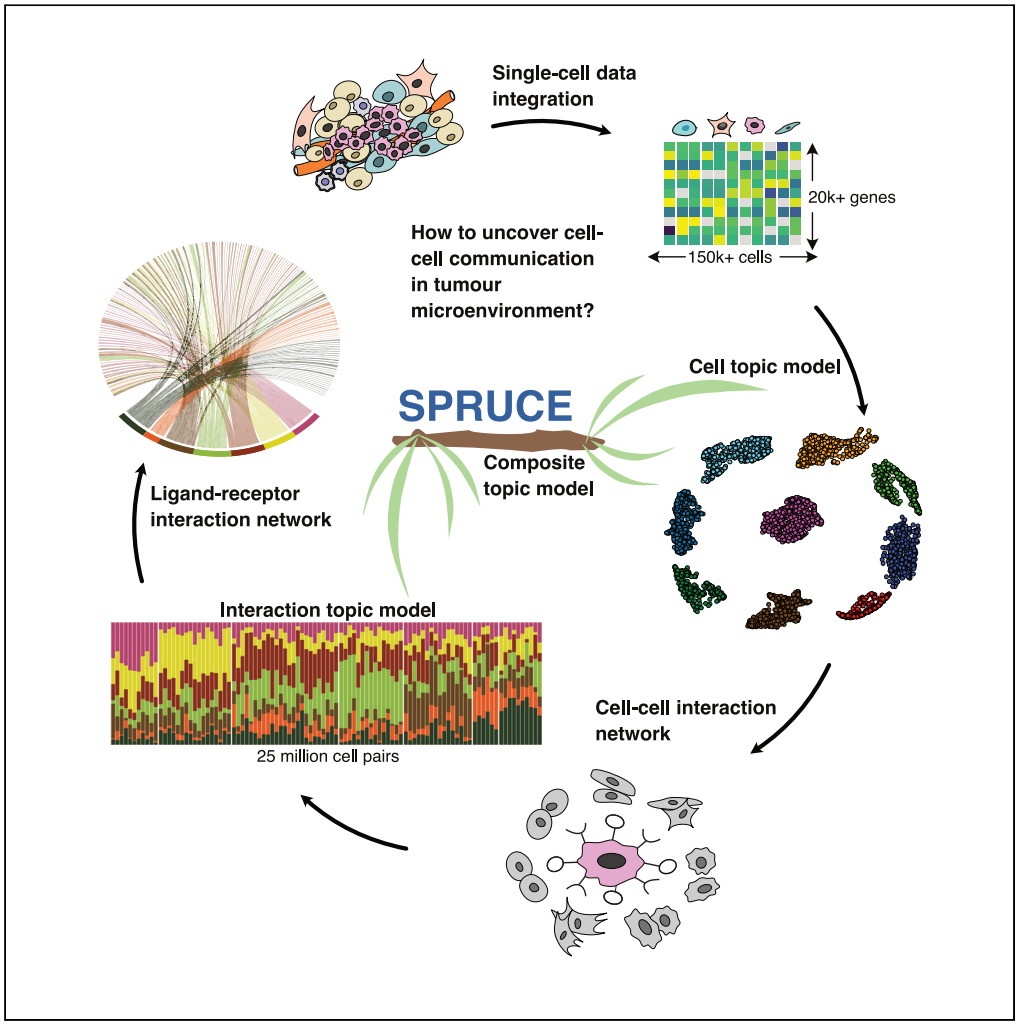
iScience

CellPress
OPEN ACCESS

Article

Single-cell pair-wise relationships untangled by composite embedding model



Sishir Subedi,
Yongjin P. Park

yongjin.park@ubc.ca

Highlights

A modeling framework to investigate intercellular communications in single-cell data

A machine learning algorithm focuses on interactions of data vectors

Interaction-driven models provide new insights into tumor heterogeneity

Subedi & Park, iScience 26, 106025
February 17, 2023
https://doi.org/10.1016/j.isci.2023.106025

iScience

CellPress
OPEN ACCESS

Article

Single-cell pair-wise relationships untangled by composite embedding model

Sishir Subedi[1,2] and Yongjin P. Park[2,3,4,5,*]

SUMMARY

In multicellular organisms, cell identity and functions are primed and refined through interactions with other surrounding cells. Here, we propose a scalable machine learning method, termed SPRUCE, which is designed to systematically ascertain common cell-cell communication patterns embedded in single-cell RNA-seq data. We applied our approach to investigate tumor microenvironments consolidating multiple breast cancer datasets and found seven frequently observed interaction signatures and underlying gene-gene interaction networks. Our results implicate that a part of tumor heterogeneity, especially within the same subtype, is better understood by differential interaction patterns rather than the static expression of known marker genes.

INTRODUCTION

The advancement in single-cell RNA-sequencing (scRNA-seq) has emerged as a new frontier in genomics. Quantification of multimodal omics at a single-cell resolution has made it possible to gain insights into different aspects of cancer biology.[1] One of the fundamental questions in cancer research is how cancer cells interact with each other in a confined heterogeneous environment such as a tumor microenvironment (TME). Studies have shown that cell-cell communication (CCC) among cell populations in the TME is crucial in cancer growth and metastatic processes.[2] Understanding the intricacies of communication among tumor and their interacting partner cells could aid in identifying a potential therapeutic avenue in cancer.

A significant technical challenge that stands in understanding the dynamics of cell-cell interactions in TME is devising a systematic approach to isolate and capture interaction signals from each interacting cell pair. A conventional approach to studying CCC involves clustering features in low-dimensional space and inferring interactions between clusters of know cell types.[3–5] Although these methods have uncovered numerous signaling mechanisms that govern cellular differentiation and pathogenesis, they assume each cluster, annotated using a limited number of marker genes, represents a cell type; hence all the cells within a cluster interact in the same manner. These methods do not account for intracluster cellular heterogeneity. Cells within a cell type may exist in multiple subtypes/states and manifest heterogeneous interaction patterns based on the type and state of the interacting partner cell, which is critical in understanding cancer progression.[2,6] In addition, interaction among cells in different contexts, such as disease states, are studied separately which loses context-specific variability information and are repetitive and computationally expensive.

Recent studies have addressed these challenges and developed methods to capture the diversity of cell interactions within the same cluster. Tensor-cell2cell[7] uses tensor-based dimensionality reduction techniques to infer context-driven CCC pattern. scTensor[8] also uses a tensor decomposition algorithm to infer many-to-many cell pair relationships as a hypergraph. These methods rely on *a priori* knowledge of cell type and aggregating cells to calculate communication scores based on the mean expression of ligand-receptor (LR) genes. SoptSC[9] calculates signaling probability between two cells based on pathway-specific LR and target genes and addresses heterogeneity of cells within the same cluster. However, the method requires a user-defined comprehensive list of pathway genes and does not scale to cohort-level studies.

Here, we present a novel computational approach termed SPRUCE, Single-cell Pair-wise Relationship Untangled by Composite Embedding, to analyze tens of millions of cell pairs in a scalable way. Adopting known ligand and receptor protein-protein interactions, we asked how and why cell pairs are localized in the proximity of latent topic space and highlighted common patterns repeatedly observed in tumor

[1]Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC, Canada

[2]BC Cancer Research, Part of Provincial Health Care Authority, Vancouver, BC, Canada

[3]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada

[4]Department of Statistics, University of British Columbia, Vancouver, BC, Canada

[5]Lead contact

*Correspondence: yongjin.park@ubc.ca

https://doi.org/10.1016/j.isci.2023.106025

microenvironments data. SPRUCE is based on an embedded topic model (ETM), a generative deep learning method built on variational autoencoder architecture, and represents single-cell vector data in low-dimension topic space with an interpretable topic-specific gene expression dictionary matrix. It has been successfully implemented in natural language processing to extract meaningful topics representing large-scale documents.[10] A recent study, scETM, showed that ETM-based techniques efficiently capture essential biological signals from sparse and heterogeneous single-cell data.[11] The key contribution of our approach is the unbiased identification of interpretable cell subtype/state across multiple datasets by characterizing LR genes-driven patterns of cell-cell interactions. Existing graph-based single-cell analysis methods often define cell-cell interaction modules as densely-connected components in a graph (an adjacency matrix). Our SPRUCE model considers cell-cell interaction patterns as a stream of edges, or a giant incidence matrix (edge by vertex or other vertex property).

RESULTS

Overview of SPRUCE model training in breast cancer study

We combined existing breast cancer datasets[12,13] and cancer-specific immune cell data[14] and constructed a comprehensive single-cell catalog for an unbiased breast cancer study, yielding a data matrix consisting of 20,288 genes and 155,913 cells. We first mapped cells from multiple datasets onto a common latent topic space (K = 50) and harmonized them based on variational autoencoder-based topic modeling, not posing additional assumptions, such as a selection bias imposed by top marker genes. Based on cosine similarity in the latent topic space between cells, we then constructed cell-cell interaction networks and performed stratified sampling so that topic-topic relationships are similarly represented in the subsequent steps in SPRUCE training. For each of these 25M+ cell pairs, we extracted gene expressions of the known 648 ligand and 672 receptor proteins[15] and used them as feature vectors for SPRUCE model training.

Multinomial probabilistic topic modeling identified 50 cell topics across 11 known cell types

We implemented a Bayesian deep learning approach to estimate embedded topic models across 155,913 cells with 50 latent dimensions (Figure 1A). We found each cell topic corresponds to a group of an average of 3118 cells (with a SD of $\pm$ 5,987) (Figure 1B). Among 50 topics, 32 of them contained more than 100 cells. The highest number of cells (24% of the dataset) was assigned to topic 37, in which 96% of cells were previously identified as immune cells (T and B cells). 98% of cancer cells from the dataset were assigned to 13 cell topics. The cancer cell proportion in 9 of 13 topics was greater than 95%. The latent cell topics with cell type annotated were visualized with UMAP, showing distinct clusters for each topic where the majority of cells belong to one of the major types of cells in the dataset (Figure 1C). These clusters show a unique set of top genes associated with each topic identified using the optimized gene loading matrix from the model (Figure 1D). We further confirmed concordance with the previous analysis conducted in original papers using cell type lineage canonical markers (Figure 1F). The estimated cell topic proportions show that the resident cell types have similar topic proportions. However, cancer cells have a different mixture of topic proportions which shows that the model identified many distinct topics of cancer cells (Figure 1E). We also tried a different number of cell topics from 10, 25, and 50 and decided to use the 50-topic model because major cell types, especially cancer cells, showed well-separated distinct clusters (Figures S1A and S1B).

