A Framework for Comparing and Selecting from Several Segmentation Methods in Absence of Ground Truth Labels

Dharani Thirumalaisamy*, Luke Strgar*, Eun Na Kim, Bipasa Bose, Luke Ternes, Guillame Thibault, Young Hwan Chang
Department of Biomedical Engineering, Oregon Health and Science University (OHSU) (*equal contribution)

Background

- Cell and nuclei segmentation is often a first step in analysis of multiplex tissue imaging (MTI) data.
- Evaluating segmentation results on user's datasets without ground truth labels is either subjective or amounts to the task of performing the time-intensive annotation.
- Published performance of the pretrained models may not guarantee satisfactory performance on the user's data.
- End-to-end pipelines such as MCMICRO are highly usable but lack ability to guide selection of the most appropriate segmentation method for a user's dataset.

Motivation

- Feature level discrepancy between segmentation methods:
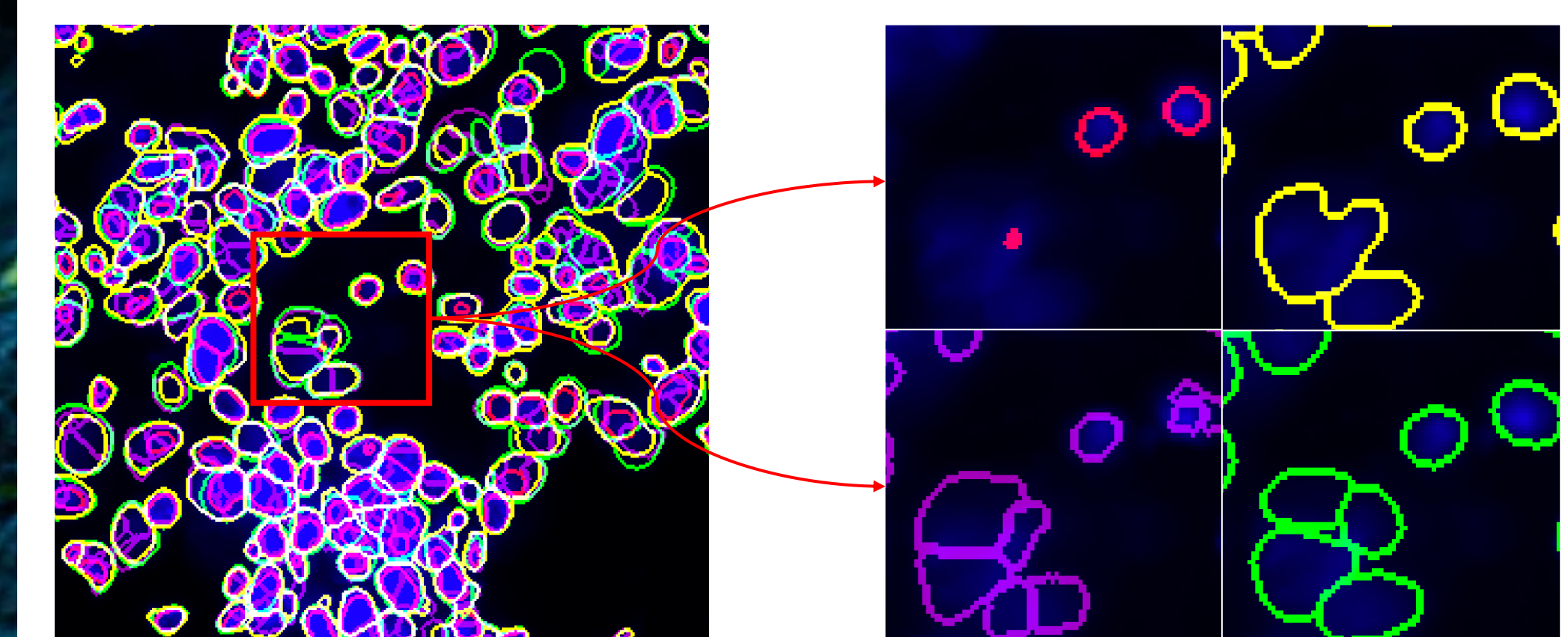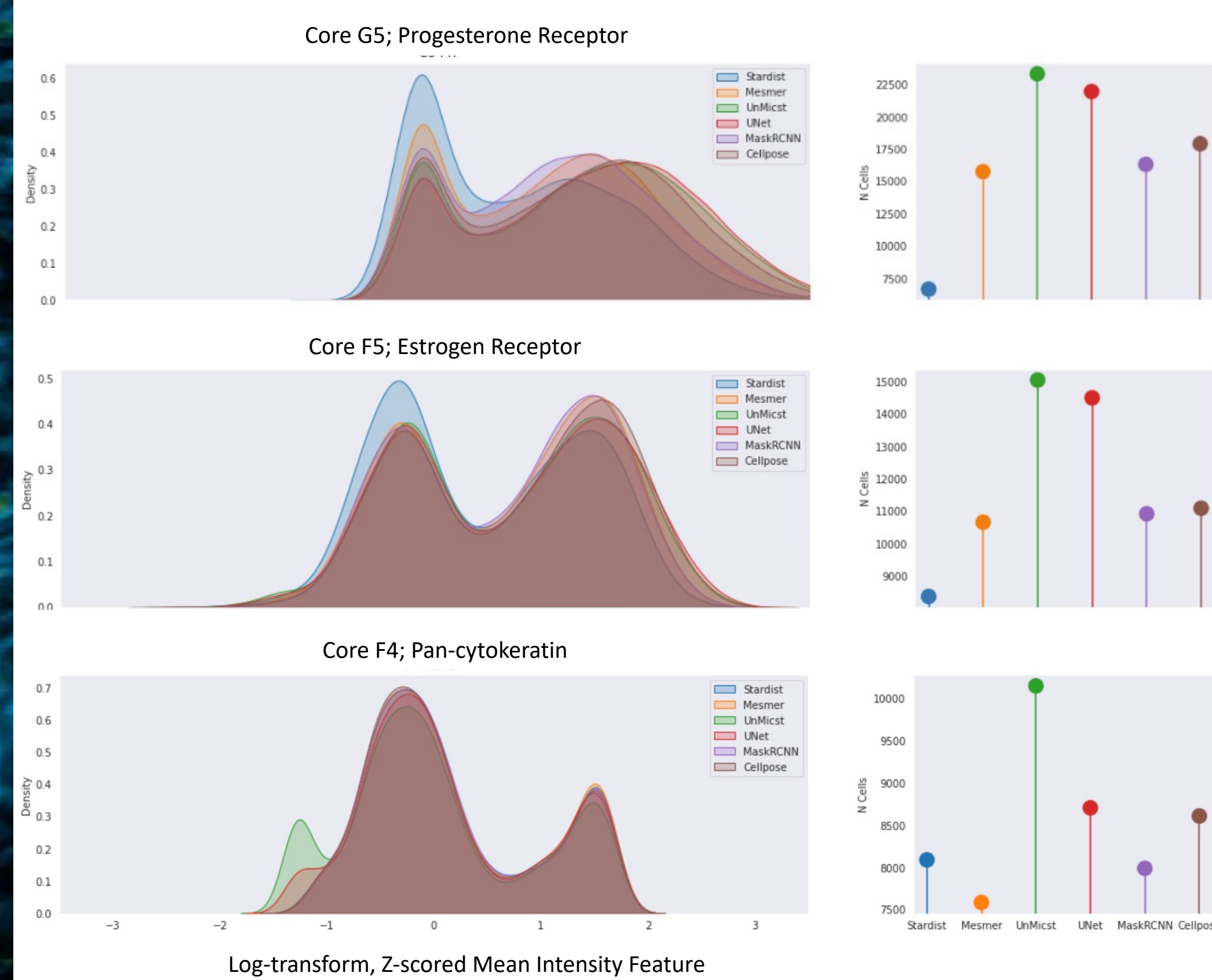
Dataset

