TITLE

Regression plane concept: analysing continuous cellular processes with machine learning

AUTHORS

Abel Szkalisity[1,2], Filippo Piccinini[3], Attila Beleon[1], Tamas Balassa[1], Istvan Gergely Varga[4], Ede Migh[1], Lassi Paavolainen[5], Sanna Timonen[5], Indranil Banerjee[6], Yohei Yamauchi[7], Istvan Ando[4], Jaakko Peltonen[8,9], Vilja Pietiäinen[5], Viktor Honti[4], Peter Horvath[1,5,10,*]

AUTHORS' AFFILIATIONS

[1] Synthetic and Systems Biology Unit, Biological Research Centre (BRC), H-6726 Temesvári krt. 62, Szeged, Hungary.

[2] Department of Anatomy and Stem Cells and Metabolism Research Program,
Faculty of Medicine, University of Helsinki, FI-00290 Haartmaninkatu 8, Helsinki, Finland.

[3] Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori (IRST) IRCCS,
I-47014 Via Piero Maroncelli 40, Meldola (FC), Italy.

[4] Institute of Genetics, Biological Research Center (BRC), H-6726 Temesvári krt. 62, Szeged, Hungary.

[5] Institute for Molecular Medicine Finland-FIMM, Helsinki Institute of Life Science-HiLIFE, University of Helsinki, FI-00014 Tukholmankatu 8, Helsinki, Finland.

[6] Indian Institute of Science Education and Research (IISER), 140306 Knowledge City, Sector 81, Mohali, India.

[7] School of Cellular and Molecular Medicine, University of Bristol, BS8 1TD University Walk, Bristol, UK.

[8] Faculty of Information Technology and Communication Sciences, Tampere University,
FI-33014 Tampere University, Tampere, Finland.

[9] Department of Computer Science, Aalto University, FI-00076 Konemiehentie 2, Aalto, Finland.

[10] Single-Cell Technologies Ltd., H-6726 Temesvári krt. 62, Szeged, Hungary.

[*] Corresponding author.

CORRESPONDING AUTHOR

Correspondence to Peter Horvath, Synthetic and Systems Biology Unit, Biological Research Centre (BRC), Temesvári krt. 62, H-6726, Szeged, Hungary. Phone: +36 62599654. Fax: +36 62433506. E-mail: horvath.peter@brc.hu

KEYWORDS

High-content screening; Single cell analysis; Regression; Machine learning; Advanced Cell Classifier; Lipid; Drosophila; Hemocyte differentiation; Cell mitosis

ABSTRACT

Biological processes are inherently continuous, and the chance of phenotypic discovery is significantly restricted by discretising them. Using multi-parametric active regression we introduce a novel concept to describe and explore biological data in a continuous manner. We have implemented *Regression Plane (RP)*, the first user-friendly discovery tool enabling class-free phenotypic supervised machine learning.

MAIN TEXT

Large-scale imaging scenarios, including high-content screening (HCS) and digital pathology imaging, have become the *de facto* tools for discovering drugs, genes and understanding tissue physiologies and pathologies, including cancer heterogeneity. This has induced a rapid growth in the amount of microscopy data, making it essential to elaborate appropriate bioinformatics tools to analyse them, and thus improve the current understanding of underlying biological processes [1,2,3].

Machine learning provides automation for analysing big data, such as that acquired in large-scale, image-based experiments, and it has been successfully utilized for phenotypic analysis tasks [4]. Although a great variety of software tools are available for performing imaging assays in a supervised manner (e.g. CellProfiler Analyst, Ilastik, CellCognition, Advanced Cell Classifier [5]), all of them rely on the assumption that the underlying biological processes have stable steady states that can be dissected into discrete phenotypic classes (**Fig. 1a**). However, biological processes are inherently continuous, and modelling them as a set of discrete states may reduce the potential to properly understand biological phenomena.

The application of traditional classification models for single cell image analysis [6,7,8] is especially unreliable when the cells of interest change their morphological features gradually in the course of time. Annotation of such data is error-prone and laborious, and even field experts tend to make faulty decisions (e.g. in the case of samples with interclass properties), often leading to arbitrary labelling. Additionally, user defined classes may obscure the real underlying distribution by inappropriate discretization.

Currently, none of the available and widely used software tools enable single-cell based image analysis in a continuous, supervised manner. Instead, unsupervised models, such as Lineage Reconstruction Techniques (LRT) [9] and Dynamic Time Warping (DTW) prevail. Cycler [8] is an LRT and embeds 5 pre-selected image-based single-cell features to a one dimensional (1D) continuous space called the cell-cycle trajectory. Similarly, Cai et al. used DTW to align mitotic cells into the *mitotic standard time* based on 6 selected features [10]. Indeed, these tools provide robust solutions for their targeted tasks, but the lack of expert interaction significantly reduces the potential to customize these methods for various purposes. Therefore, another set of tools known as Visual Analytics (VA) was developed, offering various techniques for experts to interactively change the machine learning model through a visualization interface, which is most often a continuous space (visualization map) [11,12]. CellCognition was a pioneer of supervised tools, designed with the intent to efficiently analyse biological processes, however still using classification [7].

Here, we propose a novel methodology called *Regression Plane* (RP), an interface for fully supervised, continuous machine learning appropriate for image-based single-cell analysis. The idea originates from a study of an influenza A virus entry in which histone deacetylase-mediated reorganisation of the microtubules led to various endosomal morphological and trafficking phenotypes that affected influenza infection [13]. The scatteredness of late endosomes and lysosomes (single output variable) was determined using regression instead of classification. Restricting the output to a single dimension prohibited the modelling of branching, circulating (*e.g.* cell cycle), parallel and crossing processes. Therefore, we have introduced a novel approach to utilize a 2D plane (**Fig. 1a**, **Fig. 2a, e**). Considering cellular steady-states as graph nodes and gradual changes between the states as edges, the biological systems that correspond to planar graphs can be modelled with RP. Further extension of the modelling to 3D would increase the complexity of labelling and raise the chance of annotation errors. Additionally, to improve the quality of the annotated sets and decrease the time required from experts, we have incorporated novel active learning methods appropriate for regression-based phenotyping.

Regression Plane is implemented as an open-source module of Advanced Cell Classifier (ACC) [6], and it has been available since *ACC v3.0*. RP was incorporated into traditional phenotypic classification in a hierarchical manner: each class may be extended with a distinct regression plane, allowing multiple regression planes to be incorporated into a single project. RP is easy to use, well documented and supported by video tutorials (**Suppl. Materials 1-3**). Annotation is performed by assigning continuous labels to representative cells via placing them on a 2D plane. After training, RP predicts the position of every unlabelled cell and outputs versatile and easy-to-read visual representations at single-cell, population and treatment levels (for details see **Online Methods**).

Similarly to classification, a representative Training Set (TS) is also essential for RP. Active learning algorithms are routinely used in classification to find the most efficient TS [14] but are not widely used in regression [15]. In this work, we introduce various active regression algorithms by extending those used in classical active learning tasks (**Suppl. Fig. 1a**). These methods propose cells whose automatic prediction on the regression plane is uncertain or ambiguous. Details are reported in **Online Methods**.

To analyse data discovery capabilities of RP, we have generated a synthetic HCS image dataset simulating drugs perturbing cell shape and protein expression (**Fig. 2a-c**). Details about the modelled biological processes are reported in **Online Methods**. Ten microscopy experts were asked to identify the distinct underlying processes in the experiment (**Suppl. Note 1**). The first group of five experts used *ACC v2.1* to annotate cells with discrete labels, while the other group used RP only (*ACC v3.0*). Despite the great variety of the regression planes created by the microscopists (**Suppl. Fig. 2**), the results obtained using RP significantly outperformed the classification, both in terms of precision and recall (**Fig. 2d**). Specifically, the experts using RP performed better in estimating the number of ongoing processes, and achieved, on average, an improvement of approximately 20% in precision and 5% in recall, upon defining image sets containing cells with similar behaviour.

Next, we have evaluated whether siRNA perturbations of candidate genes, previously revealed to influence blood triglyceride (TG) levels in humans in a genome-wide association study [16], would affect the morphology of lipid droplets (LDs) in cultured hepatocytes (Huh7 cell line). Regarding their continuous changes in localization, number and size, LDs form a heterogeneous population reflecting different cellular metabolic states [17]. Thus, RP was used for the analysis of neutral lipids in lipid droplets labelled with LipidToxGreen (**Suppl. Fig. 3a, b, c**). To train the model, 457 cells were placed on the regression plane by a microscopy expert (**Fig. 2e**). We found that siRNA-mediated knockdown of *TM6SF2* (a gene associated with decreased blood TGs) led to increased intracellular staining of neutral lipids, as it had been expected from the earlier evidence of *TM6SF2* affecting hepatic lipid droplet content and TG secretion [18]. In contrast, the cells transfected with siRNAs targeting *CD300LG* (a gene associated with increased blood TGs [16]) showed a decreased amount of intracellular TGs, accompanied by the disappearance of (larger) LDs. Additional biochemical analysis measuring cellular TG levels confirmed these findings (**Suppl. Fig. 3d**). These data provide the first functional evidence for the role of *CD300LG* in regulating TG metabolism in hepatocytes.

Intriguingly, the knockdown of *TM4SF5* (a gene associated with decreased blood TGs) which codes for a protein functioning as an arginine sensor and mTORC1 regulator on lysosomal membranes [19], not shown earlier to affect triglyceride levels, promoted the increase of small LDs (**Fig. 2f**). Meta-visualization and clustering of the regression planes **(Fig. 2g**, **Suppl. Fig. 3e-h)** further supplemented the findings from an earlier study [16], and suggest that *CD300LG* and *TM4SF5* may have biological effects on hepatic TG levels and LD composition, to be further addressed in future studies. Details are reported in **Online Methods**.

