The Human Protein Atlas

Deep Learning Image Classification

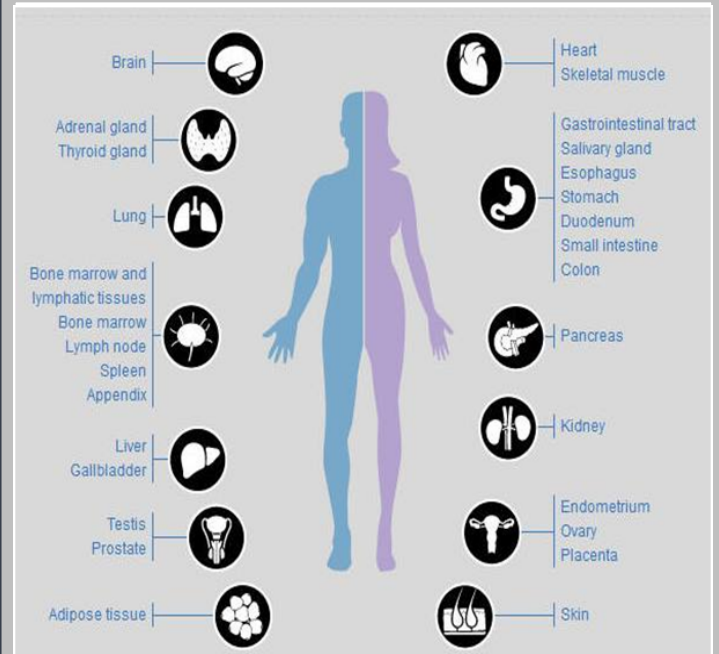
By: Alex Husted

What is the HPA?

- Swedish-based program started in 2003.

- Aims to map human proteins in cells, tissues, and organs.

- Uses the integration of various omics technologies.

