Ardigen

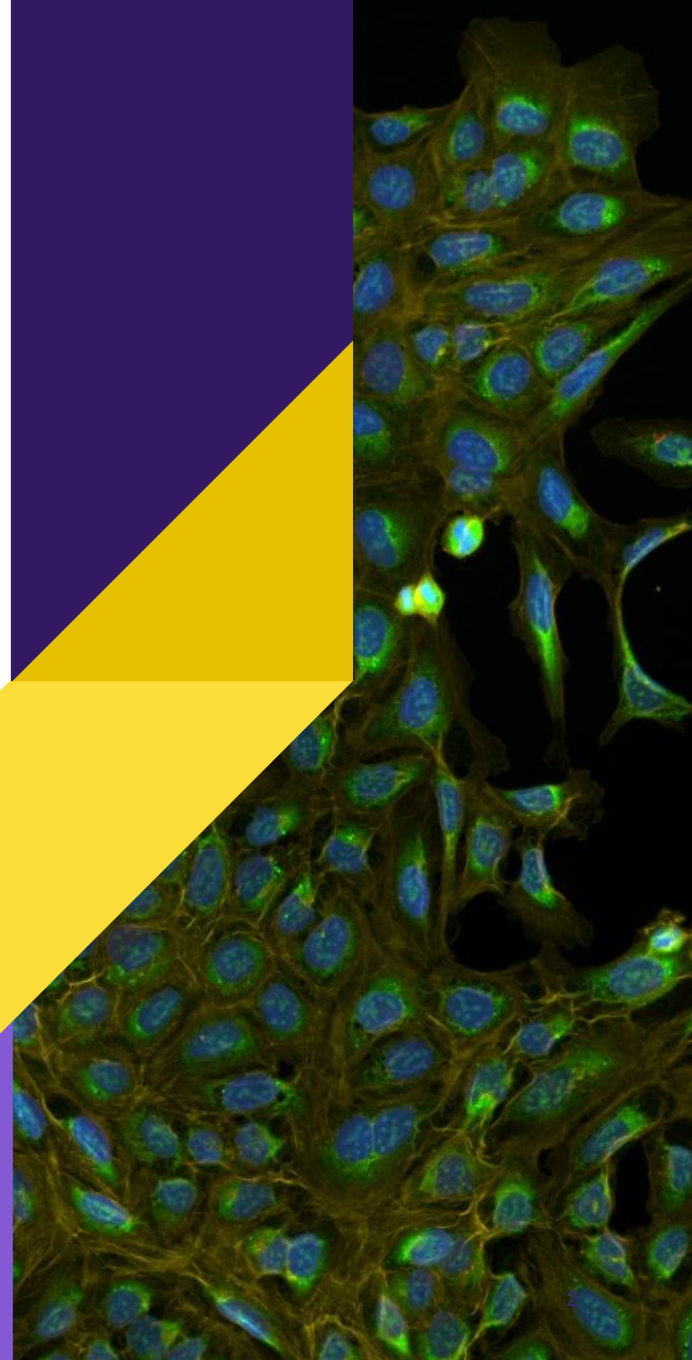
Deep Learning for effective analysis of High Content Screening

Adriana Borowa
Senior Data Scientist
8.11.2024

ML in PL
CONFERENCE 2024

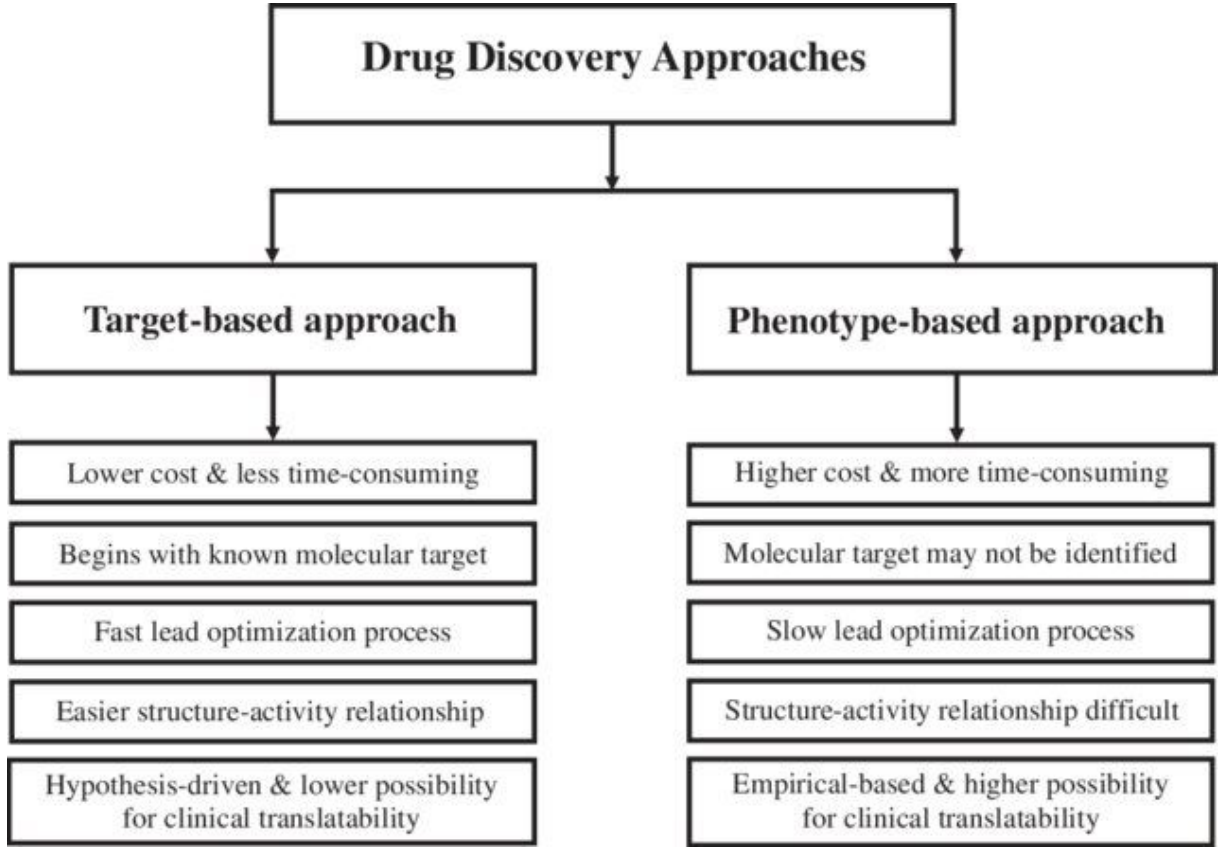
Deep Learning for effective analysis of High Content Screening: Agenda

