Identifying Potent Compounds with Low Concentration Cell Painting Images

[**Son Ha, PhD[1,2]**; Steffen Jaensch, PhD[1]; Lorena Freitas Krikler, PhD[1]; Dorota Herman, PhD[1]; Paul Czodrowski, PhD[2] ; Hugo Ceulemans, MD, PhD[1]]

1. [Janssen Pharmaceutical], 2. [Johannes Gutenberg University Mainz]

ABSTRACT

Background: Image-based models have been shown to accurately classify biodactivity in a range of assays and increase hit rates and chemical hit diversity[1]. These models use features extracted from cell images (from high-throughput screens called Cell Painting) as input, and often perform well at identifying active compounds with pIC50 >= 5 or 6 (IC50 <=10uM or 1uM).
High potency models (pIC50 >=7, IC50 <= 100nM) are also of interest in drug discovery. However, they pose a non-trivial problem due to low numbers of positive labels. We propose a method, improving on the existing image-based model, to accurately identify highly potent compounds. Our method overcomes class imbalance by using cell images acquired at different concentrations.

Methods: Firstly, we <u>train models with high concentration input images to classify compounds active at a low potency threshold</u>. There is sufficient training data available for many bioassays. The model learns to recognize image phenotypes specific to different assays. Then, <u>we perform inference with low concentration input images and evaluate the model with a higher potency threshold than training</u>. We expect bioactivity-related phenotypes are induced at low concentration for highly potent compounds, but not for less potent compounds. Hence, if low concentration input images are used to do inference, the model would more effectively identify potent compounds at a higher threshold than training.

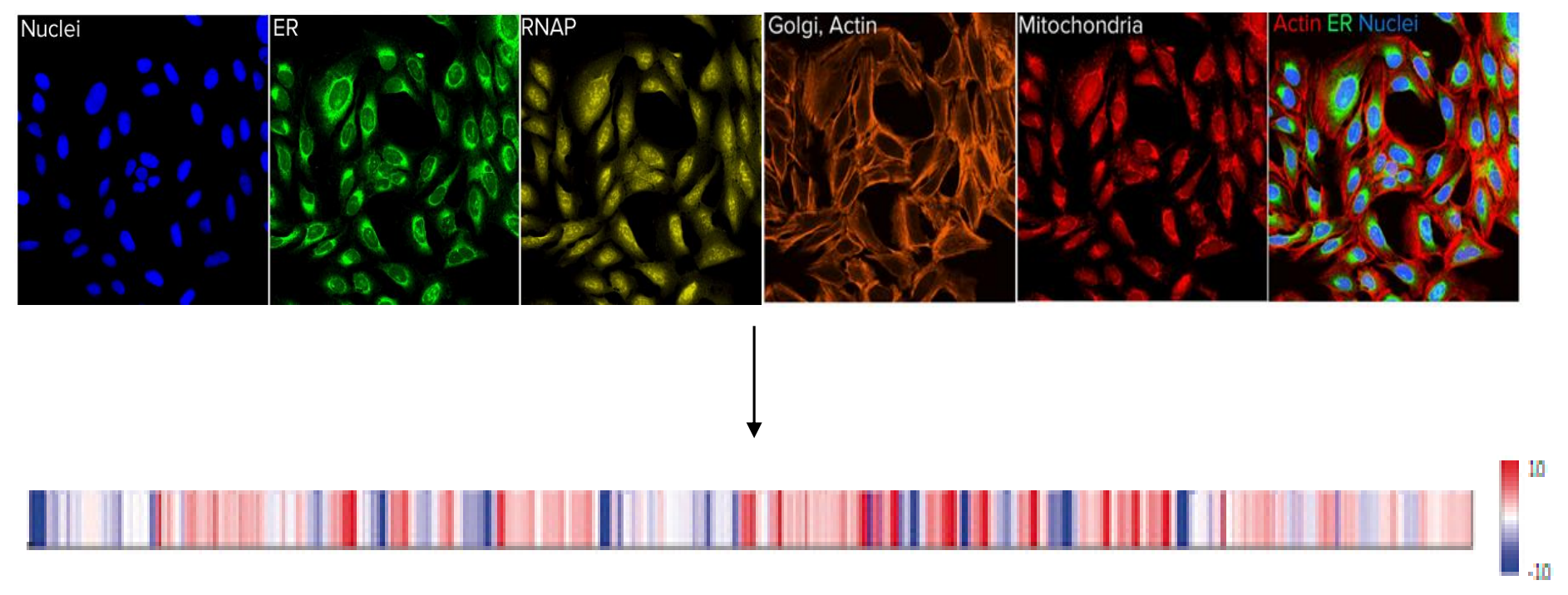
Results: Using our method, we managed to increase AUC-PR of high potency classification in ~75% of the bioassays investigated. We observed marked improvement in correctly identifying positives, compared to traditional method.

Applications: Prioritizing hits from image-based virtual screening for experimental follow-up by potency, and deprioritizing compounds with potent off-target activities in the hit-triaging phase

REFERENCES

1.Simm, J. et al. Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. Cell Chem Biol. 2018 May 17;25(5):611-618.e3. doi: 10.1016/j.chembiol.2018.01.015. Epub 2018 Mar 1. PMID: 29503208; PMCID: PMC6031326.

