Deep learning virtual screening with active signature learning improves the identification of small-molecule modulators of complex phenotypes

Anonymous Authors[1]

Abstract

Phenotypic drug discovery holds promise for developing new medicines but is limited by throughput and scalability. Current application of AI to improve screening efficiency relied on single-use models trained on a phenotype-specific high throughput screen. We introduce a generalizable deep learning framework leveraging omics data to prioritize compounds for virtually any phenotype using a single model. We also developed a novel closed-loop active signature learning procedure to optimize the omics signature associated with a target phenotype. We trained our model on over 425,000 perturbation signatures and validated it using a new 1.2M-cell transcriptomics benchmark dataset profiling 88 perturbations across 10 cell lines. Our approach outperformed published methods by 15-80% and led to a 16-19X increase in productivity in two hematology phenotypic discovery campaigns, providing the first experimental validation that deep learning and omics data can improve the productivity of phenotypic discovery in a real-world setting. We next demonstrated the ability of our active signature learning algorithm to refine hit compound prioritization and gain mechanistic insights through an integrative lab-in-the-loop framework. This approach enables rational drug design targeting complex phenotypes, ushering in a new era of drug discovery.

1. Introduction

Despite steadily increasing spending in therapeutics R&D over the past 20 years (Austin & Hayford, 2021), overall clinical trial success rates have remained stagnant, with the percentage of Phase 1 compounds reaching FDA approval estimated to be between 7.9% and 13.8 (Wong et al., 2019; Hay et al., 2014; Thomas et al., 2021). As a result, the R&D expenditure of large pharmaceutical companies per newly marketed drug has soared to \$6.7B in recent years (Schuhmacher et al., 2023). Although several causes have been suggested (Scannell et al., 2012), a recurring culprit is the reductionist target-centric drug discovery model (Zheng et al., 2013; Kell, 2013; Moffat et al., 2017), which seeks to identify a single protein implicated in a disease process and then to screen for compounds that bind that protein selectively. Despite target-based discovery being the dominant paradigm for the past 30 years of drug discovery, retrospective analysis shows that more than 65% of all approved medicines were discovered via phenotypic observations, even during years when the target-based model was most popular (Sadri, 2023). This suggests that shifting focus to improving the efficiency of the phenotypic paradigm might help address the decades-long productivity crisis in drug discovery.

The critical difference between phenotypic discovery and the target-based approach is that while a target-based effort focuses on modulating the activity of a single protein, the phenotypic approach aims to modulate the behavior of an in vitro or in vivo system that accurately models the disease biology. However, there is an inherent tradeoff between the complexity of an assay, therefore its clinical translatability, and its scalability (Moffat et al., 2017). Given the lack of methodologies to accurately predict the phenotypic activity of the 10^{60} possible drug-like compounds, many phenotypic discovery programs sacrifice complexity for scale and resort to high-throughput screens (HTS) of millions of compounds against simplistic phenotypes. This reductionist brute-force bias has been identified as a critical inefficiency in drug discovery (Scannell et al., 2012; Lowe, 2012). Higher-resolution assays that measure complex information-rich phenotypes in disease-relevant cellular systems, such as a molecular signature of the disease process can enhance clinical translation (Theodoris et al., 2021). These assays have lower throughput and are more expensive. Without tools to accurately prioritize compounds to screen, deploying these more realistic, information-rich assays and models to bridge the translational gap is unfeasible.

For over 25 years, virtual screening has been to improve the productivity of target-based discovery by predicting the binding of individual molecules to protein structures (Walters et al., 1998). Seeking to generalize this approach, several groups have proposed frameworks to apply AI and machine learning to accelerate phenotypic discovery. The first generation of AI tools to predict phenotypes were models designed to predict phenotypic activity from chemical structures directly trained on readouts from an initial HTS. This approach has led to the discovery of novel antibiotics and senolytics (Stokes et al., 2020; Liu et al., 2023; Wong et al., 2023). While these models improve hit rates compared to traditional brute-force screening methods, they necessitate retraining with large datasets for each new phenotype targeted.

To overcome this limitation, researchers have proposed leveraging omics signatures as proxies for phenotypic outcomes. In this setting, compounds are prioritized based on the probability they will induce a gene expression profile associated with the desired phenotype. An initial implementation of this approach showed promise for phenotypic screening in mice (Zhu et al., 2021). However, whether these predictions yield increased productivity compared to traditional brute-force screening has not been systematically evaluated. Furthermore, all current approaches to prioritize compounds based on gene expression profiles use statistical heuristics originally designed for other bioinformatics applications like gene set enrichment to rank compounds (Subramanian et al., 2017; Chan et al., 2019; He et al., 2023). Finally, the success of omics-based prediction depends on the input gene expression profile being sufficient to induce the target phenotype. Current approaches infer gene expression signatures from correlative associations, which may not translate to the in vitro assay used to model the disease.

Here, we introduce the first closed-loop framework for omics-based prediction of complex phenotypes using machine learning and lab-in-the-loop feedback to improve the productivity of phenotypic discovery. Rather than rank compounds using statistical heuristics, we developed the first deep learning architecture optimized to directly predict whether an input gene expression profile is likely to be induced by any of a set of compounds, such as a library of purchasable chemical matter. We then performed the first comprehensive cross-tissue benchmark of compound ranking algorithms enabled by a new 1.2M cell benchmarking dataset comprising 88 chemical perturbations in 10 cell lines, which demonstrated our neural network architecture achieves state-of-the-art performance across contexts. We then performed the first systematic evaluation of omics-based predictions for two real-world phenotypic discovery campaigns in hematology, demonstrating an order of magnitude increase in productivity compared to brute-force screening. We finally introduce the first closed-loop feedback mechanism for omics-based phenotypic prediction by integrating paired phenotypic and transcriptomic measurements to refine our target input signature. We use this feedback to characterize why the model works in some cases and not others, and we show that our refined target signature is twice as effective at prioritizing molecules that modulate the phenotype of interest. Collectively, our framework enables greater productivity in phenotypic drug discovery, empowering the use of more representative and translatable phenotypes and cellular models.

2. Results

2.1. A closed-loop predictive framework to enable phenotypic discovery using deep learning

To enable greater productivity in drug discovery using complex clinically translatable phenotypic assays, we propose a closed-loop framework to nominate compounds likely to modulate a phenotype of interest (**Figure 1**). Step 1 of this framework starts with the identification of a target omics signature from clinical datasets and calibrated to a phenotypic assay. Due to the abundance of single-cell transcriptomics data, we focus on transcriptional signatures, but this framework could be applied to other omics modalities. In Step 2, a model trained on many observations of chemically induced omics signatures is used to predict compounds that will induce the desired omics signature and thereby is predicted to effect a change in phenotype. In Step 3, a limited number of these compounds are screened experimentally for phenotypic activity and hits are identified and validated. These hits are the primary output and can be used for downstream development. In Step 4, we introduce a closed-loop feedback mechanism using joint transcriptional and phenotypic measurements of hit and non-hit compounds from the previous screen. This enables refinement of the input signature moving beyond associated changes in transcription post hoc from observational data and identifying causal changes in gene expression derived from perturbation experiments. This also provides information about why some model predictions failed to validate and about the mechanisms by which hit compounds induce a change in phenotype, which enables downstream drug development.

The core of our phenotypic discovery framework is a deep learning model to identify compounds predicted to induce a change in transcriptional measurements linked to a clinically relevant phenotype (**Figure 1**). We formalize this task as a multiclass regression problem where the goal is to predict the probability that each compound in a reference library could induce each phenotype-associated signature. To optimize a model capable of this task across contexts, we selected the the Connectivity Map (CMap) as a training data set (Subramanian et al., 2017). This data set comprises mRNA perturbation signatures for 978 landmark genes following the treatment of a diversity of compounds (**Figure 2**). We filtered the full CMap dataset to 425,242 transcriptional signatures associated with 9,597 small molecules measured at multiple doses and in numerous cell lines. Our trained model, which we named DrugReflector, is an ensemble of identical multi-layer perceptron (MLP) classifiers each trained and validated on different sets of 3-fold replicate

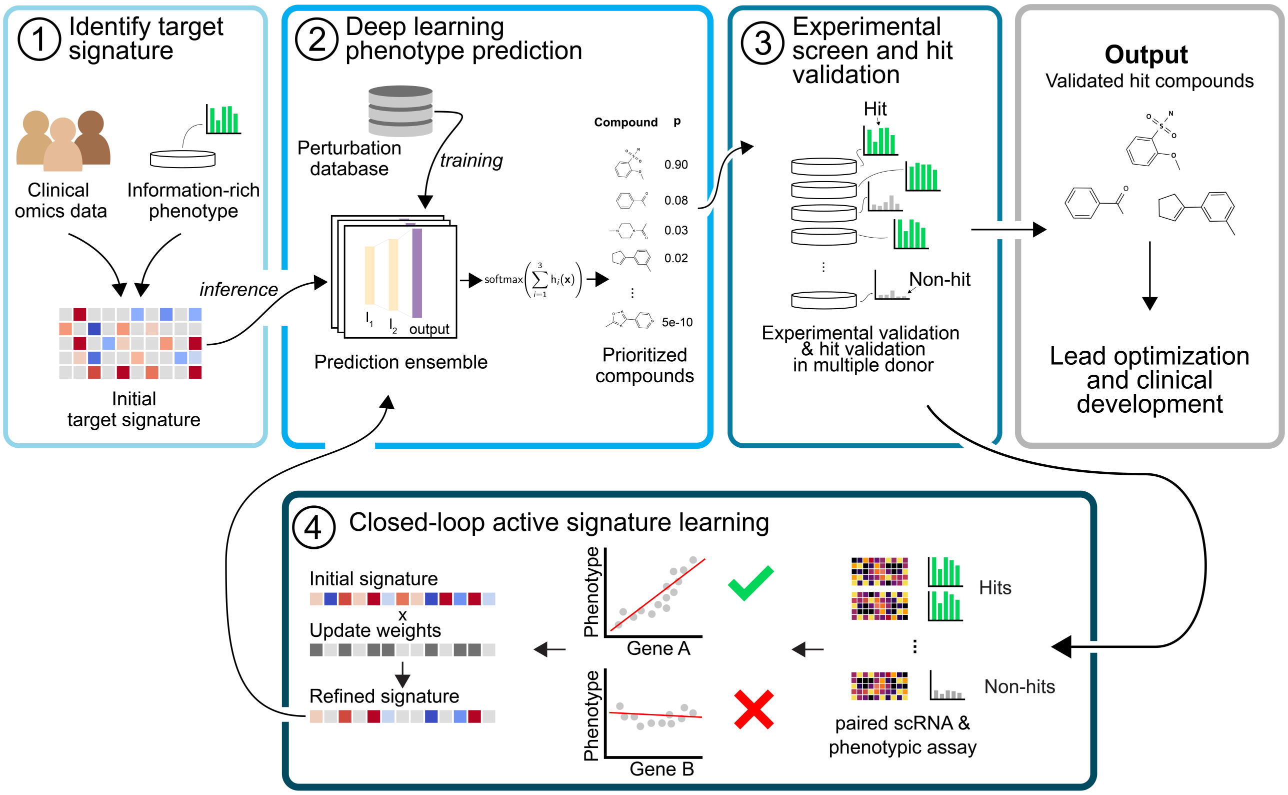

1 Identify target signature
Clinical omics data
Information-rich phenotype
inference
Initial target signature
2 Deep learning phenotype prediction
Perturbation database
training
l1 l2 output
softmax
Prediction ensemble
Compound p
0.90
0.08
0.03
0.02
5e-10
Prioritized compounds
3 Experimental screen and hit validation
Hit
Non-hit
Experimental validation & hit validation in multiple donor
Output
Validated hit compounds
Lead optimization and clinical development
4 Closed-loop active signature learning
Initial signature
x
Update weights
Refined signature
Phenotype
Gene A
Gene B
Hits
Non-hits
paired scRNA & phenotypic assay

Figure 1. A modular and generalizable framework to enable phenotypic discovery using omics-level deep learning models. (1) The first step in this framework is to identify a target omics signature based on a combination of clinical data and/or data from an information-rich clinically translatable phenotypic assay. (2) To identify compounds for screening, a deep learning model trained on perturbation signatures (such as the LINCS Connectivity Map) predicts which compounds will likely induce the target signature. (3) A limited number of compounds are then experimentally screened, compounds that induce the desired phenotype are identified, and hits are validated in multiple donors. Validated hits are the output of this discovery stage and may be used for downstream pre-clinical development. (4) The signature is iteratively refined by the lab-in-the-loop use of paired transcriptomic and phenotypic measurements, thus allowing better understanding of the mechanisms by which chemical perturbations alter target phenotypes.

splits of the training dataset using focal loss (Lin et al., 2017)

$$FocalLoss(p_t) = -(1-p_t)^{\gamma} log(p_t)$$

where p_t is the probability of the true class and γ is a tunable focusing parameter to emphasize hard-to-classify labels, aimed to increase the recall for compounds that may be observed in only a few samples or have subtle differential expression patterns.

To evaluate the performance of our model, we benchmarked DrugReflector against four approaches to match gene signatures to compounds, using top 1% compound recall as a measure of performance. The recall score is 1 if the correct compound label appears in the top 1% of all compounds predicted by the model, else 0, when the transcriptional signature of the compound is given to the model. This score is then averaged across observations for that compound in the dataset and then averaged across compounds. The comparison models include two classical baseline methods, a k-nearest neighbor (kNN) classifier and a logistic regression model. Additionally, we included two approaches that have been used to match query gene signatures to CMap perturbation signatures and for cell-type specific drug repurposing namely gene set enrichment analysis (GSEA; SigCom LINCS implementation) and Dr. Insight (Evangelista et al., 2022; Chan et al., 2019).

Our benchmarking covered three independent data sets (**Figure 2**, **Supplementary Figure 1**). First, we evaluated and compared our model on the CMap Touchstone dataset, comprising 1000 compounds tested in 9 cell lines. Our results show that DrugReflector outperformed all four algorithms, surpassing Dr. Insight by 80% and GSEA by 15% (**Figure 2c**). Second, we compared the five algorithms on the sciPlex3 dataset of 188 compounds measured in three CMap cancer cell lines (Srivatsan et al., 2020), where DrugReflector again outperformed all algorithms and outperformed Dr. Insight and GSEA by 39% and 51% on average, respectively. Finally, to examine extendibility to cell contexts not well-represented in LINCS, we generated a new scRNA-seq dataset profiling 88 compounds from CMap tested in each of 6 cancer cell lines and 4 primary cell lines, resulting in 1,737 scRNA samples with a total of 1.26M cells (**Figure 2**). The cancer cell lines are present in CMap, but the primary cell lines are either absent in the dataset or only available for a few compounds. Here, we again found that DrugReflector outperformed all algorithms, achieving an average 78% increase in recall compared with Dr. Insight and a 27% increase compared with GSEA.

