Subject Section

Identifying signaling genes in spatial single cell expression data

Dongshunyi Li¹, Jun Ding¹ and Ziv Bar-Joseph¹,²,*

¹Computational Biology Department and ²Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, United States.

* To whom correspondence should be addressed.

Abstract

Motivation: Recent technological advances enable the profiling of spatial single cell expression data. Such data presents a unique opportunity to study cell-cell interactions and the signaling genes that mediate them. However, most current methods for the analysis of this data focus on unsupervised descriptive modeling, making it hard to identify key signaling genes and quantitatively assess their impact.

Results: We developed a mixture of experts for spatial signaling genes identification (MESSI) method to identify active signaling genes within and between cells. The mixture of experts strategy enables MESSI to subdivide cells into subtypes. MESSI relies on multi-task learning using information from neighboring cells to improve the prediction of response genes within a cell. Applying the methods to three spatial single cell expression datasets, we show that MESSI accurately predicts the levels of response genes, improving upon prior methods and provides useful biological insights about key signaling genes and subtypes of excitatory neuron cells.

Availability: MESSI is available at: https://github.com/doraadong/MESSI

Contact: zivbj@cs.cmu.edu

1 Introduction

The ability to determine gene expression at the single cell level led to several novel findings and insights (Tanay and Regev (2017), Cembrowski et al. (2020)). These techniques, which primarily rely on combinatorial in situ hybridization (FISH) or in situ sequencing, can determine the levels of hundreds to thousands of genes for each cell while still retaining information on the location of the cells. Such information opens the door to much more detailed and accurate analysis of the impact of spatial location on cell-cell interactions and gene expression (Lein et al. (2017)).

There have been a number of methods that attempted to infer cell-cell interactions from scRNA-Seq data. Since such data does not contain spatial information, these methods have mainly focused on the co-expressed ligand-receptor pairs between cell types (Cabello-Aguilar et al. (2020), Efremova et al. (2020), Wang et al. (2019), Tsuyuzaki et al. (2019), Kumar et al. (2018), Zhou et al. (2017)). However, given the data used, it was hard to determine if the cell types identified as interacting were indeed close in 2D or 3D space. A number of computational methods have been proposed to analyze and model spatial single cell data. To date, most methods for the analysis of such data are unsupervised focusing on clustering or representation to infer spatial patterns for cells, genes and transcripts (Zhu et al. (2018), Edsgärd et al. (2018), Svensson et al. (2018), Sun et al. (2020)). Very few methods attempted to study interactions between cells using spatial single cell expression data. For example, Giotto (Dries et al. (2019)) identifies potential interacting neighboring cell type pairs by testing if genes in a particular cell type are differentially expressed when adjacent to another cell type. While Giotto provided some information on the set of interacting genes and proteins, these models are descriptive rather than predictive. Due to the lack of ground truth regarding ligand-receptor interactions in specific tissues or cell types, it is hard to evaluate the accuracy of such methods in a comprehensive manner and to determine whether they indeed capture the true interactions or instead just overfit the observed expression values.

To better infer the set of signaling genes within and between neighboring cells we developed a framework, Mixture of Experts for Spatial Signaling genes Identification (MESSI), that utilizes Mixture of Experts (MoE) and multi-task learning to jointly model interactions within...
and between cells from spatial single cell data. For each cell type in the data, the MESSI model uses as input a subset of inter- and intra-signaling genes to predict the expression of a set of response genes. As we show, a major advantage of the MoE framework for modeling single cell data is its ability to account for subtypes. The use of multi-task learning further enables the sharing of information among response genes via joint learning of response genes’ covariance matrices.

To test our models, we perform cross-validation analysis in which we learn on a subset of the data (in our case, a subset of the animals profiled in different studies) and test on other animals. Once confirmed, the coefficients assigned by the model to the different signaling genes are analyzed to identify key signaling genes that mediate cell-cell communication and to identify cell subtypes that differ in these and their impact on response genes’ expression.

We applied our MESSI framework to three spatial single cell datasets. As we show, the method can accurately predict gene expression by utilizing neighboring cells’ information and improves upon several other methods suggested for expression prediction in spatial and non-spatial analysis settings. Analysis of the resulting experts for one of the datasets identifies an important signaling gene that cannot be determined based on expression alone, and that can be used to define an important excitatory neuron subtype.

