nature methods

SUPPLEMENTARY INFORMATION

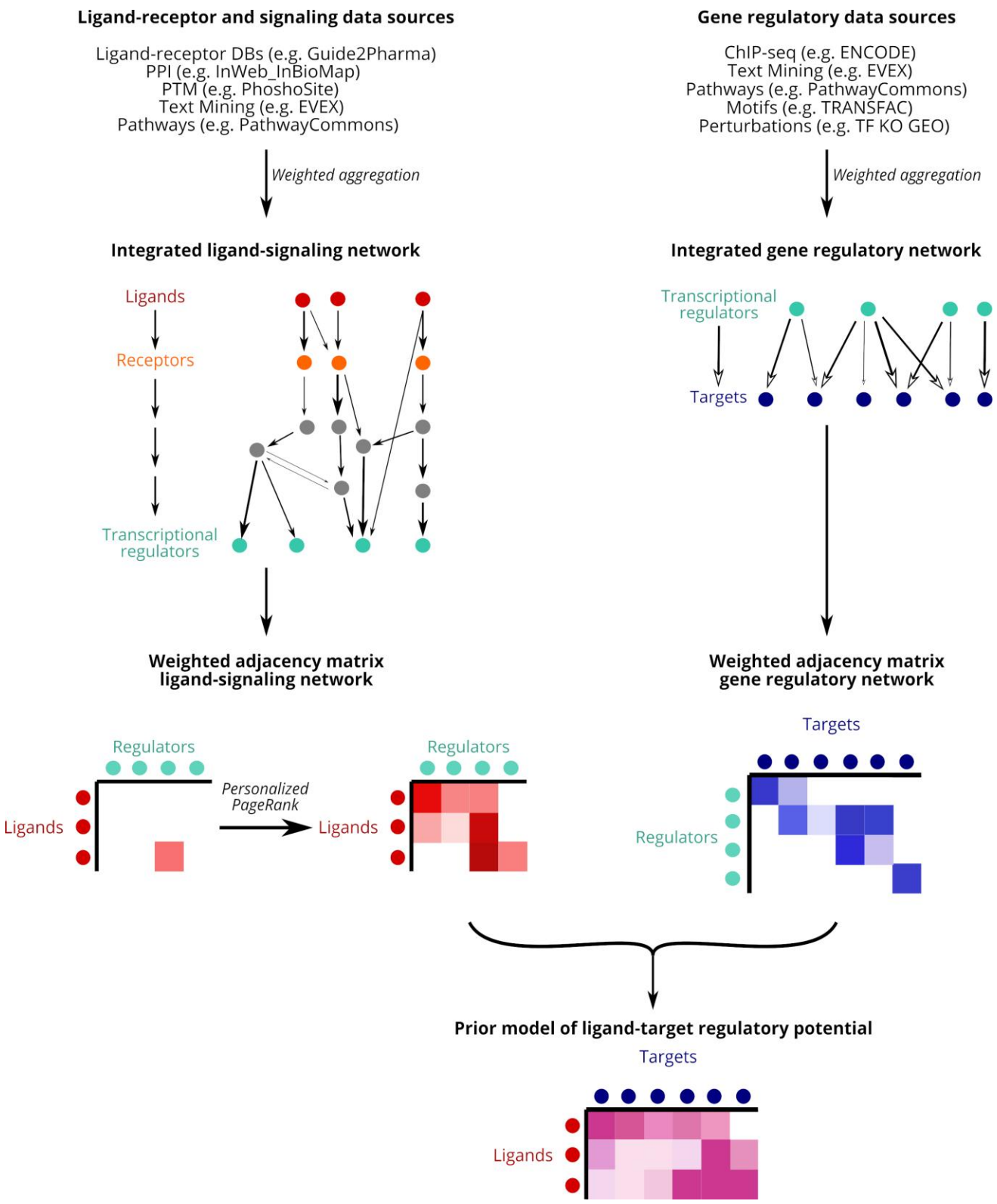
https://doi.org/10.1038/s41592-019-0667-5

In the format provided by the authors and unedited.

NicheNet: modeling intercellular communication by linking ligands to target genes

Robin Browaeys[1,2], Wouter Saelens[1,2,3] and Yvan Saeys[1,2,3]*

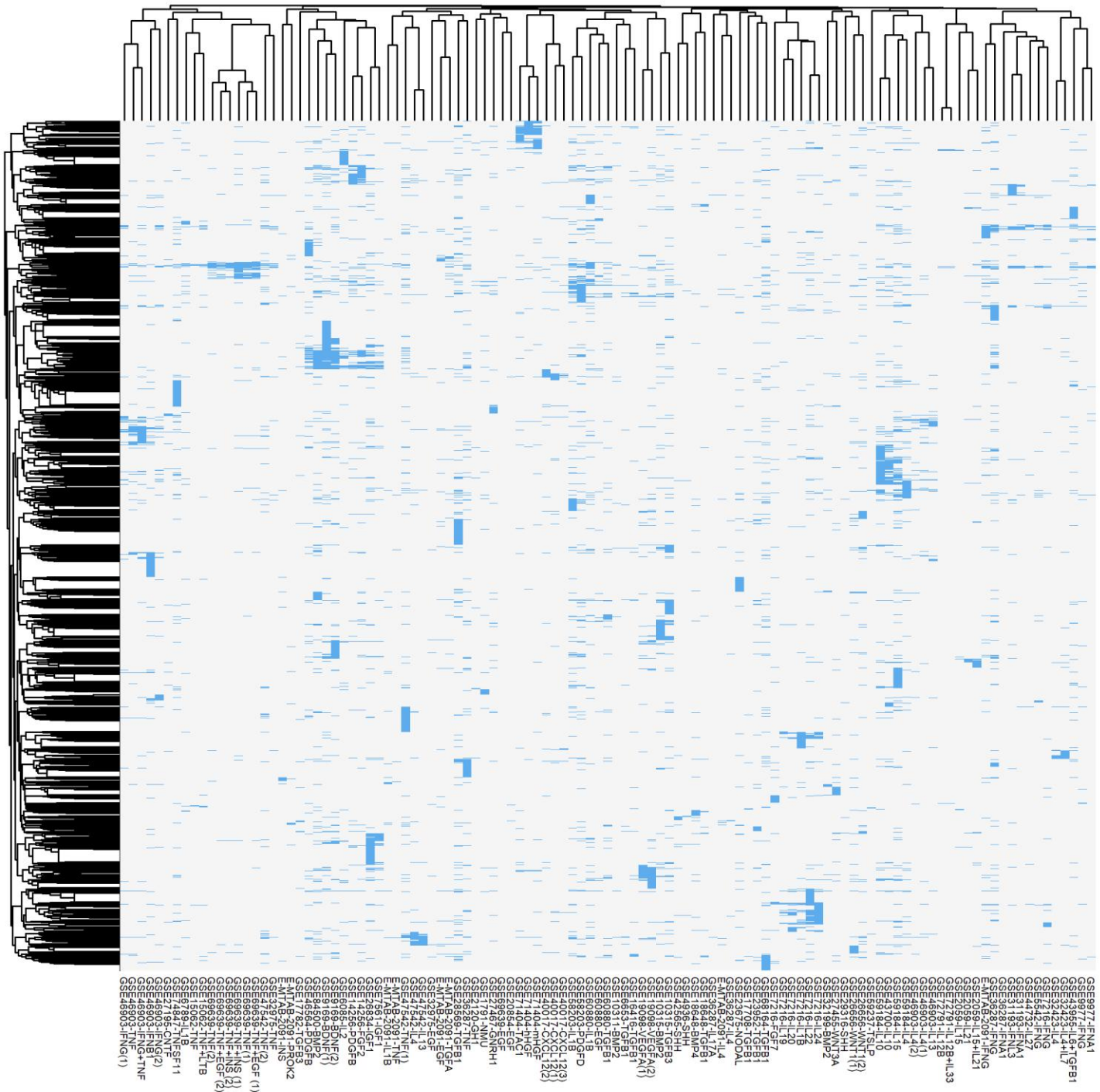
[1]Data Mining and Modeling for Biomedicine, VIB Center for Inflammation Research, Ghent, Belgium. [2]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium. [3]These authors jointly supervised this work: Wouter Saelens, Yvan Saeys.
*e-mail: yvan.saeys@ugent.be



Supplementary Figure 1

Schematic overview of the methodology proposed to predict ligand-target links through data-integration

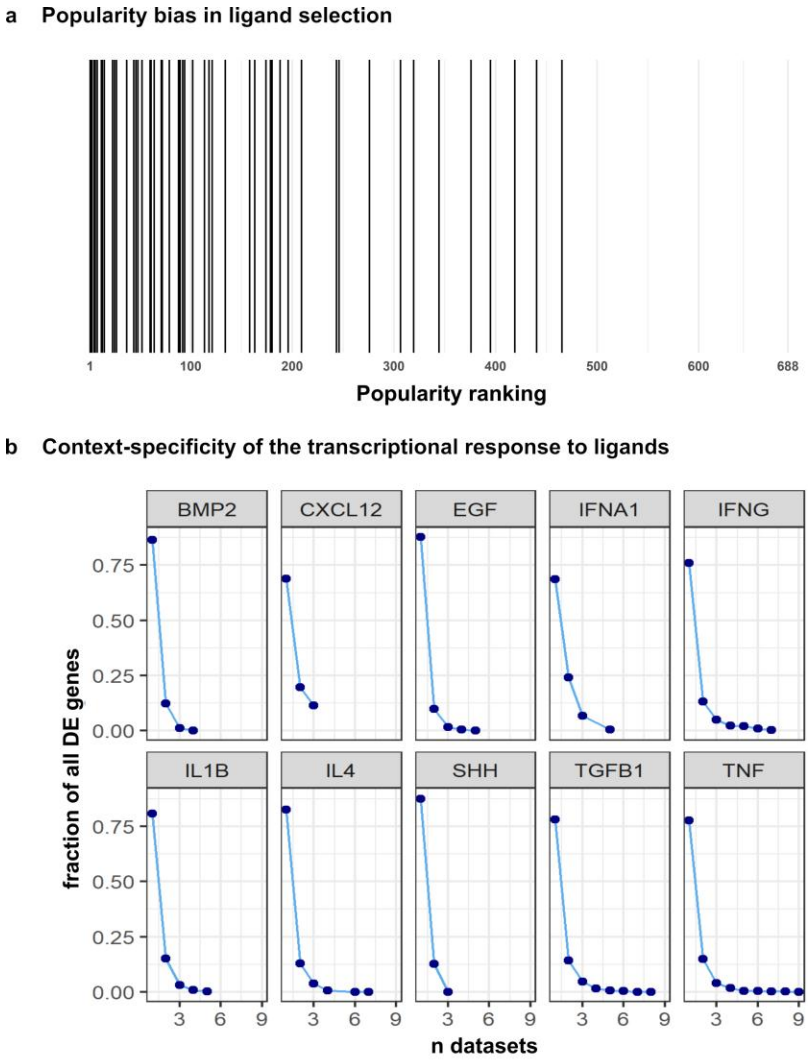
Interactions inferred from ligand-receptor, signaling, and gene regulatory data sources are aggregated in respectively a weighted integrated ligand-signaling and gene regulatory network (line thickness corresponds to edge weight in this visualization). Having these two different integrated networks instead of one network is necessary because the edges represent a different type of interaction: a protein-protein interaction in the ligand-signaling network, and a gene regulatory interaction in the gene regulatory network. Data sources are aggregated in a weighted manner, meaning that a weight is assigned to every data source proportional to how the data source contributes to the final model performance, as automatically determined via model-based optimization via mlrMBO. These networks are then used to first link ligands to downstream transcriptional regulators and secondly to link these regulators to target genes. To link ligands to regulators, the weighted integrated ligand-signaling network is used to calculate for every ligand of interest an importance score for each downstream gene. Here we applied Personalized PageRank to calculate these importance scores which can be considered as a proxy for the probability that a particular ligand signals to a particular regulator. In the second phase, the matrix of ligand-regulator signaling scores is multiplied with the adjacency matrix of the weighted gene regulatory network to obtain ligand-target regulatory potential scores.



Supplementary Figure 2

Clustering of the ligand treatment datasets by their profile of differentially expressed genes

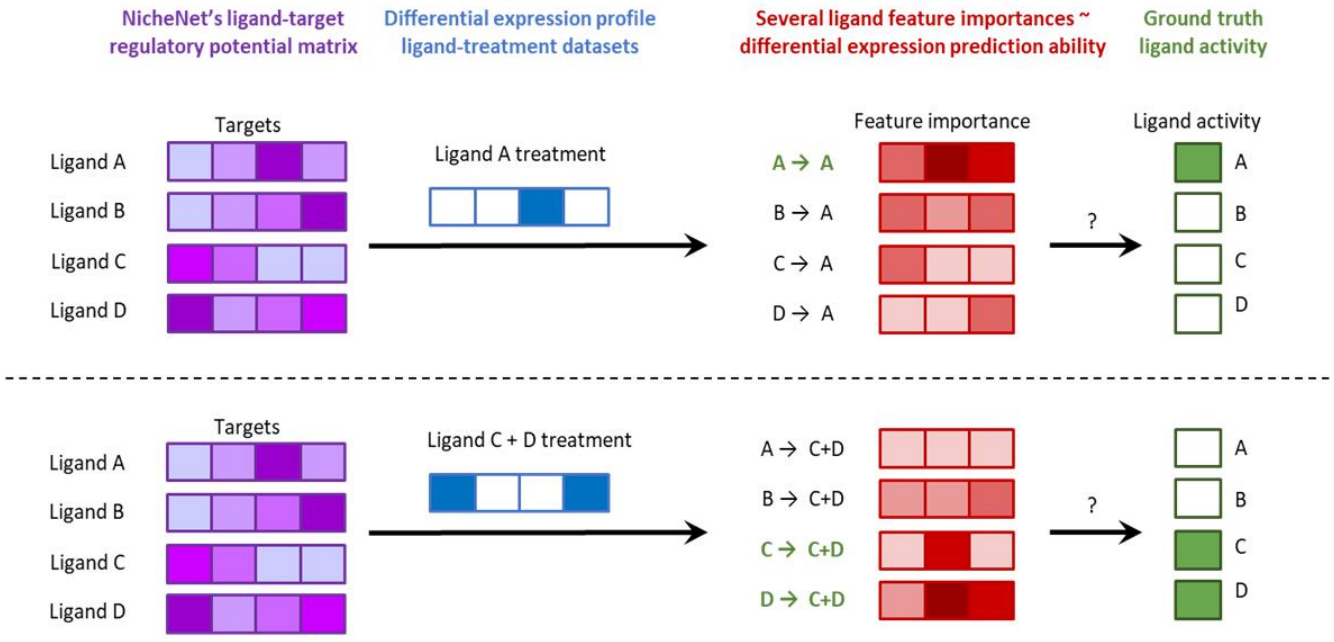
The blue color indicates whether a gene (in the rows; n = 8986) was differentially expressed in a particular dataset (in the columns; GEO accession and profiled ligand are indicated; n = 111) or not. Dendrograms are result of hierarchical clustering with complete linkage and Pearson correlation as clustering distance.



Supplementary Figure 3

Limitations of using ligand treatment datasets for validation of NicheNet's prior ligand-target model

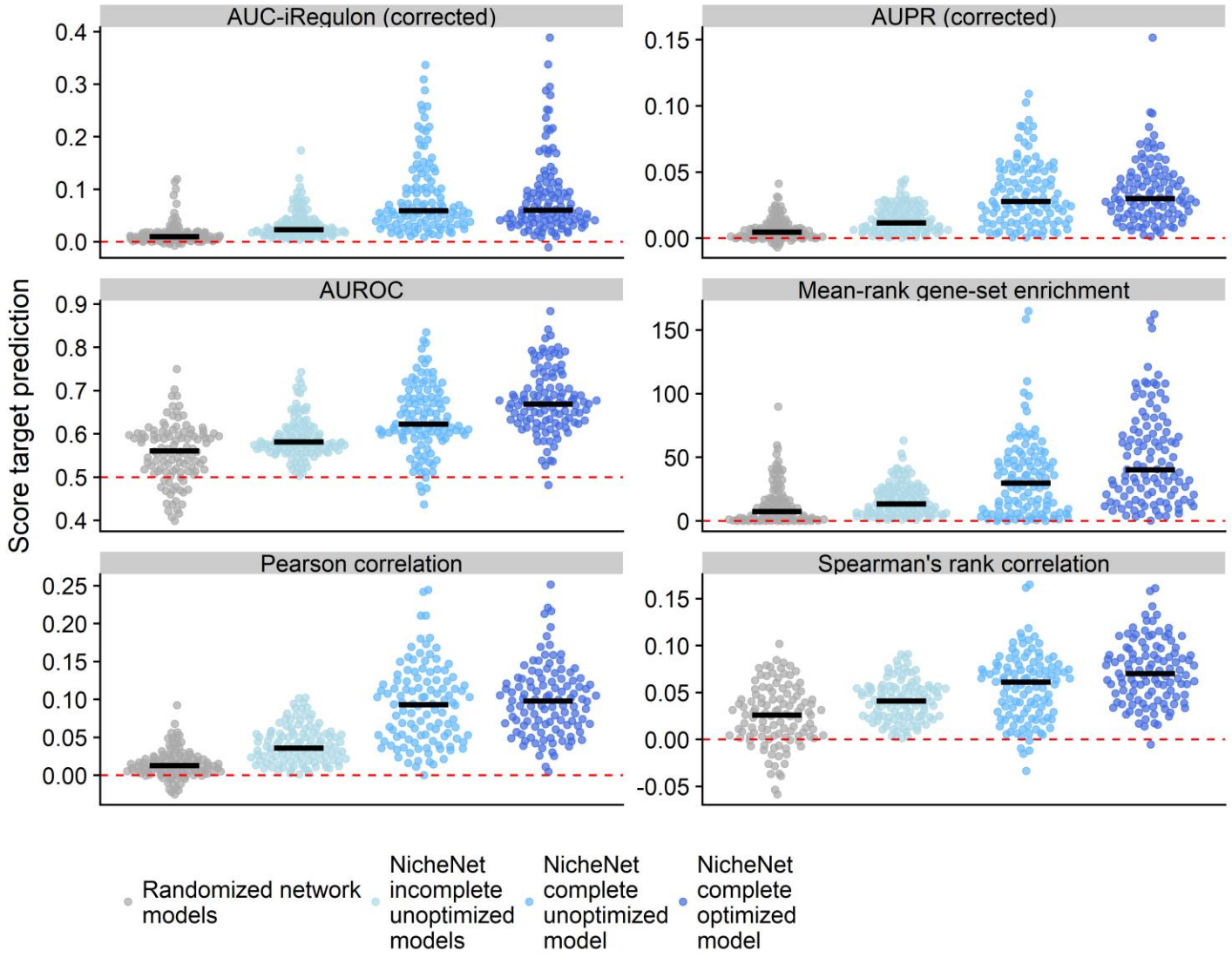
a) Popularity bias in the selection of ligands for which validation ligand treatment datasets were retrieved. All ligands in the NicheNet ligand-receptor data sources were ranked according to the number of PubMed publications in which they are described. Vertical black lines in the barcode plot are drawn when the transcriptional response to that ligand is assessed in one of the ligand treatment datasets used for validation in this study (111 datasets; 51 unique ligands). **b)** Fraction of differentially expressed (DE) genes that overlap between several ligand treatment datasets that profile the response to the same ligand. For ligands of which the transcriptional response was profiled in multiple expression datasets, we show which fraction of all DE genes of a ligand was observed in n of the ligand treatment datasets. A low overlap between different datasets of the same ligand indicate the influence of context-specificity (e.g. profiled cell type) on the transcriptional response to a ligand.



Supplementary Figure 4

Graphical overview of the procedure to assess the performance of NicheNet in predicting ligand activity from expression data

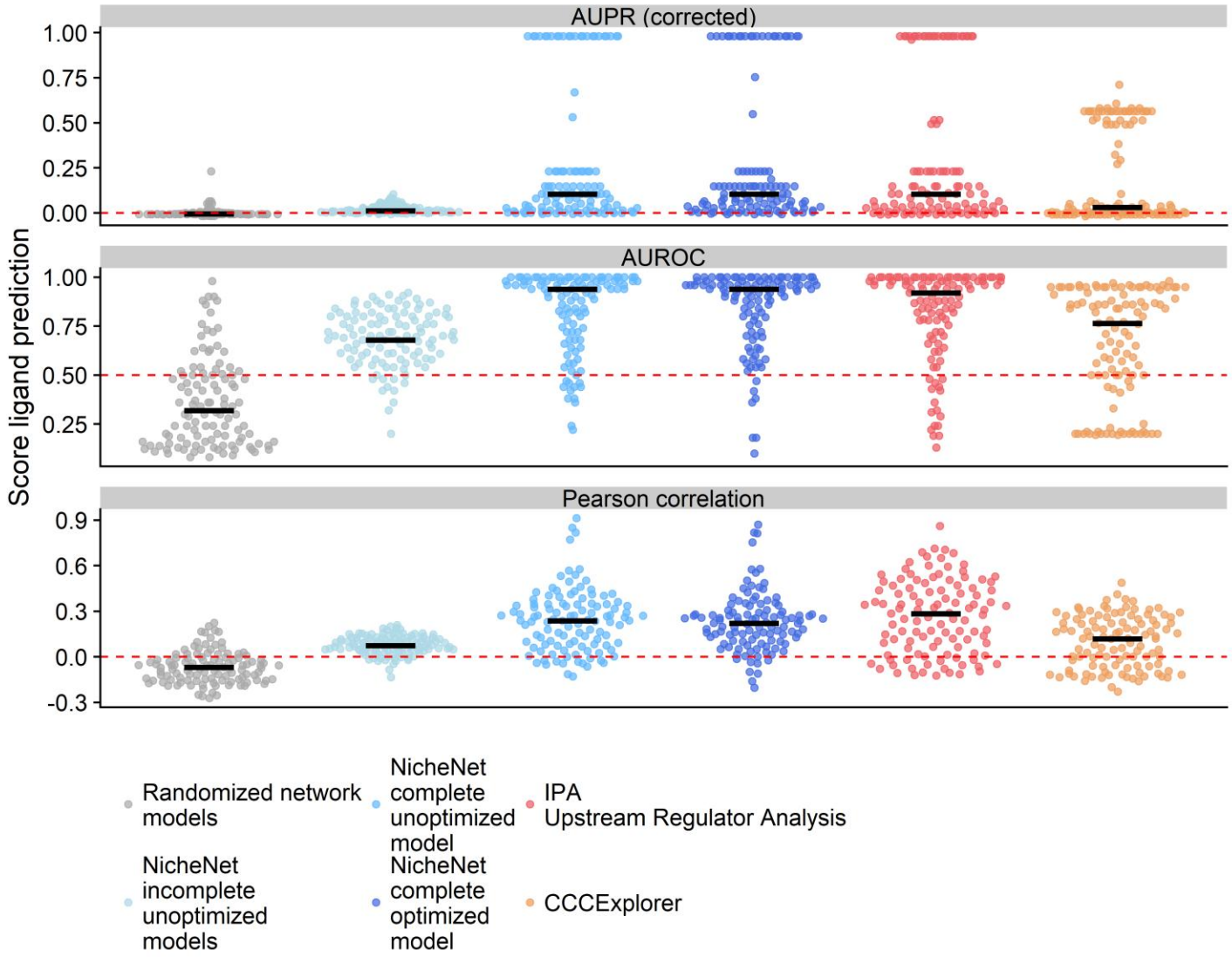
First, we collected gene expression datasets in which the transcriptional response of cells was assessed after treatment with a specific ligand or ligand combination (blue: differentially expressed target genes). Secondly, we determined ligand activity prediction accuracy, by testing the ability of the model in predicting with which ligand the cells were treated based on the expression data. In other words, we evaluated how well NicheNet prioritizes ligands according to their potential to regulate a set of affected genes . This procedure is based on the following assumption: the better a ligand predicts the transcriptional response compared to other ligands, the more likely it is that this ligand is active. Therefore, we used the ligand-target matrix (purple palette of regulatory potential scores) to calculate feature importance measures (red palette, example measures: Pearson correlation coefficient, area under the receiver operating characteristic curve) for all ligands in every expression dataset separately. These feature importance measures are proportional to the ability of a ligand to predict the observed differential expression (e.g., B -> A: feature importance measures indicate how well ligand B can predict the transcriptional response to ligand A). Then, we assessed how well each of these feature importance scores can predict whether a ligand was truly added to the cells (the activity state; green: active). The final model performance in ligand prediction is then the performance of the most predictive feature importance measure.



Supplementary Figure 5

Evaluation of the performance of NicheNet in predicting the transcriptional response: comparison to randomized networks, NicheNet models with fewer data sources, and NicheNet without parameter optimization

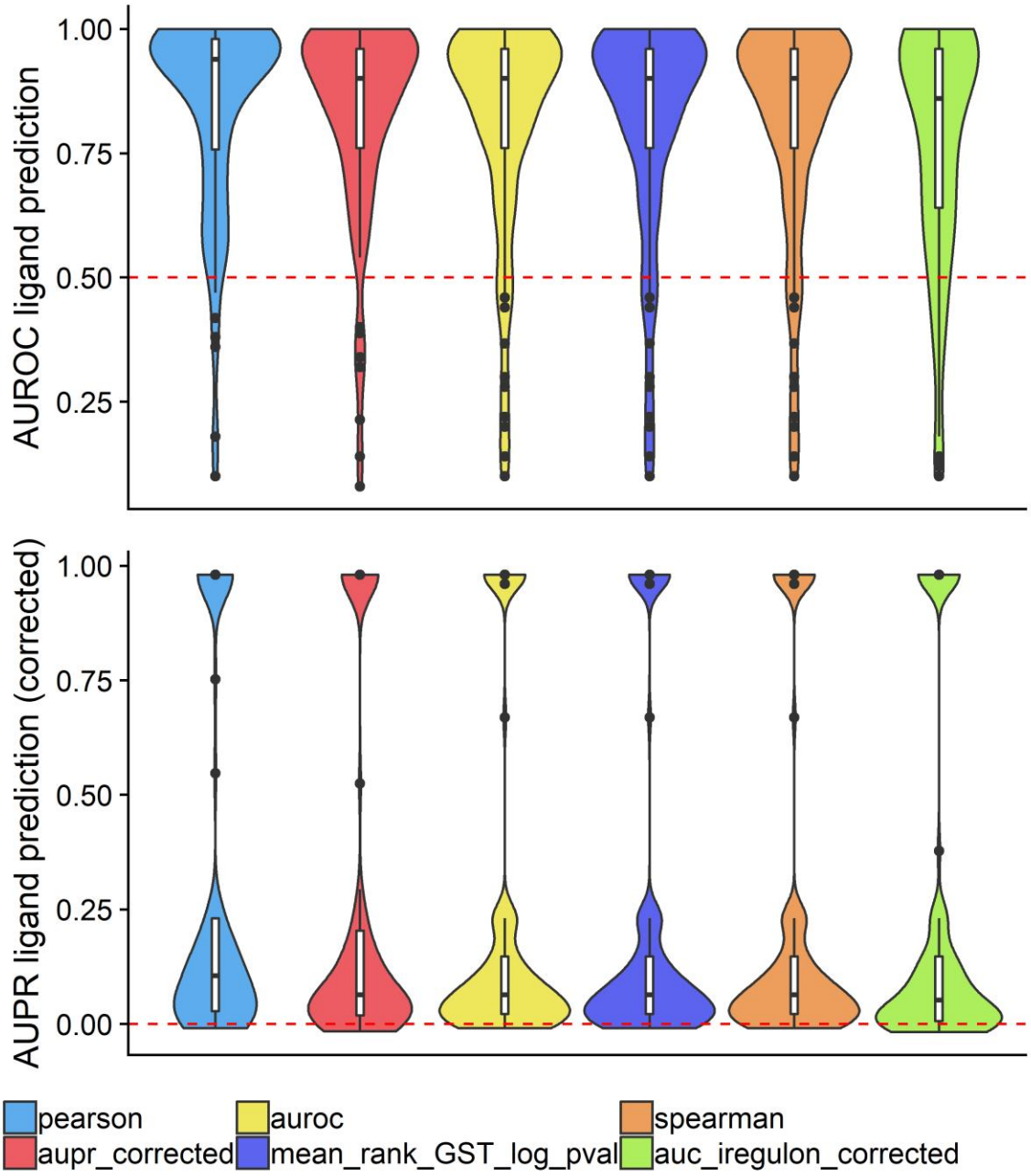
For 111 ligand treatment expression datasets, we calculated several classification evaluation metrics to assess how well NicheNet predicts the set of genes that are differentially expressed in cells upon stimulation with a particular ligand. The target gene prediction performance was compared between: 1) 100 models constructed from 100 randomized networks; 2) 280 incomplete unoptimized NicheNet one-vs-one-vs-one models for which ligand-target regulatory potential scores were calculated after using only one comprehensive ligand-receptor, one signaling, and one gene regulatory database; 3) NicheNet containing all data sources, but without parameter optimization; 4) NicheNet containing all data sources and with optimized parameters. For 1) and 2), we calculated the median performance for each ligand treatment dataset over respectively all 100 and 280 models. Each dot indicates the performance for one dataset, and the black line indicates the median performance over all datasets. Several classification evaluation metrics were calculated: AUC-iRegulon (corrected): area under the cumulative recovery curve (considering the top 3% of the ranking), corrected for random prediction; AUPR (corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under the receiver operating characteristic curve; Mean-rank gene-set enrichment: negative natural logarithm of the mean-rank gene-set enrichment p-value. Red dashed line: performance of random guessing.