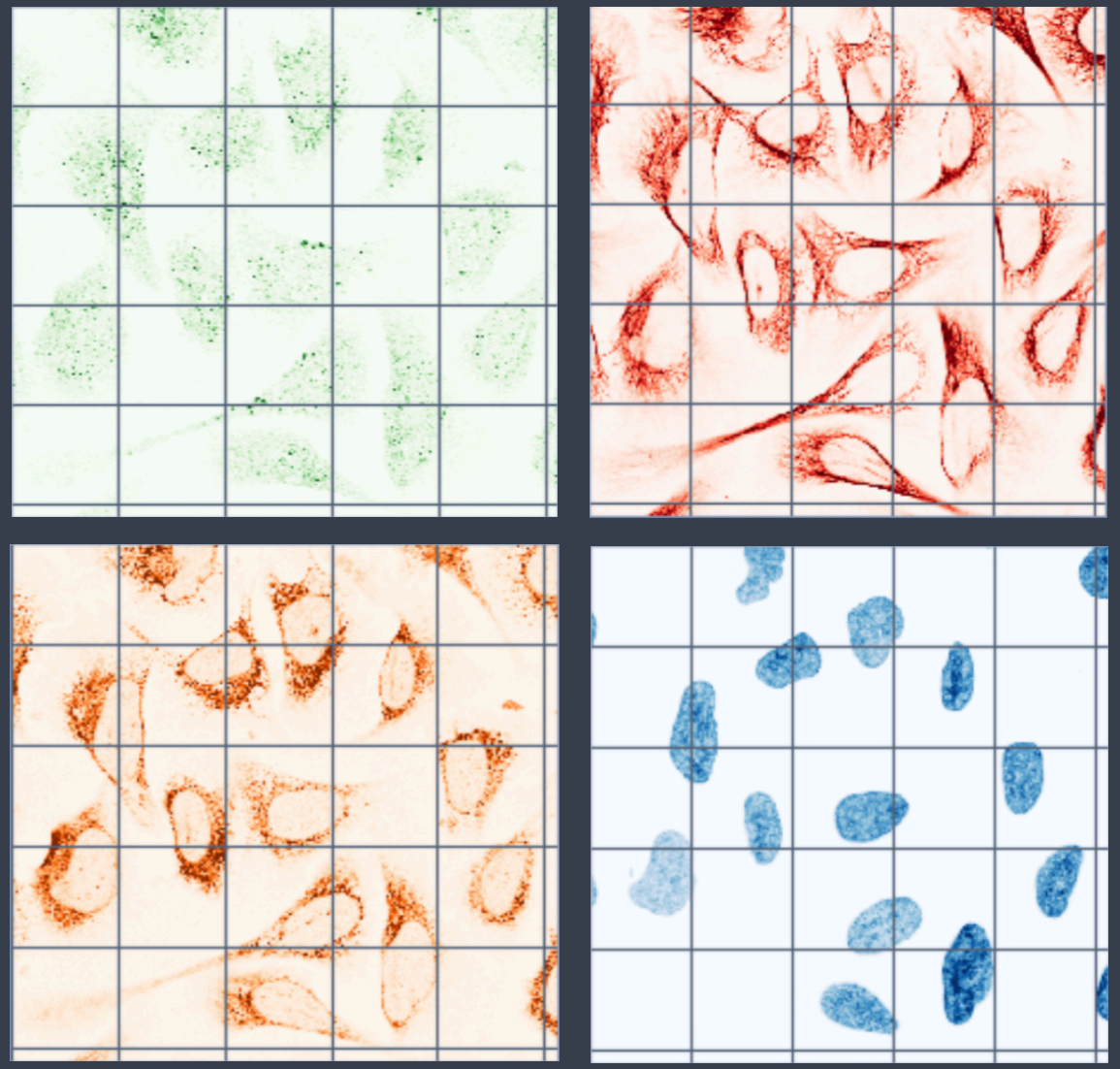


What is the scope of the project?

- Classify mixed patterns of proteins in microscope images.

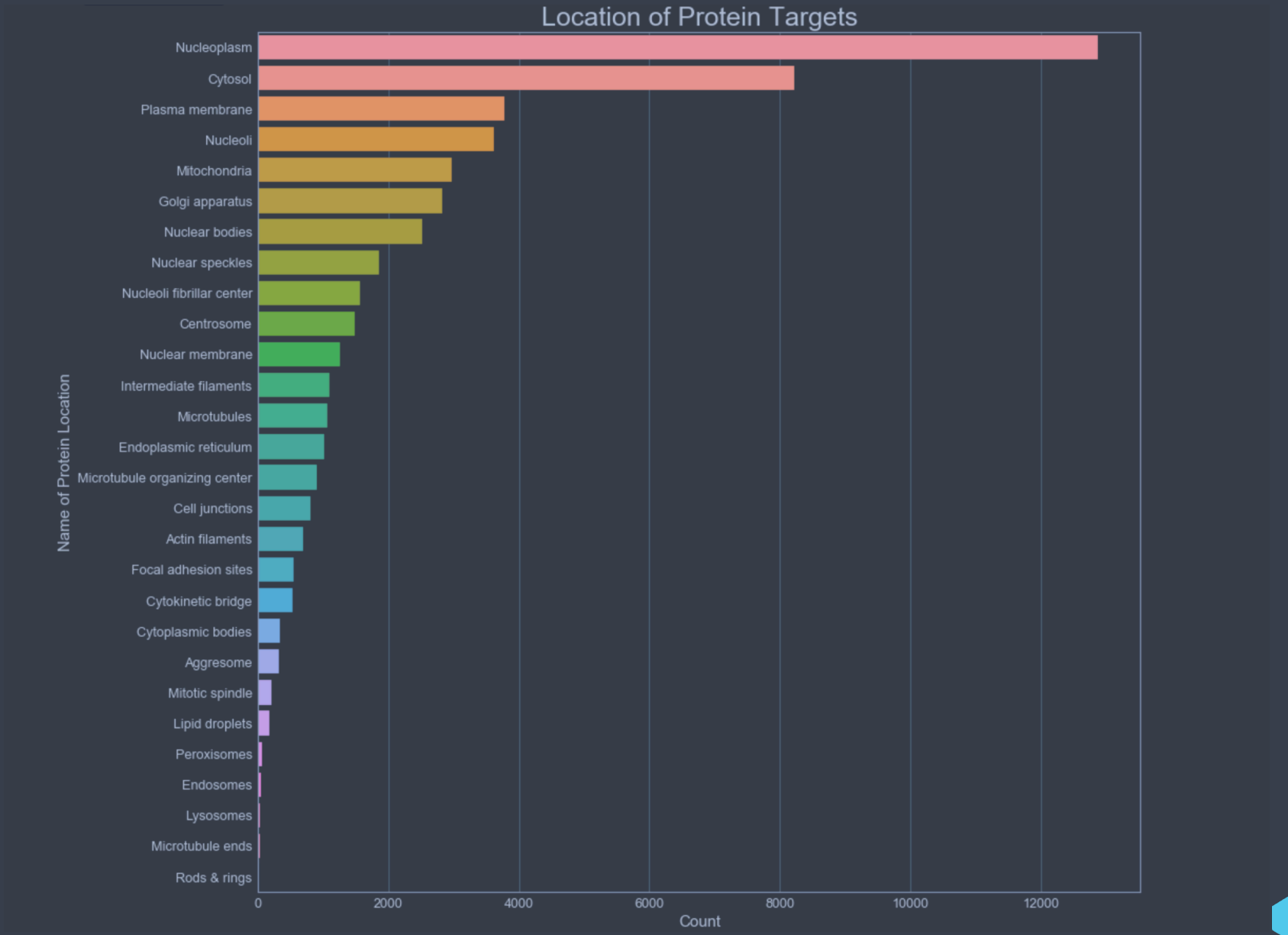
- Multi-label classification problem.

