Machine-learning prediction of drug mechanism of action from high-content cell painting images.

Authors

Max Blanck
Revvity, Inc.

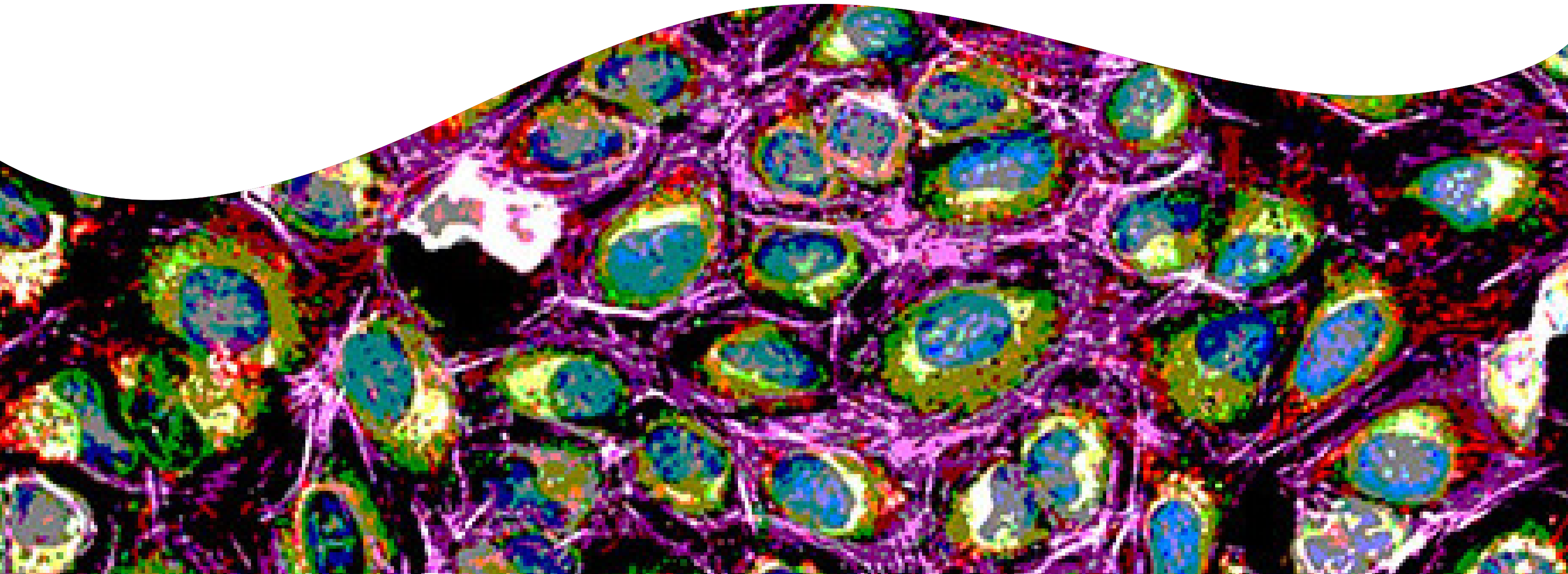
Introduction

Understanding a new therapeutic candidate's mechanism of action (MOA) helps rationalize its effects on biological systems, which is fundamental for developing effective and safe drugs. Accurate MOA identification can improve on-target efficacy and reduce potential side effects caused by off-target interactions, thereby helping to increase the success rate of clinical trials. Moreover, predicting MOAs can streamline the drug development process by identifying promising drug candidates earlier, saving time and resources.

Computational approaches like machine learning (ML) can efficiently predict MOAs, providing insights that guide experimental validation. This integration of computational and experimental methods enhances the overall efficiency and effectiveness of drug discovery[1].

ML algorithms can rapidly analyze large datasets, significantly speeding up the process. In microscopy, deep learning models can detect subtle patterns and changes in cellular morphology that might be missed by human analysis, leading to more accurate identification of MOAs.

ML can uncover new and unexpected MOAs by analyzing vast amounts of complex biological data, such as high-content imaging. ML can handle and process these large datasets efficiently, making it possible to scale up drug discovery efforts and provide novel insights into how drugs interact with biological systems. Additionally, automating the analysis of high-content imaging data reduces the need for extensive manual labor and expensive biochemical assays, lowering the overall cost of drug discovery[2].