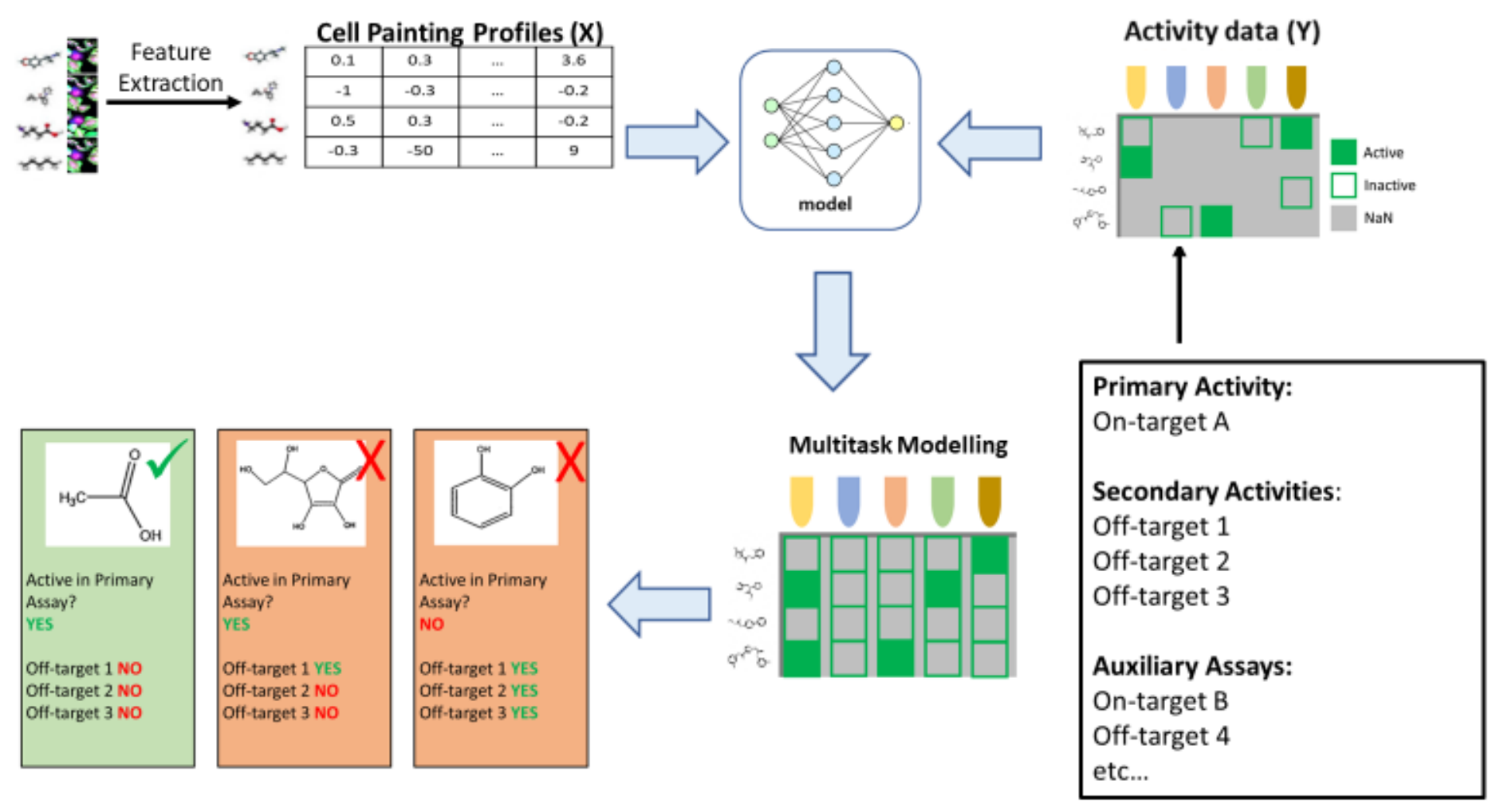
BACKGROUND: CELL PAINTING ASSAY



Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7687248/figure/anie201907853-fig-0003/, https://www.moleculardevices.com/lab-notes/cellular-imaging-systems/how-cell-painting-making-its-mark-on-drug-discovery

- High-content, multiplexed image-based assay for morphological profiling.
- **Method**: Perturb the cell (e.g. with a compound). Light up major components of the cell with up to 6 dyes. Then software measures morphological features from cell images.
  - ➢**<u>Biological 'fingerprint' characterizing compound-induced phenotypes.</u>**

