Supplementary Material

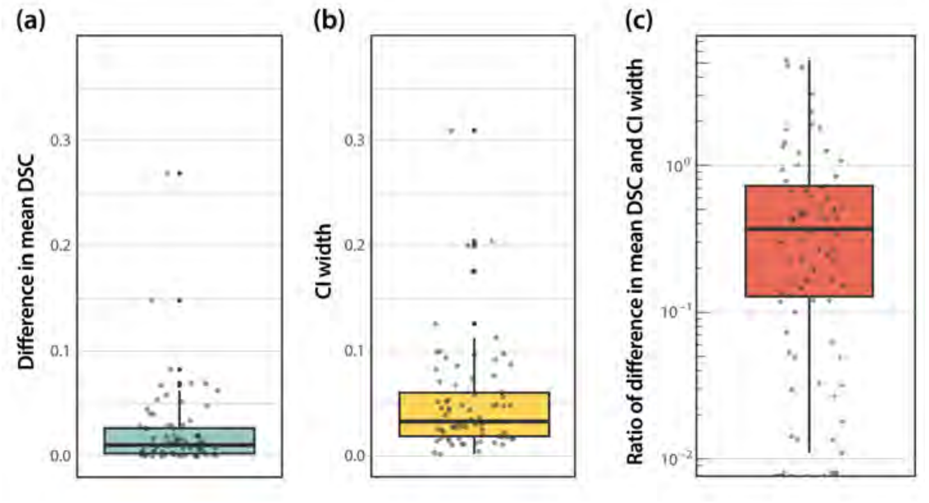

Fig. 5: **The width of confidence intervals (CIs) is mostly larger than the performance gain.** Boxplots describing: (a) the difference in mean Dice Similarity Coefficient (DSC) between the two top-ranked methods within a paper, (b) the CI width of the first-ranked method of each paper, and (c) the ratio of difference in mean DSC for the two top-ranked methods methods within a paper and CI width of the first-ranked method of each paper.