Carnegie Mellon University

Streamlined Image Processing for Protein Crystallization

MS-DAS Capstone collaboration with Bristol Myers Squibb (Spring 2024)

Richard Wang, Rachel Tang, Gayatri Chabra

Mentors: Stephen Thomas, Ankit Kanthe

Introduction

Project Description

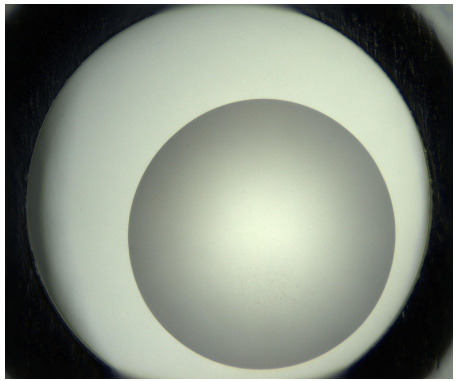
Addressing Challenges in Protein Crystallization Analysis

- Current bottleneck: conventional image processing
- Need for efficient workflows
- Potential solution: Machine Learning and Neural Networks
- Focus on Protein Crystallization

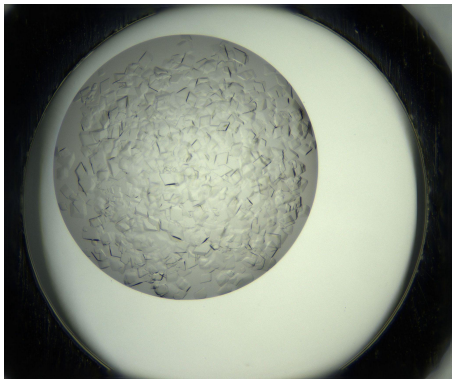
Proposed Objectives

- Identify drops showing crystal formation (currently manual)
- Generate phase diagram to understand inhibition concentrations
- Identify crystal-free, nucleation, and precipitant zones
- Utilize high throughput analysis on a 96-well plate
- Fully automated experimental process

