

SuperFeat: Quantitative Feature Learning from Single-cell RNA-seq Data Facilitates Drug Repurposing

Jianmei Zhong ^{1,#}, Junyao Yang ^{2,#}, Yinghui Song ^{3,#}, Zhihua Zhang ³,
Chunming Wang ¹, Renyang Tong ¹, Chenglong Li ³, Nanhui Yu ³, Lianhong Zou ³,
Sulai Liu ³, Jun Pu ¹, Wei Lin ^{1,*}

¹State Key Laboratory for Oncogenes and Related Genes, Department of Cardiology, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai Cancer Institute, Shanghai 200127, China

²Department of Laboratory Medicine, Xin Hua Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200092, China

³Central Laboratory of Hunan Provincial People's Hospital/The First Affiliated Hospital of Hunan Normal University, Changsha 410005, China

*Corresponding author: weilin.baylor@gmail.com (Lin W).

#Equal contribution.

Handling Editor: Jin Gu

Abstract

In this study, we devised a computational framework called Supervised Feature Learning and Scoring (SuperFeat) which enables the training of a machine learning model and evaluates the canonical cellular statuses/features in pathological tissues that underlie the progression of disease. This framework also enables the identification of potential drugs that target the presumed detrimental cellular features. This framework was constructed on the basis of an artificial neural network with the gene expression profiles serving as input nodes. The training data comprised single-cell RNA sequencing datasets that encompassed the specific cell lineage during the developmental progression of cell features. A few models of the canonical cancer-involved cellular statuses/features were tested by such framework. Finally, we illustrated the drug repurposing pipeline, utilizing the training parameters derived from the adverse cellular statuses/features, which yielded successful validation results both *in vitro* and *in vivo*. SuperFeat is accessible at <https://github.com/weilin-genomics/rSuperFeat>.

Key words: Single-cell transcriptomics; Cell state transition; Cell scoring; Drug search; Feature learning.

Introduction

Since the emergence of various high-throughput omics technologies such as microarray [1] and next-generation sequencing (NGS) [2], researchers have been generating vast amounts of molecular profiling datasets of the biological and clinical samples at an unprecedented rate. The single-cell sequencing technique has even added another degree of magnitude and a new dimension of information [3–6]. Many models and algorithms have been developed to evaluate the biological signal and characterize the sample. Nonetheless, most of the existing methods are hardly generalizable but specific to particular problems [7,8]. A universal framework for the rapid yet generic assessment of the biological/clinical samples using molecular profiling data could be very useful and efficient, especially for the omics assays at high granularity.

Human learning involves the processes of classification and quantification. The artificial neural network (ANN) has been used as a flexible framework to simulate and streamline such processes using a machine/computer. It provides a relatively simple but generalizable classification and quantification model to automatically structure the human knowledge acquired from a large body of data. The ANN nodes of the neural network reflect the qualitative or the quantitative state of a biological concept in the human mind. Most importantly, subsequent human decisions, such as clinical decisions, could be made based on the evaluation of the cellular state. Therefore, the ANN structure could be used to realize the automatic learning and evaluation of biological statuses/

features and thus facilitate making decisions more efficiently based on the input of a high volume of data from the high-throughput omics techniques. This mechanism is becoming increasingly powerful and can accomplish many unprecedented tasks, even in the biomedical field.

Previously, we developed the ANN-based cell type classifier framework, SuperCT [9]. It uses the single-cell RNA sequencing (scRNA-seq) digital expression profiles as input to characterize the canonical cell lineages. Similar cell type classification strategies have been published since then [10,11]. This work represents an early attempt to apply ANNs to high-throughput single-cell gene expression data, offering a general interface for the now prevalent deep learning models. In a fully connected neural network for a cell type classifier, the node weight reflects the empirical contribution of a gene to a certain cell type, which also makes this classifier interpretable.

In the meantime, a certain cell type could undergo a spectrum of variable cellular states reflected by a few function-related and signaling-related gene expression patterns. For instance, T cells turn into an exhaustion state in most solid tumors [12]; macrophages undergo polarization in certain biological contexts [13,14]. Nonetheless, assessing the variability of such cellular state by a couple of markers using the current single-cell transcriptomics data based on barcoded-bead (BCB) oftentimes is not robust. This is due to the stochastic nature of the RNA transcription within a single cell [15] and the limitation of detection for a specific molecule by

Received: 31 August 2023; Revised: 8 April 2024; Accepted: 8 May 2024.

© The Author(s) 2024. Published by Oxford University Press and Science Press on behalf of the Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

the current BCB capture [4]. Unless we perform a more comprehensive assessment using a group of coordinately expressed genes, the state of a single cell might not be accurately evaluated. The good news is that a certain cellular status/feature is always associated with a group of up-regulated and down-regulated genes, which thus can be reflected by a scoring strategy based on the presence/absence of a group of transcriptions.

The strategy of activating or reversing a certain cellular state of the disease-involved cell type could be employed for possible therapeutics. For example, immune checkpoint inhibition has been used in cancer therapy. It serves the purpose of reversing the T cell exhaustion [16]. Targeting helper T cellular state conversion could be used in mitigating autoimmunity [17]. Ideally, if small molecules could be used to reverse the adverse cellular state or promote the beneficial cell state. Such an idea was implemented in the Connectivity Map (CMap) bulk (but not single-cell) transcriptomics datasets. Cell lines have been used to generate the CMap perturbation transcriptomics data repository, which was used to search for repurposing drugs [18,19]. Nonetheless, prior to the application of the scRNA-seq technique to clinical samples, the CMap search was oftentimes compromised by the convoluted signals of a mixture of cell types of different roles. The high-resolution transcriptomics data like scRNA-seq allow to characterize the target cell population and thus make the CMap search more specific. signatureSearch provided a flexible tool to use ranked gene set to search a perturbation database that could be used to achieve such goal [20].

In addition to the scRNA-seq, a similar BCB-based RNA quantification strategy is applied in the 10X Genomics Visium spatial transcriptomics (ST) platform. Obviously, the cellular status/feature evaluation strategy could be applicable to this type of data and provide a very user-friendly spatial visualization of the histological slides. The landscape of the cellular statuses/features will be smoothed even with the stochastic transcriptions of the functional genes on the spots.

In this study, to efficiently perform the learning and quantitative assessment of a variety of biological features documented in the literature, we built an ANN-based framework, Supervised Feature Learning and Scoring (SuperFeat). This framework can facilitate the thorough evaluation of the critical features of the tissue samples or even the precise diagnoses of the patients with adequate details. Considering that gene set-based methods such as Seurat modular score [21], AUCell [22], singscore [23], and gene set variation analysis (GSVA) [24] can somewhat do a similar job, we performed an extensive performance comparison to understand the strength of SuperFeat.

Method

Organization and preprocessing of the training and validation datasets

The datasets for feature training are summarized in Table S1. The cell populations with the designated cellular statuses/features came from the unsupervised clustering, and were enriched with the canonical markers for the specific status/feature. The annotations were conducted by the authors of the original papers. For the training dataset, we only included the cell types in which the statuses/features were supposed to be exhibited in the model training. A total of 19,202 genes (Table S2) were taken into the input, which were derived from the total genes of the MSigDB v7.4.1 gene sets of subcategories H and C5

[25]. The missing expression values of these 19,202 genes in the training or validation dataset were filled up with zero. The gene expression values of the digital expression matrix were transformed to a binary matrix.

Implementation of the SuperFeat ANN

The ANN structures and the learning strategy were implemented using Python library keras v2.9.0 and sklearn v1.0.2. We designed a fully connected ANN model for cell type classification. Similar to what has been done for SuperCT, the inputs were transformed into the binary values of 19,202 genes. As seen in most of the flow cytometry analyses, the presence/absence of the signature gene partially contributes to the status/feature of the cell. Also, the binary signal input is compatible across most of the unique molecular index (UMI)-based scRNA-seq platforms with the robust performance of the cell type classification. The code could be found at https://github.com/weilin-genomics/rSuperFeat/blob/main/train_new_model/SuperFeat_trainingCode.py.

The input layer was connected to a hidden dense layer with only one hidden layer neuron using rectified linear unit (ReLU) activation functions. The single node in the hidden layer represented that a neuron received the signal from 19,202 inputs by different weights, only if the combined stimulation by these input signals was above a certain level, it determined the probability of the cell either in state0 cell or state1, by outputting two nodes. The input layer had the L1 regularization with a coefficient of 0.01. To avoid the under-representation of the small-sample-size cell types in the calculation of the accuracy function, we included the class weight based on the sample size of each type in the model training. The loss function was defined as categorical cross-entropy.

Evaluation of the designated cellular features using SuperFeat score

The cellular status/feature was evaluated using the following equation:

$$\text{SuperFeat score} = \sum_{n=1}^{19202} w_i \cdot g_i + b \quad (1)$$

where w_i denotes the weight value of gene i from the model trained from the training dataset to the first layer, g_i denotes whether gene i is detected in this cell or not, and b denotes the bias value for the single node of the first hidden layer.

Drug search using weight-ranked genes

In this study, we utilized the qSig function from the signatureSearch package v1.8.2 to conduct connectivity ranking, drawing on the perturbational cell gene expression signatures within the CMap and LINCS L1000 databases [26]. The top 250 genes, both positively and negatively weighted, were identified using the printTopWeights function from the rSuperFeat package and subsequently input into “signatureSearch” to identify potential drugs. Our methodology incorporated reference databases from both CMap and LINCS, employing their respective gene expression signature search (GESS) methods.

Drug search was predicated on cumulative positive or negative connectivity scores, indicating their potential to either induce or counteract shifts in cellular states or features. For LINCS, we prioritized drug–cell type pairs using the weighted

connectivity score (WTCS); while for CMap, ranking was based on the raw score metric. If the cellular transition is beneficial, we should prioritize the drug with larger positive scores. Conversely, if the cellular transition is detrimental, we should prioritize the drug with larger negative scores. When using “signatureSearch” for drug searching, different scores were employed due to the differences of two perturbation databases, and the statistical assumptions satisfied by the analysis signals varied.

