cp_measure: API-first feature extraction for image-based profiling workflows

Alán F. Muñoz [1] **Tim Treis** [2 1] **Alexandr A. Kalinin** [1] **Shatavisha Dasgupta** [1] **Fabian Theis** [2] **Anne E. Carpenter** [1] **Shantanu Singh** [1]

Abstract

Biological image analysis has traditionally focused on measuring specific visual properties of interest for cells or other entities. A complementary paradigm gaining increasing traction is image-based profiling - quantifying many distinct visual features to form comprehensive profiles which may reveal hidden patterns in cellular states, drug responses, and disease mechanisms. While current tools like CellProfiler can generate these feature sets, they pose significant barriers to automated and reproducible analyses, hindering machine learning workflows. Here we introduce cp_measure, a Python library that extracts CellProfiler's core measurement capabilities into a modular, API-first tool designed for programmatic feature extraction. We demonstrate that cp_measure features retain high fidelity with CellProfiler features while enabling seamless integration with the scientific Python ecosystem. Through applications to 3D astrocyte imaging and spatial transcriptomics, we showcase how cp_measure enables reproducible, automated image-based profiling pipelines that scale effectively for machine learning applications in computational biology.

1. Introduction

High throughput screening of biological phenomena via complex modalities, such as RNA sequencing, is prohibitively expensive, therefore microscopy is an efficient first step. Through microscopy, modern biologists use a wide array of fluorescence dyes or proteins to observe the location and distribution of cells, organelles, and other components. Nowadays, quantification of biological images is standard, with software often identifying regions of interest (such as cells) and extracting features that represent these regions, such as intensity.

Image-based profiling–also termed morphological profiling–is a technique of measuring an extensive suite of morphological features for a population of biological objects, such as cells. These features are fed into statistical or machine learning methods to identify biologically meaningful patterns. One of its biggest applications of morphological profiling is drug discovery, where scientists leverage microscopy's low acquisition cost and high throughput to accomplish many goals, such as grouping genes by function, identifying chemical compounds that target a protein, and predicting toxicity of drug candidates (Chandrasekaran et al., 2021).

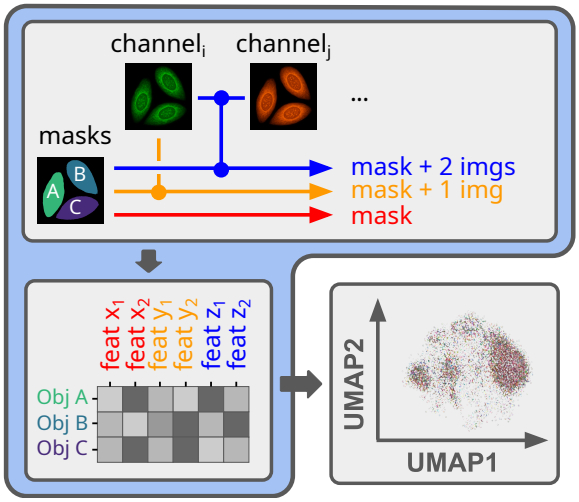

Figure 1. cp_measure generates features from images by using information in every region of interest ("object"). It can featurize the pairwise combination of all the available channels (colours) and objects. The resultant matrices represent the entire experiment and can be studied using statistical, machine, and deep learning methods.

1.1. The current state of bioimage analysis

The most widely used software for processing high-throughput biological images is CellProfiler (Stirling et al., 2021), providing experimental biologists with limited pro-

[1]Broad Institute of MIT and Harvard, United States [2]Institute of Computational Biology, Helmholtz Zentrum München, Germany. Correspondence to: Alán F. Muñoz <amunozgo@broadinstitute.org>, Shantanu Singh <shantanu@broadinstitute.org>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

gramming expertise an accessible yet powerful toolset. By contrast, it may not suit the needs of computational biologists building high throughput pipelines comprised of multiple tools. CellProfiler is ideal for creating and iteratively adapting manually-defined workflows; for low-customization high-throughput analyses such as extracting features for image-based profiles or spatial transcriptomics data, however, certain aspects become a hindrance. For example, complementing CellProfiler with other image analysis tools is a struggle, as building plug-ins to add functionality is a time and effort-consuming challenge that requires an understanding of its interfaces. Lastly, CellProfiler depends on many Python and Java packages, increasing the likelihood of dependency conflicts. Containers mitigate this problem, but not without increased complexity and caveats.

Alternative tools can extract features for image-based profiles, such as scikit-image (van der Walt et al., 2014), ScaleFex, or SpaCR (Comolet et al., 2024; EinarOlafsson, 2025), however, these approaches present trade-offs for different use cases. Their independent implementation means they generate different feature sets than CellProfiler, which can be advantageous for novel applications but complicates direct comparison with established datasets like those in the Cell Painting Gallery (Weisbart et al., 2024). Additionally, while all tools necessarily impose input requirements, some enforce rigid organizational constraints—demanding specific directory structures and naming conventions—that require extensive data reorganization before analysis, beyond the standard data type and format specifications. Some alternatives also target specific deployment scenarios, with ScaleFex designed primarily for cloud environments, limiting utility for researchers preferring local computation.

