Subset	Image Type	Setting	Image References	Annotator	Number of image tiles	Train	Test	Number of ROIs (grains)
APF_2	UAV, Handcamera	Fluvial Gravel	Mair et al. (2022), Mair et al. (2024), Chen et al. (2022), Brayshaw (2012), Litty & Schlunegger (2017)	DM	59	49	10	8090
S1_2	UAV	Fluvial Gravel Bar	Mair et al. (2024)	DM	19	15	4	2136
FH_2	Handcamera	Outcrop gravel pit	Mair et al. (2024), Garefalakis et al. (2023)	DM	8	7	1	1426
AR	UAV, Handcamera	Fluvial gravel, debris flow deposits	this study	DM	16	13	3	1011
PR	Handcamera	Fluvial Gravel	Litty & Schlunegger (2017)	DM	7	5	2	733
NZ1	Handcamera	Fluvial Gravel	this study	DM	20	18	2	2971
NZ2	Handcamera	Fluvial Gravel	this study	DM	23	19	4	1809
NB2	Handcamera	Fluvial Gravel	this study	DM, AdP, AW	42	35	7	2938
PP	Handcamera	Fluvial Gravel	this study	FV	8	7	1	3483
HP	Handcamera	Fluvial Gravel, painted	this stuy	FV	7	6	1	2744
JF	UAV	Proglacial Sediments	Hiller et al. (2023)	DM	9	8	1	549
CT	Micro-X-ray-CT	Bioclastic Sand	Fabbri et al. (2024)	DM	6	5	1	883
DV_4	X-ray-CT	Glaciogenic Diamict, drillcores	Schuster et al. (2025)	BS	19	16	3	849
IG2 (all)					**243**	**203**	**40**	**29622**

Table S1: Summary statistics for the IG2 dataset.

