STACCato: Supervised Tensor Analysis tool for studying Cell-cell Communication using scRNA-seq data across multiple samples and conditions

Qile Dai[1,2], Michael P. Epstein[2*], Jingjing Yang[2*]

1. Department of Biostatistics and Bioinformatics, Emory University School of Public Health, Atlanta, Georgia 30322, United States of America.
2. Center for Computational and Quantitative Genetics, Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia 30322, United States of America.

*Correspondence Authors: M.P.E. (mpepste@emory.edu) and J.Y. (jingjing.yang@emory.edu)

Abstract

Research on cell-cell communication (CCC) is crucial for understanding biology and diseases. Many existing CCC inference tools neglect potential confounders, such as batch and demographic variables, when analyzing multi-sample, multi-condition scRNA-seq datasets. To address this significant gap, we introduce STACCato, a **S**upervised **T**ensor **A**nalysis tool for studying **C**ell-cell **C**ommunication, that identifies CCC events and estimates the effects of biological conditions (e.g., disease status, tissue types) on such events, while adjusting for potential confounders. Application of STACCato to both simulated data and real scRNA-seq data of lupus and autism studies demonstrate that incorporating sample-level variables into CCC inference consistently provides more accurate estimations of disease effects and cell type activity patterns than existing methods that ignore sample-level variables. A computational tool implementing the STACCato framework is available on GitHub.

Introduction

Cell-cell communication (CCC) involves cells exchanging signals to coordinate physiological and developmental functions in multicellular organisms. The study of CCC events, which involves interactions between one ligand-receptor pair from one sender cell type to one receiver cell type, is important for elucidating biological processes, exploring disease mechanisms, and inspiring advancements in drug discovery. Using gene expression data produced by single-cell RNA sequencing (scRNA-seq) technology, multiple computational tools are now available to infer CCC events[1–9].

Recently, high-throughput sequencing technology advancements have significantly reduced the cost of scRNA-seq, allowing researchers to gather scRNA-seq data from multiple biological samples under multiple biological conditions[10–13], such as disease versus healthy control samples or samples from multiple tissue types. Most existing computational tools developed for CCC inference were originally designed for analyzing single-sample scRNA-seq data[1–7]. When attempting to apply these tools to multi-sample multi-condition scRNA-seq datasets, a three-step procedure is typically necessary. First, data from all samples within the same condition are combined to create an aggregated "sample" per condition. Second, communication scores are calculated for CCC events using the aggregated "samples", one per condition. Last, CCC events with significantly different communication scores across conditions are identified as condition-related CCC events. Another proposed strategy to handle such multi-sample multi-condition single-cell data is to use the tensor decomposition technique, which has been used to extract underlying lower-dimensional patterns from high-dimensional genomic data[8,9,14,15]. For example, the recently developed tool Tensor-cell2cell [11] constructs a 4-dimensional communication score tensor, with 4 dimensions corresponding to samples, ligand-receptor pairs sender cell types, and receiver cell types. Tensor-cell2cell applies unsupervised tensor decomposition to identify underlying communication patterns, and then tests if the communication patterns are significantly different across conditions.

An important drawback of both the three-step procedure and the Tensor-cell2cell tool for analyzing multi-sample and multi-condition scRNAseq data is that they ignore important sample-level variables (such as processing batch, age, gender, and ancestry) that are typically collected in such studies. These variables can have substantial impacts on both biological conditions and CCC, likely confounding the identification of condition-related CCC events. Neglecting these confounding variables may mask true biological associations between CCC events and conditions, or, even more concerning, lead to false positive associations that could result in misguided interpretations of CCC events. Therefore, the development of a CCC inference tool to effectively incorporate sample-level variables and adjust for potential confounding variables in multi-sample multi-condition scRNA-seq data becomes increasingly important.

To bridge this gap, we introduce the **S**upervised **T**ensor **A**nalysis tool for studying **C**ell-cell **C**ommunication (STACCato), that uses multi-sample multi-condition scRNA-seq dataset to identify CCC events significantly associated with conditions while adjusting for potential sample-level confounders. STACCato considers the same 4-dimentional communication score tensor as the Tensor-cell2cell tool, with 4 dimensions corresponding to samples, ligand-receptor pairs, sender cell types, and receiver cell types. Different from the Tensor-cell2cell tool, STACCato employs supervised tensor decomposition[16] to fit a regression model that considers the 4-dimensional communication score tensor as the outcome variable while treating the biological conditions (e.g., disease status, time points, tissue types) and other sample-level covariates (e.g., batch and demographic variables) as independent variables. Through this supervised tensor-based regression model, STACCato can identify CCC events and estimate the impact of conditions on CCC events, while effectively controlling for potential confounding variables.

In subsequent sections, we first introduce the analytical framework of STACCato. We then apply STACCato to two real datasets: the Systemic Lupus Erythematosus (SLE) dataset[10,11] consisting of scRNA-seq data of peripheral blood mononuclear cells (PBMC) samples from 154 SLE patients and 97 healthy controls, and the Autism Spectrum Disorder (ASD) dataset[12] consisting of snRNA-seq data of prefrontal cortex (PFC) samples from 13 ASD patients and 10 controls. Notably, the SLE dataset exhibits an unbalanced study design, resulting in batch effects being highly confounded with the disease effect. We observed dramatic changes in estimated disease effects for CCC events before and after adjusting for batch effects, leading to contrasting conclusions regarding the associations between these CCC events and SLE. These findings underscore the substantial impact of confounding variables on CCC inference, emphasizing the necessity of accounting for confounding variables in CCC studies. We further validate these observations through a simulation study considering various study designs. Finally, we conclude with a discussion.

Results

STACCato framework

We propose STACCato, a powerful tool that utilizes multi-sample multi-condition scRNA-seq data to identify condition-related CCC events while accounting for potential confounding variables. Briefly, STACCato first generates a 4D communication score tensor with four dimensions representing samples, ligand-receptor pairs, sender cell types, and receiver cell types (Figure 1A-1C). Next, STACCato employs a supervised tensor decomposition method that incorporates sample-level information (such as biological conditions or batches) to estimate a coefficient tensor, representing the effects of sample-level variables on CCC events (Figure 1C). Finally, we conduct parametric bootstrapping to assess the significance of the estimated coefficients. We describe the general supervised tensor decomposition framework below and relegate the technical details to the Methods section.

Supervised tensor decomposition of communication score tensor

With respect to an CCC event involving the interaction of ligand-receptor pair j from sender cell type k to receiver cell type l, we consider the following regression model to assess the association between the CCC event and the condition adjusting for other covariates,

$$y_{ijkl} = \beta_{1jkl}x_{i1} + \cdots + \beta_{qjkl}x_{iq} + \epsilon_{ijkl};$$

$$i = 1, \cdots I;\ \ j = 1, \cdots J;\ \ k = 1, \cdots K;\ \ l = 1, \cdots L;\ \ q = 1, \cdots Q. \quad \text{(Equation 1)}$$

Here, I, J, K, L, and Q are the total number of samples, ligand-receptor pairs, sender cell types, receiver cell types, and sample-level variables, respectively. In Equation 1, y_{ijkl} denotes the communication score representing the communication level of the CCC event involving the interaction of ligand-receptor pair j from sender cell type k to receiver cell type l in sample i (see Methods for details about communication score calculation); x_{iq} denotes the sample-level variable q, such as biological condition or batch, for sample i; β_{qjkl} denotes the effect of variable q on the communication score of the CCC event involving the interaction of ligand-receptor pair j from sender cell type k to receiver cell type l; and $\epsilon_{ijkl} \sim N(0, \sigma^2)$ denotes the random error that follows a Gaussian distribution with mean 0 and standard deviation σ.

A straightforward way to estimate β_{qjkl} is to fit a regression model with $\boldsymbol{y}_{jkl} = \left[y_{1jkl}, \cdots, y_{Ijkl}\right]^T$ as the values of the dependent variable and sample-level information matrix $\boldsymbol{X} \in \mathbb{R}^{I \times Q}$ as the design matrix for independent variables. The major limitation of this strategy is that it estimates $\boldsymbol{\beta}_{jkl} = \left[\beta_{1jkl}, \cdots, \beta_{Qjkl}\right]^T, j =$

$1, \cdots J$, $k = 1, \cdots K$, $l = 1, \cdots L$ separately for each CCC event and ignores the correlations among CCC events. For example, the interactions of the same ligand-receptor pair j across different sender and receiver cell types are dependent, and thus β_{qjkl} is dependent of $\beta_{qjk'l'}$ with $k \neq k'$ and $l \neq l'$. To consider such correlations among CCC events, we employ a supervised tensor technique to jointly estimate $\boldsymbol{\beta}_{jkl}$ for all $j = 1, \cdots J$, $k = 1, \cdots K$, $l = 1, \cdots L$. To do so, we note that Equation 1 is equivalent to the tensor model,

$$\mathcal{Y} = \mathcal{B} \times_1 \boldsymbol{X} + \mathcal{E} \qquad \text{(Equation 2)}$$

where $\mathcal{Y} \in \mathbb{R}^{I \times J \times K \times L}$ denotes the 4-dimensional communication score tensor with dimensions of I samples, J ligand-receptor pairs, K sender cell types, and L receiver cell types, with the (i, j, k, l) entry corresponding to y_{ijkl} in Equation 1 (see Figure 1A – 1C for an example communication score tensor; see Methods for details about constructing communication score tensor); $\mathcal{B} \in \mathbb{R}^{Q \times J \times K \times L}$ denotes a 4-dimensional coefficient tensor with dimensions of Q sample-level variables, J ligand-receptor pairs, K sender cell types, and L receiver cell types, with the (q, j, k, l) entry corresponding to β_{qjkl} in Equation 1; $\boldsymbol{X} \in \mathbb{R}^{I \times Q}$ in Equation 2 denotes sample-level design matrix for Q variables of I samples, with the (i, q) entry corresponding to x_{iq} in Equation 1; $\times_1$ denotes multiplying a tensor by a matrix in the tensor's first dimension; and $\mathcal{E} \in \mathbb{R}^{I \times J \times K \times L}$ denotes a 4-dimensional tensor with the (i, j, k, l) entry corresponding to ϵ_{ijkl} in Equation 1. The graphic representation of an example tensor model as in Equation 2 is shown in Figure 1C, with disease, age, and batch as example sample-level variables. The detailed illustration of how this supervised tensor technique can incorporate correlations among CCC events is described in the Methods section.