We tested the capabilities of RP on 2 different time-resolved datasets. First, RP has been demonstrated to be capable of reproducing an unsupervised mitotic time model developed in the MitoCheck project (*www.mitocheck.org*, for details see **Online Methods**). Secondly, hemocyte differentiation was evaluated in *Drosophila melanogaster*. Hemocytes are blood cells of invertebrates that play a role in immune defense. Following infestation by a parasitic wasp, the larvae of *D. melanogaster* produce a special blood cell type called lamellocyte, to isolate the invader by forming a multilayer capsule around the wasp's egg [20]. Several lineage tracing studies have indicated that these capsule forming lamellocytes differentiate from phagocytic plasmatocytes upon immune induction - which were underscored by findings of the most recent transcriptome analyses [21,22,23]. It has also been suggested that the lamellocyte pool actually consists of two cell types, including the larger type I lamellocyte and the smaller type II lamellocyte, of which only type II lamellocytes originate from plasmatocytes [24]. To resolve this contradiction, we developed an *ex vivo* method for culturing *Drosophila* hemocytes, appropriate for monitoring their differentiation with time-lapse microscopy. Blood cell types can be characterized by their morphologies and *in vivo* transgenic reporter expression pattern [24]. The regression plane was manually trained using 109 cells based on their morphology and reporter gene expression (**Fig. 2h**). The analysis revealed that 5.6% of the plasmatocytes trans-differentiated into lamellocytes upon immune induction (wounding) of the larvae (the threshold line is indicated in **Fig. 2j**). However, instead of identifying 2 clearly separated subtypes I and II), we have observed that the differentiation processes are evenly distributed on the regression plane, as reflected by specific features (**Fig. 2i, j, k**). This finding suggests that type I and type II lamellocytes, both differentiating from plasmatocytes, are not definitely distinguishable cell types, but rather they are two extreme stages of a size continuum (**Fig. 2l**). Details are reported in **Online Methods**.

Regression Plane increases the resolution of classification to represent subtle phenotypic differences by exploiting regression techniques, extended by active learning. First, using artificial datasets we have demonstrated its capability to outperform the available classification tools in phenotypic discovery. Second, we have applied RP to analyze lipid droplets in cultured hepatic cells, serving as a model of a heterogeneous population that reflects different cellular metabolic states, and have revealed genes playing a crucial role in regulating triglyceride levels in hepatocytes. Finally, we have identified the previously undiscovered continuous characteristics of hemocyte differentiation in *Drosophila melanogaster*. Our findings indicate that RP is a promising tool to explore biological data in a continuous manner, reflecting the non-discrete nature of biological processes.

METHODS

Methods, including statements of data availability and associated accession codes and references are available in the online version of the paper.

Note: Supplementary information and Source Data files are available in the online version of the paper.

ACKNOWLEDGEMENTS

The authors thank Antti Lehmussola and Pekka Ruusuvuori (Tampere University of Technology, Finland) for the information provided about the SIMCEP software; Samuli Ripatti and Ida Surakka (FIMM, Helsinki, Finland) for their valuable comments on our experiments related to the genes associated with dyslipidemia; Elina Ikonen and Anna Uro (Faculty of Medicine, University of Helsinki) for providing expertise in the biochemical quantification of lipid levels; the FIMM High Throughput Biomedicine Unit for providing access to high throughput robotics and siRNA library (HiLIFE, University of Helsinki and Biocenter Finland); Antti Hassinen and Olli Kallioniemi (FIMM, Helsinki, Finland) for their support with HC-imaging capabilities; Gabriella Tick and

Máté Görbe (BRC, Szeged, Hungary) for their help with the software documentation; Csaba Molnár (BRC, Szeged, Hungary) for his expertise on image analysis; the Finnish Grid and Cloud Infrastructure (urn:nbn:fi:research-infras-2016072533) for computational resources; Dóra Bokor (BRC, Szeged, Hungary) for proofreading the manuscript.

FUNDING

ASZ, BT, AB, EM and PH acknowledge support from the Hungarian National Brain Research Program (MTA-SE-NAP B-BIOMAG), from the LENDULET-BIOMAG Grant (2018-342), from the European Regional Development Funds (GINOP-2.3.2-15-2016-00006, GINOP-2.3.2-15-2016-00026, GINOP-2.3.2-15-2016-00037), from the H2020 (ERAPERMED-COMPASS, DiscovAIR), and from the Chan Zuckerberg Initiative (Deep Visual Proteomics). VP, LP and PH acknowledge support from the Finnish TEKES FiDiPro Fellow Grant 40294/13 and FIMM High Content Imaging and Analysis Unit (FIMM-HCA; HiLIFE-HELMI) and Biocenter Finland, Juselius Foundation, Academy of Finland Centre of Excellence in Translational Cancer Biology and Finnish Cultural Foundation. FP acknowledges support from the Union for International Cancer Control (UICC) for a UICC Yamagiwa-Yoshida (YY) Memorial International Cancer Study Grant (ref: UICC-YY/678329). VH acknowledges support from the National Research, Development and Innovation Office (OTKA K-131484). JP, LP acknowledges support from the Academy of Finland, decision numbers 295694, 313748, 327352 and 310552.

AUTHOR CONTRIBUTIONS

PH conceived and led the project. ASZ developed the Regression Plane tool. ASZ and AB developed the trajectory tool. TB debugged and released the software. ASZ, FP, TB, IGV, EM, LP, ST, VP and VH designed and performed the experiments. ASZ, FP, AB and VP tested the software tool. YY, IA, JP and PH supervised the project. FP prepared the documentation and website. ASZ, FP, VP, VH and PH, wrote the manuscript. ASZ, FP and IB prepared the figures included in the paper. All authors read and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

REFERENCES (MAIN TEXT)

1. Carragher, N., Piccinini, F., Tesei, A., Trask Jr, O. J., Bickle, M., & Horvath, P. Concerns, challenges and promises of high-content analysis of 3D cellular models. *Nature Reviews Drug Discovery*, **17**(8), 606-606 (2018).

2. 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., & Carpenter, A. E. Data-analysis strategies for image-based cell profiling. *Nature Methods*, **14**(9), 849-863 (2017).

3. Moen, E., Bannon, D., Kudo, T., Graf, W., Covert, M., & Van Valen, D. Deep learning for cellular image analysis. *Nature Methods*, **16**, 1233–12461 (2019).

4. Sommer, C., & Gerlich, D. W. Machine learning in cell biology–teaching computers to recognize phenotypes. *Journal of Cell Science*, **126**(24), 5529-5539 (2013).

5. Smith, K., Piccinini, F., Balassa, T., Koos, K., Danka, T., Azizpour, H., & Horvath, P. Phenotypic image analysis software tools for exploring and understanding big image data from cell-based assays. *Cell Systems,* **6**(6), 636-653 (2018).

6. Piccinini, F., Balassa, T., Szkalisity, A., Molnar, C., Paavolainen, L., Kujala, K., Buzas, K., Sarazova, M., Pietiainen, V., Kutay, U., Smith, K., & Horvath, P. Advanced cell classifier: user-friendly machine-learning-based software for discovering phenotypes in high-content imaging data. *Cell Systems*, **4**(6), 651-655 (2017).

7. Held, M., Schmitz, M. H., Fischer, B., Walter, T., Neumann, B., Olma, M. H., Peter, M., Ellenberg, J., & Gerlich, D. W. CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging. *Nature Methods*, **7**(9), 747-754 (2010).

8. Gut, G., Tadmor, M. D., Pe'er, D., Pelkmans, L., & Liberali, P. Trajectories of cell-cycle progression from fixed cell populations. *Nature Methods*, **12**(10), 951-954 (2015).

9. Kester, L., & van Oudenaarden, A. Single-cell transcriptomics meets lineage tracing. *Cell Stem Cell*, **23**(2), 166-179 (2018).

10. Cai, Y., Hossain, M. J., Hériché, J. K., Politi, A. Z., Walther, N., Koch, B., Wachsmuth, M., Nijmeijer, B., Kueblbeck, M., Martinic-Kavur, M., Ladurner, R., Alexander, S., Peters, J. M., & Ellenberg, J. Experimental and computational framework for a dynamic protein atlas of human cell division. *Nature*, **561**(7723), 411-415 (2018).

11. Sacha, D., Sedlmair, M., Zhang, L., Lee, J. A., Peltonen, J., Weiskopf, D., North, S. C., & Keim, D. A. What you see is what you can change: Human-centered machine learning by interactive visualization. *Neurocomputing*, **268**, 164-175 (2017).

12. Buja, A., Swayne, D. F., Littman, M. L., Dean, N., Hofmann, H., & Chen, L. Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*, **17**(2), 444-472 (2008).

13. Yamauchi, Y., Boukari, H., Banerjee, I., Sbalzarini, I. F., Horvath, P., & Helenius, A. Histone deacetylase 8 is required for centrosome cohesion and influenza A virus entry. *PLoS Pathogens*, **7**(10), e1002316 (2011).

14. Sverchkov, Y., & Craven, M. A review of active learning approaches to experimental design for uncovering biological networks. *PLoS Computational Biology*, **13**(6), e1005466 (2017).

15. Kumar, P., & Gupta, A. Active Learning Query Strategies for Classification, Regression, and Clustering: A Survey. *Journal of Computer Science and Technology*, **35**(4), 913-945 (2020).

16. Surakka, I. et al. The impact of low-frequency and rare variants on lipid levels. *Nature Genetics*, **47**(6), 589-597 (2015).

17. Olzmann, J. A., & Carvalho, P. Dynamics and functions of lipid droplets. *Nature Reviews Molecular Cell Biology*, **20**(3), 137-155 (2019).

18. Mahdessian, H., Taxiarchis, A., Popov, S., Silveira, A., Franco-Cereceda, A., Hamsten, A., Eriksson, P. & van't Hooft, F. TM6SF2 is a regulator of liver fat metabolism influencing triglyceride secretion and hepatic lipid droplet content. *Proceedings of the National Academy of Sciences*, **111**(24), 8913-8918 (2014).

19. Jung, J. W., Macalino, S. J. Y., Cui, M., Kim, J. E., Kim, H. J., Song, D. G., Song, G., Nam, S. H., Kim, S., Choi, S., & Lee, J. W. Transmembrane 4 L six family member 5 senses arginine for mTORC1 signaling. *Cell Metabolism*, **29**(6), 1306-1319 (2019).

