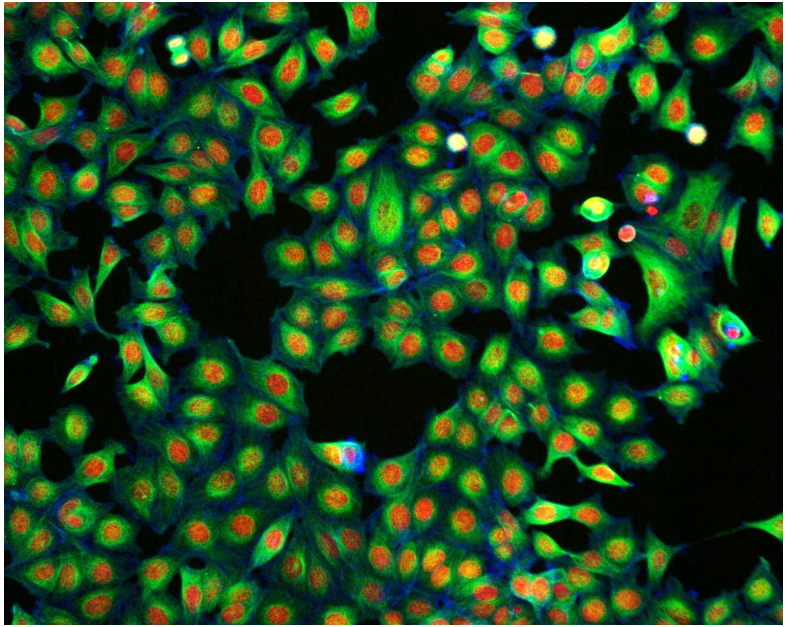
Large Scale Dataset Curation and Model Evaluation

