

Deep learning for morphological profiling: learning cell representations for biological discovery

SBI2 Short Course III: Bioimage Analysis

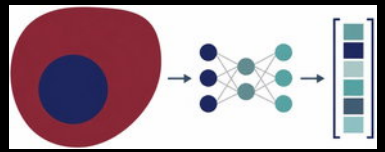
Runxi Shen, Ph.D. | Alán F. Muñoz, Ph.D.

Carpenter-Shen Lab

Department of Biological Sciences

Purdue University

Nov 2nd, 2026



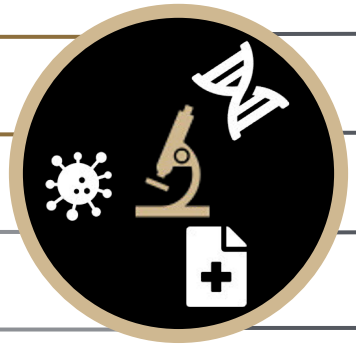
PURDUE UNIVERSITY

Download PDF of slides here: *http://runxi-shen.github.io/talks/SBI2-2026.pdf*
(talk content is non-confidential)

Curate Cell-Painting datasets and develop computational workflows to cure diseases

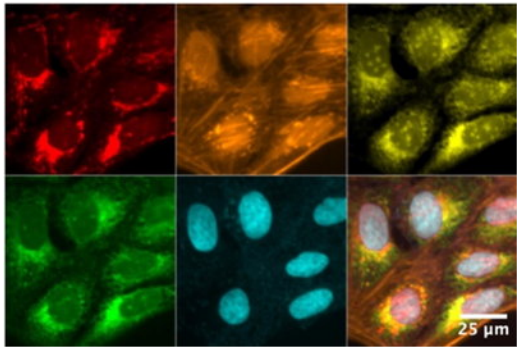
WHAT WE PERTURB

- Patient variants
- CRISPR perturbations
- Compound screens
- Disease states

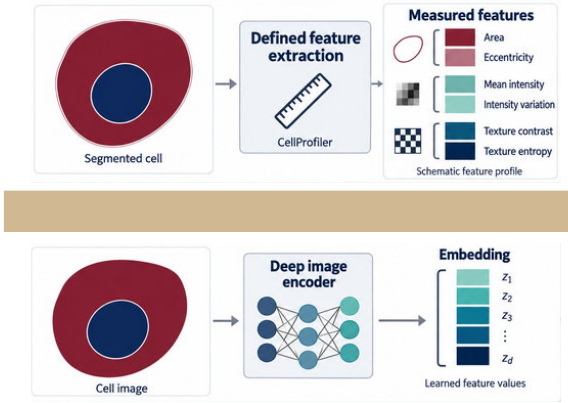


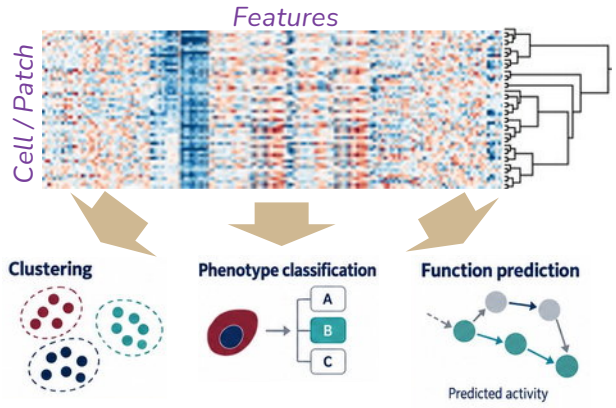
WHAT WE LEARN

- Fundamental cell biology
- Disease mechanisms
- Biomarkers & diagnosis
- Drug response & toxicity



"Cell Painting" assay





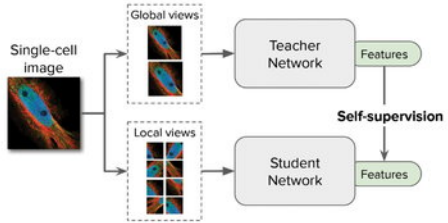
PURDUE UNIVERSITY

Cell-image representations can be learned from images alone or by combining images and biological information

Training objectives shape learned cell representations

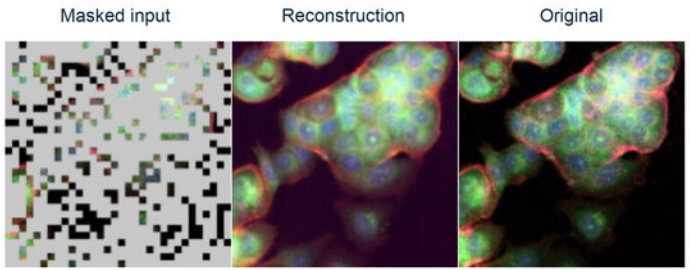
Images alone

1 Match different views



Cell-DINO

2 Reconstruct masked regions



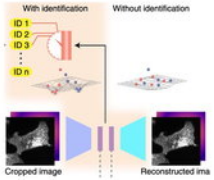
Microscopy MAE

Images + biological information

3 Predict biological identity

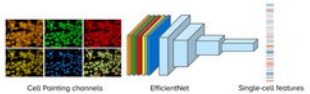
Cytoself

Protein ID

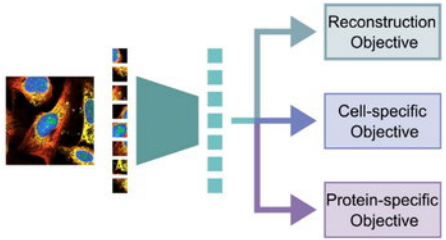


DeepProfiler

Compound ID



4 Combine training objectives

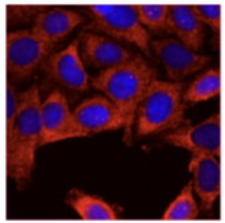


SubCell

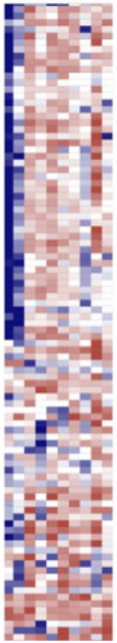
A nice benchmark of self-supervised methods by Abdelmoula et al. *Nature Methods*, 2026

Moutakanni *et al.*, *PLOS Comp Bio*, 2025; Kraus et al., *CVPR*, 2024
Kobayashi et al., *Nature Methods,* 2022; Moshkov et al. *Nature Comm*, 2024; Gupta et al. *bioRxiv*, 2025

Image-based profiling and machine learning to understand genetic variants, predict compound toxicity and accelerate drug development



Cell images



Feature sets



Modeling

Genetics

MisLocus

Characterize the functional impacts of genetic variants

Drug discovery

JUMP-lite

Identify drugs to treat the diseases

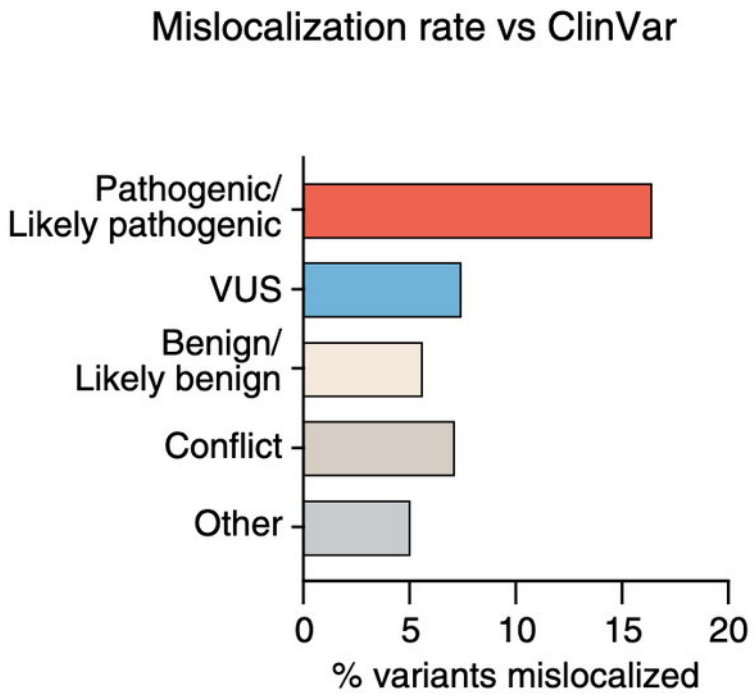
Drug activity

OASIS

Predict drug toxicity using in-vitro assays

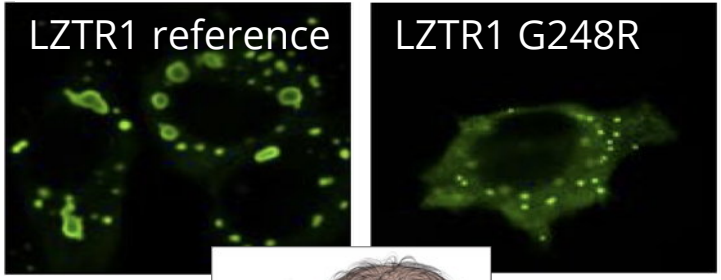
PURDUE UNIVERSITY

Protein mislocalization is a prominent phenotype of pathogenic genetic variants and can be used as a phenotype for drug screening