20. Evans, C. J., Hartenstein, V., & Banerjee, U. Thicker than blood: conserved mechanisms in Drosophila and vertebrate hematopoiesis. *Developmental Cell*, **5**(5), 673-690 (2003).

21. Honti, V., Csordás, G., Kurucz, É., Márkus, R., & Andó, I. The cell-mediated immunity of Drosophila melanogaster: hemocyte lineages, immune compartments, microanatomy and regulation. *Developmental & Comparative Immunology*, **42**(1), 47-56 (2014).

22. Cattenoz, P. B., Sakr, R., Pavlidaki, A., Delaporte, C., Riba, A., Molina, N., Hariharan, N., Mukherjee, T., & Giangrande, A. Temporal specificity and heterogeneity of Drosophila immune cells. *The EMBO Journal*, e104486 (2020).

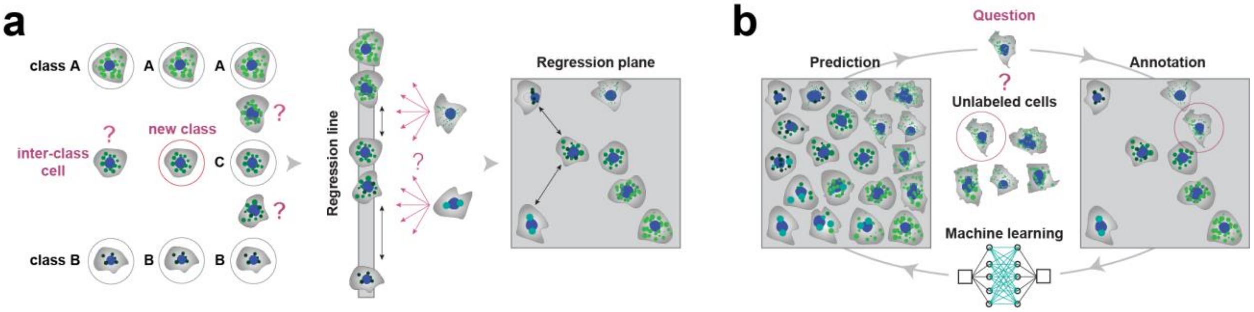
23. Tattikota, S. G., Cho, B., Liu, Y., Hu, Y., Barrera, V., Steinbaugh, M. J., Yoon, S. H., Comjean, A., Li, F., Dervis, F., Hung, R. J., Nam, J. W., Sui, S. H., Shim, J., & Perrimon, N. A single-cell survey of Drosophila blood. *Elife*, **9**, e54818 (2020).

24. Anderl, I., Vesala, L., Ihalainen, T. O., Vanha-Aho, L. M., Andó, I., Rämet, M., & Hultmark, D. Transdifferentiation and proliferation in two distinct hemocyte lineages in Drosophila melanogaster larvae after wasp infection. *PLoS Pathogens*, **12**(7), e1005746 (2016).

25. Hollandi, R., Szkalisity, A., Toth, T., Tasnadi, E., Molnar, C., Mathe, B., Grexa, I., Molnar, J., Balind, A., Gorbe, M., Kovacs, M., Migh, E., Goodman, A., Balassa, T., Koos, K., Wang, W., Caicedo, J. C., Bara, N., Kovacs, F., Paavolainen, L., Danka, T., Kriston, A., Carpenter, A. E., Smith, K., & Horvath, P. nucleAIzer: A parameter-free deep learning framework for nucleus segmentation using image style transfer. *Cell Systems*, **10**(5), 453-458 (2020).

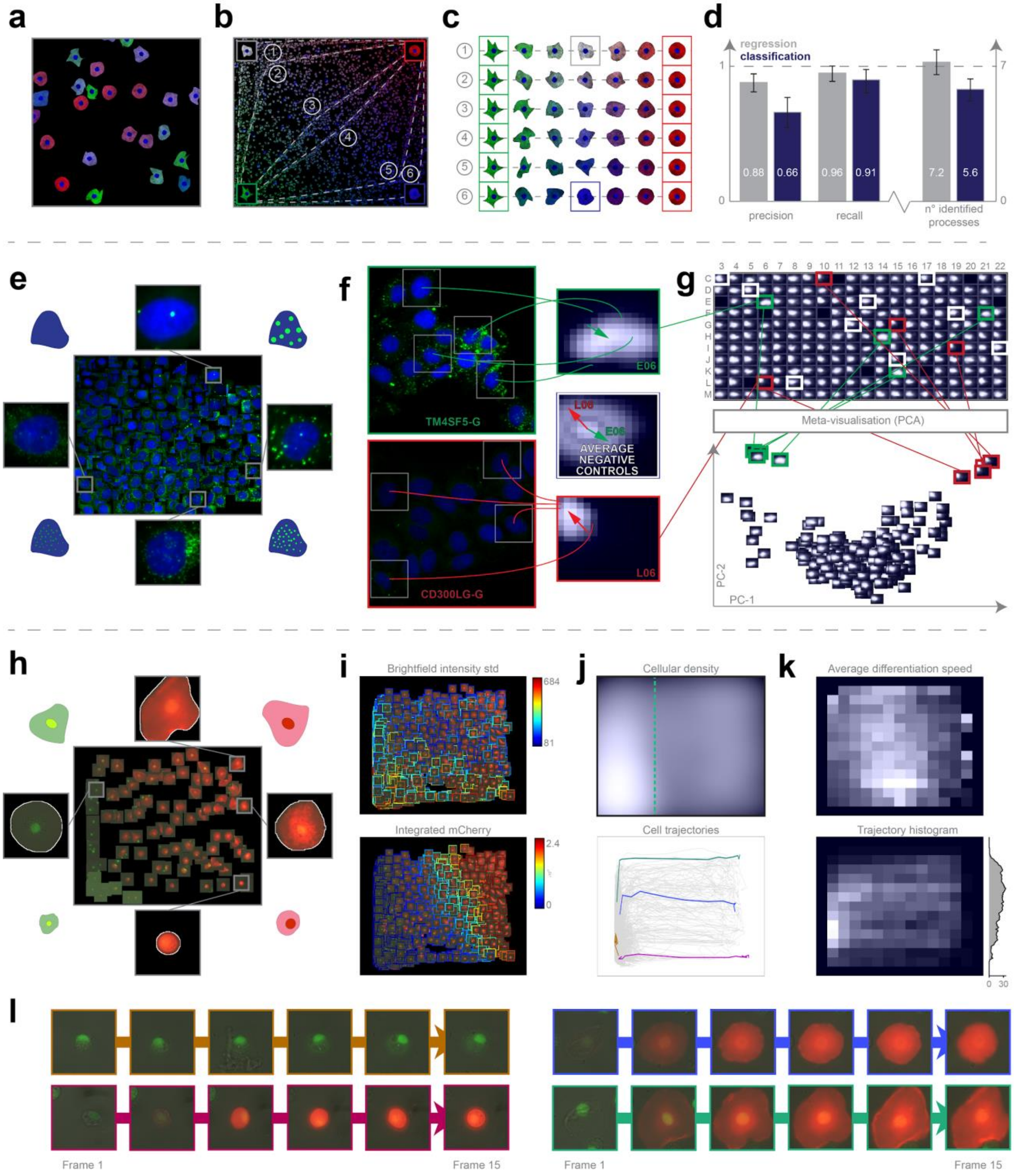
FIGURE LEGENDS

Figure 1



(**a**) **Classification *vs* regression**. The classical way to model a biological process includes the phenotypical analysis of cells (*i.e*. subdividing cells into classes). However, in a high-content screening scenario, the multitude of different phenotypes makes it extremely challenging to create a set of representative classes. A possible solution builds on using a regression line, allowing to represent a single effect without the need of discretization. Nonetheless, biological processes are typically characterized by numerous ongoing effects. Thus, the regression plane represents a good trade-off between visualization capabilities and annotation complexity. Basically, it allows to represent a biological process with the limits of a planar graph. (**b**) **Active regression**. The aim of an active regression algorithm is to improve the training set (TS) to achieve better prediction performance. It is an iterative process where a cell that is difficult to annotate is proposed to the oracle who annotates it, and by doing so moves it to the TS used to train the regression model.

Figure 2


a
b
1
2
3
4
5
6
c
1
2
3
4
5
6
d
regression
classification
1
7
0
0
0.88
0.66
0.96
0.91
7.2
5.6
precision
recall
n° identified processes
e
f
TM4SF5-G
CD300LG-G
E06
L06
E06
AVERAGE NEGATIVE CONTROLS
L06
g
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
C D E F G H I J K L M
Meta-visualisation (PCA)
PC-2
PC-1
h
i
Brightfield intensity std
684
81
Integrated mCherry
2.4
0
j
Cellular density
Cell trajectories
k
Average differentiation speed
Trajectory histogram
0 30
l
Frame 1
Frame 15
Frame 1
Frame 15

(**a**) **Synthetic dataset.** Image from the synthetic dataset, generated using SIMCEP. (**b**) **RP ground truth**. Regression plane generated by automatically placing 300 cells on it. We modelled 6 continuous biological processes, plus an extra process with uniformly distributed cells (*latent process 7*). (**c**) **Ground truth processes.** The 6 continuous processes are modelled between two fixed endpoints: green cells of highly irregular shape and red, rounded cells. To assign a colour to the middle point of each process we interpolated between white (*process 1*) and blue (*process 6*). (**d**) **Classification *vs* regression applied on synthetic data.** Comparison of the performance of regression and classification. Statistics: precision, recall and the number of identified processes. (**e**) **Lipid droplet dataset**. Regression plane of 457 cells representing various lipid morphologies, created by an expert biologist. (**f**) **RP output**. Kernel Density Estimation (KDE)-maps of the predicted regression positions for cells treated with selected siRNAs. Arrows originate from the peak of the control KDE-map, and point to the peaks of the selected KDE-maps. (**g**) **HCS analysis**. Plate-based analysis performed by comparing well-based KDE-maps. Meta-visualization is obtained by extracting the principal components (PC1 and PC2) of the flattened KDE-maps. **(h) Hemocyte dataset analysis**. 109 cells were placed on the regression plane by a microscopy expert. Cells were segmented by applying the NucleAIzer [25] deep learning method on brightfield microscopy images. **(i) Single cell features.** Colour-coded feature values overlay on the predicted cells. **(j) Density plots.** Top: Kernel Density Estimation of single cells. Bottom: 2,323 cell trajectories on the regression plane. **(k) Histogram plots.** Top: Cell differentiation speed on the regression plane. Bottom: Trajectory histogram (2D on the regression plane and 1D projection with trajectory counts) including only those trajectories that reach beyond the green line in *(j)*. **(l) Selected cell trajectories.** Representative phenotypes highlighted in *(j)*.

