APPENDIX A
CALIBRATION QUALITY (WHOLE VOLUME RESULTS)

Table I compares the calibration quality of different models calculated on the whole volume. The baselines and ensembles (M=50) trained with CE loss are compared with those that were trained with Dice loss and those that were calibrated with MC dropout.

TABLE I. CALIBRATION QUALITY FOR BASELINES TRAINED WITH DICE LOSS ($\mathcal{L}_{DSC}$) ARE COMPARED WITH THOSE THAT TRAINED WITH CROSS-ENTROPY ($\mathcal{L}_{CE}$) AND THOSE THAT WERE CALIBRATED WITH ENSEMBLING (M=50) AND MC DROPOUT. BOLD FACED INDICATES BEST RESULTS FOR EACH APPLICATION (MODEL) AND SHOWS THAT THE DIFFERENCES ARE STATISTICALLY SIGNIFICANT.

Organ (Model)	NLL (95% CI)	Brier (95% CI)	ECE% (95% CI))
Brain ($\mathcal{L}_{CE}$)	0.08 (0.01−0.38)	0.03 (0.01−0.14)	1.06 (0.12−5.69)
Brain (MCDO $\mathcal{L}_{CE}$)	0.16 (0.01−0.70)	0.07 (0.01−0.24)	2.45 (0.08−11.11)
Brain (EN $\mathcal{L}_{CE}$)	0.04 (0.01−0.16)	0.02 (0.00−0.09)	0.97 (0.29−2.50)
Brain ($\mathcal{L}_{DSC}$)	0.16 (0.03−0.56)	0.02 (0.00−0.07)	1.16 (0.17−3.32)
Brain (MCDO $\mathcal{L}_{DSC}$)	0.13 (0.02−0.63)	0.02 (0.00−0.11)	1.05 (0.14−4.81)
Brain (EN $\mathcal{L}_{DSC}$)	**0.03 (0.01−0.11)**	**0.01 (0.00−0.03)**	**0.49 (0.03−1.58)**
Heart ($\mathcal{L}_{CE}$)	0.04 (0.01−0.12)	0.02 (0.01−0.04)	0.53 (0.14−1.55)
Heart (MCDO $\mathcal{L}_{CE}$)	0.04 (0.01−0.11)	0.02 (0.01−0.04)	0.52 (0.11−5.02)
Heart (EN $\mathcal{L}_{CE}$)	**0.03 (0.01−0.06)**	**0.01 (0.01−0.03)**	0.51 (0.25−0.71)
Heart ($\mathcal{L}_{DSC}$)	0.07 (0.01−0.24)	0.02 (0.00−0.05)	1.10 (0.10−3.29)
Heart (MCDO $\mathcal{L}_{DSC}$)	0.04 (0.01−0.15)	0.34 (0.02−0.85)	47.39 (6.24−90.55)
Heart (EN $\mathcal{L}_{DSC}$)	0.04 (0.01−0.07)	0.01 (0.01−0.03)	**0.36 (0.07−1.33)**
Prostate ($\mathcal{L}_{CE}$)	0.08 (0.04−0.16)	0.04 (0.02−0.09)	2.15 (0.50−7.17)
Prostate (MCDO $\mathcal{L}_{CE}$)	0.11 (0.04−0.24)	0.06 (0.03−0.11)	1.82 (0.23−4.65)
Prostate (EN $\mathcal{L}_{CE}$)	0.07 (0.05−0.10)	0.03 (0.02−0.06)	2.62 (1.65−3.87)
Prostate ($\mathcal{L}_{DSC}$)	0.26 (0.10−0.58)	0.04 (0.02−0.08)	1.94 (0.97−4.12)
Prostate (MCDO $\mathcal{L}_{DSC}$)	0.17 (0.07−0.37)	0.04 (0.02−0.08)	1.79 (0.80−3.99)
Prostate (EN $\mathcal{L}_{DSC}$)	**0.05 (0.02−0.09)**	**0.02 (0.01−0.04)**	**0.65 (0.13−1.26)**

APPENDIX B
HAUSDORFF DISTANCE METRIC

Table II compares the segmentation performance of different models with 95^{th} Hausdorff distance (in mm). The baselines and ensembles (M=50) trained with CE loss are compared with those that were trained with Dice loss and those that were calibrated with MC dropout.

