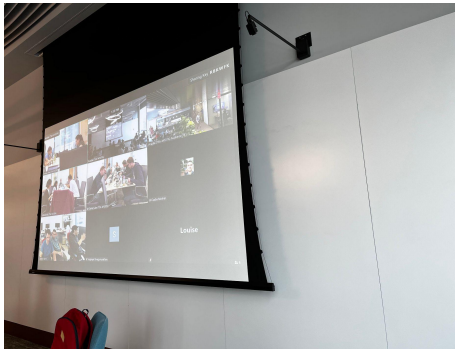
Cytodata Hackathon Highlights

Alán F. Muñoz
Carpenter-Singh Lab
Broad Institute of Harvard and MIT

By the numbers

- 10 hours of hacking
- 43 participants
- 21 participants from 16 different companies
- 13 teams
- 9 mentors/volunteers



It was the best of times

It was the worst of times

Project highlights

Possible to relate the clusters in different modalities?

We obtained connectivity scores for the compound etoposide.

pert_id	pert_iname	Connectivity score against Etoposide
BRD-K92960067	SMER-3	0.92
BRD-K13662825	Dinaciclib	0.9
BRD-K68548958	C-646	0.87
BRD-K35960502	Niclosamide	0.86
BRD-K80738081	Resveratrol	0.78
BRD-K27305650	LY-294002	0.73
BRD-K11528507	Geldanamycin	0.69
BRD-K87949131	Daunorubicin	0.69
BRD-M45964048	Verteporfin	0.68
BRD-K51313569	Palbociclib	0.65
BRD-K92093830	Doxorubicin	0.65
BRD-A79768653	Sirolimus	0.63
BRD-A04322457	Isoproterenol	0.62
BRD-K17953061	Staurosporine	0.59

Are the morphological profiles of the connected compounds related?

Still figuring it out…

- Li-Jiun Chen
- Suganya Sivagurunathan
- Ank Kumar
- Johan Fredin Haslum

Team 13 - Feature subspaces

Team 13: Ian Smith

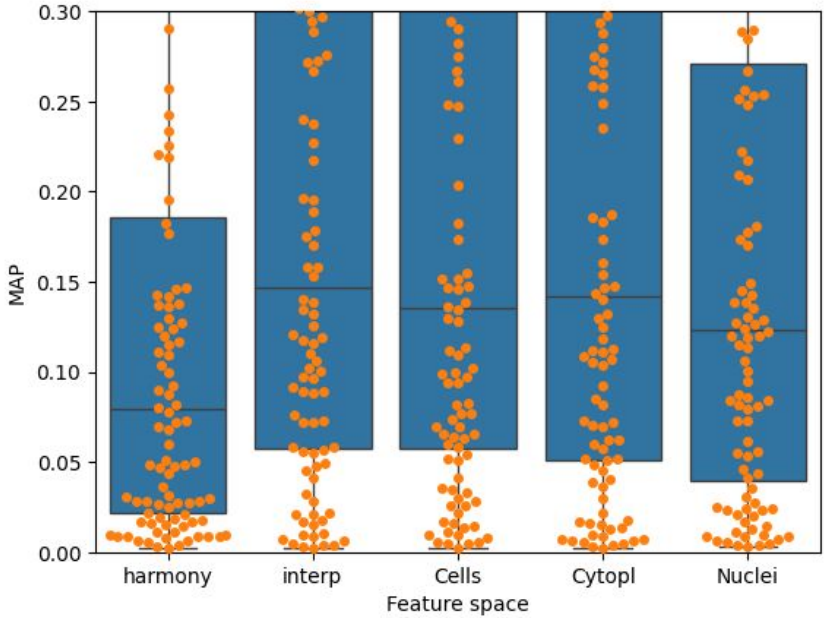
Question: Are feature subspaces as informative as the entire feature space? N = 1959 signatures.

