Directing cellular transitions on gene graph-enhanced cell state manifold

Tianze Wang[1,2,+], Yan Pan[1,3,+], Fusong Ju[1,+], Shuxin Zheng[1,+], Chang Liu[1,+], Yaosen Min[1], Xinwei Liu[1,4], Huanhuan Xia[1], Guoqing Liu[1], Haiguang Liu[1], Pan Deng[1,*]

[1]Microsoft Research AI for Science, Beijing, China. [2]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA. [3]Department of Automation, Tsinghua University, Beijing, China. [4]Department of Physics, Beijing Normal University, Beijing, China. [+]These authors contributed equally.
[*]Correspondence: pan.deng@microsoft.com

Abstract

A select few genes act as pivotal drivers in the process of cell state transitions. However, finding key genes involved in different transitions is challenging. To address this problem, we present CellNavi, a deep learning-based framework designed to predict genes that drive cell state transitions. CellNavi builds a driver gene predictor upon a cell state manifold, which captures the intrinsic features of cells by learning from large-scale, high-dimensional transcriptomics data and integrating gene graphs with causal connections. Our analysis shows that CellNavi can accurately predict driver genes for transitions induced by genetic modifications and chemical treatments across diverse cell types, conditions, and studies. It is proficient in tasks involving critical transitions such as cellular differentiation, disease progression, and drug response by leveraging a biologically meaningful cell state manifold. CellNavi represents a substantial advancement in the methodology for predicting driver genes and manipulating cell states, opening up new research opportunities in disease biology and therapeutic innovation.

Introduction

Understanding the genetic drivers of cellular transitions is crucial for elucidating complex biological processes and disease mechanisms[1–3]. However, identifying these drivers has remained intrinsically challenging due to the sheer number of genes with intricate dependencies involved in transitions, contrasted with our limited experimental capacity and accumulated knowledge. Therefore, in silico methods capable of predicting driver genes across diverse contexts are highly desirable.

Traditionally, efforts to pinpoint critical driver genes have primarily relied on network-based methodologies, with a particular focus on gene regulatory networks (GRNs) [4–8]. While GRN-centric approaches have made notable progress, they also encounter limitations that hinder their broader use. For example, deducing accurate GRNs within heterogeneous cell populations, which is more relevant to translational research, remains a challenge [9,10].

Moreover, these GRN models tend to prioritize transcription factors and may overlook non-transcriptional drivers of cellular transitions. This limitation is critical for understanding complex cellular processes such as disease progression, immune modulation, and pharmacological responses.

To this end, we developed CellNavi, a deep learning framework designed to predict driver genes and navigate cellular transitions. CellNavi constructs a driver gene predictor on top of a learned manifold that parameterizes valid cell states. This manifold is modeled by mapping raw cell state representations onto a lower-dimensional coordinate space, where the dimensions correspond to intrinsic features of cell states, and the distance reflects the biological similarity between cells. To build this manifold, CellNavi is trained on large-scale, high-dimensional single-cell transcriptomic data, along with prior causal gene graphs that reveal the underlying structure of cell states. By projecting cellular data onto this biologically meaningful space with reduced dimensionality and enhanced biological relevance, CellNavi provides a universal framework that generalizes across diverse cellular contexts, allowing robust driver gene predictions even in previously unexplored cell types or conditions.

Our results show that CellNavi excels at predicting driver genes across a wide range of biological transitions, demonstrating strong performance in quantitative tasks curated in both immortalized cell lines and primary cells. It identifies crucial regulators in T cell differentiation and uncovers key genes associated with neurodegenerative diseases. Notably, CellNavi can infer mechanisms of action (MoA) for drug compounds without the need for drug-specific training, underscoring its potential in drug discovery. In summary, CellNavi offers a powerful framework for deciphering cell state transitions and their underlying mechanisms, holding significant promise for advancing cell biology and disease research.

Overview of CellNavi

CellNavi is designed to predict driver genes for given cellular transitions, where the transcriptomic data of the source and target cells represent the initial and final states of these transitions **(Fig. 1a-c)**.

CellNavi comprises two main components: that Cell Manifold Model (CMM), which captures and represents cell states, and the Driver Gene Predictor (DGP), which identifies the key genes driving these transitions based on learned cell representations **(Fig. 1b)**.

The CMM is built to capture valid cell states across diverse biological contexts. While transcriptomes are often used to represent cell states, valid cell states do not span the entire high-dimensional transcriptomic space but instead form a lower-dimensional manifold **(Fig. 1c)**. To model this, the CMM maps transcriptomic vectors to a lower-dimensional coordinate space that represents the intrinsic features of cell states, while preserving the relative similarities between cells.

We first curated a dataset of approximately 20 million single-cell transcriptomic profiles sourced from the Human Cell Atlas[11] **(Fig. 1d)** and adapted a Transformer architecture based on attention mechanisms, known for its ability to discern complex patterns in large-scale data[12–17], to train the CMM **(Fig. 1e)**. The training process involved a self-supervised down sampling-reconstruction task. Specifically, we down sampled the read-depth of gene expression profiles by a factor ranging from one to twenty and then asked the model to reconstruct the true expression values of genes after normalization. To focus learning on cell representations rather than gene-specific representations, we developed a decoder module to reconstruct gene expression profiles from the cell coordinates—representations of cells within the coordinate space of the cell state manifold—generated by the CMM **(Fig. 1e, Fig. S1.1, Methods)**.

However, relying solely on transcriptomic data may overlook the intricate gene-gene interactions that are crucial for describing and distinguishing cell states. To address this, we incorporated 20 million cell-specific gene graphs into the CMM training process **(Fig. 1d-e)**. These graphs contain causal connections derived from a prior network that spans over 30,000 human genes and their associated signaling pathways[18] **(Methods)**, thus providing richer information about the complex dependencies among genes, which extends beyond simple transcriptomic data, hence better implying intrinsic variables spanning the valid cell space. To fully leverage these gene graphs, we enhanced the CMM by introducing a GeneGraph Attention Layer to replace the standard Transformer encoder layer **(Fig. 1d)**. These layers, inspired by attention variants tailored for graph data[19], are specifically designed to process gene networks, enabling the model to integrate critical gene-gene relationships. With these designs, the model is driven to cultivate a manifold that systematically represents cell states and effectively reflects the relationships between cells, forming an informative foundation for downstream driver gene prediction.

Building upon this manifold, the DGP was developed to predict candidate driver genes for specified cellular transitions **(Fig. S1.2, Methods)**. The DGP is trained using data from CRISPR screening experiments, which link genetic perturbations to consequent changes in cell states [20–25]. In modeling these experiments, we designated unperturbed controls and

CRISPR-perturbed cells as source and target pairs, respectively, and utilized validated perturbed genes as labels for joint training of the CMM and DGP **(Fig. 1f)**. Specifically, for each cell pair, their transcriptomic data is transformed into cell coordinates by the CMM, which are then processed by the DGP to generate a likelihood score vector indicating the probability that various candidate genes are orchestrating the transitions **(Fig. S1.2)**.

While CRISPR screenings are typically conducted on cultured cells or homogenous populations and focus on immediate genetic perturbations, CellNavi extends its applicability to more complex transitions in heterogeneous tissues and primary cells **(Fig. 1g)**. By leveraging the biologically meaningful manifold, CellNavi enables knowledge gained from CRISPR screenings generalize to previously unexplored scenarios, including transitions that are challenging to study with traditional CRISPR methodologies.

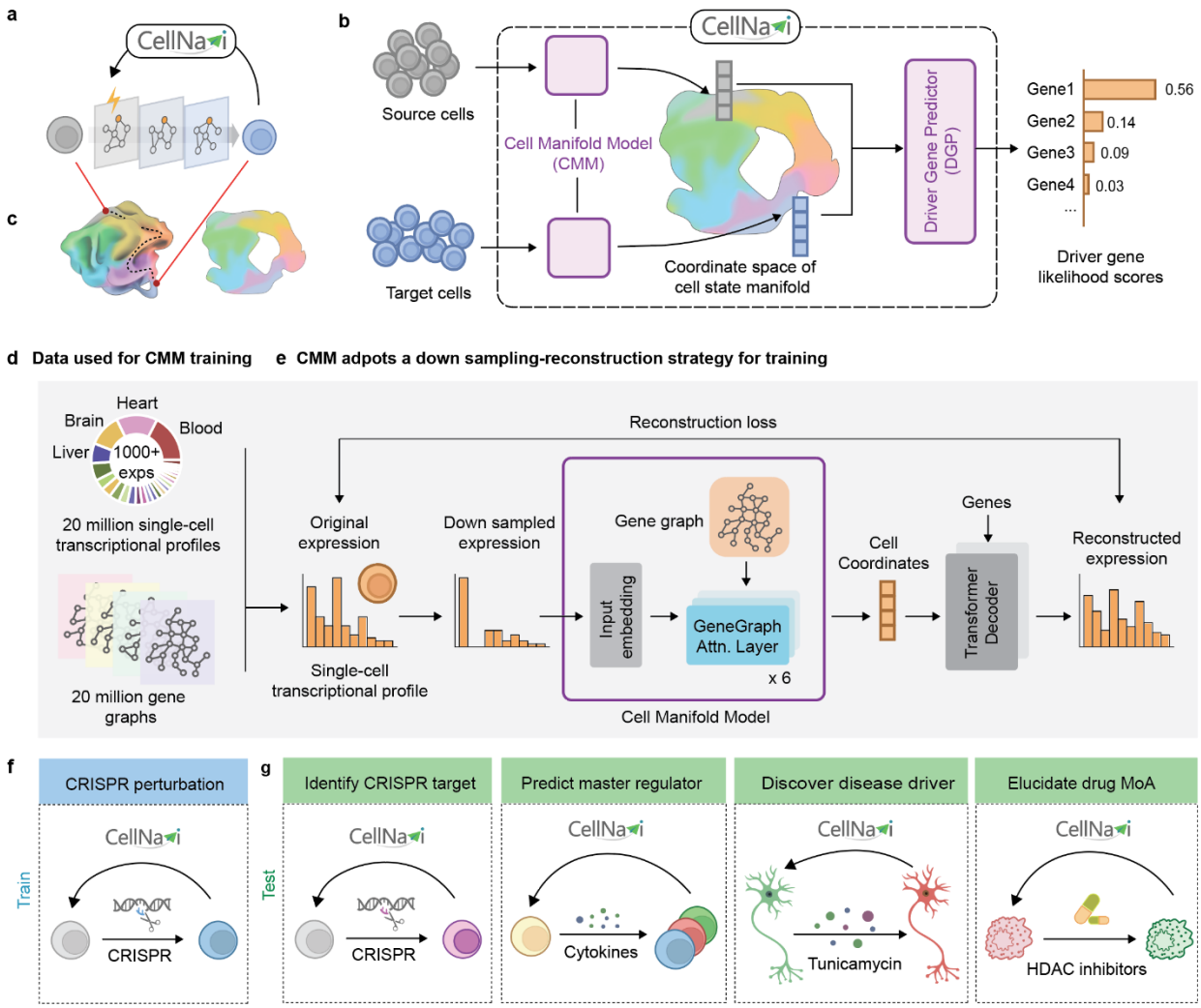


Figure 1. Overview of CellNavi. a) Conceptual illustration of CellNavi's task. Given a pair of source and target cells undergoing a transition induced by stimuli, CellNavi predicts the driver gene responsible for this transition. **b)** Workflow of CellNavi. Cell Manifold Model (CMM) maps the source and target cells onto a coordinate space of the cell manifold. The Driver Gene Predictor (DGP) then uses the cell coordinates produced by CMM to rank the candidate genes by likelihood scores. **c)** Illustration of the cell manifold and its coordinate space. **d)** Data used for CMM training. **e)** Training of the CMM. The CMM consists of six GeneGraph Attention (Attn.) layers designed to incorporate graph-based information. During training, single-cell transcriptomic profiles are randomly sampled from the curated HCA dataset and used as input. Cell embeddings generated by the model are then used by a Transformer decoder to reconstruct gene expression profiles. See Figure S1 and Methods for further details. **f)** Data used for DGP training. **g)** Application scenarios and test cases of CellNavi.

Quantitative evaluation of CellNavi

To assess the capabilities of CellNavi, we first evaluated its performance on CRISPR perturbation datasets, where driver gene information is well-established for transitions from source (unperturbed) to target (perturbed) cells.

We initially applied CellNavi to the Schmidt dataset, a CRISPR activation screen derived from human T cells that profiles 69 core regulators involved in type I cytokine production[26]. This dataset captures distinct expression profiles and molecular phenotypes across both resting and re-stimulated T cells, within and between different cell types **(Fig. S2.1)**. We trained the DGP on unperturbed and perturbed re-stimulated T cells and tested CellNavi on resting T cells **(Fig. 2a)**. This approach allowed us to evaluate CellNavi's ability to generalize across heterogeneous primary cells and predict driver genes in new cell states.

For each source-target cell pair, CellNavi prioritizes candidate genes based on their predicted likelihood scores. CellNavi achieved a strong performance across 23,047 source-target cell pairs, with an area under the receiver operating characteristic curve (AUROC) score of 0.874 **(Fig. 2b)**, an F1 score of 0.615, and a top-1 accuracy rate of 0.621 **(Fig. 2c)**. When the top-5 candidates were considered as positive hits, the accuracy increased to 0.733 **(Fig. 2c)**. Interesingly, substantial variation in top-1 accuracy was observed across perturbed genes, independent of sample size **(Fig. 2d)**. Correlation analysis between gene performance and the Local Inverse Simpson's Index (LISI) [27] suggests that CellNavi's accuracy is influenced by the degree of perturbation heterogeneity. Specifically,

perturbations with low average LISI values, indicative of a more distinct and homogeneous response, were associated with higher accuracy (top-1 accuracy > 0.8, **Fig. 2e**).

To demonstrate CellNavi's effectiveness, we compared it to two alternative methods: SCENIC+[4,5], a training-free approach that infers gene regulatory networks with a focus on master regulators from transcriptomic data, and GEARS[28], an in silico perturbation approach, which targets a partially inverse problem of cellular transition prediction **(Methods)**. Both SCENIC+ and GEARS significantly underperformed compared to CellNavi **(Fig. 2b-d)**. In addition, SCENIC+, the network-based approaches, faced challenges in identifying regulons at the single-cell level **(Fig. S2.2)**, and therefore struggled to make predictions in many cases.

Given that CRISPR perturbation directly alters the expression of target genes, it could be presumed that the models simply identify genes with significant expression changes. To dismiss this possibility, we performed differential gene expression (DGE) analysis and found that the rankings of differentially expressed genes were poorly correlated with the actual outcomes **(Fig. 2b-d)**. In addition, we conducted an ablation study by systematically removing the expression of perturbed genes from the input. Although a decrease in performance was observed, CellNavi still maintained substantial predictive accuracy, achieving an AUROC of 0.750, a top-1 accuracy of 0.206, and a top-5 accuracy of 0.377, far surpassing expectations of random prediction **(Fig. S2.3-S2.4)**. These findings indicate that CellNavi is capable of identifying driver genes from a broad gene pool.

We further tested CellNavi on the Norman dataset[29], which features a CRISPR interference screen on the K562 cell line. This dataset encompasses 105 single-gene and 131 gene pair perturbations, allowing us to assess CellNavi's performance on transitions driven by both single and multiple genes. Using the unsupervised Leiden algorithm[30], we stratified the cells into distinct states, selecting a specific cell cluster for testing while training on the remaining clusters **(Fig. 2a)**. To ensure rigorous evaluation, we excluded all multi-gene perturbations from the training set.

CellNavi continued to demonstrate strong performance on single driver gene prediction in the Norman dataset **(Fig. 2f-g)**. Since CellNavi, like many other models, can rank candidate genes directly, we focused on the predicted rankings of driver genes in the multi-gene scenario. CellNavi achieved an average ranking of 7.9 and 31.2 for the first and second perturbed genes, respectively, out of 105 candidates, significantly outperforming all other tested methods **(Fig. 2h)**.

These robust results across diverse datasets and metrics underscore CellNavi's strong capability to identify genes driving cellular changes, even within previously unexplored cell states.

