# AI-driven method for identification of hits from phenotypic screening with Cell Painting Assay

*Anna Bracha[1]*, Szymon Adamski[1]*, Adriana Borowa[1,2], Krzysztof Rataj[1], Dawid Rymarczyk[1,2], Michał Warchoł[1], Magdalena Otrocka[1],*

*1. Ardigen, Krakow, Poland*

*2. Faculty of Mathematics and Computer Science, Jagiellonian University, Krakow, Poland*

** equal contribution*

## OVERVIEW

High Content Screening is a powerful technology commonly used in phenotypic drug discovery. Currently, development of high content screening assays involves multiple technical processes, including cycles of wet-lab experiments and image/ data analysis. This process can be streamlined through the application of Cell Painting, an unbiased, target agnostic high content screening assay. The method, based on simple fluorescent staining of the 8 most important cellular compartments, generates imaging data that can be applied to identify biologically active compounds against various molecular targets without the need for development of different assays.

The key element in the process of identification of new active small molecules using Cell Painting Assay is the analysis of the results. Cell Painting generates multiparametric data from thousands of morphological features which are extracted from acquired microscopic images. Analysis of such complex datasets remains the biggest challenge that hinders researchers from using this approach for the hit identification purpose.

To tackle the problem, Ardigen developed a Deep Learning hit identification method as a part of its phenAID platform. The method is trained to predict compounds that induce the most similar phenotype to the reference compound, using images as a source of truth. The resulting output of the model is a ranking of possible hits with their probabilities.

To test this approach, we have used a JUMP-CP data [1] set containing HCS images of ~120 k compounds created by 10 different laboratories. For each compound, multiple replicates were acquired across different partner sites, in each sample plate a negative control (DMSO) and a set of 8 positive controls with known mechanisms of action were included. The positive controls were selected to ensure diversity of targets and phenotypes. Controls were used as the reference points for identifying hits in datasets coming from different sources, and to test the robustness of our method comparing hits across sources. Additionally, we explored the chemical diversity of predicted hits, to ensure the method's viability in the drug design process.

## INTRODUCTION

In this research we explore the JUMP-CP dataset, a vast set of ~120 000 molecular perturbations performed on U2OS cells, imaged in the Cell Painting protocol. We strive to find whether the Cell Painting images can be used to find new compounds that can induce a phenotype of interest. To do so we utilize some of the 8 positive controls present on each plate in the dataset. We selected 2 molecules used as positive controls: AMG900, an Aurora kinase inhibitor, and FK-866, a chemotherapeutic, due to their pronounced phenotypic effect. Our aim is to find compounds within the JUMP-CP dataset that induce a similar phenotype to the selected control molecules. We use 2 different approaches to this problem, one based on CellProfiler features and the other using Deep Learning.

## METHODS

### Cell profiler based - Cluster Hit ID

This hit selection method is based on identifying compounds whose CellProfiler [2] features exhibit the highest similarity to the positive controls (poscons). Each data point in this study represents a single well on a single plate, which means we have multiple data points for both hits and positive controls.

First, we reduce the features to a two-dimensional representation using T-SNE. Then, we cluster positive control data points and designate the centroid to each cluster. Due to the batch effects present within the JUMP-CP dataset, one positive control may end up populating multiple clusters. We exclude top 5% most distant data points from each cluster to prevent outliers. Finally, for each cluster we set a threshold equal to the maximum distance of any point within the cluster from the cluster centroid. We calculate distances of each data point to the centroid of each cluster. The data point is considered as an initial hit if its distance is lower than the threshold. In the last step, we compare the occurrences of those data points with the total number of data points for this given compound. If this ratio is below 75%, we discard this compound.

### Deep Learning based - Model Hit ID

This method is based on a Deep Learning (DL) model that was trained in Contrastive Learning [3] regime using compound images from different sources as pairs. We generate feature representation using this model and then train the Hit ID model to discriminate between a given positive control and the negative control (DMSO). This way the model learns to recognize important phenotypic changes caused by the treatment with the positive control molecule on the given cell line.

Next, we run inference on the compounds in the JUMP-CP dataset, which were never seen by the Hit ID model. Having predictions for each compound, we average scores across multiple wells, rank them, and choose the top 1% of samples. These are our hits, which have the highest probability of having similar effect to the celline as known positive control on which Hit ID model was trained.

