Supplementary Figures

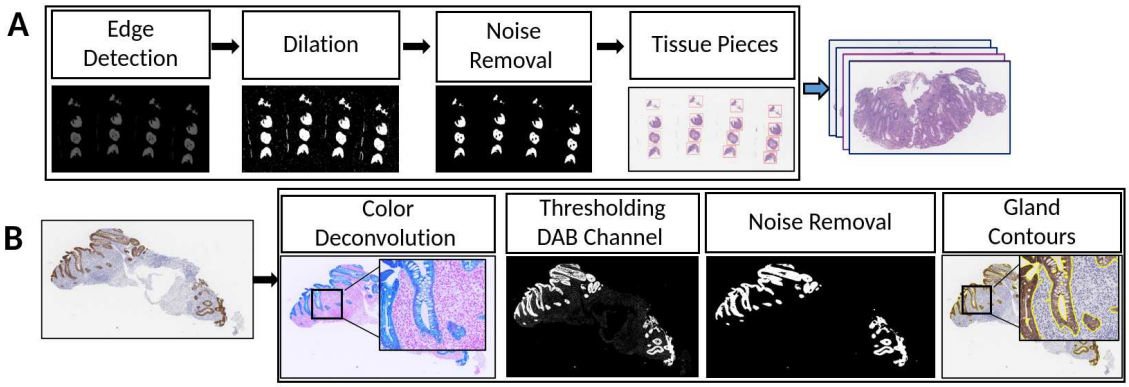

Figure SF1. Modules inside the automated pipeline for the generation of ground truth gland outlines in IHC images: A. Extraction of tissue pieces block B. IHC Gland mask generation block of Figure 1.

