DOI: 10.1007/978-3-031-98694-9_11

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] Department of Artificial Intelligence and Human Interfaces, Paris-Lodron-University Salzburg, Salzburg, Austria
[2] Department Laboratory of Immunological and Molecular Cancer Research-Salzburg Cancer Research Institute, Cancer Cluster Salzburg, Salzburg, Austria
[3] Department of Internal Medicine III with Haematology, Medical Oncology, Haemostaseology, Infectiology and Rheumatology, Oncologic Center, Paracelsus Medical University, Salzburg, Austria
[4] Department of Biosciences, Paris-Lodron-University Salzburg, Salzburg, Austria

Abstract

Accurate segmentation mask generation is critical for single-cell analysis workflows. While established semi-automated tools require expert intervention, emerging approaches aim to eliminate human guidance through fully automatic segmentation models. However, the suitability of automatically generated cell segmentation masks as reliable alternatives to expert annotations remains uncertain. This study evaluates different imaging mass cytometry (IMC) datasets by feeding them to a variety of generalist cell segmentation models and comparing the outputs with corresponding segmentation masks. Performance is assessed using instance segmentation metrics which are also viewed in the light of an upper bound determined by inter-annotator agreement.

Contents

1 Introduction 2

2 Related work on inter-annotator agreement for cell instance segmentation 2

3 Methods 3
3.1 Data . . . 3
3.1.1 IMC sample generation . . . 3
3.1.2 Mask generation . . . 3
3.1.3 Dataset statistics . . . 3
3.1.4 Patching strategy and channel aggregation . . . 3
3.2 Segmentation models . . . 4
3.3 Segmentation evaluation metrics . . . 5

4 Results and discussion 5
4.1 Inter-annotator agreement . . . 6
4.2 Automatic segmentation using generalist models . . . 6
4.3 Automatic segmentation on public datasets . . . 6
4.4 Limitations of this study . . . 7

5 Conclusion 8

This work has been partially supported by the Salzburg State Government WISS 2025 Research and Transfer Lab on "Artificial Intelligence in BioMedical Image Analysis (AIBIA)"; the Austrian Science Fund (FWF) doc.funds.connect programme "Artificial Intelligence-Driven Biomedical Imaging Innovation (REVELATION)" Grant-DOI 10.55776/DFH4791124; the WISS 2025 Cancer Cluster Salzburg, CCSII-IOS; the Province of Salzburg and FWF Grant-DOI 10.55776/P32762 to N. Zaborsky.

1 Introduction

Since its introduction[1], imaging mass cytometry (IMC) has become a widely adopted imaging methodology for downstream tasks like cell phenotyping (i.e., identifying and categorizing different cell types) and analyzing the spatial landscape of cells. IMC is well suited for tissue analysis due to its ability to simultaneously detect over 40 protein markers using metal-tagged antibodies, bypassing signal from autofluorescence and spectral overlap limitations of fluorescence-based techniques like immunofluorescence microscopy. The process involves labeling tissue samples with antibodies conjugated to rare-earth metals (e.g., Neodymium *Nd* or Iridium *Ir*), followed by laser ablation at 1 μm^2 resolution to ionize tissue regions. These ions are then quantified via mass spectrometry, generating high-dimensional spatial cell marker data. The subcellular resolution enables mapping of cellular microenvironments, used for studying tumor heterogeneity, immune cell interactions, or tissue architecture.

A crucial step in existing frameworks or pipelines [2], [3] for IMC single cell analysis, is the segmentation of cellular structures on a single cell basis (instance segmentation). While this process is increasingly recognized as prone to errors[4], [5] and some approaches even try to circumvent the process of segmentation [6], [7], it remains a focus for methodological advancements: a recent study[8] investigates IMC cell segmentation performance on partially labeled data. Other research is aimed to develop segmentation models specifically tailored for IMC data[9], [10]. It can be derived that improving single cell segmentation using IMC is still a current topic.

This study establishes a reasonable upper bound for whole-cell segmentation on IMC data by quantifying inter-annotator agreement across cell masks independently annotated by four domain experts. This baseline reflects the inherent variability in human expert interpretations of cellular boundaries. The performance of four state-of-the-art generalist cell segmentation models is systematically evaluated on both in-house and publicly available IMC datasets, with segmentation performance compared against the established human-annotator benchmark. By comparing model performance with inter-annotator agreement, this work explores current limitations of automated segmentation methods for IMC whole cell segmentation.

The remainder of this study is structured as follows: Section 2 explores related works on inter-annotator agreement for cell instance segmentation, thereby positioning the current work within the context of existing research. In Section 3, the employed data, the segmentation models and the evaluation metrics are introduced. Section 4 presents and discusses the experimental results. Finally, Section 5 concludes this study.