TABLE II. SEGMENTATION QUALITY OF BASELINES IN TERMS OF 95ˆTH HAUSDORFF DISTANCES IN MM. MODELS TRAINED WITH DICE LOSS ($\mathcal{L}_{DSC}$) ARE COMPARED WITH THOSE THAT TRAINED WITH CROSS-ENTROPY ($\mathcal{L}_{CE}$) AND THOSE THAT WERE CALIBRATED WITH ENSEMBLING (M=50) AND MC DROPOUT. FOR BRAIN APPLICATION SEGMENTS, I, II, AND III CORRESPOND TO NON-ENHANCING TUMOR, EDEMA, AND ENHANCING TUMOR, RESPECTIVELY. FOR HEART APPLICATION SEGMENTS, I, II, AND III CORRESPOND TO THE RIGHT VENTRICLE, THE MYOCARDIUM, AND THE LEFT VENTRICLE, RESPECTIVELY. FOR PROSTATE APPLICATION SEGMENT I CORRESPONDS TO THE PROSTATE GLAND. BOLD FACED INDICATES BEST RESULTS FOR EACH APPLICATION (MODEL) AND SHOWS THAT THE DIFFERENCES ARE STATISTICALLY SIGNIFICANT.

Organ (Model)	Segment I*	Segment II*	Segment III*
Brain ($\mathcal{L}_{CE}$)	59.05 (5.39−107.69)	52.93 (7.28−83.41)	56.85 (3.00−101.92)
Brain (MCDO $\mathcal{L}_{CE}$)	63.95 (5.83−110.00)	60.71 (10.20−86.34)	62.45 (3.61−103.77)
Brain (EN $\mathcal{L}_{CE}$)	36.53 (3.46−95.12)	35.63 (3.46−80.14)	41.24 (2.24−100.92)
Brain ($\mathcal{L}_{DSC}$)	40.05 (4.00−102.69)	39.22 (3.00−80.06)	40.49 (2.00−99.70)
Brain (MCDO $\mathcal{L}_{DSC}$)	44.19 (4.12−107.24)	44.83 (3.61−81.61)	45.80 (2.24−100.84)
Brain (EN $\mathcal{L}_{DSC}$)	**22.55 (2.56−93.82)**	**24.75 (2.24−71.03)**	**30.81 (2.00−94.88)**
Heart ($\mathcal{L}_{CE}$)	26.30 (7.21−135.30)	20.55 (4.00−136.38)	24.04 (2.00−154.95)
Heart (MCDO $\mathcal{L}_{CE}$)	30.60 (7.21−169.28)	22.15 (4.00−146.43)	24.64 (2.00−164.93)
Heart (EN $\mathcal{L}_{CE}$)	14.42 (5.66−30.00)	7.37 (4.00−20.69)	6.43 (2.00−20.79)
Heart ($\mathcal{L}_{DSC}$)	15.18 (2.00−79.97)	10.47 (2.00−88.23)	13.69 (2.00−126.91)
Heart (MCDO $\mathcal{L}_{DSC}$)	15.53 (2.00−79.51)	9.51 (2.00−64.18)	11.85 (2.00−104.34)
Heart (EN $\mathcal{L}_{DSC}$)	**9.50 (2.00−26.91)**	**5.90 (2.00−14.70)**	**6.30 (2.00−20.98)**
Prostate ($\mathcal{L}_{CE}$)	11.67 (5.00−25.07)	–	–
Prostate (MCDO $\mathcal{L}_{CE}$)	14.54 (6.18−28.40)	–	–
Prostate (EN $\mathcal{L}_{CE}$)	6.62 (3.54−19.91)	–	–
Prostate ($\mathcal{L}_{DSC}$)	8.22 (3.64−20.59)	–	–
Prostate (MCDO $\mathcal{L}_{DSC}$)	9.84 (4.12−23.34)	–	–
Prostate (EN $\mathcal{L}_{DSC}$)	**5.66 (3.16−18.71)**	–	–

APPENDIX C
QUANTITATIVE RESULTS

Figure 1 visually compares the baselines trained with cross entropy, $\mathcal{L}_{CE}$, Dice loss, $\mathcal{L}_{DSC}$, with those calibrated with MC dropout and ensembling over the three segmentation tasks. For each prediction map, a reliability diagram over the whole volume is provided. In rendering the reliability diagrams only bins with greater than 1000 samples are shown. Figures 2 and 3 show example cases for heart and prostate applications.