Unsupervised clustering, dimensional reduction, and data visualization in scRNA-seq analyses

For the scRNA-seq dataset downloaded from Gene Expression Omnibus (GEO) database, the unsupervised clustering, dimensional reduction, and data visualization in this study were realized by a widely used scRNA-seq analytical suite, Seurat v4.1.0. The Seurat objects were generated for each dataset with their digital expression matrices as input. The principal component analysis (PCA) was performed by Seurat RunPCA function. The *t*-distributed stochastic neighbor embedding (*t*-SNE) coordinates were calculated using Seurat RunTSNE function. The uniform manifold approximation and projection (UMAP) coordinates were calculated using Seurat RunUMAP function. The putative clusters were defined by Seurat FindClusters function using the top 10 principal components and other default parameters. If the cell annotation was provided by authors in the database, we will directly use their annotation. If not, the unsupervised clusters were re-annotated according to the enriched literature markers. The receiver operating characteristic (ROC) analysis was implemented with pROC package v1.18.0 [27].

Cell culture

The GC cell line MKN-45 was purchased from the Cell Bank of Type Culture Collection of the Chinese Academy of Sciences and cultured in Dulbecco's Modified Eagle Medium (DMEM; Catalog No. 21063029, Gibco, Carlsbad, CA) with 10% phosphate-buffered saline (PBS) and antibiotics (100 IU/ml penicillin and 100 µg/ml streptomycin). The cell culture was placed in humidified air at 37°C with 5% CO₂/95% air (v/v).

In vivo subcutaneous tumor generation and drug treatment

For *in vivo* studies, 4–6-week-old male BALB/c nude mice (Shanghai Laboratory Animal Center, Shanghai, China) were housed in a controlled environment with a 12-h light/12-h dark cycle, with free access to water and food at temperature of 21°C–23°C and humidity of 40%–60%.

Low-passage MKN-45 cells were resuspended in a 1:1 mixture of PBS and Matrigel (Catalog No. 356231, Corning, NY) at 1×10^6 cells/ml. Then, 100 µl of cell stock was injected subcutaneously on the shaved right flank of BALB/c nude mice. After 10 days of growing, the tumor volume increased to an average size of 60 mm³. WH-4023 (Catalog No. S7565, Selleck, Shanghai, China) was dissolved in dimethyl sulfoxide (DMSO; Catalog No. D2650, Sigma-Aldrich, Saint Louis, MO), and then injected into the tumors every other day for 6 days at the concentration of 0.5 mg/kg body weight. The same volume of DMSO was included as control. The mice were examined at regular time points until they were sacrificed. The tumor size was measured using a digital caliper, and the tumor volume was calculated with the following formula: volume = $0.5 \times \text{width}^2 \times \text{length}$.

ST dataset

We conducted an in-depth ST study on samples from two cases of intrahepatic cholangiocarcinoma (iCCA) using the Visium technology from 10X Genomics. These slides were simultaneously subjected to hematoxylin and eosin (H&E) staining, and different histopathological regions were independently annotated by two distinct pathology experts for various stained areas.

Tissue processing

Tumors from two iCCA cases were washed with PBS, and any liquid was soaked up with gauze. Tissues were cut into pieces of 6.5 mm³, snap-frozen in isopentane cooled with liquid nitrogen, and stored at –80°C.

Slide preparation

Slides for ST had four 6.5 × 6.5 mm² areas with 5000 bar-coded primer spots. Tissue sections were cut at 10-µm thickness and placed on specific slides from 10X Genomics.

Fixation, staining, and imaging

Sectioned slides were fixed, stained with H&E, and mounted with glycerol. Bright-field images were taken at 20× magnification and processed with VSlide software [28].

Reverse transcription and sequencing

The protocol followed was as described using the Visium platform of 10X Genomics. Sequencing was performed on an Illumina NovaSeq 6000, and initial data analysis was conducted with Space Ranger software [29], mapping to the GRCh38_release 95 human genome.

Immunofluorescence staining

The tumor tissues were harvested and fixed in 4% paraformaldehyde, embedded in paraffin, followed by cryosection with a thickness of 5–10 µm. The cryosections were fixed with 4% paraformaldehyde, permeabilized with 0.1% Triton X-100 in PBS for 15 min, and blocked with 10% fetal bovine serum (FBS)/PBS for 1 h. Then, the cryosections were stained with the primary antibodies: anti-CD34 (Catalog No. ab81289, Abcam, Waltham, MA), anti-PI16 (Catalog No. PA5-111740, Invitrogen, Carlsbad, CA), anti-COL1A1 (Catalog No. PA5-29569, Invitrogen), anti-α-SMA (Catalog No. ab124964, Abcam). The primary antibodies were diluted in 10% FBS/PBS by the dilution factor recommended by the suppliers, applied to the samples, and incubated at 37°C for 1.5 h or at 4°C overnight. The secondary antibodies were diluted at 1:1000 in 10% FBS/PBS, applied to the samples, and incubated at 37°C for 45 min. The cell nucleus was counterstained with 4',6-diamidino-2-phenylindole (DAPI) at room temperature for 5 min, and coverslips were mounted on slides with fluorescent mounting media.

Results

An efficient framework of feature learning and scoring using single-cell transcriptomics data

We constructed a fully connected ANN, SuperFeat, to learn the quantitative features with one single node in the hidden layer and two output nodes. The value of the middle-layer node will be used for the quantitative evaluation of the cellular status/feature. Figure 1 illustrates the workflow of the entire framework.

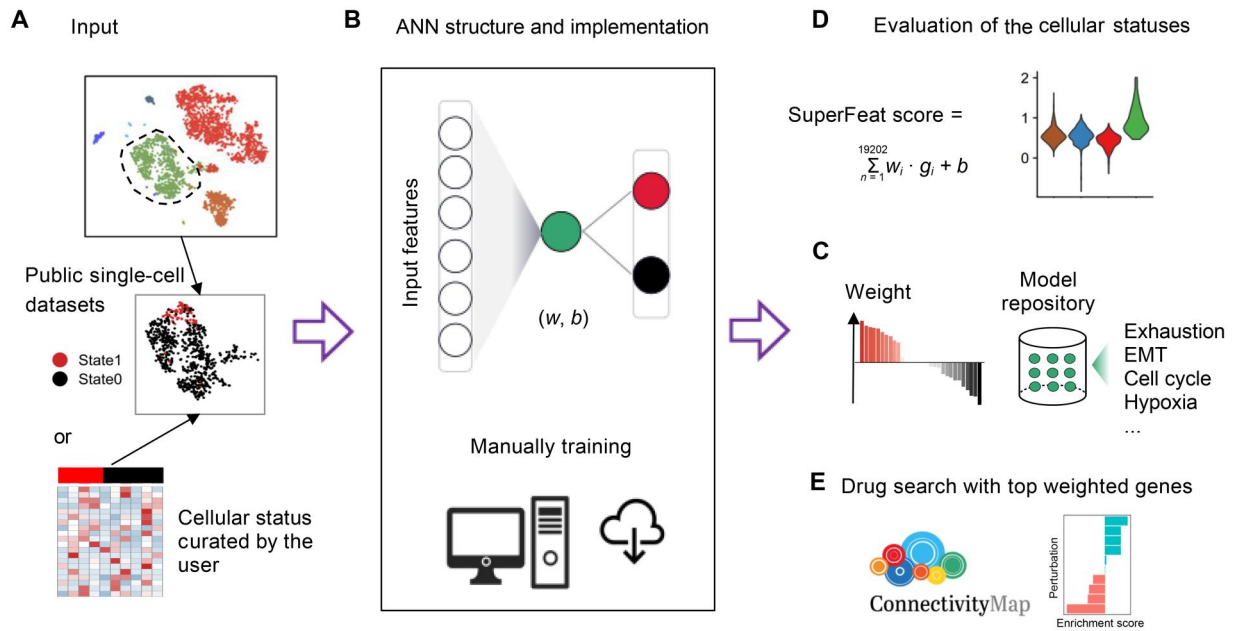


Figure 1 Workflow of SuperFeat framework

A. A diagram showing how the cell lineage with a variable cellular status/feature is defined in the training dataset. **B.** ANN structure and implementation. **C.** The parameters of the models trained from dataset and the repository of models for the canonical cellular statuses/features. **D.** Evaluation of the cellular statuses/features using SuperFeat model and visualization. **E.** Drug search using the model parameters. scRNA-seq, single-cell RNA sequencing; EMT, epithelial-mesenchymal transition; ANN, artificial neural network.

Figure 1A shows a single-cell dataset containing multiple cell types. If one of the cell types (circled by the dashed line) was annotated with a variable cellular status/feature (according to some signature genes) and can be manually divided into two annotated subpopulations, this variable cell population could be used to train the SuperFeat model for evaluating this cellular status/feature. If there are more datasets with the similar cellular state annotation, we should include the cell populations from more datasets to enhance the representativity and the model generalizability. To enable users to customize their own model, we also provided the Python code (Figure 1B) for those who likes to train their own cellular status/feature model and to assess similar transcriptomics data. Figure 1B also shows the structure of the ANN, and the green node basically provides a score to evaluate how much the transcriptional profile should be categorized to state1 (in red) or state0 (in black).

Figure 1C shows the weights of the genes contributing to the score in this ANN model. Figure 1D shows how the model could be used to calculate the score of the cellular status/feature of the cells from another dataset of scRNA-seq or spatial RNA sequencing (RNA-seq) in the applicable biological context. This part of application was implemented in an R package rSuperFeat (<https://github.com/weilin-genomics/rSuperFeat>). Later, as shown in Figure 1E, the model parameters can be further utilized to search for the CMap [18] perturbagen that potentially enhances or alleviates the stress associated with cellular state change.