- Use a Convolutional Neural Network (CNN) to differentiate images.



A look into the data:

- 512x512 sized images.
- 31,100 training files.
- 11,700 testing files.
- 28 different target proteins.
- 27 different cell types.
- Images are represented by four filters.

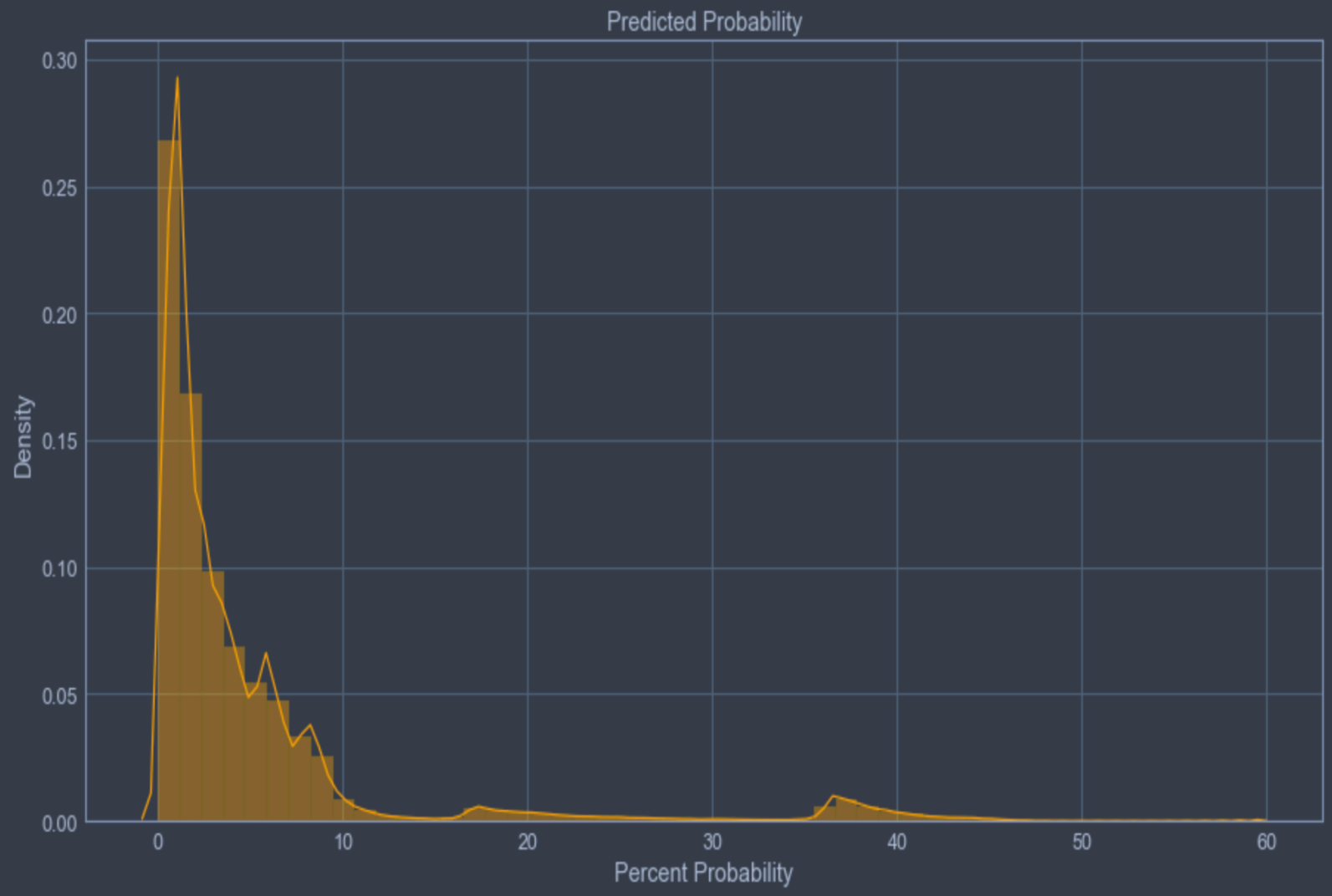


Preliminary Results:

- Accuracy Score: 94%

- Model is predicting absence of target proteins.

- Failed to predict presence of target proteins.