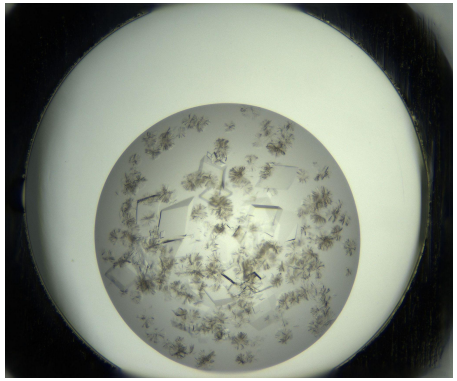
Protein Crystallization Classes



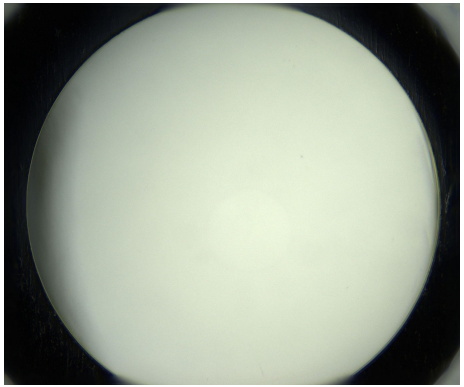
Ill-formed: Empty Drop



Well-formed: Pure Crystal



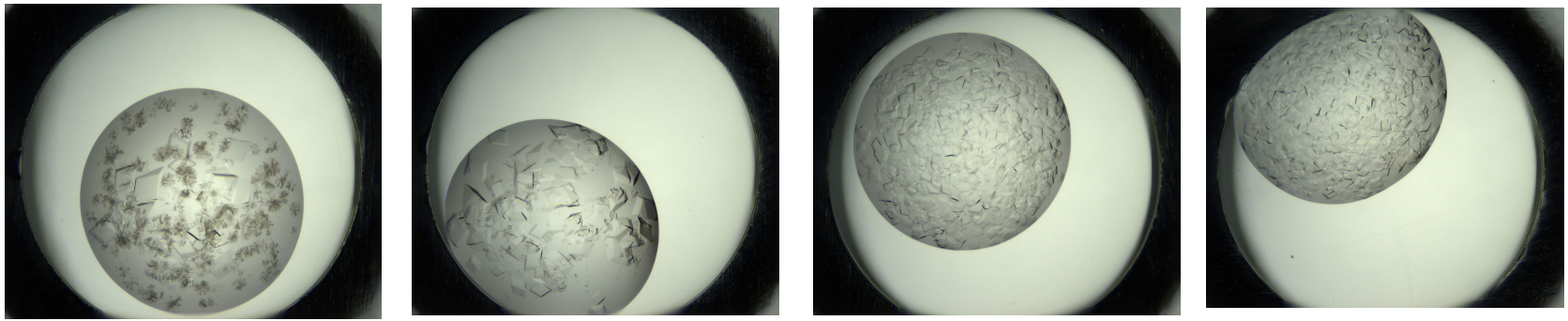
Others



Ill-formed: Clear drop

First Step: **Select Droplets**

Before we start, we want to crop the droplets from images to simplify the classification task.



Initial Circle Detection with OpenCV

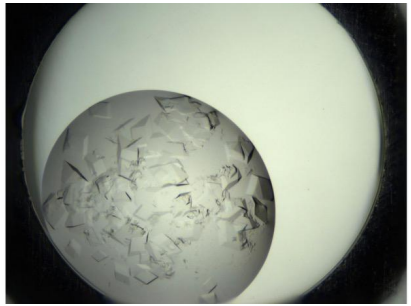
Due to the limited number of images, we start with some computer vision techniques

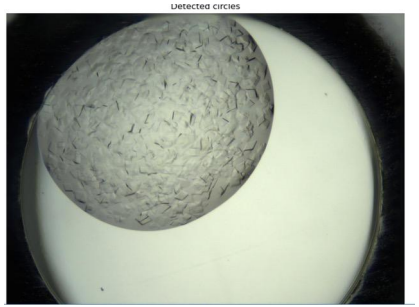
- Edge detection using Canny
- Circle detection using Hough Transform
- Contour detection using ‘cv2.findContours’

By utilizing these techniques, we were able to

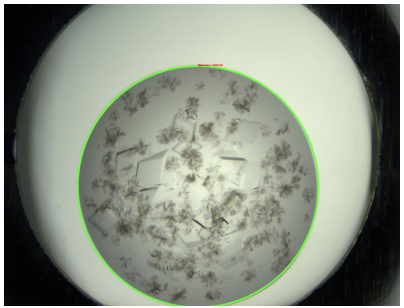
1. **Distinguish between Complete Circles and semi-Circles**