Training and evaluation of four tumor hallmark cellular statuses/features

Using our most recent version of SuperFeat training framework and the training dataset from a study of kidney renal clear cell carcinoma (KIRC) [30], we first performed the training of the T cell exhaustion model parameters and

applied the scoring model to assess the T cells. The exhaustion scores allowed us to discern the exhausted and active T/natural killer cell populations that were correlated to the immunotherapy. The distributions of the exhaustion scores of the annotated populations of T cells in training dataset are shown in Figure 2A. The layout of these cells is shown in Figure S1. The canonical marker signals of exhaustion annotated by Neal et al. [30] are shown in Figure 2B, confirming the overall concordance. Then we used the datasets including the infiltrating T cells from hepatocellular carcinoma (HCC) samples published by Zheng et al. [31] to test the model performance. The C4_CD8-LAYN population stood out (Figure 2C), which was concordant to the cluster interpretation by Zheng and colleagues [31]. The layout of the cell populations in the validation datasets is shown in Figure S2.

Epithelial-mesenchymal transition (EMT) is another example of the variable cellular state that oftentimes occurs in the epithelial cells under wound healing, organ fibrosis, and especially initiation of the metastasis of cancer progression. Here, we used a set of pancreatic ductal adenocarcinoma (PDAC) single-cell data with EMT annotation in the tumors [32] to train a EMT model. We used the EMT population from high-grade serous ovarian tumor [33] to test the performance of this model. The scores of the training cells are shown in Figure 2D. The C7 cluster, which was annotated as “EMT/mesenchymal-like” population, exhibited higher EMT signals. Both show a concordance with the published annotation and the canonical EMT markers (Figure 2E). The scores of the testing cells are shown in Figure 2F.

Hypoxia refers to a state in when the cells are deprived of adequate oxygen. Tumors oftentimes develop such a hypoxic feature. We used Darmanis et.al. dataset [34] to train the hypoxia model and used Neftel et.al. dataset [35] to evaluate

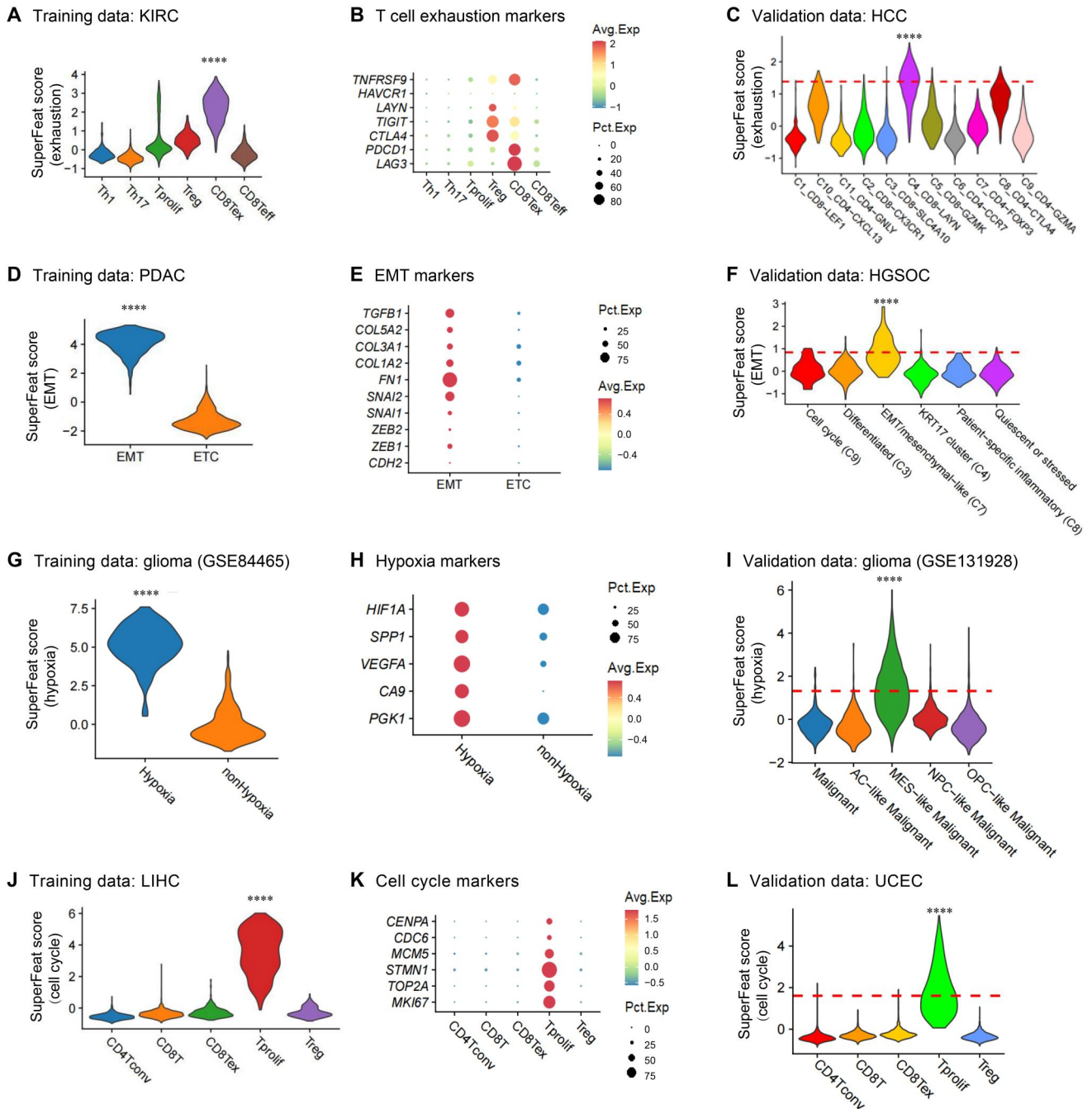


Figure 2 The tumor-related cellular statuses/features defined by SuperFeat

A. T cell exhaustion SuperFeat scores on training data. **B.** Canonical T cell exhaustion markers. **C.** SuperFeat scores validated by a HCC dataset. **D.** EMT SuperFeat scores on training data. **E.** Canonical EMT markers. **F.** SuperFeat scores validated by a HGSOC dataset. **G.** Hypoxia SuperFeat scores on training data. **H.** Canonical hypoxia markers. **I.** SuperFeat scores validated by a glioma dataset. **J.** Cell cycle SuperFeat scores on training data. **K.** Canonical cell cycle markers. **L.** SuperFeat scores validated by a UCEC dataset. All *P* values were determined by Wilcoxon test (****, *P* < 0.0001). KIRC, kidney clear cell carcinoma; HCC, hepatocellular carcinoma; PDAC, pancreatic ductal adenocarcinoma; HGSOC, high-grade serous ovarian cancer; LIHC, liver hepatocellular carcinoma; UCEC, uterine corpus endometrial carcinoma.

the feature. The results are shown in Figure 2G–I. The annotated hypoxic cell populations (the MES-like malignant cluster) in the validation dataset gave high SuperFeat scores and exhibited canonical hypoxic gene expression such as *HIF1A* and *VEGFA*.

The proliferation rate is essential to the development of tissue. For example, the immunohistochemistry (IHC) signature

Ki-67 (with gene symbol *MKI67*) has been commonly used to evaluate cell proliferation by pathologists. Here, we also trained a model of proliferation [36] and tested this model performance [37]. Figure 2J shows the proliferation scores in the training dataset, and Figure 2K shows the transcriptional signals of the canonical cell cycle genes. The proliferation scores of the testing cells are shown in Figure 2L.

SuperFeat model parameters uncover involved known and novel features

The neural network used to be considered a black-box learning machine. As our framework has a very simple network structure with a limited number of nodes in the middle layer and the input genes are trained independently, the contribution of the input nodes (genes) can be easily evaluated by the weight values. Therefore, the top positively/negatively weighted genes could be used to perform an enrichment analysis to understand how they underlie the cellular state change.

For the T cell exhaustion model trained by the Neal et al. dataset [30], the canonical marker genes such as *PDCD1*, *LAG3*, and *HAVCR2* showed high-rank weights with no surprise (Figure 3A). The enrichment analysis based on the top positively weighted genes of T cell exhaustion hit the “regulation of cell activation” (Figure 3B), while the enrichment analysis based on the top negatively weighted genes were enriched for “lymphocyte activation”. Obviously, they were two polarized states of T cells.

For the EMT model trained by the Lin et al. dataset [32], the top positively weighted genes included *SERPINE1*, *VIM*, *SNAI2*, and *ANGPTL4*, which were also expected (Figure 3C). The top positively weighted genes exactly hit the “hallmark of EMT” (Figure 3D). The top negatively weighted genes of EMT hit the “endopeptidase activity”, which has never been discussed and probably needs extra attention.

For the hypoxia model trained by the Darmanis et al. dataset [34], *VEGFA*, *CAVI*, and *CA9* emerged at the top positively weighted genes (Figure 3E), which were canonical hypoxia-related genes. The top positively weighted genes exactly hit the “hallmark of hypoxia”, which again validated the reliability of our framework (Figure 3F). Interestingly, the enrichment analysis based on the top positively weighted genes of hypoxia also hit the “hallmark of glycolysis”. This correlation has been reported in a previous study [38]. Such correlations, as byproducts, won't be naturally investigated when employing a canonical gene set-based scoring system to evaluate the cellular state.

For the cell cycle model trained by the Zhang et al. dataset [36], *MKI67*, *STMN1*, *MCM4*, and *TOP2A* showed up at the top positively weighted genes (Figure 3G), which were canonical cell cycle-related genes. These top positively weighted genes hit the terms such as “cell division” and “cell cycle checkpoint” (Figure 3H).

