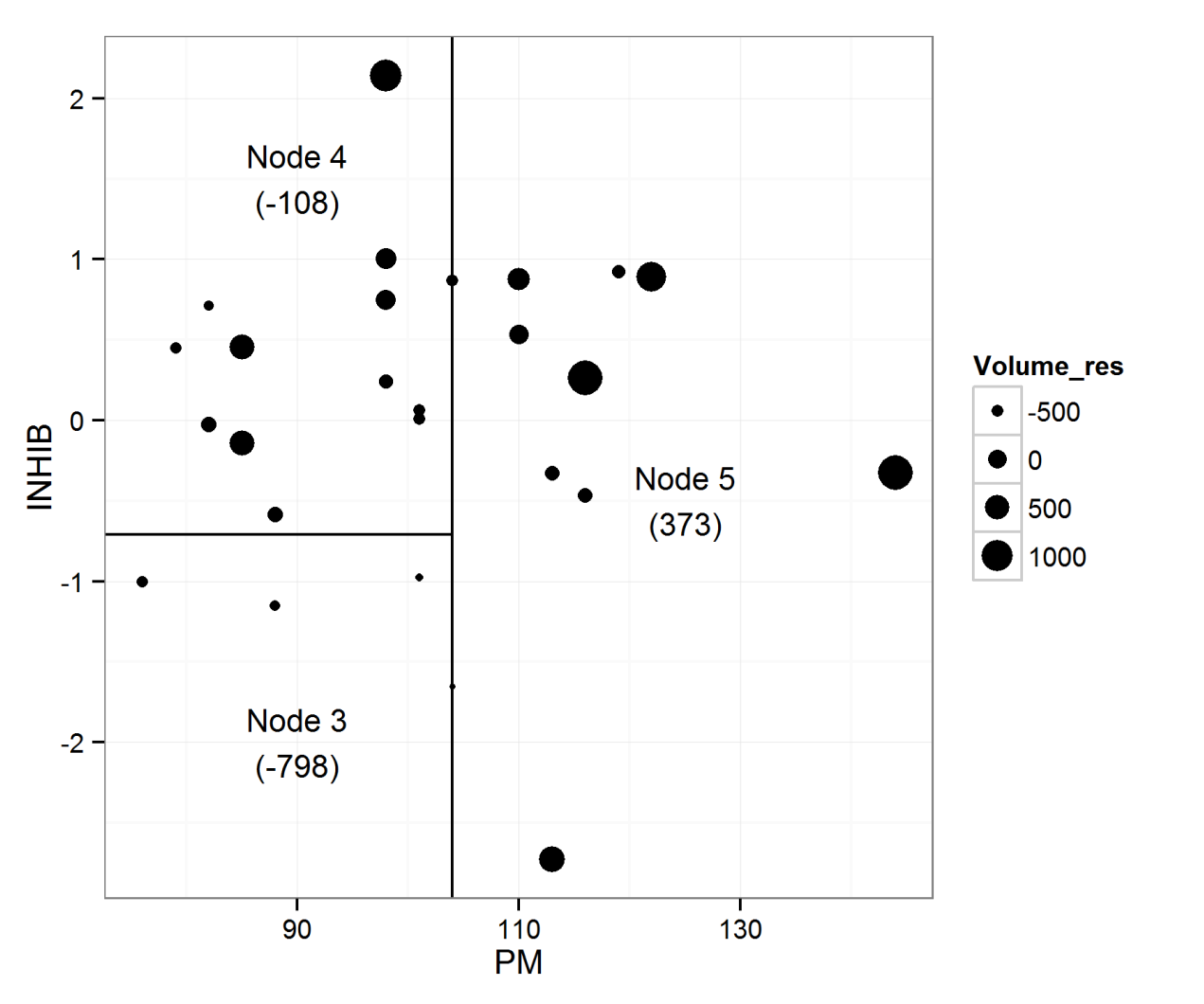5117 Right Uncinate Fasciculus

DECISION TREES

Partitions data at various cutpoints to preserve homogeneity. In this tree, individuals with Phonological Memory ≤ 104 are more similar, and those individuals are further partitioned at inhibition scores of -.711. That is, if Phono Memory ≤ 104 then Inhibition ability also matters.

