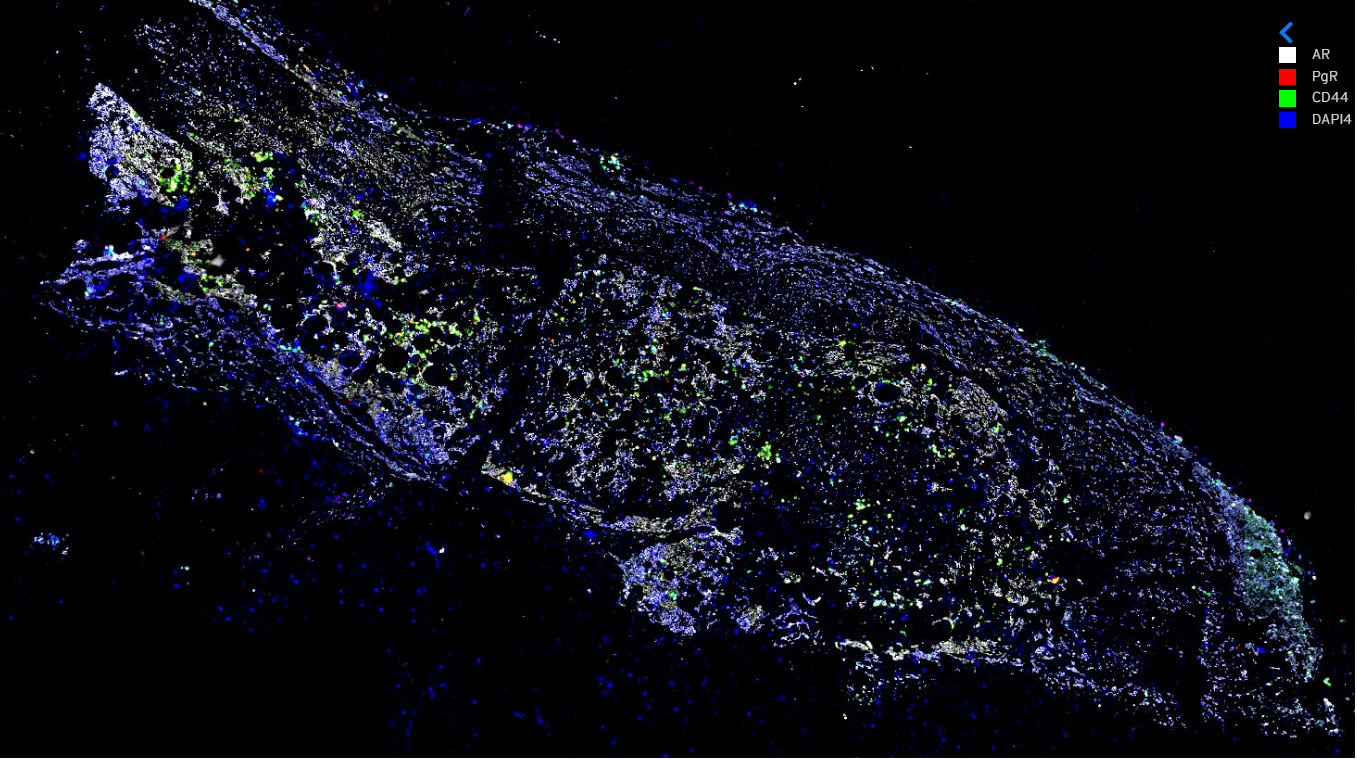
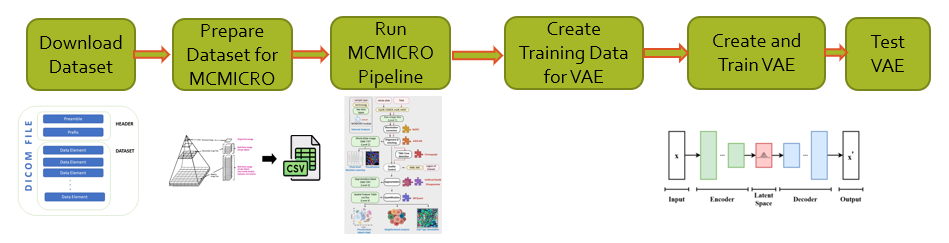
Automated Single-Cell Quantification of Breast Cancer Marker Evolution from Serial Immunofluorescence Assays



Automated Single-Cell Quantification of Breast Cancer Marker Evolution from Serial Immunofluorescence Assays



A Project by Rikhil Rao

Canyon Crest Academy

San Diego

Contents

[List of Figures 4](#_Toc129633314)

[Student Checklist/Timeline 5](#_Toc129633315)

[Abstract 6](#_Toc129633316)