Seven robust TME-specific interaction signatures were found in 25 million cell-cell pairs

We constructed LR gene expression data from 155,913 cells to construct a set of 24,790,167 cell pairs and estimated embedded interaction topic models with 25 latent dimensions (Figures 2A and 2B). The interaction topic model recapitulates two types of cell-cell communication mechanisms. First, the model captures interactions between differentially expressed ligand and receptor genes in different cell types. Second, the model considers ligand and receptor genes correlated within each cell type but may not be differentially expressed. Among 25 interaction topics representing 25 million cell pairs, seven topics (2, 4, 7, 10, 18, 22, and 24) represented 55% of the total cell pair interactions, with each topic containing 3% + cell pairs (Figure 2C). The other 18 interaction topics, each with 2% of the total cell pair interactions embedded baseline interaction signal. The most represented interaction topic was topic 22, consisting of 12% of the total cell pair interactions.

The model estimated the LR gene loadings in each interaction topic that described the relative contribution of each gene. These loadings can be ranked to identify biologically interpretable topic-specific top genes in each interaction topic (Figure 2D). Topics 22 and 24 captured immune-related interactions. Topic

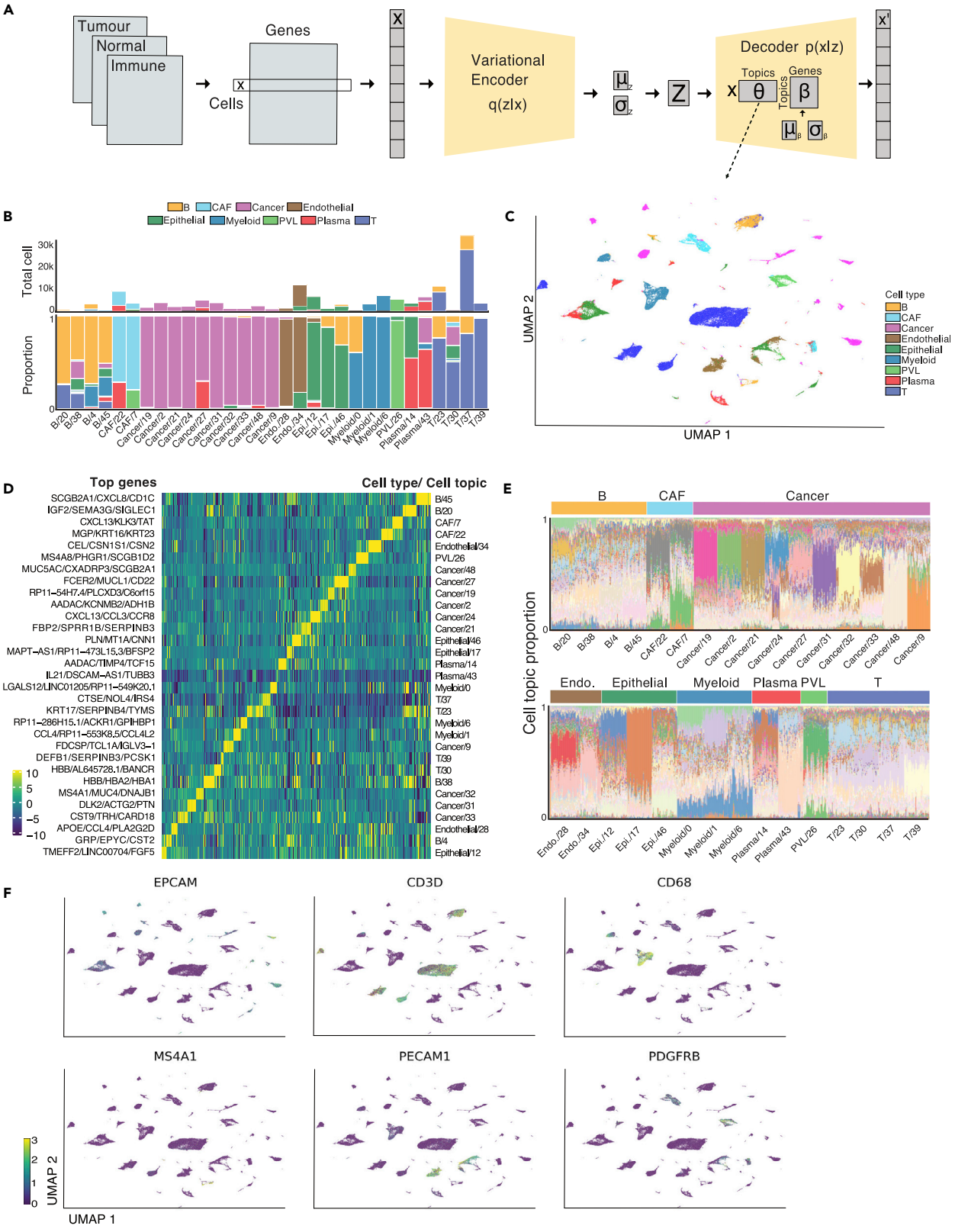

A
Tumour
Normal
Immune
Cells
Genes
x
Variational Encoder
q(z|x)
Z
Decoder p(x|z)
Topics
Genes
θ
β
x'
B
B
CAF
Cancer
Endothelial
Epithelial
Myeloid
PVL
Plasma
T
Total cell
Proportion
C
UMAP 2
UMAP 1
Cell type
D
Top genes
Cell type/ Cell topic
SCGB2A1/CXCL8/CD1C
IGF2/SEMA3G/SIGLEC1
CXCL13/KLK3/TAT
MGP/KRT16/KRT23
CEL/CSN1S1/CSN2
MS4A8/PHGR1/SCGB1D2
MUC5AC/CXADRP3/SCGB2A1
FCER2/MUCL1/CD22
RP11–54H7.4/PLCXD3/C6orf15
AADAC/KCNMB2/ADH1B
CXCL13/CCL3/CCR8
FBP2/SPRR1B/SERPINB3
PLN/MT1A/CNN1
MAPT–AS1/RP11–473L15.3/BFSP2
AADAC/TIMP4/TCF15
IL21/DSCAM–AS1/TUBB3
LGALS12/LINC01205/RP11–549K20.1
CTSE/NOL4/IRS4
KRT17/SERPINB4/TYMS
RP11–286H15.1/ACKR1/GPIHBP1
CCL4/RP11–553K8.5/CCL4L2
FDCSP/TCL1A/IGLV3–1
DEFB1/SERPINB3/PCSK1
HBB/AL645728.1/BANCR
HBB/HBA2/HBA1
MS4A1/MUC4/DNAJB1
DLK2/ACTG2/PTN
CST9/TRH/CARD18
APOE/CCL4/PLA2G2D
GRP/EPYC/CST2
TMEFF2/LINC00704/FGF5
B/45
B/20
CAF/7
CAF/22
Endothelial/34
PVL/26
Cancer/48
Cancer/27
Cancer/19
Cancer/2
Cancer/24
Cancer/21
Epithelial/46
Epithelial/17
Plasma/14
Plasma/43
Myeloid/0
T/37
T/23
Myeloid/6
Myeloid/1
Cancer/9
T/39
T/30
B/38
Cancer/32
Cancer/31
Cancer/33
Endothelial/28
B/4
Epithelial/12
E
B
CAF
Cancer
Endo.
Epithelial
Myeloid
Plasma
PVL
T
Cell topic proportion
F
EPCAM
CD3D
CD68
MS4A1
PECAM1
PDGFRB
UMAP 2
UMAP 1

Figure 1. Probabilistic topic model identifies cell topics for resident cell types and cancer subtypes

(A) Given an integrated single-cell data from breast cancer tumor microenvironment, SPRUCE cell topic model learns the latent topic representation of the cells and generates biologically interpretable topic embeddings separately over the genes.

(B) Distribution of cell types highlighting the total number of cells in each cell topic. Relative proportion of cell types in each cell topic. Cell topic is assigned to each cell based on the highest score.

(C) UMAP visualization of 50 cell topics representing an integrated dataset consisting of 155,913 cells. Each dot is a single cell, and colors represent the corresponding cell type clustered according to k-means (n = 50) algorithm on cell topic proportions and annotated using the majority voting rule with annotation from previous studies and SingleR.

(D) Heatmap of top 25 gene loadings associated with each cell topic with the total number of cells >100. Cell topics (y axis) and genes (x axis) are ordered according to hierarchical clustering (optimal leaf ordering). The cell type and topic are shown on the right y axis, while the top 3 genes for each cell topic are depicted on the left y axis.

(E) Relative proportion of cell topics from a sample(n = 50 cells per topic) of cells assigned to cell type - cell topic pairs.

(F) Log normalized (total count to 10,000 reads per cell) expression of markers genes for cell types- *EPCAM* for epithelial, *CD3D* for T-cells, *CD68* for myeloid cells, *MS4A1* for B-cells, *PECAM1* for endothelial cells, and *PDGFRB* for CAF/PVL cells.