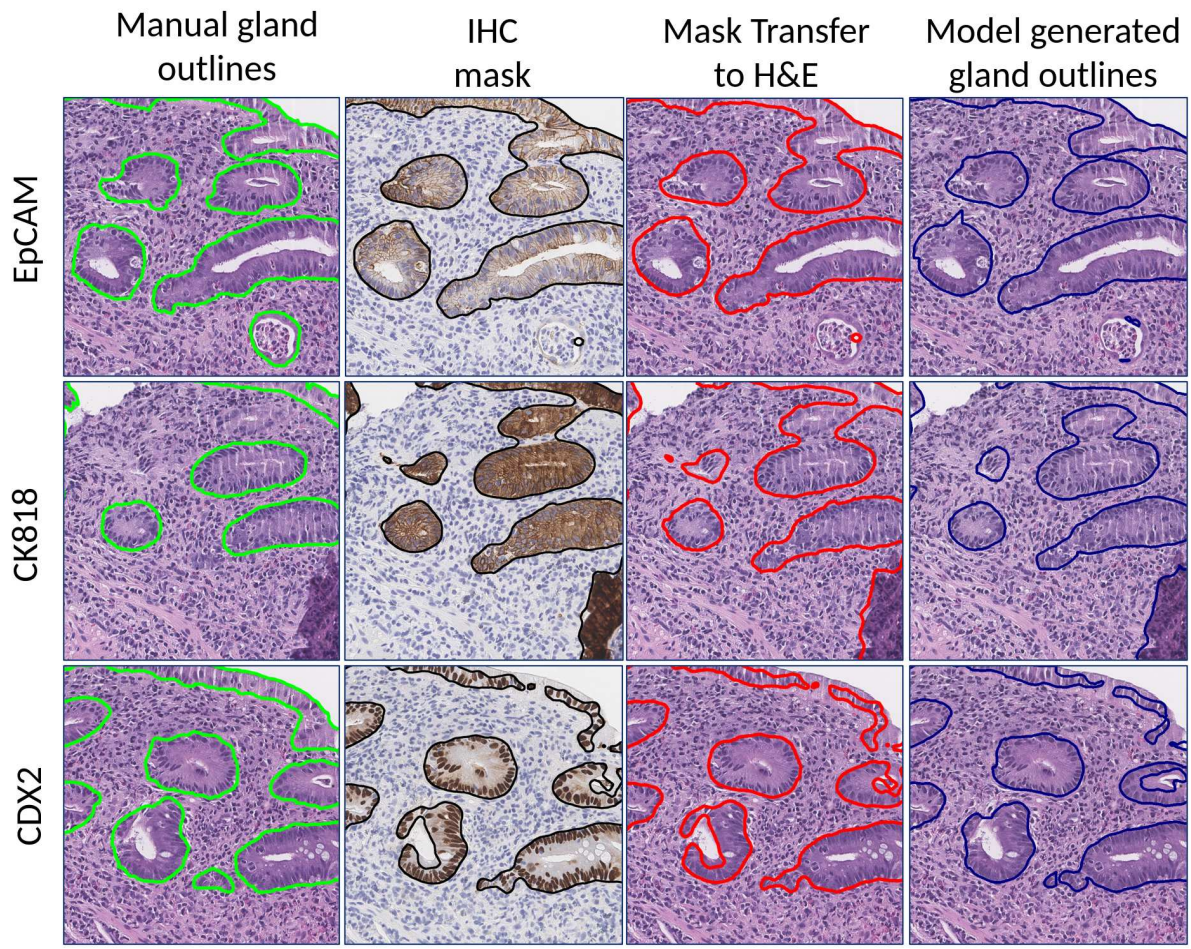

Figure SF2. Gland segmentation results. Gland outlines are generated manually (column 1), based on IHC staining (column 2), by transfer of the IHC mask to the H&E image (column 3) and by the algorithm trained on gland outlines in H&E images that are obtained by the automated transfer of IHC masks. To generate gland outlines by IHC, slides were stained with CDX2, CK8/18, or EpCAM.