[Acknowledgements 7](#_Toc129633317)

[Introduction 8](#_Toc129633318)

[Review of Literature 9](#_Toc129633319)

[Problem Statement 13](#_Toc129633320)

[Hypothesis 14](#_Toc129633321)

[Materials and Design 15](#_Toc129633322)

[Procedure 16](#_Toc129633323)

[Results 18](#_Toc129633324)

[Model Accuracy 19](#_Toc129633325)

[Model Prediction 22](#_Toc129633326)

[Summary & Conclusions 23](#_Toc129633327)

[Recommendations 24](#_Toc129633328)

[Bibliography 25](#_Toc129633329)

[Appendix A: Code 26](#_Toc129633330)

# List of Figures

[Figure 1: MCMicro Modules 10](#_Toc129633302)

[Figure 2: Procedure 16](#_Toc129633303)

[Figure 3: Slide Images used in the Study 18](#_Toc129633304)

[Figure 4: Model Loss 19](#_Toc129633305)

[Figure 5: Image Quality 19](#_Toc129633306)

[Figure 6: Mcmicro Output 20](#_Toc129633307)

[Figure 7: Quantification-1 21](#_Toc129633308)

[Figure 8: Quantification-2 21](#_Toc129633309)

[Figure 9: Model Prediction 22](#_Toc129633310)

# Student Checklist/Timeline

* 10/25/2022 Topic/Idea
* 02/01/2023 Abstract
* 02/01/2023 Acknowledgements.
* 02/20/2023 Introduction
* 03/03/2023 Review of the Literature
* 02/21/2023 Statement of the Problem or Purpose.
* 02/01/2023 Hypothesis
* 02/05/2023 Materials
* 02/05/2023 Procedures
* 02/27/2023 Results
* 02/28/2023 Data, Table, Graphs, Charts
* 02/28/2023 Data Analysis
* 02/28/2023 Conclusions
* 02/28/2023 Recommendations
* 03/10/2023 Bibliography
* 03/10/2023 Appendices
* 03/11/2023 Completed Project Notebook and Display are ready for submissions.

# Abstract

Despite advances in breast cancer diagnoses and treatment, it continues to be the second leading cause of cancer death in women. Recent interest in managing and prognosticating breast cancer has been centered around examining single-cell and spatial expression patterns on molecular imaging, such as immunofluorescence and immunohistochemistry. However, current techniques for this single-cell quantification are disjointed and steps are not formatted to be compatible.

The purpose of this project is to develop a simple, easy to use, end-to-end pipeline to quantify how much a patient’s cancer changes at different timepoints in their treatment. To accomplish this, serial cyclic immunofluorescence images acquired over the course of a patient’s breast cancer treatment were gathered from the Human Tumor Atlas Network (HTAN). Two images of different samples were chosen from time point 0 and one image from time point 1 was chosen. Each image consisted of 25 unique channels, including DAPI, HER2, and ER. These images were used to train a patient-specific Variational Auto-Encoder (VAE), which was then used to quantify how much the cancer changed. In order to train the model, individual cells were segmented using an open-source tool called MCMICRO, resulting in around 14,000 cells in each image.

By running these analyses, the tumor phenotype of the patient (e.g., HER2 negative, estrogen receptor [ER] positive) was able to be automatically determined. In addition to this the MCMICRO quantification outputs show that the marker expression changed at a later time point as the patient underwent hormone therapy for their breast cancer. For example, the ER expression decreased 2-fold and the Ki67 expression increased 12-fold. Additionally, the VAE latent space plots show significantly different distributions at different time points, implying that the network was able to encode the baseline single-cell expression patterns.

This research shows that automated single-cell quantification of marker expression in breast cancer can be important to track tumor evolution over time, which can be quantified and visualized using a VAE.

# Acknowledgements

I would like to thank our computer science teacher and guide Mr. Kevin Hare for the support and encouragement to complete this project. I would like to acknowledge the guidance and help received from Mr. Gunvant Chaudhari, who taught me cancer biology and machine learning concepts that helped me to implement this project.

Finally, I would like to thank my parents for being a source of information and providing the components, supplies needed for this project and to proofread the draft write-ups and material to evolve to the current version.

# Introduction

Cancer has become one of the main threats to human health and life. Among all types of cancer, breast cancer is very common among women and diverse in character. About 70 percent of all breast cancers in women have overexpression of receptors (proteins) that attach to estrogen or progesterone. Breast cancers without ER (Estrogen Receptor) and PR (Progesterone Receptor) are known as HR (Hormone Receptor)-negative. Estrogen receptor and progesterone receptor tests are used to help guide breast cancer treatment. Cancer cells and cells within the tumor microenvironment together determine disease progression and response to treatment.

