Final presentations

Team 1

Tiling or segmentation, that is the question

Dmitry Isaev, Martin Cottet, Félix Lavoie-Pérusse

Context

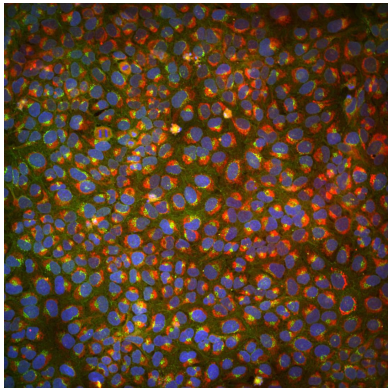
Chosen Track: tool/method
Question: How to approach image analysis of live cells: comparing the outcome of image tiling vs segmentation

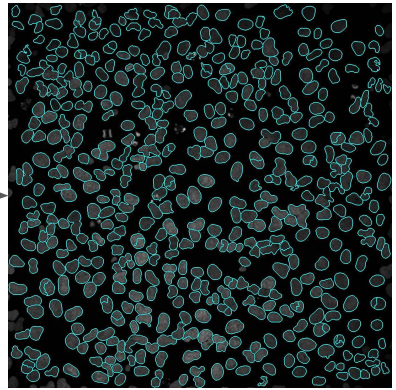
Relevance

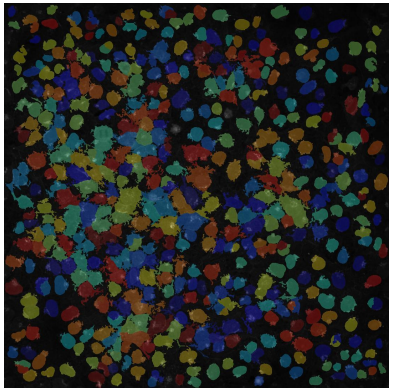
- Analyzing cell images requires cropping of images in the form of segmentation of cells or tiling of images
- Live cell segmentation with a thresholding approach that uses nuclear masks requires the use of nuclear dyes that are mostly toxic to cells
- It leaves two alternatives:
  - Segmentation with label-free images using ML models
  - Obviate cell segmentation and instead perform tiling of images

Objective: compare which approach between cell segmentation and tiling of images performs best using mAP score

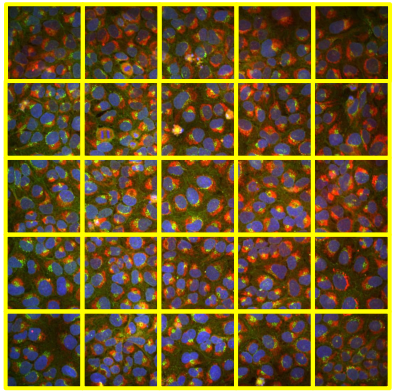
Cell segmentation vs. tiling







X objects =
single cells



Objects = tiles

Smaller tiles
-> cell-like objects
-> cell-like features?

Methodology

1. Use the Cell Painting dataset as it provides the ground truth for DAPI staining
2. Segment nuclei using simple Otsu segmentation from DAPI staining
3. Feed images with these centers to the feature extractor.
4. As a comparison, feed tiles of each image into feature extractor

   ======= stopped here =====

5. Compare resulting features with mean average precision method on a subset of compounds.

Next steps

- Is it possible to reduce the size of the tiles (224x224), while still using a pretrained image analysis model
- Can (smaller) tiles be injected as objects into CellProfiler to run standard analysis pipelines?
- Compare extracted features from Cell Painting assay using standard segmentation to the ones extracted using tiles (evaluating robustness of phenotypic profiling using mAP)
- Can this tiling method be transposed to 3D (voxels) for “segmentation” of spheroids/organoids?

Team 3

Cytodata 2024 - JUMP

Team Oktob- R -fest

Florian Heigwer
Daron Hakimian
Martina Zowada
Cornelia Redel-Smirnov

Our question

Major aim: cost reduction

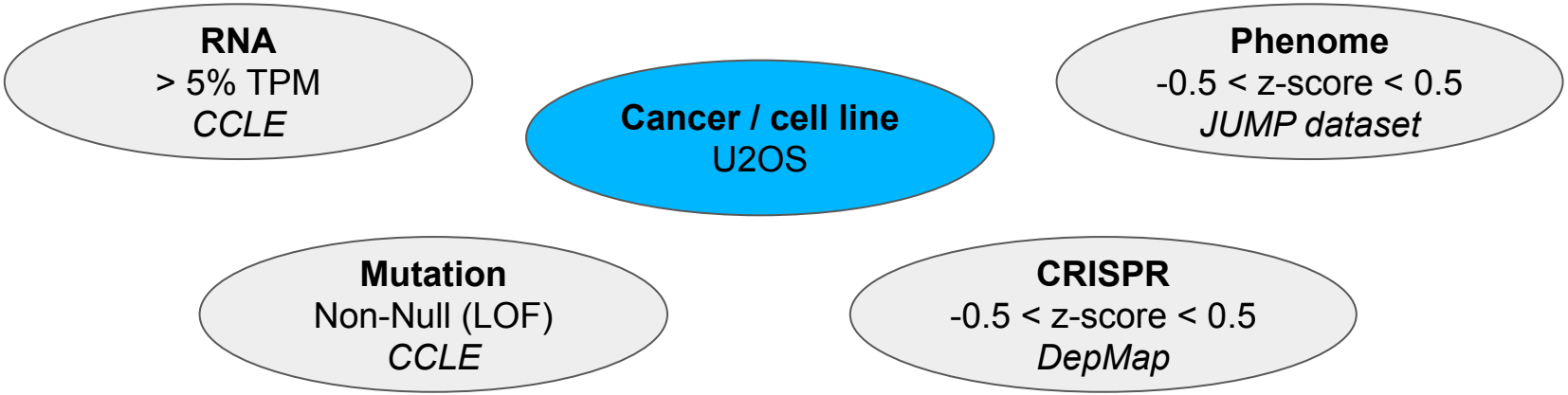
- less material
- less data storage
- less time
- …

How can we pre-select the compounds to use and can we reduce the amount of channels to use?

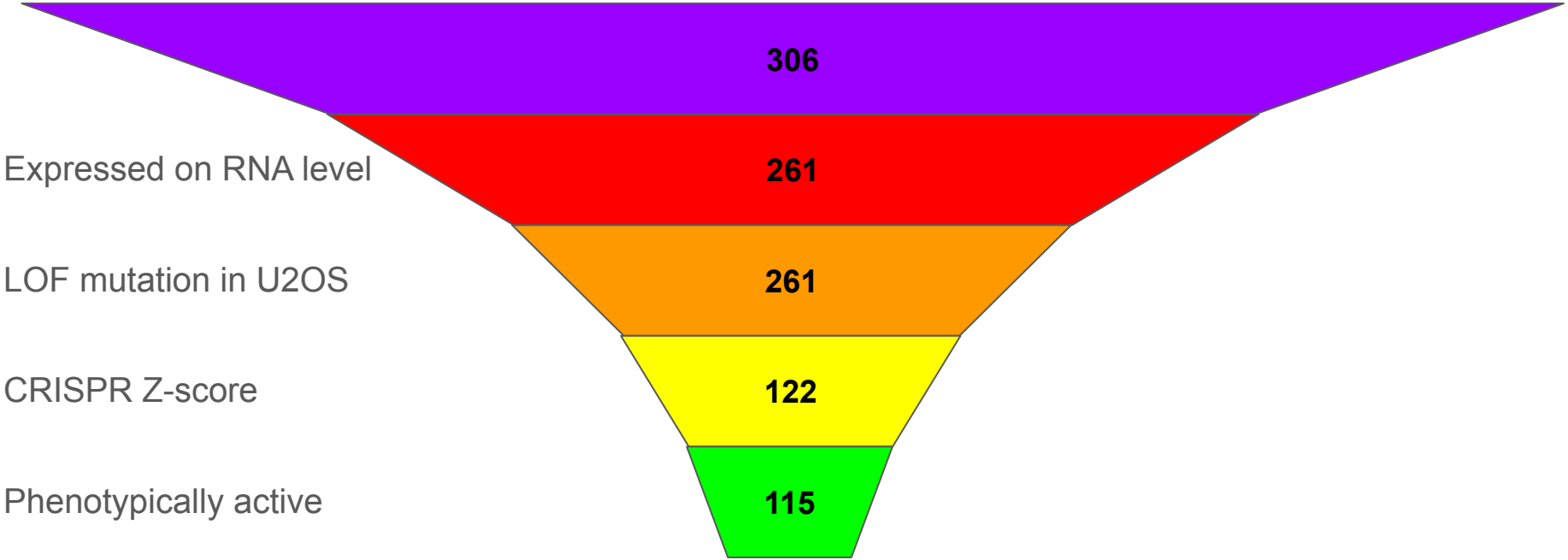
Pre-filtering / pre-selection of compounds

How can we reduce the amount of compounds?

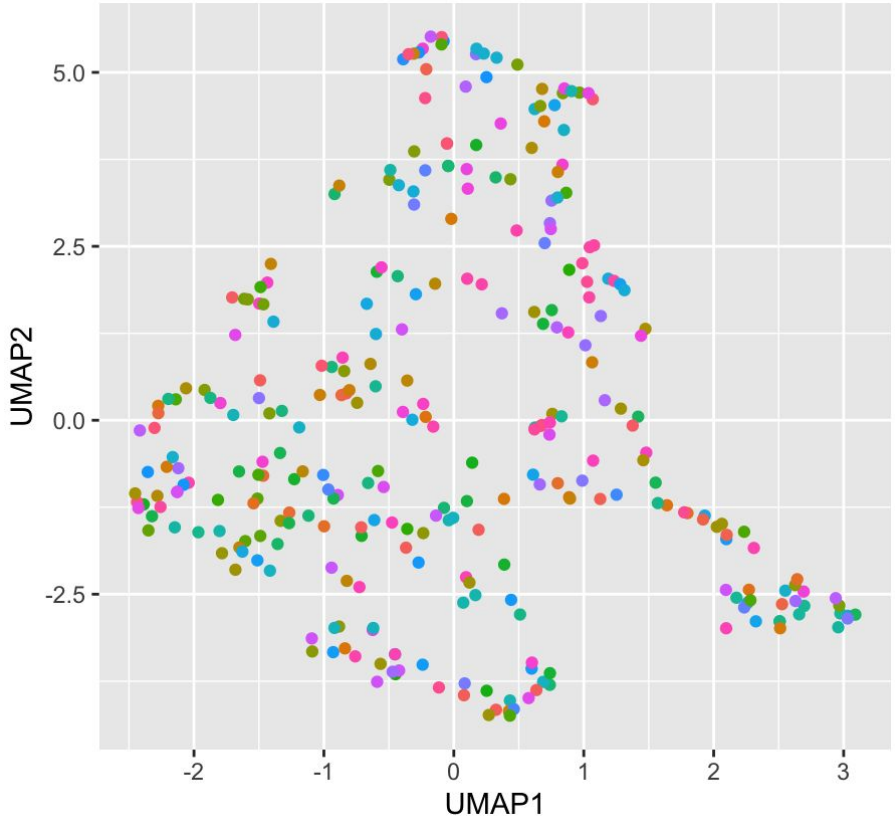
Test dataset: JUMP-Target-2-Compound plate (306 drugs)