Task: Discriminate CRISPR profiles for **100** genes from negative controls by MAP

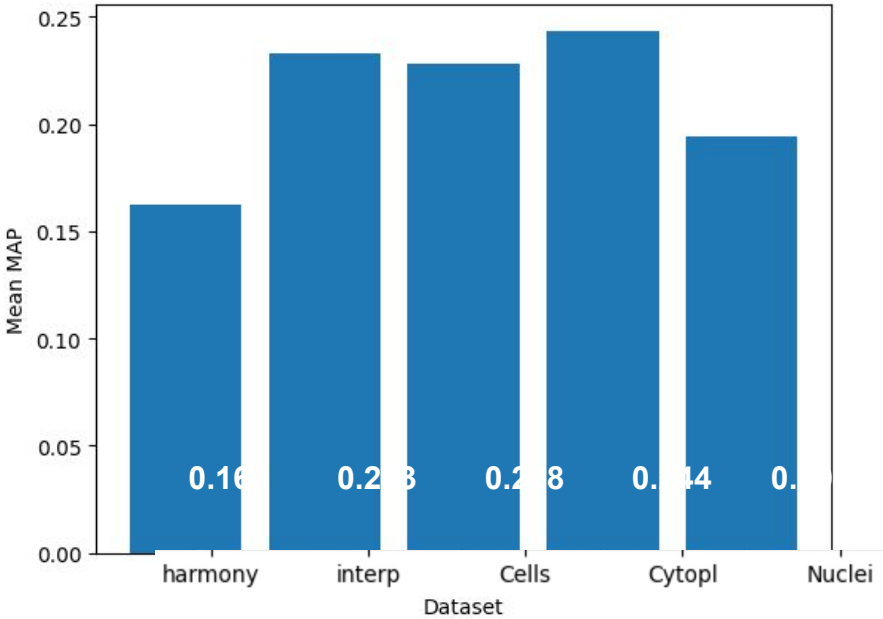
Data: Harmonized (baseline, 259); Interpretable (3651), Cells (1252), Cytoplasm (1234), Nuclei (1165)

BETTER

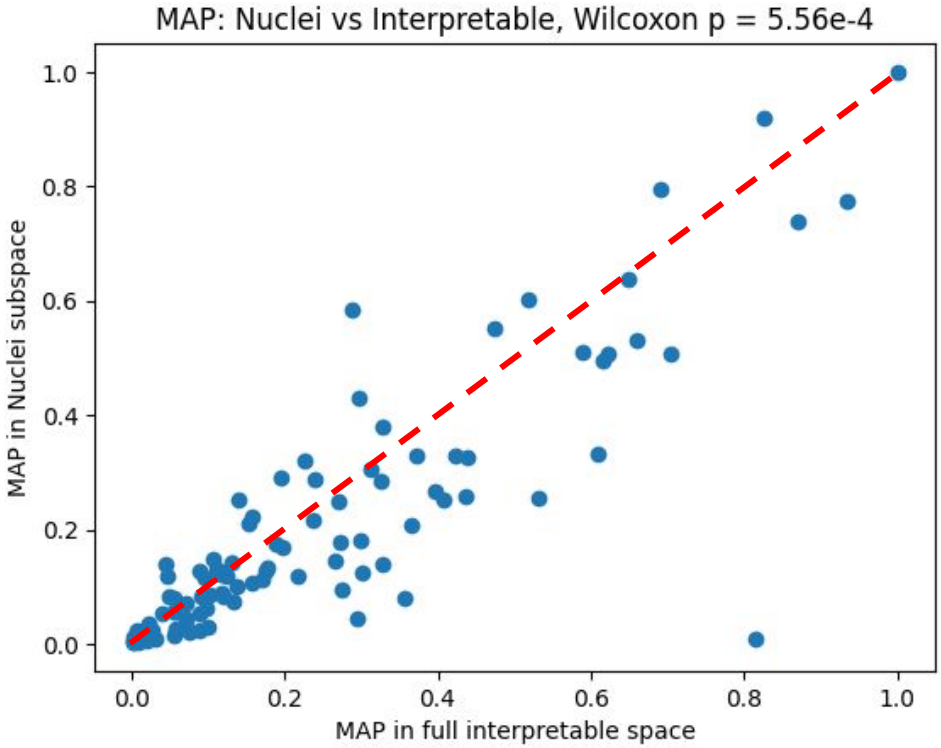
Mean average precision per gene



Mean MAP



Comparison of Nuclei and Interpretable MAP



Conclusion/Future directions:

- Using all the data may not necessarily be informative [fewer stains?]
- Different feature spaces may be more informative for different biology
- Run on more data, e.g. compound classes with known biology

Pipe dream:

Given a biological signal (e.g. MAPK inhibitors), learn a similarity function or feature transformation that maximizes recall of that signal.

