Supplemental Material: Multidimensional analysis and detection of informative features in human brain white matter

Adam Richie-Halford[1*], Jason Yeatman[2], Noah Simon[3], Ariel Rokem[4],

1 eScience Institute, University of Washington, Seattle, WA, USA
2 Graduate School of Education and Division of Developmental and Behavioral Pediatrics, Stanford University, Stanford, CA, USA
3 Department of Biostatistics, University of Washington, Seattle, WA, USA
4 Department of Psychology, University of Washington, Seattle, WA, USA

* richford@uw.edu

1 Bundle and coefficient profiles

Here, we present the bundle profiles and $\hat{\beta}$ coefficients for each dataset. Throughout this section, diffusion metrics are plotted along the length of eighteen bundles: right corticospinal (CSTR), left corticospinal (CSTL), right uncinate (UNCR), left uncinate (UNCL), left inferior fronto-occipital fasciculus (IFOL), right inferior fronto-occipital fasciculus (IFOR), right arcuate (ARCR), left arcuate (ARCL), right thalamic radiation (ATRR), left thalamic radiation (ATRL), right cingulum cingulate (CGCR), left cingulum cingulate (CGCL), callosum forceps posterior (CFP), callosum forceps anterior (CFA), right inferior longitudinal fasciculus (ILFR), left inferior longitudinal fasciculus (ILFL), right superior longitudinal fasciculus (SLFR), and left superior longitudinal fasciculus (SLFL). We display results for two different diffusion metrics: fractional anisotropy (FA) and mean diffusivity (MD), which are extracted from different diffusion models depending on the dataset: diffusion tensor imaging (DTI) for the ALS and WH datasets and diffusion kurtosis imaging (DKI) for the HBN and Cam-CAN datasets [1]. The diffusion metric is always plotted on the left y-axis while the $\hat{\beta}$ coefficients are displayed on the twin axis on the right-hand-side. The scale of the $\hat{\beta}$-axis is shared between the FA and MD metrics so that one can compare the relative imporance of each metric.

1.1 ALS bundle profiles

Figures 1 and 2 show the bundle profiles and regression coefficients for the ALS dataset FA and MD metrics, respectively. These figures reinforce the findings in the main text that ALS is localized to the corticospinal tract. In this study, SGL selected the right corticospinal tract (CSTR) as important and regularized coefficients in the CSTL. Yet, Fig. 1 also shows group FA differences in the CSTL. This highlighted a potential drawback of the SGL method, discussed in the main text in the context of age regression. Namely, SGL is not guaranteed to identify *all* important features. In this case, if the diagnostic signal in the CSTL is redundant to that in the CSTR, SGL will regularize the CSTL features, thereby reducing its sparsity penalty without any corresponding increase in loss. This parsimony cuts both ways; it is a feature of the method when one seeks an efficient predictive model, but is a disadvantage of the method when one wants an exhaustive explanation of feature importance. We will use the phrase "parsimony pitfall" to refer to the case when SGL regularizes away redundant but obviously important features.