2 Related work on inter-annotator agreement for cell instance segmentation

A recent work[11] provides an overview on how to assess inter-annotator agreement for medical image segmentation, including Kappa statistics, STAPLE-based and heatmap-based methods. While all of these methods constitute dedicated ways of expressing the inter-annotator agreement numerically, they are limited to semantic segmentation. A major challenge in applying similar techniques to instance segmentation problems on densely populated areas like cell tissue, is to find unambiguous element assignments between annotators, also considering over- and undersegmentation cases (i.e. one cell annotated as multiple cells or missing cell annotations). While for the two-annotator case, this could be treated as an assignment problem (as done for the sorted average precision metric described in section 3.3), this becomes a practically unsolvable problem quickly as the number of annotators increases. Authors in [12] approach this by distance-based consensus matching between cell centroids across multiple annotators. However, because their approach requires the average cell size as an input parameter and also their work is centered around H&E stained samples from histopathology, its applicability to other imaging modalities, such as IMC, remains uncertain. An established way[13], [14] for measuring inter-annotator agreement in the domain of cell instance segmentation is to systematically compare pairs of annotators using instance segmentation performance metrics. In [13], nuclei (as opposed to whole cells) annotations are compared using measures based on the calculation of intersection over union per element at a certain threshold. In [14], the authors evaluate the human-to-human whole cell segmentation performance for a variety of multiplexed tissue imaging modalities and tissue types, however not directly focused on IMC data. It can be deduced that assessing human-to-human performance on IMC whole cell segmentation represents a meaningful direction for systematic investigation, as current literature lacks benchmarks in this

specific domain. The term *inter-annotator agreement* in the context of this work refers to the human-to-human evaluation using segmentation performance metrics described in section 3.3.

3 Methods

3.1 Data

This section details the acquisition and preprocessing of a small-scale in-house IMC dataset and four employed external datasets, accompanied by a comprehensive overview of dataset statistics.

3.1.1 IMC sample generation

Samples were prepared from a high cell density lymphoid tissue according to standard IMC sample preparation techniques as suggested by the manufacturer [1]. The donor sample was collected according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of the Province of Salzburg: Ethics statement/Ethics number: 415-E/1287/18-2018, Version 5. Written informed consent was obtained from the donor. For sample preparation, the formalin fixed paraffin embedded tissue block was sectioned into 4 µm thick slices on a HM340E microtome (Epredia, Portsmouth, New Hampshire) and mounted to superfrost plus microscopic slides (Epredia). The mounted sections were baked at 60°C for one hour and immediately transferred to coblin jars containing UltraClear solution (VWR, Radnor, Pennsylvania) and incubated for 5 minutes. After a second 5 minute incubation in fresh UltraClear, the sample was incubated in an ethanol series of 2x100 percent, 95 percent, 80 percent ethanol and incubated for 3 minutes (first two incubations) and 2 minutes, each, followed by incubation in water for 5 minutes. For epitope retrieval the sample was transferred to pre-heated pH9 EDTA epitope retrieval buffer and heated in a KOS pathological microwave (Milestone, Valbremba, Italy) for 30 minutes at 96°C. After retrieval, the regular protocol was continued with a one hour blocking step using Superblock blocking solution (ThermoScientific, Waltham, Massachusetts). The antibody cocktail was incubated over night at 4°C. Iridium staining was performed at a dilution of 1:100 from a 125 µM stock (Standard Biotools, Markham, Ontario) for 10 minutes at room temperature.

[1] https://fluidigm.my.salesforce.com/sfc/p/#700000009DAw/a/4u000000dmhS/avZw9iOjvvsWnYP.pqXU5au3oBAZ_xuFo0uf7pWo4Nw

3.1.2 Mask generation

After laser ablation, the data was annotated by four domain experts. During the process, each annotator manually annotated 10 crops of 50 x 50 pixels resolution which were then used to extrapolate the annotations to the whole images. Ilastik[15] was used to create pixel probability maps which are subsequently transformed to single cell segmentation masks using Cellprofiler[16].

3.1.3 Dataset statistics

Additionally, four external datasets are used in this study, all of which provide IMC data and corresponding segmentation masks generated in a comparable manner as the in-house data. It should be noted, however, that the focus of these works was not to create a segmentation mask but it was merely an intermediate step towards the respective downstream task. All five datasets are listed in Table 1 along with their statistics.

3.1.4 Patching strategy and channel aggregation

Preliminary experiments suggested unstable results in terms of segmentation performance caused by the varying resolution of the individual datasets (see column 5 in Table 1). Hence, experiments in this work analyze the data in two modes:

- *whole image*: The full image is fed into the respective segmentation model and the full mask is evaluated afterwards.
- *sliding window patches*: Patches of size 256 x 256 pixels are cropped from the original image with an overlap of 128 pixels. For evaluation, a border of 28 pixels is ignored on all four sides, reducing the evaluated field of view per patch to 200 x 200 pixels. The border removal is meant to mitigate artifacts generated by partitions of cells at the border which potentially distort the results especially when using smaller crops. It should be noted, however, that from a practical perspective smaller patching is unfavorable for analysis of larger data because stitching masks is prone to errors caused by border cells.

Because all the models described in the upcoming section require both a nucleus channel and a membrane channel, IMC channels are aggregated using the arithmetic mean over a selection of markers on a per pixel basis. Table 2 presents a breakdown what markers are aggregated to which channel, along with additional details (used clone and dilution for the

Table 1: Overview datasets and their statistics. For patching, 68 samples from the Ali20 dataset and 14 samples from the Jackson20 dataset are ignored because at least one dimension is smaller than 256 pixels.