1.2. Engineered features in the era of deep learning

While computer vision has largely transitioned to deep learning-based feature extraction, biological imaging presents unique challenges that maintain the relevance of engineered features. These images differ substantially from natural images in their acquisition methods, visual properties, and analytical objectives, often limiting the effectiveness of standard pre-trained models, even those trained on biological data.

The application of deep learning to feature extraction for biological image analysis shows mixed results. When appropriately trained on domain-specific data, deep learning networks can outperform engineered features (Lafarge et al., 2019; Moshkov et al., 2022; Chow et al., 2022; Wolf et al., 2018), but not universally (Tang et al., 2024; Kim et al., 2023). Effective deep learning feature extractors for biological applications typically require substantial domain-specific datasets and careful model architecture choices, rather than simple transfer from general computer vision models.

Several practical considerations favour engineered features in biological imaging. First, interpretability remains crucial for biological discovery—researchers need to understand which morphological characteristics drive observed phenotypes. While deep learning features can be powerful predictors, their interpretation requires sophisticated techniques that may not provide the mechanistic insights biologists seek (Li et al., 2023). Engineered features have clear mathematical definitions that directly translate to biological concepts, such as protein co-localisation or nuclear morphology changes (Garcia-Fossa et al., 2023).

Second, many biological imaging laboratories lack the GPU resources, large annotated datasets, or machine learning expertise required for custom model development. While foundational models are emerging, visual foundational models can be challenged by the diversity of imaging modalities and experimental conditions (Azad et al., 2023).

Third, engineered features offer consistency and reproducibility advantages valuable for comparative studies and multi-site collaborations, producing identical results across different computing environments.

Rather than viewing these approaches as competing alternatives, the field increasingly recognizes their complementary roles. This motivates the need for robust, accessible tools that efficiently extract interpretable features while integrating seamlessly with modern computational workflows. With all this in mind, we developed cp_measure.

2. Extracting interpretable features

The library cp_measure branches off the CellProfiler codebase, adapted to include calculations for all features of an image-based profiling pipeline, while removing the user interface and anything else non-essential to the task. Measurements are defined as a collection of related features, following CellProfiler's nomenclature, and are categorized based on three kinds of input: One object (producing features such as area and eccentricity), one object + one imaging channel (these can be used to extract such measurements as intensity, texture), one object + multiple channels (e.g., for Manders correlations between intensity values across channels). A visual representation of this can be found in Figure 1.

By isolating the feature calculations from the graphical interface and orchestration components, we aimed to make them accessible to the data science community while improving reusability, testability and extensibility. This reduces the amount of time and human effort required to integrate featurization into existing and new pipelines, while still retaining the standard set of features available for multiple public datasets.

2.1. Reproducing CellProfiler measurements

We first tested whether cp_measure features match the original CellProfiler features, for a representative set of 300 images. These correspond to 150 perturbations from the JUMP Cell Painting dataset (Chandrasekaran et al., 2023), selecting a representative subset of the most significant phenotypes for each feature. We segmented images to delineate the cells and nuclei using CellProfiler, providing object masks (regions of interest) to CellProfiler and to cp_measure for feature extraction. Next, we applied cp_measure on these masks and the original images. To validate cp_measure, we compared its output directly with CellProfiler's on identical images and masks. For each feature, we computed the R^2 value between the two tools' measurements. Over 95% of features showed $R^2 > 0.9$ (Figure 2), indicating near-perfect agreement. The few outliers with lower correlations reveal edge cases handled differently by each tool; these will be addressed through comprehensive unit testing. These results indicate that cp_measure is not likely to produce massively different profiles than those calculated by CellProfiler.

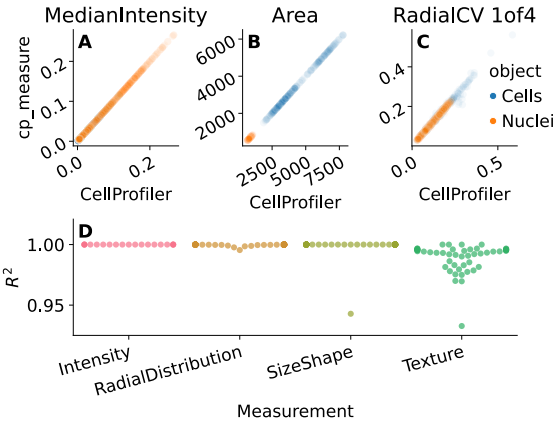

Figure 2. cp_measure features match their CellProfiler analogues. **Panels A-C.** Representative examples comparing CellProfiler feature values (x-axis) to cp_measure's (y-axis), generated using matching pairs of masks and images. **Panel D.** R^2 value of a linear fit for each individual feature, comparing cp_measure to CellProfiler.