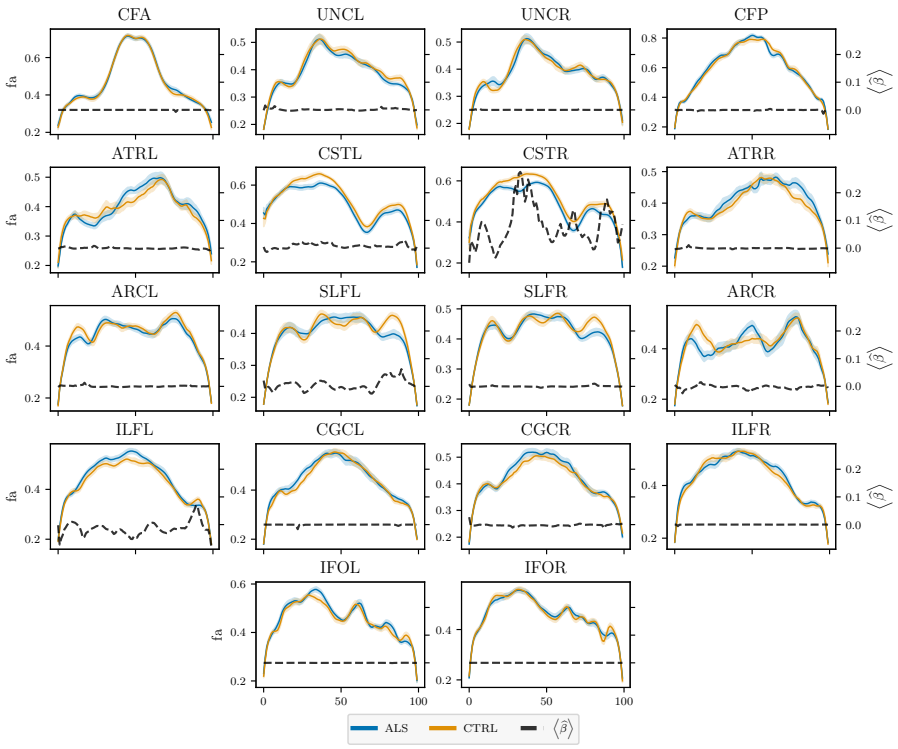

Fig 1. Fractional anisotropy (FA) bundle profiles and $\hat{\beta}$ coefficients for ALS classification.

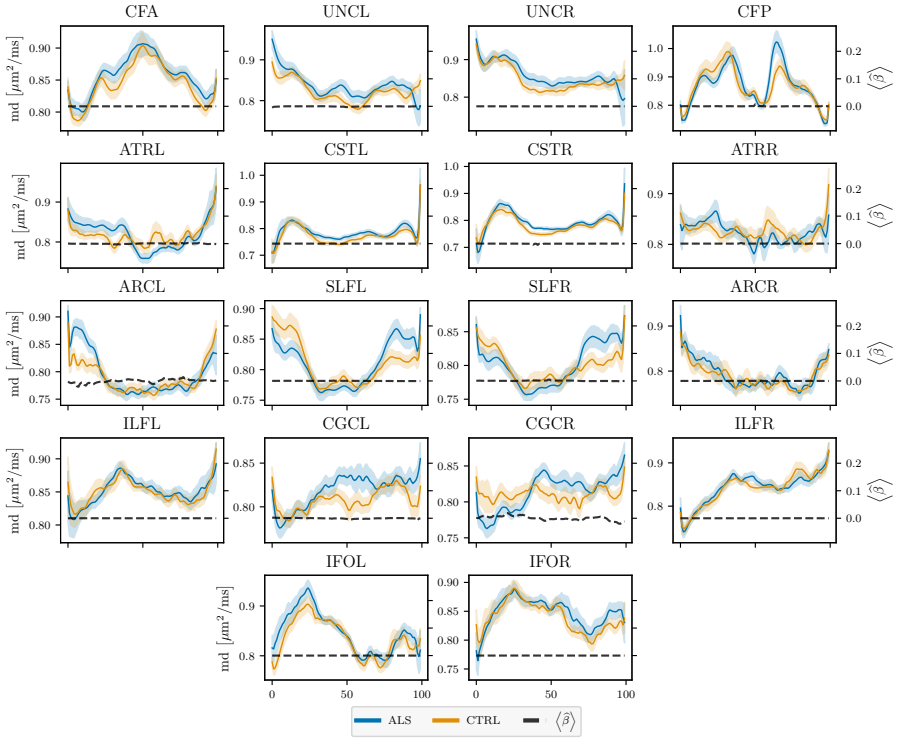

Fig 2. Mean diffusivity (MD) Bundle profiles and $\hat{\beta}$ coefficients for ALS classification.

