# Machine Learning enabled Pooled Optical Screening in Human Lung Cancer Cells

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Pooled CRISPR-based gene knockout (KO) screening has emerged as a powerful method to uncover gene effects on various phenotypes [1, 2]. Recently, an optical pooled CRISPR screening method was developed [3] in which gene targeting guide-RNA (gRNA) are determined using *in situ* sequencing coupled with microscopy imaging of cellular structure and spatial features [3–6]. Pooled optical screening is very scalable and cost-effective. It can be coupled with different imaging assays to perform large-scale high-content image-based CRISPR-based KO screens. However, development of automated and general approaches for data processing and analysis are required to unlock its full potential as a tool for drug target discovery. Here, we introduce a machine-learning enabled computational framework for *in situ* sequencing, segmentation and feature representations of cell morphology from pooled optical screens and apply it to human lung cancer cells (A549). We develop a convolutional neural network (CNN) method for gRNA sequence calling, and show that it increases the cell yield by 10% and enables automation. We suggest self-supervised single-cell embeddings as a method to create informative representations of cell morphology, moderately improving upon commonly used engineered features. We demonstrate that such embeddings, aggregated for each gene KO, are more similar for gene pairs that are known to interact and cluster genetic perturbations by their cellular components, biological pathways, and molecular functions. We also highlight ways to use the perturbation clusters to generate hypotheses about gene functions, which are consistent with results from orthogonal studies. Put together, we develop a scalable and general computational approach to process and analyze pooled CRISPR-based morphological screens that can be applied to screen for various disease relevant phenotypes.

## 1 Introduction

Pooled CRISPR KO screening technologies have been widely used for conducting large scale investigation of gene effects on diverse sets of phenotypes. Recently, Feldman et al. introduced a methodology for performing optical pooled screens in human cells [3, 7] by obtaining high-content image-based data with their corresponding perturbation identities from pooled CRISPR screens. Briefly, this approach involves transfection of a pool of cells with gRNAs to enable targeted CRISPR editing. Cells are then run through a phenotyping assay such as antibody staining and fixed. The gRNAs within the cells are then amplified using rolling circle amplification, and *in situ* sequencing is then conducted on the plates to read out the gRNA and respective CRISPR knockout within each cell. This approach was applied to study the NFkB pathway using p65 protein localization as a readout [3], and in a later work to study essential genes using intensity features derived from fluorescence markers [6], but not yet to a general morphology assay. Cell Painting [8] is a morphology imaging technique that is known to contain rich information about cell state, allowing practitioners to cluster compounds by their MOA. However, Cell Painting phenotypic screening is typically performed in arrayed format which is costly, labor intensive and is subject to batch effects [9]. Combining pooled optical screens

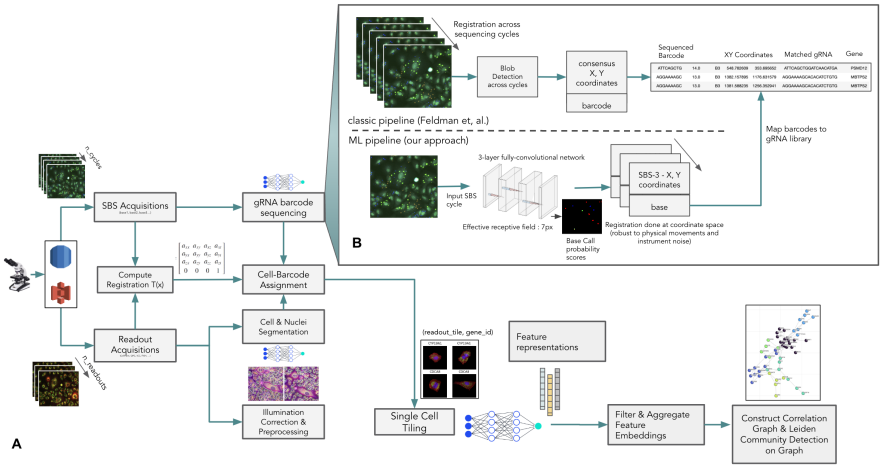


Figure 1: (A) Machine learning enabled image processing workflow for in situ sequencing, cell and nucleus segmentation and feature extraction from pooled optical screens. (B) Proposed approach for in situ sequencing using a 3-layer fully convolutional neural network followed by coordinate transformation for stitching of gRNA barcodes

