Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SigFormer: interpretable modeling of transcellular signaling and microenvironment interactions from multimodal single-cell data

Yunping Zhu

zhuyunping@ncpsb.org.cn

State Key Laboratory of Medical Proteomics, National Center for Protein Sciences (Beijing)

Mingfei Han

National Center for Protein Sciences (Beijing) https://orcid.org/0000-0002-6036-6309

Xiao Li

National Center for Protein Sciences (Beijing)

Yu Zhao

College of Life Sciences, Hebei University

Tongpeng Yue

Peking Union Medical College Hospital

Liya Wang

College of Life Sciences, Hebei University

Juan Wang

State Key Laboratory of Medical Proteomics, National Center for Protein Sciences (Beijing)

Jie Ma

National Center for Protein Sciences (Beijing) https://orcid.org/0000-0002-8934-922X

Hebing Chen

Academy of Military Medical Sciences

Article

Keywords:

Posted Date: May 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9489991/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

SigFormer: interpretable modeling of transcellular signaling and microenvironment interactions from multimodal single-cell data

Mingfei Han[1,5], Xiao Li[1,5], Yu Zhao[2,5], Tongpeng Yue[3], Liya Wang[2], Juan Wang[1], Jie Ma[1*], Hebing Chen[4*], and Yunping Zhu[1*]

[1]State Key Laboratory of Medical Proteomics, National Center for Protein Sciences (Beijing), Beijing 102206, China.

[2]College of Life Sciences, Hebei University, Baoding 071002 Hebei, China

[3]Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China.

[4]Academy of Military Medical Sciences, Beijing, 100850, China.

[5]These authors contributed equally: Mingfei Han, Xiao Li, Yu Zhao.

*Corresponding authors. Jie Ma, Email: majie729@163.com; Hebing Chen, Email: chb-1012@163.com; Yunping Zhu, Email: zhuyunping@ncpsb.org.cn.

Abstract

Signaling pathways integrate extracellular cues with intracellular signal transduction and gene regulation, providing a mechanistic foundation for interpretable perturbation analysis and drug discovery. Here we present SigFormer, a graph Transformer-based multi-task framework that integrates cell classification, network reconstruction, and in silico knockout to infer mechanistically interpretable sender – receiver pathways spanning ligands, receptors, mediators, transcription factors, and target genes. SigFormer does not require large training cohorts, enabling accurate pathway

reconstruction from both single-sample single-cell RNA sequencing and multimodal datasets. Benchmarking demonstrates that SigFormer consistently outperforms nine existing methods across five signaling stages, improving F1 scores by 8–46% over the second-best method at each stage. By applying SigFormer to cancer multi-omics data, we constructed the Cancer Cell Pathway Atlas—spanning 28 cancers and 23 tumor microenvironment cell types — facilitating cross-tumor pathway analysis, drug repurposing, and future tissue-level "white-box" modeling of signaling programs.

Introduction

Cells rarely function in isolation; they interact through diverse signaling molecules and membrane-associated structures to coordinate gene expression programs and orchestrate cellular behavior[1-4]. These integrated intra- and intercellular signaling networks provide a foundation for white-box models at both the cellular and tissue levels, facilitating high-fidelity mechanistic interpretation, systematic perturbation analysis, and the rational generation of biological hypotheses[5].

To study transcellular signaling, tools such as CellPhoneDB[6] and CellChat[7] estimate communication between cell types or clusters based on ligand and receptor expression, while more recent approaches extend beyond ligand–receptor interactions to incorporate intracellular signal transduction and downstream gene regulatory processes[8]. For example, NicheNet[9] models how ligands influence transcription factor (TF) activity and target gene expression in recipient cells, whereas exFINDER[10] considers ligand-induced signaling networks arising from both cellular and non-cellular sources. However, the multi-step and multi-dimensional signal transduction processes

connecting activated receptors to downstream TFs—including chemical catalysis, physical interactions, and transport regulation (Extended Data Fig. 1a)—remain largely underexplored. Although methods such as NicheNet[9] and scSeqCom[11] incorporate intracellular signaling proteins, they do not explicitly model how signals propagate from receptors to TFs through stepwise intermediate reactions (Extended Data Fig. 1b). This black-box treatment limits reconstruction of comprehensive tissue signaling programs and generation of interpretable therapeutic hypotheses.

Another major challenge in this field is integrating the complementary strengths of multimodal data. Single-cell RNA sequencing (scRNA-seq) quantifies mRNA abundance but does not directly capture protein activity. In contrast, single-cell proteomics (scProteomics) measures functional protein levels, single-cell ATAC sequencing (scATAC-seq) reveals chromatin accessibility and transcriptional regulatory potential, and the spatial organization of cells plays a critical role in shaping cell–cell communication[12]. Although multi-omics integration has been widely applied to tasks such as disease classification, diagnosis, and cross-modal prediction, integrating multimodal single-cell data to improve signaling inference remains challenging.

Here we present SigFormer, a graph Transformer-based multi-task framework that integrates cell classification, cell-type-specific signaling network reconstruction, and in silico knockout to infer mechanistically interpretable sender – receiver signaling pathways spanning ligand–receptor interactions, receptor-to-TF signal transduction through six classes of intracellular interactions, and TF–target gene regulation. By

constructing and learning cell-specific network topologies, SigFormer avoids reliance on large training cohorts, enabling accurate pathway reconstruction even from single-sample scRNA-seq data. Furthermore, it modularly integrates matched scProteomics, scATAC-seq, and spatial information to further refine pathway inference.

Compared with nine existing pathway inference methods, SigFormer consistently performed best across five key signaling stages, improving F1 scores by 16%, 8%, 46%, 20%, and 17% for predicting gold-standard ligands, receptors, mediators, TFs, and target genes, respectively, over the second-best method at each stage. Moreover, we demonstrate that incorporating matched scProteomics and scATAC-seq data significantly enhances the biological reliability of inferred components, while spatial constraints improve the resolution of membrane-bound communication.

Applying SigFormer to single-cell multi-omics data from the Curated Cancer Cell Atlas (3CA)[13], Single-cell Proteomics DataBase (SPDB)[14], and ATACdb[15], we constructed the Cancer Cell Pathway Atlas (CCPA), encompassing bidirectional signaling pathways between malignant cells and 22 distinct non-malignant cell types across 28 common human cancers. CCPA provides high-confidence candidate signaling components with annotated cellular origins, functional roles, and prior support from MSigDB[16]. As a case study, cross-tumor comparison of signaling pathways across 16 epithelial tumors identified 61 drug repurposing candidates. In vitro experiments further showed significant cytotoxic effects of selected tumor-cell receptor-targeting drugs in human lung adenocarcinoma A549 cells and pancreatic cancer AsPC-1 cells.

Result

Overview of SigFormer

SigFormer takes scRNA-seq dataset with cell type annotations as input and first constructs cell-specific networks by integrating genes that are highly expressed in each cell with a curated intracellular signaling library (stage 1 in Fig. 1). This library incorporates seven categories of signaling events—TF–target regulation, protein–protein interactions, catalytic reactions, transport regulation, state-change control, protein-complex formation, and phosphorylation regulation—from extensive, authoritative sources. To ensure reliability, only interactions supported by at least two independent sources are retained (see Methods 1.1 for details).

To extract high-confidence, cell type-specific signaling interactions, cell-specific networks are fed into a graph Transformer-based encoder to generate latent node embeddings (stage 2 in Fig. 1). This encoder captures the functional context of genes by dynamically attending to their interactions. The node embeddings of each cell-specific network serve two purposes: 1) an inner product decoder reconstructs the network topology by scoring potential edges between pairs of nodes based on their latent embeddings, and 2) a graph-level representation (cell embedding) is obtained via global max pooling over all node embeddings and passed through a feed-forward neural network for cell type prediction (see Methods 1.2 for details). Given that wiring of signaling pathways is highly cell type dependent[17], this process aims to learn graph representations that are consistent within cell classes but distinct between them, thereby reconstructing cell type-specific signaling networks.

