Quantifying the Impact of Segmentation Inaccuracies on Cell Phenotyping in Imaging Mass Cytometry*

Johannes Schuiki[1][0000−0002−6197−2327], Markus Steiner[2,3][0000−0003−4424−6347], Heinz Hofbauer[1][0000−0003−2969−1848], Stephan Drothler[2,3,4][0000−0001−8187−5180], Giulia Pessina[2,3], Richard Greil[2,3][0000−0002−4462−3694], Nadja Zaborsky[2,3][0000−0002−9775−185X], and Andreas Uhl[1][0000−0002−5921−8755]

[1] Department of Artificial Intelligence and Human Interfaces, Paris-Lodron-University 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
Corresponding author: jschuiki@cs.sbg.ac.at

Abstract. Cell phenotyping, the process of assigning class labels to cells, usually relies on a segmentation preprocessing step. However, segmentation is prone to errors due to 2D image limitations and tool imperfections, which can propagate inaccuracies into phenotyping. This study systematically evaluates how different segmentation methods affect phenotyping by analyzing the resulting spatial cell phenotype landscapes. Our findings suggest that inaccuracies in single cell segmentation indeed impact the resulting landscape measurably.

Keywords: imaging mass cytometry · cell segmentation · cell phenotyping · inter-annotator agreement.

1 Introduction

Spatial proteomics techniques such as imaging mass cytometry (IMC) [4] enable simultaneous measurement of dozens of protein markers at subcellular resolution, offering detailed insights into tissue organization. One of the fundamental

* 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.

tasks in analyzing such data is cell phenotyping, i.e. assigning a class label to each cell. Here, cell segmentation is a crucial prerequisite to map protein marker data on individual cells. Naturally, inaccuracies in segmentation directly propagate to phenotyping, as marker signals from neighboring cells may polute the cell to be phenotyped. Because of these propagated inaccuracies, recent work tries to circumvent the segmentation step entirely [7,17]. However, segmentation is still a standard step for the process of cell phenotyping [8,20]. A recent study [12] quantifies the inter-annotator agreement between different annotators and also compares manual annotations to a variety of automatic segmentation tools. In the light of these findings the question arises whether the differences between manual phenotyping and potentially suboptimal automated methods are quantifiably different, i.e. measurably change with better segmentation or if it is sufficient to use masks deemed "good enough". Currently, literature on quantifying the effects of segmentation inaccuracies in spatial proteomics is very scarce. There is one recent work [1] which applies artificial perturbations on cell masks for CODEX [5] data and concludes that even minor perturbations affect the phenotyping outcome based on evaluating cell feature tables. By analyzing differences in cell frequency distributions and the resulting spatial cellular arrangements, we approach the quantification of segmentation errors from a different perspective. Furthermore, the present study (i) differs in the imaging modality (IMC vs CODEX) and (ii) uses a different tissue type, which has a higher cell density. Doing so, we provide complementary information to the existing work.

To express the impact that different segmentation masks have on the phenotyping result numerically, we first strive to find a method able to replicate ground truth phenotyping and extend it automatically to unseen masks. Second, we evaluate differences not only on a per-cell basis, but analyze the resulting phenotyping landscape. Comparison of the results to instance segmentation accordance allows to derive that mask inaccuracies measurably impact the phenotyping landscape.

2 Study Design

Fig. 1 illustrates the design of this study. To quantify the impact of single cell segmentation inaccuracies on the resulting phenotype landscape, we first apply a variety of cell segmentation strategies (including manual annotation) to the data samples as described in section 3.2. Per data sample, the different single cell segmentation masks are compared in a pair-wise fashion and the correspondence is numerically expressed using the sorted average precision metric (introduced in section 3.3). Further, a cell phenotype landscape is constructed by assigning a cell type to every individual cell from a given mask. To do so, we first evaluate a variety of approaches to do phenotyping automatically. Along with an out-of-the-box foundation model for spatial proteomics, we train classification models using a reference phenotyping obtained by a domain expert. We select the method with the highest accuracy and fix it for all automatic phenotyping inference runs. Approaches for automated cell phenotypnig are detailed in section 3.4.