2.2. Developing a complex phenotypic assay with a high clinical translatability

To demonstrate the potential of our framework to identify phenotypically active compounds, we systematically applied our framework to 2 different target phenotypes in human hematopoiesis. Hematopoiesis is an essential developmental process, and aberrant hematopoiesis can lead to numerous proliferative disorders and cytopenias. In particular, we focused our screens on modulating lineage commitment in human CD34+ hematopoietic stem and progenitor cells (HSPCs) because of their high clinical translatability. HSPC transplantation treats various blood cancers and other hematological disorders including severe anemias. In addition, HSPCs can be used as a model system to study hematological disorders, including rare disease. The hematopoietic system is also an attractive choice due to an abundance of public human scRNA data from healthy and diseased individuals that can be used to identify target gene signatures associated with various hematopoietic processes.

As phenotypic targets, we aimed to modulate the differentiation of the megakaryocyte and erythroid lineages. To characterize the cell states involved with this process, we analyzed a CITE-seq dataset we previously generated for a NeurIPS Competition in 2022 (Burkhardt et al., 2022). This joint single-cell RNA + surface protein CITE-seq dataset profiles primary HSPCs from 4 healthy donors sampled at 5 time points over a 10-day time course (**Methods**). We measured the differentiation of major lineages of the myeloid lineage at varying states at multiple time points, enabling the comprehensive capture of the maturation process. We identified progenitor and early lineage-committed cell states, including cells at a range of stages of differentiation along the megakaryocyte (Mk), erythroid (Ery), eosinophil/basophil/mast (EBM), monocyte (Mono), and neutrophil (Neu) lineage trajectories (**Supplementary Figure 2**), while observing consistency in cellular differentiation across all four donors (**Supplementary Figure 3**).

To design a phenotypic assay to measure changes in lineage differentiation, we combined the joint scRNA and surface marker dataset with literature knowledge to identify surface markers that identify each lineage. To calibrate our phenotypic assay and our reference single-cell dataset, we confirmed that RNA-defined cell types expressed surface markers consistent with our assay for both lineages (**Methods**). This analysis enabled us to define a gating strategy to identify each lineage (**Supplementary Figure 4**). For each lineage, we also identified positive control compounds and established the assay's dynamic range to facilitate the identification of phenotypically active compounds (**Supplementary Figure 5, Methods**).

To establish a hit threshold for each of the two cell-type assays, we first filtered out compounds that lead to low cell viability or in which we measured an insufficient number of cells. We then calculated a significance cutoff relative to DMSO treatment, considering the variation of DMSO samples within and across plates (**Methods**). We considered perturbations that induce the target population abundance at 6 standard deviations above DMSO as hits.

2.3. Deep learning-enabled phenotypic discovery to induce megakaryopoiesis

To nominate compounds for screening in each lineage, we identified cell state transitions associated with early differentiation into megakaryocytes and erythrocyte progenitors. We used that transition as input to our model to prioritize compounds. To generate transition-associated signatures, we derived a statistic with similar properties to the Z-score representation of CMap Level 4 signatures used for model training. While CMap's Z-score quantifies differential expression effect size across plates of the L1000 assay, we computed a v-score to estimate the standardized difference in log-counts means between the two cell populations, accounting for each population's variance.

$$vscore(x, y) = \frac{\mathbb{E}(\log(1+y)) - \mathbb{E}(\log(1+x))}{\sqrt{Var(\log(1+x)) + Var(\log(1+y))}}$$

We input these v-scores to the model and ordered the top compounds from the model's output to assess their ability to induce the phenotype of interest. For this study, we relied on an inventory of 1,635 compounds from the CMap training set available to us at the time of study initiation.

To generate predictions for the megakaryocyte lineage, we focused on the bipotential megakaryocyte erythroid progenitor (MEP), the earliest cell state associated with lineage decision giving rise to the erythroid (Ery) and megakaryocyte (Mk) lineages (McDonald & Sullivan, 1993). We reasoned that this was the optimal point to intervene in differentiation

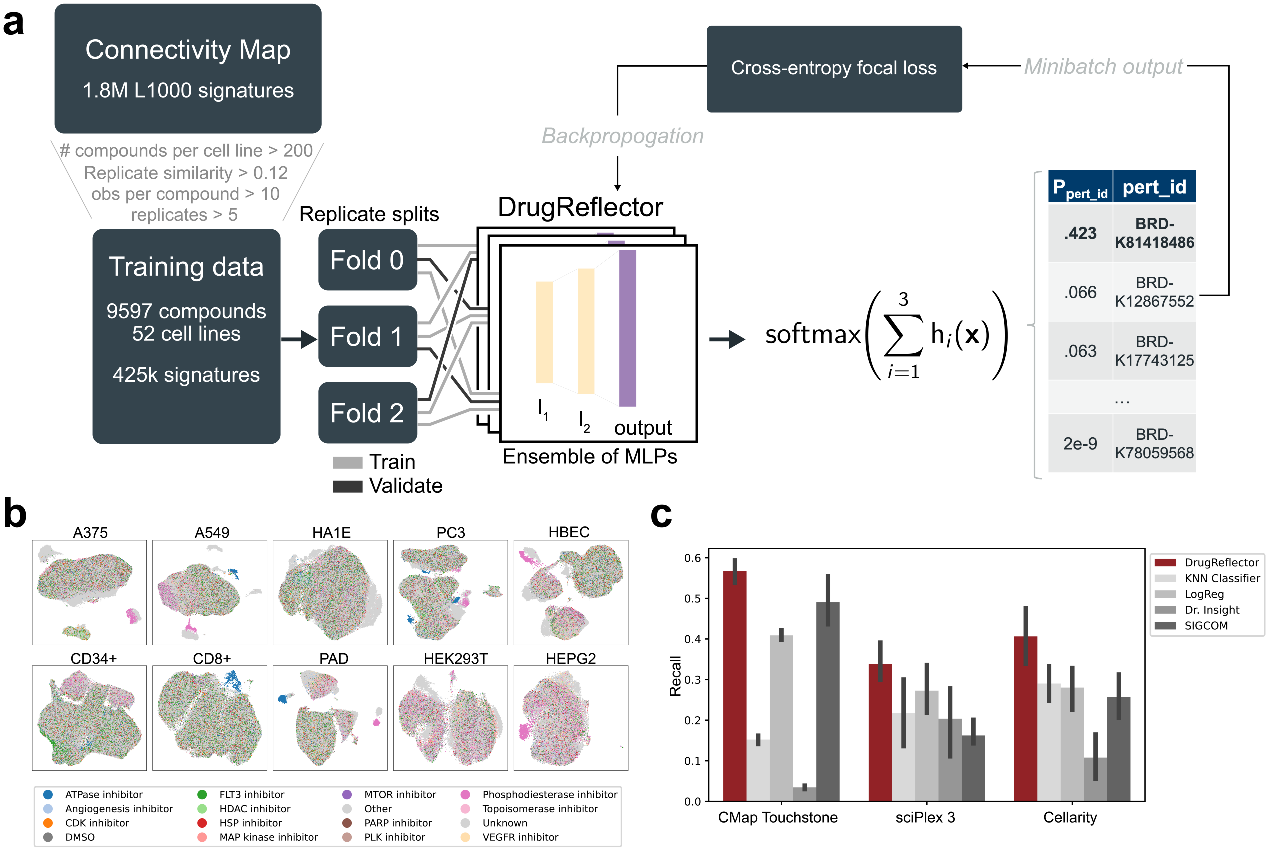


Figure 2. A deep learning approach to phenotypic virtual screening. (a) A schematic representation of the model training regime showing the input and output for a single example from CMap. (b) UMAP embeddings of all cells from our benchmarking dataset of 1.2M single-cell transcriptomes under perturbation of 88 compounds tested in 10 cell types in duplicate. Color denotes the compound mechanism of action annotated by CMap. (c) A boxplot showing performance of each algorithm on each benchmarking dataset averaged across cell lines. Error bars denote standard deviation across cell lines.

because transcriptional and metabolic changes in these cells are associated with commitment to differentiation into either lineage (Lu et al., 2018). We aimed to alter the MEP cells to adopt a transcriptional state similar to the MPC population, which are progenitors committed to differentiating towards the Mk lineage. To define a differential expression statistic for single-cell transcriptomics data with similar properties to the z-scores used in the CMap training data, we derived a v-score, short for variance-score (**Methods**). We calculated v-scores from the MEP to MPC population and used these v-scores as input to Drug Reflector to obtain a prioritized list of compounds for screening. We confirmed that the 1,635 compounds in our inventory were a representative subset of all compounds ranked by the model (**Supplementary Figure 6**).

To experimentally determine which compounds induced our target phenotype, we treated CD34+ cells with each model-nominated compound under HSPC maintenance conditions (CC100/TPO cytokines) (**Methods**). On day 7, we evaluated the induction of CD41a+ CD71- CD42b+ Mk population by flow cytometry. We tested 107 compounds with a rank less than 1,000 prioritized by our model, and to compare to the brute-force approach, we also tested a random selection of 96 compounds from the same compound inventory.

Among our 107 highly ranked DrugReflector-nominated compounds, we identified 21 above our 6 standard deviations hit threshold, resulting in a 19.6% hit rate (**Figure 3**). 2 compounds were highly active inducing more than a 4-fold increase in Mk progenitors. By contrast, we identified only 1 compound from our random selection that passed our hit threshold, resulting in a 1.1% hit rate. These results highlight that our deep learning model enriches the selection of compounds that modulate cell state transitions of interest greater than 19-fold compared to a traditional screening approach.

To confirm that these hits validated in multiple donors, we re-tested 17 DR-nominated hit compounds in 2 additional donors at the dose at which we observed maximal induction of the Mk lineage. While 2 compounds did not pass our viability or cell count criteria, 13 out of the remaining 15 hit compounds validated in both donors, demonstrating the robustness in our assay and biological translation of our chemical perturbations across different donors (**Figure 3**).

2.4. Deep learning-enabled phenotypic discovery to induce erythropoiesis

Next, we sought to demonstrate the generalizability of our framework by aiming to bias the MEP population towards an alternative fate decision: erythroid progenitor cells. Like our previous transition, we calculated v-scores between the MEP and Ery erythroid progenitor population to derive an input signature for our DrugReflector model. We then obtained the top 96 compounds from our DR-nominated compounds list and a new set of 96 random compounds for screening.

To experimentally determine which compounds induced Ery progenitors, we treated CD34+ cells with each of the model-nominated compounds and with each of the randomly selected compounds at two doses (1µM and 10µM), dropping the 100nM dose because we observed that few compounds were maximally active at that level. We again cultured and treated donor-derived CD34+ HSPCs with each compound over 7 days and measured Ery lineage abundance using flow cytometry (**Methods**).

In our DR-nominated compound screen, after removing samples failing our quality control filter, we observed 13 out of 81 compounds passing our 6-standard deviation above the DMSO hit cutoff, representing a 16% hit rate. In our randomly selected compound set, we observed only 1 out of 85 compounds inducing Ery progenitors above our cutoff, representing a 1.2% hit rate (**Figure 3**). Again, our transcriptomics-based compound prioritization significantly increased our success rate in inducing the desired phenotype ($\sim$16X). We evaluated how the identified hits performed across additional donors as part of our cross-donor validation. Out of 10 compounds passing our quality control filter in our validation experiment, 5 significantly increased Ery progenitors in both donors, and 3 more did so in one of the two. These results provide further support for the capacity of our machine learning model to increase our phenotypic hit rate across multiple experimental settings.

2.5. Closed-loop signature refinement in the megakaryocyte lineage

A significant advantage of our phenotypic discovery framework is the ability to pair transcriptional and phenotypic measurements to better understand the underlying mechanisms governing the target phenotypic response. This enables us to refine our input target signature based on transcriptional differences between hits and non-hits and to learn the changes in gene expression following compound perturbation associated with changes in phenotype. To explore the utility of this approach, we performed a scRNA-seq time course on 12 hits, 8 non-hits, a DMSO negative control, and our positive control compound with samples collected in duplicate at days 0, 1, 2, 5, and 7 for a total of 192 scRNA datasets. The non-hits were included to explore why these compounds did not induce our target phenotype despite being prioritized by DrugReflector and to use this information to refine our predictions. We also collected paired phenotypic data on day 7 from the same samples used for scRNA-seq. Across the scRNA samples, we recovered 145,157 cells with a median of 754 cells per condition. We integrated this scRNA dataset with the original time course using Harmony to facilitate comparison with the original time course (Korsunsky et al., 2019). In our transcriptional time course, we observed cells from all major expected cell types. We also observed a strong correlation between the abundance of Mk cells as determined by our scRNA-seq and phenotypic measurements (**Supplementary Figure 7**).

Using our transcriptional validation, we first sought to understand why not all prioritized compounds from the DR model were phenotypically active. We reasoned that one cause could be compounds having a cell-type specific effect in CD34+ cells that differs from the impact measured in the CMap dataset. 43% of compounds in CMap were previously reported to exhibit cell-type specific effects in the cancer cell lines (Subramanian et al., 2017), and CD34+ HSPCs were absent from the training data. To test this explicitly, we calculated for each compound the distance between the 24-hour signatures in our follow-up experiment and the 10 most similar signatures for the same compound in LINCS (**Methods**), providing an unbiased estimate of the similarity between our observed perturbation signatures in CD34+ cells and signatures for the same compounds in CMap. We found that, on average, CD34+ signatures of non-hit compounds were 11% further from their closest neighbors in CMap compared with hit compounds (**Figure 4**, p=0.037, paired t-test). Although this distance to CMap can only be calculated using L1000 landmark genes, most cell-type-specific perturbation effects fall outside the landmark gene set based on our benchmarking dataset (**Supplementary Figure 8**). This suggests that developing strategies to map transcriptional signatures associated with compounds into new cell types and across all genes will likely improve vir-

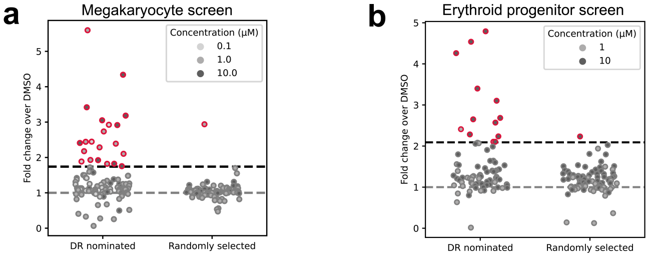


Figure 3. A deep learning approach to phenotypic virtual screening. (a) Result of experimental validation of compounds to induce Mk differentiation measured with flow cytometry following a 7-day in vitro differentiation in the presence of each compound. Each dot is a compound. The color is the dose at which the compound maximally induced Mk abundance. The grey dashed line denotes a fold-change of 1 relative to DMSO, i.e. no change. The black dashed line represents the hit significance cutoff for Mk. (b) Same as for (a), but for the Ery discovery campaign.

tual screening performance.

Next, we aimed to refine our signature to identify the gene expression patterns sufficient to induce a change in phenotype. We hypothesized that only a subset of our initial input signature was required to effect the change in phenotype, with the remainder comprising passenger genes, noise, or potentially inhibitory feedback gene expression signals. To test this hypothesis, we first performed DE analysis on the 24-hour gene expression counts summed across all cells in each cell type, also called pseudobulk expression. Here, we chose to use limma to calculate differential expression (Ritchie et al., 2015), enabling us to explicitly model experimental covariates such as library and plate. However, instead of using the compound perturbation label as the explanatory variable, we used the 7-day fold-change variable as the basis for DE (**Methods**). Because the fold-change is a scalar variable, this implementation identifies genes that are linearly associated with the induction of the Mk lineage while controlling for various technical confounding variables. We observed 672 landmark genes that are significantly associated with Mk induction (adj. $p < 0.01$), including genes previously implicated in Mk maturation, like FLI1, GATA2, and NFE2 (**Figure 4**).