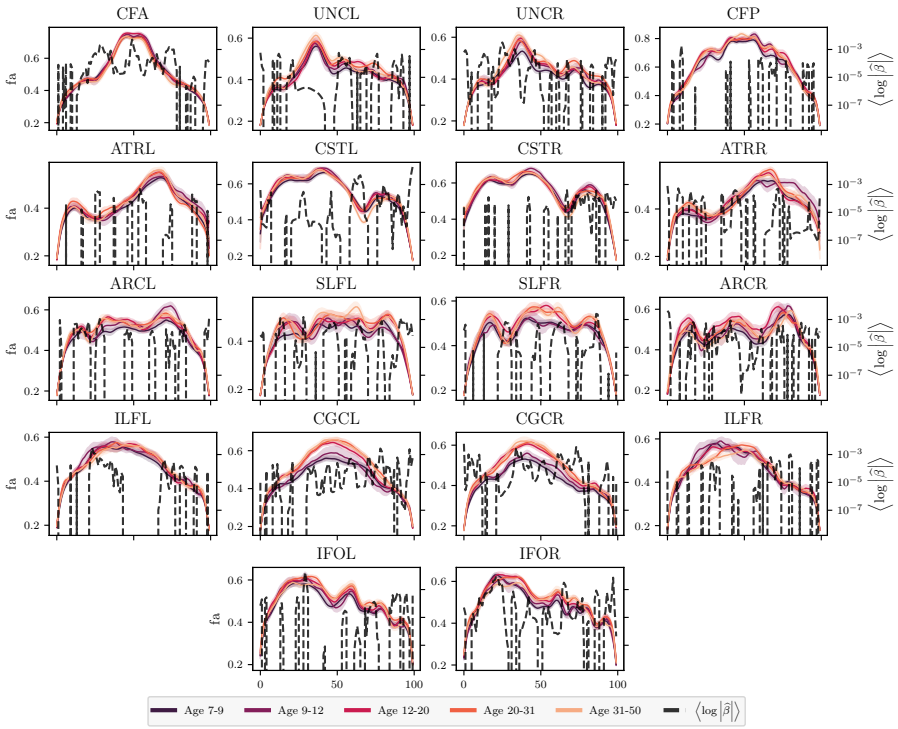

Fig 3. Fractional anisotropy (FA) bundle profiles and $\hat{\beta}$ coefficients for age regression in the WH dataset.

1.2 WH bundle profiles

Figures 3 and 4 show the bundle profiles and regression coefficients for the WH dataset. In contrast to the ALS classification case, the $\hat{\beta}$ coefficients are distributed widely through the brain, supporting the interpretation that aging is a large and continuous whole-brain process. These figures also show that SGL behaves much more like the lasso than the group lasso, as discussed in the main text. The parsimony pitfall is most evident in the IFOL and IFOR bundles in Fig. 4.

1.3 HBN bundle profiles

Figures 5 and 6 show the bundle profiles and regression coefficients for the HBN dataset. Like the WH dataset, the $\hat{\beta}$ coefficients are distributed widely through the brain and SGL behaves more like the lasso than the group lasso. In contrast the the WH results, the bundle profiles show different behaviors. For example the SLFL and SLFR bundle profiles in Fig. 4 and Fig. 6 have different concavity. This is unsurprising, however, given the differences between these datasets: (i) different diffusion models, with DTI for the WH dataset and DKI for the HBN and Cam-CAN datasets, (ii) different age ranges and distributions (which is evident in the figure legends), with HBN being a developmental dataset, while WH and Cam-CAN are lifespan maturation datasets, and (iii) different anatomical extents, with the WH streamlines truncated to remain with the bundle's bounding regions of interest (the default behavior in the legacy `mAFQ`) and the HBN and Cam-CAN streamlines allowed to retain their full extent from whole-brain tractography (the default behavior in `pyAFQ`). Thus, one should use caution when comparing bundle profiles and $\hat{\beta}$ coefficients between the WH, HBN, and Cam-CAN models. The parsimony pitfall is most evident in the UNCL, UNCR, ARCL, SLFL, and