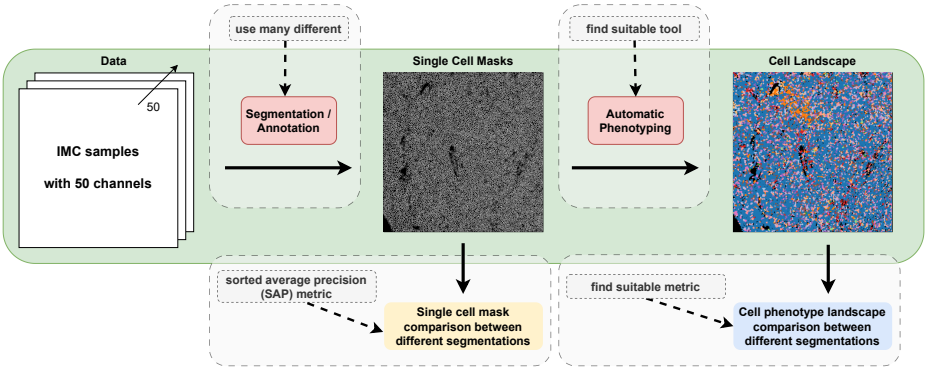

Fig. 1. Diagram illustrating the study design.

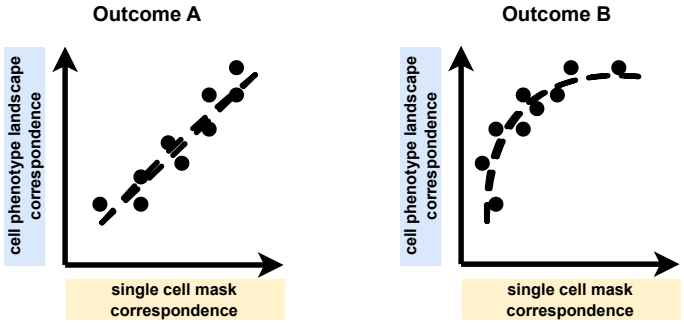

Fig. 2. Illustration of the two options for an expected outcome.

To compare resulting cell phenotype landscapes, we test two approaches for their applicability. Suitability is measured using validity checks. After selecting a suitable method to quantify the landscape similarity, we pair-wise compare cell-type landscapes per sample created using the different single cell segmentations. Ultimately, every pair of segmentation masks per sample will be compared on a single-cell and on a cell-type-landscape basis. Thus, the last objective is to identify correlations between both scores and perform a regression analysis to model the underlying trend.

Essentially, there are two reasonable outcomes of how the trend is expected to look. Fig. 2 visualizes both as an idealized sketch-diagram. Outcome A (left), a linear relationship, would indicate that even small inaccuracies in single cell masks propagate into a decrease in phenotype landscape overlap. Outcome B (right) depicts a relationship where landscape overlaps are consistently high regardless of changes in the single cell correspondence. This would indicate that errors in cell segmentation masks do not propagate as much into the phenotyping downstream task.

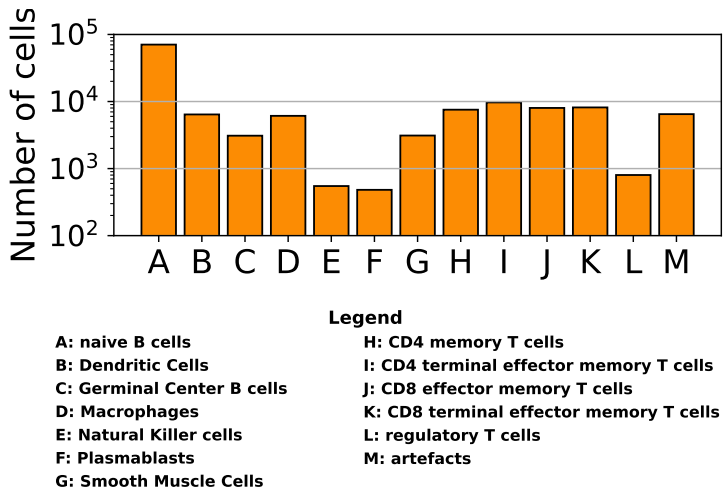

Fig. 3. Number of cells per phenotype across all samples from the dataset. The most frequent phenotype (A) occurs 70'597 times while phenotype F only appears 482 times. Log scale was applied to the y-axis.

3 Materials and Methods

3.1 Data

This study employs data drawn from a publicly available data repository, published in [12]. The data consists of 10 IMC samples captured with the Hyperion system. Each sample comprises 50 marker channels and has a resolution of 1000 × 1000 pixels. One pixel corresponds to the physical length of 1μm. All samples depict tissue sections of human lymphoid tissue. To generate ground truth for cell phenotyping, the steinbock framework [20] together with FlowSOM [18] was used to determine a phenotype per cell. The number of cells for the phenotype ground truth are shown in Fig. 3.