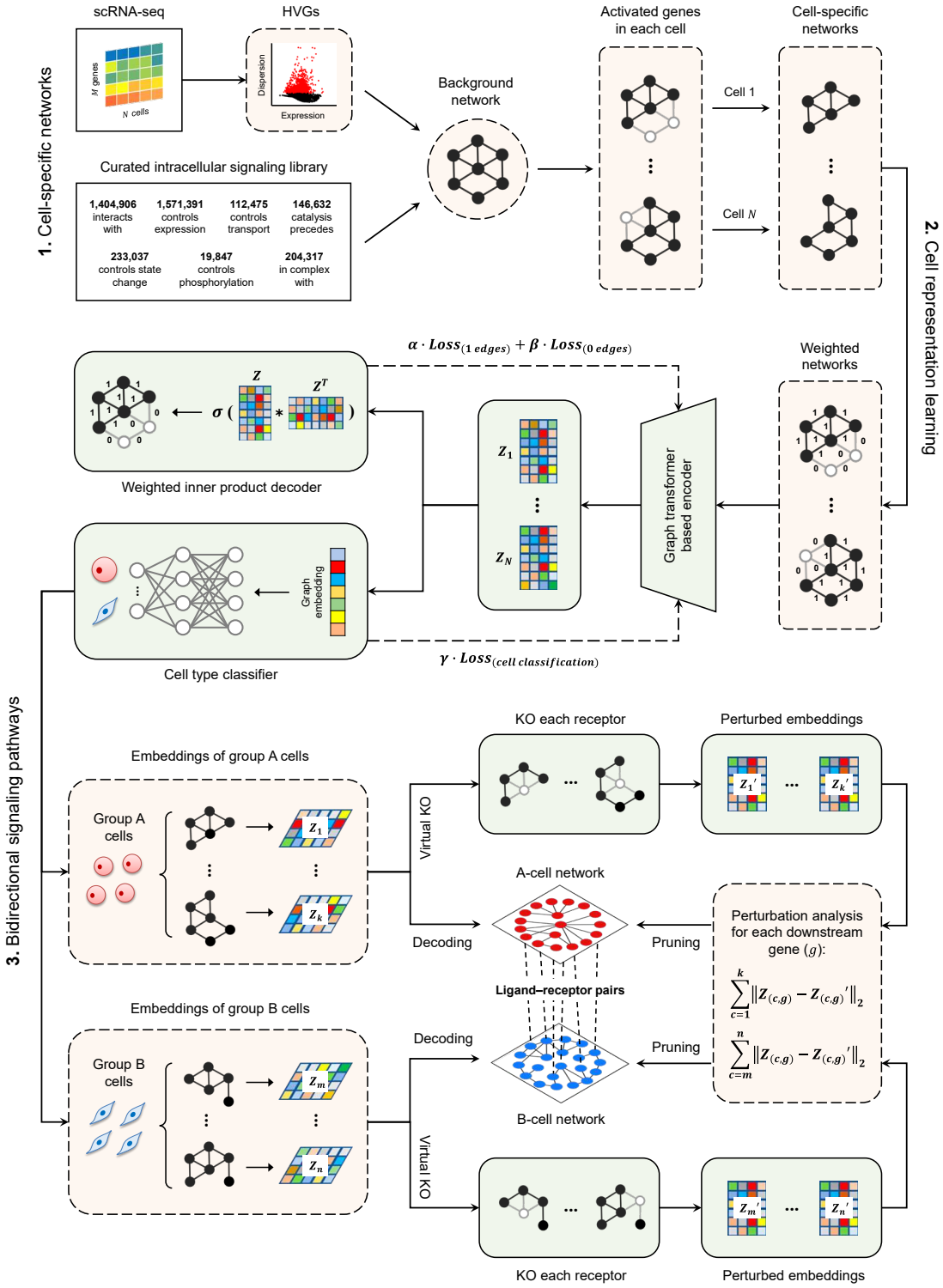


Fig. 1 | Architecture of SigFormer. SigFormer takes scRNA-seq data with cell type labels as requirted input and infers signaling pathways between user-defined index cells and other cell types through three major steps. **1. Cell-specific network inference:** A background signaling network is first constructed by integrating highly variable genes with a curated intracellular signaling library. Cell-specific activated genes and their

corresponding interactions in the background network form cell-specific networks. **2. Cell representation learning:** Cell-specific networks are then fed into a graph Transformer-based multi-task learning framework to learn both node embeddings (gene level) and graph embeddings (cell level). **3. Transcellular pathway reconstruction:** Node embeddings from cells of the same type are aggregated and decoded into a cell type-specific network. Sender and receiver networks are subsequently integrated to infer ligand–receptor-mediated cell–cell communication, while receptor-dependent downstream signaling is inferred through virtual knockout of receptors in the receiver-cell network.

After training, node embeddings from cells of the same type are aggregated using a trimmed mean to generate cell type-specific joint embeddings, which are decoded with an inner-product decoder to reconstruct the corresponding signaling network (stage 3 in Fig. 1). Finally, sender and receiver signaling networks are integrated to infer ligand–receptor-mediated cell–cell communication, while receptor-mediated downstream signaling is identified through in silico knockout of receptors in the receiver-cell network (see Methods 1.3 for details). When matched scProteomics, scATAC-seq, or spatial data are available, pathways inferred from scRNA-seq can be further refined by retaining ligands/receptors supported by protein-level evidence, TFs supported by chromatin accessibility, or sender–receiver cells constrained by spatial neighborhoods (Figs. 3a, 4a and Methods 1.5).

Pathway inference benchmark

We benchmarked SigFormer against nine state-of-the-art methods, including four representative tools for intercellular ligand–receptor inference (CellChat[7], CellPhoneDB[6], ICELLNET[18], and LRLoop[19]) and five methods capable of inferring

both ligand–receptor interactions and downstream intracellular signaling (NicheNet[9], exFINDER[10], scSeqComm[11], CellCall[20], and scMLnet[21]). Benchmarking was conducted using scRNA-seq datasets from eight tumor types—glioblastoma, leukemia, pancreatic cancer, breast cancer, colorectal cancer, lung cancer, melanoma, and ovarian cancer (Fig. 2a). Because rare cell types may not reliably support stable signaling inference, we collected all available samples for each tumor type from 3CA[13] and selected the sample exhibiting the lowest standard deviation in cell-type proportions. Details of data collection and preprocessing are provided in Methods 2.1.

By applying SigFormer and existing methods to each cancer dataset, we inferred bidirectional signaling pathways between malignant cells and each non-malignant cell type. Detailed descriptions of all methods and parameter settings are provided in the Supplementary Methods. For systematic evaluation, we decomposed each inferred pathway into five key stages—ligand, receptor, receptor-to-TF mediator, TF, and target gene—and constructed stage-specific positive reference sets using manually curated databases of cancer-associated ligand–receptor pairs, intracellular signaling interactions, and TF–target regulatory relationships (Methods 2.2). Summary statistics of these reference sets are shown in Extended Data Fig. 2, with complete lists provided in Supplementary Tables 1–5.



Fig. 2 | Evaluation of pathway inference performance. a, scRNA-seq datasets used for benchmarking were derived from eight common human cancers. Pie charts indicate cellular composition, and annotation boxes specify data source and sequencing technology. **b–p,** Recall (b,e,h,k,n), precision (c,f,i,l,o), and F1 score (d,g,j,m,p) of different methods in predicting positive ligands (b–d), receptors (e–g), mediators (h–j),

TFs (k–m), and target genes (n–p). Bars represent mean ± S.D./2 across the eight cancer datasets. Methods are grouped according to the signaling stages they cover: complete transcellular signaling (blue), lacking receptor-to-TF mediators (red), and lacking intracellular signaling (orange). Each method was evaluated only for the stages it supports.

Using these reference sets, SigFormer consistently achieved the highest recall across all five signaling stages. Compared with the second-best method at each stage, SigFormer improved recall by 17% for ligands (Fig. 2b), 10% for receptors (Fig. 2e), 173% for mediators (Fig. 2h), 22% for TFs (Fig. 2k), and 33% for target genes (Fig. 2n). Importantly, these gains did not come at the expense of precision. SigFormer maintained comparable precision across stages, with ligands, receptors, TFs, and target genes all above the stage average (Figs. 2c,f,l,o). Only mediators, with the largest recall improvement, showed slightly lower precision than the average (Fig. 2i).

Overall, SigFormer achieved substantially higher recall while maintaining balanced precision. This advantage is further reflected in the combined F1 score, where SigFormer ranked first across all evaluated stages (Figs. 2d,g,j,m,p), improving F1 by 16%, 8%, 46%, 20%, and 17% over the second-best method for ligands, receptors, mediators, TFs, and target genes, respectively.

Spatial constraints enhance inference of membrane-bound ligands

The spatial organization of cellular neighborhoods plays a critical role in shaping cell–cell communication[12]. Although spatial transcriptomic technologies enable spatially resolved measurements, their limited gene coverage restricts genome-wide signaling pathway inference. Nicheformer[22], a transformer-based foundation model for single-

cell and spatial omics, offers a promising solution to this challenge by learning spatial context directly from dissociated scRNA-seq data.

To evaluate the impact of spatial information on pathway inference, we used Nicheformer to predict spatial embeddings for eight cancer scRNA-seq datasets included in the benchmark analysis (Fig. 2a). For each malignant cell, we identified its 10 nearest non-malignant neighbors and inferred bidirectional malignant–non-malignant signaling pathways within these spatially constrained neighborhoods, which were compared with pathways inferred from all dissociated cells (Fig. 3a).

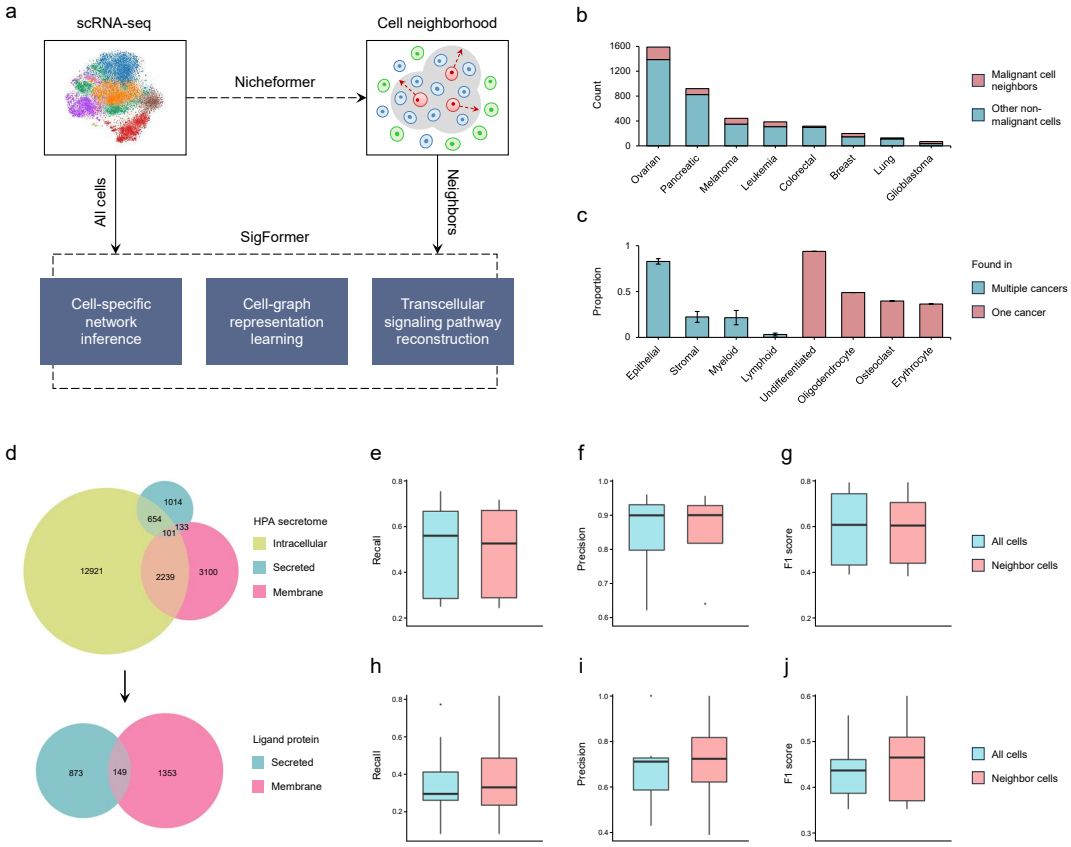


Fig. 3 | Effect of spatial information on pathway inference. a, Schematic illustrating the introduction of spatial constraints by integrating SigFormer with Nicheformer. **b,** Proportion of non-malignant cells located within malignant cell neighborhoods relative to all non-malignant cells for each cancer type. **c,** Proportion of each cell type found in malignant cell neighborhoods relative to its total abundance. Cyan bars indicate cell

types recurring as malignant cell neighbors in more than one cancer (error bars denote the standard deviation across cancers), whereas magenta bars indicate those observed in only a single cancer type. **d,** Venn diagram showing secreted and membrane-bound ligands annotated from The Human Protein Atlas. **e–g,** Recall, precision, and F1 score for secreted ligand inference, with cyan and pink boxes representing ligands inferred using all dissociated cells and spatial neighborhoods, respectively. **h–j,** Recall, precision, and F1 score for membrane-bound ligand inference under the same comparison.