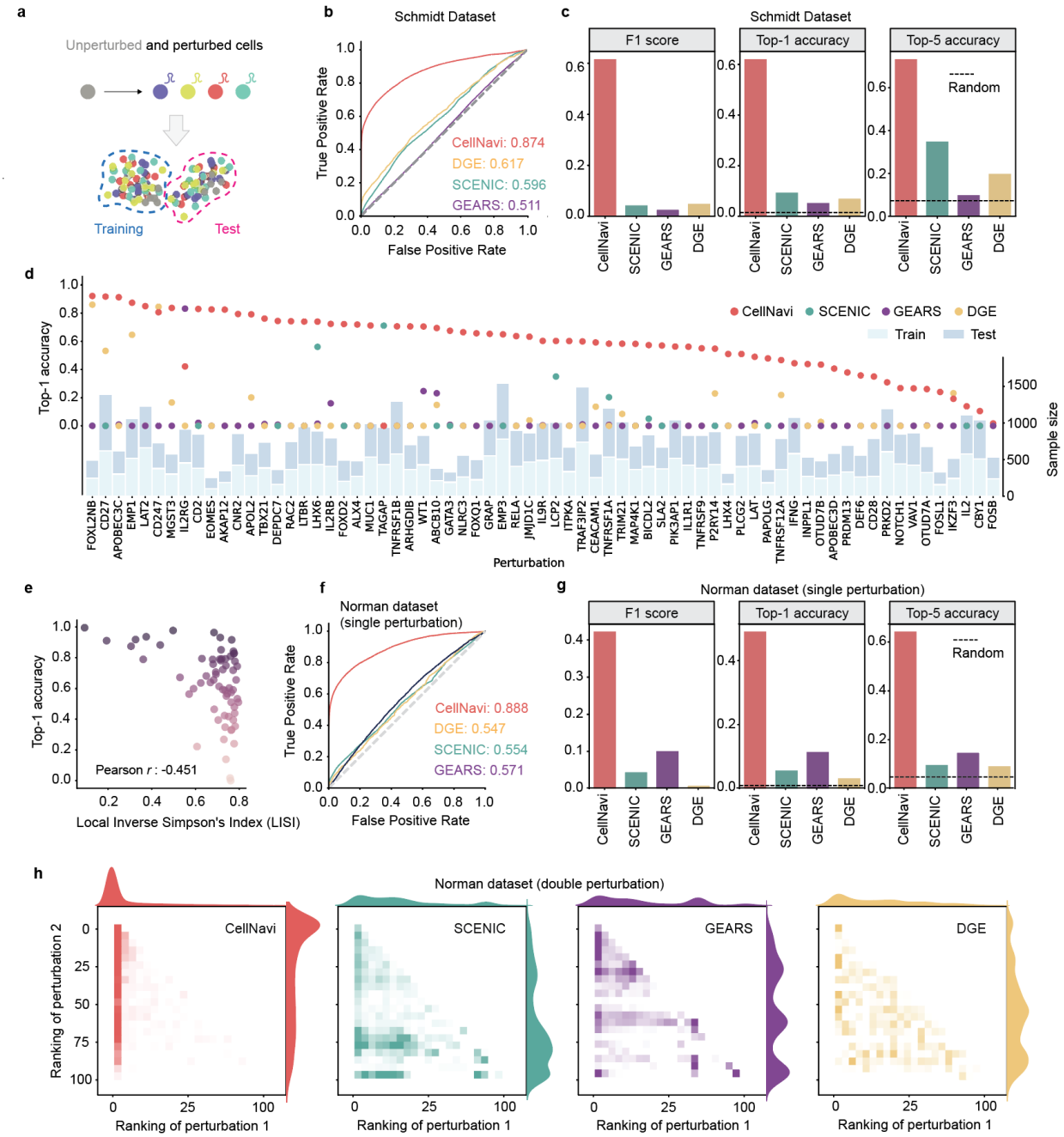


Figure 2. Quantitative assessment of CellNavi. a) Schematic of the quantitative evaluation framework. CRISPR-perturbed cells and their unperturbed controls were used for model training and evaluation, with data split by cell states for more rigorous testing. **b)** AUROC

scores for driver gene prediction in the Schmidt dataset, comparing CellNavi with alternative methods. **c)** F1 score, top-1 accuracy, and top-5 accuracy for driver gene prediction in the Schmidt dataset, comparing CellNavi with alternative methods. The dashed line indicates the performance of a random guess. **d)** Average top-1 accuracy for each gene. Left y-axis: top-1 accuracy of different methods for each gene. Right y-axis: the number of training (light blue) and test (steel blue) samples. **e)** Negative correlation between CellNavi's top-1 accuracy and the average LISI score across genes (Pearson correlation coefficient: -0.451). A LISI score of 1 indicates indistinguishable perturbation effects, while a score of 0 suggests a distinct perturbation pattern. Dot colors represent top-1 accuracy for individual genes. **f)** AUROC scores for driver gene prediction in the Norman dataset (single perturbation), comparing CellNavi with alternative methods. **g)** F1 score, top-1 accuracy, and top-5 accuracy for driver gene prediction in the Norman dataset (single perturbation), comparing CellNavi with alternative methods. The dashed line indicates the performance of a random guess. **h)** Distribution of predicted rankings for perturbed gene pairs. "Perturbation 1" represents genes ranked higher, and "Perturbation 2" represents genes ranked lower.

CellNavi identifies key genes involved in T cell differentiation

We next applied CellNavi to the Cano-Gomez dataset[31], which profiles T cell differentiation by stimulating naïve and memory CD4+ T cells in vitro with anti-CD3/anti-CD28 and a range of cytokines. During this differentiation process, external signals, such as antigens and cytokines, activate key genes responsible for modulating genetic circuits and gene expression programs, allowing T cells to adopt specialized functions. Our goal was to determine whether CellNavi could identify these essential driver genes during transitions driven by external stimuli, rather than by direct genetic perturbations, in differentiation processes.

Upon analysis of the Cano-Gomez dataset, we observed that cytokine induction led cells to differentiate into distinct states, as reflected by the enrichment of canonical marker genes [26,31–36] for several effector T cell subtypes **(Fig. S3.1)**. In particular, marker genes characteristic of IL2-high, IFNγ-high, and Th2 cells were strongly enriched **(Fig. S3.1)**. To quantify these changes, we compiled a comprehensive set of marker genes and calculated a "transition score" for each cell, which allowed us to categorize cells into specific groups based on their differentiation states **(Fig. 3a-c, Methods)**.

We examined CellNavi's ability to identify distinct driver genes across various effector T cell groups. To do this, we constructed source-target cell pairs using Th0 cells as the source and cytokine-induced cells as targets. These pairs were input into a CellNavi model trained on the Schmidt dataset, which contains extensive data on immune system-related gene programs. Next, we curated a list of established genes recognized as drivers of phenotypic transitions towards specific effector cell types, based on literature[26,35–42] **(Table S1)**, and evaluated CellNavi's performance in prioritizing these known driver genes.

CellNavi accurately ranked CD28 and VAV1, key drivers of IL2-high cells, as the top candidates in the IL2-high group defined by the transition score **(Fig. 3d)**. Similarly, high rankings were observed for *CD27* and *IL9R* in IFNγ-high cells, and GATA3 in Th2 cells **(Fig. 3d)**. We further analyzed the average rankings of these established driver genes across the different effector cell groups. As expected, the relevant driver genes consistently ranked higher in their corresponding cell groups where they are known to drive differentiation. Notably, *CD28, VAV1, CD27,* and *IL9R* achieved average rankings of 2.6, 2.9, 5.3, and 8.5, respectively, in their associated cell groups, significantly outperforming their rankings in unrelated groups **(Fig. 3e)**. These results demonstrate CellNavi's effectiveness in identifying key genes that govern distinct differentiation pathways while distinguishing between different cell fates. However, we noted that *CD28*'s dual role in IL2 and IFNγ regulation was not fully captured by the model. Additionally, although *GATA3* ranked highly in Th2 cells, its average ranking was not as strong as expected. Upon further inspection of the Th2 cluster, we observed that *GATA3* was ranked first in an aggregated subset of cells, but its ranking was more dispersed across the entire Th2 group **(Fig. 3f)**, suggesting heterogeneity within the cluster.

Next, we examined the likelihood scores generated by CellNavi for the driver genes across different cell groups. For the established driver genes, CellNavi assigned significantly elevated likelihood scores within their corresponding cell groups compared to other groups **(Fig. 3g)**. These likelihood scores could also effectively classify the different cell groups, as demonstrated by strong F1 and AUROC scores **(Fig. 3h-i, Methods)**. These findings confirm that CellNavi's likelihood scores not only prioritize key driver genes accurately but also effectively differentiate between cell states undergoing specific transitions.

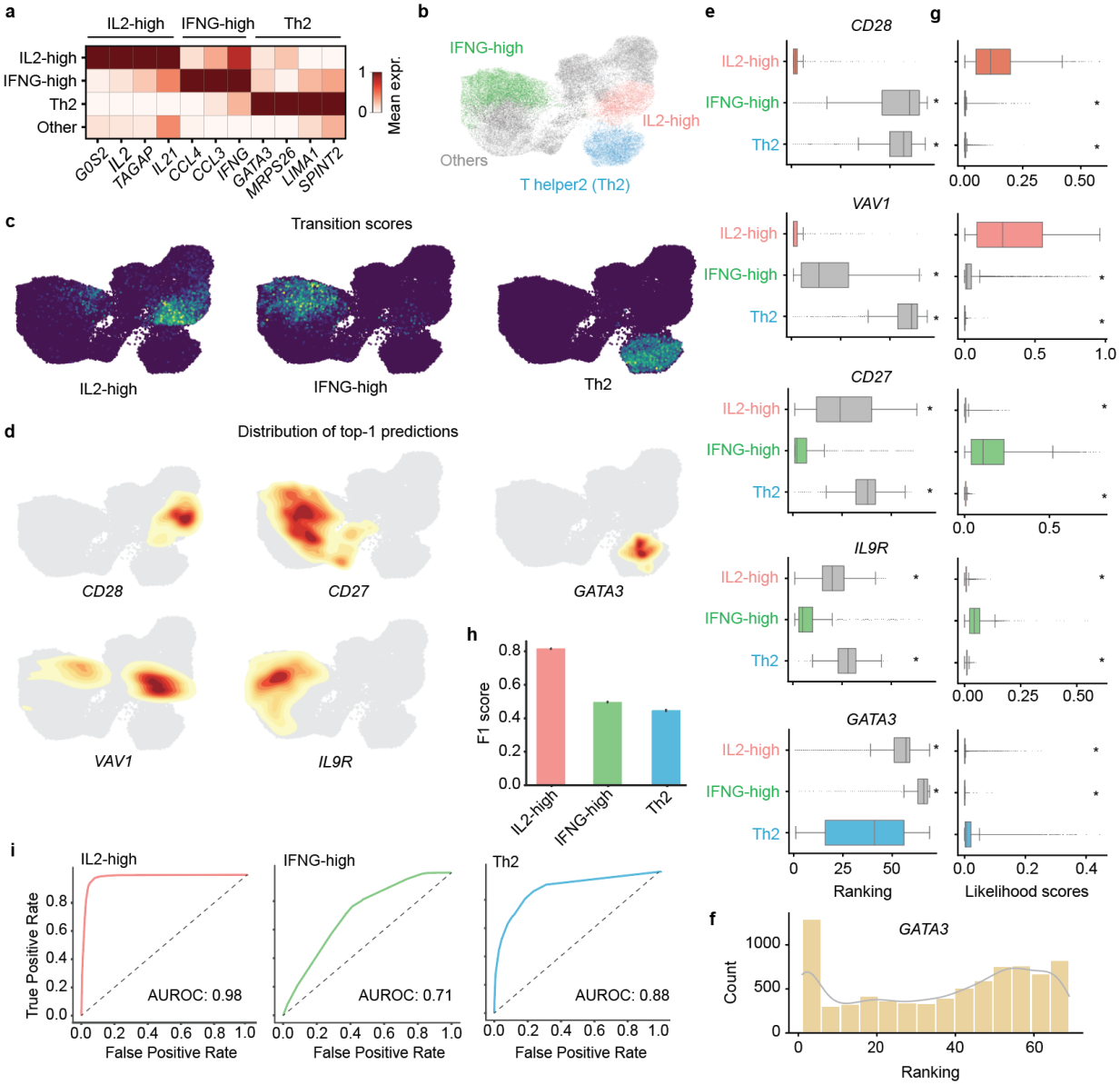


Figure 3. CellNavi identifies key genes involved in T cell differentiation. a) Changes in expression levels for canonical marker genes corresponding to specific T cell groups. **b)** Transition scores calculated based on IL2-high, IFNγ-high, and Th2-related marker genes mentioned in panel a. **c)** UMAP visualization of source-target CD4+ T cell pairs, colored by effector T cell groups classified by transition scores. Each cell pair is reduced from high-dimensional representations learned by CellNavi. **d)** Distributions of established driver genes predicted by CellNavi for IL2-high cells (*CD28*, *VAV1*), IFNγ-high cells (*CD27*, *IL9R*), and Th2 cells (*GATA3*). **e)** Predicted rankings of established driver genes across different cell groups. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x

interquartile range; points, outliers. *P*-values between were calculated with Mann–Whitney *U* test. Star (*) indicates a *p*-value smaller than 1e-6. **f)** Distribution of predicted rankings for *GATA3* in Th2 cells. **g)** Predicted likelihood scores for established driver genes in different cell groups. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers. *P*-values between were calculated with Mann–Whitney *U* test. Star (*) indicates a *p*-value smaller than 1e-6. **h)** F1 scores for using likelihood scores to infer effector T cell types. Error bars represent the standard error, calculated from ten-fold cross-validation. See Methods for details. **i)** AUROC for using likelihood scores for effector T cell type inference. See Methods for details.

CellNavi predicts key genes involved in neurodegenerative pathogenesis

We then investigated whether CellNavi could predict key genes involved in disease progression. We applied CellNavi to an *in vitro* model system of neurodegenerative diseases, specifically the Fernandes dataset[43]. This system comprises induced pluripotent stem cells (iPSC)-derived dopaminergic neurons subjected to Tunicamycin treatment. Tunicamycin induces endoplasmic reticulum (ER) stress and Parkinson's disease (PD)-like symptoms by inhibiting N-linked glycosylation[44], a process that affects many proteins post-translationally, without perturbing any single gene directly.

Prior to this analysis, CellNavi was trained on single-cell CRISPR screening data on iPSC-derived neurons from a different study, the Tian dataset[45]. While both studies investigate neurodegenerative diseases using human iPSC-derived neurons, they differ in the source of iPSCs and the differentiation protocols used, resulting in the generation of distinct neuron types[43,45,46] **(Fig. 4a)**.

To determine if CellNavi could identify genes related to the ER stress response, we utilized untreated cells from the Fernandes dataset as the source population and cells exposed to tunicamycin as the target population, and analyzed approximately 47,000 source-target cell pairs. We then asked CellNavi to prioritize 184 candidate genes, of which 5 were strongly associated with the ER stress response. CellNavi successfully pinpointed *EIF2S1*, *BAX*, and *HSPA5*, which achieved median rankings of 3, 7, and 16, respectively, among the candidate genes **(Fig. 4b)**. However, *HYOU1* and *VCP* were ranked lower. One possible explanation is that these genes might play more nuanced roles in the ER stress response or be involved in pathways not prominently activated under the experimental conditions of this study.