## RESULTS

Using our approaches we selected a number of potential hit candidates for 2 positive controls with the most distinguished phenotypic effect: AMG900, an Aurora kinase inhibitor, and FK-866, a chemotherapeutic. For AMG900 the ClusterHit method yielded 129 candidates. The DL-based approach is capable of classifying the entirety of the dataset, so for this research we are selecting the top 1% of the scored compounds, in this case 877 compounds.

### AMG900

In the Fig. 4 we can see that the phenotypic effect induced by one of the top selected hit candidates is extremely similar to the phenotype of the positive control and vastly dissimilar from the negative control. What is also important, within the top 10 selected hits 5 were already annotated in the ChEMBL [4] database. One of these compounds is also an Aurora kinase inhibitor, two other are pan-kinase inhibitors and one was reported as having a anti-proliferative effect, similar to that of the Aurora kinase. In total, the JUMP dataset contains 104 compounds with known activity towards Aurora kinase, and the hits selected using CellProfiler approach found 11 of them, while the DL approach included 21. What is also interesting, we analyzed the chemical similarity of selected hits to known Aurora inhibitors. We found that while not perfect, those similarities are significant, especially in the manner of presence of particular chemical substructures.

### FK-866

Similar results can be observed for FK-866. Its anti-proliferative effect can be visualized and the hits selected by both our methods can find novel compounds with very similar phenotype (Fig.4).

## CONCLUSIONS

In our research, we confirm that Cell Painting coupled with Machine Learning analysis is a valid method of finding new compounds with novel structures that induce a desired phenotype. We also show that, thanks to its size and molecular diversity, the JUMP-CP dataset is a very useful tool, allowing us to explore new chemistries and biological targets.

In this study, we used CellProfiler and AI-based approaches, and both are capable of delivering valid conclusions. Results obtained by using CellProfiler show a slightly higher similarity to the known chemistry for the target, yielding fewer results than Deep Learning approach. The models utilizing Deep Learning yield more results, with higher chemical diversity, allowing us to find novel chemotypes for phenotypes of interest

**References:**

1. Chandrasekaran, Srinivas Niranj, et al. "JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations." bioRxiv (2023): 2023-03.
2. Stirling DR, Swain-Bowden MJ, Lucas AM, Carpenter AE, Cimini BA, Goodman A (2021). CellProfiler 4: improvements in speed, utility and usability. BMC Bioinformatics, 22 (1), 433. . PMID: 34507520 PMCID: PMC8431850.
3. Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.
4. Mendez, David et al. “ChEMBL: towards direct deposition of bioassay data”, Nucleic Acids Research, 2019

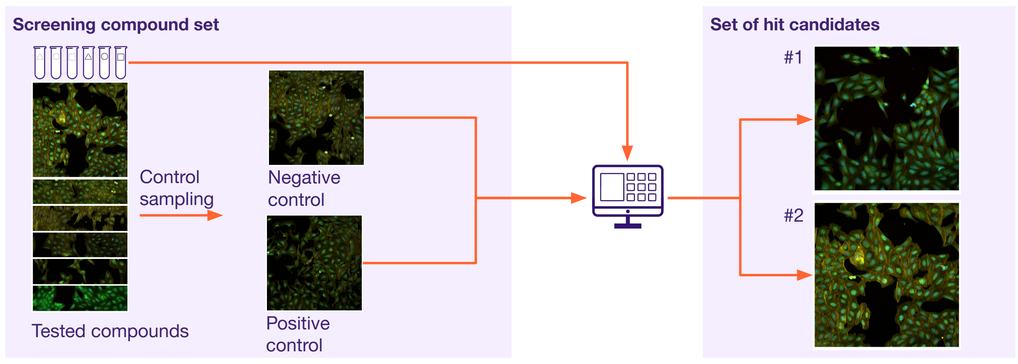


**Figure 1:** The definition of Hit Identification problem. We select positive (desired phenotype) and negative (no change in phenotype) controls to use in the training of our method. Then, from all other samples in the screening compound set we assign probability of being a hit and create an ordered list of hit candidates.

