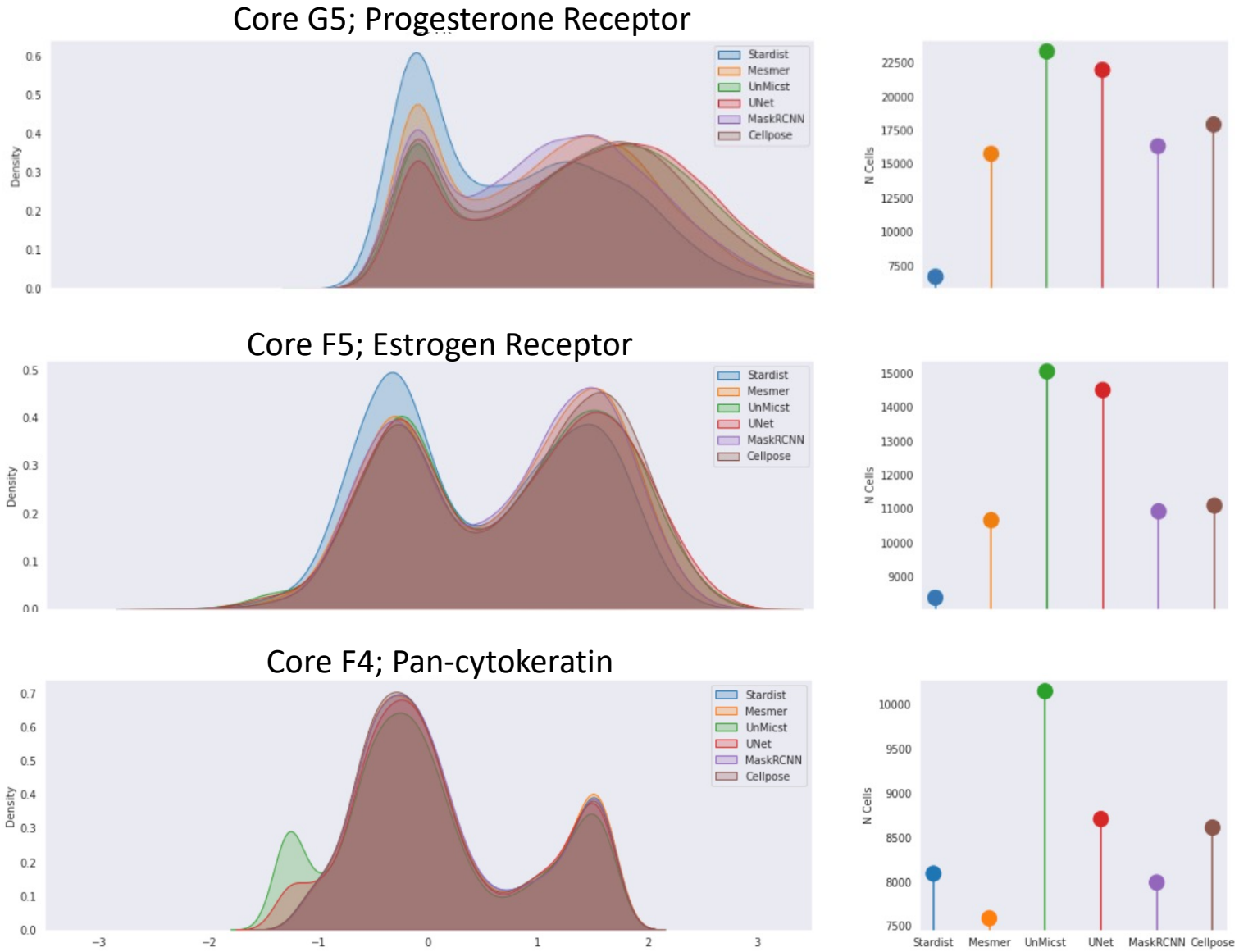
Figure 1A. Examples of feature level discrepancy between segmentation methods and individual segmentation masks