Team 13: Ian Smith

Enhancing JUMP cell painting exploration with quick and approximate methods

Tom Ouellette, Yiran Shao, Yuxin (Zoe) Zhi

(1) Download

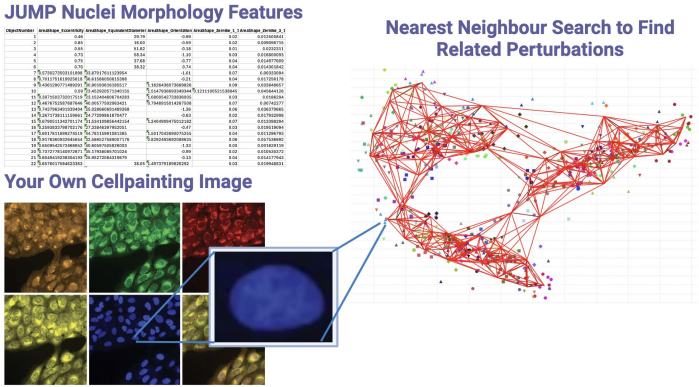
```
[ 2024-09-17 | 17:17:02 | pineapple ] Initializing JUMP download
[ 2024-09-17 | 17:17:04 | pineapple ] Detected 166 samples for downloading.
[ 2024-09-17 | 17:17:04 | pineapple ] Downloading JUMP images |==============
```

Developed a Rust CLI tool for filtering and downloading JUMP images

https://drive.google.com/file/d/1eBEyxdV9U35V1c8Zj5PGrM3rGHRY7Ia-/view?usp=drive_link

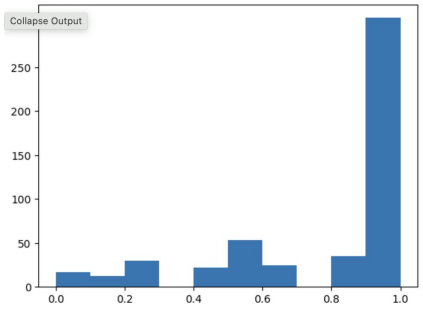
./pineapple download images --compound "KYRVNWMVYQXFEU-UHFFFAOYSA-N"

(2) Features



Iterative Loop (or build on entire dataset)

(3) Index & Query



Single-cell ANN Query Accuracy (nn = 10)
(built on DINOv2 embeddings of JUMP images)

