A Epoch 10 vs Epoch 70 Comparison of Convergence

To verify that the epoch-10 checkpoint used in Figure 1 is representative rather than an artifact of checkpoint selection, we test whether the regional ranking of instability holds, whether its magnitude holds, and whether any disagreement is systematic across our random seed models. For ranking, Spearman ρ between per-ROI U_S at epoch 10 and epoch 70 is 0.98 (95% CI [0.96, 0.99]): the spatial pattern of instability is preserved between checkpoints. Because ρ is blind to a uniform shift in absolute U_S values, we also report ICC(A,1) = 0.92 ([0.88, 0.95]), which measures absolute agreement. That it remains high, only slightly below $\rho = 0.98$, confirms that not just the ordering but the magnitude of per-ROI variability is largely preserved.

Bland-Altman analysis gives a bias of −0.03 Dice (95% LoA [−0.10, +0.05]), indicating epoch 10 is slightly more unstable on average. The differences are concentrated in low-Dice cortical regions (mean absolute difference of 0.04) while subcortical structures agree almost exactly (0.01). The offset ($\sim$ 0.03 Dice) is far smaller than the FastSurfer–FreeSurfer differences in Figure 1 and does not alter its conclusions; a more converged model would, if anything, show marginally lower cortical variability.

B Data Augmentation Tables

Table 1: Best MAE reached across repetitions (minimum of MAE across repetitions), with the repetition count r at which it occurred (in brackets).

	FastSurfer Internal			**HBN**		
Strategy	GB	RF	SVM	GB	RF	SVM
Numerical Ensemble (NE)	**7.31 [14]**	**8.54 [15]**	9.30 [15]	1.73 [15]	1.73 [11]	1.53 [15]
Gaussian Synthetic (GS)	8.19 [6]	9.14 [10]	**9.26 [14]**	**1.44 [2]**	1.63 [6]	**1.43 [11]**
NE + GS	8.15 [14]	9.02 [14]	9.33 [15]	1.50 [14]	**1.61 [14]**	**1.43 [15]**

C Cost Benefit Analysis

We conducted a preliminary cost benefit analysis (Supplementary Figure 1). In the FastSurfer internal cohort most of the MAE reduction is captured within the first 5 repetitions (22 of a possible 68 GPU-days), after which returns diminish; the HBN cohort improves more steadily, so the compute-optimal ensemble size is dataset-dependent.

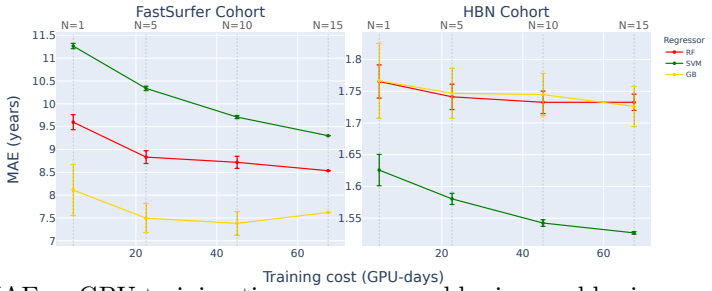

Fig. 1: MAE vs GPU training time across ensemble sizes and brain age regressors.

D Mean & Median Dice Figures

(a) Mean Sørensen-Dice variability (b) Median Sørensen-Dice variability

Fig. 2: Comparison of Sørensen-Dice variability (U_S) across FastSurfer and FreeSurfer ROIs on CoRR data (32 subjects). △ indicates regions where FastSurfer is less variable, and ▽ where it is more variable than FreeSurfer.

E Data Augmentation with MCA Ensembling

To confirm the benefit is not specific to seed-based perturbation, we repeated the experiment with an MCA ensemble at epoch 30, a training-loss minimum (fully-trained MCA models are computationally prohibitive). Supplementary Figure 3 reproduces the two main patterns of the seed ensemble: the same dataset dependence, with MCA ensembling matched or exceeded synthetic augmentation on the wide-age FastSurfer cohort, while synthetic reached lower MAE on narrow-age HBN, and, more monotonic improvement with ensemble size than synthetic augmentation across every model and dataset. This indicates the benefit arises from training-time numerical variability itself rather than the perturbation mechanism, though results at epoch 30 are indicative rather than checkpoint-matched, and the combined strategy is less consistent than for the seed ensemble.

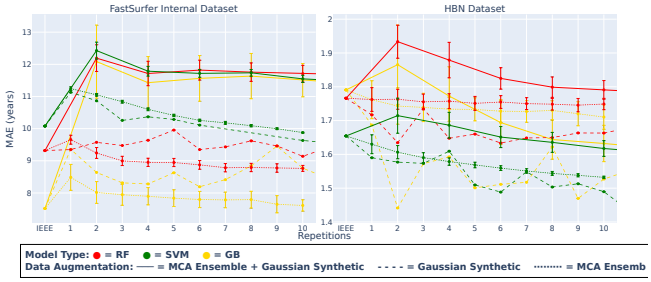

Fig. 3: Comparison of brain age regression performance for RF, SVM, and GB models across the internal FastSurfer and HBN datasets with MCA ensemble.