Immunofluorescence is a powerful approach for getting more information about cellular structures. Expressions and locations of a multitude of specific proteins can be assessed, making this process best suited for scientists to solve many cells biological questions. The extent of cellular heterogeneity in breast cancer has been hypothesized to impact diagnosis, response to therapy, and long-term outcome.

Determining tumor heterogeneity and its impact on drug response is essential in the development of personalized therapies. Scientists have recognized that images of cells contain more information than what can be extracted by the human eye. Computer aided image analysis has the potential to make complex information more accessible in diagnostic practice. Artificial Intelligence based computational pathology has shown increased diagnostic accuracy and reduced turnaround times.

Review of Literature

Breast cancer is one of the most prevalent cancers globally and is the second leading cause of cancer death among women. According to the World Health Organization, breast cancer accounts for 16% of all female cancer deaths, with an estimated 2.3 million new cases and 685,000 deaths reported in 2020. Early diagnosis and treatment of breast cancer is crucial for successful outcomes, and medical professionals have been using various methods to improve diagnostic accuracy, including mammography, ultrasound, and magnetic resonance imaging (MRI). Recently, there has been growing interest in using machine learning algorithms to assist in breast cancer diagnosis and treatment by examining single cells in biopsies.

Single cell analysis refers to the process of examining individual cells at the molecular level. This technique has many advantages over traditional bulk analysis, where entire populations of cells are analyzed together. One of the key advantages of single cell analysis is that it allows for the detection of rare cell types or subpopulations that may be missed in bulk analysis. Single cell analysis also provides a more accurate representation of the heterogeneity within a population of cells, as it reveals differences in gene expression, epigenetic modifications, and protein levels among individual cells. This level of detail can lead to a better understanding of cellular processes and disease mechanisms, as well as more precise diagnostic and therapeutic approaches. Furthermore, single cell analysis is becoming increasingly accessible and cost-effective, making it a valuable tool for a wide range of research and clinical applications.

Biomarkers are genes, proteins, or other substances that can be tested for to reveal important details about a person’s cancer. In the research paper by Alison Min-Yan Cheung, et al. [ [3](https://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-021-01475-y)], the author explains how cellular heterogeneity in breast cancer could have potential impact on diagnosis and long-term outcome. In this paper, authors assess heterogeneity within each clinical subtype of breast cancer in the context of protein marker expression signature of single cells. Image registration and single cell segmentation were performed with software package by GER for MxIF platform. This research demonstrates MxIF intensities of biomarkers in single cell showed distributions comparable to IHC (Immunohistochemistry) scoring of the tissue. Author concludes that the biomarker expression signatures evaluated on single cells, revealed heterogeneous composition and spatial arrangement of subgroups. Cancers with the same IHC score of overexpression of P53 and/or P16 exhibited a range of expression levels when measured in individual cells. This information will be helpful in characterization of breast cancers and predicting their response to therapy.

In my project, single cell analysis was accomplished by using MCMICRO, a tool that allows for segmentation and quantification of single cells from a whole slide image.

MCMICRO provides customizable software that processes whole slide microscopy data into cohesive images that can be easily visualized and quantified as single cell data [ [13](https://mcmicro.org/overview/) ]

MCMICRO modules: [ Below image is from MCMICRO [ [13](https://mcmicro.org/overview/)] ]