Across most cancers, only a small fraction of non-malignant cells were included in malignant cell neighborhoods (Fig. 3b), reflecting a stringent spatial constraint. This stringency aims to amplify differences attributable to spatial constraints. Epithelial, stromal, and myeloid cells were consistently enriched among malignant cell neighbors across cancer types (cyan bars in Fig. 3c and Extended Data Fig. 3), consistent with established tumor microenvironment (TME) organization[23-25]. We also observed cancer-specific neighborhood patterns (magenta bars in Fig. 3c), including undifferentiated cells in leukemia, oligodendrocytes in glioblastoma, and osteoclasts in melanoma, in agreement with known tissue contexts[26-28].

Ligands were classified as secreted or membrane-bound based on The Human Protein Atlas[29, 30] (Fig. 3d and Supplementary Table 6). Introducing spatial neighborhoods had minimal impact on secreted ligand inference, causing only a slight decrease in recall and a modest increase in precision, with little change in F1 score (Figs. 3e–g). In contrast, spatial constraints consistently improved recall, precision, and F1 score for membrane-bound ligand prediction (Figs. 3h–j). These results suggest that using all cells is appropriate when focusing on secreted ligand-mediated signaling, whereas

introducing spatial constraints is advantageous for studying membrane-bound ligand-mediated cell–cell communication. Although this evaluation represents only an initial step, it highlights the importance of spatial context in shaping signaling pathways and provides a foundation for further investigation as spatial models and spatial transcriptomics technologies continue to advance.

Multi-omics refinement strengthens support for inferred pathways

Compared with scRNA-seq, scProteomics and scATAC-seq provide measurements closer to protein function and regulatory activity but are limited in gene and cell-type coverage. To leverage their complementary strengths, we used scRNA-seq to ensure broad pathway coverage and integrated available scProteomics and scATAC-seq to refine signaling inference (Fig. 4a and Methods 1.5).

Applying SigFormer to human cancer scRNA-seq datasets from 3CA[13] (Methods 3.1), we constructed the Cancer Cell Pathway Atlas (CCPA), encompassing bidirectional signaling between malignant cells and 22 major non-malignant cell types across 28 cancers (Figs. 4b,c and Extended Data Figs. 4a,b). Multi-omics refinement with SPDB[14] and ATACdb[15] (Methods 3.2 and 3.3) retained high-confidence pathways across ten tumor types—pancreatic ductal adenocarcinoma (PDAC), glioblastoma, lung adenocarcinoma (LUAD), colorectal cancer, breast cancer, melanoma, ovarian cancer, hepatocellular carcinoma (HCC), acute myeloid leukemia (AML), and acute lymphoblastic leukemia (ALL)—including proteomics-supported ligands derived from 14 non-malignant sender cell types (Fig. 4d), proteomics-supported receptors expressed on 15 non-malignant cell types (Fig. 4e), and scATAC-seq-supported TFs active in 9

malignant cell types (Fig. 4f). All CCPA pathways are publicly available at https://ccpa.ncpsb.org.cn, together with annotations of cancer type, cellular origin, and prior support from MSigDB[16].

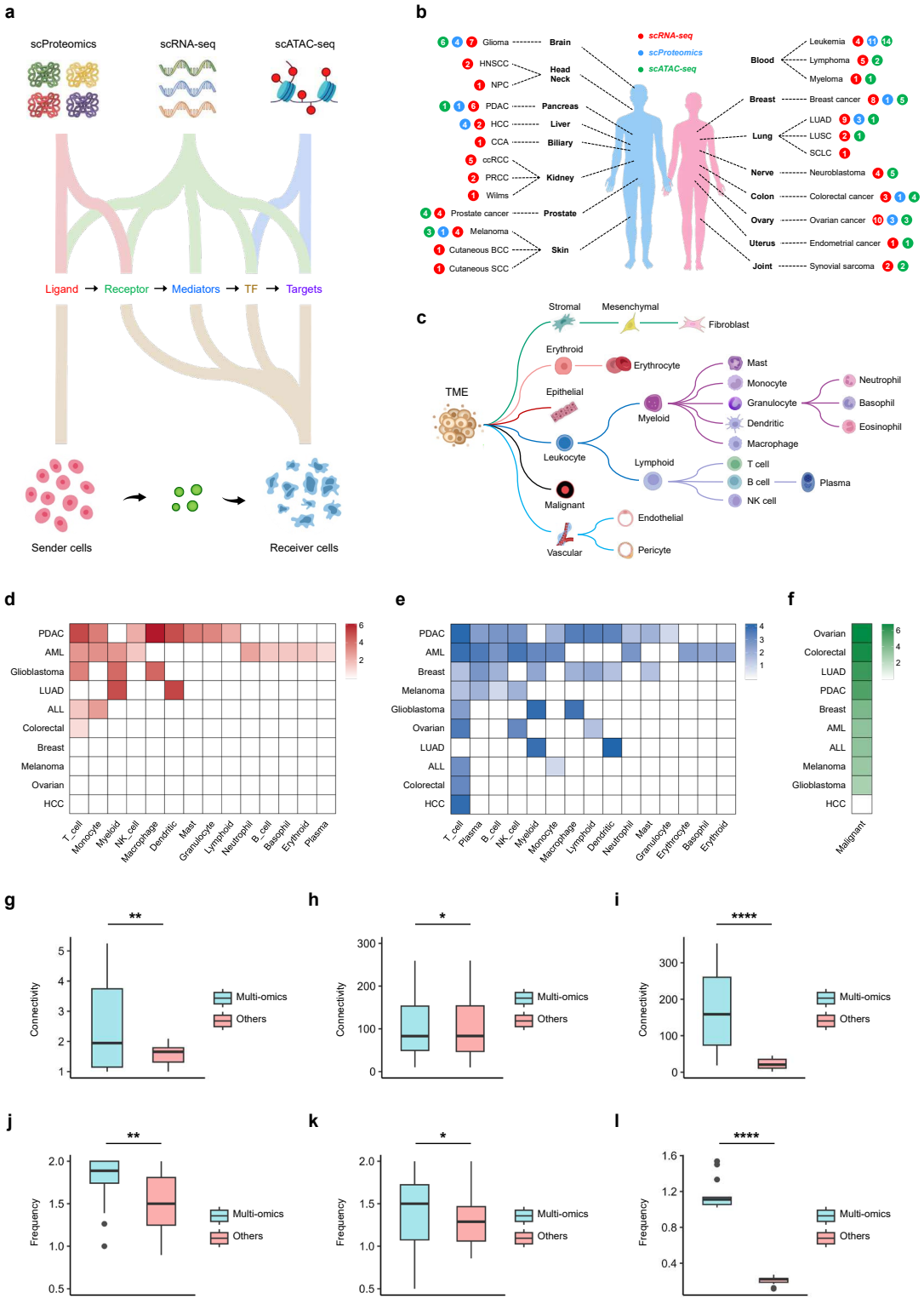


Fig. 4 | Single-cell multi-omics integration in CCPA. a, Schematic of the integration framework. scRNA-seq enables comprehensive reconstruction of transcellular signaling pathways, while scProteomics and scATAC-seq refine ligand–receptor and TF–target gene inference, respectively. **b,** Summary of scRNA-seq (red), scProteomics (blue), and scATAC-seq (green) datasets incorporated into CCPA. Numbers within circles denote the number of datasets per modality. **c,** Major tumor microenvironment cell types represented in CCPA. **d–f,** Tumor and cell-type coverage of proteomics-supported ligands (d), proteomics-supported receptors (e), and scATAC-seq-supported TFs (f) derived from scRNA-seq-inferred pathways. Only cell types represented in more than one tumor type are shown. Color intensity indicates the log2-transformed counts. **g,** Number of receptors targeted by proteomics-supported ligands (cyan) versus receptors targeted by ligands inferred exclusively from scRNA-seq (pink). **h,** Number of TFs regulated by proteomics-supported receptors (cyan) versus TFs regulated by receptors inferred exclusively from scRNA-seq (pink). **i,** Number of target genes regulated by scATAC-seq-supported TFs (cyan) versus target genes regulated by TFs inferred exclusively from scRNA-seq (pink). **j,** Frequencies of proteomics-supported ligands (cyan) versus other ligands (pink) annotated in tumor cell–cell communication databases. **k,** Frequencies of proteomics-supported receptors (cyan) versus other receptors (pink) annotated in tumor cell–cell communication databases. **l,** Frequencies of scATAC-seq-supported TFs (cyan) versus other TFs (pink) annotated in tumor regulatory factor databases.