High-content imaging in drug discovery

High-content imaging (HCI) is a transformative technique in biological research, especially in cell biology, pharmacology, and drug discovery. It allows researchers to simultaneously study multiple phenotypic responses at the cellular level, enabling large-scale analysis of complex biological processes.

The strength of HCI lies in its ability to capture high-resolution images quickly and efficiently under various experimental conditions, such as different drug concentrations and treatment times. Automated microscopy systems take images of cells in microtiter plates, which are then analyzed using algorithms to extract quantitative features like cell shape, size, texture, and intensity distributions across cellular compartments (e.g., nucleus, mitochondria, Golgi apparatus, endoplasmic reticulum, lysosomes). This results in a rich, high-dimensional dataset that allows for detailed examination of diverse cellular phenotypes[3].

Image-based profiling has advantages over traditional biochemical assays. It can measure multiple phenotypic outcomes simultaneously, offering a holistic view of cellular biology. This is particularly useful for studying complex processes like cell differentiation or apoptosis. It is also highly scalable, making it ideal for high-throughput drug screening, allowing researchers to quickly identify compounds with desirable effects. A notable image-based profiling method is cell painting. This technique uses multiple fluorescent dyes to label different cellular compartments, providing a comprehensive view of cellular architecture and function. It extracts detailed morphological profiles from cellular images to classify compounds based on their phenotypic effects[4].

Deep learning models in imaging for drug discovery

High-content imaging is a data-intensive methodology and thus presents computational and analytical challenges. Recent advancements in ML techniques, such as convolutional neural networks (CNNs), have improved the capacity and accuracy of phenotype classification by cell painting[5]. Applying semi-supervised models facilitates generalization across different experimental conditions, enhancing robustness to batch effects and variations in imaging protocols[6].

This application note discusses the implementation process of a deep learning pipeline for applying HCI in drug discovery, focusing on the study of the mechanisms of action of chemical compounds. By implementing a semi-supervised training method, accuracy is improved while preserving the ability to identify unknown MOAs, thereby providing a more accurate and flexible framework for phenotypic analysis.

Methods

The study used the ground truth labeled subset of the BBBC021 cellular imaging dataset from the Broad Bioimage Benchmark Collection[1]. The dataset comprises high-content images of MCF-7 breast cancer cells treated with 104 treatments with 13 MOAs for 24 hours (Figure 1). Cells were stained with three fluorescent markers: Hoechst 33342 for DNA, a tubulin marker for microtubules, and a fluorescent phalloidin marker for actin. Images were captured in three corresponding channels from 96-well plates processed across ten experimental batches.