We next compared the differential expression score to the original input v-score for each gene. We observed three patterns: genes that are concordantly associated (n=366), inversely associated (n=312), and unassociated with the original input v-score (**Figure 4**). We asked whether the concordantly associated genes could be used as a refined input to our model. When filtering the input v-scores to include only genes concordant between the target transition and our transcriptional validation experiment, the median rank of hit compounds improved significantly, from 1,060 to 375. To understand the significance of this shift, we generated 10,000 random gene sets with the same size as the concordant gene set as a background distribution. We measured the median hit rank after filtering to each random set. Concordant genes performed significantly better than random gene sets by median hit rank (**Figure 4**, p=0.0026 paired t-test). These results confirm our hypothesis that only part of our original input signature is necessary to prioritize hit compounds and offer a proof-of-concept strategy to identify the essential component of the signature.

3. Discussion

Here, we performed phenotypic discovery campaigns across two lineages of hematopoiesis using a state-of-the-art deep learning classifier that enables the nomination of compounds to induce cell phenotypes based on transcriptional signatures. These discovery efforts are facilitated by a modular and generalizable framework that links chemistry and phenotypic activity using omics-level data. We provide one implementation of this framework using transcriptional data that achieves a 16-19X improvement in hit rate compared with brute-force screening in head-to-head experiments. This enables us to leverage existing datasets profiling cell states across numerous tissues in health and disease contexts and large perturbational databases like the LINCS CMap. Indeed, the idea of matching compounds to transcriptomic signatures has been suggested for several applications, focusing primarily on drug repurposing. Our approach provides a less biased and more efficient method to query disease biology, providing the opportunity to find novel biology associated with a given cellular process and link it to chemical structure. Furthermore, we can use feedback from assay results to improve predictions and better understand disease biology. Such lab-in-the-loop refinement is necessary to realize the promise of AI-guided scientific discovery (Wang et al., 2023).

A key feature of our paradigm is its modular nature and the ability to optimize each component independently. For example, identifying target signatures based on healthy and diseased patient samples is a rapidly developing field, and even the most recent methods focus on differential expression between cell states (He et al., 2023), as we do here.

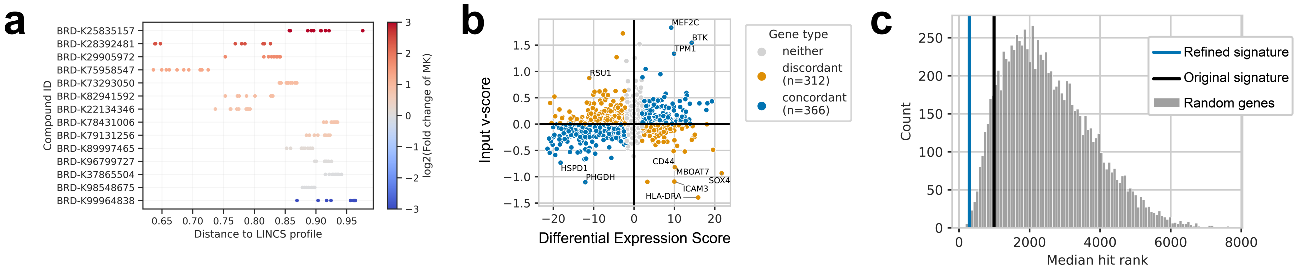


Figure 4. A deep learning approach to phenotypic virtual screening. (a) Result of experimental validation of compounds to induce Mk differentiation measured with flow cytometry following a 7-day in vitro differentiation in the presence of each compound. Each dot is a compound. The color is the dose at which the compound maximally induced Mk abundance. The grey dashed line denotes a fold-change of 1 relative to DMSO, i.e. no change. The black dashed line represents the hit significance cutoff for Mk. (b) Same as for (a), but for the Ery discovery campaign.

However, future work will likely leverage more advanced strategies, such as identifying driver genes based on fate mapping (Lange et al., 2022) or causal inference of regulatory relationships between genes (Kamimoto et al., 2023).

There is also a need for better training datasets. Although it is the largest publicly available dataset of its kind, CMap is fundamentally limited. The L1000 assay used in the CMap is noisy (Qiu et al., 2020) and only measures 978 genes. During model development, we explored whether using the inferred 11,350 for model training would improve performance. Still, we found that it led to overfitting to the training dataset and poor performance on the two test datasets (**Supplementary Figure 13**). This leads us to question the ability of models trained on CMap inferred genes to predict transcriptomic signatures for compounds not in CMap. Moreover, almost all the data is measured in cancer cell lines, which we showed can fail to generalize to primary cell types. To provide a better basis for prediction in our discovery efforts, we are building a dataset of perturbation signatures tailored to the therapeutic areas we focus on.

Considering the experimental screening and hit validation, we applied a straightforward selection strategy by picking the top-ranked compounds output by the model. However, further improvements in screening efficiency are likely to arise from more sophisticated compound selection approaches. In reinforcement learning, acquisition functions balance exploration and exploitation to identify a set of actions, such as compounds to test, to maximize a reward function, such as the hit rate. Although it may be challenging to directly apply online learning algorithms to phenotypic drug discovery due to the latency involved with doing rounds of experiments, these concepts will likely lead to more efficient screening.

Finally, methods to characterize the impact of experimental perturbations on single-cell datasets are only a few years old. Only recently have methods been proposed to learn causal relationships between genes considering perturbation data (Jiang et al., 2023). There is a vast opportunity to improve these tools to learn causal dynamics. Integrating these approaches is likely to lead to better signature refinement. Here, we took a straightforward approach using linear regression with our hit phenotype, but more sophisticated approaches are possible. For example, much as linear driver genes are identified using trajectory inference algorithms (Lange et al., 2022), we can imagine using differential driver gene analysis to identify genes associated with specific hit compounds.

This framework has broad utility for phenotypic discovery across disease settings. Thanks to a surge in single-cell datasets across diseases, it is possible to use existing single-cell atlases to derive an initial target signature for dozens of indications. Calibrating these initial signatures to the dataset used in a phenotypic assay may be necessary but will likely require less data than needed for the original atlas (Dann et al., 2022). We anticipate that this paradigm will reduce the need for brute force screening, enabling lower throughput and higher translatable models such as patient-derived organoids, tissues-on-a-chip, or even explants to drive more productive drug discovery.

References

Austin, D. and Hayford, T. Research and development in the pharmaceutical industry. Technical report, 4 2021. URL https://www.cbo.gov/publication/57126. [Online; accessed 2023-09-25].

Burkhardt, D., Luecken, M., Benz, A., Holderrieth, P., Bloom, J., Lance, C., Chow, A., and Holbrook, R. Open problems - multimodal single-cell integration competition, 2022. URL https://kaggle.com/competitions/open-problems-multimodal.

Chan, J., Wang, X., Turner, J. A., Baldwin, N. E., and Gu, J. Breaking the paradigm: Dr insight empowers signature-free, enhanced drug repurposing. *Bioinformatics*, 35(16): 2818–2826, 8 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz006.

Dann, E., Teichmann, S. A., and Marioni, J. C. Precise identification of cell states altered in disease with healthy single-cell references. 11 2022. doi: 10.1101/2022.11.10.515939. URL https://www.biorxiv.org/content/10.1101/2022.11.10.515939v1. page: 2022.11.10.515939 section: New Results.

Evangelista, J. E., Clarke, D. J. B., Xie, Z., Lachmann, A., Jeon, M., Chen, K., Jagodnik, K., Jenkins, S. L., Kuleshov, M., Wojciechowicz, M., Schürer, S., Medvedovic, M., and Ma'ayan, A. Sigcom lincs: data and metadata search engine for a million gene expression signatures. *Nucleic Acids Research*, 50(W1):W697–W709, 7 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac328.

Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., and Rosenthal, J. Clinical development success rates for investigational drugs. *Nature Biotechnology*, 32(1): 40–51, 1 2014. ISSN 1546-1696. doi: 10.1038/nbt.2786. PMID: 24406927.

He, B., Xiao, Y., Liang, H., Huang, Q., Du, Y., Li, Y., Garmire, D., Sun, D., and Garmire, L. X. Asgard is a single-cell guided pipeline to aid repurposing of drugs. *Nature Communications*, 14(1):993, 2 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36637-3. number: 1 publisher: Nature Publishing Group.

Jiang, J., Chen, S., Tsou, T., McGinnis, C. S., Khazaei, T., Zhu, Q., Park, J. H., Strazhnik, I.-M., Hanna, J., Chow, E. D., Sivak, D. A., Gartner, Z. J., and Thomson, M. D-spin constructs gene regulatory network models from multiplexed scrna-seq data revealing organizing principles of cellular perturbation response. 5 2023. doi: 10.1101/2023.04.19.537364. URL https://www.biorxiv.org/content/10.1101/2023.04.19.537364v3. page: 2023.04.19.537364 section: New Results.

Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., and Morris, S. A. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949):742–751, 2 2023. ISSN 1476-4687. doi: 10.1038/s41586-022-05688-9. number: 7949 publisher: Nature Publishing Group.

Kell, D. B. Finding novel pharmaceuticals in the systems biology era using multiple effective drug targets, phenotypic screening and knowledge of transporters: where drug discovery went wrong and how to fix it. *The FEBS Journal*, 280(23):5957–5980, 2013. ISSN 1742-4658. doi: 10.1111/febs.12268.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296, 12 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0619-0. Citation Key: korsunsky-FastSensitiveAccurate2019 publisher: Nature Publishing Group tex.copyright: 2019 The Author(s), under exclusive licence to Springer Nature America, Inc.

Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H. B., Pe'er, D., and Theis, F. J. Cellrank for directed single-cell fate mapping. *Nature Methods*, 19(2):159–170, 2 2022. ISSN 1548-7105. doi: 10.1038/s41592-021-01346-6. number: 2 publisher: Nature Publishing Group.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollar, P. Focal loss for dense object detection. pp. 2980–2988, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Lin_Focal_Loss_for_ICCV_2017_paper.html. [Online; accessed 2023-10-02].

Liu, G., Catacutan, D. B., Rathod, K., Swanson, K., Jin, W., Mohammed, J. C., Chiappino-Pepe, A., Syed, S. A., Fragis, M., Rachwalski, K., Magolan, J., Surette, M. G., Coombes, B. K., Jaakkola, T., Barzilay, R., Collins, J. J., and Stokes, J. M. Deep learning-guided discovery of an antibiotic targeting acinetobacter baumannii. *Nature Chemical Biology*, 19(11):1342–1350, 11 2023. ISSN 1552-4469. doi: 10.1038/s41589-023-01349-8. publisher: Nature Publishing Group.

Lowe, D. The brute force bias, 3 2012. URL https://www.science.org/content/blog-post/brute-force-bias. [Online; accessed 2023-11-22].

Lu, Y. C., Sanada, C., Xavier-Ferrucio, J., Wang, L., Zhang, P. X., Grimes, H. L., Venkatasubramanian, M., Chetal, K., Aronow, B., Salomonis, N., and Krause, D. S. The molecular signature of megakaryocyte-erythroid progenitors reveals a role for the cell cycle in fate specification. *Cell Rep*, 25(8):2083–2093, 11 2018. Citation Key: pmid30463007.

McDonald, T. P. and Sullivan, P. S. Megakaryocytic and erythrocytic cell lines share a common precursor cell. *Experimental Hematology*, 21(10):1316–1320, 9 1993. ISSN 0301-472X. PMID: 8135919.

Moffat, J. G., Vincent, F., Lee, J. A., Eder, J., and Prunotto, M. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature Reviews Drug Discovery*, 16(8):531–543, 8 2017. ISSN 1474-1784. doi: 10.1038/nrd.2017.111. number: 8 publisher: Nature Publishing Group.

Qiu, Y., Lu, T., Lim, H., and Xie, L. A bayesian approach to accurate and robust signature detection on lincs l1000 data. *Bioinformatics*, 36(9):2787–2795, 5 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa064. PMID: 32003771 PMCID: PMC7203754.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 4 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv007.

Sadri, A. Is target-based drug discovery efficient? discovery and “off-target” mechanisms of all drugs. *Journal of Medicinal Chemistry*, 9 2023. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.2c01737. URL https://doi.org/10.1021/acs.jmedchem.2c01737. publisher: American Chemical Society.

Scannell, J. W., Blanckley, A., Boldon, H., and Warrington, B. Diagnosing the decline in pharmaceutical r&d efficiency. *Nature Reviews Drug Discovery*, 11(3):191–200, 3 2012. ISSN 1474-1784. doi: 10.1038/nrd3681. number: 3 publisher: Nature Publishing Group.

Schuhmacher, A., Hinder, M., von Stegmann Und Stein, A., Hartl, D., and Gassmann, O. Analysis of pharma r&d productivity - a new perspective needed. *Drug Discovery Today*, 28(10):103726, 10 2023. ISSN 1878-5832. doi: 10.1016/j.drudis.2023.103726. PMID: 37506762.

Srivatsan, S. R., McFaline-Figueroa, J. L., Ramani, V., Saunders, L., Cao, J., Packer, J., Pliner, H. A., Jackson, D. L., Daza, R. M., Christiansen, L., Zhang, F., Steemers, F., Shendure, J., and Trapnell, C. Massively multiplex chemical transcriptomics at single-cell resolution. *Science (New York, N.Y.)*, 367(6473):45–51, 1 2020. ISSN 0036-8075. doi: 10.1126/science.aax6234. PMID: 31806696 PMCID: PMC7289078.

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., and Collins, J. J. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13, 2 2020. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2020.01.021. publisher: Elsevier PMID: 32084340.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., Lahr, D. L., Hirschman, J. E., Liu, Z., Donahue, M., Julian, B., Khan, M., Wadden, D., Smith, I. C., Lam, D., Liberzon, A., Toder, C., Bagul, M., Orzechowski, M., Enache, O. M., Piccioni, F., Johnson, S. A., Lyons, N. J., Berger, A. H., Shamji, A. F., Brooks, A. N., Vrcic, A., Flynn, C., Rosains, J., Takeda, D. Y., Hu, R., Davison, D., Lamb, J., Ardlie, K., Hogstrom, L., Greenside, P., Gray, N. S., Clemons, P. A., Silver, S., Wu, X., Zhao, W.-N., Read-Button, W., Wu, X., Haggarty, S. J., Ronco, L. V., Boehm, J. S., Schreiber, S. L., Doench, J. G., Bittker, J. A., Root, D. E., Wong, B., and Golub, T. R. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, 11 2017. ISSN 1097-4172. doi: 10.1016/j.cell.2017.10.049. Citation Key: subramanianNextGenerationConnectivity2017 PMID: 29195078 tex.pmcid: PMC5990023.