The ChoosR Algorithm

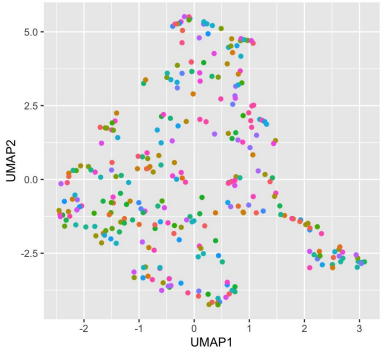


Mode of action could be clearer

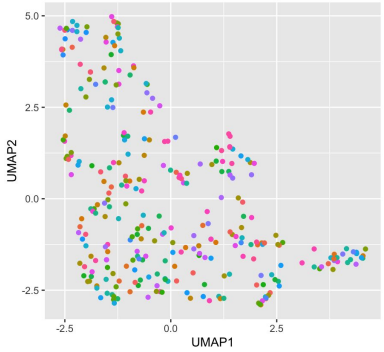


Effects of channel reduction on clustering

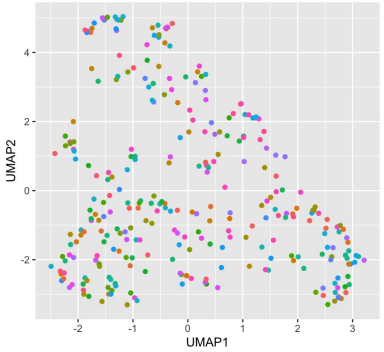
DNA+Mito+ER+RNA+AGP



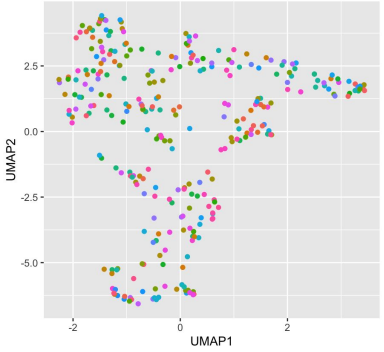
DNA+Mito+ER+RNA



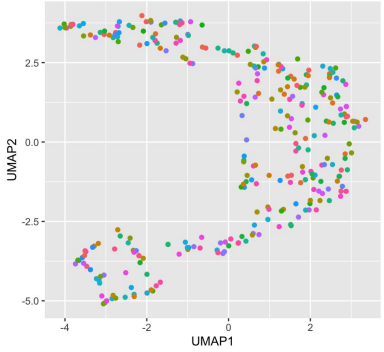
DNA+Mito+ER

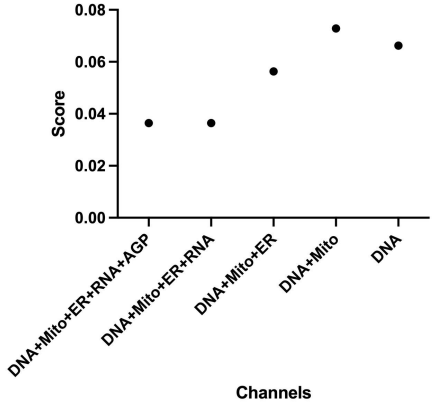


DNA+Mito



DNA only





Team 2

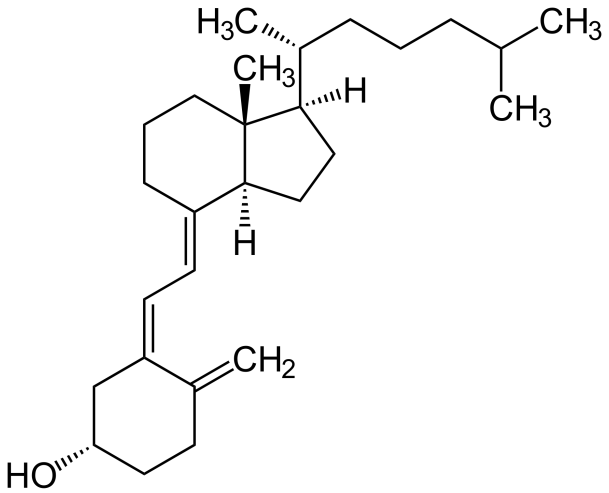
Osheen Sharma

Erik Serrano

Jesko Wagner

Question

What functional groups predict cell injuries?



Article

https://doi.org/10.1038/s41467-023-36829-x

Reference compounds for characterizing cellular injury in high-content cellular morphology assays

Received: 1 August 2022

Accepted: 20 February 2023

Published online: 13 March 2023

Check for updates

Jayme L. Dahlin[1,2,13], Bruce K. Hua[2,13], Beth E. Zucconi[3], Shawn D. Nelson Jr[4], Shantanu Singh[5], Anne E. Carpenter[5], Jonathan H. Shrimp[6], Evelyne Lima-Fernandes[7], Mathias J. Wawer[2], Lawrence P. W. Chung[2], Ayushi Agrawal[2], Mary O'Reilly[8], Dalia Barsyte-Lovejoy[7], Magdalena Szewczyk[7], Fengling Li[7], Parnian Lak[9], Matthew Cuellar[10], Philip A. Cole[3], Jordan L. Meier[6], Tim Thomas[11], Jonathan B. Baell[12], Peter J. Brown[7], Michael A. Walters[10], Paul A. Clemons[2], Stuart L. Schreiber[2] & Bridget K. Wagner[2]

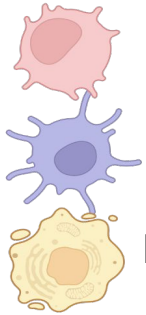
Idea

Predict compounds causing cell injuries, check chemistry for why ?

Method

Input

Labels of cellular injuries



mTOR

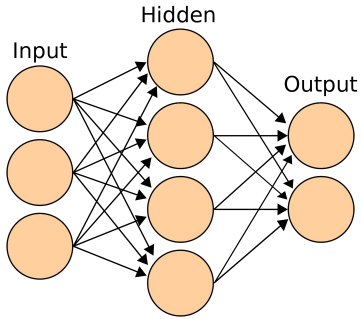
HDAC

Kinase

Well-level morphological features

Well	Compound	Cell area	Intensity	…
1	DMSO	120	438	…
2	Paclitaxel	60	746	…

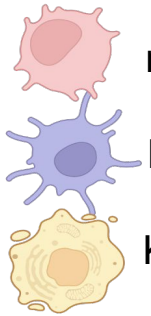
Model



Method

Input

Labels of c



Well-level morp

Well	Compound	C	Term Source 1 REF	Term Source 1 Accession	Characteristics [Cell Line]	Term Source 2 REF	Term Source 2 Accession	Experimental Condition [Treatment time (h)]	...	Cytoplasm_AreaShape
1	DMSO	120		438	…					
2	Paclitaxel	60		746	…					

Model

```
cts found within the both
a,
s,
= 'Negative'",
or",
```

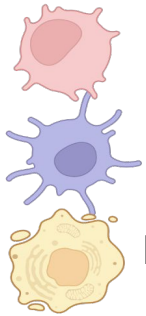
Intentando volver a conectar en 2 segundos...

Recargar ventana | Descartar | Volver a conectar ahora

Method

Input

Labels of cellular injuries



mTOR

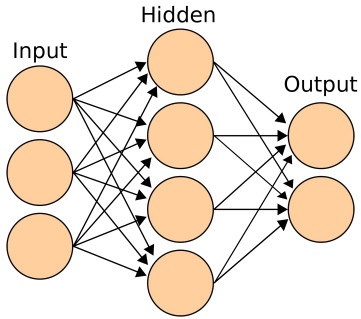
HDAC

Kinase

Well-level morphological features

Well	Compound	Cell area	Intensity	…
1	DMSO	120	438	…
2	Paclitaxel	60	746	…

Model

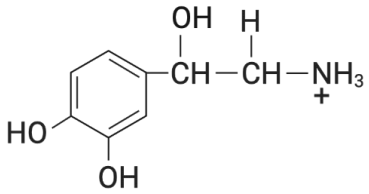


JUMP CellProfiler features

Well	Compound	Cell area	Intensity	…
1	Caffeine	37	923	…
2	Choline	421	87	…

Compound	Injury
Caffeine	Cytoskeleton
Choline	mTor

Over-representation of functional groups



Osheen Sharma

Erik Serrano

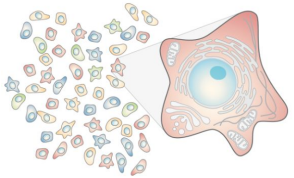
Jesko Wagner

Team 4

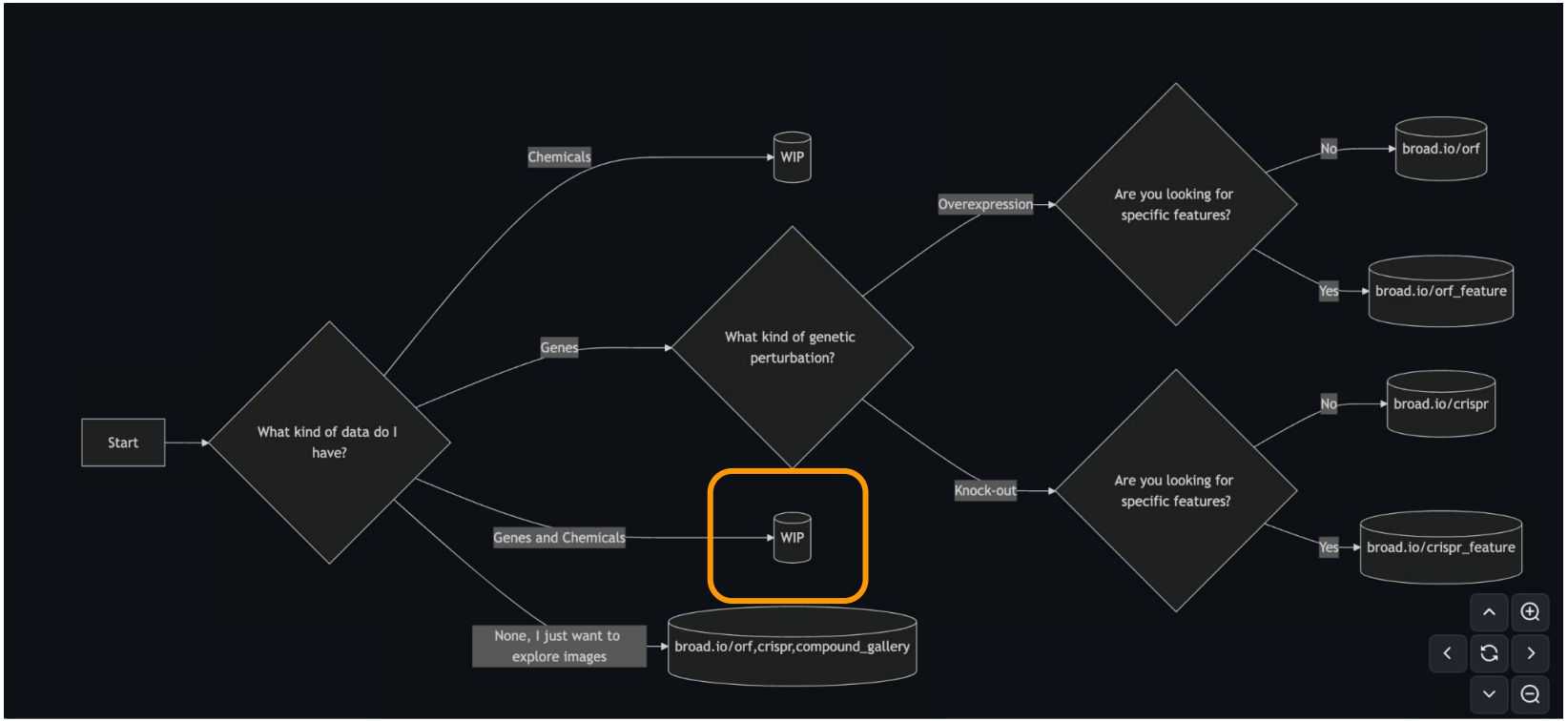
How to link genetic and chemical perturbations?

Csaba Molnar, Jingzhe Ma, Louise Morlot

Introduction



JUMP-Cell Painting Consortium
Joint Undertaking in Morphological Profiling



Which datasets?

MOTIVE: A Drug-Target Interaction Graph For Inductive Link Prediction

John Arevalo* **Ellen Su*** **Anne E. Carpenter** **Shantanu Singh**
Broad Institute of MIT and Harvard
{jarevalo, suellen, anne, shantanu}@broadinstitute.org

Abstract

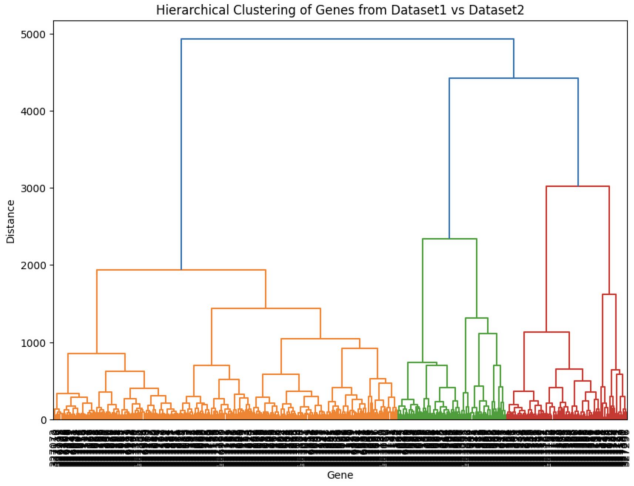
Drug-target interaction (DTI) prediction is crucial for identifying new therapeutics and detecting mechanisms of action. While structure-based methods accurately model physical interactions between a drug and its protein target, cell-based assays such as Cell Painting can better capture complex DTI interactions. This paper introduces MOTIVE, a **M**orphological c**O**mpound **T**arget **I**nteraction **Graph** dataset that comprises Cell Painting features for 11,000 genes and 3,600 compounds along with their relationships extracted from seven publicly available databases. We provide random, cold-source (new drugs), and cold-target (new genes) data splits to enable rigorous evaluation under realistic use cases. Our benchmark results show that graph neural networks that use Cell Painting features consistently outperform those that learn from graph structure alone, feature-based models, and topological heuristics. MOTIVE accelerates both graph ML research and drug discovery by promoting the development of more reliable DTI prediction models. MOTIVE resources are available at https://github.com/carpenter-singh-lab/motive.