Dataset	Abbreviation	Tissue type	# Samples whole image	avg resolution whole image (y/x)	# Samples patches	# annotators per sample	avg # cell masks per patch
in-house	Annot1 – 4	Lymphoid	10	1000.0/1000.0	360	4	823.7
Ali20 [17]	A20	Breast	548	462.8/478.0	2787	1	314.0
Rendeiro21 [18]	R21	Lung	229	1108.4/1187.5	13361	1	185.3
Jackson20 [19]	J20	Breast	746	596.5/626.7	8714	1	320.5
Hoch22 [20]	H22	Melanoma	167	993.1/963.4	6361	1	467.4

Table 2: Aggregated channels and their corresponding antibody markers.

Category	Marker	Clone	Channel	Dilution (1:n)
Membrane Channel	CD19	6OMP31	Nd142	100
	TOMM20	EPR1581-54	Nd144	100
	CD5	CLDA5-1	Nd145	30
	CD4	EPR6855	Gd156	400
	CD68	KP1	Tb159	1000
	CD20	H1	Dy161	150
	CD8a	C8/144B	Dy162	800
	CD14	EPR3653	Dy163	1000
	CD45RA	HI100	Er166	600
	B2M	B2M/961	Yb171	200
	CD45RO	UCHL1	Yb173	1000
Nucleus Channel	DNA	–	Ir191	100
	DNA	–	Ir193	100

generation of the in-house data is also mentioned for better reproducibility) on the employed markers. To ensure a consistent and fair comparison across all experiments, the markers for channel aggregation were carefully selected based on analysis of all the datasets used in this study and their available markers.

3.2 Segmentation models

In total, four generalist single cell segmentation models are employed in this work without further fine-tuning of internal parameters. In the following, each model is briefly introduced:

- Cellpose v3[21]: Cellpose was introduced as an early generalist cell segmentation model. The Cellpose model got progressively enhanced by adding more training data. Currently, the "cyto3" model is its latest iteration for whole cell segmentation, which is also the one used in this work. Unlike other segmentation models at the time, which mostly employed a U-Net architecture and trained to directly yield a semantic segmentation map of nuclei, cell borders, and background, Cellpose predicts gradient vector fields that guide pixel assignments toward cell centers, enabling instance segmentation of diverse cell morphologies.

- DeepCell[14]: The DeepCell model employs a ResNet-50 architecture as its core that is connected to a feature pyramid network. It is especially aimed at nucleus and whole cell segmentation for tissue data. The version used in this work is *0.12.10*.

- CellSAM[22]: CellSAM is based on the Segment Anything Model (SAM)[23]. It automates segmentation via a transformer-based object detector that generates bounding box prompts. CellSAM is still in development at the time of this study. The used version number is *0.0.dev1*.

- VISTA-2D[24]: VISTA-2D is a recent generalist cell segmentation model developed by NVIDIA, which also uses the SAM as its core. The model is part of the MONAI framework[25]. The model delivers gradient vector fields similarly to the Cellpose model. The final segmentation masks are derived from these vector fields using Cellpose's postprocessing implementation. While the other models come with their internal way of contrast enhancement, the data was manually rescaled using histogram percentile clipping with limits 1% and 99%.

Models differ in the way the nucleus and membrane channels are arranged as an input tensor. Cellpose and DeepCell provide exemplary usages of the order to feed the channels into the model. Because CellSAM stems from the same lab as the DeepCell model, similar usage is assumed. VISTA-2D does not

come with a defined way how the nucleus and membrane channel should be assigned to the expected RGB input. Based on preliminary experimental validation, the three input channels are ordered as follows: an empty channel (all zeros), the membrane channel, and the nucleus channel.

3.3 Segmentation evaluation metrics

The right evaluation metric to choose for segmentation experiments heavily depends on the specific scenario. Existing works on cell instance segmentation rely on various metrics such as F1 score[14] (also reported as the technically identical dice score[8]) or mean average precision[26]. While the term mean average precision is widely used for evaluating object detection/newly trained models, it can be unclear how this metric translates to masks without confidence scores per object. A recent work[27] unravels this confusion by giving an overview on how different variants of "average precision" metrics are calculated. For evaluation of segmentation masks in this work, three metrics are employed, all of which rely on the calculation of the intersection over union (IoU) of cell areas on a per object basis. In coherence with [26], [27], the calculation for the "average precision" (AP) for a specific IoU threshold t_{IoU} is shown in equation 1. Note that this formula is closely related to the F1/Dice score. After assigning cell elements from the ground truth mask to cell elements on the prediction mask (TP), leftover (without counterpart) cells on the ground truth mask are viewed as false negatives (FN) and leftover cells on the prediction mask are treated as false positives (FP).

$$AP(t_{IoU}) = \frac{TP(t_{IoU})}{TP(t_{IoU}) + FN(t_{IoU}) + FP(t_{IoU})} \tag{1}$$

The metric ap50 corresponds to AP from equation 1 at threshold $t_{IoU} = 0.5$. The metric map computes the arithmetic mean of AP values at t_{IoU} points 0.5 to 0.95 with a step size of 0.05. The sortedAP[28] (sap) metric uses linear assignment optimization for finding corresponding segmentation objects based on their IoU, thereby allowing for IoU values below 0.5 to be included in the calculation. Further, the AP metric is calculated at every available IoU point. This is the equivalent of calculating the area on the whole AP/IoU curve. In fact, this metric is similar to the map metric, but instead of calculating the AP only at points {0.5, 0.55, ..., 0.95}, the AP is calculated at every step and also the arithmetic mean is used between steps. Figure 1 illustrates how the different metrics can be visually represented. The three metrics were chosen to assess performance not only at one fixed threshold but also over a spectrum of IoU values, providing a more comprehensive evaluation. Since ap50 already provides a single-threshold score (and F1/Dice would similarly focus on one cutoff), F1/Dice was excluded to avoid redundant evaluations.