ONLINE METHODS

Synthetic dataset

We generated a synthetic dataset by modelling 6 continuous biological processes representing continuous changes from one cell state to another, plus an extra process (*latent process 7*) formed from uniformly distributed cells (**Fig. 2b**). To generate the dataset we used a customized version of SIMCEP [26], provided as **Supplementary Material 4**. Synthetic microscopy images were organized into a 24-well plate format, and the dataset was composed of 9 images/well, for a total of 216 images and approximately 10,000 cells. The images of each well were generated by considering a predominant process mixed with other ones. To model the continuous processes we fixed two endpoints: green cells of highly irregular shape, and red, rounded cells (**Fig. 2c**). The degree of cell shape deformation decreases from the green to the red endpoint. Next, for each process we selected a middle point, and assigned a colour to that, ranging from white (*process 1*) to blue (*process 6*). The colour of the cells in each process was then defined by linear interpolation between the colour of the middle point and one of the two endpoints. The generated dataset was deposited to the Broad Bioimage Benchmark Collection (BBBC), and it is freely available at: *https://data.broadinstitute.org/bbbc/image_sets.html* (dataset ID: BBBC031).

Lipid droplet dataset

Lipid droplets are storage units for neutral lipids, including triglycerides, and play a significant role in several disorders, including e.g. cardiovascular diseases. The lipid droplet dataset evaluated with RP was derived from a previous genome-wide association study, in which hepatocytes (Huh7) were transfected with 1-7 siRNAs (10 nM/gene) for 72 h to silence the expression of specific genes, allowing to examine their relationship with lipid formation. The effects of siRNAs on cellular neutral lipids (TG and cholesteryl esters) were scored by using a probe validated for quantitative analysis of neutral lipids. The cells were displaced into a 384-well plate, and after 72 h of siRNA transfection they were fixed with 4% paraformaldehyde, followed by staining for LDs with Green^{TM} (Invitrogen) and for nuclei using 300 nM DAPI (Sigma-Aldrich) for 30 min at room temperature. Finally, 9 images/well were acquired per channel for 2 identical plates with an automated epifluorescence ScanR microscope (Olympus) equipped with a 150W Mercury-Xenon mixed gas arc burner, a 20× long working distance objective (UIS2) and a digital monochrome CCD camera (Hamamatsu), yielding a total of 3,956 images of 232,084 cells (>2,200 cells per siRNA). The list of the siRNAs used and the corresponding target genes is provided as **Supplementary Material 6.** The generated dataset was deposited to FigShare, and it is permanently available at: https://doi.org/10.6084/m9.figshare.c.5067638.v1. To validate our findings, additional biochemical analysis was performed by siRNA-transfecting Huh7 cells, collected in 0.2 N NaOH, followed by extracting the lipids. TGs and cholesteryl esters were resolved on TLC plates using hexane/diethyl ether/acetic acid (80:20:1) as the mobile phase.

MitoCheck dataset

Cai *et al.* [10] analysed cell mitosis by performing time-lapse experiments to establish a canonical model for the morphological changes appearing during the mitotic progression of human cells. In particular, they reorganized the feature space according to the *mitotic standard time* instead of the *imaging time* (see *Fig. 2d* in [10]), and by applying an unbiased peak-detection method in the warped feature space they identified up to 20 mitotic stages. The model was then used to integrate dynamic concentration data of several fluorescently knocked-in mitotic proteins, and to create a generic dynamic protein atlas of human cell division.

Their public data include 3D images and segmented masks of 31 z-stacks. We intended to analyse this dataset without using prior feature information about the underlying process by exploiting regression techniques to characterize mitosis.

In our analysis, a field expert created a regression plane representing the process of mitosis, resulting in a training set of 585 cells (**Suppl. Fig. 4a**). After prediction, the cells followed the designed circular path recalling canonical mitotic phases (**Suppl. Fig. 4b-c**), while they also represented subtle phenotypic changes and single-cell differences in the regression plane. Finally, we compared the results of the original methodology presented by Cai *et al.* (Multi-dimensional Dynamic Time Warping for creating the *standard mitotic time*, **Suppl. Fig. 4d**) with the results obtained by RP (**Suppl. Fig. 4e**), and we concluded that RP is capable of reproducing a mitotic time model equivalent to the original one. This indicates that RP is able to compete with complex analysis techniques, such as DTW. Additionally, RP provides the flexibility to customize the output space, enabling higher resolution analysis of user-defined sections of the biological process.

Blood cell differentiation dataset

The fruit fly, *Drosophila melanogaster*, serves as a popular model system to study innate immune functions, such as phagocytosis, wound healing and capsule formation [20]. In the larva, these functions are executed by hemocytes, which are categorized into three main cell types: (*1*) phagocytic plasmatocytes, accounting for the majority of circulating hemocytes, (*2*) crystal cells, which play a role in melanization and wound healing, and (*3*) lamellocytes, which are large flat cells that appear only in certain tumorous genetic backgrounds or following immune induction [21]

Such an immune induction appears in nature as a result of egg-laying by a parasitoid wasp, *Leptopilina boulardi*. Following infestation, newly differentiating lamellocytes, together with plasmatocytes, eliminate the invader by forming a multilayer capsule around the wasp's egg [27,28,29]. Lamellocytes are also produced when larvae are wounded with an insect pin [30]. (**Suppl. Fig. 5c**)

Cell lineage-tracing studies revealed that plasmatocytes, which had previously been considered as terminally differentiated phagocytic cells, show plasticity, and are capable of differentiating into encapsulating lamellocytes upon immune induction [31,32,33,21]. This trans-differentiation process has been underlined by recent single-cell RNA sequencing studies [22,23]. However, the cells intermediate of the plasmatocyte-lamellocyte transition process have not been characterized morphologically in detail so far, and the routes of differentiation are still controversial. A study by Anderl *et al*. [24] described two types of lamellocytes, and suggested that only the smaller type II lamellocytes (**Suppl. Video 1**) differentiate from plasmatocytes, while the regular, flattened type I lamellocytes (**Suppl. Video 2**) originate from dedicated precursors.

To clarify the potential routes of differentiation, we set up an *ex vivo* method for hemocyte culturing and differentiation. According to Anderl *et al*. [24], for the live experiments, we used *eaterGFP* as a marker of plasmatocytes, and *MSNF9MOmCherry* as a marker of lamellocytes.

Early third instar *Me* larvae (*eaterGFP*, *MSNF9MOmCherry*; [24]) were immune induced by wounding the cuticle with an Austerlitz Insect Pin® of 0.2 mm in diameter. Wounded larvae were kept on standard *Drosophila* food at 25 °C. Circulating blood cells were isolated 12 hours after wounding. Blood samples of 10 larvae were collected, pooled in 300 µl Schneider's medium (Lonza, Cat: 04-351 Q) supplemented with 10% fetal bovine serum (FBS; Gibco, Cat: 10270) plus 0.01 mg/ml gentamicin (Sigma, Cat: G3632), 0.065 mg/ml penicillin (Sigma, Cat: P7794) and 0.1

mg/ml streptomycin (Sigma, Cat: S6501). Next it was spread into a well chamber of an 8-well μ-slide (Ibidi, Cat: 80826). Both sample storage and microscopic analysis were carried out at 25 °C.

We acquired 15-frame image sequences/field (141 fields) on 3 channels: brightfield, mCherry, and EGFP, with 2-hour-gaps between the subsequent frames. Images were acquired with a high-content screening microscope (Operetta, Perkin Elmer) equipped with a 60× high-numeric-aperture objective and a digital high resolution 14-bit CCD camera, yielding a total of 4,230 images (2 plates, 2,115 images in each). The image size was 1360×1024 pixels and 8-bit per channel, in TIFF format. The generated dataset was deposited to FigShare, and it is permanently available at: https://doi.org/10.6084/m9.figshare.c.5075093.v1.

Using the method described above, we found that 5.6% of the plasmatocytes are capable of trans-differentiation into lamellocytes (**Suppl. Videos 3-4**), which is well reflected by the expression of cell type specific transgenes. After the formation of lamellocytes, no significant alterations in their cell size were observed, indicating that all types of lamellocytes are terminally differentiated cells. Most of the plasmatocytes (94.4%), however, did not differentiate into lamellocytes, but either spread out, increasing their cell size, or kept their size and morphology during the experiment, which is in line with the results of *in vivo* studies on blood cell differentiation in *Drosophila*.

Image segmentation and feature extraction

In order to classify the cells in an image, ACC requires the position and features of each cell to be analyzed. For this purpose, we first flattened illumination distortions of the acquired images by using CIDRE [34]. Then, we used CellProfiler [35] and the NucleAIzer deep learning framework [25] to segment the cells and extract the standard features describing morphology, intensity and texture characteristics. Details of the image analysis and the regression models used in each experiment are reported in **Supplementary Note 2**.

Regression models

Regression methods, a subgroup of supervised machine learning techniques, are aiming at approximating continuous target variables. Alike for classification, various models have been proposed for regression, ranging from linear regression to neural networks and random forests [36]. The diverse set of regression models raise the problem of model selection for RP. As the RP is completely user-defined, it is impossible to have any prior assumptions on the function to be learnt, hence model selection should be data-driven. RP provides cross-validation assessment of model performance by root mean squared error measure (RMSE) and relative RMSE [37]. Additionally, two important aspects are to be considered when selecting the model.