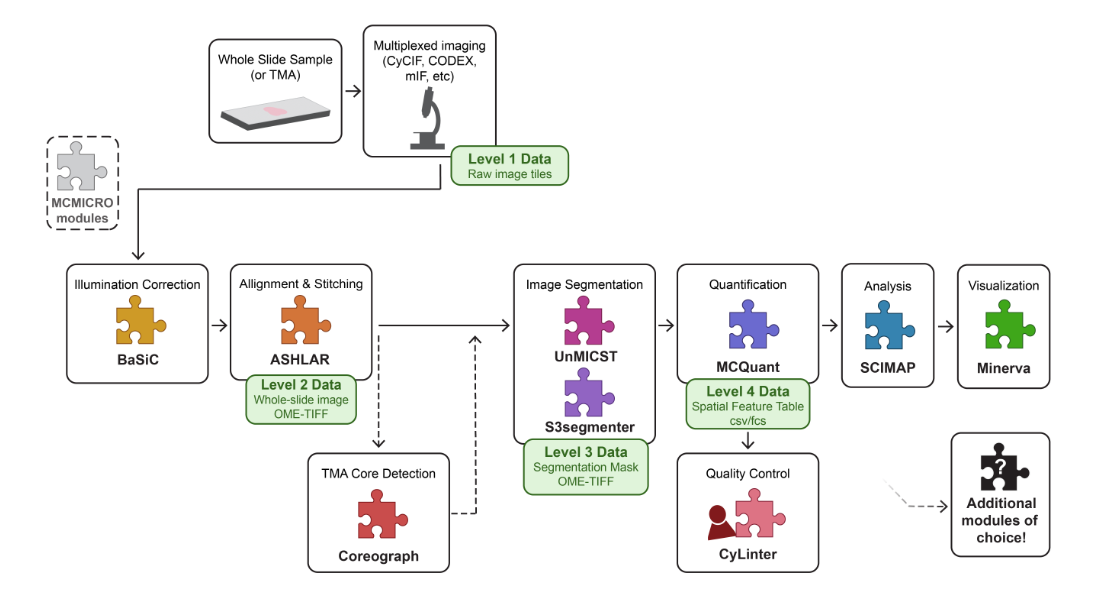


Figure : MCMicro Modules

The specific machine learning algorithm that was employed in my project was a VAE. VAEs (Variational Auto-Encoders) are a type of generative model that can learn a compressed and meaningful representation of high-dimensional data, such as images or text. They work by encoding data into a lower-dimensional latent space and then decoding it back to its original form. Unlike traditional autoencoders, VAEs use probabilistic techniques to generate new samples from the learned latent space. VAEs are important because they can be used for a variety of tasks such as image generation, data compression, and feature extraction. They also provide a way to explore and manipulate the latent space, which in this project was useful as it allowed us to quantify the difference between two time points. The latent space in the VAE is a probabilistic function that is used to generate a gaussian (normal) distribution. By training on data, VAEs modify the mean and standard deviation of the distributions in the latent space to better be representative of all the data it trained on. In the field of biomedical image analysis, VAEs have shown promising results in analyzing biopsies. By training the VAE on a dataset of biopsy images, it can learn to extract meaningful features and patterns from the images and store them in the latent space. These latent spaces can then be used to compare the distributions of the patient’s biopsy at different timepoints and draw conclusions about the progression of the cancer. VAEs have the potential to enhance the accuracy and efficiency of analyzing biopsies, leading to improved diagnostic outcomes. Additionally, VAEs are more robust to missing or incomplete data than other generative models, making them a powerful tool in unsupervised learning.

Immunofluorescence is a powerful approach for getting insight into cellular structures and processes using microscopy. Specific proteins can be assessed for their expression and location, making immunofluorescence indispensable for scientists to solve many cell biological questions. An immunofluorescence experiment is based on the following principal steps: Specific antibodies bind to the protein of interest and fluorescent dyes are coupled to these immune complexes in order to visualize the protein of interest using microscopy. The result of this is antigens binding to fluorophores which allows for the easily visualization of the proteins. An example of some proteins is DAPI, which is used to gauge the number of cells in the slide and Ki67, which is a prominent proliferation marker. By analyzing the expression of these cells at different timepoints and feeding these cells into the VAE allowed the visualization and quantification of the cancer at different timepoints.

The combination of VAEs, immunofluorescence, and single cell analysis has the potential to revolutionize the medical field. By integrating these technologies, researchers can analyze complex medical data at multiple levels, from the molecular to the cellular and beyond. This can lead to a more comprehensive understanding of disease processes, as well as the development of more effective treatments. For example, VAEs can be used to analyze large volumes of immunofluorescence data, identifying patterns that are difficult for humans to detect. Ultimately, this integrated approach has the potential to accelerate medical research and improve patient outcomes by enabling the development of more targeted and personalized treatments.

# Problem Statement

Breast cancer is the second leading cause of cancer death in women and is responsible for 40,000 deaths per year in the United States. Diagnosis and prognostication of breast cancer is done through evaluation of a tumor biopsy using immunofluorescence. Recent advances in analyzing multiplex immunofluorescence at the single cell level have shown cell-specific spatial differences of various tumor marker expressions within the same biopsy, which may be indicative of long-term outcomes. However, current methods of clinical analyses are not able to quickly quantify markers at the single-cell level. In this study, the MCMICRO tool is applied to breast cancer molecular assays in order to automatically quantify the tumor phenotype and quantify single-cell changes in expression over time using a Variational Autoencoder (VAE).

# Hypothesis

A combination of Immunofluorescence, Machine learning model by using the Variable Auto Encoder (VAE), and single cell analysis techniques is very powerful in learning about the progression/remission of cancerous tumors in breast cancer patients. I hypothesize that an automated cell segmentation and quantification algorithm can accurately quantify breast cancer marker phenotypes in single cells and can be applied to track cell expression changes over time. The results that I have generated are promising and allow a doctor to accurately review the state of cancer in patients.