Special recognition to Dr. Niranj Chandrasekaran

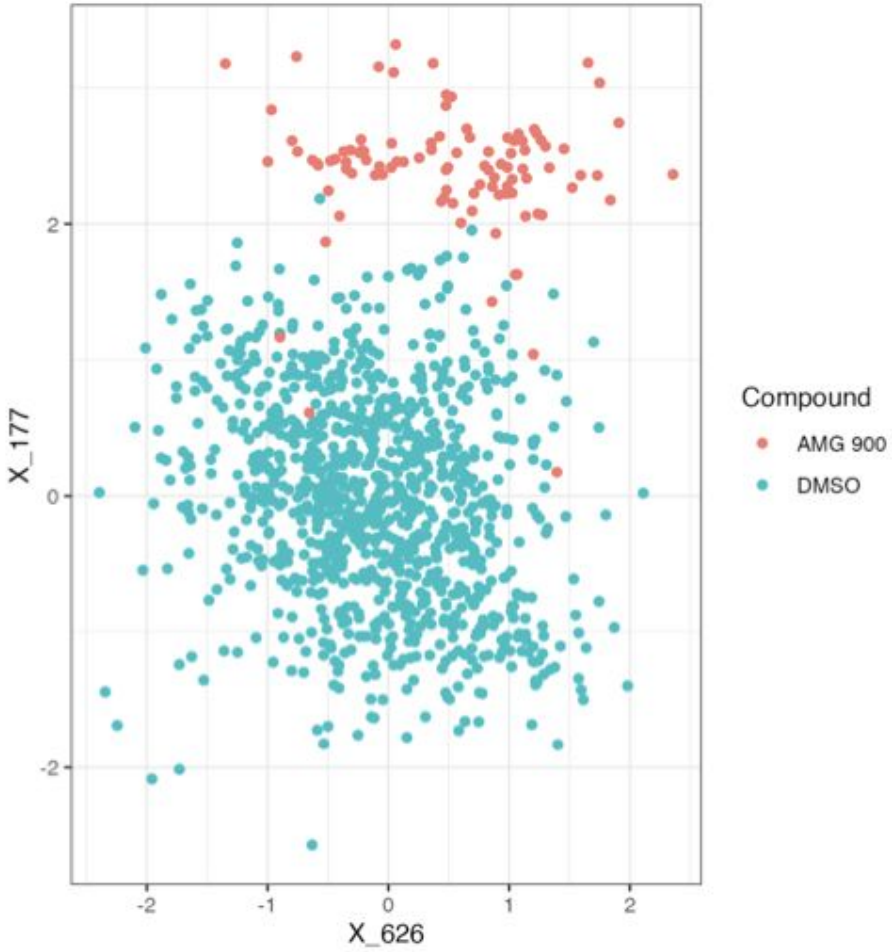
For his instrumental work in coordinating collaborators to make the JUMP-Cell Painting a reality

Team 10

- Idea: can we identify a driving signal from one of the positive controls, fit a model, and observe the effects of query compounds along this axis?
- Proof of concepting on subset of the compound data:
  - 1000 DMSO wells
  - 10,000 query wells
  - 8 x 100 positive control wells

Contributing features

- Visual inspection of top contributing features to LDA generation show ability to separate space without feature transformation



