Ruprecht-Karls-Universität Heidelberg
Fakultät für Biowissenschaften
Bachelorstudiengang Molekulare Biotechnologie

Robustness Evaluation of Cell-Cell-Communication Inference Methods

Bachelorarbeit

P. Leo Burmedi

Abgabetermin: November 2021

Die vorliegende Bachelor-/Masterarbeit wurde im Institute for Computational Biomedicine in der Arbeitsgruppe von Prof. Saez-Rodriguez an der Universität Heidelberg in der Zeit vom 09/08/2021 bis 01/11/2021 angefertigt.

Gutachter der Arbeit: Prof. Julio Saez-Rodriguez
AG Saez-Rodriguez
Institute for Computational Biomedicine

Ich erkläre hiermit ehrenwörtlich, dass:

1. ich die vorliegende Bachelorarbeit selbständig unter Anleitung verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe;
2. die Übernahme wörtlicher Zitate aus der Literatur/Internet sowie die Verwendung der Gedanken anderer Autoren an den entsprechenden Stellen innerhalb der Arbeit gekennzeichnet wurde;
3. ich meine Bachelorarbeit bei keiner anderen Prüfung vorgelegt habe.

Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

01.11.2021
Ort, Datum

Unterschrift

Hinweis:
Ohne Unterschrift, Datums- und Ortsangabe ist diese Erklärung ungültig und damit auch die Bachelorarbeit.

Acknowledgements

I would like to thank Daniel Dimitrov for his steadfast guidance and the wealth of knowledge he brought to the project, as well as the entire Saez-Rodriguez Group for their expertise, support and the welcoming environment they foster.

I would not have been able to complete this project without the support of my family, friends, and most importantly my partner Laura.

Lastly, to Ray, thanks for holding on. This one's for you.

Abstract

Cell-cell communication is vital to many biological processes but is challenging to measure at scale. Single-cell transcriptomics provides an alternative approach which infers cell-cell communication and offers easier analysis and a greater breadth of information. There are many methods that perform this type of cell-cell communication inference, but they are often difficult to benchmark and lack a gold standard. As such, their relative strengths and weaknesses remain fairly unknown. To address this, we analysed the robustness of six inference methods with regards to four common sources of noise. We found unstable clustering and the choice of resource to be the most impactful factors on method robustness while laxly curated resources and small sample sizes were of overall low impact. Moreover, the degree of cluster sensitivity and the way in which this sensitivity is assessed proved to be the main differentiating trait between the inference methods' performance, with more cluster sensitive methods being generally less robust. Lastly, our results highlight that the relationship between an inference method's design choices and its consequent qualities remains challenging to assess.

Zusammenfassung

Zellkommunikation ist von zentraler Wichtigkeit für viele biologische Prozesse, aber oft schwer in direkter Vermessung zu erfassen. Einzelzell Genexpressionsanalyse bietet eine Alternative an, um Rückschlüsse auf Zellkommunikation zu erlangen. Vorteile dieses Ansatzes sind unter anderem eine leichtere und umfassendere Analyse. Dementsprechend wurden einige Zellkommunikationsinferenzmethoden auf Basis der einzelzell Genexpressionsanalyse entwickelt, diese sind aber oft schwer zu vergleichen. Ihre jeweiligen Stärken und Schwächen sind weitgehend unbekannt. Aus diesem Grund haben wir die Robustheit von sechs Zellkommunikationsinferenzmethoden im Kontext von vier häufigen Fehlerquellen analysiert. Unsere Ergebnisse deuten, dass ein instabiles Clustering und die Wahl der Ressource einen erheblichen Einfluss auf die Robustheit hatten, wo hingegen eine geringere Probenanzahl und eine weniger strikte Ressourcekuration einen geringeren Einfluss haben. Des Weiteren unterschieden sich die Methoden in ihrer Robustheit hauptsächlich wegen ihrer Vermessung und Wertschätzung der Clusterspezifität der Zellkommunikation. Letztendlich deuten unsere Ergebnisse, dass das Verhältnis zwischen dem Design einer Zellkommunikationsinferenzmethode und deren endgültigen Eigenschaften schwer zu erfassen ist.

Table of Contents

1. Introduction

Cell-Cell-Communication ("CCC") is vital to any multicellular organism. The ability to coordinate one another's actions is central to a cellular community's survival and enables differentiation and division of labor. As such, CCC plays a role in cellular development, homeostasis, and deregulation, making it of interest to many biological disciplines (Armingol et al. 2021).

CCC is often defined as the sum of information a cell receives through its cellular receptors and sends through its secreted ligands. However, CCC can be broadened to include any biochemical signals a cell is exchanging with its environment (Armingol et al. 2021; Almet et al. 2021). For the purposes of this paper, we refer to CCC as any protein-protein mediated intercellular event a cell uses to communicate.

The direct assessment of CCC is complex, expensive, and difficult to standardize. Typical CCC validation techniques require specialized assays and imaging techniques that need to prove protein abundance and interaction partner colocalization (Armingol et al. 2021). Ideally CCC needs to be analyzed through *in vitro* and *in vivo* model systems that modulate a given interaction (using known inducer and inhibitor molecules or gene knockouts) and can then measure a direct impact on cellular functions (Armingol et al. 2021). Traditional *in vitro* methods can only assess a few cell types and a few proteins at a time, raising the question if they can do justice to the complexities of CCC (Almet et al. 2021). As such, not even large-scale studies such as Ramilowski et al. 2015 have achieved the type of depth that would allow one to measure the sum of all CCC in a given cellular context. Yet it is this previously unattainable breadth of analysis that makes up the allure and promise of CCC-Inference using single-cell transcriptomic ("scRNA") data.

Single-cell transcriptomic analyses were first published over a decade ago (Tang et al. 2009) and have significantly developed since then, expanding into other experimental setups, such as spatial transcriptomics (Longo et al. 2021) and other -omics types, such as genomics, epigenomics and proteomics (Lee et al. 2020). As the name indicates, scRNA protocols enable the measurement of gene-transcripts in individual cells (and cell types), and at a much greater depth and granularity than was previously

possible (Stark et al. 2019, Almet et al. 2021). As scRNA data becomes cheaper and more available, it is not unusual to have measured thousands of cells over tens of thousands of genes within a given tissue as in this example (Hao et al. 2021).

However, estimating CCC from scRNA has key weaknesses. For example, it uses gene transcription as a proxy for protein activity and consequently usually limits CCC to protein-protein interactions. This assumption bypasses many logistical factors and regulatory mechanisms in signaling, such as post-translational modification, the assembly of complexes and even the spatial distribution of cells (Armingol et al. 2021; Almet et al. 2021; Liu et al. 2016). CCC-Inference usually only involves binary CCC interactions and as such doesn't assess interactions between multiple cell types, which are no less biologically relevant (Cabello-Aguilar et al. 2020; Dimitrov et al. 2021). In addition, they often cannot properly assess signaling complexes (Armingol et al. 2021; Almet et al. 2021). Furthermore, CCC-Inference from scRNA-Seq is in general fundamentally limited to CCC events known to exist, as most inference approaches rely on a prior knowledge resource (Armingol et al. 2021; Almet et al. 2021).

Thus CCC-Inference should not be considered with the same level of certainty as direct CCC measurement. As the name suggests, it is better to consider these scRNA based predictions to be inferences that can consequently be experimentally validated or disproven (Almet et al. 2021; Armingol et al. 2021). As such, CCC-Inference functions as an exploratory tool that helps prioritize interactions to experimentally validate.

1.1 The Basis of CCC-Inference

There are many interpretations of CCC-Inference, but a few core elements are shared by most approaches. In order to work, CCC-Inference requires a resource, a clustered data set, and a method.

A resource is a list of genes curated through prior knowledge. Each entry is a binary protein-protein interaction that has been collected and curated through some means defined by the creator of the resource. As they are based on literature, interactions are biased in their distribution and potentially even curation standards (Dimitrov et al. 2021), nonetheless, they are essential to the process. This curated knowledge is in

principle what lends validity to the analysis of a gene pair; it is a gene pair previously known to be relevant in at least one context.

The context of a CCC event is brought in through single cell data. At minimum, this consists of scRNA data, but may contain further information such as spatial (Longo et al. 2021) or even proteomics data (Stoeckius et al. 2017). Using scRNA, the protein activity of ligands, receptors, and sometimes interaction modulators is estimated from gene transcription data. In addition, the scRNA can be used to cluster cells according to their expression profiles, which can be used as a proxy for cell type clusters in the data. This enables the estimation of cell-type specific communication and can help identify clusters that are particularly communicative.

Finally, the method one uses is at its core a scoring function that evaluates whether a communicative context exists. This analysis is performed in a granular fashion, for every interaction in the resource, and for every possible two-cluster pair in the data, a Ligand and Receptor ("LR") Score is calculated. The scoring function is thus dependent on four variables, the ligand expression, the receptor expression, the source cluster (which expresses the ligand) and the target cluster (which expresses the receptor). As such, in the most general way the methods used in this work can be expressed as:

$$LR\ Score\ (Ligand, Receptor, Source, Target)\ =$$
$$f(Ligand\ expression\ in\ Source, Receptor\ expression\ in\ Target)$$

$$LR\ Score\colon A\ continuous\ or\ binary\ score\ indicating\ communication.$$
$$Source\colon Source\ cluster\ expressing\ Ligand.$$
$$Target\colon Target\ cluster\ expressing\ Receptor.$$
$$f()\colon The\ scoring\ function\ unique\ to\ the\ method.$$

It is important to highlight this commonality in the scoring function. Since every interaction is seen as occurring between two resource-based genes and two data-driven clusters, these two major dependencies limit the accuracy of CCC prediction. An inference can only be as accurate as the underlying data clustering (Raredon et al. 2021) and resource (Dimitrov et al. 2021).

1.2 Benchmarking CCC-Inference Methods

In the years since the inception of scRNA methods, a variety of tools have been developed to estimate CCC. Unfortunately, the ground truth of CCC is usually unknown, making comparisons of these methods extremely difficult (Almet et al. 2021; Armingol et al. 2021). This is further complicated by a lack of standardization in the approaches to CCC inferences and the prior knowledge they operate on. Ultimately, without a ground truth, methods can only be compared in indirect experiments aimed at assessing individual method qualities, such as the analyses in Dimitrov et al. 2021.

In the following we analyzed the robustness of 6 CCC-Inference methods in regard to four common sources of noise in scRNA experiment. We modelled unstable cell type clustering, experiments with fewer samples, laxly curated resources, and the effects of switching resources in an experiment. In this process, we sought to emulate good benchmarking practice (Weber et al. 2019). This analysis was performed as a part of a larger comparison of these 6 methods (Dimitrov et al. 2021).

Of the CCC-Inference methods that have been developed, there are ones that aim to predict CCC Interactions, and ones that go beyond this and estimate CCC pathway activities within a sample (Dimitrov et al. 2021). We constrained ourselves to direct interaction predictions. Specifically, we focused on Connectome (Raredon et al. 2021), NATMI (Hou et al. 2020), a LIANA++ based log fold change (“LogFC”) method inspired by iTALK (Dimitrov et al. 2021; Wang et al. 2019), SingleCellSignalR (Cabello-Aguilar et al. 2020), CellPhoneDB (Efremova et al. 2020), and CellChat (Jin et al. 2021). The inner workings of these methods are explained in detail in Supplementary Note 1.

2. Methods

2.1 Data Availability

The PBMC data used in this analysis is publicly available for download from the 10x genomics website and has been used in multiple scRNA benchmarking analyses (https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz). Before being used, the raw data was processed according to the Seurat guided clustering tutorial (https://satijalab.org/seurat/articles/pbmc3k_tutorial.html). As such it went through quality control, normalization, principal component analysis, variable feature detection, and cell type clustering, among other steps.

2.2 Code Availability

The code used in this analysis is publicly available on GitHub (https://github.com/saezlab/ligrec_robustness).

2.3 Packages and Tools

The code for this analysis was written in R (R Core Team 2021) using the tidyverse package (Wickham et al. 2019), in conjunction with Python Version 3.8.5. More information on Python can be found at https://www.python.org/. We used Seurat 4.0.4 (Hao et al. 2021) to handle the scRNA data. The LIANA++ framework (Dimitrov et al. 2021) was used to run the original implementations of the methods. OmniPath (Türei et al. 2021) was used as the CCC resource in all analyses.

The methods used within LIANA++ were Connectome (Raredon et al. 2021), NATMI (Hou et al. 2020), SingleCellSignalR (Cabello-Aguilar et al. 2020), CellChat (Jin et al. 2021), CellPhoneDBv2 (Efremova et al. 2020) (implemented in Squidpy (Palla et al. 2021) and a custom LIANA++ method that was originally inspired by the basics of iTALK (Wang et al. 2019). Whenever we used these methods, we used the author-recommended settings for single-tissue analysis. In the case of CellChat, we also applied an additional 10 % expression proportion threshold.

2.4 Assessing CCC-Inference Method Robustness

We assessed the robustness of the above methods by systematically creating a set of baseline CCC predictions for each method using the standard PBMC data and OmniPath resource. For each method, we isolated its highest ranked predictions. We then manipulated the experimental setup in targeted ways and compared the overlap between the highly ranked baseline predictions and the highly ranked predictions under modified circumstances.

Overlap, is a linear metric, but also a directional one:

$$O_B^{M,B} = \frac{|B \cap M|}{|B|};\ O_M^{M,B} = \frac{|B \cap M|}{|M|}$$

$$O\text{: } Overlap\ between\ two\ sets\ of\ predictions. The\ proportion\ of\ one\ set\ that\ is\ in\ the\ other. The\ superscript\ refers\ to\ both\ sets\ being\ compared, the\ subscript\ to\ the\ norming\ factor.$$

$$B\text{: } Baseline\ predictions$$

$$M\text{: } Predictions\ under\ modified\ circumstances$$

$$|x|\text{: } Number\ of\ items\ in\ the\ set\ x.$$

It is dependent on the number of items in one of the sets as a scaling factor. Whenever our two sets of predictions varied dramatically, we used the larger set to scale the overlap, as this mimics the overlap one would see when downsampling the larger set to the size of the smaller (Supplementary Note 2).

