1. *Experimenting on Optimal Number of Atlases for BEaST*

For consistency, we have used 4 atlases for all the competing methods BEaST, SPECTRE, and MONSTR. However, it is possible to improve performance of BEaST using additional atlases. Here, we compare the performance of BEaST and MONSTR with an increasing number of atlases.

Ten subjects were randomly chosen from the `ADNI-29` dataset to serve as atlases, and these were then used to generate masks for the remaining 19 subjects. BEaST was run with 4 to 10 atlases, also randomly chosen from the atlas set of 10. As BEaST uses left-right flipped images, it effectively is using 8 to 20 atlases in this scenario. MONSTR was also run with 4 (as proposed in the paper) and 10 atlases. The following boxplot shows the Dice coefficients for BEaST with 4 to 10 atlases and MONSTR with 4 and 10 atlases.

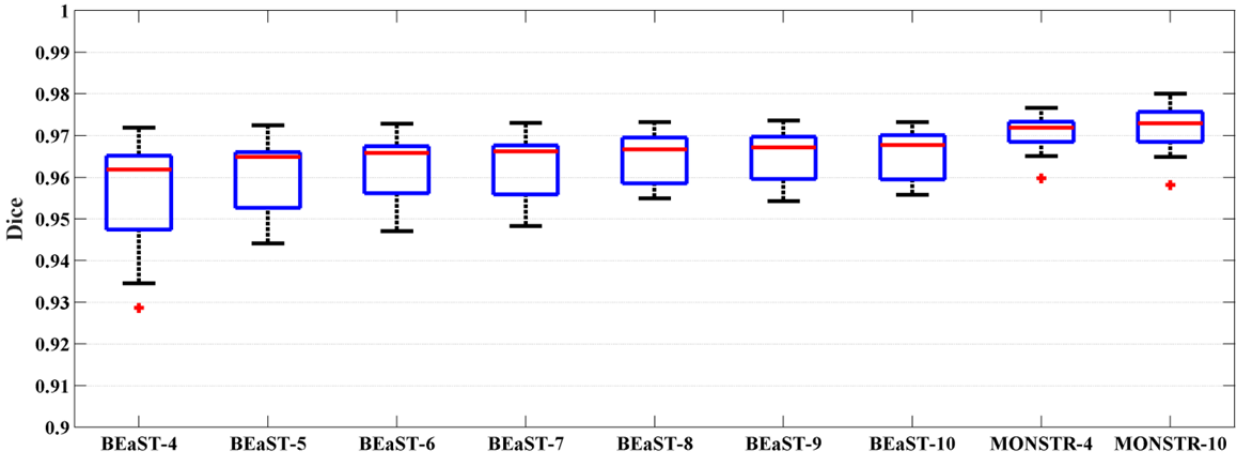

Fig. S1: Dice coefficients of BEaST and MONSTR with 4 to 10 atlases are shown. Both MONSTR results have significantly higher Dice than BEaST with 4 to 10 atlases.

There is no significant difference ($p>0.25$) between BEaST with 8, 9, and 10 atlases, while all of them are significantly different ($p<0.01$) from 5 to 7 atlases. However, the median Dice improvements are small (median Dice 0.9649 for BEaST-5 and 0.9677 for BEaST-10). Using 5 to 10 atlases also yields significant improvement ($p<0.01$) over 4 atlases, the improvement being in third decimal place (median Dice is 0.9619 for BEaST-4).

Using MONSTR with 4 or 10 atlases does not change Dice significantly (median Dices 0.9719 for MONSTR-4 vs 0.9730 for MONSTR-10, $p=0.17$). However both MONSTR-4 and MONSTR-10 have significantly higher Dice ($p=0.0001$) than BEaST-10 or less. Therefore even with 20 atlases, MONSTR outperforms BEaST with only 4 atlases.