with a general morphological profiling assay such as Cell Painting can provide an efficient and general assay for morphological screening in large genetic perturbation screens. Processing pooled optical screening data is challenging. It requires accurate gRNA sequencing, accurate segmentation of cell extents and correct association of guides to target-cells. In this work, we describe a computational framework for analyzing a screen that combines an adapted form of Cell Painting (high-content) with a pooled optical screen (high-throughput). In the following sections we present machine learning enabled methodologies for *in situ* sequencing and self-supervised feature extraction followed by the construction of a gene-gene phenotype similarity network. We demonstrate and evaluate the application of the above methodologies in learning gene similarity networks from a 300 gene (4 gRNAs per gene) pooled CRISPR knock-out screen dataset containing $\sim 1.5\ million$ cells.

## 2 Machine learning improves *in situ* sequencing

To scale up pooled optical screening we developed a fully automated pipeline for processing (Figure 1A). Hereinbelow are some of the method improvements that enabled this pipeline.

During the *in situ* sequencing step, each plate is processed to amplify the gRNA sequence present in each cell. These gRNAs are sequenced by synthesis (SBS) by labeling each nucleotide with a unique fluorophore, stripping, then relabeling with the next nucleotide in the sequence in a cyclic manner. This leads to a dataset in which the full plate is imaged several times, with stationary dots showing variable fluorescent signatures that need to be converted to sequencing base calls. A major step in optical screens is *in situ* sequencing of the gRNA. Feldman et. al [3], presents a computational methodology for *in situ* gRNA sequencing that requires manual alignment of field-of-view images during acquisition followed by local image registration and blob detection that requires manual fine tuning of parameters. Here, we propose an improved methodology for base-calling by training a 3-layer fully convolutional neural network that takes as input a sequencing-by-synthesis fluorescence base call image with channels corresponding to fluorescent nucleotide signals (A, C, T, G) and produces a probability mask corresponding to each channel (Figure 1B). We then use the probability mask to identify base locations and the corresponding base call. The base calls are stitched based on spatial correspondence across all the SBS acquisition cycles to generate a gRNA barcode readout corresponding to each spatial location in the image (the first $k$ *(k=10)* bases of the gRNA is referred to as a barcode in subsequent sections). Our method does not require manual alignment of field-of-view images at acquisition time, does not require manual parameter tuning, and increases the percentage of cells recovered with valid gRNA barcode from 68.6% to 78.79% in our test dataset (Table 1).

Table 1: The number of cells with a valid barcode recovered using different *in situ* sequencing methodologies

| Method/Metric | Number of cells with a valid barcode | % of cells recovered with a valid barcode |
|---|---|---|
| SBS cycle aligned blob detector (Feldman et. al, [3]) | 1288234 | 68.60% |
| Blob detector + coordinate space alignment (this work) | 1251669 | 66.65% |
| FCN spot detector + coordinate space alignment (this work) | 1479631 | **78.79%** |

The gRNA barcode locations computed from the above step are projected onto the Cell Painting fluorescence images using a coordinate transformation matrix constructed by image registration between the acquisitions. The Cell Painting images are then preprocessed to correct for illumination and intensity artifacts and single cell and nuclei contexts are segmented using CellPose [10]. Finally, a single cell dataset is generated by cropping tiles centered on each nucleus and masked by its corresponding cell mask. Each tile is associated with a gRNA identity based on the mapped barcode locations.