When deciding if a given interaction was highly ranked, we considered the top 500 inferred CCCs significant, to maintain a consistent selection threshold. In general, when tied interactions at the border of significance occurred, we opted to consider them significant, which sometimes made more than the top 500 items significant. In a supplementary analysis, we repeated some of our robustness tests while considering the top 1000 and top 1500 ranked interactions significant (Supplementary Note 3).

2.5 Cluster Subsetting

A baseline of CCC predictions was produced as above and compared to predictions made from a subset of the original data set. To do this, the number of cells in the entire data set were reduced, but the proportion of cells per cluster remained constant (and the number of clusters never changed). We compared the overlap of predictions for multiple stages of subsetting, removing from 5 % to 60 % of cells in 5 % intervals. Since there is randomness in the way clusters are subset which might impact results, we performed this analysis over 10 iterations. For each iteration, we used a separate seed to perform subsetting, ensuring that the subsetting was different on each repetition. We then collated the results across these iterations. We used the same iterative approach for all our robustness tests.

2.6 Cluster Reshuffling

In order to assess the robustness impacts of partially reshuffled cluster annotations, we created a baseline of CCC predictions using the default PBMC data and then compared their overlap to CCC predictions produced with the reshuffled data. For each comparison, we took a certain proportion of all of the cells in the PBMC data set. Then we systematically replaced each cell's cluster annotation with a random sample drawn from all the cluster annotations present in the original data set. The one restriction was that the cluster annotation had to change; if a cell was originally assigned to cluster A, it could not be reassigned to cluster A again.

In this manner we reshuffled the cluster annotations of part of the data set but made sure that the degree of reshuffling and degree of cluster annotation mismatch were the same. In addition, on average, cell type distribution remained the same. It is only the underlying gene expression within the new clusters that is affected by cells moving from one cluster into another. We compared baseline predictions with predictions based on reshuffled cluster annotations (5 % intervals from 5 % to 60 %). Since this process is affected by chance, we repeated this robustness test 10 times and collated the results.

2.7 Indiscriminate Resource Dilution

We compared baseline CCC predictions to predictions generated when OmniPath was systematically diluted with non-canonical interactions.

To create non-canonical interactions, we drew on the PBMC data set. We randomly paired genes from the top 2000 most variable features, making sure that no gene ever paired with itself, that there were no duplicate interactions, and that a gene marked as a receptor in one interaction was never a ligand in another interaction (and vice versa). This ensured no contradictions to the general topological rules that interactions in OmniPath follow. In a supplementary analysis, we also investigated the overlap when the topology of OmniPath interactions were mimicked more closely (Supplementary Note 4). We also repeated the analysis using generic genes from the data set to construct non-canonical interactions, rather than the highly variable genes (Supplementary Note 5).

To achieve only partial dilution of OmniPath, we filtered it to contain only interactions relevant to the data set (both genes in a given interaction must have been measured). Since all our non-canonical interactions were relevant to the data, we made sure the interactions they replaced were too. We then indiscriminately removed a given percentage of OmniPath interactions and replaced them with non-canonical ones.

Predictions from resources diluted in 5 % intervals from 5 % to 45 % were compared to the baseline. Since there is randomness to how resources are diluted, this analysis was repeated 10 times and the results collated.

2.8 Discriminant Resource Dilution

Our final robustness test followed the same procedure as above, except that during dilution we ensured that no significant interactions in the baseline predictions of any method were removed from the resource. This ensured that the focus of the robustness test was not on the rate of losing baseline predictions, but the rate of introducing non-canonical predictions.

2.9 Topology Analysis

In supplementary analysis, we investigated the degreeness of the CCC predictions by method (Supplementary Note 6).

3. Results

3.1 Cluster Subsetting

This test modelled the impact of profiling fewer cells in the scRNA data on CCC inference. To do so, we subset our original data set while preserving the relative cluster sizes. We then compared the overlap of CCC predictions with the entire data set to predictions made with the subset data (Figure 1).

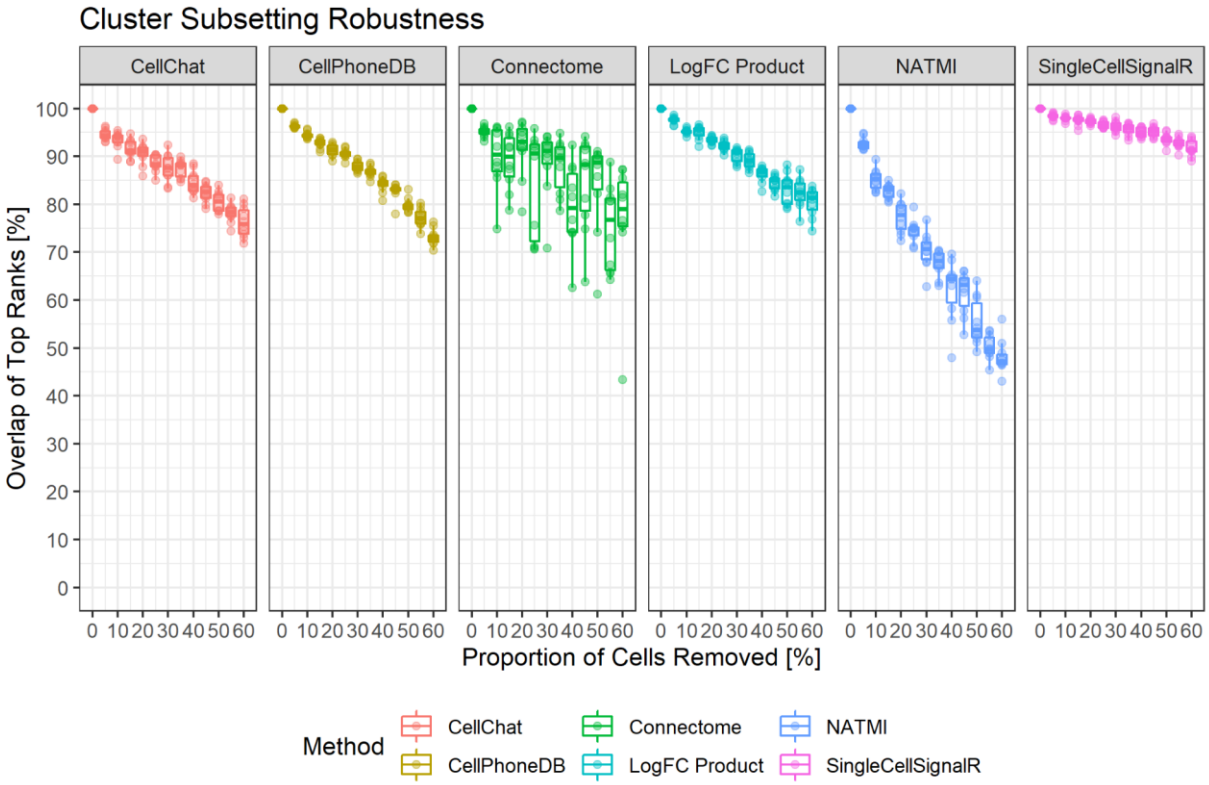

Figure 1. Cluster Subsetting Robustness. The PBMC data was subset in 5 % intervals. We assessed the overlap between predictions based on the original data and the subset data.

SingleCellSingalR proved particularly robust, showing a fairly linear relationship between the percentage of cells per cluster removed and the baseline-to-modified prediction overlap. Even at 60 % subsetting, it's mean robustness was over the 90 % mark.

Both permutation-based methods (CellChat, CellPhoneDB) showed practically identical trends in robustness, and showed a comparatively middling robustness compared to the rest of the field. Their robustness was comparable to the LogFC and

Connectome methods, all of which declined fairly linearly to 80 % robustness when 60 % of cells were removed. Unlike all the other methods, Connectome's distribution of overlaps was rather broad. While the highest and lowest values for overlap at a given level of subsetting ranged around 10 percentage points for most methods, Connectome's overlap reached over 40 percentage points in disparity at 60 % cell removal, the highest recorded overlap being close a little under 90 %, and the lowest recorded overlap being a little above 40 %.

Finally, NATMI showed very little robustness compared to the other methods. It's steep decline in prediction overlap left it at roughly 50 % robustness at 60 % subsetting.

3.2 Cluster Reshuffling

This test modelled the impact of unstable clustering on CCC predictions. To do so, we reshuffled the cluster annotations of our data set in varying proportions. This caused gradual changes in the cluster composition of the data set. We then compared the overlap of CCC predictions with the original data set and the reshuffled data set (Figure 2).

Connectome shows the highest robustness and continues to show a much larger spread in its values than the other methods, reaching an almost 30 percentage point spread at times. This spread in distribution appears to grow with the degree of cluster reshuffling. At its worst, the Connectome robustness drops to a mean at around 70 % overlap.

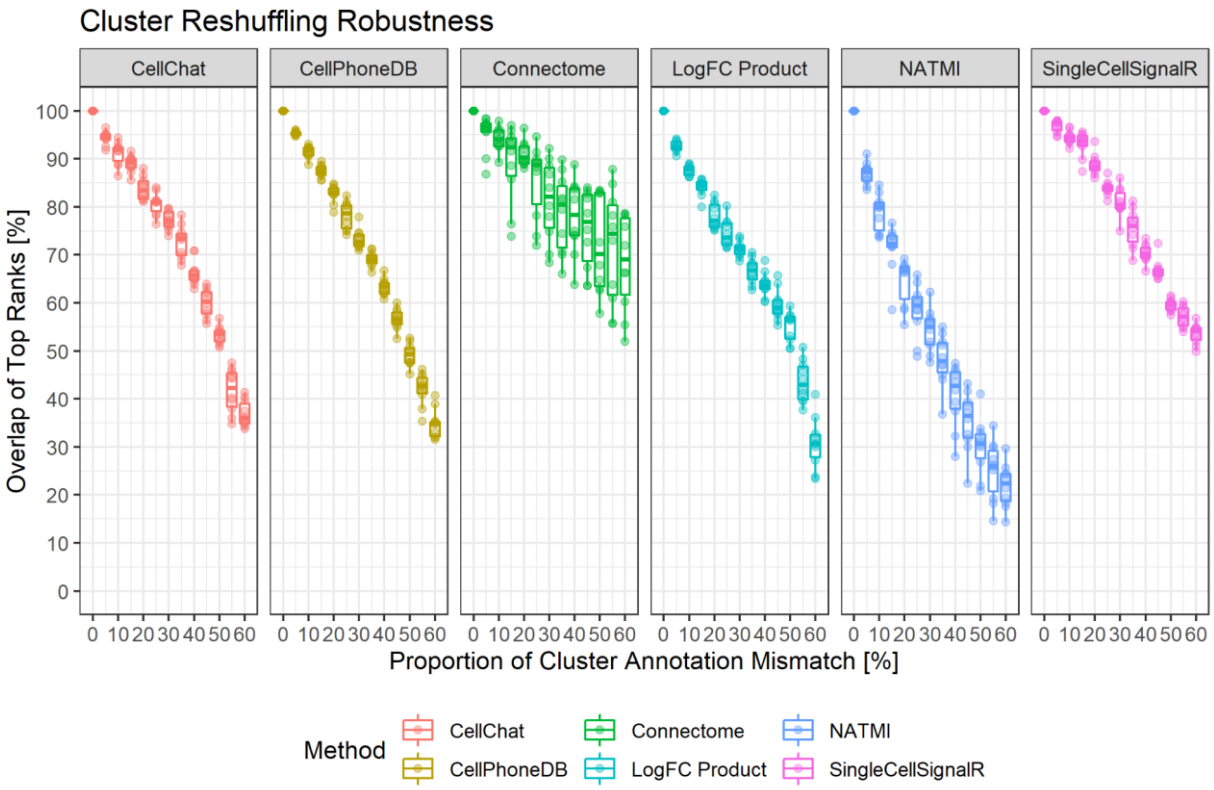

Figure 2. Cluster Reshuffling Robustness. The cluster annotations of the PBMC data were reshuffled in 5 % intervals. We assessed the overlap between predictions based on the original data and the reshuffled data.

SingleCellSignalR robustness worsens over 30 percentage points when compared to its subsetting performance, and no longer appears to have a linear relationship to the manipulation. Instead, it shows a sigmoidal decrease in robustness the more clusters are reshuffled, landing at a final 50 - 60 % robustness at 60 % cluster reshuffling. Despite this steep drop when compared to the subsetting, it still is more robust than most other methods.

Much like the other methods, the permutation-based methods (CellChat, CellPhoneDB) drop precipitously in robustness. Both show practically the same robustness trend, and dip to roughly 40 % robustness at 60 % cluster reshuffling. This puts their robustness right in the middle again, when compared to the rest of the field.

The LogFC method declines non-linearly with the degree of cluster reshuffling, falling steeply at first, plateauing somewhat at 40 % reshuffling, before dropping off steeply past this point. Ultimately, it shows 80 % robustness after 20 % reshuffling, a little over

60 % robustness after 40 % reshuffling, and roughly 30 % robustness after 80 % reshuffling, putting it just ahead of NATMI, and less robust than all other methods.

Much as in the subsetting results, NATMI was the least robust method of the field. It drops steeply, reaching less than 70 % mean robustness at 20 % reshuffling, slightly over 40 % mean robustness at 40 % reshuffling and slightly over 20 % robustness at 60 % reshuffling, the worst mean robustness in all the data-based robustness tests.

3.3 Indiscriminate Resource Dilution

This test modelled the impact of using two resources with a variable percentage of mismatch between one another. It made no effort to make sure that the interactions predicted in the baseline were protected from dilution (Figure 3).