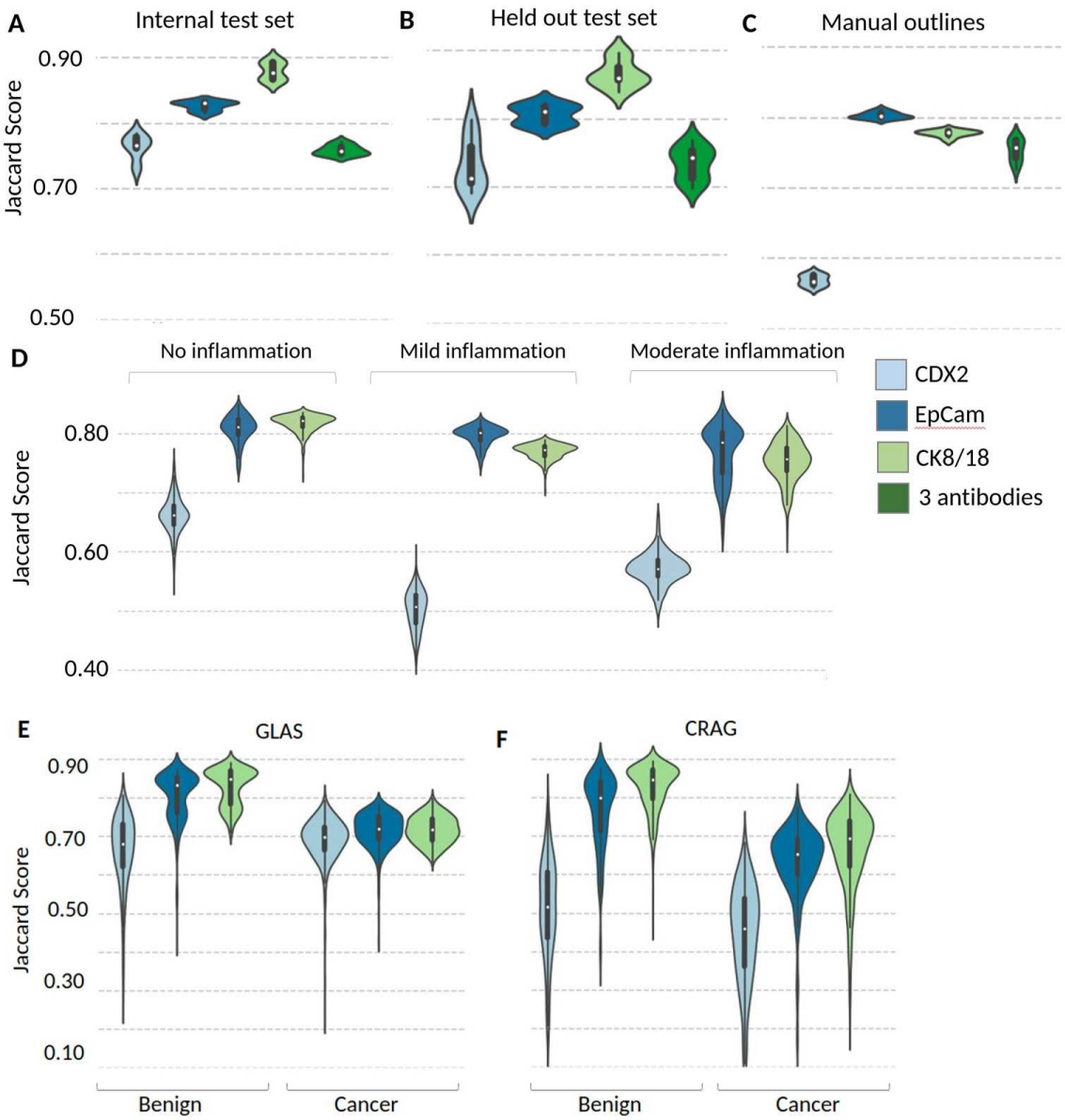

Figure SF3. Performance of gland segmentation models. A. Internal test set. Jaccard scores of models trained on ground truth masks obtained from CDX2, EpCAM, CK8/18, or on all three antibodies are shown. One tissue piece from each case is held out for training and used to evaluate the performance of the model. **B. Held out test set.** Models are applied to a held-out case and are not used for training. **C. Manual annotations.** Computer-generated gland outlines are compared to outlines obtained through manual annotations. **D. Effect of disease activity on model performance.** The amount of inflammation in each tissue piece was scored by a pathologist and classified as no, mild or moderate inflammation. **E. External GLAS dataset.** Models were tested separately on benign and cancer cases in the GLAS testing cohort. **F. External CRAG dataset.** Models were tested separately on benign and cancer cases in the CRAG testing cohort.