Supplementary Figure 6

Evaluation of the performance of NicheNet in predicting ligand activity: comparison to randomized networks, NicheNet models with fewer data sources, NicheNet without parameter optimization, CCCExplorer, and IPA Upstream Regulator Analysis

For 111 ligand treatment expression datasets, we assessed the ability in predicting with which ligand cells were treated based on the set of genes that are differentially expressed upon ligand treatment. Ligand prediction performance was compared between: 1) 100 models constructed from 100 randomized networks; 2) 280 incomplete unoptimized NicheNet one-vs-one-vs-one models for which ligand-target regulatory potential scores were calculated after using only one comprehensive ligand-receptor, one signaling, and one gene regulatory database; 3) NicheNet containing all data sources, but without parameter optimization; 4) NicheNet containing all data sources and with optimized parameters; 5) Upstream Regulator Analysis of IPA; 6) CCCExplorer (adapted to run on ligand treatment datasets). For 1) and 2), we calculated the median performance for each ligand treatment dataset over respectively all 100 and 280 models. Each dot indicates the performance for one dataset, and the black line indicates the median performance over all data sets. Several classification evaluation metrics were calculated: AUPR (corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under the receiver operating characteristic curve and Pearson correlation. Red dashed line: performance of random guessing.

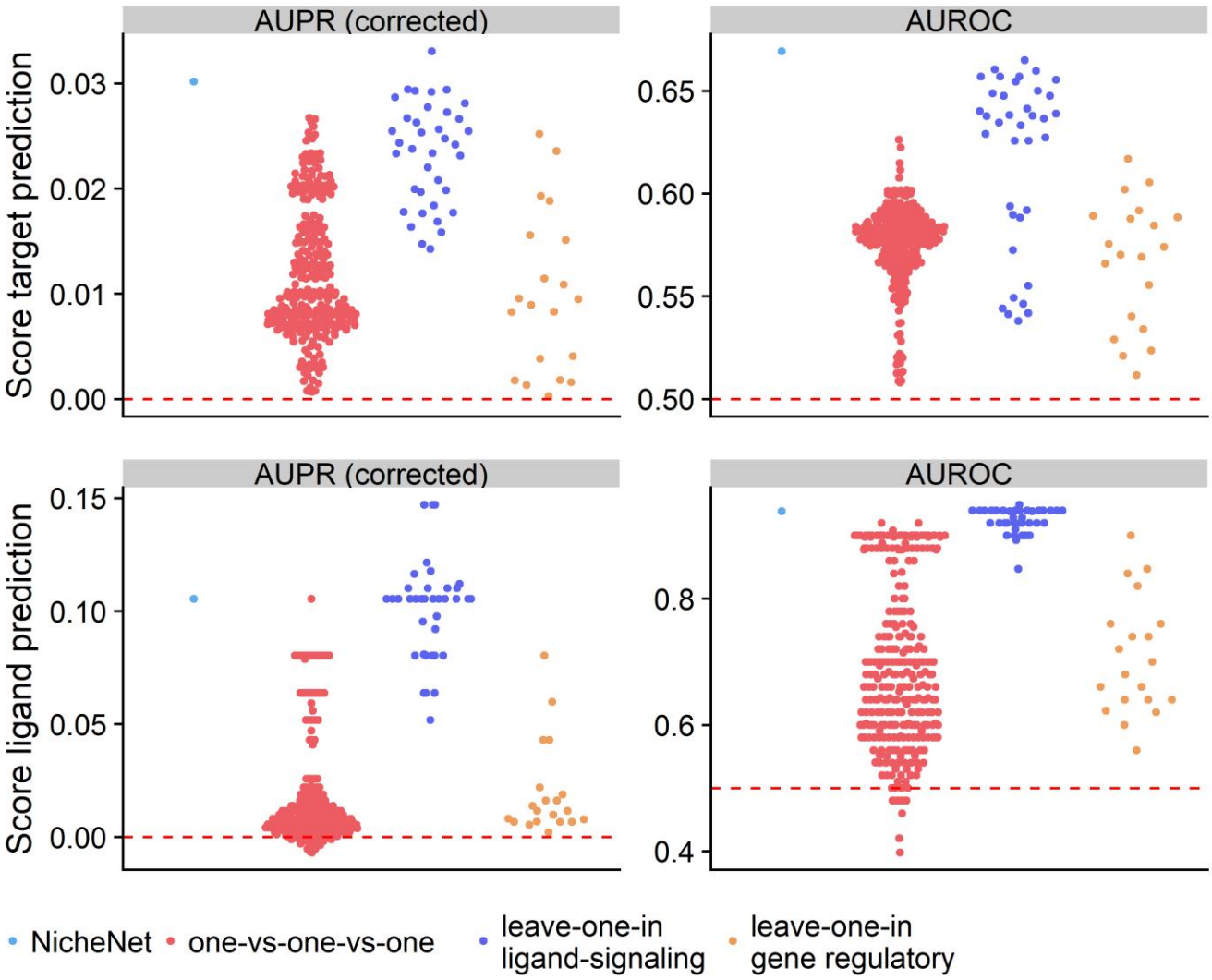


Supplementary Figure 7

Performance of several feature importance scores in predicting ligand activity

For 111 ligand treatment expression datasets, we assessed the ability of NicheNet in predicting with which ligand cells were treated based on the set of genes that are differentially expressed upon ligand treatment. For this procedure, ligand-target regulatory potential scores were used to calculate feature importance measures for all ligands in every expression dataset separately (n = 111 datasets; 51 unique ligands) and these importance measures were in a next step used to predict the ligand activity state of each ligand in each dataset. The performance of each feature importance measure in ligand activity prediction (as evaluated by the AUROC and AUPR) is shown for different feature importance measures: spearman: Spearman's rank correlation; pearson: Pearson correlation; mean_rank_GST_log_pval : negative natural logarithm of the mean-rank gene-set enrichment p-value; auc_iregulon_corrected: area under the cumulative recovery curve (considering the top 3% of the ranking), corrected for random prediction; aupr_corrected: area under the precision-recall curve, corrected for random prediction; auroc: area under the receiver operating characteristic curve. Boxplot elements: center line: median; box limits: upper and lower quartiles; whiskers: 1.5x interquartile range; points: outliers . The violin plots

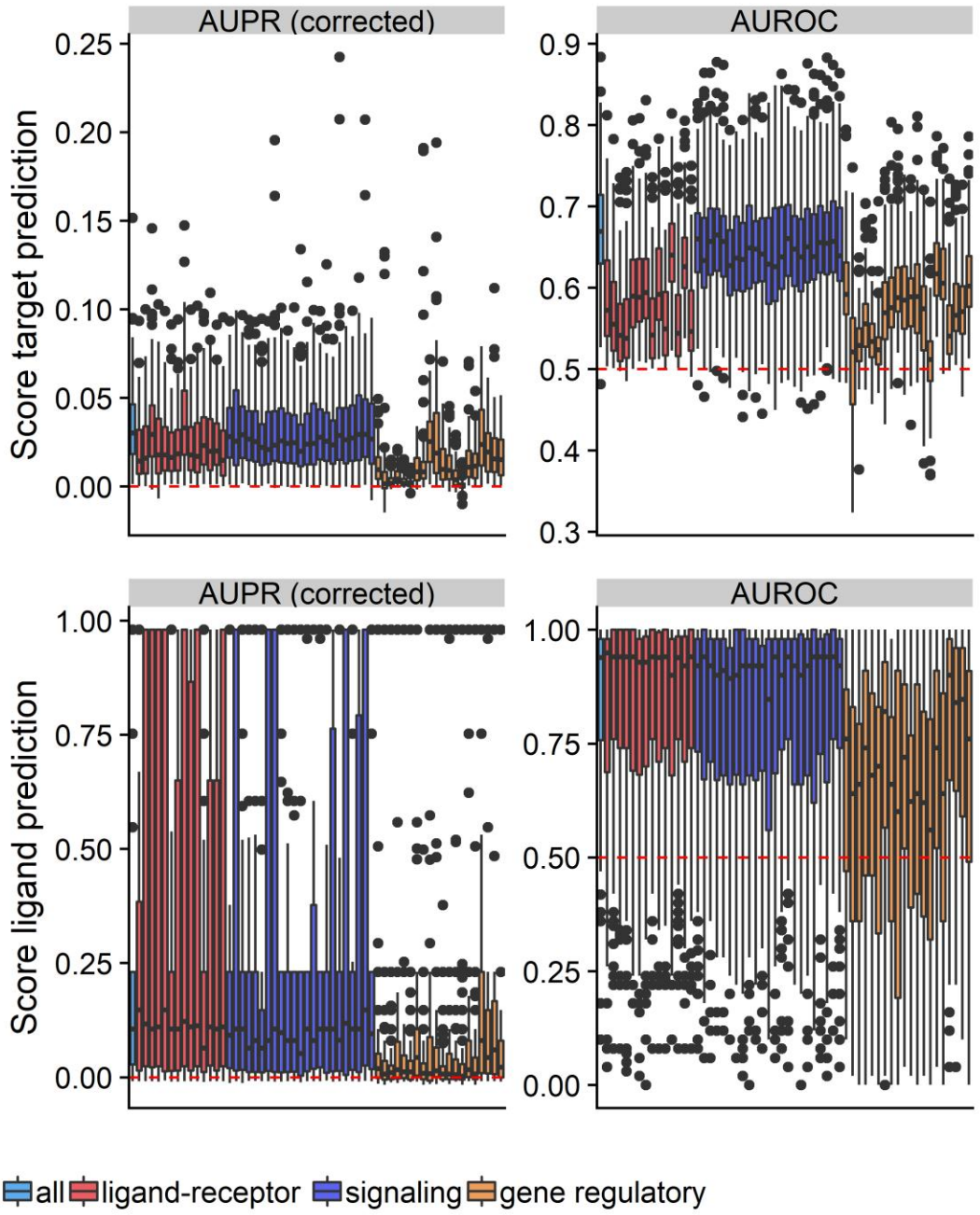
show the probability density of the data (tails of the violins are trimmed to the range of the data). Red dashed line: performance of random guessing.



Supplementary Figure 8

Comparison between NicheNet, one-vs-one-vs-one models and leave-one-in models in target gene and ligand activity prediction performance

To investigate the importance of the integration of multiple complementary data sources, we compared the performance of Niche Net to "one-vs-one-vs-one" and "leave-one-in" models. In one-vs-one-vs-one models, ligand-target regulatory potential scores were calculated after using only one comprehensive ligand-receptor, one signaling and one gene regulatory database (n = 280 one-vs-one-vs-one models in total). In ligand-signaling leave-one-in models, one ligand-signaling data source was used with all gene regulatory data sources; in gene regulatory leave-one-in models the opposite (n = 37 ligand-signaling and n = 20 gene regulatory leave-one-in models in total). For each separate one-vs-one-vs-one and leave-one-in model, we calculated the median target gene and ligand activity prediction performances over all ligand treatment datasets (111 datasets for 51 unique ligands) and compared to the median performances of the complete NicheNet model. AUPR (corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under the receiver operating characteristic curve. Red dashed line: performance of random guessing.

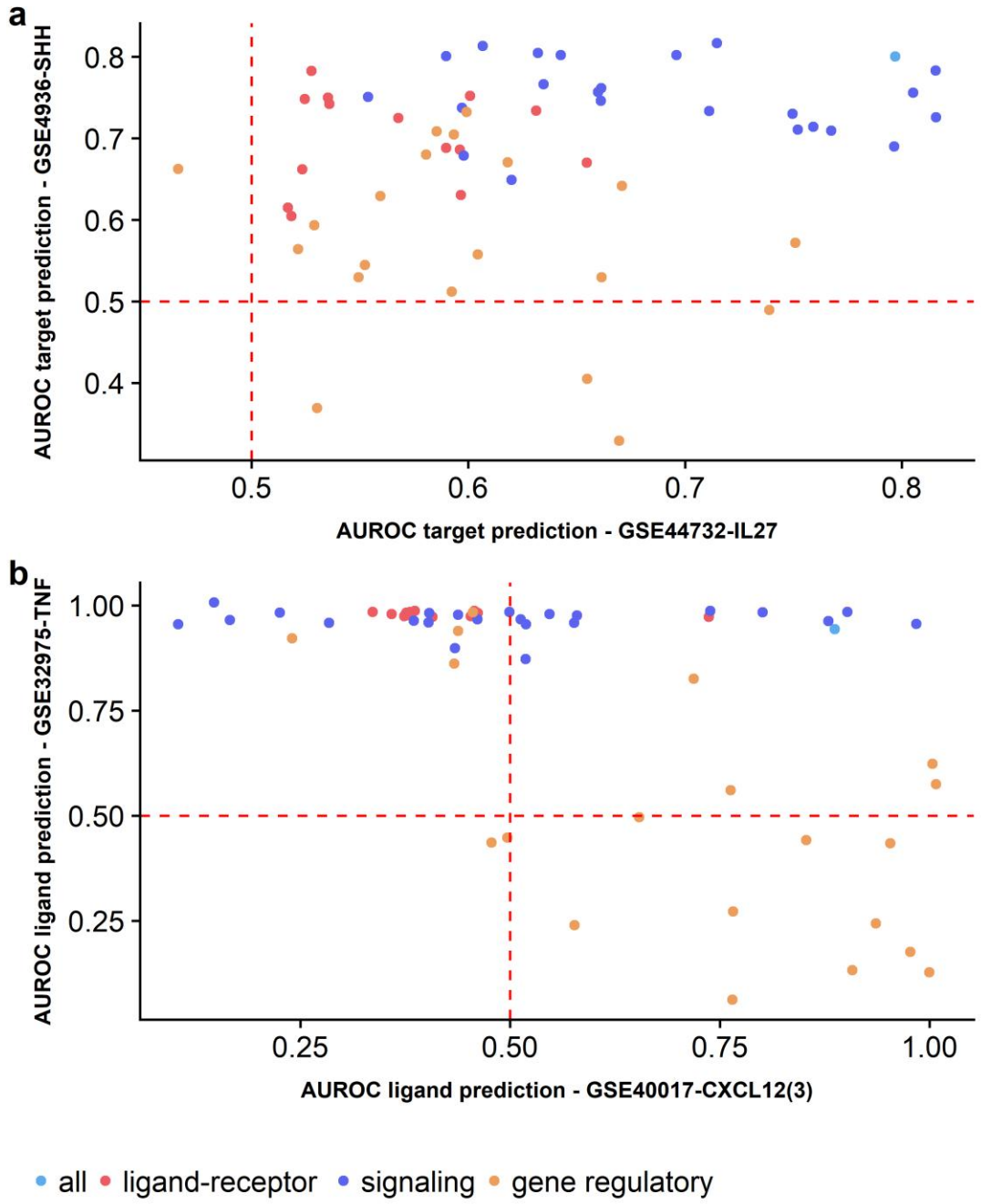


Supplementary Figure 9

Target gene and ligand activity prediction performance of NicheNet of complete and leave-one-in models

To assess the information content of individual data sources, we constructed "leave-one-in" models of ligand-target regulatory potential scores in which every data source in a specific layer (ligand-signaling or gene regulatory) was left out, except a single data source of interest. Hereby, differences in model performances between leave-one-in models of the same layer can be attributed to differences in information content between the corresponding data sources. It should be remarked that a distinction between ligand-receptor and signaling interaction data sources was in this figure only made for visualization purposes, not for the construction of leave-one-in models. The target gene and ligand activity prediction performances of the individual ligand-signaling (red and dark blue) and gene regulatory (orange) leave-one-in models were compared to the performance of the complete final model (light blue) (n = 111 ligand treatment datasets for 51 unique ligands). AUPR (corrected): area under the precision-recall curve, corrected for random prediction;

AUROC: area under the receiver operating characteristic curve. Boxplot elements: center line: median; box limits: upper and lower quartiles; whiskers: 1.5x interquartile range; points: outliers. Red dashed line: performance of random guessing.

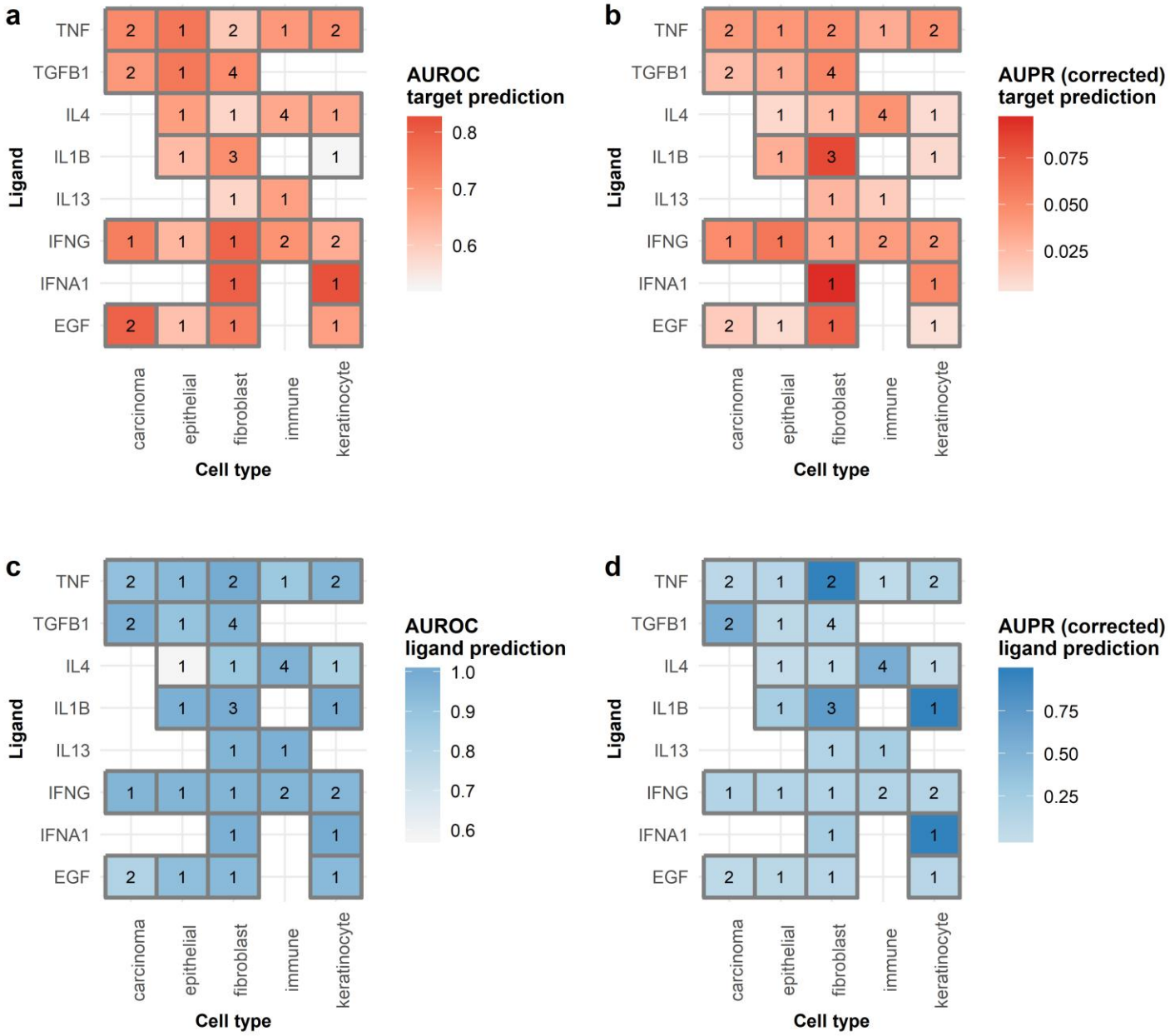


Supplementary Figure 10

The contribution of individual data sources to target gene and ligand activity prediction performance can vary over different ligand treatment datasets

To assess the effect of individual data sources on target gene and ligand activity performance prediction performance on individual ligand treatment datasets, we evaluated the performance of "leave-one-in" models of ligand-target regulatory potential scores in which every data source in a specific layer (ligand-signaling or gene regulatory) was left out, except an individual data source of interest. Hereby, differences in model performances between leave-one-in models of the same layer can be attributed to differences in information content between the corresponding data sources. It should be remarked that a distinction between ligand-receptor and signaling interaction data sources was in this figure only made for visualization purposes, not for the construction of leave-one-in models. For two pairs of ligand treatment datasets, we compared in (a) the target gene prediction performances and in (b) the ligand

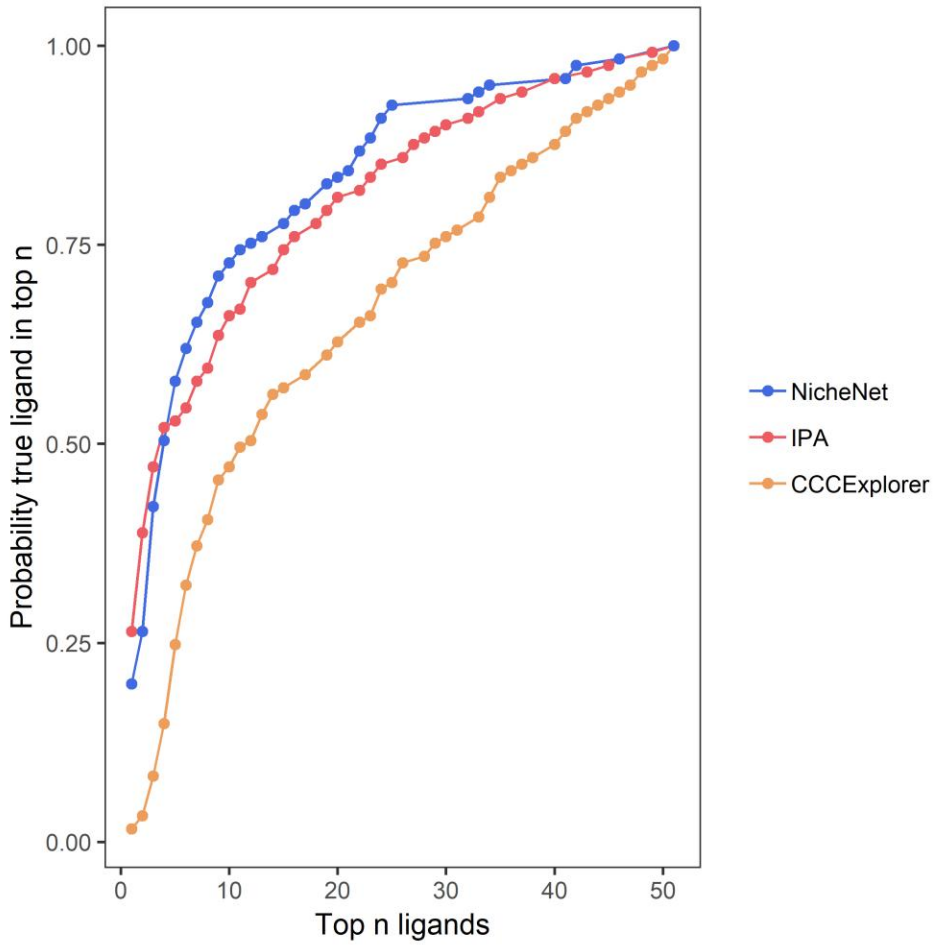
activity prediction performances of the individual ligand-signaling (red and dark blue) and gene regulatory (orange) leave-one-in models to each other and to the complete final model (light blue). AUROC: area under the receiver operating characteristic curve. Red dashed line: performance of random guessing.



Supplementary Figure 11

Evaluation of cell type bias in target gene and ligand activity prediction performance of NicheNet

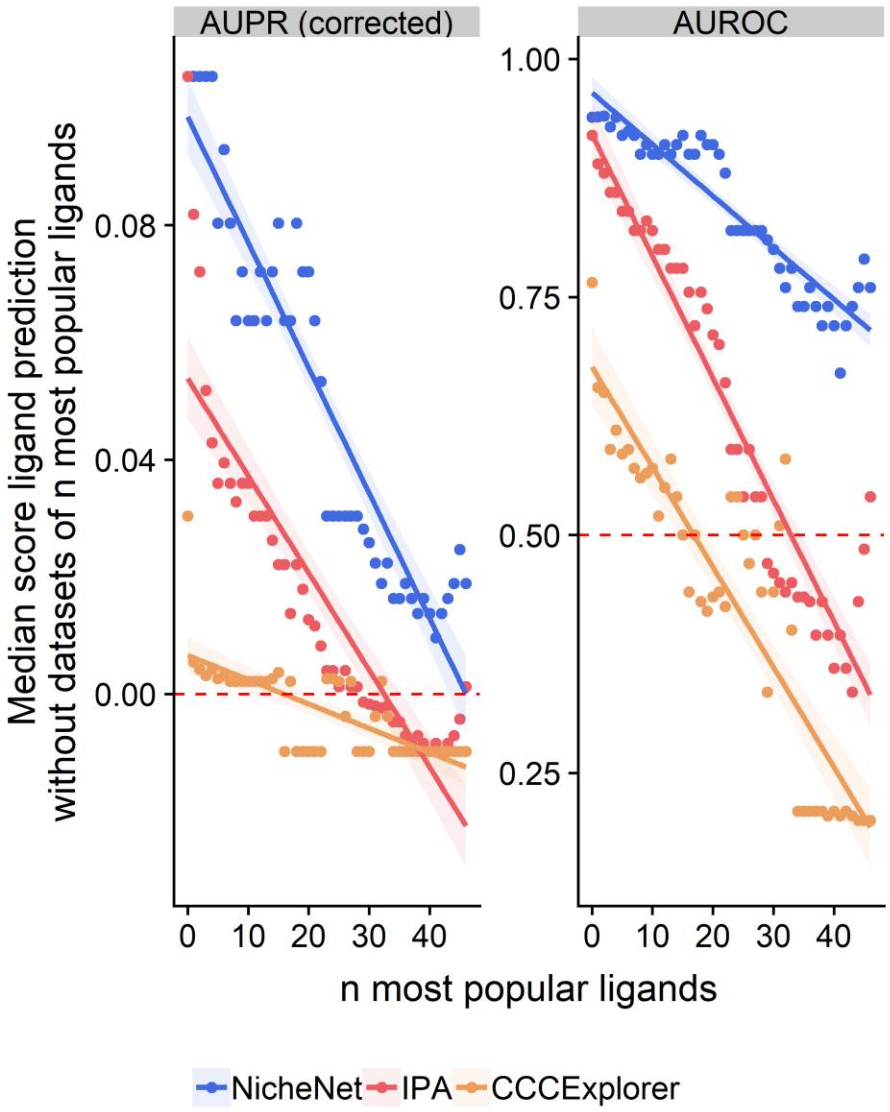
To evaluate the possible presence of a cell type bias, we assessed the predictive performance of NicheNet on ligand treatment datasets that were only selected out of the total set of 111 datasets if they measure the response to a ligand for which the response was profiled in at least two different cell types that were stimulated by at least two different ligands. For the 43 selected ligand treatment expression datasets, we show the target gene (a-b) and ligand activity (c-d) prediction performances in function of the ligand added to the cells and the cell type that was stimulated. For each ligand-cell type combination, we also indicate the number of analyzed datasets. When multiple datasets were analyzed for a specific ligand-cell type combination, the average performance is shown on the heatmap. AUROC: area under the receiver operating characteristic curve; AUPR (corrected): area under the precision-recall curve, corrected for random prediction.



Supplementary Figure 12

Ligand rankings in ligand activity prediction: comparison between NicheNet, CCCExplorer and IPA Upstream Regulator Analysis

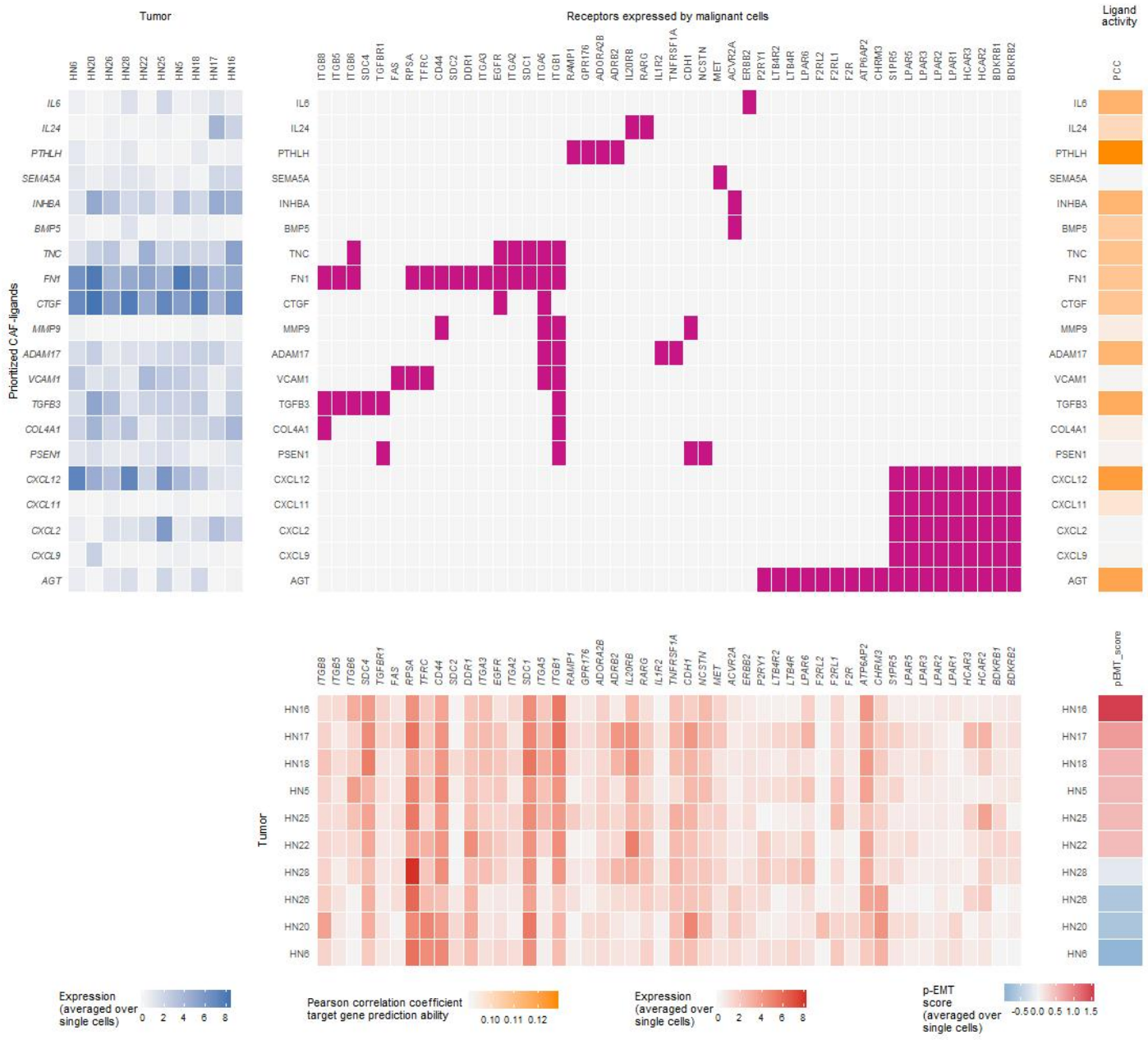
For 111 ligand treatment datasets (profiling the transcriptional response to one or two out of 51 different ligands), the 51 corresponding ligands were ranked by NicheNet according to how well they predict the transcriptional response observed in a particular dataset (as evaluated via the Pearson correlation); by IPA Upstream Regulator Analysis and CCCExplorer according to the negative logarithm of the enrichment p-value given by both methods. We plot here the fraction of datasets for which the true active ligand was present in the top n of the ranking of ligands in function of n.