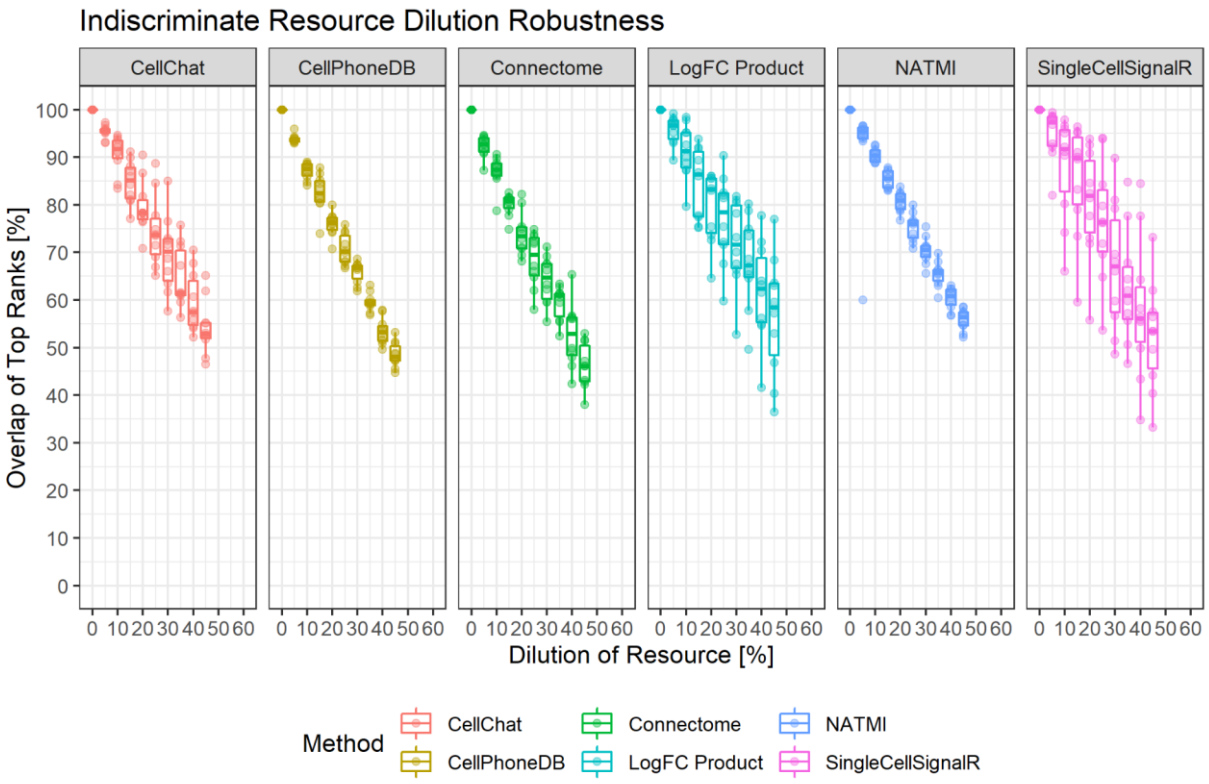

Figure 3. Indiscriminate Resource Dilution Robustness. In 5 % intervals, portions of the resource were removed and replaced with spurious interactions. We then measured the overlap between predictions based on the original resource, and the diluted resource.

Overall, all methods perform practically identically, showing a linear inverse trend with the proportion of the resource that is removed. At 20 % dilution roughly 80 % robustness is found in all methods, at 40 % dilution roughly 60 % robustness is found in all methods. At their worst, most methods showed around 55 % robustness, for 45 % percent dilution. Overall, LogFC, NATMI and Connectome performed a little better than 45 %, and Connectome performed a little worse.

The only noticeable trend, which holds somewhat for the other resource-based robustness test, is that SingleCellSignalR and the LogFC method show a broad spread in their distribution, with up to 40 percentage points of difference between the maximum and minimum value measured for a given dilution proportion. Connectome and CellChat also show a slight spread, reaching an about 25 percentage point disparity between their highest and lowest measured values.

3.4 Discriminate Resource Dilution

This test modelled the impact of having poorly curated interactions in your resource. Generally, it had less impact on the robustness than cluster reshuffling, but more than cluster dilution, and contained unique trends per method that differed from the data-based robustness tests (Figure 4).

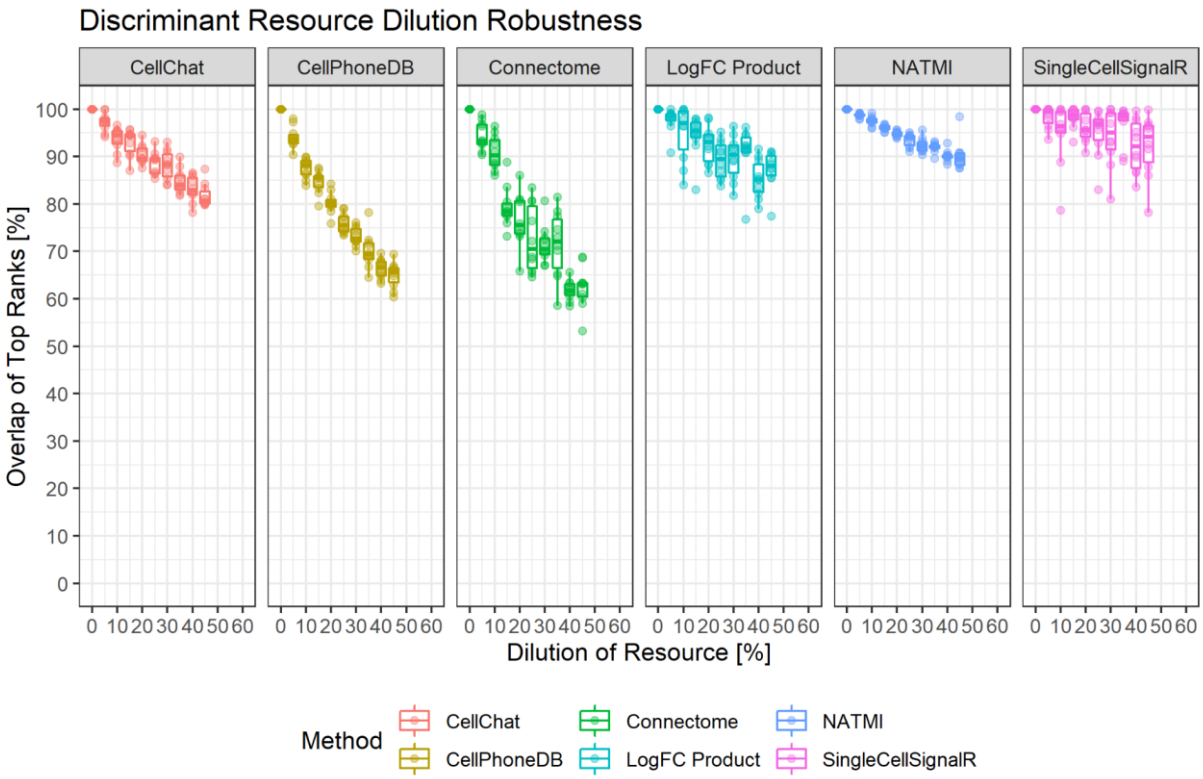

Figure 4. Discriminant Resource Dilution Robustness. We generated a baseline of predictions with each method. Then, in 5 % intervals, portions of the resource were removed and replaced with spurious interactions. In this process, none of the interactions in any of the baseline predictions were altered. We then measured the overlap between predictions based on the original resource, and the diluted resource.

SingleCellSingalR, NATMI and LogFC all performed comparably, being more robust than the other three methods. Overall, they show very little mismatch between baseline and modified predictions. All three methods ended at around 90 % robustness after 45 % resource dilution, with SingleCellSignalR coming in a little more robust than its peers, and LogFC a coming in a little less robust. Both LogFC and SingleCellSignalR show a much wider and sporadic distribution than a few of the other methods, reaching about 20 percentage point disparities between the highest and lowest measured robustness at given dilution percentages, where most other methods are in the 5 - 10 percentage point range.

Unlike the data-based robustness tests, the permutation-based methods don’t have the same trend in robustness. In this case CellChat shows a higher robustness than CellPhoneDB. It trends downward evenly with rising resource dilution, reaching 80 %

robustness at 45 % resource dilution. Accordingly, it is below average in relation to the other methods, but not in the lowest tier of robustness.

CellPhoneDB and Connectome show similar robustness. Connectome maintains it’s characteristically wide distribution of robustness values, though it is not as extreme a disparity as found in other robustness tests. Both methods decline down to just above 60 % at 45 % resource dilution.

4. Discussion

4.1 Overview

We studied the robustness of CCC inference methods in relation to common disruptive factors. Our first robustness test analyzed the robustness of CCC methods in relation to two sources of data noise. Once under the assumption that cluster subsetting mimics experiments with fewer cell samples, and once under the assumption that reshuffling cell cluster annotations mimics inaccurate clustering of the data. Unstable clustering can occur for a variety of reasons, and a strong argument can be made that the underlying biology of cells can't be evenly divided (Kiselev et al. 2019). As such, its impact on CCC prediction remains a relevant factor to observe. In addition, the cost of single cell analysis makes the number of cells profiled an important parameter to keep in mind for budgetary reasons, as such its impact on robustness are of immediate interest to many CCC inference experiments (Schmid et al. 2020).

Overall, cluster subsetting did not affect CCC prediction robustness as severely as other manipulations we performed. Methods that valued cluster specificity less tended towards higher robustness than more cluster specific ones, and the permutation-based methods had practically identical robustness. Cluster reshuffling had the largest impact of all manipulations. All methods except Connectome showed their lowest robustness over all tests. They generally follow the same trends as subsetting.

Our second robustness test analyzed the robustness of CCC-Inference in relation to two sources of faulty resources. We modeled switching between two resources in an experiment by indiscriminately replacing resource interactions with non-canonical interactions (indiscriminate dilution). We also modeled a poorly curated resource by repeating the above process but making sure that when modifying the resource significant interactions (interactions that CCC methods predicted as significant in ordinary conditions) were preserved (discriminant dilution). The results in Dimitrov et al. 2021 show a dramatic impact on scRNA CCC predictions based on the resource that is used, as well as highlighting the different curation standards and internal biases of resources. As such, the impact of resource choice on the analysis is an important influence to be aware of. Additionally, the question of which standards are best for resource curation remain unresolved.

As was to be expected, indiscriminate resource dilution (dilution that did not necessarily preserve previously significant interactions) had a uniform impact on predictions, linearly driving robustness downward equal to the proportion of the resource that had been replaced. Discriminant resource dilution had a lesser but more varied impact on robustness.

4.2 LR Magnitude vs. LR Specificity

The CCC-Inference Methods we analyzed fell into two camps, those that incorporate magnitude of LR expression, and those that additionally incorporate cluster specificity of LR expression. Our data-based robustness results indicate cluster specificity and its implementation as a relevant factor that differentiates method performance, a result other benchmarks have also suggested (Dimitrov et al. 2021). Thus, we elaborate on the difference between these two approaches here.

The main aim of the CCC-inference scoring functions is to quantify a context in which communication is likely. In general, all methods assess the magnitude of LR expression, that is, they assume that higher ligand and receptor expression indicates a higher likelihood that an interaction is occurring. It is obvious that expression is a prerequisite for communication, but some methods also assess how cluster specific an interaction is, i.e. how differently a gene is expressed in its current cluster in relation to all others. These methods would score an interaction highly if its ligand and receptor were expressed more (magnitude), and/or differently in comparison to their baseline expression in other clusters (specificity).

Of the methods we use, five use specificity as a measure of biological relevance. They argue for specificity on the basis of intuition and the notion of avoiding housekeeping communication (Hou et al. 2020) or identifying rare and highly communicative cells (Raredon et al. 2021), but they do not provide any direct evidence or sources relating to this claim. In fact, this claim still remains unconfirmed (Dimitrov et al. 2021). As such, without any evidence to support there is no way to evaluate whether specificity is helpful, harmful, or has different effects altogether.

It should also be noted that interactions that are not binary between two proteins and two cells are actively punished by specificity metric. As Cabello-Aguilar et al. 2020 and Dimitrov et al. 2021 suggest, there are multiple important signaling contexts that are not binary and would score poorly given a specificity metric, despite being of high biological relevance.

4.3 Data-Based Robustness Tests

Every method relies on mean gene expression per cluster at its core, and in the data-based robustness tests, we see this having a decisive impact. In cluster subsetting, the mean cluster expression per cluster doesn't change much, provided that the data contains stable cell clusters. As such, robustness is high for most methods. In cluster reshuffling, the mean gene expression per cluster changes drastically, and we see this have a much larger impact on CCC prediction, and hence observe lower robustness. Overall, it should be considered that this represents the best-case scenario for cluster subsetting, in which the quality of the clustering is not impacted. In a real experiment that has a low number of samples, or uses inappropriately high cluster granularity, there is a risk of unstable clustering, which as the cluster reshuffling and other literature indicates, is likely to be quite impactful (Raredon et al. 2021; Dimitrov et al. 2021).

In our setting, SingleCellSignalR was most robust for cluster subsetting, showing a 40-percentage point lead over the least robust method in the comparison, which is NATMI. This likely has to do with the fact that SingleCellSignalR depends primarily on the pooled mean expression of all genes, the mean gene expression of the ligand in the source cluster, and the receptor in the target cluster. When subsetting a well-clustered data set, neither of these cluster-specific means should change much.

The other methods all consider cluster-specificity when assessing an interaction, and as such bring in statistics that are meant to reflect the entirety of a given gene's expression to compare to its expression in the given cluster. Connectome uses the mean expression of the given gene in all samples as a reference, LogFC uses the mean expression of the given gene in all cells outside the given cluster as a reference, and NATMI uses the sum of the gene's mean expression in every cluster as a reference.

All of these metrics allow for the compounding of errors from subsetting. While subsetting is unlikely to change a single cluster mean drastically, NATMI compounds this difference over every cluster in the data. While SingleCellSignalR is dependent on two cluster means and the pooled mean to operate, NATMI considers all clusters in the data twice. As such it seems plausible that LogFC and Connectome would rank a little lower than SingleCellSignalR, as their gene-wide means are little more sensitive to noise than any of SingleCellSignalR's metrics, and NATMI would rank even lower, as it is dependent on multiple smaller means, compounding errors between them.

Similar behavior can be seen in the permutation-based methods. Both perform almost identically, indicating that the permutation approach they share is likely more important than the differences in their scoring functions. Both use the permutation approach to account for cluster specificity (Supp Note 7), which appears to be the relevant difference between the two groups of scoring methods. As such, it would make sense for it to be the driving factor in their robustness. It is not unreasonable to think that the empirical distribution generated in the permutation approach is prone to more inaccuracies when it relies on a less representative sample. Inaccuracies in the main LR score could then be compounded by these inaccuracies in the specificity.