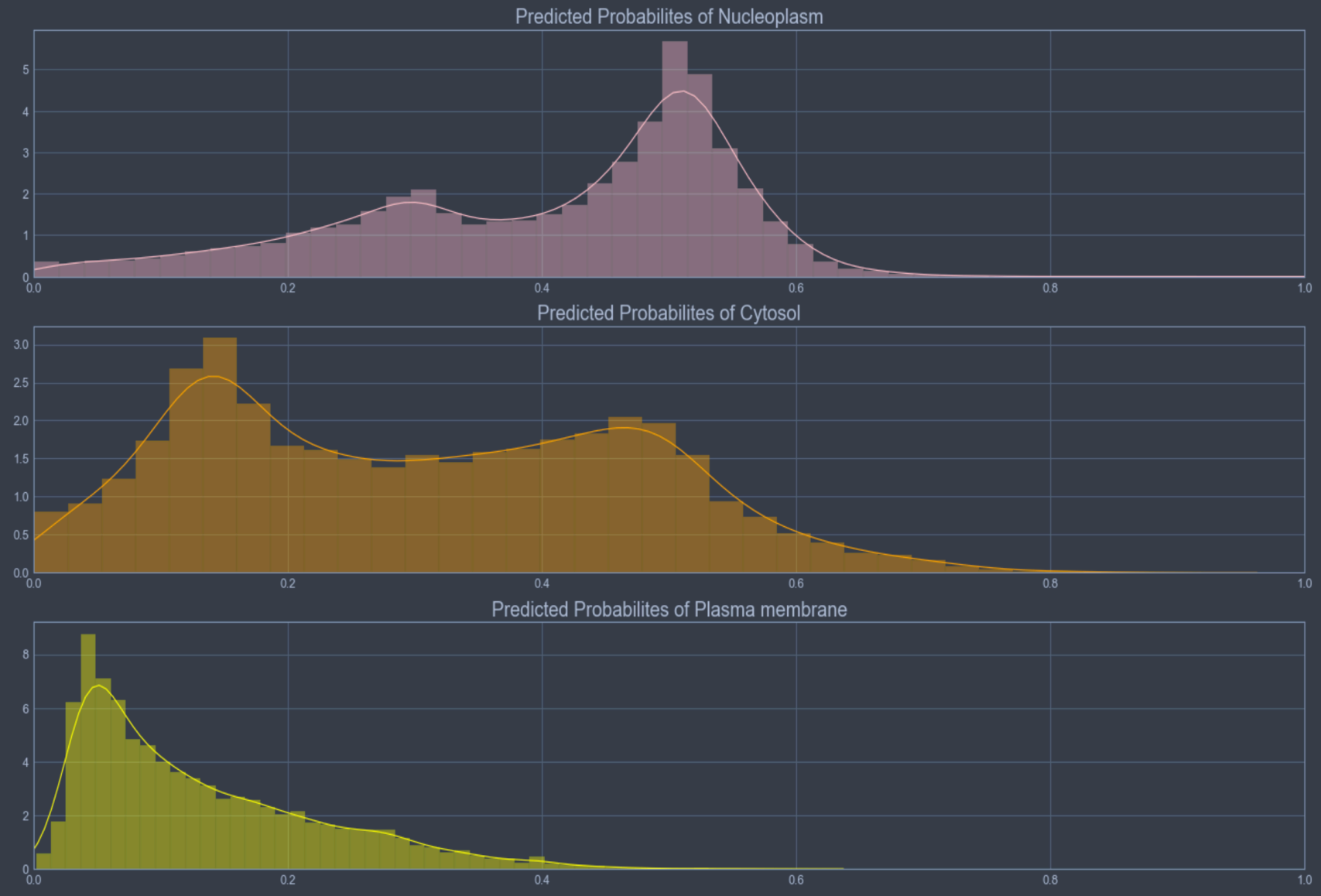


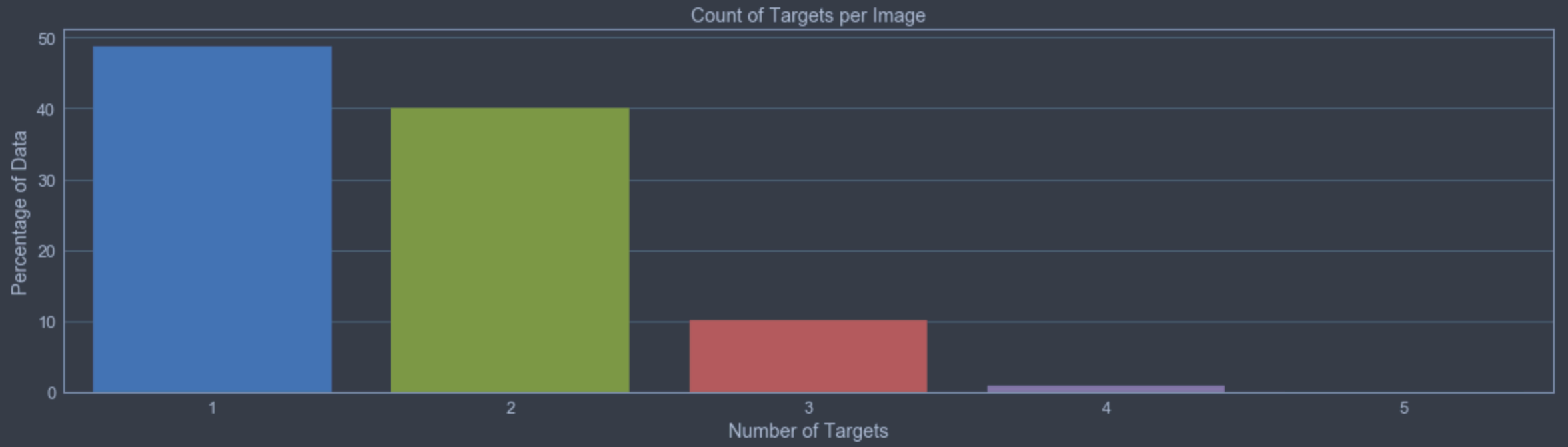
Further Results:

- Chose to model top three features.

- Nucleoplasm, Plasma membrane, Cytosol.