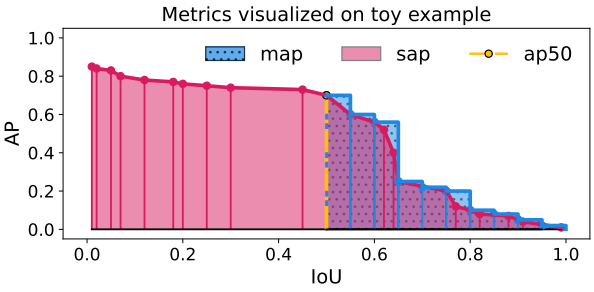

Figure 1: Exemplary depiction of the three metrics. The map metric constitutes the area under the AP/IoU curve over the interval $[0.5, 1.0]$. The sap metric corresponds to an approximation of the whole area under the AP/IoU curve. All values are exemplary.

4 Results and discussion

This section covers the experimental results. In section 4.1, the inter-annotator agreement is determined. Section 4.2 evaluates the performance of automatic segmentation via generalist models against the four expert-annotated masks obtained in a semi-automatic fashion. Section 4.3 evaluates the performance of the models on publicly available IMC datasets. The generalist models are used out-of-the-box without further fine tuning. Finally, Section 4.4 states limitations of the present study. For significance testing in Figures 4 & 5, data was checked for normality using the Shapiro-Wilk test and for homogeneity of variances using Bartlett's test. Based on these preliminary checks, if the data were normally distributed and variances were equal, an independent t-test was used; if normality was met but variances were unequal, Welch's t-test was applied; and if normality was not satisfied, the non-parametric Mann-Whitney U test was chosen. For all experiments significance tests yield p-values $< 10^{-6}$, i.e. statistically significant.

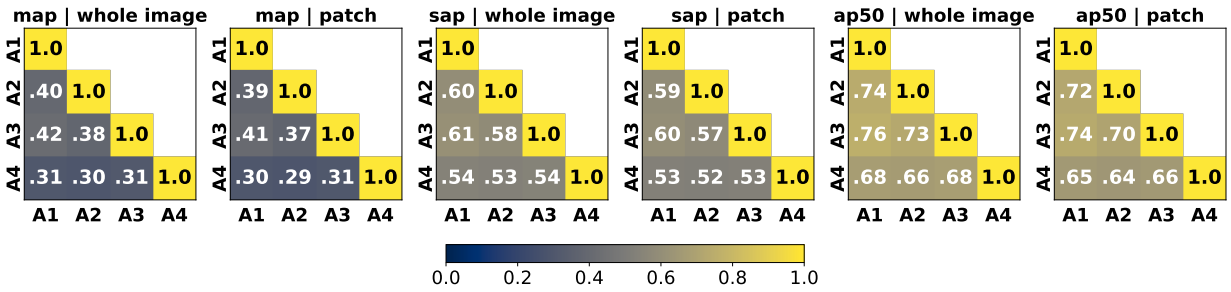

Figure 2: Pairwise inter-annotator comparison across different metrics and patch modes. Abbreviations A1 – A4 denote Annotators 1 to 4, respectively.

4.1 Inter-annotator agreement

Employing the evaluation metrics from section 3.3 and the cell masks from the in-house dataset annotated by four experts, the inter-annotator agreement is determined as a baseline. Figure 3 presents a direct comparison of each pair of annotators by displaying manually annotated single cell mask boundaries from a single patch. Although this work deals with single cell areas, the boundary is depicted here for better visualization of the differences. The first named annotator is depicted in blue, annotations from the second annotator are visualized in gold. Overlapping annotations are depicted in white. For visual context, a contrast enhanced version of the Ir191 metal tag channel is used as a background image. The patch, originally of resolution 50 x 50 pixels, is upscaled to 300 x 300 pixels using nearest neighbor interpolation for better visibility. While large parts of the respective cell boundaries are overlapping in a pixel-perfect manner, we can also observe a subjective bias where each annotator draws cellular borders. This results in minor spatial offsets (acceptable errors) but also leads to under- and oversegmentation, which constitute unfavorable errors.

Table 3: Average inter-annotator agreement over all annotators.

average map		average sap		average ap50	
whole	patch	whole	patch	whole	patch
.353	.345	.566	.557	.708	.686

Figure 2 lists detailed results for pairwise comparisons between expert annotators using an agreement matrix. Numbers in *whole image* matrices therefore show the arithmetic mean of 10 mask comparisons, numbers in the *patch* matrices represent the arithmetic mean of 360 mask comparisons. Finally, Table 3 shows the overall average over all pairwise comparisons per evaluation metric.