Quantitative analysis revealed that proteomics-supported ligands bound significantly more receptors (Fig. 4g; paired t-test, $P = 0.002$), proteomics-supported receptors regulated significantly more TFs (Fig. 4h; $P = 0.018$), and scATAC-seq-supported TFs controlled significantly more target genes (Fig. 4i; $P = 1 \times 10^{-42}$) compared with components inferred from scRNA-seq alone. We further evaluated the frequency of occurrence of each ligand, receptor, and TF in tumor interaction and regulatory

databases (Methods 3.4), demonstrating that multi-omics-supported molecules exhibited significantly stronger evidence of cancer relevance (Figs. 4j–l; $P = 0.003$, 0.03, and 5.7×10^{-109}, respectively).

Together, these results indicate that multi-omics integration preferentially identifies signaling components that are both functionally influential and biologically reliable. Although current scProteomics and scATAC-seq datasets remain limited in scale, continued technological advances are expected to support an increasing fraction of multi-omics-supported signaling pathways within CCPA.

CCPA reveals conserved signaling pathways across epithelial tumors

As a resource spanning major cell types across common human cancers, CCPA enables cross-tumor comparisons that reveal conserved signaling mechanisms and potential therapeutic opportunities. As a case study, we identified signaling pathways shared by more than half of 16 epithelial tumor types (Fig. 5a), revealing extensive bidirectional communication between malignant cells and four major non-malignant cell types—macrophages, fibroblasts, endothelial cells, and T cells (Fig. 5b and Extended Data Figs. 5–8).

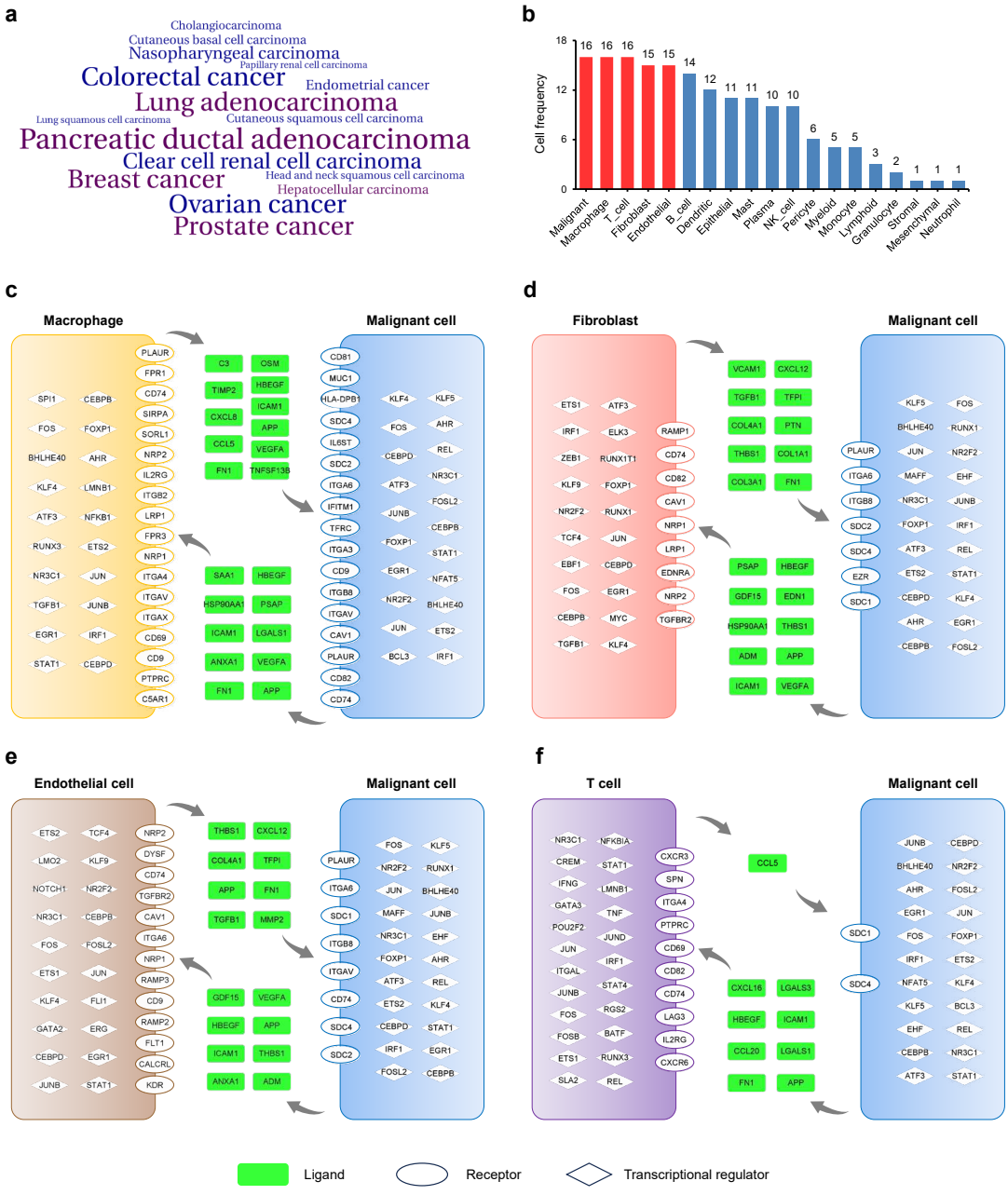


Fig. 5 | Consensus signaling pathways for epithelial tumors. a, Word cloud of the 16 epithelial tumor types, where font size reflects sample counts. **b,** Frequencies of major cell types across the 16 epithelial tumors. Cell types present in more than half of the tumors are represented by red bars. **c,** Bidirectional signaling between macrophages (yellow) and malignant cells (blue) shared by more than half of the epithelial tumors. Squares, ovals, and diamonds denote ligands, receptors, and TFs, respectively. **d,** Fibroblast–malignant cell bidirectional signaling shared by more than half of epithelial tumors. **e,** Endothelial–malignant cell bidirectional signaling shared by more than half of epithelial tumors. **f,** T cell–malignant cell bidirectional signaling shared by more than

half of epithelial tumors.

As shown in Figs. 5c–f, non-malignant cell-derived signaling involved 21 ligands targeting 19 malignant-cell receptors and regulating 28 TFs, whereas malignant cell-derived signaling comprised 17 ligands targeting 35 receptors across non-malignant cell types and influencing 57 TFs (Supplementary Table 7). Literature evidence supported the predicted cellular origins and functions of these ligands (Supplementary Table 8). Tumor-derived ligands included regulators of angiogenesis, immune suppression, and tumor growth (ADM[31], GDF15[32], VEGFA[33], LGALS1[34] and LGALS3[35]), whereas non-malignant cell ligands reflected canonical tumor microenvironment functions such as fibroblast-driven epithelial–mesenchymal transition (TGF-β1[36], CXCL12[37], COL1A1[38], and COL3A1[39]), macrophage-mediated angiogenesis and invasion (VEGFA[40], CXCL8[41]), and endothelial regulation of extracellular matrix remodeling (MMP-2[42], THBS1[43]). These results support the biological validity of CCPA-based consensus pathway analysis.

Consensus pathways across epithelial tumors prioritized multiple drug repurposing candidates, including 17 drugs targeting 14 non-malignant cell ligands, 10 drugs targeting 8 malignant-cell receptors, 18 drugs targeting 12 malignant-cell ligands, and 26 drugs targeting 21 receptors on non-malignant cells (Extended Data Fig. 9). Because all predicted targets were present in LUAD and PDAC, we evaluated representative drugs in LUAD A549 and PDAC AsPC-1 cells. As in vitro systems lack a complete tumor microenvironment, we focused on drugs targeting malignant-cell receptors (Fig. 6a).