Subset	n	Ell: a - axis (px)			Ell: b - axis (px)			Convexity			Eccentricity			IRn			Ell: b - axis Azimuth (°)		
		min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean
S1_2	1432	13.1	309.6	66.4 ± 59.2	8.8	206.6	39.6 ± 38.8	0.62	0.99	0.92 ± 0.07	0.31	0.97	0.76 ± 0.15	0.38	1.02	0.87 ± 0.12	2.6	177.1	88.1 ± 52.9
PR	537	12.0	599.8	108.6 ± 125.8	9.0	407.0	67.3 ± 81.0	0.73	1.00	0.94 ± 0.05	0.35	0.97	0.73 ± 0.15	0.50	1.01	0.88 ± 0.10	1.8	178.6	86.1 ± 54.1
NZ2	1184	21.0	713.2	157.2 ± 167.2	14.9	481.0	100.3 ± 111.3	0.75	0.99	0.94 ± 0.05	0.36	0.94	0.73 ± 0.15	0.58	1.00	0.88 ± 0.10	12.9	166.7	90.6 ± 46.2
FH_2	1002	10.1	215.2	41.1 ± 34.7	8.3	146.9	25.4 ± 21.4	0.77	0.99	0.95 ± 0.04	0.25	0.96	0.73 ± 0.15	0.56	1.04	0.93 ± 0.07	1.5	178.8	89.4 ± 58.6
NZ1	2131	12.7	250.9	48.8 ± 48.5	9.6	152.1	30.2 ± 28.8	0.82	0.99	0.95 ± 0.03	0.35	0.93	0.72 ± 0.14	0.67	1.02	0.93 ± 0.07	2.1	177.6	94.4 ± 57.3
NB2	1383	24.6	365.6	102.5 ± 80.1	15.2	242.4	61.5 ± 53.6	0.77	0.98	0.93 ± 0.05	0.40	0.94	0.74 ± 0.14	0.59	1.00	0.87 ± 0.10	8.9	175.0	89.5 ± 50.8
APF_2	4362	13.3	308.7	63.9 ± 63.1	9.6	199.4	39.1 ± 40.2	0.68	0.99	0.93 ± 0.06	0.36	0.95	0.74 ± 0.14	0.45	1.02	0.88 ± 0.12	5.8	176.6	91.8 ± 50.9
AR	531	12.6	334.1	93.0 ± 89.9	9.4	223.9	56.1 ± 55.9	0.75	0.98	0.94 ± 0.05	0.40	0.94	0.74 ± 0.15	0.60	1.00	0.88 ± 0.11	10.3	170.7	89.2 ± 50.3
JF	328	10.2	157.2	44.7 ± 37.9	8.3	104.3	27.6 ± 22.5	0.78	0.99	0.94 ± 0.04	0.35	0.93	0.71 ± 0.14	0.63	1.02	0.91 ± 0.09	3.4	173.5	84.9 ± 51.4
CT	458	9.1	32.9	15.8 ± 4.7	8.0	20.8	10.8 ± 2.7	0.88	0.99	0.95 ± 0.02	0.27	0.94	0.68 ± 0.15	0.80	1.04	0.97 ± 0.04	1.2	178.7	89.1 ± 52.4
DV_4	331	12.1	126.7	40.3 ± 33.1	8.9	84.1	25.9 ± 21.6	0.83	0.96	0.91 ± 0.04	0.45	0.91	0.73 ± 0.13	0.69	0.93	0.83 ± 0.07	11.2	169.7	90.3 ± 49.4
PP	2728	19.0	480.3	82.7 ± 64.3	12.1	291.0	51.5 ± 40.6	0.64	0.99	0.92 ± 0.06	0.26	0.96	0.73 ± 0.15	0.46	0.99	0.85 ± 0.09	0.7	178.9	91.5 ± 53.6
HP	2129	18.7	471.2	80.6 ± 62.1	11.9	296.8	50.5 ± 41.0	0.62	0.99	0.93 ± 0.06	0.26	0.97	0.74 ± 0.14	0.46	1.00	0.86 ± 0.09	1.3	179.7	95.6 ± 53.6
IG2 (all)	18536	9.1	713.2	72.7 ± 67.0	8.0	481.0	45.1 ± 43.0	0.62	1.00	0.93 ± 0.05	0.25	0.97	0.73 ± 0.14	0.38	1.04	0.89 ± 0.09	0.7	179.7	90.0 ± 52.4

Table S2: Summary of 2D metrics for grain morphometry for all manually labelled grain masks across all data splits after filtering for minimum grain size and distance to image tile boundary (see section 2.2 for details). Mean and average standard deviation (1 sigma) values are calculated for image-averaged values.