22 was labeled a lymphoid topic because top receptor genes included the known subunit of T-Cell Receptor Complex *CD3D* and killer cell lectin-like receptors *KLRC1*, *KLRC2*, and *KLRD1*. The top ligands in this topic are *HLA-E*, *CLEC2B*, and *CLEC2DC*, which are essential known modulators in cytotoxic T cells.[16] Similarly, topic 24 was labeled a myeloid topic as top genes in this topic showed enrichment of LR genes expressed by myeloid progenitors, for example - receptors such as *CD68*, *TREM2*, and *CR1*, and ligands such as *CCL23*, *CCL18*, *CCL13*, and *C1QA*.[17]

Topics 10 and 7 represented many oncogenes mutated in cancer. For instance, Topic 10 was cancer-growth associated, and genes that play a role in cancer cell survival and growth are enriched in this topic. The top receptors in this topic are growth factor receptors such as *ERBB2*, cell proliferation and growth signaling receptor *FZD10*, and immune inhibiting signaling receptor *ADORA2A*.[18,19] Similarly, Topic 7 was a cancer-metastasis topic and genes such as *NTRK3*, known to increase the metastatic potential of cancer cells, *GRPR*, which promotes EMT, and *UNC5A*, a known regulator of cancer plasticity, are enriched in this topic.[20–22]

Moreover, Topic 18 was stroma-specific and represented genes that play an integral role in regulating the extracellular matrix (ECM) of the tumor immune microenvironment. These genes are highly expressed in cancer-associated fibroblast (CAF) and perivascular-like (PVL) cells. The top ligands in these topics are *COL1A1*, *COL1A2*, *COL3A1*, and *MMP13*, and the top receptors are *ITGA11* and *SCARA5*.[23,24] Similarly, Topic 2 is enriched with genes highly expressed in endothelial cells, thus labeled an endothelial topic. Here, ligand proteins highly expressed in endothelial cells, such as *CD34*, *ANGPT2*, and *NID2* and receptors such as *APLNR* and *ESAM* are enriched.[25] Likewise, Topic 4 enriches genes involved in TME regulation processes and is thought to facilitate cancer progression and growth. The top genes in this topic include *KISS1R*/*KISS1*, which play a complex role in both restricting and promoting cancer cell survival, *IL20RB*, which promotes immunosuppressive microenvironment, and *MMP24*, which negatively regulates the aggressiveness of cancer cells.[26]

The top LR genes in the major interaction topics show enrichment of different cell type-specific functional interactions. To confirm that each interaction topic captured cell type-specific CCC, we took a closer look at the distribution of cell types of target cells in each interaction topic for all the cells in the dataset. We found that the functional role of enriched top LR genes in each interaction topic matched with the dominant cell type of target cells in that topic (Figure 2E). For example, in the cancer-growth associated, on average, 68% of target cells for all cell types were cancer cells. Similarly, 49% of target cells in the stroma topic were CAF/PVL cells, and myeloid and T cells comprised 49% and 38% of target cells in the myeloid and lymphoid topics, respectively. For the endothelial topic, the dominant target cell type stemmed from endothelial cells, comprising 18%. In contrast, for the TME-regulation topic, both epithelial and plasms cells were dominant cells consisting of 20% and 22%, respectively.

In addition, the cell type-specific enrichment of interaction topics was further corroborated by the distribution of interaction topics in each cell type. Cancer cells with epithelial, plasma, and B cells showed heterogeneous interaction patterns compared to myeloid, T, endothelial, CAF, and PVL cell types (Figure 2F). For cancer cells, the majority of interactions belonged to the cancer-growth and cancer-metastasis topics, where many of the top genes were oncogenes. Here, 55% of the total interactions were incident with a cancer-growth process, and 17% belonged to the cancer-metastasis topic, whereas the other

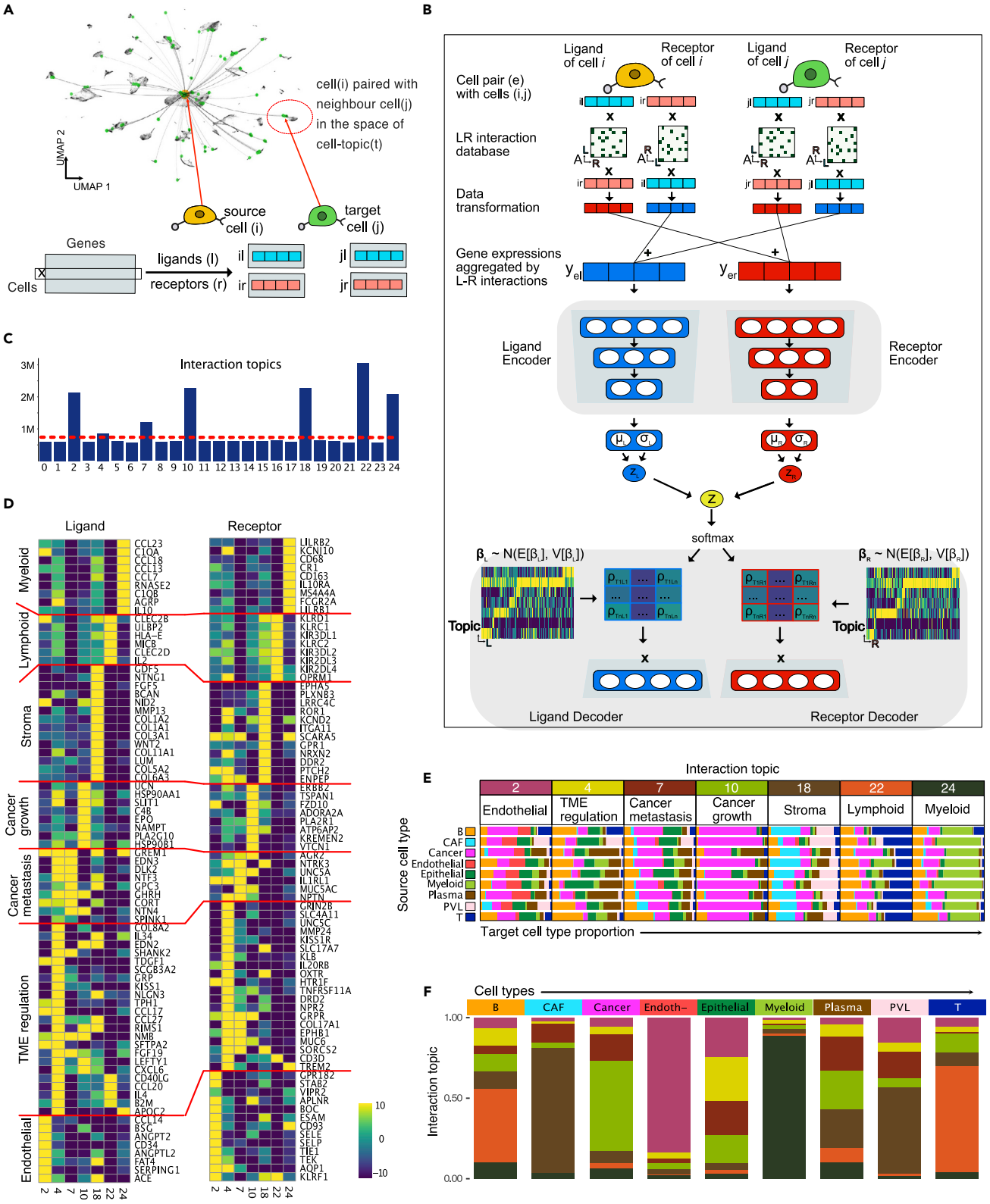

A
cell(i) paired with neighbour cell(j) in the space of cell-topic(t)
UMAP 2
UMAP 1
source cell (i)
target cell (j)
Genes
Cells
ligands (l)
receptors (r)
il
ir
jl
jr
B
Ligand of cell i
Receptor of cell i
Ligand of cell j
Receptor of cell j
Cell pair (e) with cells (i,j)
LR interaction database
Data transformation
Gene expressions aggregated by L-R interactions
Ligand Encoder
Receptor Encoder
softmax
Topic
Ligand Decoder
Receptor Decoder
C
Interaction topics
3M
2M
1M
D
Ligand
Receptor
Myeloid
Lymphoid
Stroma
Cancer growth
Cancer metastasis
TME regulation
Endothelial
E
Interaction topic
Endothelial
TME regulation
Cancer metastasis
Cancer growth
Stroma
Lymphoid
Myeloid
Source cell type
B
CAF
Cancer
Endothelial
Epithelial
Myeloid
Plasma
PVL
T
Target cell type proportion
F
Cell types
Interaction topic

Figure 2. SPRUCE model overview and the common interaction patterns identified by the model

(A) UMAP-based representation of cell pairing method where source cell is paired with four different target cells from each cell topic.

(B) Given a cell pair LR gene expression data as input, SPRUCE model transforms and aggregates interaction-driven LR data space and feeds into the neural network. The encoder learns the latent topic representation of the interaction between cells using the mixture of experts approach from ligand and receptor encoders. The decoder generates biologically interpretable topic embeddings separately over the ligand and receptor genes.

(C) The distribution of ~25 million cell pair interactions over 25 interaction topics shows seven major interaction patterns.

(D) Heatmap of top 10 gene loadings associated with seven interaction topics. Interaction topics (x axis) and LR genes (y axis) are ordered according to hierarchical clustering (optimal leaf ordering).

(E) Relative proportion of target cell type in all the source cell types associated with the major interaction patterns. Each row is source cell type (y axis) and target cell type proportions (x axis).

(F) Enrichment of interaction patterns among cell types.

five remaining topics consisted of 3–8% of interactions. Among the non-malignant cell types, the dominant interaction topic for myeloid cells was the myeloid interaction topic, and for T-cells, it was the lymphoid interaction topic. Here, 88% of myeloid cell and 65% of T cell interactions were found to be in respective interaction topics. Similarly, 83% of cell interactions with endothelial cells belonged to the endothelial topic, and 77% and 53% of interactions with CAF and PVL cells, respectively, were the stroma topic. In contrast, plasma, B, and epithelial cells showed a higher mixture of non-immune associated interaction topics.

Interaction topics provide a way to understand breast cancer heterogeneity

Next, we investigated the heterogeneity of breast cancer cells based on the unbiased transcriptomic signature captured by the cell topic model while relating the cell topics to the interaction topics characterized by the SPRUCE analysis. The interaction patterns of 25,835 cancer cells manifested all the patterns of interactions (Figure 3B). The cell topic model identified 13 cell topics for cancer cells that demonstrate a distinct pattern of interactions with their target cells (Figure 3C). As expected, the cancer growth topic was the most dominant (>57%) among 7 of 13 cell topics. For instance, 75%, 70%, and 68% of interactions for cancer cells in cell topics 24, 48, and 2 were driven by cancer-growth-related interactions. However, two cell topics were found more frequently interacting with non-cancer cells. The 66% of interactions involving (cell) Topic 9 were of stroma edges, and 60% of edges emanating from Topic 34 were assigned to endothelial interactions. Such a striking interaction heterogeneity was rarely observed in other cell types, such as T-cells, myeloid, endothelial, CAF, and PVL cell types, as they mostly interact within the same cell types. B-cells and epithelial cells, however, showed substantial variability of interaction patterns across cell topics.