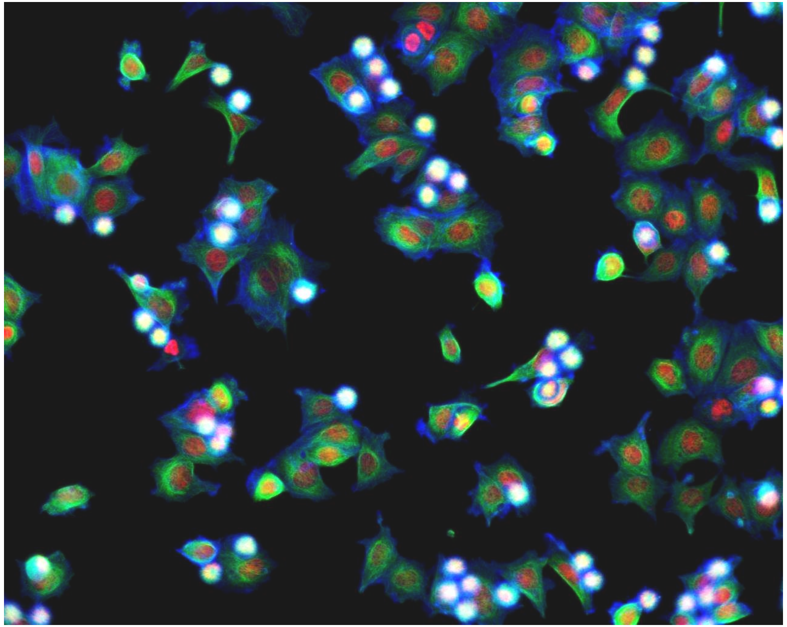
John Peters
Caicedo Lab

6/5/25

Image-based experiments

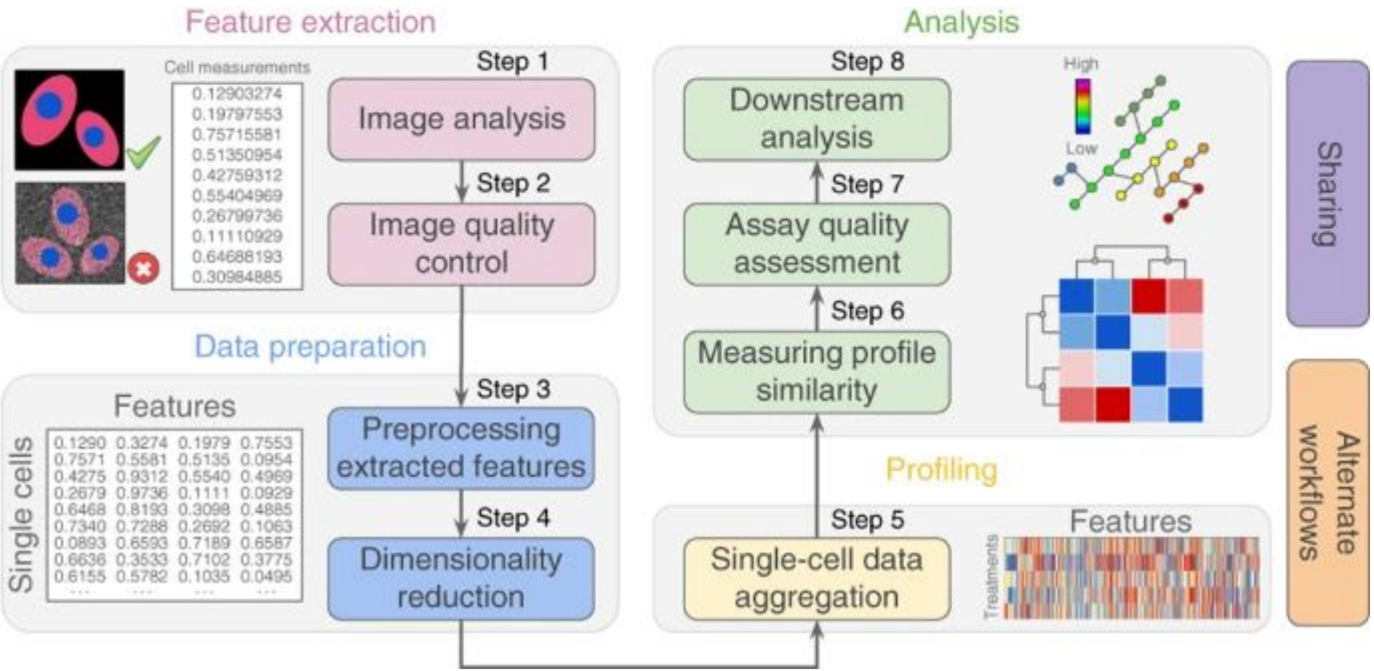


Control condition



Treated condition

Image-based profiling workflow



Feature extraction

Step 1

Image analysis

Step 2

Image quality control

Cell measurements

0.12903274
0.19797553
0.75715581
0.51350954
0.42759312
0.55404969
0.26799736
0.11110929
0.64688193
0.30984885

Data preparation

Step 3

Preprocessing extracted features

Step 4

Dimensionality reduction

Features

Single cells

0.1290	0.3274	0.1979	0.7553
0.7571	0.5581	0.5135	0.0954
0.4275	0.9312	0.5540	0.4969
0.2679	0.9736	0.1111	0.0929
0.6468	0.8193	0.3098	0.4885
0.7340	0.7288	0.2692	0.1063
0.0893	0.6593	0.7189	0.6587
0.6636	0.3533	0.7102	0.3775
0.6155	0.5782	0.1035	0.0495
...	...	...	...