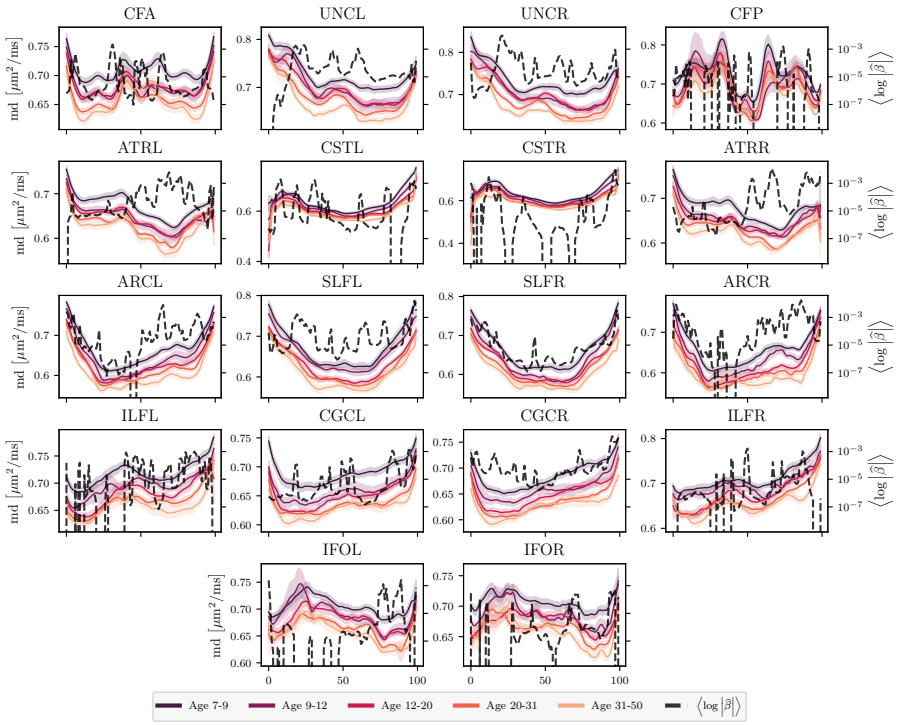

Fig 4. Mean diffusivity (MD) bundle profiles and $\hat{\beta}$ coefficients for age regression in the WH dataset.

SLFR bundles in Fig. 6.

1.4 Cam-CAN bundle profiles

Figures 7 and 8 show the bundle profiles and regression coefficients for the Cam-CAN dataset. Like the WH and HBN datasets, the $\hat{\beta}$ coefficients are distributed widely through the brain and SGL behaves more like the lasso than the group lasso. As before, one must be cautious about comparing bundle profiles and $\hat{\beta}$ coefficients between models. While the HBN and Cam-CAN datasets share the same diffusion model and refrain from clipping streamlines, the age distributions for the two are roughly disjoint, with the WH age distribution straddling the two. The parsimony pitfall is again evident in the UNCL, UNCR, ARCL, SLFL, and SLFR bundles in Fig. 8.

2 Data harmonization in the HBN dataset

Figures 10 and 11 show the bundle profiles for the HBN dataset separated by scanner site: Rutgers Rutgers University Brain Imaging Center (RU) and the CitiGroup Cornell Brain Imaging Center (CBIC). We see strong site differences in both FA and MD profiles. Conversely, Fig. 9 shows that the age distributions at each site are similar. ComBat effectively harmonizes the scanner site differences, as shown in Figs. 12 and 13.

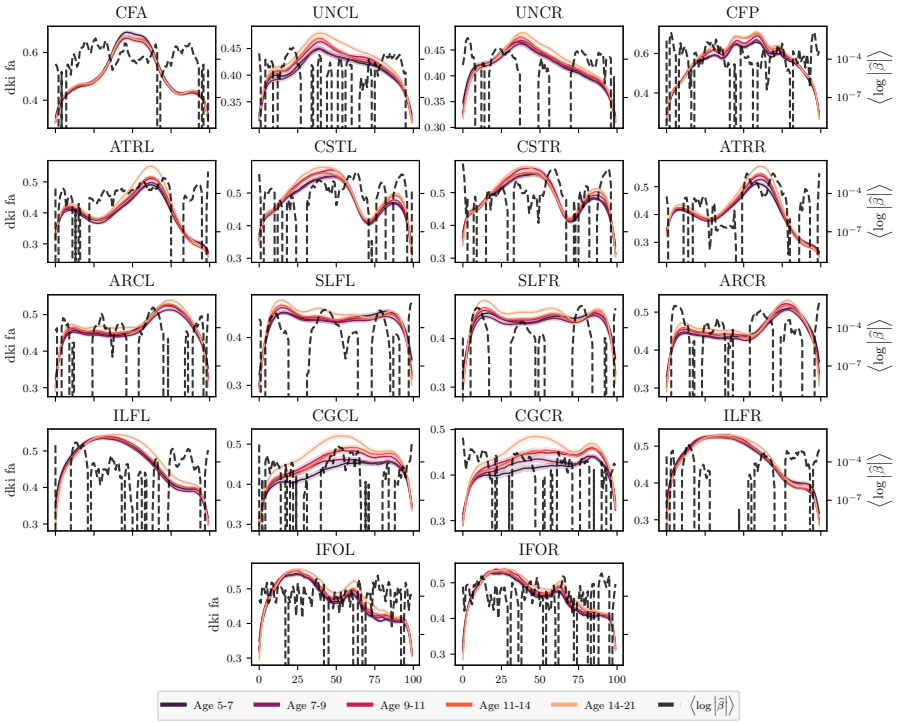

Fig 5. Fractional anisotropy (FA) bundle profiles and $\hat{\beta}$ coefficients for age regression in the HBN dataset.