2.2. Results

2.2.1. Astrocyte 3D nuclei data

We demonstrate cp_measure's utility through a biologically relevant task: tracking astrocyte maturation over time. From 433 3D images containing 831 astrocyte nuclei (Kalinin et al., 2021), cp_measure extracted morphological features that we preprocessed following standard procedures (Caicedo et al., 2017). We then trained a Gradient Boosting classifier to predict the day of cell differentiation with 87% accuracy. SHAP analysis (Sundararajan & Najmi, 2020) (3) revealed that nuclear minor axis length was the most predictive feature: nuclei became progressively wider during maturation, a phenotype consistent with known astrocyte biology.

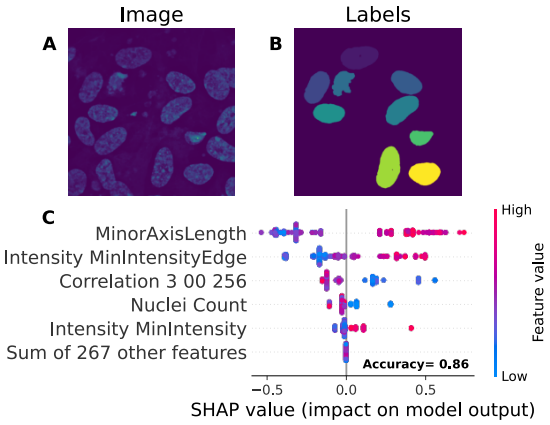

Figure 3. **Panels A and B.** Example pair of astrocyte nuclei image and masks. The 3D images were projected over the z-axis, taking the maximum value across the z-stack. **Panel C.** SHAP values of the most important features for classifying the day of differentiation (out of three days). Each point represents a single 3D image containing multiple nuclei, the magnitude of SHAP values indicate how important those features are for the classifier. Features are ranked by importance, with individual SHAP values shown for the top features and the summed impact shown for the remaining 267 features shown as a single aggregated value. Shape features, such as the minor axis length, refer to the nucleus, while intensity features refer to the DNA staining. The test data accuracy is shown in bold.

2.2.2. Spatial transcriptomics

To demonstrate the broad applicability of cp_measure, we developed a morphology featurizer for the popular spatial omics analysis toolbox Squidpy (Palla et al., 2022). Many of the recently developed spatial omics analysis technologies generate, in addition to their target omics layer, imaging data, usually H&E (hematoxylin and eosin) stained or fluorescently labelled.

Spatial transcriptomics poses unique computational challenges: images routinely contain 500,000+ cells, far exceeding CellProfiler's memory constraints and rigid data structure requirements. cp_measure's modular design enabled us to build a custom workflow integrated with Squidpy (Palla et al., 2022) that leverages the SpatialData format (Marconato et al., 2024) to process cells in streaming parallel batches. This architecture handles massive spatial datasets efficiently while bringing morphological profiling directly into the scverse ecosystem (Virshup et al., 2023), making these features immediately accessible to the single-cell community.

To illustrate the value of this workflow, we used it to identify morphologically similar regions within two human breast cancer sections profiled with the spatial transcriptomics technology Xenium (10x Genomics, 2023). We generated a pattern of tessellating hexagonal masks across both images, of which 18469 were found to overlay the tissue. These masks were then featurized with a cp_measure-based workflow: We first preprocessed the features following standard workflows (Serrano et al., 2025), then removed batch effects between the samples using Harmony (Korsunsky et al., 2019), and finally obtained the Leiden clusters via Scanpy (Wolf et al., 2018; Traag et al., 2019).

For each cluster, we extracted the cells within its hexagons and calculated the respective cell type proportions. A visual comparison of panels A and B in Figure 4 shows that morphology-only clustering largely recapitulates the underlying tissue architecture. Panel C reveals that some clusters have distinct cell-type compositions: For example, cluster 2 is dominated by fibroblasts, whereas adipocytes appear almost exclusively in clusters 5-7. Furthermore, comparing clusters 0 and 1 suggest intra-tumour differences, which could be correlated to immune-hotness (see Figure A5).

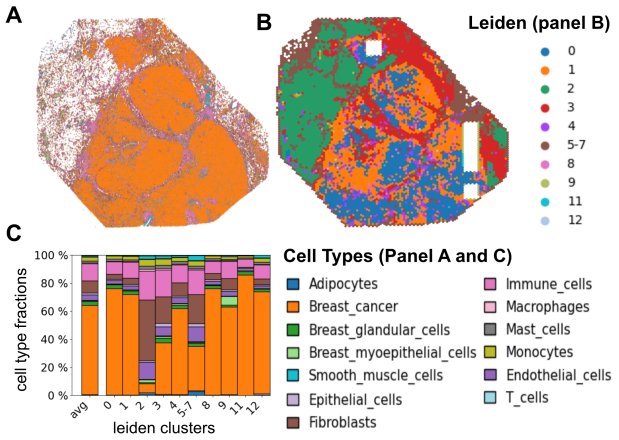

Figure 4. Morphology reveals the composition of tumour subregions. **Panel A.** Spatial map of cell types inferred from transcriptomics. **Panel B.** Morphological subregions obtained by Leiden clustering the image-derived feature embeddings. **Panel C.** Bar chart showing, for each morphological cluster, the fraction of cells assigned to each cell type. Clusters 5-7 were merged due to similarity, cluster 10 removed due to low cell count (see Figure A2B).