1. Phenotypic screening introduction
2. Data: High Content Screening
3. Data: Challenges
4. Applications
5. Summary



Phenotypic screening: introduction

Target vs phenotypic screening

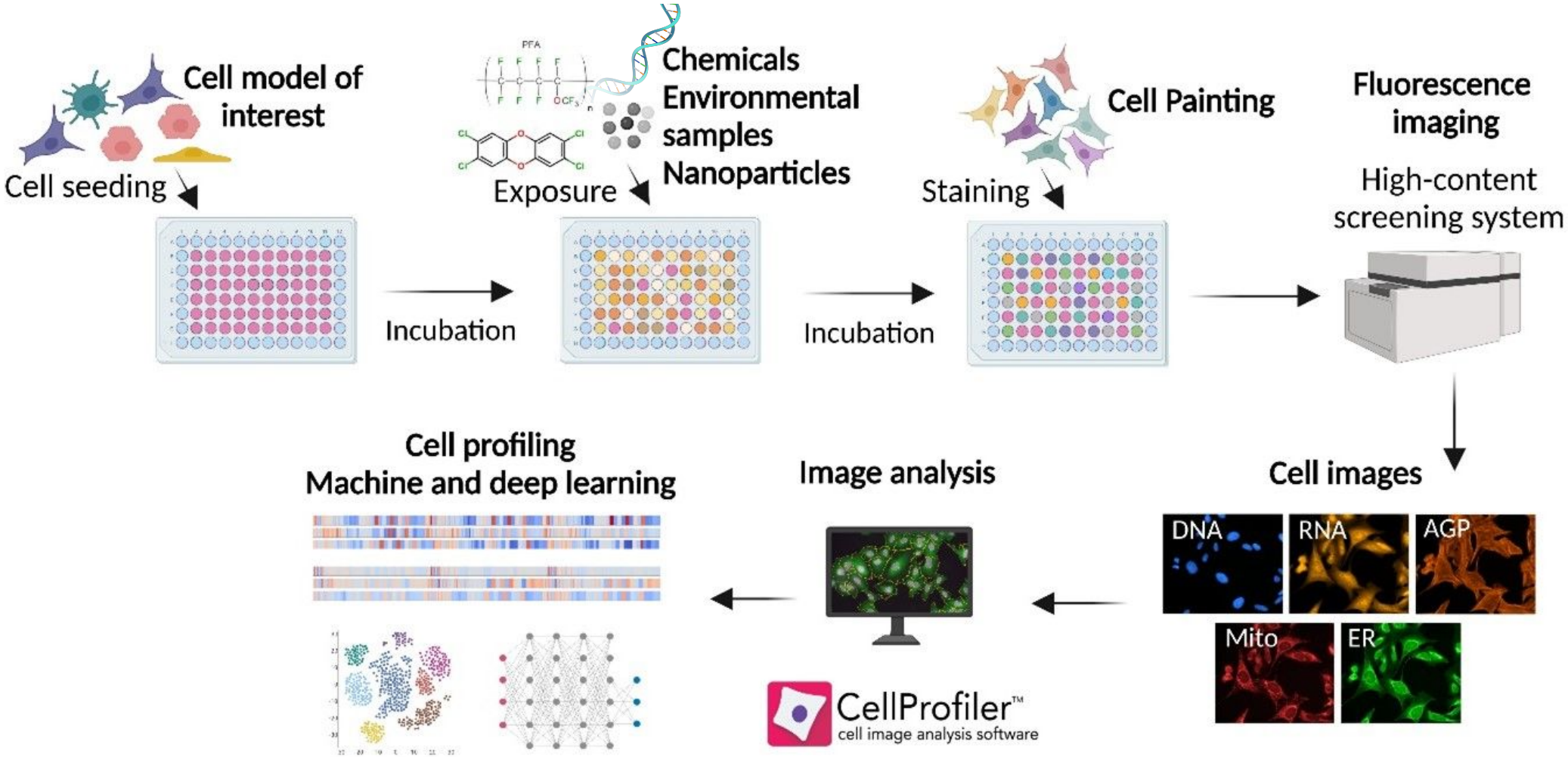


Source: Singh, Vijay & Seed, Thomas & Ayodele, Okedele. (2019). Drug discovery strategies for acute radiation syndrome. Expert Opinion on Drug Discovery. 14. 1-15. 10.1080/17460441.2019.1604674.