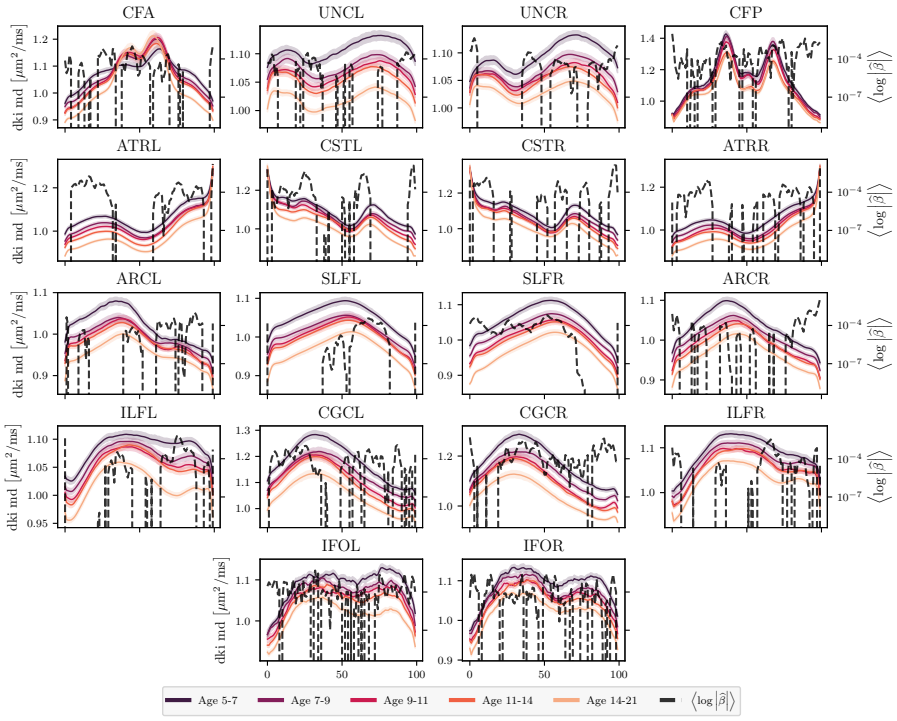

Fig 6. Mean diffusivity (MD) bundle profiles and $\hat{\beta}$ coefficients for age regression in the HBN dataset.

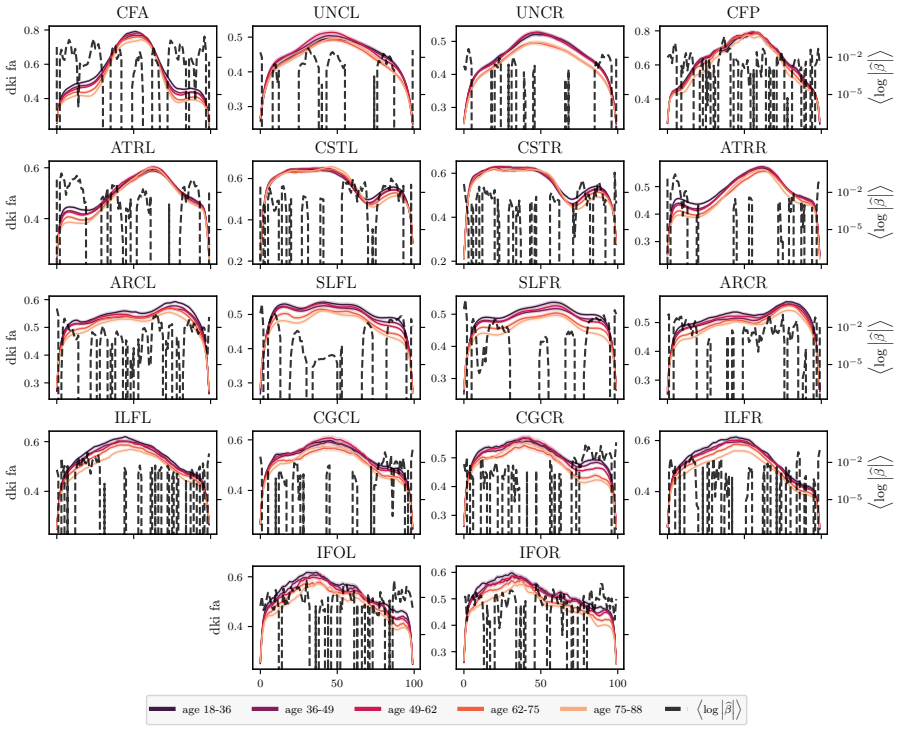

Fig 7. Fractional anisotropy (FA) bundle profiles and $\hat{\beta}$ coefficients for age regression in the Cam-CAN dataset.