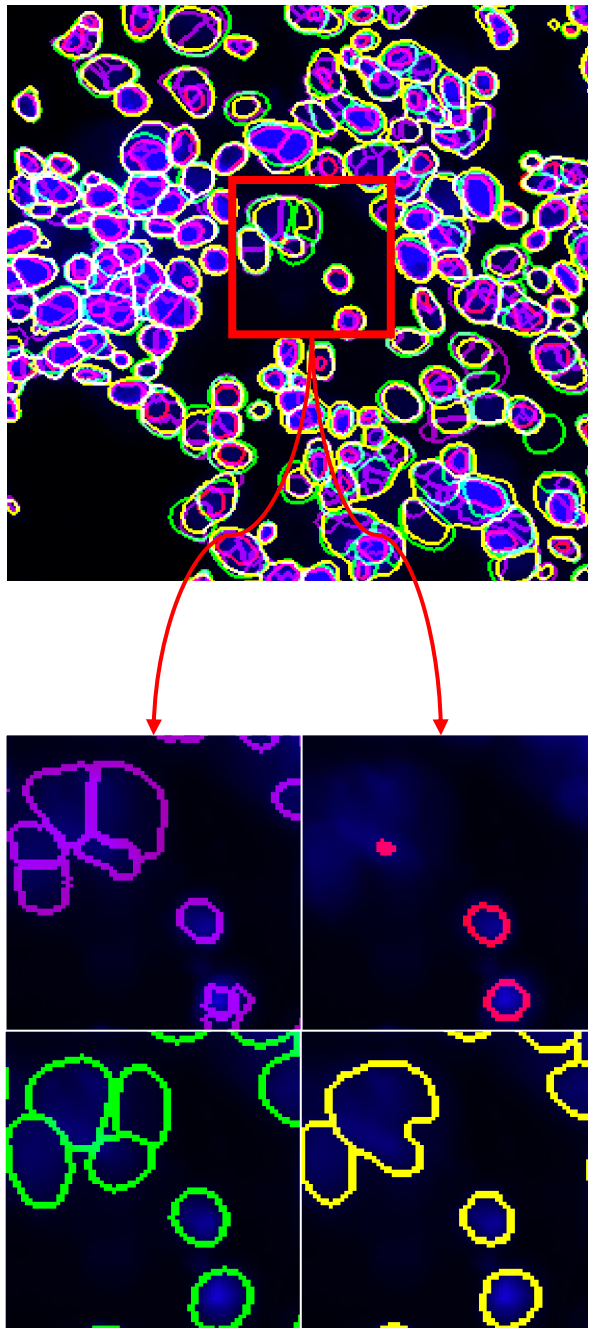