As mentioned, cluster reshuffling had a much greater impact, but overall followed similar trends to subsetting. The main difference to the cluster subsetting setup is that the mean gene expression per cluster changes dramatically, and that the mean gene-wide expression doesn't change. Thus, Connectome shows the highest robustness among all methods. It is still drastically impacted by the changing cluster means, but the global mean in its z-transformation is constant, as it is agnostic to the cluster annotations. This difference likely separates it from all the other cluster specific methods, which make up the bottom of the field.

SingleCellSignalR drops dramatically in robustness compared to subsetting but is still more robust than most of the field. This large change is likely due to the drifting cluster means, which it is entirely dependent on. Still, it doesn't incorporate any other parameters that change when the cluster distribution changes, and thus manages a comparatively high robustness.

Unlike Connectome, the rest of the cluster specificity-based methods do not use approaches that are agnostic to the cluster annotation of the cells. As such, these measures change when clusters are partially reshuffled. What is particularly problematic for these approaches, is that if a gene is cluster specific in normal conditions, and its expression gets more diffused among other clusters through reshuffling, then the signal is slowly introduced into multiple clusters, making it less cluster specific, and less likely to score well. In fact, many of the following approaches return less and less significant interactions as the degree of reshuffling goes up.

In the case of the permutation-based approaches (which performed similarly again), the expression of an interaction is measured against an empirical baseline to determine its significance. As reshuffling occurs, the signal is incorporated more and more into the baseline, and it becomes harder and harder for an interaction to be significant.

Similarly, LogFC uses the mean gene expression of cells outside a given cluster to measure cluster specificity, and NATMI uses the sum of all other cluster means. Clearly, these measures are impacted when the cluster annotation changes. In NATMI we likely see the compounding errors from drifting cluster annotation again, and in LogFC we likely see the impact of multiplying and dividing noisy parameters - at a certain point its robustness drops precipitously. Both of these methods showed the lowest robustness in this test.

Our analysis of the topology of method predictions (Supplementary Note 6) partially supports these subsetting and reshuffling results. It suggests that NATMI is the most cluster specific of all methods, as it has the highest degreeness in its predictions, making it the least likely to consider a given gene-pair significant between multiple cluster-pairs. This fits our observation that NATMI is most impacted by reshuffling and subsetting.

In summary, it appears likely that the cluster specificity of a method, and the implementation of the cluster specificity within a method are the deciding factors in their robustness to cluster based manipulations. This is intuitive, the more information about the distribution of cells in clusters is brought into a scoring function, the more it must rely on accurate clustering. It is also in line with another benchmark’s findings

(Dimitrov et al. 2021). Accordingly, the one cluster-non-specific method, SingleCellSignalR has the highest robustness in our field, and NATMI, which includes the most clustering information in its score, has the lowest robustness across the board.

4.4 Resource-Based Robustness Tests

The resource-based robustness tests were not as straightforward to evaluate. Overall, indiscriminate dilution showed a predictable, strong, and uniform impact. Discriminant dilution showed a more varied impact but was in parts unexpected and perplexing.

Robustness to indiscriminate resource dilution mostly shows the base rate of interaction replacement taking place. It is intuitive that if two resources are different in each percentage of their interactions, the prediction derived from them (using the same method each time) will differ by the same proportion. We find this linear trend in our robustness results as well. For example, when 40 % of a resource is replaced, we observe a 40 % mismatch between the baseline and original prediction. This effect appears uniform across all methods, which makes sense. If an interaction is not in the resource, it cannot be predicted, no matter the scoring function. Overall, this paints an unfortunate picture, the choice of resource likely has an unavoidable and severe impact on the results that are returned, and the more different the chosen resource is from the rest of the field, the more impactful the choice will be. This is also suggested in other analyses on the impact of resources in this context (Dimitrov et al. 2021).

Discriminant dilution showed far more varied results. By preserving all significant interactions in the baseline, we only analyse the proportion of non-canonical positives that are introduced through the addition of non-canonical interactions to the resource. This is meant to mimic a strictly curated resource (which has no non-canonical interactions) being compared to a less strictly curated resource (which might have more spurious or poorly validated interactions). It should be noted that our analysis shows the worst-case scenario for this context, as every spurious added interaction is made up of highly variable genes in the data set. When we construct our spurious interactions using generic genes in the data set, the overall impact of robustness is negligible for all methods (Supplementary Note 5).

SingleCellSingalR has a high robustness to discriminant dilution, mostly owed again to its disregard of cluster specificity. To rank well, a gene must be expressed highly, and the highly variable genes we introduce have no guarantee of meeting this criterion.

Interestingly, NATMI and LogFC also show high robustness to this form of manipulation. Even though highly variable genes are more likely to be differentially expressed between clusters, and thus cluster specific, they do not appear to have a great impact on these methods' robustness. They perform similarly to SingleCellSignalR, which is interesting as it doesn't appear that there is much that unites these three approaches in philosophy or approach.

Counter to our experience in the data-based analysis, CellChat and CellPhoneDB perform quite differently, which indicates that some of the extra information CellChat accounts for helps it differentiate between spurious and non-spurious interactions. It seems likely that CellChat's method for accounting for interaction mediators makes it hard for spurious interactions to score well, as it is less likely for a spurious gene to have its inducers expressed highly and inhibitors expressed lowly. However, OmniPath likely does not contain enough interactions with annotated mediators to fully account for this effect. Potentially CellChat's more noise-resistant trimean makes it less likely to be skewed by the highly variable genes introduced. Regardless, CellChat performs close to the level of the other methods, while CellPhoneDB considerably lags behind the rest of the field.

Overall, resource dilution's impact on non-canonical positives is difficult to interpret and does not line up cleanly to elements of the method scoring functions or other processing steps. While it is possible that magnitude-based methods and methods that include upstream signaling context are more resistant to this manipulation, a few of the results remain hard to explain. Generally, the impact on robustness is low however, indicating that a larger, more laxly curated resource isn't hugely negative to the robustness of most methods. As such it may potentially be worth it to accept a slightly higher rate of false positives in exchange for the chance at a higher number of relevant interactions being analyzed, though further research would be required in this area. Finally, these results serve as another example of the difficulties in understanding how

these methods work and how seemingly small intuitive decisions in their design can have relevant consequences that are difficult to suss out and assess.

4.5 Conclusion

Many important characteristics of CCC Inferences remain unresolved. Without a gold standard to establish CCC ground truth there is no easy way to evaluate and develop methods and their approaches (Dimitrov et al. 2021; Armingol et al. 2021; Almet et al. 2021). As such, it is currently left to less direct analyses to identify the individual strengths and weaknesses of CCC-Inference methods. In this analysis, we set out to analyse four common sources of noise in a scRNA CCC-Inference experiment and identify the robustness to this noise that each of the analyzed methods have.

Generally, we found that these sources of noise can have a dramatic impact on the CCC predictions, but that not all manipulations we employed had a similar impact on robustness. When it comes to grading which influences are most detrimental to the robustness of an analysis, unstable clustering and the choice of resource are likely to have a large impact, while a laxly curated resource or a smaller sample number are likely to have a far lesser impact. Unstable clustering can be avoided in a variety of ways (Kiselev et al. 2019), but the choice of resource is unavoidable. Without a gold standard as far as resources go, there is no good way to resolve this problem. Our only recommendation to preserve robustness would be to avoid small resources with high dissimilarity from the rest of the field, as these results are unlikely to be reproducible with other resources.

The methods we tested also had varied levels of robustness between them. Generally, our results suggest that methods that consider the cluster specificity of an interaction are more dependent on stable clustering and a high number of samples. In fact, the approach that methods take toward integrating cluster specificity into their scores, and the degree to which they value cluster specificity, appears to be the main factor in differentiating their robustness, with low-specificity methods generally being more robust than high-specificity methods. However, our results also showed trends that are not easily explained by the anatomy of their scoring function or processing steps, highlighting the gaps in our knowledge regarding their behavior. Hence, this suggests

that even small intuition-based design choices when making these methods can have unintended consequences that can be difficult to assess.

5. Outlook

Secondary analyses of CCC-Inference methods, such as this one and Dimitrov et al. 2021 can help us understand how methods work, and how they should be designed. Over time, it will likely become clearer what metrics of gene expression should be evaluated (e.g. magnitude, specificity, upstream regulation, etc.), and in what manner these metrics are best evaluated to accurately detect CCC. Even if these investigations cannot reveal a clearly superior method, then at minimum, they can indicate which methods and metrics work best in which context. In addition, there are many other approaches to CCC-Inference that break the mold of those we analyzed here, such as less binary approaches (scTensor by Tsuyuzaki et al. 2019) and approaches take into account much broader trends in transcriptomic data (NicheNet by Browaeys et al. 2020, CytoSig by Jiang et al. 2021). Ultimately, there is likely much potential for improvement in the design of CCC-Inference methods, and it is unlikely that any method tested here has reached the upper bound of predictive accuracy and potential that can be derived from scRNA data.

Of course, scRNA data is not the only source of information available. As approaches such as CITE-Seq (Stoeckius et al. 2017) coupled with CiteFuse (Kim et al. 2020) (a multi-omics integration approach) show, there is ample opportunity to integrate new data types into CCC-Inference. For example, proteomics integration and/or spatial integration would help reduce the assumptions made when proxying protein activity from gene transcription. Similarly, metabolomics integration could help CCC-Inference move beyond purely protein-protein interactions. This intersection highlights where theoretical method design and experimental innovation can fruitfully intersect to increase the performance of CCC-Inference.

Finally, there can be no true golden standard amongst methods before there is a golden standard in benchmarking, which requires innovation in direct CCC measurement. For the moment, the biological ground truth of CCC remains largely unknown, which highlights the need for new experimental approaches that can help establish one. Of course, this is easier said than done, but perhaps simple model tissues or simple cultured tissues, which would be easier to analyze, could help bridge the gap. If there were at least a few datasets available for different tissue contexts in

which the sum of relevant CCC was known and experimentally validated, one would simply need to run their CCC inference tool on the data set and see how much of the known ground truth was recovered. It would be immediately apparent which methods were suited best for which contexts and data types. Obviously, an established ground truth for a given tissue would also be invaluable to the fields studying the tissue in question. Consequently, it is not just CCC-Inference that could make use of innovation in direct CCC measurement, but the entire field of CCC as a whole.

6. References

Almet, A.A., Cang, Z., Jin, S., and Nie, Q. (2021). The landscape of cell-cell communication through single-cell transcriptomics. Current Opinion in Systems Biology 26, 12–23.

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

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

Cabello-Aguilar, S., Alame, M., Kon-Sun-Tack, F., Fau, C., Lacroix, M., and Colinge, J. (2020). SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. Nucleic Acids Res. 48, e55.

Dimitrov, D., Türei, D., Boys, C., Nagai, J.S., Ramirez Flores, R.O., Kim, H., Szalai, B., Costa, I.G., Dugourd, A., Valdeolivas, A., et al. (2021). Comparison of Resources and Methods to infer Cell-Cell Communication from Single-cell RNA Data. BioRxiv.

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.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell 184, 3573-3587.e29.

Hou, R., Denisenko, E., Ong, H.T., Ramilowski, J.A., and Forrest, A.R.R. (2020). Predicting cell-to-cell communication networks using NATMI. Nat. Commun. 11, 5011.

Jiang, P., Zhang, Y., Ru, B., Yang, Y., Vu, T., Paul, R., Mirza, A., Altan-Bonnet, G., Liu, L., Ruppin, E., et al. (2021). Systematic investigation of cytokine signaling activity at the tissue and single-cell levels. Nat. Methods 18, 1181–1191.

Jin, S., Guerrero-Juarez, C.F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., Myung, P., Plikus, M.V., and Nie, Q. (2021). Inference and analysis of cell-cell communication using CellChat. Nat. Commun. 12, 1088.

Kim, H.J., Lin, Y., Geddes, T.A., Yang, J.Y.H., and Yang, P. (2020). CiteFuse enables multi-modal analysis of CITE-seq data. Bioinformatics 36, 4137–4143.

Kiselev, V.Y., Andrews, T.S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. Nat. Rev. Genet. 20, 273–282.

Lee, J., Hyeon, D.Y., and Hwang, D. (2020). Single-cell multiomics: technologies and data analysis methods. Exp. Mol. Med. 52, 1428–1442.

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell 165, 535–550.

Longo, S.K., Guo, M.G., Ji, A.L., and Khavari, P.A. (2021). Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nat. Rev. Genet. 22, 627–644.

Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. 15, e8746.

Palla, G., Spitzer, H., Klein, M., Fischer, D.S., Schaar, A.C., Kuemmerle, L.B., Rybakov, S., Ibarra, I.L., Holmberg, O., Virshup, I., et al. (2021). Squidpy: a scalable framework for spatial single cell analysis. BioRxiv.

Ramilowski, J.A., Goldberg, T., Harshbarger, J., Kloppmann, E., Lizio, M., Satagopam, V.P., Itoh, M., Kawaji, H., Carninci, P., Rost, B., et al. (2015). A draft network of ligand-receptor-mediated multicellular signalling in human. Nat. Commun. 6, 7866.

Raredon, M.S.B., Yang, J., Garritano, J., Wang, M., Kushnir, D., Schupp, J.C., Adams, T.S., Greaney, A.M., Leiby, K.L., Kaminski, N., et al. (2021). Connectome : computation and visualization of cell-cell signaling topologies in single-cell systems data. BioRxiv.

R Core Team (2021). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

Schmid, K.T., Cruceanu, C., Boettcher, A., Lickert, H., Binder, E.B., Theis, F.J., and Heinig, M. (2020). Design and power analysis for multi-sample single cell genomics experiments. BioRxiv.

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. Nat. Rev. Genet. 20, 631–656.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods 14, 865–868.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods 6, 377–382.

Tsuyuzaki, K., Ishii, M., and Nikaido, I. (2019). Uncovering hypergraphs of cell-cell interaction from single cell RNA-sequencing data. BioRxiv.

