Supplementary Material

Table S1. Maximum Calibration Error

	ET	TC	WT	Avg
CE	0.26 ± 0.19	0.28 ± 0.18	0.23 ± 0.12	0.25 ± 0.12
CE + Ts	0.30 ± 0.16	0.29 ± 0.16	0.25 ± 0.10	0.28 ± 0.10
CE + L1-ACE	0.24 ± 0.19	0.26 ± 0.18	0.20 ± 0.15	0.23 ± 0.14
CE + L1-ACE + Ts	**0.23 ± 0.15**	**0.25 ± 0.16**	0.20 ± 0.12	**0.23 ± 0.11**
Dice	0.59 ± 0.15	0.61 ± 0.15	0.55 ± 0.13	0.58 ± 0.11
Dice + Ts	0.56 ± 0.15	0.58 ± 0.15	0.51 ± 0.14	0.55 ± 0.11
Dice + L1-ACE	0.29 ± 0.18	0.30 ± 0.17	0.23 ± 0.14	0.27 ± 0.13
Dice + L1-ACE + Ts	**0.23 ± 0.17**	**0.25 ± 0.16**	0.21 ± 0.11	**0.23 ± 0.12**
Dice-CE	0.46 ± 0.17	0.46 ± 0.17	0.36 ± 0.17	0.43 ± 0.13
Dice-CE + Ts	0.26 ± 0.19	0.32 ± 0.19	**0.19 ± 0.12**	0.26 ± 0.13
Dice-CE + L1-ACE	0.29 ± 0.21	0.30 ± 0.20	0.21 ± 0.14	0.27 ± 0.15
Dice-CE + L1-ACE + Ts	0.25 ± 0.19	0.27 ± 0.18	0.20 ± 0.11	0.24 ± 0.13

Table S2. Expected Calibration Error (×10^{-4})

	ET	TC	WT	Avg
CE	4.21 ± 7.49	5.66 ± 10.5	**9.71 ± 11.3**	6.53 ± 7.53
CE + Ts	7.77 ± 7.16	10.2 ± 10.0	19.1 ± 9.45	12.4 ± 6.93
CE + L1-ACE	4.00 ± 4.75	6.16 ± 9.41	12.0 ± 12.9	7.38 ± 6.94
CE + L1-ACE + Ts	21.1 ± 4.69	27.2 ± 6.65	62.0 ± 11.1	36.8 ± 5.69
Dice	5.09 ± 7.67	7.48 ± 16.5	15.5 ± 17.8	9.34 ± 11.1
Dice + Ts	4.95 ± 7.63	7.36 ± 16.5	15.0 ± 17.8	9.10 ± 11.1
Dice + L1-ACE	3.47 ± 7.56	4.50 ± 8.45	10.1 ± 12.8	**6.03 ± 6.84**
Dice + L1-ACE + Ts	**3.27 ± 7.17**	**4.41 ± 8.00**	12.0 ± 10.9	6.54 ± 5.84
Dice-CE	4.40 ± 7.57	5.70 ± 12.2	11.6 ± 14.4	7.21 ± 8.62
Dice-CE + Ts	3.84 ± 7.43	5.20 ± 11.8	10.2 ± 13.4	6.41 ± 8.11
Dice-CE + L1-ACE	3.91 ± 8.09	6.04 ± 14.6	10.3 ± 14.8	6.75 ± 9.84
Dice-CE + L1-ACE + Ts	4.23 ± 7.68	6.56 ± 14.0	13.9 ± 12.6	8.23 ± 8.83

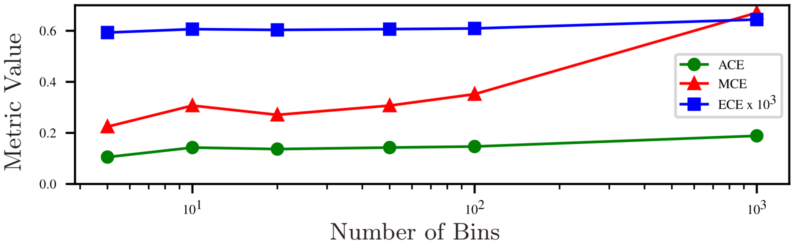

Fig. S1. Comparison of calibration metrics using different number of bins when evaluated on the DSC + mL1-ACE loss model. Both ACE and ECE are very stable with respect to number of bins (5, 10, 20, 50, 100, 1000). MCE shows a larger increase from 100 to 1000 bins, this is due to the presence of empty bins, a likely occurrence when using such a high number of bins.