3.2 Single Cell Segmentation Methods

To obtain various single cell segmentation masks for each data sample, different strategies are used. Four manual annotations, annotated by four experts, are available as download together with the data samples itself. Beyond manual annotation, segmentation masks were generated using four out-of-the-box available generalist cell segmentation models in the same fashion as done in [12]: Cellpose v3 [16], CellSAM [10], VISTA-2D [11] and DeepCell Mesmer [6]. As shown in [12], the two SAM based models fail to generate meaningful masks at full image resolution (1000 × 1000 pixels). Hence, a patching strategy is applied where 49 partly overlapping patches at resolution 250 × 250 are extracted, segmented individually and stitched afterwards. The global segmentation mask is assembled by stitching central crops from each patch. For boundary regions,

the boundary patch edges are preserved. Overlap artifacts are handled post-hoc by defining resulting connected components as single cells. Because a variety in segmentation masks is desirable for this study, inaccuracies introduced during the patch stitching procedure are a welcome effect. Therefore, to artificially increase the number of varying masks, the patching strategy is also applied on the Cellpose and DeepCell model outputs. Additionally, one DeepCell Mesmer segmentation mask is available which was generated whilst creating a phenotying ground truth. Note that although the Mesmer model was used two times, once within the steinbock framework and once as a standalone model in Python, their input configurations (most notably: which channels were aggregated into nucleus and membrane channel, respectively) slightly differ, which results in minor differences. An overview of available single cell segmentation masks is given in Table 1.

Table 1. Sources for segmentation masks and abbreviations.

Single Cell Segmentation Source	**Abbreviation**
Manual Annotator 1	M1
Manual Annotator 2	M2
Manual Annotator 3	M3
Manual Annotator 4	M4
DeepCell Mesmer in steinbock framework	DC-F
DeepCell Mesmer standalone	DC
DeepCell Mesmer standalone stitched	DC-S
CellSAM stitched	CS-S
Cellpose v3	CP
Cellpose v3 stitched	CP-S
Vista2D stitched	V2D-S

3.3 Single Cell Segmentation Comparison Metric

To evaluate the agreement of two single-cell-level masks for the same data sample, we employ the sorted average precision (SAP) as described in Chen et al.[2]. SAP relies on the calculation of the intersection over union (IoU) of cell areas on a per object basis. The IoU value per element pair (i.e. one element from either mask under evaluation) is stored in an IoU-matrix. To find a weight-optimal (yet deterministic) correspondence of elements from both masks, Hungarian-style assignment optimization is applied. The number of assignment elements are viewed as true positives (TP), leftover (without counterpart) cells in the first mask are viewed as false negatives (FN) and leftover cells on the second mask are treated as false positives (FP). Because the metric is symmetric, ordering of masks is inconsequential. The values of assigned element pairs are sorted in ascending order and viewed as "thresholds" t_n. The average precision (AP), as defined

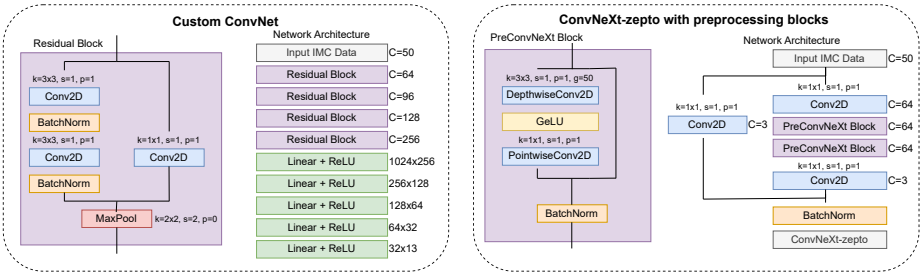

Fig. 4. (left) Custom network visualization. Dimensions of the first linear layer shown for 32×32 pixel patches; vary for 16×16. (right) ConvNeXt architecture visualization. To match our 50-dimensional data to the 3-channel input, a projection header was added upstream.

in equation 1, is evaluated at every threshold and multiplied by the threshold difference, thereby computing the area under the AP/IoU curve.