- Improved Probabilities:
  Nucleoplasm = 40.5%
  Plasma Membrane= 12.2%
  Cytosol = 28.9%





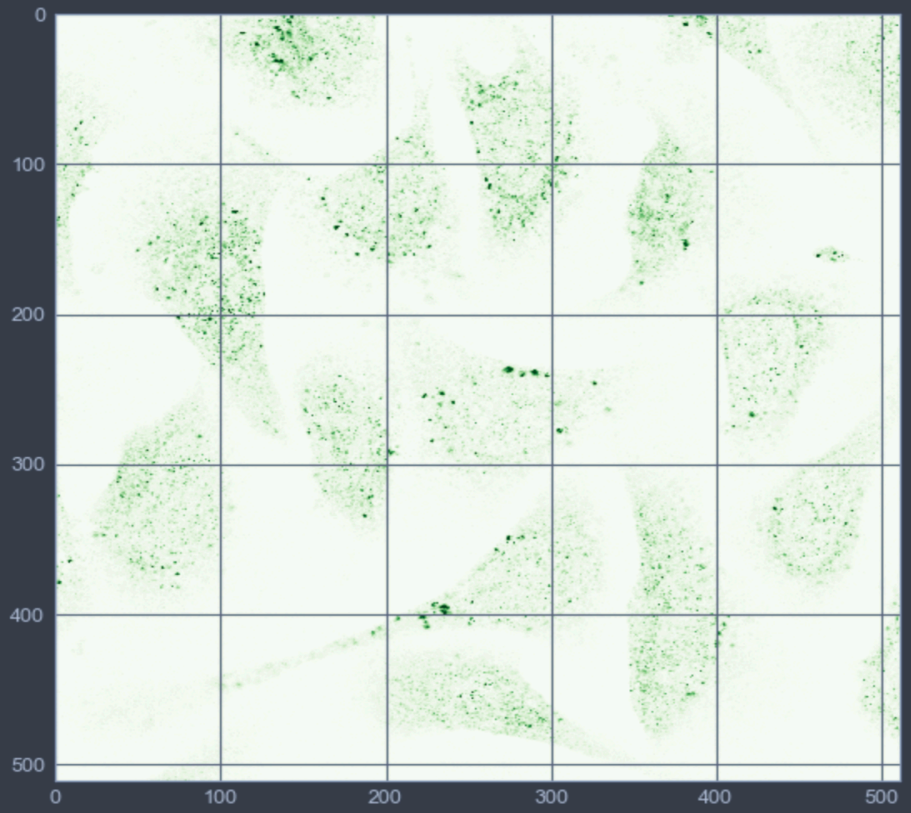
Challenges:

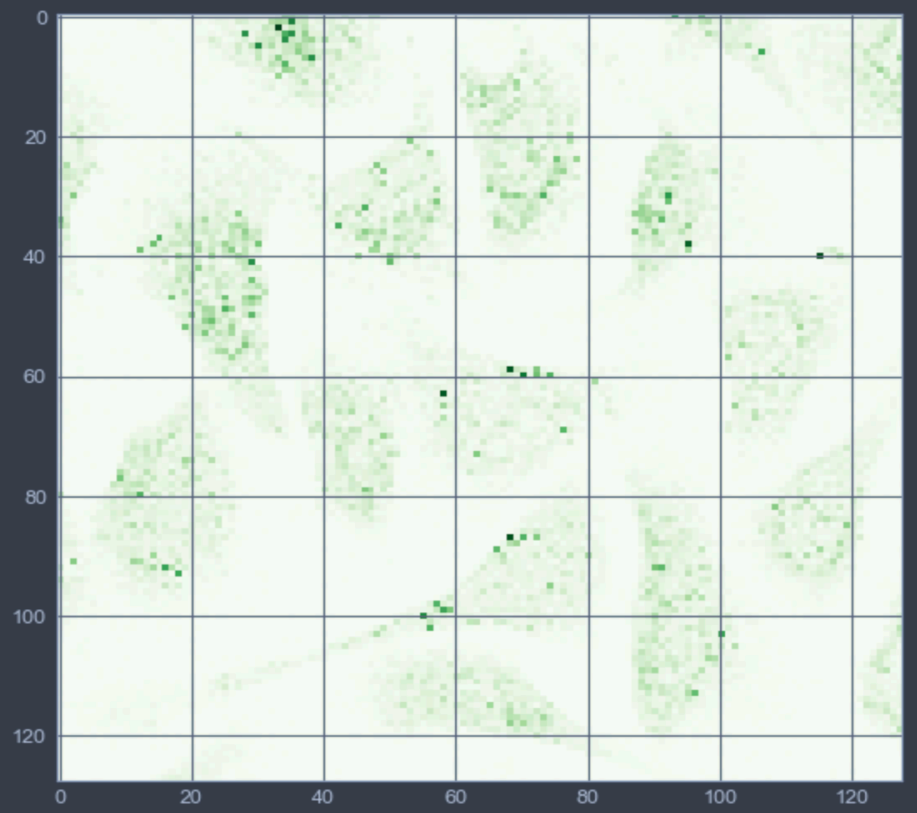
Multi-label classification is inherently difficult.

Images were taken under different filter types.

Large image size: 512x512.

Class imbalance resulted in misleading accuracy.





Solutions:

Loaded images with accurate filter type.

Trained the model based upon image batches.

Reshaped the image size to 128x128.

Chose the top three features to model.

Conclusions:

Used Deep Learning techniques to:

- Classify mixed patterns of proteins in microscope images.
- Perform multi-label classification on batch images.
- Evaluate probability scores.

Recommendations:

- Mixed protein clusters occur in various areas around the cell.
- Many clusters can be found around the nucleus.
- Proteins also form on the edges of cells, ex. the plasma membrane.

If there was more time…

- Continue building CNN model that battles inefficient class imbalances.
- Perform target group analysis using a Latent Variable model.
- Perform Bernoulli Mixture model to target groups found by clustering.
- Building visualizations that show mixed protein clusters within each cell.

THANK YOU!