First, the two-dimensional output format of RP requires the use of multi-target regression, as we require a 2D position (expressed by 2 coordinates) to be predicted. Traditionally, regression models aim at predicting a single continuous variable, which may be naturally extended for multiple dimensions by considering the outputs as independent variables, also called the single-target (ST) method [38]. On the contrary, it has been reported several times that multi-target models that exploit the possible correlation between the output variables may yield significantly better results than the ST methods [39,40]. Consequently, when a strong relationship between the output variables is evident, choosing a multi-target regression model is more appropriate.

Secondly, models that are capable of providing a probabilistic output (i.e. those that provide not only the predictive mean, but also some sort of uncertainty) are less wide-spread for regression than for classification. However, uncertainties provide valuable information to assess the model's performance, and most of the active learning strategies essentially rely on them.

Gaussian processes (GPs) can be used as non-parametric regression models with a probabilistic output [41]. Instead of providing a single prediction for each cell, GP returns a normal distribution whose mean can be used as the predicted value. More importantly, its variance is an estimate for the uncertainty of the given cell. GP is originally considered as a single-target method, however, its multi-target extensions also exist and are known as co-kriging [42,39].

Although GP is a non-parametric method (hence training is not required in principle), it still has hyperparameters (mean, covariance, likelihood, inference functions and their parameters) that can be optimized for enhanced performance. The most frequently applied iterative optimization methods (gradient descents) require initial hyperparameter settings which significantly affect the quality of the ultimate hyperparameter set. Consequently, we have designed heuristical hyperparameter initialization methods for several mean and covariance functions as described in **Supplementary Note 3**. Due to the broad selection of implementable models, RP provides an interface (via Object Oriented Programming) to facilitate the extension of implemented regression methods. By default, the package contains bridges to several models from Weka [43], Mulan [44] and Matlab’s Deep Learning Toolbox. The full list and instructions on how to include new models are provided in **Supplementary Note 4**.

Active regression

Usually, the most time consuming part of statistical learning for biomedical applications (including shallow and deep learning) is the procedure of annotation, and – as transfer learning is rarely used – it is often repeated for new experiments. Active learning [45] aims at reducing the number of training samples needed to achieve the most representative training set by automatically proposing cells for annotation. It has previously been shown by Smith and Horvath [46] that active learning reduces the time cost of annotation in HCS compared to classical labelling. Most of the active classification methods are based solely on the predicted class labels, enabling the underlying model to be freely modified. However, these methods are not directly applicable for regression, as they assume that the predicted label is discrete. Active regression methods were developed by Cohn *et al.* [47], based on variance reduction for Neural Networks, Mixture of Gaussians and Locally Weighted Regression. Here we present novel active regression methods inspired by the general active classification approaches, and a specific method for Gaussian Processes utilizing its properties (**Suppl. Fig. 1**).

Committee Members. The Committee Members approach is inspired by the *QueryByCommittee* active classification method. Similarly to cross-validation, a set of models (committee) is built up from the available training samples, and a measure of disagreement is defined for the committee. In case of regression, the classical measures cannot be applied directly for two reasons: (*1*) they rely on the fact that the output is discrete, and (*2*) they require a probabilistic model. Thus, we propose using the quadratic mean of the Euclidean distance between the committee consensus and the single committee predictions. Hence, the next cell to be labelled by the expert is defined by the following formula:

$$x^* = \underset{x}{\operatorname{argmax}} \sqrt{\sum_{i=1}^{C} \frac{\mathrm{d}(\hat{y}_i, \bar{y})^2}{C}}$$

where *C* is the size of the committee, $\hat{y}_i$ is the predicted position for *x* (a sample not taken from the TS) by the i*th* committee member, $\bar{y}$ is the mean of $\hat{y}$, and *d* is the Euclidean distance.

Empty Regions. The Empty Regions method targets the cells which were predicted to the least dense region of the regression plane in terms of training samples. This heuristic is supposed to explore those cell types that are not presented in the TS.

Out of Bounds. By design, the regression plane is represented by a unit-square, and has limits in each direction. However, this limitation was not incorporated into the regression models, consequently it is possible that cells are predicted outside of the regression plane's boundaries. Therefore, we propose a strategy that selects these cells for annotation, ranked by their distance from the edges of the regression plane.

Uncertainty Sampling. When a probabilistic regression model (such as GP) is available, then, instead of plain predictions, a posterior distribution is defined for each cell, enabling the application of active learning methods aiming at decreasing the variance of this posterior. Our proposed method targets the cell with the highest posterior variance, where the final value for the selection is determined by taking either the mean, the sum, the product, the minimum or the maximum of the 2 separate variances, calculated for each output dimension of the regression plane.

Overall Uncertainty Sampling. GP has an intriguing property, namely that the posterior distribution is independent of the actual TS positions; it only depends on the input features and the hyperparameters of the GP. In consequence, given fixed hyperparameters, it is possible to exactly calculate how the posterior variance changes, assuming that a new cell is included in the TS even without knowing its position on the regression plane. Executing this calculation for all possible candidates, the resulting cell proposed for annotation is the one that decreases overall variance the most. This approach is formulated by:

$$x^* = \underset{x}{argmin}\Big(\sum_{i}^{N} f_\sigma^x(x_i)\Big)$$

where *N* is the size of the full dataset (including the training dataset) and $f_\sigma^x(x_i)$ is the variance for *x*$_i$, supposing that the GP was trained on the available training set extended with *x*. The predictive variances for individual samples are calculated from the diagonal elements of the predictive variance matrix according to [41] by the following formula:

$$K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*))$$

where K is the kernel (covariance) function, X_* is the feature matrix of samples not yet predicted, and *X* is the feature matrix of the training set's elements.

We assessed the performance of the proposed active learning methods with 4 regression models: Random Forest, Gaussian Process, Neural Network and Support Vector Machine; on 2 of our datasets: *Lipid droplets* and *MitoCheck* containing 457 and 586 annotated cells respectively. In each scenario the experiment started with randomly isolating ⅓ of the available samples to a *test set*, leaving the remaining ⅔ in a pool. Then, 10 cells were randomly selected from the pool for initializing the *training set*, followed by iteratively extending it with 290 cells according to the active query strategy. In each iteration a regression model was trained, and the relative root mean square error (RRMSE) was calculated on the test set.

The results from 50 independent runs are displayed on **Supplementary Fig. 1b-c**. In all but one (Gaussian Process in the MitoCheck dataset) scenario there was at least one active learning technique that outperformed random sampling, despite the high variance of error values among different regression models. The Random Forest and Gaussian Process models achieved smaller RRMSE values than the other two methods inhibiting the active strategies' ability to significantly improve the performance in these cases. Still, the CommitteeMembers strategy resulted in the lowest average area under the curve value in 5 out of the 8 cases. We also note that although mean prediction error is the most widespread measure of active learning, other aspects of the model performance (e.g. model coverage) might be equally interesting for the users.

Regression Plane output

RP provides output in various formats to satisfy the diverse needs of field experts (**Suppl. Material 3**). The simplest output can be obtained by predicting an image in the main window of ACC, by clicking on a cell to see its raw regression plane position. Alternatively, in the regression plane one can select an arbitrary number of images, so that all cells in those images are going to be visualized on the regression plane with their icon at their predicted position. Importantly, these predictions can easily be added to the TS as well.

For well-based analysis, a multi-component report can be generated for each plate. The first component of the report is a pdf file containing a heatmap (simple cell count in a discretized regression plane) and a kernel density estimation (KDE) visualizing the distribution of cells on the regression plane in the particular well (**Fig 2f-g**). Besides, the difference and the most dense position shift between single wells, and the average of user-defined control wells are also included. Secondly, RP provides standard visualization tools (PCA, t-SNE [48] and NeRV [49]) for assessing the relationships among the wells. Each of these methods can generate the figure of *Plot of plots (PoP;* **Fig 2h***)*. In PoP each well is represented by its KDE/heatmap, and the distance between these representations corresponds to the difference between the wells' regression plane distributions (i.e. similar wells are close in PoP, whilst differing ones are farther from each other). In case of plates with higher well-numbers (e.g. 96 or 384) this may result in an overwhelmingly dense diagram, so the PoPs can be re-loaded to RP where they can be examined interactively. Importantly, in the RP-PoP, wells of similar perturbations (replicates) can be highlighted with colours. In addition to these tools for visualization, a *clustergram* can also be generated, providing a way to compare the perturbations by performing hierarchical clustering (**Suppl. Fig. 3**). The matrix in the middle of the clustergram visualizes pairwise Kullback-Leibler divergence between the cell-number weighted average of the replicate wells.

Additionally, RP enables the analysis of underlying image features by the *Colour Frame* (CF) module. CF works by visualizing the feature distribution of cells from the regression plane, using an artificial colour scale. In particular, the user selects a specific feature and adjusts the visualization settings to define a colour for each cell icon's frame in the regression plane. (**Suppl. Fig. 5**). Notably, CF can be used either for fine tuning of the TS, or for assessing features of interest after prediction.

Finally, the *Trajectory Plot* (TP) facilitates the assessment of live-cell data composed of time-resolved image sequences of the same fields. Organizing the corresponding single-cells into trajectories using the predicted coordinates of the regression plane enables the visualization of the dynamics of underlying processes (**Fig. 2j-k**). TP is a multifunctional visualization tool that facilitates a better understanding of the continuous aspect of biological processes, and offers several possibilities to investigate cell fates or to compare the development of particular cells as a function of time. Filtering functions help to find subgroups of phenotypes with different behaviours. Interestingly, the dynamics of the process can be perceived by animating the evolution of trajectories (**Supplementary Videos 5-9**).

Code Availability

RP is a new module of ACC (current version 3.1). ACC is written in MATLAB (The MathWorks, Inc., USA). ACC supports the most common image formats (*e.g.* tif, bmp, png) and it works under Windows 64-bit, Linux, and OS X environments. Source code and standalone versions (which do not require a MATLAB license), video tutorials, and help documentation files are publicly available at: *www.cellclassifier.org*. All the ACC materials are copyright protected and distributed under GNU General Public License version 3 (GPLv3).

Data Availability

Synthetic dataset: https://data.broadinstitute.org/bbbc/image_sets.html (dataset ID: BBBC031).