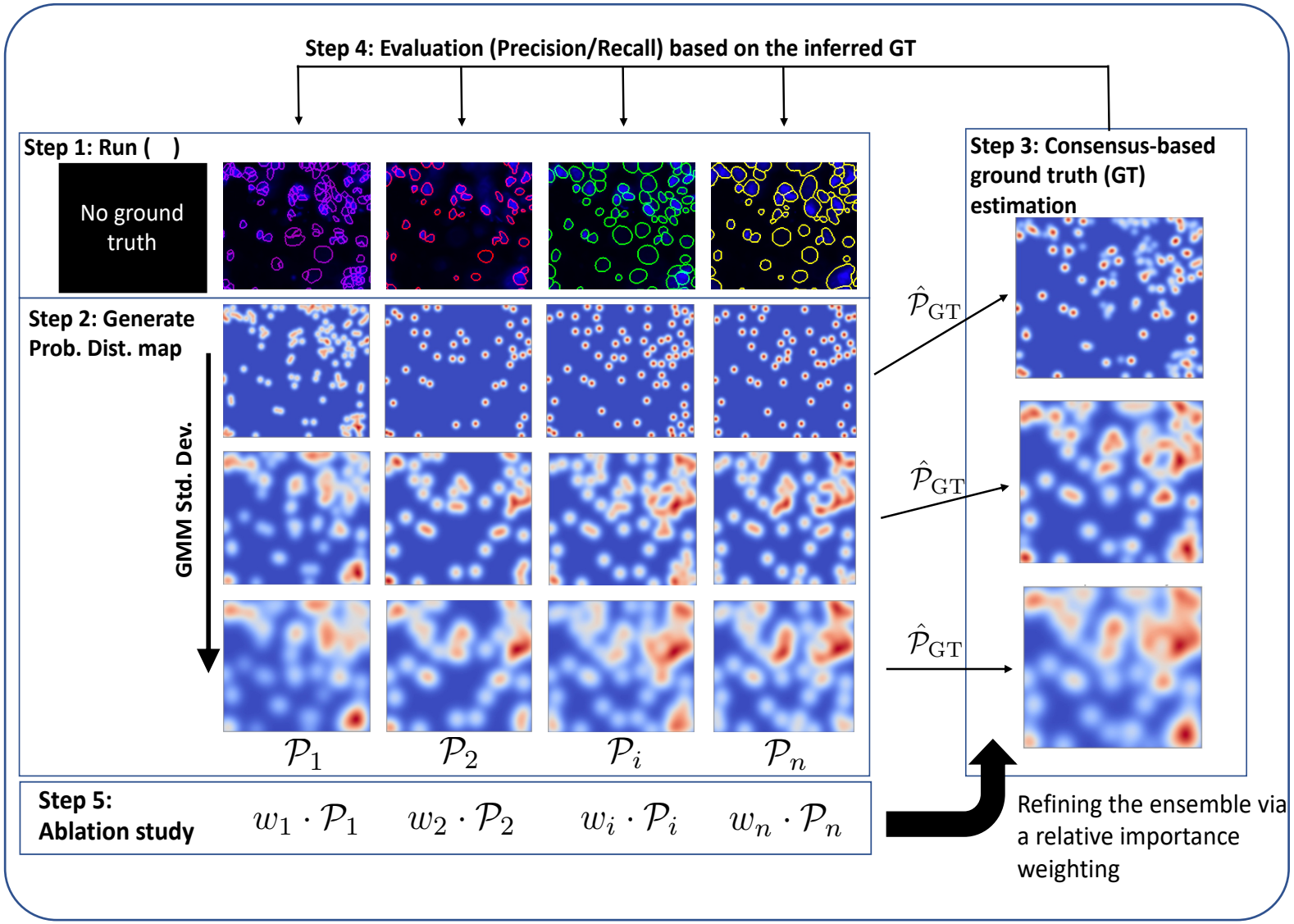