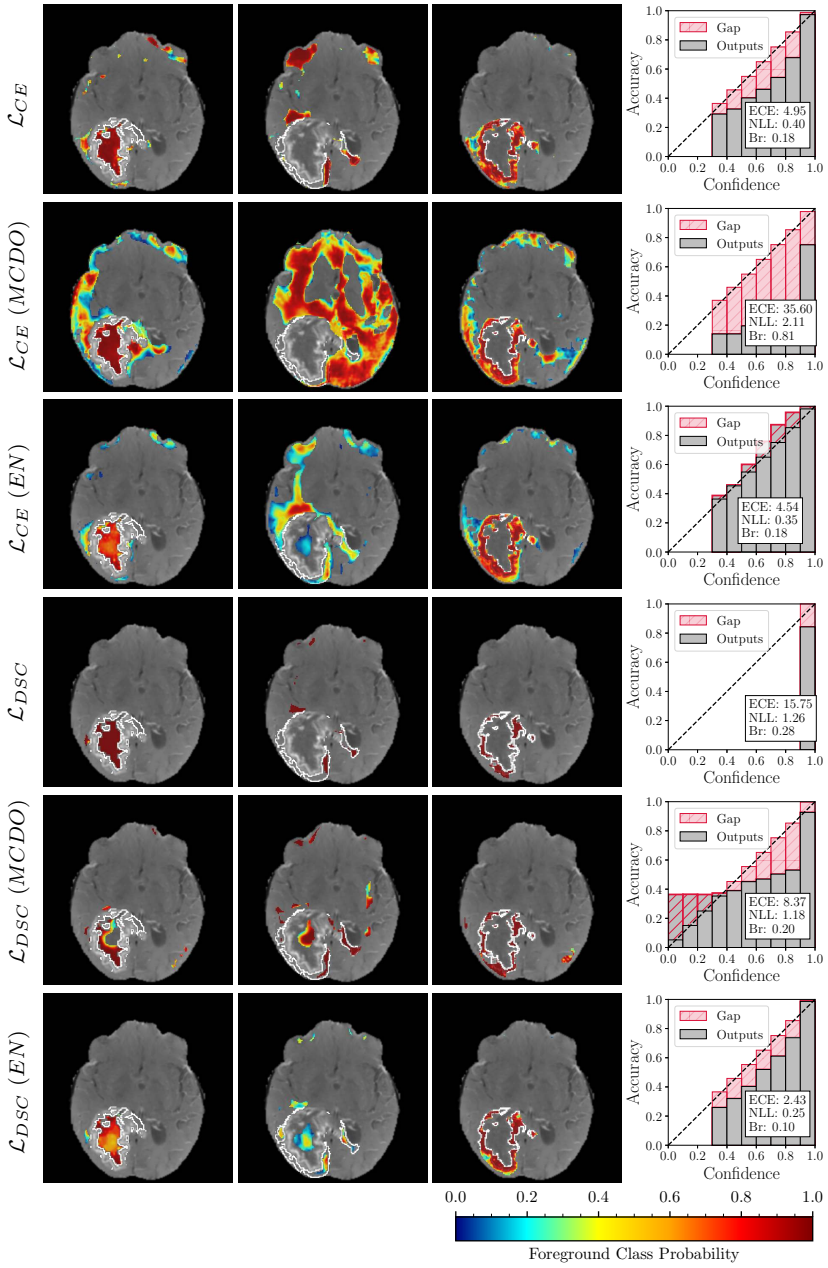

Fig. 1. Examples of uncertainty estimation quality for brain tumor segmentation using different methods. MRI images are overlaid with class probabilities, and reliability diagrams (together with ECE%, NLL, and Brier score) are given for that specific volume. In the reliability diagrams only the bins with greater than 1000 samples are shown.