To estimate $\mathcal{B}$ in Equation 2, we employ the supervised tensor decomposition technique[16] that considers $\mathcal{B}$ in Equation 2 as a core tensor $\mathcal{G}$ multiplied by 4 factor matrices $\boldsymbol{M}_Q, \boldsymbol{M}_J, \boldsymbol{M}_K, \boldsymbol{M}_L$,

$$\mathcal{B} = \mathcal{G} \times_1 \boldsymbol{M}_Q \times_2 \boldsymbol{M}_J \times_3 \boldsymbol{M}_K \times_4 \boldsymbol{M}_L.$$

where $\times_d, d = 1,2,3,4$ denotes multiplying a tensor by a matrix in the tensor's dth dimension. For the convenience of presentation, we use $\mathcal{G} \times \{\boldsymbol{M}_Q, \boldsymbol{M}_J, \boldsymbol{M}_K, \boldsymbol{M}_L\}$ to denote the above tensor-by-matrix product. Then the full supervised tensor decomposition model is given by:

$$\mathcal{Y} = \mathcal{B} \times_1 \boldsymbol{X} = \mathcal{G} \times \{\boldsymbol{M}_Q, \boldsymbol{M}_J, \boldsymbol{M}_K, \boldsymbol{M}_L\} \times_1 \boldsymbol{X} + \mathcal{E}, \quad \text{(Equation 3)}$$

where $\boldsymbol{M}_Q \in \mathbb{R}^{Q \times r_Q}$, $\boldsymbol{M}_J \in \mathbb{R}^{J \times r_J}$, $\boldsymbol{M}_K \in \mathbb{R}^{K \times r_K}$, $\boldsymbol{M}_L \in \mathbb{R}^{L \times r_L}$ are factor matrices. These factor matrices have orthonormal columns (i.e., factors), which can be thought of as the principal components for each dimension. Under the context of cell-cell communication, $\boldsymbol{M}_Q \in \mathbb{R}^{Q \times r_Q}$ contains r_Q factors, representing r_Q effect patterns of Q covariates; $\boldsymbol{M}_J \in \mathbb{R}^{J \times r_J}$ contains r_J factors, representing r_J activity patterns of J ligand-receptor pairs; $\boldsymbol{M}_K \in \mathbb{R}^{K \times r_K}$ contains r_K factors, representing r_k activity patterns of K sender cell type; $\boldsymbol{M}_L$ contains r_L factors, represents r_L activity patterns of L receiver cell type; $\mathcal{G} \in \mathbb{R}^{r_Q \times r_J \times r_K \times r_L}$ in Equation 3 denotes the core tensor whose entries show the level of interaction among the factors from different dimensions. We define the decomposition rank $\boldsymbol{r} = (r_Q, r_J, r_K, r_L)$. Details regarding the determination of $\boldsymbol{r}$ are described in the Methods section.

We use the QR-adjusted optimization algorithm proposed by Hu et al.[16] to estimate $\mathcal{B}$, $\mathcal{G}$, $\boldsymbol{M}_Q$, $\boldsymbol{M}_J$, $\boldsymbol{M}_K$ $\boldsymbol{M}_L$. The significance level of estimated coefficients in $\mathcal{B}$ are assessed using parametric bootstrap[17]. The details about the optimization algorithm and bootstrap procedure are described in Methods.

Applying STACCato to identify CCC events associated with SLE

We applied STACCato to a scRNA-seq dataset of PBMC samples from 154 SLE subjects and 97 healthy controls[10,11] to identify CCC events associated with SLE while adjusting for age, gender, self-reported ancestry, and processing batch (see Methods for details). The constructed 4-dimensional communication score tensor is a $251 \times 55 \times 9 \times 9$ tensor containing the communication scores of CCC events for 251 samples across 55 ligand-receptor pairs, 9 sender cell types, and 9 receiver cell types. The 9 cell types are B cells, natural killer cells (NK), proliferating T and NK cells (Prolif), CD4$^+$ T cells, CD8$^+$ T cells, CD14$^+$ classical monocytes (cM), CD16$^+$ nonclassical monocytes (ncM), conventional dendritic cells (cDC), and plasmacytoid dendritic cells (pDC). We used the decomposition rank $\boldsymbol{r} = (r_Q = 8, r_J = 7, r_K = 4, r_L = 4)$. We used 4,999 iterations of bootstrapping resampling to assess the significance levels of the estimated SLE disease effects. We identified disease effects with p-value < 0.05 and magnitude > 0.015 as significant disease effects (Supplementary Figure 1).

Figure 2A displays the estimated factor matrices of the sender and receiver cell type dimension, which represent the activity patterns of sender cell types and receiver cell types. The contribution of each factor to the decomposition is shown in Supplementary Figure 2 (see Methods for details about the calculation of contributions). In both sender and receiver cell type dimension, for factor 1 with the largest contribution, all cell types display scores in the same direction, indicating a critical systematic biological process that involves all cell types. Factor 2 highlights a notable contrast between the lymphocyte group (encompassing B, NK, Prolif, $CD4^+$ T, and $CD8^+$ T cells) and the monocyte group (comprising cM, nCM, cDC, and pDC cells), demonstrating opposite activities of these two groups. Factor 3 and Factor 4 unveil distinct activity patterns specific to pDC cells and B cells, respectively, shedding light on the unique roles of these two cell types.

Figure 2B displays significant disease effects corresponding to CCC events with B, $CD8^+$ T, cM, and pDC cells as the receiver cell type. The significant effects of CCC events in other receiver cell types are shown in Supplementary Figure 3. Notably, multiple ligand-receptor pairs consistently exhibit positive associations with SLE across sender and receiver cell types. For instance, ligand-receptor pairs LGALS9 – PTPRC and LGALS9 – CD44 consistently show positive associations with SLE across cell types (Figure 2B). This discovery aligns with our earlier findings that the factors representing the systematic biological process involving all cell types have the largest contributions to the decomposition.

STACCato also effectively identified CCC events with cell type specific disease effects. For instance, ligand-receptor pair CD99 – PILRA showed negative associations with SLE only with B cells and pDC cells as the receiver cell types (Figure 2B). ligand-receptor pair CD22 – PTPRC demonstrated an significant association with SLE only with B cells as the sender cell type (Figure 2B), which is consistent with the knowledge that CD22 is a B-cell-specific glycoprotein[18].

One noteworthy aspect of this SLE dataset is its highly unbalanced study design, where batch 1 included only healthy controls while batch 2 included SLE patients predominantly (Supplementary Table 1). Consequently, batch confounded the association of CCC events with SLE. We applied Tensor-cell2cell[8], which does not consider confounding variables, to the same 4-dimensional communication score tensor of the SLE dataset (Supplementary Figure 4A) and identified three factors (factor 3, 5, 7) significantly associated with SLE disease (Supplementary Figure 4B). However, we found that these factors were also strongly associated with batch (Supplementary Figure 5), suggesting that the disease effect was confounded by the batch effect in these factors (Supplementary Figure 6). For instance, healthy controls exhibited significantly larger loadings in factor 3 (Supplementary Figure 4B), indicating a negative association between factor 3 and SLE. However, when excluding batch 1 samples, the difference between SLE patients and healthy controls in other batches became minimal in factor 3 (Supplementary Figure 6). These results demonstrated that batch 1 distorted the association between factor 3 and disease in Tensor-cell2cell, leading to misleading interpretations of factor 3's role in SLE. These findings highlighted the importance of adjusting for confounding effects in CCC inference.

Evaluating the impact of confounding variables on CCC inference with the SLE dataset

To evaluate the impact of confounding variables on CCC inference, we applied STACCato to the SLE dataset with three distinct models, each incorporating different sample-level variables: Model 1, whose results were shown in Figure 2 and described above, considers sample-level variables of disease status, batches, and all other available covariates including age, gender, and ancestry; Model 2 considers disease status and batches only; and Model 3 considers disease status only. When comparing Model 1 and Model 2 to Model 3, we observed substantial changes in the estimated disease effects before and after adjusting for batch effects (Supplementary Figure 7). For example, the ligand-receptor pairs macrophage migration inhibitory factor (MIF) – CD74&CXCR4 and MIF – CD74&CD44 showed negative associations with SLE before batch adjustment but positive associations with SLE after accounting for batch effects. Monoclonal antibodies like imalumab (anti-MIF) and milatuzumab (anti-CD74) have been assessed in early phase clinical trials, demonstrating efficacy in SLE treatment[19]. This suggests a positive association between MIF – CD74 and SLE, which is consistent with the results adjusting for batch effects. These findings underscore how confounding variables can distort true associations and emphasize the importance of considering confounding variables like batches in CCC inference.