Profiling

Step 5

Single-cell data aggregation

Features

Treatments

Analysis

Step 6

Measuring profile similarity

Step 7

Assay quality assessment

Step 8

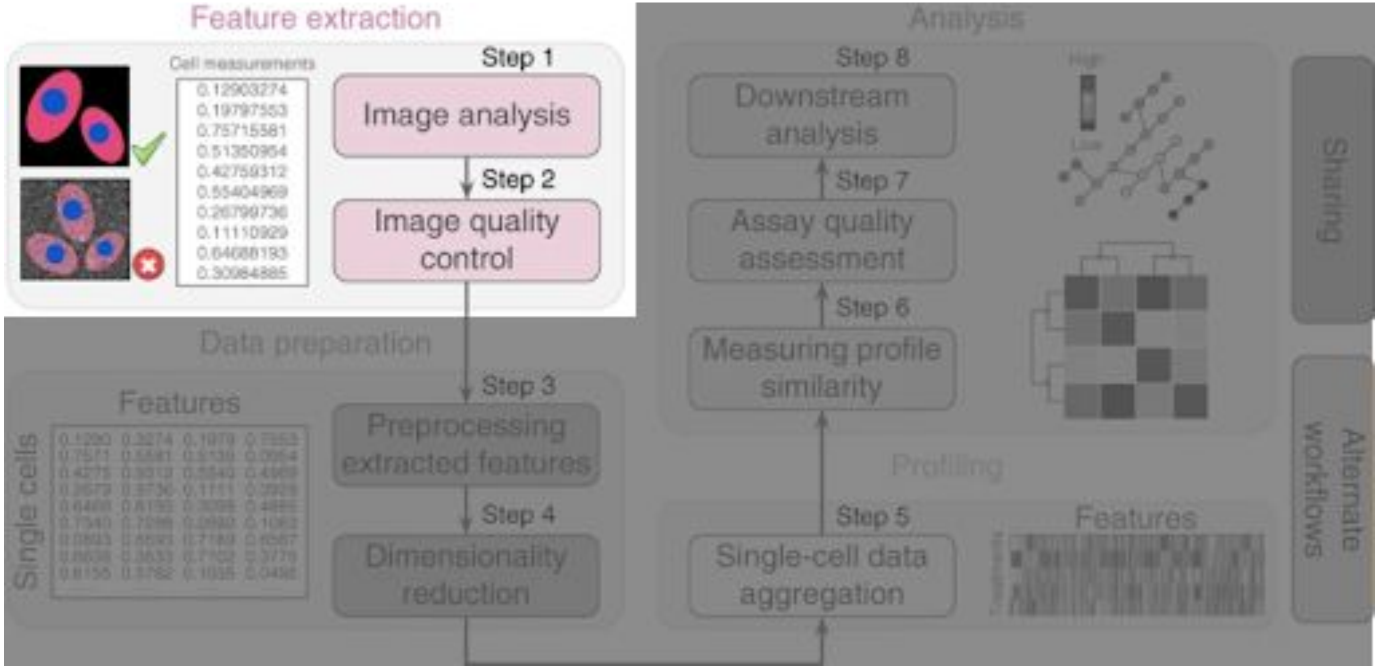
Downstream analysis

Sharing

Alternate workflows

Caicedo et al. "Data-analysis strategies for image-based cell profiling." *Nature methods* 14.9 (2017): 849-863.

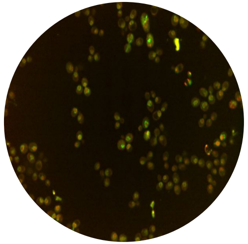
Talk Focus: Feature Extraction



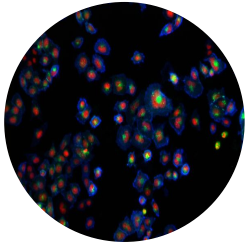
Caicedo et al. "Data-analysis strategies for image-based cell profiling." *Nature methods* 14.9 (2017): 849-863.

5 Studies to Process

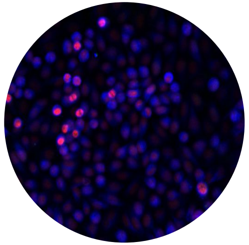
idr0007



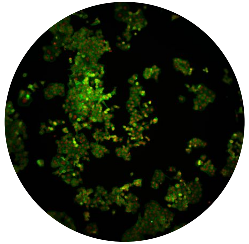
idr0009



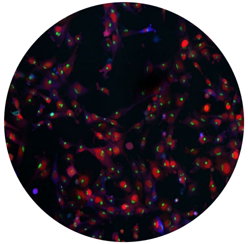
idr0020



idr0017



idr0088

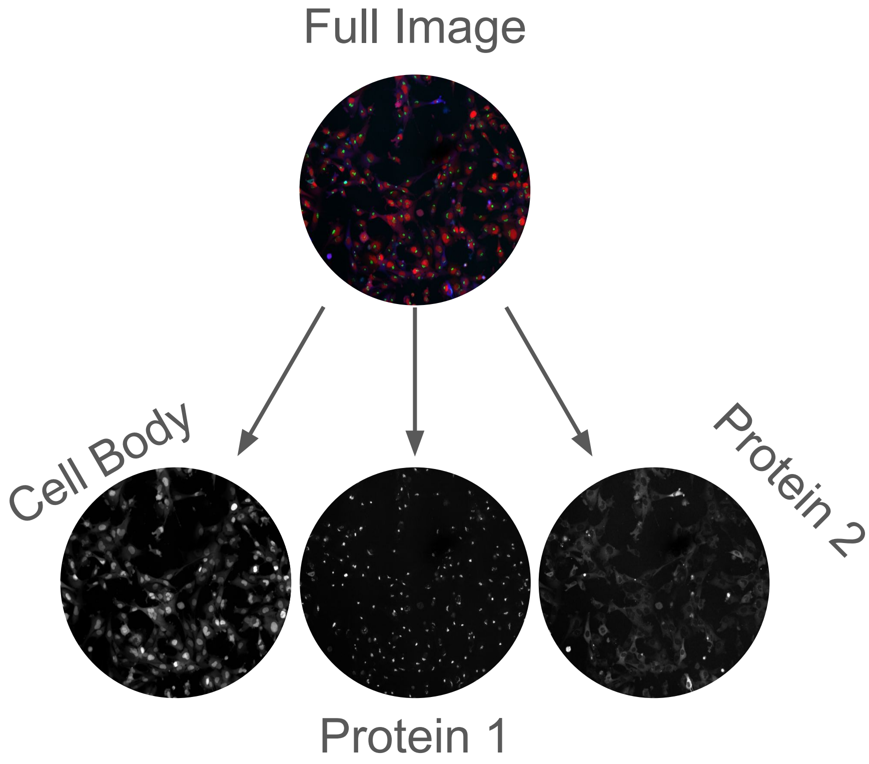


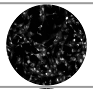
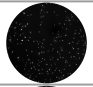
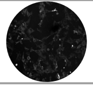
Directory Organization

🖫 Data Vault

```
Images
   |
   |--study
   |    |- study-plate_1.zip
   |    |
   |    |- study-plate_2.zip
   |            ...
   |    |- study-plate_n.zip
   |--study-2
        |- study2-plate_1.zip
        |
        |- study2-plate_2.zip
               ...
        |- study2-plate_n.zip
```