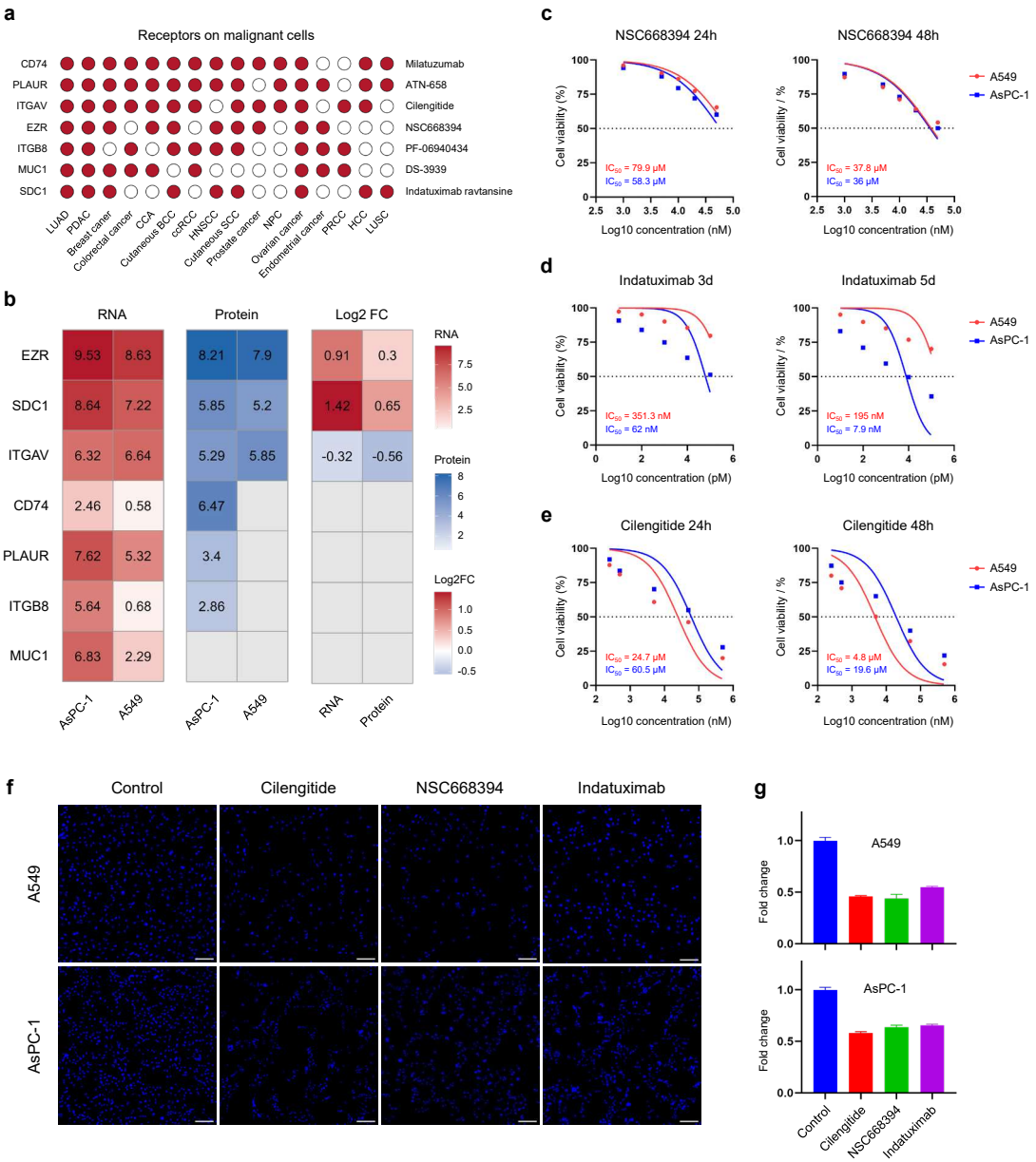


Fig. 6 | Drug repurposing for epithelial tumors. a, Malignant-cell receptors (left) and their corresponding drugs (right) shared by more than half of the 16 epithelial tumors; hotspots indicate receptor presence across cancers. **b,** Log2-transformed RNA (red) and protein (blue) expression of seven receptors in AsPC-1 and A549 cells, and log2 fold changes between AsPC-1 and A549 for receptors detected at both RNA and protein levels (left: RNA fold changes; right: protein fold changes). **c–e,** Dose–response curves of NSC668394 (c), indatuximab ravtansine (d), and cilengitide (e) in A549 and AsPC-1 cells at the indicated treatment durations. **f,** Representative fluorescence images of A549 and AsPC-1 cells treated with each drug at its IC_{50} concentration or left untreated. **g,** Quantification of cells with intact, morphologically normal nuclei. Error bars indicate

standard deviation calculated from five randomly selected non-overlapping fields.

Based on The Human Protein Atlas[30], we selected three receptors expressed at both RNA and protein levels in both A549 and AsPC-1 cell lines (Fig. 6b): EZR (targeted by NSC668394), SDC1 (targeted by indatuximab ravtansine), and ITGAV (targeted by cilengitide). Drug concentrations and treatment durations were determined according to the pharmacological properties of each compound (Methods 4). Dose–response analyses showed that all three drugs effectively reduced cell viability in both cell lines (Figs. 6c–e). Notably, drug efficacy was correlated with target expression levels: NSC668394, whose target EZR exhibited the highest RNA and protein abundance in both cell lines (Fig. 6b), showed comparable efficacy in A549 and AsPC-1 cells (Fig. 6c); indatuximab ravtansine, whose target SDC1 was more highly expressed in AsPC-1 cells, was correspondingly more effective in AsPC-1 (Fig. 6d); and cilengitide, whose target ITGAV was more highly expressed in A549 cells, showed greater efficacy in A549 (Fig. 6e).

To further validate these cytotoxic effects, A549 and AsPC-1 cells were treated with cilengitide (48 h), NSC668394 (48 h), and indatuximab ravtansine (3 days) at their respective IC_{50} concentrations, followed by quantification of cells with intact and morphologically normal nuclei (Methods 4.3). Across all treatment groups, cell numbers were reduced by approximately 50% relative to controls (Figs. 6f,g), indicating stable and reproducible drug efficacy. Collectively, these results indicate that CCPA can provide biologically informed hypotheses for drug repurposing by integrating cross-tumor signaling pathway evidence, while further validation in more

complex models will be required to assess translational relevance.

Discussion

Intra- and intercellular signaling networks are among the most complex systems in biology, characterized by redundancy, dynamic rewiring across biological states, and cyclic feedback without clear starting or end points. To cope with this complexity, existing computational models have typically adopted simplified representations: some focus on intercellular ligand–receptor interactions, others on intracellular TF–target gene regulation, while a limited number of approaches attempt to connect these layers. Nevertheless, key challenges remain.

The first challenge lies in integrating the complementary strengths of multimodal single-cell data. SigFormer addresses this through a flexible and scalable framework. When applied to scRNA-seq data from individual samples, SigFormer outperforms existing methods across multiple signaling stages. For multi-sample datasets, it performs sample-specific pathway inference followed by cross-sample integration to derive consensus signaling pathways (Methods 1.4). Therapeutic strategies targeting these recurrent signaling components may therefore be more broadly applicable across disease subtypes and stages. Moreover, we show that incorporating matched scProteomics and scATAC-seq data enhances biological support for scRNA-seq-inferred pathways, and that introducing spatial constraints specifically improves inference of membrane-bound ligand-mediated signaling.

The second challenge is reconstructing multi-step, multidimensional signal transduction from activated receptors to downstream TFs. Although SigFormer does

not fully resolve this challenge, it takes an important step forward by incorporating six classes of intracellular signaling interactions within a graph Transformer framework. This strategy substantially improved signal inference from receptors to TFs, increasing recall by 173% and F1 score by 46%. Although we primarily evaluated the therapeutic potential of predicted ligands and receptors—targets of most existing drugs—these intermediate interactions are equally important, as they expand cell-specific network topology, facilitate more informative cellular representations, and strengthen inferred associations between receptors and TFs. Future work will further investigate their roles in modeling cellular perturbations and prioritizing therapeutic targets.

Finally, we constructed the CCPA, a comprehensive resource of signaling pathways spanning 23 TME cell types and 28 common human cancers. CCPA provides high-confidence candidate signaling components together with their cellular origins and pathway contexts, and enables cross-tumor comparisons to uncover conserved signaling mechanisms and potential drug repurposing opportunities. Because current scProteomics and scATAC-seq data remain limited in tumor and cell-type coverage, most CCPA pathways were inferred from scRNA-seq data alone. Continued refinement and annotation of CCPA using expanding multi-omics datasets, curated molecular evidence, and perturbation-based analyses—both experimental and computational—will further enhance its accuracy and translational utility.

Methods

1. The SigFormer workflow

Based on user-provided scRNA-seq data with cell type annotations, SigFormer infers

bidirectional signaling pathways between a user-specified index cell type (e.g., malignant cells in the TME) and other cell types through a three-step framework: (1) cell-specific network inference, (2) cell representation learning, and (3) transcellular signaling pathway reconstruction (Fig. 1). Optional spatial, scProteomics, and scATAC-seq data can further refine the inferred pathways by contributing multimodal evidence.

1.1 Cell-specific network inference

To begin, we constructed a curated library of intracellular signaling events, comprising TF–target interactions from TRRUST[44], TFLink[45], hTFtarget[46], HTRIdb[47], and RegNetwork[48], as well as six categories of signaling interactions—protein–protein interactions, catalytic reactions, transport regulation, state-change control, protein-complex formation, and phosphorylation regulation—from STRING[49], BIOGRID[50], HuRI[51], MINT[52], and Pathway Commons[53]. These interactions can link receptors to TFs through multi-step signaling cascades. To ensure reliability, we retained only interactions supported by at least two independent databases. The full set of intracellular signaling events is available with the SigFormer toolkit, under “reference library”.

Given preprocessed scRNA-seq data (including cell and gene filtering, cell count normalization, log transformation, and cell type annotation) containing M genes across N cells, SigFormer first selects the top H (5000 by default) highly variable genes (HVGs). Based on these HVGs, a background signaling network is constructed by linking gene pairs that are connected in the curated intracellular signaling library (Stage 1 in Fig. 1). For each cell, a cell-specific network is derived by extracting a subnetwork composed of the top k (500 by default) genes expressed in that cell from

the background network. Notably, these cell-specific networks represent intermediate results rather than final outputs, as they only suggest potential signaling events within individual cells. Next, the N cell-specific networks are input into a graph Transformer model, which learns graph representations that capture both within-cell-type similarity and between-cell-type heterogeneity, ultimately reconstructing cell type-specific signaling networks.

1.2 Cell representation learning

All cell-specific networks are input into a graph Transformer-based multi-task autoencoder to learn latent embeddings at both the node (gene) and graph (cell) levels, capturing the functional state of each cell (Stage 2 in Fig. 1). The encoder uses graph Transformer layers to generate latent node embeddings (32 dimensions by default), which model the functional context of genes by dynamically attending to their interactions. These node embeddings serve two purposes: network reconstruction and cell type prediction.

For network reconstruction, the decoder uses inner-product decoders to recover network topology by scoring potential edges between node pairs based on their latent embeddings. To emphasize intercellular differences, each cell-specific network is converted into a weighted graph: edges present in the network are assigned a weight of 1, while edges absent from the network but present in other cells are assigned a weight of 0. The decoder is trained to reconstruct both types of edges. Incorporating these zero-weight edges, which appear in some cells but not others, enables the model to better capture intercellular heterogeneity.

For, a graph-level embedding is obtained by applying global max-pooling to the node embeddings of each cell and passing the pooled representation through a feed-forward neural network. This encourages the latent space to encode features that are shared within cell types but distinct across them, enabling the decoder to reconstruct cell type-specific topologies. Because this classifier is used only to guide representation learning—not to predict cell types in external datasets—techniques such as cross-validation, which are intended for assessing generalization performance, are not applied. Training is performed in a multi-task learning framework that integrates two objectives: (1) accurate reconstruction of network edges and (2) prediction of cell type. The loss function combines reconstruction error and classification error:

$$\Lambda = \alpha \cdot \Lambda_{recon_0} + \beta \cdot \Lambda_{recon_1} + \gamma \cdot \Lambda_{class},$$

Λ_{recon_0} and Λ_{recon_1} are binary cross-entropy losses for reconstructing edges with weights 0 and 1, respectively. Λ_{class} is the cross-entropy loss for cell classification. α, β, γ are weighting coefficients set to 0.5, 0.5, and 1.0 respectively. Joint optimization of these tasks ensures that node embeddings are structurally meaningful while graph-level embeddings remain biologically discriminative. The model is trained using the Adam optimizer (default learning rate = 0.001) for N (default 20) epochs. Performance is evaluated using accuracy, macro-F1, micro-F1, and weighted-F1 scores.