The averaged inter-annotator values from Table 3 can be viewed as a reasonable upper bound for automatic segmentation models. Because if a model would perform better on a set of annotated masks from a specific annotator, discrepancies would increase when evaluated against masks from other annotators. To achieve results that more accurately reflect reality, methods beyond manual annotation would be required for ground truth generation to mitigate human biases. Hence, average performance metrics are used as upper bounds in the following diagrams, with average values indicated by dashed lines. Note that these lines represent the arithmetic mean from all inter-annotator calculations and should thus be compared with the mean values from the boxplot diagrams to ensure a fair comparison.

4.2 Automatic segmentation using generalist models

Figure 4 shows the results for model performance on the in-house dataset for both modes, whole images and sliding window patches. Model outputs are always compared against segmentation masks from one annotator, indicated via abbreviations A1–A4. The SAM based models appear to face difficulties with the rather small cellular structures on the whole-image resolution. For smaller patches, the performance aligns well with the other two approaches. However, when compared to the human-to-human baseline, model output results in inferior segmentation performance across all three evaluation metrics.

4.3 Automatic segmentation on public datasets

Figure 5 presents the results of the model performance on the four external datasets. Here, it is assumed that the determined inter-annotator agree-

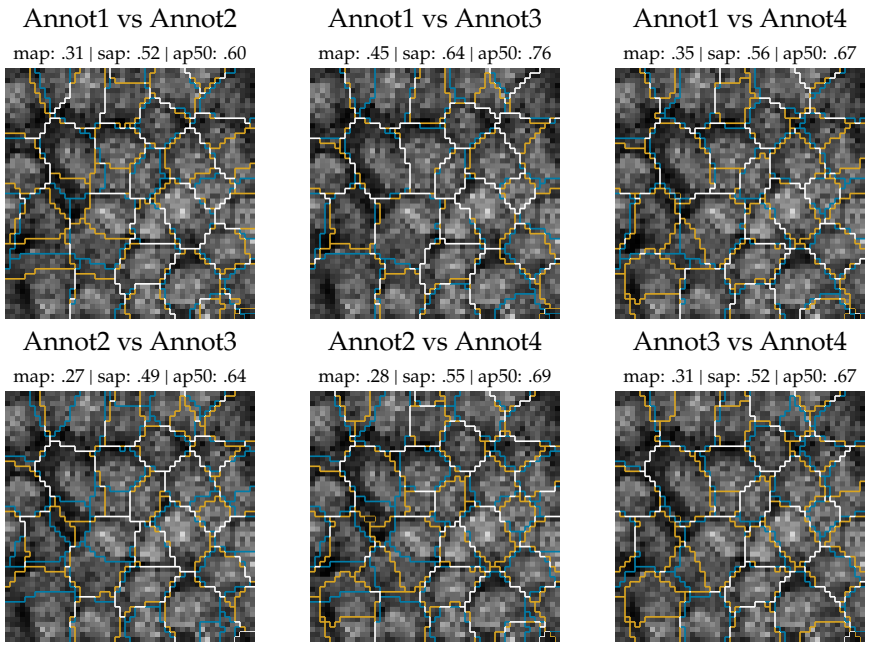

Figure 3: Visualization of annotation agreement between two annotators and the corresponding metrics per patch. Only cell borders are displayed for better visibility. Annotators colored in blue and golden. Annotation overlap in white. Best viewed in color.

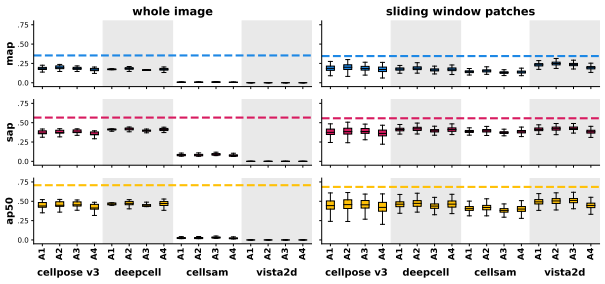

Figure 4: Model outputs evaluated against annotation masks on a per annotator basis.

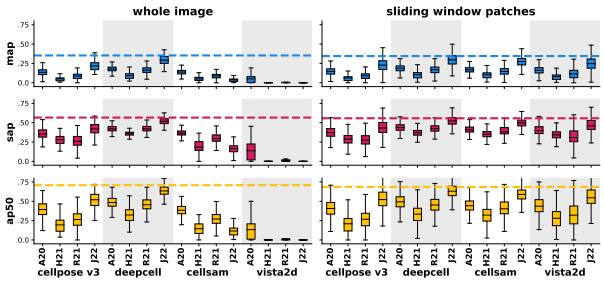

Figure 5: Model outputs evaluated on their available segmentation masks.

ment from section 4.1 can also be applied to the external datasets, although the used tissue types differ. The performance of VISTA-2D varies across datasets for the whole-image mode, which can be attributed to differences in average resolution per dataset as outlined in Table 1. The better performance of CellSAM on the whole images in comparison to the previous experiment using the in-house data can also be explained by the differences in image resolution. For the patching approach, all models perform best on the H22 dataset, yet none reach the reasonable upper bound established by inter-annotator agreement metrics, indicating systematic limitations in current segmentation algorithms. The results on external data reveal a larger variance in results compared to previous experiments, likely due to increased sample size capturing a broader biological heterogeneity.