Data: High Content Screening

Ardigen

Acquisition



Source: https://www.oru.se/english/research/research-projects/rp/?rdb=p2550

Data: High Content Screening

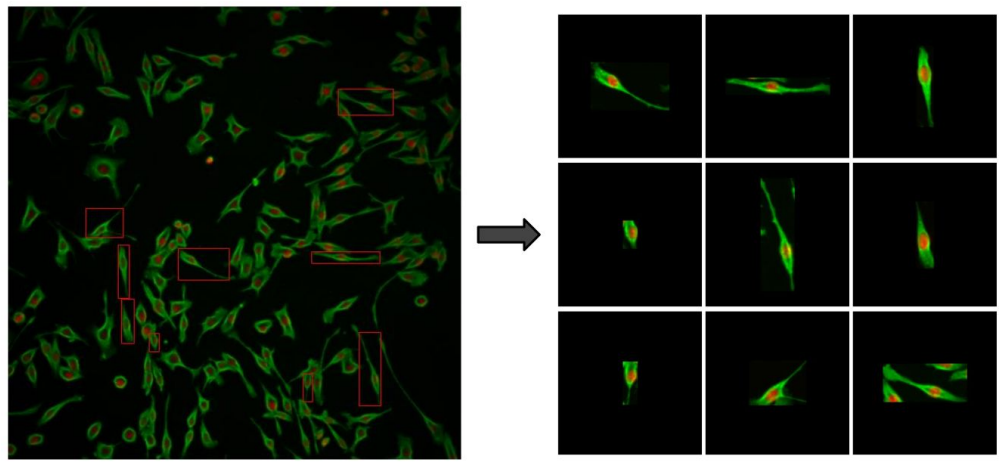
Ardigen

2-dye assay for inflammation assessment

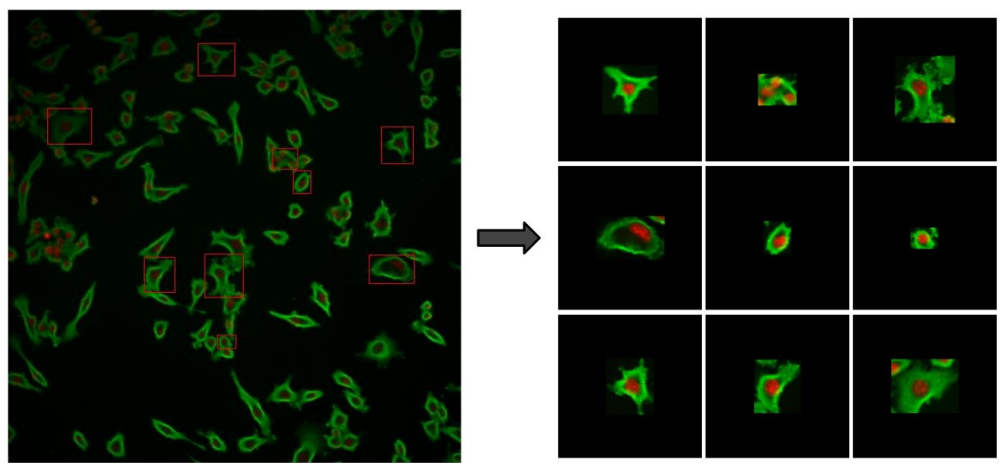
Red: nucleus

Green: plasma membrane

Control



Inflammation



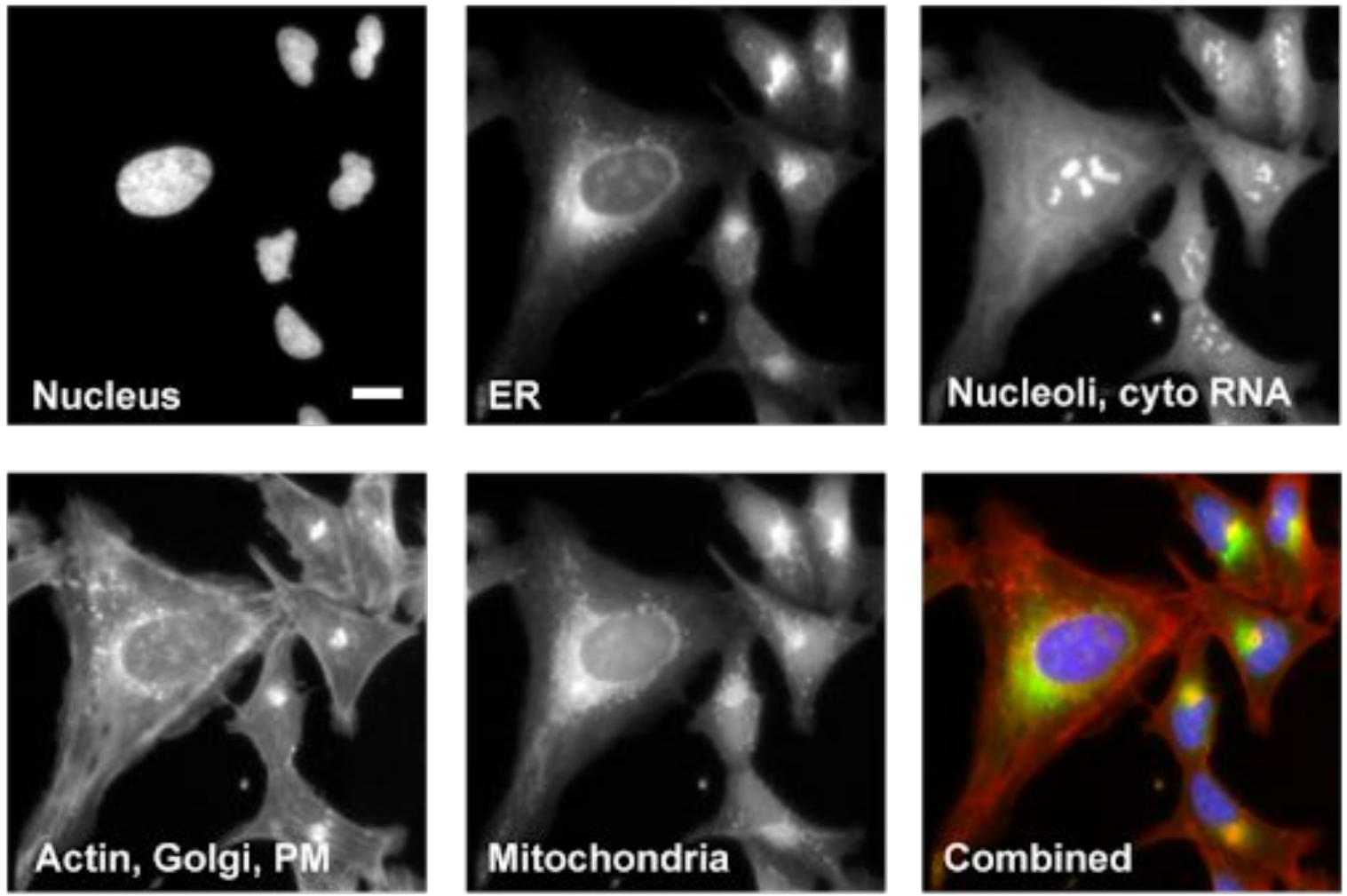
Source: Borowa et al. *Weakly-Supervised Cell Classification for Effective High Content Screening,* ICCS 2022

Data: High Content Screening

Ardigen

Cell Painting

- 6 stains
- 5 channels
- 8 cellular components



Source: https://jump-cellpainting.broadinstitute.org/

Data: Challenges

Ardigen

1. Large amounts of data: needle in a haystack
2. Standardization and reproducibility
3. Imbalanc
4. Interpreta
5. Complexi

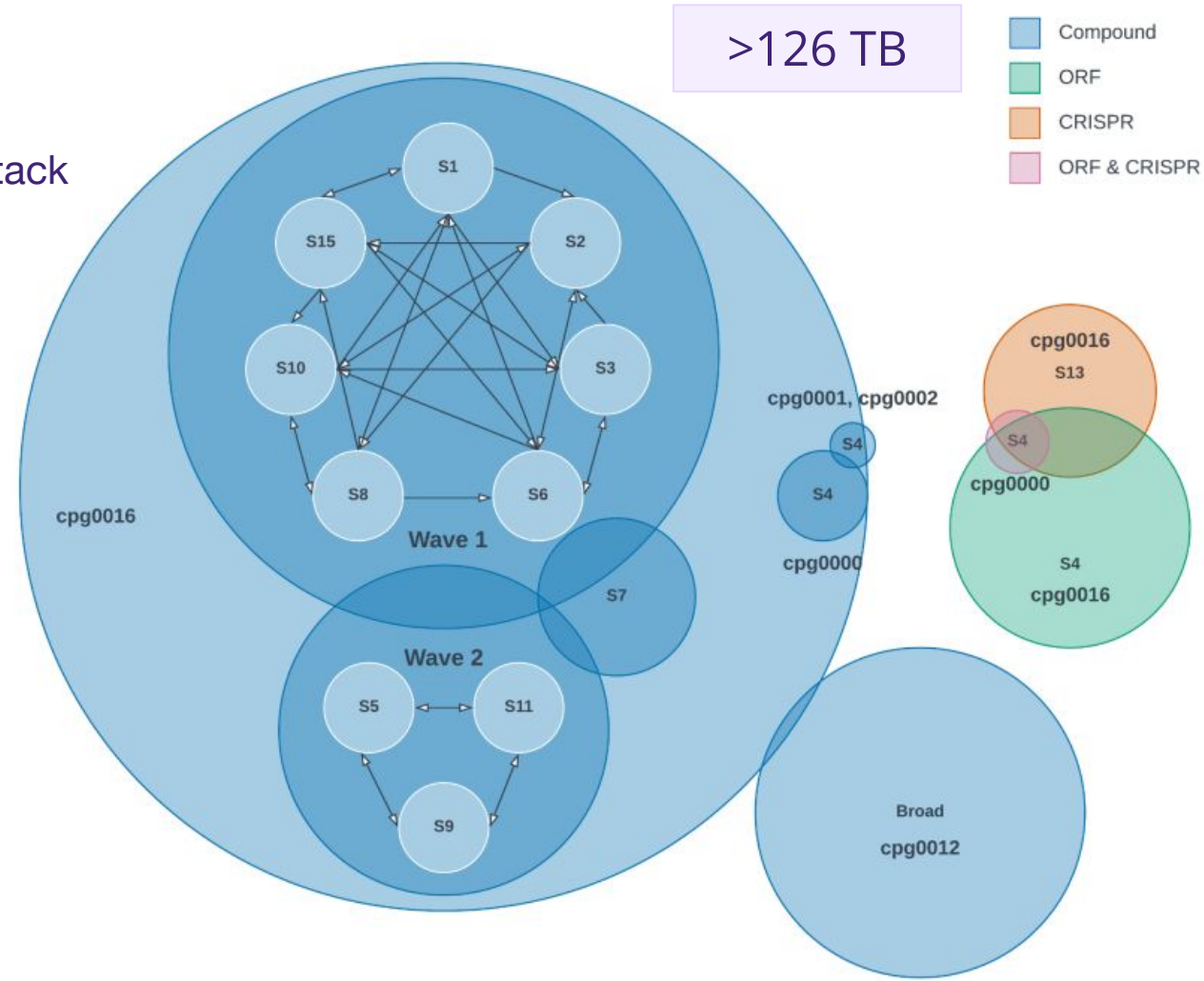
	RxRx3 Download Read More	RxRx2 Download Read More	RxRx19b Download Read More	RxRx19a Download Read More	RxRx1 Download Read More
Release Date	January 2023	August 2020	April 2020	August 2020	January 2023
Categories	Genetic Perturbations (genome-wide CRISPR) Small molecules	Large molecules	Infectious disease Small & large molecules	Infectious disease Small molecules	Genetic perturbations (siRNA)
Total Number of Images	~2.2M	131,953	70,384	305,520	125,510
Image Dimension	2048×2048×6	1024×1024×6	2048×2048×6	1024×1024×6	512×512×6
Compressed Dataset Size	~83,100 GB	~185 GB	~409 GB	~450 GB	~46 GB