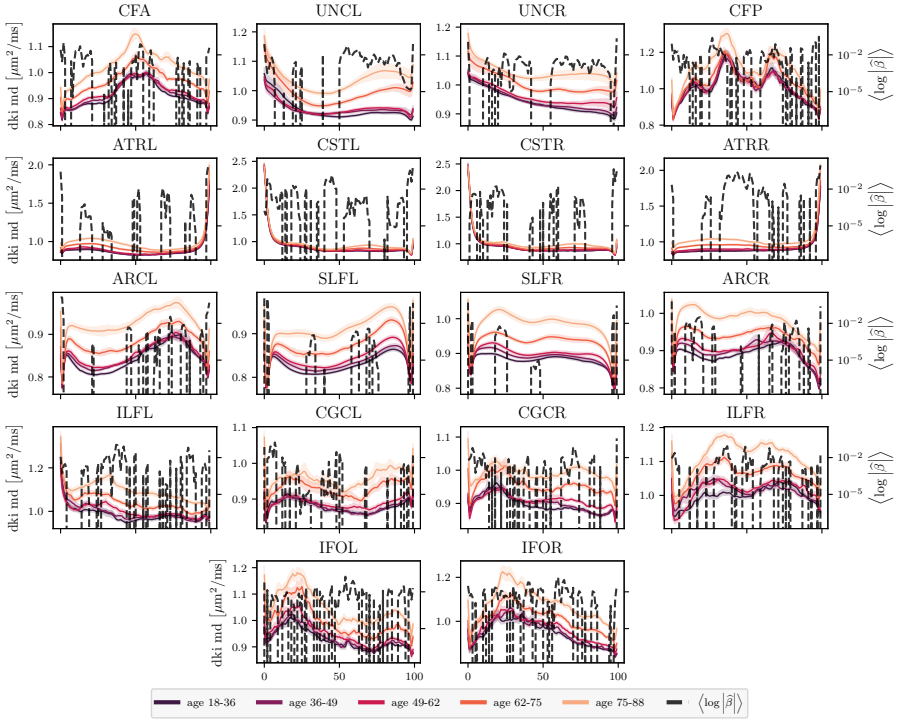

Fig 8. Mean diffusivity (MD) bundle profiles and $\hat{\beta}$ coefficients for age regression in the Cam-CAN dataset.

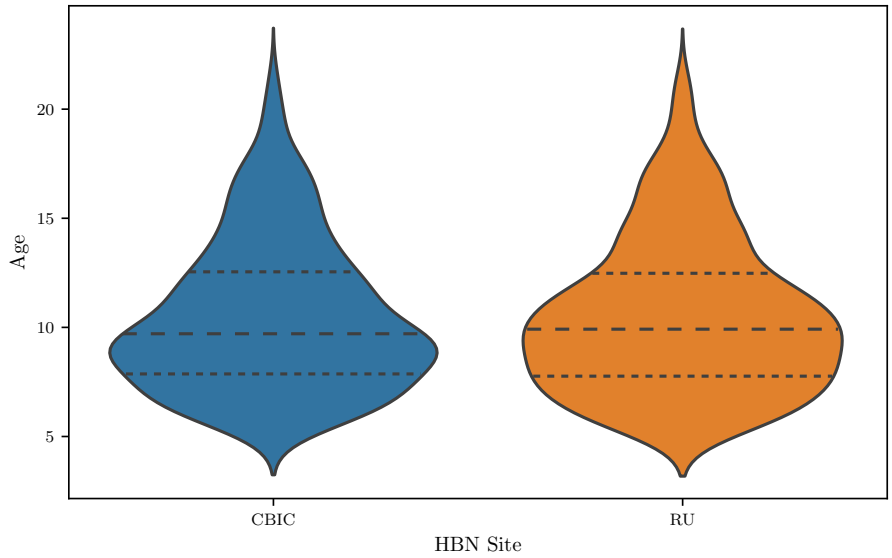

Fig 9. Age distributions are similar between the different HBN sites.