Figure 1B. An overview of consensus-based ground truth estimation and refinement



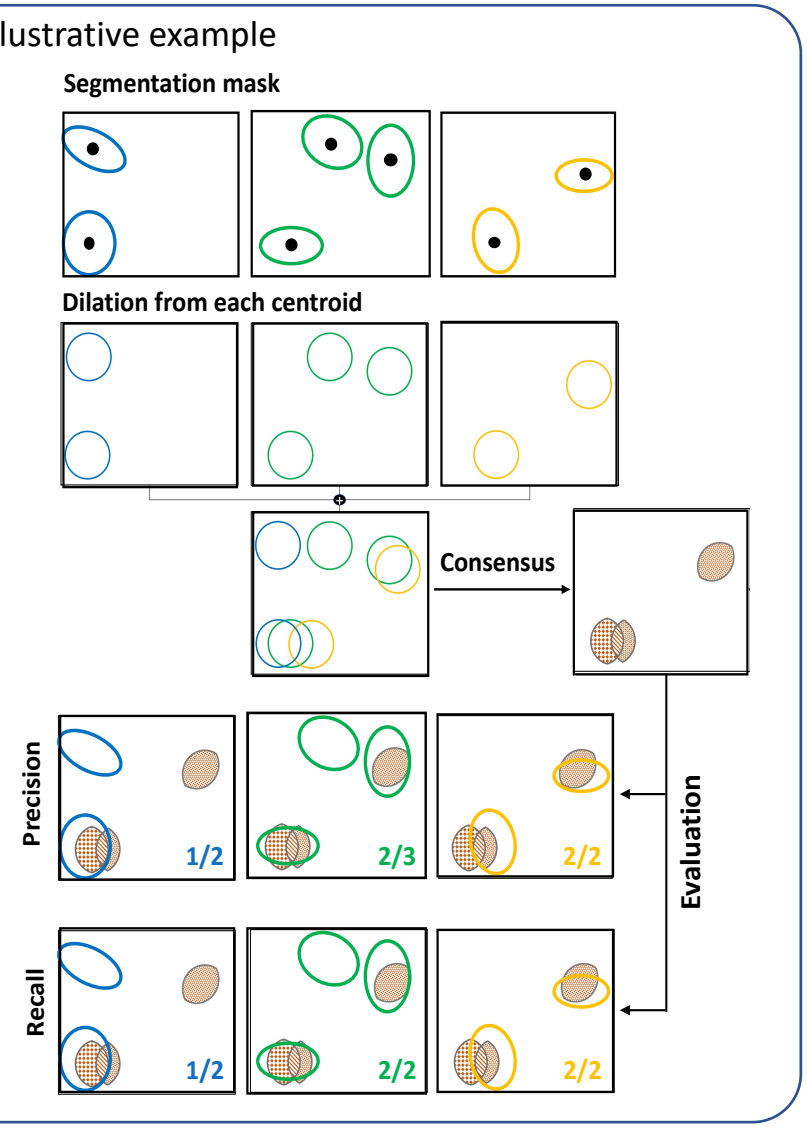


Figure 2. Method-specific weighting via ablation study avoids potential sensitivity to collective bias

Core	Dropped Method	DICE	Core	Dropped Method	DICE	Core	Dropped Method	DICE
Scene 002	Mesmer	0.772	Scene 017	Mesmer	0.755	Scene 059	Mesmer	0.731
	Stardist	0.777		Stardist	0.769		Stardist	0.737
	Cellpose	0.782		Cellpose	0.766		Cellpose	0.746
	UnMicst	0.807		UnMicst	0.798		UnMicst	0.787
Scene 003	Mesmer	0.804	Scene 049	Mesmer	0.699			
	Stardist	0.821		Stardist	0.706			
	Cellpose	0.791		Cellpose	0.718			
	UnMicst	0.838		UnMicst	0.78			

