GenTract: Generative Global Tractography

Supplementary Material

A. Statistical Analysis

Significance ($p < 0.01$) was determined using paired t-tests with one-sided bootstrapping on the test set.

Architecture (Table 1) Diffusion models significantly outperformed FM models at each architectural size ($M = 4, 6, 8$). Furthermore, the Diffusion model with $M = 8$ and dimension $n = 256$ yielded statistically superior results compared to all other tested architecture depths ($M = 4, 6$) and widths ($n = 128, 512$).

Benchmarking (Tables 2–5) GenTract demonstrated statistically significant improvements over the second-best performing methods (TractOracle or DDTracking) in both BS % P and TO-Net % P across all experiments.

B. Further GenTract Implementation Details

Input representation and preprocessing. Following standard practice [18, 35], we use an SH order of $L_{\max} = 6$, resulting in $m = 28$ SH coefficients. All streamlines are resampled to 128 points as in previous work [35]. Each of the 28 SH 3D coefficient volumes is padded to shape $(H, W, D) = (88, 112, 88)$.

Latent spaces and architecture. We use latent dimension values of $(C_z, H_z, W_z, D_z) = (4, 22, 28, 22)$ for the initial VAEs, and $(C_c, H_c, W_c, D_c) = (32, 11, 14, 11)$ for the class-conditioned encoder $\mathcal{E}^c$. Within $\mathcal{E}^c$, coefficient-index conditioning ($i \in [0, m-1]$) is implemented via a learned embedding injected into the residual downsampling blocks using the same additive mechanism commonly used for diffusion timestep embeddings. In the Transformer backbone, 8 attention heads are used for both self- and cross-attention.

Training setup and computational cost. All GenTract models are trained with a batch size of 1024 streamlines for 100 epochs. We use an initial learning rate of 5×10^{-5} with cosine annealing over the full training cycle to a final learning rate of 1.25×10^{-5}. Training proceeds in two stages: representation learning, where each of the m independent VAEs (one per SH coefficient) requires approximately 48 GPU-hours, and generative modeling, where the conditional Transformer requires 168–336 GPU-hours depending on the architecture.

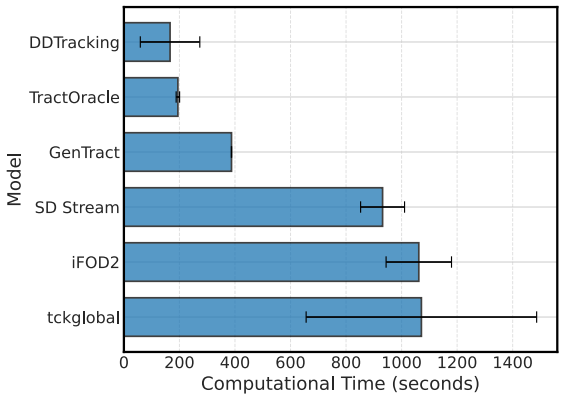

Figure 5. Computational time comparison for all methods. Bars show inference time (seconds) (mean ± std). The std for GenTract is very small.

C. Baseline Implementation Details

For each of the DL-based methods (TractOracle, DDTracking), we use the standard inference pipelines from their open source implementations. Specifically, we run TractOracle inference with parameters: noise level: 0.1, seeds per voxel: 10, number of actors: 2500, and DDTracking with parameters: step size: 1, max angle: 45, max streamline length: 200, min streamline length: 40. Seeding masks are generated by dilating Fractional Anisotropy (FA) masks at FA > 0.2 to approximate a white matter mask, which is then dilated (and subtracted) to form a White Matter / Gray Matter interface mask, as done in [35]. For the classic non-DL methods (iFOD2, SD Stream, tckglobal), we use recommended practices for their inference [37]. Specifically, iFOD2 and SD Stream are used to generate 5×10^6 streamlines before undergoing SIFT filtering to 500×10^3 streamlines, with seeding masks generated from 5ttgen [31]. tckgen is used with number of iterations: 1×10^9.