Supplementary Figure 13

Popularity bias of NicheNet, CCCExplorer and IPA Upstream Regulator Analysis in ligand activity prediction performance

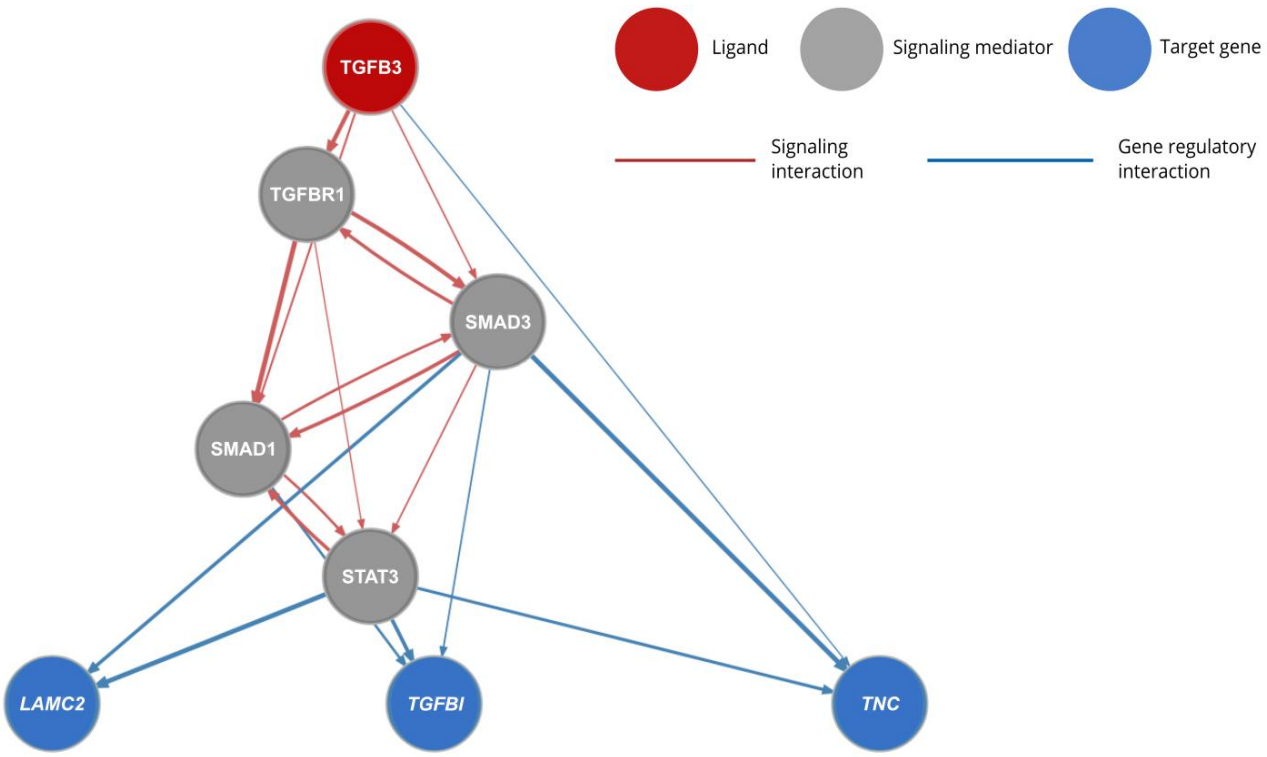
To analyze popularity bias in ligand activity prediction performance, we assessed performance after iteratively leaving out the ligand treatment datasets profiling the transcriptional response to the top n most popular ligands. Following ligand activity prediction evaluation metrics were calculated: AUPR (corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under the receiver operating characteristic curve. To analyze popularity bias and compare between NicheNet, CCCExplorer and IPA Upstream Regulator Analysis, we assessed the relation between popularity of ligands and median performance when datasets of these ligands are still included (n = 51 different ligands). The shown smoothing lines are the result of fitting a linear regression model to analyze this relation. The accompanying shaded region indicates the 95% confidence interval. The popularity of ligands is determined by the number of studies in PubMed in which a ligand is described.



Supplementary Figure 14

NicheNet can prioritize interactions between CAF-ligands and malignant cell receptors involved in the regulation of expression of p-EMT genes

Summary result of NicheNet analysis: results are shown for the 20 (out of 131) CAF-ligands best predicting the p-EMT gene set. Top left: scaled expression of ligands in CAFs, averaged per tumor. Top center: Ligand-receptor interactions between CAF-ligands and receptors strongly expressed in malignant cells. Top right: outcome of NicheNet's ligand activity prediction on the p-EMT gene set. As the ligand activity metric, we used the Pearson correlation coefficient (PCC) between regulatory potential scores and p-EMT gene set assignments (n = 6072 genes, of which 96 belong to the p-EMT program). Bottom center: scaled expression of receptors in malignant cells, averaged per tumor. Bottom right: each cell was scored for expression of genes belonging to the p-EMT program, and an average p-EMT score was calculated for every tumor.



Supplementary Figure 15

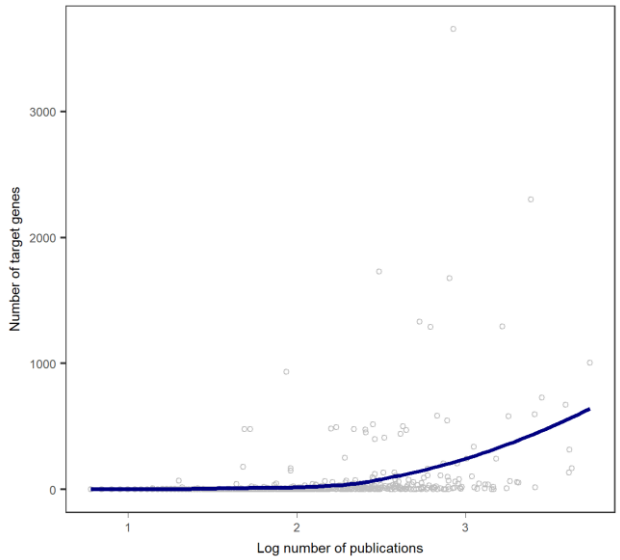
Network showing high-confidence interactions forming putative signaling paths between the ligand TGFB3 and its predicted target genes *TGFBI*, *LAMC2* and *TNC*

From NicheNet's weighted integrated signaling and gene regulatory networks, we inferred the most important signaling mediators and transcriptional regulators involved in the signaling from TGFB3 to *TGFBI*, *LAMC2*, and *TNC*. Shown here is the network of interactions (as documented by NicheNet's data sources) between ligand, signaling mediators, transcriptional regulators, and target genes. The ligand node is indicated in red, the target gene nodes in blue and nodes of signaling and transcriptional regulators in grey. Edge line thickness is proportional to the weight of the represented interaction in the weighted integrated networks. Edges representing signaling interactions are colored red, gene regulatory interactions blue.

Supplementary notes

Supplementary Note 1: data source collection

To obtain prior information on which ligands potentially influence the expression of which target genes, we did not only collect known ligand-target links from the literature but also predicted novel links. We considered this necessary because there is only a limited number of known direct ligand-target links, certainly for less-studied ligands **(Supplementary Notes Figure 1)**. Ligand-target predictions were made based on how most ligands exert their effect on gene expression in a receiving cell. Ligands first bind and activate a specific receptor on the plasma membrane of a receiver cell, and this is then followed by intracellular signaling that results in transcription factor activation and subsequent induction of target genes. Although this process is very complex, context-dependent and currently impossible to accurately model on a global scale, we hypothesized that it could be possible to predict plausible ligand-target links by collecting data sources that cover the gene-gene interactions occurring in ligand-to-target signaling (i.e., ligand-receptor, intracellular signaling, and gene regulatory interactions). We assumed that getting a comprehensive view on ligand-to-target signaling and obtaining a model that could be generally applicable to a variety of biological systems would require extensive data-integration of multiple informative and complementary data sources. For example, some data sources such as those offering curated interactions from the literature are expected to contain a small number of highly confident interactions, whereas others derived from high-throughput experimental methods are expected to be noisier, but covering a more substantial part of the whole genome. Moreover, some data sources will be more popularity biased than others, meaning that well-studied genes have more known links with other genes than less-studied genes – even though this is often not necessarily related to the biological reality[1,2] **(Supplementary Notes Figure 2)**. Interactions derived from different experimental technologies will thus differ substantially, and we expect that combining several different sources will provide a more comprehensive view.



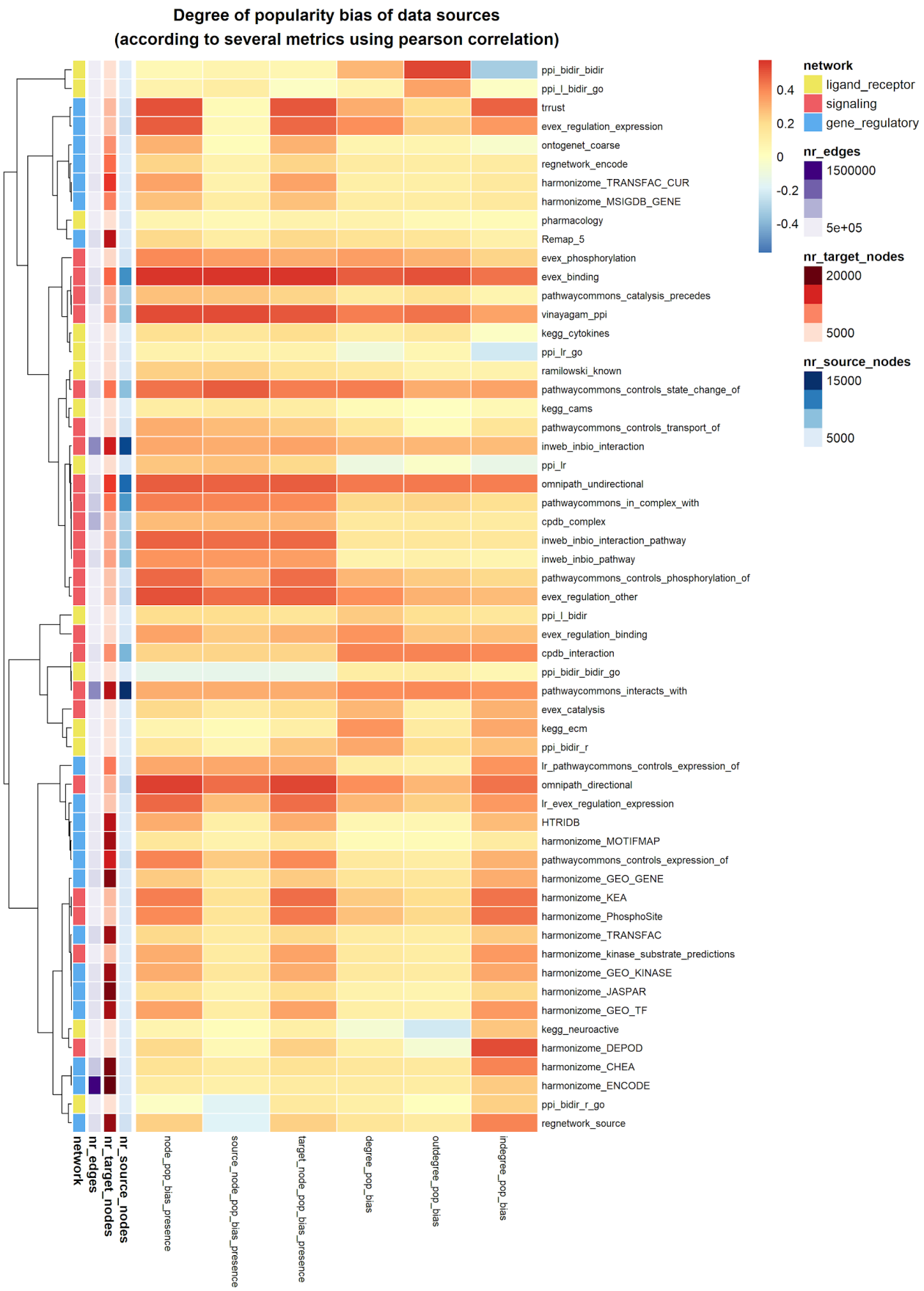
Supplementary Notes Figure 1 | Popularity bias present in known ligand-target links. For all ligands from NicheNet's ligand-receptor data sources, the number of known target genes, which were retrieved from pathway and text mining databases, is plotted versus the number of studies in PubMed in which a ligand is described (NCBI Entrez Gene2Pubmed database). The fitted smoothing line is the result of local regression (LOESS).

In this study, we collected data sources covering interactions between genes in human and mouse. Interactions between mouse genes were converted into interactions between the human one-to-one orthologs of these genes to obtain one final model. Because these organisms are well-studied, efforts were already made to create 1) databases that provide users with readily available interaction networks after processing raw experimental data and 2) databases that are an aggregation of several other data sources. Both substantially facilitated the collection of data sources required for this study.

For compiling ligand-receptor links, we collected known ligand-receptor interactions from curated databases and the literature. We also predicted some ligand-receptor interactions based on protein-protein interactions between genes functionally annotated as ligands and receptors. To link receptors to transcriptional regulators, we collected protein-protein interactions (PPI) and signaling interactions contained in aggregated databases, and complemented them with a directional PPI network, interactions inferred by text mining and interactions from post-translational modification databases. Finally, to obtain links between regulators and target genes, we combined several pathway databases, text-mining based data sources, data sources containing links inferred from expression data, motif databases, ChIP-seq, and regulator perturbation experiments. Note that some data sources, such as text mining and perturbation experiments, also contained direct links between ligands/receptors and target genes.

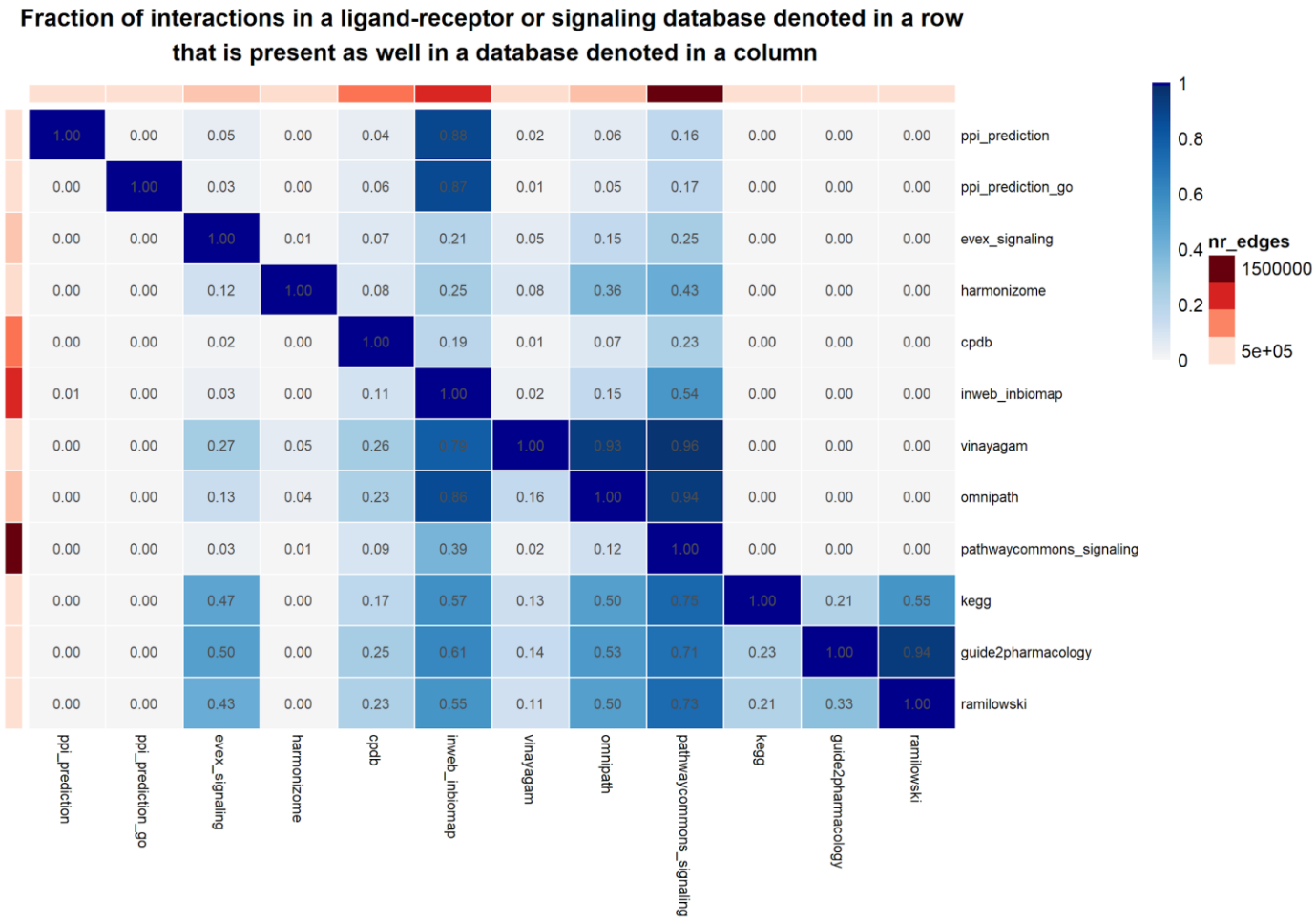
On this topic of data source collection, two critical remarks should be made. A first remark is that we ignore information regarding whether a signaling or gene regulatory interaction is activating or inhibitory - an important limitation but difficult to overcome in the framework proposed in this study. A second remark is that we divided aggregated databases according to their underlying data source or nature of interactions (data source/type) if possible, because different interaction types might have different information content. For example, an aggregated pathway database can contain non-directional protein-protein interactions and directional interactions like kinase-substrate interactions. Because kinase-substrate interactions might be more relevant for ligand-to-target signaling than common protein-protein interactions, we considered these as separate data sources such that they both can have a different impact on the final model after data source weight optimization (see **Supplementary Note 2**).

An overview of the databases collected in this study and their division into data sources/types can be found in **Supplementary Table 1a-c**. In this table and **Supplementary Notes Figure 2**, network properties such as the number of genes, the number of interactions, and several metrics related to the population bias are documented as well. An analysis of the overlap between databases **(Supplementary Notes Figure 3-4)** and data sources **(Supplementary Notes Figure 5-6)** and the relation between the number of interactions and the number of data sources supporting an interaction **(Supplementary Notes Figure 7)** indicate that collection of multiple complementary data sources resulted in substantially higher coverage, as was the objective. Since many different databases retrieve interactions from the same experiments, redundancy between some data sources could be observed as well. A detailed explanation of the preprocessing of the several data sources is described in **Supplementary Note 13.**

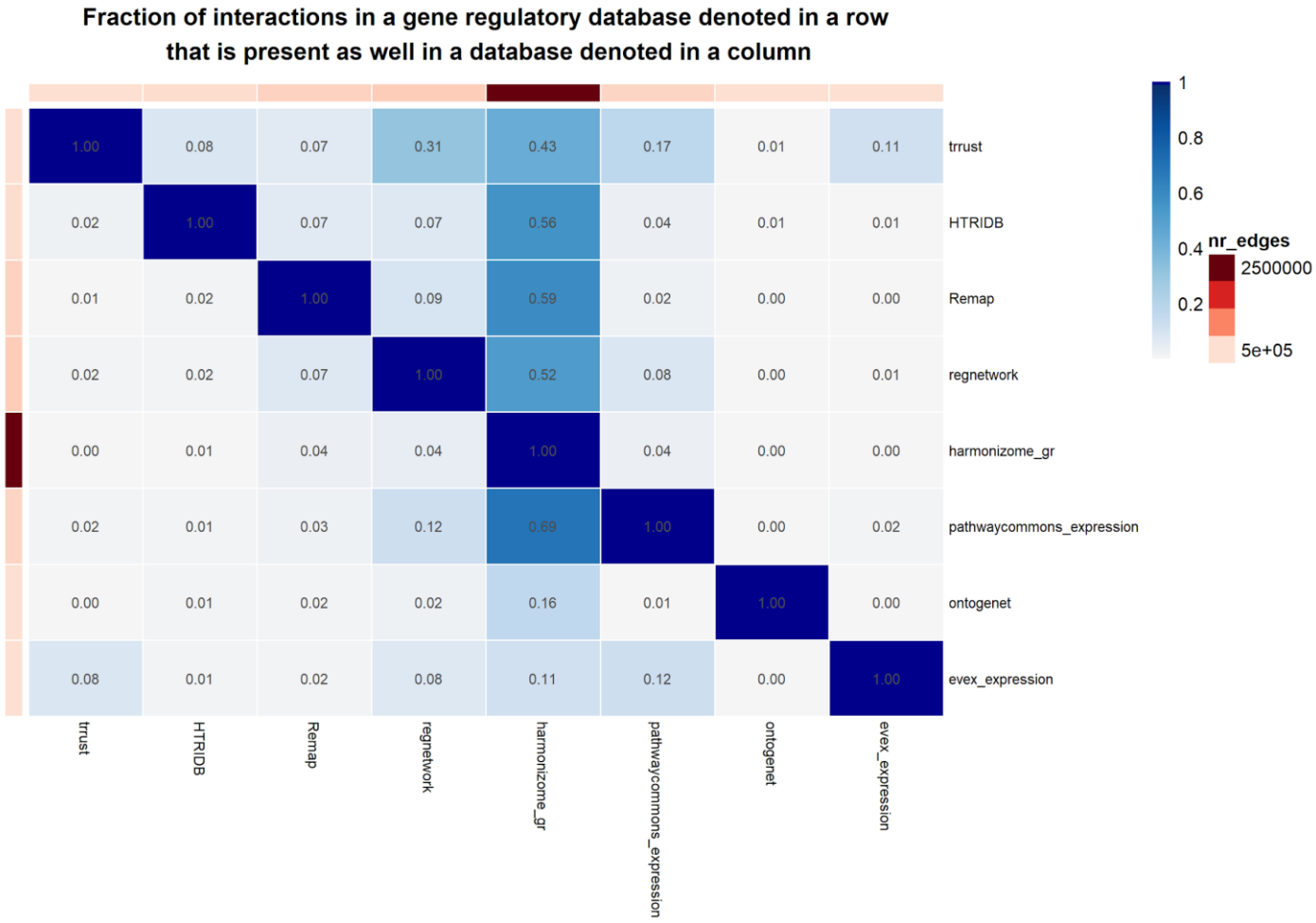


(Legend next page)

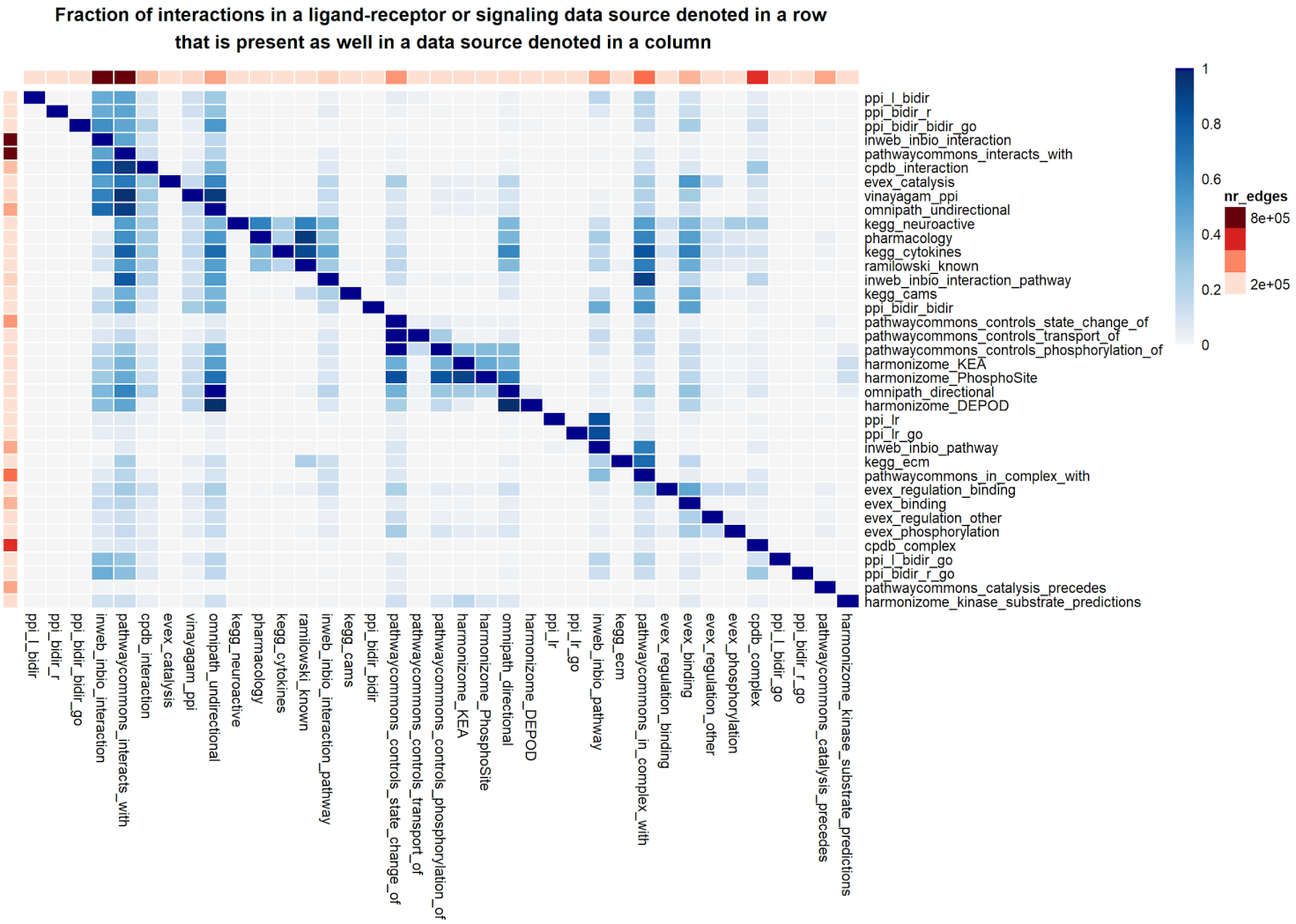
Supplementary Notes Figure 2 | Popularity bias in collected data sources. The blue-red color gradient scale denotes the popularity bias of data sources as calculated by several measures. "Node_pop_bias_presence", "source_node_pop_bias_presence" and "target_node_pop_bias_presence": indication of whether more frequently studied genes are more likely to be documented in a specific data source. For this, we calculated the Pearson correlation between the log number of studies in Pubmed in which a gene is described and whether the gene is documented in a particular data source or not (respectively all nodes, only source nodes, and only target nodes were considered). "Degree_pop_bias", "indegree_pop_bias", "outdegree_pop_bias": indication of whether more frequently studied genes are more likely to have more interactions in a specific data source. For this, we calculated the Pearson correlation between the log number of studies in Pubmed in which a gene is described and respectively the overall degree, indegree, and outdegree of a gene. Dendrogram is the result of hierarchical clustering of the data sources (n = 57) with complete linkage and Pearson correlation as clustering distance. The exact number of total nodes, source nodes, and target nodes of each individual data source can be found in Supplementary Table 1b.