2 Materials and Methods

2.1 Datasets and data pre-processing

We tested MESSI using three spatial transcriptomics datasets: the MERFISH hypothalamus data, the MERFISH U-2 OS data, and the STARmap mPFC data. The MERFISH hypothalamus data (Moffit et al. (2018)) profiled 11 million cells in the preoptic region of the mouse hypothalamus. For each cell, the expression for 155 genes was measured using a single-molecule imaging method based on combinatorial or sequential FISH labeling. Multiple animals under different behavioral conditions were profiled, each with roughly 20K to 70K cells. In our analysis, we used all genes profiled and all naive (4), parenting (4), and virgin parenting (5) female animals that were studied. The MERFISH U-2 OS data (Xia et al. (2019)) is a single cell-line dataset profiling 1368 cells in three batches. For each cell, the expression of 10050 genes is measured. We selected the most significantly DE (differentially expressed) genes in the clusters identified by Xia et al. (2019) for the MESSI modeling resulting in 742 DE genes. The STARmap data includes cells from the mouse medial prefrontal cortex (mPFC) profiled using an in situ sequencing method (Wang et al. (2018)). We used the three control samples profiled by STARmap, each containing between 1100 and 1300 cells. STARmap measures the values of 166 genes and we used all of them. See Supplementary Methods for details on how the data was processed and genes were selected.

To identify potential signaling genes, we used genes determined to be ligands or receptors, including those from Ramilowski et al. (2015). See Supplementary Methods and Table S1 for detailed information on the set of genes we used. To identify neighboring cells, we used Delaunay triangulation with a specific distance cutoff (for example, 100 micrometers for the MERFISH hypothalamus dataset, Figure 1).

2.2 Inferring inter- and intra- gene-gene interactions

Our goal is to identify genes encoding signaling molecules (termed signaling genes) used for cell-cell interactions. For this, we developed a Mixture of Experts for Spatial Signaling genes Identification (MESSI) framework that aims to predict the expression of genes (termed response genes) in a specific cell type using a combination of intra-cellular signaling molecules, specifically the ligands or receptors produced by the cell itself, and the expression of ligands in neighboring cells (also referred as the intercellular signaling genes). Figure 1 provides a high-level overview of our MESSI framework. Our method utilizes the expression of genes in neighboring cells (dark blue in the zoom-in view of Figure 1) and the expression of genes within a cell (green and light blue) to predict a subset of response genes. Predictions are performed by a set of experts that correspond to potential subtypes of cells learned by the model. Cell assignment to subtypes (experts) is determined by the signaling genes and the additional spatial information (e.g., neighboring cell types and spatial location). Below we discuss in detail how to formulate the MESSI optimization problem, how to learn parameters for the model, and how we evaluated its performance and extract biological insights from the model learned.

2.3 Mixture of Experts for Spatial Signaling genes Identification (MESSI) framework

We extended and employed a Mixture of Experts (MoE) framework proposed by Jacob and Jordan (Jordan and Jacobs (1994)) for spatial expression prediction. Denote the response gene expression vector by $y$ (with $K$ entries where $K$ is the number of response genes to be predicted). Denote the input features as $x$, which is a $D$ dimensional vector (so $D$ is the total number of input features). For each cell $i$, we denote by $y_i$ the expression values of its response genes and by $x_i$ the features for this cell. The features we considered are the expression of ligands and receptors’ in cell $i$, the expression of ligands in cell $i$’s neighbors, cell $i$’s spatial coordinates (location within the image/tissue) and cell types of its neighbor (we used the cell type annotations from the original studies). Given these definitions, for $N$ cells, we construct the output $Y(N \times K)$ and input $X(N \times D)$ matrices. Assuming that given the expression of ligands in neighboring cells, and the expression of signaling genes within the cell, the expression of response genes in a cell is independent of other cells, we can write the total conditional likelihood as $p(Y|X) = \prod_{i=1}^{N} p(Y_i|X_i)$.