D. Qualitative Results

In Figure 6 we provide additional qualitative results to show generated tractograms (with 5,000 sampled streamlines) from GenTract, along with other baseline methods. We also show two different segmented bundles: the Corpus Callosum Occipital (CC_Oc) in Figure 7, and the Left Pyramid Tract (PYT_L) in Figure 8, generated by GenTract across different subjects, showing subject-specific geometries. Additionally, we compare bundle-specific results across baseline methods to further assess sensitivity of these methods to noisy and low-resolution conditions (Figure 9).

In this figure, we present a failure mode of GenTract. We see that GenTract shows enhanced bundle attenuation under low-resolution and noisy conditions compared with baseline methods for the PYT_L bundle, but similar to the other baseline methods, it is unable to resolve the CC_Oc bundle.

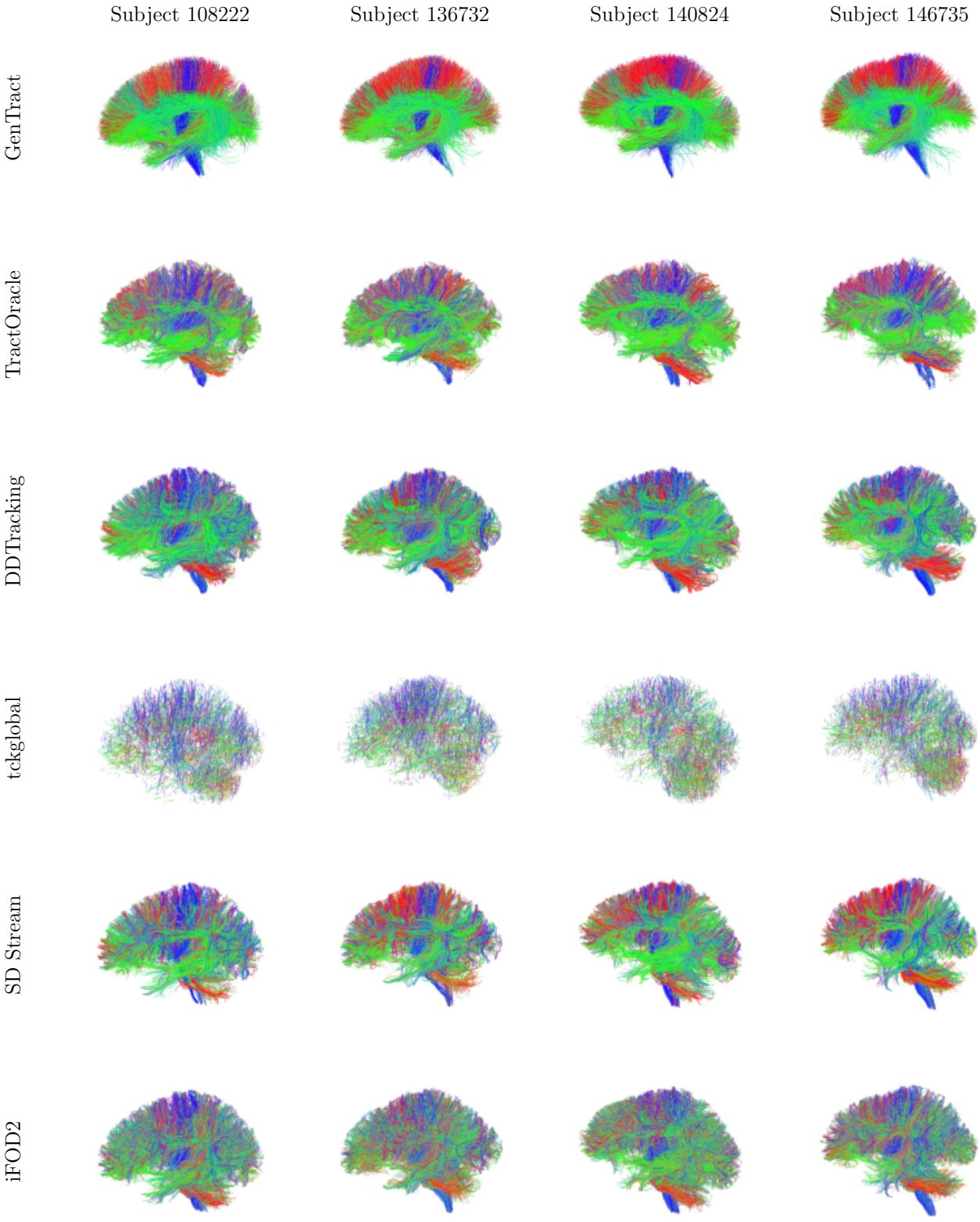

Figure 6. Qualitative result showing generated tractograms for 4 HCP test subjects for all methods on original (uncorrupted) data.