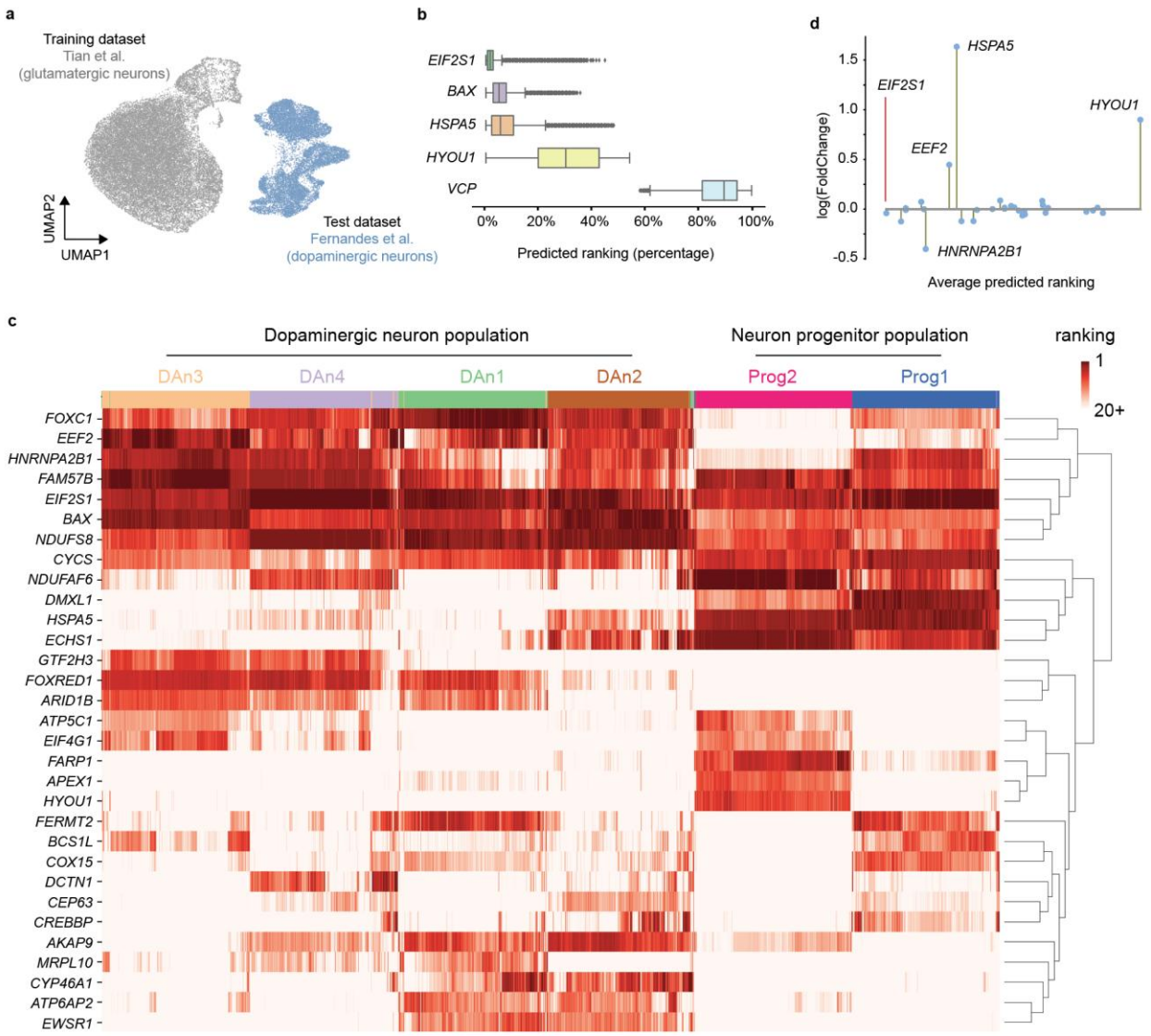


Figure 4. CellNavi predicts key genes involved in neurodegenerative pathogenesis. a) UMAP visualization of transcriptomic profiles from neurodegenerative disease-related datasets. Gray: iPSC-derived glutamatergic neurons (Tian dataset) used for model training. Blue: iPSC-derived dopaminergic neurons (Fernandes dataset). **b)** Predicted rankings for ER stress response-associated genes based on likelihood score vectors generated by CellNavi. Center line, median; box limits, upper and lower quartiles; cekers, 1.5x interquartile range; points, outliers. **c)** Distribution of the top 20 predicted genes across all cell pairs. Rows represent cell types as defined by the original publication (Fernandes et al.). Darker colors indicate higher rankings, while lighter colors indicate lower rankings. Ward's method was used for hierarchical clustering. **d)** Expression changes for the top 20 predicted genes. The

303 x-axis represents the average ranking of each gene across cell pairs, and the y-axis shows
304 the fold change in expression between target cells and source cells. See Figure S4.1 for
305 further details.

306

307 We also examined the top 20 predicted genes in each cell. A total of 31 genes showed
308 significant enrichment in this list **(Fig. 4c)**. Among these, *FAM57B*, *EIF2S1*, *NDUFS8*, *BAX*,
309 and *CYCS* consistently ranked at the top across the majority of cells. Notably, *EIF2S1* and
310 *BAX* are well-established key players in the ER stress response, while *NDUFS8* and *CYCS* are
311 likely involved in the mitochondrial stress response, which is often closely associated with
312 ER stress[47]. In parallel, Fernandes et al. previously classified iPSC-derived dopaminergic
313 neurons into six subtypes based on transcriptomic profiles. Our top 20 predictions for each
314 cell group aligned well with these subtypes, while also revealing unique gene preferences
315 within each subtype. For instance, our model suggests that *FARP1*, *CELF1*, *HYOU1*, and
316 *APEX1* may play more critical roles in progenitor cells **(Fig. 4c)**, which is subjected to further
317 investigation. Lastly, apart from *HSPA5* and *HYOU1*, the expression levels of the predicted
318 genes did not exhibit drastic changes **(Fig. 4d and Fig. S4.1)**, consistent with previous
319 observations that CellNavi can identify key genes beyond mere transcriptomic fluctuations.

320

321 **CellNavi elucidates the mechanisms of action for drug compounds**

322 Understanding the mechanisms of action (MoA) of novel drug candidates may enhance drug
323 safety and efficacy, reduce development costs, and accelerate drug discovery process.
324 However, conventional drug screening paradigms often fall short in elucidating the cellular-
325 level effects that drive biological functions and therapeutic outcomes.

326 Here, we applied CellNavi to predict key genes modulated by Histone Deacetylase (HDAC)
327 inhibitors, a class of anti-tumor drugs with promising therapeutic potential in cancer
328 treatment[48]. HDACs are enzymes integral to post-translational protein modifications and
329 interact with various oncogenic pathways to promote tumor progression[49,50]. The intricate
330 downstream pathways influenced by HDAC presents a significant challenge in fully
331 understanding mechanisms through which HDAC inhibitors exert their effects within cells.

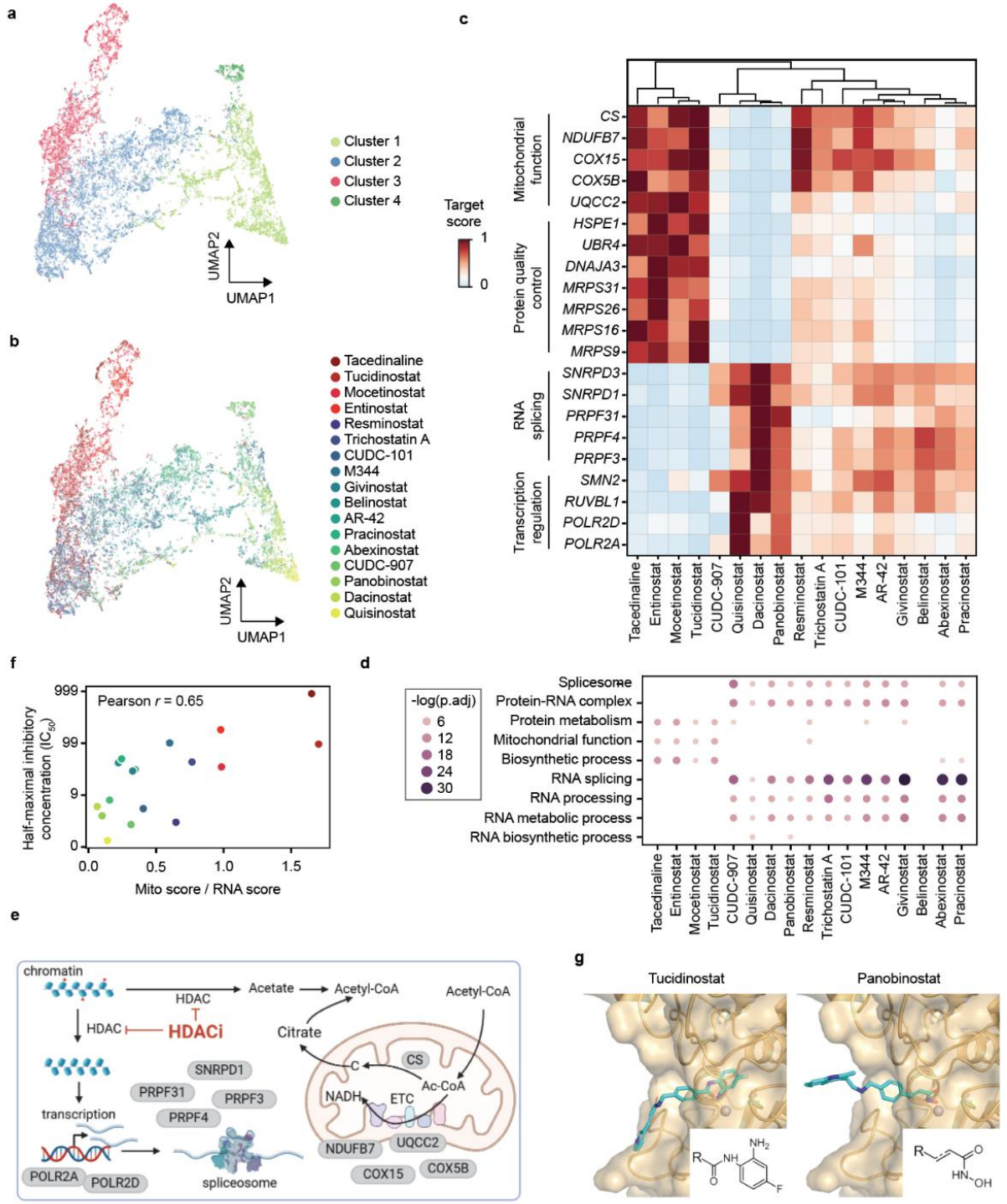
332 For this purpose, we applied CellNavi to a chemical screen that quantifies the
333 transcriptomic response of K562 cells to 17 distinct HDAC inhibitors (referred to as the
334 Srivastan dataset)[51]. In this setup, vehicle-treated cells were designated as the source, while
335 cells exposed to the HDAC inhibitors served as the target. The predicted likelihood score for

336 each of the candidate driver genes indicated whether a gene was modulated during drug
337 treatment, with higher scores suggesting a more prominent role during treatment with
338 specific HDAC inhibitors. To be noted, CellNavi was trained exclusively on genetic
339 perturbations[25], without any compound information.

340 While the transcriptomic data depicted a mixed response across the inhibitors **(Fig. S5.1)**,
341 the likelihood score vectors effectively categorized the inhibitors into distinct clusters **(Figs.**
342 **5a-b and Fig. S5.2)**. Further analysis revealed diversity in the top-ranked driver genes across
343 different inhibitors **(Fig. 5c)**. Specifically, cells treated with mocetinostat, tucidinostat,
344 entinostat, and tacedinaline, which were grouped in cluster 3 **(Fig. S5.2)**, exhibited high
345 likelihood scores for genes relevant to mitochondrial functions, such as *MRPS31* and
346 *NDUFB7*. In contrast, the majority of other compounds elicited high likelihood scores for
347 genes related to RNA splicing and transcription regulation, such as *PRPF3* and *POLR2A*.

348 Consistent with these observations, Gene Ontology (GO) enrichment analysis of the top 50
349 genes predicted for each treatment uncovered a similar pattern **(Fig. 5d, Fig. S5.3)**:
350 compounds in cluster 3 were enriched for genes involved in biosynthetic processes,
351 mitochondrial functions, and protein metabolism, whereas compounds in other clusters
352 were enriched in gene programs related to RNA splicing, processing, and metabolism. These
353 findings resonate with the established understanding that the inhibition of deacetylation
354 reduces cytoplasmic acetate levels and alters acetyl-CoA concentrations, a key metabolite
355 involved in cellular metabolism[51]. Moreover, the results suggest that certain HDAC inhibitors
356 may preferentially target chromatin regions that regulate genes associated with RNA
357 processing, which are crucial for the tumor cells proliferation [52–54] **(Fig. 5e)**.

358 Intriguingly, a correlation appears to exist between the selectivity of downstream gene
359 programs and the half-maximal inhibitory concentration (IC_{50}) values reported in the
360 literature[51] **(Fig. 5f)**. Specifically, drugs with lower IC_{50} values tend to influence RNA
361 processes, whereas those with higher IC_{50} values impact mitochondrial functions. To further
362 explore the potential molecular mechanisms underlying this variability in downstream gene
363 programs and compound efficacy, we examined the interactions between panobinostat
364 (which scores high for RNA-related genes) and tucidinostat (which prioritizes mitochondrial-
365 related genes) with the human HDAC2 protein. Despite the absence of significant
366 differences in their potential interactions with the zinc-dependent HDAC protein, as
367 suggested by molecular docking, we observed that the aniline group attached to the
368 benzamide in tucidinostat allows it to embed more deeply into the binding pocket **(Fig. 5g)**.
369 Interestingly, all four compounds within cluster 3 exhibit similar warheads, a characteristic
370 not seen in other molecules **(Fig. 5g and Fig. S5.4)**. This structural feature introduces a steric



371

372 **Figure 5. CellNavi reveals diverse downstream gene programs affected by HDAC**
373 **inhibitors. a-b)** UMAP visualization of cells treated with HDAC inhibitors. Each cell is
374 represented as a 2,057-dimensional vector consisting of likelihood scores predicted by

CellNavi for each candidate driver gene. In **a**, cells are colored by clusters identified using the Leiden algorithm, and in **b**, by HDAC inhibitor type. **c)** Average likelihood scores for top-ranked genes in each treatment group, with hierarchical clustering performed using Ward's method. **d)** Gene Ontology (GO) enrichment analysis for each treatment group. The size and darkness of the dots correlate positively with the adjusted *p*-value. See Figure S5.3 for a complete list of GO enrichment results. **e)** Schematic representation of HDAC inhibitor mechanisms. Proteins encoded by top-ranked driver genes are shown in gray, and red dots on chromatin indicate histone acetylation. ETC: electron transport chain. Ac-CoA: acetyl-CoA. Diagram created with BioRender.com. **f)** Scatter plot showing the correlation between IC50 values and the functional selectivity predicted by CellNavi. Each dot represents a compound, colored according to the clusters in **b**. Mito score: averaged likelihood scores for genes involved in mitochondrial functions. RNA score: averaged likelihood scores for genes involved in RNA regulation. **g)** Binding modes of tucidinostat and panobinostat at the active site of the zinc-dependent HDAC2 enzyme, with the enzyme represented as a surface representation and the drug compounds in stick representation. Shared warhead (or structural) motifs of different compound classes are highlighted in the bottom right corner. See Figure S5.4 for a complete list of molecular structures.

effect, which we hypothesize may influence the efficacy of compounds[55] and lead to divergent downstream response, a phenomenon known as functional selectivity frequently observed in drugs targeting G protein-coupled receptors[56] and other proteins[57–59]. This finding underscores CellNavi's potential to elucidate the intricate mechanisms of action underlying drug interventions, highlighting an approach to optimize drug efficacy and specificity for targets involving complex downstream signaling pathways.

CellNavi predicts driver genes in novel cell types

Next, we evaluated the generalization capability of CellNavi by testing its performance across distinct biological systems. We focused on a CRISPR inference screening experiment across HEK293FT and K562 cell lines[60], which are markedly different in origin and characteristics—HEK293FT cells are derived from human embryonic kidney cells, while K562 cells are derived from human chronic myelogenous leukemia **(Fig. 6a)**. In this experiment, CellNavi was trained on HEK293FT cells, with all K562 cells held out as the test set (Methods).

For the 16 perturbations involved in the cleavage and polyadenylation (CPA) regulatory machinery (**Fig. 6a**), CellNavi achieved a macro F1 score of 0.432, when only the top 1 predicted genes were considered (**Fig. 6b**). The model misclassified some genes in the CPSF and CSTF complexes, likely due to the similarity in their transcriptomic profiles post-perturbation (**Fig. 6c**). However, the model performed well in predicting *CPSF6* and *NUDT21*, which exhibit highly similar transcriptomic profiles post-perturbation. Interestingly, despite the distinct transcriptomic profiles of *RPRD1A* and *RPRD1B* post-perturbation, the model confused these genes in many cases. *RPRD1A* and *RPRD1B* can form heterodimers to dephosphorylate the RNA polymerase II C-terminal domain[61], suggesting that our model may be prioritizing functional interactions and shared pathways over transcriptomic differences, causing it to misinterpret these genes.