Lipid droplet dataset: https://doi.org/10.6084/m9.figshare.c.5067638.v1.

Drosophila dataset: https://doi.org/10.6084/m9.figshare.c.5075093.v1.

REFERENCES (ONLINE METHODS)

26. Lehmussola, A., Ruusuvuori, P., Selinummi, J., Huttunen, H., & Yli-Harja, O. Computational framework for simulating fluorescence microscope images with cell populations. *IEEE Transactions on Medical Imaging*, **26**(7), 1010-1016 (2007).

27. Nappi, A. J., Vass, E., Frey, F., & Carton, Y. Superoxide anion generation in Drosophila during melanotic encapsulation of parasites. *European Journal of Cell Biology*, **68**(4), 450-456 (1995).

28. Russo, J., Dupas, S., Frey, F., Carton, Y., & Brehelin, M. Insect immunity: early events in the encapsulation process of parasitoid (Leptopilina boulardi) eggs in resistant and susceptible strains of Drosophila. *Parasitology*, **112**(1), 135-142 (1996).

29. Lanot, R., Zachary, D., Holder, F., & Meister, M. Postembryonic hematopoiesis in Drosophila. *Developmental Biology*, **230**(2), 243-257 (2001).

30. Márkus, R., Kurucz, É., Rus, F., & Andó, I. Sterile wounding is a minimal and sufficient trigger for a cellular immune response in Drosophila melanogaster. *Immunology Letters*, **101**(1), 108-111 (2005).

31. Stofanko, M., Kwon, S. Y., & Badenhorst, P. Lineage tracing of lamellocytes demonstrates Drosophila macrophage plasticity. *PloS One*, **5**(11), e14051 (2010).

32. Kroeger Jr, P. T., Tokusumi, T., & Schulz, R. A. Transcriptional regulation of eater gene expression in Drosophila blood cells. *Genesis*, **50**(1), 41-49 (2012).

33. Honti, V., Csordás, G., Márkus, R., Kurucz, É., Jankovics, F., & Andó, I. Cell lineage tracing reveals the plasticity of the hemocyte lineages and of the hematopoietic compartments in Drosophila melanogaster. *Molecular Immunology*, **47**(11-12), 1997-2004 (2010).

34. Smith, K., Li, Y., Piccinini, F., Csucs, G., Balazs, C., Bevilacqua, A., & Horvath, P. CIDRE: an illumination-correction method for optical microscopy. *Nature Methods*, **12**(5), 404-406 (2015).

35. Carpenter, A. E. et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, **7**(10), R100 (2006).

36. Hastie, T., Tibshirani R., & Friedman J. *The Elements of Statistical Learning*. Second edition. Springer (2008).

37. Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W., & Vlahavas, I. Multi-target regression via input space expansion: treating targets as inputs. *Machine Learning*, **104**(1), 55-98 (2016).

38. Borchani, H., Varando, G., Bielza, C., & Larrañaga, P. A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **5**(5), 216-233 (2015).

39. Boyle, P., & Frean, M. Dependent gaussian processes. *In: Advances in Neural Information Processing Systems (NIPS), December 5-8, 2005, Vancouver, British Columbia, Canada*, 217-224 (2005).

40. Han, Z., Liu, Y., Zhao, J., & Wang, W. Real time prediction for converter gas tank levels based on multi-output least square support vector regressor. *Control Engineering Practice*, **20**(12), 1400-1409 (2012)

41. Rasmussen, C. E., & Williams, C. K. I. *Gaussian Processes for Machine Learning*. MIT Press (2006).

42. Cressie, N. A. C. *Statistics for Spatial Data*. Chapter 3. Spatial Prediction and Kriging, 105-209. John Wiley & Sons (1993).

43.Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, **11**(1), 10-18 (2009).

44. Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., & Vlahavas, I. Mulan: A java library for multi-label learning. *The Journal of Machine Learning Research*, **12**, 2411-2414 (2011).

45. Settles, B. Active learning literature survey. *University of Wisconsin-Madison Department of Computer Sciences* (2009).

46. Smith, K., & Horvath, P. Active learning strategies for phenotypic profiling of high-content screens. *Journal of Biomolecular Screening*, **19**(5), 685-695 (2014).

47. Cohn, D. A., Ghahramani, Z., & Jordan, M. I. Active learning with statistical models. *Journal of Artificial Intelligence Research*, **4**, 129-145 (1996).

48. Maaten, L. V. D., & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, **9**(11), 2579-2605 (2008).

49. Venna, J., Peltonen, J., Nybo, K., Aidos, H., & Kaski, S. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, **11**(2), 451-490 (2010).

SUPPLEMENTARY MATERIALS

Supplementary Material 1

Regression Plane user manual.

Supplementary Material 2

Video tutorial: "*Regression Plane: Annotation Possibilitie*s".

Supplementary Material 3

Video tutorial: "*Regression Plane: Output Possibilities*".

Supplementary Material 4

Customized version of SIMCEP, distributed as MATLAB source code under the GNU General Public License version 3.

Supplementary Material 5

MATLAB script provided to the 5 microscopists analysing the synthetic dataset using standard classification approaches.

Supplementary Material 6

Plate layout reporting the targeted genes of lipid droplet screen.

Supplementary Note 1

Description of the synthetic data experiment.

Supplementary Note 2

Experiment specific Image Analysis pipelines and Regression models.

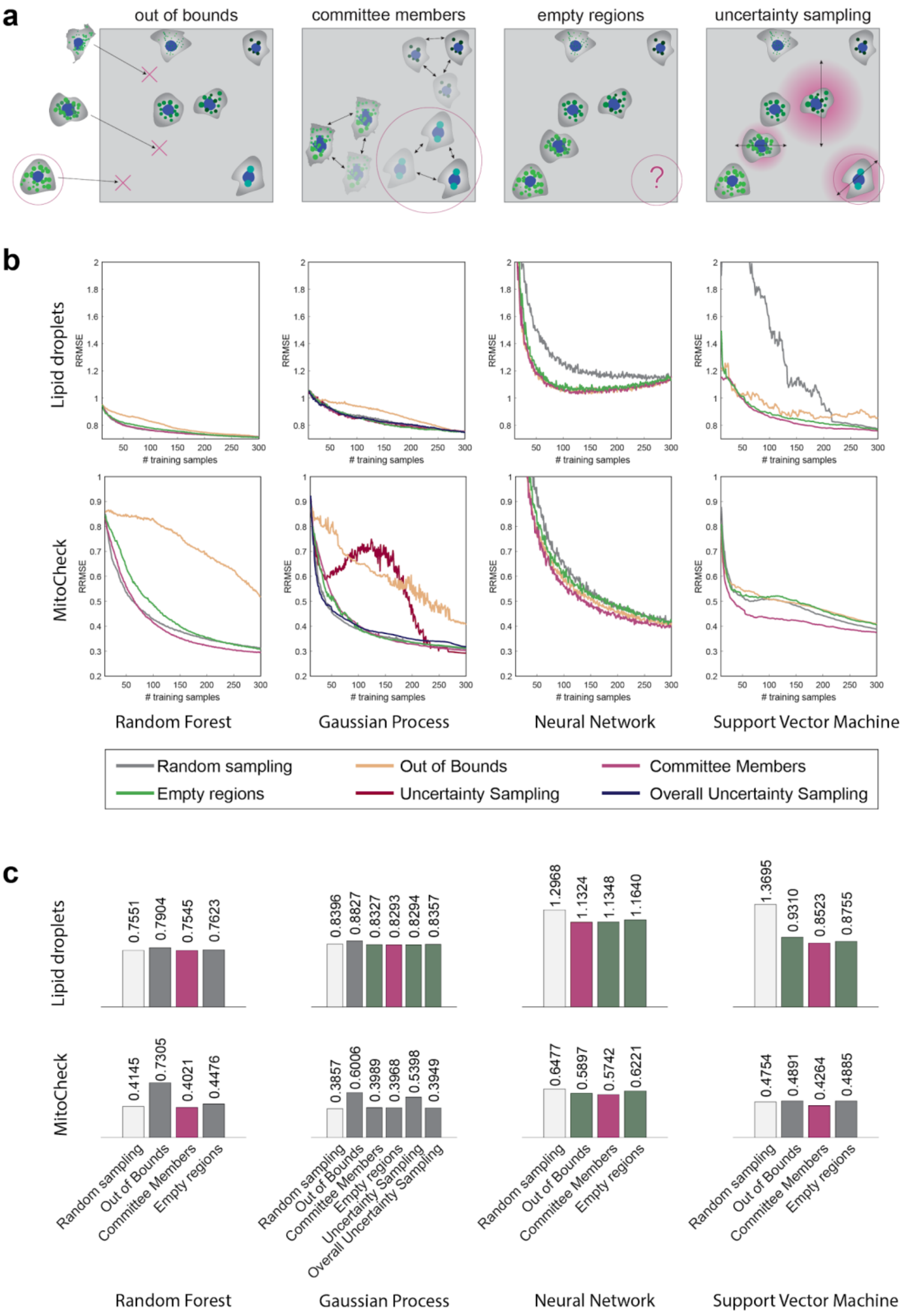
Supplementary Note 3

Technical details on hyperparameter initialization.