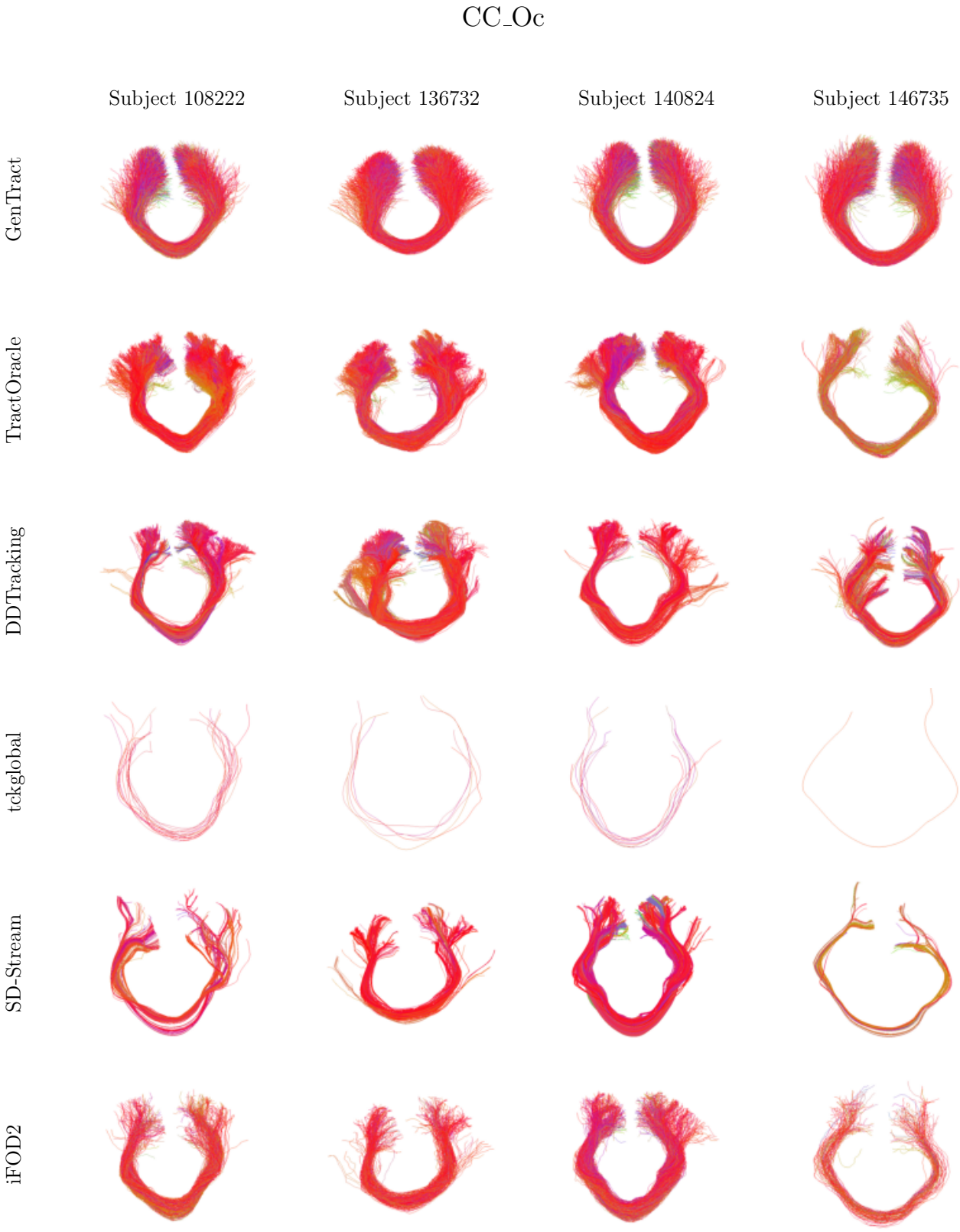

Figure 7. Qualitative result showing generated and segmented CC_Oc bundles for 4 HCP test subjects for all methods on original (uncorrupted) data.