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

3.4 Automated Cell Phenotype Classification

To transfer the phenotyping style from the manually determined ground truth cell phenotypes to unseen segmentation masks, an automatic phenotype classification method is needed. Within this study, four approaches are considered to do so:

- KRONOS feature embedding: KRONOS [13], a recent foundation model for spatial proteomics is used as a specialized pre-trained feature extractor. Although not directly trained on IMC data, the model is able to handle input of variable channel dimension. To do so, every channel is fed into the model with a channel encoding. For every channel, the model delivers a 384-dimensional vector embedding describing the input patch. The data used in this study has 50 channels, 32 of which can be used as input to the KRONOS model. The resulting 12'288-dimensional feature embedding undergoes dimensionality reduction to a 384-dimensional vector using principal component analysis (PCA). Finally, using leave one out cross validation (LOOCV), a logistic regression classifier is trained on cells from nine of the ten available images, respectively, and evaluated on the tenth. Because the class distribution is highly imbalanced, 482 cells (number of cells for phenotype F in Fig. 3) are randomly drawn from all classes for training. Multiple identical classifiers are trained and the result is averaged to account for random subsampling.
- Average marker expression feature embedding: At every cell location, every marker expression (i.e. channel) is collapsed into a single value using the arithmetic mean. Hence, every cell is described via a d-dimensional vector,

where d corresponds to the number of markers used for a certain experiment (either 32 or 50). To achieve cell phenotyping, a logistic regression classifier is trained in a similar fashion as for the PCA-reduced KRONOS embeddings.
- ConvNeXt model: As a third option, a pre-trained ConvNeXt-zepto [9], taken from the Torch Image Model library [19], is employed. To fully leverage the pre-trained weights as a starting point while only allowing 3 input channels (the model was trained on the ImageNet [3] dataset which is in RGB format), we add a few residual Conv2D blocks before to distill the relevant information into three input channels as shown in Fig. 4 (right). Additionally, we set the stride of the first ConvNeXt block to 1 to counter the rapid downscaling of our already low-resolution data samples. An image-wise LOOCV is performed, i.e. train on all cells from nine images and test on all cells from the tenth image, respectively. All parameters of the combined model are trained using for 150 epochs using the AdamW optimizer with a learning rate of 2×10^{-3} and a weight decay of 0.05. A CosineAnnealing learning rate schedule is applied. To address class imbalance, a weighted random sampler is employed during data sampling.
- Custom deep learning architecture: As a fourth variant, a custom deep learning model is trained from scratch which employs convolutional kernels and residual connections. The network architecture is shown in Fig. 4 (left). The custom model uses a similar training configuration as the ConvNeXt adaption.

All approaches except the averaged marker expression features are tested for patch sizes of 16 and 32 pixels and for both masked cells (M) and in unmasked mode (U). The averaged marker expression method is tested using all marker channels (A) and also using the same reduced marker set (R) which is also available when using the KRONOS model. Further, different channel normalization techniques are tested (int, max & none).

3.5 Phenotype Landscape Comparison Metrics

To compare phenotyping results obtained from varying segmentation masks, two strategies are evaluated. For the first approach, a graph is constructed by viewing every cell as a node and spanning edges to other nodes where the cells are adjacent in the segmentation masks. Every node receives the corresponding cell phenotype as label. Subsequently, two labeled graphs are compared using Weisfeiler-Lehman kernel (WL) graph comparison [14]. Local neighborhoods are aggregated for 3 iterations. To compute the WL similarity, the GraKeL [15] Python library is used. As a second approach, the resulting cell phenotyping landscape is treated as a semantic segmentation result and compared using the average Dice similarity coefficient per cell class (DSC). To determine which method is more suitable in our scenario, three validity checks are included: i) an artificial (random) segmentation, generated using 13k random points on the unit square to generate a Voronoi grid followed by random assignment of cell labels following a uniform distribution (RND), ii) a similar Voronoi grid strategy

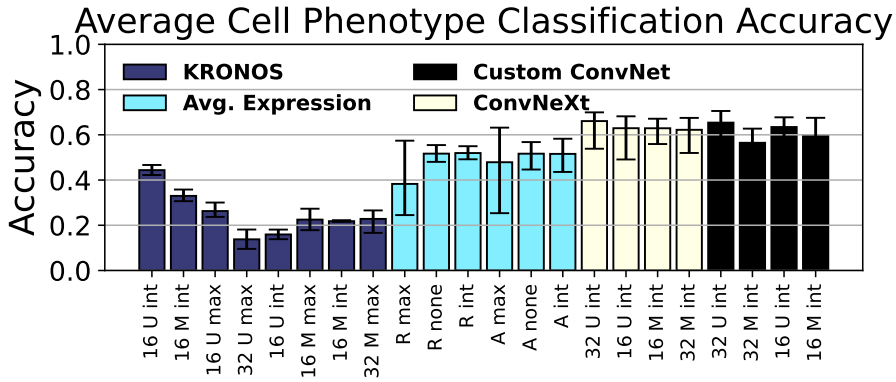

Fig. 5. Phenotype classification accuracy obtained using the classification approaches described in section 3.4.