MoE models assume that these conditional distributions are a mixture of conditional distributions produced by different experts (in our case, cell subtypes). For each cell $i$, we use $z_i$ as a categorical variable to indicate the expert (“subtype”) that generates cell $i$. Under the mixture assumption of MoE, we have for each cell $i$:

$$p(y_i|x_i) = \sum_{j=1}^{J} p(y_i|z_i = j|x_i)$$

$$= \sum_{j=1}^{J} \prod_{i=1}^{N} p(y_i|z_i = j, x_i) p(z_i = j|x_i)$$

where $J$ is the number of experts and $j$ is the $j$th expert. In other words, we can express the conditional likelihood as a weighted average of conditional likelihoods $p(y_i|x_i = j, x_i)$ produced by each $j$th expert. Note that here the contribution (weight) from expert $j$ for cell $i$, $p(z_i = j|x_i)$, is also dependent on the feature variables $x_i$.

This contribution is determined by a classifier (gate).

2.3.1 Multi-output learning model for experts

Given that we are trying to predict multiple genes, we would like to use multi-output models for each expert. These models take into account the dependency among different tasks and so are more suitable for predicting the expression of multiple genes, some of which are likely co-expressed or co-regulated, than methods assuming conditional independence. For this, we extended the MoE framework such that each expert uses a weighted version of the Multiple-output Regression with Output and Task Structures (MROTS) model (Rai et al. (2012)). For a single cell $i$, the conditional
likelihood described by an expert \( j \), as modeled by MROTS, is:
\[
p(y | z = j, x) = N(y | W^T_j x + b_j, \Omega_j, \Sigma_j)
\]
In other words, it assumes a multivariate normal distribution with \( W^T_j x + b_j \) as the mean and \( \Omega_j, \Sigma_j \) as the covariance. Here \( W_j \in \mathbb{R}^{D \times K} \) and \( b_j \in \mathbb{R}^K \) are model coefficients and intercepts. In addition, the model also assumes a prior on the distribution of the coefficients \( W_j \):
\[
p(W_j) = \prod_{i=1}^K N(w_{ij} | 0, I_D) \mathcal{M}_D \times K \{0D \times K; I D, \Sigma_j\}
\]
where \( \mathcal{M} \) denotes the matrix normal distribution. This distribution for \( W_j \) couples the coefficients of different tasks by assuming a covariance matrix \( \Sigma_j \) among tasks. The other component of the prior, \( N(w_{ij} | 0, I_D) \), can be viewed as a \( I_D \) normalization on the weights \( w_{ij} \) for each individual task \( k \). The covariance matrices \( \Omega_j, \Sigma_j \) each characterizes different aspects of the dependence among tasks, \( \Omega_j \) for the conditional covariance among the outputs and \( \Sigma_j \) for the marginal covariance among the coefficients.

2.4 Learning and inference for MESSI

Learning for MESSI is performed using an Expectation–maximization (EM) algorithm which is commonly applied to Mixture of Experts (MoE) models (Jordan and Jacobs (1994)). The algorithm aims to maximize:
\[
Q(q, \theta) = \sum_{q(z)} q(z) \log p(x, Y | X, \theta) - \sum_{q(z)} q(z) \log q(z),
\]
a lower bound of the log conditional likelihood \( \log p(Y | X) \). Here, \( x \) is the membership vector of size \( N \times 1 \), \( q(z) \) is the unknown distribution of \( z \) and \( \theta \) is the collection of all model parameters. Since the distribution \( q(z) \) is unknown, in the E-step of iteration \( t \), for each cell \( i \), we first postulate a distribution for \( z_i \) based on the parameters \( \theta^{t-1} \) learned from the previous iteration. See Supplementary Methods for the complete derivation.

Next, we perform the M-step where we maximize \( Q(q, \theta) \) w.r.t. \( \theta \). Given the modularity of the MoE framework, the M-step can be sub-divided into separate learning tasks for each expert and classifier (Supplementary Methods). For each expert, we derived and implemented an alternating minimization algorithm for a weighted version of MROTS with \( h_j(i) \) as the weight of expert \( j \) for cell \( i \). We use logistic regression for the classifier. The learning algorithm iterates between the E-step and M-step until convergence. For a new sample \( x \in \mathbb{R}^D \), the prediction \( \hat{y} \) is based on the weighted average of the predictions made by each single expert. The weight is given by the trained classifier based on the input \( x \), indicating how likely the new sample comes from an expert. See Supplementary Methods for a detailed derivation of the training algorithm and how the model makes predictions.