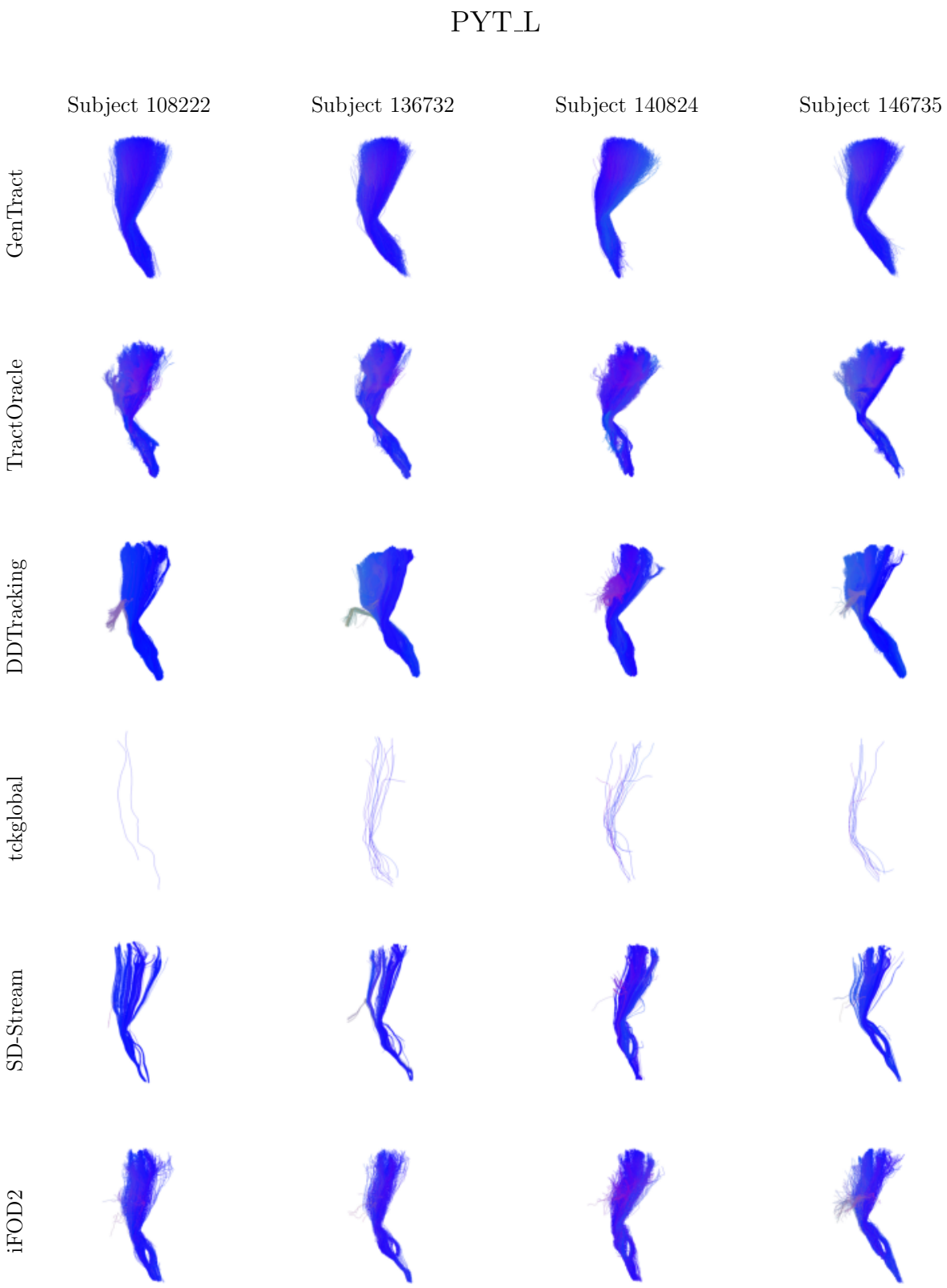

Figure 8. Qualitative result showing generated and segmented PYT_L bundles for 4 HCP test subjects for all methods on original (uncorrupted) data.