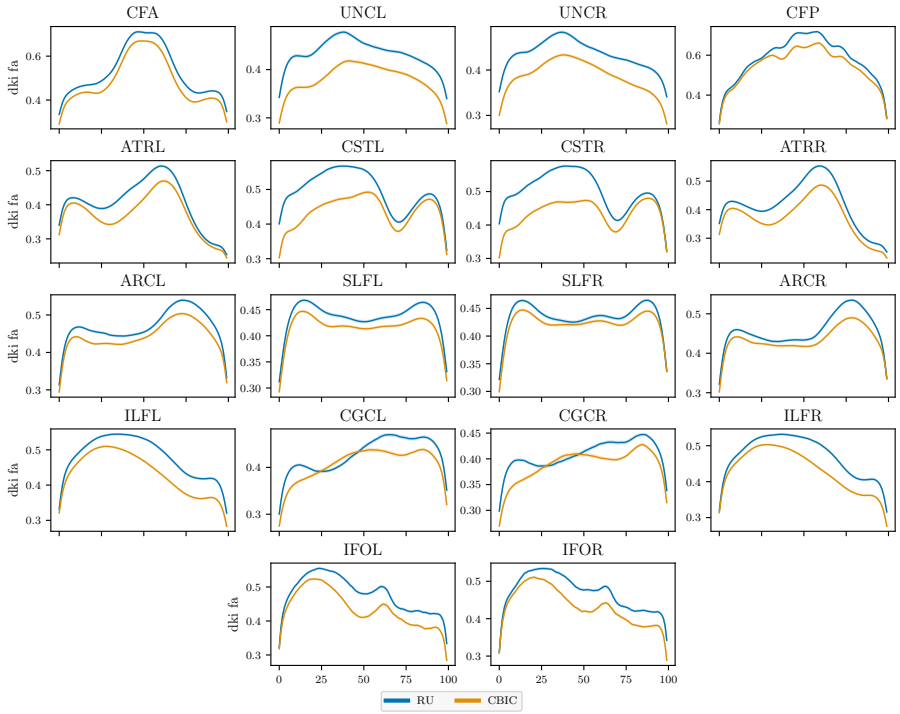

Fig 10. Fractional anisotropy (FA) bundle profiles exhibit strong site differences in the HBN dataset.

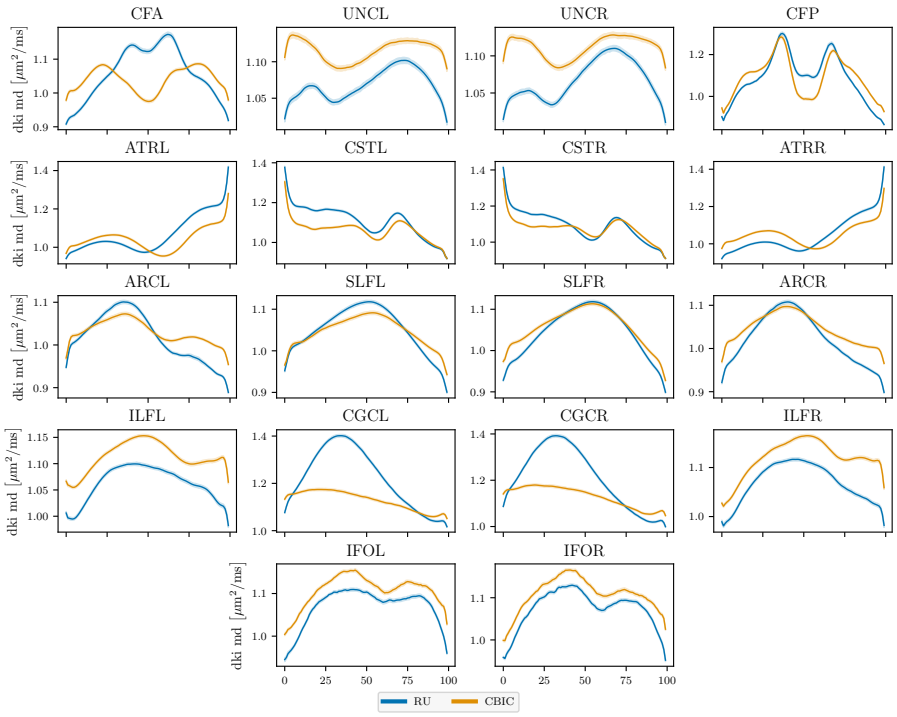

Fig 11. Mean diffusivity (MD) bundle profiles exhibit strong site differences in the HBN dataset.

