# Authors/Contributions:

**William J Van Trump**: Data analysis, scientific research/input, manuscript drafting, resource contribution (via Spring Science)

**Hope O’Donnell**: Scientific input (our immunologist who ran down the IFN AMG900 link and sources)

**Ben Komalo**: Data analysis/managment (head of ML, platform model application support)

**Michael Wiest**: Data analysis (Data Scientist that supported some of the modeling)

\*all affiliated only with ‘Spring Science’

# 

# Cell Death Model

## Figure

## 

Fig XX: A. Confusion matrix for the single-cell classification model. Numbers indicate the number of cells in the training set for each combination of Actual vs. Predicted classes. B. Example ‘Alive’ cell image crops from the training set. C. Example ‘Dead’ cell image crops from the training set. D. Difference in the predicted number of ‘Alive’ cells between untreated wells and wells containing different ORF perturbations. Gene labels provided for the most death-inducing genes. The red line indicates the cutoff threshold for inclusion in the gene set enrichment analysis. E. Dot plot showing significantly enriched gene annotations. The size of the dot indicates the number of genes in for each annotation in the gene set. The color indicates the adjusted p-value.

## Results

A single-cell phenotypic model was created for dead and dying cells. Users identified 53 dead and dying cells in the dataset and provided the supervised learning tool with another 179 living cells. The resulting classification model trained on an 80/20 training/test dataset was found to have 100% recall and 100% precision in correctly calling dead cells [Fig XX]. The model was applied to the entire dataset and aggregated to calculate the percent alive cells per well. A geneset of the genes corresponding to the ORF perturbations which demonstrates the most cell death by imposing a cutoff of 6.5% increase in the percentage of dead cells over untreated wells [Fig XX]. Of the 187 ORF genes that exceeded the cuttoff, 63 were found in the utilized annotated geneset. Nine annotation groups were found to be significantly enriched in the matching genes, five of which were related to apoptosis, and the remaining four pertained to some form of cell stress [Fig XX].

## Methods

### Single Cell PhenoSorter Models

Raw images were uploaded to Spring Engine (a cloud-based software-as-a-service product provided by Spring Science[1](https://app.readcube.com/library/a36367c8-dcd9-4fc3-ae27-b941ddcef836/all?uuid=7119114158166044&item_ids=a36367c8-dcd9-4fc3-ae27-b941ddcef836:a8aa0237-8887-42db-89fb-e3bcd738c469). The PhenoSorter application was used to generate a single-cell-based phenotypic model for dead and dying cells. Briefly, this application presents individual segmented cell crops to a user who classifies them as dead/dying cells or living cells. Once this training set has been completed, the platform generates a classification model based on the intermediate embedding layer of a pre-trained convolutional neural network using XGBoost, a gradient-boosting decision tree-based framework. Training is done using an 80/20 split in which 80% of the user-classified cell images are randomly chosen to be used in the training of the model, and the remaining 20% are tested to determine model performance. In order to accommodate for different example set sizes, the loss function is weighted so that examples from smaller classes are given greater weight such that each class is ultimately balanced despite the different sample set sizes. Once trained, the model is applied to all single-cell image crops in the entire dataset. The data for all cells in each well are aggregated to yield a percent value for each model class for every well.

### Gene Set Enrichment Analysis

The results of our ML classification models were evaluated by performing a gene set enrichment analysis on the ORF perturbations that demonstrated the greatest effect of the model. For the dead and dying cell model, we compared the fraction of cells within each well that the model scored as alive for each ORF treatment against that same fraction of cells for untreated cells. ORF treatments were then ranked by the percent reduction in alive cells. A cutoff of 6.5% increase in dead and dying cells was used to generate our gene subset for genes demonstrating increased levels of cell death. The R programming language package ‘clusterProfiler’[2](https://app.readcube.com/library/a36367c8-dcd9-4fc3-ae27-b941ddcef836/all?uuid=5349025301803984&item_ids=a36367c8-dcd9-4fc3-ae27-b941ddcef836:e12bbc99-5116-4470-b275-de71eb2b61ce) was used for gene set enriment analysis. The WikiPathways annotation set [3](https://app.readcube.com/library/a36367c8-dcd9-4fc3-ae27-b941ddcef836/all?uuid=33526736019851566&item_ids=a36367c8-dcd9-4fc3-ae27-b941ddcef836:8de04619-9f67-4b5b-918a-5a4d19932119) of the C2 subset of annotations sourced from the MSig database[4](https://app.readcube.com/library/a36367c8-dcd9-4fc3-ae27-b941ddcef836/all?uuid=5533297916181811&item_ids=a36367c8-dcd9-4fc3-ae27-b941ddcef836:f72a8b53-d8c0-407b-be35-d5efe69f8daf) was used for gene annotations.

## Discussion

We were able to generate a highly reliable model for detecting and quantifying the presence of dead vs live cells within the experimental dataset. ORF perturbations that increased our quantification of dead and dying cells showed significant enrichment for apoptosis and cell stress-related annotations. Based on this, we can conclude that our user-generated model was correctly able to detect cell state and that at least a subset of the ORF perturbations elicited a detectable increase in cell death consistent with expectations derived from known gene function.

# 

# 

# Controls Model

## Figure

## 

Fig XX: A. Confusion matrix for the aggregate well-level control compound classification model. Numbers indicate the number of wells in the training set for each combination of Actual vs. Predicted classes. B. Example well images for wells that scored as highly similar to the AMG900 class C. Example wells that score as highly dissimilar to AMG900. D. Difference in the model score for the AMG900 classifier between untreated cells and each ORF purturbation. Gene labels are provided for the most AMG900-like genes. The red line indicates the cutoff threshold for inclusion in the gene set enrichment analysis. E. Dot plot showing significantly enriched gene annotations. The size of the dot indicates the number of genes in for each annotation in the gene set. The color indicates the adjusted p-value.