BACKGROUND: MULTITASK BIOACTIVITY MODELLING

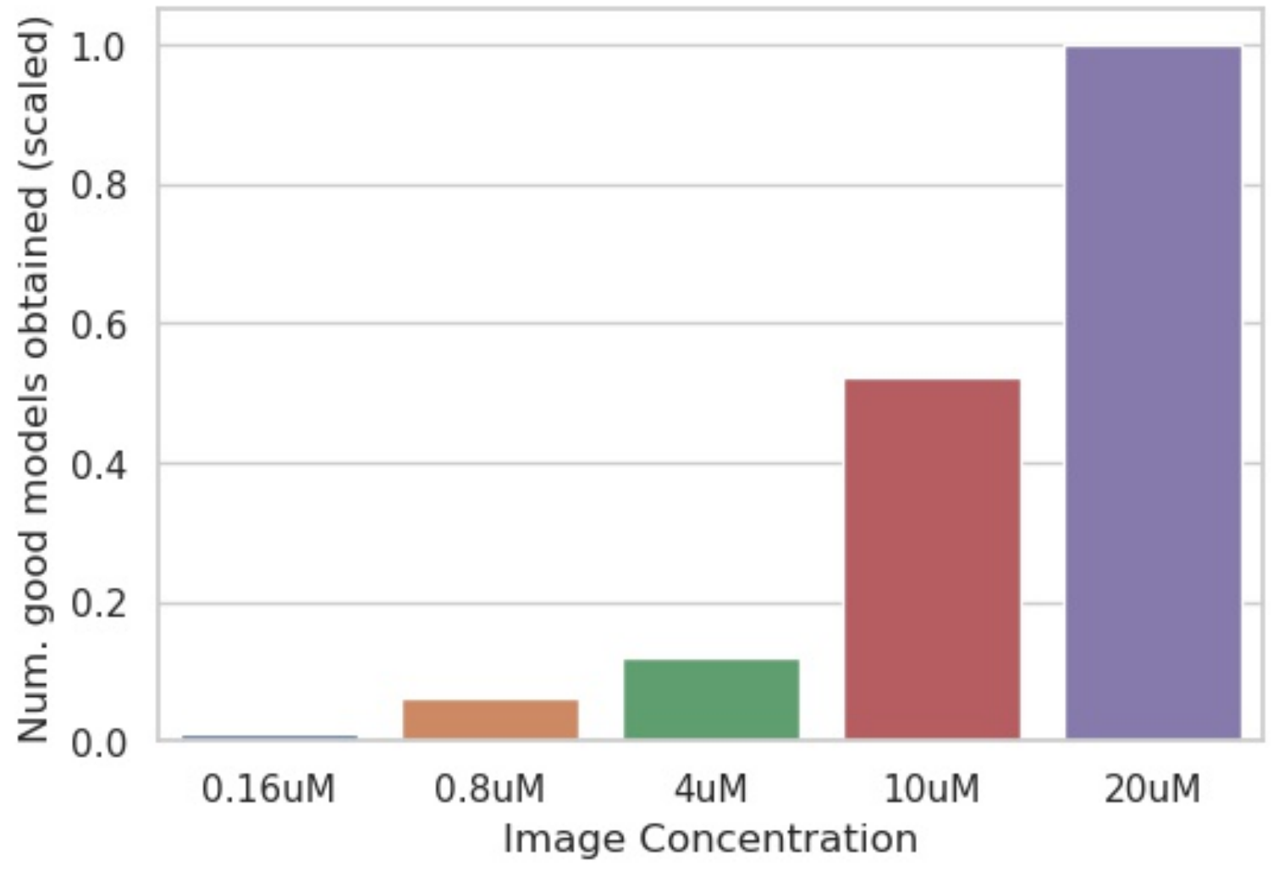


Primary Activity:
On-target A

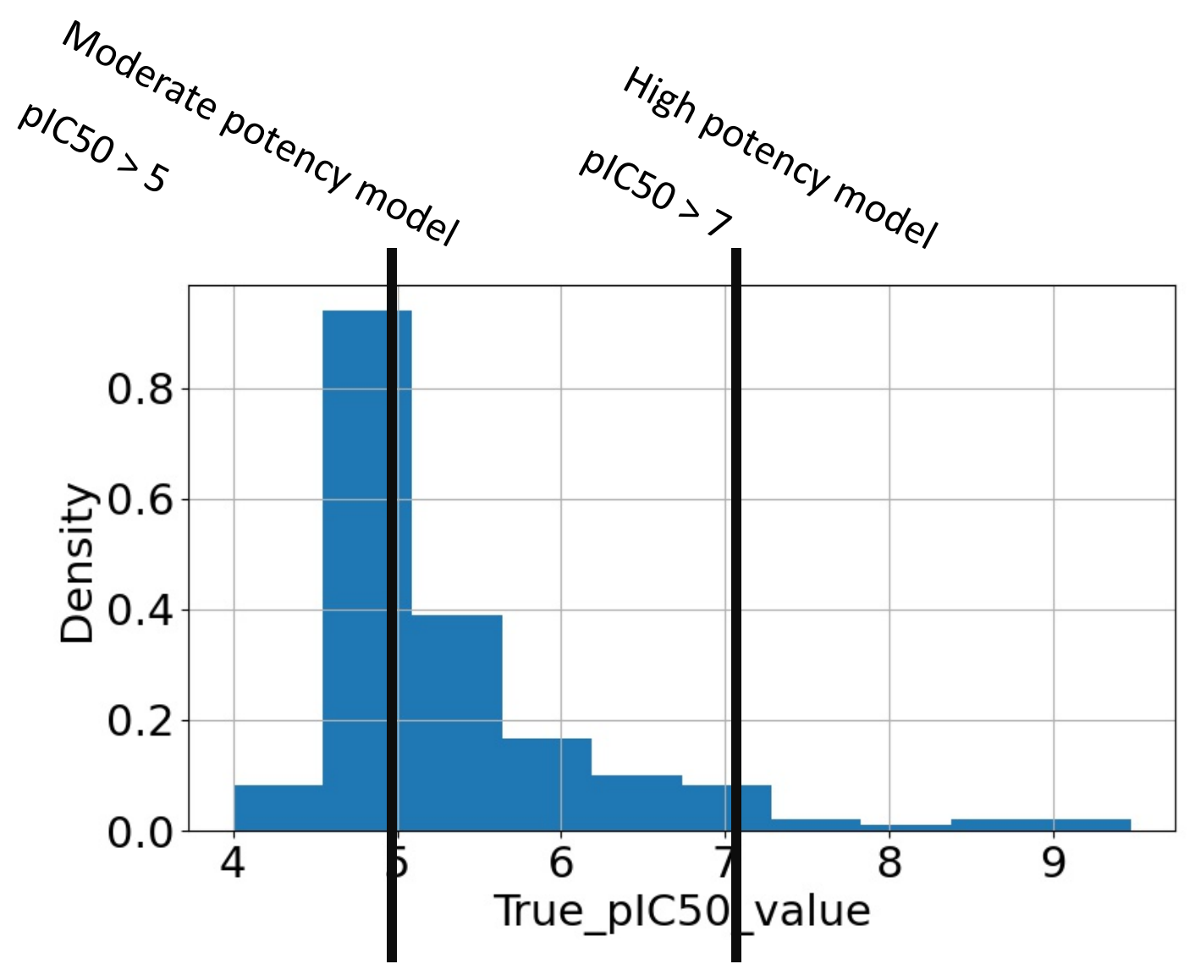
Secondary Activities:
Off-target 1
Off-target 2
Off-target 3