We also compared the factor matrices estimated with and without adjustment of batch effects by calculating the normalized chordal distance between the estimated factor matrices. Normalized chordal distance is a metric ranging from 0 to 1 for measuring distances between subspaces. A larger chordal distance indicates a greater difference between the subspaces of the estimated factor matrices (see Methods for details about chordal distance). The normalized chordal distances between the factor matrices estimated before (Model 3) and after adjusting for batches (Model 2) were 0.009 for sender cell types and 0.013 for receiver cell types, indicating minor differences. These results illustrate that confounding variables can significantly influence the estimation of disease effects in CCC events while having a relatively minor impact on the estimation of factor matrices.

Applying STACCato to identify CCC events associated with ASD

We applied STACCato on the snRNA-seq dataset of postmortem tissue samples of prefrontal cortex from 13 ASD patients and 10 controls[12] to identify CCC events associated with ASD (see Methods for details). We considered 16 sender/receiver cell types: fibrous astrocytes (AST-FB), protoplasmic astrocytes (AST-PP), Endothelial, parvalbumin interneurons (IN-PV), somatostatin interneurons (IN-SST), SV2C interneurons (IN-SV2C), VIP interneurons (IN-VIP), layer 2/3 excitatory neurons (L2/3), layer 4 excitatory neurons (L4), layer 5/6 corticofugal projection neurons (L5/6), layer 5/6 cortico-cortical projection neurons (L5/6-CC), maturing neurons (Neu-mat), NRGN-expressing neurons (Neu-NRGN-I), NRGN-expressing neurons (Neu-NRGN-II), Oligodendrocyte precursor cells (OPC), and oligodendrocytes. We applied STACCato to a $23 \times 749 \times 16 \times 16$ communication score tensor (consisting of 23 samples, 749 ligand-receptor pairs, 16 sender cell types, 16 receiver cell types) to examine associations between CCC events and ASD, while adjusting for age, gender, and processing batch. We used the decomposition rank $\boldsymbol{r} = (r_Q = 5, r_J = 5, r_K = 5, r_L = 5)$. We used 4,999 iterations of bootstrapping resampling to assess the significance levels of the estimated ASD disease effects. We identified estimated disease effects with p-value < 0.05 and magnitude > 0.015 as significant disease effects (Supplementary Figure 8).

In Figure 3A, we present the estimated factor matrices of the sender and receiver cell type dimension, which depict the activity patterns of sender and receiver cell types. The contributions of all factors are shown in Supplementary Figure 9A – 9B. Similar to our findings in the SLE dataset, we observed that factor 1 contributed the most and reflected a systematic process involving all cell types. Factors 2 through 5 for both sender and receiver cell types successfully revealed 6 cell type groups with distinct activity patterns: (1) astrocytes group including AST-FB and AST-PP; (2) Endothelial; (3) inhibitory neurons group including IN-PV, IN-SST, IN-SV2C, IN-VIP; (4) excitatory neurons group including L2/3, L4, L5/6, L5/6-CC; (5) expressing neurons group including Neu-mat, Neu-NRGN-I, and Neu-NRGN-II; (6) neuroglia group including oligodendrocytes and OPC (Figure 3A).

For each pair of sender cell type and receiver cell type, we ranked the ligand-receptor pairs by the estimated ASD disease effects and performed preranked Gene Set Enrichment Analysis (GSEA)[20] to determine if ligand-receptor pairs belonging to a particular pathway are more likely to be clustered at the top or bottom of the ranked list, and thereby identifying pathways associated with ASD (see details of pathway enrichment analysis in the Methods section). Figure 3B shows significantly enriched KEGG pathways[21] across AST-PP, Endothelial, IN-PV, L2/3, and Neu-NRGN-I cells. A total of 10 significantly enriched pathways were identified, including the axon guidance, cell adhesion molecules (CAMs), cytokine-cytokine receptor interaction, extracellular matrix-receptor (ECM-receptor) interaction, ErbB signaling, focal adhesion, MAPK signaling, notch signaling, regulation of actin cytoskeleton, and small cell lung cancer. Importantly, 8 out of these 10 pathways (axon guidance, CAMs, ECM-receptor interaction, ErbB signaling, focal adhesion, MAPK signaling, regulation of actin cytoskeleton, small cell lung cancer) have been previously identified as significantly enriched pathways with p-values $< 5 \times 10^{-7}$ for ASD[22]. The molecules related to the notch signaling pathway have been shown to have increased expression in the PFC in an animal model of autism[23], which is consistent with our observation of a positive association of the notch signaling pathway with ASD between AST-FB and L2/3 cells.

Evaluating the impact of confounding variables on CCC inference with the ASD dataset

We also examined the impact of batch information on our ASD results by fitting three distinct STACCato models with Model 1 considering disease status and all available covariates including batches, age, and gender (as shown in Figure 3), Model 2 considering disease status and batches only, and Model 3 considering disease status only. Unlike the SLE dataset, the ASD dataset exhibits a fairly balanced design (Supplementary Table 2). Consequently, batch is no longer a confounding factor. As anticipated, the estimated disease effects remain consistent before and after adjusting for batch effects (Supplementary Figure 10). Interestingly, the chordal distances between the factor matrices estimated before (Model 3) and after adjusting for batch (Model 2) were 0.384 for sender cell types and 0.438 for receiver cell types, indicating substantial discrepancies in the estimated factor matrices before and after batch adjustment. We further evaluated the relative contributions of all sample-level variables and found that batch contributed substantially to the communication tensor, indicating a non-negligible batch effect on the communication scores (Supplementary Figure 9C). This underscores a crucial point — even in datasets with balanced designs, failing to account for variables with significant impacts on the CCC can significantly impact the estimation of factor matrices and, consequently, the interpretations of cell type activity patterns.

Simulation Study

We conducted simulations to investigate how sample-level variables affect the CCC inference in different study designs. We simulated the communication score tensor $\mathcal{Y} \in \mathbb{R}^{I \times J \times K \times L}$ from the supervised tensor decomposition model as in Equations 2 and 3. We set $\mathcal{G}$, $\boldsymbol{M}_Q$, $\boldsymbol{M}_J$, $\boldsymbol{M}_K$, $\boldsymbol{M}_L$ in Equation 3 as the core tensor and factor matrices estimated from the ASD dataset and simulated $\boldsymbol{X}$ for 60 subjects with intercept, disease status, and batch variables. The elements of $\mathcal{E}$ were independently simulated from a normal distribution with mean 0 and variance $\hat{\sigma}^2$, where $\hat{\sigma} = 0.05$ was taken as the standard error of the estimation residuals from ASD data. We considered a study with 30 disease subjects and 30 healthy controls processed in two batches. We considered three study designs: (1) balanced design with 15 controls and 15 disease subjects in both batches; (2) moderate unbalanced design with 20 controls and 10 disease subjects in batch 1, and 10 controls and 20 disease subjects in batch 2; (3) extreme unbalanced design with 30 controls and 5 disease subjects in batch 1, and batch 2 only contains 25 disease subjects.

We applied STACCato with two models: Model 1 considers disease status and batch variables, and Model 2 considers only disease status. We calculated the mean squared errors (MSEs) of the estimated disease effects across 100 simulations. Figure 4A shows that neglecting confounders in an unbalanced design can generate larger estimation errors, and the MSEs of the disease effect dramatically increased as the degree of imbalance became more extreme. We also assessed the proportion of estimated disease effects with opposite directions to the assumed one (Supplementary Figure 11). We found that, before adjusting for batch, 14.7% of the disease effects had incorrect estimated directions in the extremely unbalanced design, which was significantly higher than the proportion 3.1% after adjusting for batch. Additionally, we assessed the accuracy of the estimated factor matrices by calculating the chordal distance between the estimated factor matrices and the assumed factor matrices. We observed that neglecting the batch variable resulted in decreased accuracy in estimating the factor matrices (Figure 4B), especially in balanced and moderate unbalanced design. Failing to account for the batch variable prevents the identification of factors that are solely batch-associated and not disease-associated, resulting in inaccuracies in the estimated factor matrices. Conversely, in extreme unbalanced designs where batch and disease are strongly correlated, batch-associated factors are also strongly linked to the disease. In such scenarios, neglecting the batch variable did not significantly impact the accuracy of estimating the factor matrices. These observations align with our real-data analysis findings, suggesting that regardless of whether the dataset originates from a balanced or unbalanced design, incorporating information of sample-level variables into CCC inference consistently leads to more accurate estimations of disease effects or activity patterns of cell types.

