Quantifying Inter-Annotator Agreement and Generalist Model Limitations in Imaging Mass Cytometry Single Cell Segmentation

Johannes Schuiki[1] Markus Steiner[2,3] Heinz Hofbauer[1] Stephan Drothler[2,3,4]
Giulia Pessina[2,3] Richard Greil[2,3] Nadja Zaborsky[2,3] Andreas Uhl[1]

[1]Dept. of Artificial Intelligence and Human Interfaces, Paris-Lodron-University Salzburg, Austria

[2]Cancer Cluster Salzburg, Austria

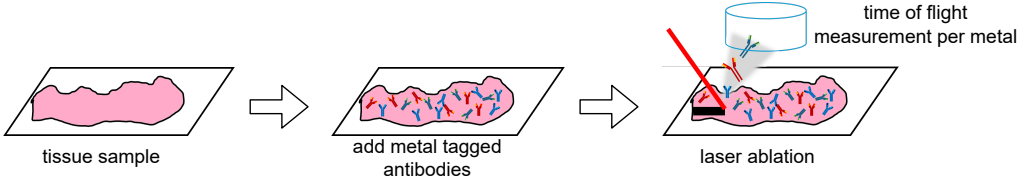
[3]Dept. of Internal Medicine III, Paracelsus Medical University, Austria

[4]Dept. of Biosciences, Paris-Lodron-University Salzburg, Austria

July 16, 2025

Background | Imaging mass cytometry

- fluorescence microscopy suffers from spectral overlap of fluorescent markers and autofluorescence
- one alternative is mass spectrometry (here: imaging mass cytometry)

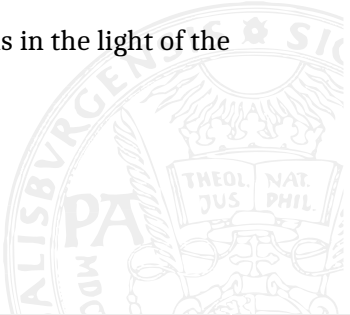


- commercialized system named “Hyperion”by Flugidim
- fixed-size resolution: 1μm per pixel
- many channels (40+)
- tissue sample gets destroyed in the process

Aim of research

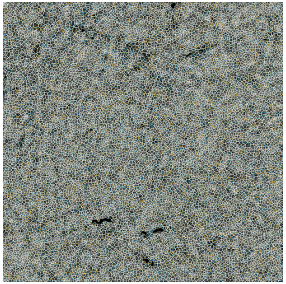
1) Assess inter-annotator agreement on a set of lymphoid tissue samples annotated by 4 experts.

2) Evaluate performance of 4 generalist cell segmentation models in the light of the results from 1) and also on four external datasets.

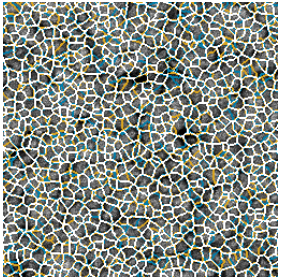


Typical workflow:

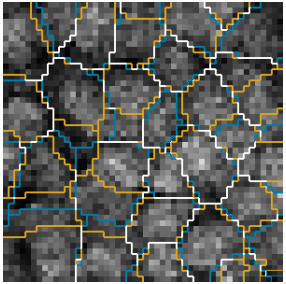
- Experts annotate patches manually using Ilastik[1] to generate pixel probability maps (background, nuclei, membrane)
- Probability maps are expanded to the whole mage using CellProfiler[2]



full 1000 x 1000 image



256 x 256 crop



50 x 50 crop (upscaled)

1 S. Berg et al. (2019). "ilastik: interactive machine learning for (bio)image analysis". In: *Nature Methods*

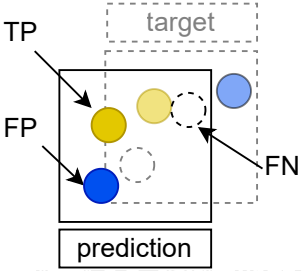
2 D. R. Stirling et al. (2021). "CellProfiler 4: improvements in speed, utility and usability". In: *BMC Bioinformatics*

- Related works rely on F1-score or ”average precision”-variant (dependent on fixed IoU)
- Side note to average precision: a recent work[3] unravels confusion

This work uses three metrics:

1. **average precision@IoU** (as cited above and used in the Data Science Bowl 2018)