Theodoris, C. V., Zhou, P., Liu, L., Zhang, Y., Nishino, T., Huang, Y., Kostina, A., Ranade, S. S., Gifford, C. A., Uspenskiy, V., Malashicheva, A., Ding, S., and Srivastava, D. Network-based screen in ipsc-derived cells reveals therapeutic candidate for heart valve disease. *Science*, 371 (6530):eabd0724, 2 2021. doi: 10.1126/science.abd0724. publisher: American Association for the Advancement of Science.

Thomas, D., Chancellor, D., Micklus, A., LaFever, S., Hay, M., Chaudhuri, S., Bowden, R., and Lo, A. W. Technical report: Clinical development success rates and contributing factors 2011-2020 | biotechnology innovation organization. Technical report, 2 2021. [Online; accessed 2023-10-13].

Walters, W. P., Stahl, M. T., and Murcko, M. A. Virtual screening—an overview. *Drug Discovery Today*, 3(4):160–178, 4 1998. ISSN 1359-6446. doi: 10.1016/S1359-6446(97)01163-X.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., Marks, D., Ramsundar, B., Song, L., Sun, J., Tang, J., Veličković, P., Welling, M., Zhang, L., Coley, C. W., Bengio, Y., and Zitnik, M. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 8 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06221-2. number: 7972 publisher: Nature Publishing Group.

Wong, C. H., Siah, K. W., and Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics (Oxford, England)*, 20(2):273–286, 4 2019. ISSN 1465-4644. doi: 10.1093/biostatistics/kxx069. PMID: 29394327 PMCID: PMC6409418.

Wong, F., Omori, S., Donghia, N. M., Zheng, E. J., and Collins, J. J. Discovering small-molecule senolytics with deep neural networks. *Nature Aging*, 3(6): 734–750, 6 2023. ISSN 2662-8465. doi: 10.1038/s43587-023-00415-z. publisher: Nature Publishing Group.

Zheng, W., Thorne, N., and McKew, J. C. Phenotypic screens as a renewed approach for drug discovery. *Drug Discovery Today*, 18(21-22):1067–1073, 11 2013. ISSN 1878-5832. doi: 10.1016/j.drudis.2013.07.001. PMID: 23850704 PMCID: PMC4531371.

Zhu, J., Wang, J., Wang, X., Gao, M., Guo, B., Gao, M., Liu, J., Yu, Y., Wang, L., Kong, W., An, Y., Liu, Z., Sun, X., Huang, Z., Zhou, H., Zhang, N., Zheng, R., and Xie, Z. Prediction of drug efficacy from transcriptional profiles with deep learning. *Nature Biotechnology*, 39(11):1444–1452, 11 2021. ISSN 1546-1696. doi: 10.1038/s41587-021-00946-z. number: 11 publisher: Nature Publishing Group.

SUPPLEMENT: Deep learning virtual screening with active signature learning improves the identification of small-molecule modulators of complex phenotypes

Methods

Data availability

Transcriptomics data from the benchmarking dataset, the HSPC reference atlas, and the paired phenotypic and transcriptomic screen is planned to be made available in GEO under a Creative Commons license at the time of publication in an archival journal.

Data availability

Code to run the DrugReflector algorithm and to reproduce the major results of this manuscript is planned to be made available on GitHub under an open-access license at the time of publication in an archival journal.

DrugReflector algorithm overview

Model Architecture

The DrugReflector classifier is an ensemble of three fully-connected neural networks implemented in PyTorch[1]. Each network has two hidden layers with the same structure but separate parameters. The input layer has 978 nodes (one for each landmark gene), and the output layer has 9,597 nodes (one for each target LINCS perturbation). The first hidden layer has 1,024 nodes, and the second has 2,048 nodes using rectified linear units (ReLU) to compute node activations.

To generate predictions, we split the data into three folds based on replicate labels from CMap and trained this model architecture independently on each fold. The three models were then ensembled for inference. The final predicted class probabilities were the softmax probabilities of the average score over all three folds. To compute final ranks, we ranked the average score across all three models. Higher scores were ranked lower (i.e., closer to 0).

Curating LINCS CMap into a training dataset

The starting point for training was the LINCS CMap 2020 Level 4 dataset[2], which we obtained from https://lincsportal.ccs.miami.edu/datasets/view/LDS-1611. The level 4 dataset contains differential expression z-scores for each compound against all values measured on the same platen. We then filtered out observations according to quality control criteria. The criteria were as follows:

1. Remove any diversity-oriented synthesis (DOS) compounds that are difficult to procure
2. Remove any compounds with fewer than 5 observations in total

3. For each compound, remove any observations with a cosine similarity <0.12 to the closest replicate
4. For each compound, select the most frequently recorded dose between 1μM and 20μM
5. Keep only measurements recorded at 6-hour or 24 hours post-treatment
6. After applying the first four filters, remove any compounds measured in fewer than 5 cell lines, more than 40 cell lines, or with fewer than 3 replicates.

We next applied the following chemical filters:

1. Molecular weight must be between 60 and 1,000 (inclusive)
2. No more than 1 covalent motif (defined by SMARTS[3])
3. No more than 9 NIBR structure flags[4]
4. Pass BRENK critieria[5]
5. Must not match 30 SMARTS patterns (exact patterns not disclosed).

Applying these filters, we retained 425,242 observations comprising 9,597 small molecules measured in 52 cell lines, with a median of 32 observations per compound and 751 compounds measured more than 100 times.

In addition, every transcriptional vector v is clipped to range [-2,2] such that is standard deviation after clipping equals 1. These gives a hybrid representation between a binarization of the data (up vs down regulation) and high-variance continuous values, while normalizing its scale.

Model inputs

The input to DrugReflector is a representation of a desired transition between two cellular states. While during training these are measured chemical perturbations (e.g. CMap data), these are differences in cell populations found in a clinical data set during model deployment (e.g. a single-cell atlas). Therefore, to obtain high performance, it is crucial to account for "domain shift" in the data, i.e. to make the clinical cell transitions look like the data the model was trained on. For this reason, we used the same principles to construct the input representation from single-cell data during model deployment as LINCS does for perturbation effects: we represented gene log fold-changes in units of standard deviations of log expression. Whereas in LINCS, these standard deviations are obtained from other wells on the same plate, single-cell data enables us to measure gene standard deviations from other cells from the same state.

Using these considerations, we defined the *v-score* for a gene measured in two states as follows:

$$\text{v-score}(x \to y) = \frac{E\big(log(1+y)\big) - E\big(log(1+x)\big)}{\sqrt{Var\big(log(1+x)\big) + Var\big(log(1+y)\big)}}$$

where x and y are the transcript counts in each state, normalized to a fixed total count per cell. A pseudocount of 1 was added to each logarithm to avoid the singularity at 0. Similar to the training data, we clip the v-score vector to range [-2,2] such that it has standard deviation 1 after clipping. This ensures the training and test data have similar scale.

Unlike the t-score, the v-score does not depend on the number of cells in each group in theexpectation. As in LINCS level 4, differences are measured in units of standard deviation. A high v-score is obtained when genes have different mean log expression between the two states but relatively low standard deviation of log expression within them.

Training regime

The training data was divided randomly into three folds, with perturbation replicates balanced across the folds. Models were independently trained on two of three folds, and the resulting class scores were averaged to give the final class ranks.

The models were trained using a focal loss function[6] on the softmax probabilities, with a focusing parameter of $\gamma = 2$.To ensure robustness of the trained models, each hidden layer randomly zeroed some of its inputs with a fixed dropout probability of 0.64. We applied batch normalization[7] during model training using momentum = 0.1. The learning rate was determined by a cosine annealing schedule with warm restarts[8], with 20 epochs before the first restart, an initial learning rate of 0.0139, and a minimum learning rate of 0.00001. Each model was trained for a total of 50 epochs.

The above dropout probability, initial learning rate, and time to first warm restart were determined via hyperparameter optimization using Optuna[9]. To evaluate a particular set of hyperparameters, we trained the parameterized model on two of the three folds, and measured recall of the held-out compound signatures in the third fold. The hyperparameters with highest average recall were used in training. A summary of the hyperparameter search is in **Table 1**.

An overview of training is provided in **Algorithm 1**.

Parameter	Range tested	Value selected
Dropout	(0.2, 0.8]	0.64
Initial learning weight	(1e-4,1e-1]	0.0139
Weight Decay	(1e-7,1e-1]	1e-5
Time to first restart	10-50	20

Table 1 – An overview of the hyperparameter search for DrugReflector.

Algorithm 1: Training DrugReflector

```
Input: Training Data $D_{all}$
Hyperparameters:
Focal loss with $\gamma = 2$
Dropout with $p = 0.64$
Batch normalization with $\text{momentum} = 0.1$
```

```
Warm restart cosine annealing with $T_0 = 20, \eta_{\min} = 1 \times 10^{-5}$, and $\eta = 0.0139$
**Output:** Trained Models $\mathcal{M}_{ensemble} = \{\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2\}$
Let training data $D_{\text{all}} = \{\mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2\}$ be split into 3 folds
Folds $\mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2$ are balanced across perturbation replicates
**for** $k = 0$ to $2$ **do**
    Training data $D_{train} \leftarrow D_{all} \setminus \{F_k, F_{(k+1)\ \%\ 3}\}$
    Train $M_k$ on $D_{train}$ for 50 epochs, with an early stop at 20 epochs
**end for**
```

Model Benchmarking

Curating the CMap Touchstone Dataset

We filtered our curated CMap level 4 training data to the 9 cell lines of the CMap touchstone dataset: A375, A549, HA1E, HCC515, HEPG2, HT29, MCF7, PC3, and VCAP. We then selected 1,000 compounds that have samples in all 9 cell lines based on number of observations per cell line. For some compounds, there was a long tail in some cell lines, so we only considered the first 30 observations per compound per cell line. We then calculated the mean number of observations per cell line and took the top 1,000 compounds. For this set of 1,000 compounds, we subsampled uniform random from each of the 9,000 combinations of cell line and compounds to generate a dataset of 9,000 samples.

For this dataset, we benchmarked ensemble models and KNN differently to ensure we did not mix test and train data.

DrugReflector and softmax regression are ensembles, each containing three models. Each model has a unique test fold in the curated CMap dataset, and is trained on the remaining two data folds. When we ran ensemble benchmarks for this dataset, we ran individual predictions absent in the data folds for each model. Then, we computed the recall per compound by cell line. If the rank of the query compound was in the top 1% of the output of the compound from the model, the recall was 1, else 0. We then averaged the recall across compounds for each cell line and averaged the recall across cell lines for each model for final ensemble performance.

To benchmark k-nearest neighbors, we tested over each of the three folds and set the reference dataset to the other two folds. We computed the recall per compound by cell line, and then averaged the recall per model for final ensemble performance.

For public methods, SigCom LINCS and Dr. Insight, we uniformly subsampled 500 observations due to long runtimes.

Curating the sciPlex3 Dataset

We curated a public GEO (Gene Expression Omnibus) dataset GSE139944. This dataset tests small molecule inhibitors on A549, MCF7, and K562 cells. We downloaded the pre-processed version of the dataset and calculated v-scores as described above between each treatment condition and DMSO. The v-scores were used as input to each model.

Generating and curating the Intervention Library Dataset

Human cell lines

A375, A549, HepG2, PC3, and HEK293T cell lines were purchased from the American Tissue Culture Collection. The human-embryonic kidney cell line, HA1E, was generously provided by the Cancer Cell Line Encyclopedia at the Broad Institute. Cells were cultured in either RPMI or DMEM with fetal bovine serum as per suppliers' recommendations. Cells were seeded into 24-well dishes and incubated for 24 hours at 37°C and then treated with compounds for 24 hours at a dose previously determined to be the maximum-tolerated dose for the six cell lines. Cells were harvested with 0.05% Trypsin and collected with serum-containing media.

Human bronchial epithelial cells

Normal human bronchial epithelial cells (HBEC) were obtained from Lonza from two healthy donors. Cells were thawed and grown in PneumaCult media for three days at 37°C, then plated in 24-well dishes for two days. Compounds were then added to cells in PneumaCult and incubated at 37°C for 24 hours. Cells were harvested with ACF Enzymatic Dissociation Solution for 7 minutes according to manufacturer's protocol and collected with media.

Human CD8$^+$ cytotoxic T cells

Peripheral blood mononuclear cells (PBMC) from two healthy donors were isolated from leukopaks using MACS Cell Separation kits from Miltenyi Biotec and frozen at 1e8 cells per vial. Cells were thawed and cells were isolated using the CD8$^+$T Cell Isolation Kit from Miltenyi Biotec. T cells were grown in RPMI supplemented with FBS and IL-2; CD3/CD28 Dynabeads from LifeTechnologies were added to activate cells for 72 hours at 37°C. Beads were removed from culture by incubating on a magnet for 5 minutes. Cells were resuspended in media containing fresh IL-2 and plated in 96-well plates. Compounds were added at 2x concentration and incubated at 37°C for 24 hours. Cells were harvested by centrifuging for 5 minutes at 300xG.

Human CD34$^+$ hematopoietic stem cells

Mobilized human peripheral blood CD34$^+$ cells from healthy donors were obtained from StemCell Technologies. Cells were thawed over PBS supplemented with 1% human serum albumin and incubated in StemSpan media containing CC100 and rhTPO at 37°C for 48 hours. Cells were resuspended in fresh media containing CC100 and rhTPO and plated in 96-well plates. Compounds were added at 2x concentration and incubated at 37°C for 24 hours. Cells were harvested by centrifuging for 5 minutes at 300xG.

Human preadipocytes

Preadipocytes from healthy, lean donors were obtained from Zen-Bio and thawed into 24-well plates in plating media. After incubation at 37°C for 24 hours, the media was changed to differentiation media and cells were incubated for 72 hours. Compounds were added to cells at a 2x concentration and then incubated at 37°C for 24 hours. Cells were harvested with 0.05% Trypsin and collected with serum-containing media.

Single-cell library generation

Harvested cells were washed with PBS and labeled with TotalSeq-B hashtag antibodies from BioLegend according to manufacturer's protocol. Briefly, cells were incubated with 250ng of TotalSeq-B antibodies (hashtags 1-10) for 30 minutes at room temperature. Cells were washed with PBS a total of three times and then ten samples with different hashtags were pooled and counted on a Luna Cell Counter. Single-cell libraries were then prepared using the Chromium Single Cell 3' Feature Barcoding Kit targeting 10,000 cells per library, according to manufacturer's protocols (10x Genomics; CG000317 Rev B).

Single-cell data preprocessing

Hashed sequencing libraries were filtered to remove cells with too few or too many counts. Specifically, each cell was assigned a score of log(cell library size) - log(mean library size per cell), and cells with a score less than -0.5 or greater than 0.75 were removed. Cells with greater than 18% mitochondrial gene counts were also filtered out. Genes were removed if they were not expressed in at least 0.5% of cells for each plate of data. Finally, raw gene counts were normalized using scanpy.pp.normalize_total with the target parameter set to 1e6, and rescaled using scanpy.log1p. The hashed libraries were then demultiplexed using a multivariate Gaussian mixture model.

Hashed sequencing libraries were filtered to include libraries ranging from 5,000 to 80,000 counts, with less than 18% mitochondrial gene counts.