We also compared STACCato to the separate regression procedure (Equation 1), where a regression model was fitted with communication scores as dependent variables and sample-level variables as independent variables separately for each CCC event. In contrast, STACCato employs the tensor technique to incorporate the correlations among CCC events and jointly estimates the effects of considered variables for all CCC events. Across all study designs, STACCato consistently achieved significantly lower MSE compared to the separate regression approach (Supplementary Figure 12), justifying the advantage of using the tensor technique to account for correlations among CCC events.

Computational Considerations

While a single STACCato decomposition only takes seconds, assessing the significance level of estimated effects by bootstrapping requires performing decompositions for a substantial number of bootstrapping iterations and takes hours of CPU time. We conducted the computational benchmarks using one Intel(R) Xeon(R) processor (2.10 GHz). For a simulated dataset comprising 100 samples, 10 sender and receiver cell types, 600 ligand-receptor pairs, and 10 sample-level covariates, 99 iterations of bootstrap resampling took around 11 minutes and ~1.3 GB memory usage on the upper-bound.

Considering that the numbers of cell types and sample-level covariates generally do not vary much in practice, we investigated how bootstrapping time and upper-bound memory usage vary with the number of samples and the number of ligand-receptor pairs. We simulated datasets with 10 sender and receiver cell types, 10 sample-level covariates, and various numbers of samples (ranging from 25 to 100) and ligand-receptor pairs (ranging from 150 to 600). With 99 iterations of bootstrap resampling, our simulation results revealed that computational time increased linearly with the number of samples (Supplementary Figure 13A) and quadratically with the number of ligand-receptor pairs (Supplementary Figure 14A). The upper bound memory usage changed approximately linearly with both the number of samples and ligand-receptor pairs (Supplementary Figures 13B, 14B).

Discussion

We present STACCato, a computational tool that utilizes multi-sample multi-condition scRNA-seq data to identify CCC events associated with conditions (e.g., disease status, multiple time points, different tissue types). STACCato utilizes supervised tensor decomposition to estimate the influence of the condition of interest on CCC events, while adjusting for potential confounding variables. Furthermore, it facilitates the identification of activity patterns among cell types involved in CCC. We applied STACCato to analyze a SLE dataset with an extremely unbalanced design[10,11] and an ASD dataset with a balanced design[12]. Additionally, we conducted simulation studies to mimic real data with different study designs. Our real data application and simulation results demonstrated STACCato's capability to incorporate available sample-level variables, thereby enabling more reliable inference regarding the associations between CCC events and conditions, as well as more robust estimations of activity patterns among cell types.

In practice, a common approach to address batch effects in scRNA-seq data is to remove batch effects before downstream analysis. This approach involves the estimation of batch effects, followed by the removal of these estimated batch effects to generate “batch-effect-free” data for downstream analysis. However, as noted by Nygaard et al.[24], this two-step procedure has a severe drawback: it relies on point estimates of batch effects while disregarding estimation errors. In this two-step process, even when the original batch effects could be eliminated, the estimation errors may introduce new batch effects. In contrast, STACCato incorporates potential confounding variables, such as batch effects, into the design matrix, and jointly estimates the effects of these confounders along with other variables in a single step. Moreover, although our application and simulation studies focused on addressing batch effects, STACCato can adjust for all potential confounding variables in biomedical research. For instance, age is often considered as a confounding factor in the identification of CCC events associated with Alzheimer's disease. By incorporating all potential confounding variables into the model, STACCato offers a comprehensive solution, allowing for simultaneous handling of multiple confounders and facilitating more accurate CCC inference.

In contrast to Tensor-cell2cell, which also employs the tensor decomposition technique for CCC inference, STACCato stands out in several key aspects. First, STACCato directly assesses the relationship between each CCC event and the condition of interest. In contrast, Tensor-cell2cell primarily provides insights into the association between the decomposed factors and conditions, without offering explicit interpretations regarding individual CCC events. Second, STACCato goes a step further by not only identifying associations but also estimating the condition effect for each CCC event and assessing the statistical significance of such an effect. In contrast, Tensor-cell2cell focuses on determining the significance of the association between factors and the condition, without providing detailed information on the magnitude of condition effects. Last, as highlighted throughout our paper, STACCato has the capability to account for confounding variables, a feature lacking in Tensor-cell2cell. Through our application of Tensor-cell2cell to the SLE dataset, we demonstrated its inability to effectively disentangle confounding effects from disease effects in the study of CCC events.

It is important to note that STACCato is a highly adaptable framework that can be seamlessly integrated with various existing CCC inference tools, each with its unique methods of constructing communication scores. Researchers have the flexibility to select any tool of interest to calculate communication scores. For example, one can use the LIANA tool[25], which incorporates a wide range of tools and resources to calculate cell-cell communication scores, to calculate communication scores for all CCC events and arrange the scores into a 3-dimensional communication score tensor per sample. The 3-dimensional tensors of all samples can subsequently be combined into the 4-dimensional communication score tensor, allowing STACCato to be applied for inferring CCC events associated with the specific condition of interest.

The STACCato framework does have its limitations. First, in scRNA-seq data, many genes may not be actively expressed in single cells, resulting in a significant proportion of zero values in the cell-cell communication score tensor. A future extension of STACCato involving sparse tensor decomposition, which imposes sparsity constraints on the ligand-receptor pairs, may inherently address this zero-inflation problem. Second, STACCato relies on a literature-curated database to perform CCC inference, limiting the identified condition-related CCC events to those documented in previous literature. Extending STACCato to identify novel ligand-receptor pairs is part of our ongoing research but falls outside the scope of this work.

To enable the use of STACCato by the public, we provide an integrated tool (see Code availability) to: (1) perform supervised tensor decomposition to estimate the effects of conditions on CCC events adjusting for covariates and infer activity patterns of cell types; (2) use bootstrapping resampling to assess the significance level of the estimated effects; (3) conduct downstream analyses including comparing significant CCC events across cell types and identifying pathways significantly associated with conditions. In conclusion, we present STACCato as a valuable tool to effectively incorporate sample-level variables and adjust for possible confounding variables in CCC inference using multi-sample multi-condition scRNA-seq data.

Methods

Construction of a 4-dimensional communication score tensor

With the matrix of gene expressions of multiple cell types from a scRNA-seq sample and the literature-curated list of ligand-receptor pairs, we can calculate the communication score for the CCC event involving the interaction of ligand-receptor pair j from sender cell type k to receiver cell type l as

$$y_{jkl} = f(\text{ligand}_k, \text{receptor}_l)$$

where y_{jkl} denotes the communication score; ligand_k denotes the expression of the ligand in sender cell type k; receptor_l denotes the expression of the receptor in receiver cell type l; and f denotes the scoring function (Figure 1A). In this study, we used the scoring function $y_{jkl} = \sqrt{\text{ligand}_k \times \text{receptor}_l}$. Other available scoring functions have been previously summarized by Armingol et al.[26] and Dimitrov et al[25].

Once we compute communication scores for a specific ligand-receptor pair j across all K sender cell types and L receiver cell types, we can create a communication score matrix (Figure 1B). In this matrix, the rows represent K sender cell types; the columns represent L receiver cell types; and the element located in the k^{th} row and l^{th} column corresponds to the value of y_{jkl}. By repeating this process for all J ligand-receptor pairs, we will get J matrices, which can be arranged into a sample-specific 3-dimensional tensor with dimensions $J \times K \times L$ (Figure 1B). Then the 3-dimensional tensor of all samples can be arranged into a 4-dimensional tensor with dimensions of I samples, J ligand-receptor pairs, K sender cell types, and L receiver cell types (Figure 1C). In the application studies of the SLE dataset and ASD dataset, we constructed the 4-dimensional tensor using the Tensor-cell2cell package[8] (see Code availability). In the final tensor, we only included ligand-receptor pairs with both ligands and receptors shared across all samples.

STACCato incorporates correlations among CCC events

Consider the full supervised tensor decomposition model in Equation 3,

$$\mathcal{Y} = \mathcal{B} \times_1 \boldsymbol{X} + \mathcal{E} = \mathcal{G} \times \{\boldsymbol{M}_Q, \boldsymbol{M}_J, \boldsymbol{M}_K, \boldsymbol{M}_L\} \times_1 \boldsymbol{X} + \mathcal{E}.$$

Elementwise, we have

$$\beta_{qjkl} = \sum_{r_1=1}^{r_Q} \sum_{r_2=1}^{r_J} \sum_{r_3=1}^{r_K} \sum_{r_4=1}^{r_L} g_{r_1 r_2 r_3 r_4} M_Q^{qr_1} M_J^{jr_2} M_K^{kr_3} M_L^{lr_4} \quad \text{(Equation 4)}$$

where $g_{r_1 r_2 r_3 r_4}$ denotes the (r_1, r_2, r_3, r_4) entry of $\mathcal{G}$, $M_Q^{qr_1}$ denotes the entry in the q^{th} row and r_1^{th} column of M_Q, similarly for $M_J^{jr_2}$, $M_K^{kr_3}$, and $M_L^{lr_4}$. Then for $k \neq k'$ and $l \neq l'$,

$$\beta_{qjk'l'} = \sum_{r_1=1}^{r_Q} \sum_{r_2=1}^{r_J} \sum_{r_3=1}^{r_K} \sum_{r_4=1}^{r_L} g_{r_1 r_2 r_3 r_4} M_Q^{qr_1} M_J^{jr_2} M_K^{k'r_3} M_L^{l'r_4} \quad \text{(Equation 5)}$$

Equations 4 and 5 represent the effects of covariate q on two different CCC events with the same ligand-receptor pair j but different sender (sender cell type k in Equation 4 and k' in Equation 5) and receiver cell types (receiver cell type l in Equation 4 and l' in Equation 5). These two equations share the same parameters

$M_J^{jr_2}, r_2 = 1, \cdots r_J$. Similarly, for CCC events with the same sender cell type k, the effects share the same parameters $M_K^{kr_3}$, $r_3 = 1, \cdots r_K$; for CCC events with the same receiver cell type l, the effects share the same parameters $M_L^{lr_4}$, $r_4 = 1, \cdots r_L$. In STACCato, the effects of covariates on correlated CCC events share parameters, enabling it to effectively incorporate the complex correlation structure among these CCC events.