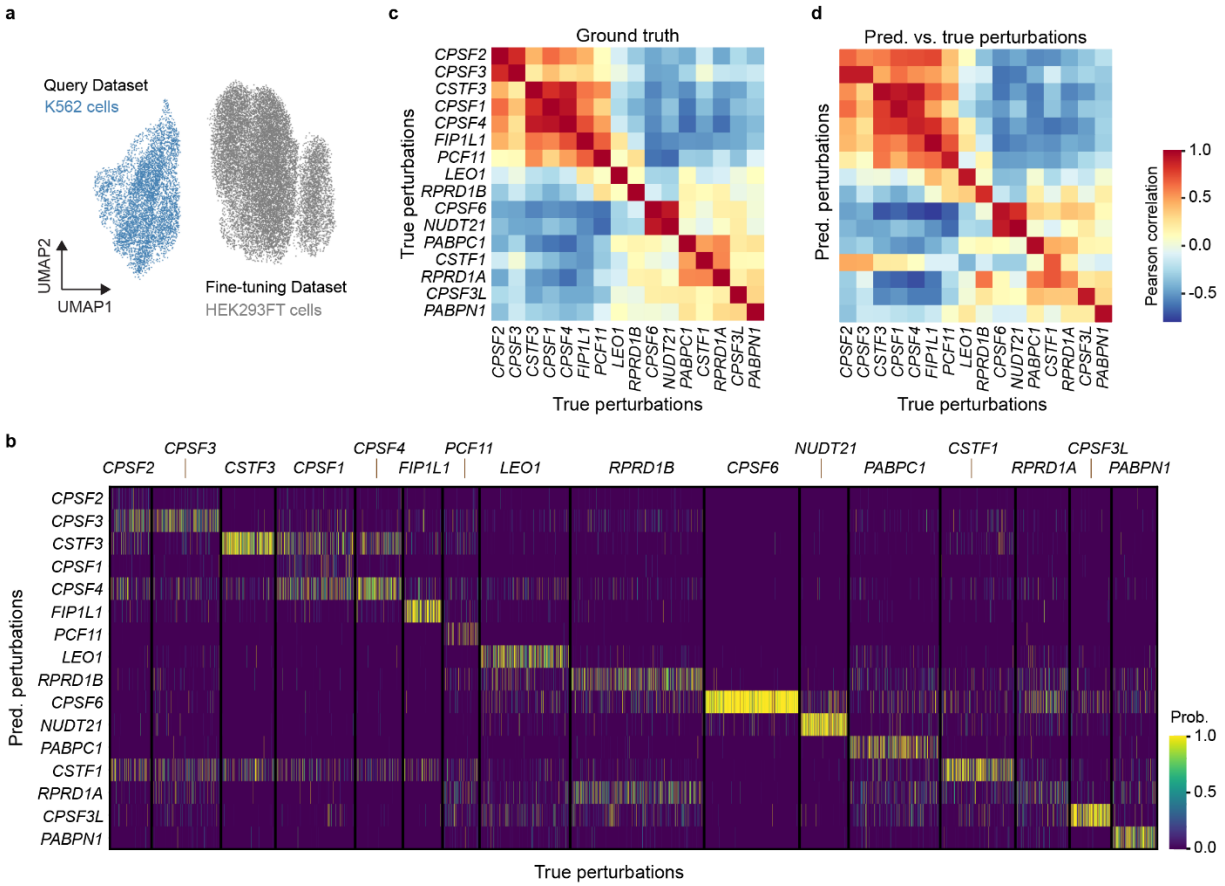


Figure 6. CellNavi predicts driver genes in novel cell types. a) UMAP visualization of transcriptomic profiles from Kowalski et al. Upper panel: Gray: HEK293FT cells used for model fine-tuning. Blue: K562 cells. Bottom panel: Each color represents a unique perturbation. **b)** Predicted perturbations vs. True perturbations in K562 cells. Each row represents a predicted perturbation, and each column represents a cell, whose true perturbation is labeled on top. **c)**

Heatmap showing average Pearson correlations over transcriptomic profiles between each pair of perturbations in K562 cells. **d)** Heatmap showing average Pearson correlations over transcriptomic profiles between predicted perturbations and true perturbations in K562 cells. Row: predicted perturbations. Column: true perturbations.

Finally, we compared the similarities between cell groups stratified by true perturbations and those stratified by predicted perturbations. We found that both the intra-perturbation and inter-perturbation correlation for predicted labels closely mirrored the correlation for true labels **(Fig. 6c-d and Fig. S6.0)**. These results indicate that cells stratified by predicted perturbations exhibited gene signatures that were highly similar to those of cells stratified by true perturbations. Though the prediction accuracy may be linked to the conservation of perturbations across cell types, overall, CellNavi makes reasonable predictions regarding driver genes in cell state transitions across very different cell types.

Discussion

Understanding the regulatory mechanisms that govern cell identity and its manipulation stands as a central focus in cell biology [6,62–65]. In this study, we introduce CellNavi, a deep learning-based framework for effective and generalized prediction of driver genes—key factors that orchestrate complex cellular transitions—across diverse biological contexts. We demonstrate that CellNavi performs well across various quantitative tasks and provides valuable predictions in real-world applications involving distinct biological processes.

The effectiveness and generalizability of CellNavi can be attributed to its accurate modeling of a cell state manifold. Describing cell states on a manifold that captures their biological dimensions has been a long-lasting endeavor [29,66–69]. However, existing methods for constructing and representing these manifolds vary widely and are often conceptual. Here, we leveraged large-scale single-cell transcriptomics data and a comprehensive prior gene graph with rich information on gene relationships and interactions, to model the cell state manifold using deep neural networks. This approach enhances the biological context of gene relationships and improved the representation of cell state similarities through gene graphs. As a result, we are able to describe cell states in a more biologically meaningful way, enabling the development of models tailored for various tasks, such as driver gene prediction, which is the focus of this study.

The integration of prior gene graphs in manifold learning sets CellNavi apart from other deep learning approaches for cell representation learning. This integration enables CellNavi to capture and represent cell states in complex biological context with greater accuracy. For instance, we benchmarked CellNavi to leading single-cell foundation models on the task of cell type annotation, a task requiring cell embeddings to accurately reflect the underlying biology. We utilized a multiple sclerosis dataset[70], which was widely employed to evaluate single-cell foundation models[12–14,17], but previous studies have reported challenges in accurately distinguishing cell types within this complex system. Using the same data splits and evaluation strategies, CellNavi demonstrated superior performance, achieving an accuracy of 0.990 and a Macro F1 of 0.987 **(Fig. S6.1)**, marking a significant advancement over existing foundation models. In addition, CellNavi places a stronger emphasis on transcription factors that are critical for defining cell states and orchestrating transitions[3,10,62] **(Fig. S6.2-6.3)**. This focus on regulatory elements provides CellNavi with a distinct advantage, enabling it to more effectively model complex biological processes.

In contrast to network-based approaches, or other traditional approaches, deep learning models like CellNavi have several unique advantages. It offers a more flexible and powerful framework by integrating diverse types of data, including single-cell CRISPR screening, to model complex gene interactions and cellular dynamics. Meanwhile, the adaptability of deep learning models enables them to refine their predictions as new data becomes available, making them exceptionally promising tools for advancing our understanding of cell state transitions. Our work marks the pioneering effort to establish benchmarks for evaluating the generalization ability of driver gene identification and rigorously assess a deep learning model on these tasks. Through our efforts, we are on the brink of achieving a new level of precision in identifying driver genes, paving the way for novel treatments targeting the root causes of diseases with unprecedented specificity.

While CellNavi marks a significant step forward, there is still room for improvement. For example, despite its robust design, the current CellNavi pipeline requires training on single-cell CRISPR screen data relevant to the system of interest. Determining the boundaries of this relevance, or how well CellNavi can generalize to different contexts, remains an open question. Quantifying such boundaries is complex due to the lack of an absolute metric for measuring the distance between cell types from different experiments and the limited availability of comprehensive data sets. As a proof of concept, we tested CellNavi's performance by fine-tuning on HEK293FT data and testing on K562 data **(Figure 6)**. Although the results were promising, the challenge may have been mitigated by the experiments being conducted within the same study by the same group. Future work should explore CellNavi's

boundaries using more diverse contexts to fully understand its generalization capabilities, while we expect the predictive power of CellNavi limited to the scope of fine-tuning dataset.

Moreover, CellNavi's ability to generalize to new genes and predict long-range transitions requires further evaluation due to the scarcity of high-quality data. Extending predictions to novel genes and diverse cellular contexts is essential for fully realizing CellNavi's potential in advancing disease understanding and treatment. Therefore, a comprehensive single-cell CRISPR experiment encompassing a broader range of target genes and cell types is highly expected.

Despite these challenges, CellNavi represents a significant advancement in our comprehension of cell state transitions and the identification of driver genes. As we continue to enhance and expand the capabilities of models like CellNavi, we move closer to realizing the full potential of deep learning in transforming biology and medicine.

Reference

1. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).

2. Li, P. *et al.* Reprogramming of T Cells to Natural Killer–Like Cells upon Bcl11b Deletion. *Science* **329**, 85–89 (2010).

3. Pioneer transcription factors: establishing competence for gene expression. https://genesdev.cshlp.org/content/25/21/2227.

4. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083–1086 (2017).

5. Bravo González-Blas, C. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods* **20**, 1355–1367 (2023).

6. Wang, P. *et al.* Deciphering driver regulators of cell fate decisions from single-cell transcriptomics data with CEFCON. *Nat Commun* **14**, 8459 (2023).

7. Fleck, J. S. *et al.* Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* **621**, 365–372 (2023).

8. Yuan, Q. & Duren, Z. Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. *Nat Biotechnol* 1–11 (2024) doi:10.1038/s41587-024-02182-7.

9. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods* **17**, 147–154 (2020).

10. Badia-i-Mompel, P. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat Rev Genet* **24**, 739–754 (2023).

11. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, e27041 (2017).

12. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* 1–9 (2023) doi:10.1038/s41586-023-06139-9.

13. Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* 1–11 (2024) doi:10.1038/s41592-024-02201-0.

14. Yang, F. *et al.* scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat Mach Intell* **4**, 852–866 (2022).

15. Zhao, W. X. *et al.* A Survey of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2303.18223 (2023).

16. Kaddour, J. *et al.* Challenges and Applications of Large Language Models. Preprint at http://arxiv.org/abs/2307.10169 (2023).

17. Hao, M. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* 1–11 (2024) doi:10.1038/s41592-024-02305-7.

18. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat Methods* **17**, 159–162 (2020).

19. Ying, C. *et al.* Do Transformers Really Perform Bad for Graph Representation? *arXiv.org* https://arxiv.org/abs/2106.05234v5 (2021).

20. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e17 (2016).

21. Adamson, B. *et al.* perturbation-A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882.e21 (2016).

22. Datlinger, P. *et al.* Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat Methods* **18**, 635–642 (2021).

553 23. Jaitin, D. A. *et al.* Dissecting Immune Circuits by Linking CRISPR-Pooled Screens
554 with Single-Cell RNA-Seq. *Cell* **167**, 1883-1896.e15 (2016).

555 24. Replogle, J. M. *et al.* Combinatorial single-cell CRISPR screens by direct guide RNA
556 capture and targeted sequencing. *Nat Biotechnol* **38**, 954–961 (2020).

557 25. Replogle, J. M. *et al.* Mapping information-rich genotype-phenotype landscapes with
558 genome-scale Perturb-seq. *Cell* **185**, 2559-2575.e28 (2022).

559 26. Schmidt, R. *et al.* CRISPR activation and interference screens decode stimulation
560 responses in primary human T cells. *Science* **375**, eabj4008 (2022).

561 27. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with
562 Harmony. *Nat Methods* **16**, 1289–1296 (2019).

563 28. Roohani, Y., Huang, K. & Leskovec, J. *GEARS: Predicting Transcriptional Outcomes of*
564 *Novel Multi-Gene Perturbations*. http://biorxiv.org/lookup/doi/10.1101/2022.07.12.499735
565 (2022) doi:10.1101/2022.07.12.499735.

566 29. Norman, T. M. *et al.* Exploring genetic interaction manifolds constructed from rich
567 single-cell phenotypes. *Science* **365**, 786–793 (2019).

568 30. Traag, V., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-
569 connected communities. *arXiv.org* https://arxiv.org/abs/1810.08473v3 (2018)
570 doi:10.1038/s41598-019-41695-z.

571 31. Cano-Gamez, E. *et al.* Single-cell transcriptomics identifies an effectorness gradient
572 shaping the response of CD4+ T cells to cytokines. *Nat Commun* **11**, 1801 (2020).

573 32. Szabo, P. A. *et al.* Single-cell transcriptomics of human T cells reveals tissue and
574 activation signatures in health and disease. *Nat Commun* **10**, 4706 (2019).

575 33. Atlasy, N. *et al.* Single cell transcriptomic analysis of the immune cell compartment
576 in the human small intestine and in Celiac disease. *Nat Commun* **13**, 4920 (2022).

577 34. Cachot, A. *et al.* Tumor-specific cytolytic CD4 T cells mediate immunity against
578 human cancer. *Sci Adv* **7**, eabe3348 (2021).

579 35. Henriksson, J. *et al.* Genome-wide CRISPR Screens in T Helper Cells Reveal
580 Pervasive Crosstalk between Activation and Differentiation. *Cell* **176**, 882-896.e18 (2019).

581 36. Kanhere, A. *et al.* T-bet and GATA3 orchestrate Th1 and Th2 differentiation through
582 lineage-specific targeting of distal regulatory elements. *Nat Commun* **3**, 1268 (2012).

583 37. Kalbasi, A. *et al.* Potentiating adoptive cell therapy using synthetic IL-9 receptors.
584 *Nature* **607**, 360–365 (2022).

585 38. Charvet, C. *et al.* Vav1 Promotes T Cell Cycle Progression by Linking TCR/CD28
586 Costimulation to FOXO1 and p27kip1 Expression1. *The Journal of Immunology* **177**, 5024–
587 5031 (2006).

588 39. Fischer, K.-D. *et al.* Defective T-cell receptor signalling and positive selection of Vav-
589 deficient CD4+CDS+thymocytes. *Nature* **374**, 474–476 (1995).

590 40. Kane, L. P., Andres, P. G., Howland, K. C., Abbas, A. K. & Weiss, A. Akt provides the
591 CD28 costimulatory signal for up-regulation of IL-2 and IFN-γ but not TH2 cytokines. *Nat*
592 *Immunol* **2**, 37–44 (2001).

593 41. Jaeger-Ruckstuhl, C. A. *et al.* Signaling via a CD27-TRAF2-SHP-1 axis during naive T
594 cell activation promotes memory-associated gene regulatory networks. *Immunity* **57**, 287-
595 302.e12 (2024).

596 42. Lim, H.-S. *et al.* Costimulation of IL-2 Production through CD28 Is Dependent on the
597 Size of Its Ligand. *J Immunol* **195**, 5432–5439 (2015).

598 43. Fernandes, H. J. R. *et al.* Single-Cell Transcriptomics of Parkinson's Disease Human
599 In Vitro Models Reveals Dopamine Neuron-Specific Stress Responses. *Cell Reports* **33**,
600 108263 (2020).

601 44. Surani, M. A. Glycoprotein synthesis and inhibition of glycosylation by tunicamycin
602 in preimplantation mouse embryos: compaction and trophoblast adhesion. *Cell* **18**, 217–
603 227 (1979).

604 45. Tian, R. *et al.* Genome-wide CRISPRi/a screens in human neurons link lysosomal
605 failure to ferroptosis. *Nat Neurosci* **24**, 1020–1034 (2021).

606 46. Tian, R. *et al.* CRISPR Interference-Based Platform for Multimodal Genetic Screens
607 in Human iPSC-Derived Neurons. *Neuron* **104**, 239-255.e12 (2019).

608 47. Quirós, P. M., Mottis, A. & Auwerx, J. Mitonuclear communication in homeostasis
609 and stress. *Nat Rev Mol Cell Biol* **17**, 213–226 (2016).

610 48. Ho, T. C. S., Chan, A. H. Y. & Ganesan, A. Thirty Years of HDAC Inhibitors: 2020
611 Insight and Hindsight. *J. Med. Chem.* **63**, 12460–12484 (2020).

612 49. Glozak, M. A. & Seto, E. Histone deacetylases and cancer. *Oncogene* **26**, 5420–5432
613 (2007).

614 50. Marks, P. A. *et al.* Histone deacetylases and cancer: causes and therapies. *Nat Rev*
615 *Cancer* **1**, 194–202 (2001).

616 51. Srivatsan, S. R. *et al.* Massively multiplex chemical transcriptomics at single-cell
617 resolution. *Science* **367**, 45–51 (2020).

618 52. Bradley, R. K. & Anczuków, O. RNA splicing dysregulation and the hallmarks of
619 cancer. *Nat Rev Cancer* **23**, 135–155 (2023).

620 53. Stanley, R. F. & Abdel-Wahab, O. Dysregulation and therapeutic targeting of RNA
621 splicing in cancer. *Nat Cancer* **3**, 536–546 (2022).

622 54. Bonnal, S., Vigevani, L. & Valcárcel, J. The spliceosome as a target of novel
623 antitumour drugs. *Nat Rev Drug Discov* **11**, 847–859 (2012).

624 55. Di Micco, S. *et al.* Structural basis for the design and synthesis of selective HDAC
625 inhibitors. *Bioorganic & Medicinal Chemistry* **21**, 3795–3807 (2013).

626 56. Bock, A. *et al.* The allosteric vestibule of a seven transmembrane helical receptor
627 controls G-protein coupling. *Nat Commun* **3**, 1044 (2012).

628 57. Perszyk, R. E. *et al.* Biased modulators of NMDA receptors control channel opening
629 and ion selectivity. *Nat Chem Biol* **16**, 188–196 (2020).

630 58. Franks, L. N., Ford, B. M., Fujiwara, T., Zhao, H. & Prather, P. L. The tamoxifen
631 derivative ridaifen-B is a high affinity selective CB2 receptor inverse agonist exhibiting anti-
632 inflammatory and anti-osteoclastogenic effects. *Toxicology and Applied Pharmacology*
633 **353**, 31–42 (2018).

634 59. Samad, S. S., Schwartz, J.-M. & Francavilla, C. Functional selectivity of Receptor
635 Tyrosine Kinases regulates distinct cellular outputs. *Front. Cell Dev. Biol.* **11**, (2024).

636 60. Kowalski, M. H. *et al.* Multiplexed single-cell characterization of alternative
637 polyadenylation regulators. *Cell* **0**, (2024).

638 61. Ali, I. *et al.* Crosstalk between RNA Pol II C-Terminal Domain Acetylation and
639 Phosphorylation via RPRD Proteins. *Molecular Cell* **74**, 1164-1174.e4 (2019).

640 62. Joung, J. *et al.* A transcription factor atlas of directed differentiation. *Cell* **186**, 209-
641 229.e26 (2023).