Türei, D., Valdeolivas, A., Gul, L., Palacio-Escat, N., Klein, M., Ivanova, O., Ölbei, M., Gábor, A., Theis, F., Módos, D., et al. (2021). Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Mol. Syst. Biol. 17, e9923.

Wang, Y., Wang, R., Zhang, S., Song, S., Jiang, C., Han, G., Wang, M., Ajani, J., Futreal, A., and Wang, L. (2019). iTALK: an R Package to Characterize and Illustrate Intercellular Communication. BioRxiv.

Weber, L.M., Saelens, W., Cannoodt, R., Soneson, C., Hapfelmeier, A., Gardner, P.P., Boulesteix, A.-L., Saeys, Y., and Robinson, M.D. (2019). Essential guidelines for computational method benchmarking. Genome Biol. 20, 125.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. JOSS 4, 1686.

7. Supplementary Notes

7.1 Supplementary Note 1 - Individual Method Details

In the following we will go over the details of the scoring functions and processing steps in each CCC inference method that we reviewed. A summary of this information can be found in Figure 5.

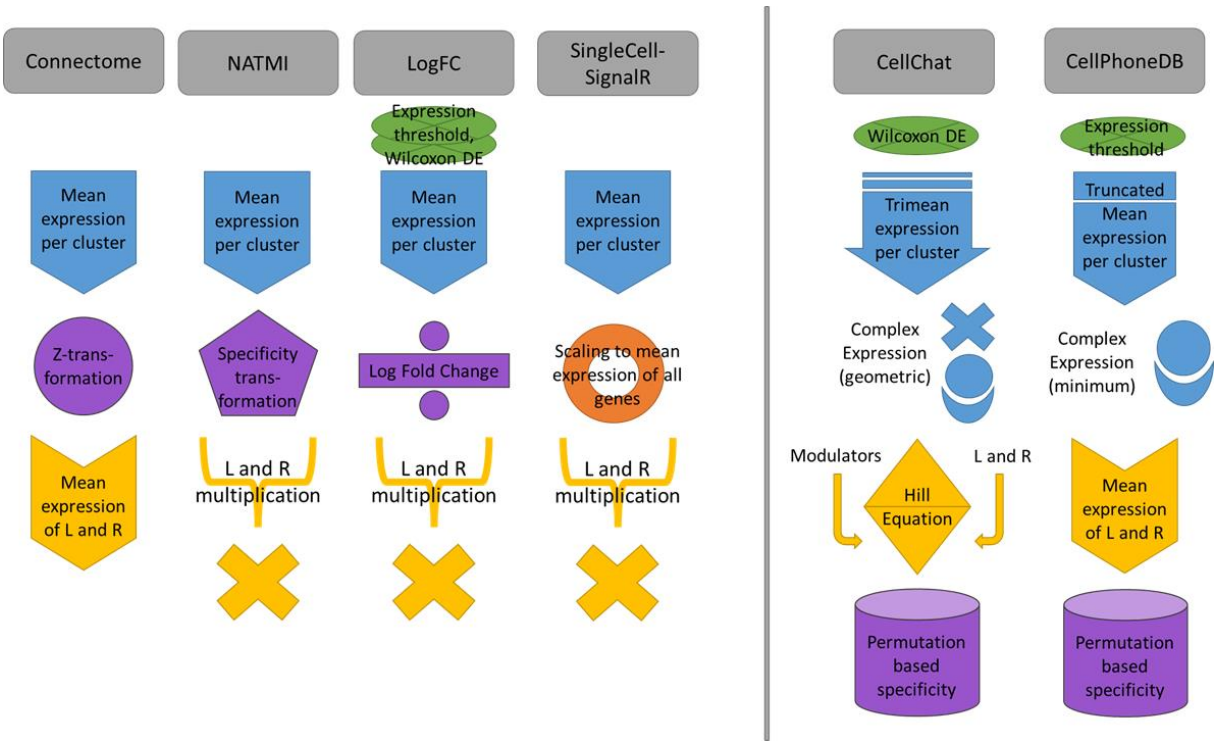

Figure 5: Anatomy of CCC-Inference Methods. This figure visually demonstrates the most relevant aspects of each method's CCC-Inference process. Every method's general approach can be read from top to bottom. Standard scoring methods are displayed on the left, permutation-based methods are displayed on the right. Steps that add cluster specificity are marked in purple. The core mathematical operator of each scoring function is marked in yellow. SingleCellSignalR's experiment wide scaling is in orange. Steps that filter out genes that don't meet certain prerequisites are marked in green. Miscellaneous processing steps are marked in blue. "L" stands for ligand expression, "R" stands for receptor expression. "Wilcoxon DE" stands for a Wilcoxon differential expression filter. "Expression threshold" refers to an expression proportion threshold. Whenever ligand expression and receptor expression are used for calculation, they have first been summarized by whatever cluster-wide mean is specified at the top of the method. The exact scoring functions and transformations are described below.

Connectome

Connectome has two different scoring functions, one to calculate W_{norm} and one to calculate W_{scale}. The W_{norm} scoring function is simply the mean of the ligand and receptor gene expression.

$$W_{norm}(L,R,S,T) = mean(mean(L_S), mean(R_T))$$

$L,R,S,T: Ligand, Receptor, Source\ Cluster, Target\ Cluster$
$L_S, R_T: L\ expression\ in\ S, R\ expression\ in\ T$

Accordingly, two highly expressed genes are seen as likely to be communicating.

However, W_{norm} is not recommended for single tissue analysis. Instead W_{scale} is recommended (Raredon et al. 2021). W_{scale} is specifically and introduce cluster specificity through the global mean of gene expression, intended to help identify small, important cell populations. To achieve this, the expression data undergoes gene-wide z-transformation. The final score is the mean of the mean z-scaled gene expression of the Ligand and Receptor in their respective clusters.

$$z = \frac{G - \underline{G}}{sd}$$

$z: The\ z-\ score\ of\ a\ given\ sample's\ gene\ expression\ of\ a\ given\ gene.$
$G: The\ gene\ expression\ of\ the\ sample.$
$\underline{G}: The\ mean\ expression\ of\ the\ gene\ across\ all\ samples.$
$sd: The\ standard\ deviation\ of\ the\ expression\ of\ the\ gene\ across\ all\ samples.$

$$W_{scale}(L,R,S,T) = mean(mean(z_{L,S}), mean(z_{R,T}))$$

$L,R,S,T: Ligand, Receptor, Source\ Cluster, Target\ Cluster$
$z_{L,S}, z_{R,T}: z-scored\ L\ expression\ in\ S, z-scored\ R\ expression\ in\ T$

In order to score well, an interaction's ligand and receptor must be highly expressed in relation to their baseline gene expression in all cells.

NATMI

NATMI calculates the mean gene expression of each gene per cluster and calculates two weights from it. Much like Connectome, the first, the mean expression edgeweight, gives no value to specificity and simply multiples mean Ligand expression with mean Receptor expression.

$$Mean\ Expression\ Edgeweight\ (L,R,S,T)\ =\ mean(L_S) * mean(R_T)$$

L,R,S,T: $Ligand, Receptor, Source\ Cluster, Target\ Cluster$

L_S , R_T : $L\ expression\ in\ S, R\ expression\ in\ T$

As in connectome, the authors suggest this weight is less informative, as it may emphasize housekeeping interactions and does not consider the interaction specificity (Hou et al. 2020). The second weight they propose aims to correct this, by transforming each mean gene expression per cluster into specificities per cluster. The mean gene expression in the given cluster is divided by the sum of its mean expression in every cluster. This measure values every cluster mean equally and integrates more information on the cluster distribution than the mean expression of a gene across all samples. In the end, specificities (and the specificity-based edgeweight) range between 0 (no expression in this cluster) and 1 (expression only in this cluster).

$$Sp_{L,S}\ =\ \frac{mean(L_S)}{\sum_i^n\ \ mean(L_i)}\ , Sp_{R,T}\ =\ \frac{mean(R_T)}{\sum_i^n\ \ mean(R_i)}$$

Sp: $The\ specificity\ of\ the\ expression\ of\ a\ gene\ in\ a\ cluster.$

$i\ to\ n$: $A\ list\ of\ all\ clusters\ in\ the\ data\ set.$

L,R,S,T: $Ligand, Receptor, Source\ Cluster, Target\ Cluster$

$$Specificity - Based\ Edgeweight\ (L,R,S,T)\ =\ Sp_{L,S} *\ Sp_{R,T}$$

L,R,S,T: $Ligand, Receptor, Source\ Cluster, Target\ Cluster$

$Sp_{L,S}$, $Sp_{R,T}$: $Specificity\ of\ L\ expression\ in\ S, specificity\ of\ R\ expression\ in\ T$

iTALK Inspired LogFC Method

This method was inspired by the basics of iTALK (Wang et al. 2019) and then expanded upon and developed in LIANA++. It first uses an expression threshold and Wilcoxon differential expression filter. The expression threshold ensures that both the ligand and receptor of any interaction evaluated must be expressed in at least 10 % of the cells in their own clusters. The Wilcoxon differential expression filter selects genes for their cluster specific differential expression. For every relevant gene, and for every two-cluster combination (including self-pairing) a Wilcoxon Rank Sum test is performed. Any gene that is not significantly differentially expressed (typically with threshold $p < 0.05$) in at least one two-cluster pair is discarded. As such, cluster specificity is brought into the analysis. Once filtering is complete, LogFC then calculates an LR score based on the fold-change between the mean gene expression in the given cluster vs all cells outside of that cluster.

$$LogFC\ (L,R,S,T)\ = log(\frac{mean(L_S)}{mean(L_{i \neq S})}) *\ log(\frac{mean(R_T)}{mean(R_{j \neq T})})$$

L,R,S,T: $Ligand, Receptor, Source\ Cluster, Target\ Cluster$

L_S , R_T : $L\ expression\ in\ S, R\ expression\ in\ T$

$i\ \neq\ S, j\ \neq\ T$: $All\ cells\ i\ that\ are\ not\ in\ S, all\ cells\ j\ that\ are\ not\ in\ T.$

This method integrates specificity directly into its score; in order to be ranked highly within their clusters both ligand and receptor must be expressed higher than their respective baseline expression outside their clusters. One unique aspect of this method is that relatively lowly expressed genes will have a negative log-fold change. If this is the case for both Ligand and Receptor, a specifically lowly expressed interaction will be ranked highly simply because it is specific. This violates the idea that expression is prerequisite to communication.

SingleCellSignalR

SingleCellSignalR does not include cluster specificity, and instead relies on the magnitude of expression of the ligand and receptor to score an interaction using the following equation.

$$LR\ Score\ (L,R,S,T)\ =\ \frac{\sqrt{mean(L_S)\ *\ mean(R_T)}}{\mu\ +\ \sqrt{mean(L_S)\ *\ mean(R_T)}}$$

L,R,S,T: $Ligand, Receptor, Source\ Cluster, Target\ Cluster$

$L_S\ , R_T\ :\ L\ expression\ in\ S, R\ expression\ in\ T$

μ: $The\ mean\ of\ the\ entire\ processed\ count-matrix.$

The μ here does not represent a gene-wide mean, it represents the pooled mean of all genes in the dataset. Hence, the $LR\ Score$ is a reference for how large the interaction's expression is in reference to this pooleed mean. If either the ligand or receptor is unexpressed, the score will be 0, it will be 0.5 when the interaction expression is at parity with the mean of all gene expression, and 1 if the interaction expression is infinitely larger than μ.

CellPhoneDB

CellPhoneDB is a permutation-based method which uses a 10 % expression threshold. This ensures that both the ligand and receptor of any interaction evaluated have to be expressed in at least 10 % of the cells in their own clusters. CellChat then calculates an LR score that quantifies the magnitude of expression.

$$LR\ Mean\ (L,R,S,T)\ = mean(CE(L_S)\ , CE(R_T))$$

L,R,S,T: $Ligand, Receptor, Source\ Cluster, Target\ Cluster$

$L_S\ , R_T\ :\ L\ expression\ in\ S, R\ expression\ in\ T$

CE: $Complex\ Expression\ of\ L_S\ and\ R_T.$

$$CE(G_{Clu}) = min(tr\_mean(S_{1,Clu}), tr\_mean(S_{2,Clu}), \ldots, tr\_mean(S_{n,Clu}))$$

G, S, Clu: $A\ given\ gene\ G\ that\ is\ a\ complex\ with\ subunits\ S_1\ to\ S_n\ , whose\ mean\ expression\ in\ cluster\ Clu\ is\ being\ evaluated. If\ G\ is\ not\ a\ complex, treat\ it\ as\ a\ complex\ with\ one\ subunit.$
tr_mean: $The\ truncated\ mean, ignoring\ the\ top\ and\ bottom\ 10\ \%\ of\ expression.$

Since the basis of CCC inference methods is the assumption of binary ligand and receptor relationships, they often struggle interpreting interactions involving complexes. CellPhoneDB is notable for resolving this issue. By using a truncated mean, the expression of a complex is estimated as the minimum expression of all its subunits.

The specificity of the $LR\ Mean$ is estimated using an empirical permutation-based approach that calculates a p-value.

$$p\ (L, R, S, T) = \frac{|\ reshuffled\ LR\ Means\ >\ original\ LR\ Mean|}{|reshuffled\ LR\ Means|}$$

p: $p - value$.
L, R, S, T: $Ligand, Receptor, Source\ Cluster, Target\ Cluster$
$|x|$: $The\ number\ of\ items\ in\ the\ set\ x$.