In summary, the top positively weighted genes are mostly involved in the pathways that drive the development of the corresponding cellular statuses/features, and the top negatively weighted genes are involved in signature pathways of the opposite cellular states. It makes the ANN-based model parameters interpretable.

Other than the canonical genes that have been included in the corresponding enriched terms, there are some highly weighted genes that have been mentioned in previous literature, suggesting that their roles in the cells exhibiting the corresponding features are corroborated by two independent studies, which is non-trivial. These genes are listed in Table S3. Notably, some of the genes with the top weights have not been previously reported in the literature. While these may represent artifacts or confounders, they hold the potential to be as significant as those canonical markers and those documented in Table S3. For comprehensive scrutiny, we listed both positively and negatively weighted genes associated with

the trained features in Table S4, available for any researchers inclined toward further exploration.

SuperFeat outperforms gene set-based scoring methods in discerning the target subpopulations

Prior to SuperFeat, people oftentimes used the gene set-based scoring methods, such as Seurat modular score [21,39,40], AUCCell [22], singscore [23], and GSVA [24], to evaluate the cellular state based on a group of canonical genes. To efficiently discern the cell subpopulations of a certain cell type that reflect the variability of a certain cellular status/feature, the separation of the signal distributions of the scores suggests a better performance. As shown in the heatmaps of median signal scores evaluated by SuperFeat, Seurat, AUCCell, singscore, and GSVA in the four validation datasets (Figure 4A–D), it appeared that SuperFeat scores made the target subpopulation stand out with better color contrast than others. We further used the Kolmogorov-Smirnov (K-S) statistics to numerically evaluate how much the subpopulations with certain cellular statuses/features stand out from their counterpart subpopulations without this feature. The results showed that SuperFeat mostly outperformed other gene set-based scoring systems by the relatively larger K-S D values and larger area under curves (AUCs) (Figure 4E–L), except being slightly lower than the Seurat scores for EMT gene set. More detailed results of the accuracy analysis are shown in Table S5. Canonical marker genes for all the specific cellular states are listed in Table S6.

Overall, it is confident to claim that SuperFeat discerns the clusters annotated with the designated cellular statuses/features more easily and accurately than other gene set-based scoring methods. Furthermore, the advantage of SuperFeat scoring resides in its independency of the canonical gene set, which saves the arbitrariness of gene selection.

In addition, SuperFeat also showed comparable performance with logistic regression (Figure S3). However, SuperFeat exhibited superior universality and flexibility as a neural network model and offered variability such as the number of hidden layers and connection structures, providing more possibilities for future efficient and artificial general intelligence.

SuperFeat scoring in ST study

The SuperFeat scoring model can also be applied to the ST data, which allows us to correlate the cellular statuses/features to the histology. We used the two 10X Genomics Visium slides from the tumor samples of a cholangiocarcinoma cancer patient to demonstrate the mapping of the proliferative signals. Figure 5A shows the original H&E staining of the two slides from the same tumor tissue. Figure 5B shows the annotation of the pathological regions of the two slides determined by an experienced pathologist. Figure 5C shows the landscape of the proliferative signals on the two slides. Figure 5D shows the violin plots of the proliferative SuperFeat scores of the two slides. From the two replicate slides of the same tissue, the high cell cycle signal was reproducibly enriched in the immune regions, suggesting the high proliferative potential. We also mapped the signals of the two signature genes, *MKI67* and *TOP2A*, on the 10X Genomics Visium slides (Figure 5E and F). The landscape of the two genes' signals appeared more stochastic than the SuperFeat proliferation signal, suggesting the advantage of a comprehensive evaluation. The other three tumor-related feature scores are shown in Figure S4.

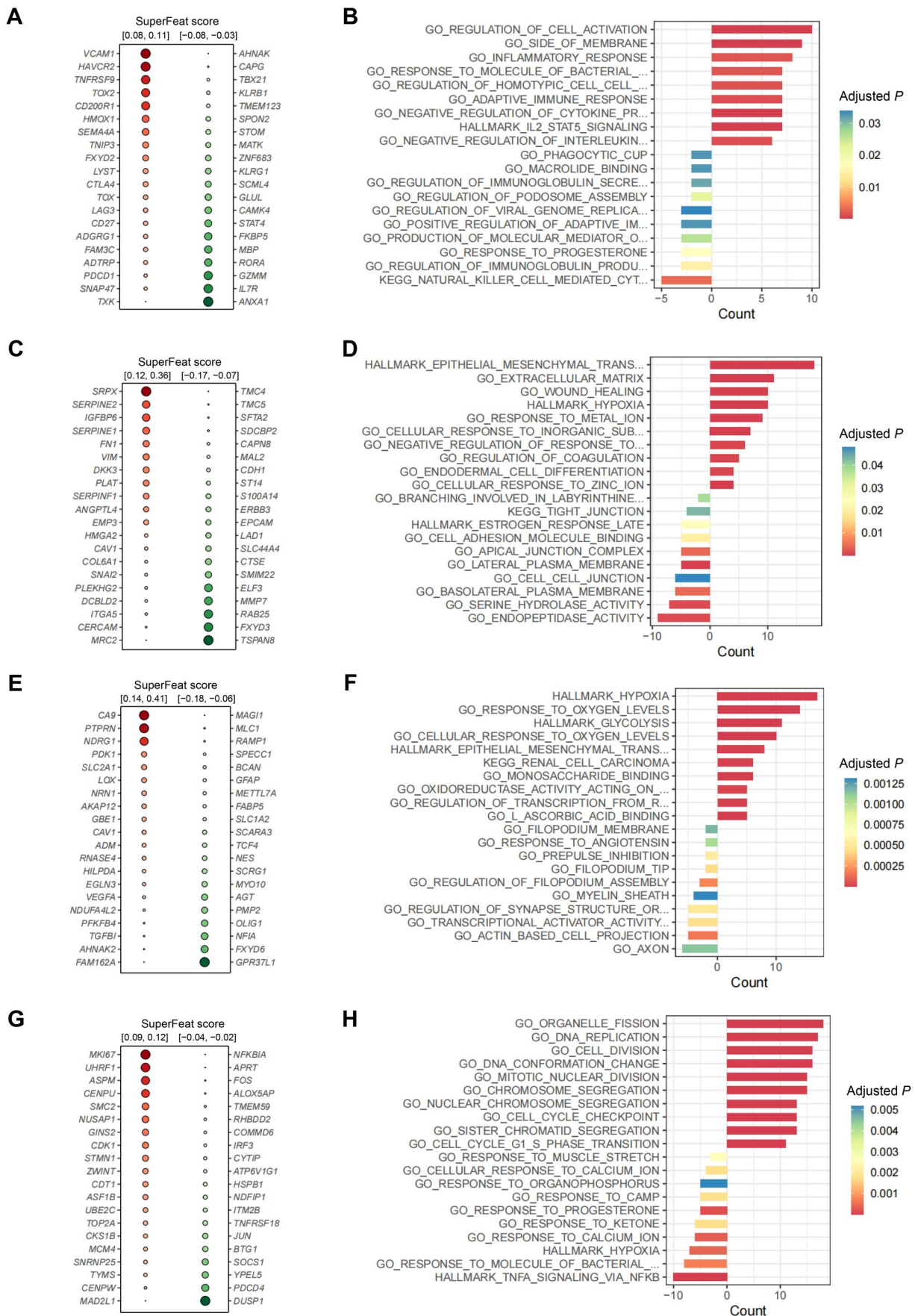


Figure 3 Interpretation of SuperFeat model parameters

A. Top 20 positively/negatively weighted genes of T cell exhaustion model. **B.** Enrichment analysis on the top 50 positively/negatively weighted genes of T cell exhaustion model. **C.** Top 20 positively/negatively weighted genes of EMT model. **D.** Enrichment analysis on the top 50 positively/negatively