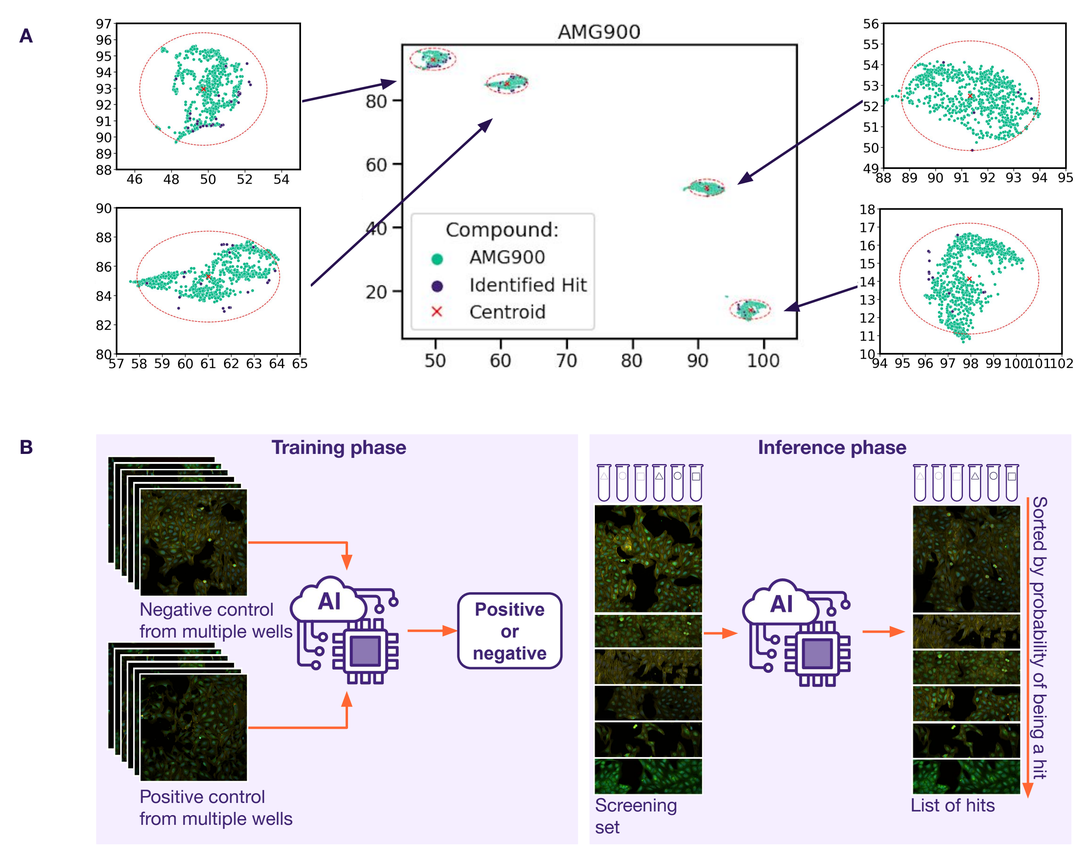


**Figure 3:** The representation of the 2 approaches used in the study. **A:** Cell profiler based - Cluster Hit ID **B:** Deep Learning based - Model Hit ID.