These results demonstrate how cp_measure can can be used outside traditional microscopy. When applied on extremely large images it produces biologically meaningful morphological clusters that validate and complement transcriptomic phenotypes. By embedding this capability within the scverse ecosystem via Squidpy, we enable researchers to seamlessly add morphology to their spatial omics workflows, unlocking new insights into tissue organization.

3. Discussion

While point-and-click interfaces open the world of image analysis to many researchers, they are not as effective for computational workflows with no human-in-the-loop. In this work we introduced our new library cp_measure, which calculates a set of widely used engineered features relevant to whole images and to regions of interest (object masks). It enables simpler automated profiling of microscopy data in short scripts and complex pipelines. The modularity it provides facilitates pipelines with better scaling capabilities for high-content microscopy, with or without cloud infrastructure.

The biologically interpretable features provided by cp_measure complement deep learning features and offer a better mechanistic understanding of the underlying biology. A potential workflow would be to use deep learning features to cluster and then use cp_measure features from these clusters. As a whole, when used in tandem with generalist image-processing tools, such as Cellpose for segmentation (Stringer et al., 2021), machine and deep learning workflows can be streamlined.

While developed for biology, engineered morphological features have proven valuable far beyond their original domain–from analysing microplastics in environmental samples (Idehara et al., 2025) to quality control in manufacturing (Ilhan et al., 2021) and agriculture automation (Xu et al., 2024). cp_measure's lightweight, API-first design makes these powerful features accessible to any field working with image data, removing the barriers of biological-specific interfaces.

4. Future work

We propose to make cp_measure an imported dependency for CellProfiler, offering several benefits. It would ensure that the results from pipelines built with either tool will always be comparable, while also providing the opportunity to formalize the inputs and outputs of all measurements. It would also mean that any changes made in cp_measure propagate to CellProfiler, benefiting the community for which it was originally developed.

We also plan to develop a comprehensive test suite to guarantee mathematical correctness, which currently CellProfiler itself is lacking. A comprehensive test suite would enable more confident and expedient optimization for the most compute-consuming features, such as granularity, by providing rapid iteration capabilities. Once tests are in place, we could add support for just-in-time compiling and GPUs. We envision that cp_measure could also be the place to develop and distribute new measurements from the community.

References

10x Genomics. FFPE human breast with custom add-on panel, in situ gene expression dataset. Xenium Onboard Analysis v1.0.2, January 2023.

Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., and Merhof, D. Foundational Models in Medical Imaging: A Comprehensive Survey and Future Vision, October 2023.

Caicedo, J. C., Cooper, S., Heigwer, F., Warchal, S., Qiu, P., Molnar, C., Vasilevich, A. S., Barry, J. D., Bansal, H. S., Kraus, O., Wawer, M., Paavolainen, L., Herrmann, M. D., Rohban, M., Hung, J., Hennig, H., Concannon, J., Smith, I., Clemons, P. A., Singh, S., Rees, P., Horvath, P., Linington, R. G., and Carpenter, A. E. Data-analysis strategies for image-based cell profiling. *Nature Methods*, 14(9):849–863, September 2017. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.4397.

Chandrasekaran, S. N., Ceulemans, H., Boyd, J. D., and Carpenter, A. E. Image-based profiling for drug discovery: Due for a machine-learning upgrade? *Nature Reviews Drug Discovery*, 20(2):145–159, February 2021. ISSN 1474-1784. doi: 10.1038/s41573-020-00117-w.

Chandrasekaran, S. N., Ackerman, J., Alix, E., Ando, D. M., Arevalo, J., Bennion, M., Boisseau, N., Borowa, A., Boyd, J. D., Brino, L., Byrne, P. J., Ceulemans, H., Ch'ng, C., Cimini, B. A., Clevert, D.-A., Deflaux, N., Doench, J. G., Dorval, T., Doyonnas, R., Dragone, V., Engkvist, O., Faloon, P. W., Fritchman, B., Fuchs, F., Garg, S., Gilbert, T. J., Glazer, D., Gnutt, D., Goodale, A., Grignard, J., Guenther, J., Han, Y., Hanifehlou, Z., Hariharan, S., Hernandez, D., Horman, S. R., Hormel, G., Huntley, M., Icke, I., Iida, M., Jacob, C. B., Jaensch, S., Khetan, J., Kost-Alimova, M., Krawiec, T., Kuhn, D., Lardeau, C.-H., Lembke, A., Lin, F., Little, K. D., Lofstrom, K. R., Lotfi, S., Logan, D. J., Luo, Y., Madoux, F., Zapata, P. A. M., Marion, B. A., Martin, G., McCarthy, N. J., Mervin, L., Miller, L., Mohamed, H., Monteverde, T., Mouchet, E., Nicke, B., Ogier, A., Ong, A.-L., Osterland, M., Otrocka, M., Peeters, P. J., Pilling, J., Prechtl, S., Qian, C., Rataj, K., Root, D. E., Sakata, S. K., Scrace, S., Shimizu, H., Simon, D., Sommer, P., Spruiell, C., Sumia, I., Swalley, S. E., Terauchi, H., Thibaudeau, A., Unruh, A., de Waeter, J. V., Dyck, M. V., van Staden, C., Warchoł, M., Weisbart, E., Weiss, A., Wiest-Daessle, N., Williams, G., Yu, S., Zapiec, B., Żyła, M., Singh, S., and Carpenter, A. E. JUMP Cell Painting dataset: Morphological impact of 136,000 chemical and genetic perturbations, March 2023.