1.3 Transcellular pathway reconstruction

Before pathway inference, SigFormer requires users to specify an index cell type. Index cells act as both senders and receivers when establishing signaling pathways with other cell types. When spatial information is available, it is used to identify the k nearest

neighbors (default $k = 10$) for each index cell (Fig. 3a). All index cells and their corresponding neighbors are then used for pathway inference. When spatial data are unavailable, all cells in the scRNA-seq dataset are included in the inference process.

1.3.1 Sender and receiver network reconstruction

For a given cell type A containing cells $\{cell_1, \dots, cell_k\}$, SigFormer first computes the trimmed mean of their node embedding matrices $\{Z_1, \dots, Z_k\}$, yielding a joint embedding matrix Z^* of dimension 5000×32:

$$Z^* = TrimmedMean_{i=1}^{k}(Z_i, \alpha = 0.1),$$

where 5000 denotes the number of HVGs, 32 denotes the dimensionality of node embeddings, and $\alpha = 0.1$ denotes the 10% trimming ratio applied element-wise across $\{Z_1, \dots, Z_k\}$. The joint embedding matrix is then passed through an inner product decoder to generate a 5000×5000 adjacency matrix. High-probability interactions (greater than 80% by default) are subsequently extracted from the adjacency matrix to generate the signaling network of that cell type. Using this approach, SigFormer reconstructs the signaling networks of sender and receiver cell types, respectively (Stage 3 in Fig. 1).

1.3.2 Virtual knockout of receiver-cell receptors

Based on sender and receiver networks, SigFormer identifies all candidate ligand–receptor pairs, with ligands derived from the sender cell network and receptors from the receiver cell network, with the aid of a curated ligand–receptor library. This library, available in the SigFormer toolkit, contains 12,276 ligand–receptor pairs supported by at least two independent databases, including CellPhoneDB, CellTalkDB, OmniPath,

and the Lewis Lab at UCSD (https://github.com/LewisLabUCSD/Ligand-Receptor-Pairs).

For receiver cell group $\{C_1, \dots, C_k\}$, SigFormer performs in silico receptor knockout in their graphs $\{G_1, \dots, G_k\}$ to identify receptor-associated downstream genes. For each receiver cell, the trained encoder $f(\cdot)$ maps the original graph G to a node embedding matrix Z. For a given receptor r, SigFormer removes all edges incident to r from G to generate a perturbed graph $G^{(-r)}$, and computes the corresponding perturbed embedding $Z^{(-r)}$. Both Z and $Z^{(-r)}$ are matrices of dimension 5000×32 with each row representing the embedding vector for each gene. The impact of receptor knockout on gene $i \in \{0, \dots, 4999\}$ is quantified as the Euclidean distance between the original and perturbed embeddings:

$$d_i = \left\| Z_i - Z_i^{(-r)} \right\|_2.$$

SigFormer then summarizes the perturbation effect of receptor r on gene i across all k receiver cells by:

$$D_i = \frac{1}{k} \sum_{c=1}^{k} d_i^{(c)}.$$

To assess significance, SigFormer constructs an empirical null distribution using 500 random knockouts. In each trial $b \in \{1, \dots, 500\}$, the same number of edges removed in the receptor knockout is randomly deleted from each matched receiver-cell graph, and the embedding shift $D_i^{(b)}$ is calculated in the same way. If D_i exceeds 95% random-knockout background, then gene i is identified as a significantly perturbed gene for receptor r. All significantly perturbed genes for receptor r are subsequently used to constrain r-derived intracellular path reconstruction.

1.3.3 Sender–receiver pathway inference

For each receptor r, SigFormer retrieves its significantly perturbed gene set PG_r identified by in silico knockout and prunes the integrated receiver-cell signaling network (generated in 1.3.1) to r-specific subgraph:

$$G_{receiver}^{sub} = G_{receiver}[\{r\} \cup PG_r].$$

Within this receptor-derived subnetwork, SigFormer searches for shortest directed paths from r to each perturbed $TF \in PG_r$, yielding intracellular signaling paths of the form

$$R \to m_1 \to \cdots \to m_i \to TF,$$

where $m_1, \ldots, m_i$ denote signaling mediators. Because all mediators and terminal TFs are required to lie within the receptor-derived perturbed subnetwork, the inferred intracellular routes are explicitly supported by receptor knockout. Finally, SigFormer appends upstream ligands and downstream target genes to reconstruct full transcellular signaling pathways:

$$L \to R \to (m_1 \to \cdots \to m_i) \to TF \to \{TG_1, \ldots, TG_j\}.$$

This perturbation-constrained strategy restricts pathway reconstruction to receptor-supported downstream routes, thereby improving the functional specificity and mechanistic interpretability of inferred transcellular signaling pathways.

1.4 Sample-specific inference and integration

Because both gene expression profiles and network wiring vary substantially across biological contexts and states, the signaling network of each individual sample is inherently context specific. Accordingly, SigFormer is designed to perform signaling

inference at the single-sample level. When scRNA-seq input data contain multiple samples, SigFormer assumes that these samples originate from a similar biological state (e.g., the same tumor type) and integrates their results to derive cross-sample consensus signaling pathways.

For a dataset containing N samples, consensus signaling pathways are constructed by retaining only those signaling interactions observed in more than T samples:

$$T = \min(10, N/2).$$

This criterion avoids overly stringent filtering in datasets with large numbers of replicates while still enforcing cross-sample consistency. Collectively, this sample-specific design enables flexible integration of multiple samples from the same biological state.

1.5 Pathway refinement through multi-omics integration

When scProteomics or scATAC-seq data with matched cell types are available, SigFormer prunes the pathways inferred from scRNA-seq by retaining sender cell ligands and receiver cell receptors supported by protein-level evidence and receiver cell TFs supported by chromatin accessibility. As a preparatory step, the scProteomics and scATAC-seq data need to be preprocessed to identify cell-type-specific proteins and activated TFs. The preprocessing procedures vary depending on the data source and type. Details for scProteomics and scATAC-seq data collection and preprocessing are described in Methods 3 and 4, respectively.

Given a protein list P_A expressed in cell type A, a protein list P_B expressed in cell type B, and a TF list TF_B active in cell type B, SigFormer refines A→B pathways by

retaining only those in which the ligand is present in P_A, the receptor is present in P_B, and the TF is present in TF_B. Because current scProteomics and scATAC-seq datasets are much smaller in scale and cover fewer cell types than scRNA-seq, only a subset of scRNA-seq-inferred pathways can receive multi-omics support. Therefore, multi-omics integration is implemented as an optional refinement module, whereas scRNA-seq-based pathway inference remains essential for analyzing cell types lacking complementary scProteomics or scATAC-seq data.

2. Evaluation experiments

To evaluate the performance of SigFormer, we benchmarked it against nine state-of-the-art methods, including four representative tools (CellCall, CellPhoneDB, and ICELLNet, and LRLoop) for intercellular ligand–receptor inference, and five methods (NicheNet, exFINDER, scSeqComm, CellCall, and scMLnet) that infer both intercellular ligand–receptor interactions and their downstream intracellular signaling pathways.

2.1 Evaluation data

The evaluation data comprise scRNA-seq datasets from eight cancer types in the 3CA database, including glioblastoma, leukemia, pancreatic cancer, breast cancer, colorectal cancer, lung cancer, melanoma, and ovarian cancer. Among numerous samples for each cancer type, we selected the sample with the lowest standard deviation in cell counts across cell types to minimize the influence of rare cell populations on the evaluation results. An overview of each tumor sample and the corresponding cell type statistics are shown in Fig. 2a.

For the gene expression matrix from each sample, we conducted per-cell normalization and logarithmic transformation using 'normalize_total' and 'log1p' in Scanpy, and identified the top 5,000 HVGs for pathway inference benchmark. To assess the impact of spatial constraints on pathway inference, Nicheformer[22] was used to generate a spatial embedding matrix of $N \times 512$ for each dataset, where N denotes the number of cells and 512 denotes the dimensionality of spatial embeddings. The processed scRNA-seq datasets with cell type labels and predicted spatial embeddings (h5ad format) are provided with the SigFormer toolkit under "test data/scRNA-seq data".

2.2 Evaluation metrics

By applying SigFormer and existing methods to scRNA-seq data from each cancer type, we inferred bidirectional signaling pathways between malignant cells and each non-malignant cell type. Detailed descriptions and parameter settings for all methods are provided in Supplementary Methods. For systematic comparison, pathways inferred by each method were decomposed into five key stages—ligand, receptor, receptor-to-TF mediator, TF, and target gene—and recall, precision, and F1 score were computed separately for each stage.

To construct reference sets for evaluation, we collected ligands, receptors, mediators, TFs, and target genes with strong evidence of involvement in specific cancer types. Using glioblastoma as an example, reference ligand–receptor pairs were obtained from two manually curated cancer ligand–receptor databases, CCIDB[54] and TICCom[55]. Reference receptor-to-TF signaling interactions were compiled from multiple sources, including cancer protein–protein interaction databases APPIC[56] and ACSN[57], pan-

cancer and glioblastoma-related pathways from Pathway Commons[53], and curated cancer signaling networks from the SIGNOR Disease Browser[58].

Reference TF–target regulatory interactions were assembled from two categories of resources. The first category includes TF-Marker[59] and Tfcancer[60], which curate TF–target interactions in human cancers; glioblastoma-associated TFs and targets were directly retrieved from these databases. The second category comprises general TF–target interaction databases, including TRRUST[44], TFLink[45], hTFtarget[46], HTRIdb[47], and RegNetwork[48]. To extract glioblastoma-relevant interactions from these broader resources, we leveraged RNA-seq data from The Cancer Genome Atlas (TCGA)[61]. Specifically, we identified genes consistently expressed across all 175 glioblastoma patients, denoted as G_{all}, and retained only TF–target pairs in which both genes belonged to G_{all}. TF–target interactions from both categories were subsequently integrated into a unified reference set of glioblastoma-related regulatory pairs.