STACCato Optimization

We first determine the number of components r_J, r_K, r_L for ligand-receptor pair, sender cell type, and receiver cell type dimension. For each dimension, we start by performing tensor unfolding to rearrange the elements of the communication score tensor into a matrix. For example, for the ligand-receptor pair dimension, we transform $\mathcal{Y} \in \mathbb{R}^{I \times J \times K \times L}$ into a matrix $Y_{(J)}$ with J rows and $I \times K \times L$ columns. Then we set r_J as the number of components that can explain more than 1% of the variation in $Y_{(J)}$. We follow the same approach to determine r_K for sender cell type dimension and r_L for receiver cell type dimension. We set r_Q as the number of sample-level variables available in $\boldsymbol{X}$.

Denoting the supervised decomposition rank $\boldsymbol{r} = (r_Q, r_J, r_K, r_L)$, we follow the optimization algorithm proposed by Hu et al. [16] to estimate $\mathcal{B}, \mathcal{G}, \boldsymbol{M}_Q, \boldsymbol{M}_J, \boldsymbol{M}_K\ \boldsymbol{M}_L$:

```
Algorithm 1:
Input: communication score tensor $\mathcal{Y} \in \mathbb{R}^{I \times J \times K \times L}$, sample-level design matrix $\boldsymbol{X} \in \mathbb{R}^{I \times Q}$, rank $\boldsymbol{r}$.
    1. Normalize sample-level design matrix via QR factorization $\boldsymbol{X} = \boldsymbol{QR}$.
    2. Project $\mathcal{Y}$ to the multilinear sample-level variable space to obtain the unconstrained coefficient tensor:
       $\tilde{\mathcal{B}} = \mathcal{Y} \times_1 \boldsymbol{Q}^{\boldsymbol{T}}$.
    3. Obtain rank-unconstrained coefficient tensor by performing a rank-$\boldsymbol{r}$ higher-order orthogonal iteration
       (HOOI)[27] on $\tilde{\mathcal{B}}$: $\hat{\mathcal{B}}^{(0)} \leftarrow HOOI(\tilde{\mathcal{B}}, \boldsymbol{r})$.
    4. Obtain estimated coefficient tensor by re-normalizing $\hat{\mathcal{B}}^{(0)}$ back to the original feature scales:
       $\hat{\mathcal{B}} = \hat{\mathcal{B}}^{(0)} \times_1 \boldsymbol{R}^{-1}$.
    5. Estimate $\mathcal{G}, \boldsymbol{M}_Q, \boldsymbol{M}_J, \boldsymbol{M}_K, \boldsymbol{M}_L$ by performing a rank-$\boldsymbol{r}$ HOOI on $\hat{\mathcal{B}}$: $\hat{\mathcal{B}} \approx \hat{\mathcal{G}} \times \{\widehat{\boldsymbol{M}_Q}, \widehat{\boldsymbol{M}_J}, \widehat{\boldsymbol{M}_K}, \widehat{\boldsymbol{M}_L}\}$.
Output: $\hat{\mathcal{B}}, \hat{\mathcal{G}}, \widehat{\boldsymbol{M}_Q}, \widehat{\boldsymbol{M}_J}, \widehat{\boldsymbol{M}_K}, \widehat{\boldsymbol{M}_L}$.
```

We also impose orthonormality on $\boldsymbol{M}_Q$, $\boldsymbol{M}_J$, $\boldsymbol{M}_K$ $\boldsymbol{M}_L$ to ensure the uniqueness of decomposition.

Parametric bootstrapping for hypothesis testing

Denote the estimated communication score tensor as $\hat{\mathcal{Y}} = \hat{\mathcal{B}} \times_1 \boldsymbol{X}$ with entry $\hat{y}_{ijkl}$ and the estimated standard error of ϵ_{ijkl} as $\hat{\sigma}$, we have residual tensor $\mathcal{S} = \mathcal{Y} - \hat{\mathcal{Y}}$ with entry $s_{ijkl} = y_{ijkl} - \hat{y}_{ijkl}$, and $\hat{\sigma}^2 = var(\text{vec}(\mathcal{S}))$, where $\text{vec}(\mathcal{S}) = [s_{1111}, \cdots, s_{ijkl}]$ denotes the vectorized version of tensor $\mathcal{S}$.

For the n^{th} bootstrap resampling[17], we generate a new tensor $\mathcal{S}^{(n)}$ with entries from $N(0, \hat{\sigma}^2)$ and construct a new communication score tensor $\mathcal{Y}^{(n)} = \hat{\mathcal{Y}} + \mathcal{S}^{(n)}$. We perform STACCato on $\mathcal{Y}^{(n)}$ to estimate a new coefficient tensor $\hat{\mathcal{B}}^{(n)}$. We repeat this procedure for N iterations to generate $\hat{\mathcal{B}}^{(1)}, \hat{\mathcal{B}}^{(2)}, \cdots, \hat{\mathcal{B}}^{(N)}$. To test the null hypothesis of $H_0\!: b_{qjkl} = 0$, we follow the guideline suggested by Hall and Wilson[28] to define the bootstrap p-value as:

$$p_{qjkl} = \frac{\sum_{n=1}^{N} I\left(\left|\hat{b}_{qjkl}^{(n)} - \hat{b}_{qjkl}\right| > |\hat{b}_{qjkl}|\right)}{N+1}$$

where $\hat{b}_{qjkl}^{(n)}$ denotes the (q, j, k, l) entry of $\hat{\mathcal{B}}^{(n)}$; $\hat{b}_{qjkl}$ denotes the (q, j, k, l) entry of $\hat{\mathcal{B}}$, which is the estimated effect of variable q on the CCC events involving the ligand-receptor pair j between sender cell type k and receiver cell type l; and p_{qjkl} is the bootstrapping p-value for $\hat{b}_{qjkl}$.

Calculation of contributions

To calculate the contributions of factors of the sender and receiver cell types, we remove each factor from the decomposition results and assess the changes in the estimated outcome. For example, for factor 1 in the sender cell type dimension, we first remove the first column of the estimated factor matrix $\widehat{\boldsymbol{M}_K}$ and construct a new factor matrix $\widehat{\boldsymbol{M}_K}^* \in \mathbb{R}^{K \times (r_K - 1)}$. We then eliminate the interactions between this factor and factors in other dimensions from the estimated core tensor $\hat{\mathcal{G}}$, creating a new core tensor $\hat{\mathcal{G}}^* \in \mathbb{R}^{r_Q \times r_J \times (r_K - 1) \times r_L}$. With the modified factor matrices and core tensor, we calculate a new predicted communication score tensor $\hat{\mathcal{Y}}^* =$

$\hat{\mathcal{G}}^* \times \{\widehat{\boldsymbol{M}_Q}, \widehat{\boldsymbol{M}_J}, \widehat{\boldsymbol{M}_K}^*, \widehat{\boldsymbol{M}_L}\} \times_1 \boldsymbol{X}$. The contribution of the removed factor is defined as the mean squared difference between the entries of $\hat{\mathcal{Y}}^*$ and the original estimated $\hat{\mathcal{Y}} = \hat{\mathcal{G}} \times \{\widehat{\boldsymbol{M}_Q}, \widehat{\boldsymbol{M}_J}, \widehat{\boldsymbol{M}_K}, \widehat{\boldsymbol{M}_L}\} \times_1 \boldsymbol{X}$.

Chordal distance between two subspaces

We use normalized chordal distance[29] to evaluate the distance between the column spaces of two factor matrices. Let $A \in \mathbb{R}^{d_1 \times d_2}$, $B \in \mathbb{R}^{d_1 \times d_2}$ as two matrices whose columns are the orthonormal bases of two subspaces $\mathbf{A}$ and $\mathbf{B}$, and $A^T B = U \Sigma V^T$ as the full singular value decomposition (SVD) of $A^T B$ with $\Sigma = diag(\sigma_1, \sigma_2, \cdots, \sigma_{d_2})$. The principal angles $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{d_2}$ between the subspaces $\mathbf{A}$ and $\mathbf{B}$ are given by:

$$\theta_i = \cos^{-1} \sigma_i, i = 1, \cdots, d_2$$

The chordal distance between the subspaces $\mathbf{A}$ and $\mathbf{B}$ is given by:

$$d(\mathbf{A}, \mathbf{B}) = \left(\sum_{i=1}^{d_2} \sin^2 \theta_i\right)^{\frac{1}{2}}.$$

Here, we use the normalized chordal distance $d^*(\mathbf{A}, \mathbf{B}) = \left(\frac{1}{d_2} \sum_{i=1}^{d_2} \sin^2 \theta_i\right)^{\frac{1}{2}}$ so that the measure is bounded within [0,1]. We used the R function *chord.norm.diff* from CJIVE package[30] (see Code availability) to calculate the normalized chordal distance.