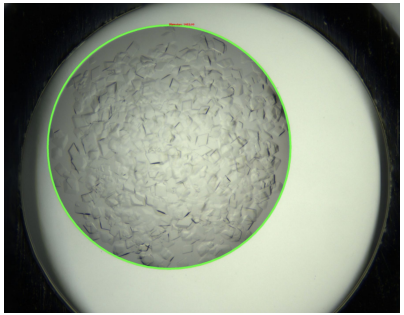
Semi-Circle Droplets



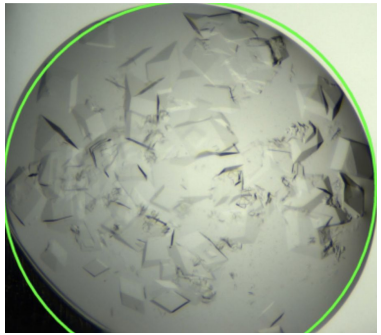


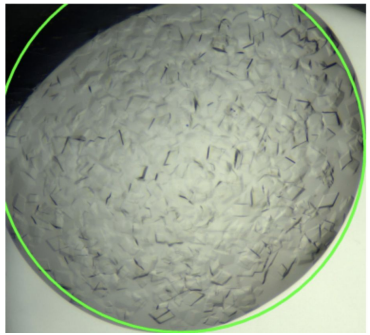
Complete-Circle Droplets

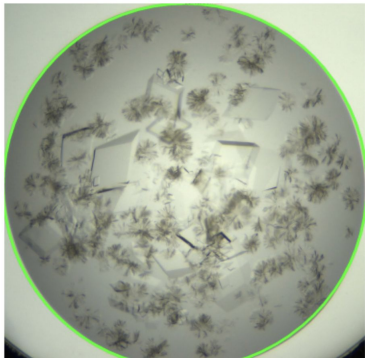


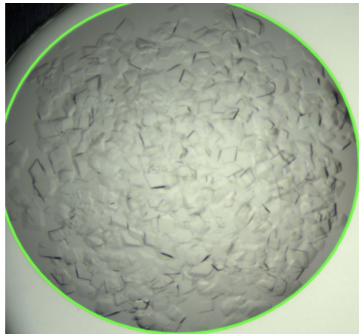


2. Crop the Droplets regardless of their shapes





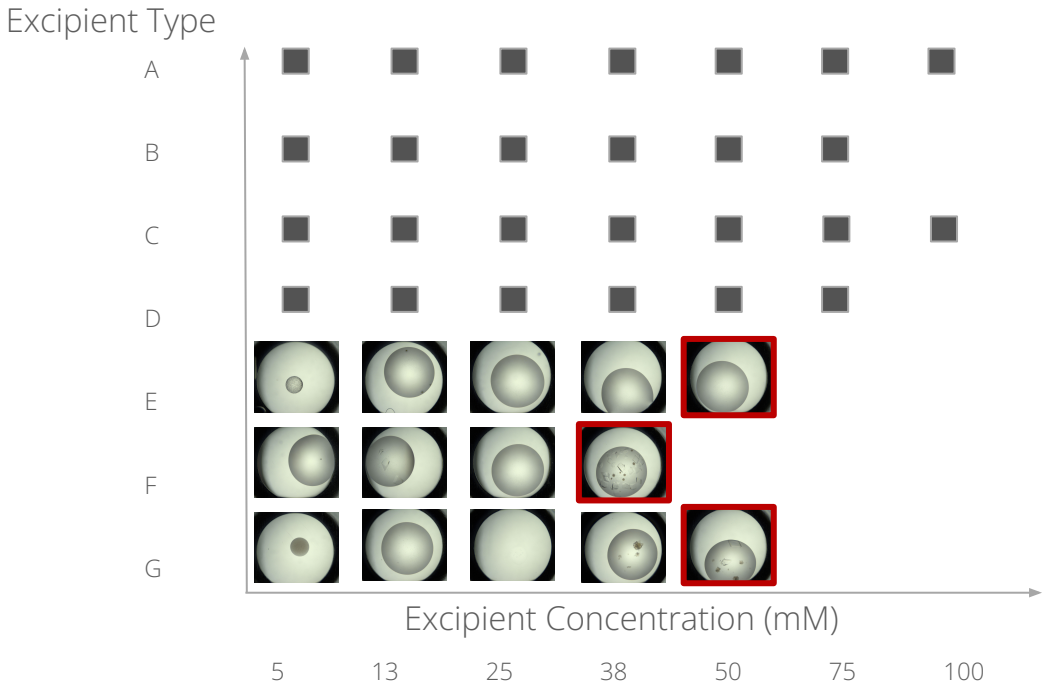




- To segment the droplets, we automatically encircle them and then create bounding rectangles using'cv2.boundingRect'.
- This method effectively isolates the droplets, regardless of their shapes.
- The resulting images are ready for the classification task

What's Our Final Goal?