Auxiliary Assays:
On-target B
Off-target 4
etc…

RESEARCH QUESTIONS



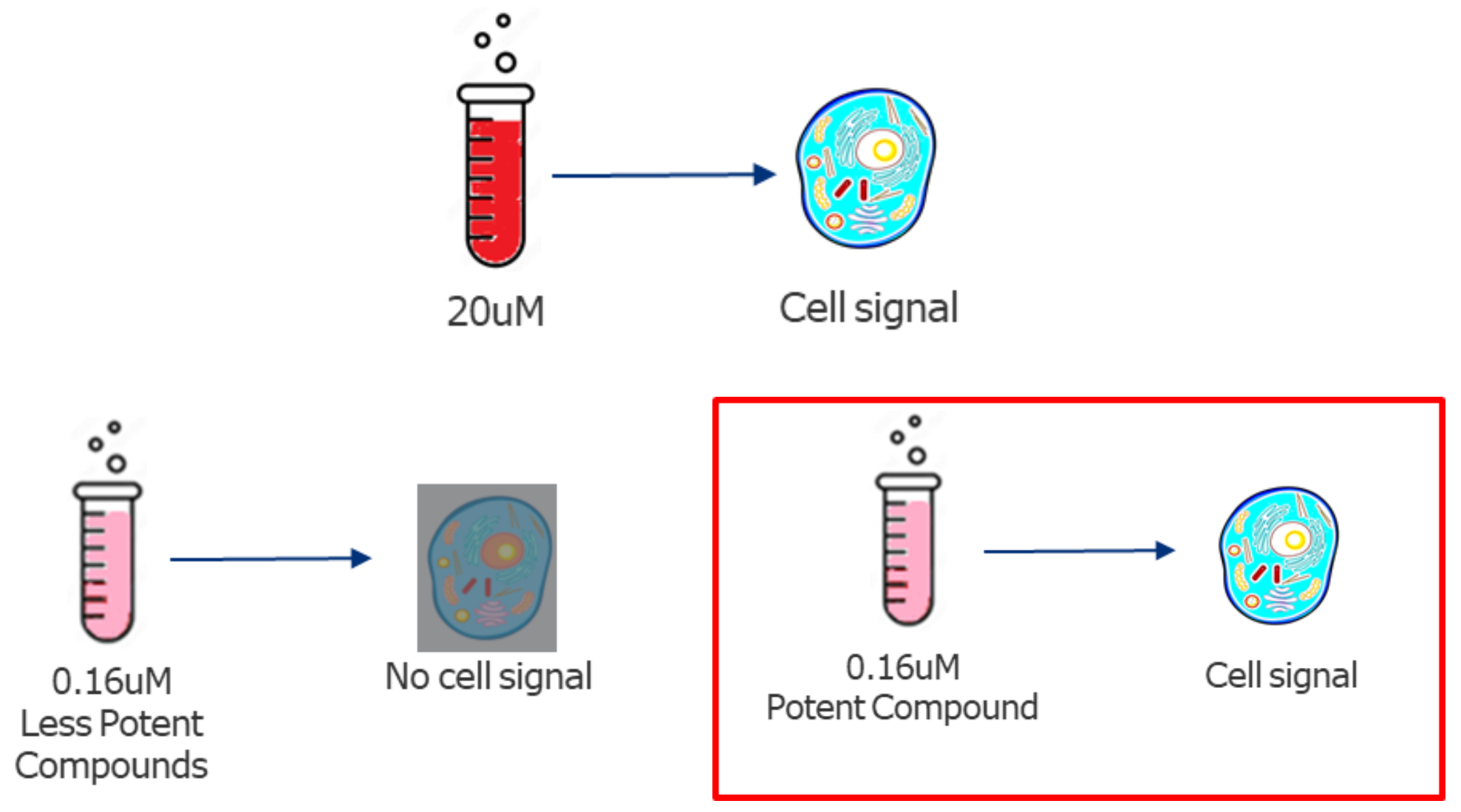
1) Using images from Cell Painting performed at concentration 10uM and 20uM leads to much higher number or good models than low concentration images. **What are the uses of low-concentration images?**