CellChat

CellChat uses a Wilcoxon differential expression filter, as LogFC does. CellChat differs from other methods in that it doesn't use a simple mean to analyze gene expression per cluster but uses a custom mean.

$$trimean(G_{Clu}) = \frac{1}{2}Q_2 + \frac{1}{4}(Q_1 + Q_3)$$

$trimean()$: $CellChat's\ noise\ resistant\ mean.$
G, Clu: $Gene\ G\ whose\ gene\ expression\ in\ cluster\ Clu\ is\ being\ evaluated.$
Q_n: $The\ n^{th}\ Quartile\ of\ the\ gene\ expression\ of\ gene\ G\ in\ cluster\ Clu.$

Otherwise, it works similarly to CellPhoneDB, as it is also a permutation-based method, and can also handle complexes. CellChat estimates the expression of a complex using the geometric mean of its subunits.

$$geomean(G_{Clu}) = (\prod_{i=1}^{n} EM(S_{i,Clu}))^{1/n}$$

G, S, Clu: $A\ given\ gene\ G\ that\ is\ a\ complex\ with\ subunits\ S_1\ to\ S_n\ , whose\ mean\ expression\ in\ cluster\ Clu\ is\ being\ evaluated.$

CellChat also makes use of a custom scoring function based on the Hill Equation (though the specific dissociation constant and the cooperativity of each interaction are unknown and therefore always the same). Using further terms based on the Hill Equation, they also expand their scoring function to include the impact of soluble agonists and antagonists. Finally, they use a linear model to quantify the impact of membrane bound receptor modulators and a scaling factor that accounts for the cell fraction being analyzed. The final value is termed the interaction probability, though it is not a probability in the traditional sense. It is limited to values between 0 and 1.

$$P\ (L,R,S,T)\ ) = \frac{geomean(L_S)\ *\ geomean(\ R_T)\ *\ \frac{1\ +A_{R,T}}{1\ +\ I_{R,T}}}{K_h\ +\ geomean(L_S)\ *\ geomean(\ R_T)\ *\ \frac{1\ +A_{R,T}}{1\ +\ I_{R,T}}}\ *\ \frac{SI_S}{K_h\ +\ SI_S}$$

$$*\ \frac{SI_T}{K_h\ +\ SI_T}\ *(1\ +\ \frac{SA_S}{K_h\ +\ SA_S})\ *\ (1\ +\ \frac{SA_T}{K_h\ +\ SA_T})\ *\frac{n_S n_T}{n^2}$$

L,R,S,T: $Ligand, Receptor, Source\ Cluster, Target\ Cluster$

P: $Interaction\ probability$

$A,I\ /\ SA,SI$: $Antagonist, Inducer\ /\ Soluble\ Antagonist, Soluble\ Inducer$

n_S , n_T , n : $Number\ of\ cells\ in\ S, number\ of\ cells\ in\ T, number\ of\ cells\ in\ the\ data\ set.$

K_h: $The\ hill\ Coefficient\ is\ always\ 0.5\ in\ this\ case.$

In this manner, CellChat incorporates more information into its LR score than most methods, and is the only method reviewed that accounts for upstream and downstream signaling of a given interaction. Similarly to cluster specificity, the intuitive utility of this is apparent, but the exact impact of this on results remains undetermined. The authors themselves mention that two major coefficients in the Hill Equation, which should be unique per interaction, are simply the same in every comparison, because they are almost always unknown constants (Jin et al. 2021). Similarly, the integration of upstream interaction mediators can violate rules of gene expression (Armingol et al. 2021).

Finally, the specificity of the Interaction Probability is represented by an empirical p-value.

$$p\ (L,R,S,T)$$
$$= \frac{|\ reshuffled\ Interaction\ Probabilities\ >\ original\ Interaction\ Probability|}{|reshuffled\ Interaction\ Probabilities|}$$

p: $p-value.$

L,R,S,T: $Ligand, Receptor, Source\ Cluster, Target\ Cluster$

$|x|$: $The\ number\ of\ items\ in\ the\ set\ x.$

7. 2 Supplementary Note 2 - Analysis of Overlap Formula

In our analysis, we always calculated the overlap between baseline and modified predictions. The overlap is a directional measure, it is scaled by one of the two sets compared. When both sets are the same or similar size, as was often the case in our analysis, the direction is negligible or irrelevant, but in some of our robustness tests, the number of predicted interactions can change dramatically. In these cases, we opted to scale the overlap by the size of the larger of the two sets. This mimics the overlap one would observe when downsampling the larger set to the size of the smaller one, as shown below.

Given a set S and a set L, where L is larger than S. If the number of interactions common between L and S are $|L \cap S|$, and on average downsampling doesn't alter the proportion of overlapping and non-overlapping signatures, then

$$|D \cap S| = |L \cap S| * \frac{|S|}{|L|}$$

D: $The\ average\ downsample\ of\ L.$
L: $Larger\ set\ of\ predictions.$
S: $Smaller\ set\ of\ predictions$
$|x|$: $Number\ of\ items\ in\ the\ set\ x.$

Since D is the same size as S, we can prove the following:

$$O_D^{D,S} = O_S^{D,S} = |D \cap S| \,/\, |D|$$

$$O_D^{D,S} = O_S^{D,S} = |L \cap S| * \frac{|S|}{|L|} * |S|$$

$$O_D^{D,S} = O_S^{D,S} = \frac{|L \cap S|}{|L|} = O_L^{L,S}$$

O: $Overlap\ between\ two\ sets\ of\ predictions. The\ proportion\ of\ one\ set\ that\ is\ in\ the\ other. The\ superscript\ refers\ to\ both\ sets\ being\ compared, the\ subscript\ to\ the\ norming\ factor.$

D: $The\ average\ downsample\ of\ L.$

L: $Larger\ set\ of\ predictions.$

S: $Smaller\ set\ of\ predictions$

$|x|$: $Number\ of\ items\ in\ the\ set\ x.$

7.3 Supplementary Note 3 - Choice of Significance

In all four of our robustness tests, we considered the top 500 scoring interactions as significant. This limit is arbitrary, and as such we repeated most of the analyses here considering the top 1000 and 1500 interactions significant.

We excluded the discriminant dilution from this supplementary analysis. In its case, any interaction that is considered significant in the baseline predictions cannot be diluted. Thus, by expanding the number of interactions considered significant we would limit the degree of dilution that could be considered.

The results for cluster subsetting can be found in Figure 6 and 7. The only deviation from the results with 500 significant interactions lies in CellChat's performance with 1500 significant interactions. Its robustness improves slightly, being roughly 10 percentage points higher at the highest degree of subsetting.

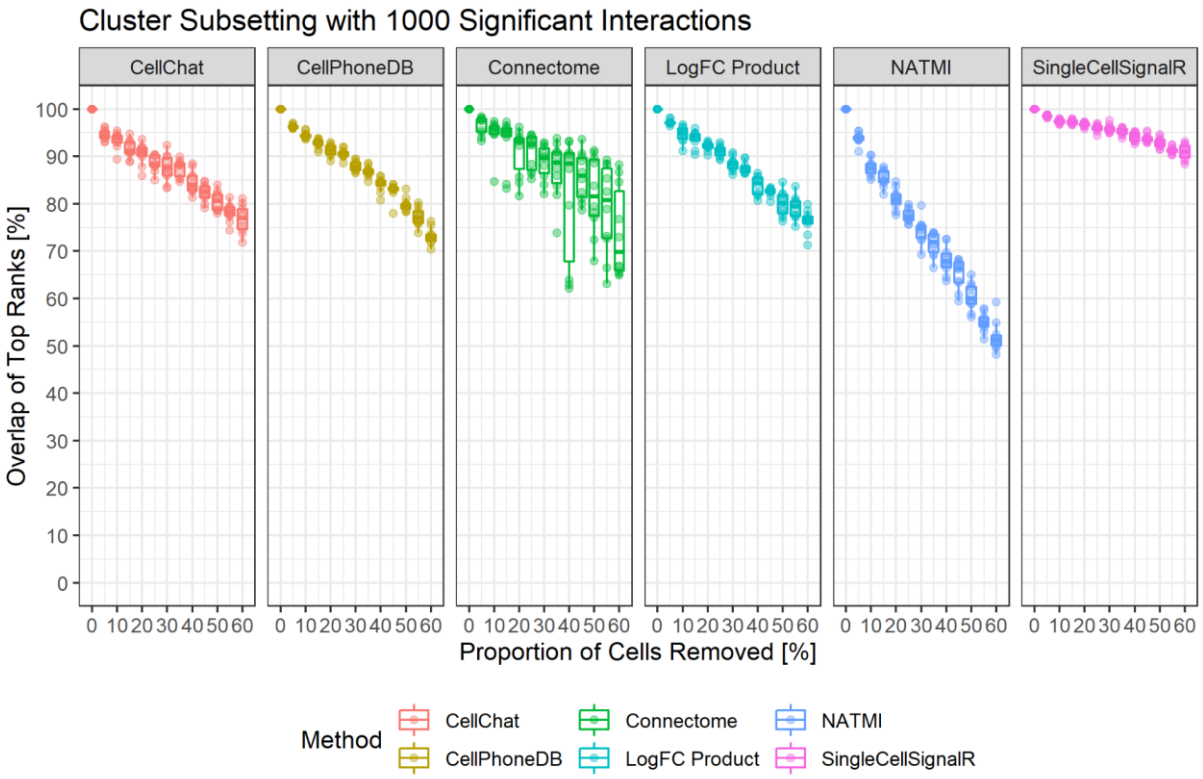

Figure 6: Cluster subsetting robustness with 1000 significant interactions. We repeated the cluster subsetting analysis described in the methods. When assessing significance in that process, we considered the 1000 highest ranked interactions for each method to be significant, instead of the top 500 highest ranked interactions. The results matched the original analysis.

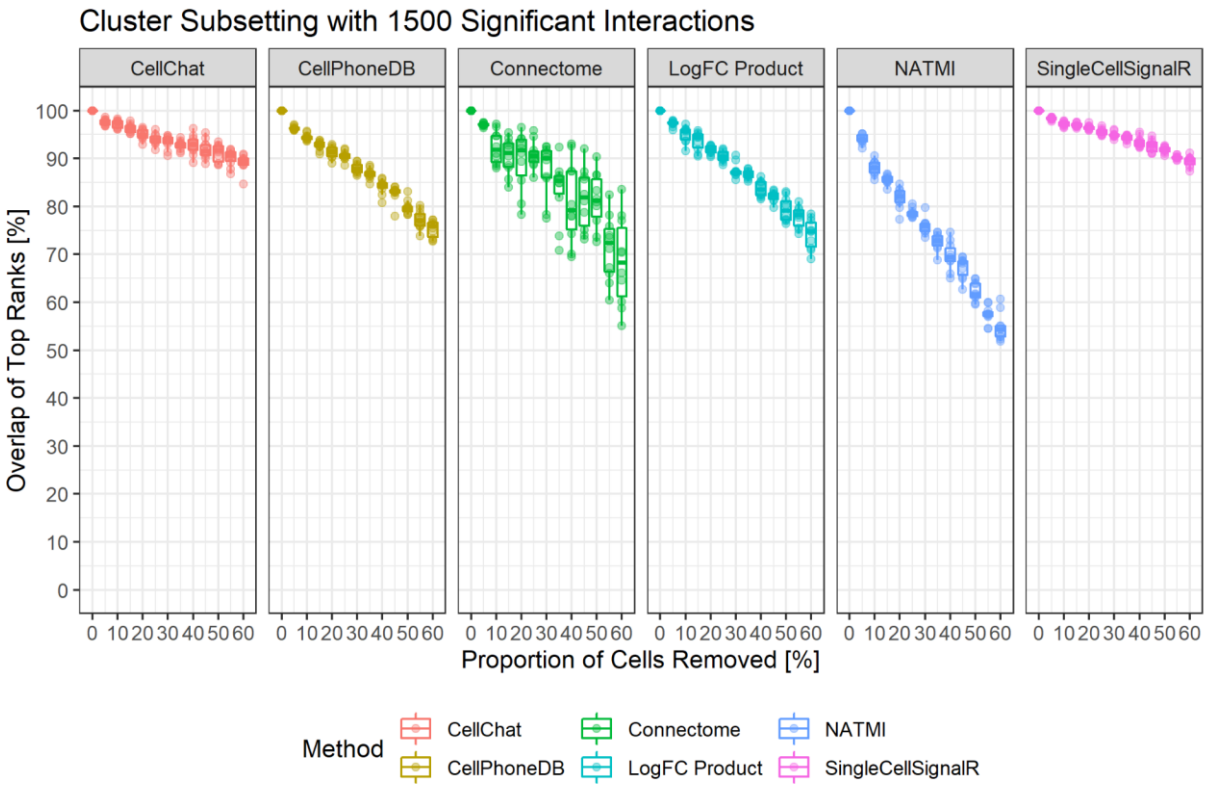

Figure 7: Cluster subsetting robustness with 1500 significant interactions. We repeated the cluster subsetting analysis described in the methods. When assessing significance in that process, we considered the 1500 highest ranked interactions for each method to be significant, instead of the top 500 highest ranked interactions. In comparison to the original analysis, CellChat's robustness increases slightly.

The results for cluster reshuffling can be found in Figure 8 and 9. The only deviation from the results with 500 significant interactions lies in CellPhoneDB's performance with 1500 significant interactions. Its robustness improves, being roughly 20 percentage points higher at the highest degree of reshuffling.

