AI-driven identification of hits from Cell Painting based screening

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

1. Ardigen, Kraków, Poland
2. Faculty of Mathematics and Computer Science, Jagiellonian University, Kraków, Poland

contact: **michal.warchol@ardigen.com**

ABSTRACT

Selection of relevant chemical starting points is crucial for the advancement of drug discovery projects. This is typically done by testing of biological activity of chemical libraries with high throughput screening. The Cell Painting Assay (CPA) was developed to measure biological activity through the detection of morphological changes in single cells in a target-agnostic manner. Here, we describe how to utilize multiple morphological features extracted from CPA images to identify compounds inducing phenotypic changes in a cell. We define this case as hit identification of any effect.

In this study, we detect phenotypes that are statistically divergent from the negative control based on two types of features: those generated by CellProfiler [1], and using a self-supervised deep learning model. We model the distribution of negative control features as a reference, and sample compounds that deviate from this neutral state of cells, measured by an anomaly score. Compounds within the top 10% of anomaly scores are considered hits.

The analysis of hits with assigned mode of action (MoA) labels demonstrated that all applied methods performed similarly, we observed substantial overlap for hits identified with different methods. Application of a further sanity check of primary hits did not identify a direct correlation between anomaly score and the cell count measured for each compound. Additionally, we identified several chemical clusters among selected hit compounds, which ultimately proves the applicability of our method for the characterization of chemical compound libraries.

We conducted the study using the JUMP-CP dataset and assessed the robustness of our method by comparing results obtained from various data sources provided by different partners. The analysis demonstrated a consistency across datasets from independent sources.

DATA

To conduct our analysis, we used the dataset cpg0016-jump [2], available from the Cell Painting Gallery on the Registry of Open Data on AWS[1]. JUMP-CP dataset is a result of a collaboration between 10 pharmaceutical companies, Broad Institute and Harvard. It is a collection of images and morphological profiles of about 120k unique compounds and more than 15k genetic perturbations in human osteosarcoma cells (U2OS). In our research, we use only Cell Painting images of chemical perturbations. Both approaches are tested using CellProfiler and deep learning features. In the validation process, we use Mode of Action annotations acquired from **ChEMBL [3] database**.

[1] https://registry.opendata.aws/cellpainting-gallery/

METHODS

We assume that a primary hit should have a phenotype statistically divergent from the negative control. Therefore, we introduce the hit identification of any effect problem as the detection of **Out-Of-Distribution (OOD)** data. We characterize a single image using extracted features, and we use the set of negative control samples to model the reference distribution. Thus, hits are defined as compounds showing anomaly to the reference distribution. That narrows down the OOD problem to the **anomaly detection** problem, where for each new example, we need to determine if it belongs to the same distribution as the negative control.

To address this problem, we compare two methods: Isolation Forest and Normalizing Flows Model. **Isolation Forest** is a tree-based method [4] which tries to identify points in the data distribution that collectively produce noticeable shorter paths in the forest and hence can be selected as anomaly. **Normalizing Flows** model is a method for constructing complex distributions by transforming a probability density through a series of invertible mappings [5]. Likelihoods learned by this model can be used as an outlier score to detect anomalies.

Our analysis comprises of three steps performed separately for data generated by each consortium partner. First, we train both models using only negative control images. Then, we run inference on all screened compounds and calculate anomaly scores for each partner. These correspond to average path lengths for the Isolation Forest and log-likelihood values for Normalizing Flows. Compounds with anomaly scores in the top 10% are considered as primary hits. We check the performance of the hit identification methods through the analysis of MoA and compound similarity. For reliability of the results, we focus on MoA labels that have been assigned to at least 15 compounds in the dataset and calculate the percentage of compounds with a given MoA that are identified as hits. We compare the results of two partners to check robustness of methods. Additionally, we performed structural analysis of hit candidates. For each of the built molecular clusters (ECFP, Tanimoto similarity, Butina clustering) we looked for a dominant MoA presented by labelled compounds.

RESULTS AND DISCUSSION

Compounds with known Mode of Action. We identify several Mechanisms of Action (MoAs) that demonstrate significant phenotypic divergence from negative controls. These MoAs include the insulin receptor, PI3 kinase, IL-1 receptor kinase, MAP activated kinase, and CDK. Table 1 showcases the percentage of compounds identified as hits, displaying predominantly divergent phenotypes.

In addition, our analysis reveals that both methods and feature types utilized for hit identification yield equivalent performance. In Fig. 4 we illustrate a correlation in hit detection across datasets generated by different partners using the example of compounds with specific MoA, thus proving the approach's robustness and reliability.