Metric	Method/Model	test/train	*IG2 (all)*	Data Split												
				S1_2	*PR*	*NZ2*	*FH_2*	*NZ1*	*NB2*	*APF_2*	*AR*	*JF*	*CT*	*DV_4*	*PP*	*HP*
AP@0.5	Cellpose 2 (IG2)	test	0.53	0.71	0.72	0.65	0.48	0.56	0.42	0.52	0.50	0.49	0.09	0.58	0.82	0.61
		train	0.52	0.64	0.64	0.56	0.49	0.40	0.40	0.52	0.43	0.35	0.09	0.51	0.75	0.66
	Cellpose 2 (IG1)	test	0.52	0.72	0.79	0.53	0.52	0.65	0.36	0.55	0.50	0.59	0.23	0.48	0.74	0.49
		train	0.50	0.69	0.66	0.44	0.56	0.36	0.40	0.57	0.37	0.38	0.18	0.36	0.65	0.60
	Cellpose-SAM (IG2)	test	**0.71**	**0.80**	**0.78**	**0.69**	**0.65**	**0.66**	**0.64**	**0.64**	**0.72**	**0.65**	**0.72**	**0.90**	**0.81**	**0.67**
		train	**0.73**	**0.79**	**0.72**	**0.74**	**0.69**	**0.57**	**0.69**	**0.70**	**0.61**	**0.73**	**0.67**	**0.83**	**0.83**	**0.78**
	Cellpose-SAM (default)	test	0.29	0.54	0.45	0.19	0.13	0.01	0.28	0.25	0.14	0.44	0.61	0.44	0.65	0.63
		train	0.14	0.33	0.39	0.15	0.15	0.02	0.11	0.21	0.14	0.24	0.59	0.45	0.59	0.40
	SAM (Vit-H)	test	0.46	0.58	0.59	0.48	0.49	0.38	0.42	0.42	0.46	0.56	0.59	0.46	0.66	0.41
		train	0.44	0.55	0.46	0.43	0.48	0.37	0.36	0.43	0.36	0.41	0.54	0.49	0.59	0.56
	SEG-SAM (default)	test	0.39	0.53	0.52	0.46	0.39	0.25	0.40	0.32	0.40	0.38	0.29	0.17	0.02	0.66
		train	0.31	0.46	0.52	0.39	0.34	0.17	0.30	0.31	0.28	0.24	0.24	0.21	0.26	0.50
	SEG-SAM (IG2)	test	0.50	0.59	0.69	0.63	0.44	0.45	0.47	0.42	0.47	0.36	0.11	0.60	0.72	0.63
		train	0.52	0.58	0.68	0.63	0.60	0.44	0.40	0.49	0.51	0.34	0.06	0.55	0.81	0.73
mAP	Cellpose 2 (IG2)	test	0.43	0.53	0.64	0.57	0.34	0.42	0.32	0.39	0.39	0.40	0.08	0.47	0.55	0.44
		train	0.38	0.48	0.55	0.44	0.36	0.29	0.31	0.38	0.32	0.26	0.06	0.39	0.54	0.47
	Cellpose 2 (IG1)	test	0.40	0.54	0.74	0.47	0.37	0.55	0.28	0.40	0.37	0.41	0.20	0.34	0.48	0.35
		train	0.35	0.53	0.57	0.37	0.42	0.26	0.31	0.42	0.27	0.28	0.14	0.23	0.45	0.42
	Cellpose-SAM (IG2)	test	**0.55**	**0.63**	**0.67**	**0.58**	**0.47**	**0.48**	**0.49**	**0.46**	**0.58**	**0.44**	**0.61**	**0.73**	**0.57**	**0.47**
		train	**0.55**	**0.61**	**0.61**	**0.58**	**0.51**	**0.41**	**0.53**	**0.52**	**0.47**	**0.51**	**0.52**	**0.74**	**0.60**	**0.56**
	Cellpose-SAM (default)	test	0.24	0.39	0.39	0.16	0.11	0.01	0.21	0.19	0.08	0.33	0.46	0.33	0.42	0.42
		train	0.11	0.24	0.33	0.12	0.12	0.01	0.07	0.17	0.09	0.14	0.40	0.34	0.42	0.29
	SAM (Vit-H)	test	0.33	0.42	0.51	0.40	0.33	0.27	0.28	0.30	0.33	0.40	0.49	0.34	0.50	0.32
		train	0.30	0.40	0.37	0.34	0.35	0.25	0.24	0.30	0.25	0.28	0.41	0.36	0.45	0.39
	SEG-SAM (default)	test	0.31	0.44	0.48	0.39	0.30	0.18	0.33	0.26	0.32	0.26	0.21	0.15	0.02	0.46
		train	0.25	0.37	0.46	0.34	0.27	0.13	0.24	0.26	0.22	0.19	0.17	0.19	0.21	0.37
	SEG-SAM (IG2)	test	0.37	0.45	0.56	0.51	0.35	0.30	0.35	0.31	0.36	0.27	0.09	0.47	0.52	0.43
		train	0.38	0.43	0.54	0.50	0.44	0.28	0.29	0.35	0.37	0.22	0.04	0.42	0.58	0.49

Table S3: Detailed segmentation performance for the respective train and test splits. Potential overfitting to training data is indicated, i.e., when AP scores for the training tiles are higher by 0.05 or more for a respective train/test pair. Please note that for the HP data, the difference between train and test images is likely due to dataset imbalance, since scores were also higher for the test image tiles for methods that have not been trained on any of the IG2 images (e.g., SAM).