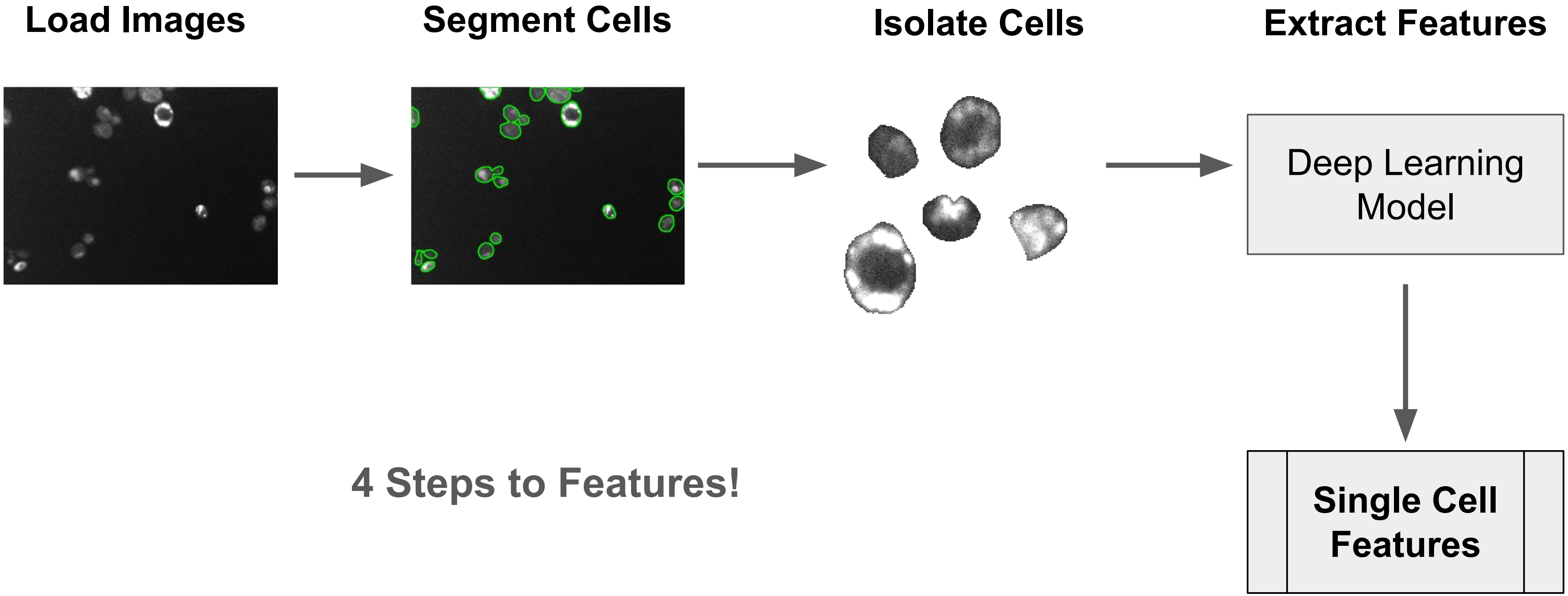
Zip Format



Multi-Channel-Id	Image-Path-In-Zip
hpa_image_0001	 .png
hpa_image_0001	 .png
hpa_image_0001	 .png

Images saved as individual channel png files within *plate_x.zip files

How do we extract features?