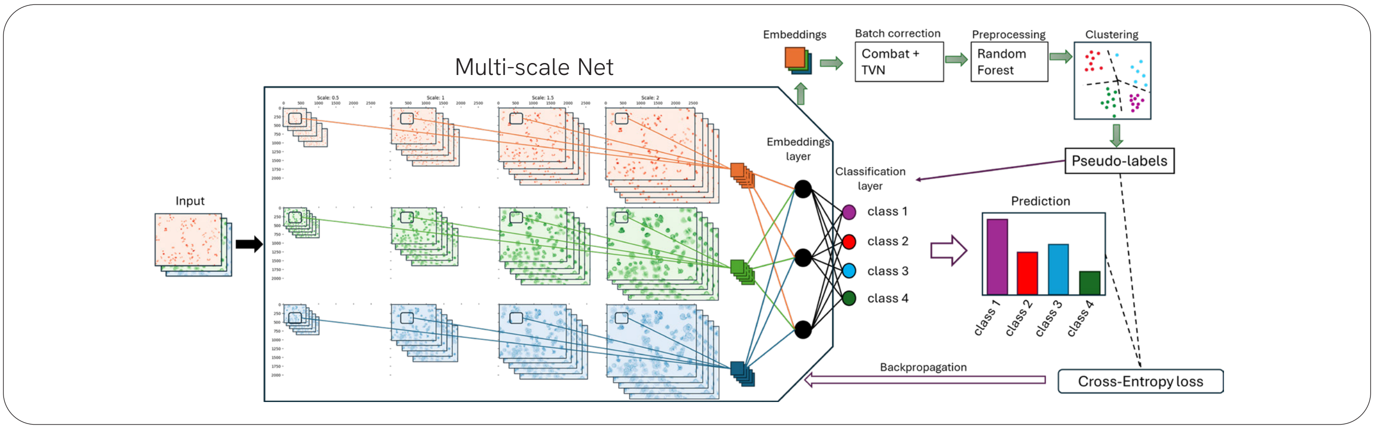


Figure 1: **Methods Summary.** Data from the Broad Bioimage Benchmark Collection dataset BBBC021, which includes 104 treatments with 13 mechanisms of action (MOAs) applied to MCF7 cells, were analyzed. The dataset comprises 39,600 image files (13,200 fields of view imaged in three channels). Pre-processing steps included data normalization and batch correction. Model training was performed using a modified version of UMM-Discovery incorporating a random forest approach to compute the proximity matrix, followed by k-medoid clustering. Model performance was evaluated based on accuracy, clustering results, and a confusion matrix.

Data processing

The Semi-supervised Multi-scale Mode-of-action Discovery (SMM-D) algorithm was applied directly to raw images without manual feature extraction, bypassing the need for prior knowledge of cellular phenotypes and making the process data agnostic[7]. SMM-D combines DeepCluster[8] and a Multi-Scale Neural Network[5] to capture both local cellular features and global effects like cell proliferation.

The neural network processed unannotated images to produce a 64/128-dimensional feature vector for each image. The features were then normalized and corrected for batch effects.

Model training

A Random Forest Classifier was trained on the batch-corrected and normalized feature vectors of a subset of the dataset using true labels for MOA, and the model was applied to the whole dataset. A proximity matrix was computed to capture the similarity between data points based on their co-occurrence in the same leaf nodes across all trees. Calculations were performed on GPU using CuPy.

The proximity matrix was fed into a k-medoid clustering algorithm to group data points based on phenotypic similarity. The number of clusters was set to the number of treatments, and k-medoids++ initialization was used to provide robust and reproducible clustering.

Cluster labels for each image were used as pseudo labels in the classification layer of the neural network and facilitated the measurement of model loss. Evaluation metrics such as Not-Same-Compound (NSC), Not-Same-Compound-and-Batch (NSCB), and Silhouette Scores were used to assess clustering quality. Combat and TVN were applied to account for batch effects across different plates and experimental conditions.

Results

The model was trained over 100 epochs, and prediction accuracy was measured by the k-nearest neighbor classifier (Figure 2). High accuracy, greater than 96%, was achieved for the NSB metric and with slightly lower performance for the NSCB metric, indicating that batch effects were largely removed.

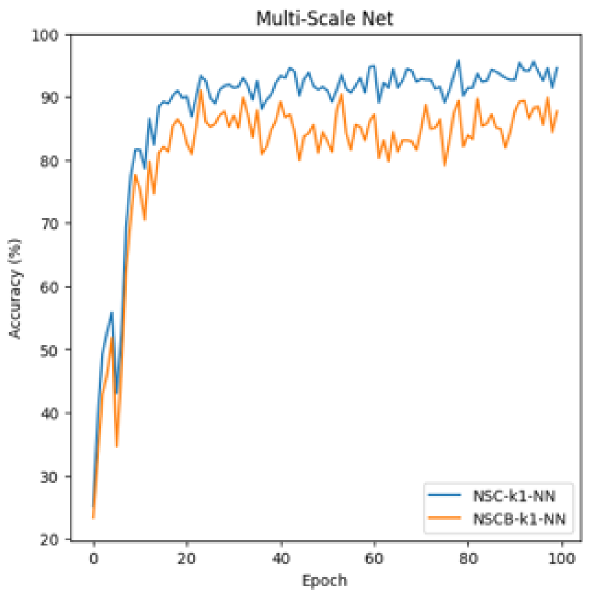


Figure 2: SMM-D accuracy by not same compound (NSC) and not same compound and batch over 100 epochs demonstrated the trained model’s high accuracy and consistency.