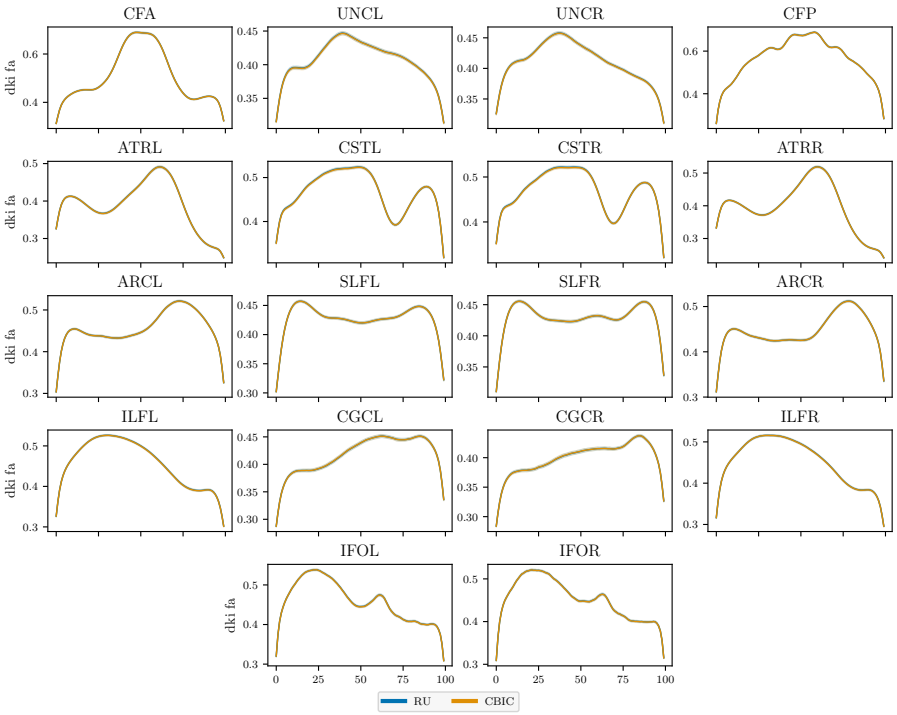

Fig 12. Site differences in fractional anisotropy (FA) do not survive ComBat harmonization.

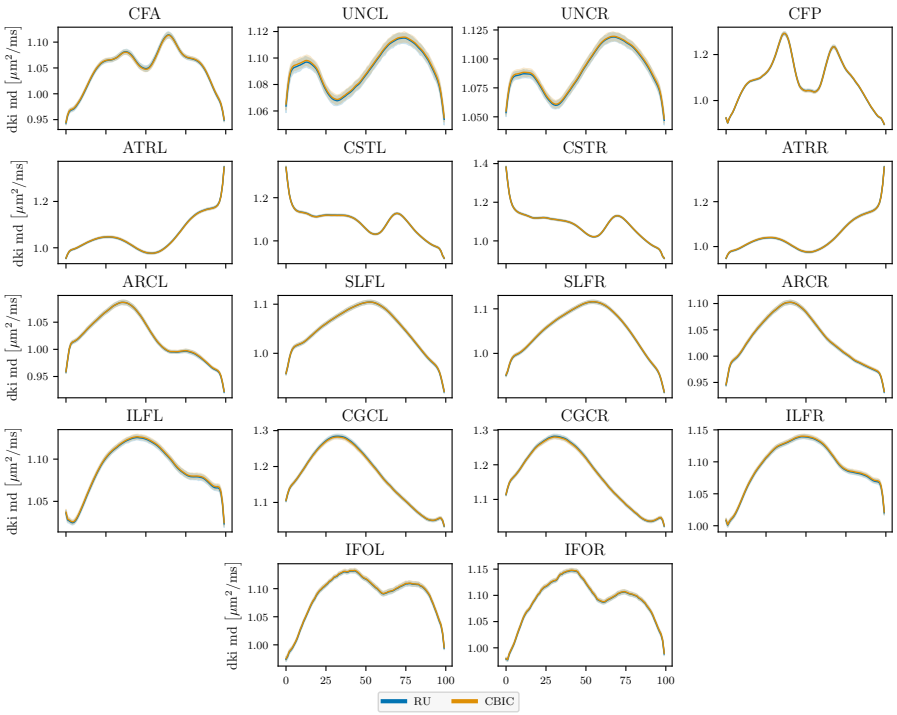

Fig 13. Site differences in mean diffusivity (MD) do not survive ComBat harmonization.

References

1. Jensen JH, Helpern JA, Ramani A, Lu H, Kaczynski K. Diffusional kurtosis imaging: The quantification of non-gaussian water diffusion by means of magnetic resonance imaging. Magnetic Resonance in Medicine. 2005;53(6):1432–1440. doi:https://doi.org/10.1002/mrm.20508.