Load Images

Segment Cells

Isolate Cells

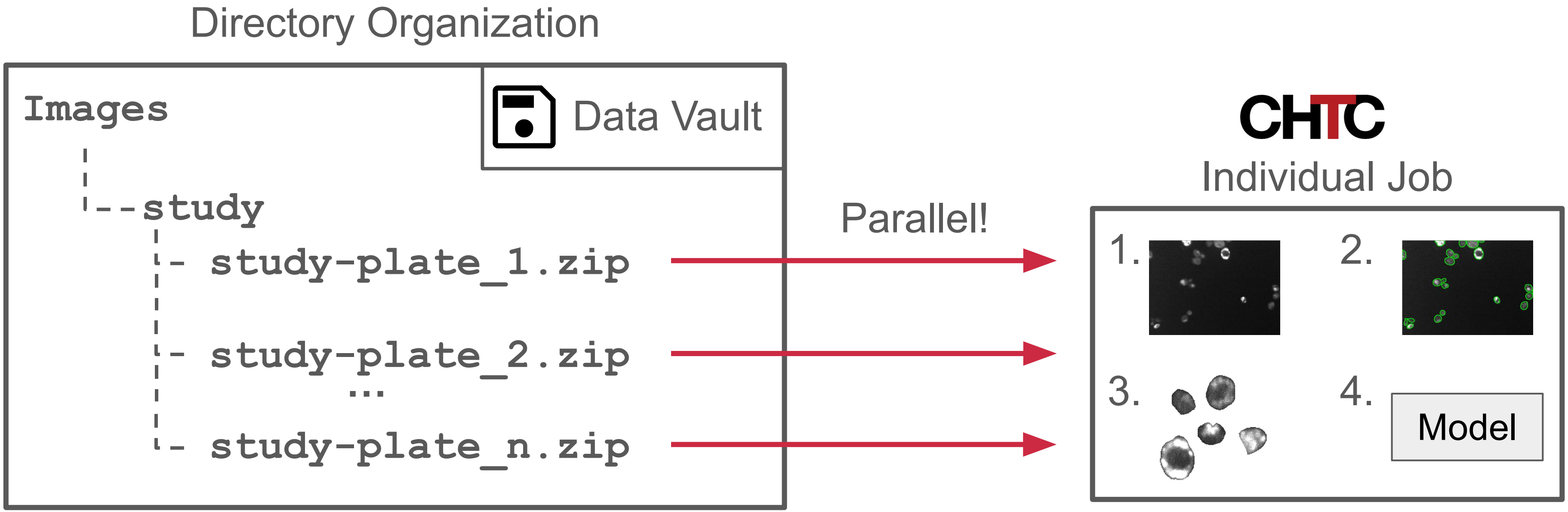
Extract Features

Deep Learning Model

Single Cell Features

4 Steps to Features!

How can we leverage CHTC?



Directory Organization

Data Vault

```
Images
   |
   |--study
      |- study-plate_1.zip
      |
      |- study-plate_2.zip
               ...
      |- study-plate_n.zip
```

Parallel!

CHTC

Individual Job

1.

2.

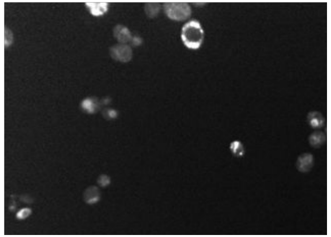
3.

4. Model

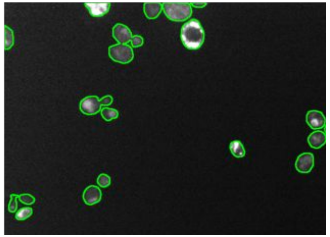
Ankit Gupta et al. bioRxiv 2024.
Kraus Oren et al. CVPR 2024.

Reusable Segmentations – Multiple Models

Load Images

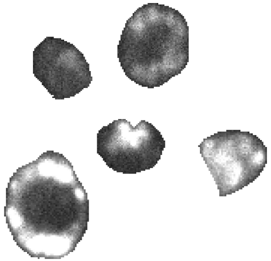


Segment Cells

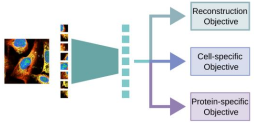


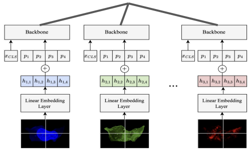
Can cache!