Subset	n	Ell: a - axis (px)			Ell: b - axis (px)			Mean Δ IRn (%)	Mean Δ Convexity (%)	Mean Δ Eccentricity (%)	Mean Δ Azimuth (%)
		GSDs ident. to GT (%)*	Mean Perc. Δ (px)	Mean Δ Diam. (%)	GSDs ident. to GT (%)*	Mean Perc. Δ (px)	Mean Δ Diam. (%)				
S1_2	1348	100.0	-0.9 ± 4.0	-1.1 ± 6.2	100.0	-1.3 ± 2.6	-3.1 ± 7.9	-1.4 ± 1.2	-1.4 ± 0.8	1.8 ± 1.4	0.1 ± 2.9
PR	475	100.0	3.7 ± 5.4	5.9 ± 6.3	100.0	-2.0 ± 6.1	-1.4 ± 11.0	-0.4 ± 1.1	-0.7 ± 0.6	2.8 ± 2.2	0.4 ± 5.4
NZ2	1081	73.9	31.9 ± 60.5	16.4 ± 16.0	78.3	18.7 ± 48.8	11.5 ± 20.8	-1.5 ± 3.2	-1.2 ± 1.8	1.8 ± 5.5	-0.8 ± 12.9
FH_2	876	100.0	-3.8 ± 2.7	-9.3 ± 6.3	62.5	-2.6 ± 1.5	-10.4 ± 5.8	-0.8 ± 1.1	-0.7 ± 0.5	2.0 ± 0.8	-4.0 ± 7.7
NZ1	1792	75.0	-1.0 ± 8.5	-0.3 ± 18.8	70.0	-0.6 ± 5.3	-0.1 ± 18.3	0.0 ± 1.5	-0.3 ± 0.8	0.0 ± 3.5	-5.9 ± 19.9
NB2	1248	73.8	11.2 ± 43.1	3.9 ± 17.2	64.3	2.2 ± 17.3	0.3 ± 16.4	-0.3 ± 2.4	-0.6 ± 1.2	2.9 ± 5.2	-2.4 ± 11.2
APF_2	3941	83.1	-3.7 ± 8.5	-5.8 ± 13.6	76.3	-3.9 ± 5.3	-10.0 ± 13.7	-1.2 ± 2.8	-1.1 ± 1.0	2.7 ± 2.9	0.1 ± 8.2
AR	447	50.0	-7.7 ± 28.0	-12.4 ± 24.1	50.0	-7.9 ± 12.6	-14.6 ± 19.4	-0.8 ± 2.7	-0.8 ± 1.5	0.0 ± 6.9	1.6 ± 8.0
JF	293	88.9	-0.8 ± 8.4	-4.7 ± 14.4	77.8	-1.3 ± 2.3	-6.4 ± 8.9	1.1 ± 1.3	0.2 ± 0.4	0.2 ± 3.9	-6.0 ± 18.3
CT	403	100.0	0.1 ± 0.4	0.5 ± 2.5	100.0	0.1 ± 0.1	0.7 ± 1.3	2.2 ± 1.0	1.1 ± 0.4	-0.2 ± 0.6	-0.7 ± 3.6
DV_4	303	94.7	-2.0 ± 3.4	-4.7 ± 6.7	84.2	-1.4 ± 2.3	-4.7 ± 6.7	-10.8 ± 5.8	-3.0 ± 1.4	-0.2 ± 2.3	-2.5 ± 12.3
PP	2708	100.0	2.9 ± 5.3	3.0 ± 4.7	100.0	0.0 ± 1.7	-0.8 ± 4.0	-4.4 ± 1.2	-1.6 ± 0.6	2.6 ± 1.0	0.1 ± 2.1
HP	2274	100.0	2.4 ± 6.3	3.1 ± 8.1	100.0	0.1 ± 4.4	-0.5 ± 8.9	-3.5 ± 2.0	-1.5 ± 0.9	3.2 ± 0.4	2.2 ± 4.2
IG2 (all)	17189	87.6	2.5 ± 14.2	-0.4 ± 11.2	81.8	0.0 ± 8.5	-3.0 ± 11.0	-1.7 ± 2.1	-0.9 ± 0.9	1.5 ± 2.8	-1.4 ± 9.0