Implementation of algorithms for benchmarking

To compare DrugReflector to baseline algorithms (k-Nearest Neighbors, Logistic Regression) and published methods for prioritizing LINCS algorithms (SIGCOM, Dr. Insight), we calculated the top 1% recall for each algorithm across three datasets.

To establish fair comparisons, DrugReflector and the baseline algorithms were trained on the same LINCS training dataset; they took in the same 978 landmark transcripts and predicted the same 9,597 compound labels. Published methods were given all transcripts in the dataset, and predictions were filtered to mutual compound labels.

k-Nearest Neighbors (kNN) implementation

To construct our model, we first balanced features for each compound signature in our curated LINCS training dataset. We scaled v-scores to target a standard deviation of 1, and then clipped values outside of [-2,2].

To make predictions, we used sklearn’s pairwise cosine similarity to compare our two datasets: the reference LINCS training dataset as input X and a benchmark test dataset as input Y. The output will contain all pairwise similarities between X and Y. Next, we wanted to group similarities by compound, as each compound has many signatures in our training dataset. For each observation in Y, our test dataset, we grouped all similarities with X by compound. We then took the average for each group to get a mean similarity for each unique compound. To interpret the similarity results, we ranked each compound by mean similarity for each test observation. Values of cosine similarity ranged from [-1,1], where values increase as similarity increases. To see if we successfully matched a compound label to a test observation, we checked to see if the label was within the lowest 100 ranked, most similar, compounds.

Multinomial Logistic (Softmax) Regression Model

To construct our model, we first partitioned the curated LINCS training dataset into three folds. These folds matched DrugReflector’s training folds. We trained each of three sklearn multiclass logistic regression models on a unique combination of two of three folds. All models had the same hyperparameters: a regularization penalty of L2, an inverse regularization strength of 1, no class weights, and a limited memory BFGS solver.

To make predictions for a benchmark observation, each model computes probability estimates of compound classes. The resulting classes were ranked by probability, where lower rank indicates higher probability. Each model then contributed a vote to the ensemble rank; we took the mean rank across all three models. To finalize the ensemble rank, we ranked the mean rank. An observation is predicted successfully if the compound label is in the lowest 100 rank, highest probability, compounds.

SigCom LINCS

Public method SigCom LINCS is accessible by LoopBack API. It is hosted by the Ma’ayn Laboratory at https://maayanlab.cloud/sigcom-lincs/.

To run predictions, we first identified all relevant compound signatures in the SigCom Database. Relevant signatures have a clearly identifiable compound that is predictable by DrugReflector and our baseline algorithms. Compound identifiers are maintained by the LINCS consortium. They start with a “BRD-” followed by 9 alphanumeric characters. We parse these identifiers from the “cmap_id” signature metadata field.

Next, we prepared our benchmark data for signature search. For each observation, genes were sorted by v-score. The highest and lowest 250 gene values were passed into up and down entities of the API “ranktwosided” enrichment query. We requested the server maximum limit of 10,000 up and down chemical perturbation signatures. The server returned the score, z-sum, and rank of the top 10,000 mimicker and reverser signatures.

To convert our signatures scores into compound ranks, we selected the signature with the maximum z-sum of each relevant compound. We then ranked compounds from low to high with increasing z-sum, increasing similarity. A benchmark observation is predicted successfully if the compound label of the observation is within the top 91 ranks. Note we threshold at 91, because there are only 8,701 relevant compound classes in SigCom from our 9,597 classes in our baseline algorithms.

Dr. Insight

Since publication, Dr. Insight has been removed from the CRAN repository, as the package is no longer maintained. We chose to include it because it was used in a recently published article describing a strategy for drug repurposing based on transcriptomics data[10]. We obtained an archived version of the software from https://cran.r-project.org/src/contrib/Archive/DrInsight/DrInsight_0.1.2.tar.gz.

This archival version matches signatures to an early CMap dataset comprising 6,100 signatures of 1,309 compounds at varying concentrations on three cell lines: MCF7, PC3, and HL60. To run Dr. Insight, we used the following parameters. Repurposing unit = “drug”, connectivity = “positive”, and the CEG.threshold to 0.05. Because the Dr. Insight reference dataset only includes 1,309 compounds, we only reported results for the intersection in each dataset and the Dr. Insight reference.

Generating a single-cell time course of hematopoiesis

Generating CITE-seq data in human CD34+ cells

Mobilized (Neupogen) peripheral blood CD34+ cells (mPB CD34+) were purchased from AllCells (vendor website)**.** mPB34+ cryopreserved cells from four healthy donors were thawed and cultured in StemSpan SFEM supplemented with CC100 (Stem cell Technologies) and TPO (100ng/ml) at a density of 300K/ml. Cells were incubated at 37 ºC over a period of 12 days with media changes every 2- to 3 days. Cell collections were done across five time points over a ten-day period (Days 2, 3, 4, 7, and 10). On days of collection, cells were processed for CITE staining using Biolegend TotalSeq antibody cocktail protocol (TotalSeq-B Human Universal Cocktail, V1.0) with minor modifications. Labeled cells were processed for single cell RNA sequencing using the 10x Genomics Single Cell Gene Expression with Feature Barcoding technology. Libraries were prepared using the Chromium Single Cell 3’ Reagent Kit v3.1 (10x Genomics, 1000268), and sequenced on an Illumina NovaSeq 6000 platform, generating paired-end reads. Raw reads were demultiplexed using bcl2fastq (v2.20.0.422) and processed using Cell Ranger software (v5.0.1, 10x Genomics). Reads were aligned to the human reference genome (GRCh38) using STAR aligner (v2.7.0a).

Developing a plate-based flow cytometry assay to measure multiple lineages in CD34+ differentiation

Hematopoietic Stem Cell *in vitro* differentiation assay

Dual-Mobilized (Neupogen and Mozobil) peripheral blood CD34+ cells were purchased from AllCells. Cryopreserved cells were thawed and cultured in flasks in StemSpan SFEM with TPO (100ng/ml) and 1x CC100 supplement (Stem Cell Technologies). On day 0 (48 hours post-thaw), cells were plated into 96-well plates in the same medium. Plating conditions were optimized for each lineage as follows: megakaryocyte lineage differentiation 60K/well in round bottom plates and erythroid differentiation 30K cells/well in flat bottom plates. Compound treatment was performed on days 0, 2, and 5 of culture. Cells were passaged at a ratio 1:4 on day 2, media was refreshed on day 5. On day 7, immunophenotype of differentiated cells was evaluated using flow cytometry. Compound treatment and media changes were performed using Integra Viaflo384.

Compound treatment

Compounds were purchased from Frontier Scientific compound management company at 10mM in DMSO and arrayed onto microplates at 0.1, 1, or 10mM in triplicate using Hamilton Microlab Star liquid handler and stored at -80°C. On day of treatment, compound plates were thawed at 37°C for 10 minutes, diluted with IMDM (ThermoFisher), and then added to cells using Integra Viaflo384.

Flow cytometry

For megakaryocyte lineage experiments, on day 7 of differentiation, cultures were washed and incubated with antibodies (Supplemental table #) in Cell Staining Buffer (BioLegend, 420201). For erythroid lineage experiments, cells were fixed after antibody staining. Briefly, plates were washed with DPBS, and incubated in DPBS containing viability dye (1:1000), followed by wash with Cell Staining buffer and fixation with Cytofix Fixation buffer (BD Biosciences, 554655). All incubations were performed for 25 minutes at 4°C in the dark. Cells were then washed and resuspended in Cell Staining Buffer and analyzed on NovoCyte Quanteon flow cytometer (Agilent).

Channel compensations were performed using single stained UltraComp beads (ThermoFisher, 01-2222-41) or cells. All antibodies were purchased from BioLegend, eBiosciences, or Invitrogen. Titrations were performed to assess optimal antibody concentration. Flow cytometry data were analyzed using FlowJo (Tree Star). Viability was determined using either viability dye or FSC/SSC gate in FlowJo. The following antibody panels were used to define cell populations. Megakaryocytes: CD41a+ CD71- CD42b+, Early erythroid progenitors: CD41a- CD71+ CD36+ CD235a-, late erythroid progenitors: CD41a- CD71+ CD36+ CD235a+.

Phenotypic data analysis and hit calling

For each set of screening experiments targeting a lineage, this analysis was applied to identify which compounds were hits. For each plate in this set of experiments, the mean percent population of DMSO (N=8 wells) was calculated. This was the mean DMSO value. The percent population value of every well (N=96) was divided by the mean DMSO value. This was the normalized value. Across both experiments for each lineage (random compounds and predicted compounds), the normalized values of the DMSO wells were compiled and the mean and standard deviation were calculated. There were 128 DMSO wells in the erythroid lineage effort and 230 DMSO wells in the megakaryocyte lineage effort. The hit-calling cutoff was equal to the mean + 6 standard deviations. For a compound to be called a hit, the average of the normalized values across replicates needed to be greater than the cutoff.

For hit validation experiments, significance was determined via a heteroscedastic one-way t-test between normalized DMSO and test compound sample values.

Paired transcriptomic and phenotypic measurements of Mk-inducing compounds

Single cell sequencing with lipid-based time course of MK differentiation

HSPCs were differentiated according to Mk assay conditions described above in the presence of test compounds. On days 1, 2, 5, and 7 of differentiation, samples were multiplexed (hashed) with cell multiplexing oligos (CMOs, 10X Genomics) according to manufacturer's protocol. Briefly, cells were washed with Cell Staining Buffer (CSB, Biolegend), counted, and incubated with CMOs for 5 min at room temperature. After incubation, cells were washed with 4% HSA (Grifols) three times. Libraries containing 12 samples each tagged with individual CMOs were pooled by combining approximately 100k cells from each well. Libraries were washed once in 4% HSA and counted, then resuspended in CSB at 1.2x10^6 cells/mL. Each test compound was sequenced in duplicate, where duplicates were spread across libraries. Each library contained a positive and negative control as well as both hit and non-hit compounds.

Processing and integration of perturbational scRNA-seq dataset

Single-cell RNA sequencing data from one CD34 donor treated at 1uM with each respective compound or DMSO was collected on Days 1, 2, 5, and 7, in biological duplicates, with paired flow cytometry readouts on Day 7. Libraries were prepared using the Chromium Single Cell 3' Reagent Kit v3.1 (10x Genomics, 1000268), and sequenced on an Illumina NovaSeq 6000 platform, generating paired-end reads. Raw reads were demultiplexed using bcl2fastq (v2.20.0.422) and processed using Cell Ranger software (v5.0.1, 10x Genomics). Reads were aligned to the human reference genome (GRCh38) using STAR aligner (v2.7.0a).

Hashed sequencing libraries were filtered to include libraries ranging from 5,000 to 80,000 counts, with less than 20% mitochondrial gene counts. Pre-filtered hashed libraries were then demultiplexed using a Gaussian mixture model and then filtered to singlets. Additional filtering was performed to remove cells with <2,500 or >60,000 counts, and cells with <1,600 or >9,000 genes. Total counts per cell were normalized to 10,000 and natural log transformation was applied using functions from Scanpy (v1.9.3)[11]. To maintain a consistent embedding of hematopoiesis, we used SymphonyPy (v0.2.1)[12] for reference mapping and label transfer between our reference time course CITE-seq dataset and the perturbation time course dataset. In brief, harmony was used to create a batch corrected PC space. The query dataset was then projected into the reference PC space and integrated in the reference's harmony-corrected PC space. Last, label transfer was conducted using SymphonyPy's K-nearest neighbors (KNN) classifier, leveraging the shared latent space to transfer cell type annotations from the reference to the new dataset. The robustness of label transfer was validated by examining the weighted Mahalanobis distance of query cells to mapped reference clusters, the cosine similarity across highly variable genes between reference and query cell types, and expression of cell type markers in the query dataset labels achieved from label transfer.

Differential abundance testing

To test whether differences in cell type proportions in the perturbed samples relative to the control DMSO condition were due to random sampling, we used the python implementation of scProportionTest (v0.1.2)[13]. scPropotionTest uses a permutation testing framework, appropriate for high dimensional data where standard parametric assumptions may not be suitable. For each comparison, compound vs. DMSO, the proportion of each cell type was calculated. Combined cells for each group were then shuffled to randomize group labels while keeping group size constant and the proportions were recalculated. The process was repeated 1,000 times to generate a p-value for significance between the permuted groups.

Using transcriptional readout to refine the Megakaryopoiesis signature

Leveraging our single-cell time course experiment, we aimed to refine our understanding of the transcriptional changes necessary and sufficient to induce megakaryopoiesis, providing closed-loop feedback for the model. To this end, we first identified the transcriptional changes in our single-cell time course that were consistently associated with megakaryocyte induction. Using limma[14], we regressed the pseudobulked gene expression of perturbed HSPCs from day 1 of the scRNA-seq time course against change in megakaryocyte abundance as measured by flow cytometry at day 7. For each gene, the model fits the following equation:

$$\text{expr} = \beta_0 + \beta_1 FC_{mk} + \beta_2 I_{library}$$

Where β_0 is an intercept term, β_1 quantifies the relationship between gene expression and the fold change FC_{mk} of late megakaryocytes, and β_2 corrects for library effects. Internally, limma estimates means and variances for each coefficient while correcting for differences in coverage and the inherent sparsity of transcriptomic data.

For each gene, the model outputs an FDR-adjusted *p*-value indicating the significance of the correlation between that gene’s expression at day 1 and MK induction at day 7. We converted this *p*-value into a score by taking the negative base-10 logarithm and multiplying by the sign of the association (positive if the gene is correlated with fold change, and negative if it is anticorrelated). Genes with FDR-adjusted p-value <0.01 were considered significantly associated with phenotype.

We divided the genes into three classes:

- *Concordant* genes were significantly associated with phenotype in the same direction as indicated by the input *v*-score.
- *Discordant* genes were significantly associated with the phenotype in the *opposite* direction as indicated by the input v-score.

- The remaining genes had no significant association with phenotype.

We hypothesized that the concordant genes drive model performance, whereas the discordant genes reduce it. To test this, we modified the input by setting all but the concordant genes to zero, or all but the discordant genes to zero, and measuring the impact on hit prioritization. We found that masking the input to only concordant genes improved the prioritization of megakaryocyte inducers as measured by flow cytometry, and performed better than masking to random sets of genes of the same size (**Figure 5e**).

Because the compounds used to classify genes as concordant or discordant were the same as those used to test model performance, this strategy may bias the model in favor of these compounds. We therefore performed stratified 5-fold cross-validation to see whether signature refinement can improve recall of *unseen* hits. The procedure was as follows:

1. We divided the profiled compounds randomly into five folds, dividing hits as evenly as possible. We designated one-fold as the test set and the remaining four as the training set.
2. Using the training set only, we identified concordant and discordant genes as described above.
3. We masked the input signature to concordant or discordant genes, ran the masked signature through DrugReflector, and recorded the rank of the *test* compounds.
4. We repeated steps 2-3 four more times, designating each fold as the test set in turn.
5. We repeated steps 1-4 ten times with different random seeds, and reported for each compound its mean cross-validation rank over all seeds.