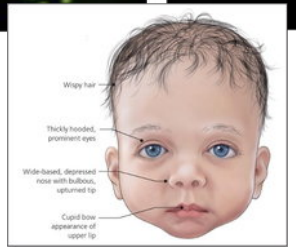


Lacoste, Haghighi et al., Cell, 2024

Identify the phenotypes



Noonan syndrome



Bhambhani & Muenke, Am Fam Physician. 2014

Screen drugs to reverse them



Mikko Taipale
U Toronto



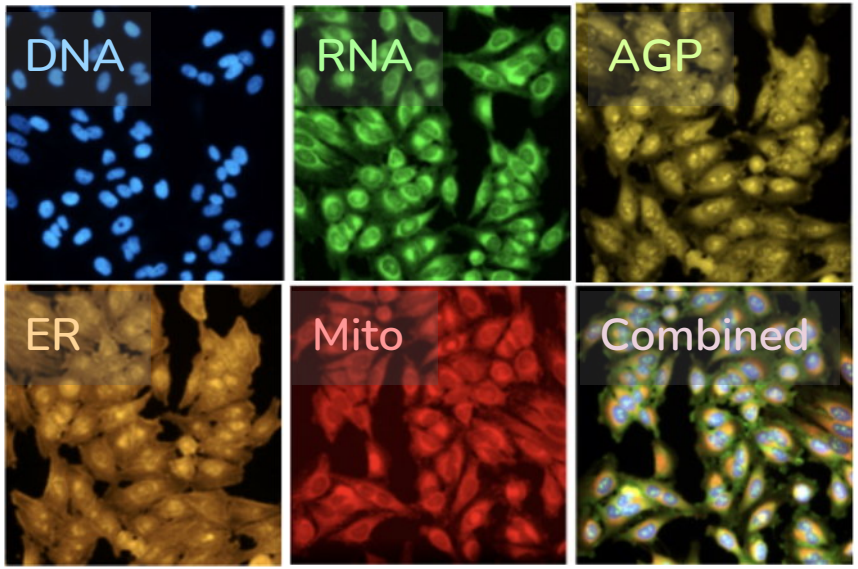
Jessica Lacoste



Marzieh Haghighi

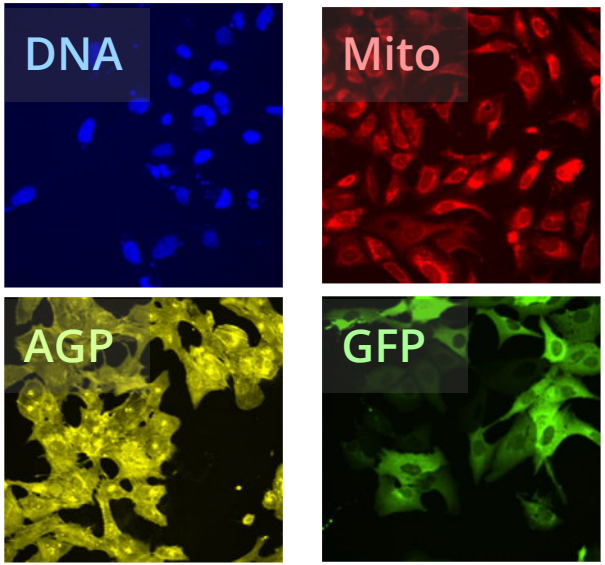
PURDUE UNIVERSITY

The Variant Painting assay can profile the complex phenotypes of genetic variants



Standard Cell Painting

(general cell structures)



Variant Painting

(Morphology + Localization)

Different imaging channels and features were used to capture alterations in protein localization and cell morphology

GFP: Protein localization

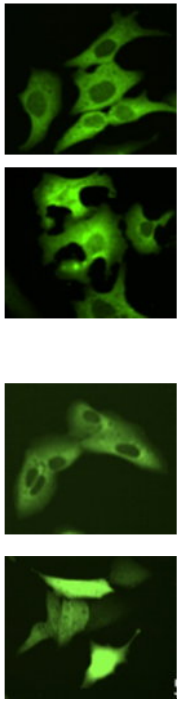
DNA, Mito, AGP: Cell morphological profiles

PURDUE UNIVERSITY

Lacoste, Haghighi et al., Cell, 2024

An automated machine learning workflow detects variant cells with protein mislocalization using the CellProfiler single-cell features

Imaging wells



Control

Control

Gene Ref

Gene Var

Single-cell profiles for reference and variant cells



Train

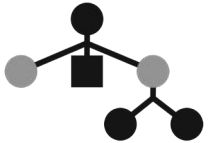
Test

4-fold split



Train

Test

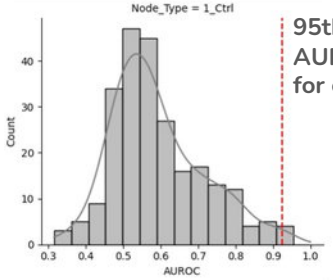


XGBoost classifier

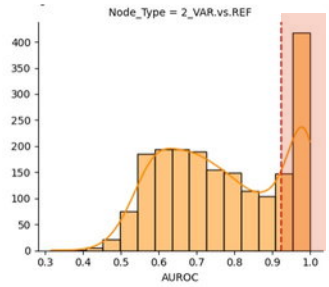
Model AUROC

Imaging Phenotype Score

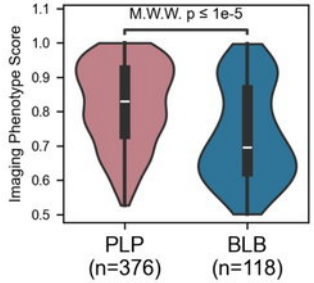
Distribution of AUROC scores between Ctrl-Ctrl and Ref-Var wells



95th percentile AUROC cutoff for calling hits



Hits for protein mislocalization



Pathogenic variants have higher scores than benign

Imaging Phenotype Score: is there a significant protein mislocalization between the Variant and Reference alleles? *(based on the null distribution of AUROC from comparing the same control alleles from different wells)*