- BC TMA: (fully annotated) 5 cores for method validation
- BC TNP-TMA: 24 antibodies (tumor panel) and 88 cores without annotation:

R1	R2	R3	R4	R5	R6	R7	R8
CD3	CCND1	Ecad	EGFR	Ki67	CK14	LamABC	PCNA
pERK	Vim	ER	pRB	CD45	CK18	AR	PanCK
Rad51	aSMA	PR	HER2	p21	CK17	H2Ax	CD31

An overview of consensus-based ground truth estimation and refinement
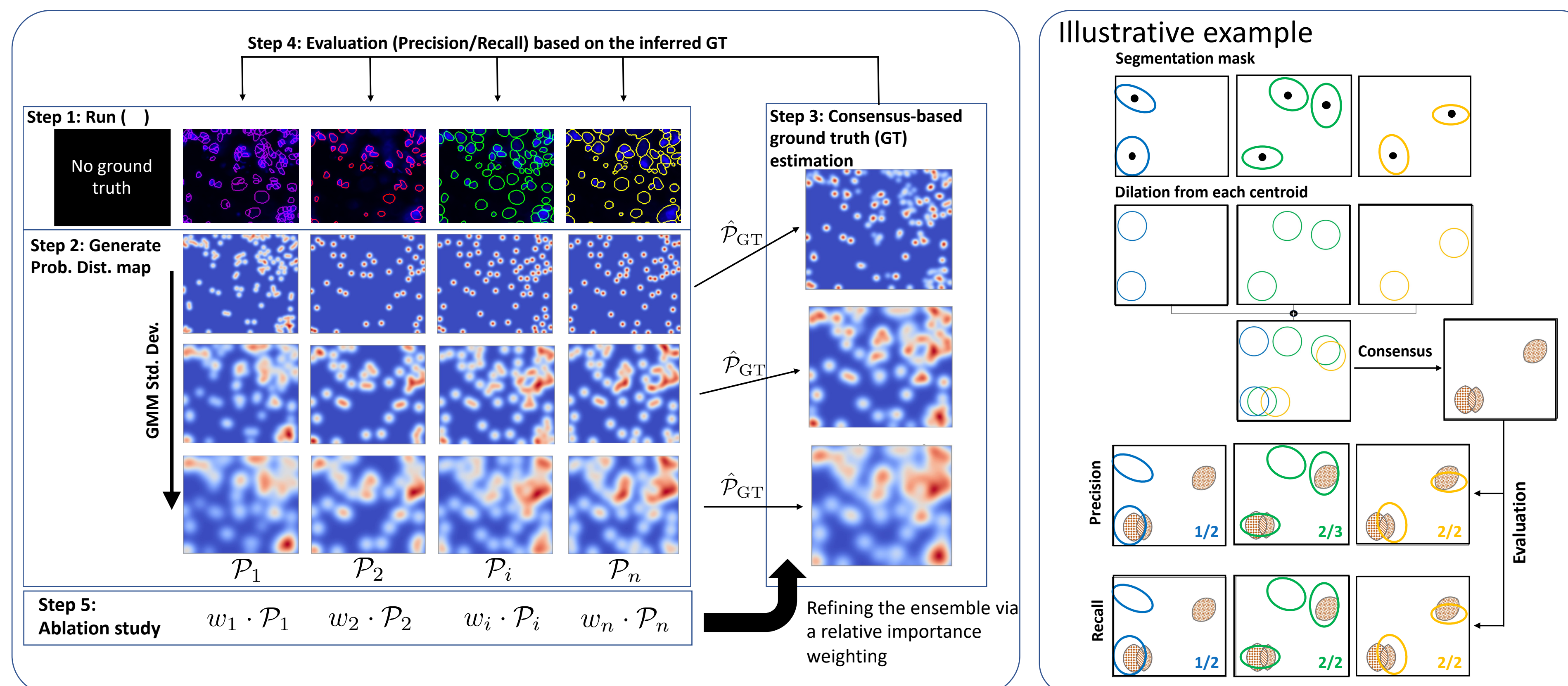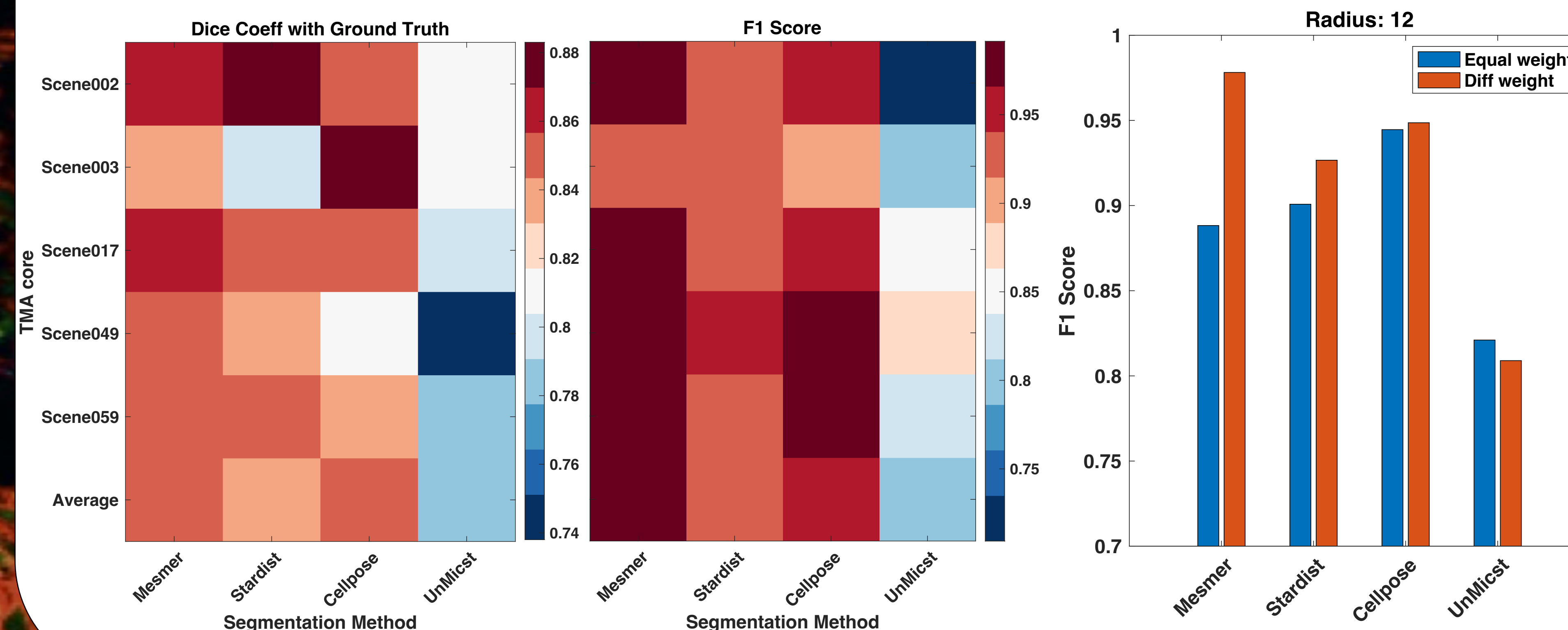

