# Cell identification and phenotyping using classical machine learning and deep learning

**Theses Of The Ph.D. Dissertation**


**Tamás Balassa**

*Supervisor:*

**Péter Horváth, PhD.**

Eötvös Loránd University

Doctoral School of Informatics

Information Systems Doctoral Program



December, 2021


**DOI: 10.15476/ELTE.2021.207**

## Introduction

One of the greatest scientific achievements of the past century is the complete sequencing of the human genome. Today, many believe that the next great challenge in biology lays in "phenomics" – the quantification of the set of phenotypes that completely characterizes an organism (the "phenome") to define its behavior, morphology or various properties. By collecting and analyzing rich phenotypic data, we expect to improve our understanding of how genetic and environmental factors give rise to changes in organisms or in their behavior. Building on the insight into pthenotypic information we could better predict important outcomes such as fitness, reproduction, crop yield, disease development, cancerogenesis, resistance, or mortality. However, in contrast to the genome, a complete understanding of the phenome is impossible with current technologies, as the complexity and information content of the phenome vastly exceeds that of the genome. Thus, it is crucial to intelligently choose what to measure and which phenomics tools to use, and thereby to develop

novel, high-throughput methods capable of in-depth phenotypic data analysis.

Image-based cellular analysis is a possible approach to accomplish such a task. Recent image analysis and machine learning technologies have enabled the characterization of biology and human health in multiple dimensions, even at the single cell level. By revealing unique characteristics of any cells of interest, single cell analysis offers significant advantages in various fields of science from cancer or brain research through drug development to personalized and targeted treatments. Usually the first step in most microscopy-based single cell analyses task is the identification of nuclei. Precise localization and identification are fundamental for the accurate assessment of important cell functions. Previously the leading approaches were based on classical image processing methods which usually require prior knowledge to accurately fine-tune the parameters assessed.

Recently, the revolutionary approach of deep learning has been incorporated into image analysis, and

has fundamentally contributed to achieve groundbreaking results in single cell analysis, including image classification, image registration, object detection and recognition, and its applications such as computer-aided diagnosistics.

This thesis consists of two parts: phenotypic analysis (with machine learning) and single cell analysis (with deep learning). In the first part, I have focused on answering major questions in phenotypic data analysis, including how to discover the whole dataset effectively or how to increase the accuracy of classification. In the second part the primary focus is on deep learning based approaches. During my research I used and developed various software tools and created algorithms to support in-depth single cell analysis. All of these are aimed to support a higher level of single cell detection and segmentation. This part covers some major research tasks in which I have participated, thus a full automation of a patch-clamping system previously driven manually and a standalone tool to help the user analyze object

occurrences on specific regions of an image are presented here.

**Thesis 1**

*We showed that machine learning is capable of identifying and phenotyping single cells.*

Machine learning has become the key to solve many challenges in bioinformatics. It has several types, the most commonly used ones are either based on existing samples for training (supervised learning) or lack training samples (unsupervised learning). Meanwhile, microscopy image analysis offers a wide ground for the application of machine learning methods [I, II]. Our main achievements in the field of machine learning-based software development include three machine learning approaches applied to support phenotypic analysis: the Advanced Cell Classifier software [III, IV, V], the Phenotype Finder module and the Regression Plane [VI].

Specifically, we implemented a software for cell-based phenotyping, called the Advanced Cell Classifier (ACC). We aimed to enhance phenotype-based research by an easy-to-use tool. It contains a package of numerous machine learning methods (SVM, logistic regression, multilayer perceptron, etc.) and offers multiple modules to support wide utility. One of its modules, the Phenotype Finder, is also a major contribution to the field. It offers the full exploration of a given dataset. By combining a supervised method (one-class classifier, step 1) and an unsupervised method (hierarchical clustering, step 2), the user will receive a cluster of either never seen cells or particular cells that belong to an already existing class. This method allows exploring the whole dataset. In addition, using the similar cell searching module the user can extend the under-represented classes by adding more training samples to make the training process more precise.