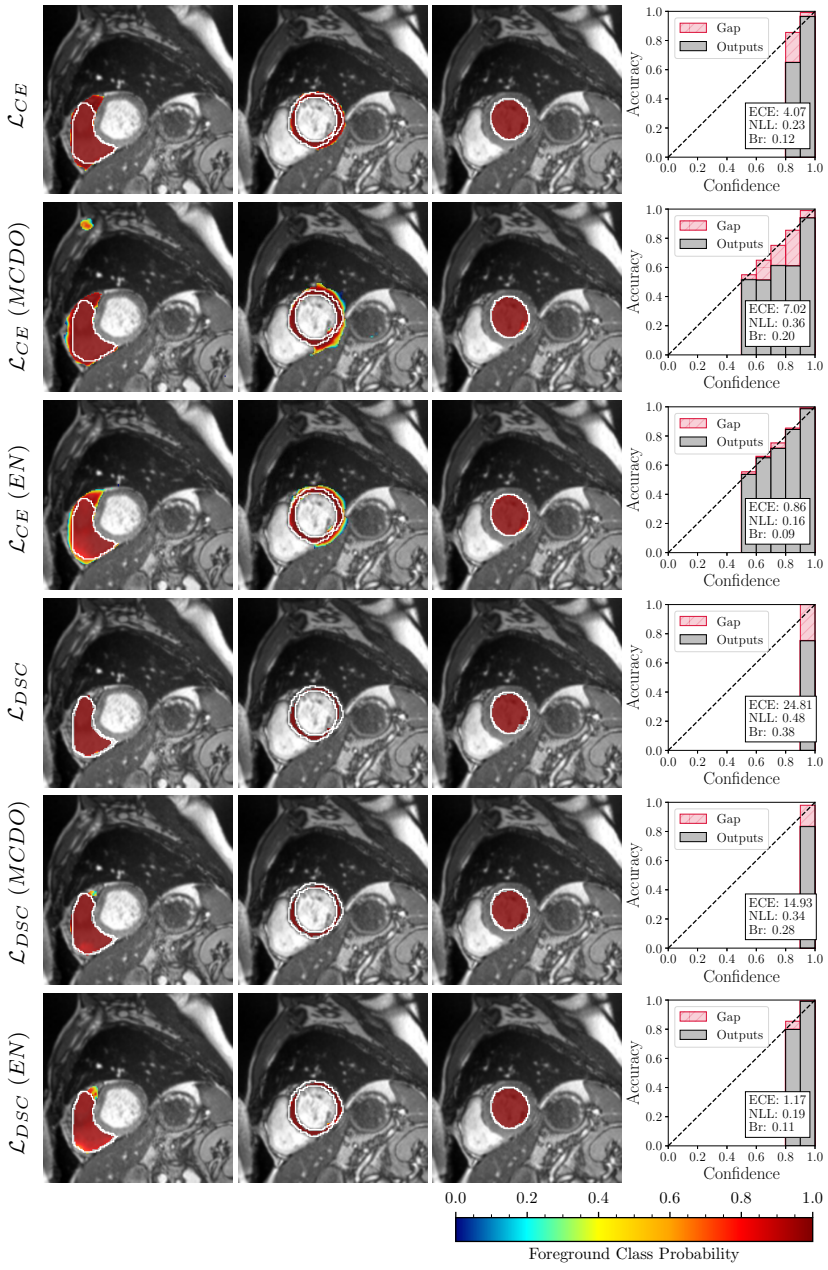

Fig. 2. Examples of uncertainty estimation quality for heart segmentation using different methods. MRI images are overlaid with class probabilities, and reliability diagrams (together with ECE%, NLL, and Brier score) are given for that specific volume. In the reliability diagrams only the bins with greater than 1000 samples are shown.