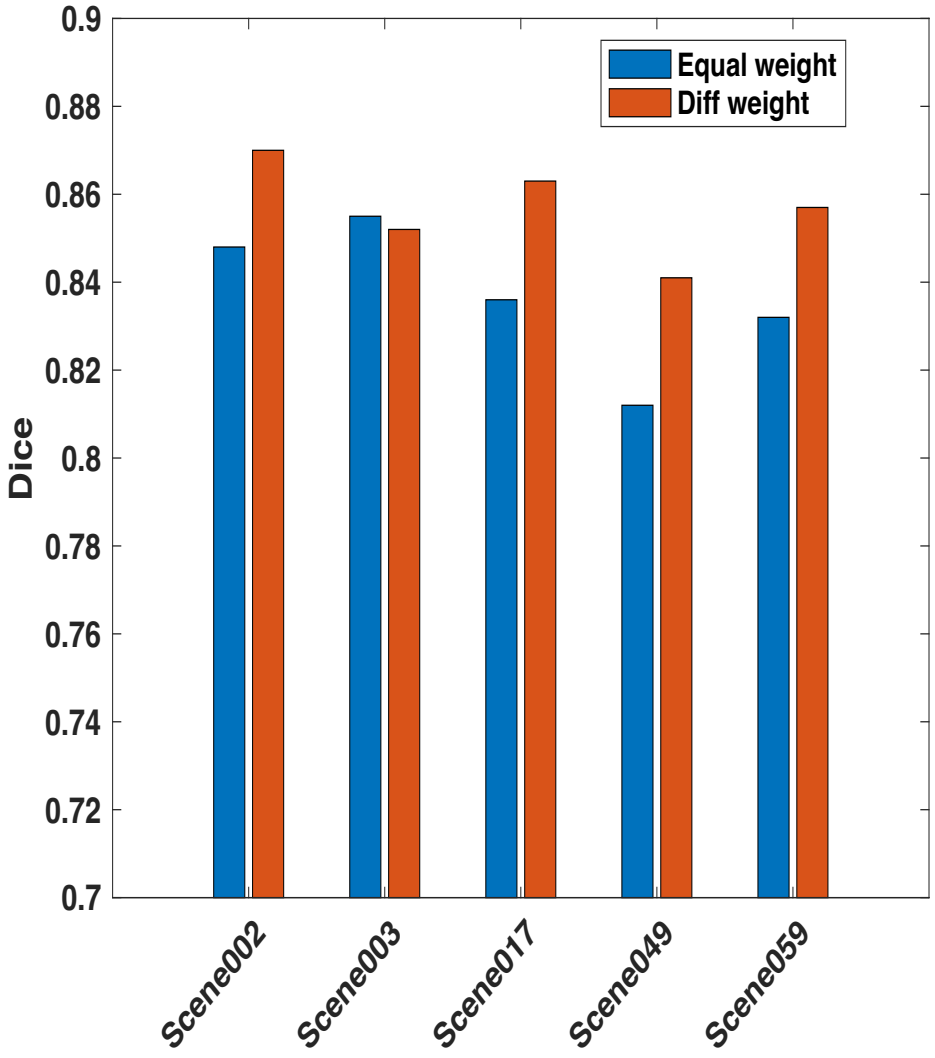


Figure 3. Refined ensemble-derived scores align with labeled ground truth

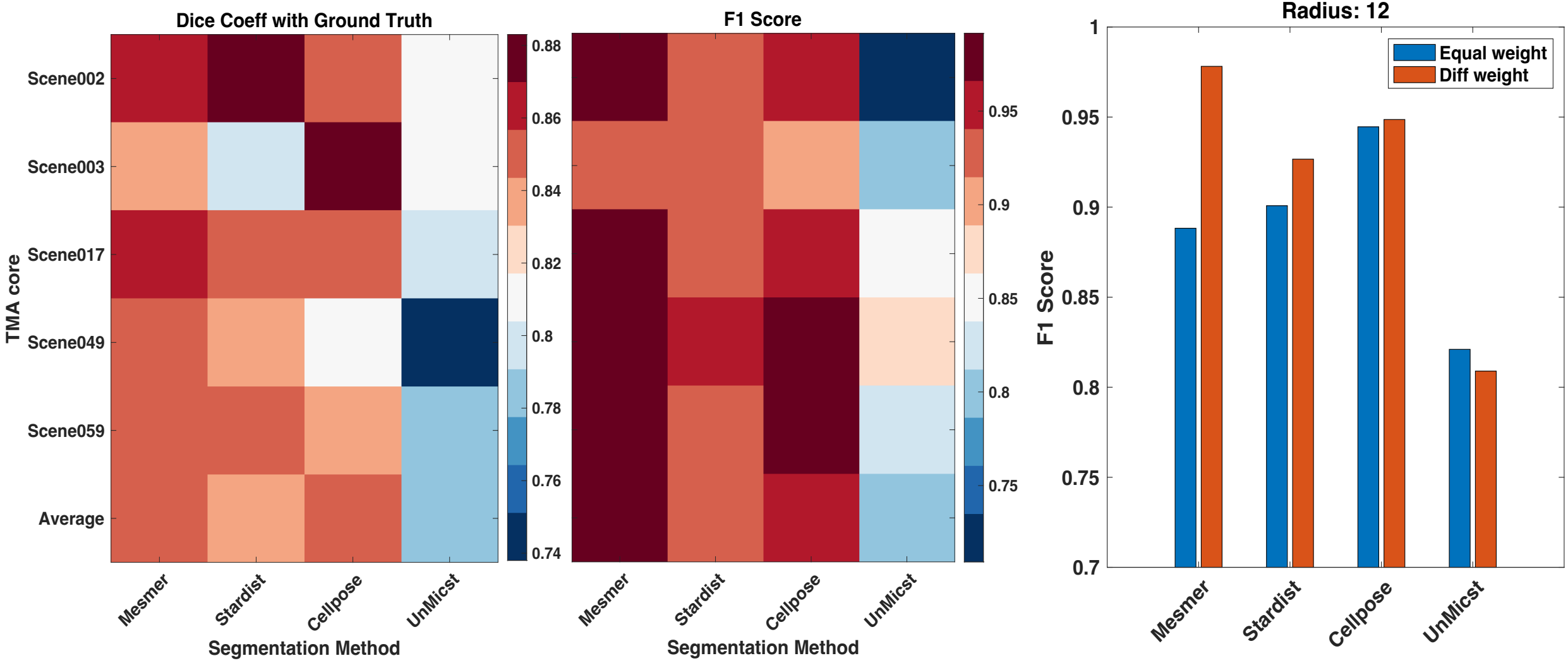