Source: https://www.rxrx.ai/

Data: Challenges

Ardigen

1. Large amounts of data: needle in a haystack
2. Standardization and reproducibility
3. Imbalanced datasets and lack of labels
4. Interpretability
5. Complexity of data and off-targets



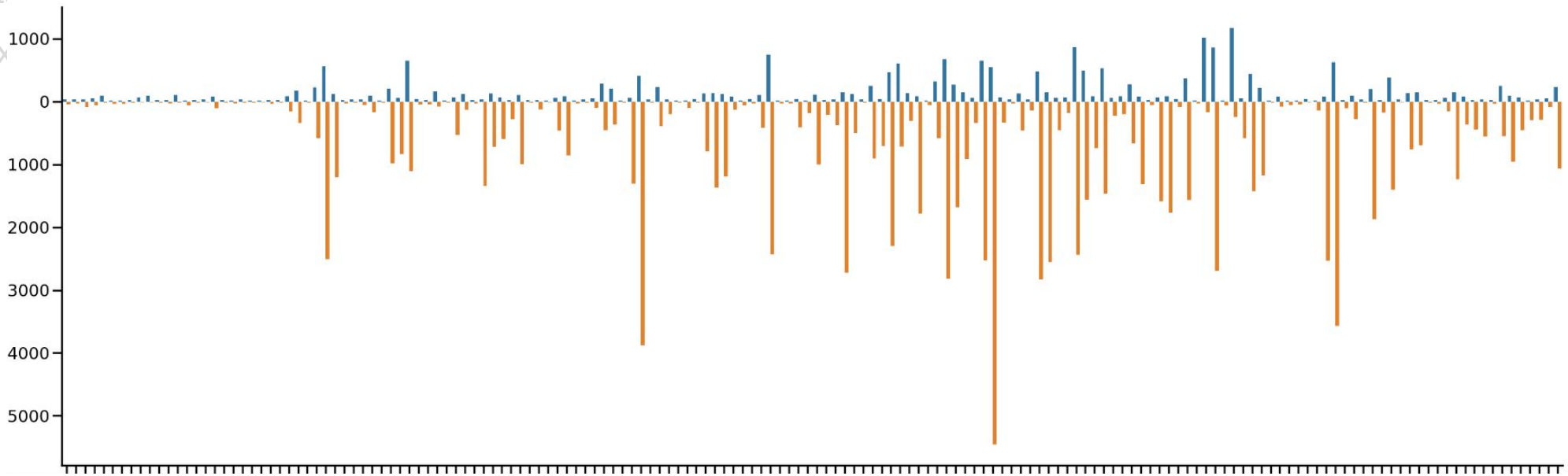
Source: Chandrasekaran et al. *JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations*
https://www.biorxiv.org/content/10.1101/2023.03.23.534023v2.full.pdf

Data: Challenges

Ardigen

1. Large amounts of data: needle in a haystack
2. Standardization and reproducibility
3. Imbalanced datasets and lack of labels
4. Interpretability
5. Complexity

Public labels for the dataset of 30,000 compounds

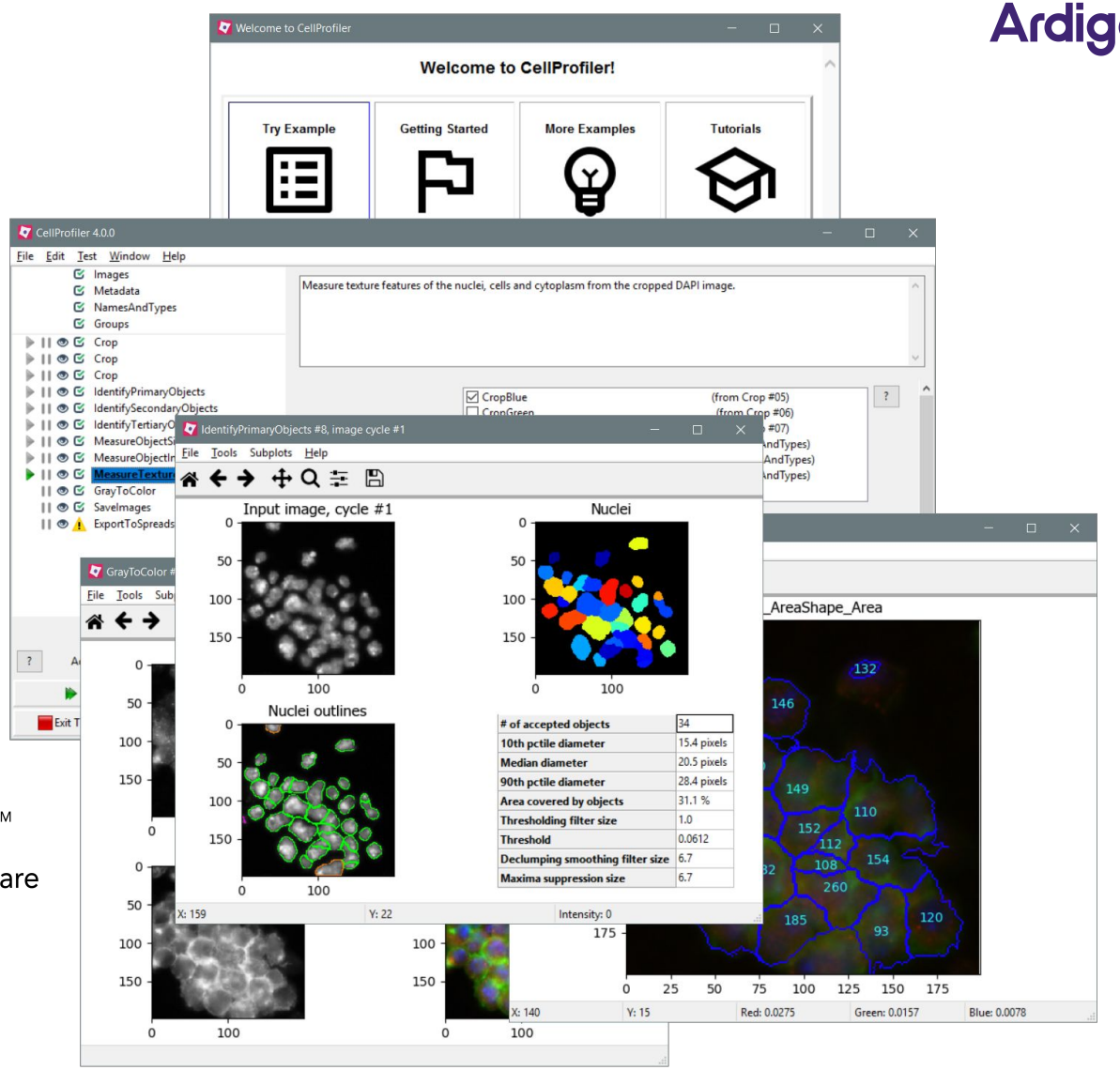


Bray et al. *A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay* GigaScience, Volume 6, Issue 12, December 2017, giw014, https://doi.org/10.1093/gigascience/giw014

Data: Challenges

Ardigen

1. Large amounts of data: needle in a haystack
2. Standardization and reproducibility
3. Imbalanced datasets and lack of labels
4. Interpretability
5. Complexity of data and off-targets