Compound structural similarities. When comparing similarity of identified hits structures, we discover several chemical clusters. Interestingly, we notice that some of the clusters are predominantly populated with compounds expressing one or several MoAs with high precedence. Each cluster also contained compounds with unknown activity. Figure 7 presents examples of compounds for three clusters with known MoA: histone deacetylase, CNS receptor, and Cytochrome p450 together with similar structures of unknown activity.

Toxicity. To perform a further sanity check of selected hits, we compared anomaly scores with the respective Cell Count obtained from CellProfiler metadata for each hit. The results are visualised in Fig. 6. Although we noticed a group of hits with low cell count and high anomaly score, however anomaly score is not solely driven by toxicity (threshold set for 200 cells), as toxic compounds were distributed over whole hit population.

Conclusions. Utilizing hit identification as a versatile tool holds potential for the comprehensive characterization and profiling of an entire chemical library. The inclusive definition of a hit, encompassing any anomaly from the control, effectively reveals diverse biological activities within a single experiment and analysis.

Nevertheless, it is necessary to conduct additional analysis together with a hit triage, to identify compounds that are promising starting points for drug development projects.

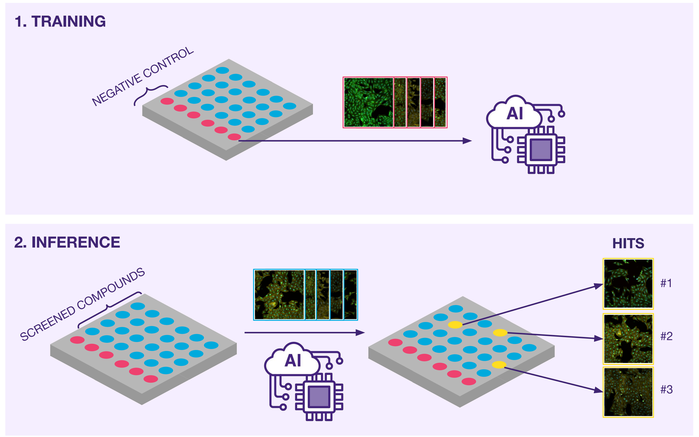

Figure 1: The definition of **hit identification of any effect** problem. We use images of negative controls to model the reference distribution. Then, for each new example in the screening compound set, we identify if it belongs to the same distribution as the negative control using the anomaly score. Samples within the top 10% of highest anomaly score are considered hits.

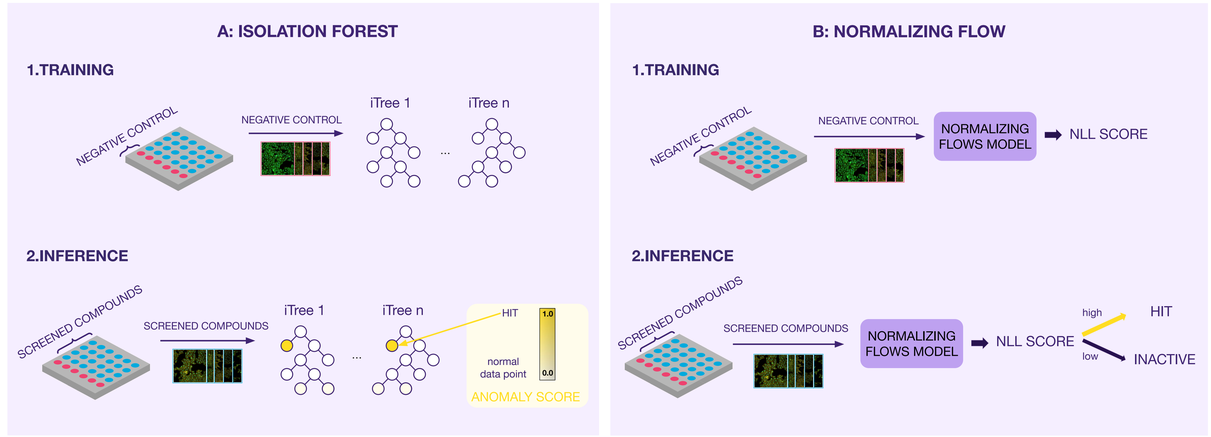

Figure 2: Computational approaches for hit identification of any effect: Isolation Forest and Normalizing Flows. Isolation Forest builds trees based on negative controls and identifies points with noticeably shorter path lengths than anomalies. Normalizing Flow model constructs complex distributions by transforming a probability density through a series of invertible mappings and uses likelihoods to detect anomalies.

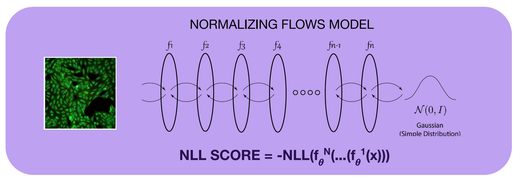

Figure 3: Normalizing Flow model is a deep learning technique that takes a raw image to encode it into a complex multidimensional space through an invertible mapping to constitute their distribution based on the negative log likelihood loss.