Chow, Y. L., Singh, S., Carpenter, A. E., and Way, G. P. Predicting drug polypharmacology from cell morphology readouts using variational autoencoder latent space arithmetic. *PLOS Computational Biology*, 18(2):e1009888, February 2022. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1009888.

Comolet, G., Bose, N., Winchell, J., Duren-Lubanski, A., Rusielewicz, T., Goldberg, J., Horn, G., Paull, D., and Migliori, B. A highly efficient, scalable pipeline for fixed feature extraction from large-scale high-content imaging screens. *iScience*, 27(12), December 2024. ISSN 2589-0042. doi: 10.1016/j.isci.2024.111434.

EinarOlafsson. SpaCr, April 2025.

Garcia-Fossa, F., Cruz, M. C., Haghighi, M., de Jesus, M. B., Singh, S., Carpenter, A. E., and Cimini, B. A. Interpreting Image-based Profiles using Similarity Clustering and Single-Cell Visualization. *Current Protocols*, 3(3):e713, 2023. ISSN 2691-1299. doi: 10.1002/cpz1.713.

Idehara, W., Haga, Y., Tsujino, H., Ikuno, Y., Manabe, S., Hokaku, M., Asahara, H., Higashisaka, K., and Tsutsumi, Y. Exploring Nile Red staining as an analytical tool for surface-oxidized microplastics. *Environmental Research*, 269:120934, March 2025. ISSN 1096-0953. doi: 10.1016/j.envres.2025.120934.

Ilhan, I., ten Klooster, R., and Gibson, I. Effects of process parameters and solid particle contaminants on the seal strength of low-density polyethylene-based flexible food packaging films. *Packaging Technology and Science*, 34 (7):413–421, 2021. ISSN 1099-1522. doi: 10.1002/pts.2567.

Kalinin, A. A., Hou, X., Ade, A. S., Fon, G.-V., Meixner, W., Higgins, G. A., Sexton, J. Z., Wan, X., Dinov, I. D., O'Meara, M. J., and Athey, B. D. Valproic acid-induced changes of 4D nuclear morphology in astrocyte cells. *Molecular Biology of the Cell*, 32(18):1624–1633, August 2021. ISSN 1059-1524. doi: 10.1091/mbc.E20-08-0502.

Kim, V., Adaloglou, N., Osterland, M., Morelli, F. M., Halawa, M., König, T., Gnutt, D., and Zapata, P. A. M. Self-supervision advances morphological profiling by unlocking powerful image representations, April 2023.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R., and Raychaudhuri, S. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods*, 16(12): 1289–1296, December 2019.

Lafarge, M. W., Caicedo, J. C., Carpenter, A. E., Pluim, J. P. W., Singh, S., and Veta, M. Capturing Single-Cell Phenotypic Variation via Unsupervised Representation Learning. In *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, pp. 315–325. PMLR, May 2019.

Li, X., Zhang, Y., Wu, J., and Dai, Q. Challenges and opportunities in bioimage analysis. *Nature Methods*, 20 (7):958–961, July 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01900-4.

Marconato, L., Palla, G., Yamauchi, K. A., Virshup, I., Heidari, E., Treis, T., Vierdag, W.-M., Toth, M., Stockhaus, S., Shrestha, R. B., Rombaut, B., Pollaris, L., Lehner, L., Vöhringer, H., Kats, I., Saeys, Y., Saka, S. K., Huber, W., Gerstung, M., Moore, J., Theis, F. J., and Stegle, O. SpatialData: an open and universal data framework for spatial omics. *Nat. Methods*, pp. 1–5, March 2024.

Moshkov, N., Bornholdt, M., Benoit, S., Smith, M., McQuin, C., Goodman, A., Senft, R. A., Han, Y., Babadi, M., Horvath, P., Cimini, B. A., Carpenter, A. E., Singh, S., and Caicedo, J. C. Learning representations for image-based profiling of perturbations, August 2022.

Palla, G., Spitzer, H., Klein, M., Fischer, D., Schaar, A. C., Kuemmerle, L. B., Rybakov, S., Ibarra, I. L., Holmberg, O., Virshup, I., Lotfollahi, M., Richter, S., and Theis, F. J. Squidpy: A scalable framework for spatial omics analysis. *Nature Methods*, 19(2):171–178, February 2022. ISSN 1548-7105. doi: 10.1038/s41592-021-01358-2.