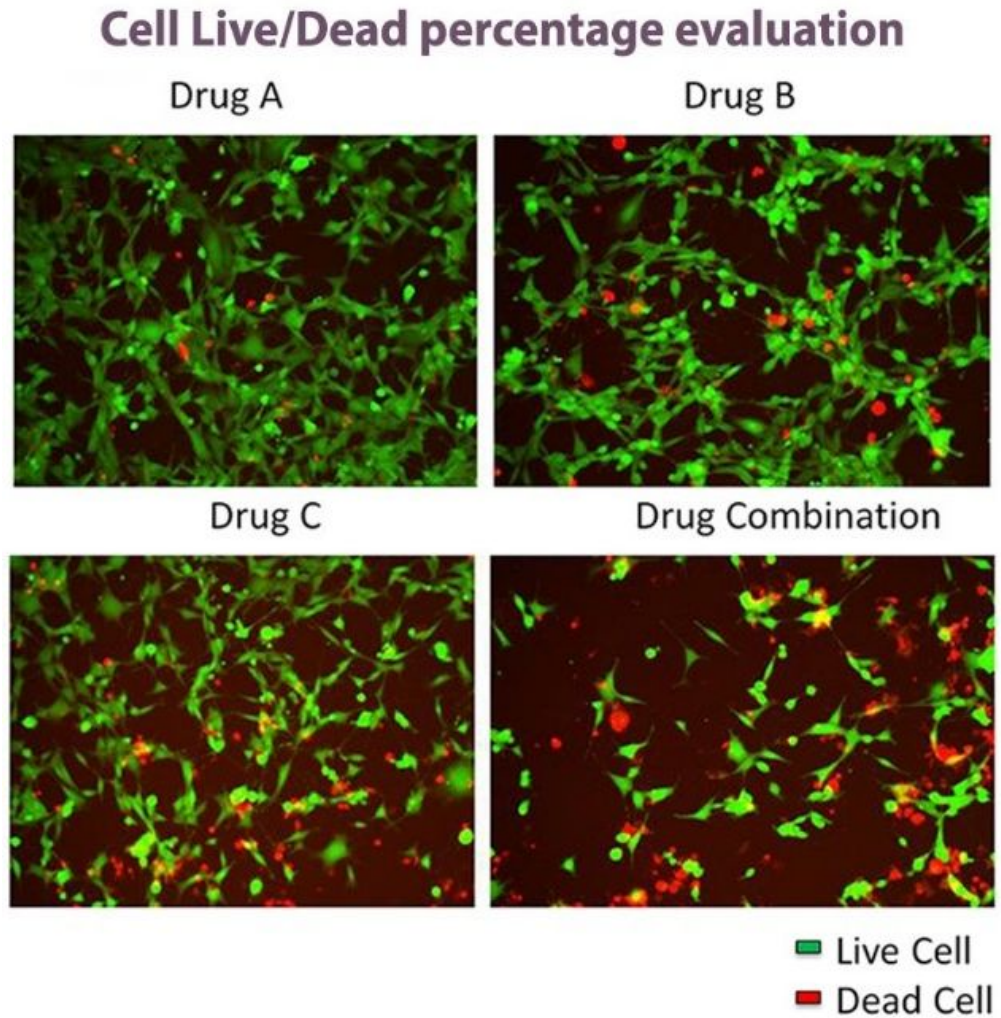
CellProfiler™
cell image analysis software

Source: https://cellprofiler.org/

Data: Challenges

Ardigen

1. Large amounts of data: needle in a haystack
2. Standardization and reproducibility
3. Imbalanced datasets and lack of labels
4. Interpretability
5. Complexity of data and off-targets

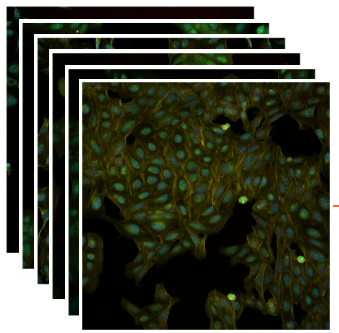


Source: https://idea-bio.com/automating-toxicology-studies-high-content-screening/