Results: Method-specific weighting via ablation study avoids potential sensitivity to collective bias

| 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			
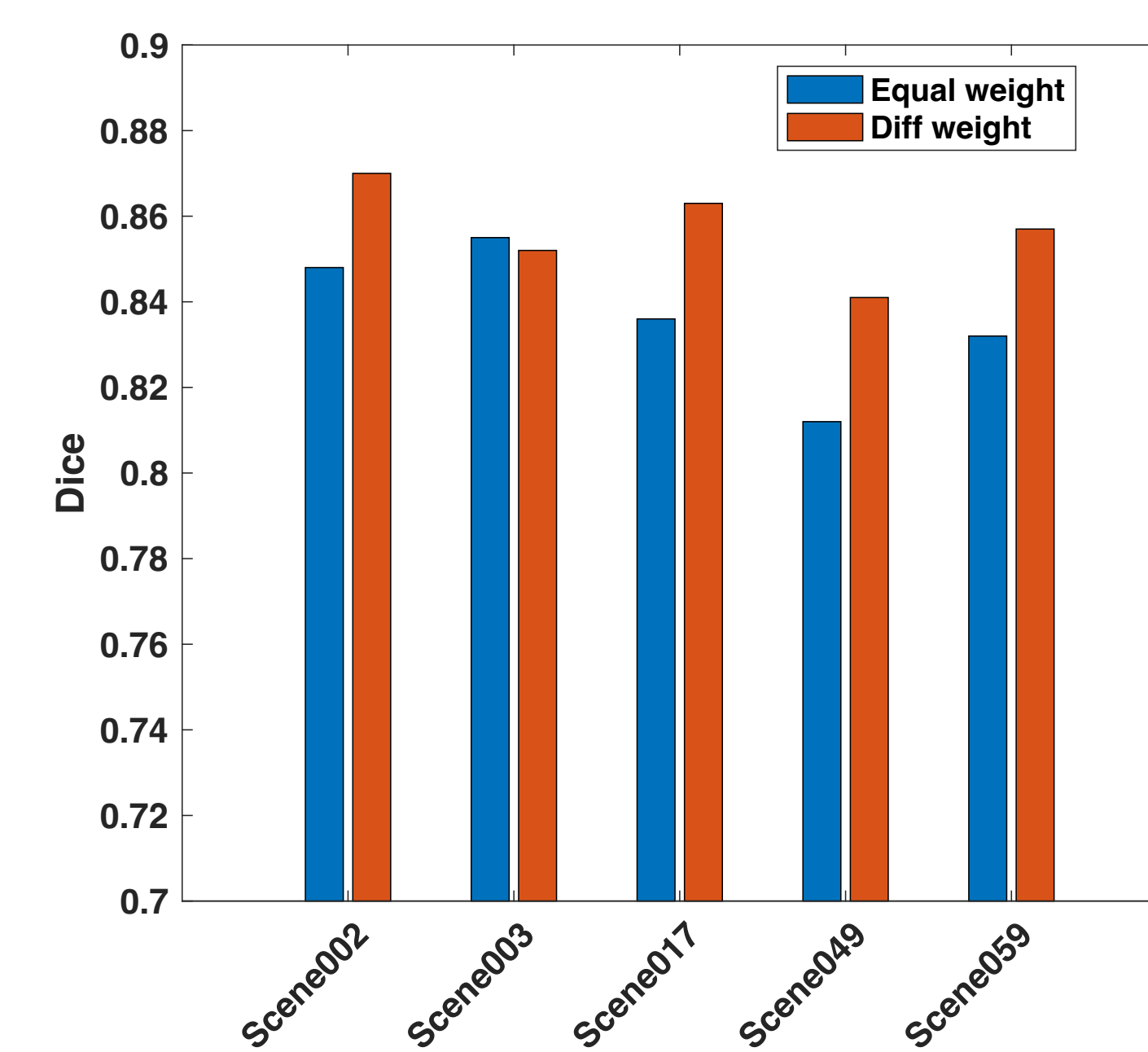
Results: Refined ensemble-derived scores align with labeled ground truth
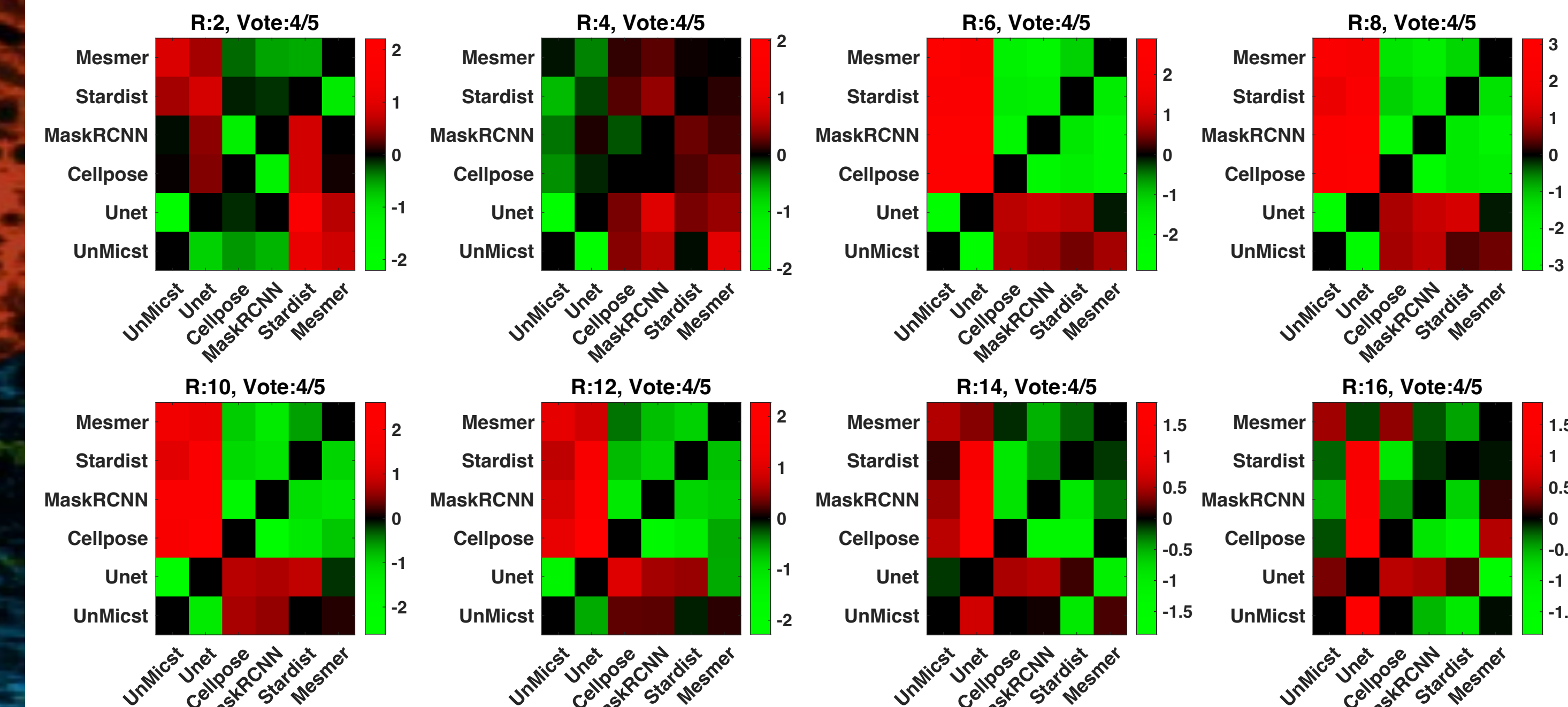