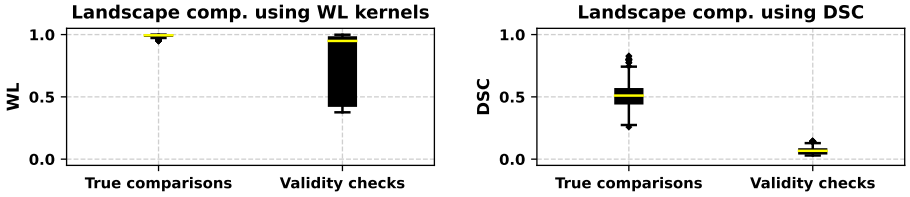

Fig. 6. Score distribution of true comparisons and validation checks for the WL kernel comparison (left) and DSC comparison (right).

but cell types are assigned via sampling from the same cell type distribution as the ground truth (RND-I; *informed*) and iii) random permutation of the ground truth phenotype assignment on the original mask (GT-P; *permuted*).

4 Results

Within this section, both the tool selection results as well as the main result are presented. Sections 4.1 and 4.2 include results for tool selection and section 4.3 presents the the main results of the study.

4.1 Comparison of Automated Cell Phenotyping Methods

As described in section 3.4, four main variants are tested to transfer the expert-labeled phenotyping to unseen masks in a variety of parameter setups. Results are shown in Fig. 5. Logistic regression classifiers for the KRONOS embeddings and average marker expressions per cell were trained multiple times because of the under-sampling for balanced training. Deep learning network training was

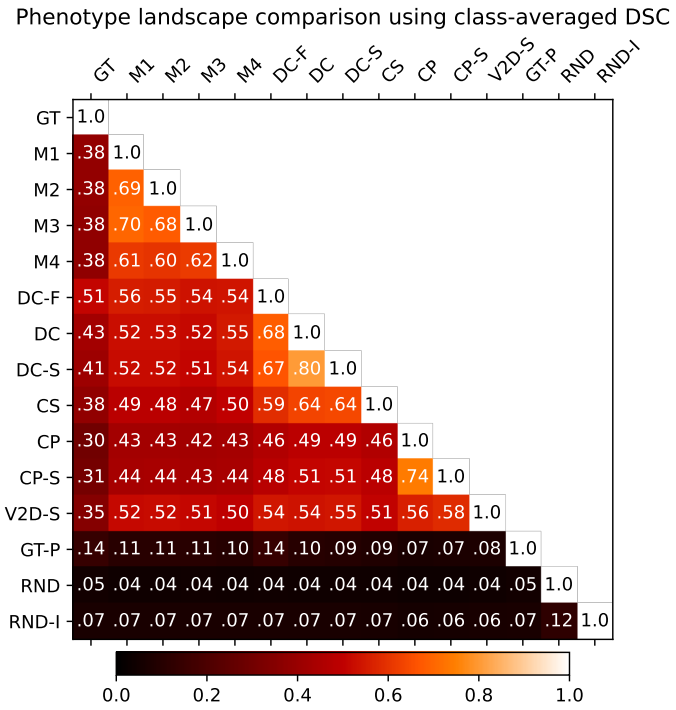

Fig. 7. Pairwise comparison of phenotype landscape using DSC averaged over all phenotypes and all 10 samples per value.

done once in a leave one sample out fashion and results are averaged over all 10 samples. The deep learning models outperform other approaches in terms of classification accuracy. Among them, the unmasked 32-pixel version for the ConvNeXt model shows achieves the highest average accuracy of 66.06%. Whilst not achieving a perfect accuracy score, the model is far from random (random performance for 13 classes would be 7.69%) and the train accuracy reached almost 100% which implies that the models, in theory, would have enough capacity for the dataset but rather indicates label inaccuracies. The inaccuracies were deemed acceptable for our experiments and, thus, this model was chosen for phenotyping on all single cell masks in this study.