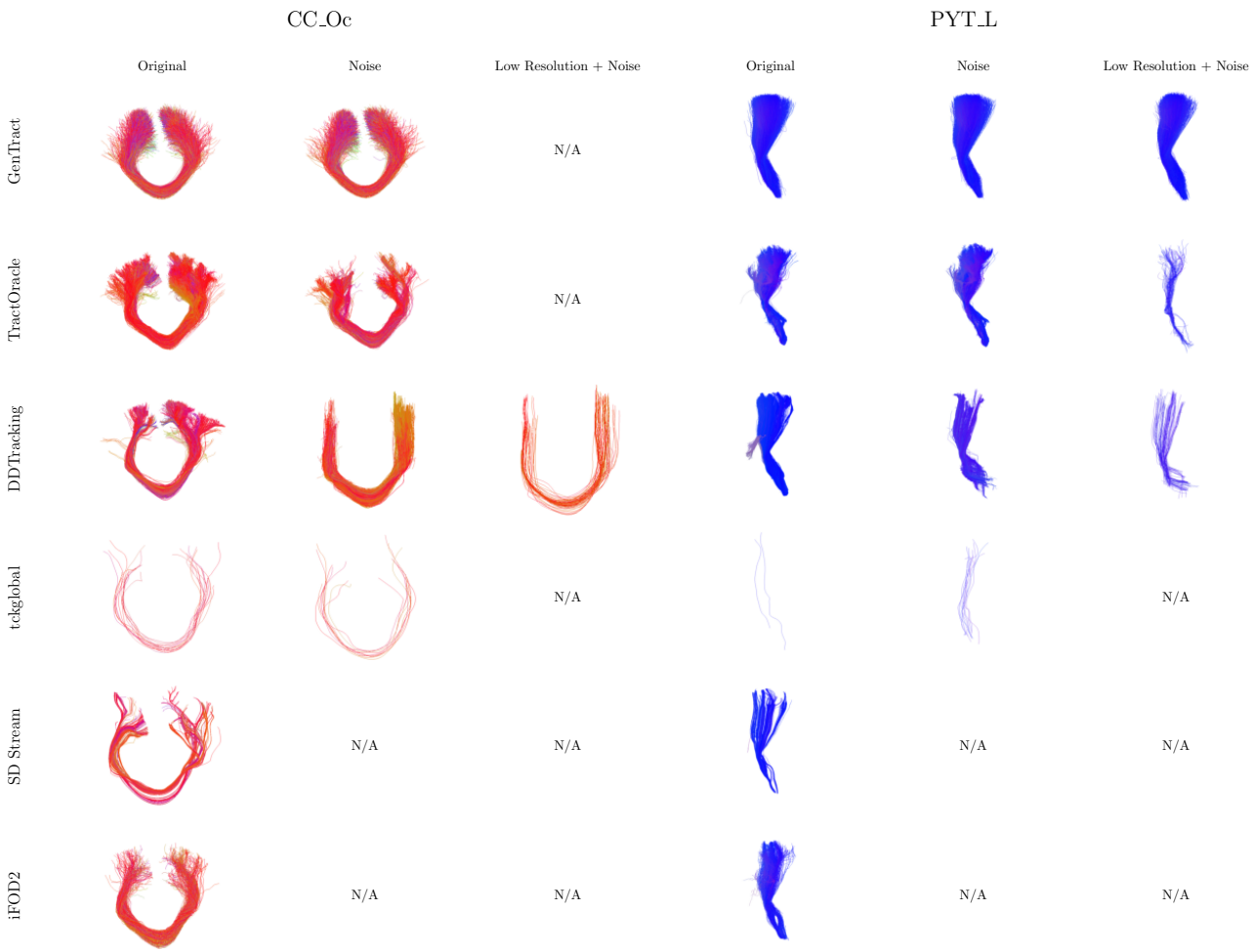

Figure 9. Qualitative result showing generated and segmented CC_Oc and PYT_L bundles for a single HCP test subject (108222) for all methods on original, noisy, and low-resolution and noise settings.