## 

## Results

A classifier model was trained for each of the four small molecule compound controls along with DMSO and untreated wells. The trained model performed very well at correctly categorizing each control with an overall accuracy of 98% [Fig XX]. The least performant class was Quinidine, which suggests that the resultant phenotype of this compound treatment was relatively difficult to distinguish. Both the DMSO and Untreated cells classes demonstrated very good performance during our validation testing with the predominant confusion being between those two classes. ORF perturbations were interrogated by this model and assigned a score (ie probability) for each control class. Nearly all ORF perturbations exhibited the highest probability for either the DMSO or untreated cells classes (data not shown).

Most of the control classifier scores were not found to correlate significantly with any particular annotations in our geneset enrichment analysis (data not shown). The few annotations that demonstrated a statistical increase in enrichment were generally related to cell death, suggesting that the controls impose a degree of cellular stress or toxicity above that experienced by cells in the untreated wells (data not shown). AMG900 was an exception to this trend. Numerous non-death-related annotations were found to be enriched in the geneset corresponding to ORF perturbations that demonstrated increased AMG900 classifier model scores [Fig XX].

## Methods

### Compound Controls Models

Classification models were generated for each of the small molecule control compounds in addition to DMSO and untreated cell groups using the cloud-based Spring Science platform[1](https://app.readcube.com/library/a36367c8-dcd9-4fc3-ae27-b941ddcef836/all?uuid=46480826535993713&item_ids=a36367c8-dcd9-4fc3-ae27-b941ddcef836:a8aa0237-8887-42db-89fb-e3bcd738c469). These models were generated by extracting the embedding layer of a pre-trained convolutional neural network applied to individually segmented cell crops for the entire dataset. These single-cell embedding vectors were mean aggregated over each well, resulting in a single embedding vector for the well. Well-level embeddings were then Z-score normalized over each individual plate to reduce plate-based batch effects. Models were subsequently built from the normalized well-aggregated embeddings using XGBoost and a K-fold cross-validation training approach. As with the single-cell phenotype models, class training sets are weighted to accommodate for the different class sizes. Once trained, the trained compound classifiers were applied to the ORF wells. The purpose of this approach was to force the AI to attempt to score each ORF perturbation as either one of the small molecule control compounds, DMSO controls, or untreated wells. For each well, the classifier will assign a score (as a probability) of that well belonging to each of the trained classes. It is important to note that ‘none of these classes’ is not a valid choice for the model. It will, therefore, be forced to assign each ORF perturbation to one of the compound controls, DMSO, or untreated well classes. While the vast majority of the ORF perturbations will be assigned to the untreated well, the confusion of the classifier with the various control compounds can provide an exceedingly sensitive mechanism for detecting similar phenotypic responses.

### Gene Set Enrichment Analysis

The results of our compound control models were evaluated by performing a gene set enrichment analysis for each compound control to understand if increased model scores corresponded to genes that shared similar annotations. Enrichment analyses were performed for each compound control class independently. Analysis was performed using the same annotation set and methods as the Dead Cell models.

## Discussion

AMG900 is an Aurora Kinase inhibitor often used to suppress cellular proliferation in tumors[5](https://app.readcube.com/library/a36367c8-dcd9-4fc3-ae27-b941ddcef836/all?uuid=08172335176810475&item_ids=a36367c8-dcd9-4fc3-ae27-b941ddcef836:6f72d6fa-a40b-467a-9416-960022d253ed). Interestingly, the ORF perturbations that were confused with the AMG900 class were enriched for Interferon signaling genes, including IFNA14, IFNB1, and IFNA8. At first glance, the association between Aurora Kinase inhibition and interferons isn’t apparent. However, in addition to their role as immune modulators, interferons have been shown to have potent antiproliferative effects, similar to Aurora Kinase inhibition[6–8](https://app.readcube.com/library/a36367c8-dcd9-4fc3-ae27-b941ddcef836/all?uuid=11030696348601099&item_ids=a36367c8-dcd9-4fc3-ae27-b941ddcef836:e0f41239-68c4-429b-8b4b-aac4dd1b5408,a36367c8-dcd9-4fc3-ae27-b941ddcef836:377c1e95-0c9e-425a-982f-66597cf04dcf,a36367c8-dcd9-4fc3-ae27-b941ddcef836:9b14aa0c-0c7c-417a-a43f-e83f39934ce9). Though more work is needed, these results hint at a similar phenotype in U2OS cells between the effects of AMG900 and that of upregulating interferon production.

[1. Spring Science. https://app.springscience.com/.  
2. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.  *Innov.* **2**, 100141 (2021).  
3. Pico, A. R. et al. WikiPathways: Pathway Editing for the People. *PLoS Biol.* **6**, e184 (2008).  
4. Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).  
5. Tang, A. et al. Aurora kinases: novel therapy targets in cancers. *Oncotarget* **8**, 23937–23954 (2017).  
6. Sangfelt, O., Erickson, S. & Grander, D. Mechanisms of interferon-induced cell cycle arrest. *Front. Biosci.* **5**, d479 (2000).  
7. Ambjørn, M. et al. IFNB1/interferon-β-induced autophagy in MCF-7 breast cancer cells counteracts its proapoptotic function. *Autophagy* **9**, 287–302 (2013).  
8. Bekisz, J., Baron, S., Balinsky, C., Morrow, A. & Zoon, K. C. Antiproliferative Properties of Type I and Type II Interferon. *Pharmaceuticals* **3**, 994–1015 (2010).](https://app.readcube.com/library/?style=Nature+%7B%22language%22:%22en-US%22%7D)