2020-09-21

AstraZeneca AI SWEDEN

Adipocyte Cell Imaging Challenge

Problem Formulation

This competition will help AstraZeneca to accelerate the drug development process. While we hope that you will enjoy the challenge, you should know that the solutions created will be used by AstraZeneca scientists to treat failing hearts, fibrotic lungs, diabetes, liver diseases, neurodegeneration, and cancer.

AstraZeneca AI SWEDEN

Novel therapeutics, background

Drugs bind to targets in the body. Historically, most drugs have been small molecules that depend on molecular interactions to recognize binding pockets, often on protein surfaces. The problem with this is that one drug might bind to several similar proteins, creating unwanted side effects. A more specific way to change protein function is therefore to target the sequence of the proteins. Protein sequences are coded in DNA (sequences of deoxyribonucleic acids) that are transcribed to messenger RNA (sequences of ribonucleic acid) and then finally translated to a protein (sequences of amino acids). AstraZeneca is now creating therapeutics that control protein function in a way that is not possible using traditional small molecule therapeutics, by targeting specific RNA sequences.

Nanomedicines

To develop this type of drug it is necessary to deliver sequences of therapeutic nucleotides that interact in different ways with cellular RNA. The delivery systems used to do this are called nanomedicines and, they are designed to efficiently transport RNA cargo through the body while protecting the fragile RNA from degradation. These nanomedicines are also small engineered machines made from biomolecules that interact with cell membranes and machinery in order to deliver RNA into cells where it can affect their function.

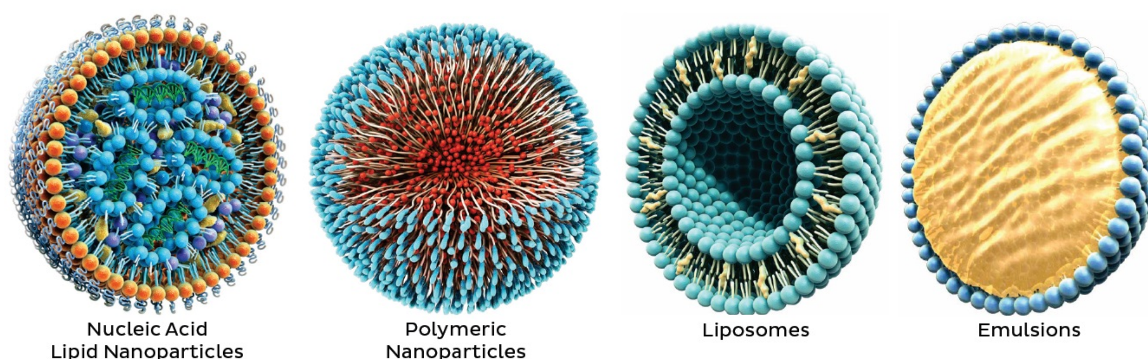

Figure 1 The figure above shows some representative nanoparticle types. These are very small and similar in size to viruses. These particles are used to deliver therapeutic agents (like mRNA) into cells. (Image from Drug Discovery World and Precision Nanosystems)

Dosing to the skin – Studying Adipocyte Cells

To make treatments with nanomedicines easier for patients we would like to be able to inject the nanomedicines into the skin (this is called a subcutaneous injection or s.c.). Following s.c. injections one of the most important cellular targets are the fat cells (adipocytes) that we all have in our skin. To avoid having to constantly inject human test subjects we can create adipocyte cell cultures, from stem cells. These allow us to test many nanomedicines cheaply and quickly without using animals or human volunteers. One of the most important methods to investigate the uptake of nanomedicines in cells (adipocytes in particular) is cell imaging which is the topic of this challenge.

The Adipocyte Dataset

The end goal of cell imaging is to extract relevant information about the cell structures that can guide pharmaceutical development. A high-level description of this process can be seen in Figure 1. The image outputs from the microscope are contained in a number of channels. To permit imaging of different cell structures, fluorescence microscopy is used to label specific parts of the cell. Each fluorescence channel contains images that have been labeled with different fluorescence staining techniques, with the resulting images containing different information about the cell. Note that the images are gray-scale 16-bit images with 65535 levels of gray. The "color" of these images is determined by the spectral filters in the microscope so that while they are gray-scale images, they represent different parts of the color spectrum.