2) High potency model tends to underperform due to data imbalance (low positive samples). **How to obtain more good high potency model?**

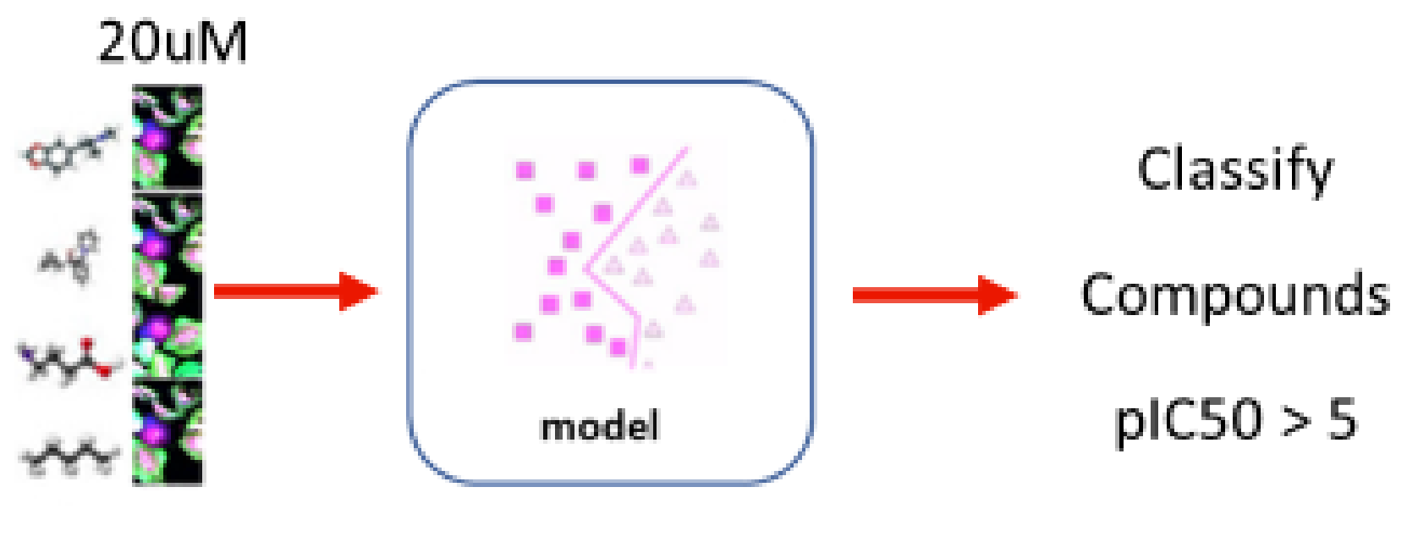
METHODS

<u>Intuition: Low vs high concentration CP images</u>



<u>Approach: Repurposing a good potency model for classification of higher potency compounds</u>

Training:



Classify

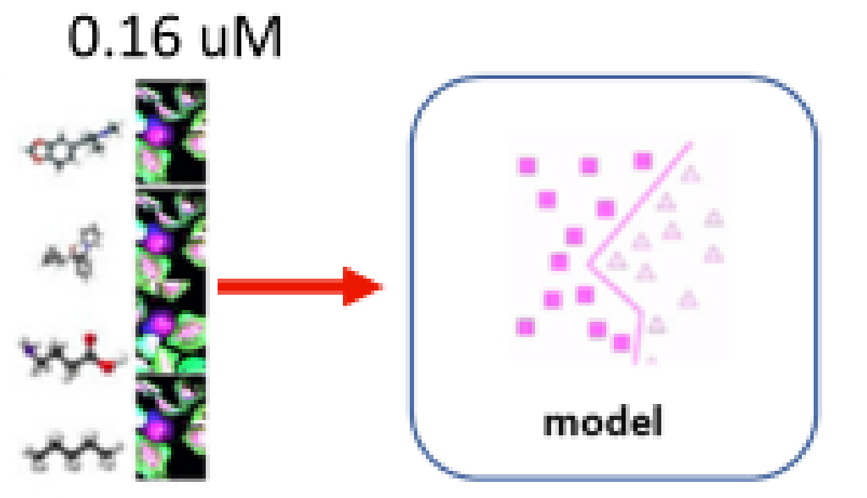
Compounds

pIC50 > 5

(Abundance Cell Signal)

(No data imbalance)

Inference:

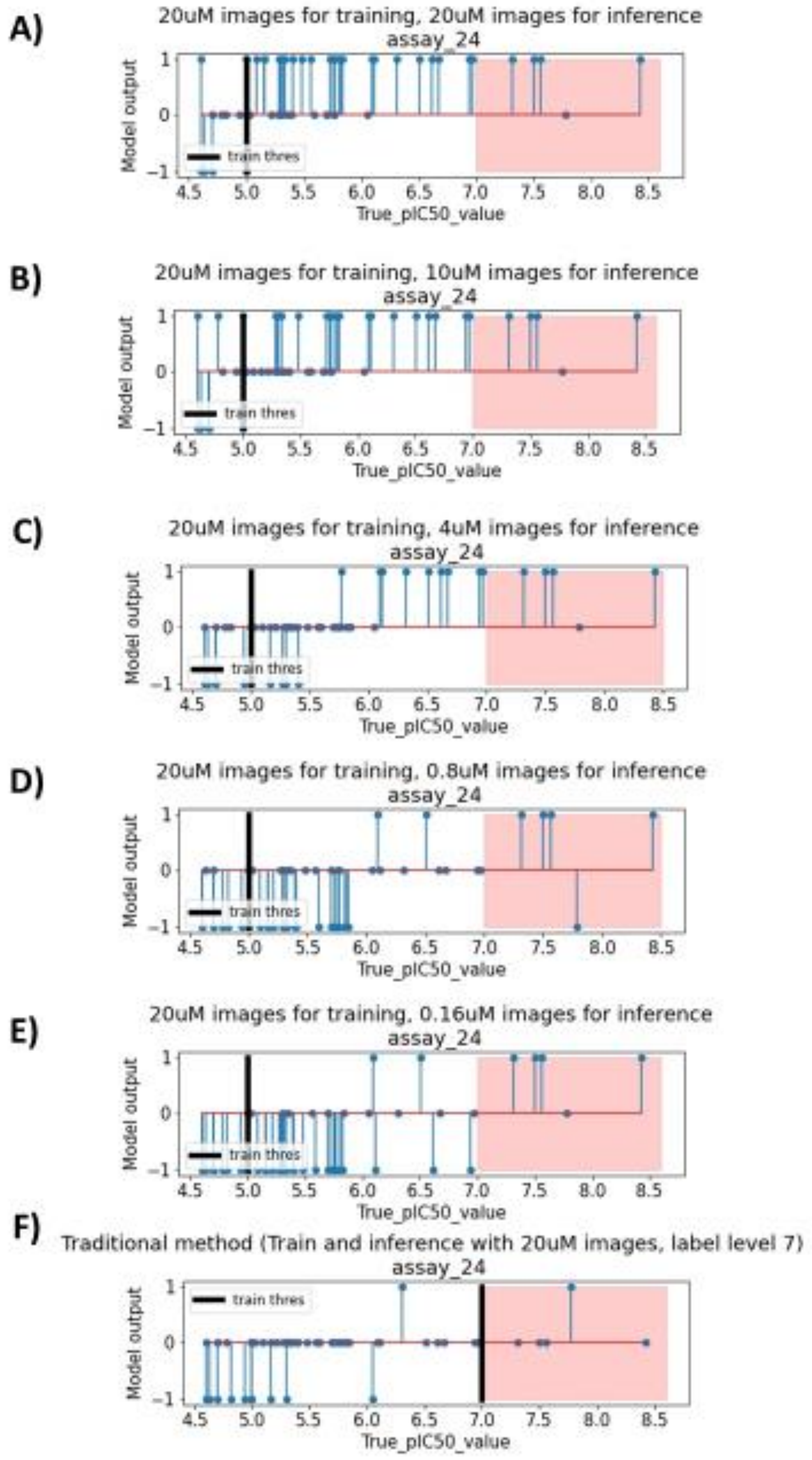


(Only classify highly potent compounds as active)

pIC50>7

RESULTS & DISCUSSION

<u>Figure 1.</u> Stem plots showing model output against the true pIC50 value. For each plot, the vertical black line denotes the potency threshold of the label the model trained on. The red area denotes the range of pIC50 which we consider highly potent (in this case it is pIC50≥7). The model behaves as a normal pIC50≥5 classifier in plot A). But when using low concentration images for inference, the model specifically retrieves highly potent compounds in the red region, and skip over the moderately potent compounds. This behavior is particularly clear in plot D) and E). In fact, these two plots show, out of five highly potent compounds, our method manages to retrieve four. Whereas the traditional method in plot F) can only retrieve one, due to data imbalance adversely affecting model training.



<u>Figure 2.</u> High potency precision heatmap recording high potency precision of each model across 57 assays. As the inference image concentration increases, the model should be more precise at classifying highly potent compounds, resulting in a lighter color. This color gradient is consistent across all assays, suggesting that in all cases the model can be repurposed to a high potency classifier by using low concentration images for inference.