2. Is single contrast MONSTR better than other T_1 based methods?

We compared MONSTR with only T_1 against other methods on both the NAMIC-20 and the TBI-19 datasets. For both datasets, MONSTR using only the T_1-weighted contrast yielded a significantly higher Dice compared to the 4 competing methods. The following table shows the comparison. An asterisk indicates significantly better Dice coefficients ($p<0.001$ Wilcoxon signed rank test) for MONSTR with T_1+T_2 and MONSTR with only T_1 over the 4 competing methods. MONSTR with T_1+T_2 outperforms MONSTR with T_1 ($p<0.01$) for both datasets as well.

Table S1: Comparison between four competing T_1 based methods and MONSTR using only the T_1-weighted image. MONSTR with only T_1 outperforms all other methods on both NAMIC-20 and TBI-19 datasets.

	BEaST	**SPECTRE**	**OptiBET**	**ROBEX**	**MONSTR (T_1 only)**	**MONSTR (T_1+T_2)**
NAMIC-20	0.9713	0.9427	0.9583	0.9558	**0.9783***	**0.9833***
TBI-19	0.9425	0.9316	0.9602	0.9563	**0.9674***	**0.9811***

3. Dependence of MONSTR on Atlas Choice

To determine if MONSTR is affected by our particular choice of 4 atlases, we performed a 2-fold cross validation on the TBI-19 dataset. Four subjects were randomly chosen from the 19 subjects, and brain masks were generated for the remaining 15. This was performed 10 times. Figure S2 shows Dice coefficients of 15 subjects for 10 permutations of different atlases. For any two pairs, none of the Dice coefficients exhibited statistical differences ($p>0.10$). Therefore even with TBI atlases having various degrees of lesions, MONSTR demonstrated robustness to the choice of atlases. Note that for this dataset, OptiBET had the highest median Dice (0.9602) amongst the other competing methods, but this value was still significantly lower ($p<0.0001$) than any of the permutations.

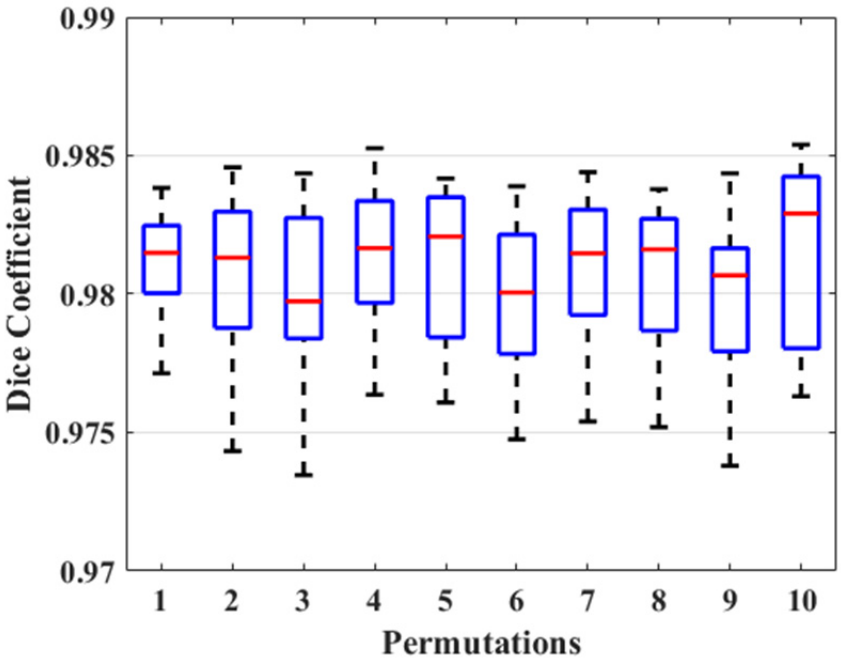

Fig. S2: A two fold cross validation on `TBI-19` dataset is performed, where 4 atlases are randomly chosen from 19 subjects. None of the Dice coefficients have any significant difference with one another.