Reference ligands, receptors, mediators, TFs, and target genes were further integrated with glioblastoma scRNA-seq data to define positive and negative evaluation sets. We first selected the top 5,000 HVGs as background genes. The intersections between the reference sets and the background genes were defined as positive ligands, receptors, mediators, TFs, and target genes, respectively. Positive sets for the remaining seven cancer types were generated using the same procedure (see Supplementary Tables 1–5 for complete gene lists). Let L denote the ligands predicted by a given pathway inference method and L_p denote the positive ligand set. The recall, precision, and F1 score are defined as follows:

$$Recall = \frac{|L_p \cap L|}{|L_p|},\ Precision = \frac{|L_p \cap L|}{|L|},\ F1\ score = \frac{2 \times Recall \times Precision}{Recall + Precision}.$$

This evaluation strategy assigns all non-positive background genes to the negative set. This constitutes a stringent criterion, as many such cases may reflect incomplete biological knowledge rather than true false discoveries. As a result, the reported precision values likely underestimate the true precision. Importantly, this bias is applied uniformly across methods and therefore does not affect their relative performance rankings. The same approach was used to calculate the recall, precision, and F1 scores for receptors, mediators, TFs, and target genes.

3. Pathway inference with multi-omics data

3.1 scRNA-seq data collection and processing

We downloaded scRNA-seq datasets from 740 human tumor samples available in the Curated Cancer Cell Atlas (3CA, https://www.weizmann.ac.il/sites/3CA/). The 3CA resource provides high-quality data with standardized preprocessing and cell type annotation, which are directly applicable for large-scale transcellular pathway inference. Each sample in 3CA contains four data files: (1) a sample metadata file describing sample origin, tumor type, and patient information; (2) a cell metadata file including cell ID, cell type, and sample of origin; (3) a gene list; and (4) a sparse gene expression matrix.

Based on sample metadata, we classified 740 samples into 41 tumor categories, including 22 primary tumors, 13 metastatic tumors, and 6 hematologic tumors (Extended Data Fig. 4b). Each sample was subjected to standardized preprocessing, including cell and gene filtering, library-size normalization, and log transformation of

gene expression values. We then retained only samples containing malignant cells and at least one non-malignant cell type. Given that samples from the same tumor type may correspond to distinct molecular subtypes or disease states, SigFormer was applied to each remaining sample to infer bidirectional signaling pathways between malignant cells and each non-malignant cell type. For multiple samples from the same tumor type, cross-sample consensus pathways were constructed following the procedure in Methods 1.4.

3.2 scProteomics data collection and processing

We downloaded human cancer scProteomics datasets from SPDB[14] (https://scproteomicsdb.com/). For each sample, SPDB provides processed cell-specific protein expression data in RDS format. Using the R function readRDS, we identified the cell types present in each sample and extracted proteins expressed in each cell type (intensity > 0). For tumor types with multiple samples, we identified proteins consistently expressed in the same cell type across more than half of the samples. In total, this approach yielded proteins expressed in 16 cell types across 17 tumor types (blue squares in Extended Data Fig. 4b). The processed data are provided with the SigFormer toolkit under “test data/scProteomics data”.

3.3 scATAC-seq data collection and processing

We downloaded scATAC-seq footprint data of human tumor cell lines from ATACdb[15] (https://bio.liclab.net/ATACdb/). Among the 41 tumor types profiled by scRNA-seq, scATAC-seq data were available for malignant cells in 27 tumor types (green triangles in Extended Data Fig. 4b). To identify potentially activated TFs, we used TF footprint

information provided by ATACdb, which includes the number of binding sites, Protection_score, and TC_score. TFs with at least one binding site were retained, and Protection_score and TC_score were Z-score normalized within each sample, then summed to generate a combined activity score. The top 50 TFs with the highest scores were considered potentially activated in that sample. For each tumor type, we retained TFs that were activated in more than half of the samples. This step ensures that cross-sample consensus signaling pathways for each cancer are refined based on TFs that show consistent activation across samples. The processed data are provided with the SigFormer toolkit under "test data/scATAC-seq data".

3.4 Evaluation of multi-omics-supported pathways

To evaluate the contribution of multi-omics integration, we tested whether ligands and receptors supported by proteomics, as well as TFs supported by scATAC-seq, exhibit greater functional impact and reliability than those inferred exclusively from scRNA-seq data.

For a given cancer type, let $L = \{l_1, \dots, l_n\}$ denote the set of ligands inferred from scRNA-seq data, and let $L^+ = \{l_1, \dots, l_k\}$ denote the subset supported by proteomics. We defined the number of receptors bound by each ligand as its impact score. The impact scores of proteomics-supported ligands and remaining ligands are denoted by S^+ and S^-, respectively. For each cancer type, we computed the mean impact scores of S^+ and S^-. Let M^+ and M^- denote the mean impact scores of proteomics-supported ligands and other ligands across cancers. A paired t-test was performed to assess whether the difference between M^+ and M^- was statistically significant.

Using the same framework, we defined the number of TFs regulated by each receptor and the number of target genes regulated by each TF as their respective impact scores, and applied identical statistical procedures to evaluate receptors and TFs.

To assess reliability, we defined the frequency with which each ligand (or receptor) appears in tumor cell–cell interaction databases (CCIDB[54] and TICCom[55]) as its confidence score. The same analytical strategy was used to test whether proteomics-supported ligands and receptors exhibit significantly higher confidence scores than their scRNA-seq-only counterparts. Similarly, for TFs, we defined the frequency of occurrence in the tumor TF–target databases described in Methods 2.2 as the confidence score and compared scATAC-seq-supported TFs with other TFs using the same statistical framework.

4. Cell experiments

4.1 Cell culture

A549 and AsPC-1 cells were cultured in DMEM/F12 and RPMI-1640 media, respectively, supplemented with 10% fetal bovine serum (FBS) and 1% penicillin–streptomycin. Cells were maintained at 37 °C in a humidified incubator with 5% CO_2. Prior to each experiment, cells were washed twice with phosphate-buffered saline (PBS, pH 7.4), detached using 0.25% trypsin–EDTA, and centrifuged at 800 × g for 5 min. The supernatant was discarded, and the cell pellet was resuspended in fresh complete medium for subsequent experiments.

4.2 CCK-8 Assay

A549 and AsPC-1 cells were seeded into 96-well plates at a density of approximately

5,000 cells per well and incubated for 24 h at 37 °C with 5% CO_2 to allow cell attachment. Cilengitide, Indatuximab ravtansine, and NSC668394 were purchased from MedChemExpress. After removing the culture medium, cells were treated with drug solutions prepared in complete medium at the indicated concentrations. Drug treatment conditions were selected based on the pharmacological properties of each compound. Cilengitide was administered for 24 and 48 h at concentrations of 0, 1, 5, 10, 20, and 30 μM. NSC668394 was tested for 24 and 48 h at concentrations of 0, 2, 4, 6, 8, and 10 μM. Indatuximab ravtansine was applied for 3 and 5 days at concentrations of 0, 1, 2, 4, 8, and 10 nM. Each concentration-time combination was tested in triplicate for each cell line.

Following drug treatment, the medium was aspirated and replaced with 100 μL of CCK-8 working solution, prepared by diluting the reagent 1:10 (v/v) in complete medium. Plates were incubated at 37 °C for 2 h, and absorbance was measured at 450 nm using a multifunctional microplate reader. Cell viability was calculated relative to untreated controls (set to 100%). Dose–response curves were generated, and IC_{50} values were determined using GraphPad Prism 10.

4.3 Hoechst 33342 Staining and Quantitative Analysis

To further assess drug-induced cytotoxicity, A549 and AsPC-1 cells were treated with Cilengitide, NSC668394, and Indatuximab ravtansine at their respective IC_{50} concentrations for 48 h, 48 h, and 3 days. After treatment, cells were stained with the cell-permeable nuclear dye Hoechst 33342 for morphological and quantitative analysis. Briefly, culture medium was removed, and cells were washed twice with pre-chilled

PBS. Cells were then incubated with Hoechst 33342 working solution (5 μg/mL in PBS) at 37 °C for 15 min in the dark. After staining, cells were washed twice with PBS to remove excess dye. Fluorescence images were acquired using a fluorescence microscope. For each condition, including drug-treated and control groups, at least five randomly selected non-overlapping fields were imaged. Quantitative analysis was performed using ImageJ software by counting cells with intact and morphologically normal nuclei. Cell viability was calculated as:

$$Viability\ (\%) = \frac{mean\ cell\ count\ in\ treated\ group}{mean\ cell\ count\ in\ control\ group} \times 100.$$

All experiments were performed in three independent biological replicates, and data are presented as mean ± standard deviation (SD).

Data Availability

The pan-cancer scRNA-seq datasets were downloaded from the Curated Cancer Cell Atlas (3CA, https://www.weizmann.ac.il/sites/3CA/). The matched scProteomics and scATAC-seq footprint data were sourced from SPDB (https://scproteomicsdb.com/) and ATACdb (https://bio.liclab.net/ATACdb/), respectively. All CCPA pathways are publicly available at https://ccpa.ncpsb.org.cn.

Code Availability

The SigFormer framework is implemented in Python and is released at GitHub (https://github.com/free1234hm/SigFormer) with test data and detailed instructions.

Acknowledgements

We thank National Center for Protein Sciences (Beijing) Bioinformatics Platform for

providing computing resources for data processing and model training. Additionally, we thank members of the Zhu laboratory for insightful discussions and feedback on this manuscript.

Author contributions

M.F.H. and X.L. designed and implemented the SigFormer model. M.F.H., X.L., Y.Z., and L.Y.W. contributed to the application and evaluation of the software. Y.Z., T.M., and J.W. developed the CCPA website. T.P.Y. and J.M. conducted the cell experiments. Y.P.Z. and H.B.C conceived and supervised the project. M.F.H., H.B.C., Y.P.Z., and J.M. drafted the manuscript, to which all the authors contributed.