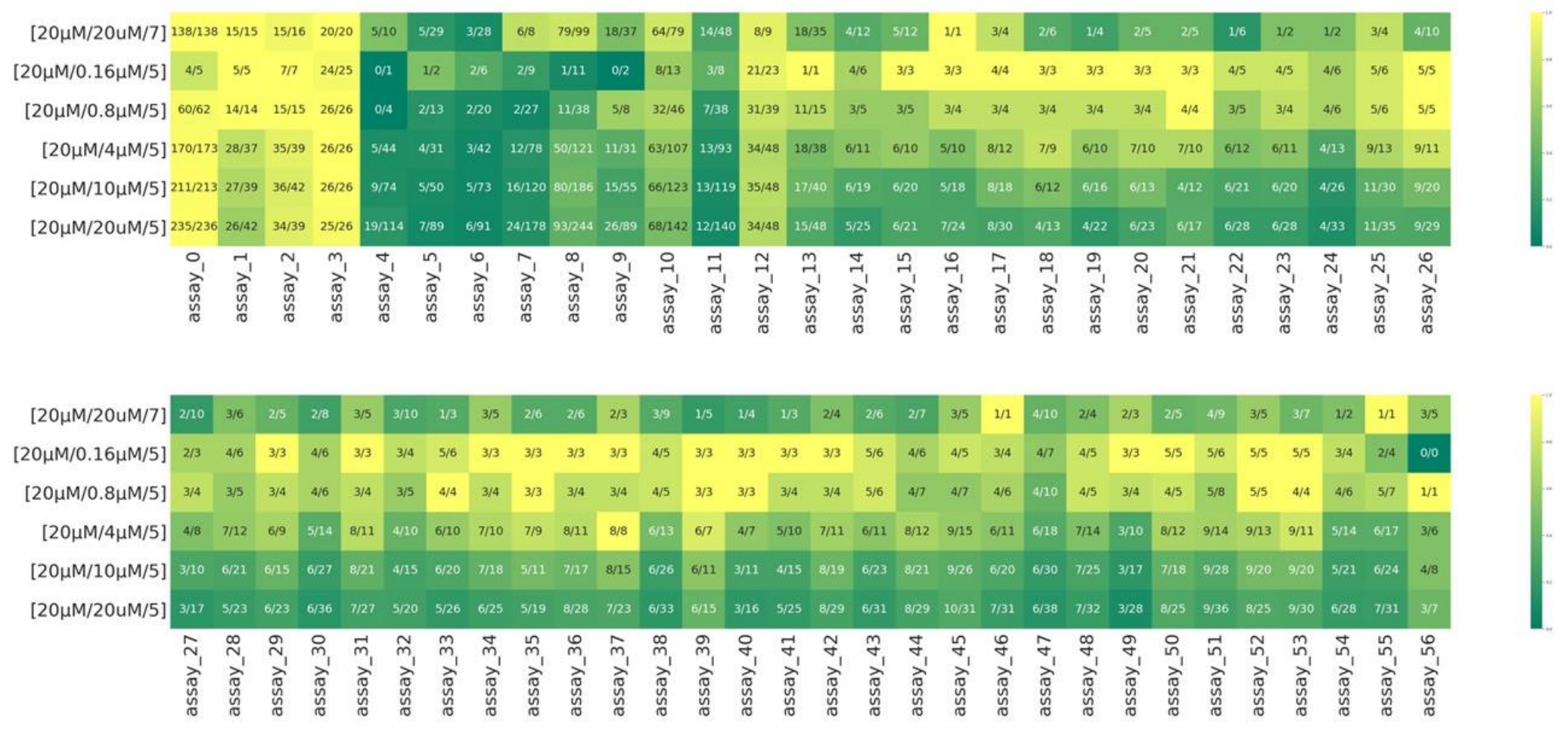


Figure 3. How much does AUC-PR improve when using our method vs the traditional method? Results across 57 assays. Our method manages to improve AUC-PR in 75% of assays investigated, with improvements around 0.2 to 0.5.