642 63. Rukhlenko, O. S. *et al.* Control of cell state transitions. *Nature* **609**, 975–985 (2022).

64. Wang, J., Zhang, K., Xu, L. & Wang, E. Quantifying the Waddington landscape and biological paths for development and differentiation. *Proceedings of the National Academy of Sciences* **108**, 8257–8262 (2011).

65. Hormoz, S. *et al.* Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. *cels* **3**, 419-433.e8 (2016).

66. Rafelski, S. M. & Theriot, J. A. Establishing a conceptual framework for holistic cell states and state transitions. *Cell* **187**, 2633–2651 (2024).

67. Qiu, X. *et al.* Mapping transcriptomic vector fields of single cells. *Cell* **185**, 690-711.e45 (2022).

68. Liang, S. *et al.* Single-cell manifold-preserving feature selection for detecting rare cell populations. *Nat Comput Sci* **1**, 374–384 (2021).

69. Yao, Z., Li, B., Lu, Y. & Yau, S.-T. Single-cell analysis via manifold fitting: A framework for RNA clustering and beyond. *Proceedings of the National Academy of Sciences* **121**, e2400002121 (2024).

70. Schirmer, L. *et al.* Neuronal vulnerability and multilineage diversity in multiple sclerosis. *Nature* **573**, 75–82 (2019).

71. Goodfellow, I. J. *et al.* Generative Adversarial Networks. Preprint at https://doi.org/10.48550/arXiv.1406.2661 (2014).

72. Peidli, S. *et al.* scPerturb: harmonized single-cell perturbation data. *Nat Methods* **21**, 531–540 (2024).

Methods

Input embeddings

In CellNavi, we use single-cell raw count matrices as the only input. Specifically, the single-cell sequencing data is processed into a cell-gene count matrix, $\mathbf{X} \in \mathbb{R}^{N \times G}$, where each element $\mathbf{X}_{n,g}$ represents the expression of the n-th cell and the g-th gene (or read count of the g-th RNA).

To better describe a gene's state in a cell, we involve both gene name and gene expression information in its input embeddings. Formally, the input embedding of a token is the concatenation of gene name embedding and gene expression embedding.

Gene name embedding. We use a learnable gene name embedding in CellNavi. The vocabulary of genes is obtained by taking the union set of gene names among all datasets. Then the integer identifier of each gene in the vocabulary is fed into an embedding layer to obtain its gene name embedding. Additionally, we incorporate a special token CLS in the vocabulary for aggregating all genes into a cell representation. The gene name embedding of cell n can be represented as $\mathbf{h}_n^{(\mathrm{name})} \in \mathbb{R}^{(G+1) \times H}$:

$$\mathbf{h}_n^{(\mathrm{name})} = \left[\mathbf{h}_{n,\mathrm{CLS}}^{(\mathrm{name})}, \mathbf{h}_{n,1}^{(\mathrm{name})}, \mathbf{h}_{n,2}^{(\mathrm{name})}, \dots, \mathbf{h}_{n,G}^{(\mathrm{name})}\right],$$

where H is the dimension of embeddings.

Gene expression embedding. One major challenge in modeling gene expression is the variability in absolute magnitudes across different sequencing protocols[13]. We tackled this challenge by normalizing the raw count expression for each cell using the shifted logarithm, which is defined as,

$$\widetilde{\mathbf{X}}_{n,g} = \log\left(L \frac{\mathbf{X}_{n,g}}{\sum_{g'} \mathbf{X}_{n,g'}} + 1\right),$$

where $\mathbf{X}_{n,g}$ is the raw count of gene g in cell n, L is a scaling factor and we used a fixed value $L = 1 \times 10^4$ in this study, and $\widetilde{\mathbf{X}}_{n,g}$ denotes the normalized count. Finally, a linear layer was applied on the normalized expression $\widetilde{\mathbf{X}}_{n,g}$ to obtain the gene expression embedding. For the CLS token, we set it as a unique value for gene expression embedding. The gene expression embedding of cell n can be represented as $\mathbf{h}_n^{(\mathrm{expr})} \in \mathbb{R}^{(G+1) \times H}$:

$$\mathbf{h}_n^{(\mathrm{expr})} = \left[\mathbf{h}_{n,\mathrm{CLS}}^{(\mathrm{expr})}, \mathbf{h}_{n,1}^{(\mathrm{expr})}, \mathbf{h}_{n,2}^{(\mathrm{expr})}, \dots, \mathbf{h}_{n,G}^{(\mathrm{expr})}\right].$$

The final embedding of cell n is defined as the concatenation of $\mathbf{h}_n^{(\mathrm{name})}$ and $\mathbf{h}_n^{(\mathrm{expr})}$:

$$\mathbf{h}_n = \text{CONCAT}\left(\mathbf{h}_n^{(\text{name})}, \mathbf{h}_n^{(\text{expr})}\right) \in \mathbb{R}^{(G+1)\times 2H}.$$

Cell Manifold Model

Model architecture

The Cell Manifold Model, is composed of 6 layers of a Transformer variant that is designed specifically for processing graph-structured data (GeneGraph Attention Layers) [19]. The encoder takes the input embeddings to generate cell representations and uses only genes with non-zero expressions. To further speed up training, also as an approach of data augmentation, we performed a gene sampling strategy by randomly selecting at most 2,048 genes as input. It should be noted that the strategy is only applied during training, all non-zero genes are included at inference stage to avoid information loss. We use $\mathbf{h}_n^{(l)}$ to represent the embedding of cell n at the l-th layer, where $\mathbf{h}_n^{(l)}$ is defined as:

$$\mathbf{h}_n^{(l)} = \begin{cases} \mathbf{h}_n, & l = 0, \\ \text{GeneGraphAttnLayer}\left(\mathbf{h}_n^{(l-1)}\right), & l \in [1,6]. \end{cases}$$

The multi-head attention module in each GeneGraph Attention Layer consists of 3 components. In addition to a self-attention module, a centrality encoding module and a spatial encoding module are also incorporated to modify the standard self-attention module for graph data integration.

We start by introducing the standard self-attention module. Let N_{head} be the number of heads in the self-attention module. In the l-th layer, i-th head, self-attention is calculated as:

$$\mathbf{Q}_n^{(l,i)} = \mathbf{h}_n^{(l)}\mathbf{W}^{(\text{qry},i)}, \mathbf{K}_n^{(l,i)} = \mathbf{h}_n^{(l)}\mathbf{W}^{(\text{key},i)}, \mathbf{V}_n^{(l,i)} = \mathbf{h}_n^{(l)}\mathbf{W}^{(\text{val},i)},$$

$$\mathbf{A}_n^{(l,i)} = \frac{\mathbf{Q}_n^{(l,i)}\left(\mathbf{K}_n^{(l,i)}\right)^{\top}}{\sqrt{D}}, \text{Attn}_n^{(l,i)} = \text{softmax}\left(\mathbf{A}_n^{(l,i)}\right)\mathbf{V}_n^{(l,i)},$$

$$\mathbf{h}_n^{(l)\prime} = \text{CONCAT}\left(\text{Attn}_n^{(l,1)}, \cdots, \text{Attn}_n^{(l,N_{\text{heads}})}\right)\mathbf{W}^{(\text{out})} \in \mathbb{R}^{(G+1)\times 2H},$$

where $\mathbf{W}^{(\text{qry},i)}$, $\mathbf{W}^{(\text{key},i)}$, $\mathbf{W}^{(\text{val},i)} \in \mathbb{R}^{2H\times D}$ are learnable matrices that project input embedding $\mathbf{h}_n^{(l)}$ of cell n in to $\mathbf{Q}_n^{(l,i)}$, $\mathbf{K}_n^{(l,i)}$, $\mathbf{V}_n^{(l,i)}$, the symbol $\mathbf{W}^{(\text{out})} \in \mathbb{R}^{(D\,N_{\text{heads}})\times 2H}$ is a learnable linear projection that refines the output of multi-head attention, and D is the feature dimension for each attention head which satisfies $D\,N_{\text{heads}} = 2H$. The output of multi-head attention $\mathbf{h}_{\text{n}}^{(l)\prime}$ is then passed through a layer normalization layer and a multi-layer perceptron model (MLP), producing the final output $\mathbf{h}_n^{(l+1)}$ as the input to the next layer.

The standard attention mechanism only processes features on each single gene, while the gene graph provides the relational information among the genes. To incorporate the gene graph information into the model, the centrality encoding module projects the relational information into the regulatory activity feature of each single gene, and the spatial encoding module directly incorporates the relational information with the attention mechanism. To be more specific, we define $\mathbf{z}^{-}_{\deg^{-}(\mathcal{G},g)}$ and $\mathbf{z}^{+}_{\deg^{+}(\mathcal{G},g)}$, learnable embeddings describing in-degree $\deg^{-}$ and out-degree $\deg^{+}$ of gene g on the gene graph $\mathcal{G}$. We add these embeddings to the gene embeddings to update cell encoding:

$$\mathbf{h}^{(l)}_{n,g} = \mathbf{h}'^{(l)}_{n,g} + \mathbf{z}^{-}_{\deg^{-}(\mathcal{G},g)} + \mathbf{z}^{+}_{\deg^{+}(\mathcal{G},g)}.$$

This cell encoding update by the centrality encoding module is applied before the self-attention module.

The spatial encoding module aims to capture regulation relations between genes from the gene graph. For this purpose, we generate the distance matrix $\mathbf{S} \in \mathbb{N}^{G\times G}$ containing shortest distances between gene pairs on the gene graph $\mathcal{G}$. We assign each element in $\mathbf{S}$ as a learnable bias added to attention weights:

$$\mathbf{A}'_{g_1,g_2} = \mathbf{A}_{g_1,g_2} + b\left(\mathbf{S}_{g_1,g_2}\right),$$

where b is a learnable scalar-valued function of the distance $\mathbf{S}_{g_1,g_2}$. It assigns a special value to genes that are not connected to the graph. We use $\mathbf{A}'$ in place of the original attention weights $\mathbf{A}$ in the standard self-attention module when computing self-attention in our model. In our implementation, we apply layer normalization and an MLP before computing multi-head self-attention. The cell representations output from the CMM $\mathbf{h}^{(6)}_{n}$ is taken as the cell coordinate for cell n, denoted as $\mathbf{CRD}_n$.

CMM Pretraining Task

The CMM is expected to generate cell coordinates that parameterize the intrinsic features/variables (that are much less than the dimensions in the raw gene expression profile representation) of a cell state and maintain cell similarity in the vector space, so as to provide a concise and biologically relevant representation for the DGP to consume. To achieve this, we design a down-sampling reconstruction pretraining task, which asks the CMM to produce a cell coordinates of a down-sampled gene expression $\mathbf{X}^{(\mathrm{ds})}_{n}$ of a cell n, that allows a separate decoder model to reconstruct the original gene expression $\mathbf{X}_n$ of that cell as accurate as possible. In order to achieve this, the CMM is enforced to capture the co-varying patterns among the raw gene expression dimensions, hence helping the CMM to extract the underlying intrinsic variables.

Specifically, for the down-sampling process, we down-sample the raw count expression of each gene via a binomial distribution. The down sampled expression $\mathbf{X}_{n,g}^{(\mathrm{ds})}$ of the n-th cell and the g-th gene is produced by:

$$\mathbf{X}_{n,g}^{(\mathrm{ds})} \sim B\left(\mathbf{X}_{n,g}, \frac{1}{r^{(\mathrm{ds})}}\right),$$

where $\mathbf{X}_{n,g}$ is the raw count of gene g in cell n, $r^{(\mathrm{ds})}$ is the down sample rate which is uniformly sampled from $[1, 20)$, and B denotes the binomial distribution. The decoder is an MLP consisting of two linear layers. For each down-sampled gene expression, the decoder concatenates the cell coordinates $\mathbf{CRD}_n$ of $\mathbf{X}_n^{(\mathrm{ds})}$ produced by the CMM and the embedding of that gene as the direct input to the MLP. The MLP output comes in the same shape as $\mathbf{X}_n$.

The learning objective for reconstructing the original gene expression profile $\mathbf{X}_n$ from the down-sampled version $\mathbf{X}_n^{(\mathrm{ds})}$ is:

$$\mathcal{L}_{\mathrm{recons}} = \frac{1}{N}\sum_{n=1}^{N} \left\| \mathrm{DEC}\left(\mathbf{CRD}\left(\mathbf{X}_n^{(\mathrm{ds})}\right)\right) - \mathbf{X}_n \right\|^2,$$

where $\|\cdot\|^2$ represents the squared 2-norm of a vector. Both the CMM and the decoder are optimized. After pretraining, the CMM is to be used for driver gene prediction, while the decoder is discarded.

Driver Gene Predictor

The driver gene classifier is an MLP consisting of two linear layers. It is optimized to predict the perturbed genes from a pair of cell coordinates output by the Cell Manifold Model. To be more specific, transcriptomes of source cell $\mathbf{X}_{\mathrm{src}}$ and target cell $\mathbf{X}_{\mathrm{tgt}}$ are mapped to cell coordinates $\mathbf{CRD}_{\mathrm{src}}$ and $\mathbf{CRD}_{\mathrm{tgt}}$ with the Cell Manifold Model. For the direct input features, DGP concatenates the two cell coordinates and then proceeds with an MLP, which outputs the logits of genes. We use the cross-entropy loss for training DGP:

$$\mathcal{L}_{\mathrm{driver_gene}} = \mathrm{CE}\left(\mathrm{DGP}\left(\mathrm{CONCAT}\left(\mathbf{CRD}(\mathbf{X}_{\mathrm{src}}), \mathbf{CRD}(\mathbf{X}_{\mathrm{tgt}})\right)\right), g_{\mathrm{drv}}\right),$$

where $\mathrm{CE}(\mathbf{l}, g) = \frac{\mathbf{l}_g}{\log \sum_{g'} \exp(\mathbf{l}_{g'})}$ is the cross-entropy loss, and g_{drv} denotes the driver gene corresponding to $\mathbf{X}_{\mathrm{src}}$ and $\mathbf{X}_{\mathrm{tgt}}$. The loss is finally averaged over all $(\mathbf{X}_{\mathrm{src}}, \mathbf{X}_{\mathrm{tgt}}, g_{\mathrm{drv}})$ tuples in the dataset. The pretrained CMM used to produce $\mathbf{CRD}_{\mathrm{src}}$ and $\mathbf{CRD}_{\mathrm{tgt}}$ is also finetuned together with the DGP by this loss.

Optional Training Schema for CellNavi to Mitigate Batch Effect

In some cases, cellular transition data for training and testing the DGP in CellNavi may originate from different sequencing platforms or experimental techniques. Such technical discrepancies can introduce batch effects that obscure biological insights. We hence propose two additional finetuning strategies for the option to account for the batch effect.

Cell Coordinate Distribution Alignment

The batch effect may introduce a shift of the cell coordinate distributions over two batches, due to biologically irrelevant factors which we do not want the cell coordinate to keep. For this purpose, we adopt the adversarial generative training strategy[71]. The intuition is to ask the CMM to produce cell coordinates that confuse a batch label classifier, i.e., worsen the batch label classification accuracy. In this way, the cell coordinates are forced to discard any information about the batch it comes from, hence eliminating the distribution shift between the cell coordinates of the two batches.

Specifically, we use a two-layer MLP for the auxiliary batch label classifier model, also called a discriminator. The input of the MLP is the concatenation of the cell coordinates of a pair of cells from the same batch, and the output is a scalar representing the predicted logit for the binary classification of which batch does the pair come from. The loss function for the discriminator is hence the binary cross-entropy loss:

$$\mathcal{L}_{\text{discr}} = \text{CE}\left(\text{DISCR}\left(\text{CONCAT}\left(\mathbf{CRD}(\mathbf{X}_{\text{src}}), \mathbf{CRD}\left(\mathbf{X}_{\text{tgt}}\right)\right)\right), y_{\text{batch}}\right),$$

where y_{batch} represents the binary label indicating whether the input cell pair $\mathbf{X}_{\text{src}}$ and $\mathbf{X}_{\text{tgt}}$ comes from the test batch (in contrast to the training batch). The discriminator is asked to minimize the loss in order to try to extract any batch-related information from the cell coordinates, while the CMM is asked to maximize the loss with the help of the discriminator in order to reduce batch-related information to eliminate the batch effect. In practice, the discriminator is optimized alternately with the optimization of CMM.

Finetuning Stabilization by Cell Label Classifier

The adversarial training process is known unstable, and could degenerate the cell coordinates to a trivial representation that does not contain even biologically relevant information. To combat this, we also introduce an auxiliary cell label classifier, which uses the cell coordinates produced by CMM to predict a biologically related cell label. To allow accurate prediction, CMM is enforced to keep biologically related information in the cell coordinates. The cell label classifier is also a two-layer MLP, and is trained by the following loss function:

$$\mathcal{L}_{\text{cell}} = \text{CE}\left(\text{CELLCLS}\left(\mathbf{CRD}(\mathbf{X})\right), y_{\text{cell_label}}\right),$$

where $y_{\text{cell_label}}$ denotes the cell label of cell $\mathbf{X}$. The loss is applied exclusively on the test data set, and optimizes both the CMM and the cell label classifier CELLCLS.