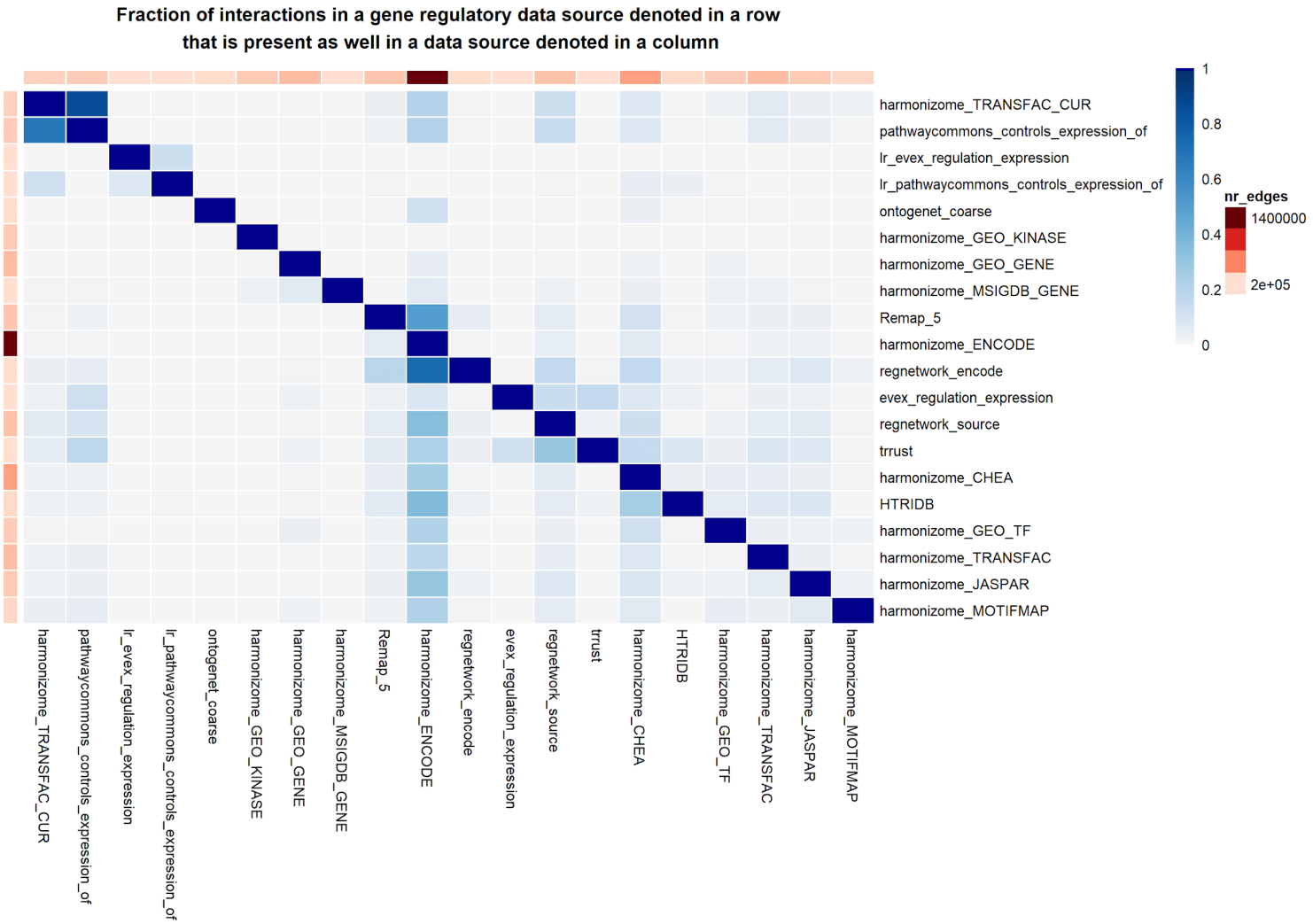
Supplementary Notes Figure 3 | Overlap in interactions between the databases providing ligand-receptor and signaling interactions. The fraction of interactions in a database denoted in a row that is present as well in a database denoted in a column (blue color scale).



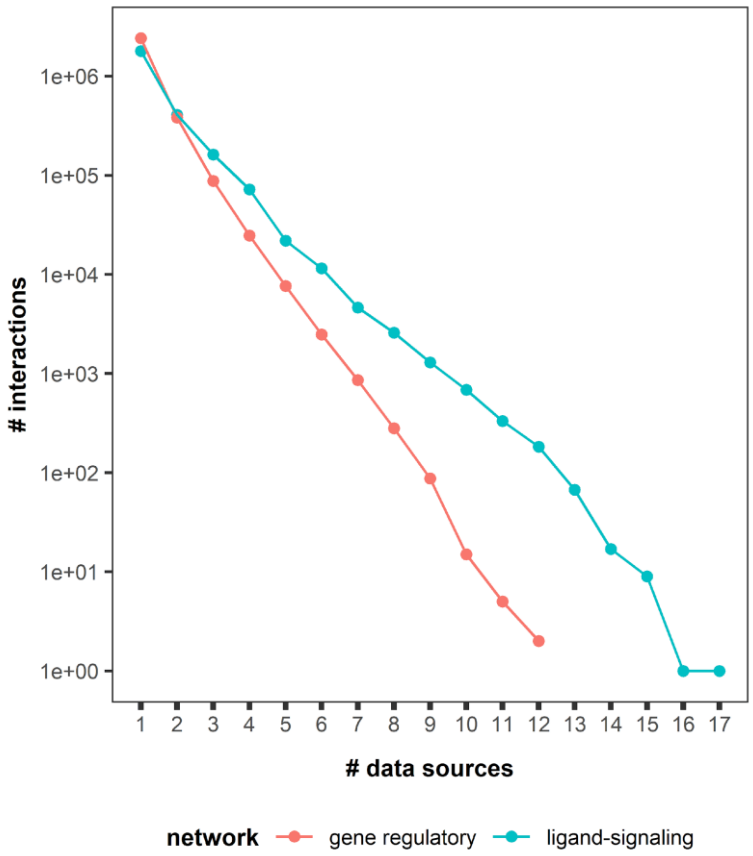
Supplementary Notes Figure 4 | Overlap in interactions between the databases providing gene regulatory interactions. The fraction of interactions in a database denoted in a row that is present as well in a database denoted in a column (blue color scale).



Supplementary Notes Figure 5 | Overlap in interactions between the data sources providing ligand-receptor and signaling interactions. The fraction of interactions in a data source denoted in a row that is present as well in a data source denoted in a column (blue color scale).



Supplementary Notes Figure 6 | Overlap in interactions between the data sources providing gene regulatory interactions. The fraction of interactions in a data source denoted in a row that is present as well in a data source denoted in a column (blue color scale).



Supplementary Notes Figure 7 | The majority of interactions are documented by only one or a few data sources. The number of interactions (on a log scale) is plotted versus the number of data sources that support these interactions. In total, there are 20 gene regulatory data sources and 37 ligand signaling data sources.

Supplementary Note 2: weighted aggregation of data sources

All collected ligand-receptor and signaling data sources were combined in a directed weighted ligand-signaling network and all gene regulatory data sources in a directed weighted gene regulatory network. Having these two different integrated networks instead of one network is necessary because the edges represent a different type of interaction. Namely, a protein-protein interaction in the ligand-signaling network, and a gene regulatory interaction in the gene regulatory network.

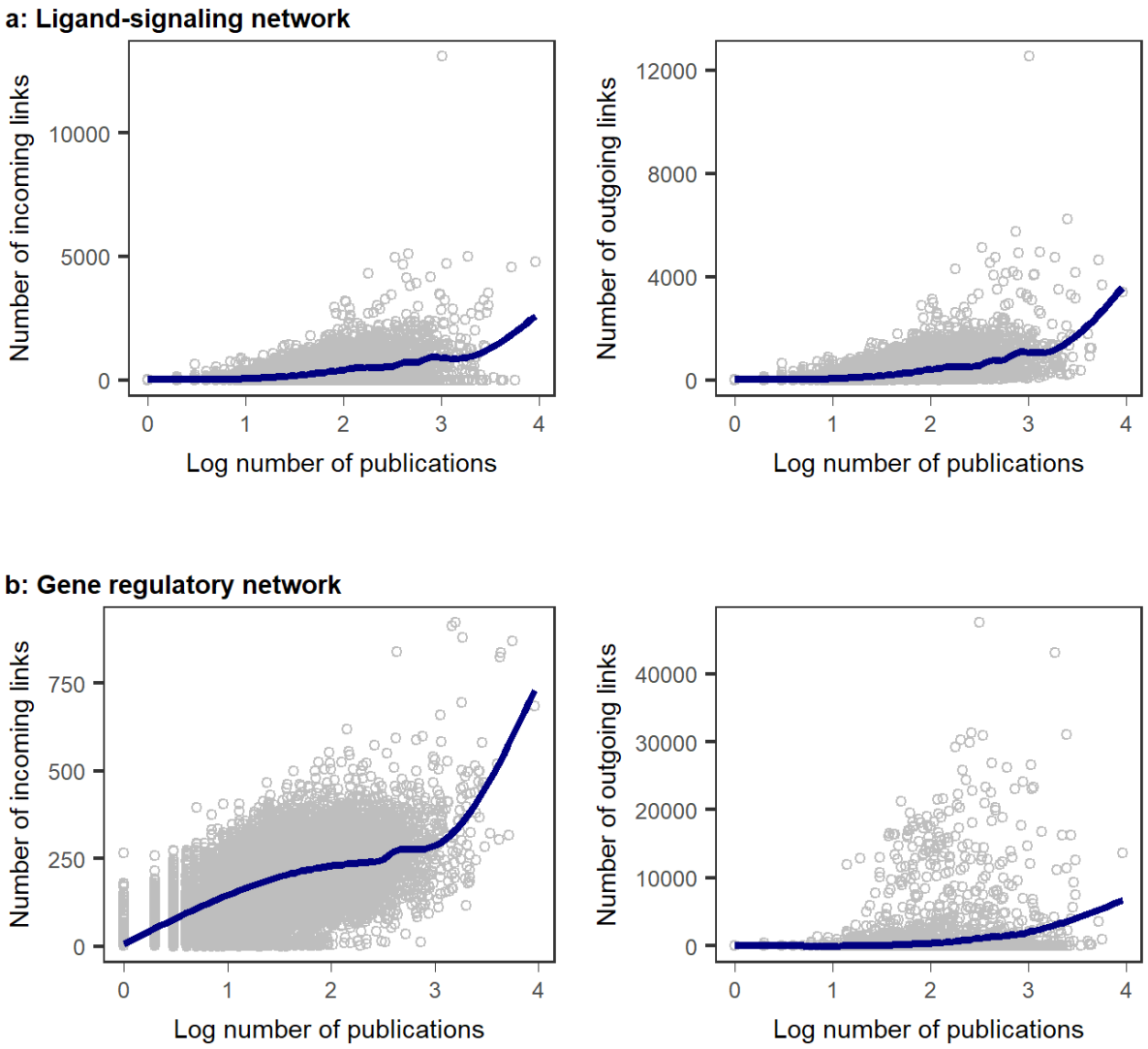
The weight of an edge between a source and target node represents a score related to the amount of evidence that the source node signals to the target node. In the most simple, unoptimized setting, the networks of all individual sources are just aggregated to form an integrated network in which the weight of an edge is calculated as the number of data sources supporting that interaction. In this setting, every single data source contributes equally to the final model. However, interactions from some data sources are possibly more informative than others, due to differences in information content, amount of false positives, popularity biases or coverage. To account for this, we assigned a weight to each data source according to its contribution to the model performance. After definition of the weights by automatic parameter optimization, the integrated networks are constructed by a weighted aggregation of the source-specific networks; every edge in the integrated layer-specific networks receives thus a weight that is the sum of the weights assigned to the data sources in which this interaction is documented. A primary advantage as a consequence is that noisy non-informative data sources can receive a weight of zero and will then be removed from the model.

We chose for this approach of data source weight optimization to increase final model performance over alternative approaches that define gene-gene interaction weights based on the accuracy by which data sources provide true interactions (for example via probabilistic integration as described by Lee et al[3]). We made this choice because the main purpose of this study is to have optimal ligand-target predictions and not necessarily optimal gene-gene interactions in the underlying integrated networks. For example, a less accurate kinase-substrate data source might be more informative to link ligands to target genes than a very accurate data source providing protein-protein interactions. This is because kinase-substrate interactions are crucial in the ligand-to-target signaling of many ligands, and many common protein-protein interactions are involved in other processes not related to ligand-to-target signaling. However, a noteworthy disadvantage of our approach is that biases present in the evaluation of the model performance might lead to biases in the final model. For this study, this means that data source weight optimization could lead to an increase in popularity bias because predictions of almost only well-studied ligands could be evaluated.

Data source weights were determined via multi-objective parameter optimization. More specifically, we optimized both the AUROC and AUPR of both the target gene and ligand activity prediction performance **(Online Methods)**. For this optimization problem, we were confronted with several challenges: 1) a large number of parameters: 57 data source weights and 4 hyperparameters **(Online Methods)**, resulting in a huge combinatorial problem; 2) interdependence of different data source weights and hyperparameters; 3) the model cannot be formulated or approximated as a derivable function such that for every parameter setting the model needs to be constructed and subsequently evaluated, which requires 20-30 minutes. To overcome all these challenges, we applied mlrMBO[4], a framework for state-of-the-art model-based optimization, which addresses expensive black-box optimization problems by approximating the objective function through a surrogate regression model.

Of note, the proposed weighted aggregation and optimization strategy could also be useful for other bioinformatics methodologies generating predictions based on an integrated network constructed from several data sources.

In **Supplementary Table 1d**, network properties such as the number of genes and the number of interactions in the integrated weighted networks are documented. In **Supplementary Notes Figure 8**, the popularity bias present in the integrated weighted networks is shown.



Supplementary Notes Figure 8 | Popularity bias present in the integrated weighted networks of NicheNet. To assess the popularity bias in the ligand signaling (a) and gene regulatory network (b), we plotted the number of incoming and outgoing interactions of every gene in each network versus the number of times the gene is mentioned in PubMed. The smoothing lines are the result of fitting a generalized additive model (GAM).

Supplementary Note 3: calculation of ligand-target regulatory potential scores

After inference of the weighted integrated ligand signaling and gene regulatory networks, we combined these networks into a single model to infer regulatory potential scores between ligands and target genes. Calculating ligand-target regulatory potential scores was based on the idea of propagation of the signal from a ligand to its receptor(s), downstream signaling proteins and finally transcriptional regulators and their target genes. Because the regulation of target genes by a ligand can be abstracted as a two-step process, namely a signaling phase from ligands to transcriptional regulators and a gene regulatory phase from regulators to target genes, we first linked ligands to regulators and then these regulators to target genes.

Linking ligands to transcriptional regulators

The ligand-signaling network was used to give for a particular ligand, a signaling importance score for every "downstream" protein. This score corresponds to the importance of a protein in the signaling path that starts from the particular ligand. The idea is that the ligand itself gets the highest score, followed by its receptors, which are followed by the signaling proteins and transcriptional regulators activated by these receptors. To calculate this ligand-regulator matrix, we did not simply use the adjacency matrix of the weighted ligand-signaling network, because direct links between many ligands and transcriptional regulators are not known and documented in pathway databases (except for well-studied ligands). Therefore, ligand-regulator scores should be inferred by graph algorithms that provide a measure related to the potential that a ligand signals to a regulator.

One possible algorithm would be the shortest path length (SPL), because the shorter the path between a ligand and a regulator, the more likely it can be considered that the ligand might signal to this regulator. Because it is argued that SPL may not be very informative for biological networks because the shortest path between two nodes is commonly short as a result of the presence of hubs (i.e., nodes with many links)[5], we applied Google's Personalized PageRank (PPR), a random-walk based algorithm. Whereas the traditional "global" PageRank is a measure to rank web pages based on their importance as inferred from the global linkage structure of the web, the PPR was developed to incorporate personal preferences of a specific web user as well[6]. The idea of the PageRank measure to define node importances is the following. Consider a random walk on a graph. This walker, starting at a uniformly chosen node, continues its walk along one of the outgoing edges of the current node with probability *d* (known as the "damping factor") whereas the walker "teleports" with probability *1-d* to a random node of the graph (uniformly selected). It should be noted that for a weighted network, the probability that an outgoing edge is selected during the walk will not be uniformly distributed but dependent on the edge weight. The PageRank of a node *v* is then the steady-state probability that the walker is at the node *v.* The higher this PageRank value, the more important the node *v* is in relation to the linkage structure of the graph. In PPR this idea is extended to take into account "personal preference". Here, the random walker will teleport with probability *1-d* to a random node of the graph, but this occurs not uniformly such that some preferred nodes, called seed nodes, will be more likely selected than other nodes. Because the random walker teleports to the seed node(s) with higher probability than to the other nodes, the steady-state probability will be higher for the seed node(s) and the nodes in the neighborhood of the seed node(s), resulting in a higher PPR importance score for these. By taking a ligand of interest as seed node, the ligand itself, followed by regulators closely downstream of a ligand will get higher importance scores than regulators located less close in the network. Conceptually, this is similar to the following: a signal will diffuse in the signaling network starting from a ligand, and the further a signal has to diffuse until it reaches a transcriptional regulator, the lower the probability that

this ligand activates this regulator. An important property of the PPR measure is that its computation is very scalable because it can be well approximated even for very large networks[7].

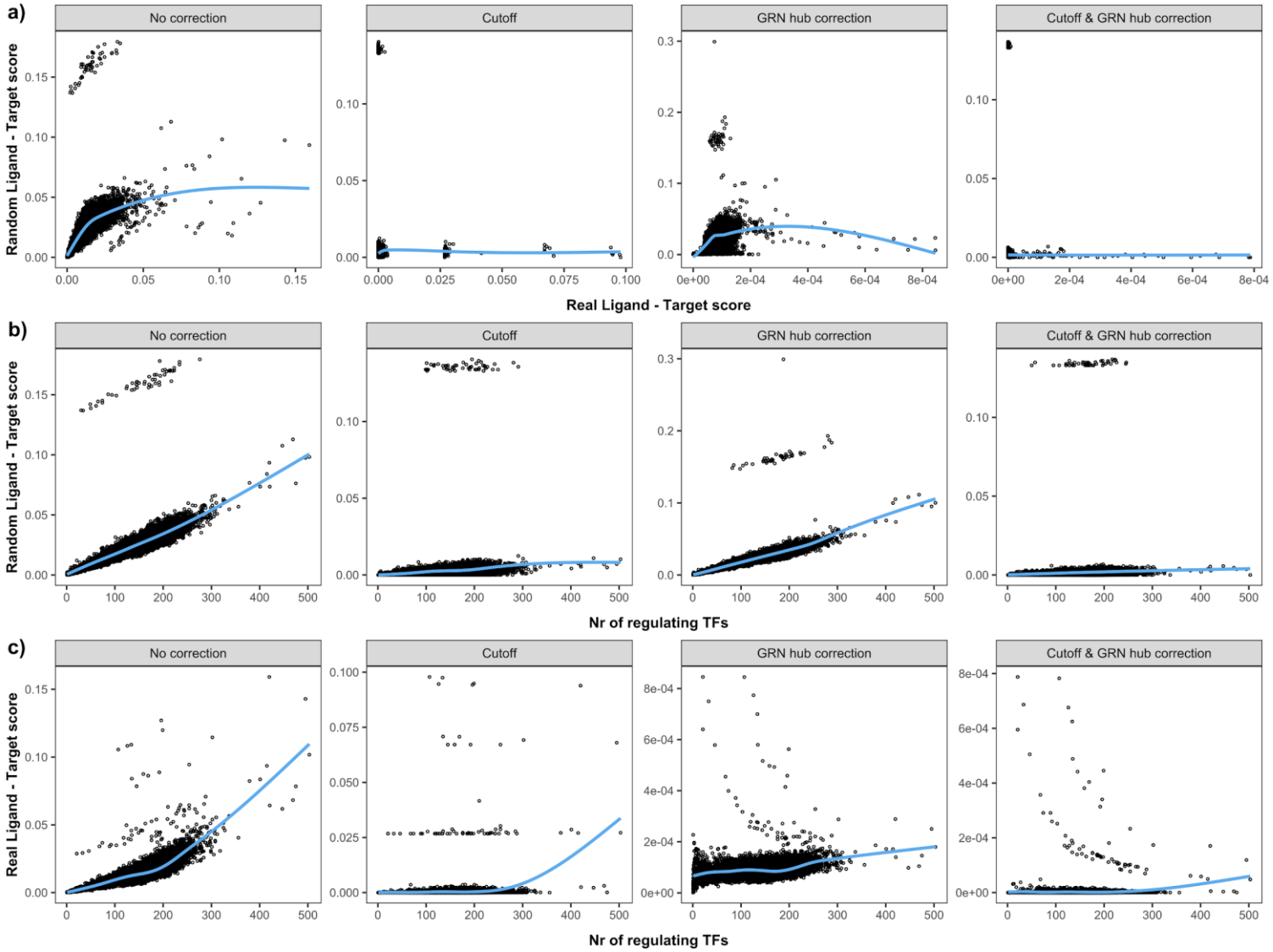
Linking transcriptional regulators to target genes

The previous paragraph described how we propagated the signal from ligands to downstream transcriptional regulators. To further propagate the signal from transcriptional regulators to their target genes, the ligand-regulator matrix was multiplied with the regulator-target matrix of the gene regulatory network (i.e., the weighted adjacency matrix). As a result of this, known target genes of the ligand itself, its receptors and closely downstream regulators will get high ligand-target regulatory potential scores because they are more likely to be regulated by the ligand than target genes of regulators that are not activated upon signaling of this ligand. This approach thus guarantees that the highest scoring target genes of a ligand are genes that are both a known target of a ligand (inferred by text mining) and target of transcription regulators to which the ligand can signal (inferred by signaling data sources).

Additional note about the use of PPR

A consequence of using the PPR algorithm is that regulators not close to the ligand receive a minimal score. However, this score is not zero as should be for regulators not part of the signaling pathways activated by a ligand. A disadvantage of this is that genes that are regulated by many TFs (i.e., "hub" target genes in the gene regulatory network), even if these are not closely downstream of the ligand of interest, can still receive high ligand-target regulator potential scores for that ligand as a consequence of the applied matrix multiplication. In **Supplementary Notes Figure 9**, it can be observed that target genes with a lot of incoming gene regulatory links tend to have very high ligand-target regulatory potential scores, both for true networks and networks randomized by node shuffling to preserve node degree. In an attempt to mitigate this problem, we applied a cutoff on the ligand-regulator scores inferred via PPR such that only closeby downstream regulators with an importance score larger or equal than the defined cutoff receive a score higher than 0. The cutoff we used, is the score larger or equal than a specific quantile between the 0.9 and 0.999 quantiles, as determined via automatic parameter optimization **(Online Methods)**. As an additional solution, we applied a gene regulatory network hub correction factor to decrease the weight of gene regulatory interactions proportionally to the indegree of the target gene node **(Online Methods)**.

To motivate the choice of PPR in NicheNet, we compared the performance of different variants of NicheNet: the default variant PPR, SPL, an adapted version of the adjacency matrix of the weighted ligand-signaling network and a model only considering direct targets (**Supplementary Note 5**).



Supplementary Notes Figure 9 | Relation between target gene score and the number of TFs regulating the target gene without and with correction. For the ligand CSF1, we plot the ligand-target regulatory potential scores derived from NicheNet's true weighted integrated networks and networks randomized by edge swapping against each other and against the number of incoming gene regulatory interactions of a target gene (n target genes = 12701). Cutoff: correction by applying a cutoff on the ligand-regulator matrix: genes with a signaling importance score smaller than the 0.999 quantile receive an importance score of 0 before calculating ligand-target regulatory potential scores. GRN hub correction: correction by decreasing the weights of gene regulatory interactions by dividing them by the indegree of the target gene node. The smoothing lines are the result of fitting a generalized additive model (GAM).

Supplementary Note 4: limitations of using ligand treatment expression datasets for validation

A first significant drawback of using publicly available ligand treatment expression data is that the nature of ligands for which ligand treatment datasets are available is biased: there is an overrepresentation of secreted ligands over membrane-bound ligands and an overrepresentation of well-studied ligands such as cytokines and growth factors over less-studied ligands. This latter popularity bias in the set of validated ligands is shown in **Supplementary Figure 3a**, where one can see a clear enrichment of datasets of ligands that are most frequently mentioned in the literature. All conclusions drawn from the evaluation should thus be considered with caution since they might not be or only partially be valid for less-studied ligands.

As can be seen in the overview tables of the used ligand treatment datasets in **Supplementary Table 2a**, the time points after which the response of the cells to the ligand stimulation is measured, varies substantially. In some studies, only the early response to the ligand is measured, whereas in other studies also more long-term effects are assessed. Although almost exclusively differential expression of the primary target genes (i.e., genes directly regulated by the ligand) may be expected in the former, a lot of secondary targets (i.e., target genes of TFs that are differentially expressed as a direct response to a ligand) may be expected in the latter. Because ligand-target regulatory potential scores are inferred for primary target genes, the presence of secondary targets among the differentially expressed genes is another limitation of this validation procedure.

A next main limitation of this validation approach is that the response to extracellular signals depends on cell type and state. This context-specificity is not taken into account in the general model of NicheNet. Some genes might, for example, be a real target of a ligand in specific cell types or conditions, but not in others due to closed chromatin state and/or absence of crucial signaling and transcriptional regulators. Hence, some non-affected genes with high regulatory potential are seen as false positives in the evaluation, but might actually be true target genes that are just not affected due to context-specific reasons. In **Supplementary Table 2c-d** and **Supplementary Figure 3b**, it is shown that the transcriptional response to the same ligand in different cell types can vary heavily, resulting in a limited fraction of genes that for the same ligand are differentially expressed in multiple conditions. For example, we observed that less than 25% of affected genes overlap between two datasets that profile the response to the same ligand.

However, we still think that the ligand treatment datasets provide a useful gold standard of ligand-target links because you know that differentially expressed genes are differentially expressed only because of the effect of the ligand on the target cell. Even though not all possible targets of a ligand will be affected in a specific condition and not all differentially expressed genes are primary targets, the set of differentially expressed genes is - to our knowledge - the best available set of experimentally determined target genes of a ligand. Nevertheless, the limitations of using these ligand treatment datasets as gold standard should be taken into account when interpreting the results of the validation study.

Supplementary Note 5: comparison of different possible variants of NicheNet

As discussed in **Supplementary Note 3**, there are several other possible ways to infer ligand-gene signaling importance scores in addition to PPR. Therefore, we compared the performance of NicheNet when using PPR against alternative approaches to link ligands to downstream signaling regulators.

A first alternative approach is to calculate the shortest path length (SPL). Over the weighted ligand-signaling network, the shortest path was defined as the path of which the summed weight of the edges is minimal, after inverting the edge weights such that more informative interactions get lower weights. The SPL distances between the ligand and the other nodes were calculated by applying Dijkstra's algorithm for weighted networks[8]. Because we wanted to assign higher importance scores to genes with shorter shortest path lengths, we subtracted the shortest path length between a ligand and a gene from the maximal observed shortest path length. Similar to PPR, a cutoff on the SPL scores was applied.

A second alternative approach is to allow at most one intermediate step between a ligand and a target gene (instead of propagating the signal multiple steps through the signaling network). This was implemented by simply taking the adjacency matrix of the weighted ligand-signaling network and giving the ligand the maximal score of all ligand-gene links as signaling importance score because the ligand needs to have the highest signaling importance score.

A third alternative approach is to use only direct known regulatory links between ligands and targets. This was implemented by setting the damping factor of the PPR to zero.

In **Supplementary Notes Figures 10 and 11**, we compare respectively the target gene prediction performance and the popularity bias in this performance between the different approaches. In **Supplementary Notes Figures 12 and 13**, we compare the performance and popularity bias of ligand activity prediction.

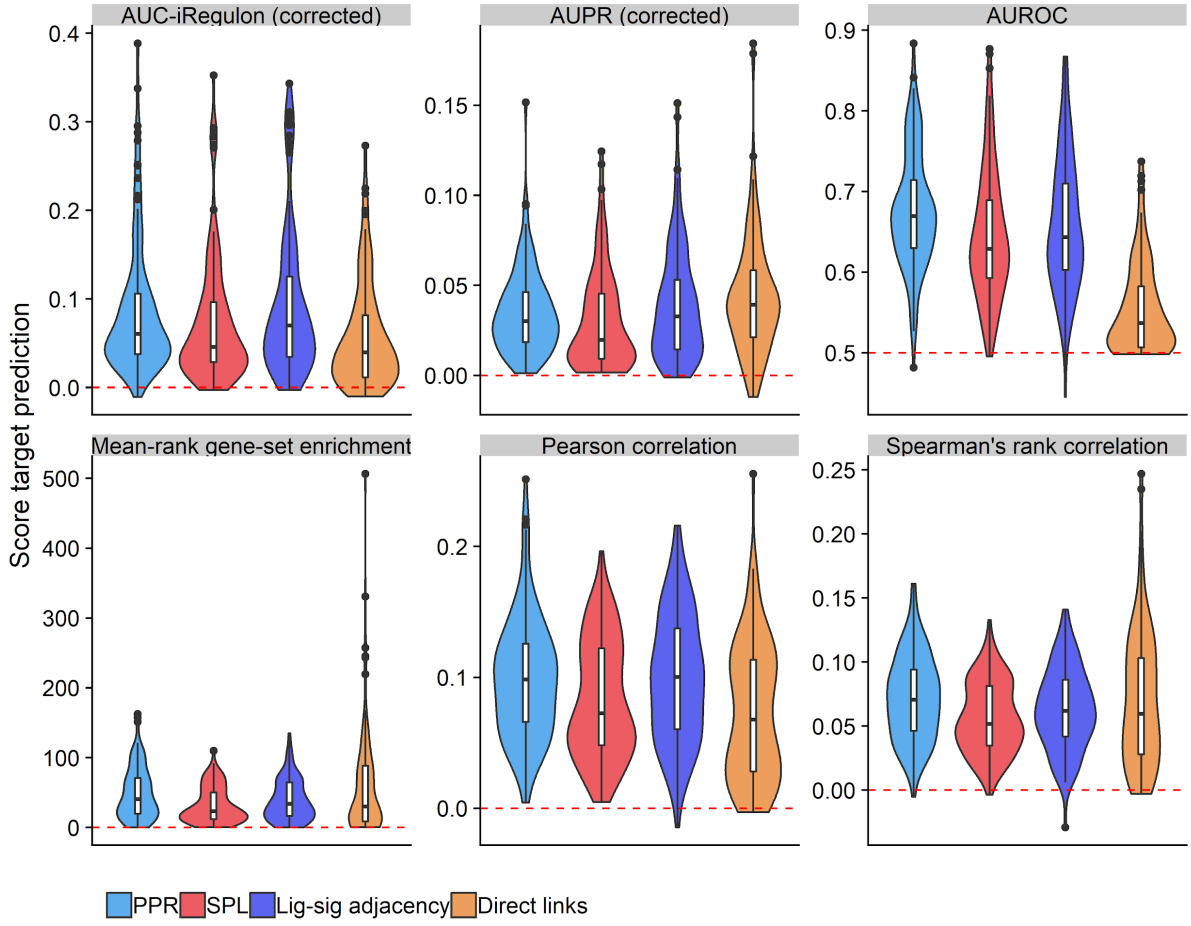
We found that using PPR outperforms SPL in both target gene and ligand activity prediction, supporting the choice of PPR over SPL.