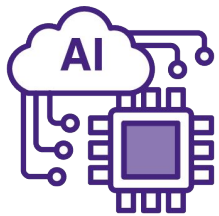
Applications

Hit identification

Divergent from negative reference

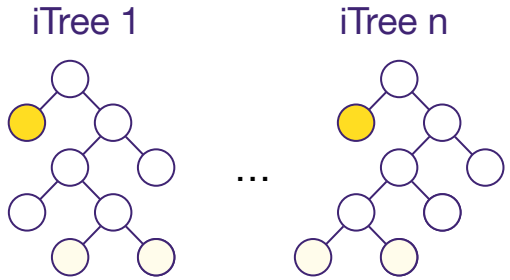


Negative control from multiple wells



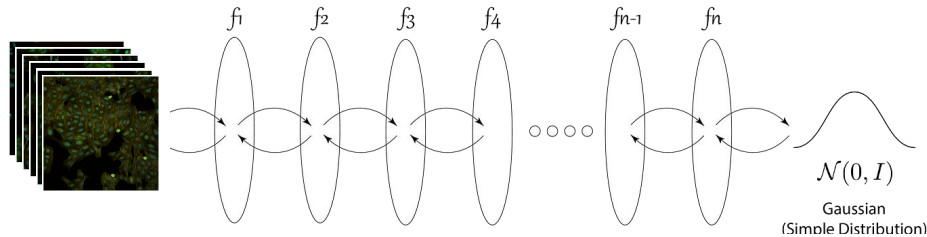
Divergence from negative control

Isolation Forest

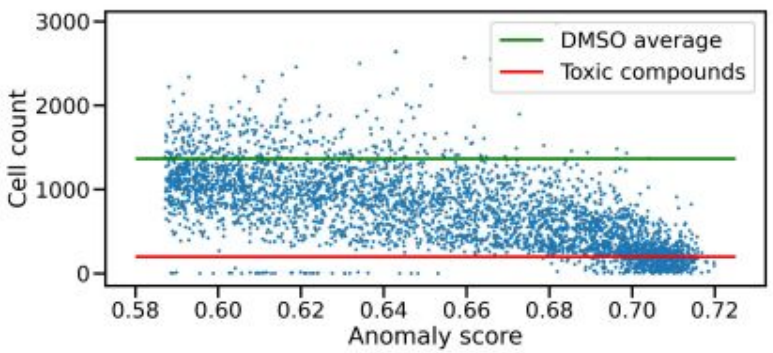


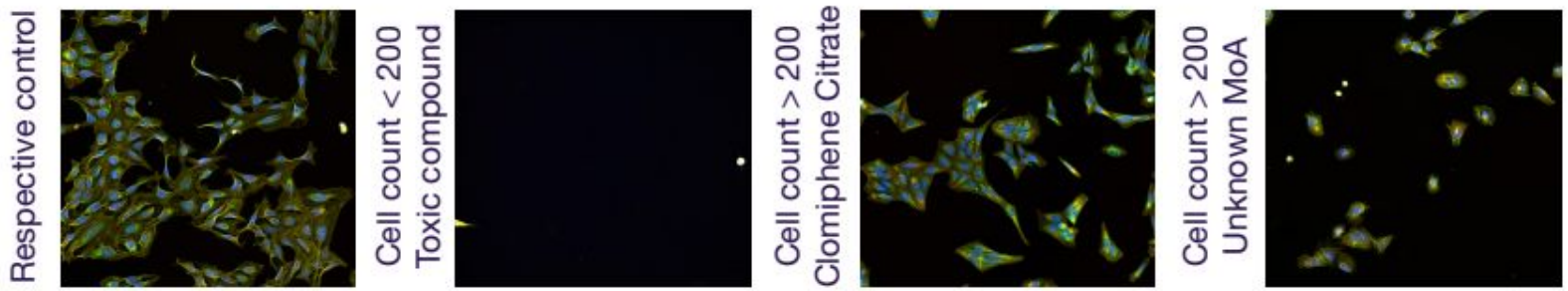
Ardigen

Sequential Flows



NLL SCORE = $-\text{NLL}(f_\theta^N(\ldots(f_\theta^1(x))))$



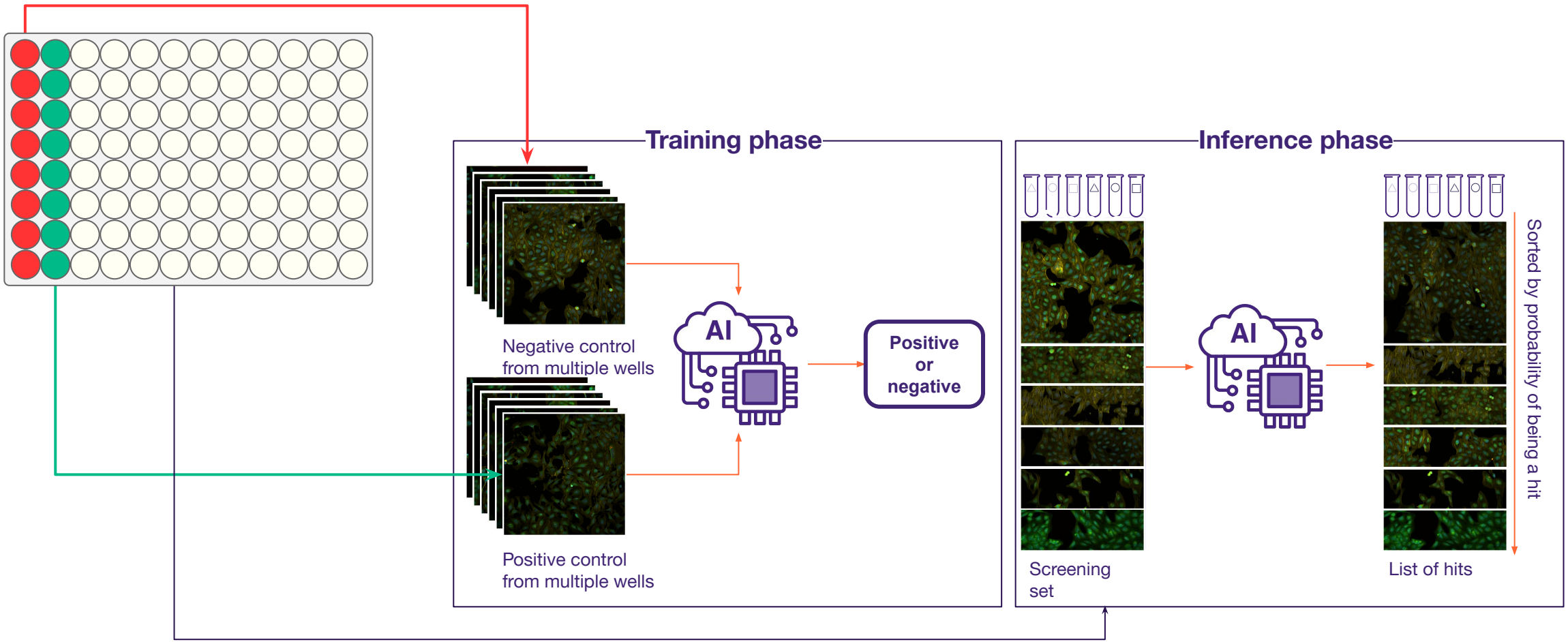


Source: Adamski et al. *AI-driven identification of hits from Cell Painting based screening*, Cytodata 2023

Applications

Hit identification

Ardigen

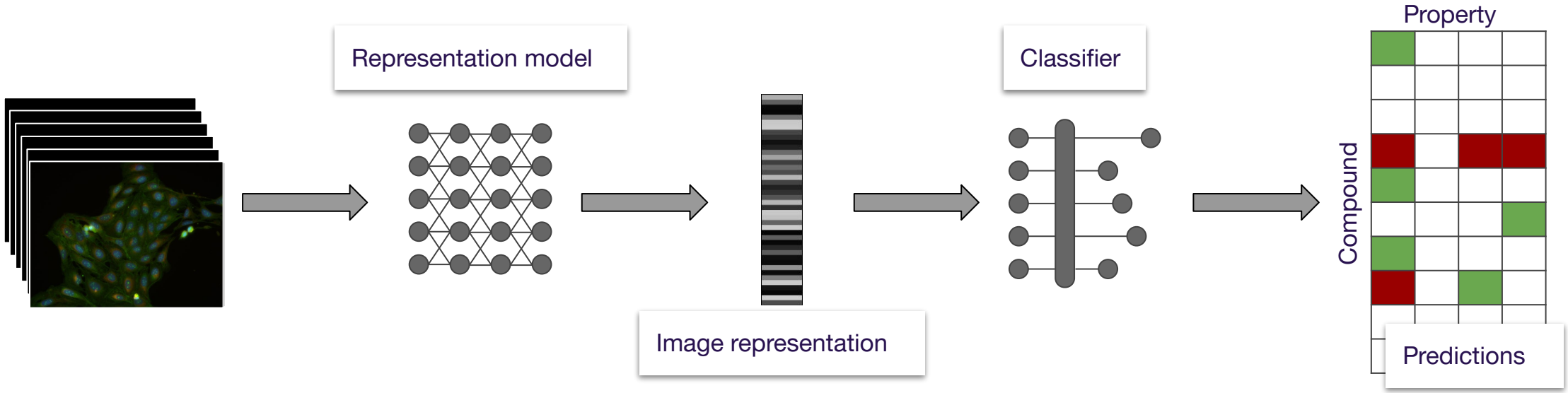


Source: Bracha et al. *AI-driven method for identification of hits from phenotypic screening with Cell Painting Assay*, SLAS Europe 2023