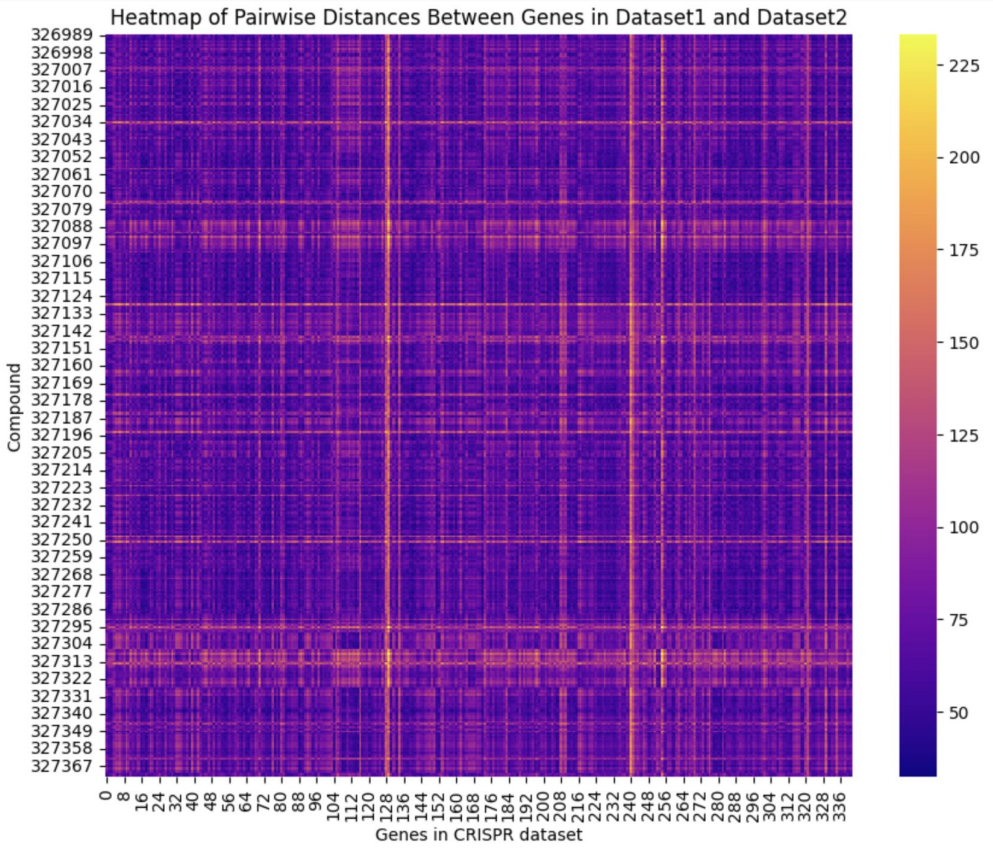
- Cell Profiler features at the well level for genetic and chemical perturbations
- > 11000 genes
- > 3000 compounds

- Sampled the data
  - one source, one plate
  - filtering based on ground truth data

Results

- Calculate the euclidean distances between each gene and each compound
- Hierarchical clustering on distances

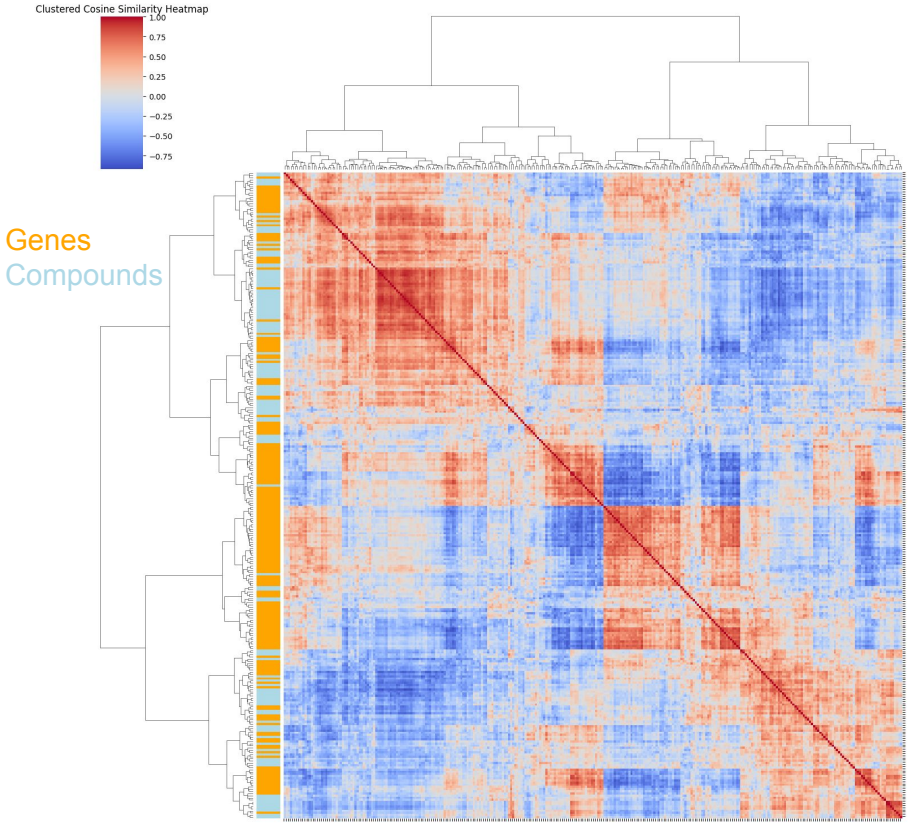




Reverse approach - sampling by know interactions

- ➢ 1000 randomly selected compound gene pairs
- ➢ Select shared features
- ➢ Aggregating features
- ➢ Calculating cosine similarity
- ➢ Clustering

- no correction for batch effect



Future directions

- Extract features from raw images (instead of CellProfiler features)
- No sampling of the data
- Use the ground truth to set distance threshold
- Use positive and negative controls
- Contrastive learning with positive interactions and predict random pairs
- Align feature sets into a shared latent space

Team 5

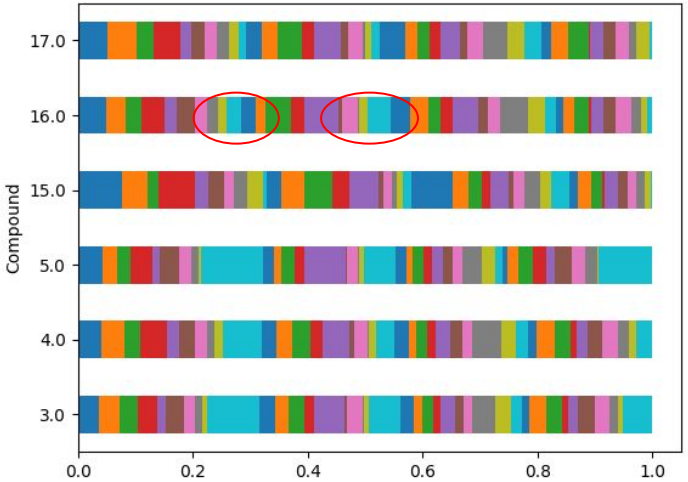
Cytodata team: Zoe Wefers, Josiane Bourque, Ibrahim Bilem

Scope of the study:

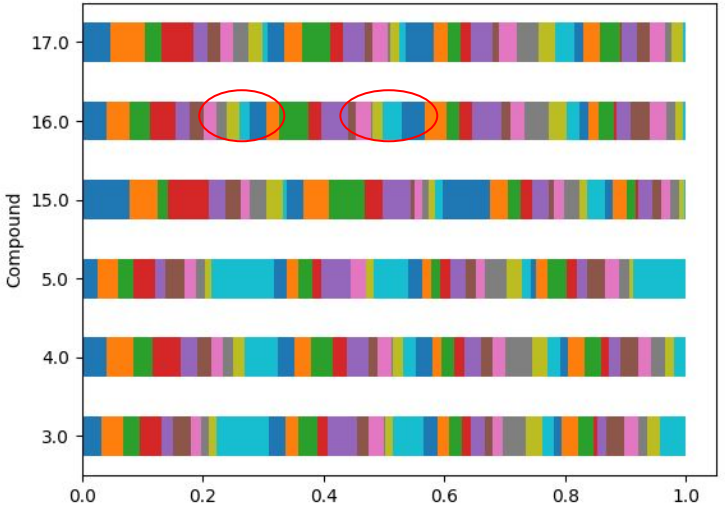
In this study, a small screen of 16 compounds was conducted, using ChromaLive, which is a non-toxic dye with high-density biological information, allowing for Live cell phenotypic profiling.

Biological question:

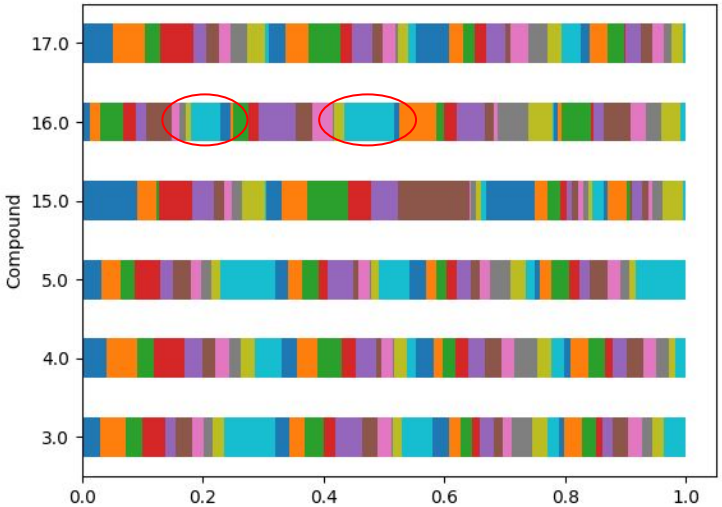
Could Live kinetic imaging help capture subtle and transient phenotypic changes at the single-cell level, across time?



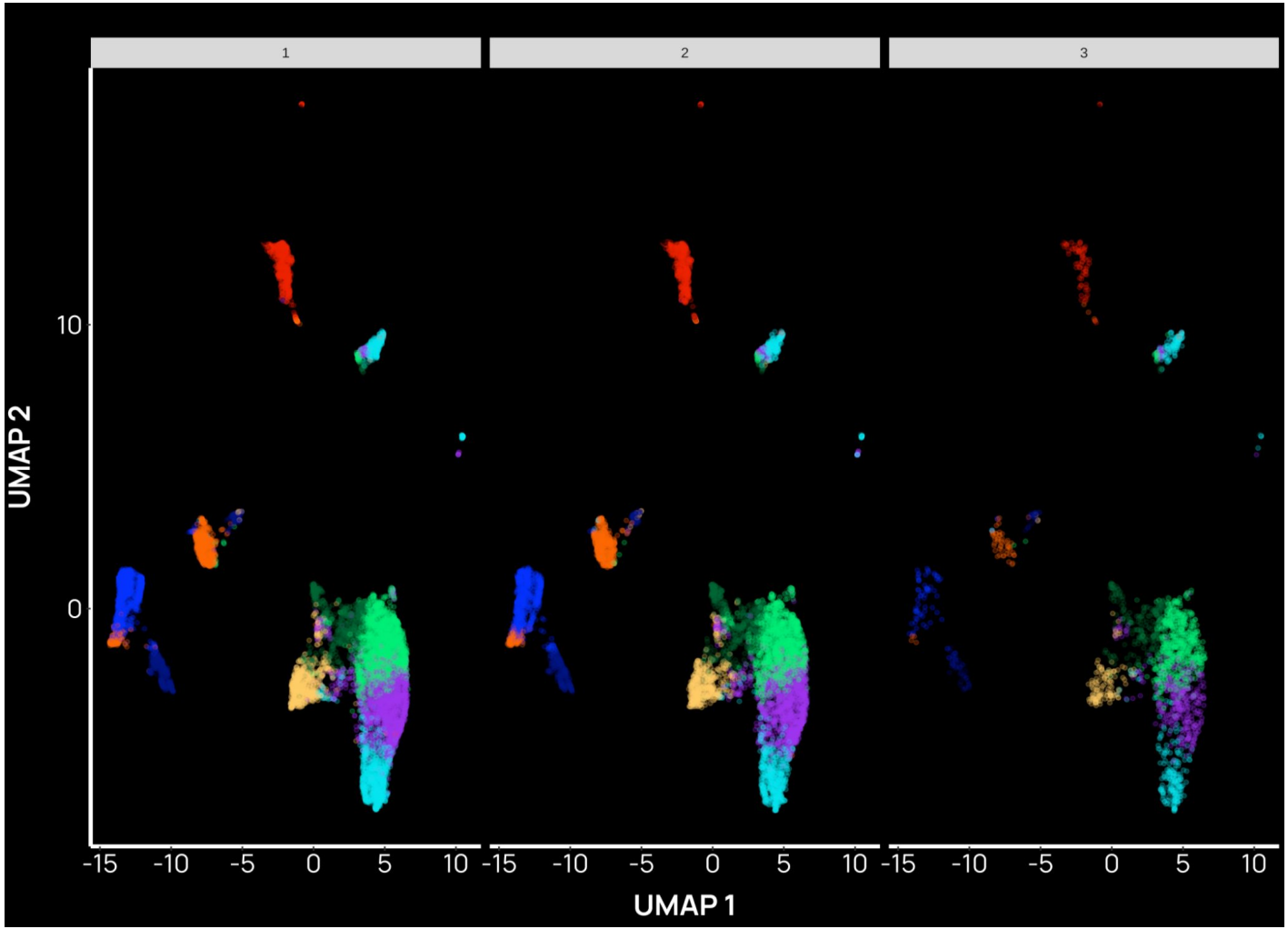
3 hrs



6 hrs



9 hrs


1
2
3
10
0
UMAP 2
-15
-10
-5
0
5
10
UMAP 1

Team 6

Exploring JUMP compound data

Resource | Published: 18 May 2020

Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker

Miquel Duran-Frigola, Eduardo Pauls, Oriol Guitart-Pla, Martino Bertoni, Víctor Alcalde, David Amat, Teresa Juan-Blanco & Patrick Aloy

Nature Biotechnology **38**, 1087–1096 (2020) | Cite this article

A Publisher Correction to this article was published on 21 May 2020

This article has been updated

Abstract

Small molecules are usually compared by their chemical structure, but there is no unified analytic framework for representing and comparing their biological activity. We present the Chemical Checker (CC), which provides processed, harmonized and integrated bioactivity data on ~800,000 small molecules. The CC divides data into five levels of increasing complexity, from the chemical properties of compounds to their clinical outcomes. In between, it includes targets, off-targets, networks and cell-level information, such as omics data, growth inhibition and morphology. Bioactivity data are expressed in a vector format, extending the concept of chemical similarity to similarity between bioactivity signatures. We show how CC signatures can aid drug discovery tasks, including target identification and library characterization. We also demonstrate the discovery of compounds that reverse and mimic biological signatures of disease models and genetic perturbations in cases that could not be addressed using chemical information alone. Overall, the CC signatures facilitate the conversion of bioactivity data to a format that is readily amenable to machine learning methods.