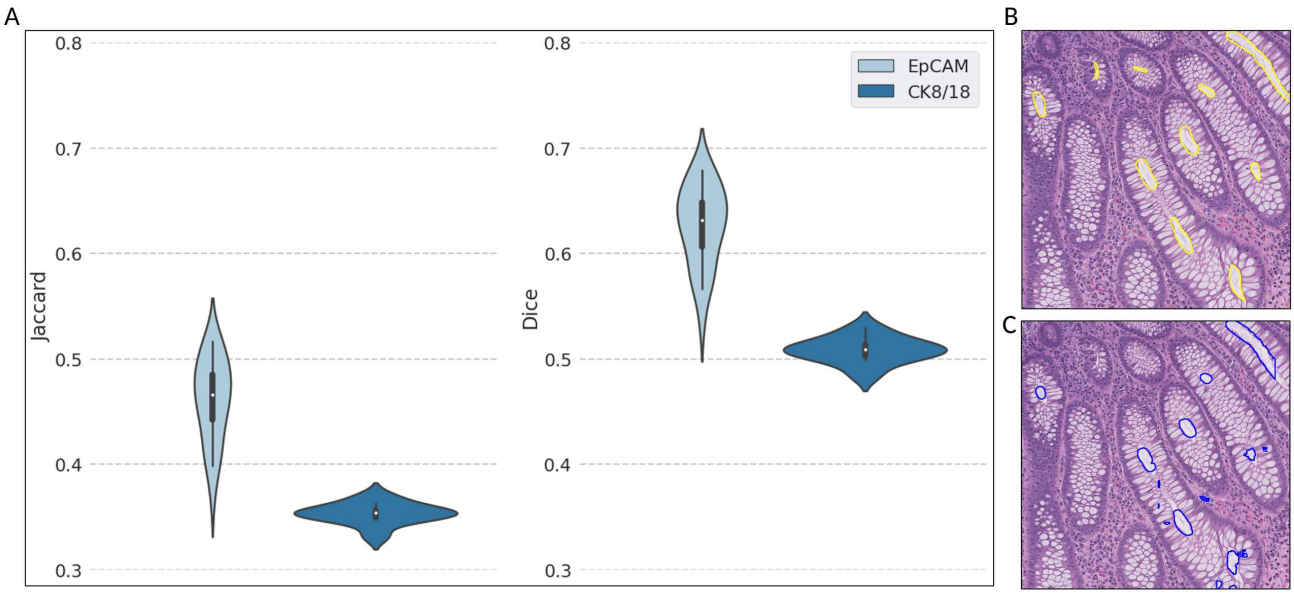

Figure SF4. Lumen Segmentation. A. Performance of models (Jaccard and Dice) trained on lumen ground truths obtained from EpCAM and CK8/18. B. Manual annotation of gland lumens. C. Representative tile with an output of lumen segmentation generated by the model trained on EpCAM ground truth data.