Well-level embeddings of the best epoch were visualized by UMAP projection (Figure 3) and showed that treatments exhibiting the same MOA clustered closely together. Highly accurate clustering was further confirmed by plotting the confusion matrix (Figure 4), which showed that 99 of 103 treatments were assigned to the correct MOA.

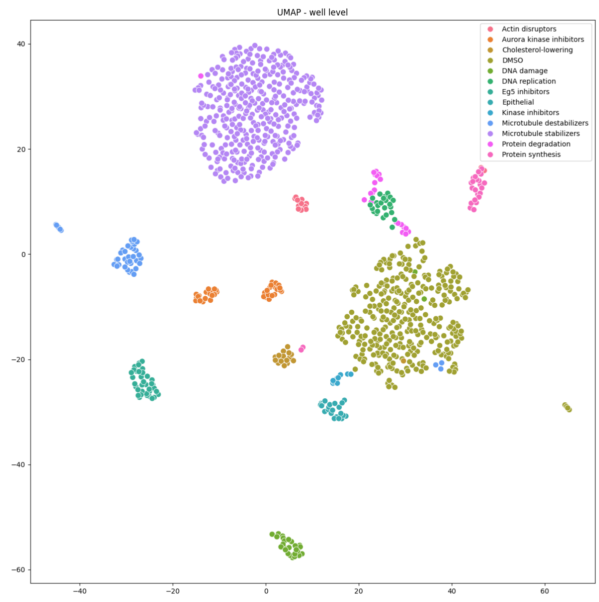
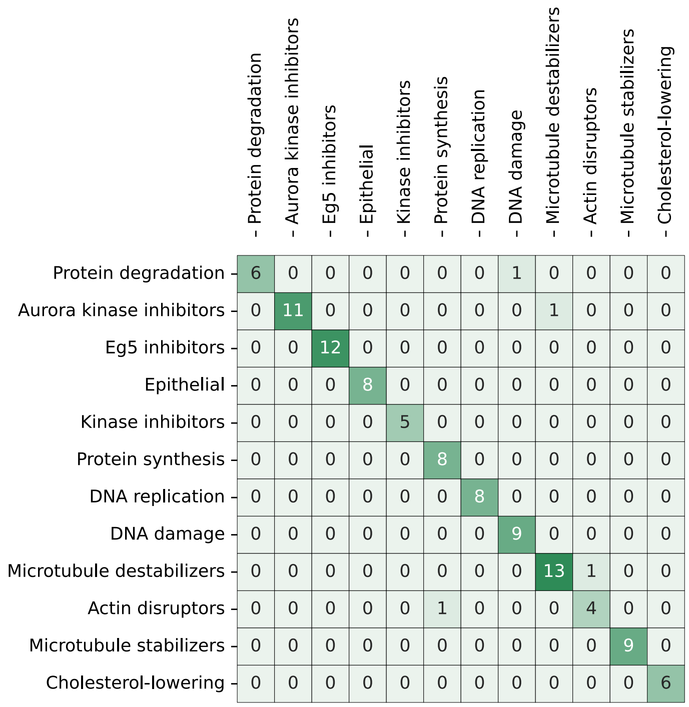


Figure 3: Best epoch UMAP projection of well-level embeddings colored by MOA showed good separation of clusters.



	Protein degradation	Aurora kinase inhibitors	Eg5 inhibitors	Epithelial	Kinase inhibitors	Protein synthesis	DNA replication	DNA damage	Microtubule destabilizers	Actin disruptors	Microtubule stabilizers	Cholesterol-lowering
Protein degradation	6	0	0	0	0	0	0	1	0	0	0	0
Aurora kinase inhibitors	0	11	0	0	0	0	0	0	1	0	0	0
Eg5 inhibitors	0	0	12	0	0	0	0	0	0	0	0	0
Epithelial	0	0	0	8	0	0	0	0	0	0	0	0
Kinase inhibitors	0	0	0	0	5	0	0	0	0	0	0	0
Protein synthesis	0	0	0	0	0	8	0	0	0	0	0	0
DNA replication	0	0	0	0	0	0	8	0	0	0	0	0
DNA damage	0	0	0	0	0	0	0	9	0	0	0	0
Microtubule destabilizers	0	0	0	0	0	0	0	0	13	1	0	0
Actin disruptors	0	0	0	0	0	1	0	0	0	4	0	0
Microtubule stabilizers	0	0	0	0	0	0	0	0	0	0	9	0
Cholesterol-lowering	0	0	0	0	0	0	0	0	0	0	0	6

Figure 4: Confusion Matrix - NSC. The accuracy of the best epoch was 99%. This showed an impressive ability to distinguish between drug MOAs

Hierarchical clustering was performed on treatment-level embeddings to assess the similarity of treatments targeting the same MOA (Figure 5). This demonstrated that treatments targeting the same MOA showed high similarity, achieving close to perfect grouping of treatments by MOA.