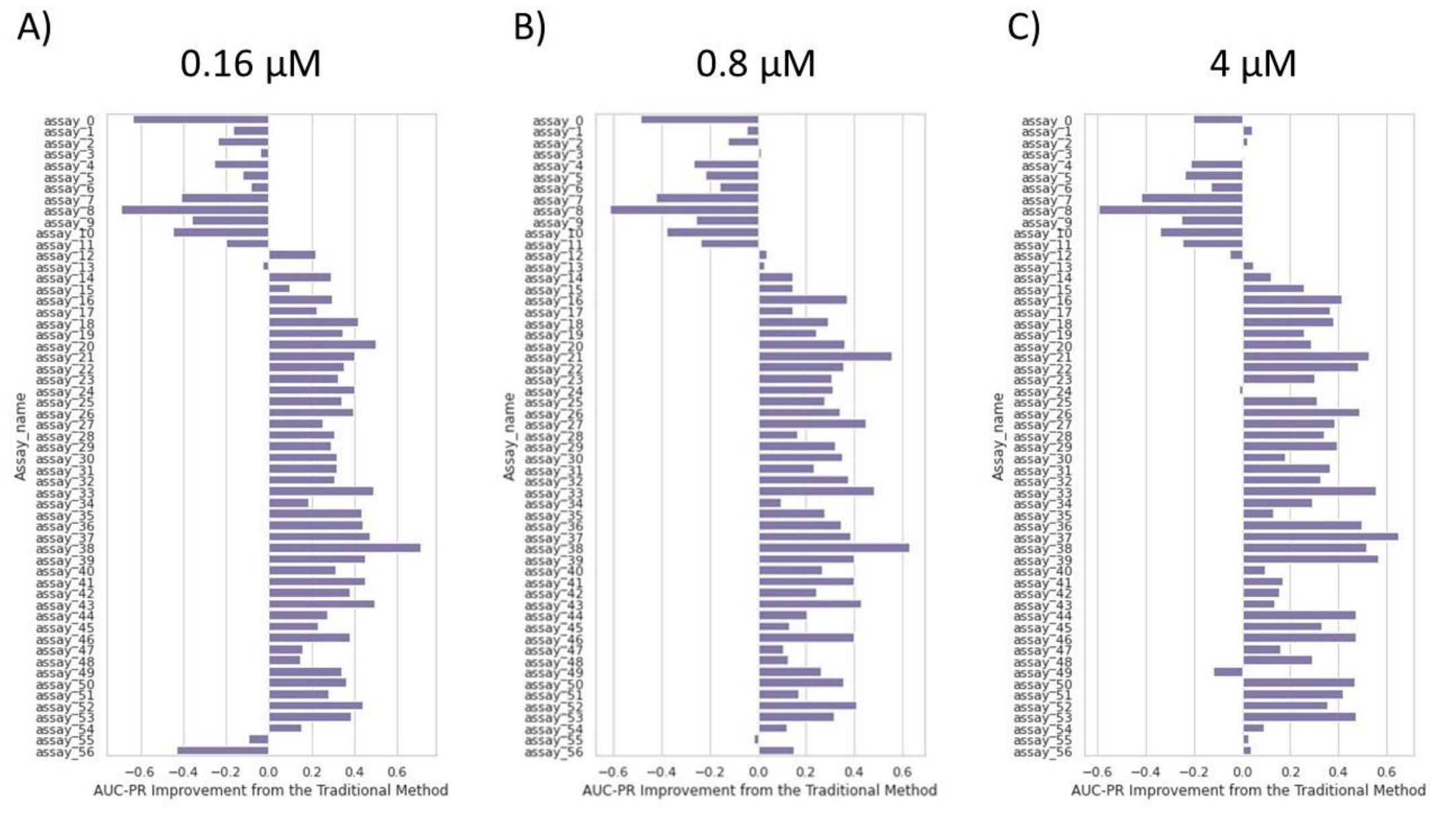


Figure 4. How much does AUC-ROC improve when using our method vs the traditional method? Results across 57 assays. Our method manages to improve AUC-PR in 65% of assays investigated, with improvements around 0.1 to 0.2.

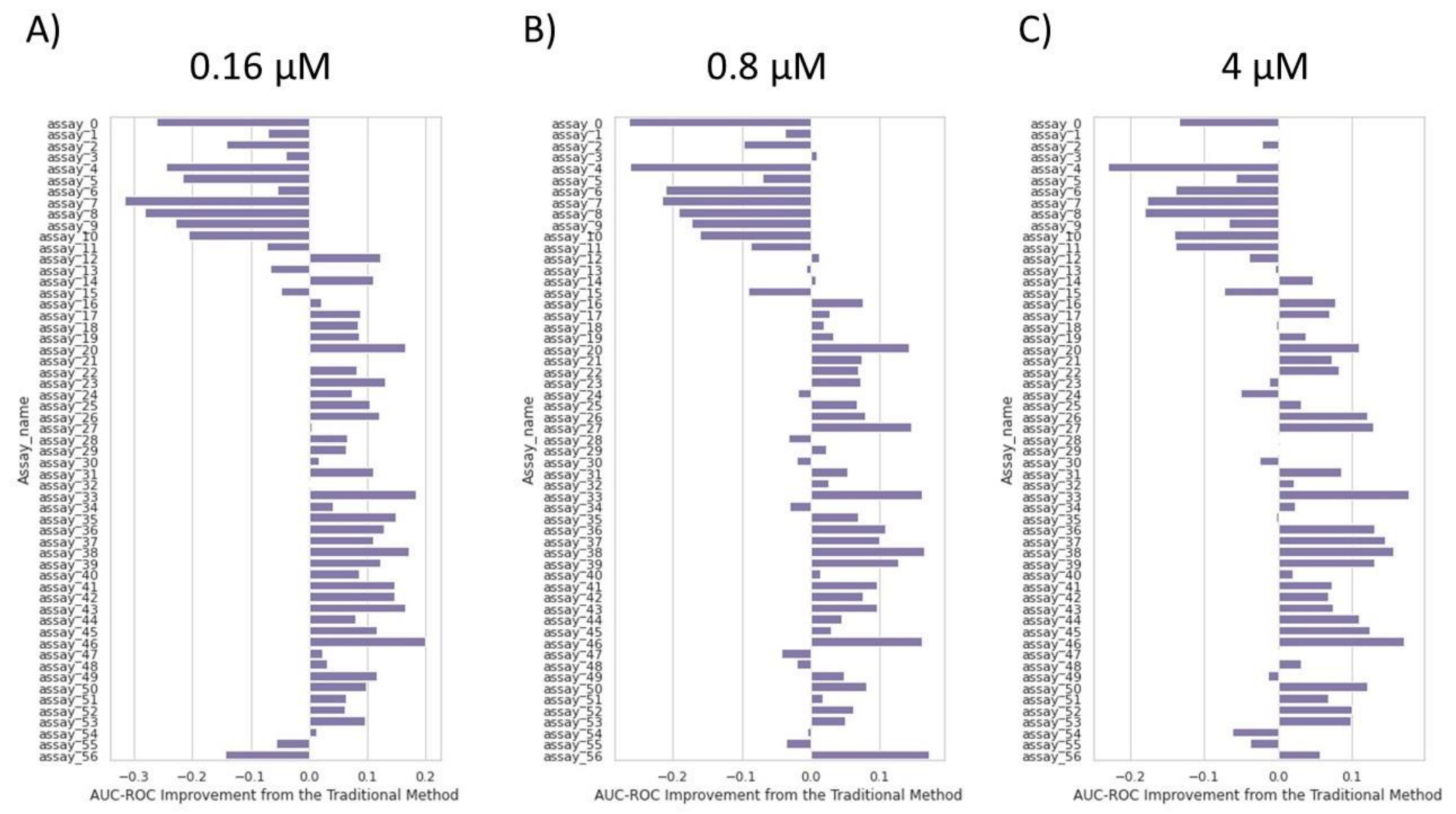


Image Credit: Jon Cenna, Biologics Discovery, Therapeutics Discovery
Image Description: Live cell image of tumor cells being killed by human T cells activated by a bispecific antibody directed against a solid tumor based target