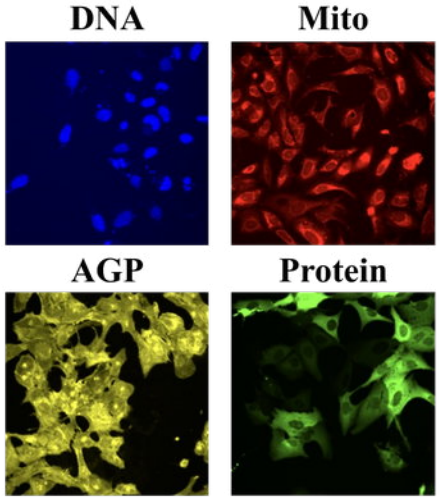


Jessica Ewald

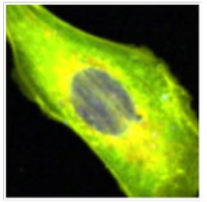
PURDUE UNIVERSITY

MisLocus: a single-cell image benchmark for protein *mislocalization*

Variant Painting:
Morphology + Localization



Segmentation
+
Quality control





Johan Fredin Haslum

Out-of-box models
CellProfiler
MorphEm
SubCellPortable

Retrained/finetuned models
Cytoself
Finetuned subcell-embed

Data cleaning
- ❑ Postprocessing
- ❑ Plate normalization
- ❑ Feature selection

1. Mislocalization detection
- ❑ Supervised model: ***XGBoost***
- ❑ Unsupervised method: ***copairs***

2. Predicting clinical pathogenicity
- ❑ Variant effect to clinical relevance

ClinVar *AlphaMissense* *ESM1b* *REVEL* *MutPred2*

3. Protein subcellular localization
- ❑ Distinguish protein localization pattern

The Human Protein Atlas

github.com
runxi-shen/MisLocus/

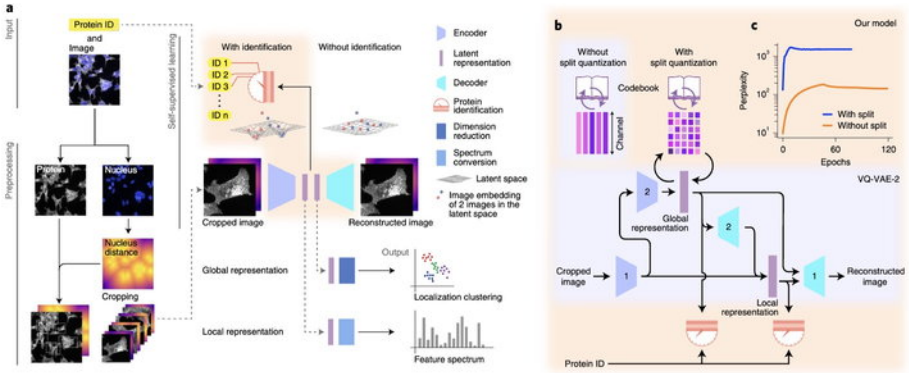
huggingface

PURDUE UNIVERSITY

Three models for protein localization were benchmarked using their out-of-box, fine-tuned, or re-trained versions

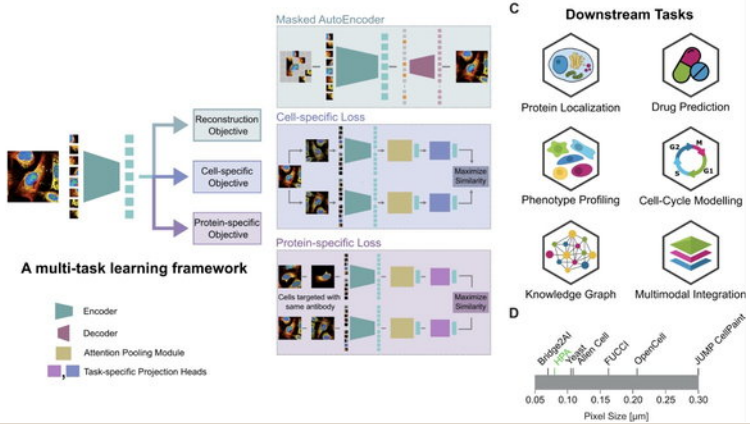
Cytoself: GFP + Nucleus, localization-specific

Retrained on our own data



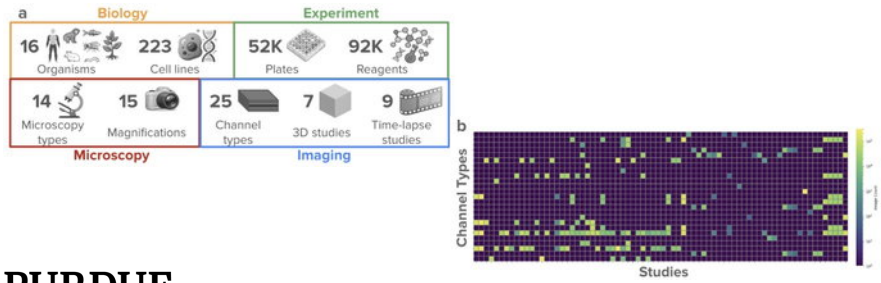
SubCell: GFP + HPA channels, localization+morph

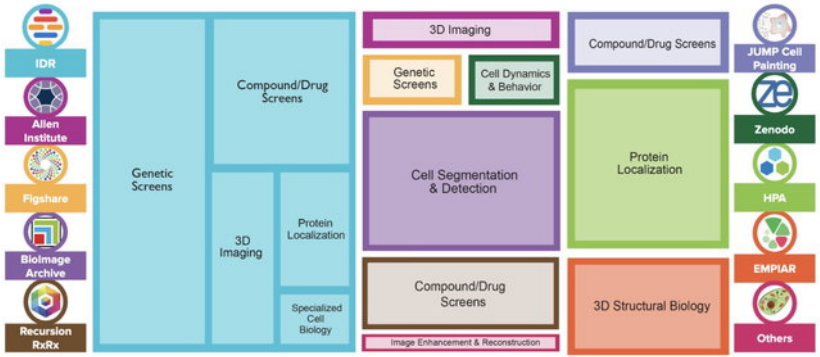
Both out-of-box model and finetuned model were used



MorphEm: channel-agnostic model, localization+morph

Out-of-box model trained from CHAMMI-75 dataset





PURDUE UNIVERSITY

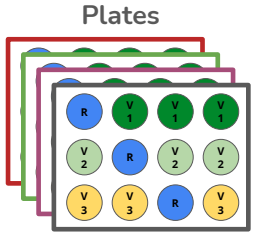
Kobayashi *et al.*, *Nature Methods,* 2022; Gupta et al. *bioRxiv*, 2025; Agrawal et al. *ICLR*, 2026

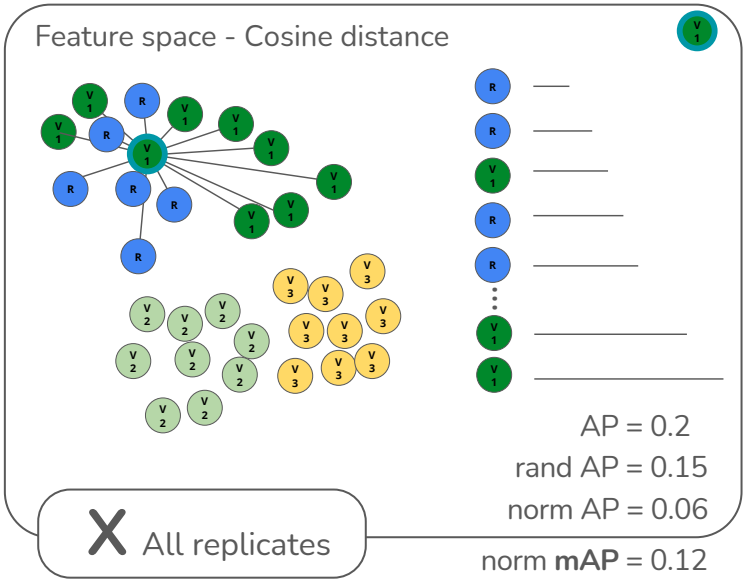
Similarity-based scoring detects variant–reference differences without classifier training: faster and less risk of overfitting

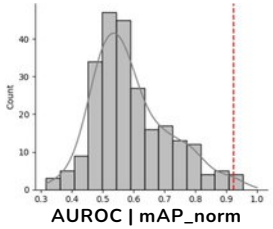
- **Can we distinguish variant from reference?**
  Distance between replicates of VAR and REF
  Per variant calculate the mean Average Precision

- **Per variant score**
  Normalized mean Average Precision across variant profiles
  Permutation based significance scoring - mAP by chance?