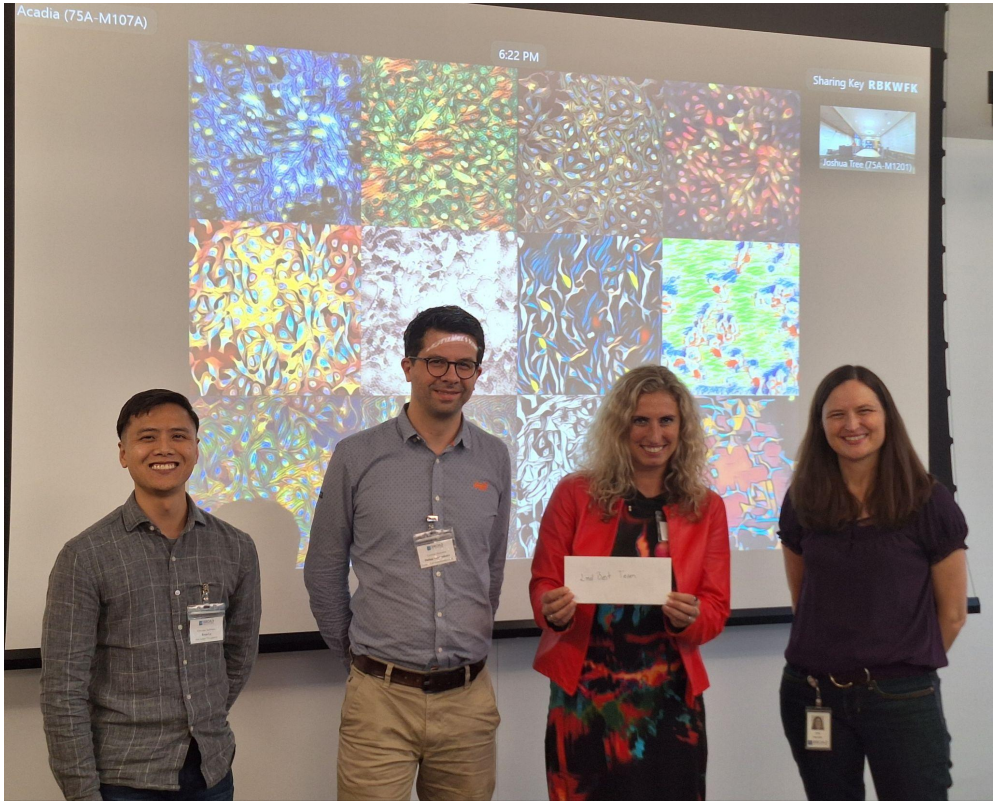
Team 10

- Idea: can we identify a driving signal from one of the positive controls, fit a model, and observe the effects of query compounds along this axis?
- PoC steps:
  - Identified AMG 900 cluster
  - Fit LDA model
  - Projected query compounds
  - Identified shared mechanism of action
- Challenges / limitations
  - File I/O – struggles with working with very large JUMP-CP data
  - Overfitting – LDA model likely overfits and may not generalize as well as hoped to query compounds (instead modelling random noise leading to false positives)

Second place winners: Team 10

- Alzbeta Srovnalova
- Brian Le
- Philipp Mergenthaler
- Mutaamba Maasha

(Dr. Anne Carpenter in the picture)



Team “Batch Breakers”

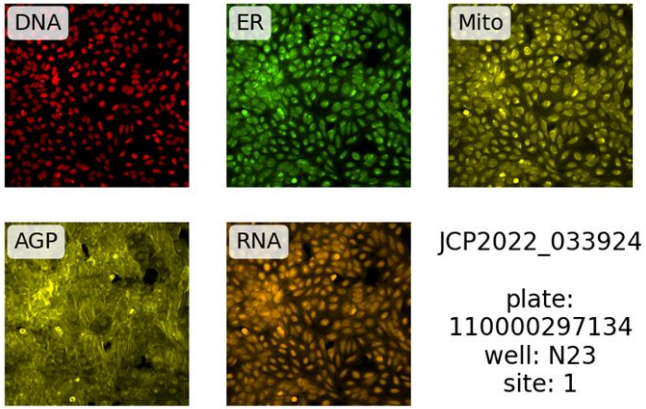
- Project Idea: Evaluate feature importance before batch correction on defined control dataset (1 neg control – 8 pos controls)

- Methods: Random forest feature importance (Gini impurity) and Shapley
  - Random forest: for multi-class
  - Shapley: binary classification on source_6 (runtime)

Note: to interpret the features, we cannot use Harmony batch correction.

Negative Control (JCP2022_033924)

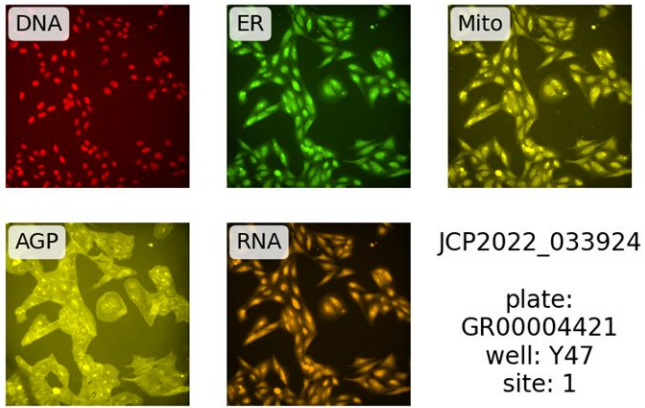
Source_6



JCP2022_033924

plate:
110000297134
well: N23
site: 1

Source_9

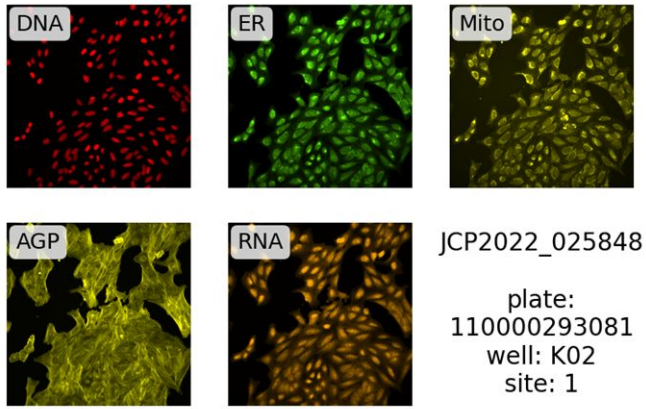


JCP2022_033924

plate:
GR00004421
well: Y47
site: 1

Positive Control (JCP2022_025848)