RNA-seq data processing

For all scRNA-seq datasets used in the study, we filtered out genes expressed in fewer than 4 cells and utilized the provided cell type labels from the metadata. For each sample in the dataset, we aggregate gene expression from single cells/nuclei into cell types by calculating the fraction of cells with non-zero counts within each cell type. Therefore, the aggregated cell-type specific gene expression is bounded within [0,1]. This approach is endorsed by Tensor-cell2cell for the accurate representation of genes with low expression levels[8,31], which is common among genes responsible for encoding surface proteins[32].

Literature-curated lists of ligand-receptor pairs

We downloaded the human list of 2,005 ligand-receptor pairs from a public available compendium of lists of ligand-receptor pairs (see Data availability). This list of ligand-receptor pairs was originally curated by Jin et al[1].

scRNA-seq dataset of SLE patients and controls

The SLE scRNA-seq dataset collects multiplexed scRNA-seq of 264 PBMC samples from 162 SLE patients and 99 healthy controls[10,11]. The data in h5ad format was obtained from NCBI's Gene Expression Omnibus[33] with GEO accession number 174188 (see Data availability). From the h5ad data, we extracted the raw UMI counts of 32,738 genes across 1,263,676 cells from 264 samples and 99 technical replicates. We reduced the dataset down to one sample per subject by selecting the sample with the largest number of cells.

The metadata, which was also extracted from the h5ad data, includes the information of age, processing batch, ancestry, and gender of subjects. 107 (41%) subjects are Asian, 149 (57%) subjects are European, 3 (1%) subjects are African American, and 2 (1%) subjects are Hispanic. We filtered out 5 samples of African American or Hispanic history, and only kept samples containing 9 main cell types: B, NK, Prolif, CD4$^+$ T cells, CD8$^+$ T cells, cM, ncM, cDC, and pDC cells. The remaining 251 samples include 154 SLE patients and 97 healthy controls from 4 processing batches. The constructed CCC tensor for the SLE dataset resulted in a 4-dimensional tensor with 251 subjects, 55 ligand-receptor pairs, 9 sender cell types, and 9 receiver cell types.

scRNA-seq dataset of ASD patients and controls

For the ASD dataset, we downloaded the log2-transformed UMI counts of PFC samples and the corresponding metadata from the UCSC Cell Browser[34] (see Data availability). The raw dataset contains the expression levels of 36,501 genes across 62,166 cells from 13 ASD patients and 10 healthy controls[12]. The constructed CCC tensor for the ASD dataset resulted in a 4-dimensional tensor with 23 subjects, 749 ligand-receptor pairs, 16 sender cell types, and 16 receiver cell types.

Gene set enrichment analysis

We follow the procedure proposed in Tensor-cell2cell to conduct the GSEA. A ligand-receptor pair is considered in a pathway if all the genes participating in the ligand-receptor pair are in the pathway. We consider the 22 KEGG pathways selected by Tensor-cell2cell (see Data availability). For one pair of sender cell type and receiver cell type, we first rank ligand-receptor pairs by their estimated disease effects, and then use the *prerank* module in the Python package GSEApy[35] (see Code availability) with 4999 permutations, gene sets with at least 15 elements, and a score weight of 1 to calculate the enrichment p-value and normalized enrichment score. We then combined the results from all tested pairs of cell types, and performed false discovery rate (FDR) correction to adjust for multiple comparisons. Pathways with FDR q-value < 0.05 were identified as pathways significantly associated with disease.

Acknowledgements

This work was supported by National Institutes of Health grant awards R35GM138313 (QD, JY) and RF1AG071170 (QD, MPE).

Data availability

The human list of 2,005 ligand-receptor pairs was downloaded from https://github.com/LewisLabUCSD/Ligand-Receptor-Pairs/blob/master/Human/Human-2020-Jin-LR-pairs.csv. The processed data of the SLE dataset in h5ad format was downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE174188. The log2-transformed UMI counts of the ASD dataset was downloaded from https://cells.ucsc.edu/autism/downloads.html. The KEGG pathways selected by Tensor-cell2cell to perform GSEA was downloaded from https://codeocean.com/capsule/9737314/tree/v2/data/LR-Pairs/CellChat-LR-KEGG-set.pkl.

Code availability

Source code for STACCato is available from https://github.com/daiqile96/STACCato. Source code for CJIVE is available from https://cran.r-project.org/web/packages/CJIVE/index.html. Source code for Tensor-cell2cell is available from https://github.com/earmingol/cell2cell. Source code for GSEApy is available from https://github.com/zqfang/GSEApy.

Reference

1. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat Commun* **12**, 1088 (2021).
2. 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).
3. Raredon, M. S. B. *et al.* Computation and visualization of cell–cell signaling topologies in single-cell systems data using Connectome. *Sci Rep* **12**, 4187 (2022).
4. Hu, Y., Peng, T., Gao, L. & Tan, K. CytoTalk: De novo construction of signal transduction networks using single-cell transcriptomic data. *Science Advances* **7**, eabf1356.
5. Wang, Y. *et al.* iTALK: an R Package to Characterize and Illustrate Intercellular Communication. 507871 Preprint at https://doi.org/10.1101/507871 (2019).
6. Hou, R., Denisenko, E., Ong, H. T., Ramilowski, J. A. & Forrest, A. R. R. Predicting cell-to-cell communication networks using NATMI. *Nat Commun* **11**, 5011 (2020).
7. Cabello-Aguilar, S. *et al.* SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Research* **48**, e55–e55 (2020).
8. Armingol, E. *et al.* Context-aware deconvolution of cell–cell communication with Tensor-cell2cell. *Nat Commun* **13**, 3665 (2022).
9. Tsuyuzaki, K., Ishii, M. & Nikaido, I. scTensor detects many-to-many cell–cell interactions from single cell RNA-sequencing data. 2022.12.07.519225 Preprint at https://doi.org/10.1101/2022.12.07.519225 (2022).
10. Thompson, M. *et al.* Multi-context genetic modeling of transcriptional regulation resolves novel disease loci. *Nat Commun* **13**, 5704 (2022).
11. Perez, R. K. *et al.* Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).
12. Nassir, N. *et al.* Single-cell transcriptome identifies molecular subtype of autism spectrum disorder impacted by de novo loss-of-function variants regulating glial cells. *Human Genomics* **15**, 68 (2021).

13. Liao, M. *et al.* Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat Med* **26**, 842–844 (2020).

14. Hore, V. *et al.* Tensor decomposition for multi-tissue gene expression experiments. *Nat Genet* **48**, 1094–1100 (2016).

15. Jung, I., Kim, M., Rhee, S., Lim, S. & Kim, S. MONTI: A Multi-Omics Non-negative Tensor Decomposition Framework for Gene-Level Integrative Analysis. *Frontiers in Genetics* **12**, (2021).

16. Hu, J., Lee, C. & Wang, M. Generalized Tensor Decomposition With Features on Multiple Modes. *Journal of Computational and Graphical Statistics* **31**, 204–218 (2022).

17. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap*. (CRC Press, 1994).

18. Kelm, S., Gerlach, J., Brossmer, R., Danzer, C.-P. & Nitschke, L. The Ligand-binding Domain of CD22 Is Needed for Inhibition of the B Cell Receptor Signal, as Demonstrated by a Novel Human CD22-specific Inhibitor Compound. *J Exp Med* **195**, 1207–1213 (2002).

19. Bilsborrow, J. B., Doherty, E., Tilstam, P. V. & Bucala, R. Macrophage migration inhibitory factor (MIF) as a therapeutic target for rheumatoid arthritis and systemic lupus erythematosus. *Expert Opin Ther Targets* **23**, 733–744 (2019).

20. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).

21. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**, 27–30 (2000).

22. Wen, Y., Alshikho, M. J. & Herbert, M. R. Pathway Network Analyses for Autism Reveal Multisystem Involvement, Major Overlaps with Other Diseases and Convergence upon MAPK and Calcium Signaling. *PLoS ONE* **11**, (2016).

23. Zhang, Y. *et al.* The Notch signaling pathway inhibitor Dapt alleviates autism-like behavior, autophagy and dendritic spine density abnormalities in a valproic acid-induced animal model of autism. *Prog Neuropsychopharmacol Biol Psychiatry* **94**, 109644 (2019).

24. Nygaard, V., Rødland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* **17**, 29–39 (2016).

25. Dimitrov, D. *et al.* Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nat Commun* **13**, 3224 (2022).

26. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell–cell interactions and communication from gene expression. *Nat Rev Genet* **22**, 71–88 (2021).

27. Kolda, T. G. & Bader, B. W. Tensor Decompositions and Applications. *SIAM Rev.* **51**, 455–500 (2009).

28. Hall, P. & Wilson, S. R. Two Guidelines for Bootstrap Hypothesis Testing. *Biometrics* **47**, 757–762 (1991).

29. Ye, K. & Lim, L.-H. Schubert Varieties and Distances between Subspaces of Different Dimensions. *SIAM J. Matrix Anal. & Appl.* **37**, 1176–1197 (2016).