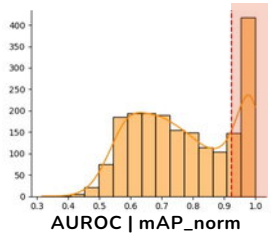
- **Hit calling**
  Well-position effect scoring - 95th percentile cutoff
  Above 95th percentile & significant mAP score







Null distribution for technical variations caused by well-position effects
95th percentile cutoff for calling hits



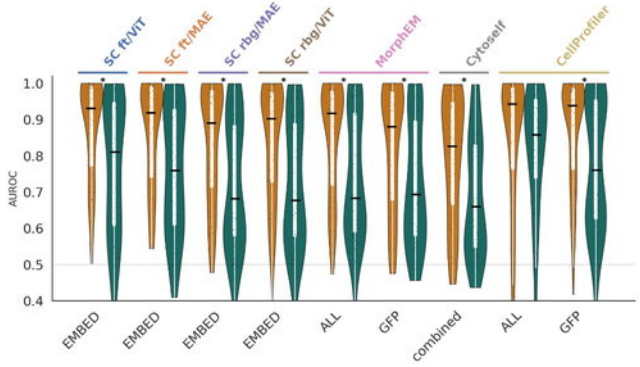
Hits for protein mislocalization

PURDUE UNIVERSITY

Kalinin *et al.*, *Nature Communications*, 2025

CellProfiler has competitive mislocalization hit rates with the learned representations under the supervised but not the unsupervised framework

Variants labeled as mislocalized by *Lacoste et al.* shift toward higher supervised AUROC scores than variants not labeled as mislocalized

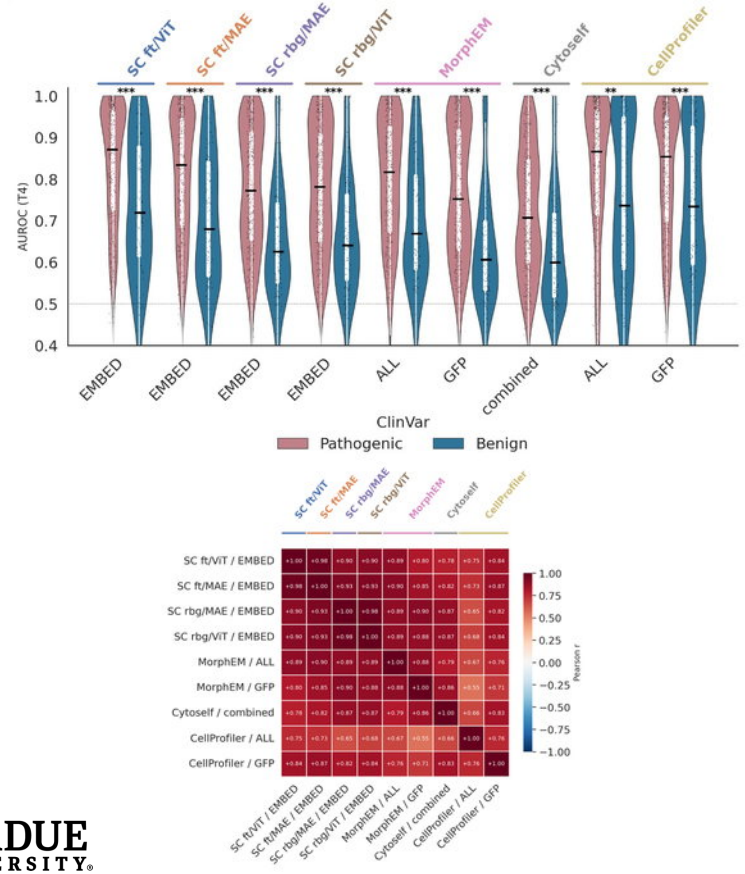


		Mislocalization hits & Imaging phenotype scores							Mislocalization hits vs ClinVar classifications					
		XGB				PA			XGB			PA		
Representation	Channel	Hits	Hit rate	AUROC	Null p_{95}	Hits	Hit rate	mAP_{norm}	Precision	F_1	Fisher p_{BH}	Precision	F_1	Fisher p_{BH}
SC ft/ViT	EMBED	274	29.2%	0.800	0.928	349	34.0%	0.354	81.6%	0.521	0.022	84.3%	0.529	0.002
SC ft/MAE	EMBED	275	29.3%	0.773	0.905	315	30.7%	0.336	81.3%	0.526	0.025	88.2%	0.508	1e-4
SC rbg/MAE	EMBED	226	24.1%	0.738	0.903	319	31.1%	0.291	83.5%	0.464	0.011	87.8%	0.493	1e-4
SC rbg/ViT	EMBED	200	21.3%	0.739	0.915	309	30.1%	0.274	81.6%	0.403	0.071	86.5%	0.452	0.001
MorphEM	ALL	170	18.1%	0.757	0.946	183	17.8%	0.164	76.7%	0.330	0.472	83.0%	0.315	0.085
MorphEM	GFP	334	35.6%	0.712	0.824	277	27.0%	0.287	87.0%	0.609	1e-5	89.0%	0.498	1e-4
Cytoself	combined	365	39.0%	0.694	0.790	203	19.8%	0.130	84.9%	0.617	1e-4	89.4%	0.357	1e-3
CellProfiler	ALL	292	31.1%	0.796	0.952	37	3.6%	0.002	77.1%	0.505	0.322	94.7%	0.091	0.062
CellProfiler	GFP	219	23.3%	0.796	0.952	49	4.8%	0.034	84.2%	0.449	0.008	92.0%	0.114	0.062

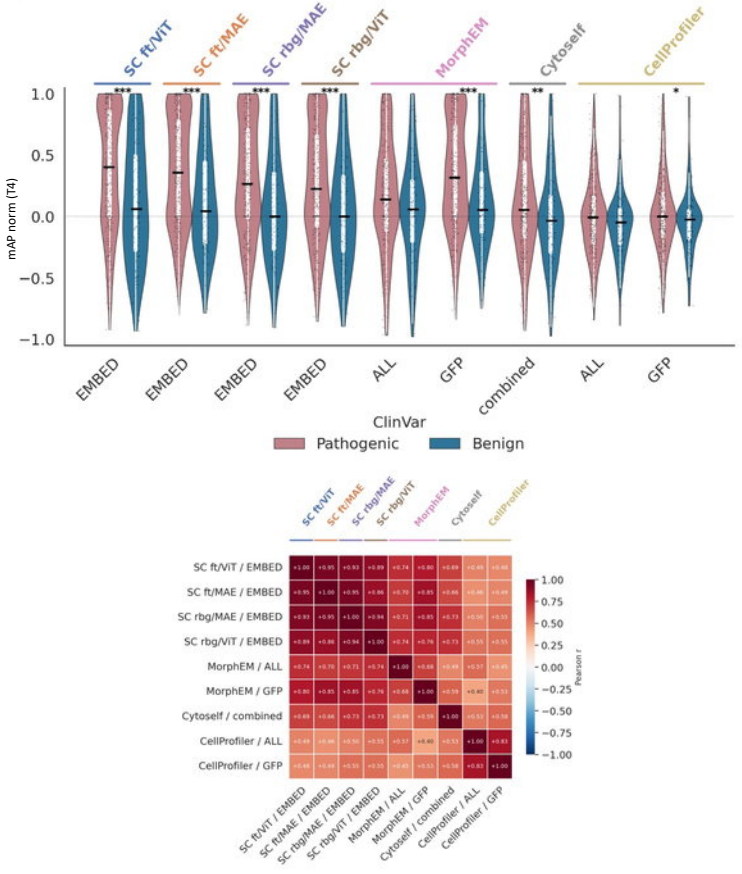
Lacoste and Haghighi, et al. *Cell* 2024

Imaging phenotype scores are associated with ClinVar variant classifications and are highly correlated between methods