Article | Open access | Published: 24 June 2021

Bioactivity descriptors for uncharacterized chemical compounds

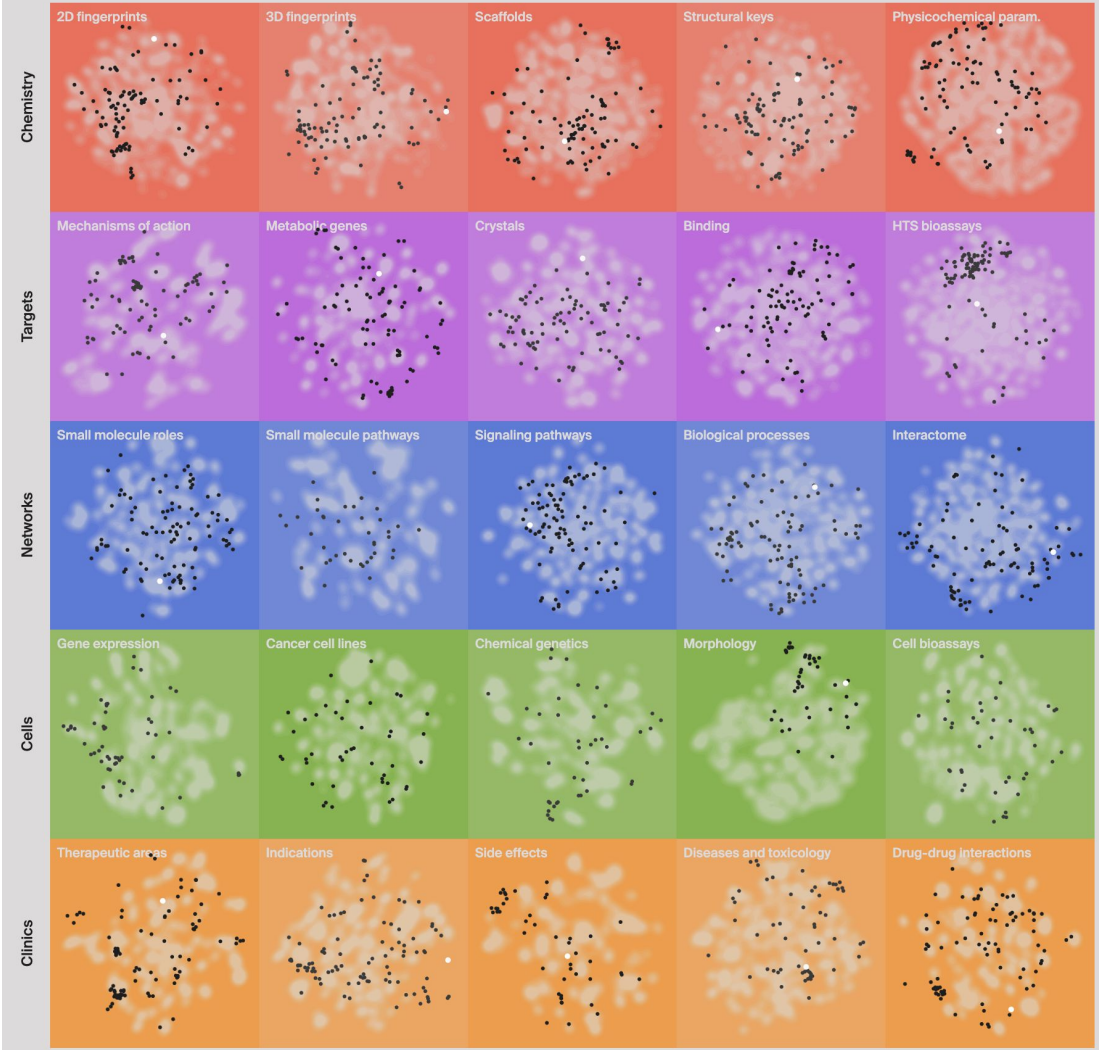
Martino Bertoni, Miquel Duran-Frigola, Pau Badia-i-Mompel, Eduardo Pauls, Modesto Orozco-Ruiz, Oriol Guitart-Pla, Víctor Alcalde, Víctor M. Diaz, Antoni Berenguer-Llergo, Isabelle Brun-Heath, Núria Villegas, Antonio García de Herreros & Patrick Aloy

Nature Communications **12**, Article number: 3932 (2021) | Cite this article

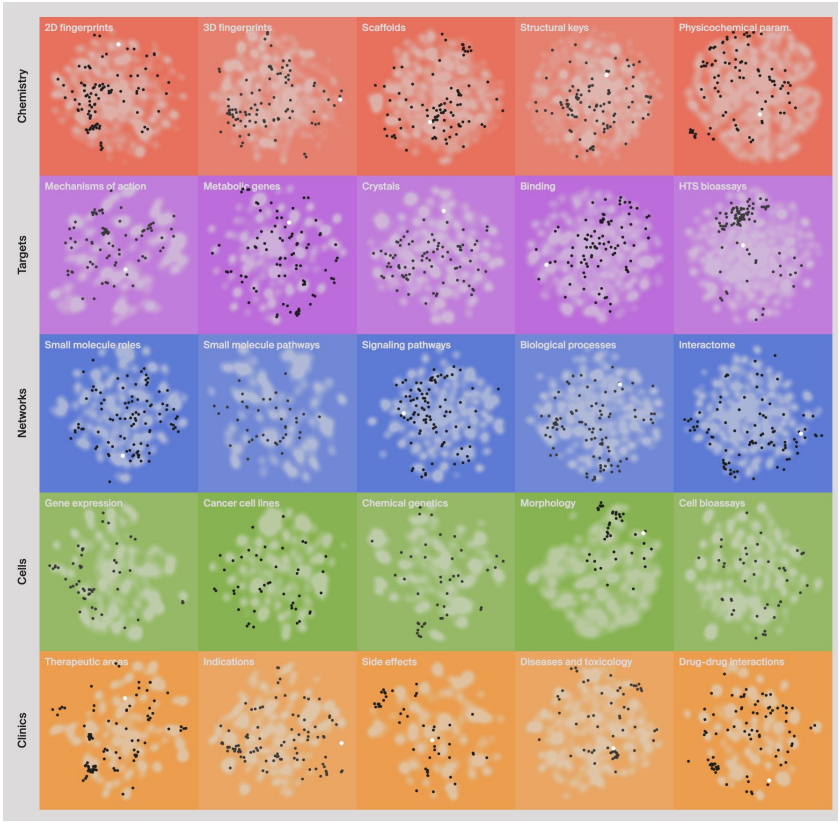
Abstract

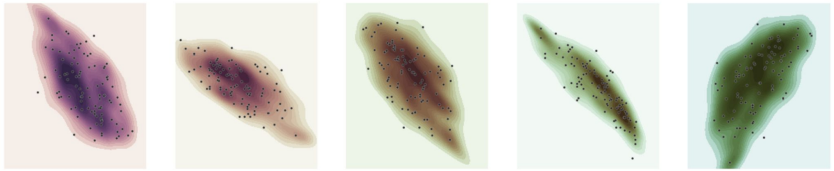
Chemical descriptors encode the physicochemical and structural properties of small molecules, and they are at the core of chemoinformatics. The broad release of bioactivity data has prompted enriched representations of compounds, reaching beyond chemical structures and capturing their known biological properties. Unfortunately, bioactivity descriptors are not available for most small molecules, which limits their applicability to a few thousand well characterized compounds. Here we present a collection of deep neural networks able to infer bioactivity signatures for any compound of interest, even when little or no experimental information is available for them. Our signaturizers relate to bioactivities of 25 different types (including target profiles, cellular response and clinical outcomes) and can be used as drop-in replacements for chemical descriptors in day-to-day chemoinformatics tasks. Indeed, we illustrate how inferred bioactivity signatures are useful to navigate the chemical space in a biologically relevant manner, unveiling higher-order organization in natural product collections, and to enrich mostly uncharacterized chemical libraries for activity against the drug-orphan target Snail1. Moreover, we implement a battery of signature-activity relationship (SigAR) models and show a substantial improvement in performance, with respect to chemistry-based classifiers, across a series of biophysics and physiology activity prediction benchmarks.

Use chemical similarity to explore JUMP compound Data in a richer context


Chemistry
2D fingerprints
3D fingerprints
Scaffolds
Structural keys
Physicochemical param.
Targets
Mechanisms of action
Metabolic genes
Crystals
Binding
HTS bioassays
Networks
Small molecule roles
Small molecule pathways
Signaling pathways
Biological processes
Interactome
Cells
Gene expression
Cancer cell lines
Chemical genetics
Morphology
Cell bioassays
Clinics
Therapeutic areas
Indications
Side effects
Diseases and toxicology
Drug-drug interactions

JUMP channels





But then bugs happened!

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…

Team 8

Enhancing JUMP cell painting exploration with quick and approximate methods

Tom Ouellette, Yiran Shao, Yuxin (Zoe) Zhi

(1) Download

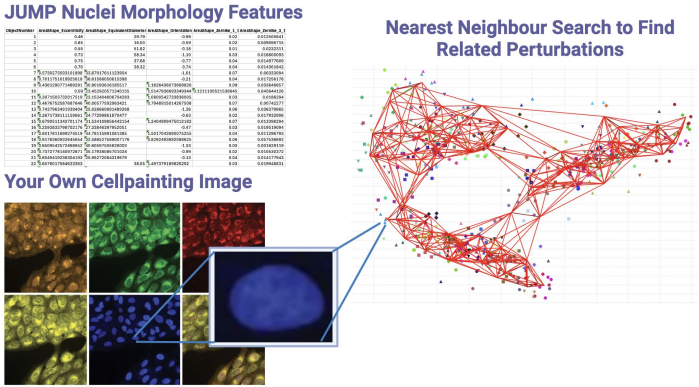
```
[ 2024-09-17 | 17:17:02 | pineapple ] Initializing JUMP download
[ 2024-09-17 | 17:17:04 | pineapple ] Detected 166 samples for downloading.
[ 2024-09-17 | 17:17:04 | pineapple ] Downloading JUMP images |==============
```

Developed a Rust CLI tool for filtering and downloading JUMP images

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

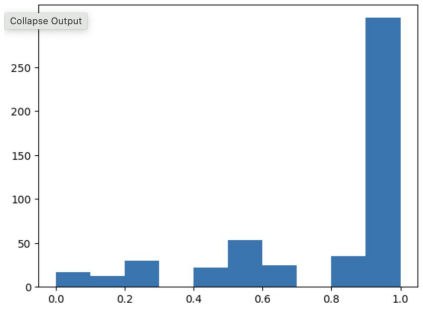
./pineapple download images --compound "KYRVNWMVYQXFEU-UHFFFAOYSA-N"

(2) Features



Iterative Loop (or build on entire dataset)

(3) Index & Query



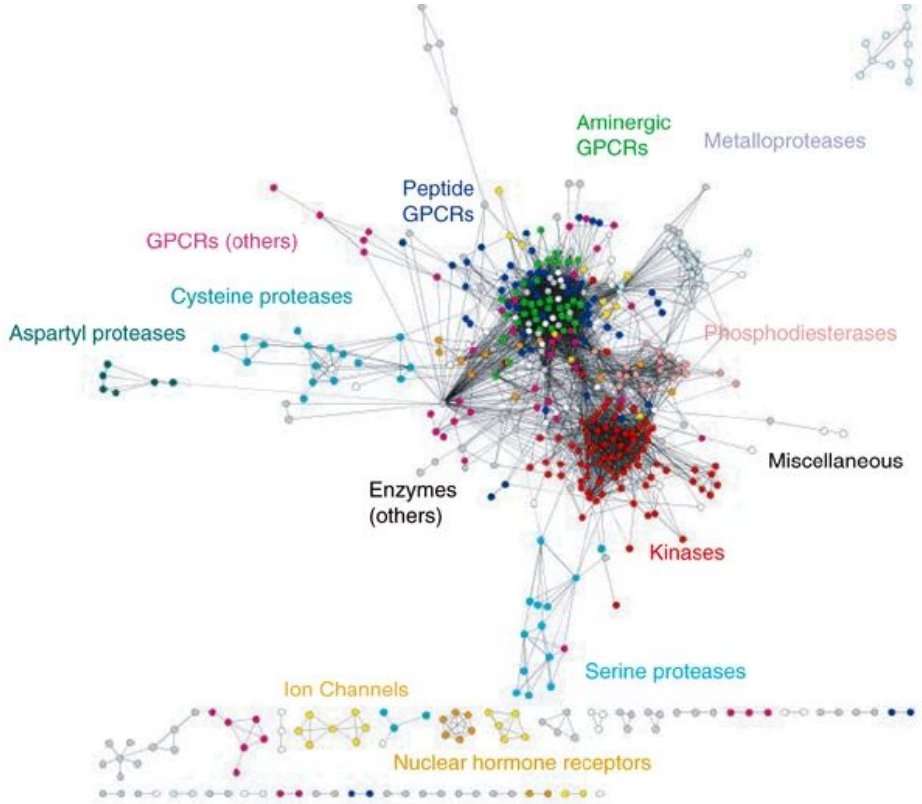
Single-cell ANN Query Accuracy (nn = 10)
(built on DINOv2 embeddings of JUMP images)