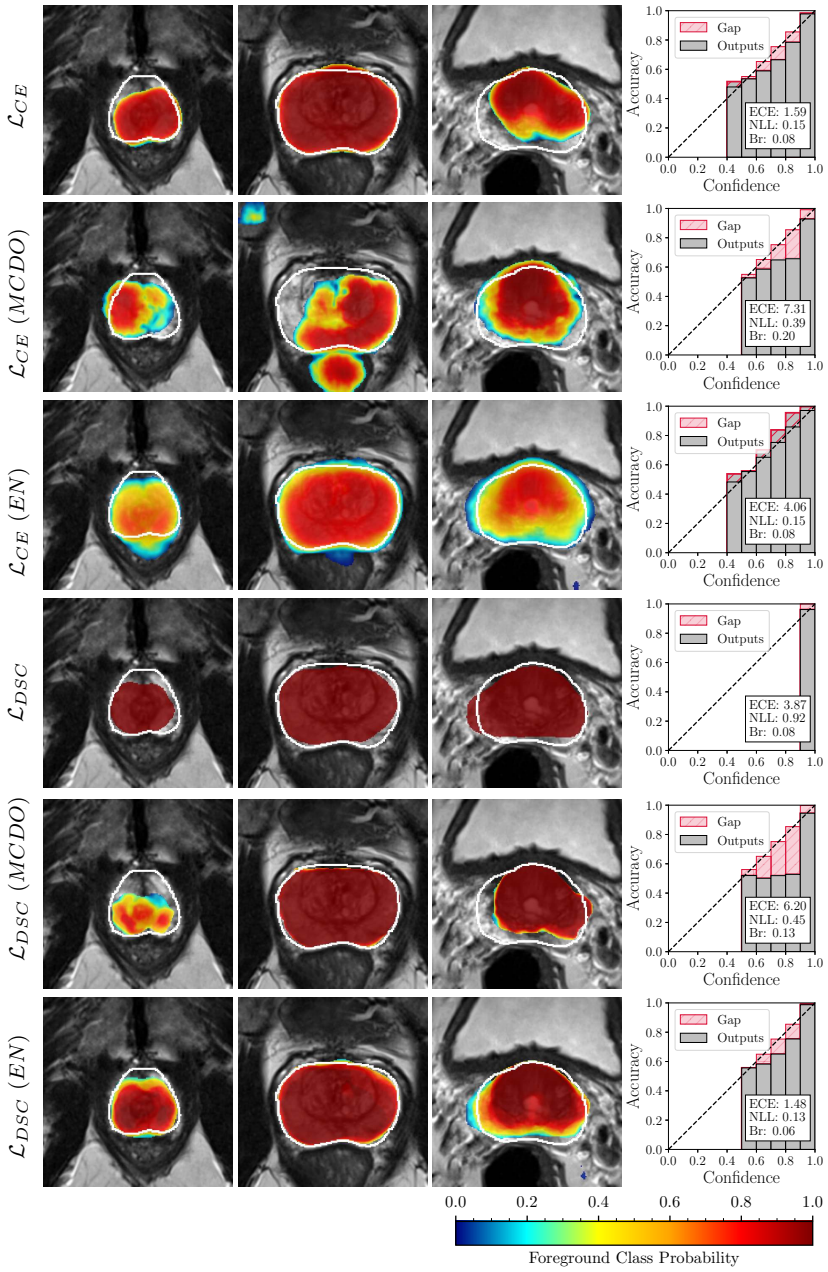

Fig. 3. Examples of uncertainty estimation quality for prostate segmentation using different methods. MRI images are overlaid with class probabilities, and reliability diagrams (together with ECE%, NLL, and Brier score) are given for that specific volume. In the reliability diagrams only the bins with greater than 1000 samples are shown.

APPENDIX D
NUMBER OF MODELS IN ENSEMBLE

The three graphs in Figure 4 show the quantitative improvement in the calibration as a function of the number of models in the ensemble, M, with 0.95 CI. The images in Figure 5 and 6 qualitatively illustrate this calibration improvement by ensembling for models trained with Dice loss and CE loss, respectively.

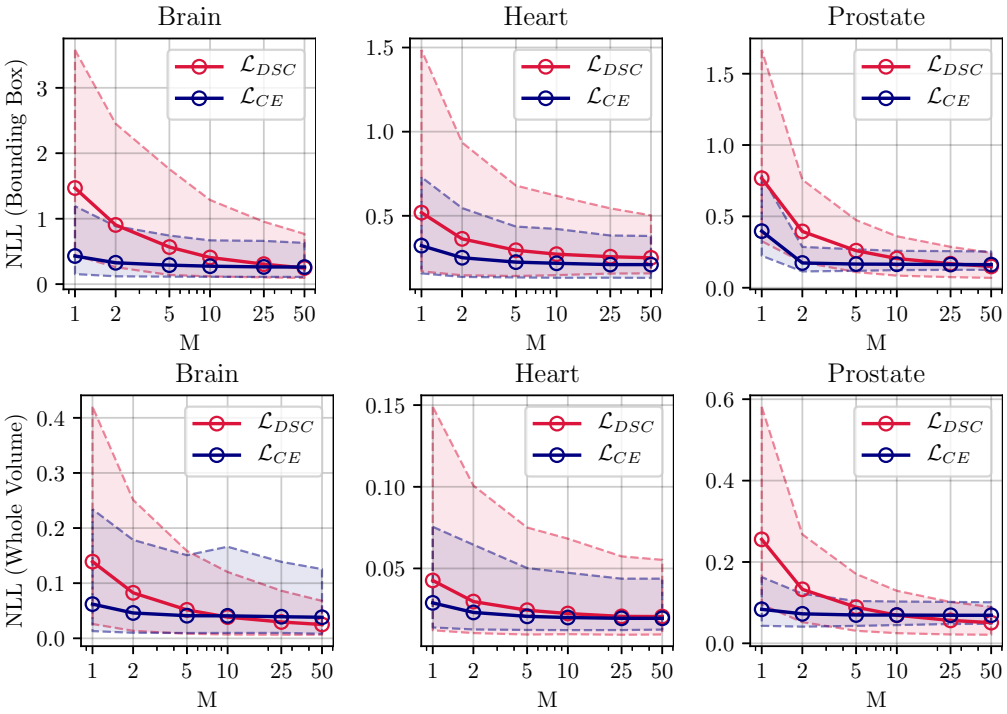

Fig. 4. Improvements in calibration as a function of the number of models in the ensemble. Calibration quality in terms of NLL as number of models M increases for prostate, heart, and brain tumor segmentation.