Final Goal: **Phase Diagram**



Identify **concentration** for every excipient type that

- Inhibits crystallization

We want to identify the concentration that

- Clear drop
- Crystal + snowflake
- Crystal + precipitant

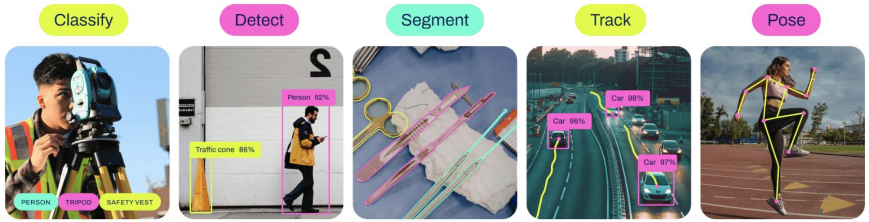
Our Ideas

Idea #1: **YOLOv8**

What is YOLOv8?

YOLO (you only look once) is an open source computer vision model

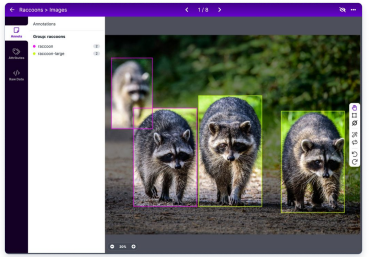
We mainly want to focus in **classification** task



Roboflow

A platform we can use to hand select images and label them

We can use it to select crystal droplets and label them



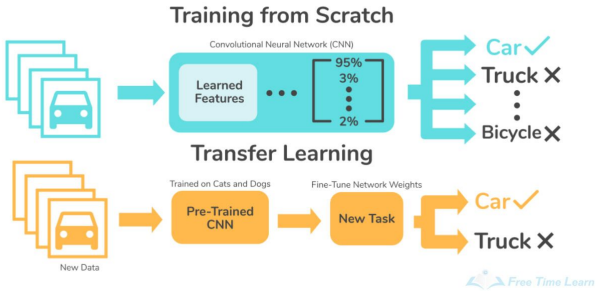
Carnegie Mellon University

Idea #2: Transfer Learning (Pytorch)

Transfer Learning is to use a pretrained model and keep training it on a new dataset while retaining previously learned weights

Ideal for tasks without much data to work with

Many examples of transfer learning to reference to in Pytorch



Idea #3: Train a **CNN**

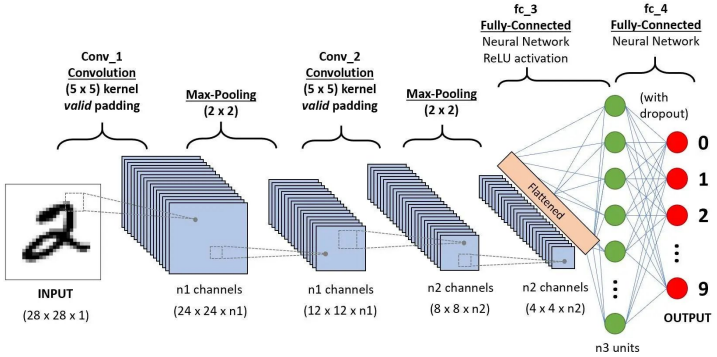
Convolutional neural network is great for image classification

Potential concern: *lack of data*

- We have a total of 1344 images
- Some are unwell crystals

Potential fixes

- Data Augmentation
- Tensorflow ImageDataGenerator



Benefits

- Simplicity in implementation
- Easier to understand errors and improvements
- Freedom in parameter tuning

Workflows

Our task is essentially a **image classification task**

- Given an image of a droplet
- Classify if it is
  - Ill droplet
  - Pure crystal
  - Clear droplet
  - Crystal + snowflake
  - Crystal + precipitant

We need to create a **metadata** on every excipient

- Concentration type
- Control environment type
- Excipient type

We compare model result with metadata, and observe the concentration threshold for every excipient type

Finally, construct phase diagram

Thank You