Interestingly, the use of only direct known links between ligands and targets gives similar accuracy in ligand activity prediction compared to PPR, although target gene prediction accuracy is lower. The ligand activity prediction performance when using only direct links is probably high because the presence of a few well-known targets of a ligand within a gene set of interest can often be informative enough to predict the correct active ligand. This high performance indicates thus that using only direct known links could be sufficient for ligand prioritization on expression data of interacting cells. However, we expect that a much smaller number of affected targets of these prioritized ligands would be predicted because of the limited number of known direct targets (compared to complementing known direct targets with predicted targets based on signal propagation, as in PPR). The lower target gene prediction accuracy indicates this as well. Another disadvantage of using only direct links compared to PPR is the popularity bias. Because these direct links were extracted from literature and pathway databases, well-studied ligands tend to have more direct known target genes than less-studied ligands (**Supplementary Notes Figure 1**). As could be expected as a consequence of this, ligand activity prediction performance for the direct-links model seems to be more popularity biased than for the PPR model (**Supplementary Notes Figure 13**).

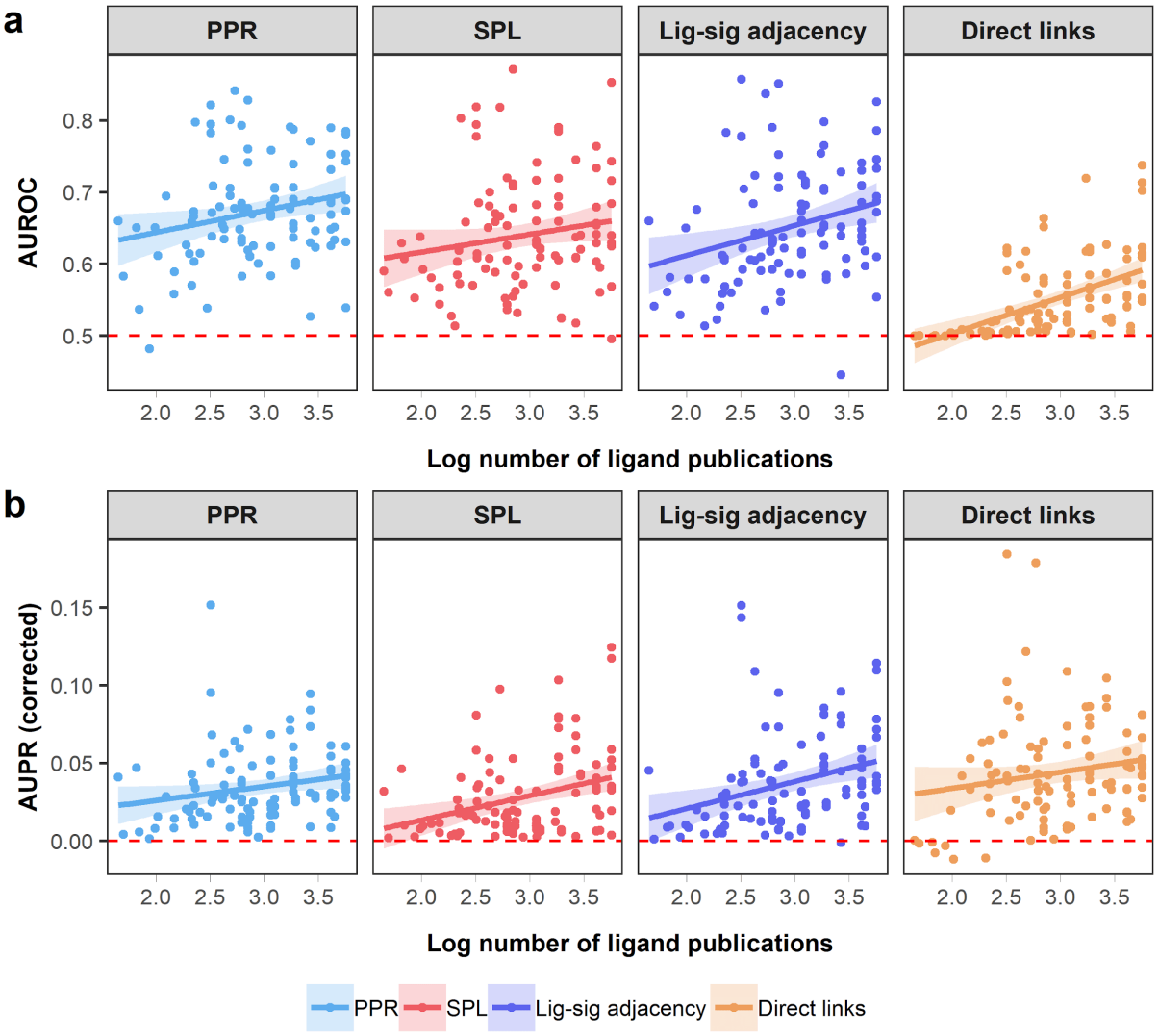
Using the adapted adjacency matrix of the weighted ligand-signaling network surprisingly resulted in performances that were in general very similar to using PPR (only slightly lower AUROC in target and ligand prediction performance and a bit more popularity biased in ligand prediction). One possible

reason for this could be that several pathway databases provide direct ligand-downstream regulator links for relatively well-studied ligands (making network propagation not necessary) and mainly only well-studied ligands were evaluated (**Supplementary Note 4**).

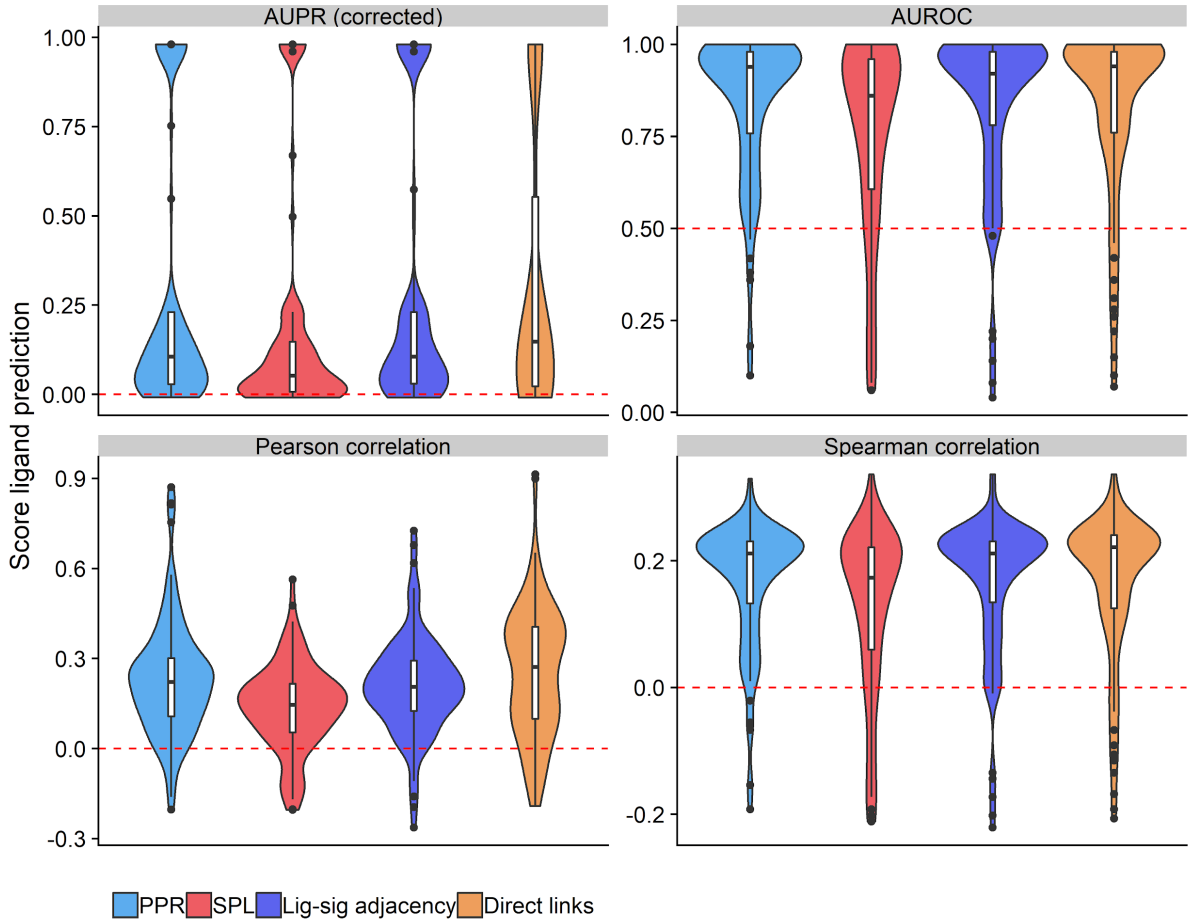
In conclusion, PPR was chosen as the final approach because it consistently resulted in the best target and ligand prediction performances overall, while being less popularity biased as well.



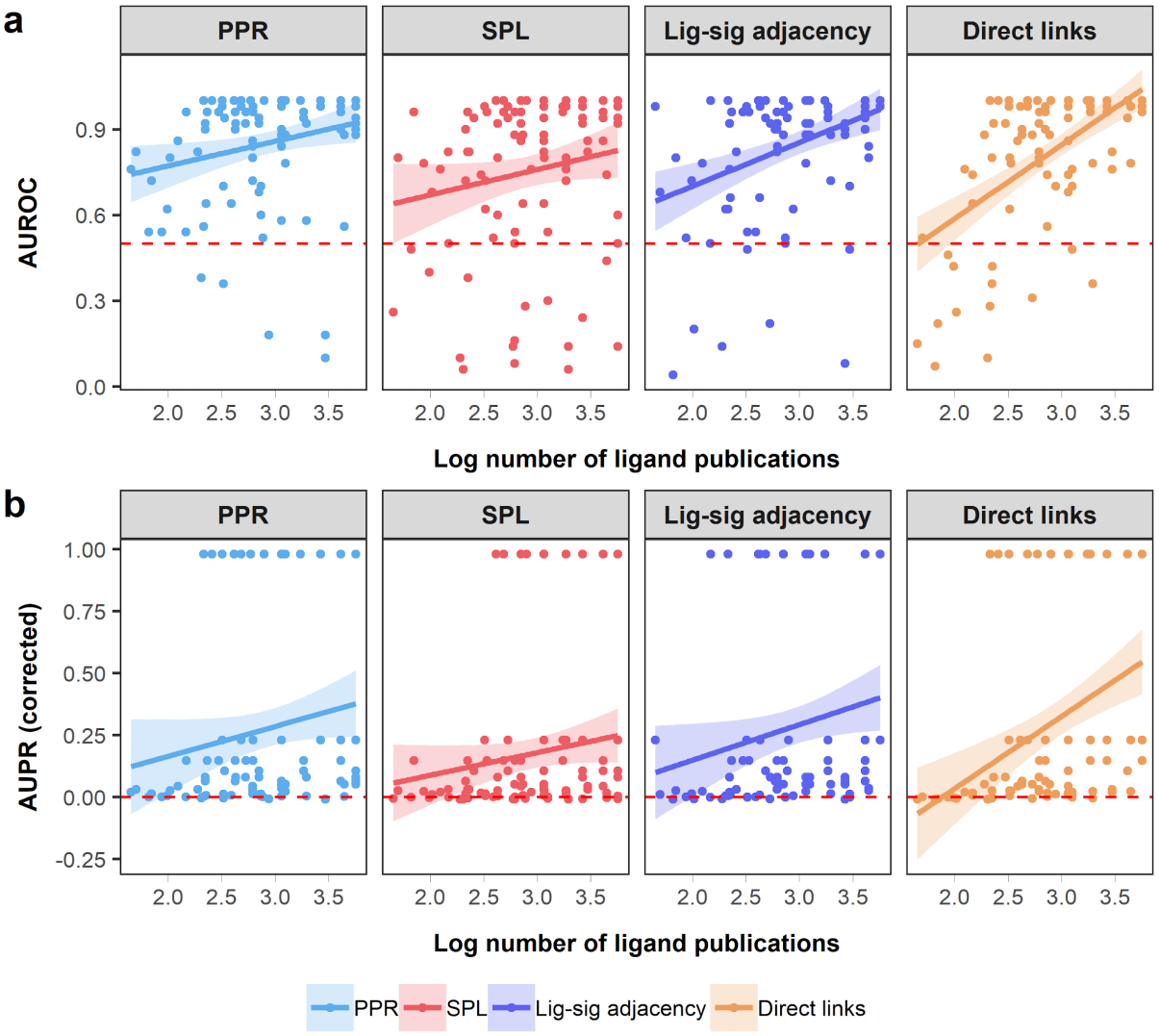
Supplementary Notes Figure 10 | Evaluation of the performance of NicheNet in predicting the transcriptional response: comparison to alternative approaches. To validate the use of Personalized PageRank (PPR) in the NicheNet methodology to calculate ligand-gene signaling importance sores, we compared it against alternative approaches. A first alternative approach was applying the shortest path length algorithm (SPL) and assigning higher importance scores to genes with shorter shortest path lengths. A second approach consisted of simply taking the adjacency matrix of the weighted ligand-signaling network (Lig-sig adjacency). A third approach uses PPR with a damping factor of 0 such that in the final ligand-target model, only direct ligand-target links are included. For 111 ligand treatment expression datasets, we calculated several classification evaluation metrics to compare these different variants of NicheNet in target gene prediction performance. All models constructed by these alternative approaches were optimized in the same way as the main variant of NicheNet. AUC-iRegulon (corrected): area under the cumulative recovery curve (considering the top 3% of the ranking), corrected for random prediction; AUPR (corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under receiver operating characteristic curve; Mean-rank gene-set enrichment: negative natural logarithm of the mean-rank gene-set enrichment p-value. Boxplot elements: center line: median; box limits: upper and lower quartiles; whiskers: 1.5x interquartile range; points: outliers. The violin plots show the probability density of the data (tails of the violins are trimmed to the range of the data). Red dashed line: performance of random guessing.



Supplementary Notes Figure 11 | Popularity bias of NicheNet in predicting the transcriptional response to ligands in ligand treatment datasets: comparison to alternative approaches. For 101 ligand treatment expression datasets profiling the transcriptional response to a single ligand, we plotted the performance of several variants of NicheNet in predicting the transcriptional response to a particular ligand in function of its popularity (the decimal logarithm of the number of studies in PubMed in which a ligand was described). Following variants to calculate ligand-gene signaling importance scores were compared: 1) the default NicheNet approach of applying Personalized PageRank (PPR); 2) applying the shortest path length algorithm (SPL) and assigning higher importance scores to genes with shorter shortest path lengths; 3) using the adjacency matrix of the weighted ligand-signaling network (Lig-sig adjacency); 4) PPR with damping factor of 0 such that in the final ligand-target model only direct ligand-target links are included. AUPR(corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under the receiver operating characteristic curve. The smoothing lines are the result of fitting a linear regression model. The accompanying shaded region indicates the 95% confidence interval. Red dashed line: performance of random guessing.



Supplementary Notes Figure 12 | Evaluation of the performance of NicheNet in predicting ligand activity: comparison to alternative approaches. We assessed the ability of NicheNet in predicting with which ligand the cells were treated based on the set of genes that are differentially expressed after ligand treatment (111 datasets, 51 unique ligands). Ligand prediction performance was compared between different variants of NicheNet: 1) the default approach of applying Personalized PageRank (PPR); 2) applying the shortest path length algorithm (SPL) and assigning higher importance scores to genes with shorter shortest path lengths; 3) using the adjacency matrix of the weighted ligand-signaling network (Lig-sig adjacency); 4) PPR with damping factor of 0 such that in the final ligand-target model only direct ligand-target links are included. All models constructed by these alternative approaches were optimized in the same way as the main variant of NicheNet. To assess performance, we calculated several classification evaluation metrics: AUPR(corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under the receiver operating characteristic curve; Pearson and Spearman's rank correlation. Boxplot elements: center line: median; box limits: upper and lower quartiles; whiskers: 1.5x interquartile range; points: outliers. The violin plots show the probability density of the data (tails of the violins are trimmed to the range of the data). Red dashed line: performance of random guessing.



Supplementary Notes Figure 13 | Popularity bias of NicheNet in predicting ligand activity. For 101 ligand treatment expression datasets profiling the transcriptional response to a single ligand, we plotted the performance of several variants of NicheNet in predicting the activity state of a ligand in function of its popularity (the decimal logarithm of the number of studies in PubMed in which a ligand was described). Following variants to calculate ligand-gene signaling importance scores were compared: 1) the default NicheNet approach of applying Personalized PageRank (PPR); 2) applying the shortest path length algorithm (SPL) and assigning higher importance scores to genes with shorter shortest path lengths; 3) using the adjacency matrix of the weighted ligand-signaling network (Lig-sig adjacency); 4) PPR with damping factor of 0 such that in the final ligand-target model only direct ligand-target links are included. AUPR(corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under the receiver operating characteristic curve. The smoothing lines are the result of fitting a linear regression model. The accompanying shaded region indicates the 95% confidence interval. Red dashed line: performance of random guessing.

Supplementary Note 6: results of parameter optimization

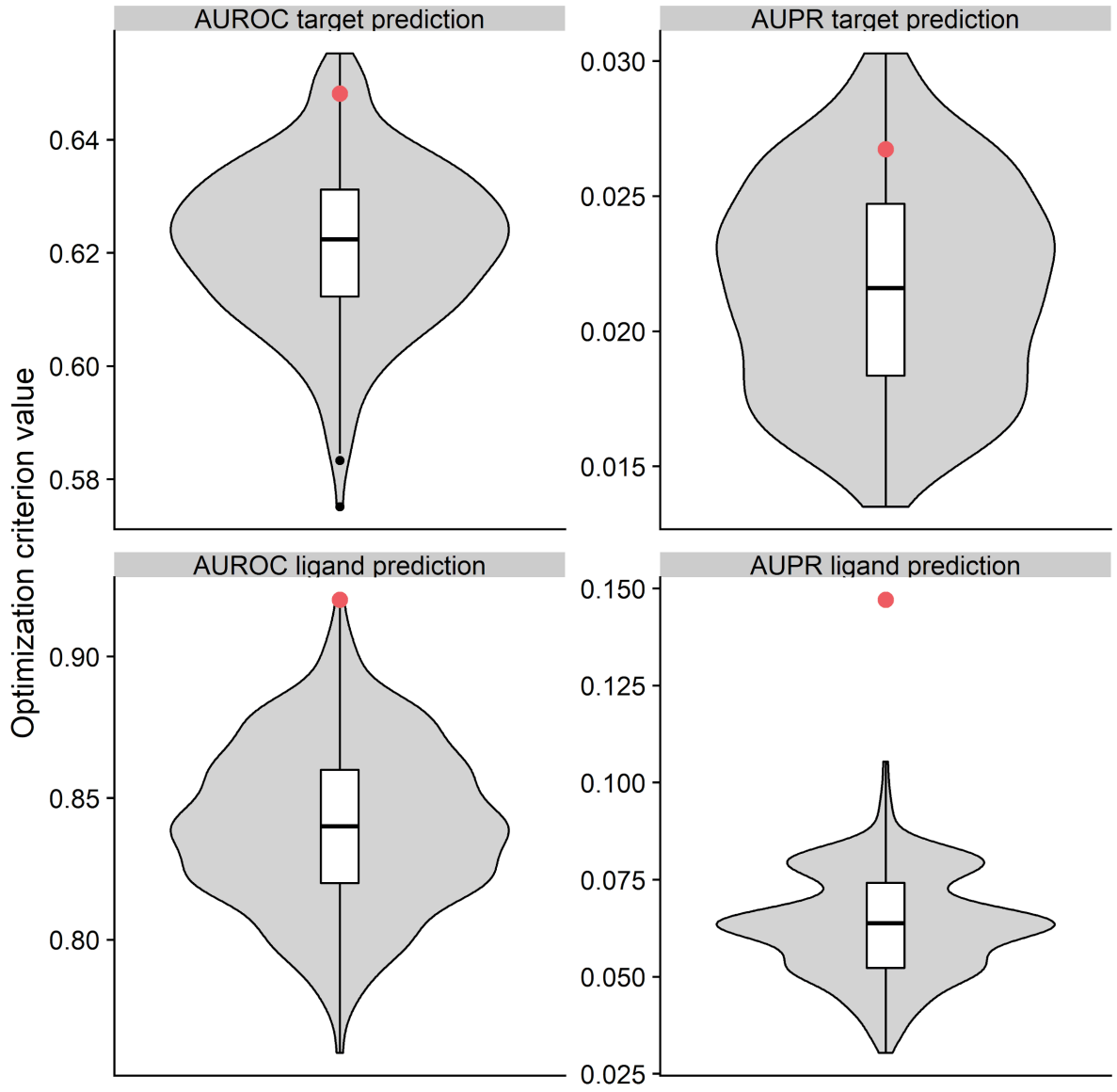
As described in **Supplementary Notes 2 and 3**, several parameters need to be defined: the weights of the 57 data sources; and the "hyperparameters": ligand signaling and gene regulatory network hub correction factors; the PPR vector cutoff; and the damping factor for PPR. We performed multi-objective model-based optimization to maximize the following four criteria: the AUROC and AUPR (after correction for random prediction) of target gene prediction (aggregated average per ligand over all ligand treatment datasets) and AUROC and AUPR (after correction for random prediction) of ligand prediction (aggregated median per ligand over all ligand treatment datasets). Optimization started with 240 parameter designs, and during each of 50 subsequent iteration rounds, 24 new designs were proposed and evaluated.

In **Supplementary Notes Figure 14**, one can see that the data source weights and parameter values can have a considerable influence on the model's performance and that optimization improves the quality of ligand-target predictions. It can also be observed that no parameter setting was optimal for all proposed optimization criteria and that a trade-off needed to be made **(Online Methods)**.

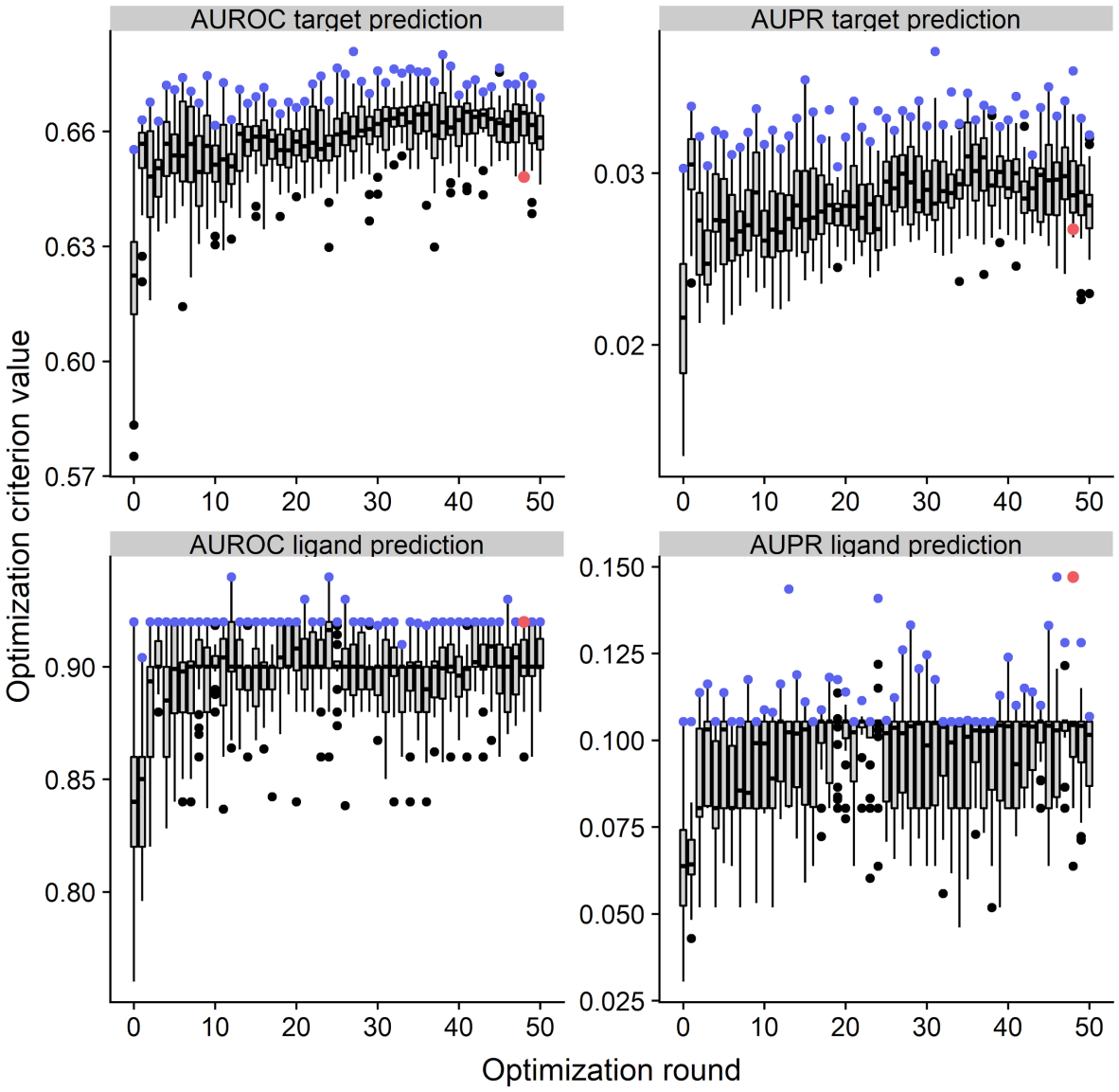
The evolution and convergence of performance over all optimization rounds are shown in **Supplementary Notes Figure 15**.

When we compared a model with optimized hyperparameters to a model with both optimized data source weights and hyperparameters, data source weight optimization seemed slightly beneficial for target gene prediction (**Supplementary Notes Figure 16**). The specific, optimized weights of the data sources are reported in **Supplementary Table 4b**.

Furthermore, we assessed the difference in performance between optimized models evaluated without and with cross-validation (**Supplementary Notes Figures 17-18 and Online Methods**). We did not observe clear signs of overfitting.

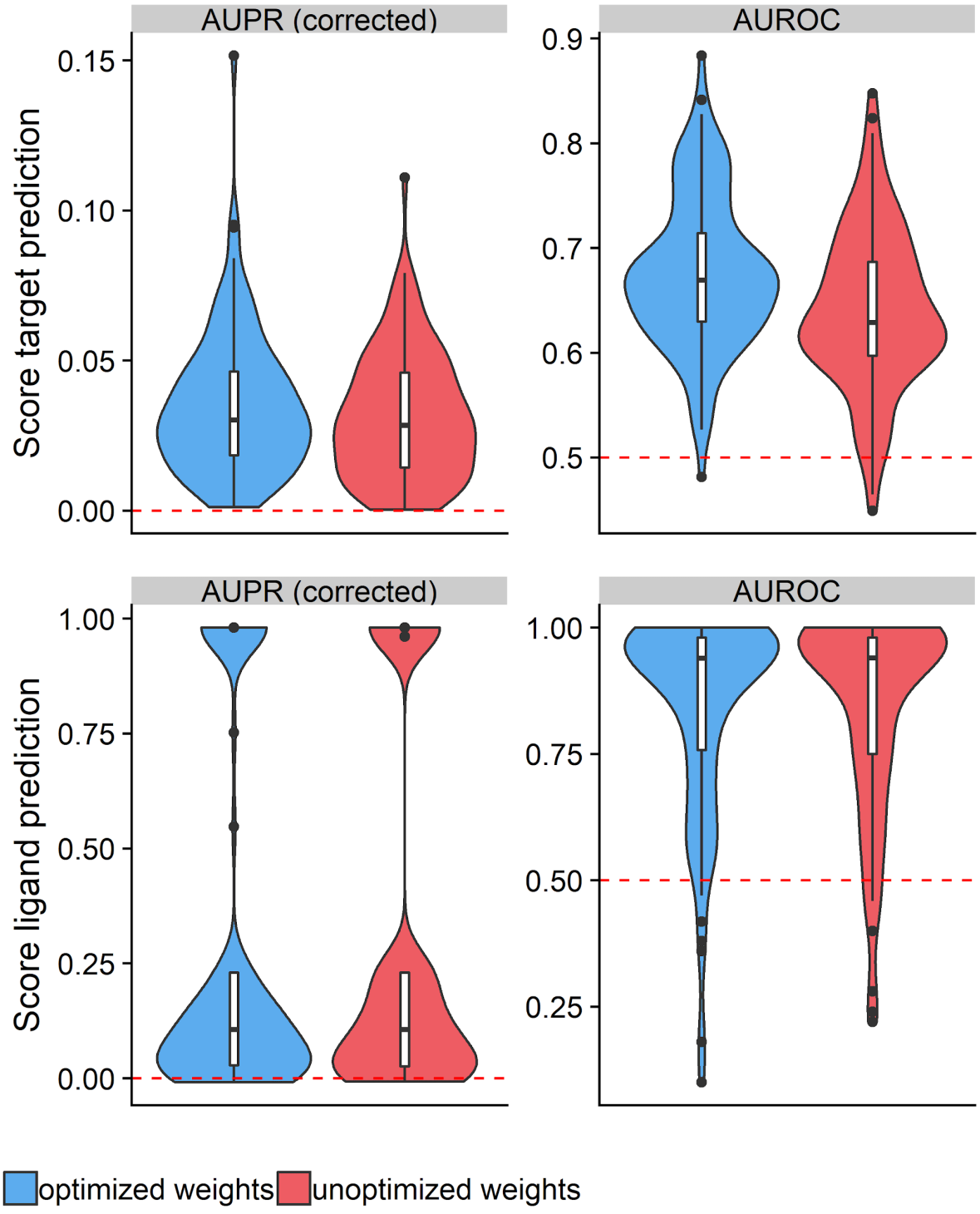


Supplementary Notes Figure 14 | Target gene and ligand activity prediction performances of all models generated during the optimization process. To assess the effect of different data source weights and parameter values on the model's performance, we show median ligand prediction and average target prediction performances over 111 ligand treatment datasets of all 1440 models evaluated during the mlrMBO multi-objective optimization procedure. The red dot indicates the performance of the final model with the most optimal parameter settings. AUPR(corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under the receiver operating characteristic curve. Boxplot elements: center line: median; box limits: upper and lower quartiles; whiskers: 1.5x interquartile range; points: outliers. The violin plots show the probability density of the data (tails of the violins are trimmed to the range of the data).