Figure 4A. Metrics computed with equal method weighting

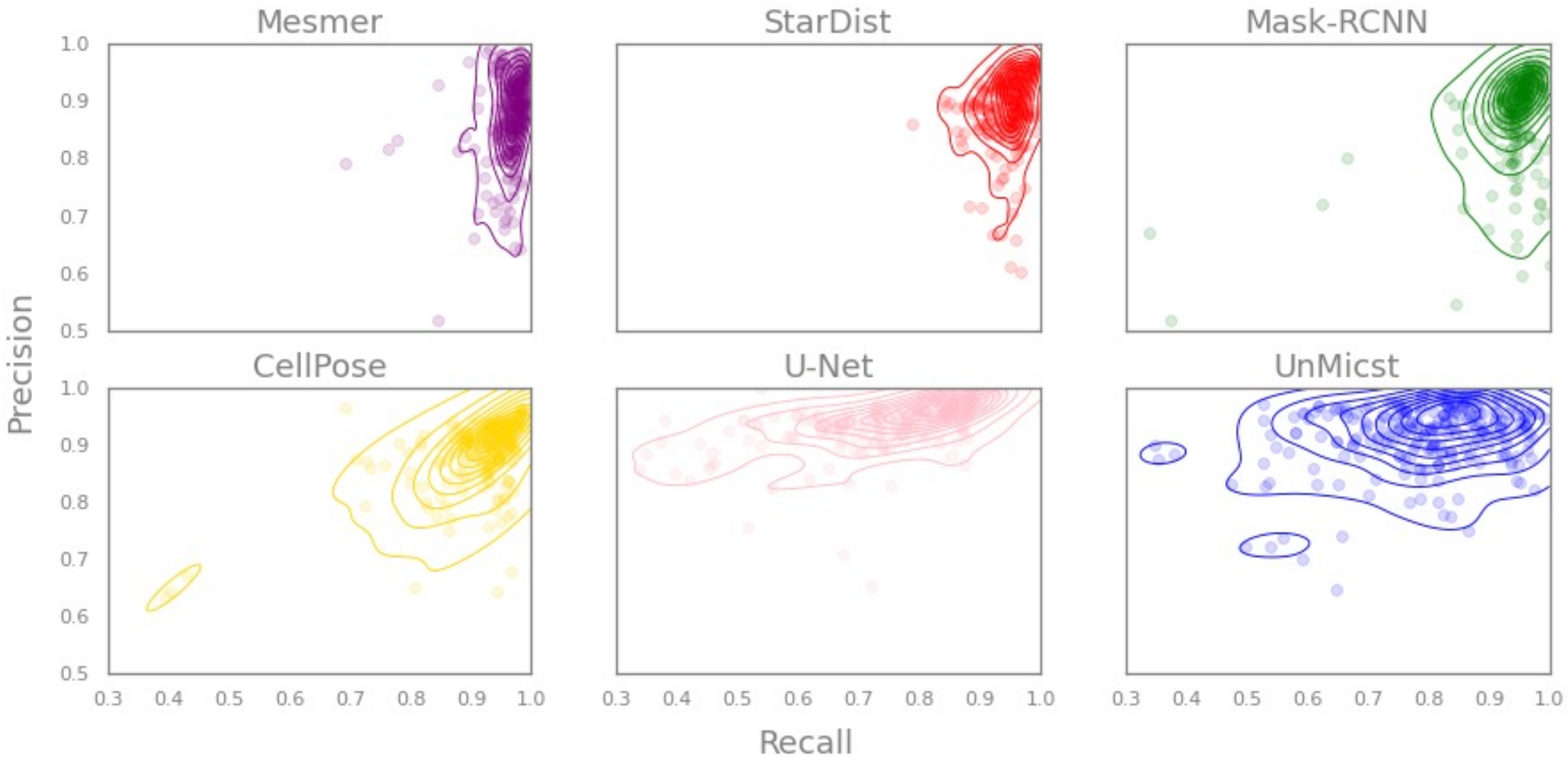


Figure 4B. Ablation study determines