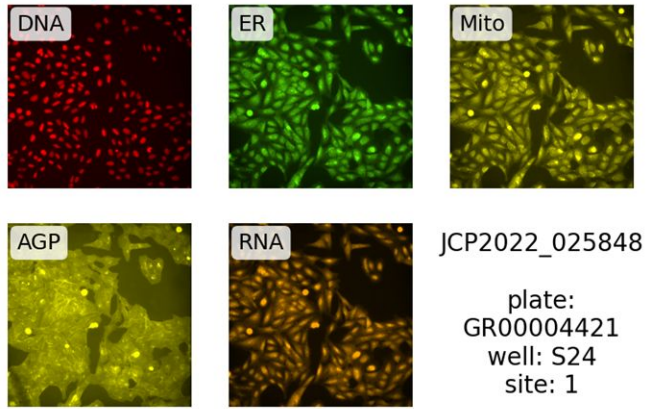
Source_6



JCP2022_025848

plate:
110000293081
well: K02
site: 1

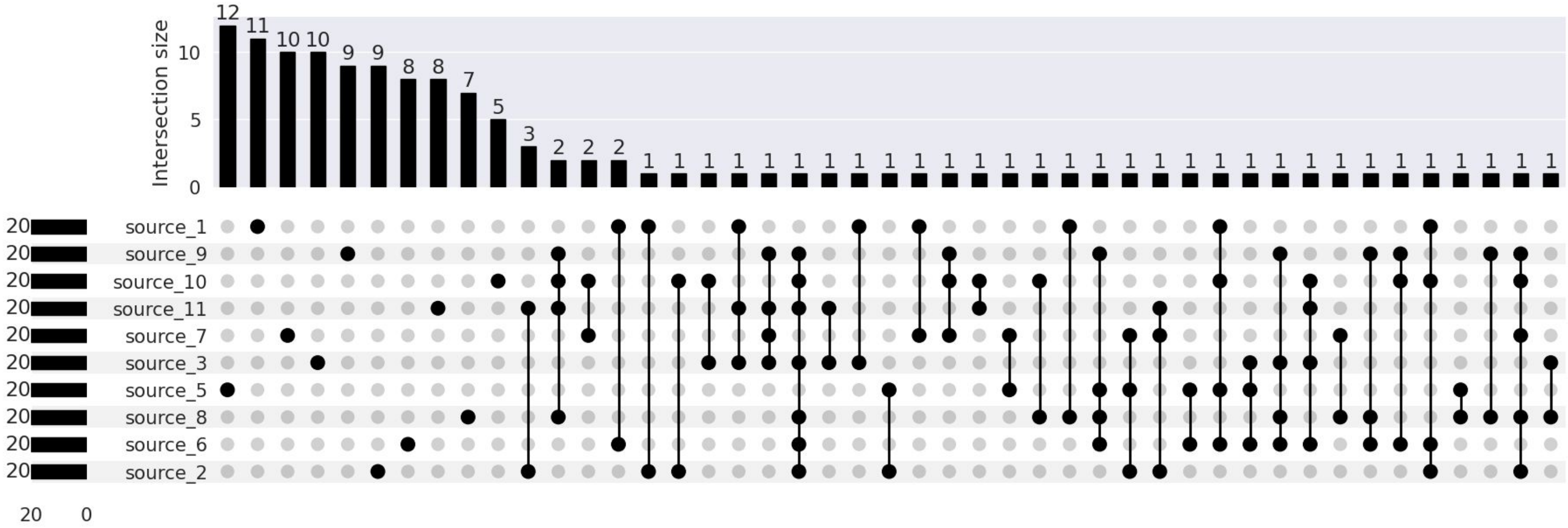
Source_9



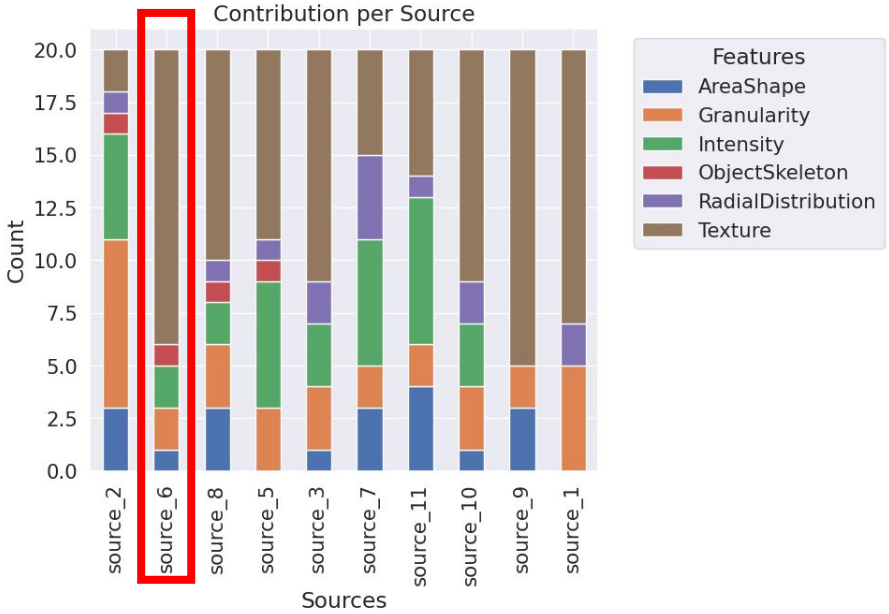
JCP2022_025848

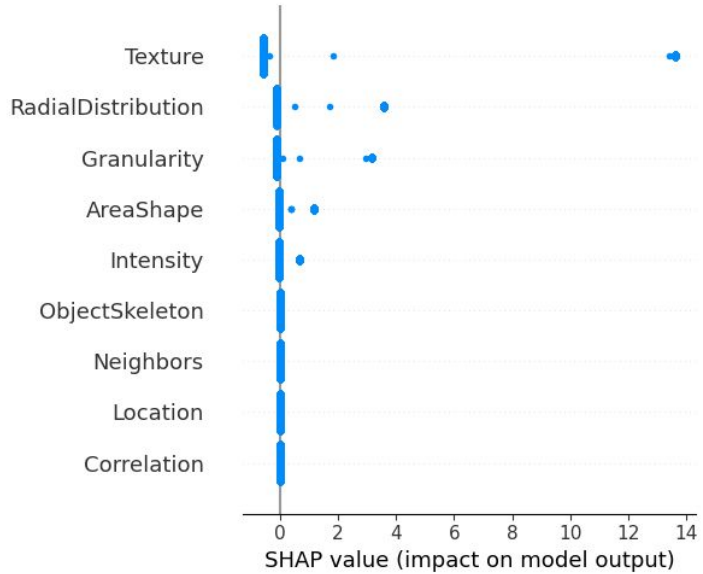
plate:
GR00004421
well: S24
site: 1