30. Murden, R. J., Zhang, Z., Guo, Y. & Risk, B. B. Interpretive JIVE: Connections with CCA and an application to brain connectivity. *Frontiers in Neuroscience* **16**, (2022).

31. Booeshaghi, A. S. & Pachter, L. Normalization of single-cell RNA-seq counts by log(x + 1)† or log(1 + x)†. *Bioinformatics* **37**, 2223–2224 (2021).

32. Baccin, C. *et al.* Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. *Nat Cell Biol* **22**, 38–48 (2020).

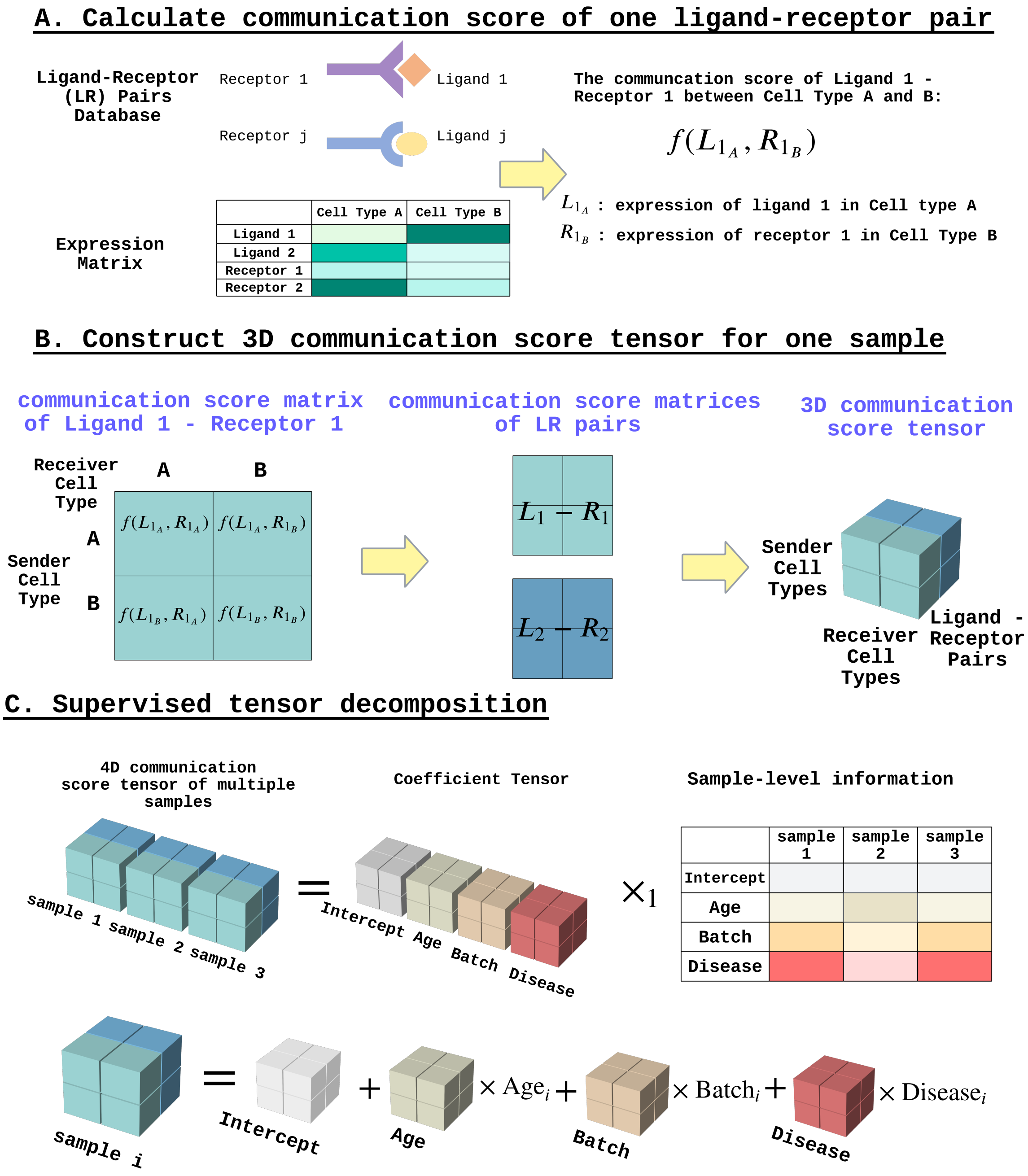
33. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207–210 (2002).

34. Speir, M. L. *et al.* UCSC Cell Browser: visualize your single-cell data. *Bioinformatics* **37**, 4578–4580 (2021).

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

Figures and Tables

Figure 1. STACCato analytic framework. (A) Cell-cell communication (CCC) score is given by a function of the expression levels of ligand 1 in sender cell type A (L_{1_A}) and receptor 1 in receiver cell type B (R_{1_B}). (B) CCC scores are calculated for a specific ligand-receptor pair across all sender and receiver cell types. CCC scores are then organized into a communication score matrix with sender cell types as rows and receiver cell types as columns. Communication score matrices are repeatedly calculated for all ligand-receptor pairs and organized into a 3-dimensional communication score tensor. (C) The 3-dimensional communication score tensors are repeatedly constructed for all samples and then combined into a 4-dimensional communication score tensor. STACCato then uses subject-level information to estimate the coefficient tensor representing the


A. Calculate communication score of one ligand-receptor pair
Ligand-Receptor (LR) Pairs Database
Receptor 1
Ligand 1
Receptor j
Ligand j
The communcation score of Ligand 1 - Receptor 1 between Cell Type A and B:
$f(L_{1_A}, R_{1_B})$
L_{1_A} : expression of ligand 1 in Cell type A
R_{1_B} : expression of receptor 1 in Cell Type B
Expression Matrix
Cell Type A
Cell Type B
Ligand 1
Ligand 2
Receptor 1
Receptor 2
B. Construct 3D communication score tensor for one sample
communication score matrix of Ligand 1 - Receptor 1
communication score matrices of LR pairs
3D communication score tensor
Receiver Cell Type
A
B
Sender Cell Type
A
B
$f(L_{1_A}, R_{1_A})$
$f(L_{1_A}, R_{1_B})$
$f(L_{1_B}, R_{1_A})$
$f(L_{1_B}, R_{1_B})$
$L_1 - R_1$
$L_2 - R_2$
Sender Cell Types
Receiver Cell Types
Ligand - Receptor Pairs
C. Supervised tensor decomposition
4D communication score tensor of multiple samples
Coefficient Tensor
Sample-level information
sample 1
sample 2
sample 3
$=$
Intercept
Age
Batch
Disease
$\times_1$
sample 1
sample 2
sample 3
Intercept
Age
Batch
Disease
sample i
$=$
Intercept
$+$
Age
$\times \text{Age}_i +$
Batch
$\times \text{Batch}_i +$
Disease
$\times \text{Disease}_i$

effects of subject-level variables on CCC events. While this example tensor contains only 2 cell types and 2 ligand-receptor pairs, the framework is generalizable to any number of cell types and ligand-receptor pairs.

Figure 2. STACCato results with the SLE dataset. (A) Bar plots of the estimated values in the factor matrices of sender and receiver cell types. Each color represents one cell type. (B) Estimated significant disease effects with p-values < 0.05 and magnitudes > 0.015 for communication events with B, CD8$^+$ T, cM, and pDC cells as receiver cell types. Positive disease effects are colored in red while negative disease effects are colored in blue. Positive disease effects indicate positive associations between CCC events and SLE, while negative disease effects indicate negative associations.