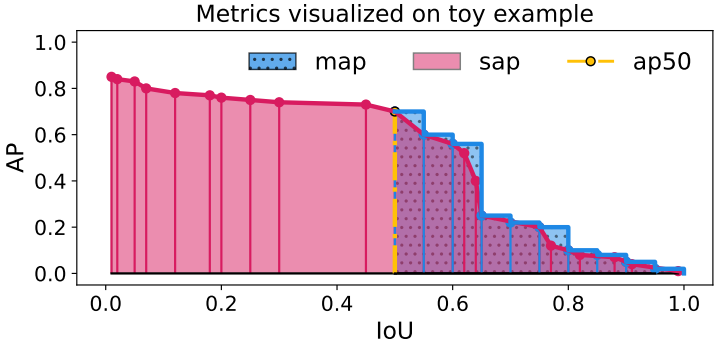
$$AP(t_{IoU}) = \frac{TP(t_{IoU})}{TP(t_{IoU}) + FN(t_{IoU}) + FP(t_{IoU})}$$



2. **mean average precision** average AP values over IoU = [0.5, 0.55, ..., 0.95]

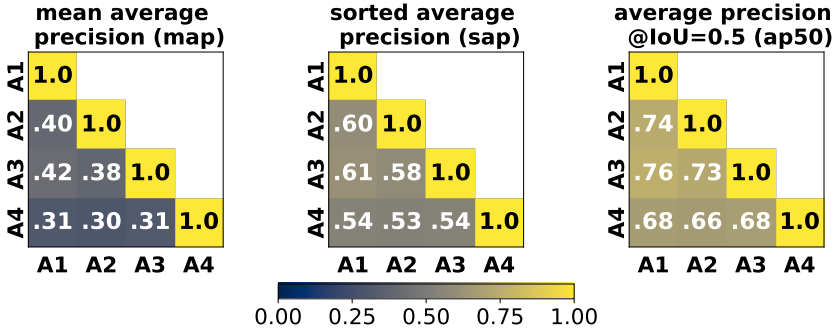
[3] D. Hirling et al. (2024). “Segmentation metric misinterpretations in bioimage analysis”. In: *Nature Methods*

3. **sorted average precision** [4]
   1. Calc IoU between all corresponding cell instances of two images
   2. Determine matching objects by treating this as an *assignment problem* (optimization, e.g. scipy.optimize → linear_sum_assignment)
   3. sort pairs according to their IoU and calculate AP at every point



[4] L. Chen et al. (2023). "SortedAP: Rethinking Evaluation Metrics for Instance Segmentation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*

Results | Inter-annotator agreement



average map	average sap	average ap50
.353	.566	.708

Model Name	Version	Year	Backbone Architecture
Cellpose[5]	v3 / cyto3	2024	Residual U-Net
Deepcell/Mesmer[6]	0.12.10	2021	ResNet-50 + FPN
CellSAM[7]	0.1.0	2023	SAM
VISTA-2D[8]	–	2024	SAM

- Models expect RGB input including membrane and nucleus channel. 11 channels are collapsed into membrane channel; 2 channels are collapsed into nucleus channel.

[5] C. Stringer et al. (2025). "Cellpose3: one-click image restoration for improved cellular segmentation". In: *Nature Methods*

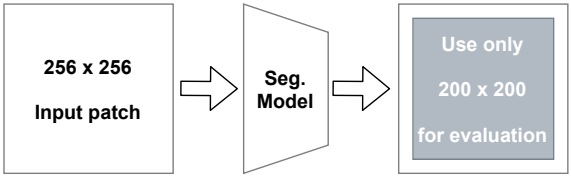
[6] N. F. Greenwald et al. (2021). "Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning". In: *Nat Biotechnol*

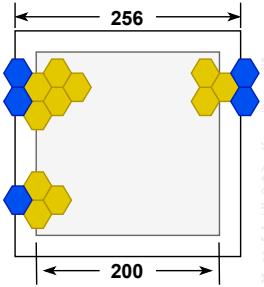
[7] U. Israel et al. (2023). ***A Foundation Model for Cell Segmentation.*** Preprint: biorxiv

[8] NVIDIA (2024). ***VISTA-2D: A foundational model for cell segmentation in spatial omics workflows.*** https://github.com/Project-MONAI/VISTA/tree/main/vista2d. Version 0.3.0

Methods | Patching strategy

- Preliminary experiments showed that full images often result in bad segmentation results
- Hence, sliding window patching strategy:





Methods | Overview data

Dataset	Abbrev.	Tissue type	# Samples whole image	avg resolution whole image (y/x)	# Samples patches	# annotators per sample	avg # cell masks per patch
in-house	A1 – 4	Lymphoid	10	1000.0/1000.0	360	4	823.7
Ali20 [9]	A20	Breast	548	462.8/478.0	2787	1	314.0
Rendeiro21 [10]	R21	Lung	229	1108.4/1187.5	13361	1	185.3
Jackson20 [11]	J20	Breast	746	596.5/626.7	8714	1	320.5
Hoch22 [12]	H22	Melanoma	167	993.1/963.4	6361	1	467.4