Decision Tree with 3 Partitions

Data Distribution for each Partition

Decision Rules

Node 3: If Phono Memory ≤ 104 AND If Inhibition ≤ -.711 then mean Tract Volume in LSLF-Temporal is -798.
Node 4: If Phono Memory ≤ 104 AND If Inhibition > -.711 then mean Tract Volume in LSLF-Temporal is -108.
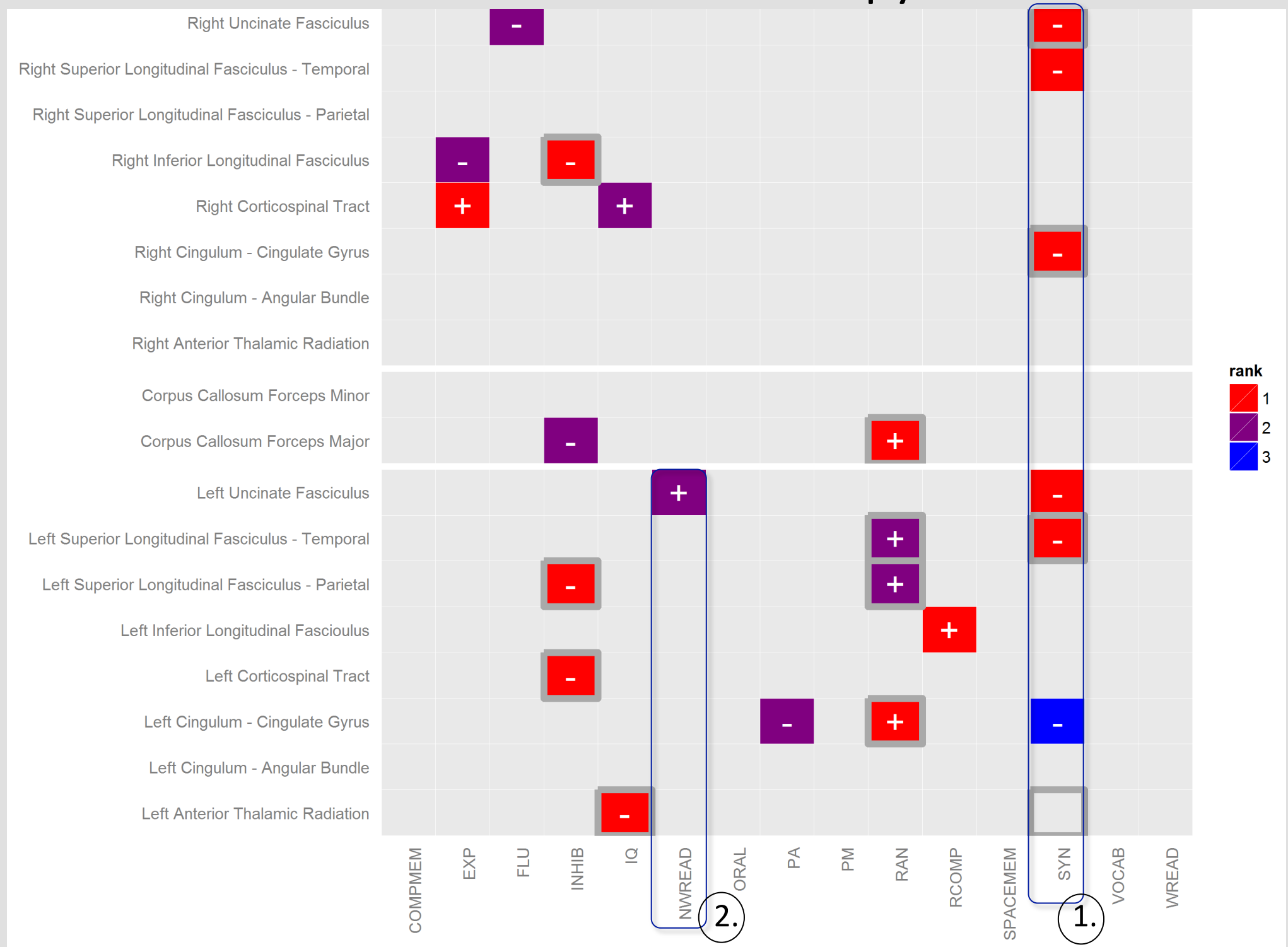Node 5: If Phono Memory > 104 then mean Tract Volume in LSLF-Temporal is 373.
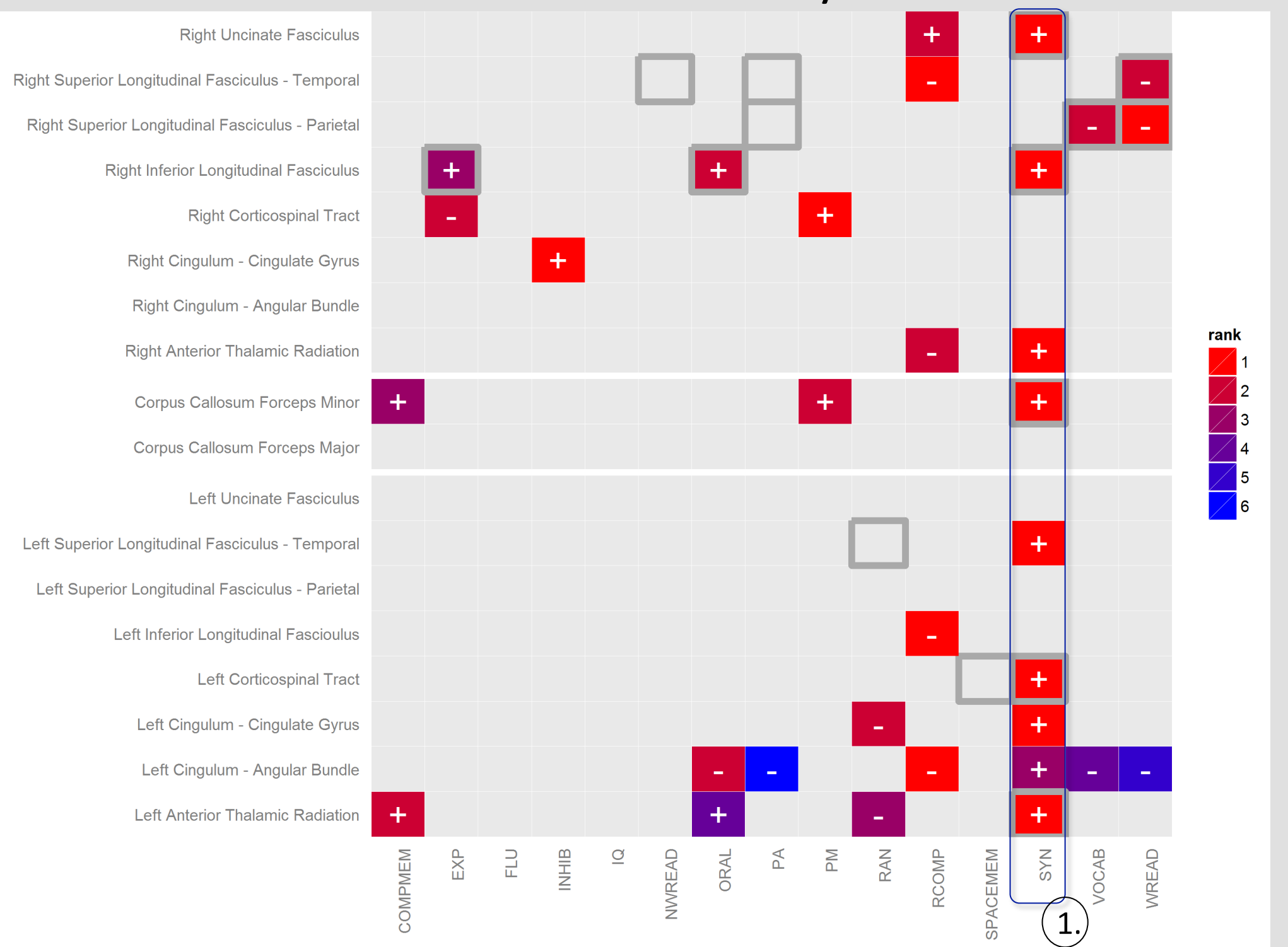
RANDOM FOREST METHOD