Applications

Property prediction

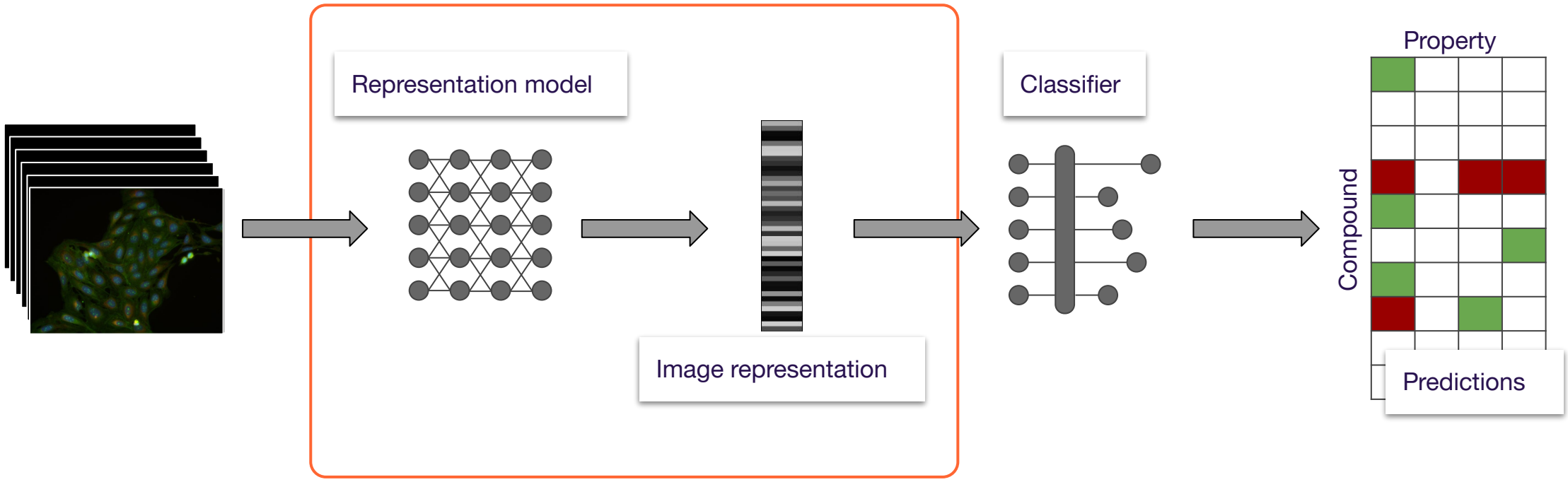
Ardigen



Applications

Property prediction

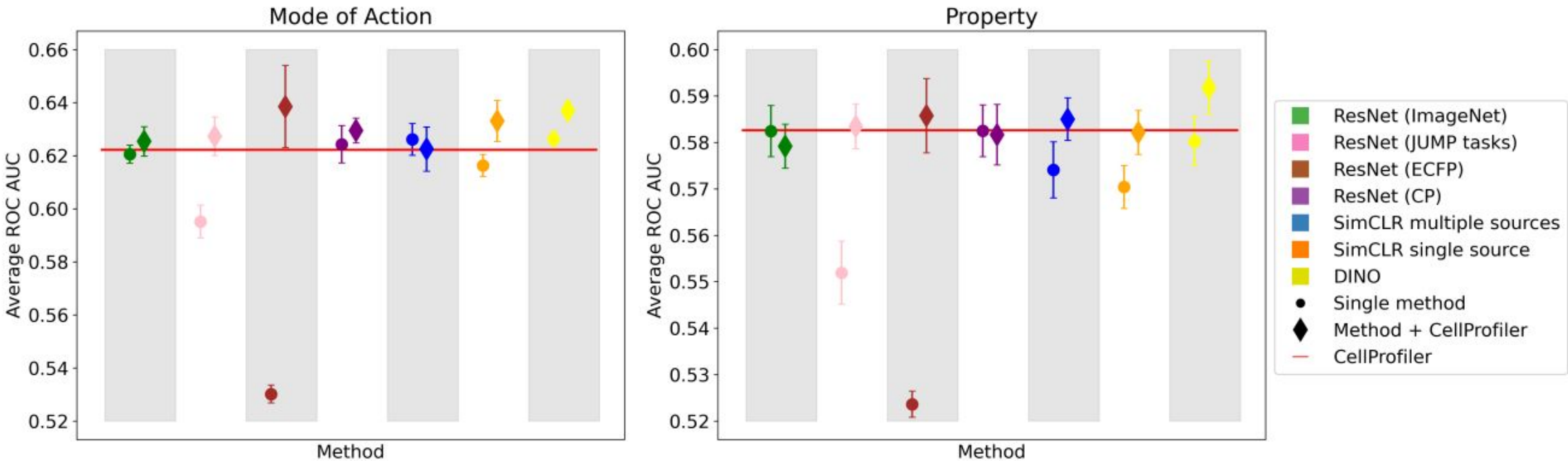
Ardigen



Applications

Image representation

Ardigen

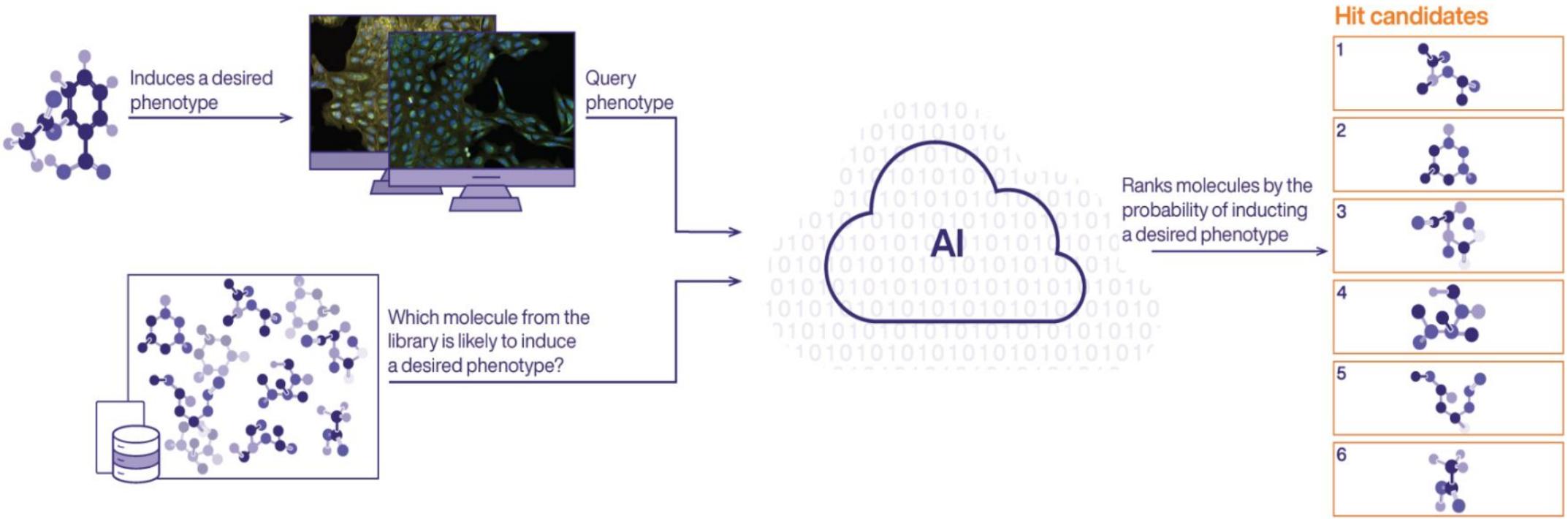


Source: Borowa et al. *Decoding phenotypic screening: A comparative analysis of image representations*, Computational and Structural Biotechnology Journal, Volume 23, 2024