the relative importance of weighting

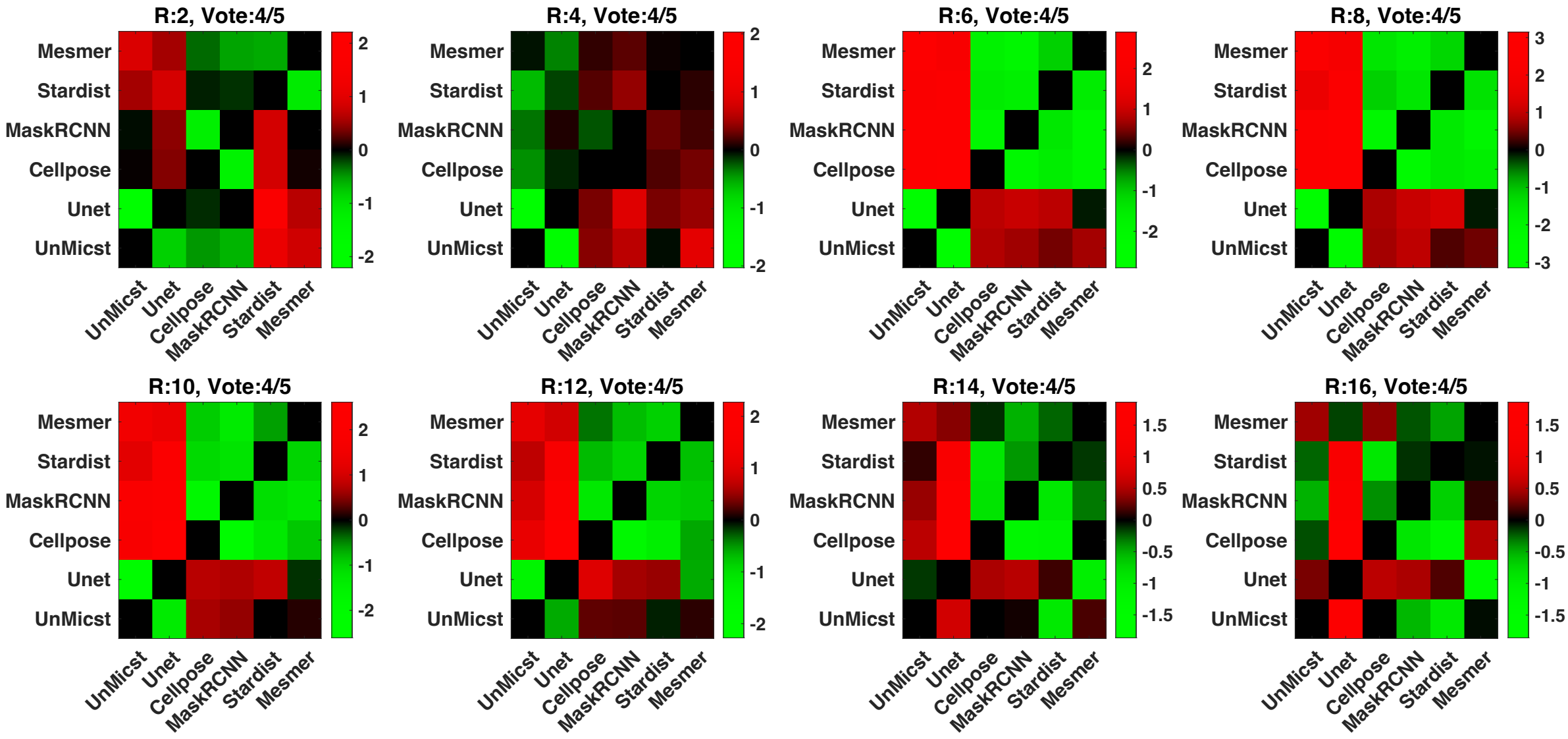


Figure 4C. Metrics refined with an un-equal weighting scheme