Funding

This work was supported by the National Key Research and Development Program of China [2024YFA1307701], the independent research fund of the State Key Laboratory of Medical Proteomics [SKLP-Y202404], and the National Natural Science Foundation of China [32570774].

Competing interests

The authors declare no competing interests.

References

1. Sgro, A.E. et al. From intracellular signaling to population oscillations: bridging size- and time-scales in collective behavior. *Mol Syst Biol* **11**, 779 (2015).
2. Dang, Y., Grundel, D.A.J. & Youk, H. Cellular Dialogues: Cell-Cell Communication through Diffusible Molecules Yields Dynamic Spatial Patterns. *Cell Syst* **10**, 82-98 e87 (2020).
3. Stent, G.S. Cellular communication. *Sci Am* **227**, 43-51 (1972).
4. Armingol, E., Baghdassarian, H.M. & Lewis, N.E. The diversification of methods for studying cell-cell interactions and communication. *Nat Rev Genet* **25**, 381-400 (2024).
5. Bunne, C. et al. How to build the virtual cell with artificial intelligence: Priorities and

opportunities. *Cell* **187**, 7045-7063 (2024).
6. Efremova, M., Vento-Tormo, M., Teichmann, S.A. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat Protoc* **15**, 1484-1506 (2020).
7. Jin, S. et al. Inference and analysis of cell-cell communication using CellChat. *Nat Commun* **12**, 1088 (2021).
8. Antebi, Y.E., Nandagopal, N. & Elowitz, M.B. An operational view of intercellular signaling pathways. *Curr Opin Syst Biol* **1**, 16-24 (2017).
9. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat Methods* **17**, 159-162 (2020).
10. He, C., Zhou, P. & Nie, Q. exFINDER: identify external communication signals using single-cell transcriptomics data. *Nucleic Acids Res* **51**, e58 (2023).
11. Baruzzo, G., Cesaro, G. & Di Camillo, B. Identify, quantify and characterize cellular communication from single-cell RNA sequencing data with scSeqComm. *Bioinformatics* **38**, 1920-1929 (2022).
12. Fischer, D.S., Schaar, A.C. & Theis, F.J. Modeling intercellular communication in tissues using spatial graphs of cells. *Nat Biotechnol* **41**, 332-336 (2023).
13. Tyler, M. et al. The Curated Cancer Cell Atlas provides a comprehensive characterization of tumors at single-cell resolution. *Nat Cancer* **6**, 1088-1101 (2025).
14. Wang, F. et al. SPDB: a comprehensive resource and knowledgebase for proteomic data at the single-cell resolution. *Nucleic Acids Res* **52**, D562-D571 (2024).
15. Wang, F. et al. ATACdb: a comprehensive human chromatin accessibility database. *Nucleic Acids Res* **49**, D55-D64 (2021).
16. Castanza, A.S. et al. Extending support for mouse data in the Molecular Signatures Database (MSigDB). *Nat Methods* **20**, 1619-1620 (2023).
17. Hu, Y., Peng, T., Gao, L. & Tan, K. CytoTalk: De novo construction of signal transduction networks using single-cell transcriptomic data. *Sci Adv* **7** (2021).
18. Noel, F. et al. Dissection of intercellular communication using the transcriptome-based framework ICELLNET. *Nat Commun* **12**, 1089 (2021).
19. Xin, Y. et al. LRLoop: a method to predict feedback loops in cell-cell communication. *Bioinformatics* **38**, 4117-4126 (2022).
20. Zhang, Y. et al. CellCall: integrating paired ligand-receptor and transcription factor activities for cell-cell communication. *Nucleic Acids Res* **49**, 8520-8534 (2021).
21. Cheng, J., Zhang, J., Wu, Z. & Sun, X. Inferring microenvironmental regulation of gene expression from single-cell RNA sequencing data using scMLnet with an application to COVID-19. *Brief Bioinform* **22**, 988-1005 (2021).
22. Tejada-Lapuerta, A. et al. Nicheformer: a foundation model for single-cell and spatial omics. *Nat Methods* (2025).
23. Mendez-Lopez, L.F. Revisiting Epithelial Carcinogenesis. *Int J Mol Sci* **23** (2022).
24. Naoi, Y. et al. Close Spatial Interactions between Cancer Cells and Cancer-Associated Fibroblasts Suppress Antitumor Immunity. *Cancer Immunol Res* **13**, 1471-1484 (2025).
25. Maisel, B.A. et al. Spatial Metrics of Interaction between CD163-Positive Macrophages and Cancer Cells and Progression-Free Survival in Chemo-Treated Breast Cancer. *Cancers (Basel)* **14** (2022).

26. Yao, Y., Li, F., Huang, J., Jin, J. & Wang, H. Leukemia stem cell-bone marrow microenvironment interplay in acute myeloid leukemia development. *Exp Hematol Oncol* **10**, 39 (2021).
27. Hide, T. et al. Oligodendrocyte Progenitor Cells and Macrophages/Microglia Produce Glioma Stem Cell Niches at the Tumor Border. *EBioMedicine* **30**, 94-104 (2018).
28. Perez, M. et al. Melanoma cells stimulate osteoclastogenesis, c-Src expression and osteoblast cytokines. *Eur J Cancer* **37**, 629-640 (2001).
29. Uhlen, M. et al. The human secretome. *Sci Signal* **12** (2019).
30. Uhlen, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
31. Larrayoz, I.M., Martinez-Herrero, S., Garcia-Sanmartin, J., Ochoa-Callejero, L. & Martinez, A. Adrenomedullin and tumour microenvironment. *J Transl Med* **12**, 339 (2014).
32. Winter, L.M. et al. Molecular basis of GDF15 induction and suppression by drugs in cardiomyocytes and cancer cells toward precision medicine. *Sci Rep* **13**, 12061 (2023).
33. Zhang, Y. & Brekken, R.A. Direct and indirect regulation of the tumor immune microenvironment by VEGF. *J Leukoc Biol* **111**, 1269-1286 (2022).
34. Thijssen, V.L. et al. Tumor cells secrete galectin-1 to enhance endothelial cell activity. *Cancer Res* **70**, 6216-6224 (2010).
35. Raiter, A. et al. Galectin-3 secreted by triple-negative breast cancer cells regulates T cell function. *Neoplasia* **60**, 101117 (2025).
36. Yoon, H. et al. TGF-beta1-mediated transition of resident fibroblasts to cancer-associated fibroblasts promotes cancer metastasis in gastrointestinal stromal tumor. *Oncogenesis* **10**, 13 (2021).
37. Ma, Z. et al. CXCL12 alone is enough to Reprogram Normal Fibroblasts into Cancer-Associated Fibroblasts. *Cell Death Discov* **11**, 156 (2025).
38. Li, M. et al. Microenvironment remodeled by tumor and stromal cells elevates fibroblast-derived COL1A1 and facilitates ovarian cancer metastasis. *Exp Cell Res* **394**, 112153 (2020).
39. Szabo, P.M. et al. Cancer-associated fibroblasts are the main contributors to epithelial-to-mesenchymal signatures in the tumor microenvironment. *Sci Rep* **13**, 3051 (2023).
40. Lin, E.Y. et al. Vascular endothelial growth factor restores delayed tumor progression in tumors depleted of macrophages. *Mol Oncol* **1**, 288-302 (2007).
41. Tatsuno, R. et al. Pivotal role of IL-8 derived from the interaction between osteosarcoma and tumor-associated macrophages in osteosarcoma growth and metastasis via the FAK pathway. *Cell Death Dis* **15**, 108 (2024).
42. Kargozaran, H. et al. A role for endothelial-derived matrix metalloproteinase-2 in breast cancer cell transmigration across the endothelial-basement membrane barrier. *Clin Exp Metastasis* **24**, 495-502 (2007).
43. Kaur, S. et al. Functions of Thrombospondin-1 in the Tumor Microenvironment. *Int J Mol Sci* **22** (2021).
44. Han, H. et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res* **46**, D380-D386 (2018).
45. Liska, O. et al. TFLink: an integrated gateway to access transcription factor-target gene interactions for multiple species. *Database (Oxford)* **2022** (2022).
46. Zhang, Q. et al. hTFtarget: A Comprehensive Database for Regulations of Human

Transcription Factors and Their Targets. *Genomics Proteomics Bioinformatics* **18**, 120-128 (2020).

47. Bovolenta, L.A., Acencio, M.L. & Lemke, N. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics* **13**, 405 (2012).
48. Li, B., Wang, C., Wang, Y., Li, P. & Liu, Z.P. RegNetwork 2025: an integrative data repository for gene regulatory networks in human and mouse. *Nucleic Acids Res* (2025).
49. Szklarczyk, D. et al. The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Res* **53**, D730-D737 (2025).
50. Stark, C. et al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535-539 (2006).
51. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402-408 (2020).
52. Licata, L. et al. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* **40**, D857-861 (2012).
53. Rodchenkov, I. et al. Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Res* **48**, D489-D497 (2020).
54. Noh, J.Y. et al. CCIDB: a manually curated cell-cell interaction database with cell context information. *Database (Oxford)* **2023** (2023).
55. Xie, Y. et al. A global database for modeling tumor-immune cell communication. *Sci Data* **10**, 444 (2023).
56. Ahn, B. et al. The Atlas of Protein-Protein Interactions in Cancer (APPIC)-a webtool to visualize and analyze cancer subtypes. *NAR Cancer* **7**, zcae047 (2025).
57. Kuperstein, I. et al. Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis* **4**, e160 (2015).
58. Lo Surdo, P. et al. SIGNOR 4.0: the 2025 update with focus on phosphorylation data. *Nucleic Acids Res* **54**, D682-D690 (2026).
59. Xu, M. et al. TF-Marker: a comprehensive manually curated database for transcription factors and related markers in specific cell and tissue types in human. *Nucleic Acids Res* **50**, D402-D412 (2022).
60. Huang, Q. et al. Tfcancer: a manually curated database of transcription factors associated with human cancers. *Bioinformatics* **37**, 4288-4290 (2021).
61. Cancer Genome Atlas Research, N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120 (2013).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.xlsx
- SupplementaryMaterials.pdf