**within 95% confidence ($p \geq 0.05$ for a two-sample Kolmogorov–Smirnov test)*

Table S4: Summary of differences in 2D morphometry metrics between predicted grain masks from Cellpose-SAM and ground truth ROIs across all data splits. Mean and average standard deviation (1 sigma) values are calculated for image-averaged values. GSD = grain size distribution, Perc. Δ = Percentile difference across percentiles of a GSD, GT = ground truth, IR_n = normalized isoperimetric ratio.

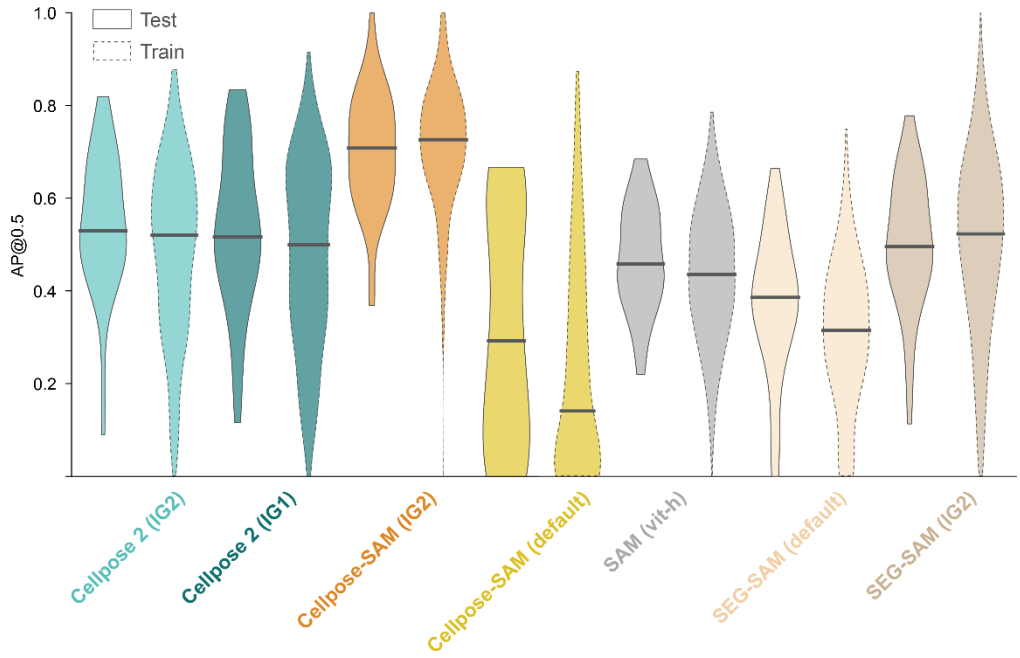
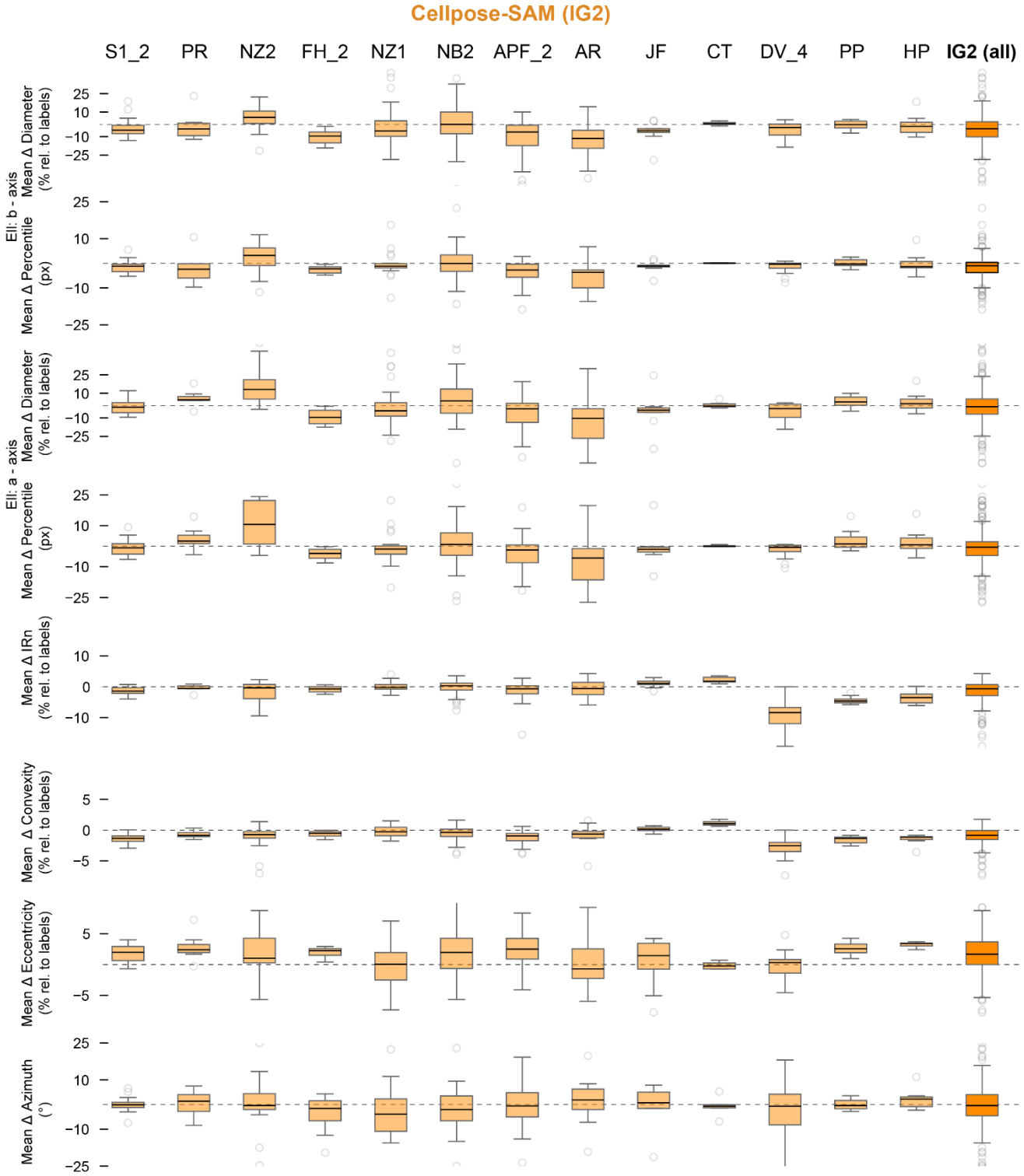