Supplementary Note 4

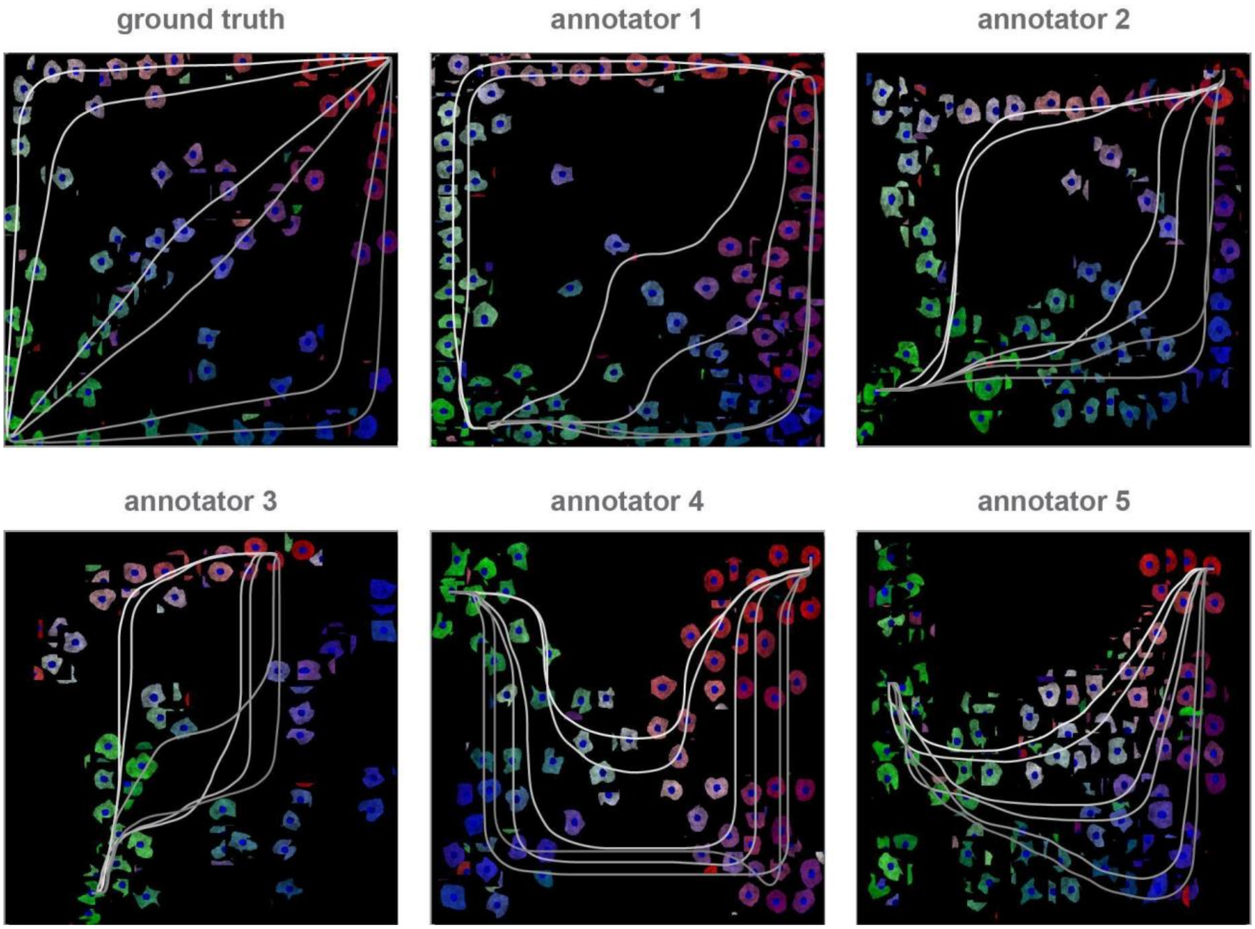
Software developer guide for extending the framework with new Predictors and Active Learning methods.

Supplementary Figure 1



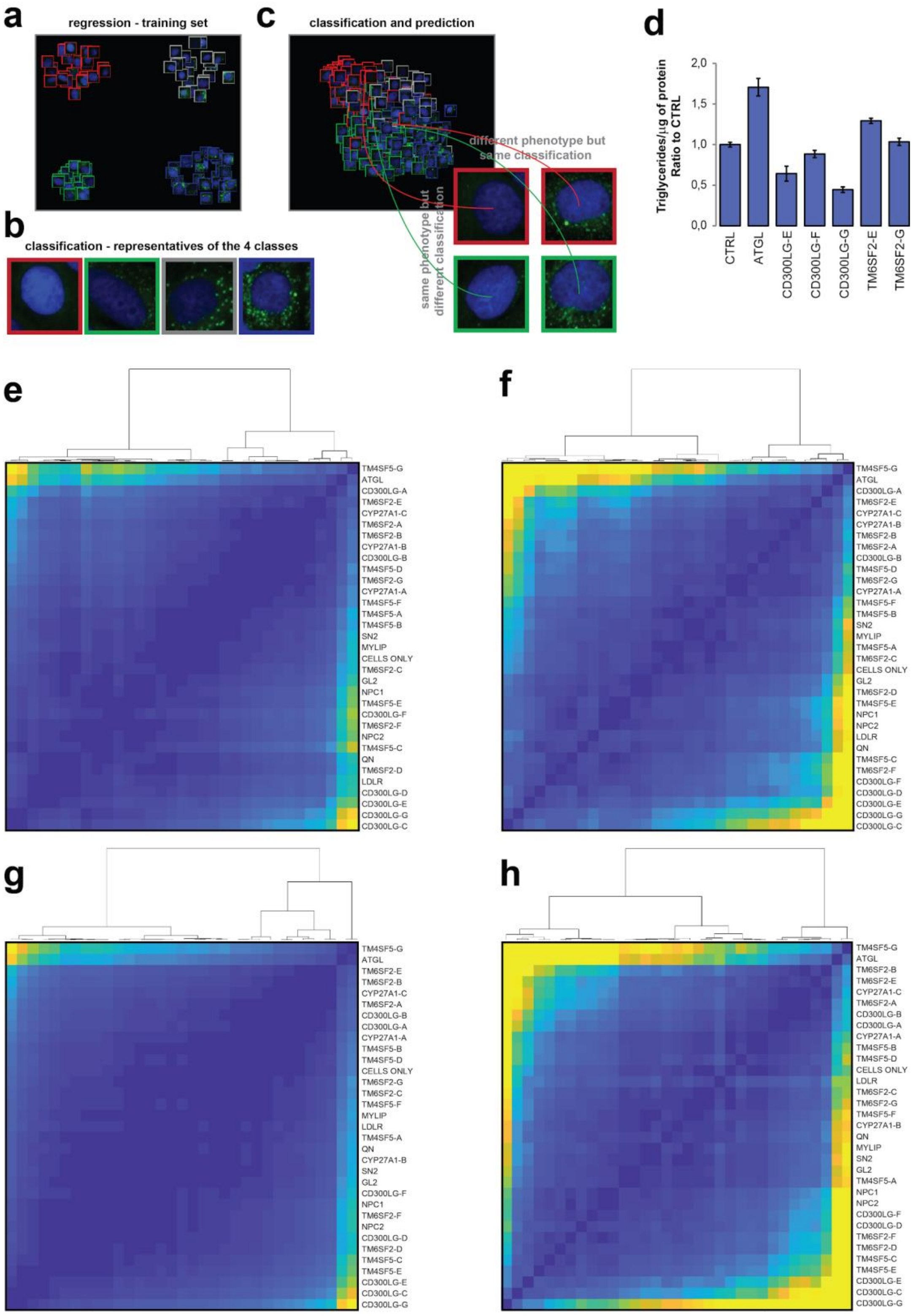
Active regression. (**a**) Schematic representation of four active regression algorithms implemented in ACC. **(b)** Performance of the proposed methods measured as *Relative Root Mean Squared Error* (RRMSE) **(c)** Performance of the proposed methods measured as the average *Area Under RRMSE Curve* (lower is better). Methods showing superior performance to random sampling are highlighted with green, and the best among these with pink. The plots in both *(b)* and *(c)* represent the mean from 50 independent runs. Gaussian Processes were trained with constant mean function, squared exponential covariance function with automatic relevance determination (covSEard) for Lipids and with isotropic distance (covSEiso) for MitoCheck. The Neural Networks were trained with a single layer containing 30 nodes with log-sigmoid activation function. Random Forest and Support Vector Machine were trained with default parameters from Weka. The size of the committee in the CommitteeMembers method was 3.

Supplementary Figure 2



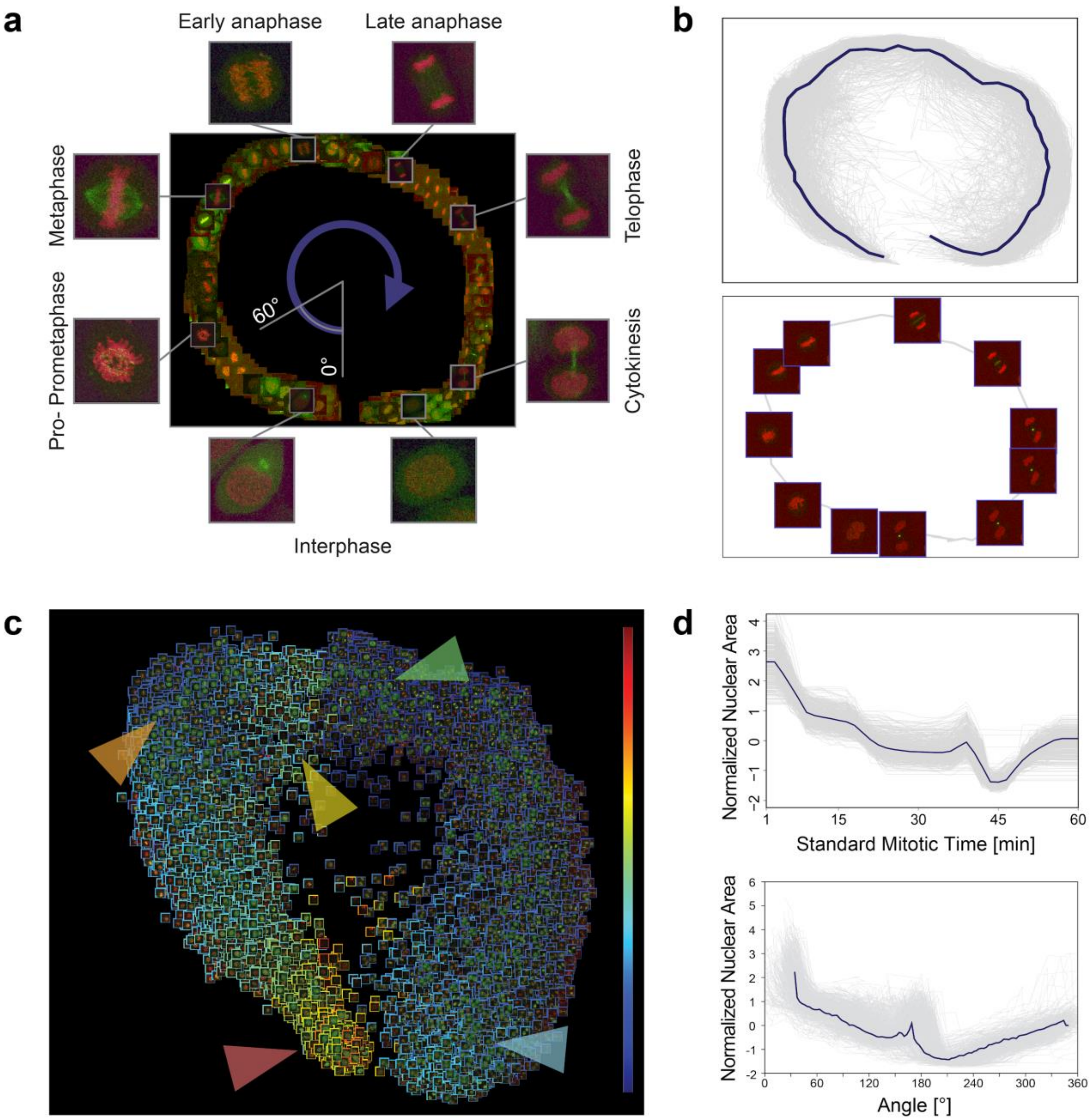
Synthetic dataset: Regression Planes. The ground truth regression plane and the annotations created by the 5 microscopy experts (*i.e.* annotators) who analysed the synthetic dataset with RP. The gray lines represent the identified processes and have been computed using a Kernel Density Estimation function and an energy minimization algorithm for finding the shortest path between process endpoints, using Dijkstra's algorithm. Despite the great variety of the regression planes generated by the annotators, in all the cases except for *annotator 1*, the six non-latent continuous processes are represented by separated lines.