Team 9

CytoData Hackathon

Team 9: Sarolt Magyari, Ta Dinh Nguyen Vo, and Marcus Chin

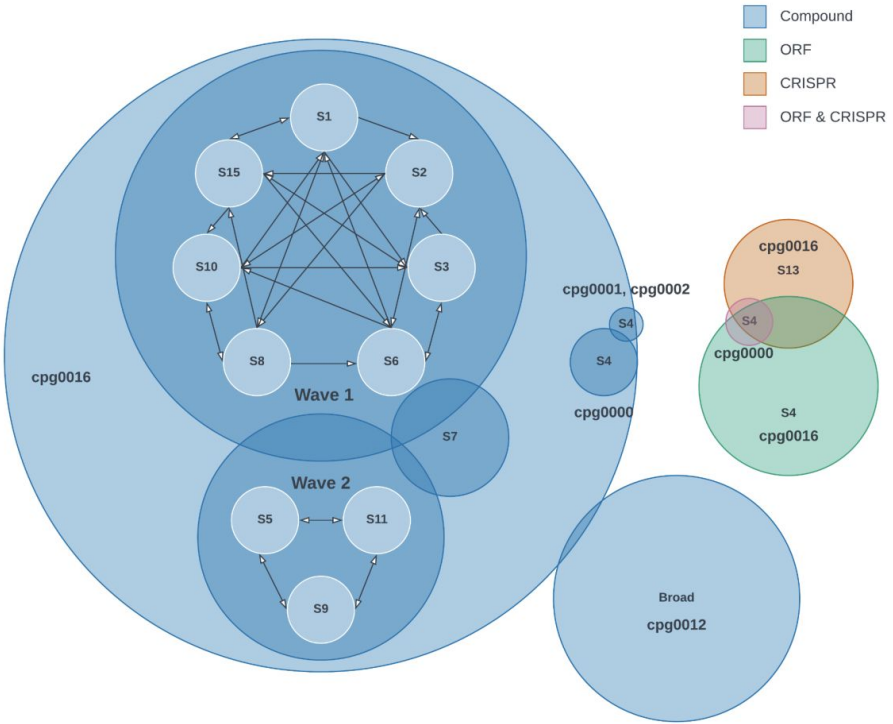
Polypharmacology in drug discovery



Key Scientific Question:

To determine whether compound mode of action (i.e. phenotypic signature) is generalizable across different cell types.

Approach: Harnessing CP-JUMP Pilot dataset



Imported pilot cell painting data was preprocessed before with pycytominer.

CP0000 (pilot) with U2OS and A549 cells

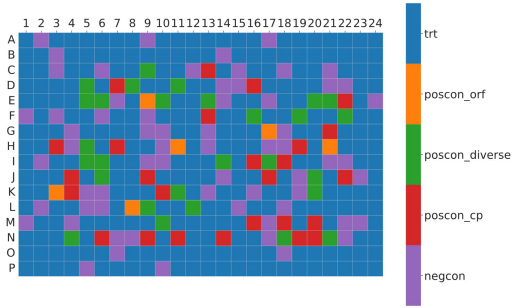
https://www.biorxiv.org/content/10.1101/2023.03.23.534023v2

Methodology

Measure differences for all features from control

Group the differences

Compare cell types

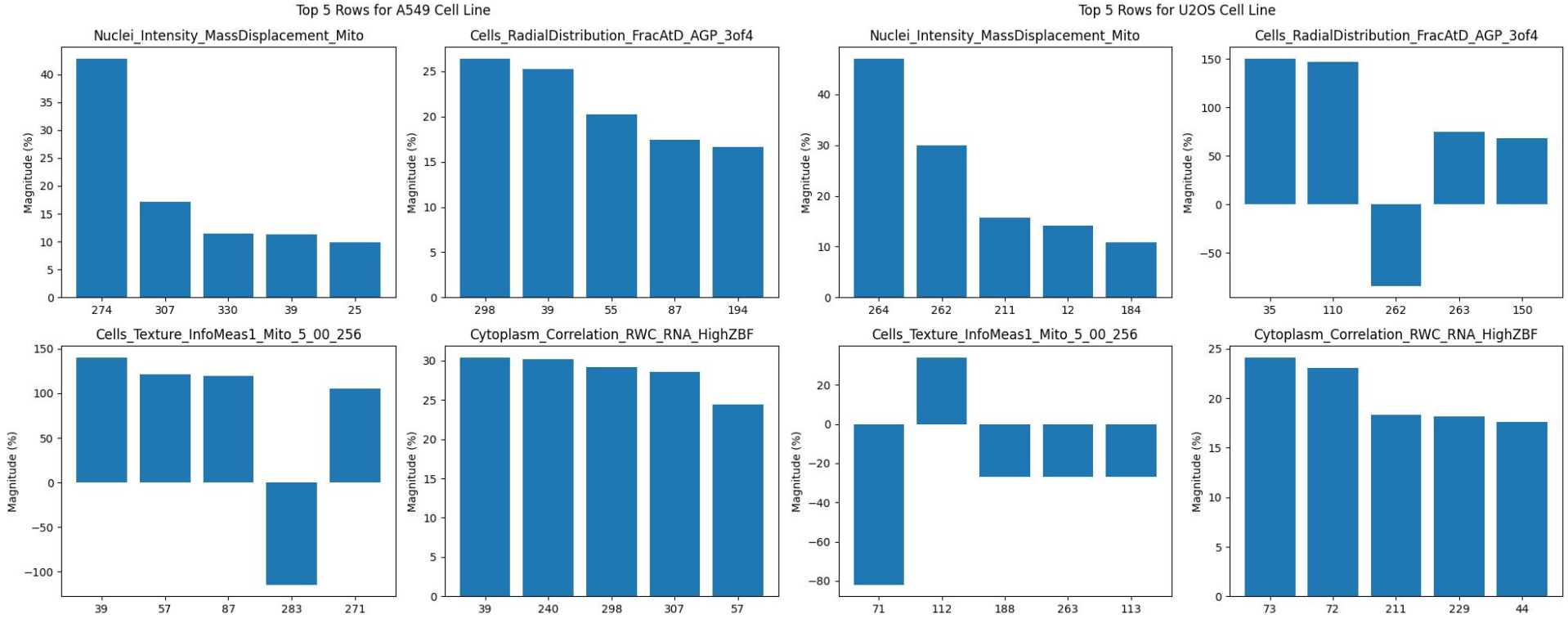



```
a549_fold_change = (a549_values - a549_control_mean) / a549_control_mean
u2os_fold_change = (u2os_values - u2os_control_mean) / u2os_control_mean
```

	Cells_AreaShape_BoundingBoxMaximum_Y	Cells_AreaShape_Compactness	Cells_AreaShape_Eccentricity	Cells_AreaShape_Extent	Cells_AreaShape_Solidity	Cells_
0	-7.368060	0.747378	-1.814903	0.840129	0.933566	
1	-37.148959	-0.143933	-2.365008	-0.678651	-0.029367	
2	1.099478	4.604615	-2.981228	-2.702426	-12.888896	
3	-9.497288	0.722248	-2.169707	-0.575357	-3.679001	
4	-52.631631	15.502080	-0.499170	-4.542555	-41.915153	
...	...	...	...	...	...	
379	34.322556	0.133016	-2.561910	-1.202757	0.234656	
380	26.291979	-3.366796	-1.541571	1.796539	6.649020	
381	38.392593	-2.233734	-0.877767	-0.120119	4.552128	
382	-8.641547	-0.179191	-1.882808	1.815651	2.486632	
383	15.596663	-1.644563	-3.149315	0.075001	10.810811	

384 rows × 904 columns

Examples compounds highest feature FC compared to control



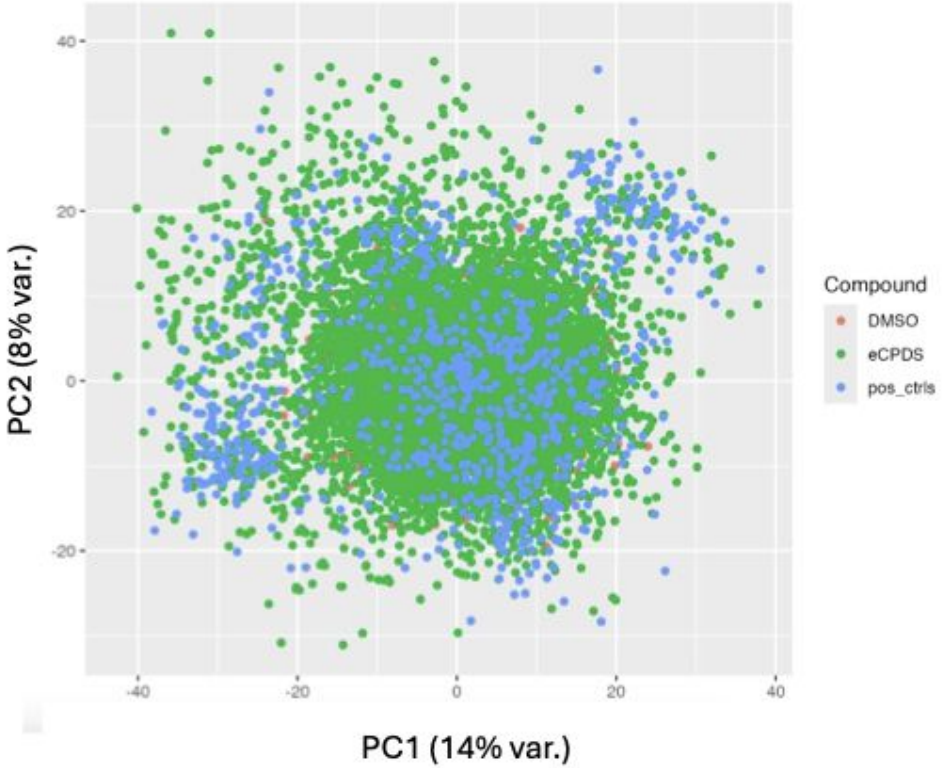
Team 10

Alzbeta Srovnalova, Brian Le, Mutaamba Maasha, Philipp Mergenthaler

Team 10

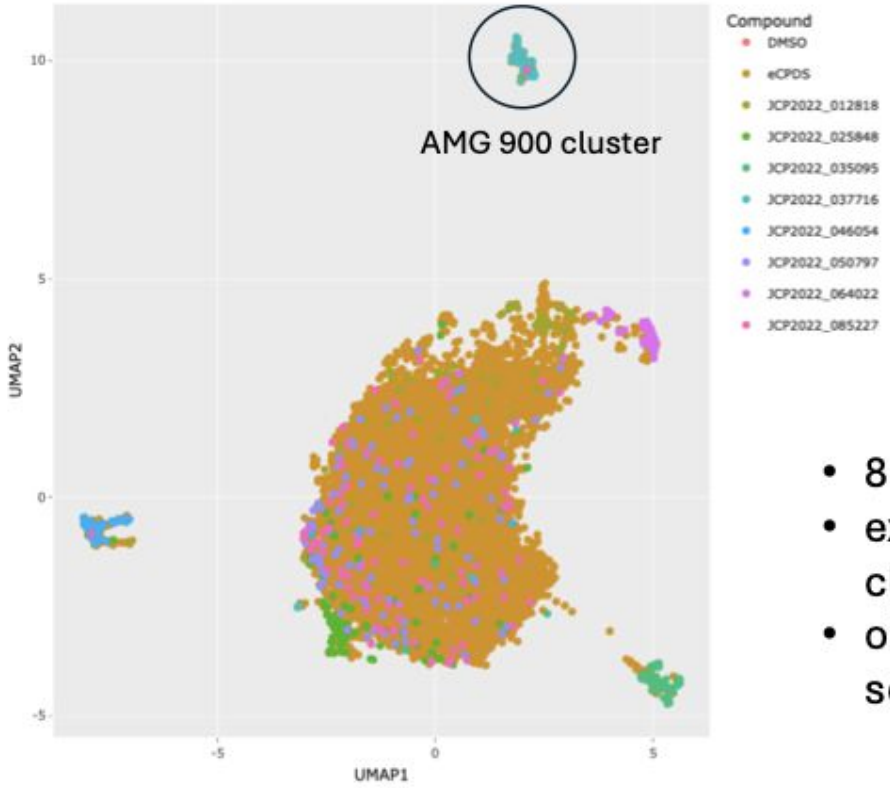
- Idea: can we identify a driving signal from one of the positive controls, fit a model, and observe the effects of query compounds along this axis?
- Proof of concepting on subset of the compound data:
  - 1000 DMSO wells
  - 10,000 query wells
  - 8 x 100 positive control wells