Concordance of known megakaryopoiesis markers with measured MK induction

To better understand the action of hit compounds, we examined their effect on transcription factors involved in megakaryopoiesis, and on marker genes of MKs. We obtained a list of nine transcription factors and four MK marker genes from previous literature[15,16] and examined their differential expression patterns in day 1 HSPCs from the transcriptional validation screen. We also used limma to model their association with MK abundance as described above.

All nine of the transcription factors were significantly associated with megakaryocyte abundance ($p<0.05$) and showed significant differential expression in MK inducers, suggesting that our hit compounds bias HSPCs towards the megakaryocyte lineage at an early time point (Figure 5C). Only one of the MK markers was significantly associated with induction, which is unsurprising given that the sample consisted of HSPCs. In addition, all but two of these genes have positive score in the model input signature, showing that the signature captures at least some of the known biology. The two genes with negative score would be filtered out by signature refinement due to the disagreement between input v-score and observed association with MKs, as described in the previous section.

Relating model performance to CD34-relevance

To quantify the similarity between a compound's effect in CD34 and in LINCS, we calculated for each compound the distance between 24-hour signatures in our experiment and the 10 most similar signatures for the same compound in LINCS. We measured similarity using cosine distance over the 940 landmark genes that were shared between the two assays. Perturbational response in CD34s was represented by a vector of differential expression scores for each shared gene, defined as:

$$\text{DES } = -\log(\text{FDR pvalue}) * \text{sign}(\text{FC}).$$

Response in LINCS was represented by the level 4 z-score vector.

We observed that predicted hits (with MK fold change > 2 at day 7) tend to have smaller differences between CD34 and LINCS response than predicted non-hits. To assess the significance of this result, we performed a two-sample independent t-test comparing the mean 10-NN LINCS similarities in hits to those in non-hits, yielding a p-value of 0.02.

Calculating cell-type specific pseudobulked differential expression

To reduce the noise in scRNA-seq we aggregated cells by summing counts across cells within a technical replicate of each cell type and day to form pseudobulks. Prior to differential expression test, we removed genes that were expressed by less than 0.5% of cells. We then analyzed the aggregated read counts using Limma, with perturbation condition as the main variable, technical replicate as covariate, and the DMSO condition as reference. We performed the differential expression test using Limma on each cell type and day independently to calculate the gene expression logged fold changes (logFC).

Then, for each cell type and day, we performed principal component analysis on the logFC results of the perturbation conditions.

Identifying GO terms associated with PC1 and PC2

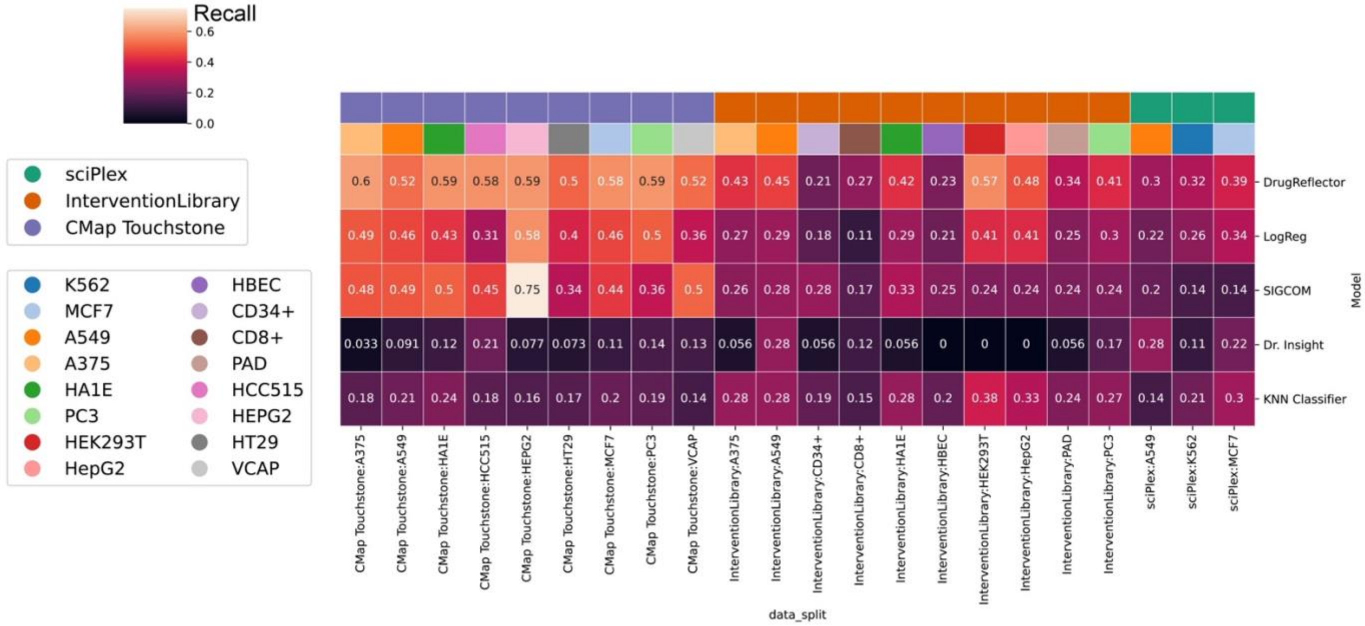
Using GSEAPy prerank[17], we identified GO terms most strongly associated with the PC1 loadings in the HSPC population at Day 1 using the GO Biological Process 2023 gene set. We sorted the term on their net enrichment score after filtering the results to terms where the Tag% was more than 0.4 and ranked. We repeated this process for each of PC1 and PC2.

Calculating rolling-window expression of GO terms over infered pseudotime in the Mk lineage

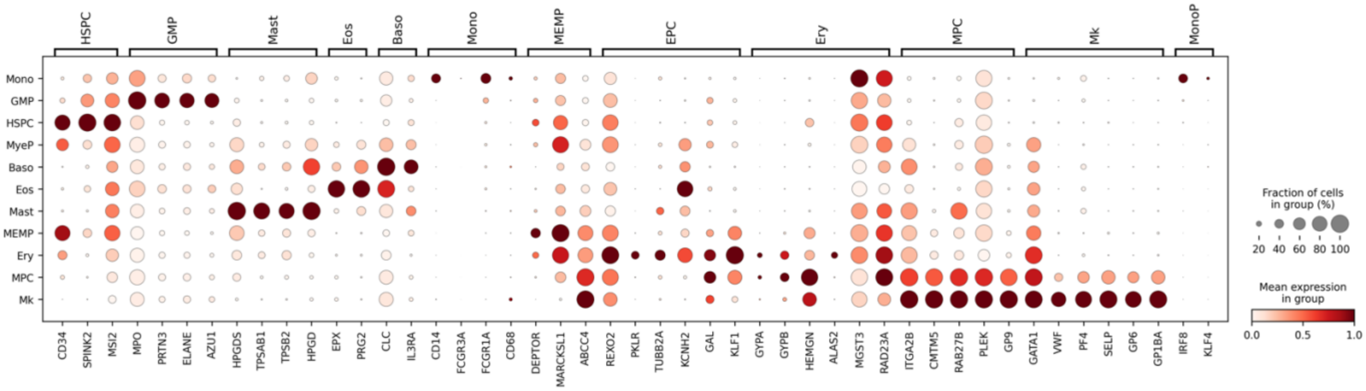
We inferred a unified developmental pseudotime for cells of the Mk lineage (HSPC, MEMP, MkP, and MPC) in all perturbation and day conditions, using the `dpt_pseudotime` function implemented in scanpy[11] version 3.9.

Using the Scanpy score gene function, we calculated a gene activity score for the GO terms of interest in each cell. To find general trends in the change of GO term scores across pseudotime, we performed smoothing for each perturbation by ordering the cells in pseudotime and averaging the scores in rolling windows of 600 cells. We also transformed the pseudotime values using the same rolling windows.

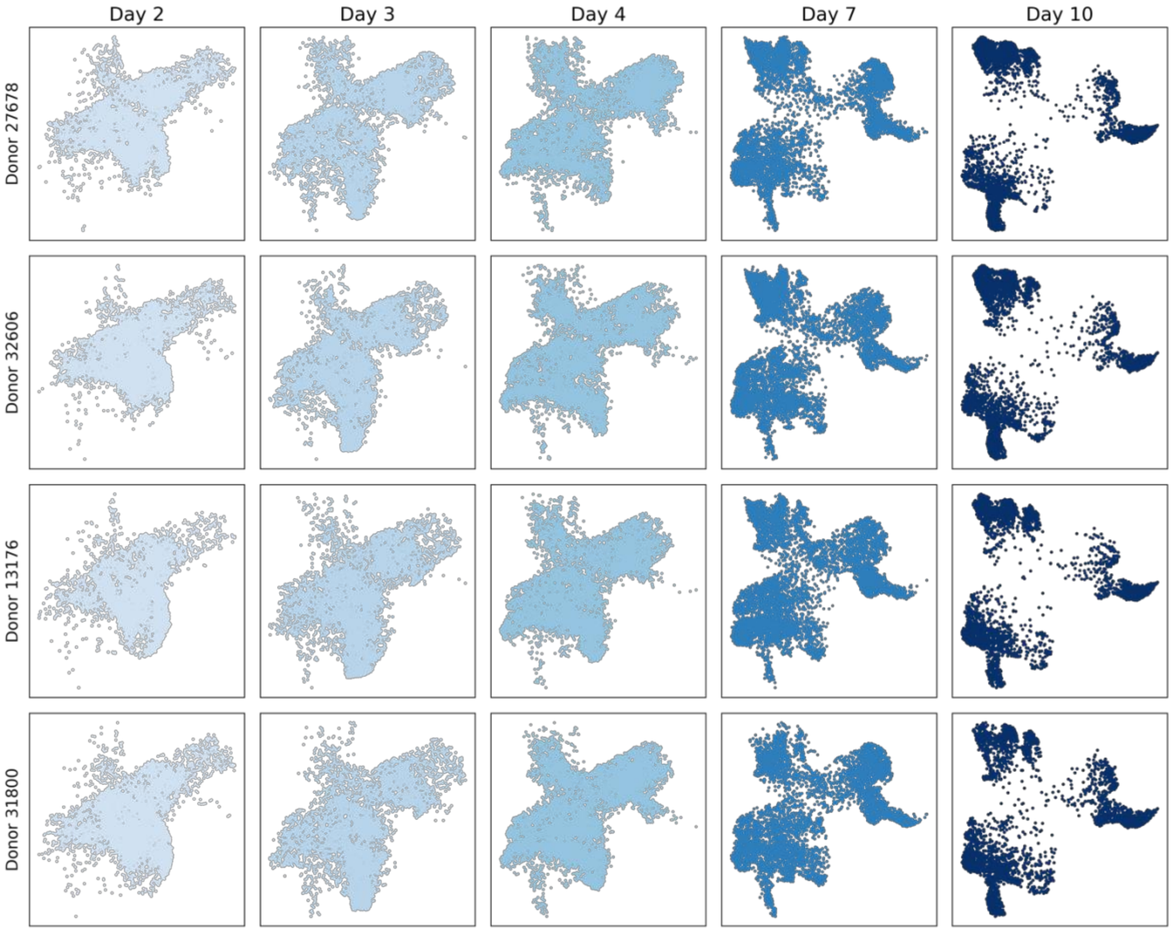
Supplementary figures



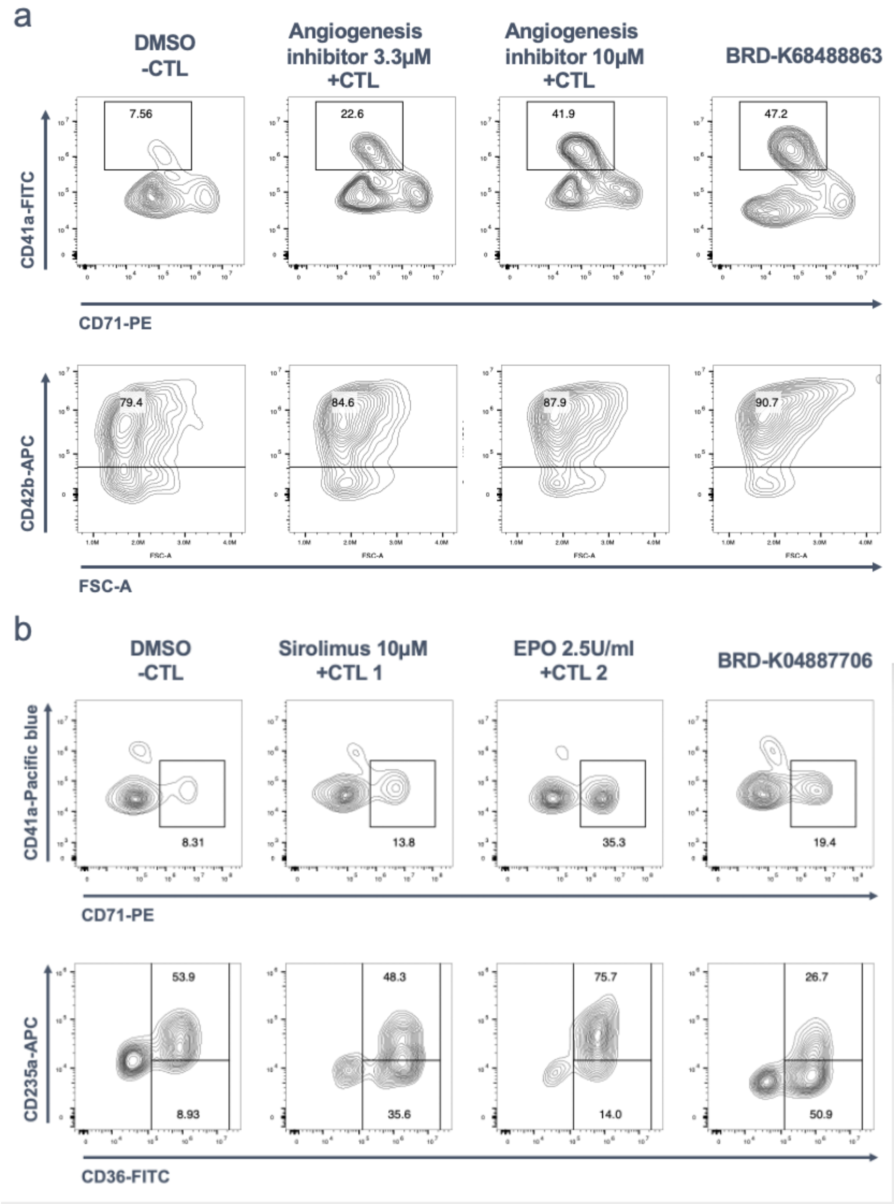
Supplementary Figure 1 – Results of full benchmarking. – Results of the benchmark split out by dataset and cell type. Continuous values in each cell denote the proportion of compounds recalled at or below the 1% of all compounds by each model in each cell line/dataset.