Supervised *XGBoost*



Unsupervised *copairs*

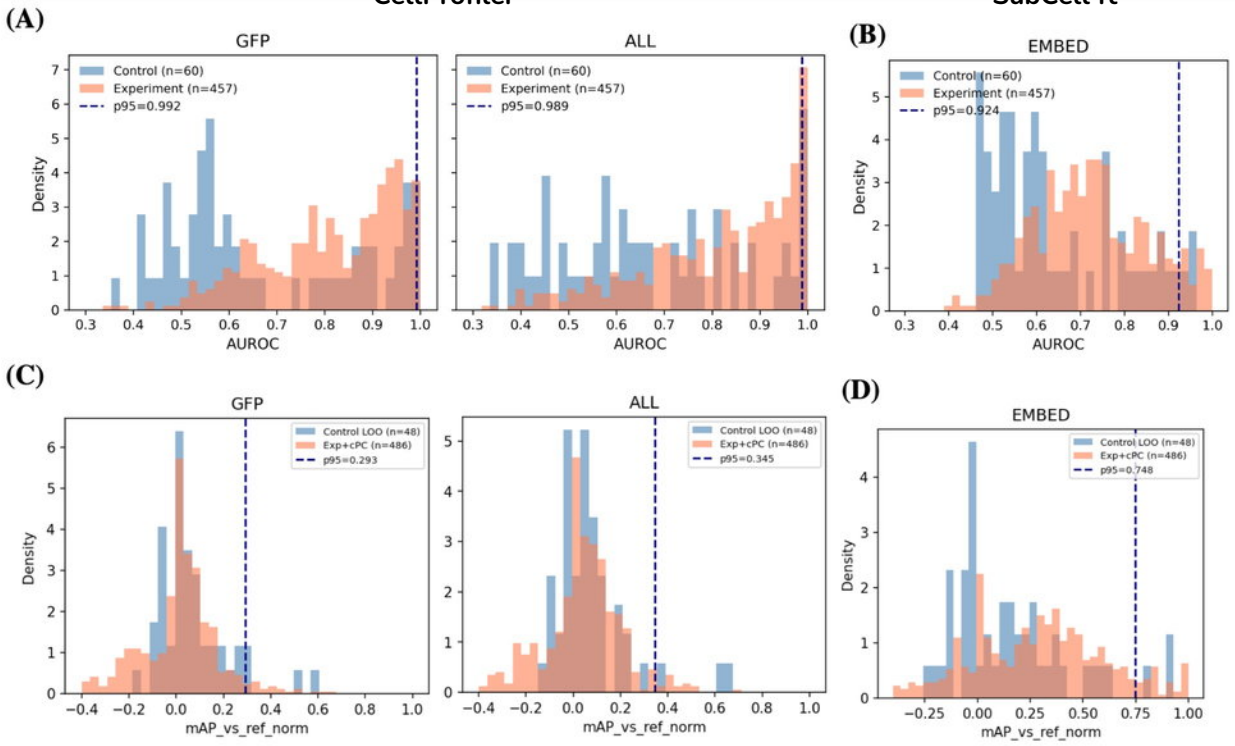


PURDUE UNIVERSITY

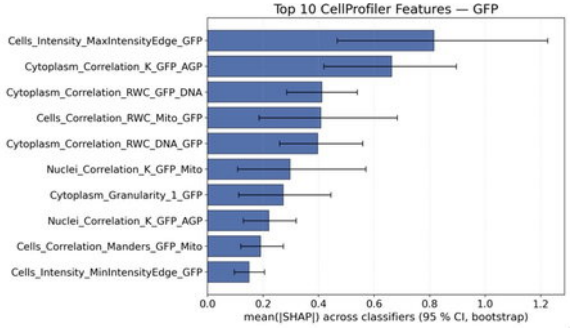
Nuisance variation in CellProfiler features reduces mislocalization hit detection when comparing experiment vs control classifiers

Lower mislocalization hit-calling rates with CellProfiler features

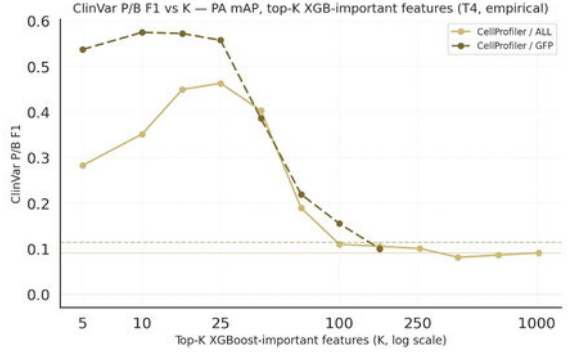
CellProfiler | **SubCell ft**



CellProfiler control-null classifiers are dominated by protein-channel intensity features



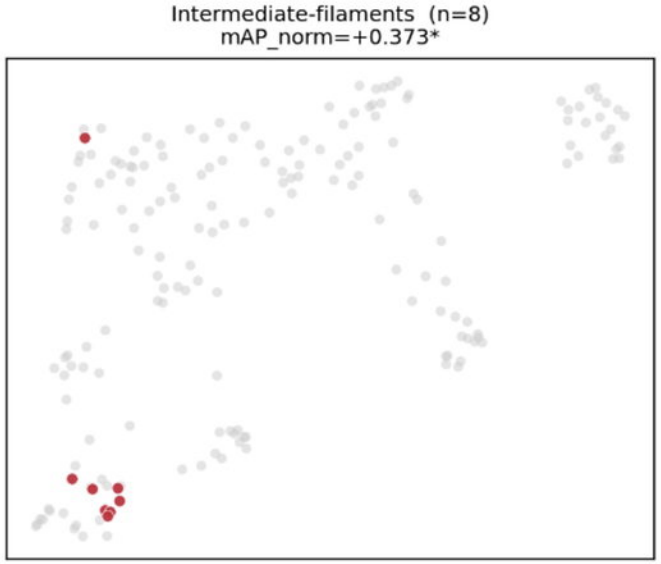
Mislocalization info is captured by representations, but obscured by other features when using *copairs*



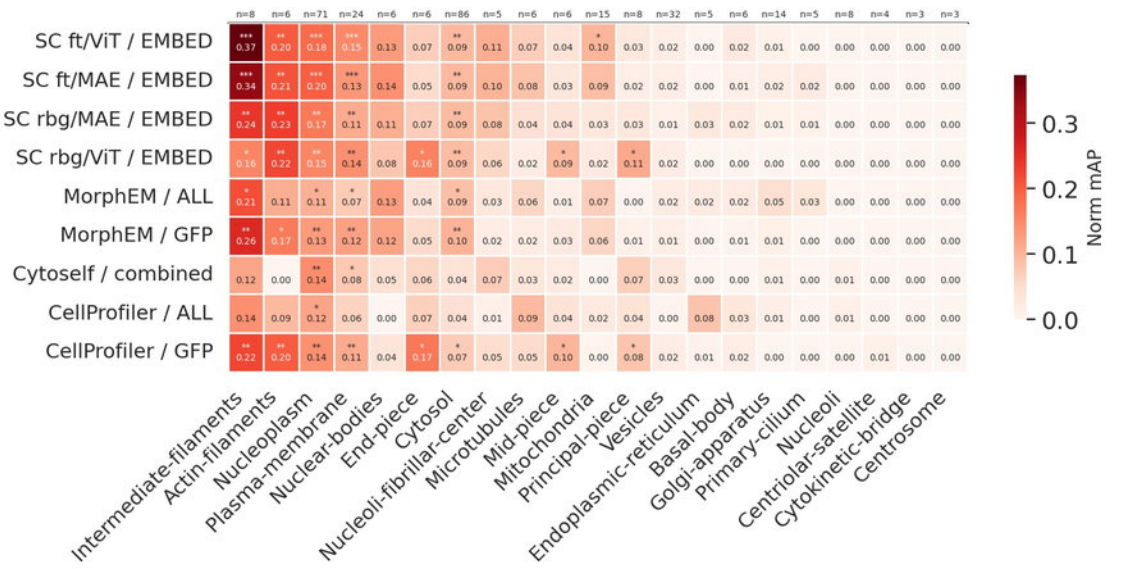
PURDUE UNIVERSITY

Learned embeddings better preserve protein-localization structure

- **Feature space grouping by localization patterns**
- **Finetuned** models better reproduce the localization grouping in HPA
- **All models** are significant in at least one group, indicating meaningful feature spaces for encoding protein-localization information.