Breast cancer cells are classified into subtypes based on the genomics and pathology of the disease, and different subtypes often result in markedly different clinical outcomes.[27] All three different subtypes of cancer cells indeed exhibit diverse interaction patterns where TNBC (triple-negative breast cancer) cells were more heterogeneous compared to HER2+ and ER + subtypes (Figure 3D). Here, more than 85% of interactions of the HER2+ subtype consisted of cancer-related topics, such as 82% for the cancer-growth and 5% for cancer metastasis topics. Similarly, for the ER + subtype, more than 80% of interactions were cancer-related: 61% for the cancer-growth and 20% for the cancer-metastasis topics. In contrast, TNBC subtype cells were more diverse in interactions, with 42% for cancer growth, 15% for the cancer-metastasis, 15% for stroma, and 10% for myeloid topics.

In addition, our approach identified a specific group of cells (cell topics) within these cancer subtypes that demonstrate topic-specific interaction patterns. For instance, TNBC cell topics show higher heterogeneity in interaction patterns at the cell topic level compared to the ER+ and HER2+ subtypes. The distribution of interaction patterns among cell topics is correlated with the expression pattern of LR genes enriched in each interaction topic. For example, TNBC cancer cells in cell topic 9 show higher expression of LR genes enriched in the myeloid interaction topic. In contrast, the cancer-growth LR genes are dominant among TNBC cancer cells in cell topic 24 (Figure 3E).

Interaction topics uncover underlying cancer-subtype-specific gene-gene interaction networks

Our approach also reveals topic-specific interaction patterns in the model parameter matrix (Figure 4), with which gene-gene correlation networks can be estimated (ligand versus receptor). For instance, gene networks in the cancer metastasis topic (Figure 4A, burgundy strips; Figure 4B, the first panel) consisted of a group of structural genes *CLDN4*, *LSR*, and *DSG2* involved in cell transformation and migration and

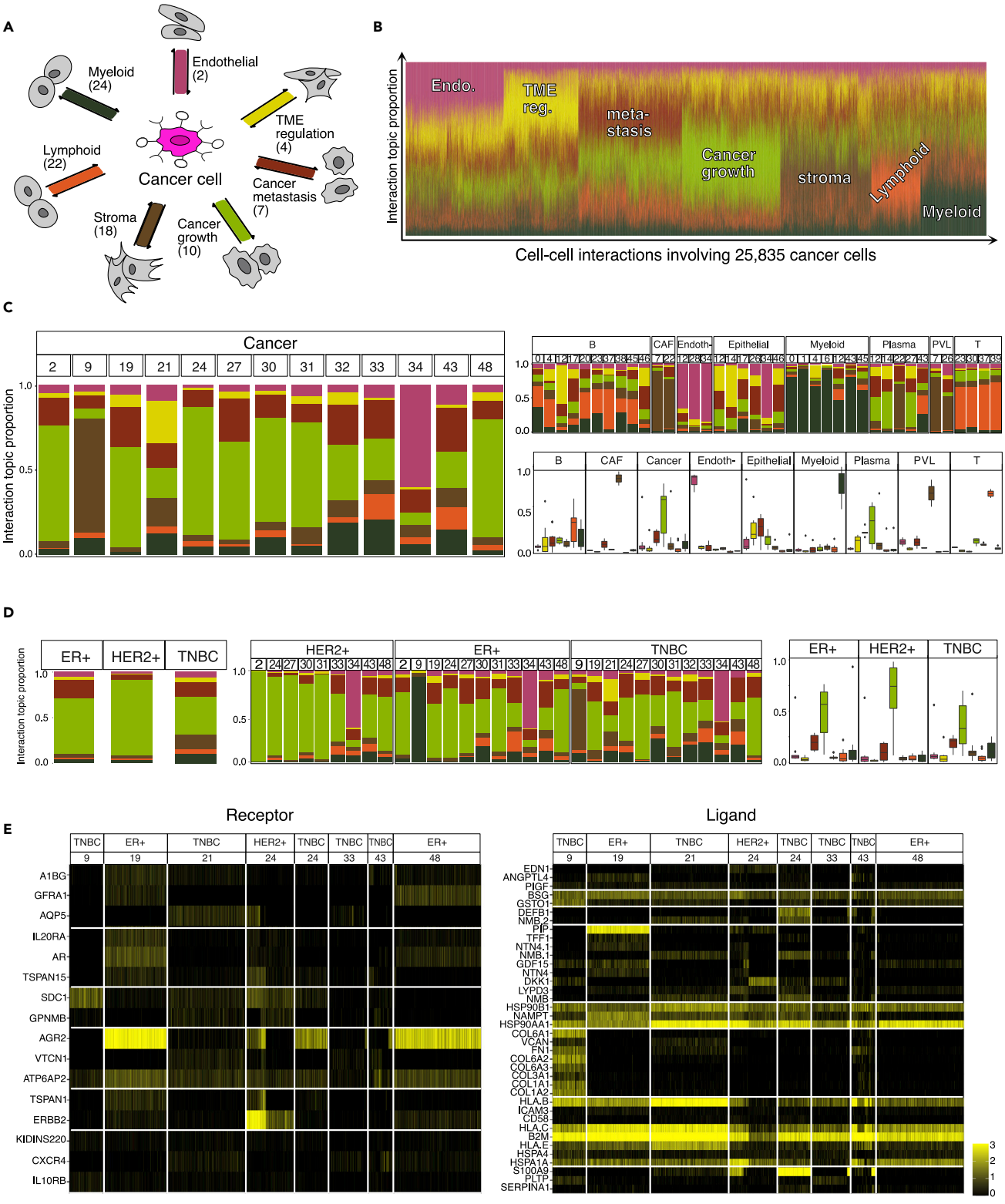


Figure 3. Heterogeneity of cancer cells revealed by the cell-topic specific interaction patterns

(A) Representation of interaction topics of cancer cells with surrounding cell types in the TME.

(B) Structure plot showing the variability of interaction topic proportion estimates among 25,835 cancer cells in the dataset. Proportion of seven interaction topics (y axis) and cancer cell pair (x axis) with one representative target cell among 159 source-target cell pairs are clustered using k-means(n = 7) clustering algorithm using interaction topic proportions.

Figure 3. *Continued*

(C) The proportion of interaction topics of cancer and non-malignant cells associated with cell topics. Boxplots show the distribution of interaction topic proportions for each interaction topic across all cell topics.

(D) The proportion of interaction topics of cancer subtypes associated with cell topics. Boxplots depict the estimated interaction topic proportions across all cell topics.

(E) Log normalized (total count to 10,000 reads per cell) expression of LR genes from all the cells associated with respective cancer subtype and cell topic. The genes (y axis) are top 25 LR genes from seven interaction topics with expression values >0.05.

signaling pathways *GPR37*, *CD151*, and *CD63* that are active in proliferation and migration, including epithelial-mesenchymal transition.[28,29] In the cancer growth topic (Figure 4A, green strips; Figure 4B, the second panel), the resulting gene network primarily consisted of known oncogenes, such as *PTPRF*, *FGFR1*, *ERBB2*, and *TNFRSF1A*[30] and developmental genes, such as *LAMP1*, *ITGB1*, *RPSA*, *CANX*, *ATP6AP2*, and *MCFD2*.[31] It is worth noting that the interaction networks adjacent to cancer cells generally include known cancer-related genes and other genes involved in cellular developmental process. Our analysis put them together in the same network modules, implicating a potential role of these interactions for cancer cells to hijack a normal cellular process. Similarly, immune-modulatory receptors primarily expressed in myeloid lineage cells (Figure 4A, dark green; Figure 4B, the third panel) *TREM2*, *CSF1R*, *CSF2R*, *LILR*, and *IL3R* and signaling pathways *LTBR* and *TYROBP* required for the activation of myeloid cells are associated with myeloid-associated interaction topic. This topic captures the interactions between cancer cells and myeloid cell progenitors such as tumor-associated macrophages (TAM) in the tumor microenvironment, suppressing T cells and facilitating tumor growth.[32]

In the lymphoid topic (Figure 4B, the fourth panel), T cell receptor (TCR) complex (*CD3D*, *CD3G*, *CD2*, and *CD247*) and TCR signaling pathway genes (*PTPRC*, *CD45*, and *CD53*) are highly expressed both in cancer and surrounding immune cells.[33] Other chemokine receptor genes, such as *CXCR3* and *CXCR4*, were also found highly co-activated in this interaction topic, corroborating the pivotal role of crosstalk between T cells and cancer cells in promoting cancer growth, immune evasion, and metastasis.[34] Of interest, other killer-cell lectin-like receptors, which also co-occurred in this module, namely *KLRC1*, *KLRD1*, and *KLRF1*, are known to restrict T-cell's antitumor immunity.[35] The stroma topic (Figure 4A, brown strips; Figure 4B, the first panel of the second row) represented gene networks that capture the interaction of cancer cells with surrounding cells that promote its vascularization. It consisted of *NOTCH3*, *AVPR1A*, *MYLK* and integrin-mediated *ITGA1*, *ITGA5*, *ITGA7*, and *ITGB1* signaling pathways that play vital roles in tumor cell adhesion and progression.[36] The genes *MCAM*, *ENPEP*, *EDNRA*, and *DCBLD2* that promote blood vessel formation and enhance tumorigenesis are enriched in this topic.[37] Similarly, genes enriched in the endothelial topic recapitulate interactions of cancer cells in developing tumor vascular networks, especially in conjunction with endothelial cells. Previously known that *PECAM1*, *CALCR*, *ADGRL4*, and *CD93* genes are predominantly expressed in endothelial cells and regulate angiogenesis in tumor cells.[38] In addition, TME-regulation associated primarily consisted of genes mixture of endothelial-associated and stroma-associated topics with enrichment of distinct genes known to control tumor growth and promote stemness of cancer cells in microenvironment such as *KCNN4*, *IL6ST*, and *CD1B*.[39] Furthermore, we observed a significant overlap between the gene-gene interactions derived from the interaction topic model and the gene network in the STRING database.[40] Here, 51%, 40%, 18%, 25%, 13%, 12%, and 13% of ligand-receptor pairs (Z score>4.0 and Pearson correlation coefficient >3.0) matched with gene pairs in the STRING database (combined score >0.3) from lymphoid, myeloid, stroma, endothelial, TME regulation, cancer metastasis, and cancer growth topics, respectively (Figure S3).