Figure S1: Summarized segmentation performance for the respective train and test splits. AP@0.5 = average precision evaluated at intersection over union (IoU) threshold of 0.5; SEG-SAM = Segmenteverygrain.



25 **Figure S2:** Image-averaged differences in size and shape metrics of predicted grain masks compared to ground-truth ROIs.

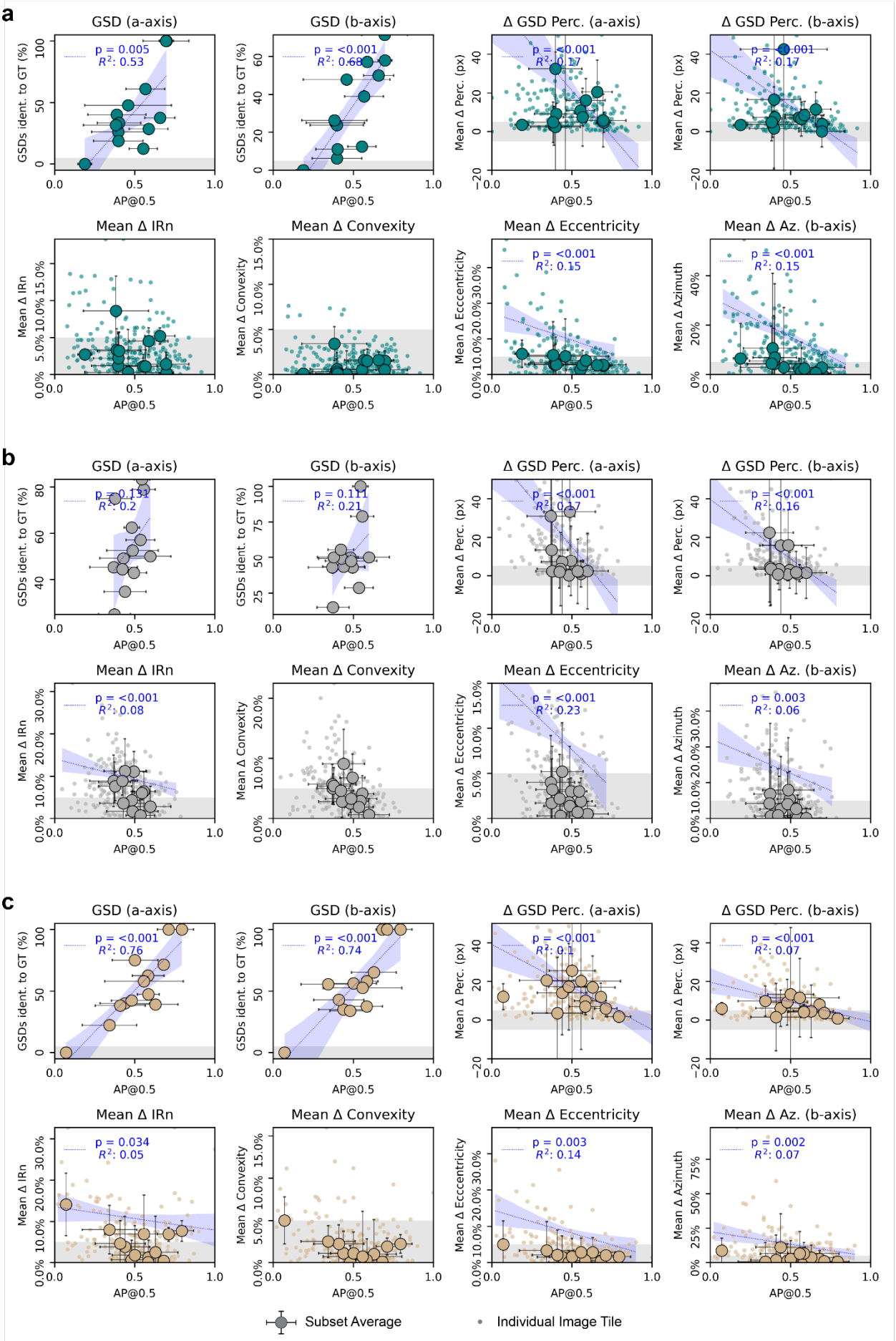


Figure S3: Comparison of segmentation performance metric (AP@0.5) with relative differences in grain size and shape between predicted grains and the ground truth ROIs for several benchmark default models: **a)** Cellpose 2 (trained on the entire IG2 dataset), **b)** SAM (Vit-H),

and **c)** Segmenteverygrain (trained on the entire IG2 dataset). Grey areas indicate very low differences between predicted grain masks and ground-truth ROIs, with differences being within ± 5px and < 5%, respectively. Only statistically significant correlations ($p \leq 0.05$, $R^2 \geq 0.05$) for individual images are indicated. For shape metrics only values with Δ values > 5% were considered for correlation. R^2 = coefficient of determination. Please note that the y-axes in panels for the Δ GSD percentile differences are cropped for a better visualization of the bulk of results.