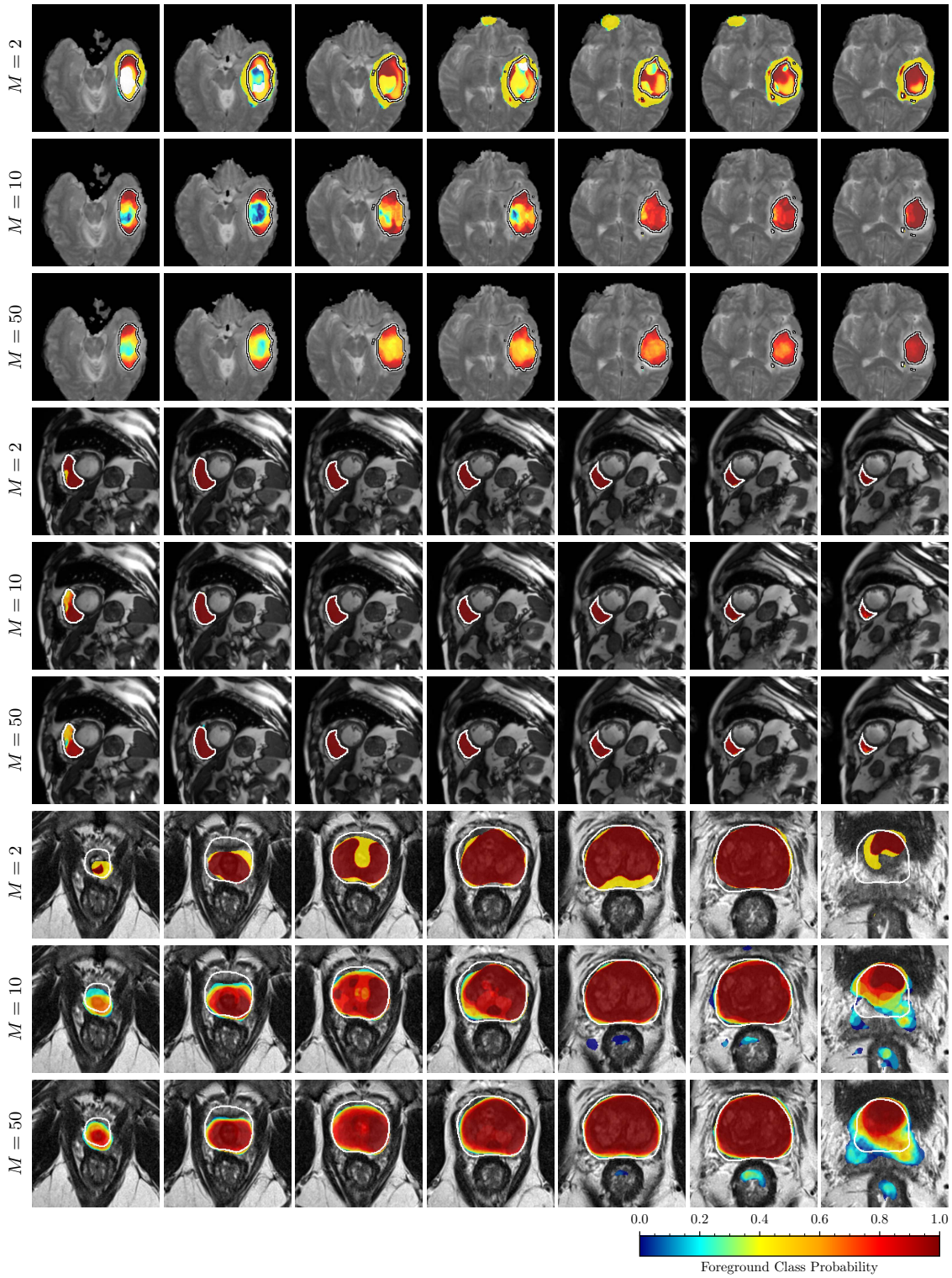

Fig. 5. Qualitative examples of improvements in calibration and segmentation as a function of the number of models M in the ensemble of models trained with Dice loss. The overlaid probability maps show the results of inference for an ensemble of size M=2, M-20, and M=50. White line shows the ground truth boundary of the structures.