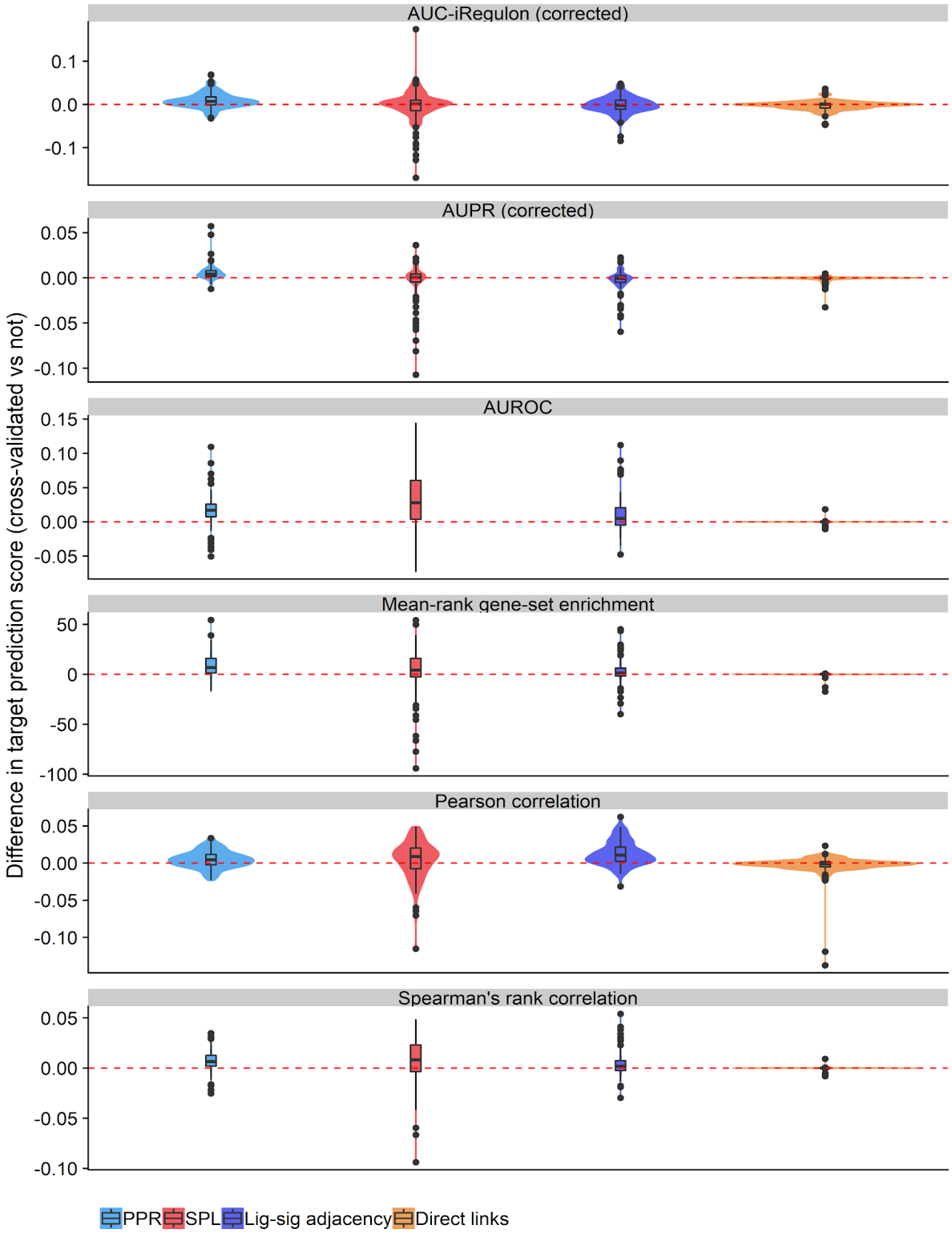


Supplementary Notes Figure 15 | Evolution of target gene and ligand activity prediction performances during the optimization process. For all models evaluated in an individual optimization round during the mlrMBO multi-objective optimization procedure, boxplots are shown of the median ligand prediction and average target prediction performances over 111 ligand treatment datasets. Boxplots summarize results for 240 models at initialization and 24 models for every optimization round. The red dot indicates the performance of the final model with the most optimal parameter setting. AUPR (corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under the receiver operating characteristic curve. Boxplot elements: center line: median; box limits: upper and lower quartiles; whiskers: 1.5x interquartile range; points: outliers.

.

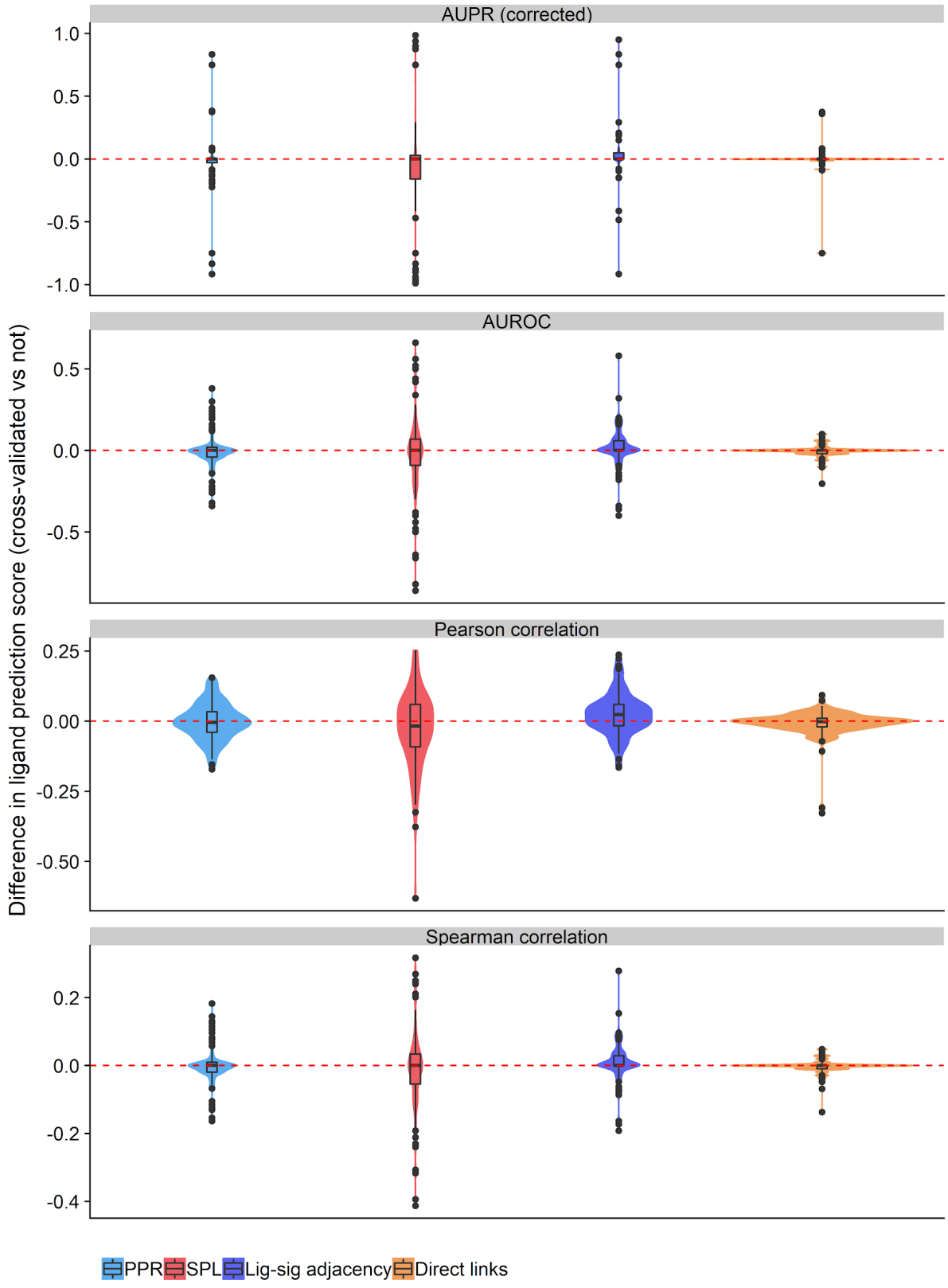


Supplementary Notes Figure 16 | Target gene and ligand activity prediction performances: comparison between NicheNet with and without optimized data source weights. To assess the influence of data source weight optimization, we compared the target gene and ligand activity prediction performance of NicheNet in which data sources receive optimized data source weights to NicheNet in which each data source receives the same weight (111 ligand treatment datasets for 51 unique ligands). AUPR (corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under the receiver operating characteristic curve. Boxplot elements: center line: median; box limits: upper and lower quartiles; whiskers: 1.5x interquartile range; points: outliers. The violin plots show the probability density of the data (tails of the violins are trimmed to the range of the data). Red dashed line: performance of random guessing.



Supplementary Notes Figure 17 | Target gene prediction performance: difference between models optimized without and with cross-validation. To assess the importance of performing cross-validation during model optimization, we subtracted the transcriptional response prediction performance on each ligand treatment dataset after optimization with threefold cross-validation by the performance after optimization without cross-validation (transcriptional response prediction performance was assessed on 111 ligand treatment datasets for 51 unique ligands). Effect of cross-validation was compared between different variants of NicheNet: 1) the default approach of applying Personalized PageRank (PPR); 2) applying the shortest path length algorithm (SPL)

and assigning higher importance scores to genes with shorter shortest path lengths; 3) using the adjacency matrix of the weighted ligand-signaling network (Lig-sig adjacency); 4) PPR with damping factor of 0 such that in the final ligand-target model only direct ligand-target links are included. To assess performance, several classification evaluation metrics were calculated: AUC-iRegulon (corrected): area under the cumulative recovery curve (considering the top 3% of the ranking), corrected for random prediction; AUPR (corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under receiver operating characteristic curve; Mean-rank gene-set enrichment: negative natural logarithm of the mean-rank gene-set enrichment p-value. Boxplot elements: center line: median; box limits: upper and lower quartiles; whiskers: 1.5x interquartile range; points: outliers. The violin plots show the probability density of the data (tails of the violins are trimmed to the range of the data). Red dashed line: performance of random guessing.



Supplementary Notes Figure 18 | Ligand activity prediction performance: difference between models optimized without and with cross-validation. To assess the importance of performing cross-validation during model optimization, we subtracted the ligand activity prediction performance after optimization with threefold cross-validation by the performance after optimization without cross-validation (performance was assessed on 111 ligand treatment datasets for 51 unique ligands). Effect of cross-validation was compared between different variants of NicheNet: 1) the default approach of applying Personalized PageRank (PPR); 2) applying the shortest

path length algorithm (SPL) and assigning higher importance scores to genes with shorter shortest path lengths; 3) using the adjacency matrix of the weighted ligand-signaling network (Lig-sig adjacency); 4) PPR with damping factor of 0 such that in the final ligand-target model only direct ligand-target links are included. To assess performance, we calculated several classification evaluation metrics: AUPR (corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under the receiver operating characteristic curve. Boxplot elements: center line: median; box limits: upper and lower quartiles; whiskers: 1.5x interquartile range; points: outliers. The violin plots show the probability density of the data (tails of the violins are trimmed to the range of the data). Red dashed line: performance of random guessing.

Supplementary Note 7: comparison of NicheNet to CCCExplorer and IPA Upstream Regulator Analysis for ligand activity prediction

We compared the ability of NicheNet to rank ligands as possible upstream regulators responsible for observed differential expression to IPA® Upstream Regulator Analysis[9] and CCCExplorer[10]. IPA's Upstream Regulator Analysis is a state-of-the-art proprietary software tool that is built on a manually curated knowledgebase and allows to prioritize possible factors (including ligands) regulating a specific gene set. CCCExplorer combines signaling pathway information and bulk transcriptomics data to predict ligand-to-target crosstalk pathways between interacting cell types.

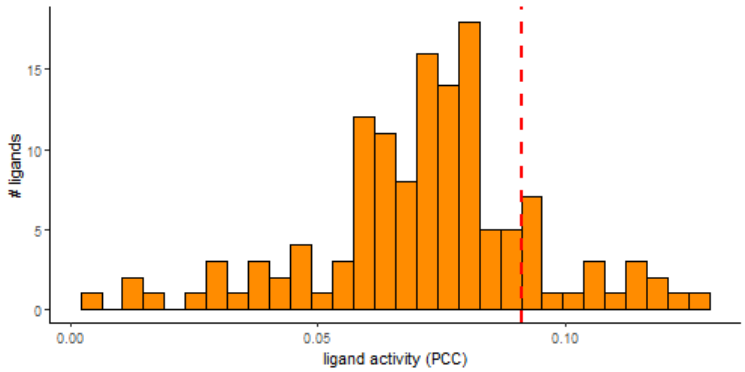
As possible active ligands for IPA® Upstream Regulator Analysis and CCCExplorer, we only considered the 51 ligands of which the transcriptional response was profiled in at least one of the 111 ligand treatment datasets (**Supplementary Table 2**). It should, however, be noticed that IPA® Upstream Regulator Analysis, in contrast to NicheNet, includes much more possible upstream regulators than only protein ligands: also metabolites, chemicals, and TFs are listed. Comparison between NicheNet and IPA was thus only performed to benchmark the inferred ligand-target predictions to ligand-target links present in IPA's knowledgebase by comparing the ranking of the considered extracellular protein ligands.

Another advantage of IPA Upstream Regulator Analysis, in addition to providing information on non-protein ligands as upstream regulators, is that predictions are also made concerning the activating/inhibitory state of possible upstream regulators.

Supplementary Note 8: p-EMT target gene prediction performance and ligand activity analysis

Because we wanted to investigate how well CAF-ligands can induce p-EMT target genes according to the NicheNet model, we considered the set of p-EMT genes defined by Puram et al. as affected target genes, and the other genes expressed in malignant cells as non-affected background genes[11]. For all ligands individually, we calculated several classification evaluation metrics that indicate how well NicheNet's ligand-target scores could predict the p-EMT genes (**Supplementary Table 5a**). To predict which CAF-ligands might be the most important in regulating the p-EMT program, NicheNet ligand activity prediction analysis ranks these ligands based on how well they predict the p-EMT gene set. Here we assume that the better a ligand predicts the observed transcriptional response compared to other ligands, the more likely it is that this ligand is the active ligand. To define ligand activity and prioritize top ligands, we used the Pearson correlation coefficient because the validation study on ligand treatment datasets showed that this metric ranks ligands according to their possible activity better than other metrics (**Supplementary Figure 7**). We see here that the top-ranked ligands can predict the p-EMT genes reasonably: e.g., for the top-ranked ligand PTHLH, the AUROC, AUPR and Pearson correlation coefficient are respectively 0.67, 0.07 and 0.13 (values respectively equal and higher than the average target gene prediction performance on the validation ligand treatment datasets; **Supplementary Table 3**). These prediction performance values indicate that signals received from CAF-ligands can explain at least a part of the p-EMT program within these cells. Because we observe reasonable target gene prediction performance, ligand prioritization based on this may be relevant in this case study. However, for other datasets, it is possible that the target gene prediction performance of the top-ranked ligands for some gene sets would not be better than random prediction. In that case, the prioritization of ligands would likely not be trustworthy. This also indicates that NicheNet does not find any evidence that the two cells under study are signaling to each other. Hence, NicheNet allows defining how confident predictions for a specific case study are, which is an important advantage.

In addition to considering each ligand individually, we trained random forest models that use the NicheNet ligand-target regulatory potential scores of CAF-ligands as features to predict the p-EMT gene set. Performances assessed via cross-validation are shown in **Supplementary Table 5c** when we considered all 131 ligands expressed by CAFs, and in **Supplementary Table 5d** when we only considered the 20 ligands with highest ligand activity (determined as described in the previous paragraph; see **Supplementary Notes Figure 19** for a distribution of ligand activity scores to justify the choice of 20 ligands). The main conclusion is that random forest models that use multiple ligands as features give slightly better p-EMT target gene predictions compared to considering ligands individually (according to AUROC, AUPR and Pearson and Spearman correlation). As an alternative for the Pearson correlation coefficient of individual ligands to define ligand activity, variable importances of a random forest can be used as well (e.g., mean decrease in Gini index, **Supplementary Table 5b**). AGT, PTHLH, and TGFB2 were the top 3 ligands according to the mean decrease in Gini index (compared to PTHLH, CXCL12, and AGT, which were the top 3 ligands according to the Pearson correlation coefficient).

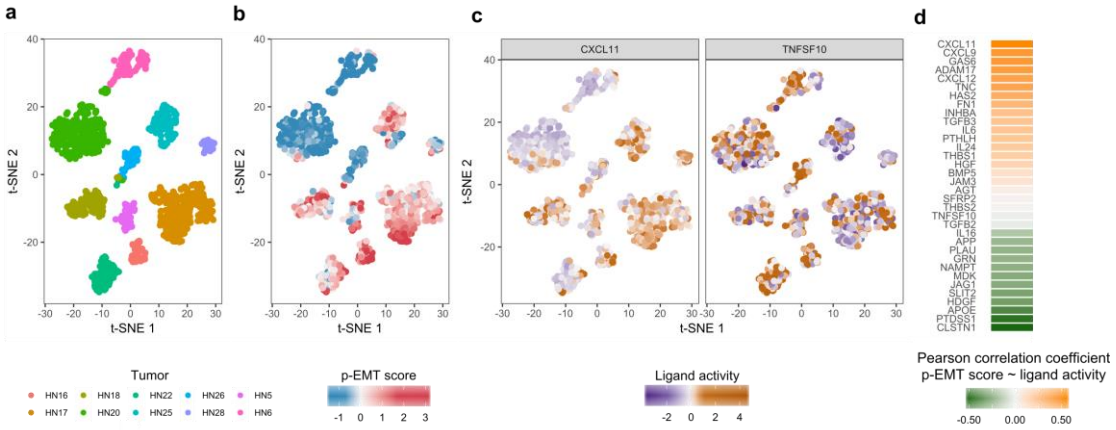


Supplementary Notes Figure 19 | Distribution of ligand activity scores for regulation of the p-EMT gene set. To investigate how CAF-ligands might promote the p-EMT program in neighboring malignant cells, NicheNet's ligand activity prediction was performed on the p-EMT gene set. Here, we show a histogram of the ligand activity scores for all 131 expressed CAF-ligands. Ligand activity is the Pearson correlation coefficient (PCC) between NicheNet's regulatory potential scores and p-EMT gene set assignments. The red dashed line indicates the ligand activity score of the ligand with rank 20.

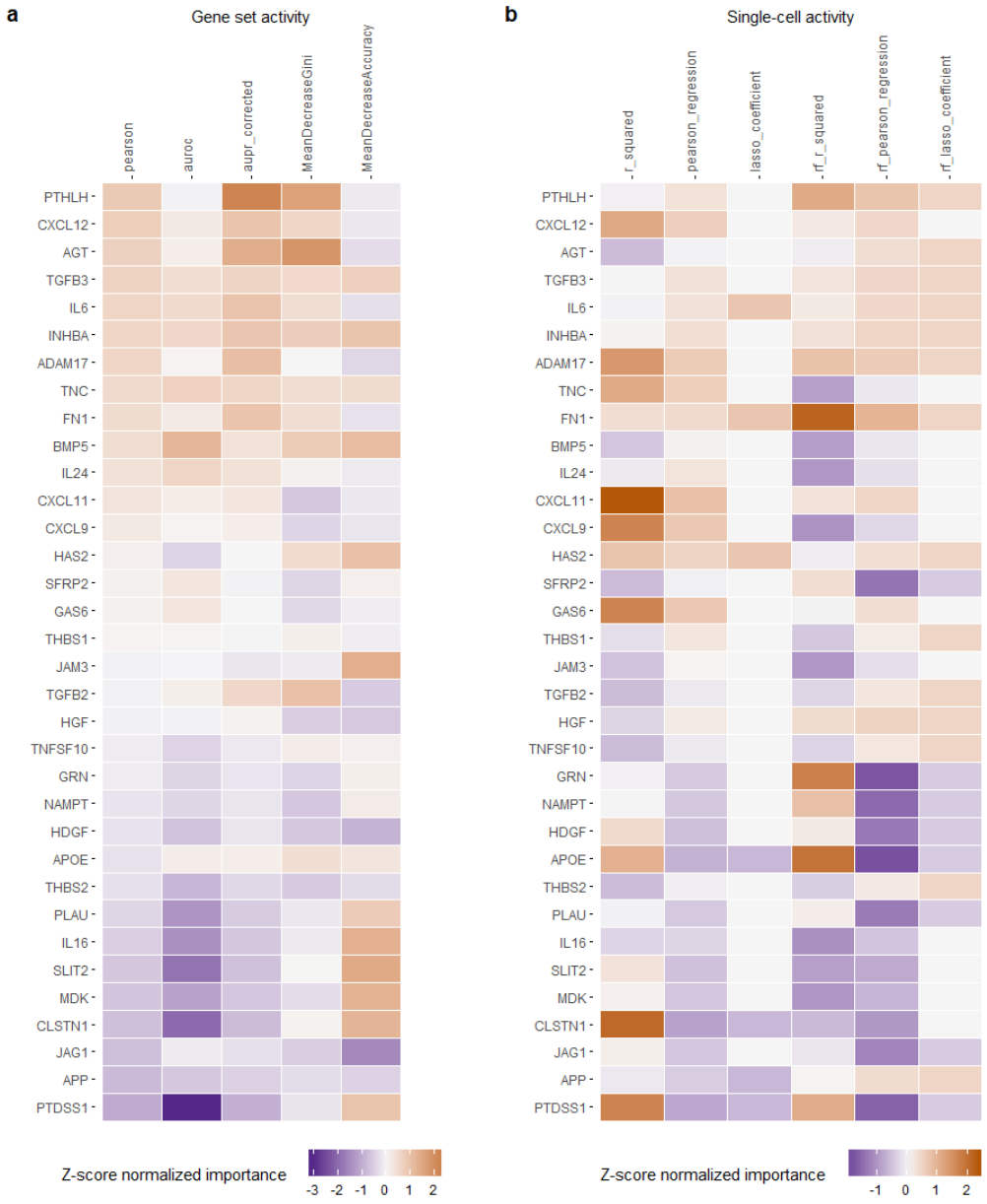
Supplementary Note 9: single-cell ligand activity prediction and correlation between ligand activity scores and p-EMT scores

Ligand activity prediction can not only be performed on predefined gene sets, but also at the single-cell level. This allows to to analyze the heterogeneity in ligand activity between individual cells. For this, ligand activity prediction was performed for every malignant cell on the set of genes most strongly expressed in that cell compared to other cells (**Online Methods**). This under the assumption that differential expression between malignant cells may - at least partially - be attributed to the effect of CAF-ligands. The obtained ligand activities were made comparable between different cells by z-score normalization. When a ligand has a high ligand activity in a cell, this means that putative target genes of a ligand tend to be stronger expressed than other genes in this cell compared to other cells. All further analyses described here were performed twice for two different ligand activity types: we used once the normalized Pearson correlation coefficient when considering ligands individually and once the normalized mean decrease in Gini index when using random forest.

Next, we scored malignant cells on the extent to which they expressed genes of the p-EMT program[11]. Subsequently, we assessed to what extent the p-EMT score of a cell can be predicted by the ligand activity of a ligand of interest. For this, we performed linear regression and correlation analysis between p-EMT scores of cells and activity scores for every CAF-ligand separately (**Supplementary Notes Figures 20-21 and Supplementary Table 8a-b**). Regression and correlation coefficients for ligands can be interpreted as follows. High t-statistics/Pearson correlations indicate that higher ligand activities are associated with higher p-EMT scores, which suggests that these ligands might be positive regulators of the p-EMT program. Negative coefficients, on the contrary, might suggest a possible role as a negative regulator for those ligands. By ranking all ligands with respect to the correlation between their activities and p-EMT scores over individual cells, we can thus rank ligands based on their potential p-EMT regulatory role. As can be expected, many top-ranked ligands according to this analysis were top-ranked as well according to the analysis discussed in the main text and Supplementary Note 8 (11 ligands were in the top 20 of both rankings and the Pearson correlation between both rankings is 0.69; **Fig. 2a**, **Supplementary Notes Figures 20d-21**).



Supplementary Notes Figure 20 | Application of NicheNet to specifically analyze the heterogeneity in CAF-ligand activity between individual malignant cells. Here, we applied NicheNet to single-cell RNA-seq data of malignant cells and CAFs to investigate how CAF-ligands might promote the p-EMT program in neighboring malignant cells. **(a)-(c)** tSNE plots of malignant cells, colored according to their tumor of origin **(a)**, p-EMT score **(b)** and ligand activity **(c)**. NicheNet's ligand activity prediction was for every cell performed on the genes that are most strongly expressed in the cell of interest compared to other cells. Shown are ligand activity scores after normalization to make them comparable between different cells. **(d)** Ligands can be ranked based on how well the activity of the ligand of interest in a cell correlates with the p-EMT score of that cell (as represented by the PCC of this correlation analysis) (n = 1388 cells).



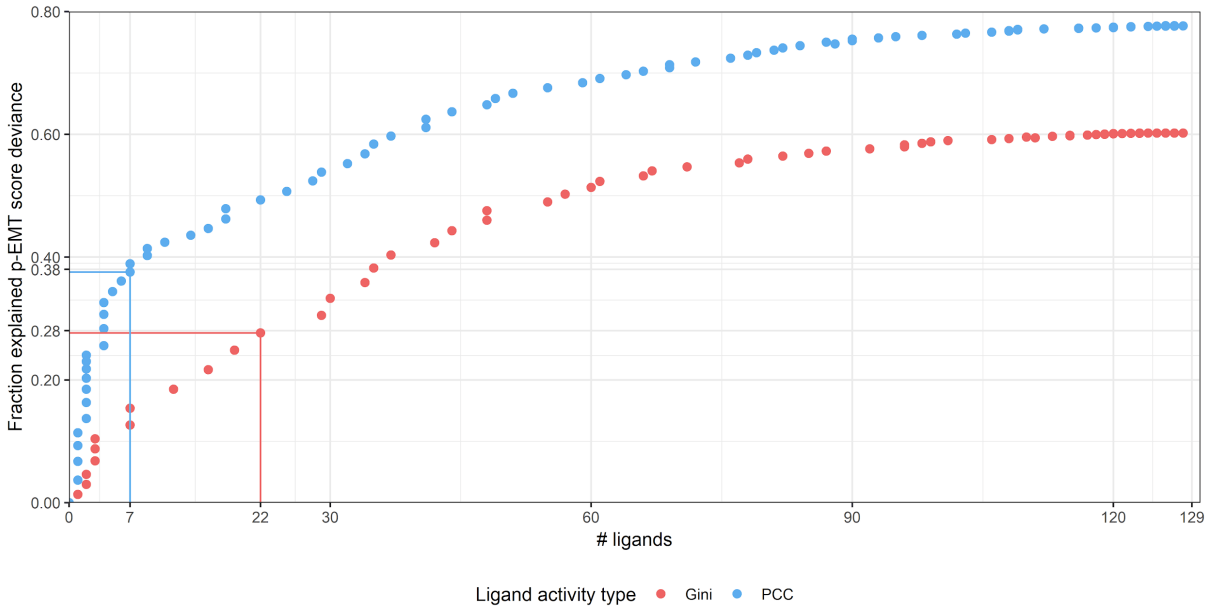
Supplementary Notes Figure 21 | Concordance between ligands prioritized by NicheNet's ligand activity prediction on the p-EMT gene set and by correlating single-cell p-EMT scores to single-cell ligand activities. (a) Results of NicheNet's ligand activity prediction on the p-EMT gene set. Ligand activities shown here are the normalized Pearson correlation coefficient, AUROC and AUPR (corrected for random prediction) for p-EMT target gene prediction considering every ligand separately and the normalized mean decrease in classification accuracy and mean decrease in gini index of a ligand when using a random forest trained with all ligands as features (n = 6072 genes, of which 96 belong to the p-EMT program). **(b)** Results of correlation, standard linear regression, and lasso regression analysis between a cell's p-EMT score and the activity of a ligand of interest for that cell (n = 1388 cells). NicheNet's ligand activity prediction was for every cell performed on the genes that are most strongly expressed in the cell of interest compared to other cells. Ligand activity was here defined as the normalized pearson correlation coefficient for target gene prediction or the mean decrease in Gini index of a ligand when using random forest ("rf") with all ligands as features for target gene prediction. For both (a) and (b) we show results for 34 out of 131 CAF-ligands. To be selected, a ligand needed to be among the 5 best predictive ligands according to the Pearson correlation coefficient, AUROC, AUPR, mean decrease in accuracy or mean decrease in gini index metrics for p-EMT gene set prediction or among the 5 ligands with the highest R squared for regression between p-EMT scores and ligand activity scores or among the ligands with a non-zero coefficient in lasso regression.