# Materials and Design

This project is implemented in Python. The following software packages and tools were used in the implementation of this project.

**Software:**

* **Visual Studio** – IDE and code editor for software developers
* **Google Collab** – Allows anybody to write and execute arbitrary python code through the browser.
* **Python 3.9** – A high level, general-purpose programming language
* **Keras** – High level API of TensorFlow
* **MCMICRO** – Open-source software that processes whole slide microscopy data into cohesive images that can easily be visualized and quantified as single cell data.
* **Google Cloud CLI** – Set of tools to create and manage Google Cloud resources.
* **Breast Cancer database** – the Human Tumor Atlas Network (HTAN) database from which imaging data was downloaded.

# Procedure

To create and train the AI model, the following steps were performed.

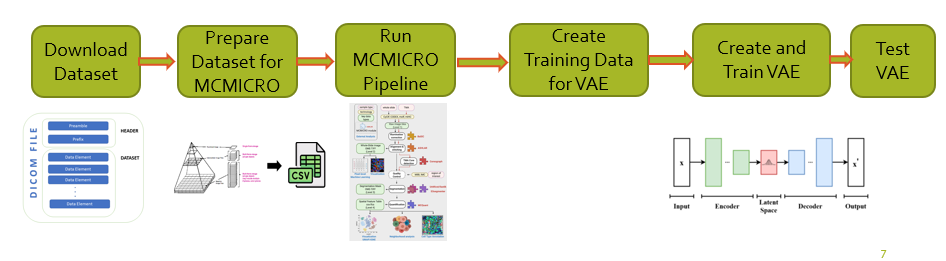


Figure : Procedure

**Downloading the Dataset:**

Data was downloaded from the Human Tumor Atlas Network (HTAN) website. The files were downloaded via the Google cloud command line interface (gCloud CLI). All the files were downloaded in DICOM format, the common file type for medical images. Each dicom file was a slide of the complete biopsy and represented a different marker.

**Prepare Dataset for MCMICRO:**

To run the data through the MCMICRO pipeline, a csv file that contains information on the marker associated with each channel in the DICOM image was made. Each dicom file that was downloaded was associated with a singular marker, so the CSV file was needed to compile all the different markers that were present in the biopsy. All the DICOM files that were downloaded were parsed and used to generate a csv file.

**Run MCMICRO Pipeline on Dataset:**

Using nextcloud, the MCMICRO pipeline was run on the dataset with the generated csv file.

**Create training data for VAE:**

To create training data for the VAE, the segmentation of the dataset that was performed by the MCMICRO pipeline was used to generate small images (24 by 24 pixels) for each of the cells that were found. The image size is 24 x 24 because it was the maximum size of any given cell that was segmented so it would for sure capture each of the cells that were segmented. The data was then saved into a numpy array.

**Create and Train VAE:**

The Variation Auto Encoder was created using Keras, a neural network library. The model that was created has 4 2D convolutional layers, 2 dense layers, 1 flatten layer, and 1 2D Convolutional up sampling layer. To train the model, saved data was loaded in the form of a numpy array and then partitioned it so that 75% of the data stored in the file was used for training while the other 25% is used for testing.

**Test VAE:**

The VAE used data generated from two different timepoints of the biopsy so that the latent vectors of each run could be tested to quantify how much the cancer changed. The KL divergence algorithm was used to quantify the difference in the probability distributions.

# Results

The result is a pipeline that takes in a set of molecular imaging files (often 10-100 GB in size) and outputs the single cell quantification for each channel of the imaging. In the training dataset, the pipeline took an input of 30 channels (DAPI, HER, ER, PCNA, Ki67, etc.) and accurately segmented ~ 16,000 cells from which cell-specific marker expression quantifications were obtained. Execution of the pipeline on the testing datasets at the same and later time points resulted in ~21,000 and ~4000 identified and quantified cells, respectively.