4.4 Limitations of this study

While this work advances understanding of inter-annotator agreement and generalist model cell segmentation on IMC data, several constraints merit consideration for future research:

- Antibody channel aggregation: The impact of channel selection for nucleus & membrane/cytoplasm aggregation remains unexplored. While multi-channel IMC data inherently captures diverse biomarkers, the interplay

between channel combinations could influence segmentation robustness. Future studies should systematically evaluate how channel aggregation strategies affect model performance. Another recent approach[29] tries to process all channels separately, thereby avoiding the channel aggregation step entirely.

- Specialized IMC segmentation models: This study focused on generalist models. While generalist models provide a convenient way due to their out-of-the-box usage potential, IMC specific cell segmentation approaches such as [9] also hold great potential.

- IMC specific pre-processing: Optional processing steps such as hot pixel removal, spillover correction[30] and denoising[31] were not incorporated into the evaluation pipeline.

- Scale sensitivity in large images: Discrepancies between whole-image and sliding-window analyses suggest resolution-dependent detection challenges for small cells. Optimal crop sizing and super-resolution techniques (e.g. SpiDe-Sr[32]) could address this limitation by preserving fine-grained structures without sacrificing contextual information.

- Emerging generalist architectures: Some advances in foundation models like μSAM[33] and SAMCell[34], were not included in this study. Incorporating these models would strengthen the findings of this work due to their SAM based nature.

These limitations, however, do not undermine the core findings regarding inter-annotation agreement and IMC segmentation performance using generalist models but highlight pathways to refine segmentation pipelines.

5 Conclusion

This study first assessed the inter-annotator agreement for IMC single cell segmentation by comparing cell masks independently obtained by semi-automatic annotation from four domain experts. This presents a reasonable upper bound for automatic segmentation approaches. Afterwards, four generalist models are utilized to apply single cell segmentation of the same data. Results indicate that there is room left for improvement in the light of this upper bound. Additionally, models also were tasked to segment publicly available IMC data. Under assumption that this upper bound still holds true for the external data, similar conclusions regarding the demand for further improvement of the accuracy can be drawn. While this study includes a systematic analysis of inter-annotator agreement and generalist model performance on IMC data, there are things yet to be addressed such as employing specialized segmentation models.

Data and code availability.

External data sets can be found through their corresponding citations. The newly generated data and the accompanied annotations are publicly available at https://zenodo.org/records/15511299. The code is available at https://gitlab.cosy.sbg.ac.at/wavelab/imc-interannotator-agreement.

References

[1] C. Giesen, H. A. O. Wang, D. Schapiro, N. Zivanovic, A. Jacobs, B. Hattendorf, P. J. Schüffler, D. Grolimund, J. M. Buhmann, S. Brandt, Z. Varga, P. J. Wild, D. Günther, and B. Bodenmiller, "Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry," *Nature Methods*, vol. 11, no. 4, 2014.

[2] B. Hunter, I. Nicorescu, E. Foster, D. McDonald, G. Hulme, A. Fuller, A. Thomson, T. Goldsborough, C. M. U. Hilkens, J. Majo, L. Milross, A. Fisher, P. Bankhead, J. Wills, P. Rees, A. Filby, and G. Merces, "Optimal: An optimized imaging mass cytometry analysis framework for benchmarking segmentation and data exploration," *Cytometry Part A*, vol. 105, no. 1, 2023.

[3] J. Windhager, V. R. T. Zanotelli, D. Schulz, L. Meyer, M. Daniel, B. Bodenmiller, and N. Eling, "An end-to-end workflow for multiplexed image processing and analysis," en, *Nat. Protoc.*, vol. 18, no. 11, 2023.

[4] Y. Amitay, Y. Bussi, B. Feinstein, S. Bagon, I. Milo, and L. Keren, "Cellsighter: A neural network to classify cells in highly multiplexed images," *Nature Communications*, vol. 14, no. 1, 2023.

[5] Y. Lee, E. L. Y. Chen, D. C. H. Chan, A. Dinesh, S. Afiuni-Zadeh, C. Klamann, A. Selega, M. Mrkonjic, H. W. Jackson, and K. R. Campbell, "Segmentation aware probabilistic phenotyping of single-cell spatial protein expression data," *Nature Communications*, vol. 16, no. 1, 2025.

[6] S. Gutwein, D. Lazic, T. Walter, S. Taschner-Mandl, and R. Licandro, *Interpretable embeddings for segmentation-free single-cell analysis in multiplex imaging*, 2024. arXiv: 2411.03341 [eess.IV].

[7] S. Sultan, M. A. J. Gorris, E. Martynova, L. L. van der Woude, F. Buytenhuijs, S. van Wilpe, K. Verrijp, C. G. Figdor, I. J. M. de Vries, and J. Textor, "Immunet: A segmentation-free machine learning pipeline for immune landscape phenotyping in tumors by multiplex imaging," *Biology Methods and Protocols*, vol. 10, no. 1, 2024. eprint: https://academic.oup.com/biomethods/article-pdf/10/1/bpae094/61701720/bpae094.pdf.