DISCUSSION

No single cell can exist alone in human tissues. We propose a novel machine learning framework that systematically dissects tens of millions of cell-cell pairs and uncovers common patterns of how cells talk to each other concerning cell surface ligand-receptor protein interactions. In particular, our analysis focused on finding commonly used communication channels in breast cancer progression and metastasis by reanalyzing state-of-the-art single-cell genomics datasets. Our approach, built on probabilistic topic modeling and variational autoencoder model, specifically demonstrated that a part of cancer heterogeneity could be understood in diverse and context-specific interaction partners of cancer cells. We found that many ligand-receptor interactions can occur in a subtype-specific manner, although cells are largely clustered as cancer cells in conventional single-cell analysis. Along the line, our results suggest that cell types and states are

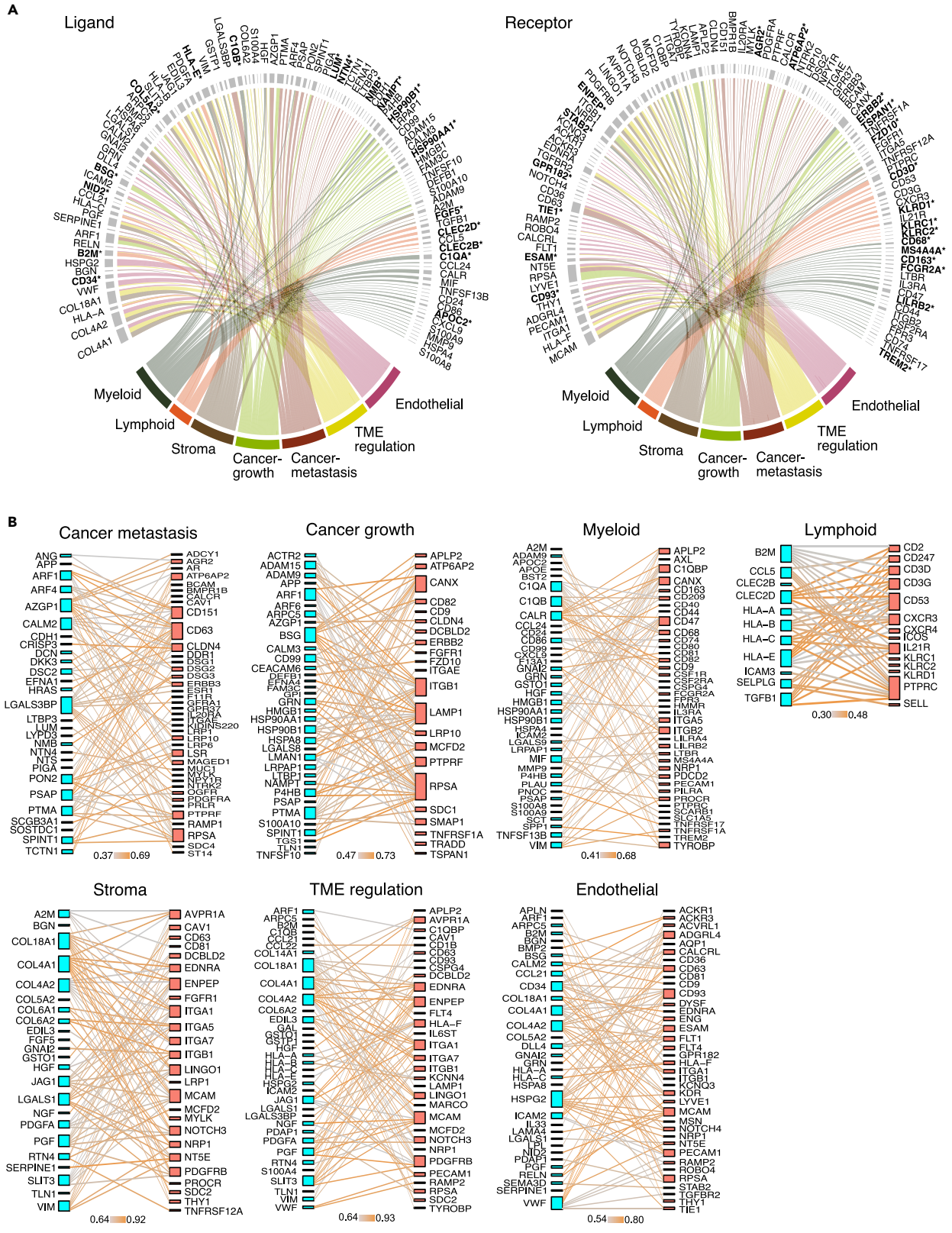


Figure 4. Gene network derived from the interaction topics

(A) Circular chord diagram of the interaction topic-gene network representing significantly (*Z* score >4.0) enriched LR gene loadings in each interaction topic. The edge between LR genes and interaction topics shows the occurrence of a gene in LR pairing and if a gene is unique to or common among different interaction topics. The genes with (*) symbol are among the top 10 LR genes based on loadings estimated by the interaction topic model.

(B) Ligand-receptor bipartite network for each interaction topics depicting significantly (*Z* score >4.0) enriched LR genes. The edge in the graphs represents the magnitude of Pearson correlation coefficients. The top 100 LR edges in each interaction topic with a correlation coefficient >3.0 are selected.

better understood, and the definitions of cell types can be refined while considering cell-cell communication patterns.

Our proposed approach generalizes existing bioinformatics methods and does not rely on prescribed cell-type annotations/clustering results, which may introduce unwanted biases in downstream analysis. Moreover, if cells within a cluster are not homogeneous as anticipated, a clustering-based cell-cell commutation method can easily result in confounded correlation statistics, clearly violating necessary assumptions, such as independent and identification distributed expression values. Here, we differently formulate cell-cell communications analysis from an edge's perspective, whereby a single edge (interaction) is a data point, and the feature vector can be engineered by exploiting the information of both endpoints of the edge (ligand and rector expression values). Such a novel formulation is better suited for the analysis of large-scale single-cell data and also easily extends to principled data integration strategies. For instance, if cell-cell interaction pairs were already constructed by spatial transcriptomics data, we can easily construct feature vectors by combining two gene expression vectors (one from the source and the other from target cells). For the multiomics data integration tasks, we can concatenate multiple data modalities to investigate the co-occurrence of multimodal expressions, such as DNA accessibility, histone modifications, and metabolomics.

Limitations of the study

We acknowledge that our SPRUCE approach relies on several specialized modeling assumptions. One of which is that we assume that known ligand-receptor protein-protein interaction networks serve as a superset/backbone of topic-specific interaction networks. Considering that most protein-protein interactions were experimentally discovered *in vitro* by error-prone high-throughput methods, there is room for improvements in terms of the precision and specificity of the interaction analysis method. Here, we only focused on immediate surface protein interactions. However, establishing causal effects on the downstream genes emanating from surface protein singling pathways can further enrich our understanding of disease etiology.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Material availability
  - Data and code availability
- METHOD DETAILS
  - Data preprocessing
  - Topic modelling
  - Notations
  - Cell-level probabilistic topic modelling
  - Bayesian autoencoder model for topic modelling
  - Variational inference algorithm
  - Celltype labelling by propagating within topic clusters
  - Cell-cell interaction topic modelling
  - Interaction topic model
  - Topic-specific gene co-expression network
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.106025.

ACKNOWLEDGMENTS

This work was supported by the BC Cancer Foundation and NSERC Discovery Grant (YPP). We acknowledge financial support from UBC Four Year Fellowship (SS). Much of the computation was supported by Cascadia Data Alliance Award (title: Eavesdropping Communications between Cancer and Immune Cells).

AUTHOR CONTRIBUTIONS

Conceptualization, Y.P.P.; Methodology, Y.P.P. and S.S.; Investigation, S.S.; Writing – Original Draft, S.S. and Y.P.P.; Writing – Review and Editing, S.S. and Y.P.P.; Funding Acquisition, Y.P.P.; Resources, Y.P.P; Supervision, Y.P.P.

DECLARATION OF INTERESTS

Nothing to declare.

INCLUSION AND DIVERSITY

We recognize the importance of inclusion and diversity in research. We uphold the value by encouraging the participation of traditionally underrepresented groups (both authors) while empowering long-standing academic journeys. We also strive to make projects transparent and intuitively understandable to avoid unnecessary disruptions that new researchers may encounter.

Received: September 26, 2022
Revised: November 24, 2022
Accepted: January 17, 2023
Published: January 23, 2023

REFERENCES

1. Teichmann, S., and Efremova, M. (2020). Method of the year 2019: single-cell multimodal omics. Nat. Methods *17*, 2020.

2. Tan, K., and Naylor, M.J. (2022). Tumour microenvironment-immune cell interactions influencing breast cancer heterogeneity and disease progression. Front. Oncol. *12*, 876451.

3. Almet, A.A., Cang, Z., Jin, S., and Nie, Q. (2021). The landscape of cell–cell communication through single-cell transcriptomics. Curr. Opin. Syst. Biol. *26*, 12–23.

4. Jin, S., Guerrero-Juarez, C.F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., et al. (2021). Inference and analysis of cell-cell communication using CellChat. Nat. Commun. *12*, 1088.

5. Efremova, M., Vento-Tormo, M., Teichmann, S.A., and Vento-Tormo, R. (2020). CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nat. Protoc. *15*, 1484–1506.