4.2 Comparison of Phenotype Landscape Comparison Metrics

To compare two phenotyping landscapes stemming from different underlying segmentations, two approaches were tested as described in section 3.5. Suitability was assessed by looking at the score distribution of *true* comparison scores (i.e. obtained scores from comparing real samples) and scores obtained via comparing real samples with validation checks. Fig. 6 depicts the score distributions of real

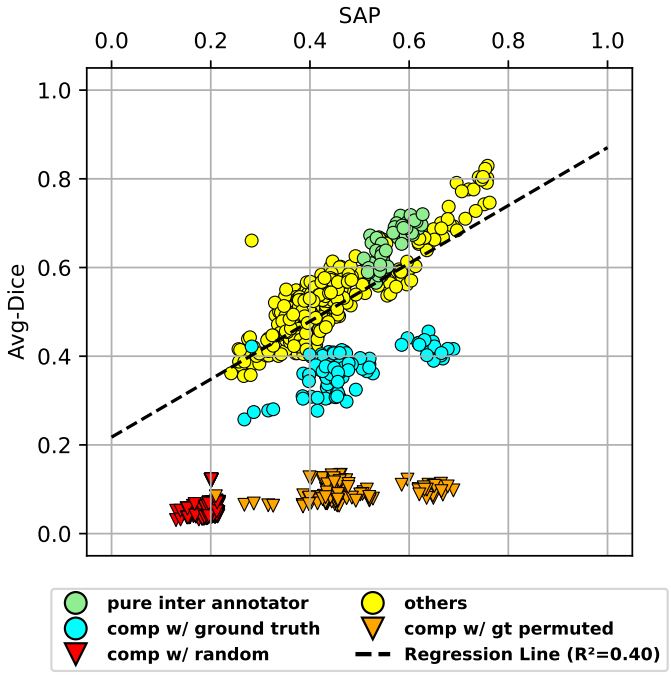

Fig. 8. Relation of DSC and SAP single cell segmentation overlap.

comparisons and validation checks for both considered approaches. For the WL kernel approach (left), the true comparison scores are altogether extremely high (almost reaching 1.0) and also the validation check scores spread into the same range. For the DSC approach (right), true scores are spread across the range with a median of 0.5. More importantly, the validation check scores are consistently in a lower range. Given the DSC approach's superior ability to distinguish between real samples and validation checks while also keeping the true scores in a broader, possibly more interpretable, range, we selected this method for the remainder of the study. An overview of all pair-wise scores (averaged over all 10 samples) is shown in Fig. 7. Ground truth (GT) here means the expert-labeled phenotyping. Note that DC-F used the same segmentation mask as GT but the resulting phenotyping landscape differs due to the inaccuracies in the automatic phenotyping method (section 4.1). One can see that similar masks achieve higher similarity scores (e.g. M1-M4 or the full vs stitched variants), while unrelated mask sources have lower correspondences. All the validity checks (GT-P, RND, RND-I) have low correspondence to any other phenotyping result. Thus, the DSC metric was chosen for landscape comparison in this study.

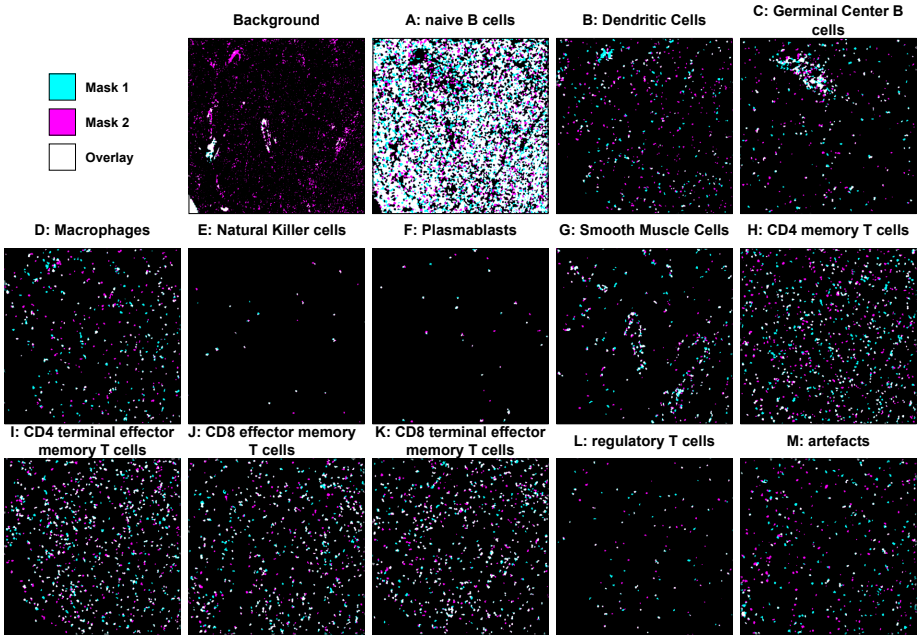

Fig. 9. Exemplary phenotype landscape visualization using results from two segmentation masks on the same sample. Source mask 1: Annot1, Source mask 2: deepcell; sample id: 50. SAP: 0.44, avg-DSC: 0.44.