[8] K. M. Bird, X. Ye, A. M. Race, and J. M. Brown, "Pushing the limits of cell segmentation models for imaging mass cytometry," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 2024.

[9] X. Xiao, Y. Qiao, Y. Jiao, N. Fu, W. Yang, L. Wang, R. Yu, and J. Han, "Dice-xmbd: Deep learning-based cell segmentation for imaging mass cytometry," *Frontiers in Genetics*, vol. 12, 2021.

[10] Y. Scuiller, P. Hemon, M. Le Rochais, J.-O. Pers, C. Jamin, and N. Foulquier, "Youpi: Your powerful and intelligent tool for segmenting cells from imaging mass cytometry data," *Frontiers in Immunology*, vol. 14, 2023.

[11] F. Yang, G. Zamzmi, S. Angara, S. Rajaraman, A. Aquilina, Z. Xue, S. Jaeger, E. Papagiannakis, and S. K. Antani, "Assessing inter-annotator agreement for medical image segmentation," *IEEE Access*, vol. 11, 2023.

[12] A. Capar, D. A. Ekinci, M. Ertano, M. K. K. Niazi, E. B. Balaban, I. Aloglu, M. Dogan, Z. Su, F. V. Aker, and M. N. Gurcan, "An interpretable framework for inter-observer agreement measurements in tils scoring on histopathological breast images: A proof-of-principle study," *PLOS ONE*, vol. 19, no. 12, 2024.

[13] C. Yapp, E. Novikov, W.-D. Jang, T. Vallius, Y.-A. Chen, M. Cicconet, Z. Maliga, C. A. Jacobson, D. Wei, S. Santagata, H. Pfister, and P. K. Sorger, "Unmicst: Deep learning with real augmentation for robust segmentation of highly multiplexed images of human tissues," *Communications Biology*, vol. 5, no. 1, 2022.

[14] N. F. Greenwald, G. Miller, E. Moen, A. Kong, A. Kagel, T. Dougherty, C. C. Fullaway, B. J. McIntosh, K. X. Leow, M. S. Schwartz, C. Pavelchek, S. Cui, I. Camplisson, O. Bar-Tal, J. Singh, M. Fong, G. Chaudhry, Z. Abraham, J. Moseley, S. Warshawsky, E. Soon, S. Greenbaum, T. Risom, T. Hollmann, S. C. Bendall, L. Keren, W. Graf, M. Angelo, and D. Van Valen, "Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning," en, *Nat Biotechnol*, vol. 40, no. 4, 2021.

[15] S. Berg, D. Kutra, T. Kroeger, C. N. Straehle, B. X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, M. Rudy, K. Eren, J. I. Cervantes, B. Xu, F. Beuttenmueller, A. Wolny, C. Zhang, U. Koethe, F. A. Hamprecht, and A. Kreshuk, "Ilastik: Interactive machine learning for (bio)image analysis," *Nature Methods*, vol. 16, no. 12, 2019.

[16] D. R. Stirling, M. J. Swain-Bowden, A. M. Lucas, A. E. Carpenter, B. A. Cimini, and A. Goodman, "Cellprofiler 4: Improvements in speed, utility and usability," *BMC Bioinformatics*, vol. 22, no. 1, 2021.

[17] H. R. Ali, H. W. Jackson, V. R. T. Zanotelli, E. Danenberg, J. R. Fischer, H. Bardwell, E. Provenzano, M. Al Sa'd, S. Alon, S. Aparicio, G. Battistoni, S. Balasubramanian, R. Becker, B. Bodenmiller, E. S. Boyden, D. Bressan, A. Bruna, B. Marcel, C. Caldas, M. Callari, I. G. Cannell, H. Casbolt, N. Chornay, Y. Cui, A. Dariush, K. Dinh, A. Emenari, Y. Eyal-Lubling, J. Fan, E. Fisher, E. A. González-Solares, C. González-Fernández, D. Goodwin, W. Greenwood, F. Grimaldi, G. J. Hannon, O. Harris, S. Harris, C. Jauset, J. A. Joyce, E. D. Karagiannis, T. Kovačević, L. Kuett, R. Kunes, A. Küpcü Yoldaş, D. Lai, E. Laks, H. Lee, M. Lee, G. Lerda, Y. Li, A. McPherson, N. Millar, C. M. Mulvey, F. Nugent, C. H. O'Flanagan, M. Paez-Ribes, I. Pearsall, F. Qosaj, A. J. Roth, O. M. Rueda, T. Ruiz, K. Sawicka, L. A. Sepúlveda, S. P. Shah, A. Shea, A. Sinha, A. Smith, S. Tavaré, S. Tietscher, I. Vázquez-García, S. L. Vogl, N. A. Walton, A. T. Wassie, S. S. Watson,

S. A. Wild, E. Williams, J. Windhager, C. Xia, P. Zheng, X. Zhuang, S.-F. Chin, and C. I. G. C. Team, "Imaging mass cytometry and multi-platform genomics define the phenogenomic landscape of breast cancer," *Nature Cancer*, vol. 1, no. 2, 2020.