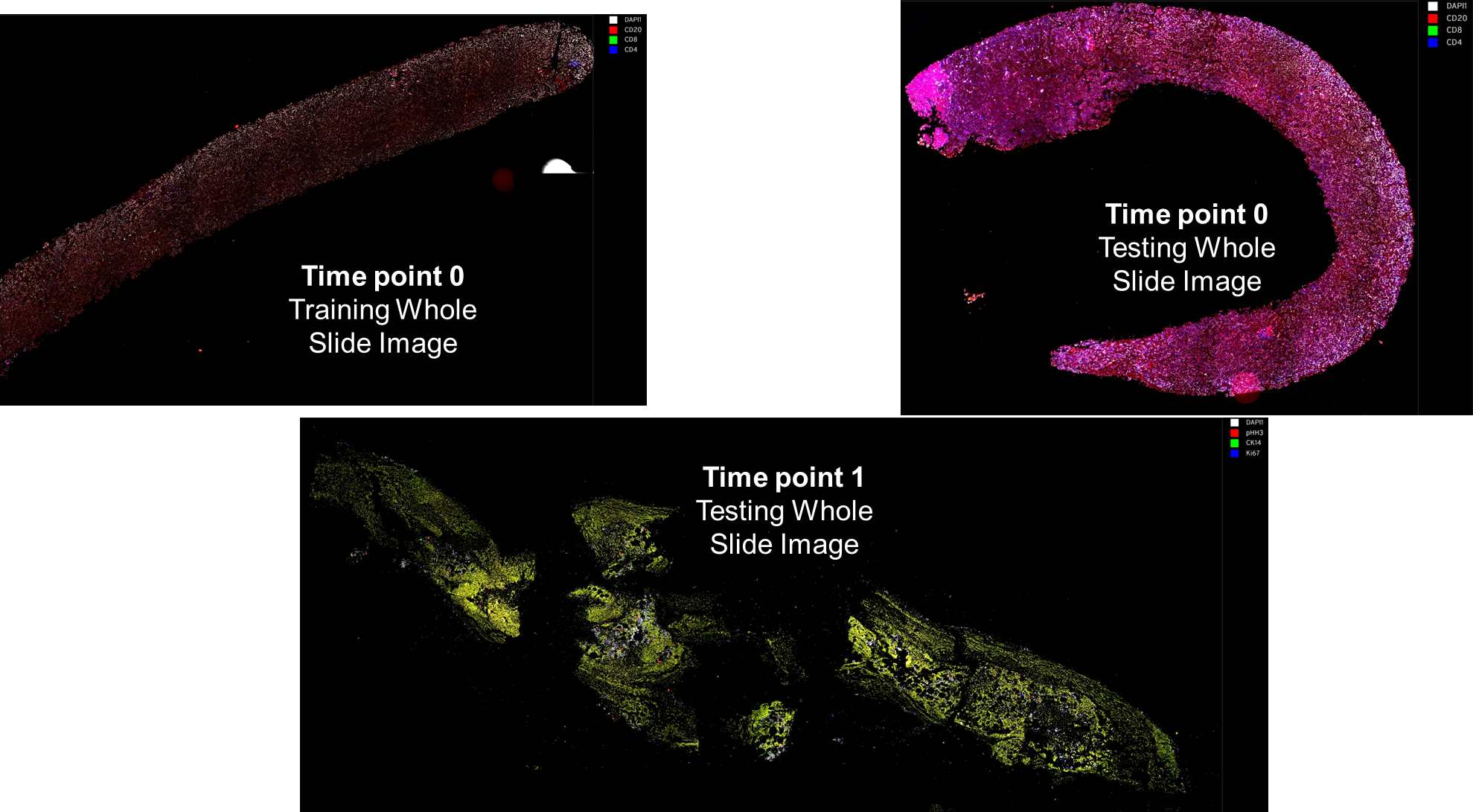


Figure : Slide Images used in the Study

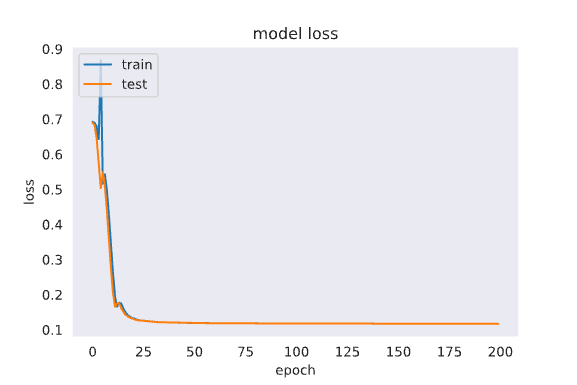
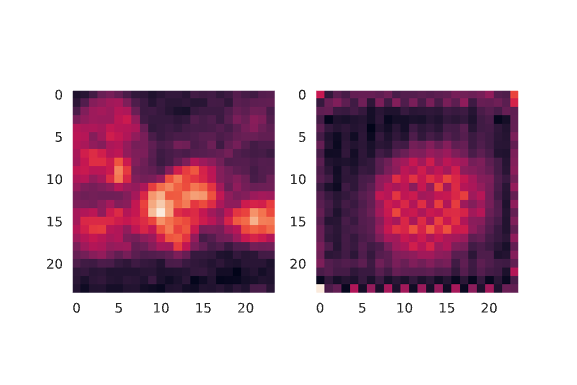
Model Accuracy

Figure : Model Loss

The above image shows the model loss of our trained VAE. Model loss = latent loss + reconstruction loss. Reconstruction loss is the loss that is experienced when rebuilding the image and latent loss is the loss that is experienced by the latent space.