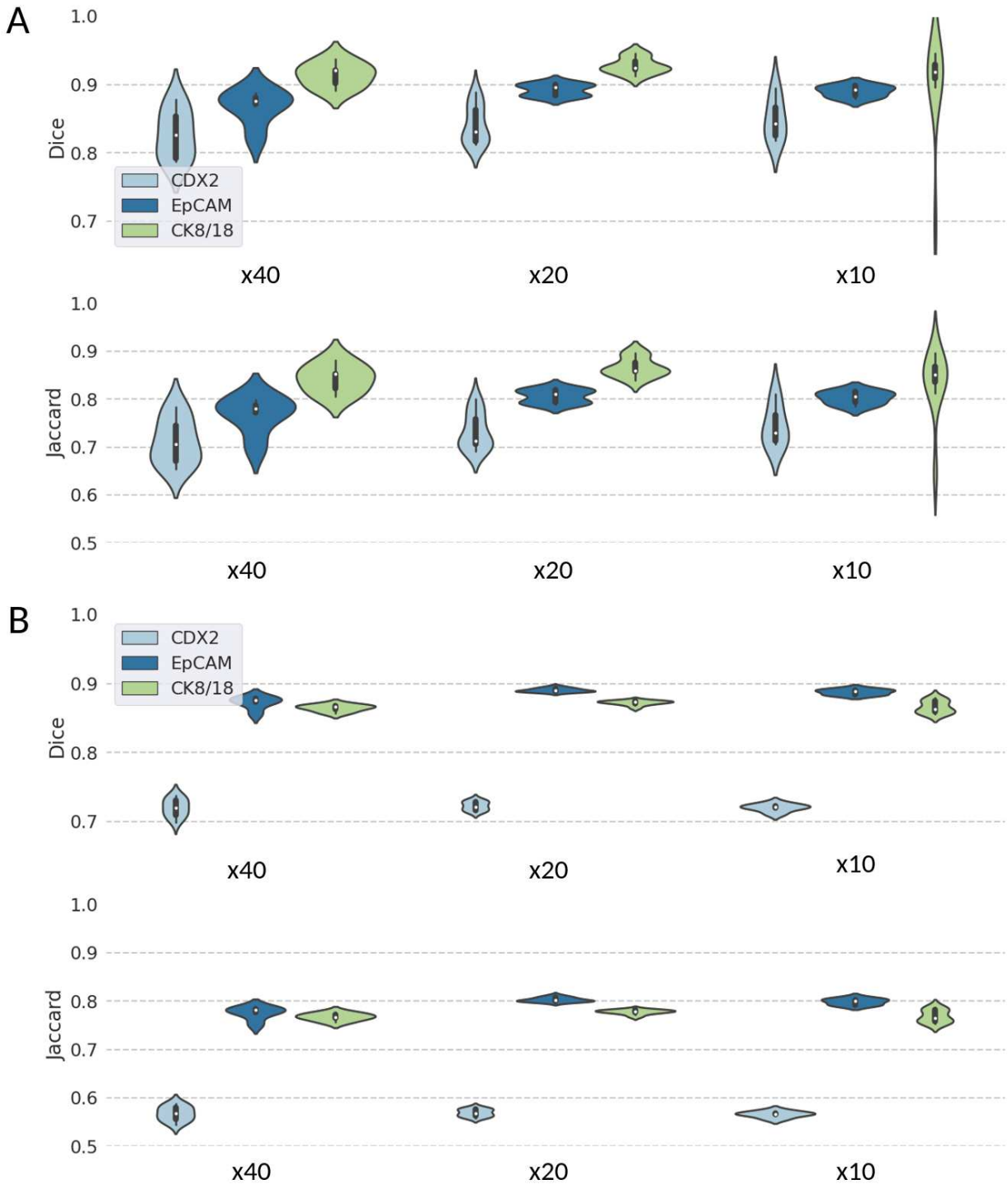

Figure SF5. Effect of pixel size on gland segmentation performance. Gland segmentation performance of models trained on different magnification patches (x-axis). Patch size to the model is fixed at 256x256 for all these experiments. Jaccard and Dice are used to compare performance metrics. A. Held out patient test set. B. Manual segmentation. Based on these findings, x20 is the optimal magnification for gland segmentation. Similar experiments demonstrate that the optimal magnification for lumen segmentation input data is x5.

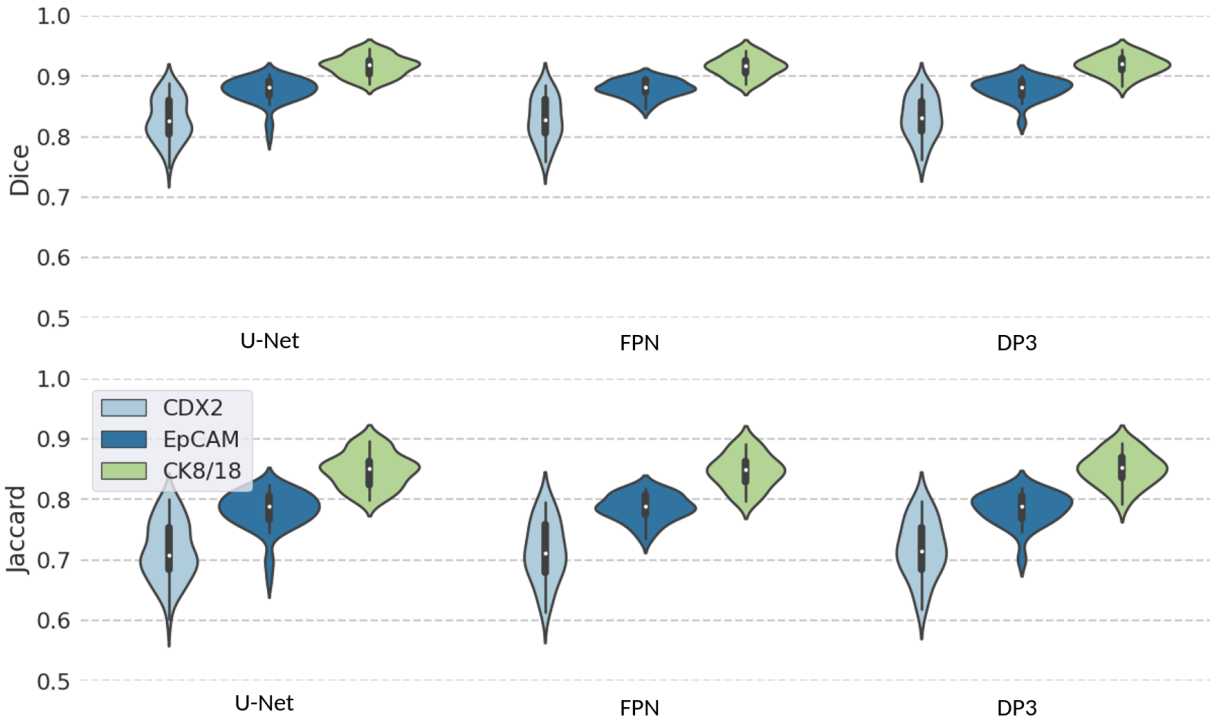

Figure SF6. Selection of model framework for the gland segmentation task. U-Net[30], Feature pyramid network[31] and DeepLabv3[32] are trained on gland outlines generated by CDX2, EpCAM, and CK8/18. All results are from one held-out patient using x20 magnification as the patch size. Models trained with DeepLabv3 (DP3 in the figure) slightly outperform in U-Net model when mean performance is compared but is not statistically significant.