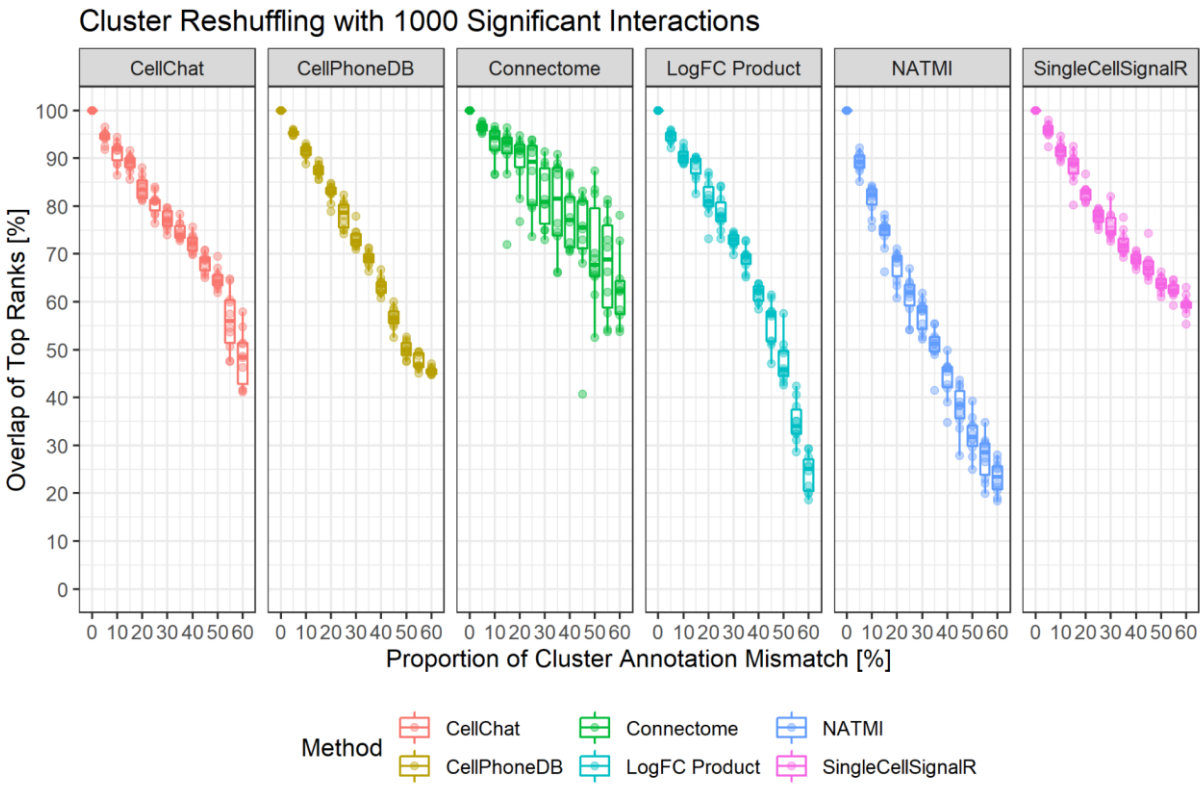

Figure 8: Cluster reshuffling robustness with 1000 significant interactions. We repeated the cluster reshuffling analysis described in the methods. When assessing significance in that process, we considered the 1000 highest ranked interactions for each method to be significant, instead of the top 500 highest ranked interactions. The results matched the original analysis.

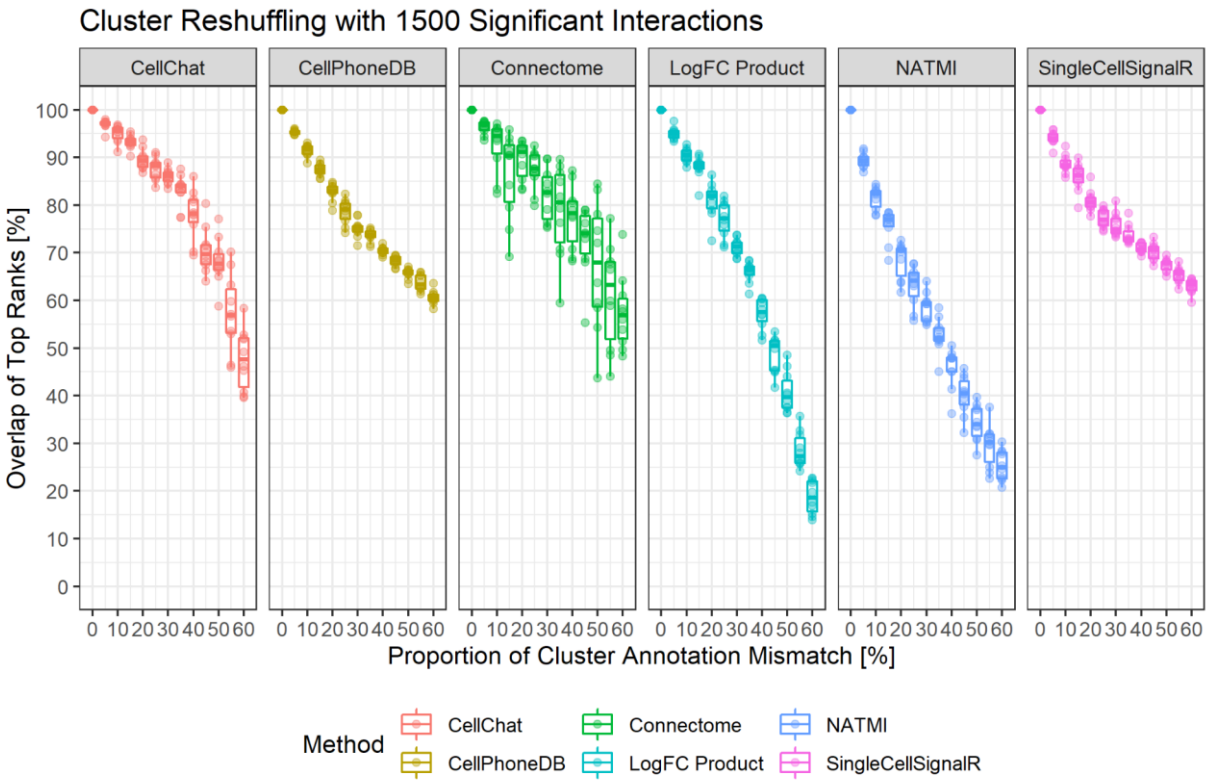

Figure 9: Cluster reshuffling robustness with 1500 significant interactions. We repeated the cluster reshuffling analysis described in the methods. When assessing significance in that process, we considered the 1500 highest ranked interactions for each method to be significant, instead of the top 500 highest ranked interactions. In comparison to the original analysis, CellPhoneDB's robustness increases.

These effects in cluster subsetting and reshuffling can be explained by the nature of these permutation method's predictions. They both produce a heavily skewed p-value distribution, including a large number of interactions all tied for a p-value of 0 (in our example, ca. 1300 interactions have a p-value of 0 in CellChat, and ca. 1900 interactions have a p-value of 0 in CellPhoneDB). Since we include equally ranked interactions at the border of significance, the number of interactions that will usually be included in the baseline is the number of interactions all tied for a p-value of 0, and not the number specified in the experiment (1300 or 1900 instead of 500). This changes when the number of interactions considered significant rises above the number of interactions tied for a p-value of 0, and this in turn impacts the robustness. In addition, subsetting and reshuffling can reduce the number of significant interactions that these methods predict, which can also affect robustness in a similar way when a high number of interactions is considered significant. Essentially, when the top 500 or top 1000

interactions are considered significant, the baseline and the modified predictions will likely all have a p-value of 0, but at 1500 the p-values can be somewhat higher on one or both sides.

Overall, this effect highlights the difficulties in comparing CCC-Inference methods with different approaches but does not greatly change our interpretation of the results. We find it more apt to use the parameters that predominantly compare inferences with a similar significance (500, 1000). Given how many interactions with p-values of 0 are produced, we find it unlikely that in a practical setting other interactions with p-values of 0.15-0.25 would be considered (which is what happens if a significance limit of 1500 is used). Additionally, when considering 1500 interactions significant, CellChat and CellPhoneDB are unevenly affected, as they each predict a different number of interactions with a p-value of 0.

The results for indiscriminate dilution can be found in Figure 10 and 11 They remain constant when considering the top 500, top 1000 and top 1500 interactions.

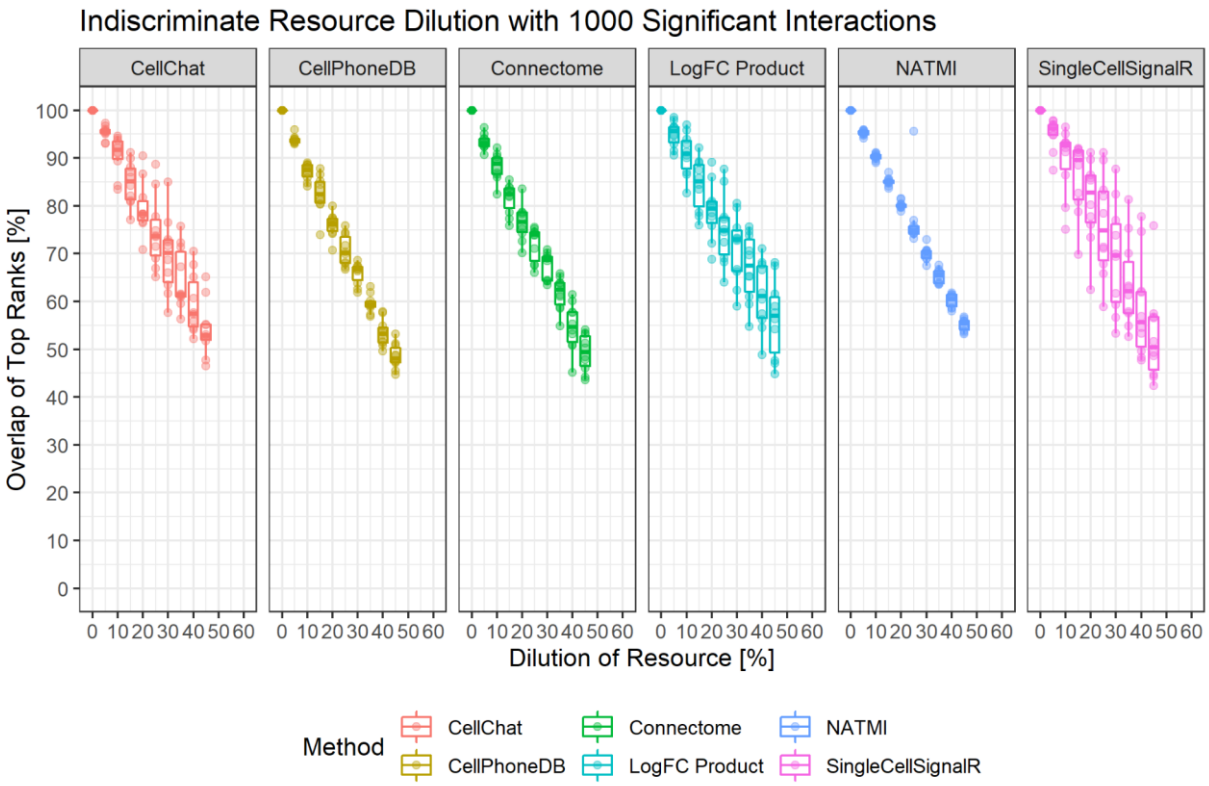

Figure 10: Indiscriminate resource dilution robustness with 1000 significant interactions. We repeated the indiscriminate resource dilution analysis described in the methods. When assessing significance in that process, we considered the 1000 highest ranked interactions for each method to be significant, instead of the top 500 highest ranked interactions. The results matched the original analysis.

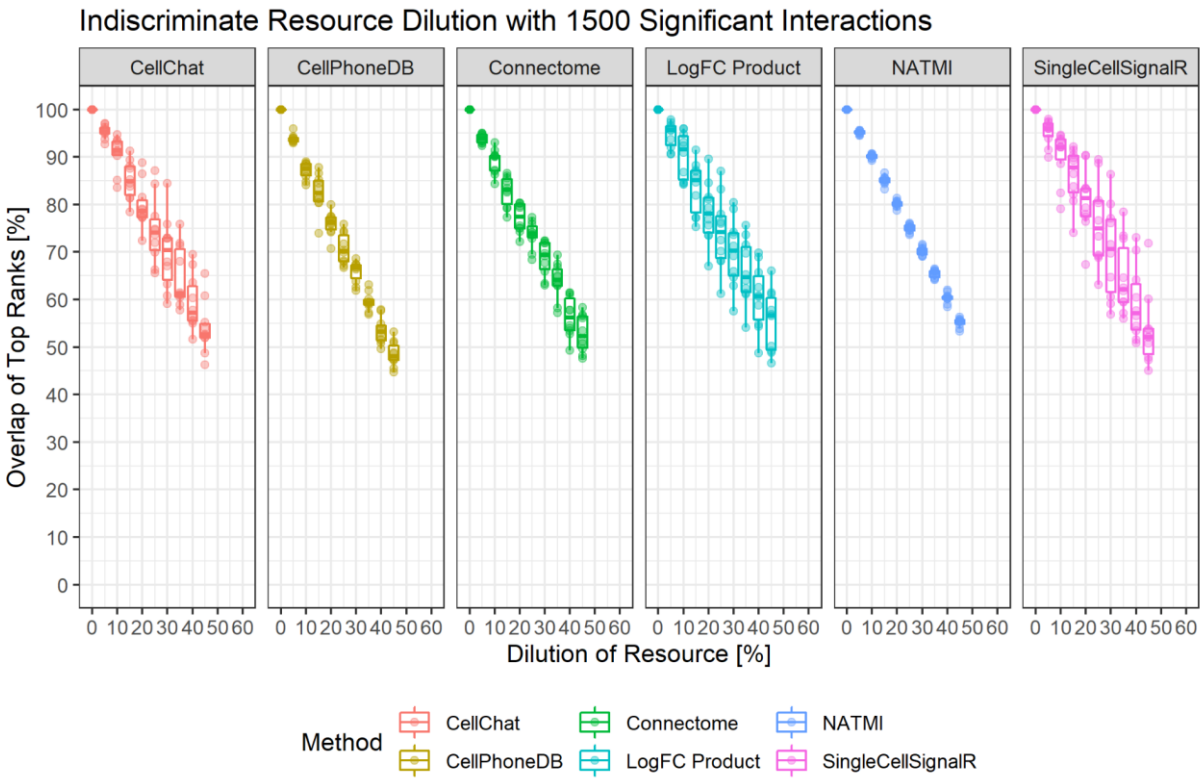

Figure 11: Indiscriminate resource dilution robustness with 1500 significant interactions. We repeated the indiscriminate resource dilution analysis described in the methods. When assessing significance in that process, we considered the 1500 highest ranked interactions for each method to be significant, instead of the top 500 highest ranked interactions. The results matched the original analysis.

7.4 Supplementary Note 4 - Discriminant Dilution with Preserved Resource Topology

In our original discriminant dilution, we replaced OmniPath interactions with spurious interactions constructed at random. We ensured they followed some of the general trends of the interactions in the OmniPath resource (no self-pairing, no duplicate interactions, no crossover between ligands and receptors). In this analysis, we performed discriminant dilution but ensured a much more faithful representation of the topology of the OmniPath resource.