Rebuilt image by VAE

Image fed into VAE

Figure : Image Quality

The quality of the rebuilt image in Figure 5 shows the disadvantage to only using 2 latent vectors to represent the entire 24x24 image.

**MCMICRO output**

Passing the image on the left through the MCMICRO pipeline generates the images on the right. The images on the right are the masks of the cells that were in the region of interest.

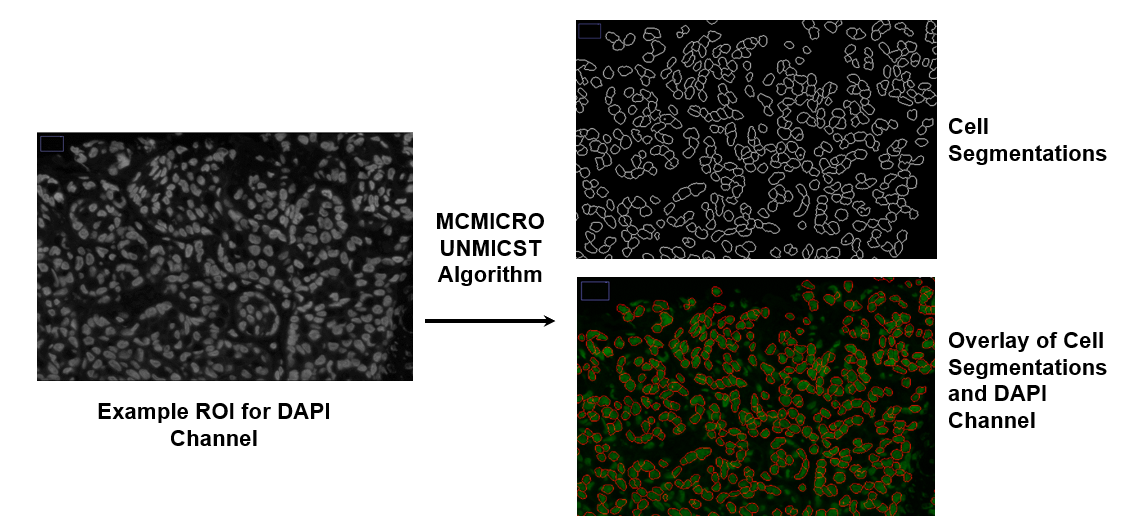


Figure : Mcmicro Output

**Quantification:**

Looking at the quantification of the markers at different timepoints allows us to make claims about the cancer progression. For example, the increase in the expression of Ki67 indicates to us that the cancer is proliferating at a much higher rate.

