

# **AI-based fluorescent labeling for cell line development**

**Hanna Pankova**

Master thesis

Date of issue:	01. April 2022
Date of submission:	29. August 2022
Reviewers:	Prof. Dr. Markus Kollmann Dr. Wolfgang Halter



## **Erklärung**

Hiermit versichere ich, dass ich diese Master thesis selbstständig verfasst habe. Ich habe dazu keine anderen als die angegebenen Quellen und Hilfsmittel verwendet.

Düsseldorf, den 29. August 2022

---

Hanna Pankova



## **Abstract**

Cell line development is an expensive and time-consuming process, however that is the most modern approach for producing the proteins needed in pharmaceuticals.



## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Notation . . . . .	1
<b>2</b>	<b>Background</b>	<b>1</b>
2.1	Biology . . . . .	1
2.2	ML . . . . .	2
<b>3</b>	<b>Prediction of cell organelles</b>	<b>4</b>
3.1	Overfitting . . . . .	4
<b>4</b>	<b>Summary</b>	<b>5</b>
	<b>References</b>	<b>6</b>
	<b>List of Figures</b>	<b>7</b>
	<b>List of Tables</b>	<b>7</b>





# 1 Introduction

## 1.1 Motivation

## 1.2 Notation

# 2 Background

## 2.1 Biology

Cell line development is a process of generating single cell-derived clones that produce high and consistent levels of target therapeutic protein. ([pharma.lonza.com/offerings/mammalian/cell-line-development](https://pharma.lonza.com/offerings/mammalian/cell-line-development))

Differential interference contrast (DIC) microscopy is an optical microscopy technique used to enhance the contrast in unstained, transparent samples ([wikipedia](https://en.wikipedia.org/wiki/Differential_interference_contrast_microscopy)).

Proteins are large biomolecules and macromolecules that comprise one or more long chains of amino acid residues (<https://en.wikipedia.org/wiki/Protein>).

Fluorescent labelling is the process of covalently binding fluorescent dyes to biomolecules such as nucleic acids or proteins so that they can be visualized by fluorescence imaging (<https://www.nature.com/subjects/fluorescent-labelling>). A fluorophore is a chemical compound that can reemit light at a certain wavelength. These compounds are a critical tool in biology because they allow experimentalists to image particular components of a given cell in detail. (O'reilly life sciences p113)

Cell line development (CLD) is the process by which the cellular machinery is co-opted to manufacture therapeutic biologics or other proteins of interest. One can use different expression systems for cell line development: bacterial, plant-based, yeast, mammalian. (copy paste from <https://www.beckman.de/resources/product-applications/lead-optimization/cell-line-development>) Chinese hamster ovary (CHO) cells are the most popular mammalian cells used for protein production. (doi:10.1016/B978-0-08-100623-8.00007-4)

First step of CLD is the introduction of the gene of interest (GOI or a DNA vector) to CHO cells. This process is called a transfection. It is important to transfect a GOI into an optimal site of genome to secure a high protein expression over time during protein production, however practically transfection mostly results in a vector being inserted into a random site within a host cell genome. In case the gene was transfected in the inactive site of genome (and the majority of genome is not transcriptionally active) the cell will likely not express the gene. (doi:10.1016/B978-0-08-100623-8.00007-4) (doi:10.1016/j.coche.2018.08.002)

The second step is the selection of cell pools that have successful and stable gene integrations. The reason why not all of them are suitable for cloning is the following: during the transfection only 80% of cells receive the vector of GOI (doi:10.1016/B978-0-08-100623-8.00007-4), only the small percent of which actually integrate a vector into the genome and, as mentioned above, only a fraction of those cells are able to stably express

a protein. (Reference needed). Such selection could be done with bulk sorting algorithm. (doi:10.1016/B978-0-08-100623-8.00007-4)

The third step in CLD is to clone the cells. The chosen stable pools of cells are phenotypically and genetically diverse - meaning they have different growth rates, metabolic profile and etc. This is not ideal for industrial production - all the cells used for protein production should be derived from the same clone ([25] here doi:10.1016/B978-0-08-100623-8.00007-4). In order to choose single best cells for further cloning one assesses several parameters like cell size, granularity, cell surface protein expression and etc. This can be done with Fluorescent Activated Cell Sorting (FACS) technology. (<https://doi.org/10.1517/14712598.4.11.1821>). Unfortunately fluorescence labeling is expensive and may ruin the cell due to its phototoxicity (<https://doi.org/10.1371/journal.pone.0007497>). There is a limited number of available fluorescent channels in microscopes as well as such labels can also be inconsistent, depend a lot on reagent quality, and require many hours of lab work. Therefore there exists a need for fluorescent labeling in silico - without intervening into the cell.

Once the cells are cloned, phenotypical and genetical heterogeneity is reduced, the next step is to characterize clonally-derived cells based on the following criteria: cell size, growth rate, protein quality, titer, metabolites and etc. With this one can estimate clones productivity and titer. Such observations may take up to 90 days after which one can determine which cells are stable and therefore suitable for production. This is the last step of CLD process and consumes a lot of time and maintaining costs for feeding and cloning the cells. Predicting the stability of the cells directly from DIC images would reduce this time significantly allowing to escape this process completely.

However there are also some disadvantages of this approach. First, it can be less accurate than skilled cells staining performed manually. Extreme or unusual clones and phenotypes might be challenging if they were not used in the training set of images.