## 3 Self-supervised models generate biologically informative embeddings

High-content image-based screens using Cell Painting have been shown to be useful in learning representations and morphological profiling of perturbation effects in cells [9, 11]. Funk et. al [6] demonstrated that simple fluorescence intensity and cell shape features derived from pooled optical screens can be useful in defining the functional landscape of human essential genes. While explicit features such as intensity and shape features can be useful, they need to be manually engineered and do not capture all the kinds of variation that can occur in a perturbation dataset. Self-supervised learning methods [12, 13] have been shown to improve the quality of learned representations compared to supervised learning methods. Recently, SimCLR [12] and DINO-ViT [13] have achieved state-of-the-art performance in learning representations from natural images. Here, we utilize these frameworks to learn single-cell phenotype representations that can be used to create gene-gene phenotype similarity networks. The process is as follows: 1) we extract single-cell representations 2) represent each genetic perturbation by the median over all cells of that gene perturbation. 3) reduce the dimensionality of these to the top 200 principal components. 4) form a correlation matrix 5) threshold to keep significant correlations 6) cluster with a community detection method (Leiden).

To assess the performance of the methodology in learning biologically meaningful feature representations, we obtained evaluation metrics based on the overlap of our learned gene-gene phenotype network with publicly available gene-network and ontology (STRING DB and Gene Ontology) databases. For STRING DB [14] evaluation, we used the protein-protein interaction network (combined_score > 900 as positive interaction, combined_score = 0 as no evidence of interaction) dataset as the ground truth dataset. For gene ontology evaluation, we constructed a ground truth network by adding an edge between a pair of genes if they belong to a common gene ontology (GO) set in each of the datasets (Cellular Component, Biological Process and Molecular Function) as obtained from MSigDB [15] (gene sets containing > 25 genes (out of the 300 genes in the screen) were not considered for evaluation). For each of the above ground truth datasets, we computed the area under the receiver operating characteristic curve (AUC) of the overlap between the latent space correlation matrix and the ground truth gene network. We trained and evaluated self-supervised models using the evaluation metric on the 300-gene perturbation dataset: 1. SimCLR model trained using a resnet-50 backbone, 2. DINO model trained with vision transformer backbone (vit-small, patch_size=16) (DINO-ViT), 3. DINO-ViT with positive pairs sampled from samples having the same gene perturbation, 4. DINO-ViT with positive pairs sampled from samples having the same gRNA barcode, 5. DINO-ViT model trained on ImageNet pretrained weights and fine-tuned with positive pairs sampled from samples having the same gRNA barcode. We also compare the above models against an ImageNet pretrained DINO-ViT model and explicitly engineered cell intensity and morphology features as baselines. The results show that pre trained or fine tuned DINO-ViT features outperform the commonly used engineered features (Figure 2, Table A1).

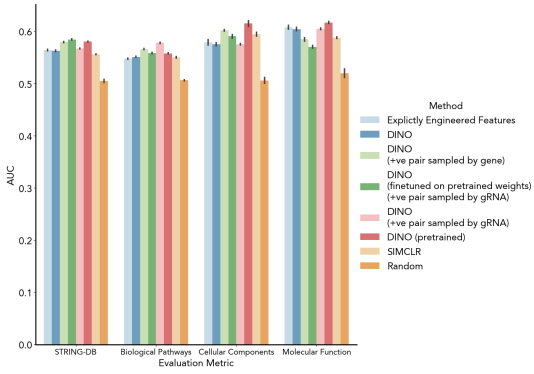


Figure 2: Comparison of feature embedding methodologies based on their ability to represent known gene-gene relationships as measured by area under the receiver operating characteristic curve (AUC)