Combined Loss for Finetuning

The overall loss function for the driver gene prediction finetuning in the case of batch effect is:

$$\mathcal{L}'_{\text{driver_gene}} = \alpha_{\text{driver_gene}}\mathcal{L}_{\text{driver_gene}} + \alpha_{\text{cell}}\mathcal{L}_{\text{cell}} - \alpha_{\text{discr}}\mathcal{L}_{\text{discr}},$$

where $\alpha_{\text{driver_gene}}$, α_{cell} and α_{discr} are hyperparameters that control the weights of the respective losses. This loss optimizes the CMM, DGP, and the cell label classifier jointly, in alternation with the optimization of the discriminator which minimizes $\mathcal{L}_{\text{discr}}$, as mentioned. Note that the CMM aims to maximize $\mathcal{L}_{\text{discr}}$ (hence the minus sign) with the help of the discriminator. This multi-objective setup promotes alignment across batches while maintaining robust biological representation in the data.

Baselines

SCENIC / SCENIC+. For each test dataset, SCENIC+ inferred a gene regulatory network (GRN), identified regulons $\mathbf{W}_r \in \mathbb{R}^{N_r \times N_g}$, and computed regulon activity $\mathbf{W}_a \in \mathbb{R}^{N_c \times N_r}$ in the cells, where N_r, N_g, N_c represent the number of identified regulons, genes, and cells in the test dataset, respectively. $\mathbf{W}_r$ is a learnt matrix containing the weights of genes for different regulons, and $\mathbf{W}_a$ indicates the regulon activities for each cell. Then we used $\mathbf{W}_g = \mathbf{W}_a\mathbf{W}_r$ to represent the regulatory importance of each gene in cells. Then, genes in each cell were ranked based on the value of their regulatory importance (elements in $\mathbf{W}_g$), as high value indicates a possible role in controlling cellular identity. We applied SCENIC+ to Norman et al. and Schmidt et al. datasets. Only genes appeared in perturbation pools in these datasets participated in the ranking based on W_g. Hyperparameters of GRN inference, regulon identification, and regulon activation were set to default. Cells with no regulon activated were removed from our analysis. SCENIC+ analysis was realized by `pyscenic 0.12.1`.

In silico perturbation. In silico perturbation (ISP) methods, such as GEARS, are capable of predicting transcriptomic outcomes of genetic perturbations. We trained GEARS model on the datasets mentioned in the corresponding tasks. For evaluation, we computed the cosine similarity between predicted transcriptomic profiles upon different perturbations, and that of cells from test datasets. Driver genes were predicted based on the similarities, and high values in similarity indicate the potential to be driver genes. GEARS analysis was realized by `cell-gears 0.1.1`. Data processing and training followed the data processing tutorial (https://github.com/snap-stanford/GEARS/blob/master/demo/data_tutorial.ipynb) and

training tutorials (https://github.com/snap-stanford/GEARS/blob/master/demo/model_tutorial.ipynb).

DGE analysis. Differential gene expression (DGE) analysis is the most frequently used method to reveal cell-type specific transcriptomic signature. Initially, cells from the test datasets were normalized and subjected to logarithmic transformation. Subsequently, we applied the Leiden algorithm, an unsupervised clustering method, to categorize the target cells into distinct groups. The number of clusters for each test dataset was set to range from 20 to 40, ensuring that cellular heterogeneity was maintained while providing a sufficient number of cells in each group for robust statistical analysis. We selected source cells to serve as a reference for comparison and performed DGE analysis on each target cell group against this reference. The Wilcoxon signed-rank test was employed to determine statistical significance. Then, significant genes were ranked according to their log-fold changes in expression as potential driver genes. Both unsupervised clustering and DGE analysis were conducted using the package `scanpy 1.9.6`.

The prior gene interaction graph

We constructed our prior gene interaction graph by combining a gene regulatory network and a cellular signaling network proposed by NicheNet. These networks were integrated to form a comprehensive representation of gene interactions across various tissues. In our approach, a connection was established between two genes if they were linked in either of the individual networks. The resulting integrated graph features 33,354 genes, each represented by a unique human gene symbol, and includes 8,452,360 edges that signify the potential interactions. The unweighted versions of NicheNet networks were used in our approach. For each cell, we remove the gene nodes with values of 0 in the raw count matrix of the single-cell transcriptomic profile, to construct the cell type-specific gene graph.

Datasets

Human cell atlas (HCA). We downloaded all single-cell and single-nucleus datasets sourced from contributors or DCP/2 analysis in Homo sapiens up to March 2023, accumulating approximately 1.5 TB of raw data. We retained all experiments that included raw count matrices and standardized the variables to gene names using a mapping list obtained from Ensembl (https://www.ensembl.org/biomart/martview/574df5074dc07f2ee092b52c276ca4fc).

Norman et al. This dataset (GEO accession number GSE133344) measures transcriptomic consequences of CRISPR-mediated gene activation perturbations in K562 cell line. We filtered this dataset, removing cells whose counts number is lower than 3500. After filtering, this dataset contains 105 perturbations targeting different genes, and 131 double

perturbations targeting two genes simultaneously. We used unperturbed cells (with non-targeting guide RNA) as source cells and perturbed cells as target cells.

Schmidt et al. This dataset (GEO accession number GSE190604) measures the effects of CRISPR-mediated activation perturbations in human primary T cells under both stimulated and resting conditions. For our analysis, we excluded cells not mentioned in metadata and removed genes appeared in less than 50 cells. Gene expression levels of single-guide RNA were deleted in order to avoid data leakage. We used unperturbed cells as source cells and perturbed cells as target cells. We also excluded cells without significant changes after perturbation following the procedure proposed by `Mixscape` tutorial via package `pertpy`. Default parameters were used for `Mixscape` analysis.

Cano-Gamez et al. This dataset (EGA accession number EGAS00001003215) comprises naïve and memory T cells induced by several sets of cytokines. With cytokine stimulation, T cells are expected to differentiate into different subtypes. We took cells not treated by cytokines as source cells and cytokine-stimulated cells as target cells. This experimental set reflects the differential process of human T cells. For our analysis, clusters 14-17 were excluded as their source cell could hardly be decided.

Fernandes et al. This dataset comprises heterogeneous dopamine neurons derived from human iPSC. These neurons were exposed to oxidative stress and ER stress, representing PD-like phenotypes. We followed preprocessing procedures as mentioned in the original GitHub repo (https://github.com/metzakopian-lab/DNscRNAseq/blob/master/preprocessing.ipynb).

Tian et al. This dataset comprises iPSC-derived neurons perturbed by more than 180 genes related to neurodegenerative diseases. CRISPR inference experiments with single-cell transcriptomic readouts were conducted by CROP-seq. For our analysis, we removed genes appeared in less than 50 cells. We used unperturbed cells as source cells and perturbed cells as target cells.

Srivatsan et al. This dataset (GEO accession number GSE139944) contains transcriptomic profiles of human K562, A549, and MCF7 cell lines perturbed by 188 different small molecules under four dosages. Specifically, in the analysis of Srivatsan et al., chemical and mechanistic properties of histone deacetylases (HDAC) inhibitors were explored. For our study, we reserved K562 cell line cells perturbed by HDAC inhibitors. We used unperturbed cells as source cells, and chemically perturbed cells as target cells. This set represents the process of cellular transition caused by drugs.

Kowalski et al. This dataset (GEO accession number GSE269600) measures the transcriptional consequences of CRISPR-mediated perturbations in HEK293FT and K562

cells. A series of cleavage and polyadenylation regulators were knocked down. Cells with minimal perturbation effects were removed from downstream analysis. For our analysis, we excluded perturbations which consist of less than 200 cells. We used cells from control group as source cell, and perturbed cells as target cells.

T cell differentiation analysis

Identification of cellular phenotypic shift. We computed 'transition score' to identify cellular phenotypic shift on the transcriptomic level. We selected canonical marker genes associated with IFNG and IL2 secretion, and Th2 differentiation. Then, we computed transition score based on the mean expression level of these marker genes. That is:

$\mathrm{CTS}_{ij} = \frac{1}{K_i}\sum_{k_i=1}^{K_i}(g_{j_1k_i} - g_{j_0k_i})$, where CTS_{ij} is the transition score of phenotypic shift type i in source-target cell pair j, $g_{j_1k_i}$ and $g_{j_0k_i}$ are normalized gene expression levels of marker gene k for cell-state transition type i in target cell j_1 and its source cell j_0. The total number of maker genes for phenotypic shift type i is represented with K_i. Then, classes of phenotypic changes were annotated based on transition score. Transition scores are calculated via function `tl.score_genes` from `scanpy` package with default parameters.

Cell type classification with predicted driver genes. We selected a series of genes related to transition mentioned above from previous studies (**Table S1**). We used the term 'likelihood scores' to describe the probability of a gene to be a driver factor predicted by the model, that is:

$$\mathrm{LS}_{jk}^{mod} = p_{jk}^{mod}$$

where LS_{jk}^{mod} means the likelihood score for gene k in source-target cell pair j from model mod, and p_{jk}^{mod} represents the probability predicted by model mod. For our analysis, the mod could be CellNavi, or baseline models.

Then, we aggregate likelihood scores into 'prediction scores' to evaluate the performance of different models:

$$\mathrm{PS}_{ij}^{mod} = \frac{1}{m_i}\sum_{k_i=1}^{m_i}\mathrm{LS}_{jk}^{mod}$$

where PS_{ij}^{mod} is the prediction score of cell-state transition type i in source-target cell pair j predicted by model mod. The number of candidate driver genes for each phenotypic changing type i is m_i. Ideal prediction should reflect similar patterns as shown by cellular transition score mentioned above. To evaluate it quantitatively, we trained decision tree classifiers with prediction scores as input to test whether predictions scores would faithfully

demonstrate cell-state transition types. Classifiers were trained for each method independently, and 10-fold cross-validation was conducted. Classifiers were implemented via shallow decision trees using the `sklearn` package.

Gene Ontology enrichment analysis

We used gene ontology (GO) enrichment analysis to explore drugs' mechanism of actions. For each drug compound, the top 50 genes with highest scores predicted by CellNavi were used for GO enrichment analysis. The significant level was chosen to be 0.05, and the Benjamini-Hochberg procedure was used to control the false discovery rate. For implementation, we used package `goatools` for GO enrichment analysis.

Molecular docking

We performed molecular docking for panobinostat and tucidinostat, with a reference protein structure obtained from the PDB entry 3MAX. The ligand structures from PDB entries 3MAX and 5G3W were used to guide the initial placement of panobinostat and tucidinostat, ensuring the pose correctness of the warheads and major scaffolds. Based on such initial poses, local optimizations were performed with AutoDock Vina. PyMol was used for structure visualization.

Foundation models

scGPT. We utilized the recommended whole-human pre-trained model (https://drive.google.com/drive/folders/1oWh_-ZRdhtoGQ2Fw24HP41FgLoomVo-y?usp=sharing) provided by scGPT and fine-tuned it using datasets provided in the scGPT documentation. The fine-tuning process adhered to the default parameters outlined in the annotation tutorial (https://github.com/bowang-lab/scGPT/blob/main/tutorials/Tutorial_Annotation.ipynb).

Geneformer. For the cell type annotation task, we fine-tuned the Geneformer model using datasets provided in the scGPT documentation (https://github.com/bowang-lab/scGPT/tree/main/data) without any parameter frozen. The fine-tuning process followed the official guidelines provided in the Geneformer documentation (https://huggingface.co/ctheodoris/Geneformer/tree/main/examples). The pretrained model used in this task was downloaded from Geneformer repository (https://huggingface.co/ctheodoris/Geneformer/blob/main/model.safetensors).

Data Availability

HCA data for Cell Manifold Model training is downloaded from the HCA data explorer (https://explore.data.humancellatlas.org/projects). The Norman et al.[29], Tian et al[45]. and Srivatsan et al.[51] datasets were downloaded from the scPerturb project[72] on Zenodo (https://zenodo.org/records/10044268). The raw count data of Schmidt et al.[30] dataset was downloaded from the National Institutes of Health GEO with accession number GSE190604, and its metadata was downloaded from Zenodo (https://zenodo.org/records/5784651). The Cano-Gamez et al.[31] dataset was downloaded from the Open Target Platform of this project (https://www.opentargets.org/projects/effectorness). The Fernandes et al.[43] dataset was downloaded from ArrayExpress with accession number E-MTAB-9154. The preprocessed Kowalski et al[60] dataset was downloaded was downloaded from Zenodo (https://zenodo.org/records/7619593#.Y-P7Zi1h2X0).

Code availability

https://github.com/DLS5-Omics/CellNavi

Acknowledgements

We thank Dr. Zeyu Chen, Dr. Sida Shao and Zeyu Tang for constructive suggestions and feedback on the manuscript and Tie-Yan Liu for the supervision. We thank Jingyun Bai for graphic designs and illustrations.

Author Contributions

Conceptualization, P.D., T.W., S.Z., and C.L.; Data curation, T.W. and P.D.; Methodology, Y.P., F.J., T.W., S.Z., C.L., and P.D.; Model implementation – pre-train, F.J., Y.P., G.L., and H.X.; Model implementation – fine-tuning and inference, Y.P., F.J., T.W., and C.L.; result analysis, T.W., Y.P., F.J., P.D., Y.M., and X.L., result interpretation, T.W., P.D., Y.P., F.J., Y.M.; Writing-original draft, P.D., T.W., C.L., S.Z., F.J., Y.M., and Y.P.; Writing-revision, P.D., C.L., S.Z., H.L., Y.P., H.X.; Supervision, H.L.. All authors have read and approved the manuscript.

Competing interest

P.D., F.J., S.Z., C.L., Y.M., H.X., G.L., H.L., and T.L. are paid employees of Microsoft Research. The remaining authors declare no competing interests.

Supplementary materials

Table S1: Driver genes related to cytokine secretion and Th2 differentiation

Gene	Function	Reference
CD28	IL2 secretion, IFNG secretion	doi: 10.1126/science.abj4008 doi: 10.4049/jimmunol.1500707 https://www.nature.com/articles/ni0101_37
VAV1	IL2 secretion, IFNG secretion	doi: 10.1126/science.abj4008 https://www.nature.com/articles/374474a0 https://doi.org/10.4049/jimmunol.177.8.5024
CD27	IFNG secretion	doi: 10.1126/science.abj4008 https://doi.org/10.1016/j.immuni.2024.01.011
IL9R	IFNG secretion	doi: 10.1126/science.abj4008 https://www.nature.com/articles/s41586-022-04801-2
GATA3	Th2 differentiation	doi: 10.1126/science.abj4008 https://doi.org/10.1016/j.cell.2018.11.044 https://www.nature.com/articles/ncomms2260

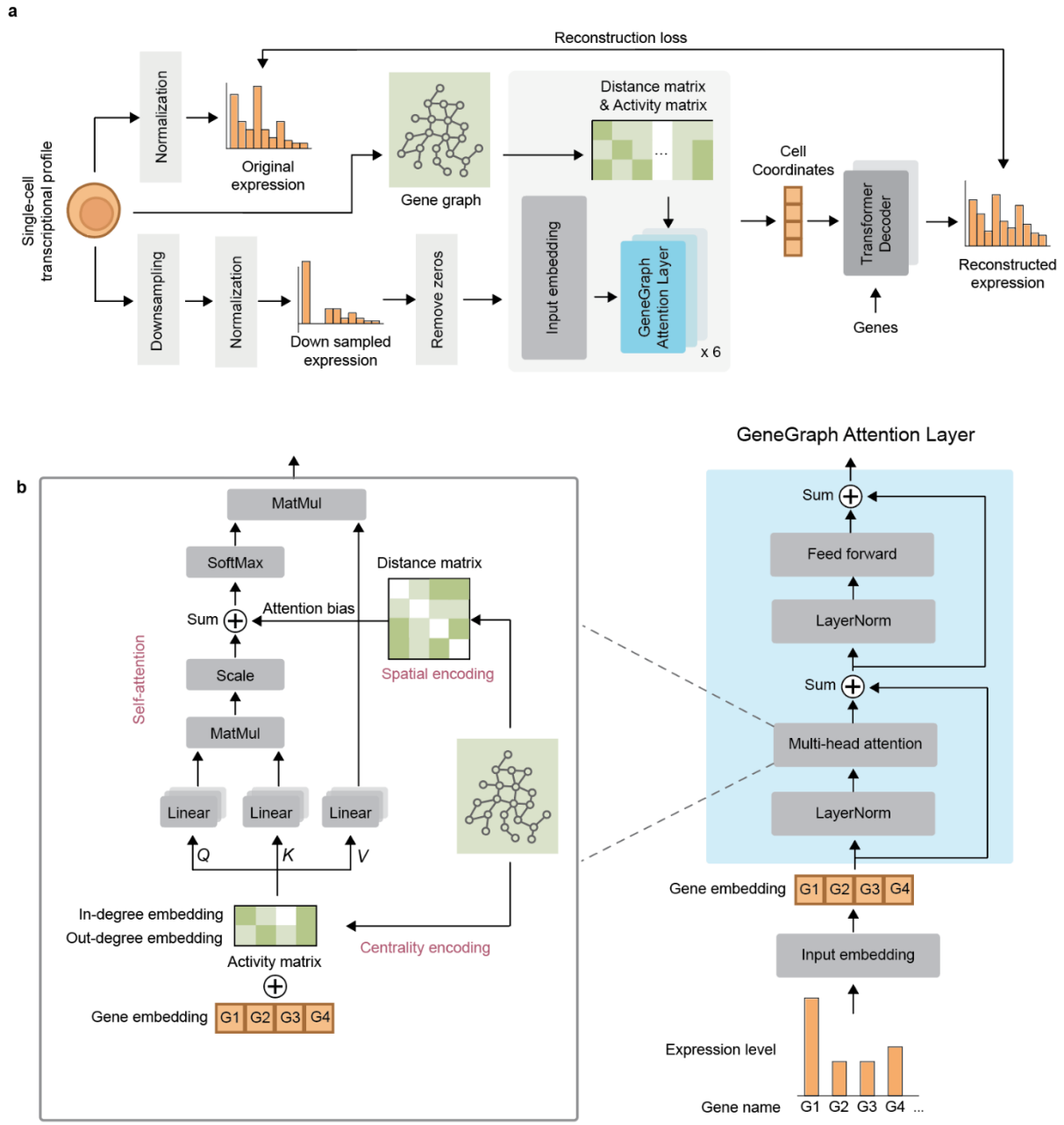


Figure S1.1. Details of Cell Manifold Model (CMM) implementation. a) CMM is designed to reconstruct gene expression profiles by leveraging a Transformer variant composed of GeneGraph Attention Layers. **b)** GeneGraph Attention Layer integrates prior gene graph via its multi-head attention layer.