In the latest version of Advanced Cell Classifier, we have introduced the concept of the Regression Plane. Using ACC on a daily basis we realized that we lack a tool

that deals with objects which are not optimal to be classified into distinct classes. Regression Plane was developed to overcome this limitation. It is a 2D plane on which the user can place the cells manually, and regression-based machine learning methods are applied to analyze continuous data to perform phenotyping in a continuous manner.

**Thesis 2**

*We showed that Deep Learning is capable of finding cells at high precision.*

The past decade is highlighted by unprecedented IT innovations, with technical novelties opening the path for the realization of a more advanced machine learning method. On the other hand, the gradual development of the theoretical basis for deep learning had already attained to a level of excellence by that time. A unique fusion of these top scientific theories and hardware technologies has recently introduced deep learning into various fields of scientific research and IT applications. Among the

complex solutions it offers, we must highlight its outstanding value in image analysis by revolutionizing image classification, face and object recognition tasks, as well as scene segmentation. Thus, deep learning has quickly become the best approach for many challenges in the IT field.

To keep up with the evolution of technologies, we learned and then realized the benefits deep learning can offer. Accordingly, we have turned our focus on deep learning based software developments [VII], resulting in novel software tools to improve single cell analysis: Autopatcher [VIII] – an automatic deep learning-driven patch clamp system, FindMyCells [IX] – an approach for astrocyte detection, and the upgraded version of FindMyCells for detecting and classifying microglia [X].

Autopatcher is an automatic patch clamp system. In the usual process of patch clamping an expert focuses on too many tasks simultaneously. An essential step is usually the careful selection of the best possible neuron from a 3D tissue for patch-clamping. For the human eye it is not so convenient to search for the target neurons, but

our method was capable to perform this task with high precision.

FindMyCells (FMC) is another deep learning based software we have developed. To find and count thousands and thousands of cells in microscopic images is usually exhausting and time-consuming. FMC was created to replace the human expert's job within this task. It is a flexible, standalone, platform-independent deep learning based software tool which has proven to be almost as precise as a human expert can be, as demonstrated by detecting astrocytes in bright-field microscopy images.

As an upgrade to the original version of FMC we have extended it to be capable of classifying objects after finding them. We have demonstrated its competence by detecting and classifying microglia in brain tissue images, where it was able to maintain detection accuracy of the original version, with the addition of an extremely precise object classification.

In summary, we have successfully demonstrated how to apply machine learning methods effectively by implementing flexible software tools appropriate for phenotyping. Thereby, we have successfully accomplished the main aim of our research to improve time-efficiency of biological image-based data analysis for human experts. We have also demonstrated that deep learning, a more advanced machine learning technique, utilized in various approaches is capable of finding cells at high precision. Thereby, it has the potential to support automatizing and replacing human tasks in single cell research.

[I] Kevin Smith, Filippo Piccinini, **Tamas Balassa**, Krisztian Koos, Tivadar Danka, Hossein Azizpour, Peter Horvath (2018). Phenotypic image analysis software tools for exploring and understanding big image data from cell-based assays. *Cell systems*, *6*(6), 636-653.

[II] Timea Toth, **Tamas Balassa**, Norbert Bara, Ferenc Kovacs, Andras Kriston, Csaba Molnar, Lajos Haracska, Farkas Sukosd, Peter Horvath (2018). Environmental properties of cells improve machine learning-based phenotype recognition accuracy. *Scientific reports*, *8*(1), 1-9.

[III] Filippo Piccinini*, **Tamas Balassa***, Abel Szkalisity, Csaba Molnar, Lassi Paavolainen, Kaisa Kujala, Krisztina Buzas, Marie Sarazova, Vilja Pietiainen, Ulrike Kutay, Kevin Smith, Peter Horvath (2017). Advanced cell classifier: user-friendly machine-learning-based software for discovering phenotypes in high-content imaging data. *Cell systems*, *4*(6), 651-655.