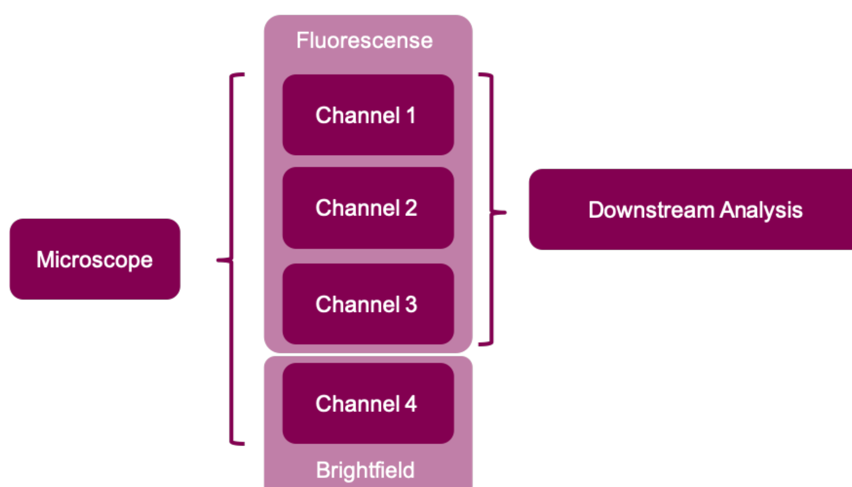

Figure 2. A high-level description of cell imaging and analysis. The output from the microscope is divided into different channels, each containing images with different information content. For this project, we consider channels containing brightfield and fluorescence images. Although the brightfield images contain information about e.g. lipid droplets, the downstream analysis is mainly fed with fluorescence images.

Along with the fluorescence channels the microscope can also capture transmission light microscopy images, in this case, brightfield. These images do provide information about cellular organization, but they lack the clear contrast of fluorescence labeling which limits their use in subsequent downstream analysis.

In contrast to transmission light microscopy which requires little preprocessing of the cell cultures, fluorescence imaging requires sample preparation to label different cell structures. In addition to being time-consuming, labor intensive, and expensive, the materials used to label the cells are often toxic, making the methods invasive. This hinders the collection of reliable longitudinal data which is often needed to study e.g. drug intake over time.

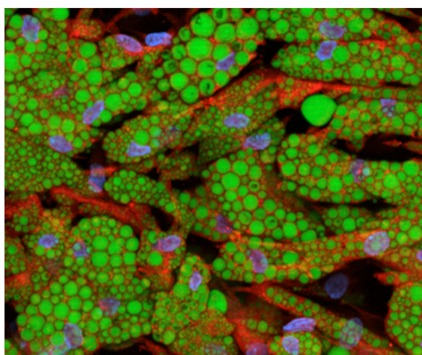

Figure 3. Superimposed fluorescence images of cell nuclei (blue), lipid droplets (green) and cytoplasm (red). Each color is a single channel image.

The dataset that forms the basis for this competition consists of pictures of stem-cell derived human adipocyte cell cultures (please see also the related presentation). The cells have been imaged, using a robotic confocal microscope, at three different magnifications, using both brightfield and fluorescence imaging. Images at each magnification are contained in separate datasets. The cells were grown in multi-well plates and each well is imaged at multiple, radially arranged, sites (also called fields of view, FOV). Each FOV is also imaged at multiple heights, creating optical slices (z-slices). This creates voxels consisting of pixels arranged in a horizontal x,y matrix (image), and a vertical direction (z, series of images). For the fluorescence (reference) images these are projected down to a single image using maximum intensity projections in the vertical direction (figure 3). Brightfield images (figure 4) are supplied as 7 individual z-slices. All cells are imaged using blue, green, and red fluorescence for nuclei, lipid droplets, and cell cytoplasm respectively. This is real data, so the nuclear channel also has quite a lot of background staining in the cell cytoplasm, which is normal and might be helpful. The red cytoplasmic stain was quite weak, so this channel contains rather noisy images as a result.