UpSet plot – no overlap between important features in the different sources



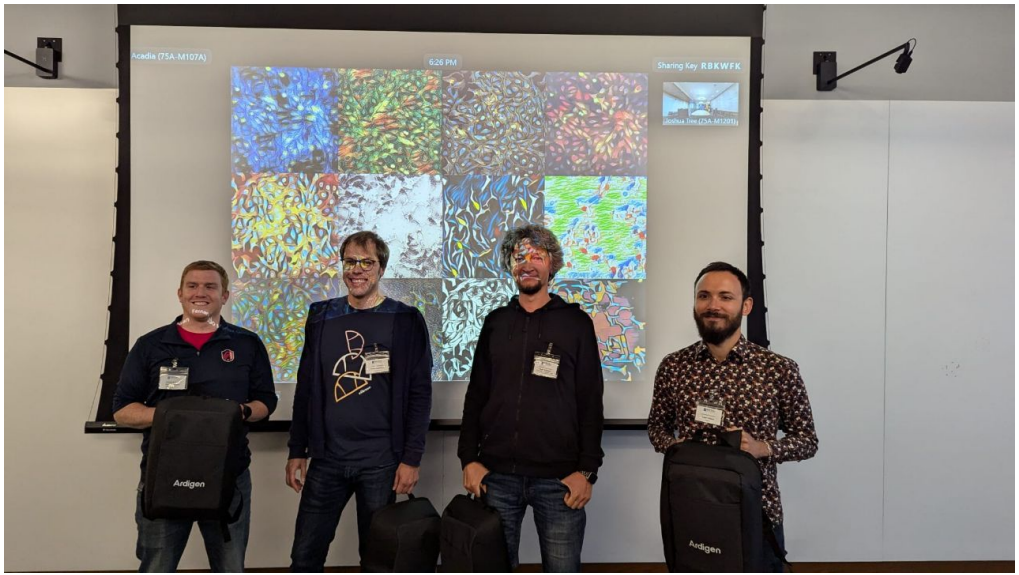
Feature importance consistent over both methods for binary classification on source_6





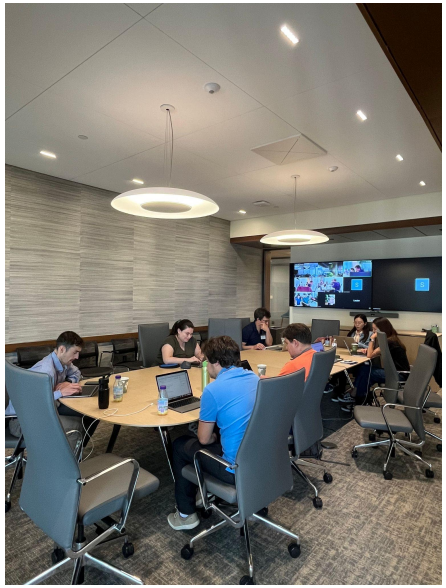
First place winners: “Batch Breakers”

- Taras Redchuk
- Daniel Siegsimund