Serrano, E., Chandrasekaran, S. N., Bunten, D., Brewer, K. I., Tomkinson, J., Kern, R., Bornholdt, M., Fleming, S. J., Pei, R., Arevalo, J., Tsang, H., Rubinetti, V., Tromans-Coia, C., Becker, T., Weisbart, E., Bunne, C., Kalinin, A. A., Senft, R., Taylor, S. J., Jamali, N., Adeboye, A., Abbasi, H. S., Goodman, A., Caicedo, J. C., Carpenter, A. E., Cimini, B. A., Singh, S., and Way, G. P. Reproducible image-based profiling with Pycytominer. *Nature Methods*, 22(4):677–680, April 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02611-8.

Stirling, D. R., Swain-Bowden, M. J., Lucas, A. M., Carpenter, A. E., Cimini, B. A., and Goodman, A. CellProfiler 4: Improvements in speed, utility and usability. *BMC Bioinformatics*, 22(1):433, September 2021. ISSN 1471-2105. doi: 10.1186/s12859-021-04344-9.

Stringer, C., Wang, T., Michaelos, M., and Pachitariu, M. Cellpose: A generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106, January 2021. ISSN 1548-7105. doi: 10.1038/s41592-020-01018-x.

Sundararajan, M. and Najmi, A. The Many Shapley Values for Model Explanation. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9269–9278. PMLR, November 2020.

Tang, Q., Ratnayake, R., Seabra, G., Jiang, Z., Fang, R., Cui, L., Ding, Y., Kahveci, T., Bian, J., Li, C., Luesch, H., and Li, Y. Morphological profiling for drug discovery in the era of deep learning. *Briefings in Bioinformatics*, 25 (4):bbae284, July 2024. ISSN 1477-4054. doi: 10.1093/bib/bbae284.

Traag, V. A., Waltman, L., and van Eck, N. J. From louvain to leiden: guaranteeing well-connected communities. *Sci. Rep.*, 9(1):5233, March 2019.

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., and Yu, T. Scikit-image: Image processing in Python. *PeerJ*, 2:e453, June 2014. ISSN 2167-8359. doi: 10.7717/peerj.453.

Virshup, I., Bredikhin, D., Heumos, L., Palla, G., Sturm, G., Gayoso, A., Kats, I., Koutrouli, M., Scverse Community, Berger, B., Pe'er, D., Regev, A., Teichmann, S. A., Finotello, F., Wolf, F. A., Yosef, N., Stegle, O., and Theis, F. J. The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nat. Biotechnol.*, 41(5):604–606, May 2023.

Weisbart, E., Kumar, A., Arevalo, J., Carpenter, A. E., Cimini, B. A., and Singh, S. Cell Painting Gallery: An open resource for image-based profiling. *Nature Methods*, 21 (10):1775–1777, October 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02399-z.

Wolf, F. A., Angerer, P., and Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, February 2018. ISSN 1474-760X. doi: 10.1186/s13059-017-1382-0.

Wu, B., Zhang, B., Li, B., Wu, H., and Jiang, M. Cold and hot tumors: from molecular mechanisms to targeted therapy. *Signal Transduct. Target. Ther.*, 9(1):274, October 2024.

Xu, S., Xu, R., Ma, P., Huang, Z., Wang, S., Yang, Z., and Liao, Q. Design of a Non-Destructive Seed Counting Instrument for Rapeseed Pods Based on Transmission Imaging. *Agriculture*, 14(12):2215, December 2024. ISSN 2077-0472. doi: 10.3390/agriculture14122215.

5. Acknowledgements

This work was supported by GSK and the National Institutes of Health, grant (R35 GM122547 to AEC). The authors would like to thank Nodar Gogoberidze, Beth Cimini, Minh Doan and Eliot McKinley for their valuable feedback and discussions during the course of this project.

6. Appendix

5.1. Methods

5.1.1. Data and software

The codebase for cp_measure is available on https://github.com/afermg/cp_measure. All code to reproduce the analyses and figures, alongside links to the original data, is available on the GitHub repository https://github.com/afermg/2025_cpmeasure.

5.1.2. Details on the spatial use case

The following figures show the processing steps for the spatial data used to generate Figure 4 in more detail. Raw data from two Xenium Spatial Transcriptomics available from the 10x Genomics Datasets homepage (name: "FFPE Human Breast with Custom Add-on Panel", release date: 2023-01-22) were downloaded and converted to the ‘SpatialData‘ format. The data comprises two samples, one infiltrating ductal carcinoma ("sample 1") and one invasive lobular carcinoma ("sample 2"). H&E images from the same dataset were also downloaded, aligned to the respective reference images, and saved to the data object. These can be seen in Figure A3A and Figure A3B. A pattern of tessellating hexagons was generated for each sample, spanning the entire tissue section. These hexagons were subselected to only those overlaying the specimens as visible in the H&E image, resulting in 10321 hexagons for sample 1 and 8148 for sample 2 respectively. These hexagonal masks were used to automatically extract the underlying H&E crops which were then featurized using ‘Squidpy‘, resulting in 945 features each. These features were then pre-processed using ‘Pycytominer‘, resulting in 252 features. We then removed batch effect using Harmony. UMAP projections of these features are shown in Figure A1.