6. Zhang, Y., Liu, T., Hu, X., Wang, M., Wang, J., Zou, B., Tan, P., Cui, T., Dou, Y., Ning, L., et al. (2021). CellCall: integrating paired ligand–receptor and transcription factor activities for cell–cell communication. Nucleic Acids Res. *49*, 8520–8534.

7. Armingol, E., Baghdassarian, H.M., Martino, C., Perez-Lopez, A., Aamodt, C., Knight, R., et al. (2022). Context-aware deconvolution of cell–cell communication with tensor-cell2cell. Nat. Commun. *13*, 3665.

8. Tsuyuzaki, K., Ishii, M., and Nikaido, I. (2019). Uncovering hypergraphs of cell-cell interaction from single cell RNA-sequencing data. Preprint at bioRxiv. https://doi.org/10.1101/566182.

9. Wang, S., Karikomi, M., MacLean, A.L., and Nie, Q. (2019). Cell lineage and communication network inference via optimization for single-cell transcriptomics. Nucleic Acids Res. *47*, e66.

10. Dieng, A.B., Ruiz, F.J.R., and Blei, D.M. (2020). Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics *8*, 439–453.

11. Zhao, Y., Cai, H., Zhang, Z., Tang, J., and Li, Y. (2021). Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. Nat. Commun. *12*, 5261.

12. Wu, S.Z., Al-Eryani, G., Roden, D.L., Junankar, S., Harvey, K., Andersson, A., Thennavan, A., Wang, C., Torpy, J.R., Bartonicek, N., et al. (2021). A single-cell and spatially resolved atlas of human breast cancers. Nat. Genet. *53*, 1334–1347.

13. Bhat-Nakshatri, P., Gao, H., Sheng, L., McGuire, P.C., Xuei, X., Wan, J., Liu, Y., Althouse, S.K., Colter, A., Sandusky, G., et al. (2021). A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells. Cell Rep. Med. *2*, 100219.

14. Zheng, L., Qin, S., Si, W., Wang, A., Xing, B., Gao, R., Ren, X., Wang, L., Wu, X., Zhang, J., et al. (2021). Pan-cancer single-cell landscape of tumor-infiltrating t cells. Science *374*, abe6474.

15. Shao, X., Liao, J., Li, C., Lu, X., Cheng, J., and Fan, X. (2021). CellTalkDB: a manually curated database of ligand–receptor interactions in humans and mice. Briefings Bioinf. *22*, bbaa269.

16. Dufva, O., Pölönen, P., Brück, O., Keränen, M.A.I., Klievink, J., Mehtonen, J., Huuhtanen, J., Kumar, A., Malani, D., Siitonen, S., et al. (2020). Immunogenomic landscape of hematological malignancies. Cancer Cell *38*, 424–428.

17. Hussain, K., Cragg, M.S., and Beers, S.A. (2021). Remodeling the tumor myeloid landscape to enhance antitumor antibody immunotherapies. Cancers *13*, 4904.

18. Miller, M.L., Reznik, E., Gauthier, N.P., Aksoy, B.A., Korkut, A., Gao, J., Ciriello, G., Schultz, N., and Sander, C. (2015). Pan-cancer analysis

of mutation hotspots in protein domains. Cell Syst. *1*, 197–209.

19. Cekic, C., and Linden, J. (2014). Adenosine A2A receptors intrinsically regulate CD8+ t cells in the tumor MicroenvironmentAdenosine maintains CD8+ t cells in solid tumors. Cancer Res. *74*, 7239–7249.

20. Zhang, Z., Yu, Y., Zhang, P., Ma, G., Zhang, M., Liang, Y., et al. (2021). Identification of NTRK3 as a potential prognostic biomarker associated with tumor mutation burden and immune infiltration in bladder cancer. BMC Cancer *21*, 458.

21. Elshafae, S.M., Hassan, B.B., Supsavhad, W., Dirksen, W.P., Camiener, R.Y., Ding, H., Tweedle, M.F., and Rosol, T.J. (2016). Gastrin-releasing peptide receptor (GRPr) promotes EMT, growth, and invasion in canine prostate cancer. Prostate *76*, 796–809.

22. Padua, M.B., Bhat-Nakshatri, P., Anjanappa, M., Prasad, M.S., Hao, Y., Rao, X., Liu, S., Wan, J., Liu, Y., McElyea, K., et al. (2018). Dependence receptor UNC5A restricts luminal to basal breast cancer plasticity and metastasis. Breast Cancer Res. *20*, 1–18.

23. Primac, I., Maquoi, E., Blacher, S., Heljasvaara, R., Van Deun, J., Smeland, H.Y., Canale, A., Louis, T., Stuhr, L., Sounni, N.E., et al. (2019). Stromal integrin α11 regulates PDGFR βsignaling and promotes breast cancer progression. J. Clin. Invest. *129*, 4609–4628.

24. Bansal, R., Nakagawa, S., Yazdani, S., Van Baarlen, J., Venkatesh, A., Koh, A.P., Song, W.-M., Goossens, N., Watanabe, H., Beasley, M.B., et al. (2017). Integrin alpha 11 in the regulation of the myofibroblast phenotype: implications for fibrotic diseases. Exp. Mol. Med. *49*, e396.

25. Wu, X., Giobbie-Hurder, A., Liao, X., Connelly, C., Connolly, E.M., Li, J., Manos, M.P., Lawrence, D., McDermott, D., Severgnini, M., et al. (2017). Angiopoietin-2 as a biomarker and target for immune checkpoint therapy. Cancer Immunol. Res. *5*, 17–28.

26. Cvetković, D., Babwah, A.V., and Bhattacharya, M. (2013). Kisspeptin/KISS1R system in breast cancer. J. Cancer *4*, 653–661.

27. Horr, C., and Buechler, S.A. (2021). Breast cancer consensus subtypes: a system for subtyping breast cancer tumors based on gene expression. NPJ breast cancer *7*, 136.

28. Shang, X., Lin, X., Alvarez, E., Manorek, G., and Howell, S.B. (2012). Tight junction proteins claudin-3 and claudin-4 control tumor growth and metastases. Neoplasia *14*, 974–985.

29. Wang, J., Xu, M., Li, D.-D., Abudukelimu, W., and Zhou, X.-H. (2020). GPR37 promotes the malignancy of lung adenocarcinoma via TGF-β/smad pathway. Open Med. *16*, 024–032.

30. Butti, R., Das, S., Gunasekaran, V.P., Yadav, A.S., Kumar, D., and Kundu, G.C. (2018). Receptor tyrosine kinases (RTKs) in breast cancer: signaling, therapeutic implications and challenges. Mol. Cancer *17*, 1–18.

31. Going, C.C., Tailor, D., Kumar, V., Birk, A.M., Pandrala, M., Rice, M.A., Stoyanova, T., Malhotra, S., and Pitteri, S.J. (2018). Quantitative proteomic profiling reveals key pathways in the anticancer action of methoxychalcone derivatives in triple negative breast cancer. J. Proteome Res. *17*, 3574–3585.

32. Molgora, M., Esaulova, E., Vermi, W., Hou, J., Chen, Y., Luo, J., Brioschi, S., Bugatti, M., Omodei, A.S., Ricci, B., et al. (2020). TREM2 modulation remodels the tumor myeloid landscape enhancing anti-PD-1 immunotherapy. Cell *182*, 886–900.e17.

33. Shah, K., Al-Haidari, A., Sun, J., and Kazi, J.U. (2021). T cell receptor (TCR) signaling in health and disease. Signal Transduct. Target. Ther. *6*, 412–426.

34. Kuo, P.T., Zeng, Z., Salim, N., Mattarollo, S., Wells, J.W., and Leggatt, G.R. (2018). The role of CXCR3 and its chemokine ligands in skin disease and cancer. Front. Med. *5*, 271.

35. Hu, Z., Xu, X., and Wei, H. (2021). The adverse impact of tumor microenvironment on NK-cell. Front. Immunol. *12*, 633361.

36. Price, J.C., Azizi, E., Naiche, L.A., Parvani, J.G., Shukla, P., Kim, S., Slack-Davis, J.K., Pe'er, D., and Kitajewski, J.K. (2020). Notch3 signaling promotes tumor cell adhesion and progression in a murine epithelial ovarian cancer model. PLoS One *15*, e0233962.

37. Wragg, J.W., Finnity, J.P., Anderson, J.A., Ferguson, H.J.M., Porfiri, E., Bhatt, R.I., Murray, P.G., Heath, V.L., and Bicknell, R. (2016). MCAM and LAMA4 are highly enriched in tumor blood vessels of renal cell carcinoma and predict patient outcome. Cancer Res. *76*, 2314–2326.

38. Sheldon, H., Bridges, E., Silva, I., Masiero, M., Favara, D.M., Wang, D., Leek, R., Snell, C., Roxanis, I., Kreuzer, M., et al. (2021). ADGRL4/ELTD1 expression in breast cancer cells induces vascular normalization and immune SuppressionELTD1 is angiogenic and immunosuppressive in breast cancer. Mol. Cancer Res. *19*, 1957–1969.

39. Fan, J., Tian, R., Yang, X., Wang, H., Shi, Y., Fan, X., Zhang, J., Chen, Y., Zhang, K., Chen, Z., and Li, L. (2022). KCNN4 promotes the stemness potentials of liver cancer stem cells by enhancing glucose metabolism. Int. J. Mol. Sci. *23*, 6958.

40. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613.

41. Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163–172.

42. Amezquita, R.A., Lun, A.T.L., Becht, E., Carey, V.J., Carpp, L.N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., et al. (2020). Orchestrating single-cell analysis with bioconductor. Nat. Methods *17*, 137–145.

43. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (New York: Springer-Verlag).

44. Minoura, K., Abe, K., Nam, H., Nishikawa, H., and Shimamura, T. (2021). A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. Cell Rep. Methods *1*, 100071.

45. Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent dirichlet allocation. J. Mach. Learn. Res. *3*, 993–1022.