Isolate Cells



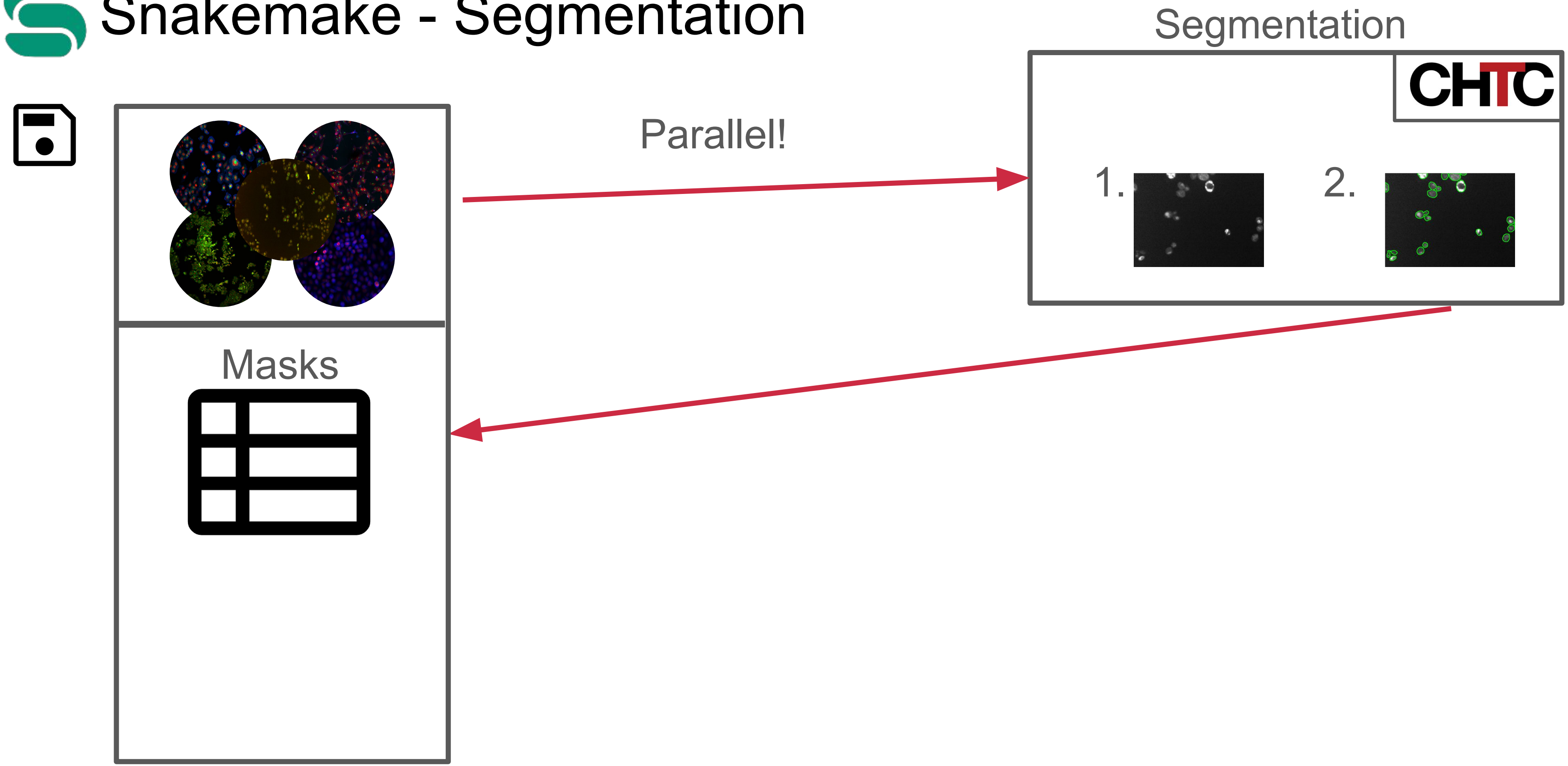
Extract Features





Single Cell Features

Snakemake - Segmentation



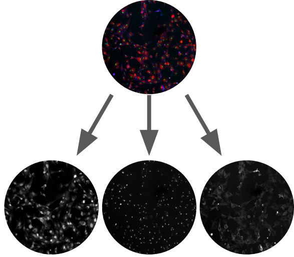
Segmentation

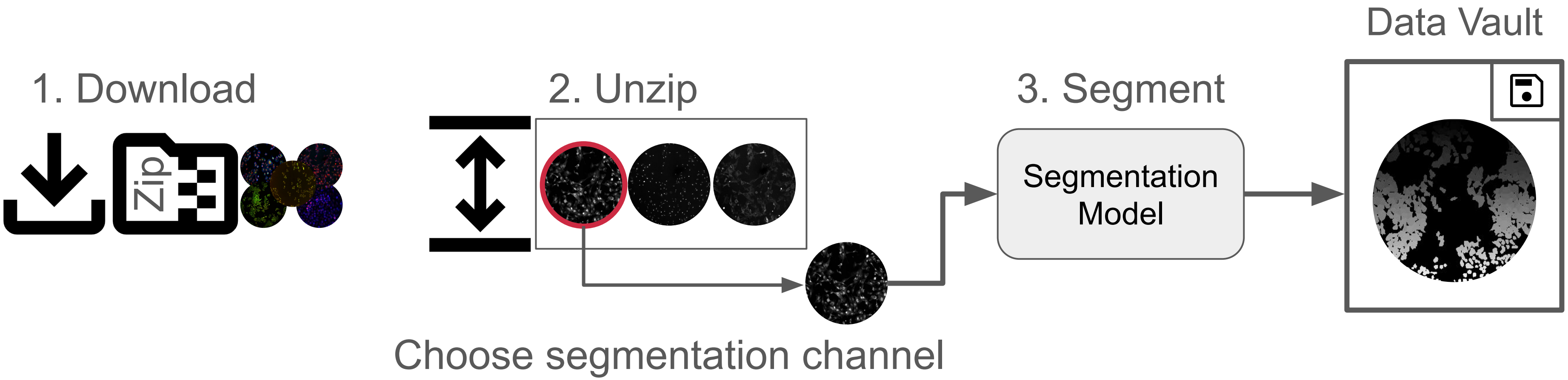
Masks

Segmentation in Detail

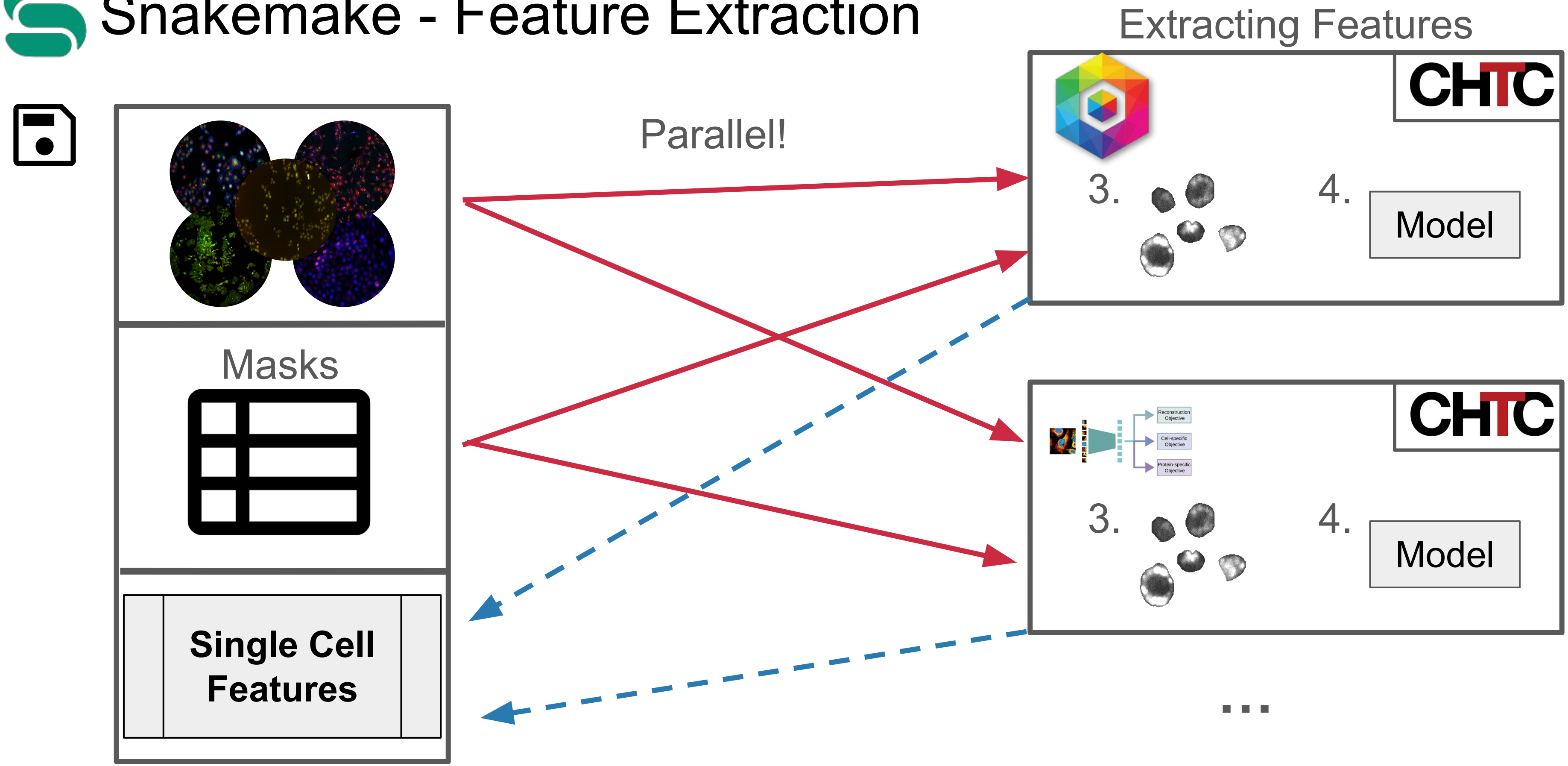
CHTC

Reminder Images are Multichannel:





Snakemake - Feature Extraction



Extracting Features

Parallel!

Masks

Single Cell Features

CHTC

3.

4.

Model

Reconstruction Objective

Cell-specific Objective

Protein-specific Objective

CHTC

3.

4.

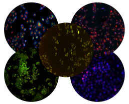
Model

...