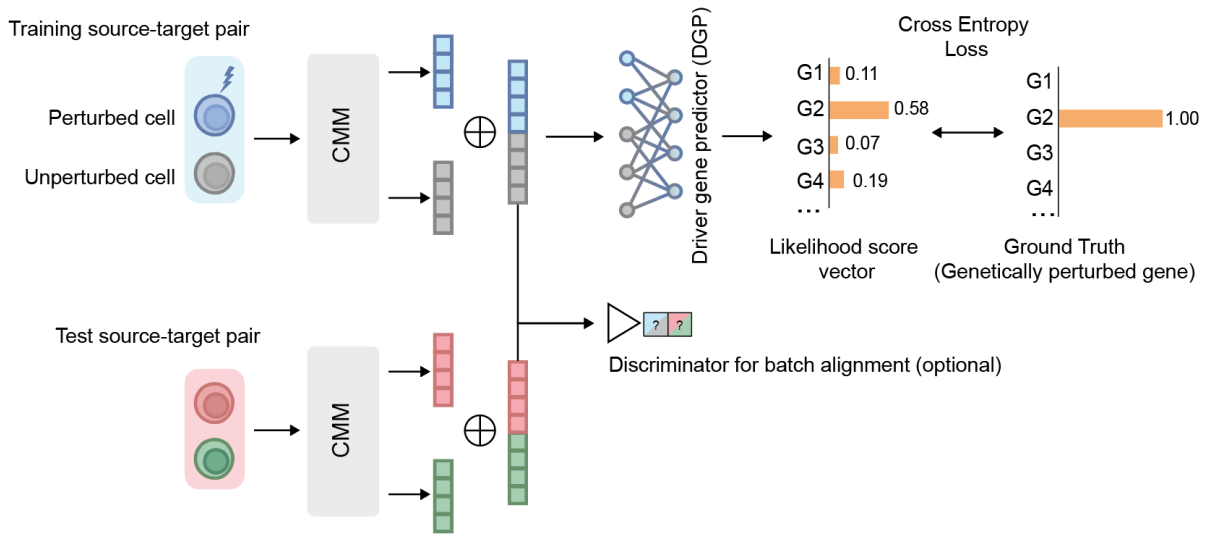


Figure S1.2. Details of Driver Gene Predictor (DGP) implementation. The DGP takes concatenated cell coordinates pairs output from the CMM to predict driver genes by generating a likelihood score vector. An optional discriminator is introduced to align training and test data, ensuring consistency and accuracy in predictions. See Methods for details.

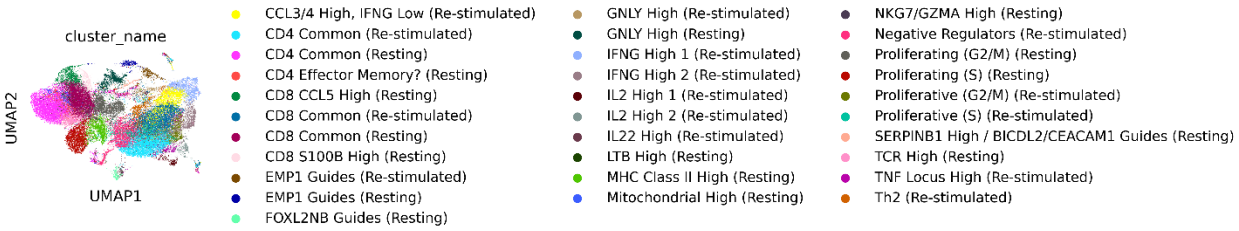


Figure S2.1. UMAP visualization of perturbed resting T cells and restimulated T cells sequenced by Schmidt et al, colored by cell clusters identified in the original study.

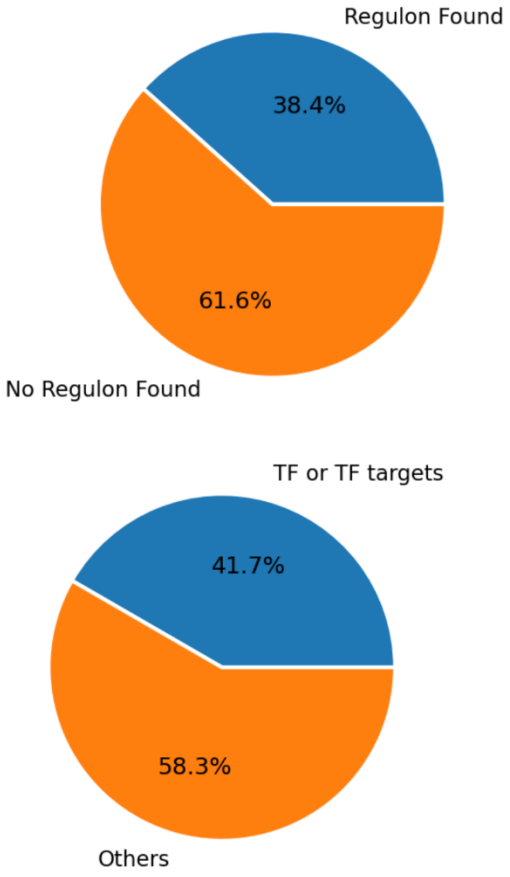


Figure S2.2. Alignment of test dataset to SCENIC/SCENIC+. Upper panel: 61.6% samples cannot be predicted by SCENIC/SCENIC+ because of missing regulons. Bottom panel: 58.3% of candidate driver genes are not included because they are neither transcription factor (TF) nor TF-target genes.

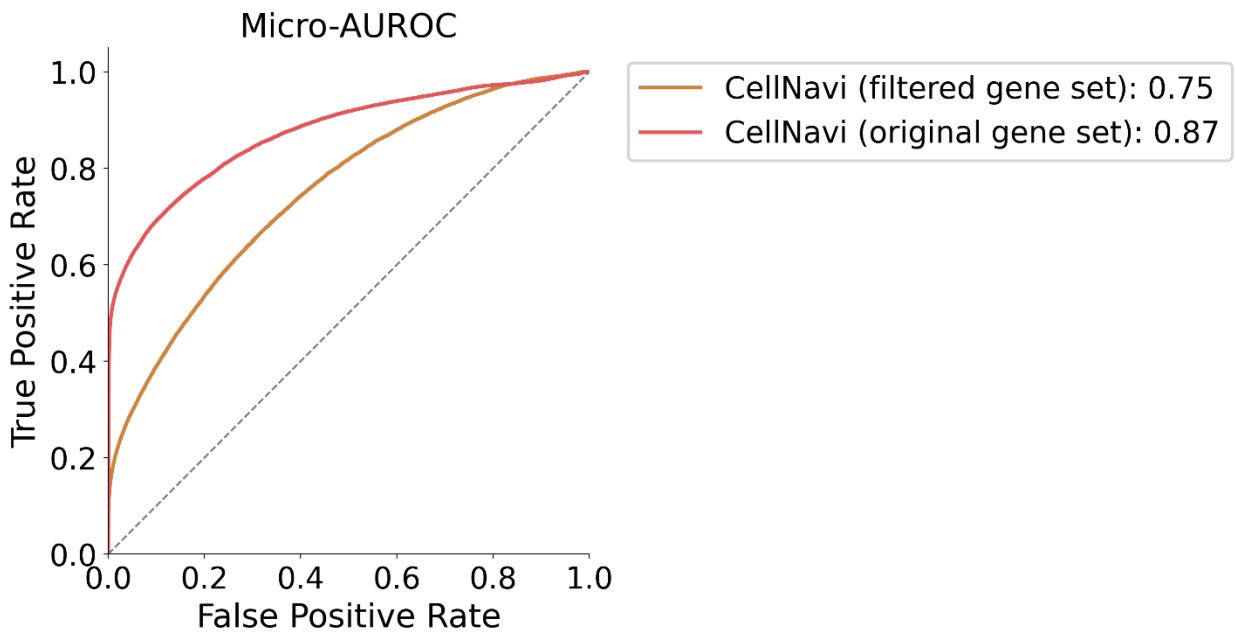


Figure S2.3. Micro-AUROC of CellNavi applied to the original single-cell transcriptomic profile and the filtered transcriptomic profile with perturbed genes excluded.

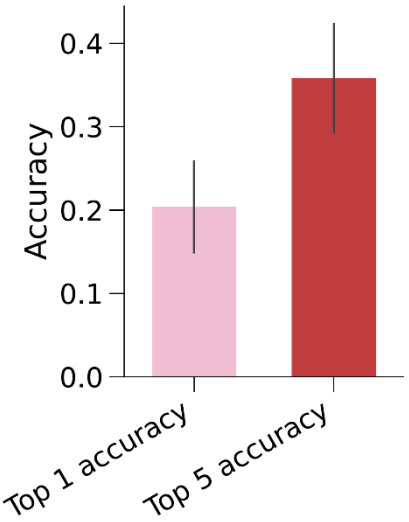


Figure S2.4. Top K (K = 1 or 5) accuracy of CellNavi on transcriptomic data with perturbed gene removed. Error bar, standard error.

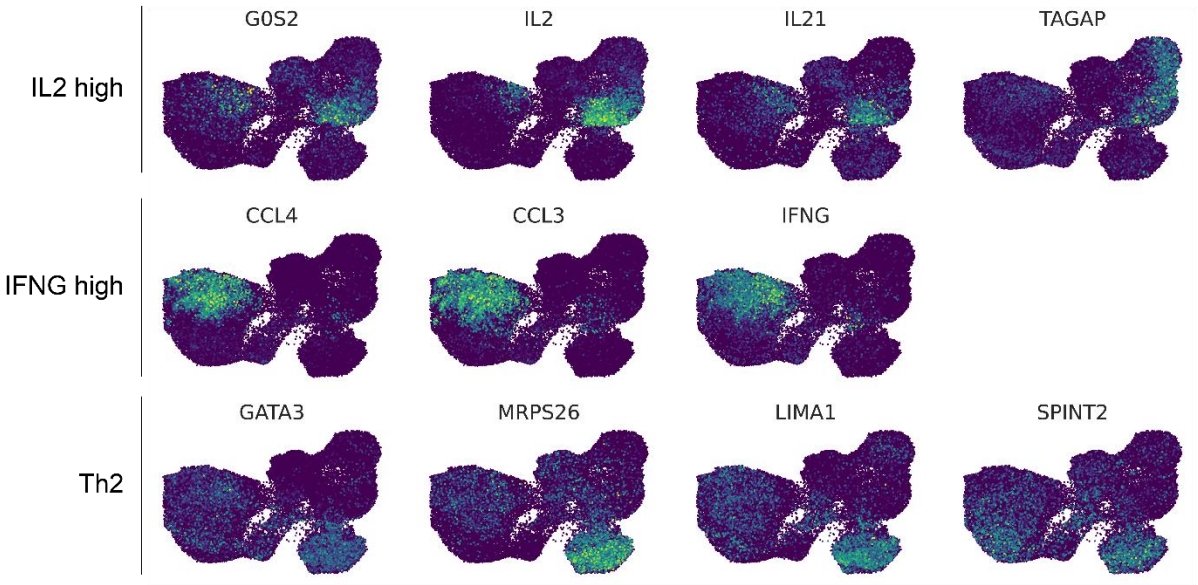


Figure S3.1 Transcriptional changes of canonical marker genes in IL2-high, IFNG-high, and Th2 cell types. IL2-high marker genes**:** G0S2, IL2, IL21, and TAGAP. IFNG-high marker genes: IFNG, CCL3, CCL4, and CCL3L3. Th2 marker genes: GATA3, MRPS26, LIMA1 and SPINT2.

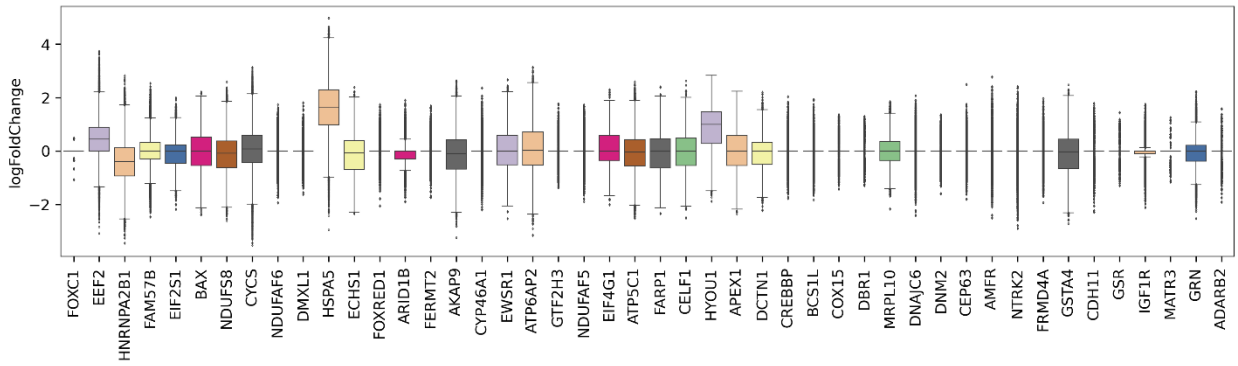


Figure S4.1 Expression changes for top-20 predicted genes across cell pairs. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers.