We removed a subset of OmniPath that had not been significant in the baseline in order to dilute it. We went through these interactions, and for each unique gene name that appeared, we reassigned it to a gene name of one of the top 2000 most variable genes in the data set. In this manner the topology of the part we had diluted was entirely preserved, meaning that the overall topology of OmniPath was semi-preserved.

The results of this analysis can be seen in Figure 12. The results are consistent with the original discriminant dilution.

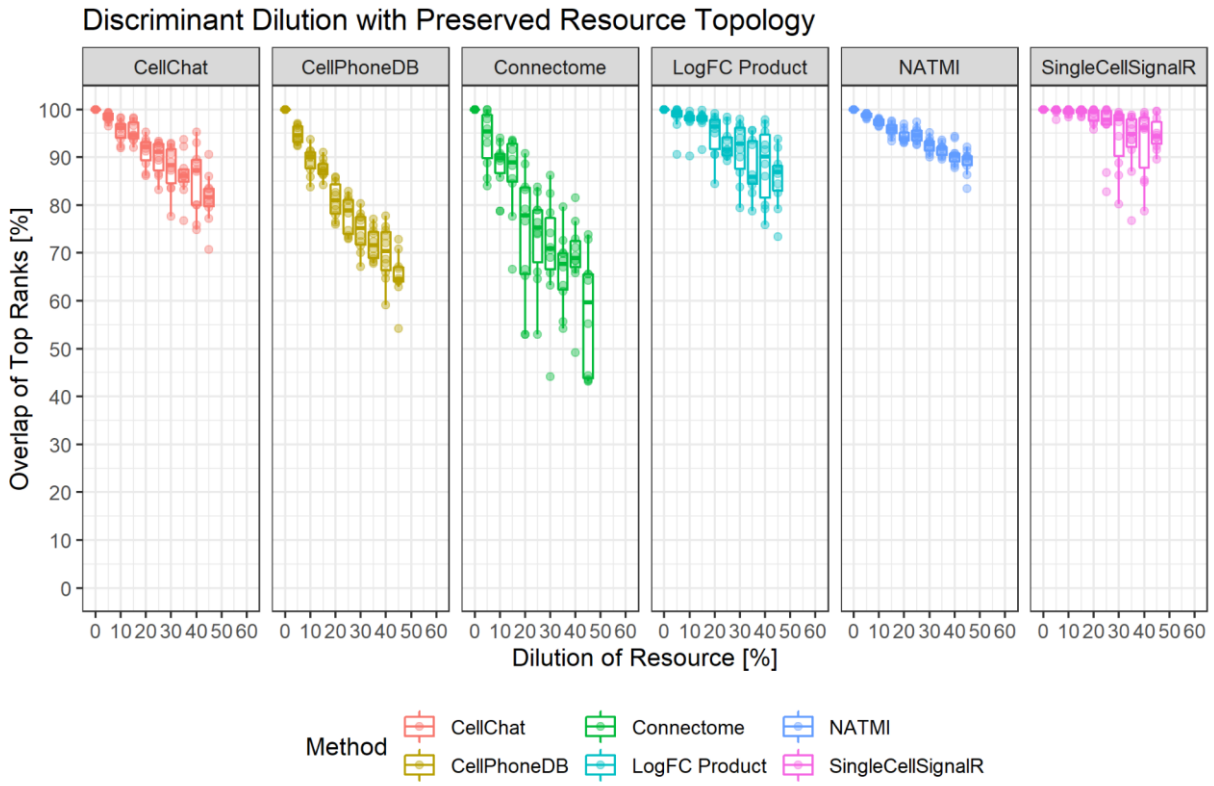

Figure 12: Discriminant resource dilution robustness with preserved resource topology. We repeated the discriminant resource dilution analysis described in the methods. When constructing spurious interactions to use for dilution, we replicated the topology of OmniPath. As such, the resource topology was semi-preserved. The results matched the original analysis.

7.5 Supplementary Note 5 - Discriminant Dilution with Generic Genes

In our original discriminant dilution, every added spurious interaction was made of highly variable genes. This may not always be a fair comparison to a laxly curated resource. Some of its spurious interactions could simply be made of generic or lowly expressed genes. As such, our main analysis of discriminant dilution paints a somewhat worse picture than might always be accurate. Here, we repeated the analysis, creating spurious reactions from any gene that was profiled in the data, rather than only highly variable ones (Figure 13)

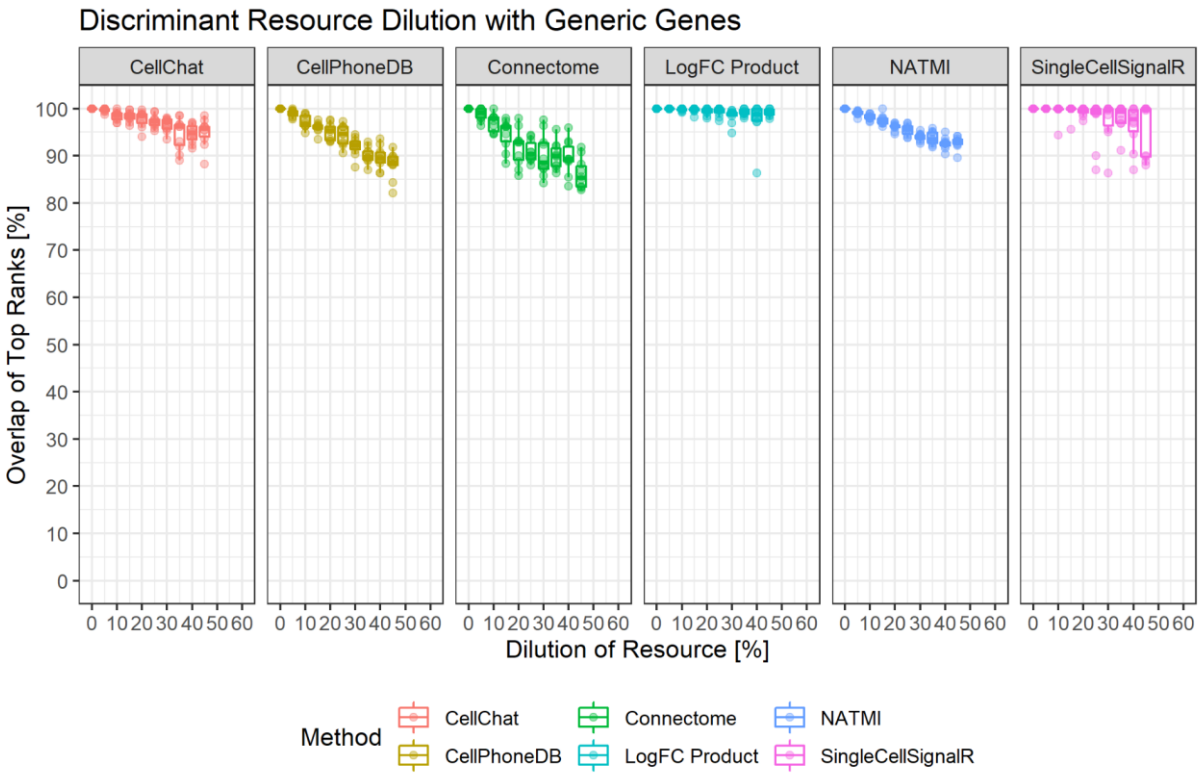

Figure 13: Discriminant resource dilution robustness with generic genes. We repeated the discriminant resource dilution analysis described in the methods. When constructing spurious interactions to use for dilution, we used genes drawn at random from all the genes that were profiled in the data, rather than only drawing from the 2000 most high variance genes. The results show a marked increase in robustness for every method.

This change had a much lower impact on the robustness, with all methods staying well above the 80 % robustness mark. This likely has to do with the sparsity of the average gene's expression scRNA data.

7.6 Supplementary Note 6 - Degreeness of Method Predictions

We analyzed the degreeness of the 6 CCC-Inference methods in our comparison. We ran all 6 of our methods and considered their top 500 highest ranked interactions significant. In cases of interactions tied in value at the edge of significance, we chose to include the higher number of interactions. This means that for CellChat actually ca. 1300 interactions were analyzed, and for CellPhoneDB ca. 1900 interactions were analyzed. We took these significant interactions and counted the unique number of LR pairs that were predicted (degreeness). We then divided the degreeness by the number of interactions that had been considered significant. This provides a measure

to approximate the cluster specificity of the method's predictions; very cluster specific methods would likely predict more unique interactions that are not significant in multiple cluster pairs, thus having a higher degreeness, and a higher degreeness per number of significant interactions profiled.

The results of this analysis can be found in Figure 14.

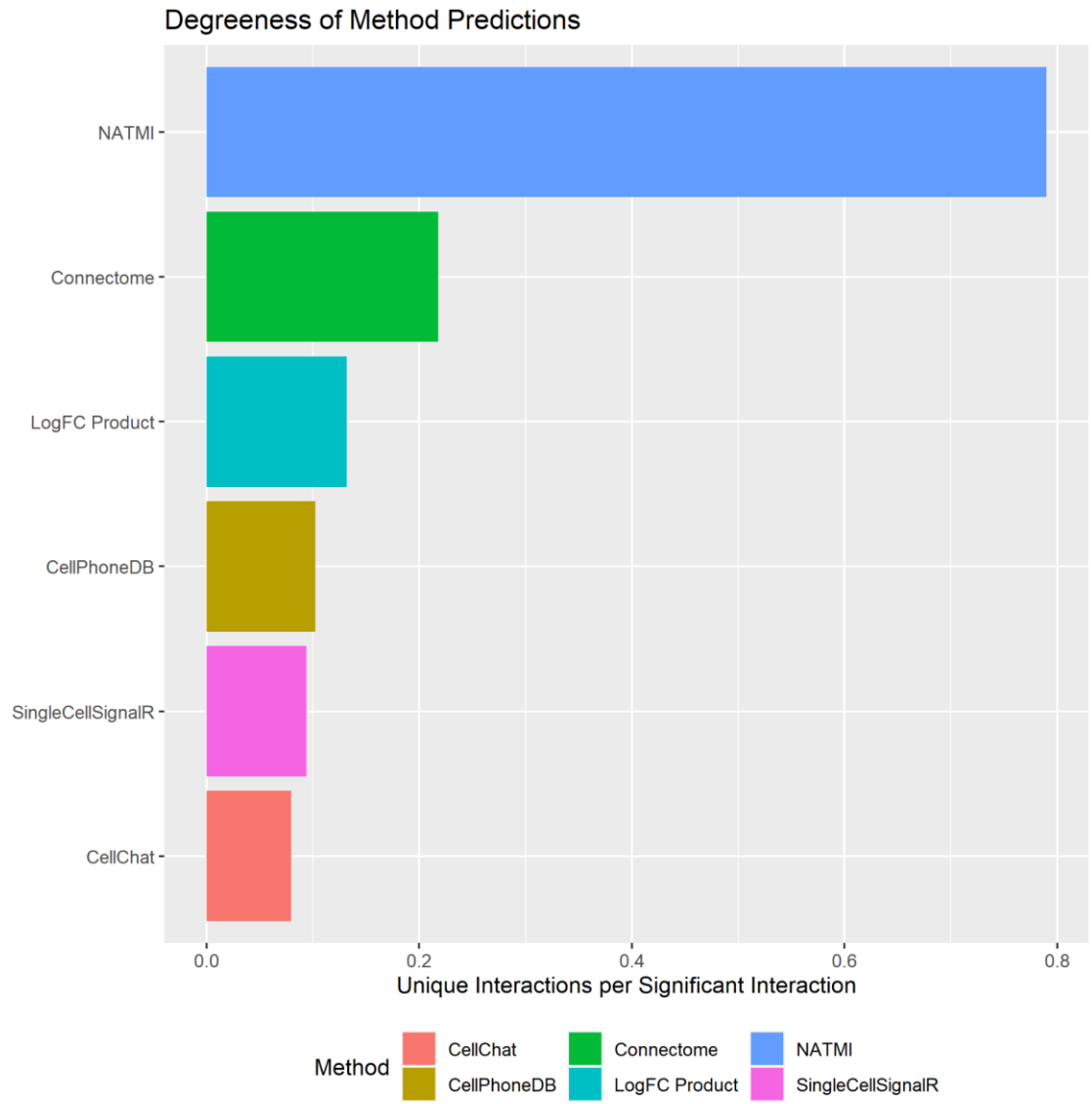

Figure 14: Degreeness of significant CCC-Inference Method predictions. We used every CCC-Inference method to produce CCC-predictions using our data set. We considered the top 500 highest ranked interactions of each method significant, including interactions tied for the 500th rank. We then counted the number of unique interactions within the significant interactions and divided them by the number of significant interactions. These results indicate that NATMI is by far the most cluster specific of all methods.

7.7 Supplementary Note 7 - Permutation vs. Scoring Methods

Two main types of methods were analyzed, scoring methods, and permutation methods. Scoring methods evaluate how communicative the cells and genes within an interaction are using one score. They include the magnitude and often specificity of the ligand and receptor expression into one calculation, and then rank each interaction based on its score.

Permutation-based methods use a scoring function to assess the magnitude of ligand and receptor expression in an interaction using an LR score. They then use a permutation-based approach to assess its significance. They achieve this by completely reshuffling the cluster annotations of the data set many times. For each reshuffled data set, they calculate a reshuffled LR score. Together, the reshuffled LR scores provide an empirical distribution to compare the original score to. Using the original score and the reshuffled distribution, a p-value for a given interaction is calculated.

$$p = \frac{|reshuffled\ scores > original\ scores|}{|reshuffled\ scores|}$$

$p: p - value.$

$$|x|: The\ number\ of\ items\ in\ the\ set\ x.$$

In order to have a low p-value, an interaction must have a higher score between the two-cluster pair being measured than any other two-cluster pair. As such, it should be considered a quantitative measure of the specificity of the interaction but not of its biological relevance. A p-value of 0.05 does not indicate that there is a 95 % chance an interaction is taking place. Instead, it indicates the interaction is fairly specific, and according to the argumentation of many methods, more likely to be taking place. But there isn't necessarily a linear connection between the two. In fact, there is a chance that there is no universal relationship between the specificity-based p-value and the biological false positive rate (Cabello-Aguilar et al. 2020).