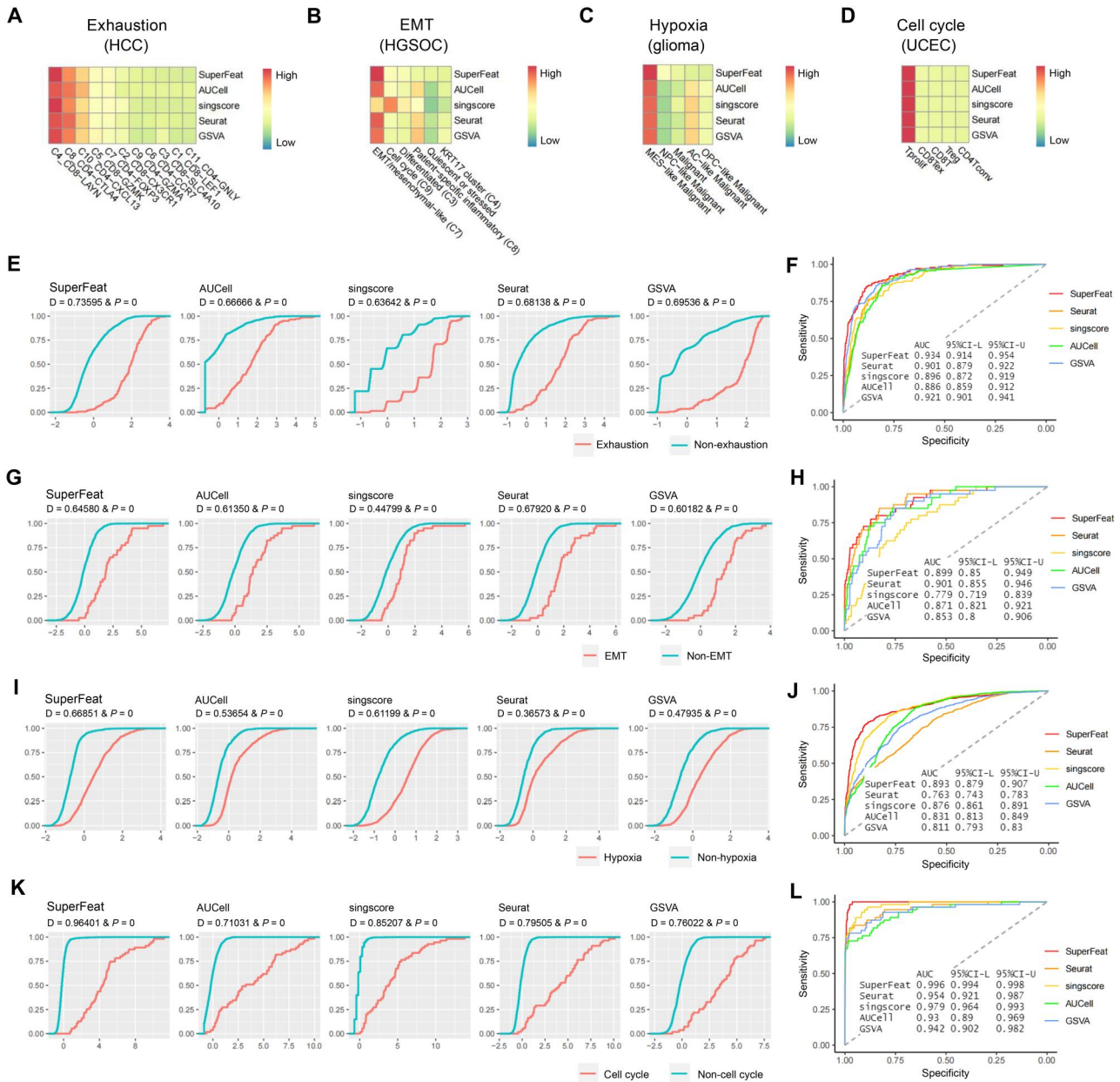


Figure 4 Comparison of cellular status/feature scoring methods

Heatmaps for median signal scores evaluated by SuperFeat, AUCell, singscore, Seurat, and GSVA for T cell exhaustion (A), EMT (B), hypoxia (C), and cell cycle (D). E. K-S statistics on T cell exhaustion state. F. ROC analysis on T cell exhaustion state. G. K-S statistics on EMT state. H. ROC analysis on EMT state. I. K-S statistics on hypoxia state. J. ROC analysis on hypoxia state. K. K-S statistics on cell cycle state. L. ROC analysis on cell cycle state. K-S, Kolmogorov-Smirnov; ROC, receiver operator characteristic curve; AUC, area under curve; GSVA, gene set variation analysis; CI, confidence interval; CI-L, the lower bound of CI; CI-U, the upper bound of CI.

Repurposing drug search using SuperFeat model parameters

As the malignant cellular states such as tumor proliferation and EMT could be targeted for cancer therapy, similar to CMap strategy, we were able to perform the search based on

the gene weights in the trained cellular state model against the perturbation databases such as CMap and LINCS L1000. Without confounding the variability derived from the heterogeneous composition of the cell populations in the samples, we intuitively believed that the drug search using the top

Figure 3 Continued

weighted genes of EMT model. E. Top 20 positively/negatively weighted genes of hypoxia model. F. Enrichment analysis on the top 50 positively/negatively weighted genes of hypoxia model. G. Top 20 positively/negatively weighted genes of cell cycle model. H. Enrichment analysis on the top 50 positively/negatively weighted genes of cell cycle model. Negative counts indicate enrichments for negatively weighted genes and the positive counts indicate positively weighted ones.

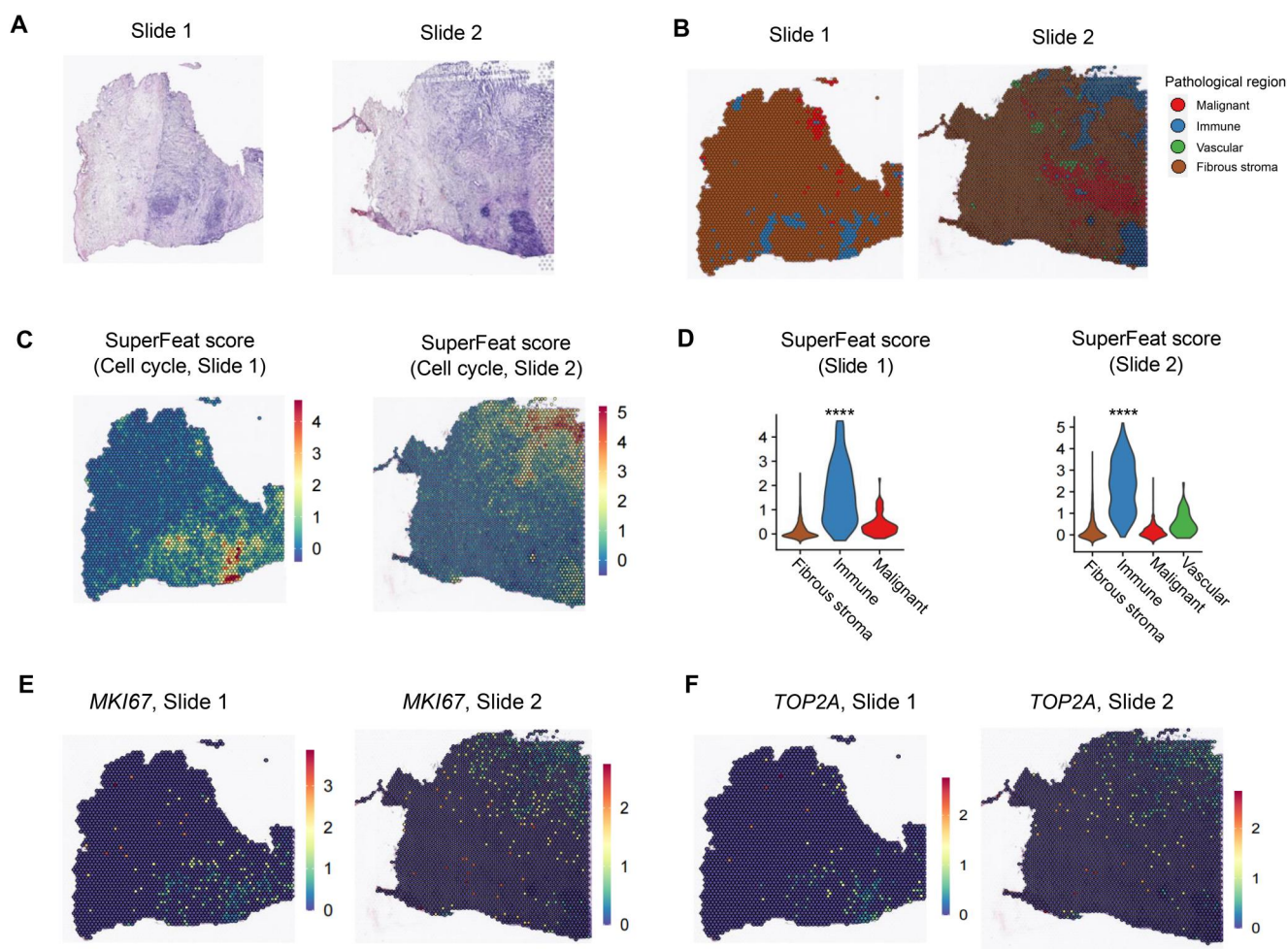


Figure 5 SuperFeat scoring on two ST slides

A. H&E staining for two slides from the same tumor tissue. **B.** Histological annotation by the pathologist. **C.** SuperFeat cell cycle scores of the dots on the two replicate slides of 10X Genomics Visium platform. **D.** Violin plots of SuperFeat cell cycle scores of the dots on the two replicate slides. **E.** *MKI67* signals of the dots on the two replicate slides. **F.** *TOP2A* signals of the dots on the two replicate slides. The significance was determined by Wilcoxon test (****, $P < 0.0001$). ST, spatial transcriptomics; H&E, hematoxylin and eosin.

weight-ranked genes derived from single-cell data could be more specific to the cellular state change than the differential genes found in bulk RNA transcriptomic data but more general in the application targeting the detrimental cellular state.

Using the CMap search strategy over the LINCS L1000 data repositories implemented by the signatureSearch package, we tested such an idea and got interesting results.

Figure 6A shows the drug search based on the top weighted genes derived from the cell cycle features. Among the top 10 hits on LINCS L1000, AZD8055 [41], palbociclib [42], NVP-BE235 [43], naproxol [44], ivermectin [45], and oxindole-I [46] are the cell cycle arrest agents that have been previously reported.

The drug search based on the top weighted genes derived from the EMT features also gives meaningful results. Among the top 10 hits on LINCS L1000, saracatinib has been reported to be involved in the restoration of E-cadherin expression [47], and lapatinib plays a role in EMT [48] (**Figure 6B**). The top 1 hit on CMap, menadione (vitamin K3) has been reported to be involved in suppression of EMT [49] (**Figure 6C**).

Stability of model parameters and reproducibility of SuperFeat results

As the model parameters are the major determinants of the drug search, in order to evaluate the robustness of the methods, we compared the top weighted genes and the output drugs derived from different training datasets using the most established cellular feature, *i.e.*, proliferation.

The datasets were retrieved from GEO databases: GSE140228 for HCC proliferation and GSE110686 for T cell proliferation. Although different cell types were involved in the same cellular state, our results showed that 55%–60% of the top positively weighted genes were reproducible in two datasets (**Figure 7A**). There were also 21%–28% of negatively weighted genes overlapped, whose roles remain elusive. The enrichment analysis of the top positively weighted genes mostly hit the same Gene Ontology (GO) terms (**Figure 7B**). At last, we compared the output drugs. For the CMap search, 7 out of 20 top selections were reproducible. For the LINCS L1000 search, 10 out of 20 top selections were reproducible (**Figure 7C**). It was also found that naproxol and palbociclib were always among the top selections of both datasets in the LINCS L1000 search.