The lipid droplets make imaging challenging. They work like small lenses and can obscure or physically crowd other cellular components, creating atypical shapes. Unlike the nuclei, the lipid droplets are, however, clearly visible in both fluorescent (figure 3) and brightfield images (figure 4).

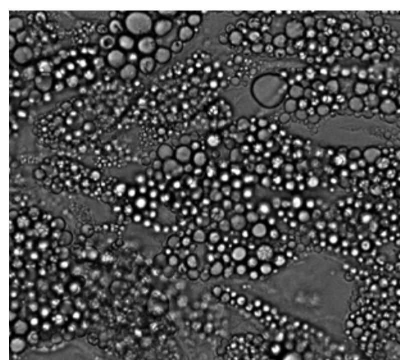

Figure 4. Bright field image of the same field of view as in Fig. 3.

There are also multiple cell types in every well, and about 15% of the cells do not contain lipid droplets. These are likely cells that have differentiated towards fibroblasts rather than adipocytes. The nuclei of these cells are larger and rounder.
During our analysis of the fluorescence images of these cell cultures we can extract a large number of measurements from each channel and for each cell: morphology (shape, size, compactness, axial parameters, radial symmetries, etc.), spatial relationships (distances between objects, clustering, etc.), intensity measurements, textural measurements (spots, ridges, valleys, holes, saddles, bright, dark) at various scales. The data from these measurements (hundreds or thousands of parameters) form patterns called phenotypes that allow us to distinguish effective nanomedicines from ineffective ones. Ideally, we would like to be able to extract this information using only the information contained in the brightfield images.

Example of bright field and fluorescence light images

A more detailed example of bright field and fluorescence images can be seen in Figure 5 where C01, C02, and C03 show falsely colored fluorescence images of cell nuclei, lipid droplets, and cytoplasm respectively. The corresponding bright field images are titled C04_Z0X, where X (ranging from 1-7) indicates different focal planes which can be used to extract three-dimensional information from the sample. Please see the associated Data Description presentation for additional information about the data.

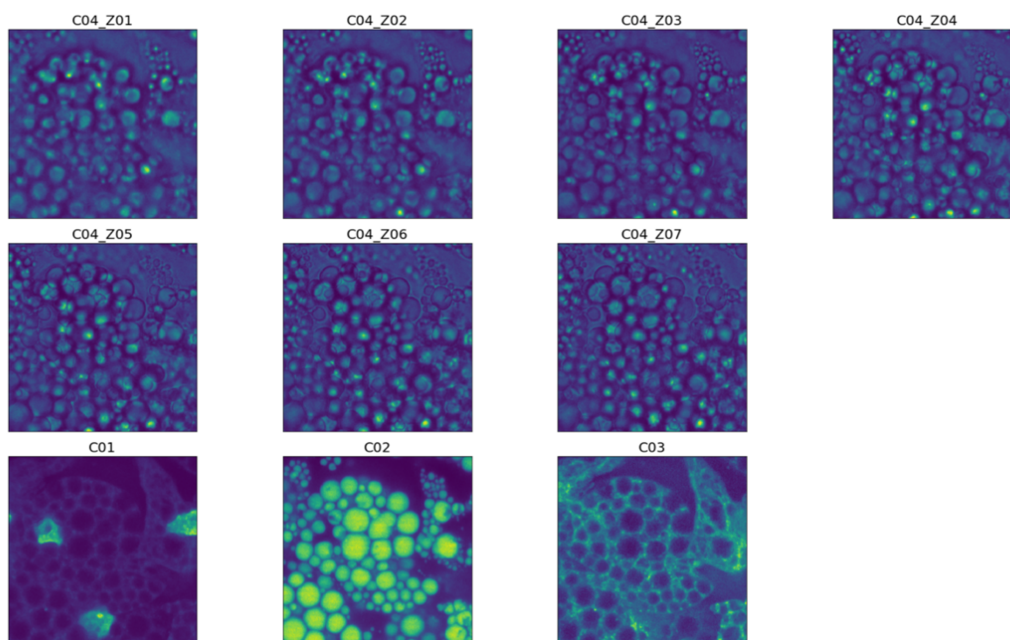

Figure 5. Example of image data produced by the microscope. The bright field images (C04_Z0x) are taken with different focal planes indicated by x. The fluorescence channels C01, C02, and C03 show cell nuclei, lipid droplets, and cell cytoplasm respectively. To obtain the information in C01-C03, the cells are tagged with fluorescent markers which target different cell structures. All images above show the same field of view.