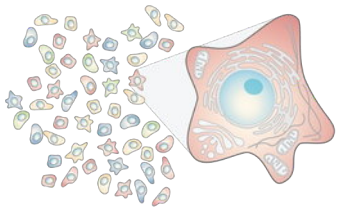


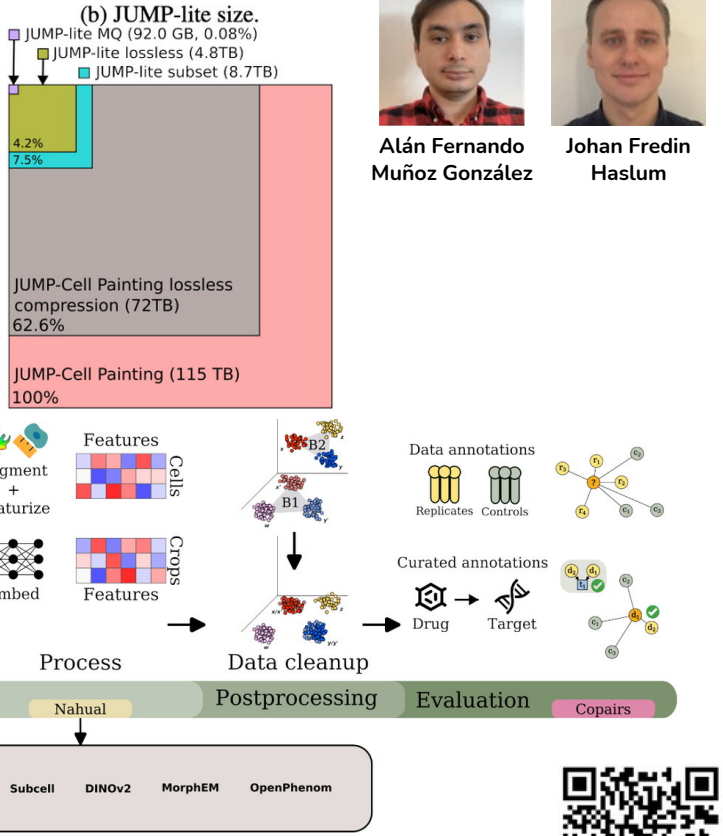
UMAP of SubCell ViT Finetuned embeddings



PURDUE UNIVERSITY

JUMP-lite: Compact, reproducible benchmark of cell representations





Alán Fernando Muñoz González

Johan Fredin Haslum

- **Cell Painting assay in U2OS cells, 5 replicates**
- **13,000 overexpressed genes**
- **8,000 CRISPR knockouts**
- **116,000 (+30,000) small molecules**

1.6 billion single cells (entire Cell Painting Gallery: 2 billion)
2,378 plates (384 well) = 913,152 wells
200 TB of images, similar amount of extracted features
A huge dataset - hard to be benchmarked and re-used
Now a compact lite version exists →

PURDUE UNIVERSITY

https://arxiv.org/abs/2608.07632

A compressed and well-annotated JUMP-CP subset for benchmarking



Select from full JUMP
115 TB → 8.7 TB

Compress the selected images
8.7 TB → 92 GB
~94× smaller with lossy JPEG XL (MQ).

All CRISPR and ORF · Most annotated compounds

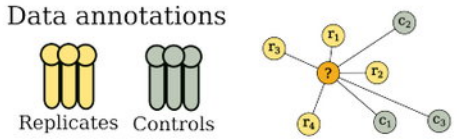
24,347 treatments retained after quality control.

Adapted from Muñoz, Fredin Haslum et al., JUMP-lite (2026).

Benchmarking on well-defined biological tasks: activity and consistency of compounds and their relationship to genetic perturbations

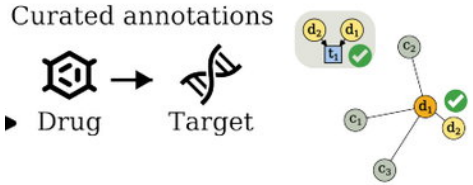
Activity: JUMP treatment and control labels

Replicate CRISPR, ORF and compound wells versus same-plate controls.

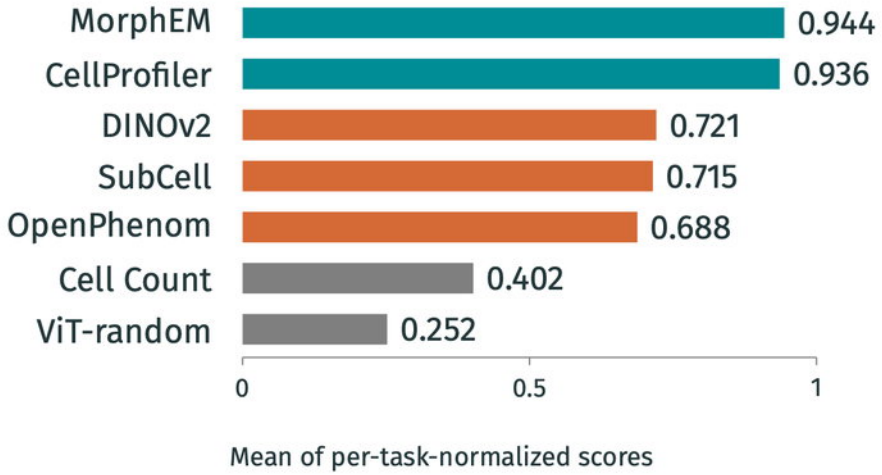


Consistency: RefChemDB compound–target annotations

Compounds sharing a target versus compounds with different targets.



Models: MorphEM and CellProfiler lead on raw images



Make comparison reproducible

Nahual isolates model dependencies.

Aggregate scores from Muñoz, Fredin Haslum et al., JUMP-lite (2026), Fig. 5. *CellProfiler uses more sites than the learned representations.

Moderate compression trades a little signal for much less storage

	HQ	MQ	D20
JPEG XL distance	1	3	20
Image storage	237.7 GB	92 GB	16.2 GB
Storage reduction	~97.3%	~98.9%	~99.8%
Mean task-score change	−1.2%	−6.5%	−19.3%

Choose the trade-off for your task.

Poster 1, the manuscript is available

arxiv.org/abs/2608.07632

Work accepted at WACV 2027

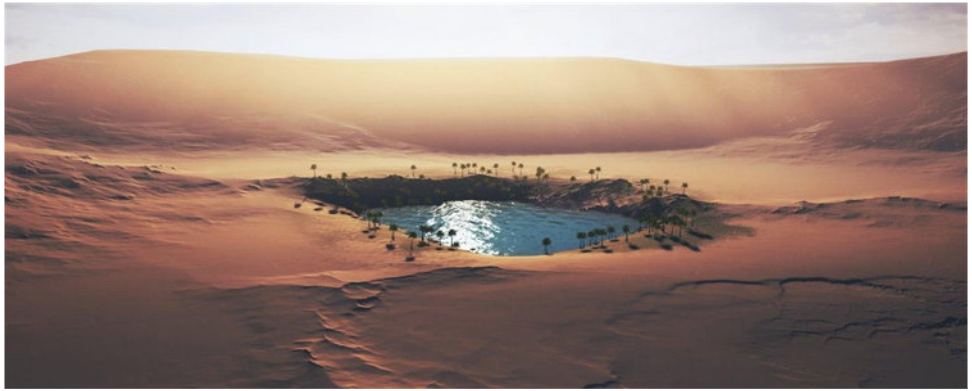
PURDUE UNIVERSITY

Omics for **A**ssessing **S**ignatures for **I**ntegrated **S**afety (OASIS) Consortium

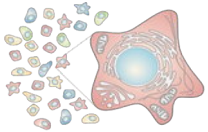
OASIS

Omics for Assessing Signatures for Integrated Safety Consortium

"Successful drug discovery is like finding oases of safety and efficacy in chemical and biological deserts." – *Scannell, Nat Rev Drug Discovery 2022*

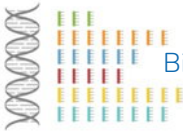


Cell Painting (images)



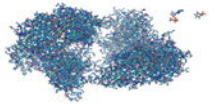
Broad

Transcriptomics (mRNA)



BioSpyder

nELISA (protein)



Nomic bio

Predict compound toxicity in humans

PURDUE UNIVERSITY

Rouquié, et al. Tox Sci 2025

OASIS Consortium for toxicity assessment

Public and private compounds with associated in vivo and/or human data

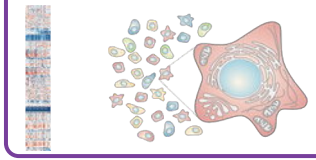
- ❑200 donated from industry partners
- ❑1,400 from public domain

Cell model systems *(of increasing complexity, physiological relevance, and cost)*

- ❑U2OS human osteosarcoma immortal cell line
- ❑HepaRG human hepatocyte-like cell line
- ❑Primary human hepatocytes
- ❑Primary rat hepatocytes
- ❑HepG2
- ❑Liver organoids, Liver-on-a-chip

Multi-omic profiling assays

Cell Painting (images)



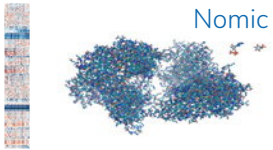
Transcriptomics (mRNA)

BioSpyder



nELISA (protein)

Nomic bio



Data analysis

- ❑Translating *in vitro* to *in vivo* concentrations (qIVIVE)
- ❑Distinguishing & interpreting mechanisms
- ❑Integrating evidence across multiple data types/cell models