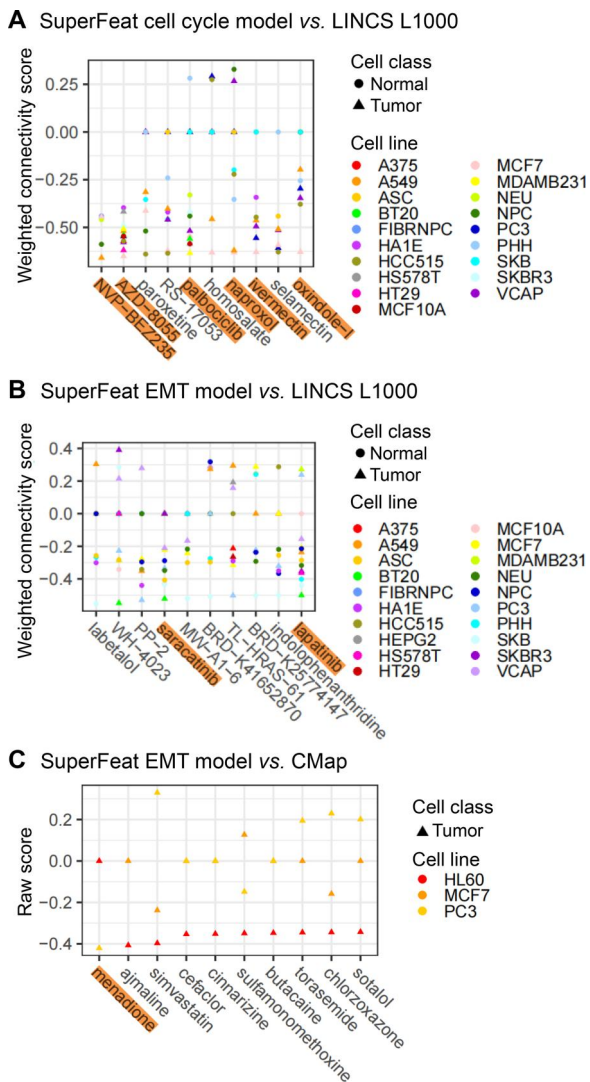


Figure 6 Drug search using weight-ranked genes of SuperFeat

A. Drug search result based on the top positively/negatively weighted genes of the SuperFeat cell cycle model and the LINCS L1000 database. **B.** Drug search results based on the top positively/negatively weighted genes of the SuperFeat EMT model and the LINCS L1000 database. **C.** Drug search results based on the top positively/negatively weighted genes of the SuperFeat EMT model and the CMap database. CMap, the Connectivity Map.

Such reproducible results suggest that, despite the training datasets being different, similar genes that contribute to the canonical cellular statuses/features will be assigned high weights. These genes will also subsequently provide the leads necessary to discover drug candidates that either promote beneficial cellular states or inhibit adverse ones.

Figures S5 and S6 display additional results that assess the stability of the features for hypoxia and T cell exhaustion, respectively. The overlap of high-weight genes remains significant for both features, underscoring the robustness of our findings. However, the limited number of cells in some published datasets could restrict our validation of reproducibility. We were unable to identify a sufficient number of cells to extend our reproducibility validation further, such as EMT. This limitation is evidenced in Table S1, which illustrates that the state1 cell populations in certain datasets are too small to train a robust model. Such dataset could still be used for score testing though.

In vitro and *in vivo* validation of drug effects on adverse cellular state development in tumor

In a collaborative study of the cancer-associated fibroblasts (CAFs), we discovered a state transition of CAFs in a subcutaneous melanoma mouse model. The dataset included 101,684 cells from the tumor tissue. It allows to delineate nine distinct tumor micro-environment (TME) types: T cells, B cells, dendritic cells, neutrophils, macrophages, fibroblasts, endothelial cells, melanoma cells, and stem/progenitor-like cells. Focusing on CAFs, our analysis revealed a novel subpopulation marked by $CD34^+PI16^+$, which gave rise to the classical $ACTA2^+MCAM^+$ CAFs in tumor, also known as myCAF (data unpublished). The lineage connection of these CAFs with these markers were confirmed using tdTomato marker.

Similar CAF subpopulations were characterized in both human and mouse tissues in the published datasets [50] (Figure 8A and B). The developmental trajectory of CAFs from the $CD34^+$ fibroblast progenitor state to the $ACTA2^+$ classical CAF state was defined and validated in our collaborative study. State1 ($CD34^+$ CAF) represents a more stemness state characterized by the signature genes such as *Cd34* and *Pdgfra*. State0 ($CD34^-$ CAF) represents a more differentiation state with the tumor-prone signature marker such as *Acta2* (Figure 8C).

Using SuperFeat framework, we were hence able to train a new stemness scoring model to evaluate the progression of CAF differentiation in a tumoral biopsy. A total of 4594 cells which were composed of $CD34^+PI16^+$ progenitor cells and $ACTA2^+MCAM^+$ myCAF cells were used for training (Table S1).

Using the subsequent search, we aimed to find the drugs that potentially suppress the conversion from the progenitor State1 to the tumor-prone State0 (Figure 8D and E). The top 20 drug candidates from LINCS L1000 are shown in Figure 8F. By validation experiments on these 20 candidates, four of them showed very promising outcomes. Figure 8G shows the outcome of MKN-45 (gastric cancer cell) subcutaneous tumor model with the treatment of WH-4023, which is one of the four drugs. The tumors significantly shrank in five replicates.

To confirm that the drug indeed inhibits tumor growth via the suppression of CAF development, we performed immunofluorescence staining of gastric subcutaneous tumors for signature markers, including $CD34$ and $PI16$ for the progenitor state, and $ACTA2$ (also known as α -SMA) and $COL1A1$ for the classic tumor-prone state. Notably, the WH-4023 treatment group exhibited lower α -SMA signals but higher $CD34$ and $PI16$ signals in comparison with the control group (Figure 8H).

Discussion

In this study, we have established and demonstrated an innovative framework designed to assess cellular states, tissue development, and even individual patient conditions. This framework comprises a few critical elements: (1) an R package that facilitates the evaluation of canonical cellular states via a scoring system; (2) a data repository containing model parameters of various documented cellular states from previous research; and (3) Python training code for user-defined cellular states. This setup allows for the efficient evaluation of canonical cellular states previously identified in research. Users can leverage this framework to simply assess cellular

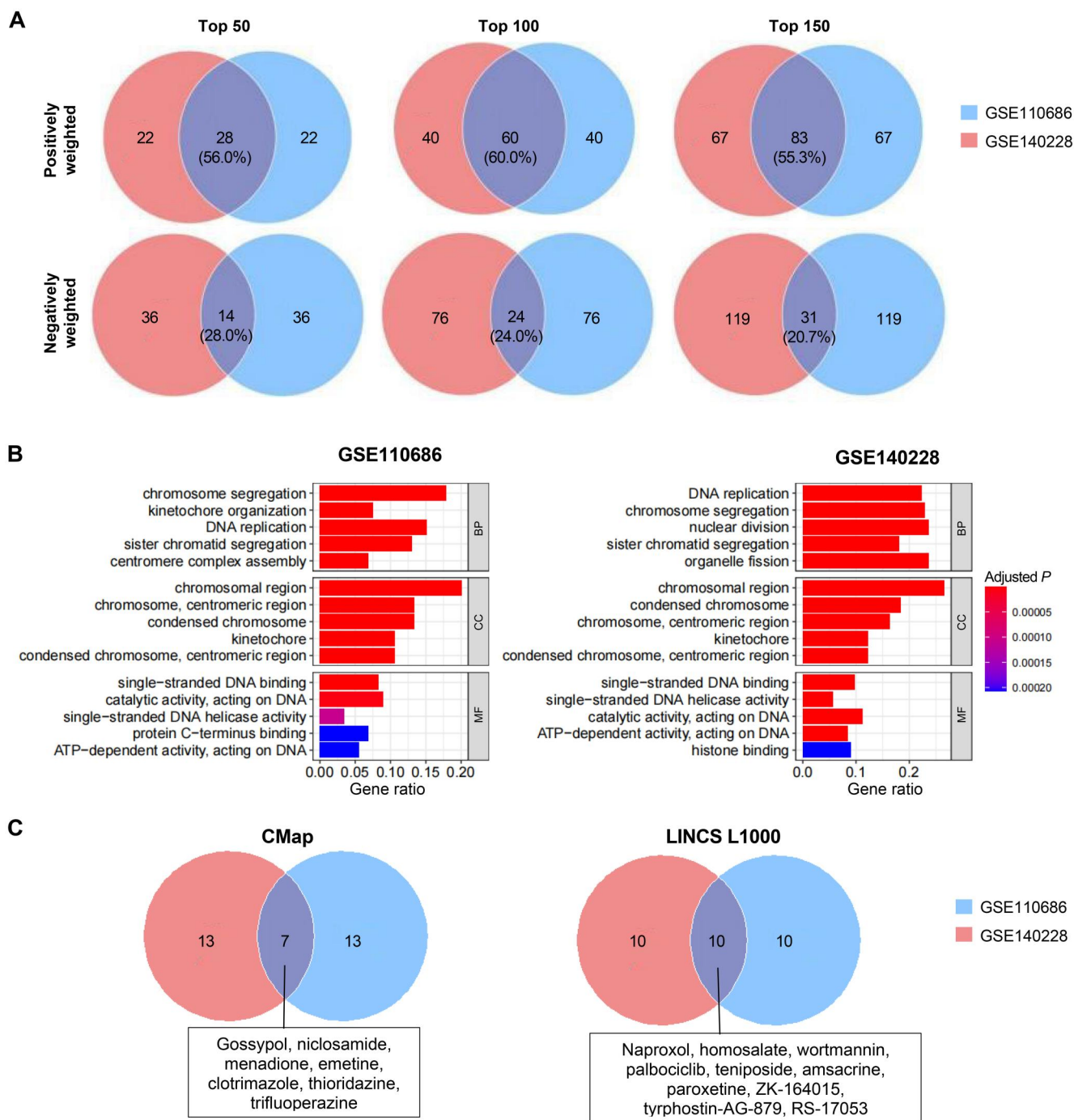


Figure 7 Stability of model parameters and reproducibility of SuperFeat results of cell cycle state/feature
A. Venn diagrams showing stable weighted genes between different datasets. The number indicates gene count. **B.** Bar plots displaying similar GO terms. Only top 5 categories are shown here. **C.** Venn diagrams showing the drugs repeatedly searched in CMap and LINCS L1000. The detailed drugs are listed in the boxes. The number indicates drug count. GO, Gene Ontology; BP, biological process; CC, cellular component; MF, molecular function.

states using existing models, or train their own SuperFeat scoring models, and potentially propagate their own cellular state models.