4.3 Effect of Cell Segmentation Errors on Phenotype Landscape

To answer the central research question of this study, we explore the relation between SAP and DSC for all pairs of masks. A scatter-plot can be seen in Fig. 8. The validation checks (any mask compared with random masks or the cell-type-permuted ground truth mask; color-coded in red and orange respectively) are also included as downward facing triangles for reference. The general trend suggests that indeed a lower single cell segmentation mask correspondence leads to differences in the phenotyping landscape which behaves almost linear. This corresponds to *outcome A* as defined in Fig. 2. The results demonstrate a positive correlation (r=0.63) between single-cell segmentation accuracy and the resulting phenotyping landscape agreement. With segmentation accuracy explaining approximately 40% of the variance in landscape similarity (R2=0.40), these data suggest that higher single cell segmentation precision yields a measurable improvement in the correspondence of the spatial landscape. To convey some intuition how these landscapes look, Fig. 9 visualizes the resulting cell types of two single cell segmentation sources and color-codes them. The illustration comprises the resulting landscape for one manual annotation (M1, color coded in cyan) and resulting landscape obtained from the deepcell mask (DC, color coded in magenta). Overlaps colored in white.

5 Conclusion

This study investigated the effects of single cell segmentation inaccuracies on the downstream task of cell phenotyping in IMC. Results indicate a trend which suggests that even slight deviations in segmentation masks result in measurably different phenotyping results. These findings imply that it is not sufficient to have masks "good enough" but having precise segmentation masks is indeed desirable. This outcome also aligns well with the findings from the related literature. However, current evaluation pipeline inaccuracies (most notably automatic phenotyping accuracy trained in expert labeled ground truth) only allow to derive a preliminary trend and will be addressed in future work.

Data and code availability.

The data and manual single cell annotations, initially published in [12], are openly available at https://zenodo.org/records/15511299. Core scripts used for this study are available at https://gitlab.cosy.sbg.ac.at/wavelab/imc-impact-seg-error-on-pheno.

References

1. Bruhns, M., Schleicher, J.T., Wirth, M., Zago, M., Babaei, S., Claassen, M.: Effects of segmentation errors on downstream-analysis in highly-multiplexed tissue imaging. PLOS Computational Biology **21**(9), 1–18 (09 2025). https://doi.org/10.1371/journal.pcbi.1013350
2. Chen, L., Wu, Y., Stegmaier, J., Merhof, D.: Sortedap: Rethinking evaluation metrics for instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 3923–3929 (October 2023)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
4. Giesen, C., Wang, H.A.O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P.J., Grolimund, D., Buhmann, J.M., Brandt, S., Varga, Z., Wild, P.J., Günther, D., Bodenmiller, B.: Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Nature Methods **11**(4), 417–422 (Apr 2014). https://doi.org/10.1038/nmeth.2869
5. Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., Nolan, G.P.: Deep profiling of mouse splenic architecture with codex multiplexed imaging. Cell **174**(4), 968–981.e15 (Aug 2018). https://doi.org/10.1016/j.cell.2018.07.010
6. Greenwald, N.F., Miller, G., Moen, E., Kong, A., Kagel, A., Dougherty, T., Fullaway, C.C., McIntosh, B.J., Leow, K.X., Schwartz, M.S., Pavelchek, C., Cui, S., Camplisson, I., Bar-Tal, O., Singh, J., Fong, M., Chaudhry, G., Abraham, Z., Moseley, J., Warshawsky, S., Soon, E., Greenbaum, S., Risom, T., Hollmann, T., Bendall, S.C., Keren, L., Graf, W., Angelo, M., Van Valen, D.: Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. Nat Biotechnol **40**(4), 555–565 (Nov 2021)