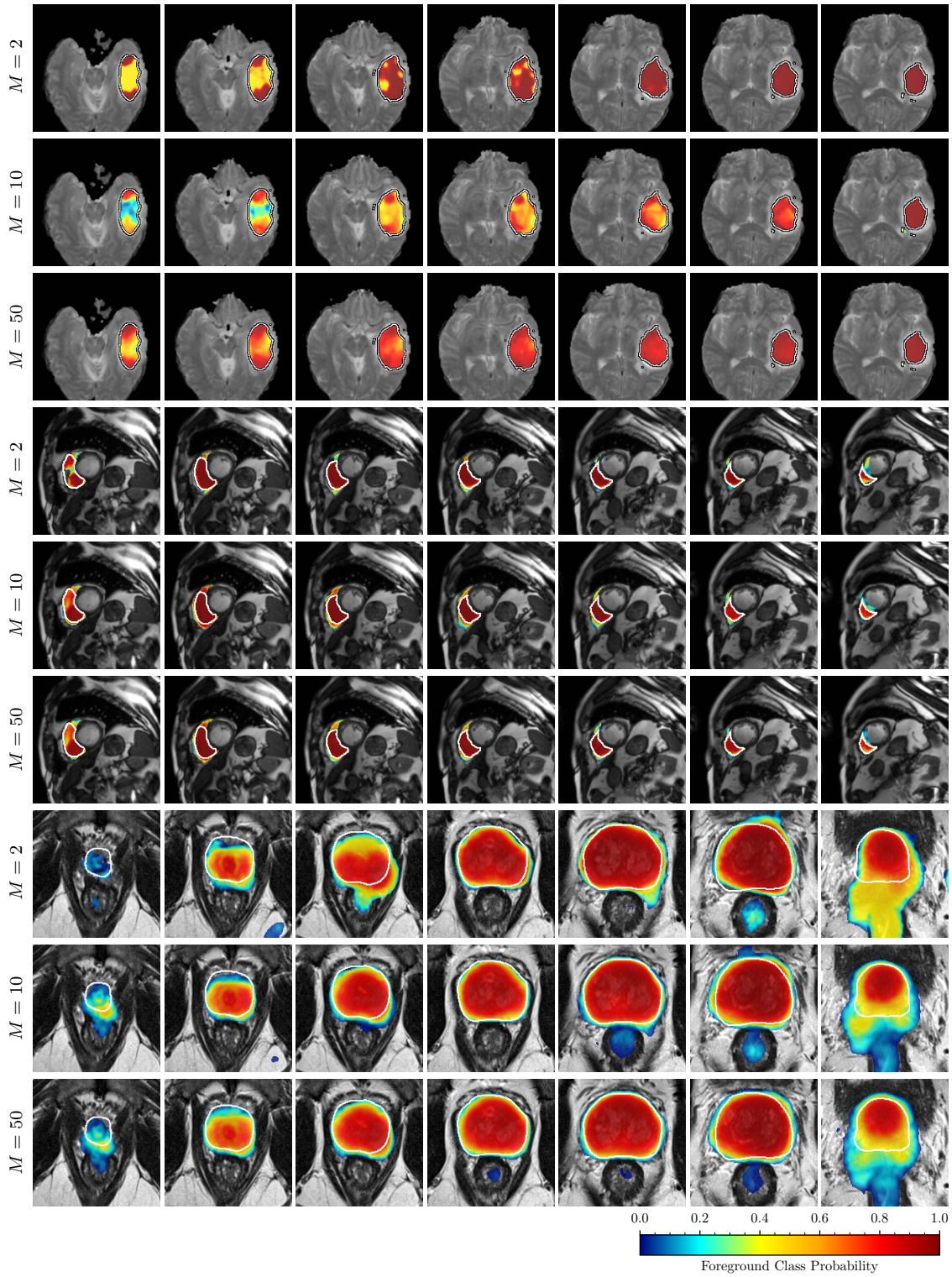

Fig. 6. Qualitative examples of improvements in calibration and segmentation as a function of the number of models M in the ensemble of models trained with cross entropy loss. The overlaid probability maps show the results of inference for an ensemble of size M=2, M-20, and M=50. White line shows the ground truth boundary of the structures.

APPENDIX E
SEGMENT-LEVEL PREDICTIVE UNCERTAINTY

Figure 7 shows examples of predictions with different levels of confidence for brain and heart application.

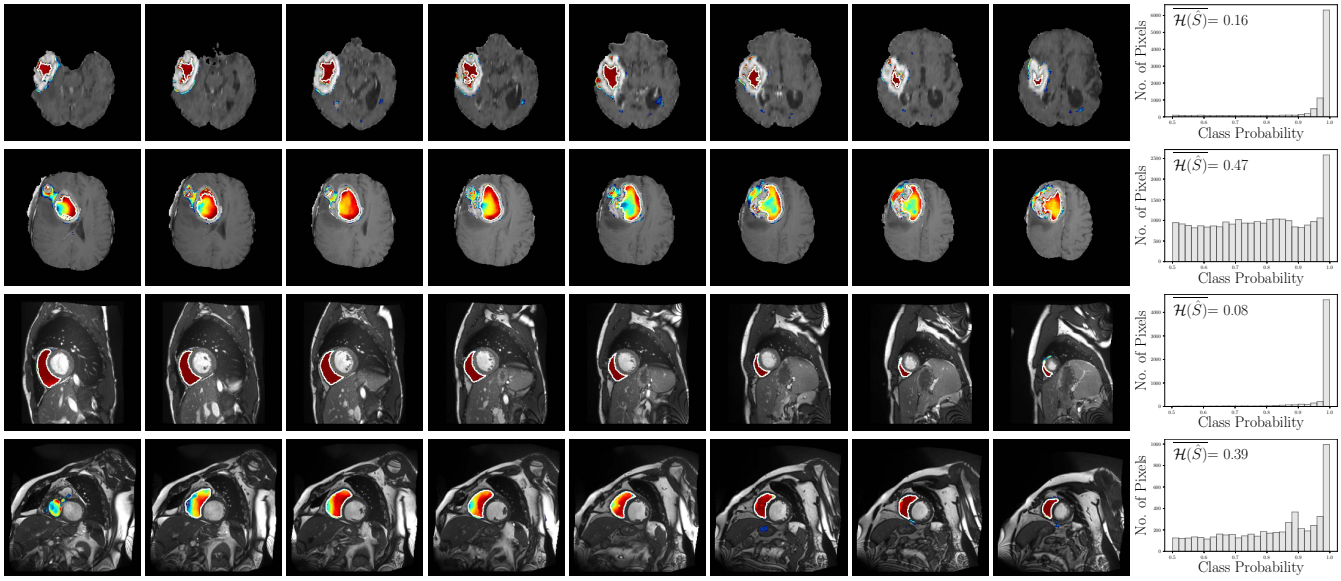

Fig. 7. Examples from tumor segmentation and right ventricle segmentation tasks.