Method \ MoA	ISOLATION FOREST		NORMALIZING FLOWS	
fts	CP features	DL features	CP features	DL features
Insulin receptor	93 %	85 %	90 %	90 %
pi3 kinase	88 %	79 %	63 %	83 %
IL-1 receptor kinase	80 %	86 %	77 %	88 %
MAP activated kinase	85 %	73 %	85 %	74 %
CDK	79 %	77 %	75 %	81 %

Table 1: The most prominent MoAs of our hits. We compare MoAs by percentage of compounds identified by our approach, i.e. displaying predominantly divergent phenotypes. Our approaches identified multiple MoAs with high fidelity. The analysis shows that both methods and both feature types are equally efficient in terms of their performance.

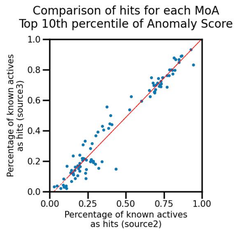

Figure 4: Comparison of overlapping MoAs between partners 2 and 3. It shows that there is a strong correlation between partners and MoAs of high divergence, suggesting high robustness of our approach.

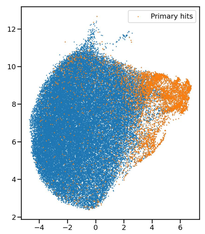

Figure 5: UMAP representation of the CellProfiler feature space for all compounds in the JUMP-CP dataset. The compounds with anomaly scores in the top 10th percentile from one of the sources is shown in orange.

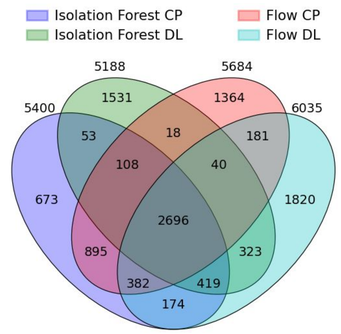

Figure 6: Venn diagram of the intersections between the top 10th percentile of anomalous compounds between various methods and features.

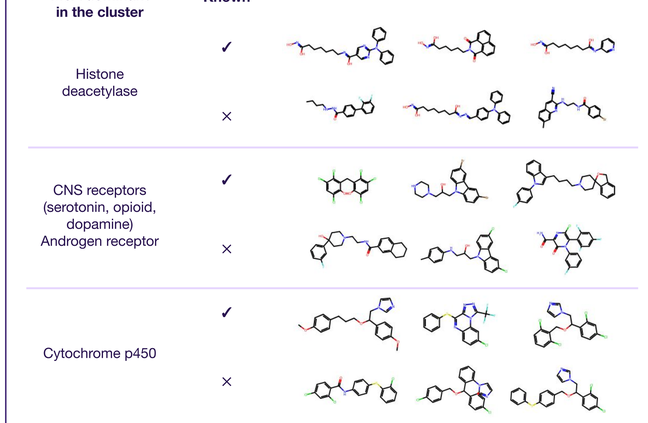

Figure 8: Examples of structures from 3 clusters obtained in the compound similarity analysis. For each cluster, we present compounds with a given MoA and compounds of unknown activity identified as hits. This analysis shows the potential of hit identification of any effect as valuable tool to discover novel drugs.

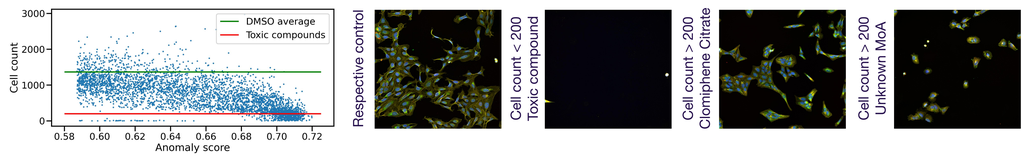

Figure 7: Plot of cell count and anomaly scores, and exemplary images of phenotypes induced by molecules with high anomaly scores. We identify that cell counts smaller than 200 corresponds to compounds that can be toxic while the higher than 200 cell count reveals bioactive hits, either known (i.e. Clomiphene Citrate) and unknown.

REFERENCES

[1] 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.
[2] Chandrasekaran, Srinivas Niranj, et al. "JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations." bioRxiv (2023): 2023-03.
[3] Mendez, David et al. "ChEMBL: towards direct deposition of bioassay data", Nucleic Acids Research, 2019
[4] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in Proc. 8th IEEE Int. Conf. Data Mining. IEEE, 2008, pp. 413–422
[5] Rezende, D. and Mohamed, S, "Variational inference with normalizing flows", Proceedings of The 32nd International Conference on Machine Learning, pages 1530–1538, 2015