Problem owner name

Gavin Rice

Problem title

CryoFilter - A deep learning solution for high resolution protein structures.

What is the problem you want to solve?

Current methods for creating new medicines consist of high throughput screens of hundreds to thousands of drug candidates against a target and are enormously wasteful both in terms of material and researcher's time. For the moment, an entirely computational drug discovery platform is not possible, but steps are being made on multiple fronts to bring this to fruition. One major step that needs to be reached in order to create better medicines through computational drug discovery is the creation of high resolution 3D models of drug targets. In recent years, hardware advances in the space of cryo-electron microscopy have brought this into the realm of possibility, but the computational structure analysis workflow is still wrought with complications.

Cryo-electron microscopy datasets consist of thousands to millions of images of individual particles that can be used to create high resolution 3D models of important drug targets. A significant challenge currently in this workflow is how to effectively remove bad images such as noise images or broken particles from the data. In a dataset of thousands of particles, just a few bad particles can significantly hamper the ability to generate a high quality 3D reconstruction. Current approaches to particle sorting rely on significant manual effort and are overall ineffective in producing clean data.

Why do you want to solve this problem?

Creating an open source tool to filter out broken particles from cryo-EM data would both make solving 3D structures faster, and improve the spatial resolution of the final models. Such a tool could be used by structural biologists around the world to aid in high resolution structure determination, a major step towards creating a fully computational drug discovery pipeline.

What do you envision as the ideal solution for this problem?

Software using machine learning to detect and remove bad particles from sets of images of particles that have been previously classified as similar by a maximum likelihood classification algorithm would cut out the time-consuming step of manually filtering data resulting in faster and better 3D reconstructions.

What sort of Open Source solution do you think can be created in 48 hours, by a small team of developers, designers and data analysts?

A basic program that uses machine learning and is trained on good and bad particles. The program would then be given the result of one image class from maximum likelihood 2D classification, that is, a set of thousands of particles that the previous algorithm determined are identical and be tasked with removing bad particles based on the training set. From here, the images could be directly fed back into the next step of the typical analysis workflow. Such a program would cut out hundreds to thousands of hours of researcher's time spent manually filtering such data via computationally expensive classification algorithms.

Are there datasets or people with domain knowledge that you will be bringing to work with? What/who are they?

I will be contributing my own domain knowledge of cryo-electron microscopy data analysis. Additionally I will provide training data consisting of >1000 good and bad reference images and a ground truth image as well as test data consisting of >200,000 images to classify.

What are the current solutions for handling this problem?

The current workflow following 2D classification is repeated rounds of 2D classification with manual filtering in between rounds. For example if you sort 500,000 particle images into 200 classes and then select the best 50 classes which contain 300,000 particle images, you would then sort these 300,000 particle images into 200 classes and continue until the particles are evenly distributed across the classes and there are no "bad" classes. Due to the computational cost of 2D classification, each round of 2D classes takes between 12 hours and 5 days. As a result, this stage of the workflow can take weeks to months of a researcher's time and still has the issue that even without "bad classes" there will still be "bad" particles hidden within "good" classes.

Summary for website (up to ~ 1 page)

PROBLEM

What do the top ten most commonly prescribed medications, including treatments for hyperthyroidism, asthma, heart disease, and ADHD all have in common? They all target membrane proteins. In fact, despite membrane proteins making up only a third of proteins in the human body, over half of all medications target them including treatments currently in development for COVID-19. However, the structural and functional mechanisms of many of these proteins remains a mystery making the effective development of new medicines highly wasteful and difficult.

In 2017 the Nobel Prize in chemistry was awarded to Jacques Dubochet, Joachim Frank and Richard Henderson for their contributions to the development of cryo-electron microscopy, a technique that has led to a revolution in finding structures of membrane proteins by allowing researchers to capture 2D images of individual molecules and use these to create high resolution 3D models. These high resolution 3D models can then be used as the basis for a computational drug discovery platform. But despite these advances, building a high quality model of a membrane protein is not a trivial task.

All datasets have anomalies, in cryo-electron microscopy data these come in the form of broken particles. One key challenge in the analysis workflow is sorting high quality 2D images from images of broken particles or other false positives. Current platforms use a maximum likelihood approach to 2D classification that groups similar looking images into classes that can then be manually determined to be good, and used in further analysis, or bad and removed from the dataset. While useful on a very large scale, this method takes a significant amount of time and worse, it never truly removes all the bad images and just a few bad images in a dataset of thousands of particles can significantly hamper high quality 3D reconstruction.

This is where 'CryoFilter' will come in. We envisage a machine learning based software tool that can automatically detect and remove bad particles from the datasets, leaving only the good particles to rapidly create the 3D high resolution structure of the membrane protein. CryoFilter will integrate seamlessly into existing workflows while removing the crucial step of manual sorting. Once in use, CryoFilter will save biomedical researchers in hundreds of groups around the world weeks to months of time which they now spend manually selecting images, significantly speeding up the discovery time.