A Forest of Decision Trees is created from multiple random samples of data and predictors. Predictor importance was assessed via a procedure that randomly shuffles the predictor among the trees to determine how its presence or absence affects data modeling. Rankings shown below are aggregated over 10,000+ decision trees.
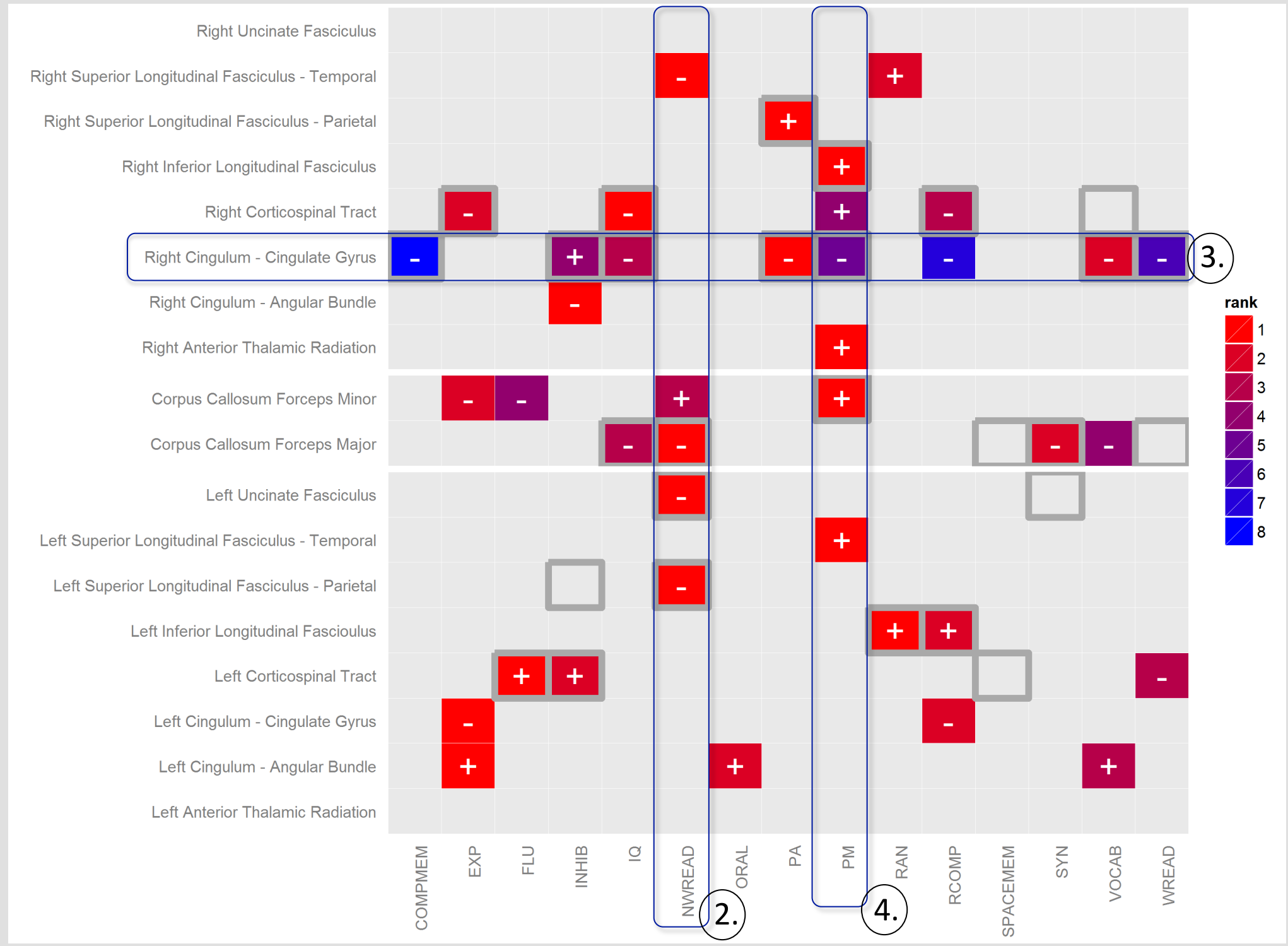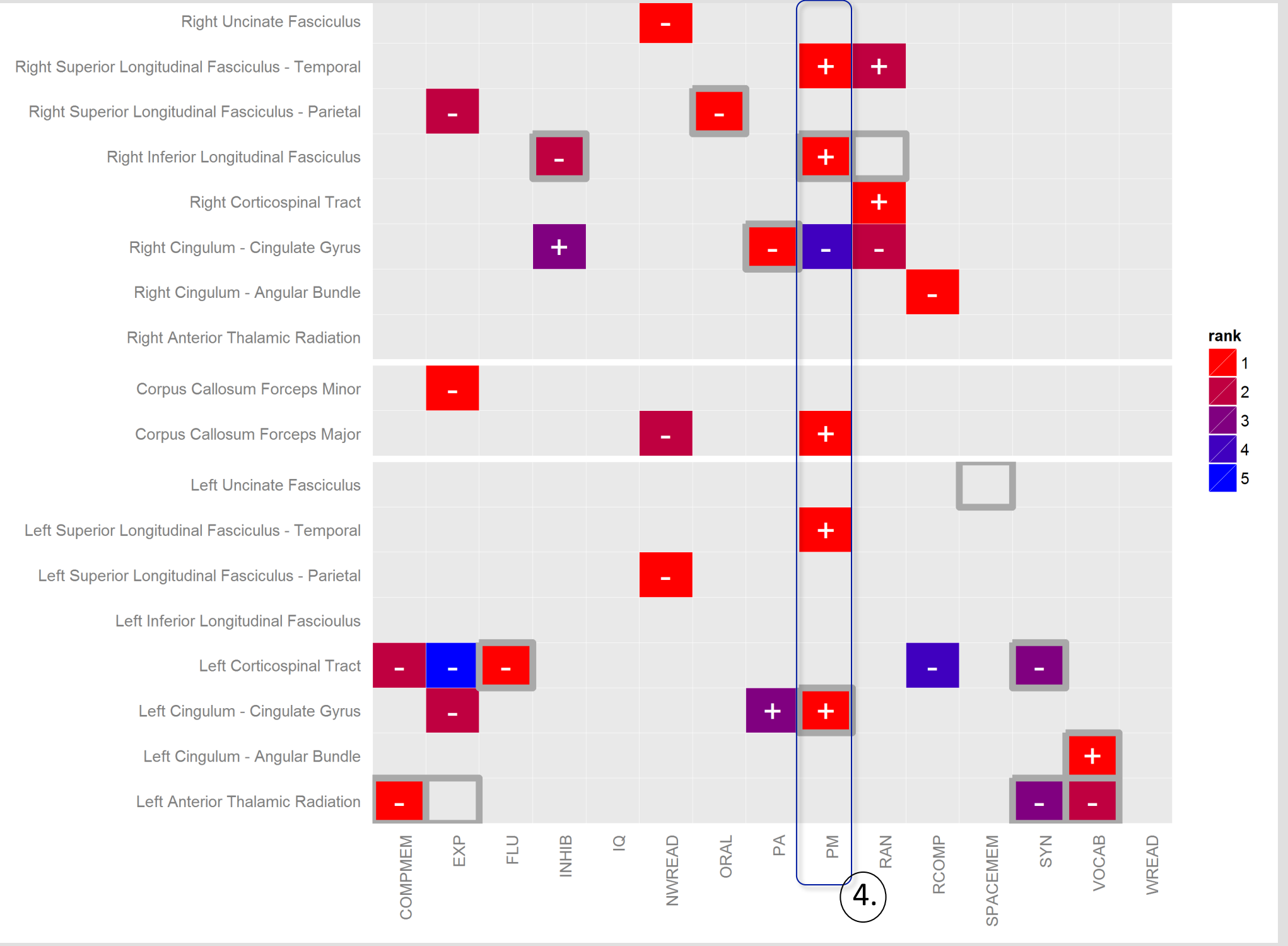
RESULTS