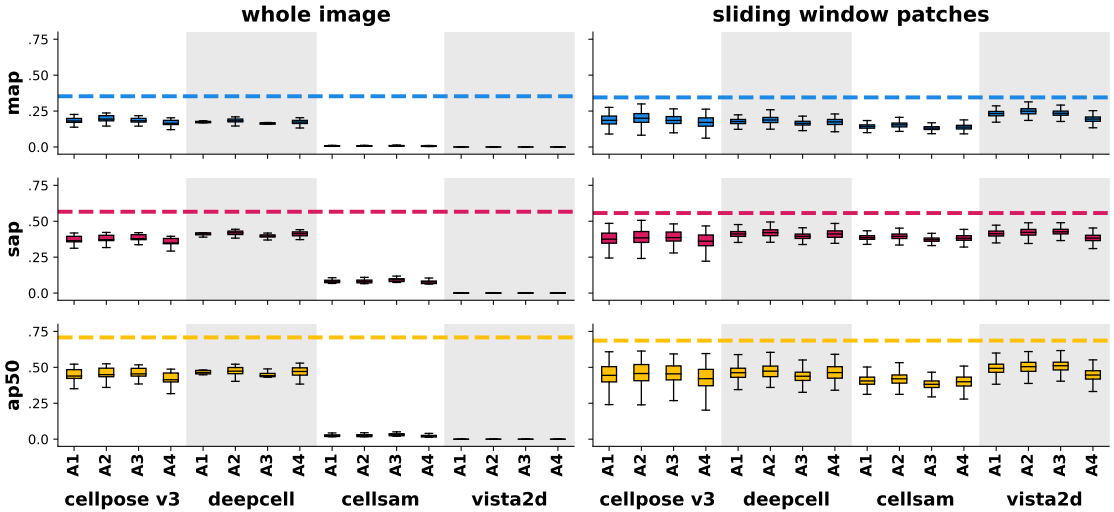
[9] H. R. Ali et al. (2020). "Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer". In: *Nature Cancer*

[10] A. F. Rendeiro et al. (2021). "The spatial landscape of lung pathology during COVID-19 progression". In: *Nature*

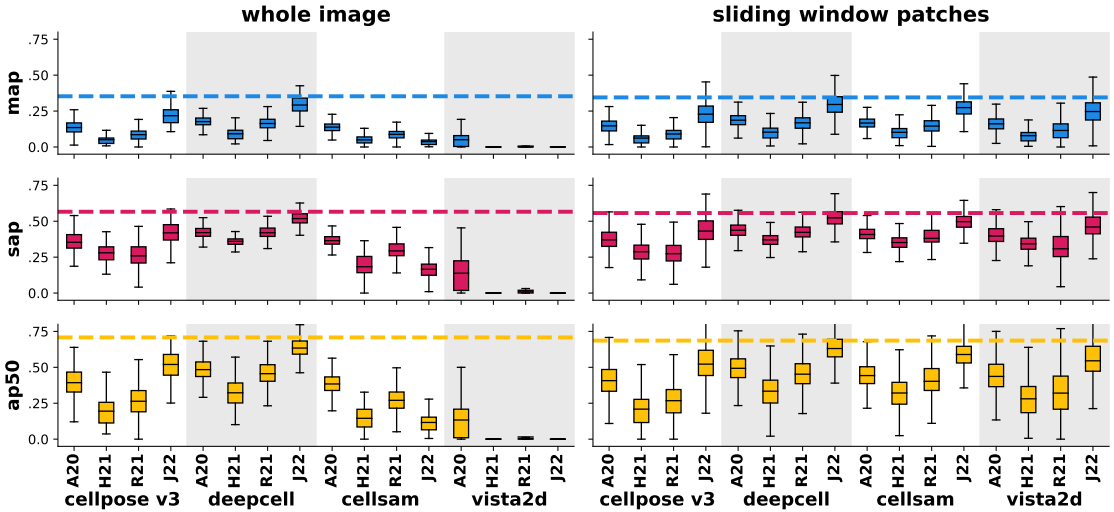
[11] H. W. Jackson et al. (2020). "The single-cell pathology landscape of breast cancer". In: *Nature*

[12] T. Hoch et al. (2022). "Multiplexed imaging mass cytometry of the chemokine milieus in melanoma characterizes features of the response to immunotherapy". In: *Sci. Immunol.*

Results | Model output vs. individual annotators



Results | Model output vs. external datasets



Limitations of this study / Future work

- Transfer of lymphoid tissue upper bound to other tissue types is debatable
- This study focused on generalist models
- Channel aggregations can be evaluated using ablation study

Conclusion

This study did:

- Quantification of inter-annotator agreement between four annotators; used as upper bound for seg. model performance
- Evaluate performance of four generalist models on in-house data and external datasets; View results in light of this upper bound

Conclusions:

- Within this experimental setup, no tested model was able to reach this upper bound
- SAM based models tend to fail at arbitrary sized images

Q & A

Find resources here:

Data

Code