To learn hyper-parameters, for example, the number of experts, we employ a nested cross-validation (CV) strategy where the inner loop selects the best model, and the outer loop evaluates the performance of the model on a left-out animal/replicate. For each iteration of the inner loop, we split the training data into a training and validation set and conduct a grid search for the values of the hyperparameters using the validation set. After looping through all training and validation sets pairs, we select the best values for the hyperparameters using majority vote. We then retrain using the entire training dataset using the selected combination of hyperparameter values, and the trained model is then applied to the left-out data. See Supplementary Methods for more details.

2.5 Baseline and comparison methods used to benchmark MESSI

The baseline model we compared to only uses information about the location of the cell and its type to predict response gene expression. Thus, it does not use any information from neighbor cells. All other methods we compared to used the same set of input features as MESSI. Chen et al. (2016) describes that researchers in LINCS program (http://www.lincsproject.org/) predict gene expressions by building a linear regression model for each response gene assuming response genes are conditionally independent. XGBoost (Li et al. (2019), Chen and Guestrin (2016)) uses a boosting idea for the prediction. Chen et al. (2016) learns a multi-layer perceptron (MLP) model. We consider both single- and multi-node output MLP models, where the latter enables learning across response genes. See Supplementary Methods for implementation details of these methods.
3 Results

3.1 MESSI achieved high prediction accuracy compared to baseline and comparison methods

We first tested our model on the MERFISH data of the mouse hypothalamus preoptic region (Moffitt et al., 2018), where for each cell, the number of RNA molecules of a set of marker genes was measured by multiplexed error robust fluorescence in situ hybridization (MERFISH). The MERFISH hypothalamus data contains 11 million cells, each with 155 genes profiled from multiple animals in both sexes under naïve and several different behavioral conditions, see Methods for details. To determine the ability of our method to accurately predict gene expression, we performed nested cross-validation (CV) analysis to learn model parameters and hyper-parameters (See Table S3 for the selected hyperparameter values). We next used the learned model to predict each cell’s expression in the samples of the left-out animal and compared the predicted results to the actual expression levels to determine the accuracy of the model. We compared MESSI’s prediction accuracy with a baseline model that only uses cell types and spatial locations of the cells, but not the expression of genes. We also compared MESSI to regression models that have been previously used to predict gene expression based on a subset of genes for non-spatial data. These include linear regression (LR), multi-layer perceptron (MLP) with single- or multi-node output (Chen et al., 2016) and XGBoost which used a boosting algorithm to learn regression trees (Li et al., 2019). Finally, we compared MESSI to the Multiple-output Regression with Output and Task Structures (MROTS) model (Rai et al., 2012), which, unlike MESSI, uses a single expert rather than a Mixture of Experts.

Results are presented in Figure 2. As can be seen, MESSI performed best, achieving the lowest mean absolute error (MAE) when averaged over cells from eight cell types (median of MAE: MESSI: 0.3568 vs. 0.3607 for MLP single and 0.3596 for XGBoost, the next 3 performed worse). Cell type specific results are shown in Figure 3. Again, MESSI performed best for the four major cell types. For example, for excitatory neurons, the median of MAE for MESSI is 0.3757 vs. 0.3801 for MLP single and 0.3824 for XGBoost. These improvements over other models are significant, as shown in Figure S12 and S13. In contrast, the baseline model performed the worst for all cell types (the median of MAE averaged over eight cell types is 0.5477). This indicates that while cell type and location are important features, the specific expression of genes in a cell is significantly affected by intra- and inter-signaling molecules. See Figures S18-S21 for the other cell types.