## 2.2 ML

### Background on Unet and ML in general

Convolutional neural network is a neural network that is based on convolutional layers. It is a powerful tool for image processing and is used in medical imaging. Convolution is a linear operation used in convolutional layers that can be performed by applying a kernel (a 2d matrix) across a bigger input matrix called tensor, which can be 3d. Element-wise product between them is calculated and summed, this value will be an element of the output 2d matrix. Kernel slides across all locations of the input tensor. In case if several different kernels were used then a 3d tensor will be created.

Main advantage of convolutional neural networks is weight sharing. Kernel has learnable weights however these weights are shared across all locations of the kernel on the input tensor, this strongly reduces the number of parameters needed.

CNNs also use non-linearities like RELU, ELU, Tanh, Sigmoids and etc. They are also often combined with max pooling layers and dropouts to escape overfitting.

Overfitting is one of the most often problems in deep learning that prevents model to

generalize well for unseen data. This can happen when the model is too big for the amount of training data given, it was not regularized well or there is just not enough data for training.

U-Net architecture is widely used for segmentation purposes. It is a convolutional neural network with the following architecture: [img] . It first performs image downsampling and upsampling afterwards. [https://arxiv.org/pdf/1505.04597.pdf] Conrad (1997) hat ein Buch geschrieben. Es gibt auch andere Arbeiten (Patterson and Hennessy, 2004) die referenziert sind. In Abbildung 1 ist ein Sachverhalt dargestellt.

1 Autor: Conrad (1997) (Conrad, 1997)

2 Autoren: Liebeck and Conrad (2015) (Liebeck and Conrad, 2015)

3 Autoren: Liebeck et al. (2016) (Liebeck et al., 2016)

Online resource: ILSVRC2016 (2017)



Figure 1: Gallien zur Zeit Caesars

Jahr	Erster Consul	Zweiter Consul
1	C. Caesar	L. Aemilius Paullus
2	P. Vinicius	P. Alfenus Varus
3	L. Aelius Lamia	M. Servilius
4	Sex. Aelius Catus	C. Sentius Saturninus
5	L. Valerius Messalla	Cn. Cornelius Cinna
suff.	C. Vibius Postumus	C. Ateius Capito
6	M. Aemilius Lepidus	L. Arruntius

Table 1: Römische Konsulen

### 3 Prediction of cell organelles

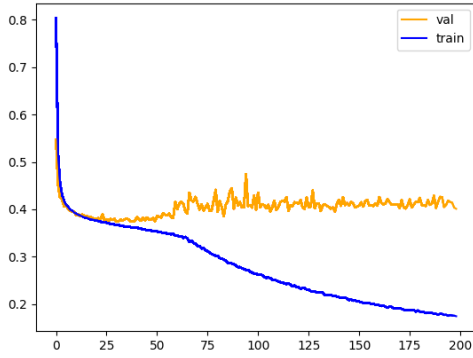


Figure 2: Not regularized

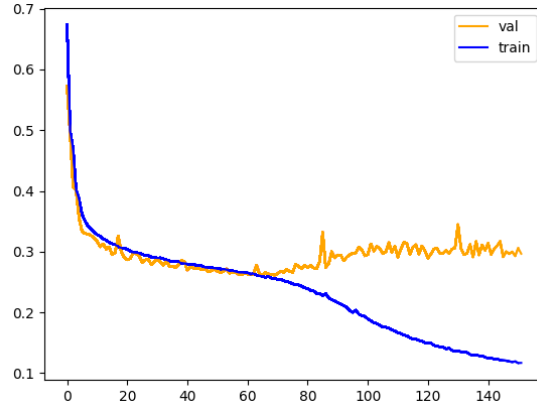


Figure 3: Regularized

#### 3.1 Overfitting

**Definition 3.1** (Overfitting). "Hypothesis overfits the training samples if some other hypothesis that fits the training samples less well actually performs better over the entire distribution of instances" (p67 Mitchell Machine Learning 1997).

Overfitting prevents model to generalize well on the unseen data and in order to avoid fitting to closely to the training dataset one has several options:

Regularize the model. This could be done either by reducing the number of the parameters, therefore by changing the model architecture. Or by adding regularization layers like BatchNorm or Dropout. Regularization can also be done by putting more restrictions on weights by adding their  $L1$  or  $L2$  norms to the loss function.

From the first training one can clearly see an overfit happening around epoch 30. Although one could use early-stopping and chose any epoch before overfit has happened, the better solution would be to regularize a network better. Therefore the *BatchNormalization* layers have been added after the first Convolution layer in each ConvBlock and *TransposedConvBlock*. The results of training the regularized network is presented in Figure 3.

## 4 Summary

## References

- Stefan Conrad (1997). *Föderierte Datenbanksysteme: Konzepte der Datenintegration*. Springer Verlag.
- ILSVRC2016 (2017). *ILSVRC 2016*. URL: <http://image-net.org> (visited on 07/17/2017).
- Matthias Liebeck and Stefan Conrad (July 2015). “IWNLP: Inverse Wiktionary for Natural Language Processing”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, pp. 414–418.
- Matthias Liebeck, Katharina Esau, and Stefan Conrad (Aug. 2016). “What to Do with an Airport? Mining Arguments in the German Online Participation Project Tempelhofer Feld”. In: *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*. Association for Computational Linguistics, pp. 144–153.
- David A. Patterson and John L. Hennessy (2004). *Computer Organization and Design*. Morgan Kaufmann Publishers.

## **List of Figures**

1	Gallien zur Zeit Caesars . . . . .	3
2	Not regularized . . . . .	4
3	Regularized . . . . .	4

## **List of Tables**

1	Römische Konsulen . . . . .	3
---	-----------------------------	---