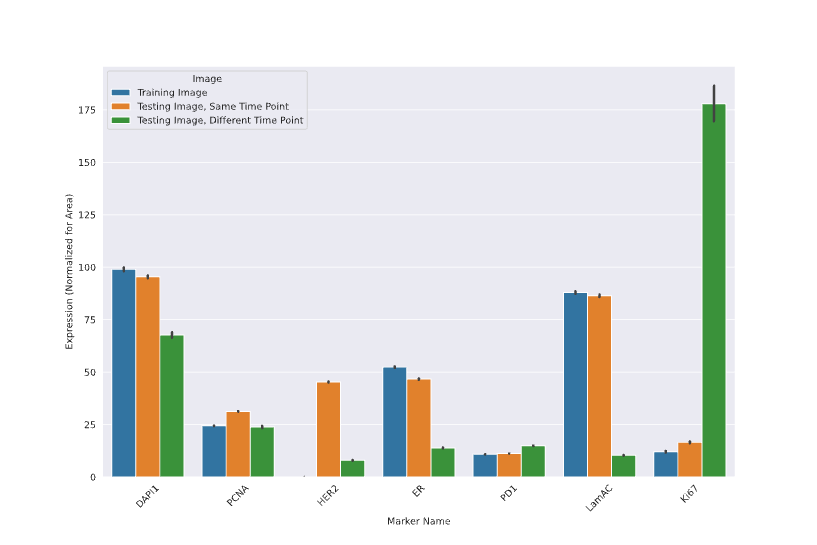
****

Figure : Quantification-1

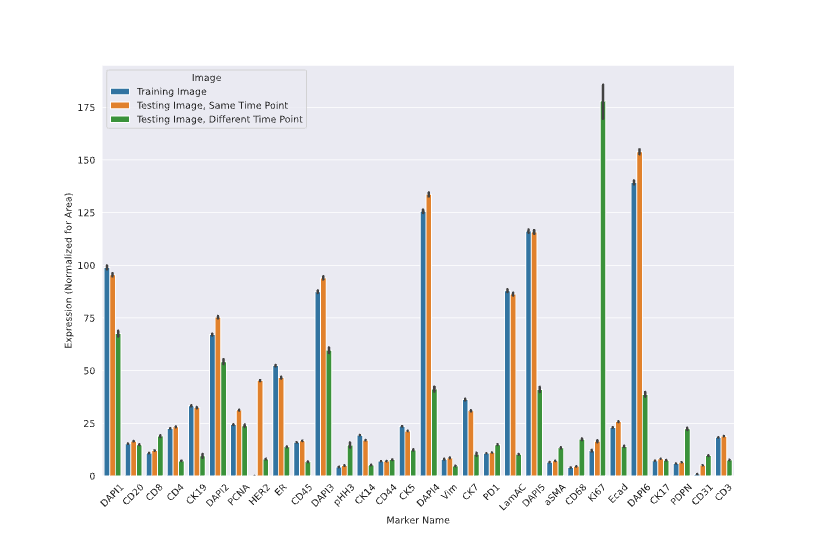


Figure : Quantification-2

## Model Prediction

The larger KL Divergences for the comparisons *between* timepoints than those *within* a timepoint suggests that the VAE is learning the patient’s baseline tumor phenotype.

A picture containing graphical user interface

Description automatically generated

Figure : Model Prediction

# Summary & Conclusions

This project applied a Deep Learning-based segmentation module to quantify tumor marker expression in breast cancer molecular imaging over time as a patient underwent breast cancer treatment. A VAE trained on baseline data showed that it was able to track and quantify the differences in single-cell expression over time. Future research on this topic should focus on validating this pipeline in multiple other subjects who have serial biopsies and better understanding the implications in tumor progression from the latent space changes.

# Recommendations

Future work of the project will be focused on better examining the latent space of the VAE. Many of the shortcomings of my project have to do with sacrificing the quality of the results for interpretability of the results. For example, adding another dimension to the latent vector would have increased the accuracy of the regenerated image but would make it much harder to quantify the difference between two timepoints as the data would be 3 Dimensional. Along with that, my project is only able to quantify how much the cancer changed and not if it changed for worse or for better. With more analysis of the latent space combined with other analyses performed on the biopsy, it would be possible to better understand how much the cancer changed and why the cancer changed.

# Bibliography

1. Characterizing advanced breast cancer heterogeneity and treatment resistance through serial biopsies and comprehensive analytics <https://www.nature.com/articles/s41698-021-00165-4>
2. Impact of mutational profiles on response of primary oestrogen receptor-positive breast cancers to oestregen deprivation: <https://www.nature.com/articles/ncomms13294>
3. Quantitative single-cell analysis of immunofluorescence protein multiplex images illustrates biomarker spatial heterogeneity within breast cancer subtypes: <https://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-021-01475-y>
4. Biological and Medical Importance of Cellular Heterogeneity Deciphered by Single-Cell RNA Sequencing – Rishikesh Kumar Gupta and Jacek Kuznicki <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7463515/>
5. Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer - <https://www.nature.com/articles/s41467-021-22801-0>
6. Cancer Ecosystems - <https://www.frontiersin.org/articles/10.3389/fonc.2019.00718/full>
7. Breast Cancer Hormone Receptor Status - [https://www.cancer.org/cancer/breast- cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-hormone-receptor-status.html](https://www.cancer.org/cancer/breast-%20cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-hormone-receptor-status.html)
8. Estrogen Receptor, Progesterone Receptor tests - [https://medlineplus.gov/lab-tests/estrogen- receptor-progesterone-receptor-tests/](https://medlineplus.gov/lab-tests/estrogen-%20receptor-progesterone-receptor-tests/)
9. Artificial Intelligence: How is It Changing Medical Sciences and Its Future? - <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7640807/>
10. Artificial Intelligence in Medicine: Applications, implications, and limitations - [https://sitn.hms.harvard.edu/flash/2019/artificial-intelligence-in-medicine-applications-implications- and-limitations/](https://sitn.hms.harvard.edu/flash/2019/artificial-intelligence-in-medicine-applications-implications-%20and-limitations/)
11. The Principle of Immunofluorescence Assays - [Immunofluorescence Assays | Principle | ibidi](https://ibidi.com/content/364-the-principle-of-immunofluorescence-assays)
12. Understanding Variational Autoencoders (VAEs) <https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73>
13. MCMICRO <https://mcmicro.org/overview/>
14. MCMICRO: a scalable, modular image-processing pipeline for multiplexed tissue imaging <https://www.nature.com/articles/s41592-021-01308-y>

Appendix A: Code