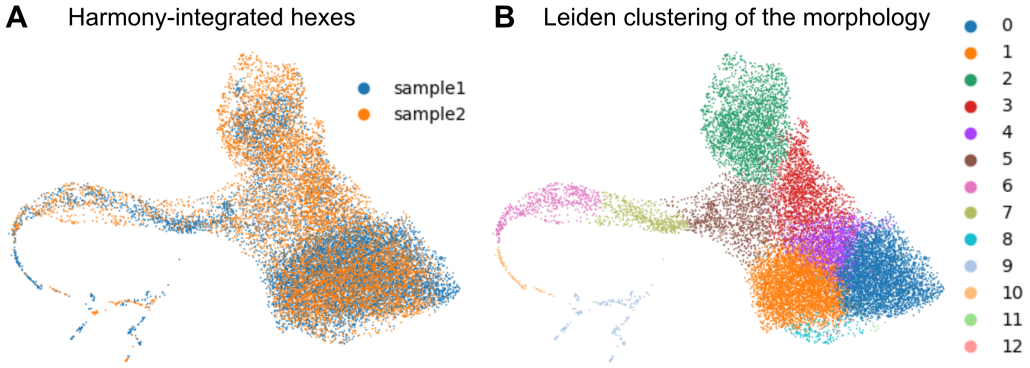

Figure A1. UMAP projection of the processed morphology features. **Panel A.** Coloured by sample of origin, we see good mixing after the Harmony integration. **Panel B.** Coloured by Leiden clustering.

Due to the low cell count in Leiden cluster 10 for sample 2, as shown in Figure A2B the cluster was excluded for some analysis steps.

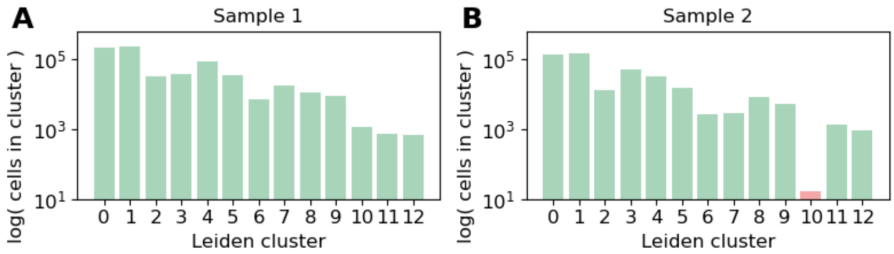

Figure A2. Morphology reveals the composition of tumour subregions. **Panel A.** Spatial map of cell types inferred from transcriptomics. **Panel B.** Morphological subregions obtained by applying Leiden clustering to image-derived feature embeddings. **Panel C.** Bar chart showing, for each morphological cluster, the fraction of cells assigned to each cell type (colour legend as shown). The top left quarter of Panel A shows that the morphological approach worked to predict regions while the expression annotations did not.

Then, the gene expression data from the Xenium experiment was used to annotate cell types for both samples. For every Leiden cluster, as defined by the Leiden clustering of the joint morphology features from both samples, the cells covered by the respective hexagons were extracted. Figure A3 shows the raw H&E images, the distribution of cell types across these, and the hexagons coloured by their respective Leiden cluster. Of note is that the H&E images both seemed to be partially corrupted, resulting in blacked-out sections. While the data contains cell segmentation masks for these sections, we filtered out hexagons overlaying these corrupted regions for the morphology featurization.

For every morphology Leiden cluster, we counted all cells whose centroids fell into the respective hexagons and calculated each cell type's share of the total cells per cluster. This data is shown in Figure A4.