- Loan Vulliard
- Nathan Carter



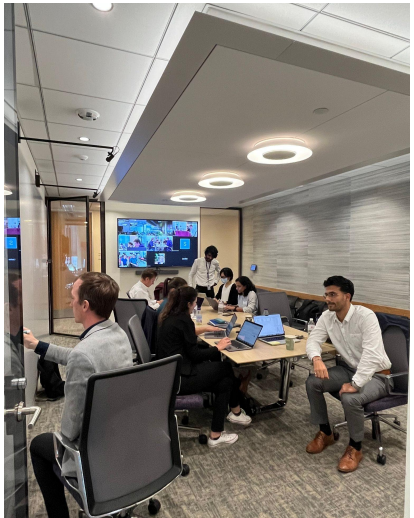
Acknowledgements

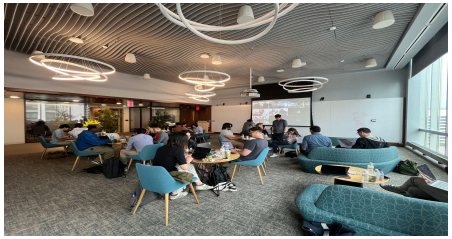
CytoData

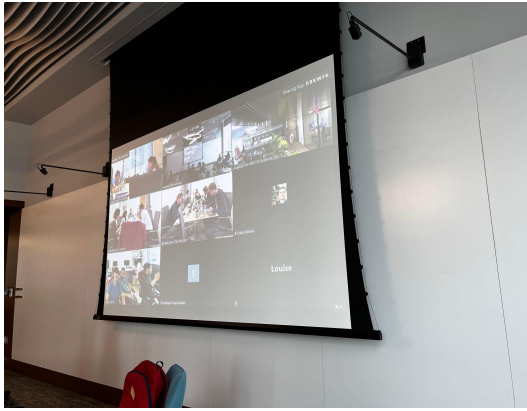


Slides and materials used will be uploaded to this github repo!

github.com/afermg/2024_09_hackathon






Louise