Results: Ablation study used to determine method specific weights in TNP-TMA dataset

(1) Metrics computed with equal method weighting
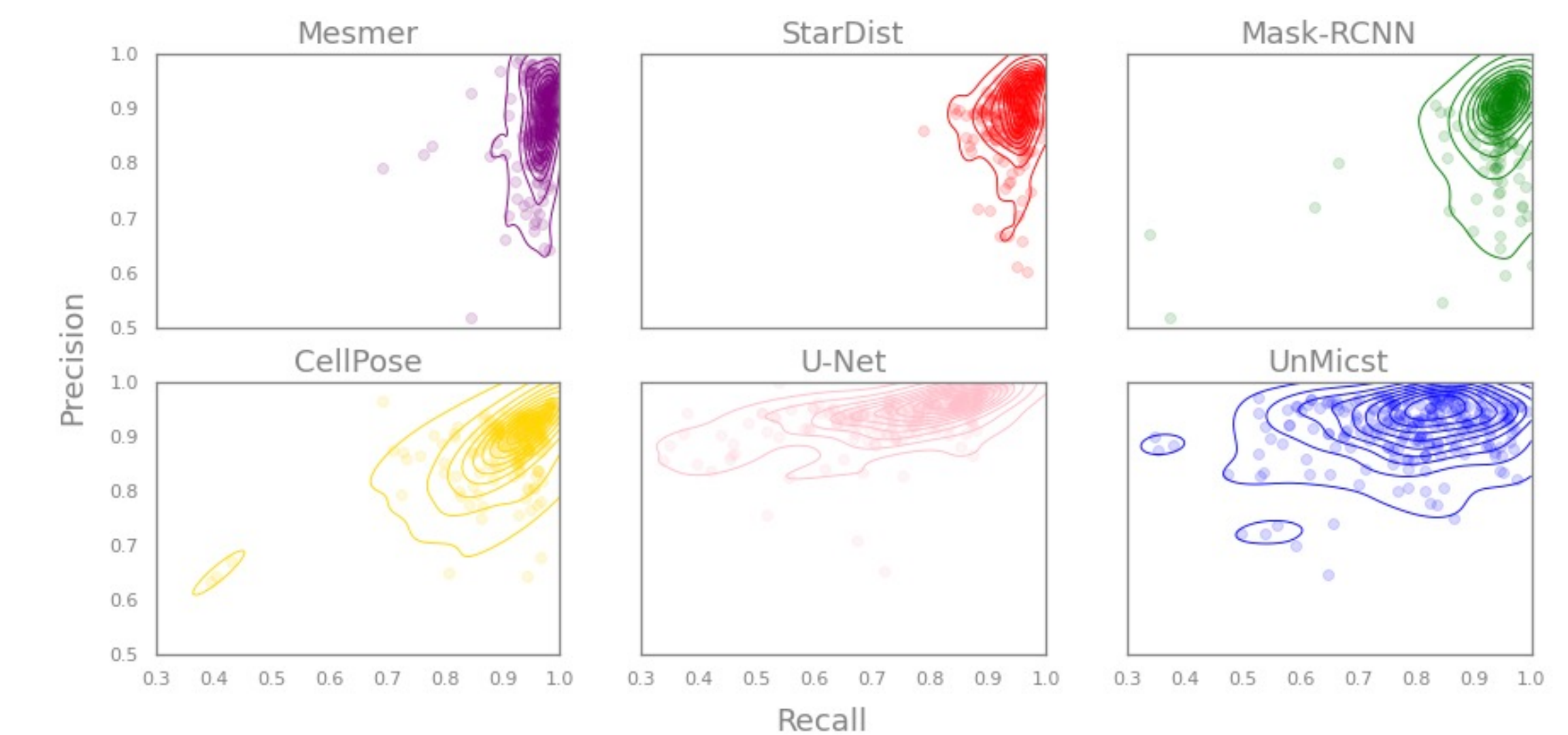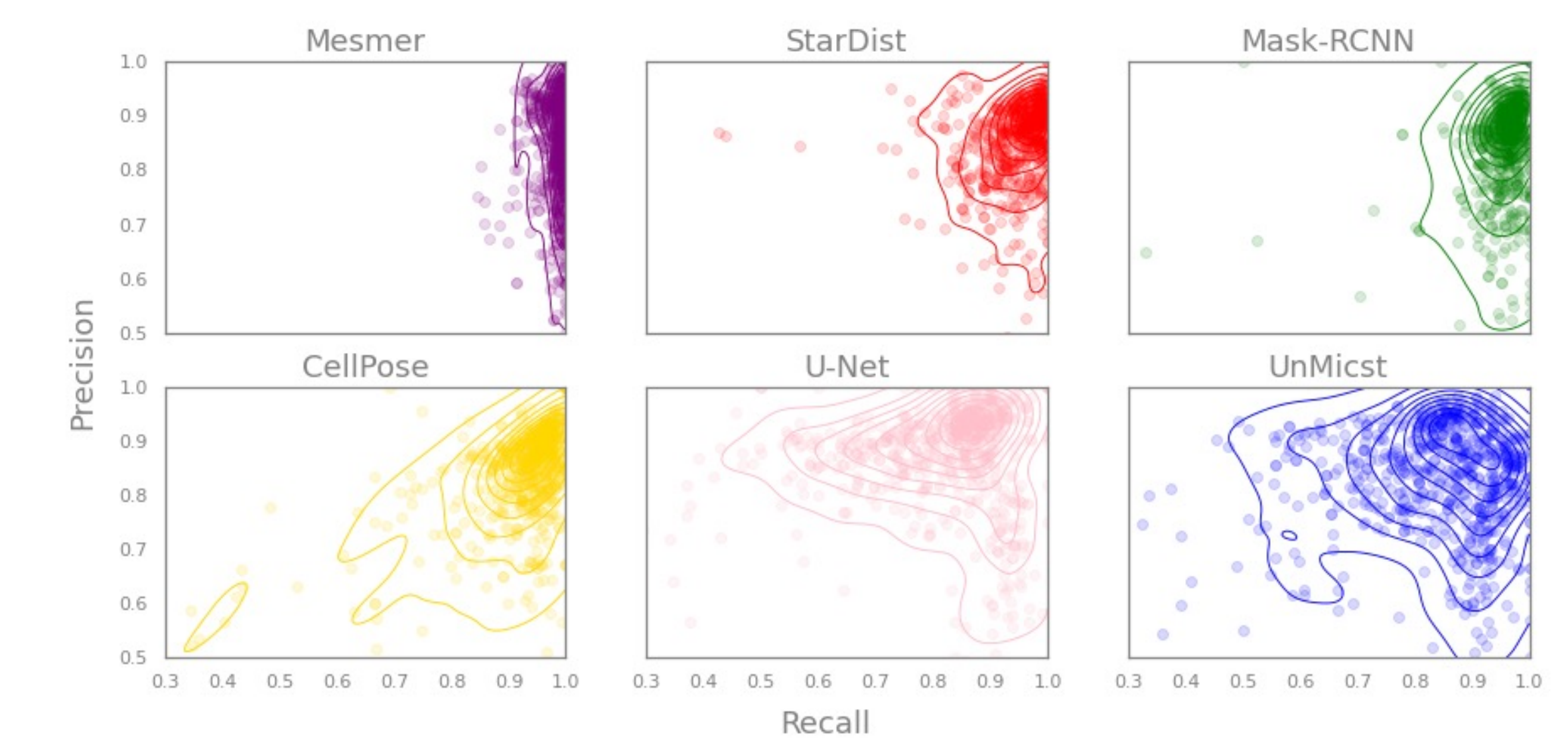

(2) Ablation study determines relative importance weighting

Example: F1 Score variation

(3) Metrics refined with un-equal weighting scheme

Summary & Discussion

We propose a methodological approach for evaluating MTI nuclei segmentation methods by scoring relative to a larger ensemble of segmentations. We demonstrate feasibility and accuracy of the proposed approach by using a small dataset (breast cancer 5 TMA cores) with ground truth labels. We validate the use of systematic model ablations to assign importance weighting scores to different segmentation methods, which further improve the ensemble-method's predictions. Lastly, we report results for 6 segmentation methods on an unlabeled TNP-TMA dataset and provide decision guidelines for the general user to easily choose the most suitable segmentation methods for their own dataset.