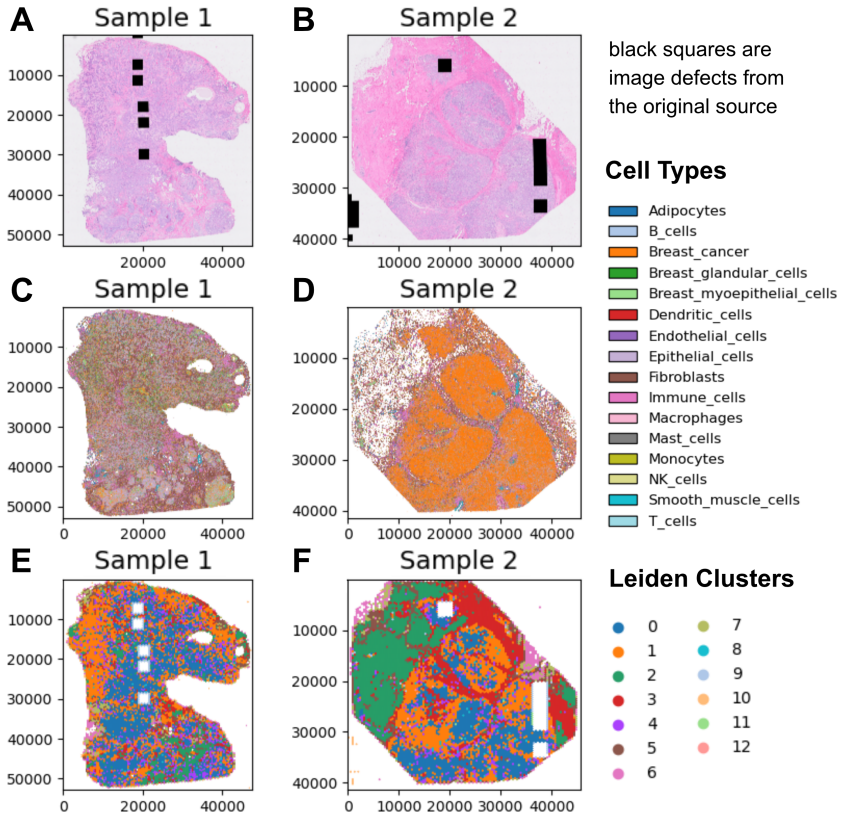

Figure A3. Overview about the spatial distribution of cell types and morphology clusters. **Panel A and B.** H&E images of both samples. **Panel C and D.** Cell segmentation masks colored by cell type. **Panel E and F.** Hexagons colored by their morphology Leiden cluster.

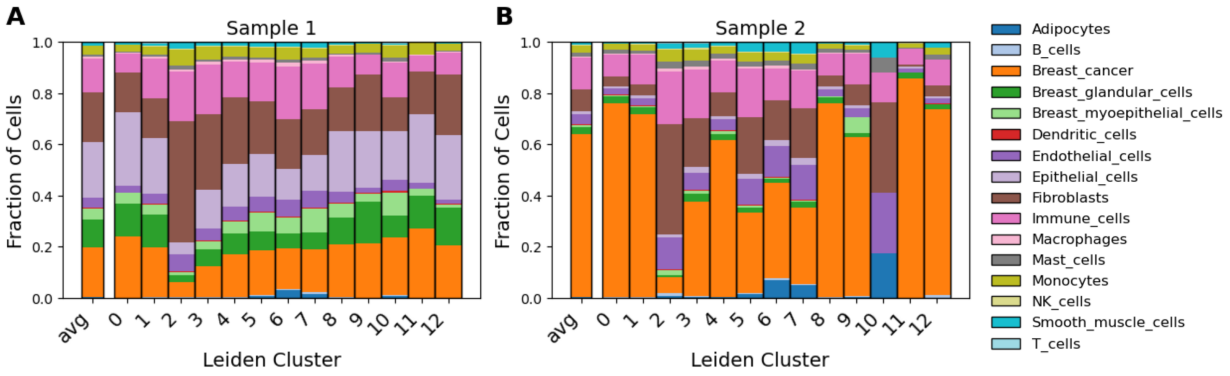

Figure A4. Barplots showing the per-cluster cell type composition for both samples. Additionally, the average per sample is shown.

Based on (Wu et al., 2024) we define an immune-hot signature comprising B cells, dendritic cells, NK cells, and T cells. For every Leiden cluster, we aggregated the respective cell type counts of this signature for both samples. These are shown in Figure A5B and provide initial evidence for the difference in clusters 0 and 1, despite both being visually mixed as can be seen in Figure A5A.

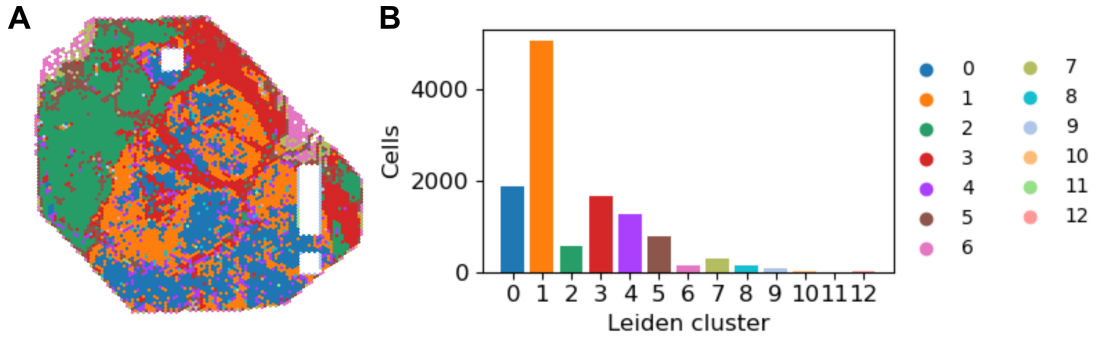

Figure A5. Potential explanation of the differences between cluster 0 and cluster 1. **Panel A** Hexagons in sample 2 coloured by Leiden cluster. **Panel B** Cumulative counts of B cells, dendritic cells, NK cells, and T cells.