Fractional Anisotropy — Mean Diffusivity — Tract Volume (corrected) — Average Path Length (corrected)

WHY USE RANDOM FORESTS?

- Efforts to relate extensive skill batteries to DTI measures suffer from high risk of model overfitting.
- Reading-related skills are highly collinear. Condition number is an index of collinearity and should be < 30 (Dormann et al.) Condition number for this (typically-sized) Assessment Battery is ≈ 118.
- Data compression methods (e.g., PCA) are not always interpretable.

	PC1	PC2	PC3
COMPMEM	0.0126	0.0108	0.0137
INHIB	-0.0032	-0.0041	-0.001
EXP	-0.0029	0.0298	-0.0032
RCOMP	0.0016	0.0253	0.0031
IQ	0.0757	0.5147	-0.0145
VOCAB	-0.0309	0.6396	0.1016
WREAD	0.0077	0.0209	0.0148
NWREAD	0.0124	0.0108	0.0103
FLU	0.0615	0.5425	0.0855
ORAL	0.0021	0.1236	-0.004
SYN	0.0068	0.0163	0.0036
PA	0.6185	-0.0516	0.2326
PM	0.7788	-0.0278	-0.1846
RAN	-0.0017	-0.1027	0.9452

- Factor 1 ≈ Phonological Ability
- Factor 3 ≈ Rapid Naming
- Factor 2 ≈ ? (subjectivity)

In this data, no correlations of PCA Factors with DTI measures were significant. Many individual assessments were significant in *uncorrected* analyses (outlined boxes).

- Correlations surviving correction: Tract Volume– corpus callosum forceps minor: NWREAD and SYN. Mean Diffusivity– Left Anterior Thalamic Radiation: SYN

FEATURED RESULTS

Confirmatory Results (Consistent with previous findings)
1. Syntax is highest ranked measure related to FA and MD in Uncinate and Superior Longitudinal Fasciculi (e.g., Friederici & Gierhan, 2013; Wilson et al., 2011).
2. Nonword reading is highest ranked measure related to Tract Volume in Left Uncinate and Corpus Callosum (Welcome & Joanisse, 2014)

Hypothesis Generation (Novel Findings)
3. Tract volume in Right Cingulate Gyrus is related to many reading skills.
4. Phonological Memory is a highly ranked predictor for the size-related dependent measures (Volume, Path Length) of multiple tracts.
5. Many other skills are identified as highly ranked predictors of particular tracts. These warrant additional confirmatory testing.

CONCLUSIONS

The Random Forests method provides a valuable tool for exploring the relationship between reading-related skills and white matter tracts.
- Especially valuable due to collinearity of skill assessments, which is pervasive in reading and language research.
- This approach provides a means of maximizing the usefulness of comprehensive assessment batteries in neurobiological research, and can lead to novel hypotheses about the function of white matter in language processing and reading disability.

REFERENCES & ACKNOWLEDGEMENTS

Breiman (2001). *Machine Learning, 45*; Dormann, et al. (2012) *Ecography, 35*; Friederici & Gierhan (2013). Current Opinion in Neurobiology, 23; Matsuki et al. (in press). Scientific Studies of Reading; Pierpaoli et al. (2010). ISMRM 18th annual meeting, #1597; Welcome & Joanisse (2014). *NeuroImage, 96*; Wilson et al. (2011). *Neuron, 72*(2); Yendiki et al., (2011). *Frontiers in neuroinformatics, 5*.