When we then specifically looked at the ligand activities within tumors, we found that for some ligands such as CTGF, the activity was strongly correlated to the cell's p-EMT score for all tumors (**Supplementary Notes Figure 22**). However, for other ligands, such as CXCL12, the relation between activity and p-EMT score depended on the tumor sample. This might suggest some intertumoral heterogeneity in crosstalk between CAFs and malignant cells. Applying NicheNet on single-cell data can thus be used to assess heterogeneity in intercellular signaling within cell populations and between samples.

Supplementary Notes Figure 22 | Single-cell NicheNet's ligand activity prediction can be applied to predict which CAF-ligands might induce a p-EMT program in malignant cells in head and neck squamous cell carcinoma. (a) tSNE plots of malignant cells from the HNSCC tumors HN16, HN26, HN17, HN6, and HN5. Cells were colored according to their p-EMT score, which shows the extent to which they express genes of the p-EMT program, and ligand activity for the p-EMT correlated ligands CTGF and CXCL12 and the non-correlated ligand TNFSF10. NicheNet's ligand activity prediction was for every cell performed on the genes that are most strongly expressed in the cell of interest compared to other cells. **(b)** Scatterplot showing the relation between p-EMT scores and CTGF, CXCL12 and TNFSF10 ligand activities over all cells from a specific tumor. The smoothing lines are the result of fitting a linear regression model (n = 369 cells for tumor HN6, 171 for HN26, 201 for HN5, 990 for HN17, and 168 for HN16). The accompanying shaded region indicates the 95% confidence interval.

In the following analysis, the objective was to find a limited set of ligands of which the ligand activities together explain a substantial part of the variation in p-EMT scores over cells (instead of looking at individual ligands as discussed in the previous paragraphs). Therefore, we performed lasso-regularized linear regression with all ligands as possible model parameters. To find a tradeoff between the number of model parameters (i.e., ligands) and predictive performance of the model, we selected the lasso model that explains half of the maximally explainable deviance. **Supplementary Notes Figure 23** shows the fraction of deviance in p-EMT scores that can be explained by ligand activities in function of the number of ligands used in the model. The regression coefficients of the finally selected models (and thus the prioritized ligands) are shown in **Supplementary Notes Figure 21b** and **Supplementary Table 8a-b**, and can be interpreted as explained above.

In conclusion, NicheNet can be applied to prioritize CAF-ligands possibly involved in the regulation of the p-EMT program in malignant cells, by assessing how strong their single-cell activities correlate to p-EMT scores. However, some limitations of ligand activity prediction at the single-cell level need to be kept in mind when interpreting these results: 1) the noisy nature of single-cell transcriptomics data; 2) that a substantial fraction of upregulated genes might be upregulated because of biological processes unrelated to intercellular crosstalk; 3) only ligands causing upregulation of their target genes will get a high activity ligand activity, repressing ligands not.



Supplementary Notes Figure 23 | Lasso regression to predict p-EMT scores of cells by ligand activities. Lasso-regularized linear regression with all CAF-ligands as possible model parameters was performed to assess to what extent the p-EMT score of a cell can be predicted by ligand activities of a subset of ligands. As ligand activity, we considered once the normalized Pearson correlation coefficient when considering ligands individually (PCC) and once the normalized mean decrease in Gini index when using random forest (Gini). The colored lines indicate the number of ligands that are necessary to explain half of the maximally explainable deviance.

Supplementary Note 10: Immune cell - T cell interactions in the T cell area in the lymph node

In a second case study, we applied NicheNet to NICHE-seq data that was generated by Medaglia et al. to characterize immune cell composition in the T cell area of inguinal lymph nodes (both in steady-state and 72 hours after lymphocytic choriomeningitis virus (LCMV) infection)[12]. To identify cell populations present in both steady-state and after viral infection, we processed the data according to the Seurat alignment pipeline. In **Supplementary Notes Figure 24**, we show that Seurat alignment is necessary to assign defined cell clusters to cell types present in the T cell area before and after LCMV infection. We observed several cell types: CD4 T cells (including regulatory T cells), CD8 T cells, B cells, NK cells, dendritic cells (DCs), and inflammatory monocytes.

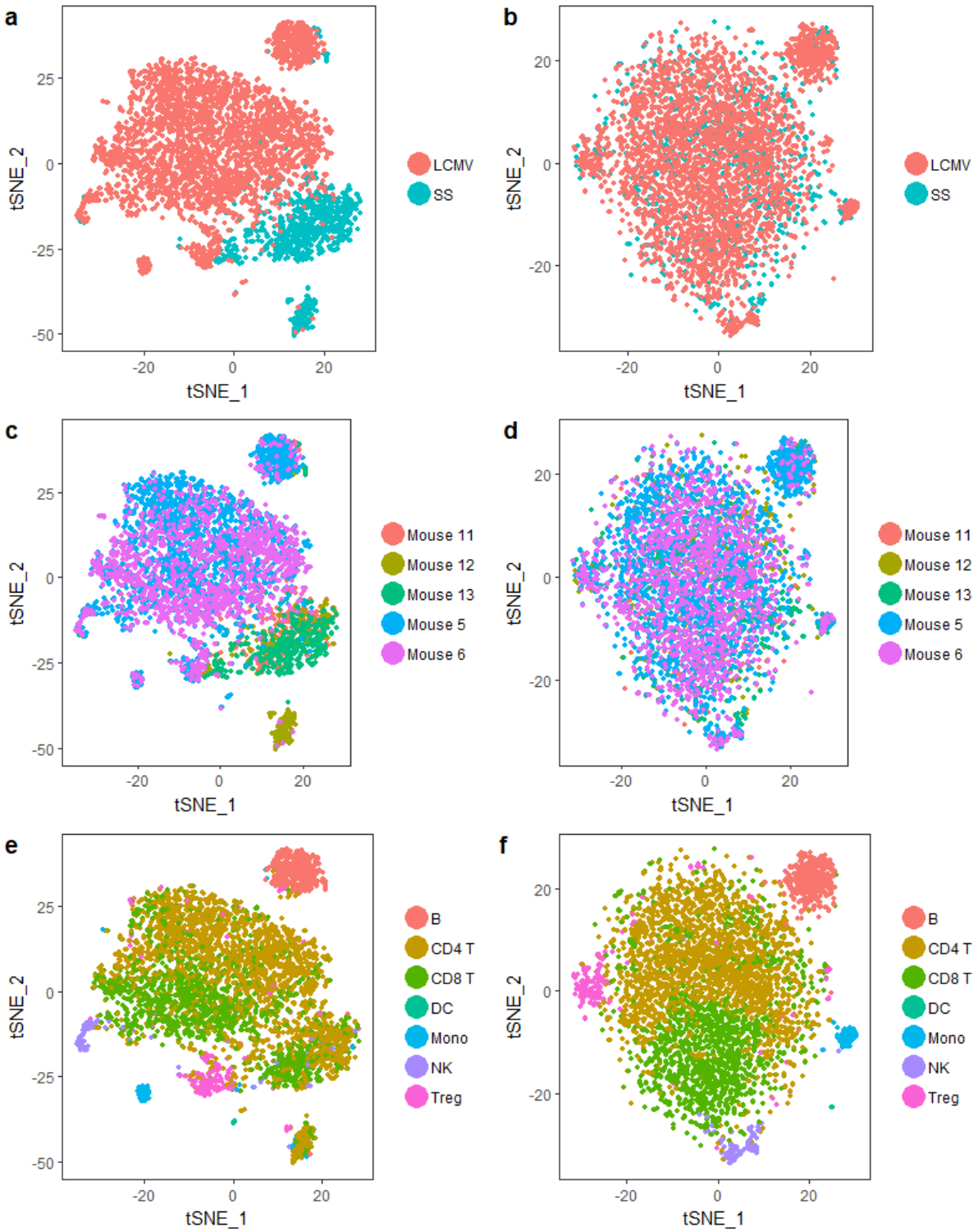
In this study, we wanted to investigate the regulation of the anti-LCMV response in CD8 T cells by ligands expressed by other cells in the T cell niche. As ligands of interest, we looked at ligands expressed by B cells, NK cells, monocytes, and DCs in steady-state and/or after LCMV infection. As affected target genes, we considered the set of genes differentially expressed before and after LCMV infection. Before applying NicheNet on this mouse data, we needed to convert the human ligand and target gene names of the ligand-target matrix to their one-to-one orthologs in mouse (**Supplementary Note 13: Conversion of gene identifiers**).

For all ligands individually, we calculated several classification evaluation metrics that indicate how well NicheNet's ligand-target scores could predict the transcriptional response (**Supplementary Table 9a**). We found that Il27, Ebi3, Il12a, Ifng, and Il18 were the five ligands with the highest ligand activity (**Supplementary Notes Figure 25)**. When we ranked ligands based on ligand activity, we found that ligands belonging to the Gene Ontology (GO) category "defense against viral infection" were enriched at the top of the ranking (the AUROC is 0.78 for predicting whether a ligand belongs to this GO category based on the ligand activity score). Notably, some top-ranked ligands like Ebi3, Il12a, and Il18 do not belong to the GO category "defense against viral infection" but are still known to be involved in the antiviral response in CD8 T cells[13]. Interestingly as well is that 4 of the top 10 ligands (Il27, Clcf1, Ebi3 and also Il12a) have a predicted interaction with Il6st, a receptor that plays a role in antiviral response (**Supplementary Notes Figure 26)**[13]. Another interesting observation is that monocytes, but only after viral infection, seem to be the niche cells expressing most of the top-ranked ligands. Altogether, these results indicate that NicheNet prioritizes ligands that are functionally important in the process of interest.

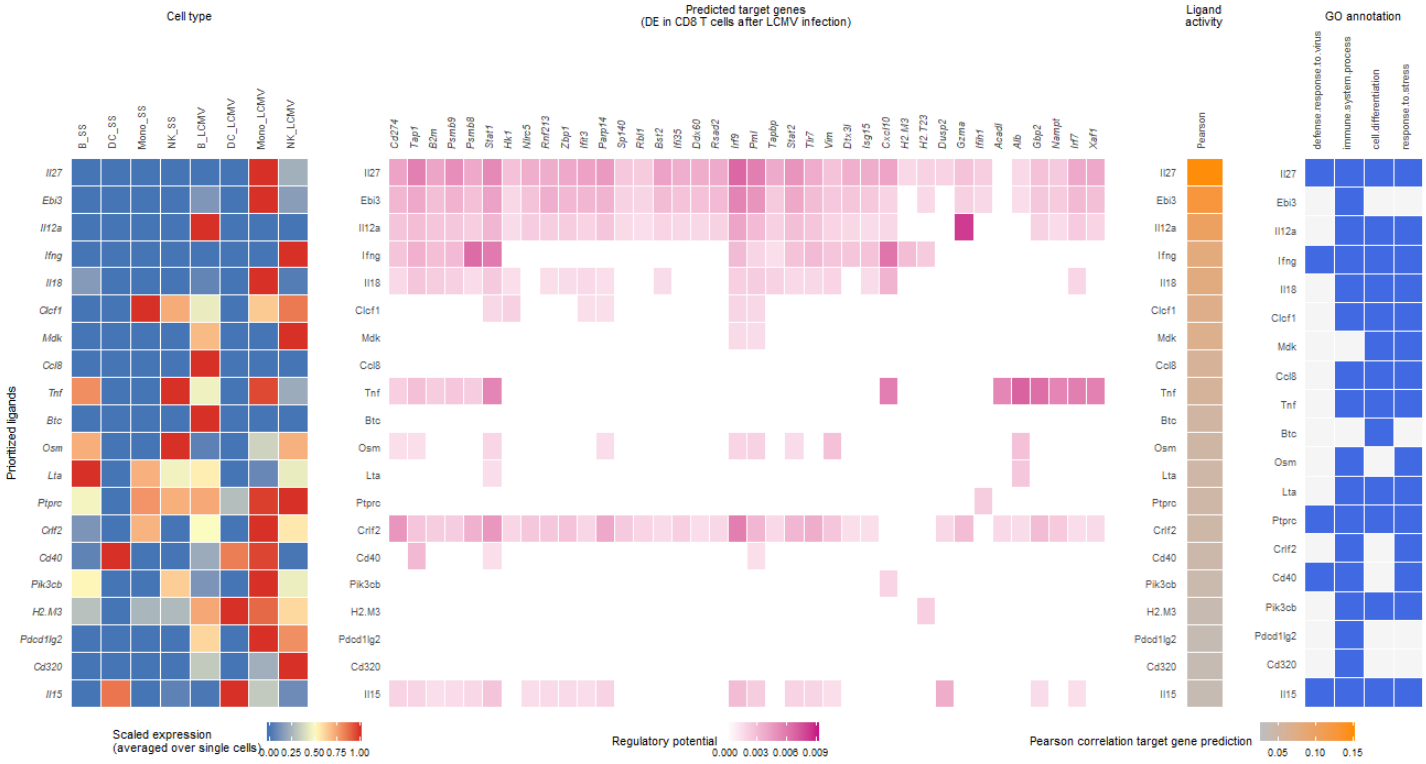
In addition to considering each ligand individually, we also trained a random forest that uses the ligand-target regulatory potential scores of multiple niche ligands as features to predict the anti-LCMV transcriptional response. Performances estimated via cross-validation are shown in **Supplementary Table 9d** when we considered all 179 ligands expressed by at least one of the T cell niche cell types, and in **Supplementary Table 9e** when we only considered the 20 ligands with highest ligand activity (determined as described in the previous paragraph; see **Supplementary Notes Figure 27** for a distribution of ligand activity scores to justify the choice of 20 ligands). These random forests with both all and top 20 ligands give better predictions compared to considering ligands individually: the average AUPR and Pearson correlation over all cross-validation rounds are respectively three and two times higher than of Il27, the ligand with highest AUPR and Pearson correlation (respectively 0.187 ± 0.007 vs 0.062 and 0.296 ± 0.003 vs 0.148 for all ligands, 0.187 ± 0.017 vs 0.062 and 0.307 ± 0.014 vs 0.148 for the top 20 ligands). When we then considered the 5% most strongly predicted target genes for the

model with the top 20 ligands, we observed that 48.4% of the anti-LCMV response genes were among these top 5% target genes, compared to 4.5% of genes not part of the anti-LCMV response (p-value Fisher's Exact Test: 4.6×10^{-39}) (**Supplementary Table 9f-h**). This suggests that top-ranked T cell niche ligands are potential regulators of a substantial set of anti-LCMV response genes. Finally, we also calculated variable importances of a random forest using all ligands as features as alternative ligand activity measure (**Supplementary Table 9c)**. Il27, Mdk, Ebi3, Btc, and Ccl5 were the top five ligands according to the mean decrease in Gini index.

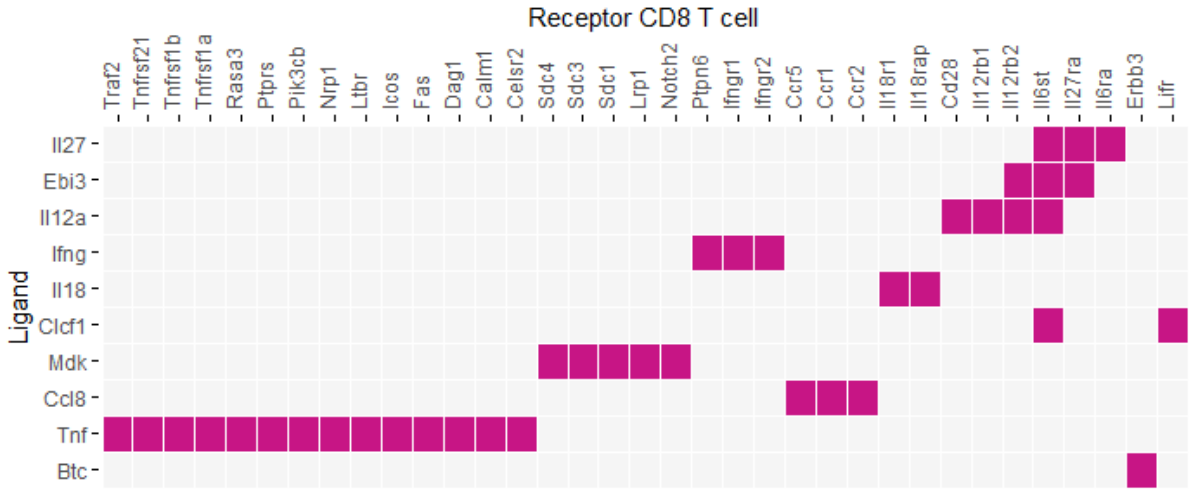
As a final remark on this case study, we discuss here the observation that no type I interferons were found back at the top of the ranking, which was in contrast to what we expected. Further research into this observation learned that no type I interferons were expressed in the niche cells at the moment of sampling (and were therefore not included in the ligand activity analysis). However, ligand activity analysis for type I interferons showed that a clear type I interferon signature was present in the anti-LCMV response in CD8 T cells (**Supplementary Table 9b)**. It is thus possible that type I interferons were expressed by other non-analyzed cell types, or by the cell types analyzed here, but before the sampled time point (which was 72 hours after LCMV infection). This observation illustrates one of the limitations of applying NicheNet if data of both interacting cells was generated at the same time point. Namely that there is a time delay between the moment of expression of ligands expressed by the sender cell type and the transcriptional response to them in the receiver cell type. Therefore, a measured transcriptional response is the consequence of the ligands expressed by niche cells some time before the transcriptional response was measured.



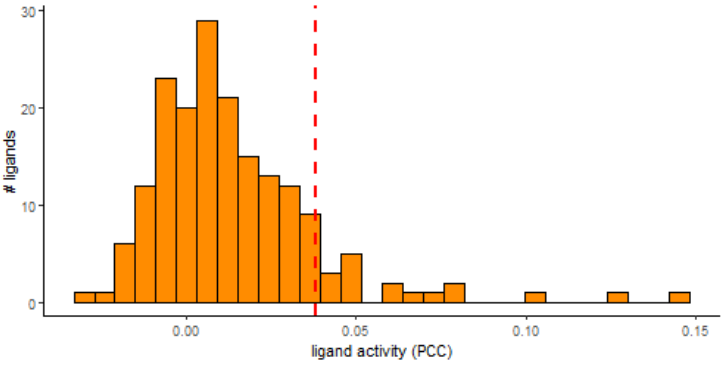
Supplementary Notes Figure 24 | Seurat alignment is necessary to find shared cell populations in the T cell area before and after LCMV infection. Dimensionality reduction was performed without (**a,c,e**) and with (**b,d,f**) the Seurat alignment procedure. tSNE plots of immune cells in the T cell area of the inguinal lymph node (n = 5027), colored according to whether NICHE-seq was performed in steady-state ("SS") or after LCMV infection (**a,b**); according to mouse of origin (**c,d**); and according to cell type (**e,f**). Seurat clusters could only be accurately assigned to cell types after alignment.



Supplementary Notes Figure 25 | NicheNet predicts regulatory interactions between ligands expressed by immune cells in the lymph node and genes affected in CD8 T cells after lymphocytic choriomeningitis virus (LCMV) infection. Summary result of NicheNet analysis: results are shown for the 20 (out of 179) T cell niche ligands best predicting the anti-LCMV transcriptional response in CD8 T cells. Left: scaled expression of ligands in B cells, dendritic cells ("DC"), monocytes ("Mono") and NK cells; for every cell type averaged per condition: steady-state ("SS") or after LCMV infection. Center left: NicheNet's ligand-target matrix denoting the regulatory potential between these niche ligands and anti-LCMV response genes in CD8 T cells. Center right: outcome of NicheNet's ligand activity prediction on the set of genes differentially expressed ("DE") after LCMV infection. As the ligand activity ranking metric, we used the Pearson correlation coefficient between prior regulatory potential scores and anti-LCMV-response gene set assignments (n = 9508 genes, of which 110 belong to the anti-LCMV-response gene set). Right: indication which of these ligands belong to specific Gene Ontology (GO) biological process categories. Ligands are ordered according to the Pearson correlation between their target predictions and observed transcriptional response; target genes are ordered based on hierarchical clustering of the ligand-target matrix.



Supplementary Notes Figure 26 | Prioritized ligand-receptor network involved in inducing the observed anti-LCMV response in CD8 T cells. Ligand-receptor interactions are shown for the 10 (out of 179) niche ligands best predicting the anti-LCMV transcriptional response in CD8 T cells (as determined by the Pearson correlation in ligand-target predictions and observed response). Shown order of receptors (n = 35) was determined via hierarchical clustering.



Supplementary Notes Figure 27 | Distribution of ligand activity scores for regulation of the anti-LCMV response in CD8 T cells. To investigate how T cell niche ligands might regulate the anti-LCMV response in CD8 T cells, NicheNet's ligand activity prediction was performed on the set of genes differentially expressed after LCMV infection. Here, we show a histogram of the ligand activity scores for all 179 expressed niche-ligands. Ligand activity is the Pearson correlation coefficient (PCC) between NicheNet's regulatory potential scores and anti-LCMV response gene set assignments. The red dashed line indicates the ligand activity score of the ligand with rank 20.

Supplementary Note 11: Studying cell-cell communication from gene expression data: comparison between basic ligand-receptor network inference methods, CellphoneDB and NicheNet

NicheNet and ligand-receptor network inference methods, such as CellPhoneDB, can all be used to study intercellular communication, but it is important to note that they address different questions. Basic ligand-receptor network inference methods (as used in several studies[11,14–17]) and CellPhoneDB[18] predict communicating pairs of cells and the facilitating ligand-receptor interactions by looking at the expression of interacting ligands and receptors. On the contrary, NicheNet predicts which of these inferred ligand-receptor links are possibly the most functional based on gene expression changes that are induced in the target cell type. Therefore, the choice of which method to use depends on the type of data and research question. We summarized the main similarities and differences between these methods in **Supplementary Notes Table 1**.

Supplementary Notes Table 1 | Comparison between basic ligand-receptor network inference methods, CellPhoneDB, and NicheNet.

	Basic ligand-receptor network inference methods	CellphoneDB	NicheNet
Main goal of the method and output	**Ligand-receptor network inference**: inference of possible ligand-receptor interactions between cell types of interest, based on the expression of ligands in one cell type and receptors in another cell type.	**Cell-type-specific ligand-receptor network inference**: inference of possible ligand-receptor interactions between cell types of interest, based on the expression of ligands in one cell type and receptors in another cell type. Moreover, one can search for ligand-receptor interactions that are enriched between a specific pair of cell types.	**Ligand-receptor network inference, followed by prioritization of ligand-receptor interactions based on gene regulatory effects.** **Predicted affected target genes** of these prioritized ligand-receptor interactions. **Predicted signaling paths** between prioritized ligands and their target genes.
Working	Integration of expression data of interacting cells with **prior information on potential ligand-receptor interactions.**	Integration of expression data of interacting cells with **prior information on potential ligand-receptor interactions.** Statistical framework to infer cell-type-specific interactions. Prior information on ligand-receptor interactions (i.e., the CellPhoneDB repository) **incorporates information on subunit architecture**, representing heteromeric complexes accurately.	Integration of expression data of interacting cells with **prior information on potential ligand-receptor, signaling, and gene regulatory interactions.**
Prioritization of ligand-receptor interactions	**No** explicit computational prioritization.	Prioritization of ligand-receptor interactions based on **cell-type specific expression** of ligands and expression**.**	Prioritization of ligand-receptor interactions based on **gene regulatory effects on a target cell type.**
Data to use on	**Bulk or single-cell** gene expression data from putatively interacting cell types. **Can be "steady state"** data.	**Single-cell** gene expression data from putatively interacting cell types. **Can be "steady state"** data.	**Bulk or single-cell** gene expression data from putatively interacting cell types. **Cannot be "steady state"** data**.** Because NicheNet predicts which ligands influence gene expression in a target cell type, there should be information in the data about gene expression changes in the target cell type that are possibly caused by interacting cells (differential expression).

Supplementary Note 12: Modularity of NicheNet

While applying the general model of NicheNet might already generate interesting hypotheses, the modularity of NicheNet permits to construct a cell-type-specific model for the biological system of interest. For example, it could be possible to include predicted gene regulatory interactions from network inference methods applied on an extensive expression compendium of the cell type of interest or directly on the single-cell transcriptomics data. One could also include expression data of intracellular regulators and chromatin data of the cell type of interest (e.g., from ATAC-seq). This would allow to filter out ligand-target predictions if an intermediary signaling mediator would not be expressed or the target gene promoter would not be accessible due to a closed chromatin state.

Context specificity may not only be established by including specific data sources: it could also be of additional value to perform context-specific data source weight optimization since in this study we showed that some data sources might be more informative for some biological systems than for others (**Supplementary Figure 10**). Without context-specific optimization, a particular interaction that is very important in the system of interest but is only supported by one data source might be less prominent compared to all other interactions. In such an approach, we would thus use the integrated networks not to construct a general model of ligand-target links but to create a context-specific model optimally explaining the observed data in the system of interest. However, to avoid overfitting, cross-validation would likely be necessary.

We consider it as an advantage of the proposed NicheNet framework that we provide a general model that is not only directly applicable to a broad set of biological systems but which can also be extended and adapted for better cell-type-specific predictions. In the software we provide, code and documentation are included such that a user can personalize the data sources within the NicheNet framework and construct their own models easily. Hereby, a user can include additional data sources, such as context-specific ones or from databases not included in NicheNet (e.g., the CellPhoneDB ligand-receptor repository), and remove specific data sources that the user considers to be noisy or irrelevant.

Supplementary Note 13: data source processing specifics

Ligand-receptor interaction data sources

From KEGG, the following pathways – for both human and mouse – were extracted and converted to ligand-receptor networks via XML parsing: "Cytokine-Cytokine receptor interaction", "Neuroactive ligand-receptor interaction", "ECM-receptor interaction" and "Cell adhesion molecules"[19]. Interactions from the IUPHAR Guide to Pharmacology database were retrieved via Harmonizome[20,21]. Ligand-receptor interactions from the database constructed in the study of Ramilowski et al. were collected as well: for our study, we only kept the literature-confirmed interactions and not the predicted ones[22].

In addition to these interactions, we predicted novel human and mouse interactions by querying protein-protein interaction databases for interactions between genes annotated as ligands and receptors. Firstly, genes were annotated as ligand, receptor or "bidirectional" (if the protein can function both as ligand and receptor) in two ways: based on the collected ligand-receptor interactions and based on GO annotations[23]. The biological function GO terms used to annotate genes as ligand were: "cytokine activity", "growth factor activity", "hormone activity" or "receptor ligand activity"; to annotate genes as receptors: "cytokine receptor activity" or "receptor activity"; and to annotate genes as bidirectionals: "homophilic cell adhesion via plasma membrane adhesion molecules". To exclude nuclear receptors as potential receptors for protein ligands, receptors should be annotated to at least one of the following GO terms for subcellular localization: "external side of plasma membrane", "integral component of membrane" or "plasma membrane". To resolve ambiguities between gene annotations based on the collected data sources, the annotation inferred from Guide to Pharmacology was preferred over KEGG, which was preferred over Ramilowski et al. To resolve ambiguities between gene annotations based on GO, genes annotated as both ligand and receptor/bidirectional were considered as ligand if the gene is linked to the subcellular localization GO term "extracellular space"; genes annotated as both receptor and bidirectional were considered as bidirectional. Secondly, we searched for ligand-receptor, ligand-bidirectional, bidirectional-receptor, and bidirectional-bidirectional interactions in the protein-protein interaction networks of ConsensusPathDB, InWeb_InBioMap, and Omnipath[24–26]. Of all links found, we only kept links not documented in one of the collected publicly available ligand-receptor data sources.

Signaling interaction data sources

We retrieved binary PPIs and protein complexes from the comprehensive interaction database ConsensusPathDB[24]. We converted known complexes of proteins to all pairwise PPIs between the members of a complex and considered binary PPI as another data type than PPI between members of a protein complex. Interactions in "default" Omnipath, a second comprehensive interaction database containing a collection of several (curated) human signaling pathways, were split up in directional and non-directional interactions[26]. From InWeb_InBioMap, another comprehensive interaction database, interactions were collected and divided according to whether the interactions were retrieved from an interaction database, pathway database, or both[25]. From PathwayCommons, a comprehensive pathway database, all signaling interactions were retrieved except for the gene regulatory ones (denoted by "controls-expression-of") and divided according to the data type they are documented in PathwayCommons ("interaction", "interaction-complex", "controls change of state of","controls phosphorylation of",...)[27].

In addition to interactions contained in comprehensive databases, signaling interactions were collected as well from post-translational modification databases (phosphorylation and dephosphorylation). Via Harmonizome, we collected interactions from KEA, PhosphoSite and DEPOD[28–

[30]. Kinase-substrate predictions from Harmonizome were collected as well; here only interactions with a probability score > 0.25 were collected. This arbitrary cutoff was chosen to avoid too many predicted interactions.