Publicly accessible data

Published methodologies

- ❑Community of Practice

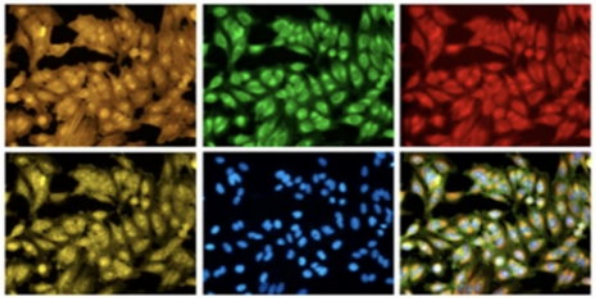


PURDUE UNIVERSITY

OASIS

Omics for Assessing Signatures for Integrated Safety Consortium

An automated computational workflow to predict drug bioactivity and liver toxicity in humans and rats from simple cell cultures

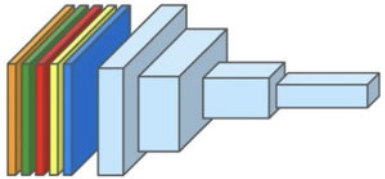


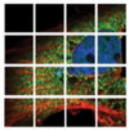
Cell Painting channels

CellProfiler™
cell image analysis software



Count, Shapes, Sizes, Intensities, Textures, Learned Features, etc.





Tokenized image

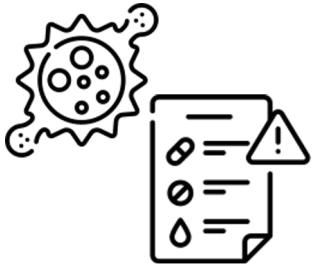
Abstract numerical vectors describe the image representation



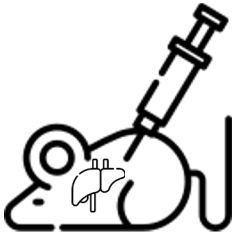
Single-cell features



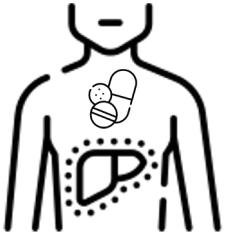
Machine Learning Model



Cytotox + MOA
ToxCast



Rat in-vivo liver tox
ToxValDB, TG-GATES, DrugMatrix



Human DILI
DILIrank 2.0

PURDUE UNIVERSITY

Diverse biological endpoints for cytotoxicity, bioactivity and hepatotoxicity were curated from multiple databases

ToxCast Mode-of-Action

- ❑ **963** compounds
- ❑ **412** endpoints
- ❑ Original data collected in ~20 different human cell types[1]
- ❑ Binary yes/no for ML
- ❑ Have hit call, AC50, and cell type information

Rat *in vivo* liver outcomes

- ❑ **471** compounds
- ❑ **63** endpoints
- ❑ ToxValDB[2], TG-GATES[3], DrugMatrix[4]
- ❑ 4 categories
- ❑ Binary yes/no
- ❑ Associated Benchmark Dose/Lowest Observed Adverse Effect Level

Human DILI

- ❑ **627** compounds
- ❑ **4** endpoints
- ❑ DILIrank 2.0[5]
- ❑ **No,** Less, **More**, Ambiguous
- ❑ Observational human data: limited info on dose and frequency
- ❑ Very broad biological category

- ❑ **Current curation is mostly based on manual data preprocessing and cleaning**
- ❑ **Automated toxicity endpoint data curation using large-language models and AI agents is an ongoing effort led by Ewald lab**



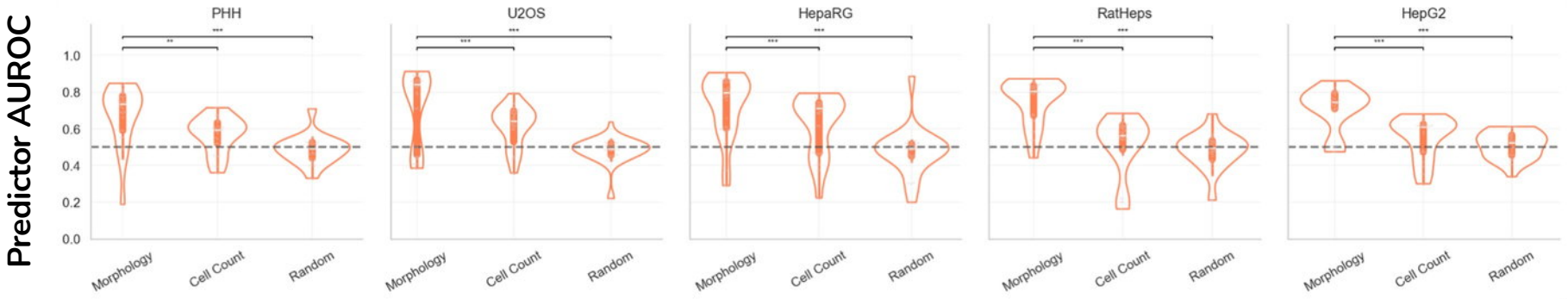
Jess Ewald

PURDUE UNIVERSITY

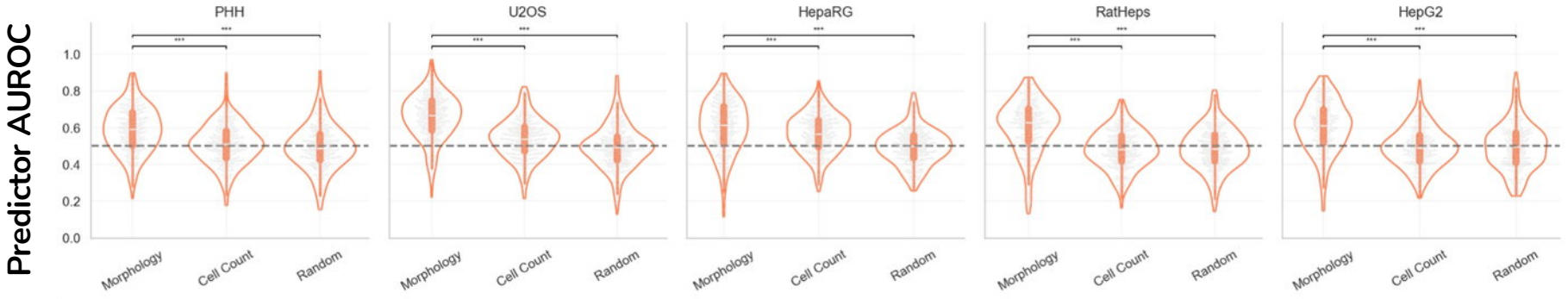
[1] U.S. EPA. ToxCast & Tox21 Summary Files from invitrodb_v4.1; [2] Patlewicz, G., et al. (2025). *Development of the Toxicity Values Database*; [3] Igarashi, Y., et al. (2015). *Nucleic Acids Research*; [4] Ganter, B., et al. (2005). *Nucleic Acids Research*; [5] Olubamiwa et al., 2025, *Drug Discov Today*

Morphological profiles outperform cell count and random baseline models for predicting cytotoxicity and drug mode-of-action

In-vitro cytotoxicity



Mode-of-action (cell-based assays)