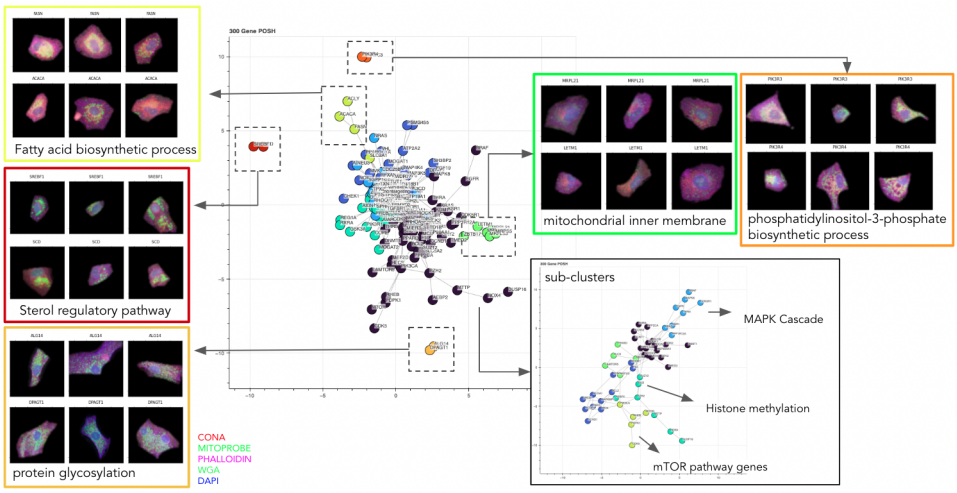


Figure 3: Community detection on pretrained DINO-ViT gene embedding correlation graph clusters genes by biological process/pathways

The DINO-ViT representation of the information-rich Cell Painting assay allows reconstruction of gene networks, and identification of new pathway components in a hypothesis-free way. Specifically, we were able to reconstruct the genetic modifiers of PI3K/Akt activation, protein glycosylation, fatty acid biosynthesis, sterol regulatory pathway, mitochondrial-inner membrane genes and more (Figure 3). Interestingly, the network was capable of clustering key components of the lipogenesis pathways. Namely, the core fatty acid synthesis enzymes (ACLY: ATP Citrate Lyase, ACACA: Acetyl-CoA Carboxylase Alpha, and FASN: Fatty Acid Synthase), upstream AKT signaling regulators (PIK3R3: PI3K Regulatory Subunit 3, PIK3R4: PI3K Regulatory Subunit 4), and downstream palmitate and mevalonate pathway regulators (SCD: Stearoyl-CoA Desaturase, SREBF1: Sterol Regulatory Element Binding Transcription Factor 1) all contribute to lipogenesis [16]; our screening and ML approach was capable of grouping these regulators in an unsupervised manner using a generic morphological readout. While typical CRISPR screens would be capable of finding genes that increase or decrease lipogenesis, this model appears to have achieved a higher level of granularity by producing these sub-clusters. Another striking observation is the clustering of CDK5 with genes from the mTOR pathway (RHEB, mTOR and PDPK1; Figure 3, bottom left). CDK5 was recently identified to phosphorylate S6 [17], our result supports its role as an mTOR pathway regulator. In contrast to that NRAS, which is closely related to the KRAS, has a distinct morphological phenotype, different from the KRAS cluster (KRAS, BRAF and EGFR; Figure 3, top right). This observation may reflect the different function of these Ras isoforms [18].

In summary, streamlined data processing pipelines and *in situ* sequencing methods via automation allow for increased scale, gRNA coverage, and improved gene network reconstructions. This technique opens the possibility of whole-genome screening for numerous imaging-based phenotypes in a variety of cellular models.

# References


[1] Ophir Shalem, Neville E Sanjana, Ella Hartenian, Xi Shi, David A Scott, Tarjei S Mikkelsen, Dirk Heckl, Benjamin L Ebert, David E Root, John G Doench, et al. Genome-scale crispr-cas9 knockout screening in human cells. *Science*, 343(6166):84–87, 2014.

[2] Tim Wang, Jenny J Wei, David M Sabatini, and Eric S Lander. Genetic screens in human cells using the crispr-cas9 system. *Science*, 343(6166):80–84, 2014.

[3] David Feldman, Avtar Singh, Jonathan L Schmid-Burgk, Rebecca J Carlson, Anja Mezger, Anthony J Garrity, Feng Zhang, and Paul C Blainey. Optical pooled screens in human cells. *Cell*, 179(3):787–799, 2019.