Top 2 PCs



- Start with PCA: very little signal

UMAP visualization

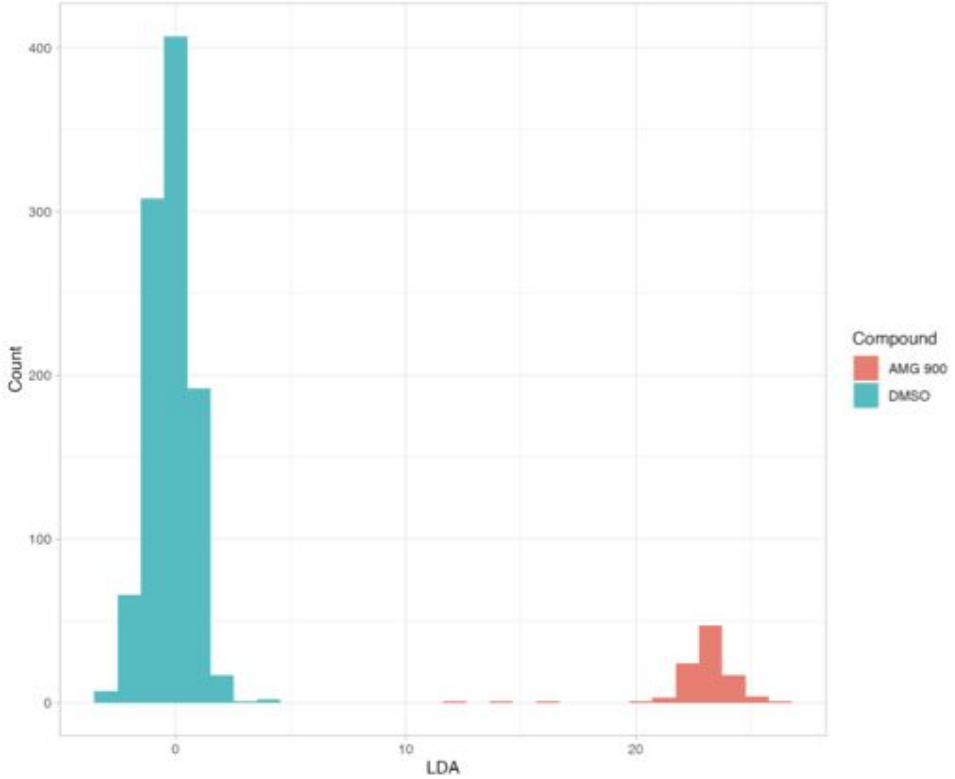


AMG 900 – aurora kinase inhibitor, characterized by unique nuclear morphology such as formation of multinuclei

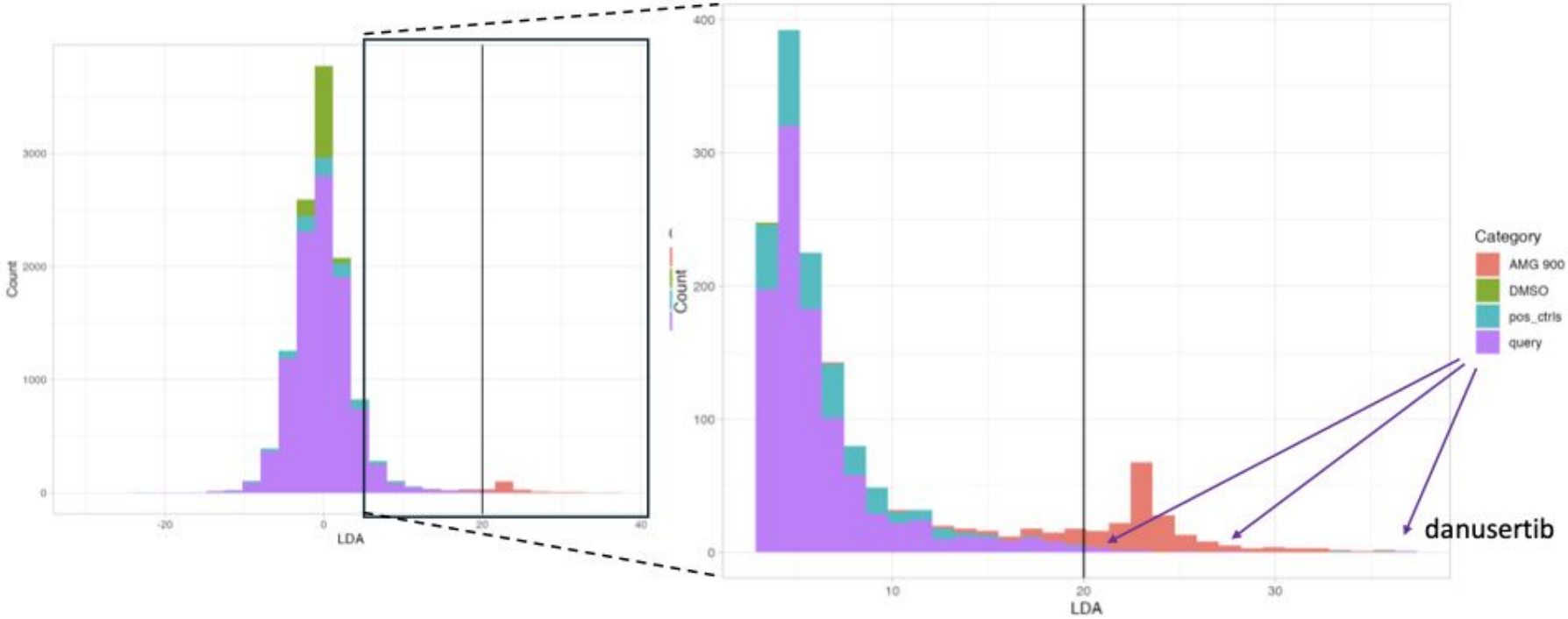
- 8 positive controls
- expected 1 large cluster + 8 small clusters
- observed 4 small clusters, primary separating is cluster of AMG 900

Fit LDA model

- LDA fit to maximize separation between DMSO wells and AMG 900 wells
- Top contributing LDA features: X_626, X_426, X_177

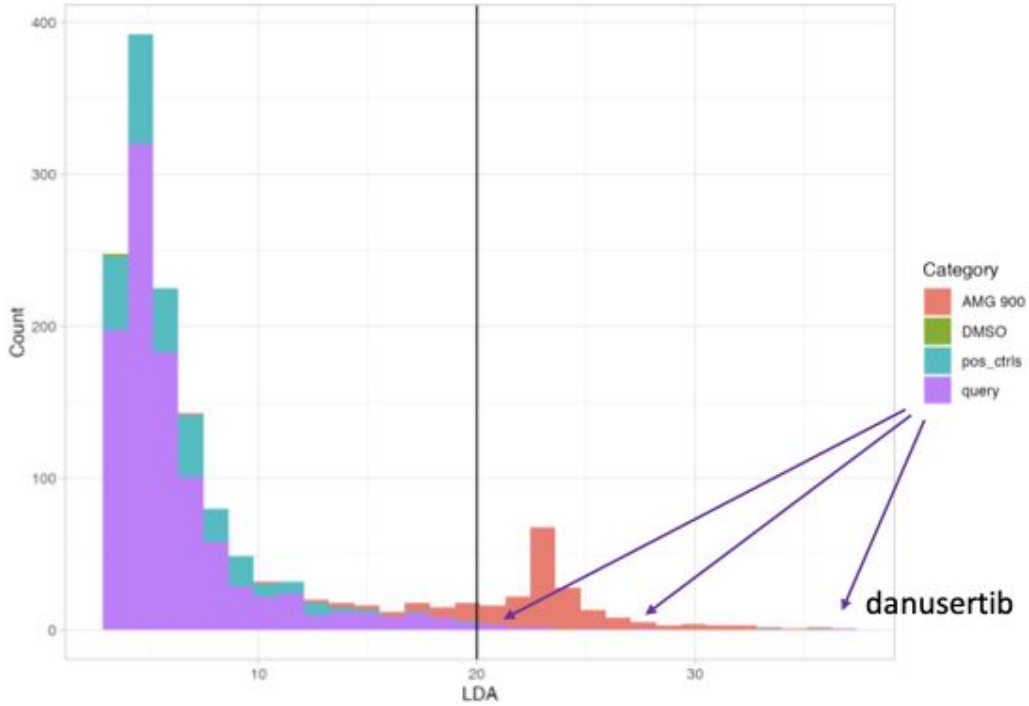


Projecting query compounds to LDA space



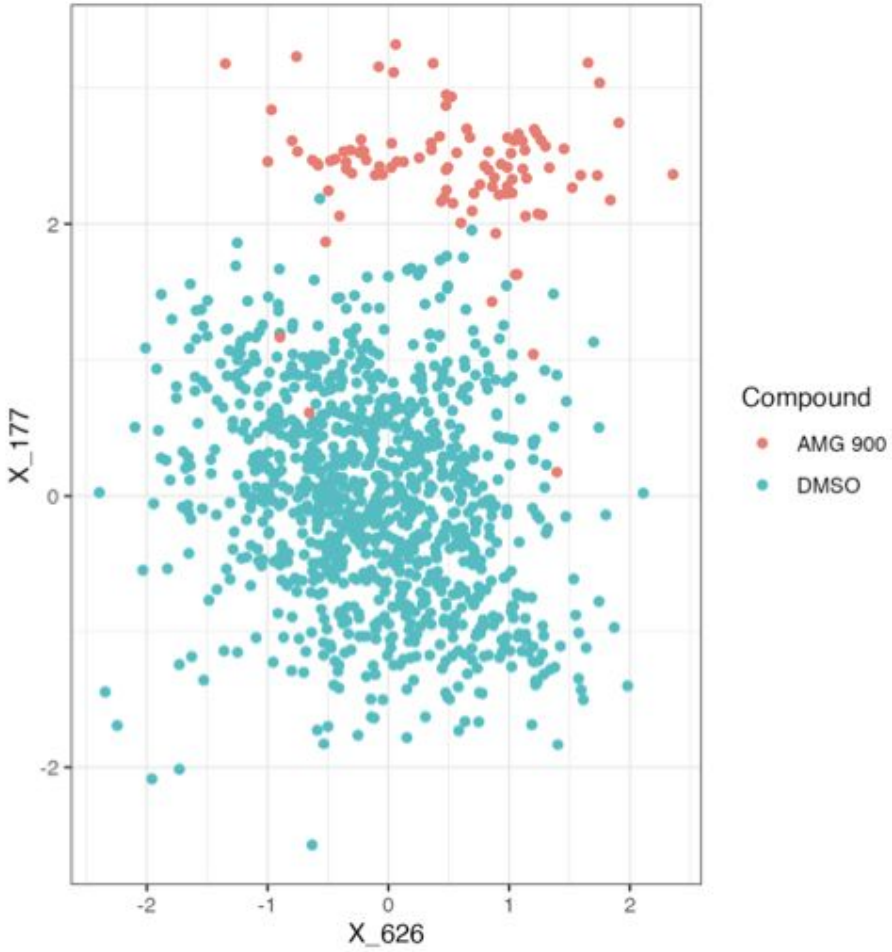
Projecting query compounds to LDA space

- danusertib – also an aurora kinase inhibitor like AMG 900



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)

Team 11

“Batch breakers”

Taras Redchuk, Daniel Siegsimund, Loan Vulliard, Nathan Carter

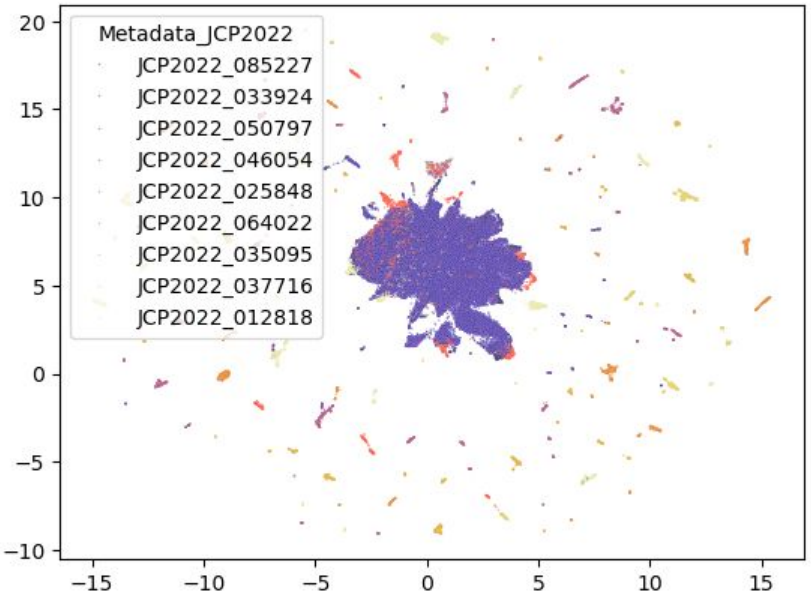
Feature interpretability: JumpCP Compound data

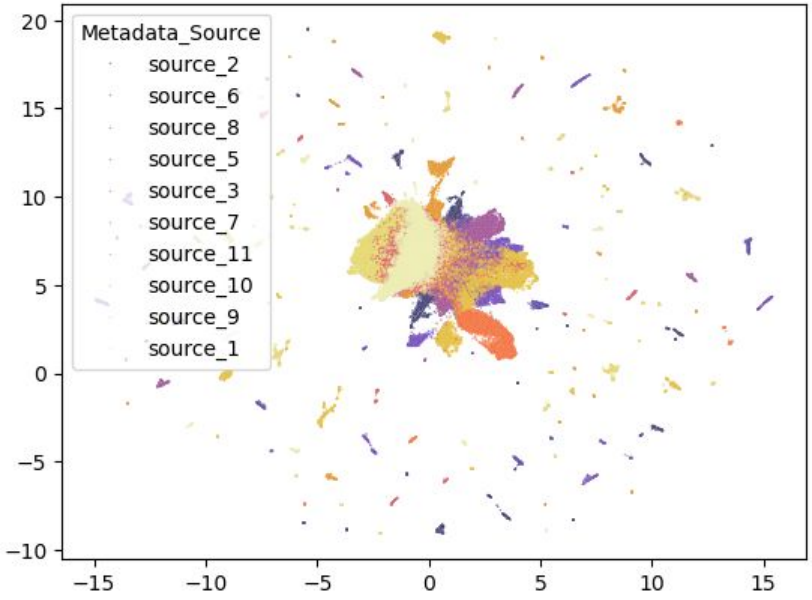
- Project Idea: Evaluate feature importance before batch correction on defined control dataset (1 neg control – 8 pos controls)

- 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.