46. Griffiths, T.L., and Steyvers, M. (2004). Finding scientific topics. Proc. Natl. Acad. Sci. USA *101*, 5228–5235.

47. Kingma, D.P., and Welling, M. (2013). Auto-encoding variational bayes. Preprint at arXiv. https://doi.org/10.48550/arXiv.1312.6114.

48. Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning representations by back-propagating errors. Nature *323*, 533–536.

49. Kingma, D., and Ba, J. (2014). Adam: a method for stochastic optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6980.

50. Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163–172.

51. De Kanter, J.K., Lijnzaad, P., Candelli, T., Margaritis, T., and Holstege, F.C.P. (2019). CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. Nucleic Acids Res. *47*, e95.

52. ANNOY library. https://github.com/spotify/annoy.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Single-cell breast cancer Data #1	Wu et al. (2021)[12]	GSE176078
Single-cell breast cancer Data	Bhat-Nakshatri et al. (2021)[13]	GSE164898
Pan-cancer tumor-infiltrating T cells	Zheng et al. (2021)[14]	GSE156728
Software and algorithms		
Scanpy	Scanpy Development Team	https://scanpy.readthedocs.io/en/stable/
PyTorch	PyTorch Development Team	https://pytorch.org/
Numpy	Python package repository (PIP)	https://numpy.org/
Scipy	Python package repository (PIP)	https://scipy.org/
Pandas	Python package repository (PIP)	https://pandas.pydata.org/
Igraph	Python package repository (PIP)	https://igraph.org/
Seaborn	Python package repository (PIP)	https://seaborn.pydata.org/
ANNOY (Approximate Nearest Neighbors Oh Yeah)	Github repository (Spotify)	https://github.com/spotify/annoy
Celldex	Aran et al. (2019)[41]	http://bioconductor.org/packages/release/data/experiment/html/celldex.html
SingleR	Aran et al. (2019)[41]	https://bioconductor.org/packages/release/bioc/html/SingleR.html
SingleCellExperiment	Amezquita et al. (2020)[42]	https://bioconductor.org/packages/release/bioc/html/SingleCellExperiment.html
Ggplot2	Hadley Wickham[43]	https://ggplot2.tidyverse.org/
Pheatmap	Raivo Kolde	https://cran.r-project.org/web/packages/pheatmap/index.html
Circlize	Zuguang Gu	https://cran.r-project.org/web/packages/circlize/
bipartite	Carsten F. Dormann	https://cran.r-project.org/web/packages/bipartite/index.html
SPRUCE topic	This work	https://doi.org/10.5281/zenodo.7508044

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Yongjin Park (ypp@stat.ubc.ca or yongjin.park@ubc.ca).

Material availability

The study did not generate new unique reagents.

Data and code availability

We share the full working directory of model estimation and statistical analysis via the public repository: https://doi.org/10.5281/zenodo.7508044. We also made our source code and datasets available in the public repository: https://github.com/causalpathlab/spruceTopic.

METHOD DETAILS

Data preprocessing

We constructed a dataset to represent an immune-enriched breast cancer microenvironment by combining cancer cells with immune cells and healthy cells from three recent breast cancer-related studies. The breast

cancer dataset consists of 100k cells from a single-cell atlas of human breast cancers.[12] The source of the normal dataset consisting 48k cells is a single-cell atlas of the healthy breast tissues.[13] The third dataset composed of immune cells is 6k subset of breast cancer CD4 and CD8 T-cells from a pan-cancer atlas of tumour-infiltrating T cells profiled across 21 cancer types and 316 donors.[14] The total number of cells in the combined dataset is 155,913. We filtered out genes detected in less than three cells along with mitochondrial and spike genes, leading to 20,265 genes in the final dataset.

Topic modelling

SPRUCE takes a topic modelling approach to identify cell subtypes/states and define their interaction pattern based on cell-cell communication in the tumour microenvironment (TME). The model consists of two types of autoencoder-based topic models, each with a pair of encoder-decoder networks. The first model takes gene expression single-cell data as input and models cell topic and topic-specific gene loadings. Next, we assigned a cell topic to each cell based on the highest topic proportion from the cell topic model. The topic assignment was used to construct a set of target cells for each cell such that one cell is paired with the five nearest target cells from each topic. The LR gene expression data from each cell pair were transformed into each other's space by a binary cell-interaction database–CellTalkDB. The transformed ligand and receptor data were treated as two independent modules with separate encoder and decoder modules in the model. The latent variables with encoded information from two modules were combined by taking their average to obtain a final interaction topic variables as a mixture of experts[44] from the ligand and receptor latent space. The second ETM, the interaction topic model, uses transformed LR data from source-target cell pairs as input and models interaction topic for each cell pair with topic-specific LR gene loadings.

Notations

The following notation will be used to describe the data and model. Notations for gene expression and interaction data:

- $i \in [N]$: an integer index for a cell i of total N cells
- $g \in [G]$: an integer index for a gene g of total G genes
- X_{ig}: gene expression (non-negative) count data measured on a gene g in a cell i

Notations for the cell topic model:

- $k, t \in [K]$: an index for a topic t of total K topics
- θ_{it}: cell topic proportion of i th cell for t th topic ($\theta_{it} > 0$ and $\sum_{t=1}^{K} \theta_{it} = 1$ for all i)
- β_{tg}(cell topic model): gene proportion of t th cell topic for g th gene

Notations for the interaction topic model:

- e: an index for a cell pair, e.g., $e = (i, j)$ for a cell i and j.
- X_{li}: expression count of a ligand protein l in a celli
- X_{ri}: expression count of a receptor protein r in a celli
- Y_{el}: transformed and aggregated count data of a ligand protein l in a cell paire
- Y_{er}: transformed and aggregated count data of a receptor protein r in a cell paire
- θ_{et}: interaction topic proportion of e th cell pair for t th topic
- β_{tg} (interaction topic model): gene proportion of t th interaction topic for a gene g, which can be either a ligand protein/gene l or a receptor protein/gene r.

Cell-level probabilistic topic modelling

Firstly, we designed a topic model for cell type annotations, treating cells as documents and genes as vocabulary, built on the Embedded Topic Model framework.[10] ETM generally outperforms traditional topic

modeling approaches, such as Latent Dirichlet Allocation,[45] relying on tailored variational inference[45] and collapsed Gibbs sampling inference,[46] and fits naturally in a variation autoencoder (VAE) framework[47] while providing a scalable GPU-based inference algorithm.

Letting $\mathbf{x}_i = (X_{i1}, \ldots, X_{iG})$ be a vector of gene expression counts on G genes for each cell i, a topic modeling assumes that $\mathbf{x}_i$ were generated by multinomial distribution parameterized by a normalized gene expression frequency vector, namely $\rho_i = (\rho_{i1}, \ldots, \rho_{iG})$, achieving a scale-invariant property across different cells, batches, and datasets:

$$p(\mathbf{x}_i|\boldsymbol{\rho}_i) \propto \prod_g \rho_{ig}^{X_{ig}}$$

In the original ETM formulation,[10] ρ is directly modelled by transforming each cell's topic proportion θ_{it} in a topic space to a gene space as a linear combination of topic-specific probabilities, $\rho_{ig} = \sum_t \theta_{it}\beta_{tg}$, where β_{tg} captures a topic t specific frequency of a gene g. The latent cell topic proportion θ_{it} is drawn from Logistic Normal distribution with an auxiliary Gaussian vector $\mathbf{z}_i = (Z_{i1}, \ldots, Z_{iK})$:

Assuming $\mathbf{z}_i \sim \mathcal{N}(0, I)$ *a priori*, the encoder network will first generate

$$\mathbf{z}_i \leftarrow \boldsymbol{\mu}_i(\mathbf{x}_i) + \sqrt{\boldsymbol{\nu}_i(\mathbf{x}_i)} \circ \boldsymbol{\varepsilon}$$

where the mean μ and variance ν functions were modelled by deep neural networks, taking expression data$\mathbf{x}_i$, and the stochastic vector were simply generated by $\mathcal{N}(0, 1)$ independently. We can then project the Gaussian latent states into the desired topic space:

$$\theta_{it} = \exp(Z_{it}) \Big/ \sum_{k=1}^{K} \exp(Z_{ik}).$$

Since a Gaussian random variable can be generated by a reparameterization trick, which then separates model parameters from stochastic variables, the latent variables θ and model parameters β are seamlessly integrated in a neural network model; they can be optimized by back-propagation algorithm[48] implemented in PyTorch library (https://pytorch.org/).

Bayesian autoencoder model for topic modelling

Instead of directly modelling ρ, we introduce the Dirichlet prior on the gene frequency and parameterize the Dirichlet as a generalized linear model(GLM) with linear combinations of topic-specific probabilities:

$$p(\rho_i|\lambda_i) = \frac{\Gamma\left(\sum_g \lambda_{ig}\right)}{\prod_g \Gamma(\lambda_{ig})} \prod_{g \in [G]} \rho_{ig}^{\lambda_{ig}-1}$$

where $\Gamma(\cdot)$ is the Euler's gamma function,

$$\lambda_{ig} = \exp\left(\sum_{t=1}^{K} \theta_{it}(\beta_{tg} + b_g)\right),$$

and

$$\beta_{tg} \sim \mathcal{N}(0, I).$$

Exploiting the conjugacy between the multinomial and Dirichlet distributions, we can integrate out the unknown parameters ρ. Then, the marginal likelihood of a single cell data$\mathbf{x}_i$ then becomes:

$$p(\mathbf{x}_i|\cdot) = \frac{\Gamma\left(\sum_g \lambda_{ig}\right)}{\sum_g \Gamma(\lambda_{ig})} \frac{\Gamma\left(\sum_g \lambda_{ig} + X_{ig}\right)}{\sum_g \Gamma(\lambda_{ig} + X_{ig})},$$

where $\Gamma(\cdot)$ is the Euler's gamma function.