Supplementary Figure 3



Lipid Droplet data analysis. Classification *vs* regression applied on real-world data. (a) Starting from a regression plane including 457 cells created by a microscopy expert, we have automatically selected 25 cells that were the closest to the center of the 4 quadrants of the regression plane, representing the 4 main cell phenotypes. **(b)** To visualize the cells belonging to the 4 different classes, we used borders of different colours (i.e. red, green, gray, blue). **(c)** Next, we classified the unannotated cells, and simultaneously predicted their position in the regression plane. The test revealed several cases of misclassifications: some cells with the same phenotype were classified into different classes, while several cells with a clearly different phenotype were classified into the same class. **(d) Biochemical analysis**. Intracellular TG levels in cultured hepatocytes (Huh7). siRNA-mediated knockdown of *TM6SF2* gene led to an increased level of TGs. In contrast, siRNAs targeting *CD300LG* decreased intracellular TG levels. **(e-h) Discovery tool: clustergram.** *Clustergrams* obtained by calculating symmetric Kullback-Leibler divergence for the KDE-maps (regression) / class probability distributions (classification) of the wells treated with different siRNAs. Bright (yellow) values indicate high divergence, meaning that the cells in the wells compared to each other have different morphologies. Clustergrams from regression show higher variation in the divergence values, better capturing subtle differences in the cell populations treated with siRNAs. **(e)** *Plate 01*, classification; **(f)** *Plate 01*, regression; (**g**) *Plate 02*, classification results; **(h)** *Plate 02*, regression results.

Supplementary Figure 4

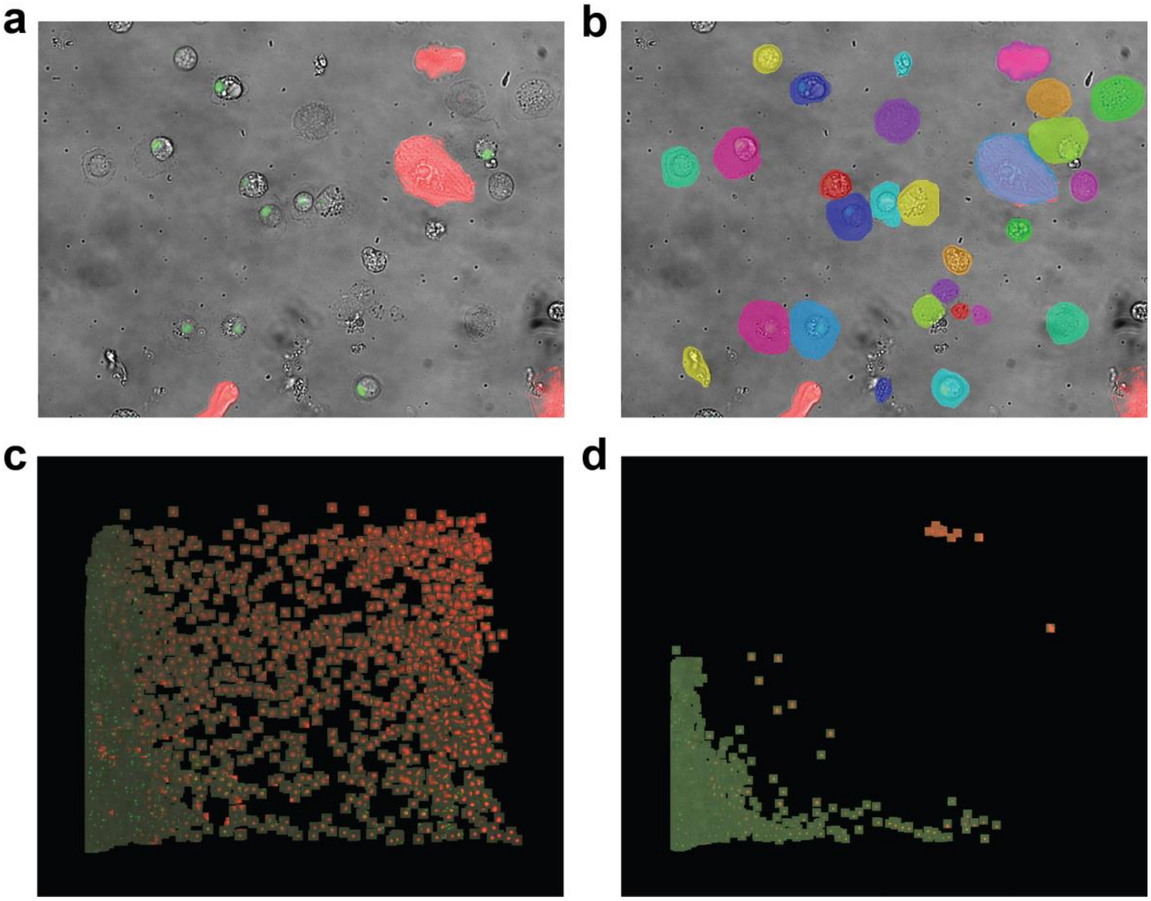


Mitosis data analysis. (a) Regression plane of 585 cells annotated by a microscopy expert. **(b)** *Top:* 505 trajectories for all the predicted cells. The median curve is shown in solid blue. *Bottom:* Example of a single-cell trajectory with representative cell icons visualized. **(c)** Regression plane with all (n = 20200) predicted cells. The borders of the cell icons correspond to their nuclear area (Colour Frame module). Highlighted regions: Early prophase region, large nuclear area (red).

Metaphase region, nuclear area decreased (orange). Early-anaphase region, nuclear area is increasing as spindle fibers are pulling chromosomes apart (yellow). Anaphase, nuclear area dropped as the nucleus is considered as two separate objects with half the area (green). Late-telophase, nuclear area increasing up to half of the initial value (blue).

(**d**) *Top:* Trend for the normalized nuclear area according to *standard mitotic time*. Gray lines represent single cell trajectories. *Bottom:* Trend for the normalized nuclear area according to the regression plane. Gray lines represent single cell trajectories. The coordinates predicted by RP were converted to 1D by taking the angle argument of the polar coordinate representation as illustrated in *(a)*.

Supplementary Figure 5



Blood Cells data validation and segmentation. (a) Exemplary composite image with 3 channels: brightfield (gray), GFP (green), mCherry (red). **(b)** Corresponding results of deep-learning segmentation performed on the brightfield channel. **(c)** Prediction of all the immune induced cells on the regression plane. **(d)** Prediction of all cells from the control experiment on the regression plane.

Supplementary Videos 1-9

The supplementary videos are dynamically visualizing the regression plane analysis of live-cell experiments. The rectangle in the videos represents the regression plane itself and the trajectories are derived from the live-cells' *predicted positions* on the plane. Each individual trajectory is assigned to a single-cell and the animation shows how the cells traverse on the regression plane as the live-cell screening progresses (the path between the actual frames were linearly interpolated).

Supplementary Video 1

Drosophila Plasmatocyte Differentiating into Type II Lamellocyte

Supplementary Video 2

Drosophila Plasmatocyte Differentiating into Type I Lamellocyte

Supplementary Video 3

Drosophila Plasmatocyte Differentiation, dynamic visualization of trajectories using protein expression. Trajectories are coloured dynamically, visualizing changes in 2 selected cell features.

Head-colour: integrated intensity value of eaterGFP representing its expression level.

Tail-colour: integrated intensity value of MSNF9MOmCherry representing its expression level.

Expression of eaterGFP was observed in a fraction of both type I and type II lamellocytes, however, type II lamellocytes express GFP more frequently and at a higher level.

Supplementary Video 4

Drosophila Plasmatocyte Differentiation, differentiation speed. Trajectories are coloured dynamically visualizing the speed of the cells on the Regression Plane. According to the defined training strategy, this reflects the speed of differentiation. Colours are ranging from blue (slow) to red (fast). Following immune induction, type II lamellocytes start differentiation later than type I lamellocytes, however type II lamellocytes differentiate faster and in a continuous manner.

Supplementary Video 5

Dynamic Visualization of Mitosis. Trajectories show how the cells are traversing on the Regression Plane. The highlighted 4 cells are reported in detail in further supplementary videos.

Supplementary Videos 6-9

Separate, single-cell videos of the highlighted cells in Supplementary Video 5.