[18] A. F. Rendeiro, H. Ravichandran, Y. Bram, V. Chandar, J. Kim, C. Meydan, J. Park, J. Foox, T. Hether, S. Warren, Y. Kim, J. Reeves, S. Salvatore, C. E. Mason, E. C. Swanson, A. C. Borczuk, O. Elemento, and R. E. Schwartz, "The spatial landscape of lung pathology during covid-19 progression," *Nature*, vol. 593, no. 7860, 2021.

[19] H. W. Jackson, J. R. Fischer, V. R. T. Zanotelli, H. R. Ali, R. Mechera, S. D. Soysal, H. Moch, S. Muenst, Z. Varga, W. P. Weber, and B. Bodenmiller, "The single-cell pathology landscape of breast cancer," *Nature*, vol. 578, no. 7796, 2020.

[20] T. Hoch, D. Schulz, N. Eling, J. M. Gómez, M. P. Levesque, and B. Bodenmiller, "Multiplexed imaging mass cytometry of the chemokine milieus in melanoma characterizes features of the response to immunotherapy," en, *Sci. Immunol.*, vol. 7, no. 70, 2022.

[21] C. Stringer and M. Pachitariu, "Cellpose3: One-click image restoration for improved cellular segmentation," 2024.

[22] U. Israel, M. Marks, R. Dilip, Q. Li, C. Yu, E. Laubscher, S. Li, M. Schwartz, E. Pradhan, A. Ates, M. Abt, C. Brown, E. Pao, A. Pearson-Goulart, P. Perona, G. Gkioxari, R. Barnowski, Y. Yue, and D. Van Valen, "A foundation model for cell segmentation," 2023.

[23] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, *Segment anything*, 2023. arXiv: 2304.02643 [cs.CV].

[24] NVIDIA, *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, 2024.

[25] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, V. Nath, Y. He, Z. Xu, A. Hatamizadeh, A. Myronenko, W. Zhu, Y. Liu, M. Zheng, Y. Tang, I. Yang, M. Zephyr, B. Hashemian, S. Alle, M. Z. Darestani, C. Budd, M. Modat, T. Vercauteren, G. Wang, Y. Li, Y. Hu, Y. Fu, B. Gorman, H. Johnson, B. Genereaux, B. S. Erdal, V. Gupta, A. Diaz-Pinto, A. Dourson, L. Maier-Hein, P. F. Jaeger, M. Baumgartner, J. Kalpathy-Cramer, M. Flores, J. Kirby, L. A. D. Cooper, H. R. Roth, D. Xu, D. Bericat, R. Floca, S. K. Zhou, H. Shuaib, K. Farahani, K. H. Maier-Hein, S. Aylward, P. Dogra, S. Ourselin, and A. Feng, *Monai: An open-source framework for deep learning in healthcare*, 2022. arXiv: 2211.02701 [cs.LG].

[26] J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin, M. Rohban, S. Singh, and A. E. Carpenter, "Nucleus segmentation across imaging experiments: The 2018 data science bowl," *Nature Methods*, vol. 16, no. 12, 2019.

[27] D. Hirling, E. Tasnadi, J. Caicedo, M. V. Caroprese, R. Sjögren, M. Aubreville, K. Koos, and P. Horvath, "Segmentation metric misinterpretations in bioimage analysis," *Nature Methods*, vol. 21, no. 2, 2024.

[28] L. Chen, Y. Wu, J. Stegmaier, and D. Merhof, "Sortedap: Rethinking evaluation metrics for instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2023.

[29] T. Goldsborough, A. O'Callaghan, F. Inglis, L. Leplat, A. Filby, H. Bilen, and P. Bankhead, "A novel channel invariant architecture for the segmentation of cells and nuclei in multiplexed images using instanseg," *bioRxiv*, 2024.

[30] Y. Bai, B. Zhu, X. Rovira-Clave, H. Chen, M. Markovic, C. N. Chan, T.-H. Su, D. R. McIlwain, J. D. Estes, L. Keren, G. P. Nolan, and S. Jiang, "Adjacent cell marker lateral spillover compensation and reinforcement for multiplexed images," *Frontiers in Immunology*, vol. 12, 2021.

[31] P. Lu, K. A. Oetjen, D. E. Bender, M. B. Ruzinova, D. A. C. Fisher, K. G. Shim, R. K. Pachynski, W. N. Brennen, S. T. Oh, D. C. Link, and D. L. J. Thorek, "Imc-denoise: A content aware denoising pipeline to enhance imaging mass cytometry," *Nature Communications*, vol. 14, no. 1, 2023.

[32] R. Chen, J. Xu, B. Wang, Y. Ding, A. Abdulla, Y. Li, L. Jiang, and X. Ding, "Spidesr: Blind super-resolution network for precise cell segmentation and clustering in spatial proteomics imaging," *Nature Communications*, vol. 15, no. 1, 2024.

[33] A. Archit, L. Freckmann, S. Nair, N. Khalid, P. Hilt, V. Rajashekar, M. Freitag, C. Teuber, G. Buckley, S. von Haaren, S. Gupta, A. Dengel, S. Ahmed, and C. Pape, "Segment anything for microscopy," *Nature Methods*, vol. 22, no. 3, 2025.

[34] A. D. VandeLoo, N. J. Malta, E. Aponte, C. van Zyl, D. Xu, and C. R. Forest, "Samcell: Generalized label-free biological cell segmentation with segment anything," *bioRxiv*, 2025.