Note that all images in Figure 5 show the same field of view, but the information content in each channel is very different.

Challenge Description

In this competition, we want to utilize machine learning to combine the advantages of bright field and fluorescence imaging while at the same time avoiding the toxic effects of cell labeling. We hypothesize that it is possible to predict the content of the fluorescence images from the corresponding bright field images using deep learning. The high-level architecture of such a solution can be seen in Figure 6, where the actual fluorescence channels have been replaced by the predictions of the machine learning model.

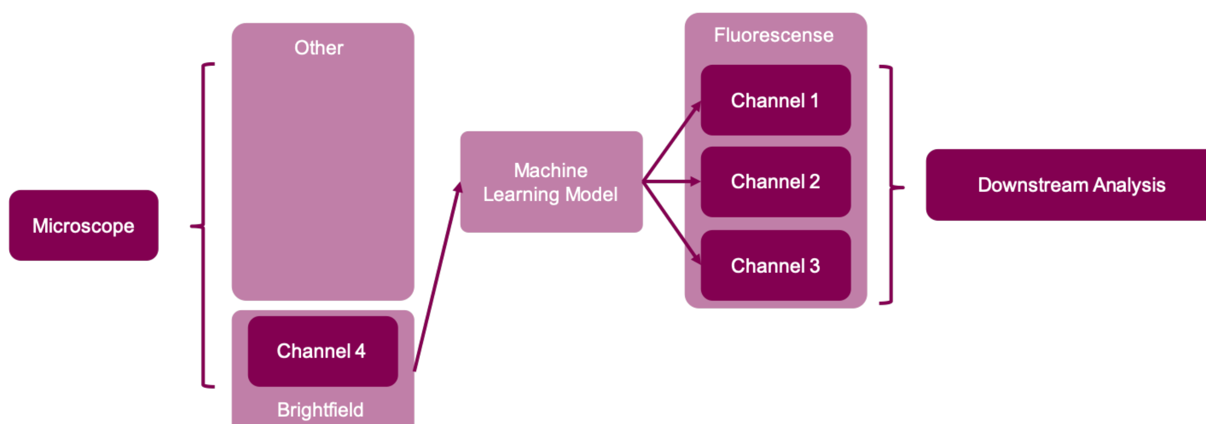

Figure 6. The machine learning augmented output from the microscope. The fluorescence images from the microscope are replaced by predicted outputs from one or several machine learning model(s). The machine learning model(s) take as input the bright field images from the microscope.

As alluded to in the previous section, information about the cell cultures, more specifically how they respond to drug exposure, is extracted from fluorescence images using image analysis. The quality of the predicted images in Figure 6 will be gauged by how useful they are for reliably extracting the outcome of a set of measurements in the downstream analysis pipeline as discussed above. In the hackathon, the downstream analysis pipeline will contain a set of predefined image measurements. An example is shown in Figure 7, where the measurements are related to features describing droplet morphology, intensity distribution, and droplet texture from a lipocyte image.