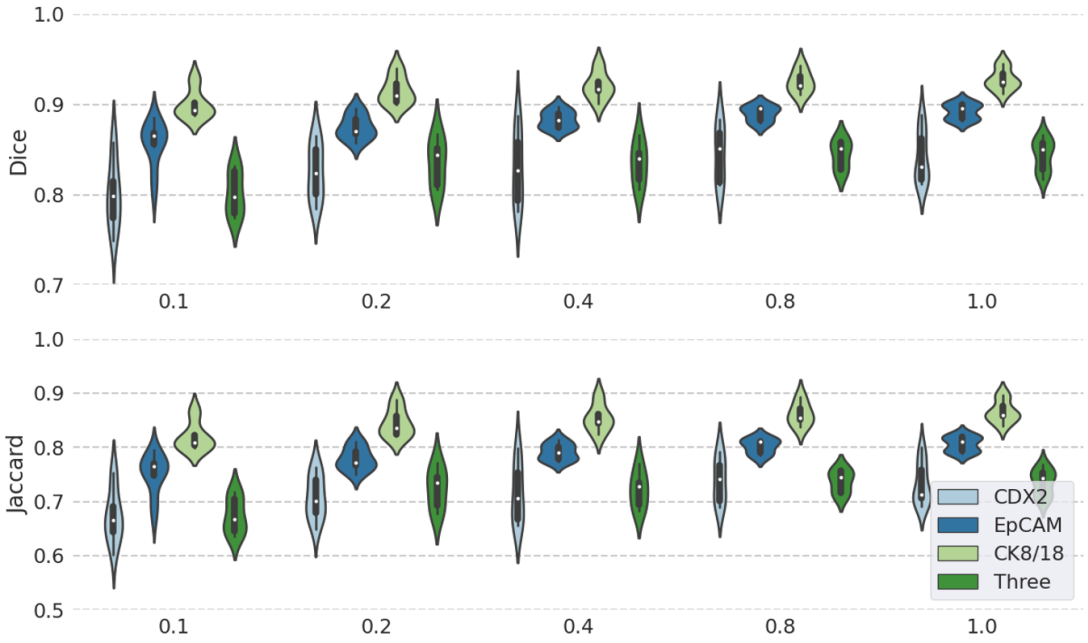

Figure SF7. Amount of training data affects model performance. The amount of input data is indicated on the x-axis. Numbers depict the percentage of data used from each patient for training. Tiles are x20 magnification. A held-out test set is used to evaluate the model performance. With an increase in dataset size, we see an increase in the mean Dice/Jaccard score and a decrease in the variation on the internal held-out test set.

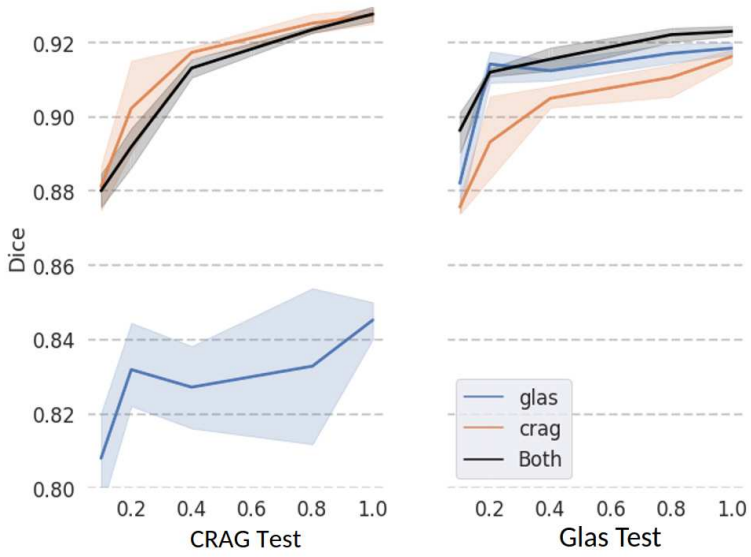

Figure SF8. Performance variation when using data from one vs both external datasets for mixed training. Performance comparison on external testing datasets when either one of the external datasets vs both are used for mixed training. This result indicates that the model always attempts to learn site/dataset-specific features; however, as the training dataset's variability increases, the model learns more generalized features.