Figure 8 provides scatter plots of Dice coefficient vs. the proposed segment-level predictive uncertainty metric, $\overline{\mathcal{H}(\hat{\mathcal{S}})}$ (Equation 7), for models trained with CE loss and calibrated with ensembling (M=50). Similar to the charts in Figure 3, a reverse correlation holds between the average entropy over the segmented foreground and the Dice scores.

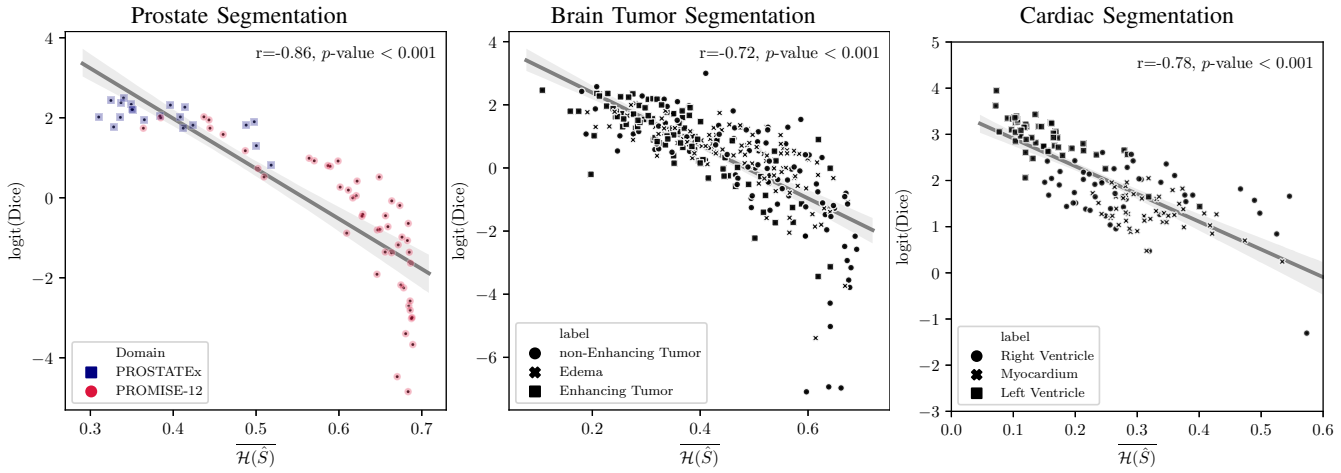

Fig. 8. Segment-level predictive uncertainty estimation for ensembles of models trained with cross entropy loss. Scatter plots and linear regression between Dice coefficient and average of entropy over the predicted segment.

APPENDIX F
3D FCN

To test the generalizability of the proposed methods on 3D CNNs, we run limited experiments on prostate gland segmentation. Prostate images were resampled to resolution of $1 \times 1 \times 1$ mm and all the 3D volumes were then cropped at the center to create images of size $112 \times 112 \times 112$ pixels as the input size of the FCN. Image intensities were normalized to be within the range of [0,1]. We constructed a 3D FCN with same number of kernels and depth as the 2D network described in Section V. 2D Convolutional, max pooling and upsampling layers were replaced 3D layers. Training parameters were kept the same except that the model was trained 10 times with cross entropy and 10 times with Dice loss, and 10 times with dropout layers for MC dropout experiments. Table III compares the calibration quality and segmentation performance of baselines and ensembles (M=50) trained with CE loss with those that were trained with Dice loss and those that were calibrated with MC dropout.

TABLE III. OBSERVED AVERAGE CALIBRATION QUALITY AND SEGMENTATION PERFORMANCE FOR 3D FCNS TRAINED FOR PROSTATE GLAND SEGMENTATION.

	Calibration Quality (bounding boxes)			Segmentation Performance	
Model	NLL	Brier	ECE%	Dice Score	95^{th} HD
$\mathcal{L}_{CE}$	0.24	0.15	8.25	0.84	9.25
MCDO $\mathcal{L}_{CE}$	0.26	0.16	6.84	0.81	12.12
EN $\mathcal{L}_{CE}$	0.21	0.12	9.23	0.87	5.62
$\mathcal{L}_{DSC}$	0.40	0.13	5.32	0.89	6.06
MCDO $\mathcal{L}_{DSC}$	0.41	0.11	5.27	0.88	7.47
EN $\mathcal{L}_{DSC}$	0.19	0.08	2.82	0.90	4.61