Variational inference algorithm

We resolved the posterior distribution of latent variables and model parameters, $p(\{\boldsymbol{\theta}_i\},\{\beta_{gt}\}\big|\{\mathbf{x}_i\})$, by finding variational/approximating distributions $q(\boldsymbol{\theta}_i|\mu(\mathbf{x}_i),\nu(\mathbf{x}_i))$ and $q(\beta_{tg})$. We defined $q(\boldsymbol{\theta}|\bullet)$ as before using deep neural networks for Logistic Normal distributions. For the topic-specific gene matrix, β_{tg}, we used mean-field approximation: $q(\beta_{tg}) \sim \mathcal{N}(\mu^{\beta}_{tg},\nu^{\beta}_{tg})$. To minimize the Kullback-Leibler (KL) divergence between the true posterior and the approximate posterior, we maximize the evidence lower bound (ELBO) of the log-likelihood $\mathcal{L}$:

$$\mathcal{L} \stackrel{\Delta}{=} \sum_{i=1}^{N} \mathbb{E}\left[\log p(\mathbf{x}_i|\boldsymbol{\theta}_i,\beta)\right] + \sum_{i=1}^{N} \mathbb{E}\left[\log \frac{p(\mathbf{z}_i)}{q(\mathbf{z}_i|\bullet)}\right] + \sum_{tg} \mathbb{E}\left[\log \frac{p(\beta_{tg})}{q(\beta_{tg}|\bullet)}\right],$$

where the expectations were taken with respect to the variational distribution. The expectation operators can be well-approximated by summing over different $\mathbf{z}_i$ and β values sampled by the reparameterization tricks.[47]

The encoder for the cell topic model consisted of a 4-layered neural network with two hidden layers of size 200 and two with sizes 100 and 50. We used an Adam optimizer[49] with a learning rate 0.01 and optimized the model for convergence for 1000 epochs with a minibatch size of 128 (Figure S1D). The cell topics are almost ubiquitously present across cells from different datasets (Figures S1C and S1E).

Celltype labelling by propagating within topic clusters

We generated a reference cell-type label for each cell in the dataset by combining the annotations from previous studies and cell type predictions from single-cell annotation tools. The cell annotations for breast cancer cells and immune cells from pan-cancer dataset were obtained from the previous studies.[12,14] For the annotation of normal breast cells, we used the reference-based cell type identification method SingleR with a tumour microenvironment reference dataset from CHETAH.[50,51] Next, we applied our proposed cell-topic model to an integrated dataset consisting of 155,913 cells and 20,265 genes. After unsupervised training of the model, all cells were mapped in the reduced latent cell topic space. We performed clustering on the reduced latent dimension using *k*-means algorithm (with *k* matched to the number of topics in cell topic model). Then each cluster was mapped to cell type using the majority rule on the reference cell type labels of cells assigned to the respective cluster, followed by relabeling of cells if needed for the downstream analyses.

Cell-cell interaction topic modelling

Construction of feature vectors for cell-cell interaction analysis

We assigned a cell topic to each cell in the dataset based on the highest proportion value from a vector of topic proportions inferred by the trained model i.e. topic assignment t_i for cell i is $\text{argmax}(\theta_{i1},\theta_{i2},\ldots,\theta_{iK})$ where θ_{it} is cell topic proportion of i th cell for t th topic. Next, a set of target cells was constructed for each cell using the topic assignment from the cell topic model. For each cell, the five closest target cells from each topic were calculated using python package[52] with angular distance on the cell topic space. An annoy model was created for each topic with a total cell count greater than 100 (32 out of 50 cell topics). We generated 159 target cell pairs for each cell - 32 topics and 5 target cells from each topic, excluding a self-target cell. Finally, a cell pair data matrix was constructed with 155,913 source cells times 159 target cells, consisting of 24,790,167 unique cell pairs.

The raw LR gene expression data from each cell pair is transformed into each other's space using a binary ligand-receptor interaction matrix generated from a publicly available database–CellTalkDB.[15] Here, let A be a l x r binary matrix with its entries as $A_{lr} = 1$ if and only if a ligand l binds with a receptor protein r in the cell interaction database; otherwise, $A_{lr} = 0$. For each cell pair $e \equiv (i,j)$, we aggregated expression counts for ligand and receptor proteins included in the A matrix by the reciprocal cell-cell interactions between i and j. For the activity for a receptor $r \in [R]$, we combined expression values emanating from relevant ligand proteins/genes:

$$Y^{(R)}_{er} = \sum_{l \in \text{ ligands}} \left[X_{il}A_{lr}X_{jr} + X_{jl}A_{lr}X_{ir}\right].$$

Similarly, for a ligand $l \in L$, we aggregated the values on the receptor side:

$$Y_{el}^{(L)} = \sum_{r \in \text{receptors}} \left[X_{il} A_{lr} X_{jr} + X_{jl} A_{lr} X_{ir} \right].$$

Interaction topic model

The interaction topic model uses transformed ligand and receptor (LR) gene expression data from source-target cell pairs to model the interaction topic for each cell pair and identify LR genes enriched in each interaction topic. The model follows the same architecture of the cell topic model, where each cell pair is considered a document and LR genes in the cell pair as words. We define the joint likelihood of the interaction topic model as product of two likelihood functions, each derived from the ligand and receptor topic models:

$$p\left(\mathbf{y}_e^{(L)} \middle| \boldsymbol{\rho}_e^{(L)}\right) p\left(\mathbf{y}_e^{(R)} \middle| \boldsymbol{\rho}_e^{(R)}\right) = \prod_{l \in [L]} \rho_{el}^{Y_{el}} \prod_{r \in [R]} \rho_{er}^{Y_{er}}.$$

The decoder (data-generating) model for each side was defined as the same Multinomial-Dirichlet hierarchical model as the cell topic model. In order to map the ligand and receptor activities to a shared topic space, we formed a mixture of experts[44] by equally mixing the outputs of two encoder networks, henceforth generating two sides of data with the same topic proportions. For each cell pair $e \in [m]$, we generate the ligand and receptor activities and optimize the model parameters by the back-propagation in the following steps:

- Combine $\mathbf{y}_e^{(L)}$ using $\mathbf{x}_i$ and $\mathbf{x}_j$ for ligand genes and $\mathbf{y}_e^{(R)}$ for receptor genes
- Generate latent state parameters, the mean $\mu(\mathbf{y}_e^{(L)})$ and variance $\nu(\mathbf{y}_e^{(L)})$ based on the ligand feature activities and the mean $\mu(\mathbf{y}_e^{(R)})$ and variance $\nu(\mathbf{y}_e^{(R)})$ based on the receptor features
- Sample $\mathbf{z}_e^{(L)} \leftarrow \mu^{(L)} + \sqrt{\nu}^{(L)} \circ \boldsymbol{\varepsilon}$ for the ligand and $\mathbf{z}_e^{(R)} \leftarrow \mu^{(R)} + \sqrt{\nu}^{(R)} \circ \boldsymbol{\varepsilon}$ receptor networks
- Transform them into a common topic proportion vector $/2\boldsymbol{\theta}_e \leftarrow \text{softmax}(\mathbf{z}_e^{(L)}/2 + \mathbf{z}_e^{(R)})$
- Sample $\beta_{tl}^{(L)} \sim \mathcal{N}(\mu_{tl}^{\beta(L)}, \nu_{tl}^{\beta(L)})$ for ligand topics and $\beta_{tr}^{(R)} \sim \mathcal{N}(\mu_{tr}^{\beta(R)}, \nu_{tr}^{\beta(R)})$ for receptor topics
- Compute the following composite ELBO objective $\mathcal{L}^{LR}$, take stochastic gradients, and optimize by the Adam optimizer.

$$\begin{aligned} \mathcal{L}^{LR} \triangleq\ & \mathbb{E}\left[\log p\left(\mathbf{y}_e^{(L)} \middle| \boldsymbol{\theta}_e, \boldsymbol{\beta}^{(L)}\right)\right] + \mathbb{E}\left[\log p\left(\mathbf{y}_e^{(R)} \middle| \boldsymbol{\theta}_e, \boldsymbol{\beta}^{(R)}\right)\right] \\ & - \mathrm{D}_{\mathrm{KL}}\left(q\left(\beta^{(L)}\right) q\left(\mathbf{z}^{(L)}\right) \parallel p\left(\beta^{(L)}\right) p\left(\mathbf{z}^{(L)}\right)\right) \\ & - \mathrm{D}_{\mathrm{KL}}\left(q\left(\beta^{(R)}\right) q\left(\mathbf{z}^{(R)}\right) \parallel p\left(\beta^{(R)}\right) p\left(\mathbf{z}^{(R)}\right)\right) \end{aligned},$$

where we denote Kullback-Leibler divergence between q and p, i.e., $\mathbb{E}_q[\log q/p]$, by $\mathrm{D}_{\mathrm{KL}}(q \parallel p)$.

The encoder for the cell topic model consisted of a 3-layered neural network with hidden layers of sizes 200, 100, and 25. We used an Adam optimizer with a learning rate of 0.01 and optimized the model for convergence for 500 epochs with a minibatch size of 5088 cell pairs (32 individual cells in a batch) (Figure S2A). The interaction topics represent cell pair interactions across different data sets (Figure S2B).

Topic-specific gene co-expression network

For each interaction topic t and LR gene g, we calculated a Z score s_{tg} as $\mathbb{E}[\beta_{tg}]/\sqrt{\mathbb{V}[\beta_{tg}]}$, where β_{tg} is topic-specific gene frequency estimated by the interaction topic model. The highest proportion estimate was used to assign an interaction topic to all cancer cell pairs. A gene co-expression network for each interaction topic was constructed using significantly enriched LR genes (Z score > 4). The edge in the network represents the magnitude of Pearson correlation coefficient between LR gene pairs based on expression data in all the cancer cell pairs assigned to respective interaction topic.

QUANTIFICATION AND STATISTICAL ANALYSIS

We performed deep learning using PyTorch library (https://pytorch.org/) as specified in the above section. We assumed the data were generated from multinomial distributions, of which uncertainty information was automatically incorporated into our deep learning models. Standard deviations of the model parameters are available for each topic-specific gene variable. We implemented custom-built visualization scripts in R language.

ADDITIONAL RESOURCES

We shared the full working directory of model estimation and statistical analysis via the public repository: https://doi.org/10.5281/zenodo.7508044.