We have specifically demonstrated how the SuperFeat model assesses several key cellular states or features of tumors, such as cell proliferation, T cell exhaustion, hypoxia, and EMT, all of which are hallmark cancer signals. These examples underscore the simplicity, generalizability, and efficiency of the new method. We strived to demystify the “black

box” of the SuperFeat ANN model by visualizing and interpreting the model parameters, specifically the gene weights potentially contributing to a specific cellular state. Compared to the signature gene set-based scoring methods, SuperFeat demonstrates superior performance because it considers not only the up-regulation of signature genes, but also suppressed genes. Such contributions are reflected by positive and negative weights, respectively. By examining the weights of genes, researchers may acquire additional insights.

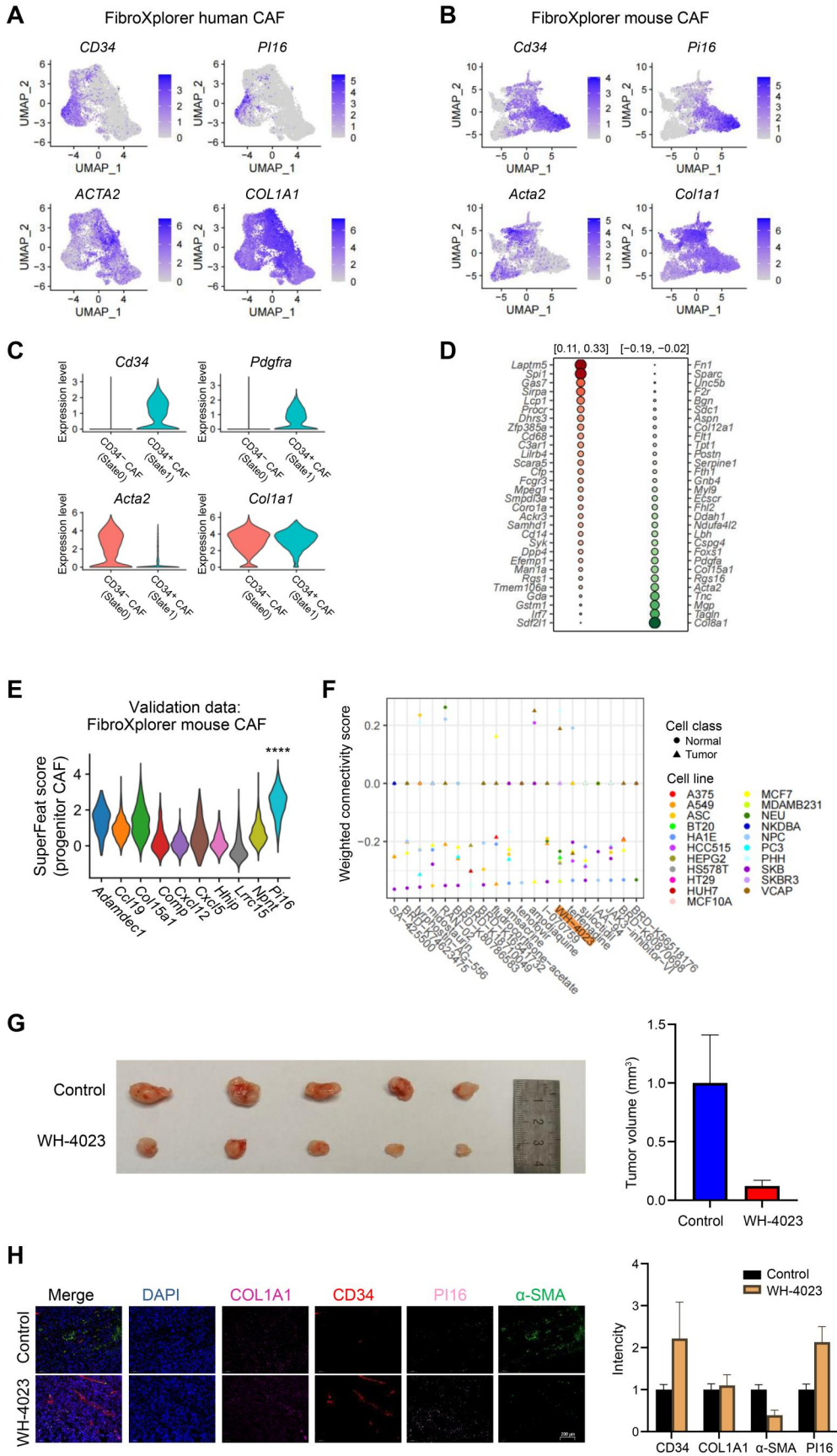


Figure 8 *In vitro* and *in vivo* validation of a drug candidate by SuperFeat and signatureSearch
A. Fibroblast subpopulations in human tissues from FibroXplorer database. **B.** Fibroblast subpopulations in mouse tissues from FibroXplorer database.
C. Violin plot for the fibroblast subpopulations in our CAF development study in subcutaneous tumor mice. **D.** Top 30 positively/negatively weighted

Essentially, SuperFeat provides additional learning framework and a scoring strategy that streamline the rapid assessment of states or features in tissue, many of which have been extensively explored and acknowledged in existing literature. Moreover, compared with other traditional methods, SuperFeat renders such assessment more intuitive, generalizable, and user-friendly, without relying on the details of the gene set. While an ANN model bypasses the uncertainties of statistical assumptions in a rigorous mathematical model, the weight parameters of genes derived from such a training still align remarkably well with expectations. Logically, the metric values of the SuperFeat cellular status/feature reflect the resemblance between the putative cells of interest in the training dataset and the cells in the new study, in comparison to their counterparts. Therefore, the model's reliability largely depends on the quality of the training dataset. As our understanding of cells and single-cell transcriptomics datasets keep growing, the SuperFeat framework is set to make the automated learning and assessment of cellular statuses/features increasingly reliable and efficient.

An even more thrilling advantage of this framework lies in its potential for downstream applications, *e.g.*, repurposing drug searches. The other gene set-based approaches of cellular state evaluation are not comparable in this aspect. As we know, CMap was based on the connectivity between gene expression profiles and drug perturbations. Previous CMap perturbation datasets were derived from bulk RNA-seq. The recent progress of the scRNA-seq technique makes it practical to generate such connectivity at single-cell resolutions. The advantage of our strategy could heighten the precision and success rate of expression profile-based drug searches. We successfully conducted the validation experiments that substantiate the feasibility of such strategy. Another advantage of neural network-based strategy pertains to the convenience of the interconnection with other models. It makes integration of drug search tasks based on multiple data sources possible.

In brief, our effort aims to build a preliminary and versatile framework that allows for the rapid and automatic evaluation of the conditions of multicellular samples by investigating single-cell transcriptomics, eventually finding solutions to more complex biological and clinical problems. ANN-based SuperFeat architecture not only can be used independently, but also enables us to generate small network modules that can be assembled and integrated. We have shown results suggesting high consistency between the single-node SuperFeat model and logistic regression, which makes SuperFeat model interpretable. However, considering the integrability of multiple models, an ANN-based model offers greater versatility and compatibility to build a more complex network structure and to accomplish more intricate learning tasks, especially when the current machine learning programming architecture is predominantly based on Python libraries such as TensorFlow, PyTorch, or Keras. More interesting learning strategy can therefore be realized. For example, the pre-trained SuperCT and SuperFeat models can be easily

assembled into a deeper learning network model that can be further trained for new tasks in a transfer learning fashion (by “freezing” pre-trained parameter sub-networks in new model training). Such a deeper network model enables the solving of more complex problems involving cell type classification and feature quantitation with better interpretability of SuperCT and SuperFeat components. We anticipate further developments in the biological field driven by the efforts of AI scientists with more careful consideration of the benefits from SuperFeat and SuperCT components in single-cell transcriptomics studies.

Ethical statement

All animal breeding, housing, and experimentation were conducted according to the guidelines of the Institutional Committee of Shanghai Jiao Tong University School of Medicine for Animal Research, China (Approval No. XHEC-F-2020-026).

Code availability

The framework is implemented in R package rSuperFeat and can be accessed at <https://github.com/weilin-genomics/rSuperFeat>.

CRedit author statement

Jianmei Zhong: Methodology, Investigation, Visualization, Writing – review & editing. **Junyao Yang:** Investigation, Visualization, Writing – review & editing. **Yinghui Song:** Investigation, Visualization. **Zhihua Zhang:** Investigation, Visualization. **Chunming Wang:** Investigation. **Renyang Tong:** Investigation. **Chenglong Li:** Investigation. **Nanhui Yu:** Investigation. **Lianhong Zou:** Investigation. **Sulai Liu:** Investigation. **Jun Pu:** Supervision, Writing – review & editing. **Wei Lin:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. All authors have read and approved the final manuscript.

Supplementary material

Supplementary material is available at *Genomics, Proteomics & Bioinformatics* online (<https://doi.org/10.1093/gpbjnl/qzae036>).

Competing interests

The authors have declared no competing interests.