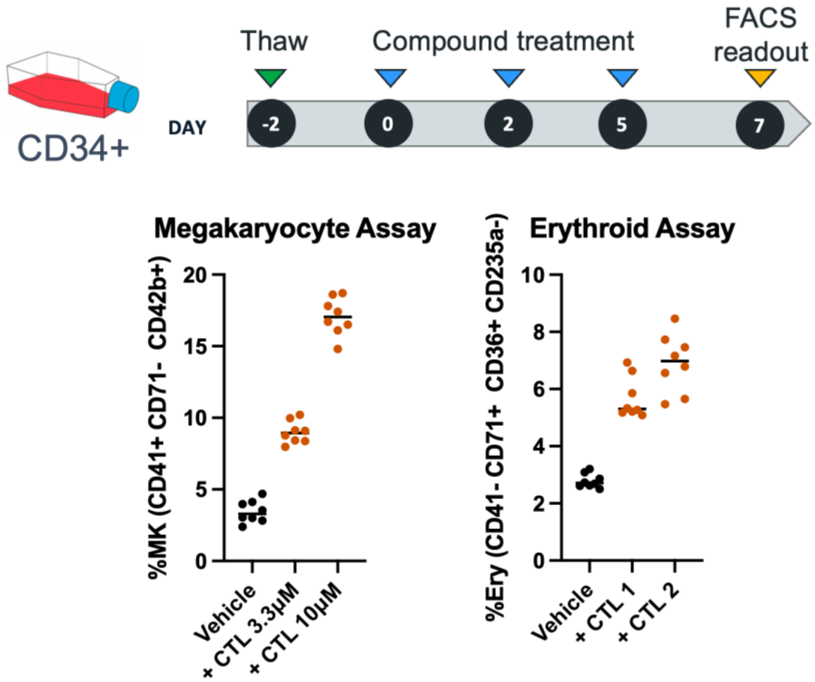
Supplementary Figure 2 – Marker gene expression in HSPC atlas cell types. A dotplot showing the normalized expression of marker genes in each annotated cell state (y-axis). Genes are organized based on prior knowledge associated each gene with distinct cell states (x-axis).



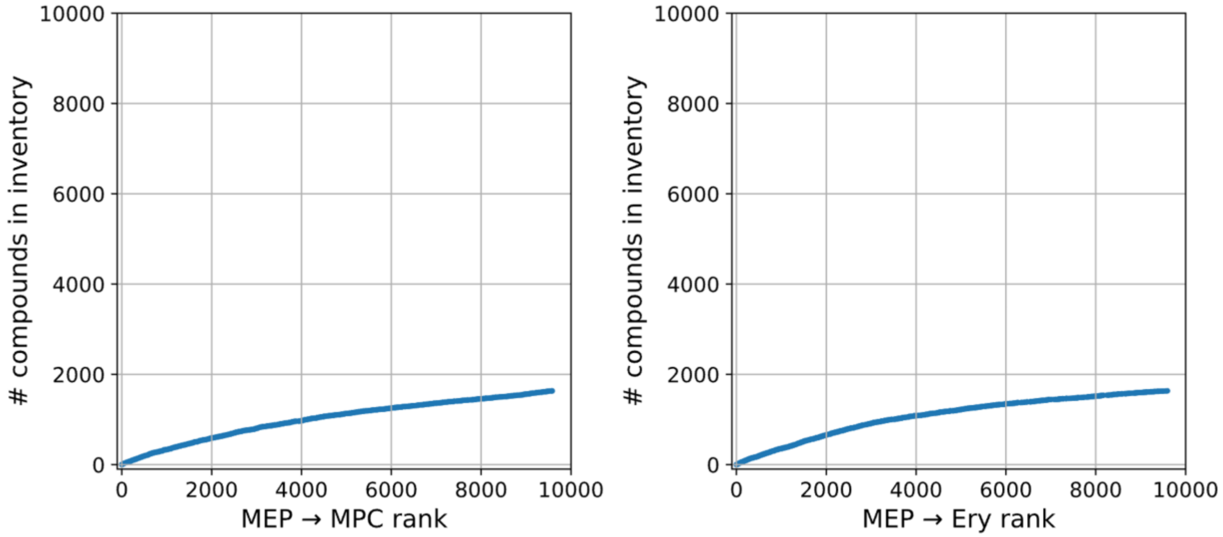
Supplementary Figure 3 – Consistent differentiation across time and donors. Comparison of cell type density across time points and donors using UMAP. No batch correction was performed on this dataset, aside from the selection of highly variable genes that are consistent across at least 2 donors. The UMAP embedding was calculated once, and then subsets were plotted on each subplot.



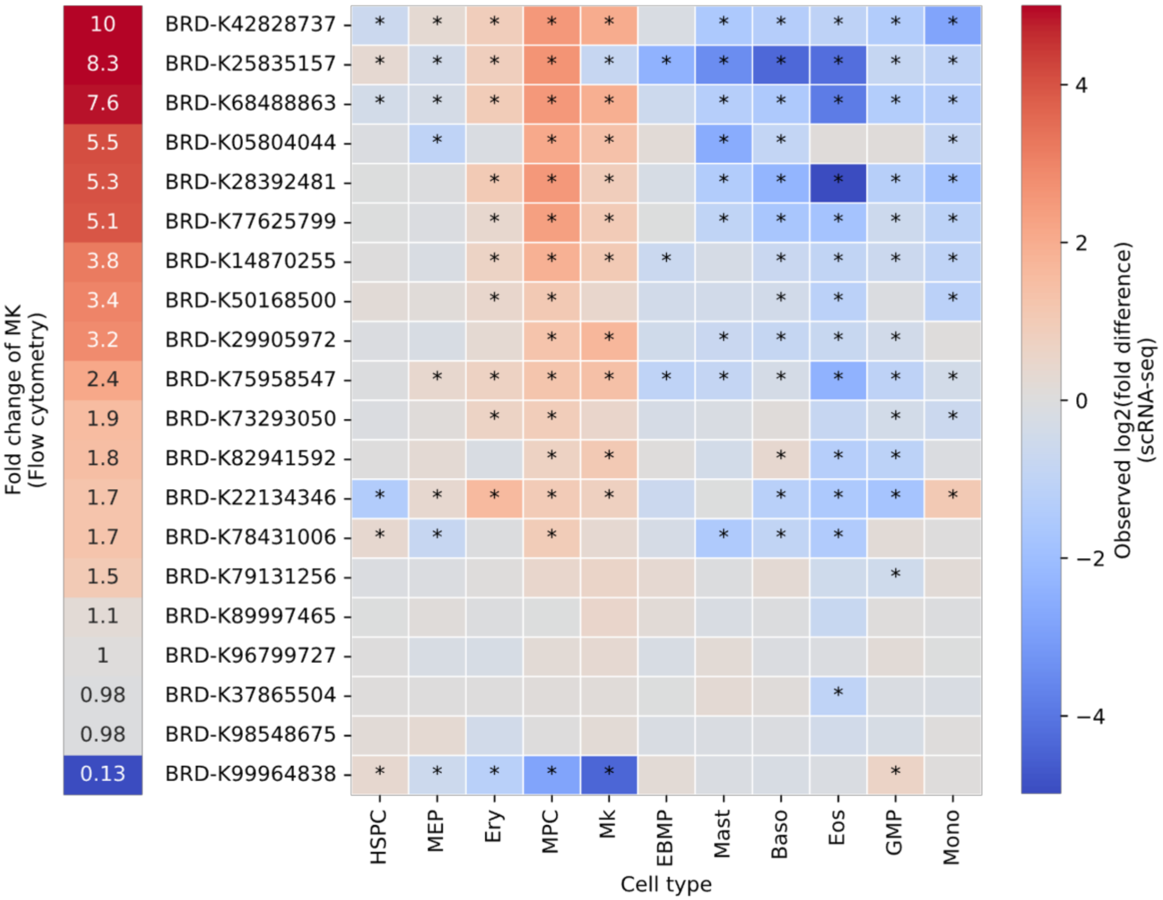
Supplementary Figure 4 – Gating strategy for flow cytometry analysis. HSPCs were differentiated in the presence of compounds as described in Methods and population abundance was quantified by flow cytometry on day 7. (**a**) For megakaryocyte differentiation, Angiogenesis Inhibitor was used as a positive control for CD41a+ CD71- CD42b+ late MK induction. BRD-K68488863 represents a typic hit compound. (**b**) For erythroid differentiation, Sirolimus 10µM and EPO 2.5U/ml were used as positive controls for induction of CD41a-low, CD71+, CD36+, CD235a- early erythroid population. BRD-K04887706 represents a typical hit compound.



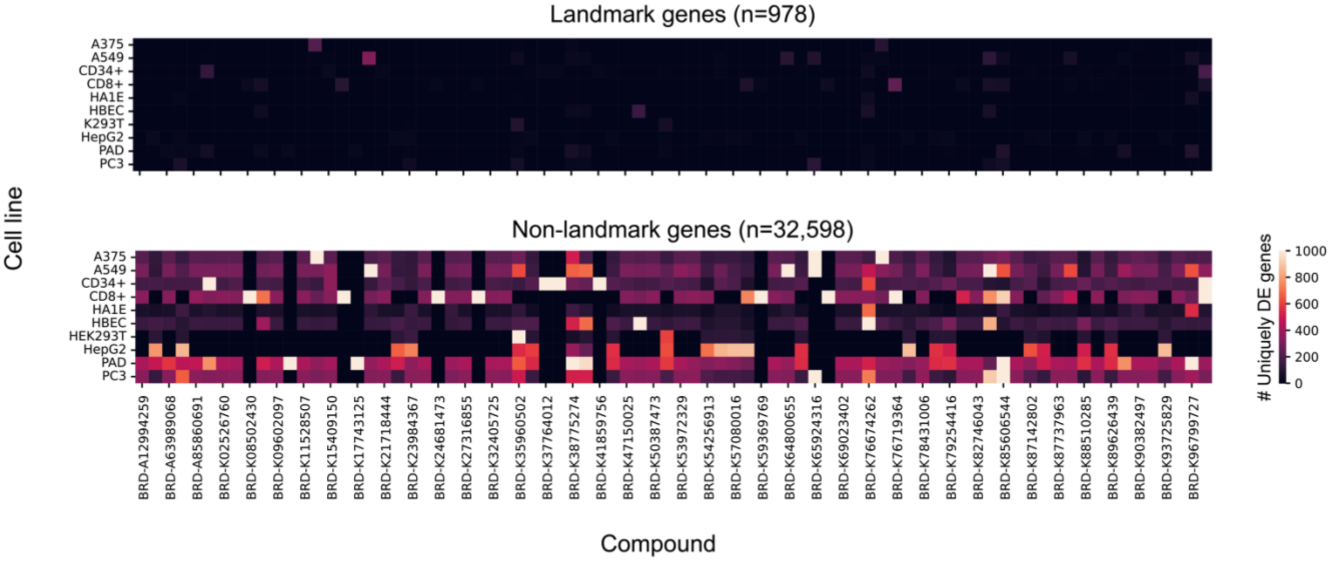
Supplementary Figure 5 – Phenotypic assay schematic and controls. To validate the reproducibility of our assay, we measured the abundance of each lineage in negative and positive control conditions. Top, we show a cartoon schematic of our phenotypic assay, in which cells are dosed with compounds on days 0, 2, and 5. Flow cytometry readout is measured at Day 7 post-treatment. Below, we show the abundance of each target population in replicate samples of DMSO Vehicle negative control and under positive controls of a representative plate from our screen. For the Mk assay, angiogenesis inhibitor (BRD-K08502430) is the positive control. For the Ery assay, CTL 1 is sirolimus (BRD-K89626439) at 10μM and CTL 2 is erythropoietin (EPO) at 2.5 U/mL.



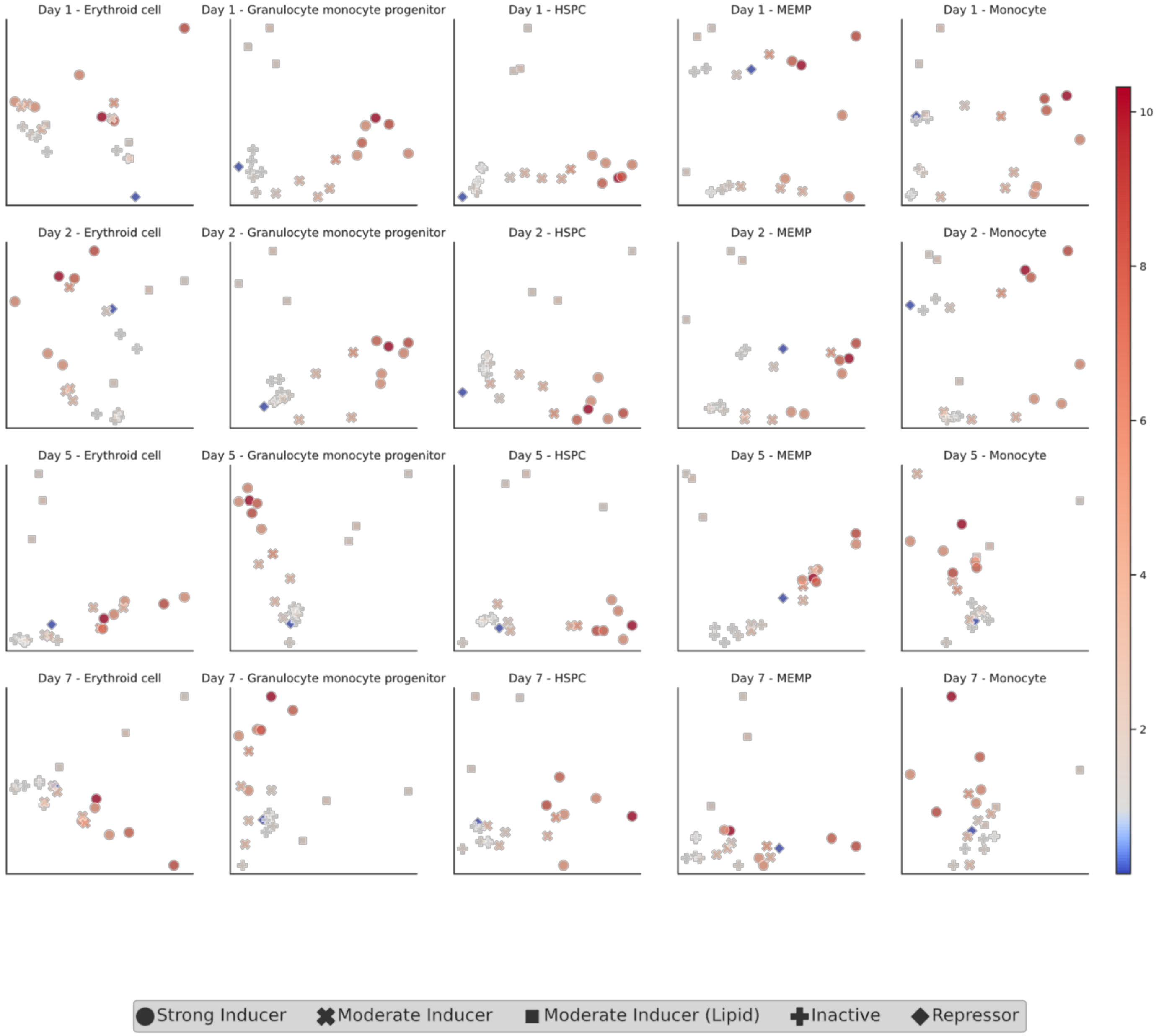
Supplementary Figure 6 – Sampling of DrugReflector ranks covered by available compounds. We observed a representative sampling of compounds across ranks for both sets of virtual screens. The x-axis shows the rank output of the DrugReflector model. The y-axis shows the cumulative number of compounds at that rank or lower out of 1,635 compounds available in our inventory at the time of study initiation.