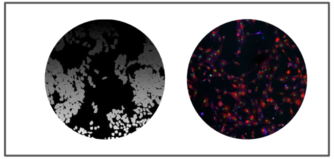
Feature Extraction in Detail

CHTC

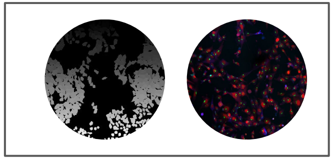
1. Download

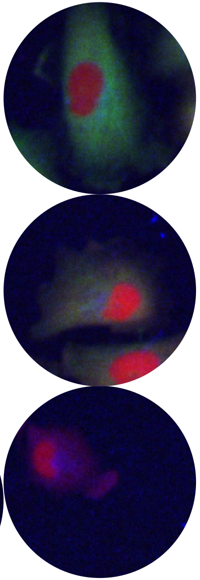


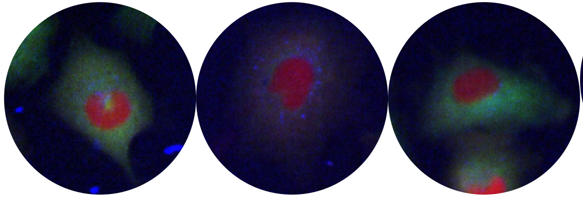
2. Unzip



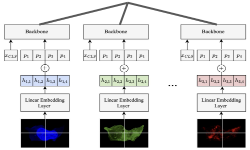
3. Isolate Cells







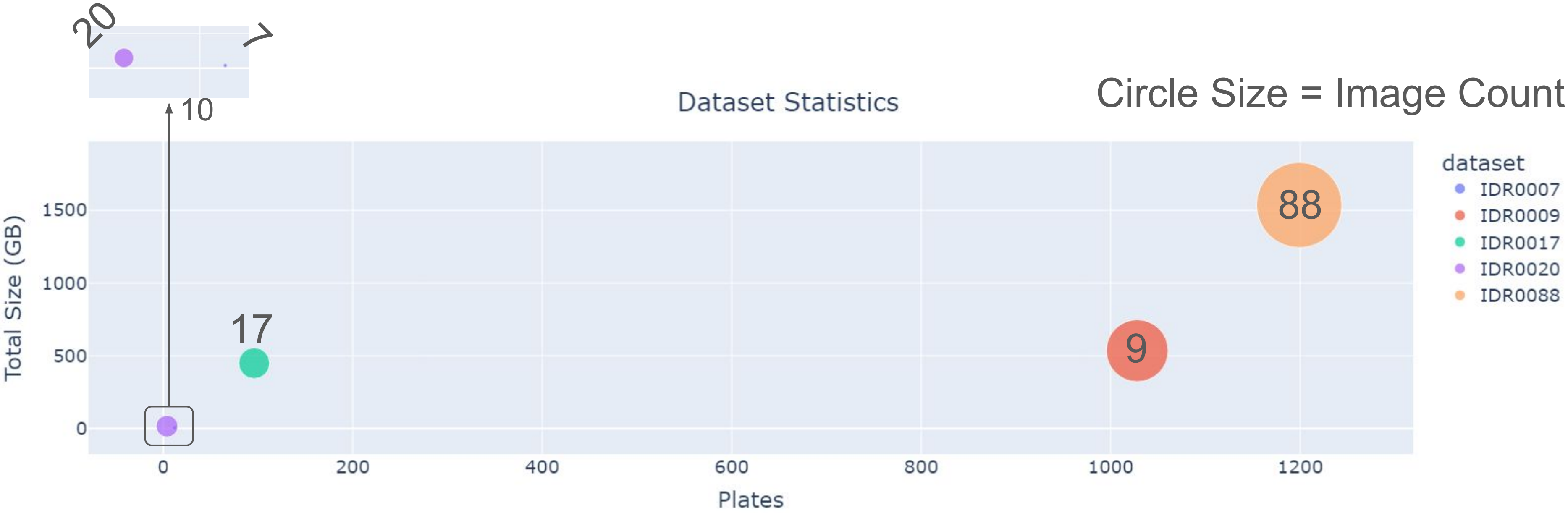
4. Extract Features



Single Cell Features

How large was our data?

Larger studies (GB) with less plates are slower on CHTC



Dataset Statistics

Circle Size = Image Count

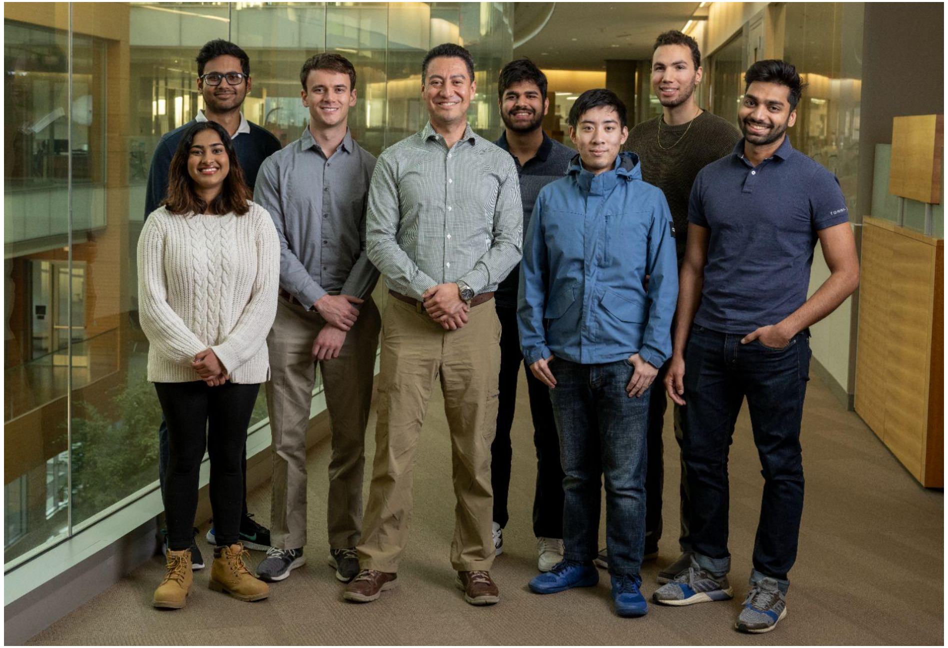
Takeaways

🚀 Small zips work very well on CHTC

🐍 Snakemake allows for local testing before running on CHTC

📦 Caching Saves tons of time with large reruns

Thanks!



Special thanks to Justin and Ian🙂

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

MORGRIDGE
INSTITUTE FOR RESEARCH

NSF

NIH Grant #5T15LM007359