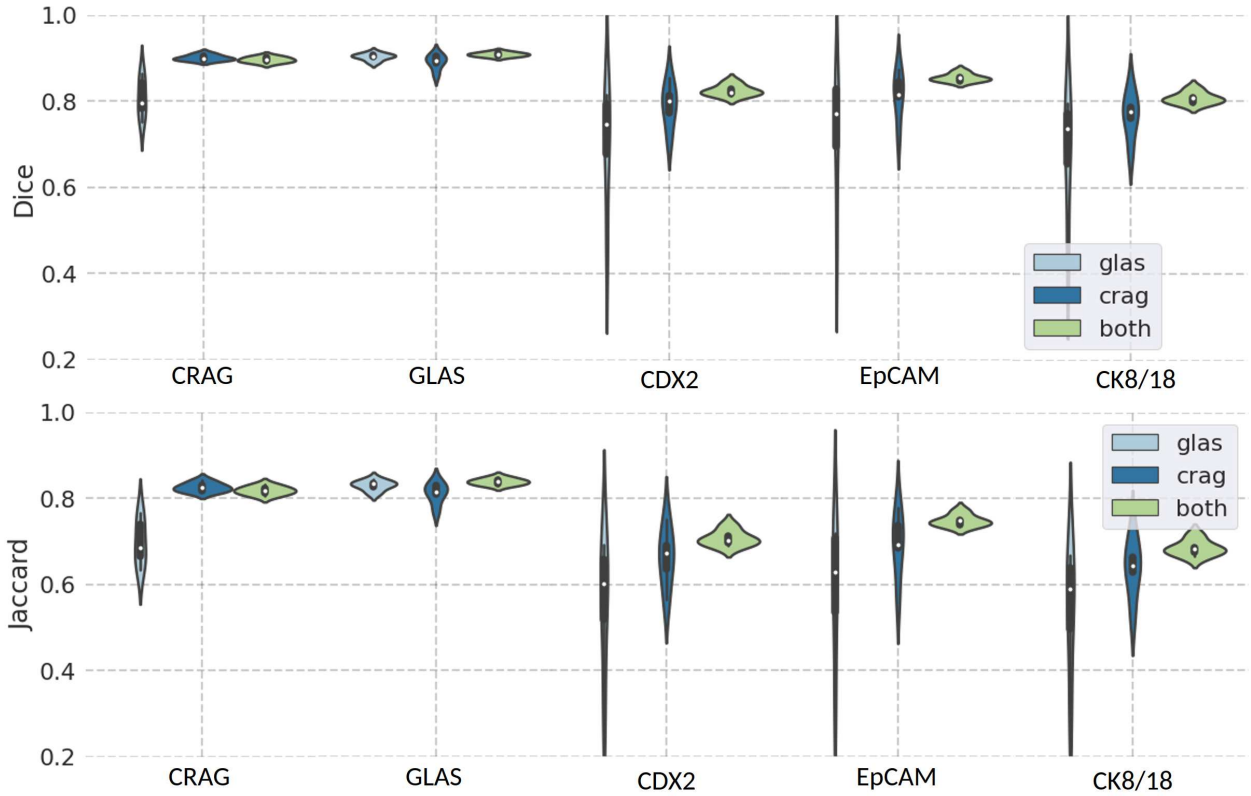

Figure SF9. Models trained on External Data. Performance comparison when models trained on Public Data(CRAG[15] and GLAS[16]) are used out of box for internal IBD data. All performance metrics are done for manually annotated tissue pieces. Three violin plots show a comparison between models trained on using only one of the external dataset vs both. Given the considerable performance decrease, it is clear that mutlisource data is better and does not degrade in performance caused due to domain shift.