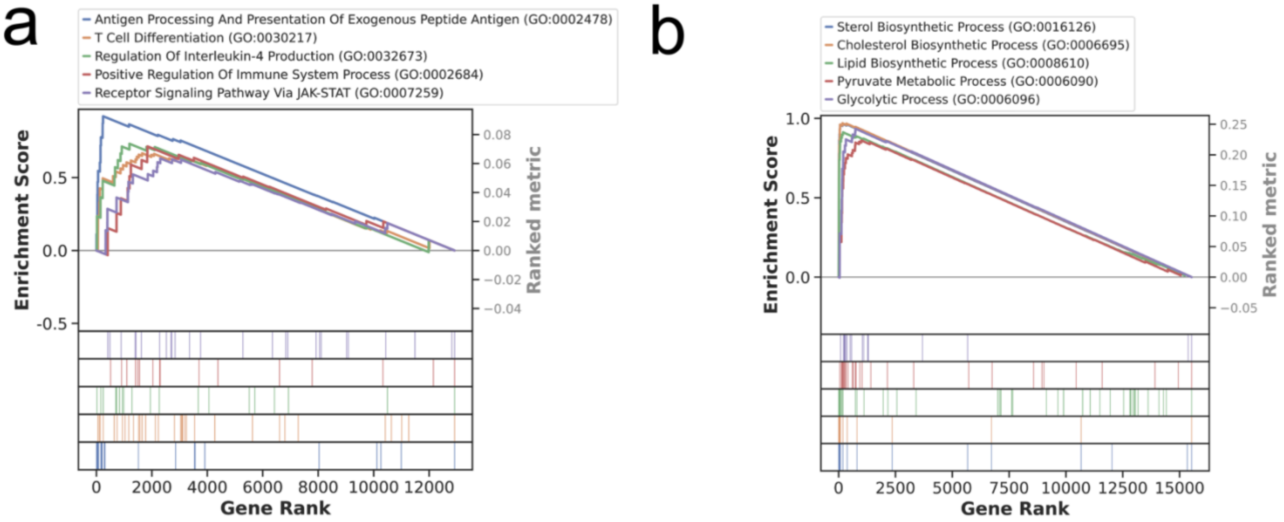
Supplementary Figure 7 – Abundance of cell types in scRNA and flow cytometry following 7 days of chemical perturbation. Differential abundance in scRNA-defined populations aligns with phenotypic assay and confirms lineage-specific induction of the Mk population. Left, the fold-change in Mk was measured via flow cytometry for each compound. Right, the fold-change in the abundance of the various annotated cells in the scRNA data relative to DMSO. Asterisks denote significance from a permutation test with FDR correction using the Benjamini-Hochberg procedure (adj p value < 0.05).



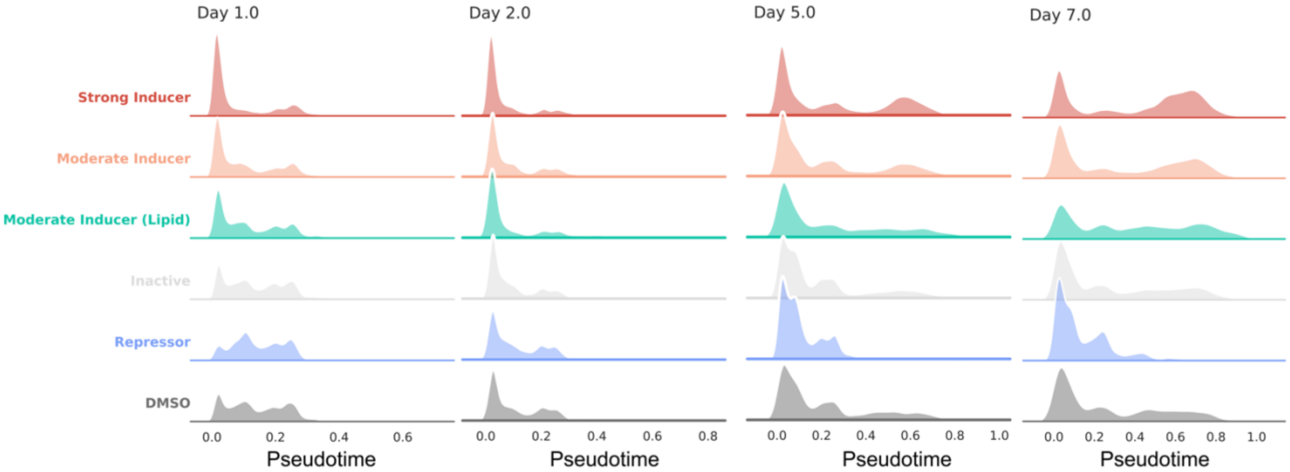
Supplementary Figure 8 – Most cell-type specific DE genes are not in the landmark gene set. These heatmaps show the number of genes that are uniquely differentially expressed in each cell line for each compound perturbation in the Intervention Library dataset. Values above 1,000 are clipped to 1,000. The top heatmap shows the number of uniquely DE genes within the landmark gene list (n=978) and the bottom shows the number of uniquely DE genes for all non-landmark genes (n=32,598).



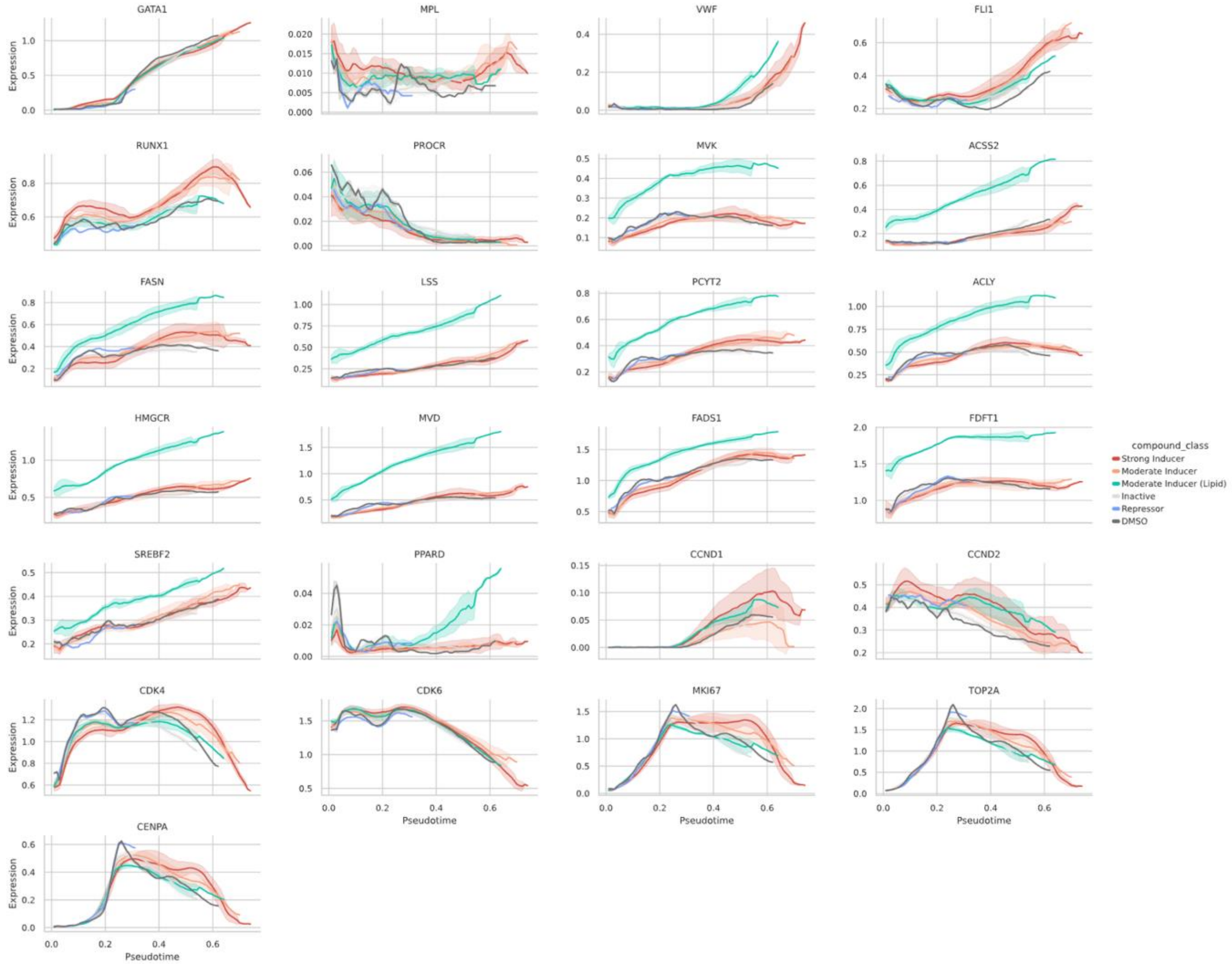
Supplementary Figure 9 – Variation across chemical perturbations per cell type and time point. Cell type-specific variation in differential expression across cell types and time points. Pseudobulked gene expression was used as input to LIMMA to calculate differential expression per cell type and DMSO at each time point. Markers denote post-hoc annotated compound classes.



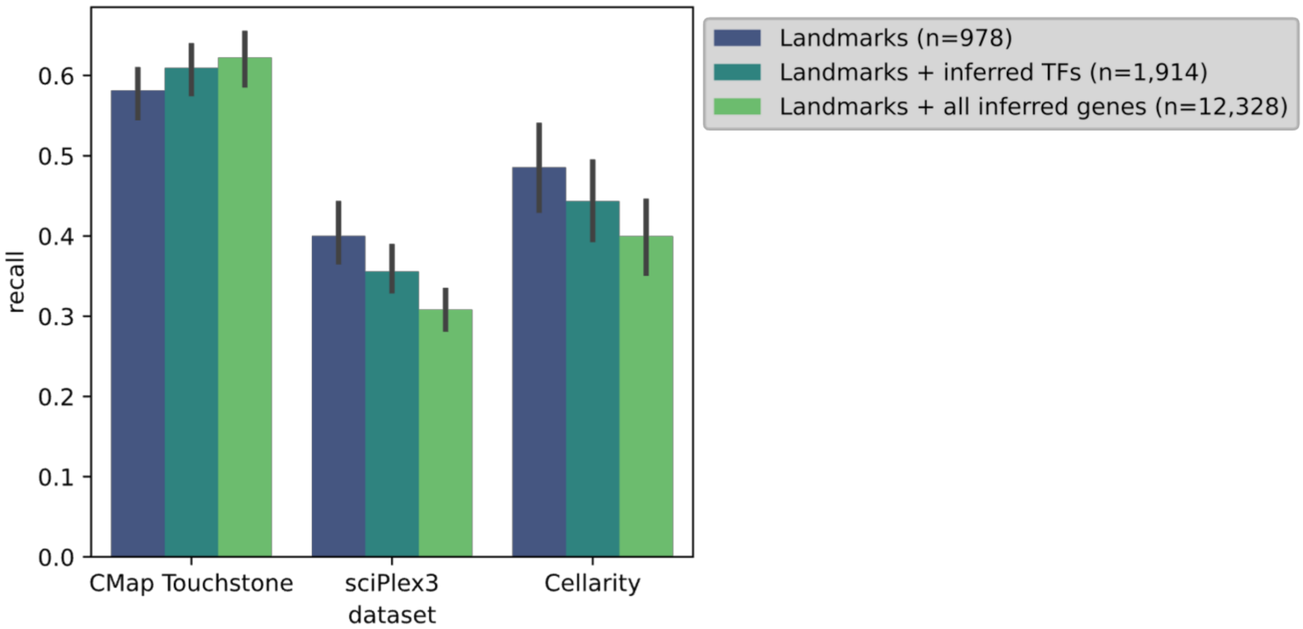
Supplementary Figure 10 – GO Term enrichment along PC1 and PC2 of Day 1 HSPC DE signatures. (a) Gene sets strongly associated with PC1 loading are enriched for antigen presentation and JAK/STAT signaling pathways associated with Mk induction, further supporting our conclusion that hit compounds induce *bona fide* megakaryopoiesis. (b) PC2 genes are enriched for lipid and cholesterol biosynthesis, leading us to label the three compounds separated by PC2 as lipid-inducing compounds.



Supplementary Figure 11 Density of cells across pseudotime for each compound class and day. The x-axis is inferred pseudotime along the Mk lineage, and the y-axis is the density of cells at each point along pseudotime averaged across all samples per compound class.



Supplementary Figure 12 – Expression of genes in rolling windows normalized by pseudotime. Expression of genes associated with GO terms from main figure 6 ordered by pseudotime and aggregated across cells per compound class. Y-axis is the expression of each gene. Shaded area represents the standard deviation across compounds within each compound class.



Supplementary Figure 13 – **Adding inferred genes in model training improves cross-validation performance but worsens generalization to new datasets.** DrugReflector was trained with the same hyperparameters as described for the final model changing only the number of input nodes to adjust for different feature sets of CMap. Recall on CMap, sciPlex3, and the Cellarity benchmarking dataset is shown. Error bars denote standard deviation across cell lines.

References

1. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. Preprint at https://doi.org/10.48550/arXiv.1912.01703 (2019).
2. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).
3. Daylight Theory: SMARTS - A Language for Describing Molecular Patterns. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.
4. Schuffenhauer, A. *et al.* Evolution of Novartis' Small Molecule Screening Deck Design. *J. Med. Chem.* **63**, 14425–14447 (2020).
5. Brenk, R. *et al.* Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases. *ChemMedChem* **3**, 435–444 (2008).
6. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollar, P. Focal Loss for Dense Object Detection. in 2980–2988 (2017).
7. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Preprint at https://doi.org/10.48550/arXiv.1502.03167 (2015).
8. Loshchilov, I. & Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. Preprint at https://doi.org/10.48550/arXiv.1608.03983 (2017).
9. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. Preprint at https://doi.org/10.48550/arXiv.1907.10902 (2019).
10. He, B. *et al.* ASGARD is A Single-cell Guided Pipeline to Aid Repurposing of Drugs. *Nat. Commun.* **14**, 993 (2023).

11. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY : large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

12. Petrova, K. Symphonypy. (2023).

13. Miller, S. A. *et al.* LSD1 and Aberrant DNA Methylation Mediate Persistence of Enteroendocrine Progenitors That Support *BRAF* -Mutant Colorectal Cancer. *Cancer Res.* **81**, 3791–3805 (2021).

14. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

15. Guo, T. *et al.* Megakaryopoiesis and platelet production: insight into hematopoietic stem cell proliferation and differentiation. *Stem Cell Investig.* **2**, 3 (2015).

16. Chiu, S. K. *et al.* Shared roles for Scl and Lyl1 in murine platelet production and function. *Blood* **134**, 826–835 (2019).

17. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, btac757 (2023).