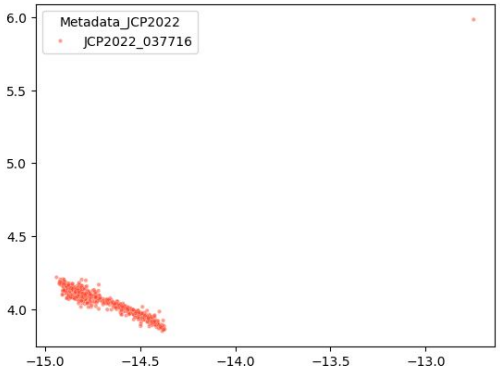
Multi-class: UMAP plots reveal severe batch effects (as expected)





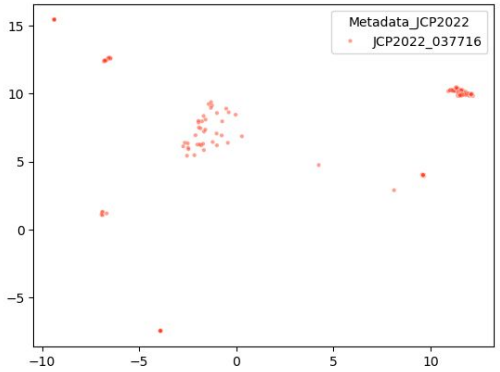
Multi-class: UMAP plots reveal severe batch effects (as expected)

Source 6: Good separation




```
array([[ 199,    0,   17,    0,    0,    0,    0,    0,    0],
       [   0,  220,    6,    0,    0,    0,    0,    0,    0],
       [   1,    0, 3542,    0,    0,    0,    2,    0,    0],
       [   0,    0,    0,  217,    0,    0,    0,    0,    0],
       [   0,    0,    0,    0,  224,    0,    0,    0,    0],
       [   0,    0,    0,    0,    0,  212,    0,    0,    0],
       [   0,    0,  164,    0,    0,    0,   49,    0,    0],
       [   0,    0,    0,    0,    0,    0,    0,  228,    0],
       [   0,    0,  197,    0,    0,    0,    0,    0,    3]])
```

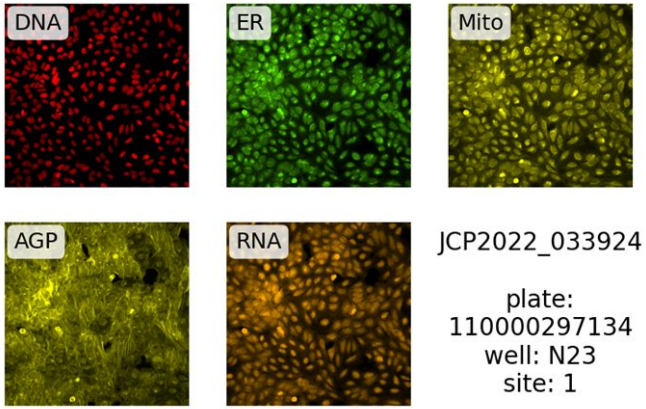
Source 9: Bad separation




```
array([[ 342,    0,   65,    0,    0,    0,    0,    0,    0],
       [   1,  262,  150,    0,    0,    0,    1,    0,    1],
       [   1,    2, 3700,    0,    0,    1,    1,    0,    3],
       [   0,    0,   11,    0,    0,    0,    0,    0,    0],
       [   1,    0,    8,    0,  395,    0,    0,    0,    0],
       [   0,    0,   15,    0,    0,  398,    0,    0,    0],
       [   0,   34,  358,    0,    0,    0,   27,    0,    2],
       [   1,    0,   12,    0,    0,    0,    0,  404,    0],
       [   0,    0,  390,    0,    1,    0,    0,    0,   28]])
```

Negative Control (JCP2022_033924)

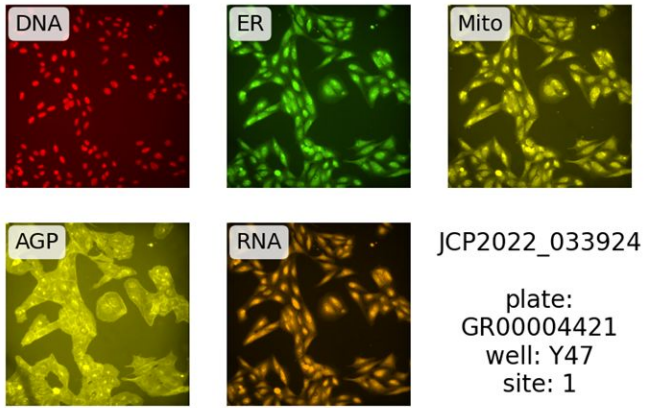
Source_6



JCP2022_033924

plate:
110000297134
well: N23
site: 1

Source_9

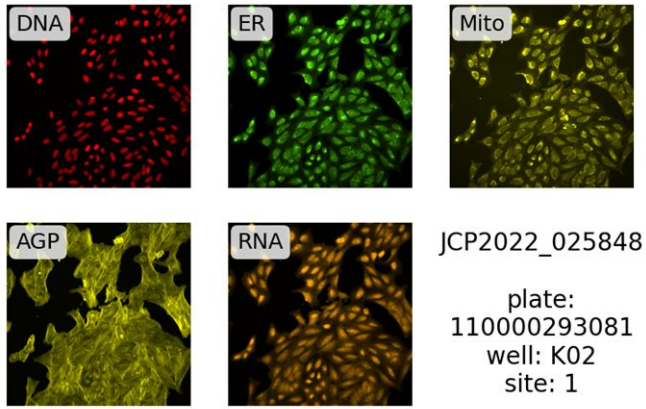


JCP2022_033924

plate:
GR00004421
well: Y47
site: 1

Positive Control (JCP2022_025848)