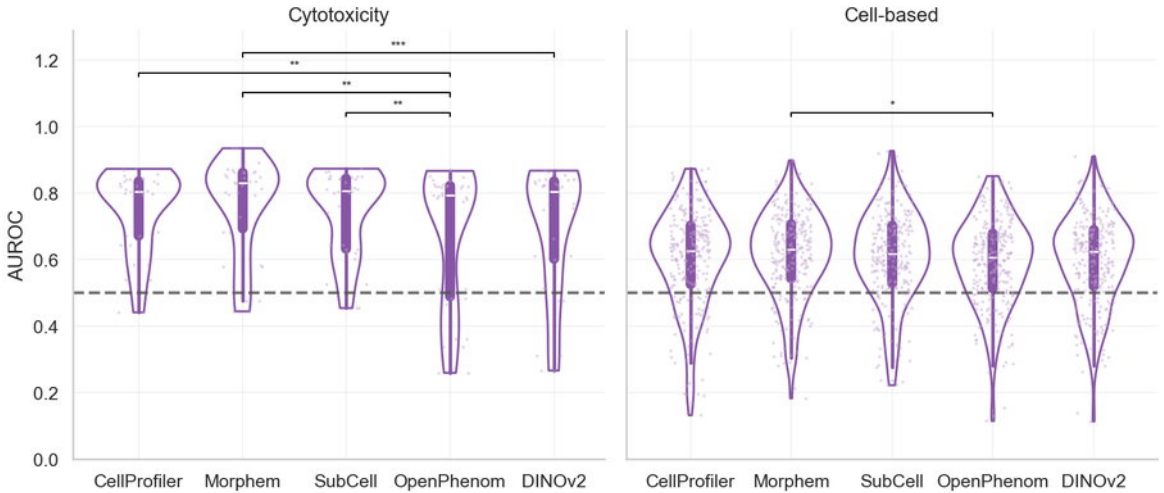


PURDUE UNIVERSITY

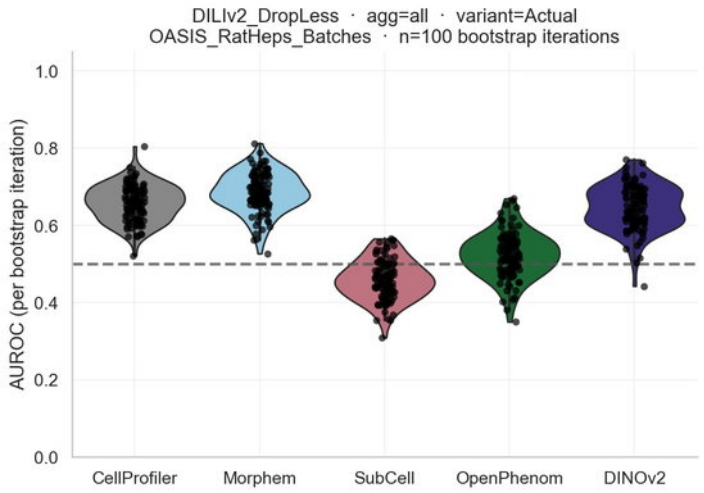
Ewald et al. (2026) Cell Systems

MorphEm and CellProfiler lead in predicting cytotoxicity, mode-of-action and human drug-induced liver injury

Cytotoxicity and Mode-of-action predictions



Human DILI predictions



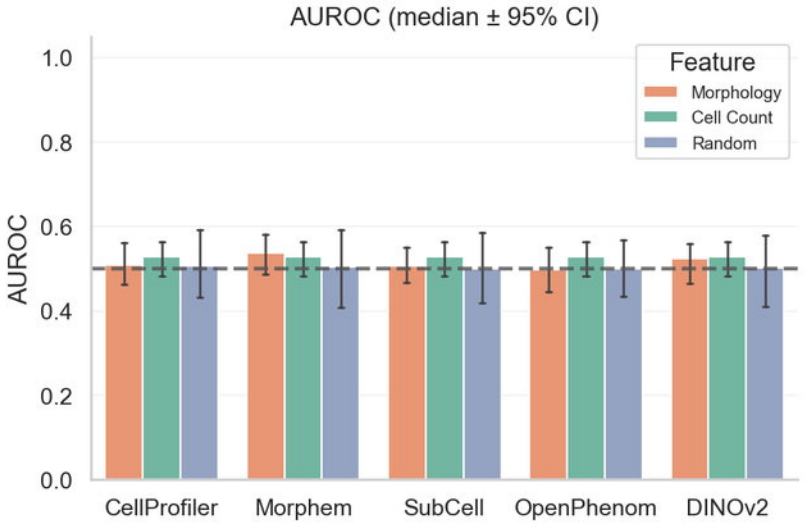
PURDUE UNIVERSITY

Results are preliminary: the toxicological labels currently curated are still not “perfect” and we are still cleaning them up with LLM and agentic workflows.

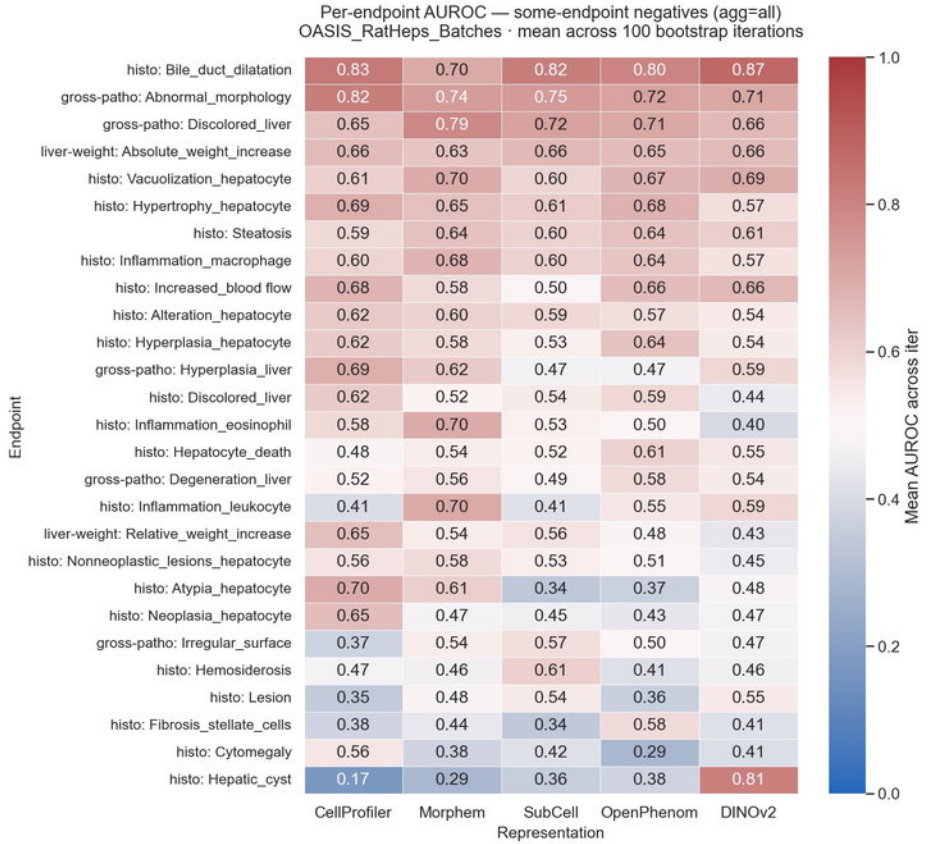
Gupta et al. *bioRxiv*, 2025; Agrawal et al. *ICLR*, 2026; Kraus et al., *ICLR*, 2025; Oquab et al. *Transactions on Machine Learning Research*, 2024

Deep learning features uniquely capture a few rat tox endpoints

Prediction of any rat in-vivo hepatic endpoint



Distinguish rat hepatotoxicity endpoints from each other



Results are preliminary: the toxicological labels currently curated are still not “perfect” and we are still cleaning them up with LLM and agentic workflows.

Gupta et al. *bioRxiv*, 2025; Agrawal et al. *ICLR*, 2026; Kraus et al., *ICLR*, 2025; Oquab et al. *Transactions on Machine Learning Research*, 2024

Lessons learned

- Figure out your biological questions
- Define your evaluation tasks clearly – what do you want to achieve?
- Pick the models that best fit your needs and choose the trade-off for your tasks (e.g., raw vs compressed images)
- Fine-tuning the models helps if the model's training datasets don't perfectly match your own data
- There is no model that excels at all the tasks – try a few with our benchmarking framework before you commit to a single one ☺
- Think about how to interpret your results at the end

Carpenter–Shen Lab



Anne E. Carpenter



Runxi Shen

Many thanks to two decades of Carpenter–Singh lab alumni and to our many collaborators, especially in Beth Cimini's lab!



Shantanu Singh



Alán Fernando Muñoz González



Jonne Rietdijk



Johan Fredin Haslum



Kazumasa Shimofuruta

Hiring at Purdue University!
Postdoc openings in data science/ML
Carpentershenlab.org/recruiting/

NIH NHGRI

iGVF

OASIS
Omics for Assessing Signatures for Integrated Safety Consortium

PURDUE UNIVERSITY

Download PDF of slides here: http://runxi-shen.github.io/talks/SBI2-2026.pdf