[IV] Csilla Brasko, Kevin Smith, Csaba Molnar, Nora Farago, Lili Hegedus, Arpad Balind, **Tamas Balassa**, Abel Szkalisity, Farkas Sukosd, Katalin Kocsis, Balazs Balint, Lassi Paavolainen, Marton Z Enyedi, Istvan Nagy, Laszlo G Puskas, Lajos Haracska, Gabor Tamas, Peter Horvath (2018). Intelligent image-based in situ single-cell isolation. *Nature communications*, *9*(1), 1-7.

[V] Zoltán Farkas, Dorottya Kalapis, Zoltán Bódi, Béla Szamecz, Andreea Daraba, Karola Almási, Károly Kovács, Gabor Boross, Ferenc Pál, Peter Horvath, **Tamas Balassa**, Csaba Molnár, Aladar Pettko-Szandtner, Éva Klement, Edit Rutkai, Attila Szvetnik, Balazs Papp, Csaba Pál (2018). Hsp70-associated chaperones have a critical role in buffering protein production costs. *Elife*, *7*, e29845.

[VI] Abel Szkalisity, Filippo Piccinini, Attila Beleon, **Tamas Balassa**, Istvan Gergely Varga, Ede Migh, Csaba Molnar, Lassi Paavolainen, Sanna Timonen, Indranil Banerjee, Elina Ikonen, Yohei Yamauchi, Istvan Ando, Jaakko Peltonen, Vilja Pietiäinen, Viktor Honti, Peter Horvath (2021). Regression plane concept for analysing continuous cellular processes with machine learning. *Nature communications*, *12*(1), 1-9.

[VII] Reka Hollandi, Abel Szkalisity, Timea Toth, Ervin Tasnadi, Csaba Molnar, Botond Mathe, Istvan Grexa, Jozsef Molnar, Arpad Balind, Mate Gorbe, Maria Kovacs, Ede Migh, Allen Goodman, **Tamas Balassa**, Krisztian Koos, Wenyu Wang, Juan Carlos Caicedo, Norbert Bara, Ferenc Kovacs, Lassi Paavolainen, Tivadar Danka, Andras Kriston, Anne Elizabeth Carpenter, Kevin Smith, Peter Horvath (2020). nucleAIzer: a parameter-free deep learning framework for nucleus segmentation using image style transfer. *Cell Systems*, *10*(5), 453-458.

[VIII] Krisztian Koos, Gáspár Oláh, **Tamas Balassa**, Norbert Mihut, Márton Rózsa, Attila Ozsvár, Ervin Tasnadi, Pál Barzó, Nóra Faragó, László Puskás, Gábor Molnár, József Molnár, Gábor Tamás, Peter Horvath (2021). Automatic deep learning-driven label-free image-guided patch clamp system. *Nature communications*, *12*(1), 1-11.

[IX] Ilida Suleymanova*, **Tamas Balassa***, Sushil Tripathi, Csaba Molnar, Mart Saarma, Yulia Sidorova, Peter Horvath (2018). A deep convolutional neural network approach for astrocyte detection. *Scientific reports*, *8*(1), 1-7.

[X] Brigitta Dukay, Fruzsina R Walter, Judit P Vigh, Beáta Barabási, Petra Hajdu, **Tamás Balassa**, Ede Migh, András Kincses, Zsófia Hoyk, Titanilla Szögi, Emőke Borbély, Bálint Csoboz, Péter Horváth, Lívia Fülöp, Botond Penke, László Vígh, Mária A Deli, Miklós Sántha, Melinda E Tóth (2021). Neuroinflammatory processes are augmented in mice overexpressing human heat-shock protein B1 following ethanol-induced brain injury. *Journal of neuroinflammation*, *18*(1), 1-24.

[XI] Filippo Piccinini, **Tamas Balassa**, Antonella Carbonaro, Akos Diosdi, Timea Toth, Nikita Moshkov, Ervin A Tasnadi, Peter Horvath (2020). Software tools for 3D nuclei segmentation and quantitative analysis in multicellular aggregates. *Computational and structural biotechnology journal*.

* These authors contributed equally to the paper.