Following human and mouse interactions were retrieved as signaling interactions from EVEX, a database containing interactions derived from applying text mining on PubMed abstracts and PubMed Central full texts: "catalysis (of acetylation, ubiquitination, hydroxylation, glycosylation)", "catalysis of phosphorylation", "regulation of binding", "binding", "regulation of other processes (localization, catabolism)"[31]. Of these data types, both direct and indirect regulations were merged together. To avoid too many less confident interactions, we only considered the 33% most confident interactions.

As final signaling data sources, we collected the PPI network from Vinayagam et al[32]. In contrast to the other PPI networks collected, interactions had a predicted directionality.

Gene regulatory interaction data sources

A lot of gene regulatory interactions were retrieved via Harmonizome: GEO signatures of differentially expressed genes for gene/kinase/transcription factor perturbations and MSigDB signatures of differentially expressed genes for cancer gene perturbations as perturbation data sources[33,34]; JASPAR, TRANSFAC, TRANSFAC curated and MOTIFMAP as motif data sources[35–37]; ENCODE and CHEA as ChIP-seq data sources[38,39].

As comprehensive gene regulatory data source, we used RegNetwork and retrieved all human and mouse interactions[40]; as data sources of experimentally verified transcriptional regulation interactions, we used HTRIdb and TRRUST (human and mouse)[41,42].

The pathway database PathwayCommons and text mining database EVEX contain not only signaling but gene regulatory interactions as well (denoted respectively by "controls-expression-of" and "(indirect) regulation of expression/transcription")[27,31]. To avoid too many less confident interactions, we only considered the 33% most confident interactions in EVEX. Because both data sources provide direct gene regulatory interactions between ligands or receptors and target genes as well, we split the interaction in two data types: one data type with ligand or receptor – target interactions and one with other interactions between other genes (like kinases and TFs) and targets. We chose to make this distinction because text mining-inferred direct ligand-target links could potentially be more informative than text mining-inferred transcription factor target links.

The ImmGen consortium collected expression data of immune cells and clustered genes in fine-grained and coarse-grained modules[43]. Via the Ontogenet algorithm[44], they predicted transcriptional regulators for each of the fine and coarse modules. Here, we collected the member genes of each coarse module and the predicted regulators for each coarse module and merged both to obtain regulator-target interactions predicted from co-expression between regulators and targets.

As a final data source, we used ReMap (ChIP-seq)[45]. In contrast to other data sources, the provided data had still to be processed into a network format. For each of the 237 TFs of which binding sites are contained in the ReMap data source, we inferred the possible target genes as follows. For each of all genes annotated in the human hg19 genome, the potential to be regulated by the TF (i.e. regulatory potential) was calculated as described by Tang et al. (i.e. the sum of the nearby TF-binding sites that are each weighted by their distance to the transcription start site of a gene)[46]. The regulatory potential of every gene was then transformed into a z-score for normalization. From this z-score, a p-value was estimated and corrected for multiple testing by controlling the FDR as described previously[47]. Transcription factor-target interactions with adjusted p-value < 0.05 were kept.

Conversion of gene identifiers

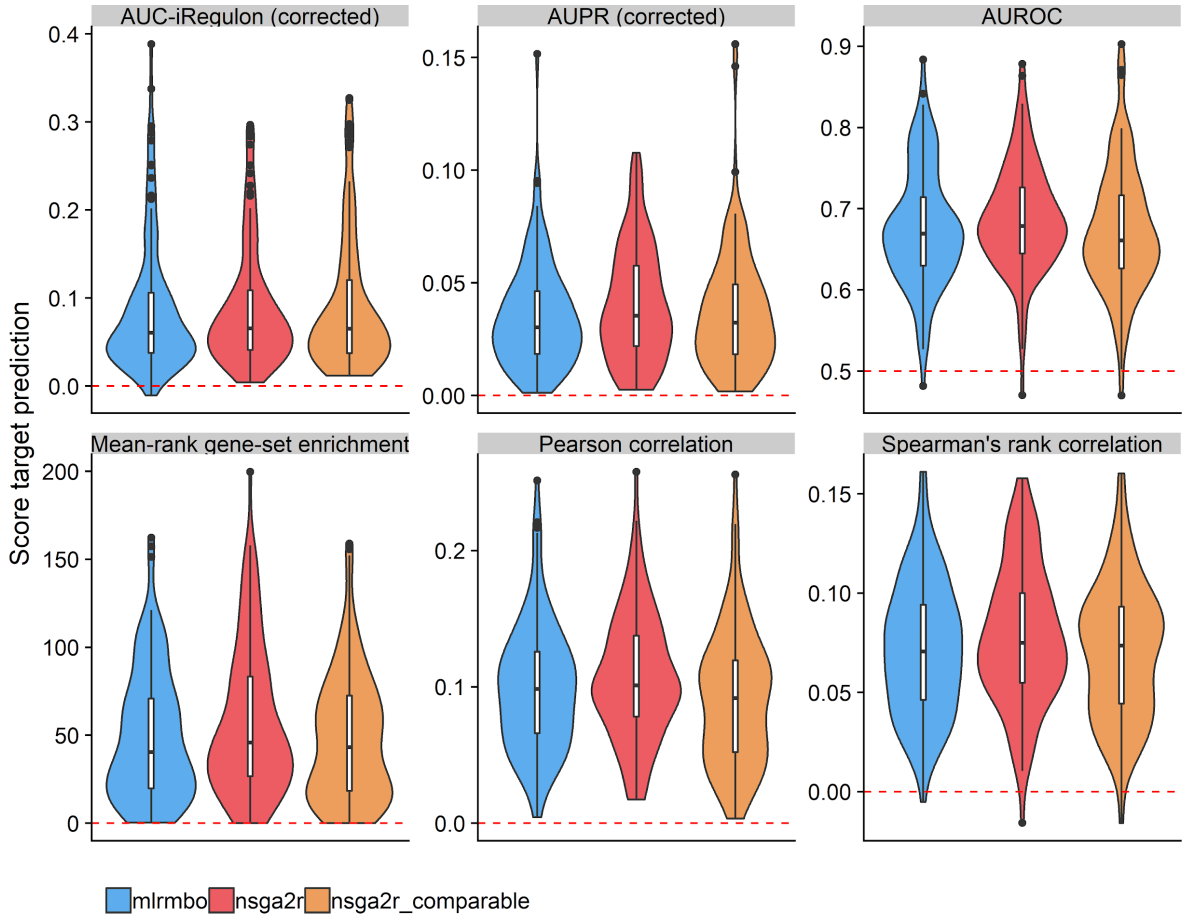
Most data sources provided interactions in human and a minority in mouse. To combine these in one final model (in HGNC human gene symbols or MGI mouse gene symbols), an interaction between mouse genes was converted to an interaction between their human one-to-one orthologs and vice versa (mainly via NCBI HomoloGene, additional one-to-one orthology relations were retrieved from ENSEMBL via *biomaRt*). For all data sources, interactions were only kept if both genes participating in an interaction were annotated in the database of the *org.Hs.eg.db* R package. This database was also used to map Entrez ID, Uniprot and ENSEMBL identifiers to gene symbols if the original network did not provide interactions in gene symbols.

Supplementary Note 14: Comparison between model-based optimization and NSGA-II

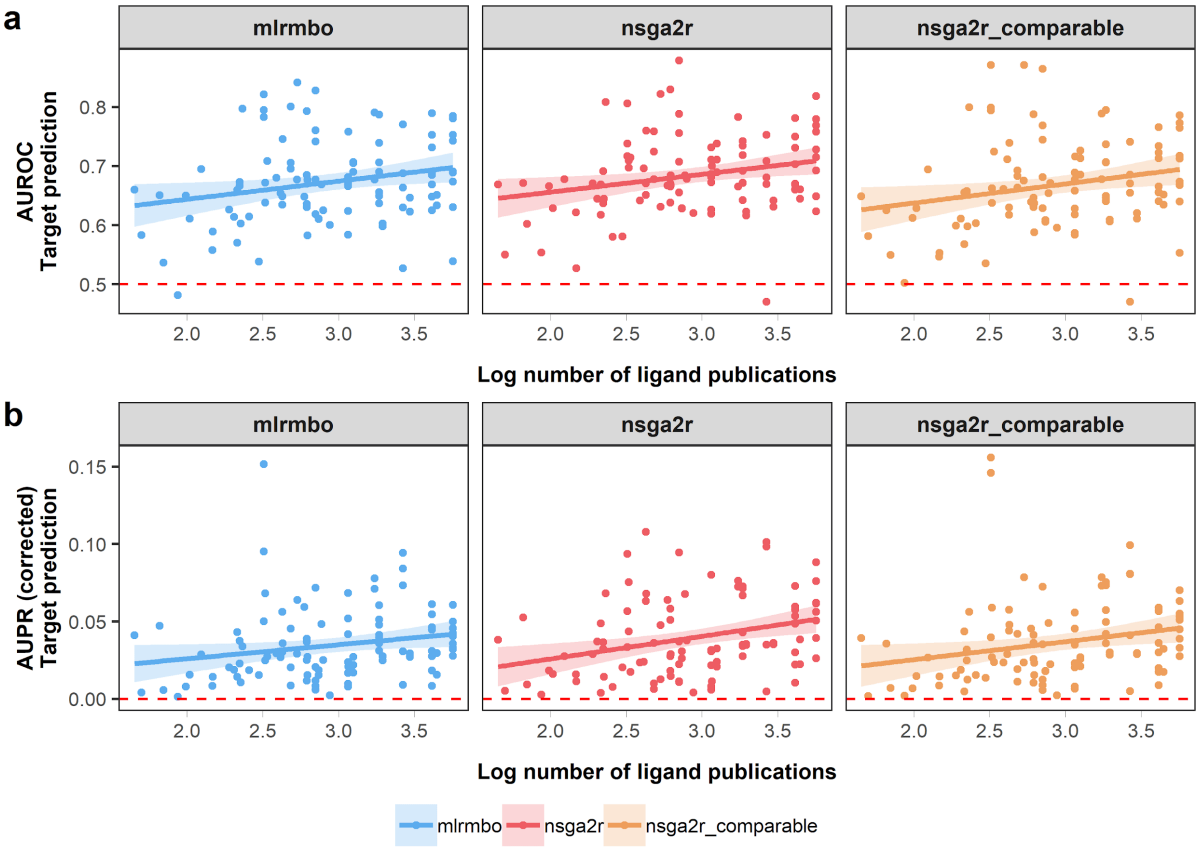
For parameter optimization, we also tested multi-objective optimization via NSGA-II[48] to compare with optimization via mlrMBO. The optimization scheme and evaluation strategy were the same as for optimization via mlrMBO (Online Methods: "Parameter optimization via mlrMBO"). For NSGA-II optimization with a population size of 120, we once selected the optimal parameter setting after 12 iterations, and once after 60 iterations. Selecting a parameter setting after 12 iterations was done to calculate performance of NSGA-II for which the same number of parameter designs were evaluated as during the mlrMBO optimization. Selecting a parameter setting after 60 iterations was done to assess the possible benefit of optimizing over more iterations.

We compared performance and popularity bias in target gene and ligand activity prediction between an mlrMBO-optimized model and the two NSGA-II-optimized models. Conclusion is that the 3 different models have very similar performance, with slightly better performance in target prediction for NSGA-II, and better performance in ligand prediction for mlrMBO (**Supplementary Notes Figures 28-31**). The geometric average of the 4 optimization objectives (i.e., mean of AUROC and AUPR target prediction and median of AUROC and AUPR ligand activity prediction; evaluated on the test sets) is 0.220 for mlrMBO, 0.213 for NSGA-II-60-iterations and 0.207 for NSGA-II-12-iterations. Noteworthy, the NSGA-II-60-iterations seems to be more popularity biased with respect to the ligand prediction performance (**Supplementary Notes Figure 31**).

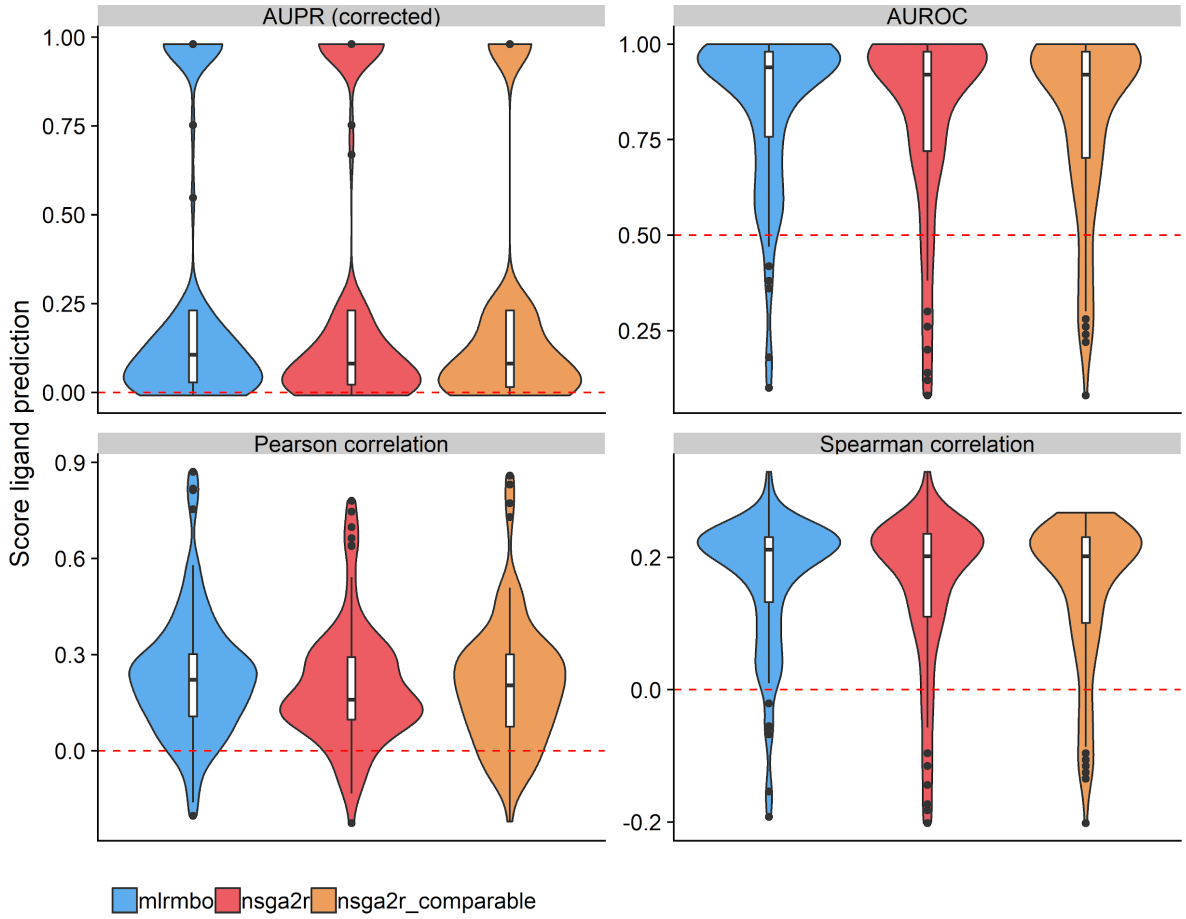
Because NSGA-II did not result in a better model than mlrMBO and the mlrMBO R software package is better documented, maintained and flexible to use than the nsga2R R package, we decided to use mlrMBO. Nevertheless, NSGA-II could have been a valid choice for optimization as well.



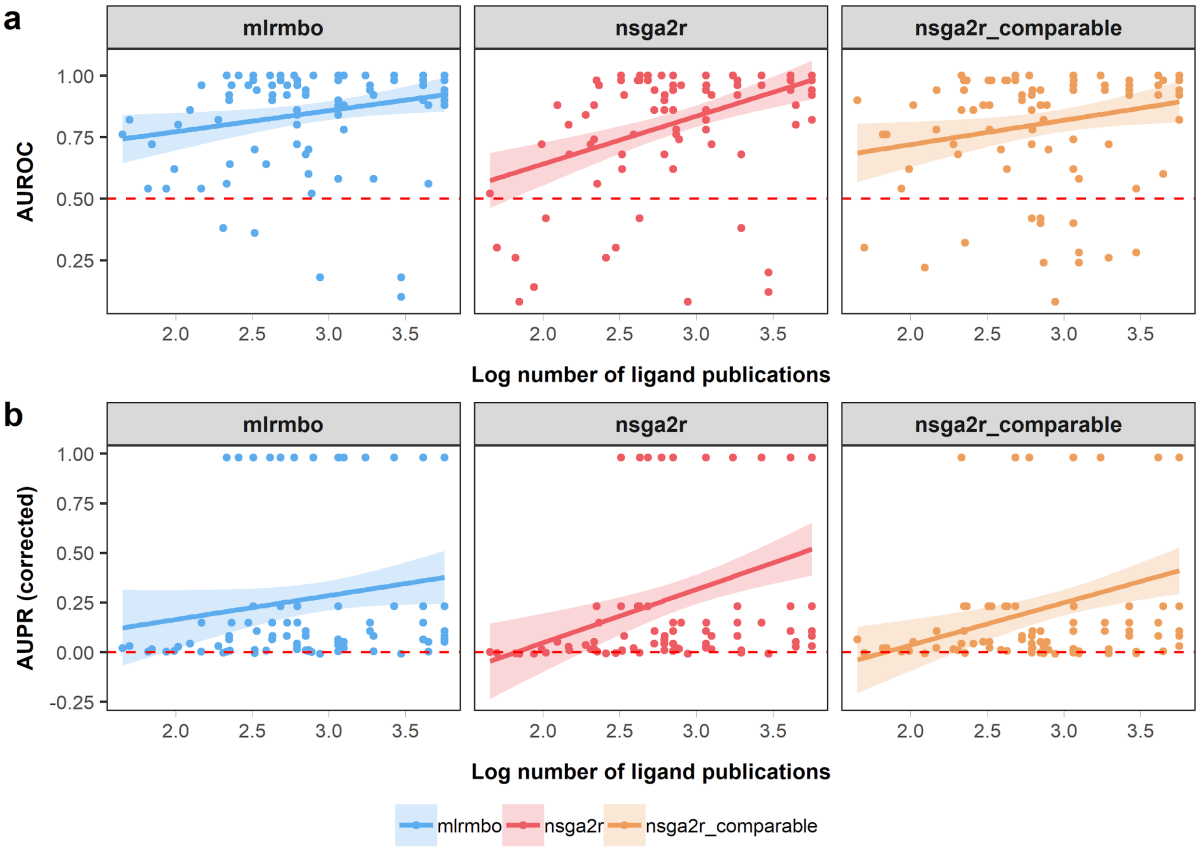
Supplementary Notes Figure 28 | Evaluation of the performance of NicheNet in predicting the transcriptional response: comparison between different optimization methods. For all 111 ligand treatment expression datasets, we compared the performance of several optimized variants of NicheNet in predicting the transcriptional response to a ligand. Following optimization variants were compared: 1) mlrmbo: 1440 evaluated parameter designs: optimization started with 240 parameter designs and during each of 50 subsequent iteration rounds, 24 new designs were proposed and evaluated; 2) nsga2r: 7200 evaluated parameter designs and 60 iterations; 3) nsga2r_comparable: 1440 evaluated parameter designs and 12 iterations. AUPR(corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under the receiver operating characteristic curve. Boxplot elements: center line: median; box limits: upper and lower quartiles; whiskers: 1.5x interquartile range; points: outliers. The violin plots show the probability density of the data (tails of the violins are trimmed to the range of the data). Red dashed line: performance of random guessing.



Supplementary Notes Figure 29 | Popularity bias of NicheNet in predicting the transcriptional response to ligands in ligand treatment datasets: comparison between different optimization methods. For 101 ligand treatment expression datasets profiling the transcriptional response to a single ligand, we plotted the performance of several optimized variants of NicheNet in predicting the transcriptional response to a ligand in function of its popularity (the decimal logarithm of the number of studies in PubMed in which a ligand was described). Following optimization variants were compared: 1) mlrmbo: 1440 evaluated parameter designs: optimization started with 240 parameter designs and during each of 50 subsequent iteration rounds, 24 new designs were proposed and evaluated; 2) nsga2r: 7200 evaluated parameter designs and 60 iterations; 3) nsga2r_comparable: 1440 evaluated parameter designs and 12 iterations. AUPR(corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under the receiver operating characteristic curve. The smoothing lines are the result of fitting a linear regression model. The accompanying shaded region indicates the 95% confidence interval. Red dashed line: performance of random guessing.



Supplementary Notes Figure 30 | Evaluation of the performance of NicheNet in predicting the ligand activity: comparison between different optimization methods. For all 111 ligand treatment expression datasets, we compared the performance of several optimized variants of NicheNet in predicting the activity state of a ligand. Following optimization variants were compared: 1) mlrmbo: 1440 evaluated parameter designs: optimization started with 240 parameter designs and during each of 50 subsequent iteration rounds, 24 new designs were proposed and evaluated; 2) nsga2r: 7200 evaluated parameter designs and 60 iterations; 3) nsga2r_comparable: 1440 evaluated parameter designs and 12 iterations. AUPR(corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under the receiver operating characteristic curve. Boxplot elements: center line: median; box limits: upper and lower quartiles; whiskers: 1.5x interquartile range; points: outliers. The violin plots show the probability density of the data (tails of the violins are trimmed to the range of the data). Red dashed line: performance of random guessing.



Supplementary Notes Figure 31 | Popularity bias of NicheNet in predicting ligand activity: comparison between different optimization methods. For 101 ligand treatment expression datasets profiling the transcriptional response to a single ligand, we plotted the performance of several optimized variants of NicheNet in predicting the activity state of a ligand in function of its popularity (the decimal logarithm of the number of studies in PubMed in which a ligand was described). Following optimization variants were compared: 1) mlrmbo: 1440 evaluated parameter designs: optimization started with 240 parameter designs and during each of 50 subsequent iteration rounds, 24 new designs were proposed and evaluated; 2) nsga2r: 7200 evaluated parameter designs and 60 iterations; 3) nsga2r_comparable: 1440 evaluated parameter designs and 12 iterations. AUPR(corrected): area under the precision-recall curve, corrected for random prediction; AUROC: area under the receiver operating characteristic curve. The smoothing lines are the result of fitting a linear regression model. The accompanying shaded region indicates the 95% confidence interval. Red dashed line: performance of random guessing.

Supplementary Note 15: Motivation for the use of random forest for target gene prediction via multi-ligand classification models

To assess how well ligands can predict which genes belong to a gene set of interest, we did not only look at prediction performances for individual ligands, but we also trained classification models that use the NicheNet ligand-target regulatory potential scores of all potentially active ligands as features to predict the gene set of interest. We tested performance of several popular classifiers on the gene sets from both discussed case studies: random forest, Linear Discriminant Analysis, logistic regression, Naive Bayes classification and Partial Least Squares. Random forest most often performed best and was always among the best performers (**Supplementary Notes Tables 2 and 3**).

Supplementary Notes Table 2 | p-EMT target gene set prediction performances (averaged over of five rounds of 5-fold cross-validation)

Classifier	AUROC - all ligands	Pearson correlation- all ligands	AUROC - top20 ligands	Pearson correlation - top20 ligands
Random forest	0.726	**0.167**	**0.744**	**0.176**
Linear discriminant analysis	**0.739**	0.123	0.638	0.086
Logistic regression	0.725	0.154	0.671	0.084
Naive Bayes Classifier	0.688	0.090	0.729	0.136
Partial Least Squares	0.683	0.096	0.632	0.039

Supplementary Notes Table 3 | Anti-viral response gene set target prediction performances (averaged over of five rounds of 5-fold cross-validation)

Classifier	AUROC - all ligands	Pearson correlation- all ligands	AUROC - top20 ligands	Pearson correlation - top20 ligands
Random forest	0.793	0.296	**0.795**	0.307
Linear discriminant analysis	**0.850**	**0.389**	0.781	0.329
Logistic regression	0.795	0.343	0.753	**0.343**
Naive Bayes Classifier	0.659	0.045	0.743	0.074
Partial Least Squares	0.766	0.104	0.774	0.116

References

1. Gillis, J., Ballouz, S. & Pavlidis, P. Bias tradeoffs in the creation and analysis of protein-protein interaction networks. *J. Proteomics* **100**, 44–54 (2014).
2. Stoeger, T., Gerlach, M., Morimoto, R. I. & Amaral, L. A. N. Large-scale investigation of the reasons why potentially important genes are ignored. *PLOS Biol.* **16**, e2006643 (2018).
3. Lee, I., Date, S. V., Adai, A. T. & Marcotte, E. M. A Probabilistic Functional Network of Yeast Genes. *Science* **306**, 1555–1558 (2004).
4. Bischl, B. *et al.* mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions. *ArXiv170303373 Stat* (2017).
5. Marbach, D. *et al.* Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods* **13**, 366–370 (2016).
6. Page, L., Brin, S., Motwani, R. & Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web.* (Stanford InfoLab, 1999).
7. Iván, G. & Grolmusz, V. When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks. *Bioinformatics* **27**, 405–407 (2011).
8. Dijkstra, E. W. A Note on Two Problems in Connexion with Graphs. *Numer Math* **1**, 269–271 (1959).
9. Krämer, A., Green, J., Pollard, J. & Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30**, 523–530 (2014).
10. Choi, H. *et al.* Transcriptome analysis of individual stromal cell populations identifies stroma-tumor crosstalk in mouse lung cancer model. *Cell Rep.* **10**, 1187–1201 (2015).
11. Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e24 (2017).
12. Medaglia, C. *et al.* Spatial reconstruction of immune niches by combining photoactivatable reporters and scRNA-seq. *Science* eaao4277 (2017). doi:10.1126/science.aao4277
13. Cox, M. A., Kahan, S. M. & Zajac, A. J. Anti-viral CD8 T cells and the cytokines that they love.

Virology **435**, 157–169 (2013).

14. Camp, J. G. *et al.* Multilineage communication regulates human liver bud development from pluripotency. *Nature* **546**, 533 (2017).
15. Pavličev, M. *et al.* Single-cell transcriptomics of the human placenta: inferring the cell communication network of the maternal-fetal interface. *Genome Res.* **27**, 349–361 (2017).
16. Zepp, J. A. *et al.* Distinct Mesenchymal Lineages and Niches Promote Epithelial Self-Renewal and Myofibrogenesis in the Lung. *Cell* **170**, 1134-1148.e10 (2017).
17. Cohen, M. *et al.* Lung Single-Cell Signaling Interaction Map Reveals Basophil Role in Macrophage Imprinting. *Cell* (2018). doi:10.1016/j.cell.2018.09.009
18. Vento-Tormo, R. *et al.* Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* **563**, 347 (2018).
19. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw1092
20. Pawson, A. J. *et al.* The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic Acids Res.* **42**, D1098–1106 (2014).
21. Rouillard, A. D. *et al.* The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016**, baw100 (2016).
22. Ramilowski, J. A. *et al.* A draft network of ligand–receptor-mediated multicellular signalling in human. *Nat. Commun.* **6**, 7866 (2015).
23. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
24. Kamburov, A., Stelzl, U., Lehrach, H. & Herwig, R. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.* **41**, D793–D800 (2013).
25. Li, T. *et al.* A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat. Methods* **14**, 61–64 (2017).
26. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-

curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).

27. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39**, D685–D690 (2011).
28. Lachmann, A. & Ma'ayan, A. KEA: kinase enrichment analysis. *Bioinformatics* **25**, 684–686 (2009).
29. Linding, R. *et al.* NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res.* **36**, D695–D699 (2008).
30. Duan, G., Li, X. & Köhn, M. The human DEPhOsphorylation database DEPOD: a 2015 update. *Nucleic Acids Res.* **43**, D531-535 (2015).
31. Van Landeghem, S. *et al.* Exploring Biomolecular Literature with EVEX: Connecting Genes through Events, Homology, and Indirect Associations. *Adv. Bioinforma.* **2012**, e582765 (2012).
32. Vinayagam, A. *et al.* A Directed Protein Interaction Network for Investigating Intracellular Signal Transduction. *Sci Signal* **4**, rs8–rs8 (2011).
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. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
35. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142–D147 (2014).
36. Matys, V. *et al.* TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
37. Xie, X., Rigor, P. & Baldi, P. MotifMap: a human genome-wide map of candidate regulatory motif sites. *Bioinformatics* **25**, 167–174 (2009).
38. Consortium, T. G. O. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049–D1056 (2015).
39. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide

ChIP-X experiments. *Bioinformatics* **26**, 2438–2444 (2010).

40. Liu, Z.-P., Wu, C., Miao, H. & Wu, H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* **2015**, bav095 (2015).
41. Bovolenta, L. A., Acencio, M. L. & Lemke, N. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics* **13**, 405 (2012).
42. Han, H. *et al.* TRRUST: a reference database of human transcriptional regulatory interactions. *Sci. Rep.* **5**, (2015).
43. Painter, M. W., Davis, S., Hardy, R. R., Mathis, D. & Benoist, C. Transcriptomes of the B and T lineages compared by multi-platform microarray profiling. *J. Immunol. Baltim. Md 1950* **186**, 3047–3057 (2011).
44. Jojic, V. *et al.* Identification of transcriptional regulators in the mouse immune system. *Nat. Immunol.* **14**, 633–643 (2013).
45. Griffon, A. *et al.* Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res.* **43**, e27–e27 (2015).
46. Tang, Q. *et al.* A comprehensive view of nuclear receptor cancer cistromes. *Cancer Res.* **71**, 6940–6947 (2011).
47. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100**, 9440–9445 (2003).
48. Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **6**, 182–197 (2002).