Applications

Ardigen

Phenotype induction: compound library screening

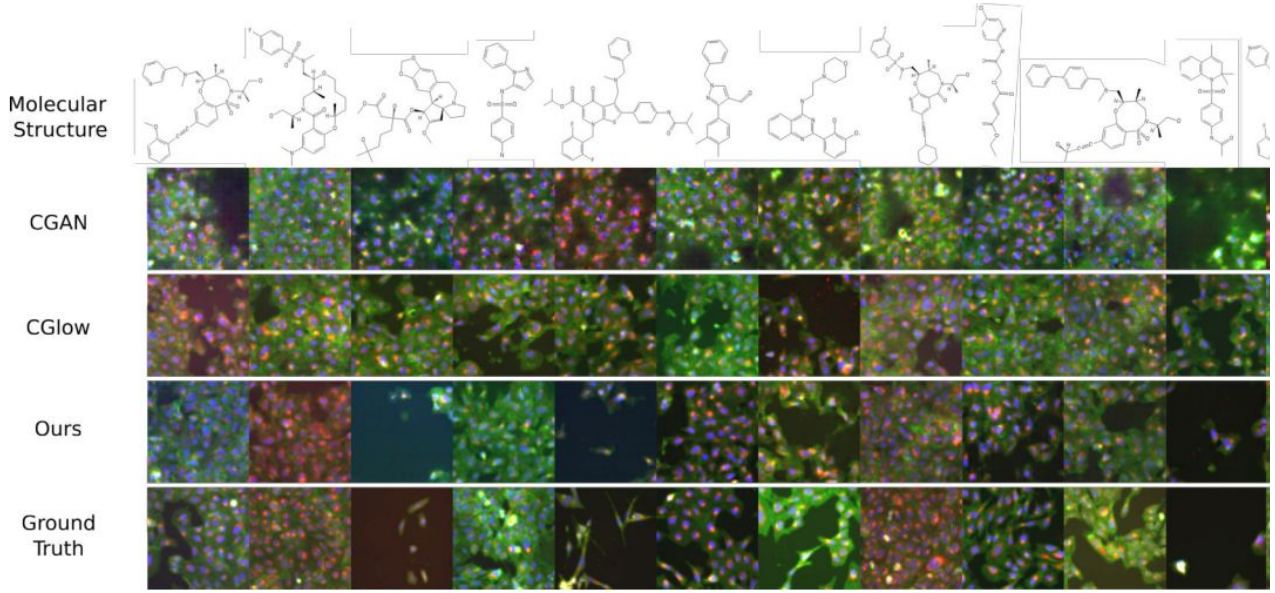


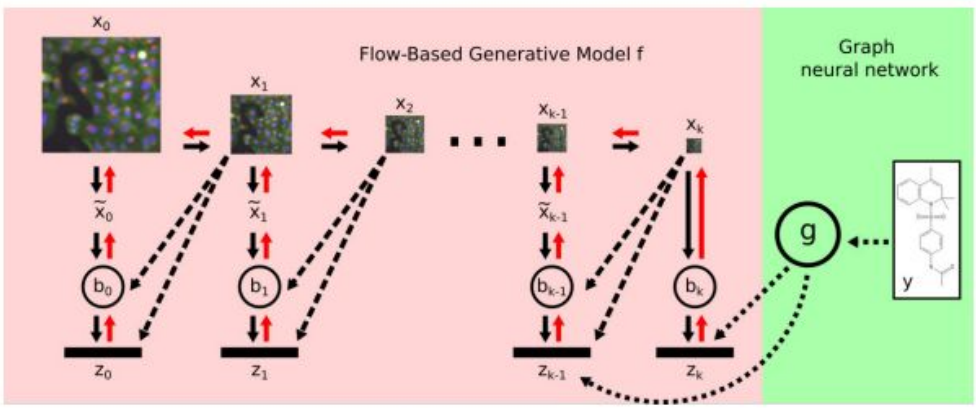
Source: Gaiński et al. *Artificial Intelligence enriches phenotype-guided Virtual Screening and proposes diverse hit candidates*, SLAS 2023

Applications

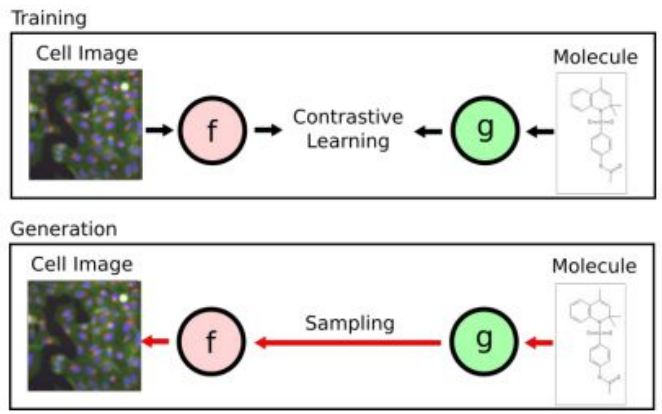
Ardigen

Phenotype prediction from molecule





(a) Model Architecture

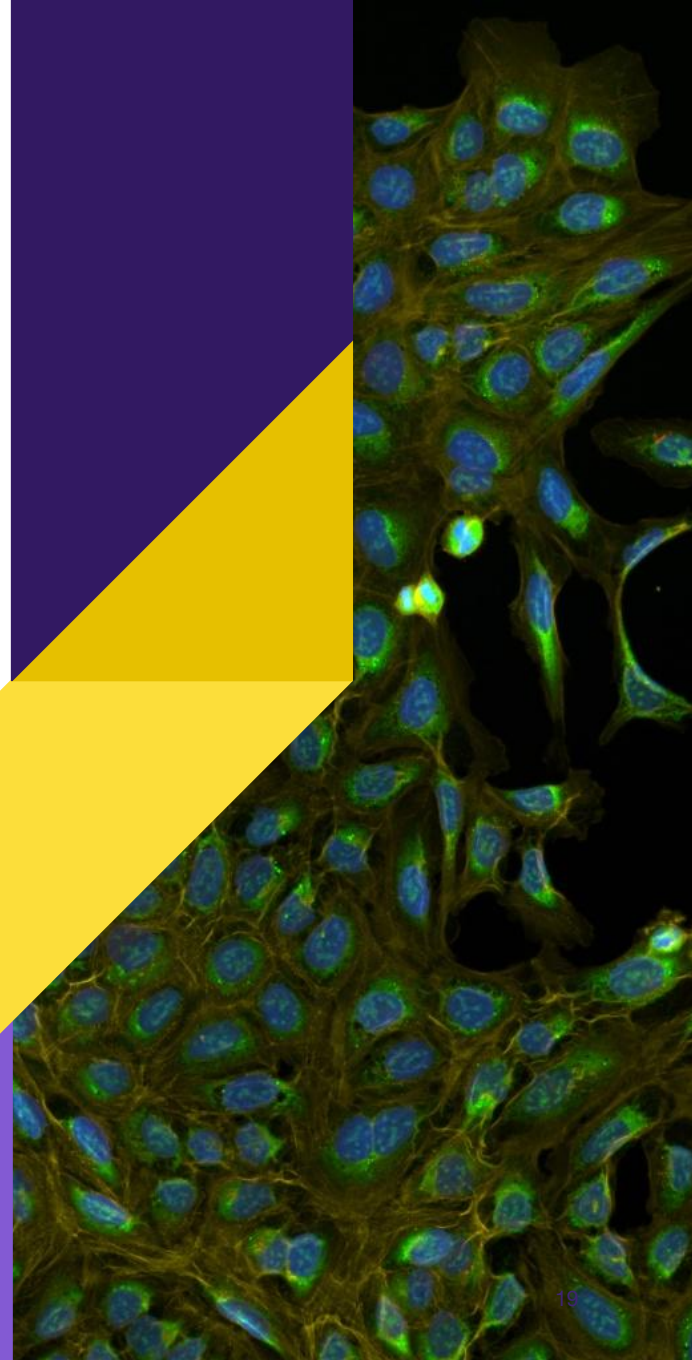


(b) Training Strategy

Source: Yang et al., *Mol2Image: Improved Conditional Flow Models for Molecule to Image Synthesis*, CVPR 2021

Summary

- Scalability and efficiency: Deep Learning methods accelerate analysis of massive amounts of data generated for phenotypic screening
- Cost and time reduction
- Problem of small amount of labeled public data persists
- There is still a way to go for interpretability



Check out our other talks during ML in PL!

Friday:

Session 2 / Lecture Hall B / 10:35

Deep learning for effective analysis of high content screening
Adriana Borowa

Session 4 / Lecture Hall A / 14:30

Efficient fine-tuning of LLMs: exploring PEFT methods and LORA-XS insights
Klaudia Bałazy

Session 5 / Lecture Hall B / 14:30

Current trends in intrinsically interpretable Deep Learning
Dawid Rymarczyk

Neural rendering: the future of 3D modeling
Przemysław Spurek

JAGIELLONIAN UNIVERSITY IN KRAKÓW

group of machine gmum learning research

Saturday:

Session 7 / Lecture Hall A / 12:00

AdaGlimpse: Active Visual Exploration with Arbitrary Glimpse Position and Scale
Adam Pardyl

Session 8 / Lecture Hall B / 12:00

Augmentation-aware Self-supervised Learning with Conditioned Projector
Marcin Przewięźlikowski

gmum.net

Thank you!

Adriana Borowa
Senior Data Scientist
Ardigen SA

adriana.borowa@ardigen.com