7. Gutwein, S., Lazic, D., Walter, T., Taschner-Mandl, S., Licandro, R.: Nextmarker: Contrastive learning for marker-level interpretability in single-cell multiplex imaging. In: Proceedings of the COMputational PAthologY and muLtimodal data MICCAI Workshop (2025), accepted paper
8. Hunter, B., Nicorescu, I., Foster, E., McDonald, D., Hulme, G., Fuller, A., Thomson, A., Goldsborough, T., Hilkens, C.M.U., Majo, J., Milross, L., Fisher, A., Bankhead, P., Wills, J., Rees, P., Filby, A., Merces, G.: Optimal: An optimized imaging mass cytometry analysis framework for benchmarking segmentation and data exploration. Cytometry Part A **105**(1), 36–53 (Oct 2023). https://doi.org/10.1002/cyto.a.24803
9. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11976–11986 (June 2022)
10. Marks, M., Israel, U., Dilip, R., Li, Q., Yu, C., Laubscher, E., Iqbal, A., Pradhan, E., Ates, A., Abt, M., Brown, C., Pao, E., Li, S., Pearson-Goulart, A., Perona, P., Gkioxari, G., Barnowski, R., Yue, Y., Van Valen, D.: Cellsam: a foundation model for cell segmentation. Nature Methods **22**(12), 2585–2593 (Dec 2025). https://doi.org/10.1038/s41592-025-02879-w
11. NVIDIA: VISTA-2D: A foundational model for cell segmentation in spatial omics workflows. https://github.com/Project-MONAI/VISTA/tree/main/vista2d (2024), version 0.3.0
12. Schuiki, J., Steiner, M., Hofbauer, H., Drothler, S., Pessina, G., Greil, R., Zaborsky, N., Uhl, A.: Quantifying inter-annotator agreement and generalist model limitations in imaging mass cytometry single cell segmentation. In: Ali, S., Hogg, D.C., Peckham, M. (eds.) Medical Image Understanding and Analysis. pp. 146–159. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-98694-9_11
13. Shaban, M., Chang, Y., Qiu, H., Yeo, Y.Y., Song, A.H., Jaume, G., Wang, Y., Weishaupt, L.L., Ding, T., Vaidya, A., Lamane, A., Shao, D., Zidane, M., Bai, Y., McCallum, P., Luo, S., Wu, W., Wang, Y., Cramer, P., Chan, C.N., Stephan, P., Schaffenrath, J., Lee, J.L., Michel, H.A., Tian, C., Almagro-Perez, C., Wagner, S.J., Sahai, S., Lu, M.Y., Chen, R.J., Zhang, A., Gonzales, M.E.M., Makky, A., Lee, J.Y.J., Cheng, H., Ahmar, N.E., Matar, S., Haist, M., Phillips, D., Tan, Y., Nolan, G.P., Burack, W.R., Estes, J.D., Liu, J.T.C., Choueiri, T.K., Agarwal, N., Barry, M., Rodig, S.J., Le, L.P., Gerber, G., Schürch, C.M., Theis, F.J., Kim, Y.H., Yeong, J., Signoretti, S., Howitt, B.E., Loo, L.H., Ma, Q., Jiang, S., Mahmood, F.: A foundation model for spatial proteomics (2025), https://arxiv.org/abs/2506.03373
14. Shervashidze, N., Schweitzer, P., van Leeuwen, E.J., Mehlhorn, K., Borgwardt, K.M.: Weisfeiler-lehman graph kernels. Journal of Machine Learning Research **12**(77), 2539–2561 (2011), http://jmlr.org/papers/v12/shervashidze11a.html
15. Siglidis, G., Nikolentzos, G., Limnios, S., Giatsidis, C., Skianis, K., Vazirgiannis, M.: Grakel: A graph kernel library in python. Journal of Machine Learning Research **21**(54), 1–5 (2020)
16. Stringer, C., Pachitariu, M.: Cellpose3: one-click image restoration for improved cellular segmentation. Nat. Methods **22**(3), 592–599 (Mar 2025). https://doi.org/10.1038/s41592-025-02595-5
17. Sultan, S., Gorris, M.A.J., Martynova, E., van der Woude, L.L., Buytenhuijs, F., van Wilpe, S., Verrijp, K., Figdor, C.G., de Vries, I.J.M., Textor, J.: Immunet: a segmentation-free machine learning pipeline for immune landscape phenotyping in

tumors by multiplex imaging. Biology Methods and Protocols **10**(1), bpae094 (12 2024). https://doi.org/10.1093/biomethods/bpae094
18. Van Gassen, S., Callebaut, B., Van Helden, M.J., Lambrecht, B.N., Demeester, P., Dhaene, T., Saeys, Y.: FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. Cytometry A **87**(7), 636–645 (Jan 2015)
19. Wightman, R.: Pytorch image models. https://github.com/rwightman/pytorch-image-models (2019)
20. Windhager, J., Zanotelli, V.R.T., Schulz, D., Meyer, L., Daniel, M., Bodenmiller, B., Eling, N.: An end-to-end workflow for multiplexed image processing and analysis. Nat. Protoc. **18**(11), 3565–3613 (Nov 2023)