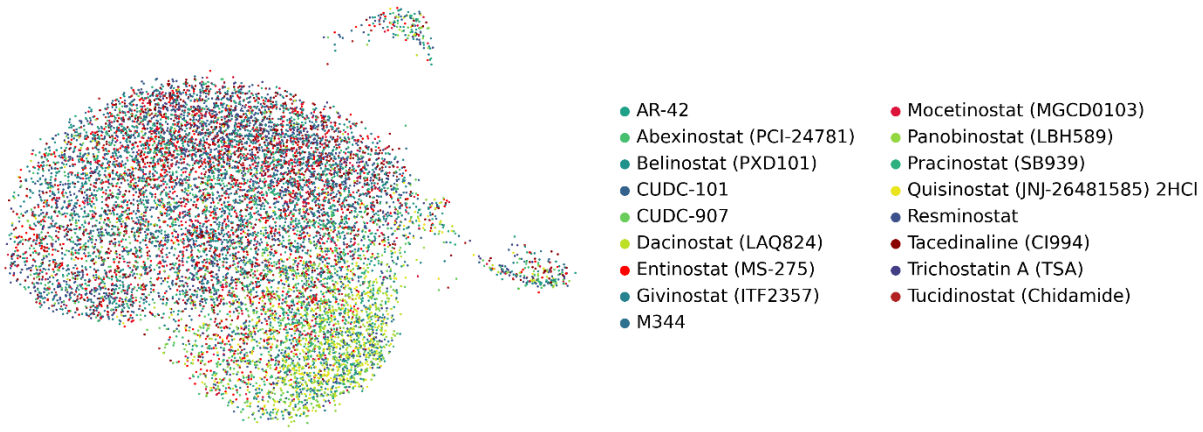


Figure S5.1. UMAP visualization of transcriptomic profiles in K562 cells treated with 17 different HDAC inhibitors

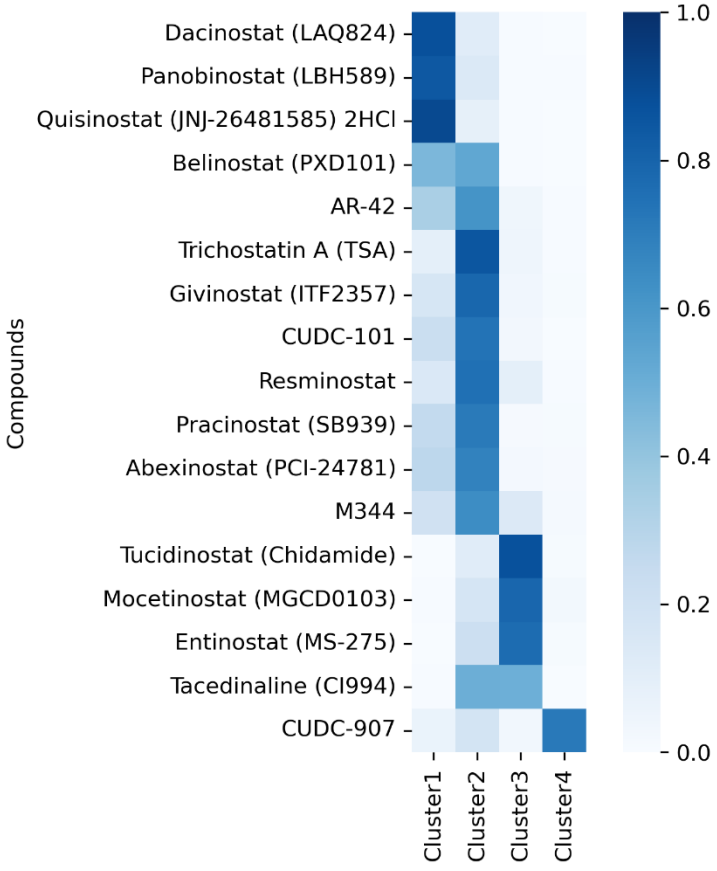


Figure S5.2 Percentage of cells treated with different drug compounds within each identified cluster as shown in Figure 5a.

Figure S5.3 GO enrichment for each treatment group

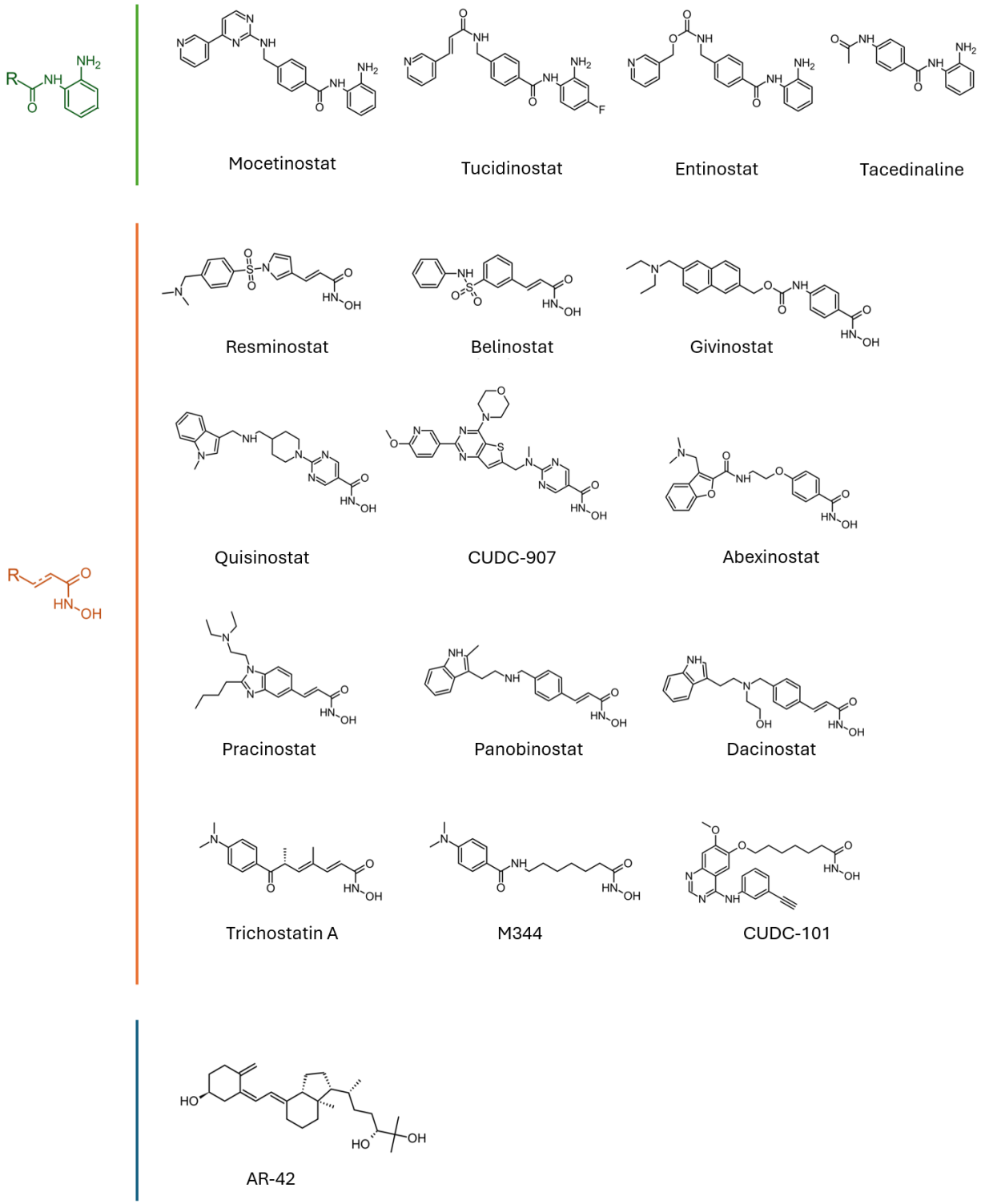


Figure S5.4 Molecular structures of the 17 HDAC inhibitors. Except from AR-42, other molecules share 2 different types of warheads.

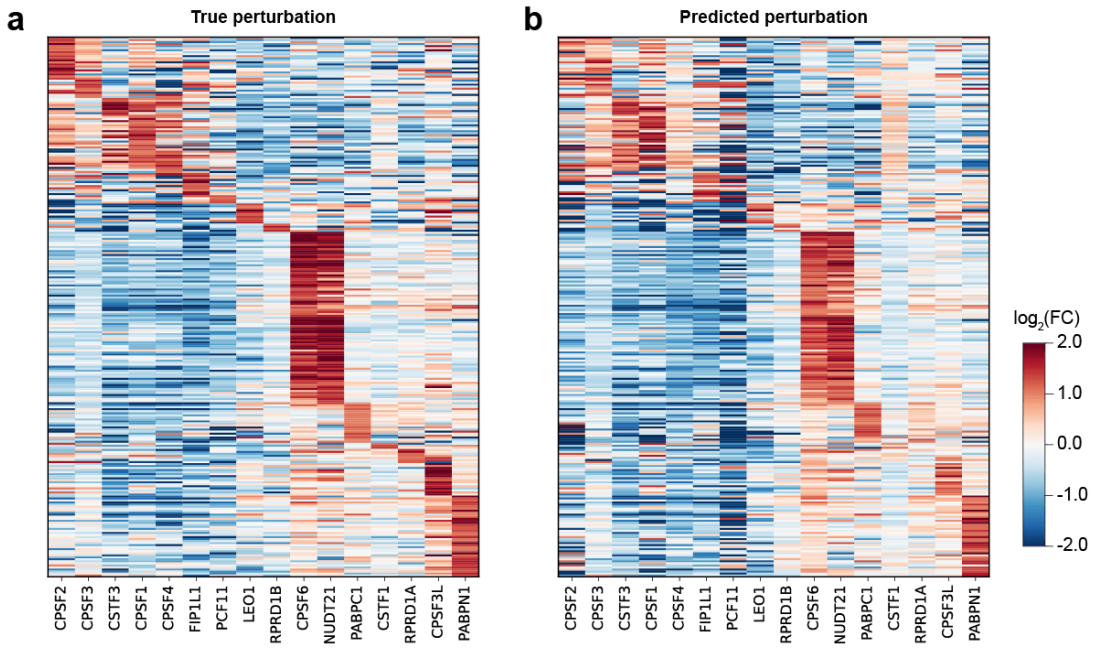


Figure S6.0 Heatmap showing gene expression in different perturbation groups for K562 cells. Rows represent genes, and columns represent true perturbations **(a)** or predicted perturbations by CellNavi **(b)**.

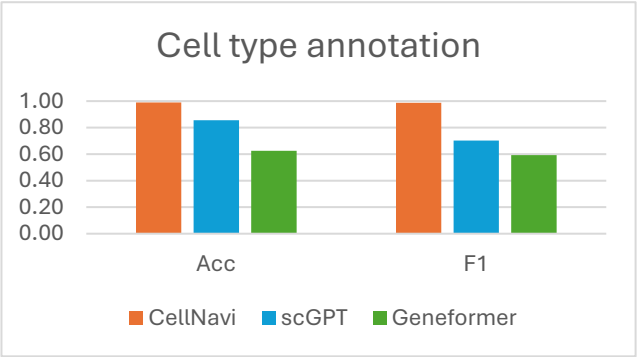


Figure S6.1 Accuracy and Macro F1 scores for cell type annotation task. CellNavi was benchmarked with scGPT and Geneformer.

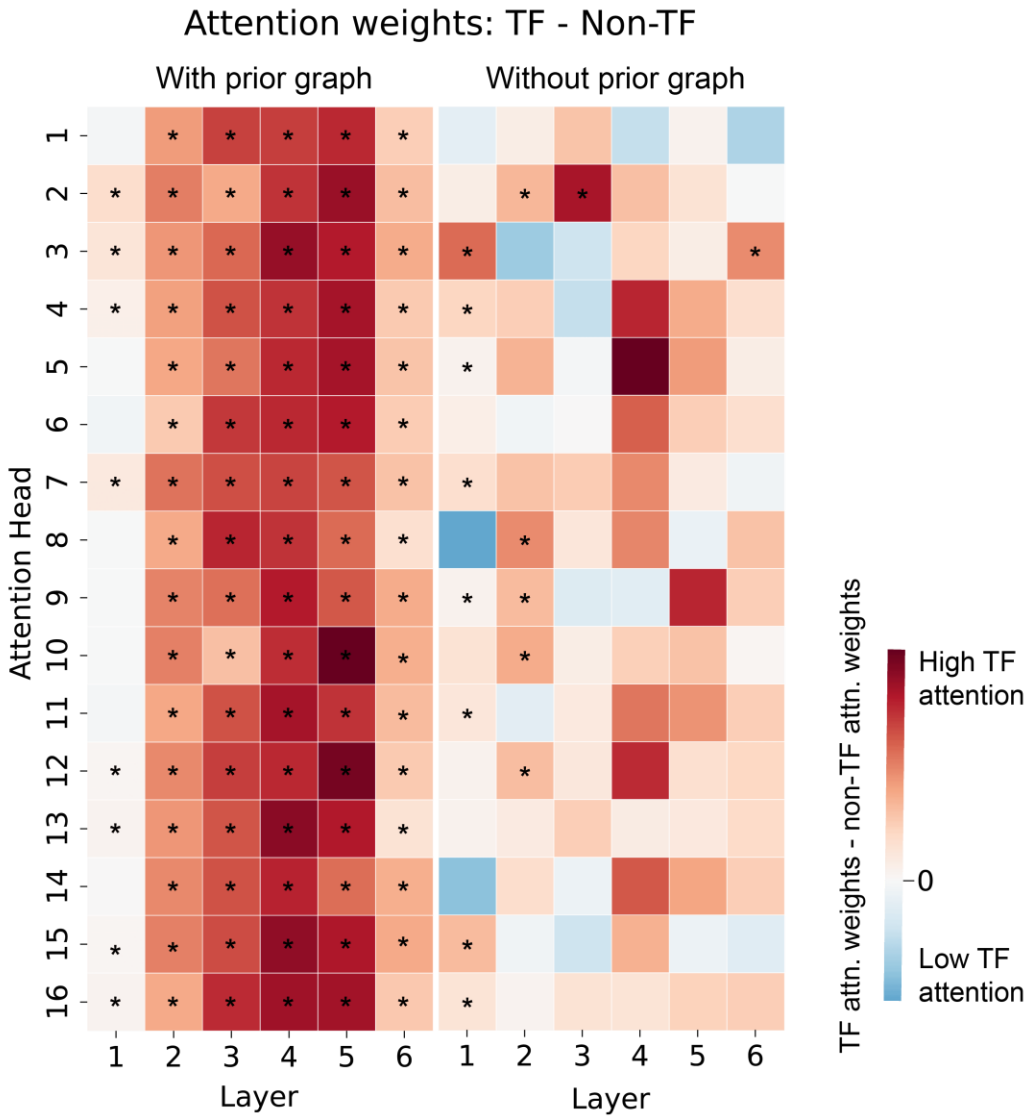


Figure S6.2. Differential attention weights between TFs and non-TFs. The heatmap displays the contrast in attention weights directed towards transcription factors (TFs) versus non-transcription factors (non-TFs) when the prior gene-gene interaction graph is included (left) or excluded (right). The color gradient represents the magnitude of the difference in attention weights. In the presence of the gene-gene graph, 88 attention heads demonstrate a statistically significant bias towards TFs, while without the graph, 15 attention heads show this preference ($P < 0.05$, Wilcoxon rank sum test, with false discovery rate (FDR) adjustment using the Benjamini–Hochberg procedure).

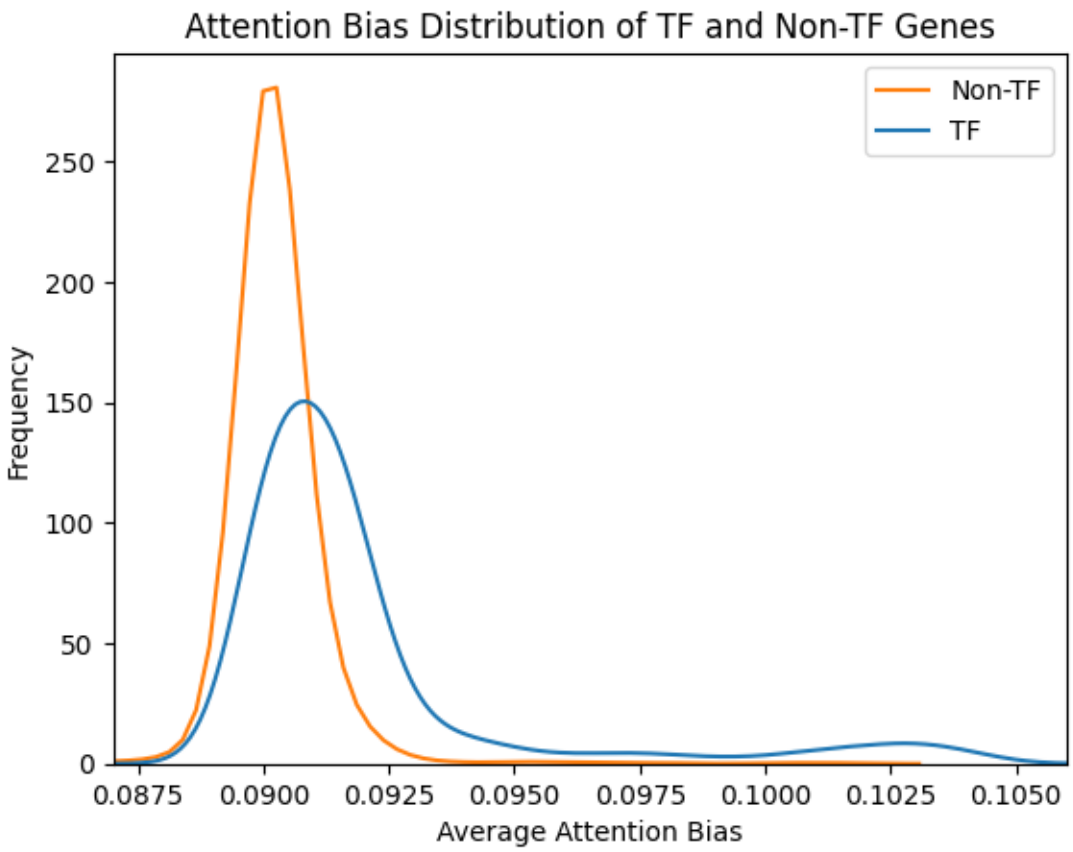


Figure S6.3. The attention bias learned from the gene graph exhibited a notably higher magnitude for TFs compared to non-TFs in distribution ($P < 0.05$, Wilcoxon rank sum test). To aid visualization, the sample sizes for both TFs and non-TFs have been equalized through random sampling.