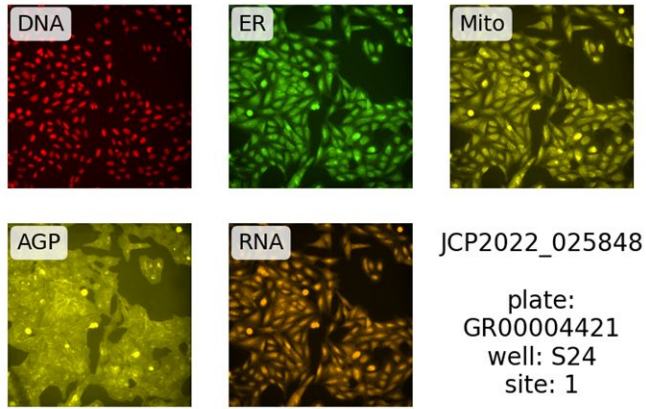
Source_6



JCP2022_025848

plate:
110000293081
well: K02
site: 1

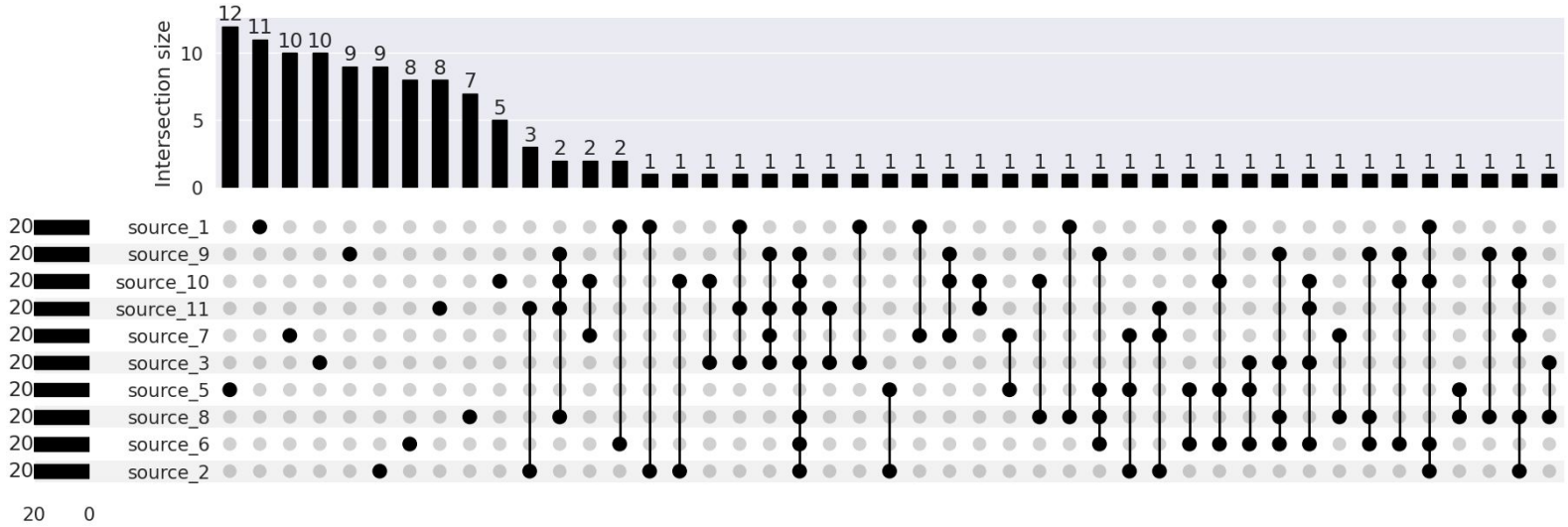
Source_9



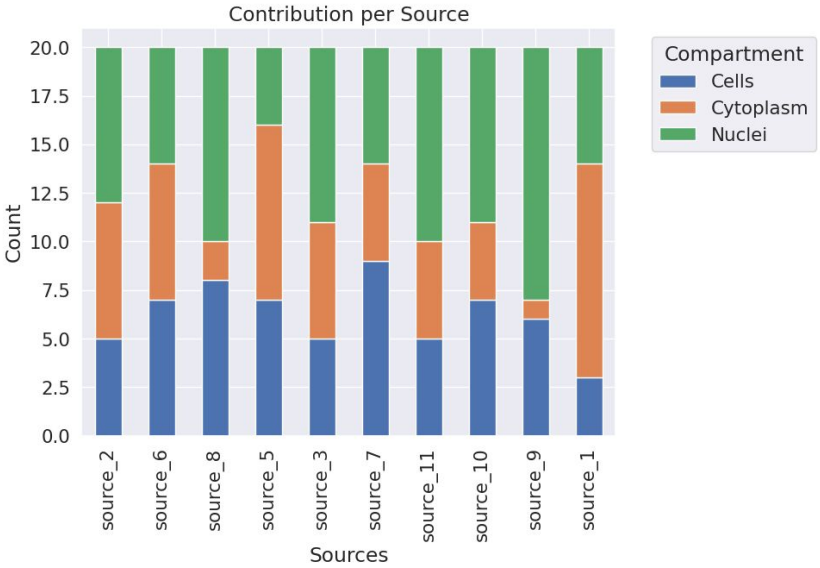
JCP2022_025848

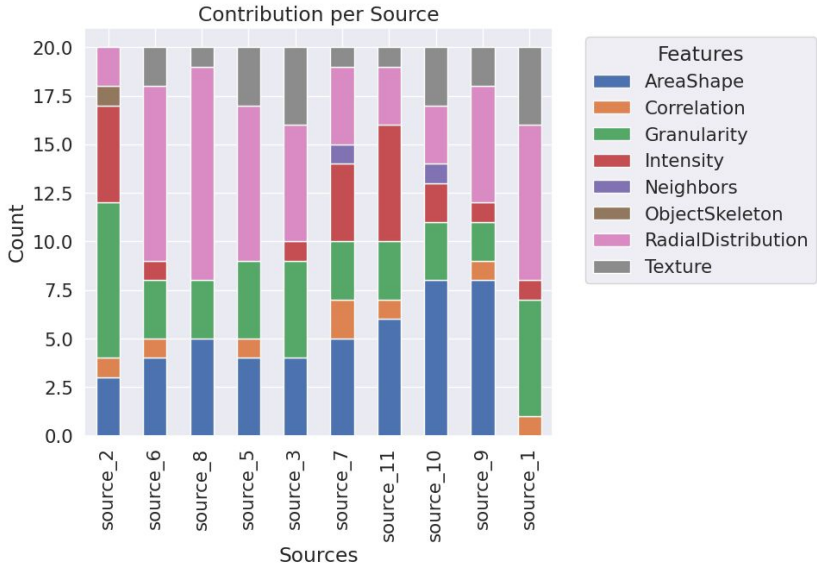
plate:
GR00004421
well: S24
site: 1

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



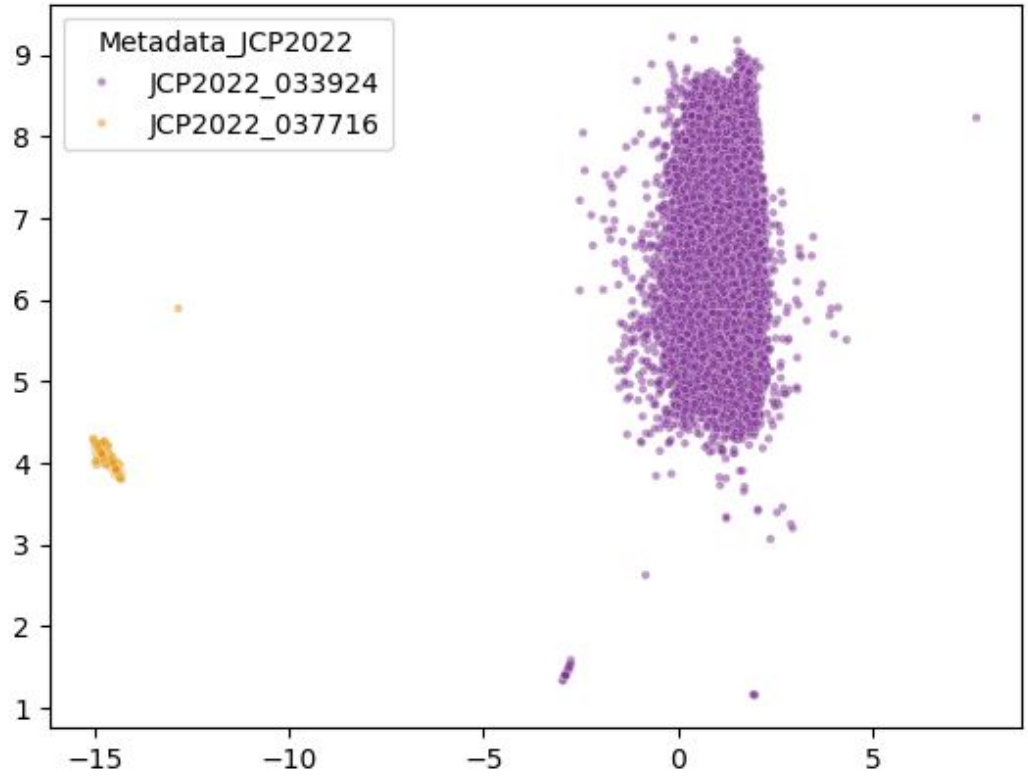
RadialDistribution features conserved in different sources

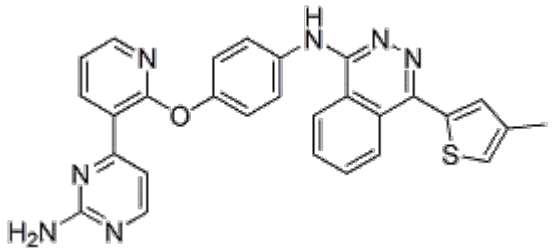




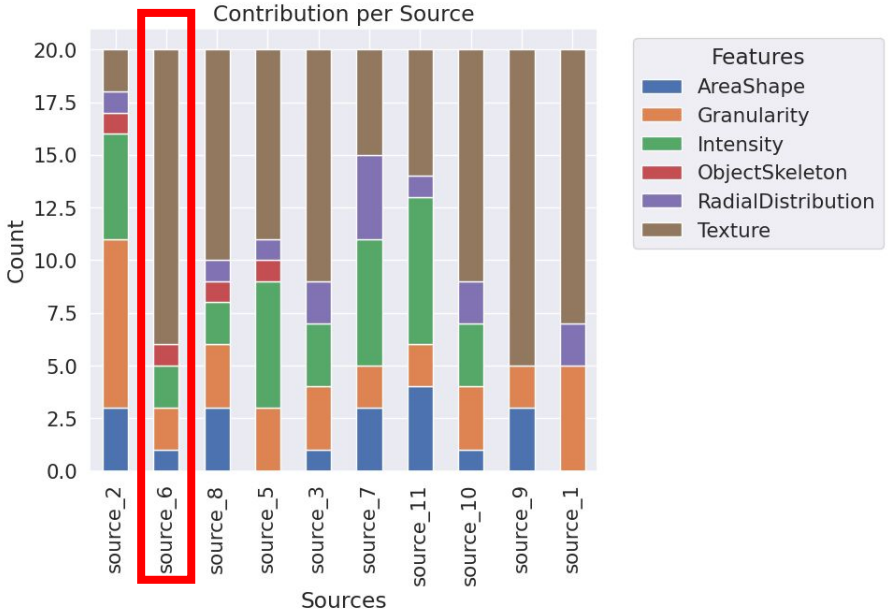
Binary: source 6,

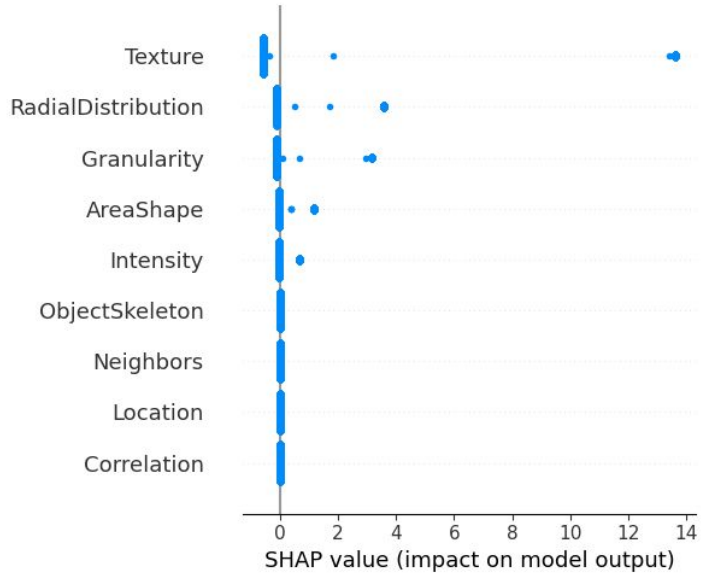
AMG 900 is a potent and highly selective pan-Aurora kinases inhibitor for Aurora A/B/C





Feature importance consistent over both methods for binary classification on source_6





Team 13

Data-driven similarity functions/feature spaces

Ian Smith

Motivation

Current perturbational analysis:

- One-size-fits-all similarity function (e.g. cosine similarity)
- Unified feature space for all analysis

Can different feature spaces or similarity functions uncover different biology?

Given a biological signal, can researchers choose a feature space or similarity function best powered to discriminate that signal?

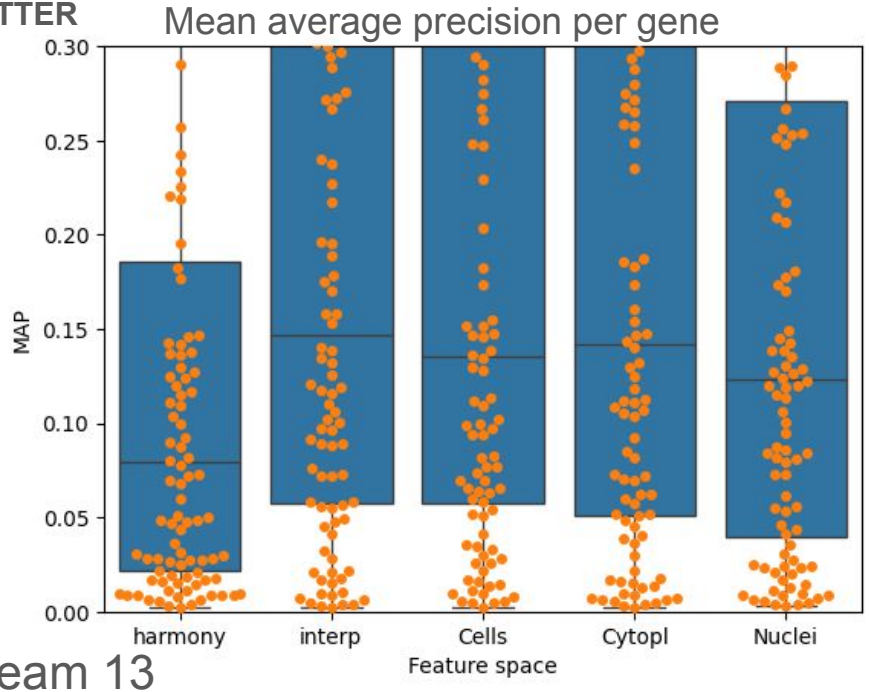
Team 13 - Feature subspaces

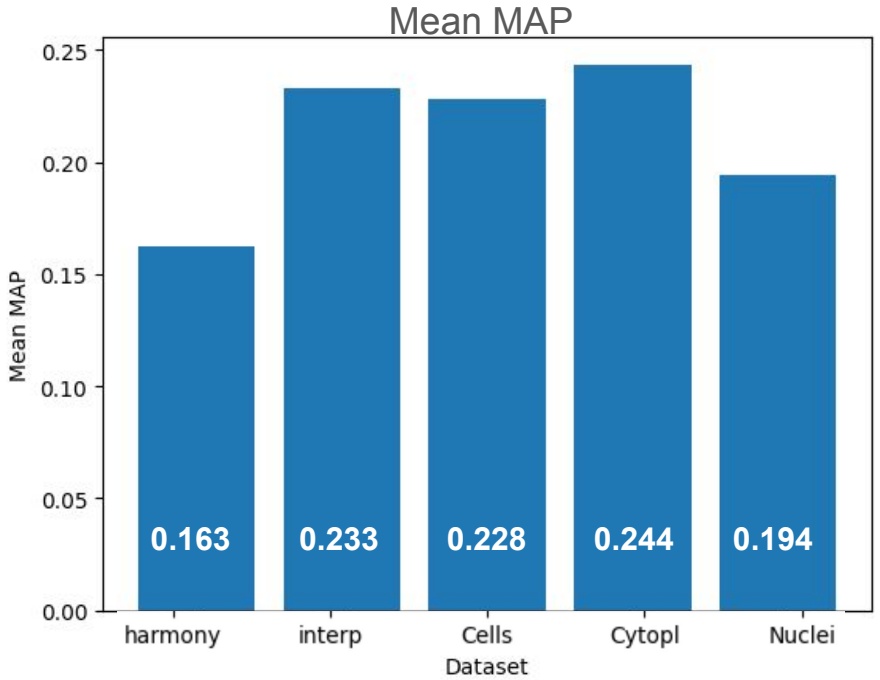
Question: Are feature subspaces as informative as the entire feature space? N = 1959 signatures.

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

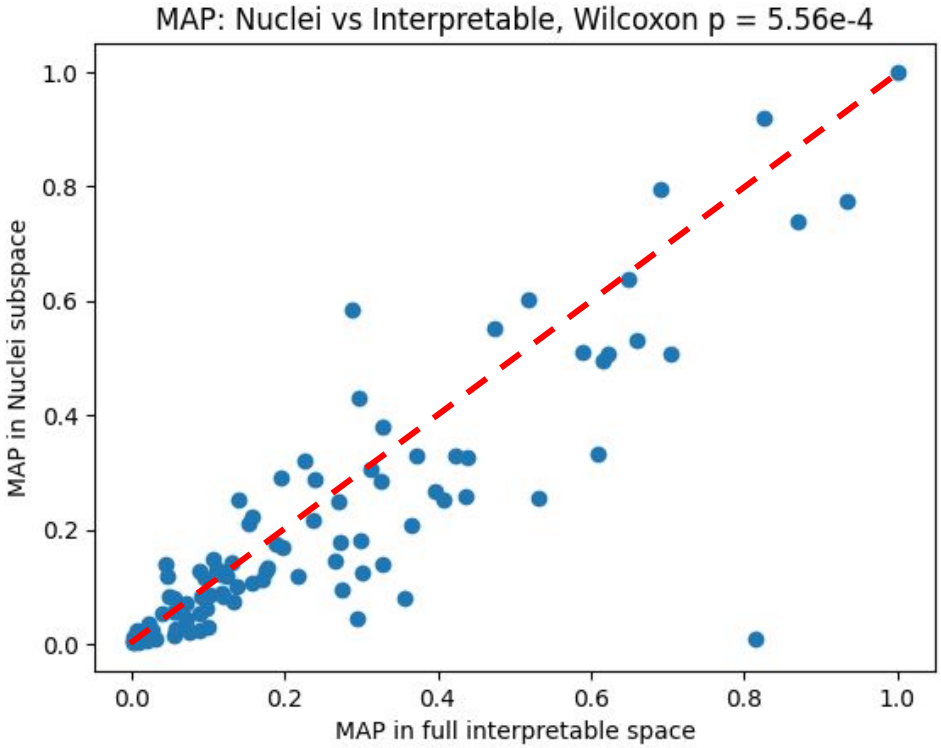
Data: Harmonized (baseline, 259); Interpretable (3651), Cells (1252), Cytoplasm (1234), Nuclei (1165)

BETTER





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.