Acknowledgments

This study was supported by grants from the Shanghai Jiao Tong University, the Renji Hospital Start-up funding for New PI, the Natural Science Foundation of Shanghai Science and Technology Innovation Action Plan (Grant No.

Figure 8 Continued

genes for CAF development SuperFeat model. **E.** Validation of CAF development SuperFeat model using FibroXplorer mouse dataset. The significance was determined by Wilcoxon test (****, $P < 0.001$). **F.** Top 20 drug search candidates. **G.** The effect of WH-4023 on tumor shrinkage. **H.** The effect of WH-4023 on fibroblast development using immunofluorescence staining. Scale bar, 200 μm . FibroXplorer is available at <https://www.fibroexplorer.com/home>. UMAP, Uniform Manifold Approximation and Projection; CAF, cancer-associated fibroblast.

21ZR1441500), the Young Talent of Hunan (Grant No. 2020RC3066), the Hunan Natural Science Fund for Excellent Young Scholars (Grant No. 2021JJ20003), and the China Postdoctoral Science Foundation (Grant No. 2021T140197).

ORCID

0000-0002-2240-8651 (Jianmei Zhong)
 0000-0003-3307-704X (Junyao Yang)
 0000-0002-4746-0536 (Yinghui Song)
 0000-0003-0457-5699 (Zhihua Zhang)
 0000-0002-8548-3843 (Chunming Wang)
 0000-0002-2429-6702 (Renyang Tong)
 0000-0001-5281-3175 (Chenglong Li)
 0000-0001-5474-2198 (Nanhui Yu)
 0000-0002-0964-7447 (Lianhong Zou)
 0000-0002-5257-3922 (Sulai Liu)
 0009-0004-2319-4634 (Jun Pu)
 0000-0002-7506-3466 (Wei Lin)

References

- [1] Hoheisel JD. Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet* 2006;7:200–10.
- [2] Mardis ER. Next-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto Calif)* 2013;6:287–303.
- [3] Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;6:377–82.
- [4] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 2015;161:1202–14.
- [5] Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 2015;161:1187–201.
- [6] Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* 2014;11:163–6.
- [7] Li C, Menoret A, Farragher C, Ouyang Z, Bonin C, Holvoet P, et al. Single-cell transcriptomics-based MacSpectrum reveals novel macrophage activation signatures in diseases. *JCI Insight* 2019;5:e126453.
- [8] Gulati GS, Sikandar SS, Wesche DJ, Manjunath A, Bharadwaj A, Berger MJ, et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* 2020;367:405–11.
- [9] Xie P, Gao M, Wang C, Zhang J, Noel P, Yang C, et al. SuperCT: a supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles. *Nucleic Acids Res* 2019;47:e48.
- [10] de Kanter JK, Lijnzaad P, Candelli T, Margaritis T, Holstege F. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res* 2019;47:e95.
- [11] Ma F, Pellegrini M. ACTINN: automated identification of cell types in single cell RNA sequencing. *Bioinformatics* 2020;36:533–38.
- [12] Wherry EJ. T cell exhaustion. *Nat Immunol* 2011;12:492–99.
- [13] Liu YC, Zou XB, Chai YF, Yao YM. Macrophage polarization in inflammatory diseases. *Int J Biol Sci* 2014;10:520–9.
- [14] Murray PJ. Macrophage polarization. *Annu Rev Physiol* 2017;79:541–66.
- [15] Corrigan AM, Tunnacliffe E, Cannon D, Chubb JR. A continuum model of transcriptional bursting. *Elife* 2016;5:e13051.
- [16] Guo Y, Xie YQ, Gao M, Zhao Y, Franco F, Wen M, et al. Metabolic reprogramming of terminally exhausted CD8⁺ T cells by IL-10 enhances anti-tumor immunity. *Nat Immunol* 2021;22:746–56.
- [17] Zhang R, Miao J, Zhu P. Regulatory T cell heterogeneity and therapy in autoimmune diseases. *Autoimmun Rev* 2021;20:102715.
- [18] Lamb J. The Connectivity Map: a new tool for biomedical research. *Nat Rev Cancer* 2007;7:54–60.
- [19] Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;313:1929–35.
- [20] Duan Y, Evans DS, Miller RA, Schork NJ, Cummings SR, Girke T. signatureSearch: environment for gene expression signature searching and functional interpretation. *Nucleic Acids Res* 2020;48:e124.
- [21] Hao Y, Hao S, Andersen-Nissen E, Mauck WR 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. *Cell* 2021;184:3573–87.e29.
- [22] Aibar S, Gonzalez-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;14:1083–6.
- [23] Foroutan M, Bhuvu DD, Lyu R, Horan K, Cursons J, Davis MJ. Single sample scoring of molecular phenotypes. *BMC Bioinformatics* 2018;19:404.
- [24] Hanzelmann S, Castelo R, Guinney J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 2013;14:7.
- [25] Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* 2015;1:417–25.
- [26] Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. *Cell* 2017;171:1437–52.e17.
- [27] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;12:77.
- [28] Gheban BA, Colosi HA, Gheban-Rosca IA, Georgiu C, Gheban D, Crişan D, et al. Techniques for digital histological morphometry of the pineal gland. *Acta Histochem* 2022;124:151897.
- [29] Du J, Yuan X, Deng H, Huang R, Liu B, Xiong T, Long X, et al. Single-cell and spatial heterogeneity landscapes of mature epicardial cells. *J Pharm Anal* 2023;13:894–907.
- [30] Neal JT, Li X, Zhu J, Giangarra V, Grzeskowiak CL, Ju J, et al. Organoid modeling of the tumor immune microenvironment. *Cell* 2018;175:1972–88.e16.
- [31] Zheng C, Zheng L, Yoo JK, Guo H, Zhang Y, Guo X, et al. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell* 2017;169:1342–56.e16.
- [32] Lin W, Noel P, Borazanci EH, Lee J, Amini A, Han IW, et al. Single-cell transcriptome analysis of tumor and stromal compartments of pancreatic ductal adenocarcinoma primary tumors and metastatic lesions. *Genome Med* 2020;12:80.
- [33] Hu Z, Artibani M, Alsaadi A, Wietek N, Morotti M, Shi T, et al. The repertoire of serous ovarian cancer non-genetic heterogeneity revealed by single-cell sequencing of normal fallopian tube epithelial cells. *Cancer Cell* 2020;37:226–42.e7.
- [34] Darmanis S, Sloan SA, Croote D, Mignardi M, Chernikova S, Samghababi P, et al. Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Rep* 2017;21:1399–410.
- [35] Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell* 2019;178:835–49.e21.
- [36] Zhang Q, He Y, Luo N, Patel SJ, Han Y, Gao R, et al. Landscape and dynamics of single immune cells in hepatocellular carcinoma. *Cell* 2019;179:829–45.e20.
- [37] Wu TD, Madireddi S, de Almeida PE, Banchereau R, Chen YJ, Chitre AS, et al. Peripheral T cell expansion predicts tumour infiltration and clinical response. *Nature* 2020;579:274–8.

- [38] Kierans SJ, Taylor CT. Regulation of glycolysis by the hypoxia-inducible factor (HIF): implications for cellular physiology. *J Physiol* 2021;599:23–37.
- [39] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;36:411–20.
- [40] Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;33:495–502.
- [41] Chen Y, Lee CH, Tseng BY, Tsai YH, Tsai HW, Yao CL, et al. AZD8055 exerts antitumor effects on colon cancer cells by inhibiting mTOR and cell-cycle progression. *Anticancer Res* 2018;38:1445–54.
- [42] Pennycook BR, Barr AR. Palbociclib-mediated cell cycle arrest can occur in the absence of the CDK inhibitors p21 and p27. *Open Biol* 2021;11:210125.
- [43] Li C, Xin P, Xiao H, Zheng Y, Huang Y, Zhu X. The dual PI3K/mTOR inhibitor NVP-BEZ235 inhibits proliferation and induces apoptosis of burkitt lymphoma cells. *Cancer Cell Int* 2015;15:65.
- [44] Kim MS, Kim JE, Lim DY, Huang Z, Chen H, Langfald A, et al. Naproxen induces cell-cycle arrest and apoptosis in human urinary bladder cancer cell lines and chemically induced cancers by targeting PI3K. *Cancer Prev Res (Phila)* 2014;7:236–45.
- [45] Zhang P, Zhang Y, Liu K, Liu B, Xu W, Gao J, et al. Ivermectin induces cell cycle arrest and apoptosis of HeLa cells via mitochondrial pathway. *Cell Prolif* 2019;52:e12543.
- [46] Kamal A, Ramakrishna G, Raju P, Rao AV, Viswanath A, Nayak VL, et al. Synthesis and anticancer activity of oxindole derived imidazo[1,5-*a*]pyrazines. *Eur J Med Chem* 2011;46:2427–35.
- [47] Huang RY, Wong MK, Tan TZ, Kuay KT, Ng AH, Chung VY, et al. An EMT spectrum defines an anoikis-resistant and spheroidogenic intermediate mesenchymal state that is sensitive to E-cadherin restoration by a Src-kinase inhibitor, saracatinib (AZD0530). *Cell Death Dis* 2013;4:e915.
- [48] Umbreit C, Erben P, Faber A, Hofheinz RD, Schultz JD, Hoermann K, et al. Lapatinib-induced mesenchymal–epithelial transition in squamous cell carcinoma cells correlates with unexpected alteration of beta-catenin expression. *Oncol Lett* 2016;11:2715–24.
- [49] Kishore C, Sundaram S, Karunakaran D. Vitamin K3 (menadi-one) suppresses epithelial–mesenchymal transition and Wnt signaling pathway in human colorectal cancer cells. *Chem Biol Interact* 2019;309:108725.
- [50] Buechler MB, Pradhan RN, Krishnamurty AT, Cox C, Calviello AK, Wang AW, et al. Cross-tissue organization of the fibroblast lineage. *Nature* 2021;593:575–9.