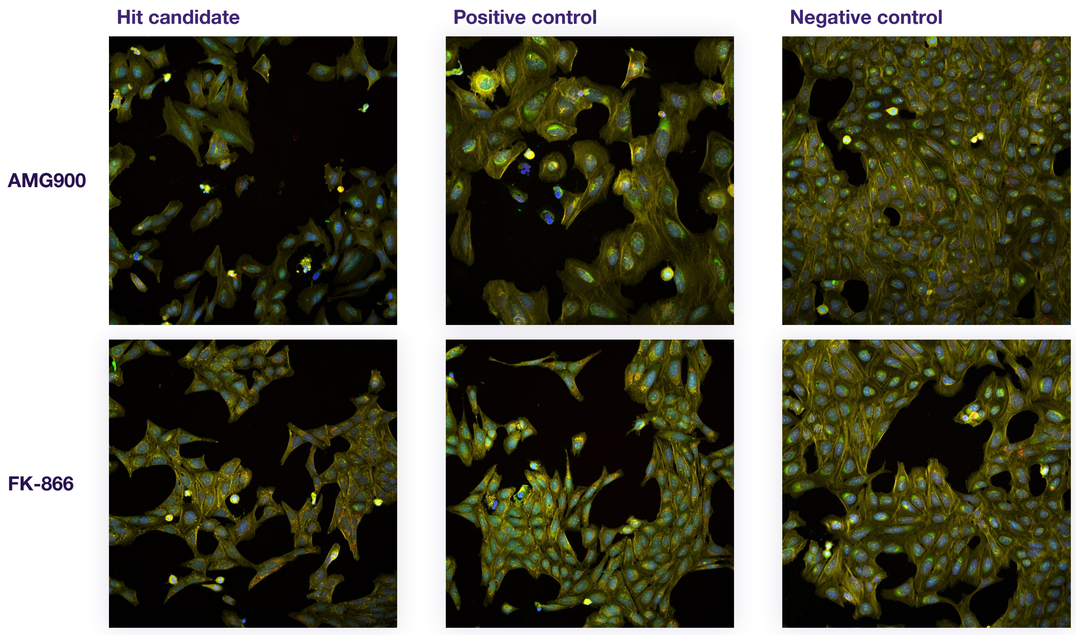


**Figure 4:** Visualization of the phenotypes for positive control, negative control, and found hits for AMG900 and FK-866.

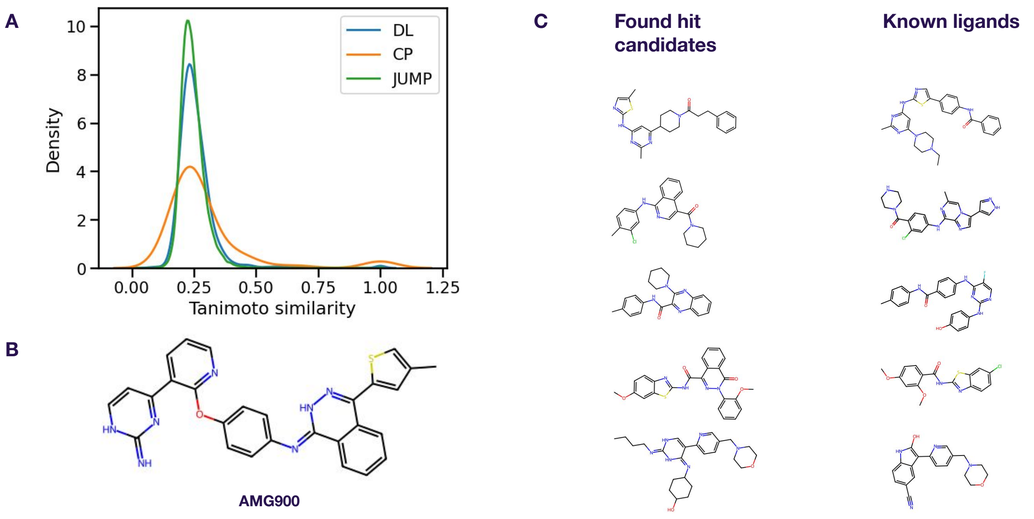


**Figure 5:** Cheminformatics analysis of the found hit candidates for AMG900.
**A:** The distribution of Tanimoto similarity between the found hits and known Aurora kinase inhibitors from ChEMBL. The graph shows distributions for the CellProfiler-based cluster hit detection method (CP), the Deep Learning model based detection (DL), and the distribution of the entirety of the JUMP-CP dataset. **B:** Structure of the positive control. **C:** Found hit candidates and similar known Aurora inhibitors. We can see, that while none of the compounds are directly similar to the AMG900 compound, many of the hit candidates are similar to other, already known Aurora kinase inhibitors.

**Acknowledgements:** The works are carried out under contract no. POIR.01.01.01-00-0878/19-00, as: „HiScAI - Development of cell-based phenotypic platform based on high content imaging system integrated with artificial intelligence data analysis for neuroinflammatory and fibrosis drug discovery”, co-financed by the European Regional Development Fund under the Smart Growth Operational Programme, Submeasure 1.1.1.: Industrial research and development work implemented by enterprises.