[4] Xiaowei Yan, Nico Stuurman, Susana A Ribeiro, Marvin E Tanenbaum, Max A Horlbeck, Christina R Liem, Marco Jost, Jonathan S Weissman, and Ronald D Vale. High-content imaging-based pooled crispr screens in mammalian cells. *Journal of Cell Biology*, 220(2), 2021.

[5] Michael Lawson and Johan Elf. Imaging-based screens of pool-synthesized cell libraries. *Nature Methods*, 18(4):358–365, 2021.

[6] Luke Funk, Kuan-Chung Su, David Feldman, Avtar Singh, Brittania Moodie, Paul C Blainey, and Iain M Cheeseman. The phenotypic landscape of essential human genes. *bioRxiv*, 2021.

[7] David Feldman, Luke Funk, Anna Le, Rebecca J Carlson, Michael D Leiken, FuNien Tsai, Brian Soong, Avtar Singh, and Paul C Blainey. Pooled genetic perturbation screens with image-based phenotypes. *Nature Protocols*, 17(2):476–512, 2022.

[8] Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature protocols*, 11(9):1757–1774, 2016.

[9] Marzieh Haghighi, Shantanu Singh, Juan Caicedo, and Anne Carpenter. High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. *bioRxiv*, 2021.

[10] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021.

[11] Nikita Moshkov, Michael Bornholdt, Santiago Benoit, Claire McQuin, Matthew Smith, Allen Goodman, Rebecca Senft, Yu Han, Mehrtash Babadi, Peter Horvath, et al. Learning representations for image-based profiling of perturbations. *bioRxiv*, 2022.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[13] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[14] Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021.

[15] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[16] T Mashima, H Seimiya, and T Tsuruo. De novo fatty-acid synthesis and related pathways as molecular targets for cancer therapy. *British journal of cancer*, 100(9):1369–1372, 2009.

[17] Abul Arif, Jie Jia, Belinda Willard, Xiaoxia Li, and Paul L Fox. Multisite phosphorylation of s6k1 directs a kinase phospho-code that determines substrate selection. *Molecular cell*, 73(3):446–457, 2019.

[18] Oliver Rocks, Anna Peyker, and Philippe IH Bastiaens. Spatio-temporal segregation of ras signals: one ship, three anchors, many harbors. *Current opinion in cell biology*, 18(4):351–357, 2006.

# A Appendix

Table A1: Comparison of feature embedding methods based on their ability to represent known gene-gene relationships as measured by area under the receiver operating characteristic curve of the feature correlation matrix overlap with gene relationships obtained from the respective database (STRING-DB, GO-BP = Gene Ontology Biological Process, GO-CC = Gene Ontology Cellular Components, GO-MF = Gene Ontology Molecular Function)

| Embeddings/ Evaluation Metric | STRING-DB overlap (AUC) | GO-BP overlap (AUC) | GO-CC overlap (AUC) | GO-MF overlap (AUC) |
|---|---|---|---|---|
| Engineered Explicit Features | 0.5643 | 0.5446 | 0.5882 | 0.6082 |
| SIMCLR | 0.5615 | 0.5531 | 0.6063 | 0.5812 |
| DINO-ViT | 0.5712 | 0.5512 | 0.5956 | 0.6137 |
| DINO-ViT (+ve pair sampled by gene) | 0.5815 | 0.5675 | 0.6050 | 0.5822 |
| DINO=ViT (+ve pair sampled by gRNA) | 0.5668 | **0.5754** | 0.5888 | 0.6062 |
| DINO-ViT (fine-tuned on pretrained weights) (+ve pair sampled by gRNA) | **0.5877** | 0.5617 | 0.5884 | 0.5710 |
| DINO-ViT (pretrained) | 0.5842 | 0.5553 | **0.6073** | **0.6182** |
| Random | 0.5084 | 0.4920 | 0.5465 | 0.5010 |