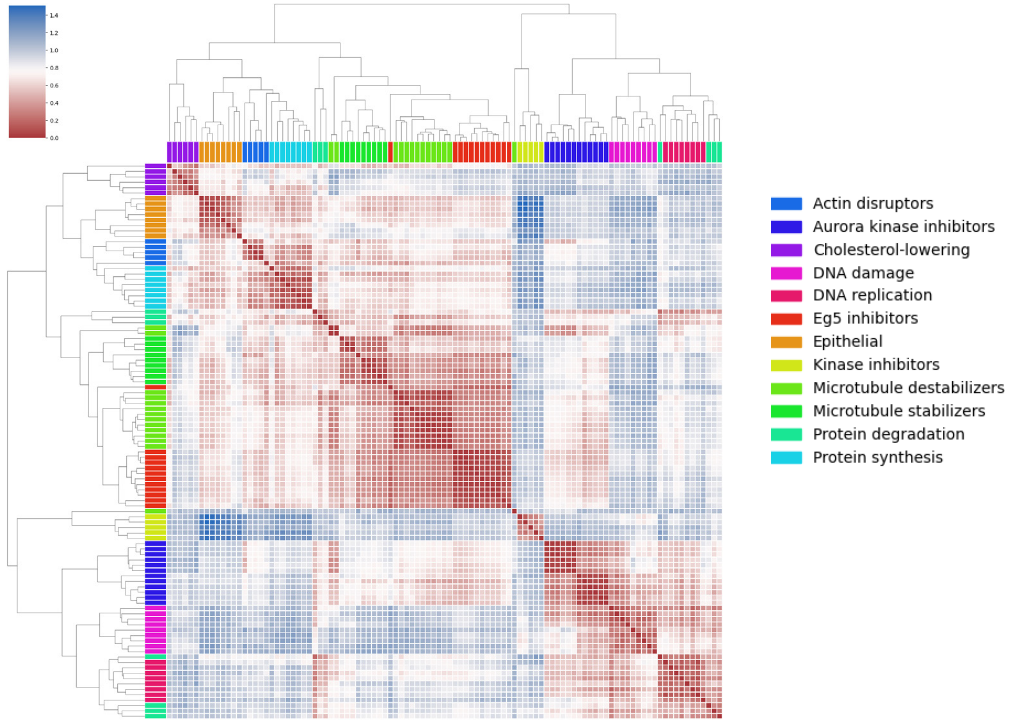


Figure 5: Best epoch hierarchical clustering of treatment level distances showed highly accurate grouping of treatments.

Conclusions

The developed model demonstrates strong potential for accurate drug mechanism of action (MOA) prediction and clustering. The semi-supervised approach, utilizing a random forest-based proximity matrix, yielded the best performance.

Future work will focus on training the model on larger cell painting datasets and further assessing the impact of batch correction methods. Additionally, the model will be expanded to include clustering of small molecules and matched genetic perturbations to enhance its utility for high-throughput drug discovery.

This application note summarizes the Master of Bioinformatics dissertation authored by Max Blanck and presented to Cranfield University.

References

1. Trapotsi et al 2022. RSC Chem. Biol.
2. Yu et.al. 2024 Nat Comput Sci
3. Caie et al. 2010. Mol Cancer Ther
4. Caicedo et al. 2022. Mol Biol Cell
5. Godinez et al 2017. Bioinformatics
6. Chandrasekaran et al. 2021 Nat Rev Drug Discov
7. Jansenns et al 2021. Bioinformatics
8. Caron et al 2018. arXiv:1807.05520

www.revvity.com

For research use only. Not for use in diagnostic procedures.