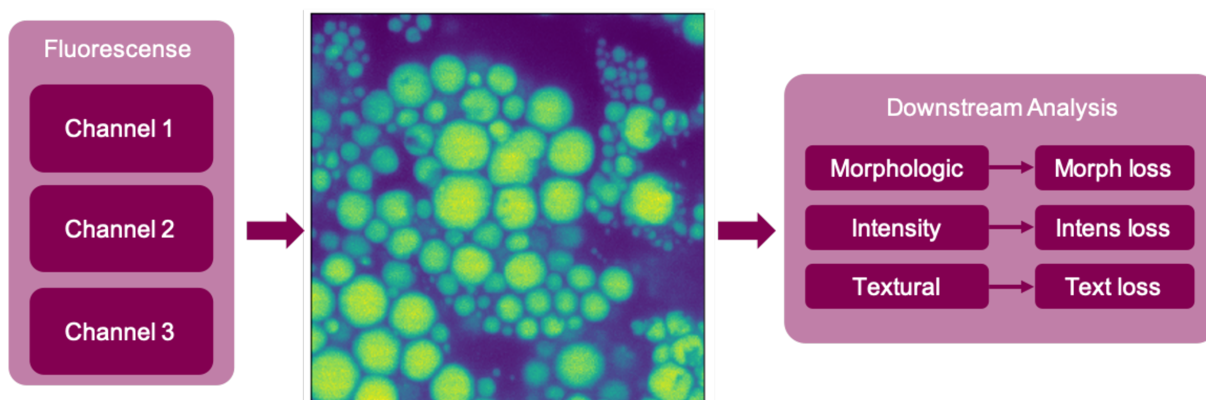

Figure 7 The fluorescence images are used to analyze the cell cultures under study. The quality of the generated images will be evaluated against such a set of pre-defined analysis tasks. Here we show a high-level example where an image from channel two is used for three different analysis tasks related to morphological, intensity, and textural biomarkers. For each task, there is a corresponding loss that will gauge the difference in task output when using the real and generated image respectively.

The different measurements from the downstream analysis pipeline will be provided by AstraZeneca in the form of code with a corresponding text which describes each underlying measurement. For each measurement, the corresponding code will output a fidelity metric i.e. the difference in measurement output between the generated image and its corresponding target. The end goal of the hackathon will be to generate images such that as many measurements as possible can be made reliably and with high fidelity. To

achieve this, we envision that a combination of custom loss functions and analysis tools need to be developed.

The different fidelity metrics form the basis for a final evaluation metric that will be used to compare the different contributions. More information about the measurements, fidelity, and evaluation metrics will be provided at the start of the hackathon.

Description of Data
The images for the challenge will be provided by AstraZeneca in the form of tif files. There are three sets of images corresponding to three different magnification settings (20x, 40x, and 60x) of the microscope. For each field of view, there will be seven bright field images for different values of focal plane, and three different fluorescence images corresponding to labeling of nuclei, lipocytes, and cell matrix respectively. There will be on the order of 50-100 such sets for each magnification setting. Each image is approximately 2156 by 2556 pixels in size, using 16 bits to represent each pixel value. Please see the associated Data Description presentation for additional information about the data.

How to enter
The number of teams that can enter the challenge will be capped due to infrastructure limitations. To enter the hackathon, we require a written application (maximum 1500 words) containing a proposed solution method, including identified risks and mitigation strategies. A resumé describing the relevant background and experience of the team is also required. Based on the written applications, eight teams will be chosen to participate in the challenge.

We acknowledge that to construct a solution strategy, it is often necessary to have access to the data. Hence, a smaller, representative data set will be made available before the competition through AI Sweden.

Timeline
The challenge will last for two weeks (November 2 –15), during the challenge teams will have access to computational resources from AI Sweden. Q&A sessions will be held with researchers from AstraZeneca during the two weeks. Results including code and quality metrics should be sent in no later than November 15. All teams will present their solution for the jury at the final event on November 19. All events will be held online.

Evaluation of Contribution

The challenge contributions will be evaluated according to the following

1) A set of pre-defined quality metrics will be used to measure the objective quality of the generated images. These metrics as well as the code to calculate them will be released at the start of the hackathon.

2) A jury with members from both academia and industry will assess the quality of the different contributions based on the presentation given by each group at the end of the hackathon. The jury will take into account different aspects of the contributions apart from the quality measured in 1) e.g., presentation skills, solution strategy, and code readability.

Computational Resources

The teams will have access to an NVIDIA A100 (one GPU per team) and support through AI Sweden.

Usage of Results

The results and resulting code of the hackathon should be openly published and free to use for non-commercial purposes.

The selected teams will be required to sign the license terms for the dataset as well as a user agreement for the provided infrastructure before the competition.

Prices

The total prize sum is $5000 sponsored by AstraZeneca.

Questions?

Please contact: johanna.bergman@ai.se