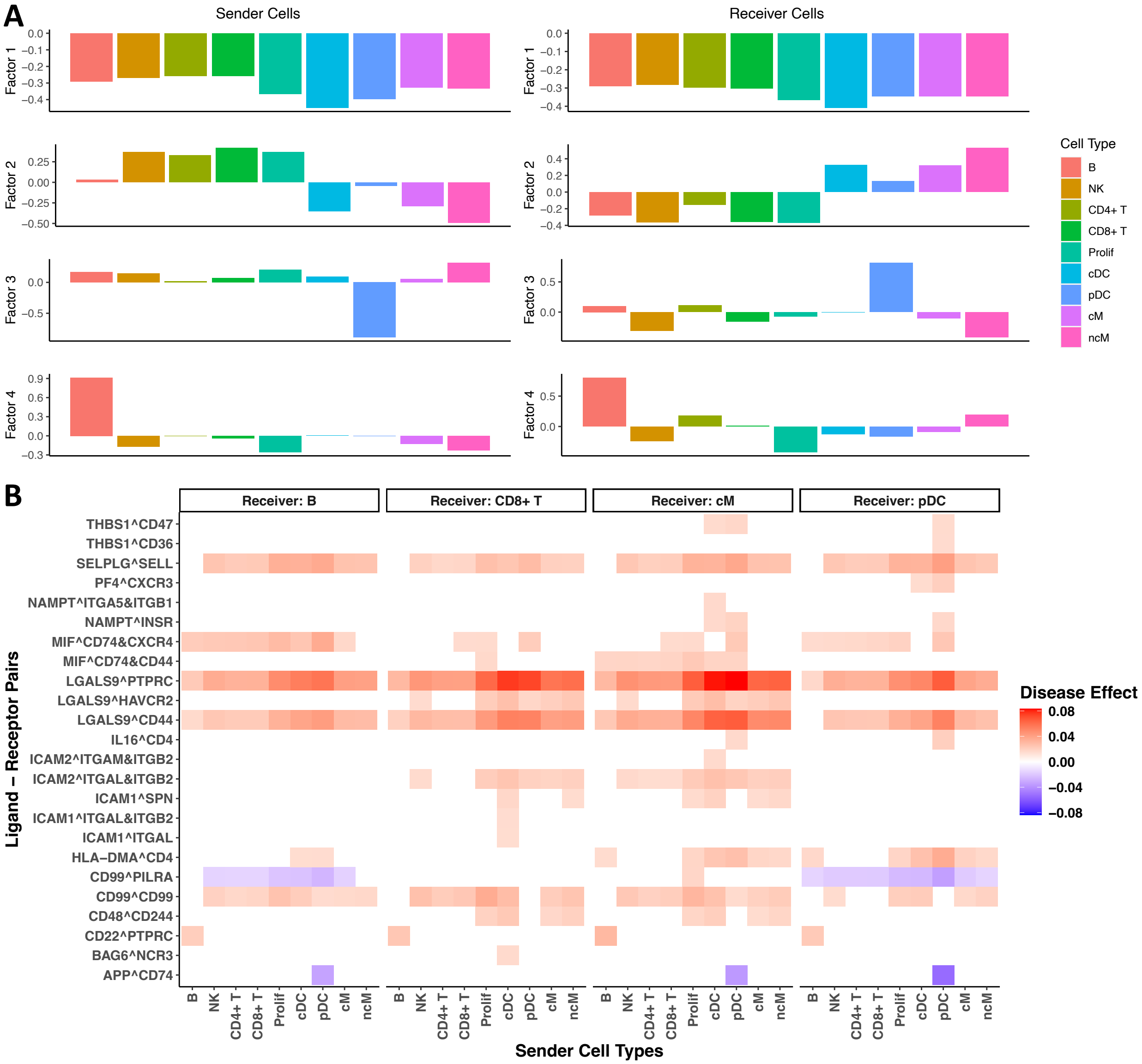
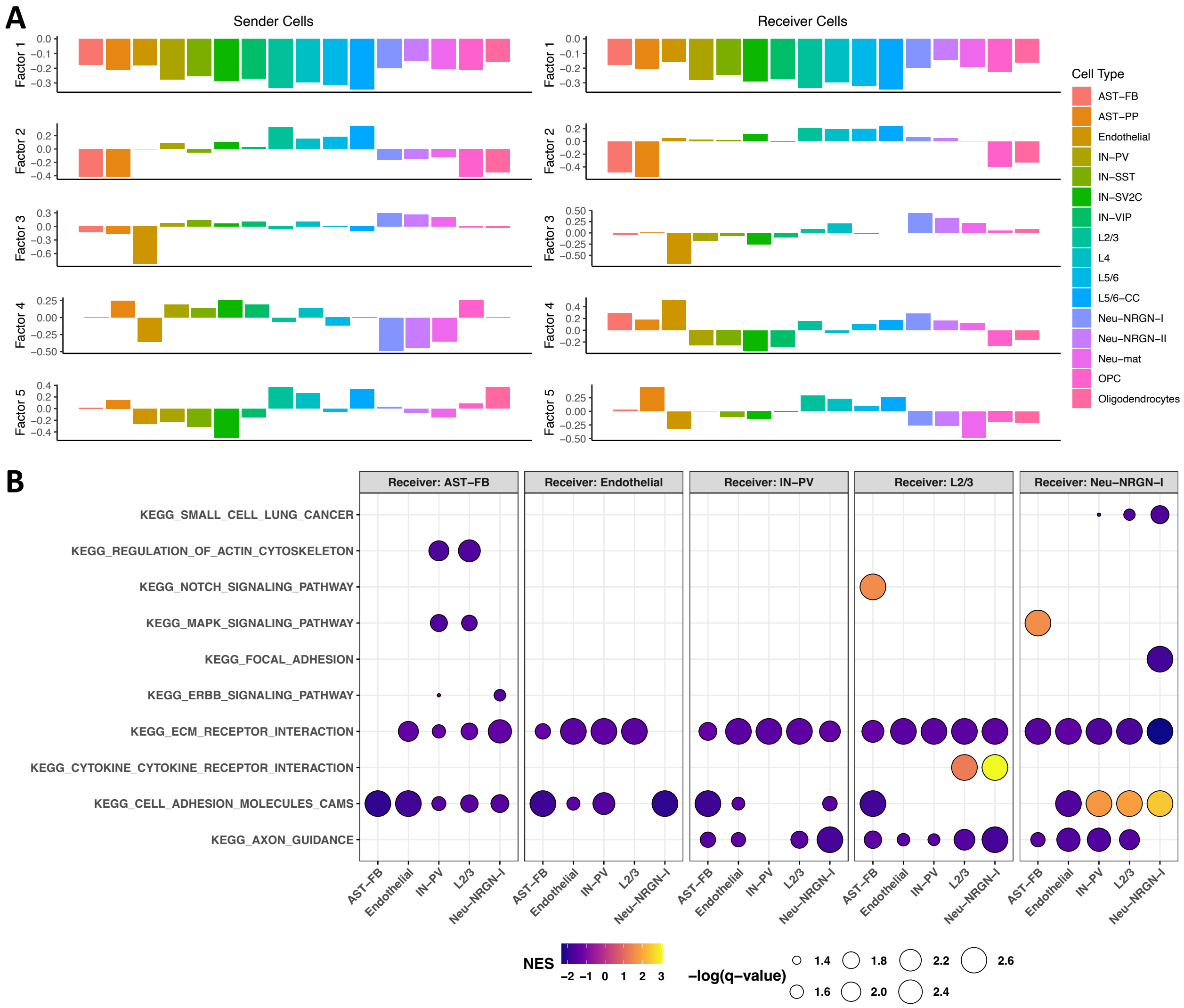
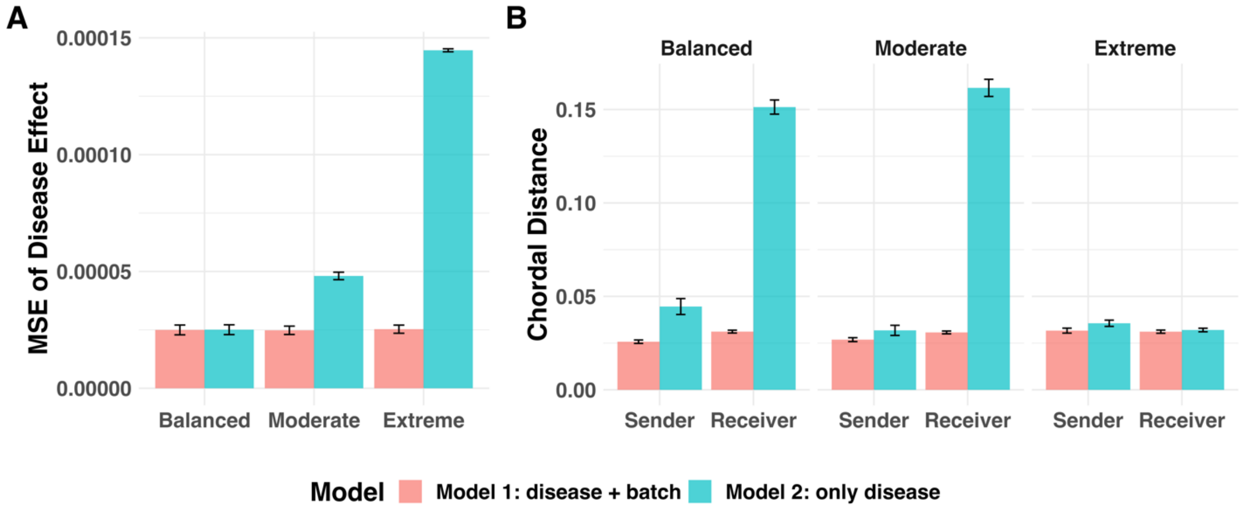


Figure 3. STACCato results with the ASD dataset. (A) Bar plots of the estimated values in the factor matrices of sender and receiver cell types. Each color represents one cell type. (B) Significantly enriched KEGG pathways with false discovery rate (FDR) adjusted p-value (q-value) < 0.05 across AST-PP, Endothelial, IN-PV, L2/3, and Neu-NRGN-I sender and receiver cell types. Colors represent the normalized enrichment scores. Positive enrichment scores indicate positive associations with ASD, while negative enrichment score indicate negative associations with ASD.



614 **Figure 4. STACCato simulation results:** MSE of estimated disease effects (A) and chordal distance of
615 estimated factor matrices (B) in balanced, moderate unbalanced, and extreme unbalanced scenarios. The bar
616 plot shows the average MSEs across 100 simulations from Model 1 considering disease status and batch (red
617 bars) and Model 2 considering disease status only (green bars) with black error bars showing standard errors.


A
MSE of Disease Effect
Balanced
Moderate
Extreme
B
Balanced
Moderate
Extreme
Chordal Distance
Sender
Receiver
Sender
Receiver
Sender
Receiver
Model
Model 1: disease + batch
Model 2: only disease