To test the impact of neighboring cells, we repeated the analysis above using the same methods, but this time only relying on the expression of intra-cellular signaling genes (i.e., genes in the same cell). Comparison between a model that uses neighboring cell expression levels and one that does not use them is presented in Figures 2 and Figure 3. As can be seen, the application of models that ignore the expression of signaling genes in neighboring cells results in significantly higher MAE for all methods. Similarly, all top-performing methods showed improved prediction performance when using intra-cellular signaling genes for most of the major cell types (Figure 3). For example, for excitatory neurons, the median of MAE using neighboring genes versus not is 0.3757 versus 0.3789 for MESSI, 0.3824 versus 0.3853 for XGBoost, and 0.3801 versus 0.3814 for MLP single.

3.2 Impact of sample size on performance

While MESSI performed well on cell types for which many cells were profiled (for example, for the 53K inhibitory cells), it was not the top performer for other cell types (for example, the 6K endothelial 1 cells). We hypothesized that reduction in performance is not due to the cell type...
identity but rather to the lack of data, which may lead to overfitting due to
the increase in the number of parameters being fitted by MESSI. To
test this, we down-sampled the largest two cell types profiled (inhibitory
to excitatory) and compared the performance when training on a reduced
number of cells to performance discussed above when using the full set for
training. In the smaller training sets, we used a training sample size ranging
from 0.7K to 6K compared to 19K to 55K for the full set. Results for
such down-sampling are presented in Figure 4. While MESSI outperforms
XGBoost when using all available data (Figure 3), when using the reduced
dataset, XGBoost is better. More generally, using more samples resulted
in an obvious decrease in prediction error for all methods, except for
the baseline model. Thus, when enough cells are sampled, models that
take into account the heterogeneity within cell types perform better than
simple regression models relying only on cell types and spatial locations.

3.3 Testing MESSI on additional datasets

To test the generality of our method, we further tested MESSI on two
other spatial cell expression datasets, both with a smaller sample
size. The first is another dataset profiled by MERFISH that only focused
on a single cell type (U-2 OS cell line) (Xia et al. 2019). This dataset
included about 13K cells though many more genes (about 10K) were
profiled for each cell. The second was from a method called STARmap,
which profiled mouse medial prefrontal cortex (mPFC) cells (Wang et al.
2018). Again, this dataset was much smaller than the first MERFISH
dataset, with about 2K cells profiled from the biggest cell type: excitatory
neurons. STARmap provides information on the expression of 166 genes
per cell. Since we use the prediction accuracy of response genes to validate
the signaling genes and subtypes MESSI learned, we have also assessed
MESSI’s performance when using different sizes of response gene sets.
For this, we used the smaller MERFISH data and varied the number of top
DE genes selected as response genes between 38 and 693. The nested CV
resulted in the selection of a single expert (essentially equivalent to the
MROTS method) for all responses sets. Same as for the first MERFISH
data, MESSI performed best for this data as well, as shown in Figure 5
when using 38 responses and Figure S7 for all settings (see Supplementary
Results 2.1 for details). MLP single-node output with the number of hidden
nodes equal to 2 performed much worse than the other models and are not
shown in these plots (see Supplementary Results Figure S10). For the

Fig. 4. Impact of training sample size on accuracy. We down-sampled inhibitory
and excitatory cells to compare the performance of methods when using a smaller dataset.
While MESSI and MLP single performed better than XGBoost when using all available data,
using the reduced dataset for training likely led to overfitting for these two methods.
MLP single: MLP with single-node output.

Fig. 5. CV results of the MERFISH U-2 OS cell line and the STARmap data. Left: Results for
the analysis of the MERFISH U-2 OS cells data. Right: Results for the analysis of the
STARmap data. ***: p-value below 1e-3; $: more than half of the CV groups show non-
significant improvement when using MESSI; MLP single: MLP with single-node output.

Fig. 6. Comparison of naive and behavior models when the behavior is virgin parenting.
Results are presented for the four major cell types. Naive model – predictions based on
the naive model. Behavior raw – predictions based on learning from the raw values in the
corresponding behavioral samples. Behavior change - predictions based on learning from
the raw values in the corresponding behavioral samples subtracted by the predictions from
the naive model. See Supplementary Methods for details. ***: p-value below 1e-3; **: p-
value below 1e-2; *: p-value below 5e-2; ---: p-value larger than 5e-2; $: more than half
of the CV groups show non-significant improvement when using behavioral models.

3.4 MESSI identified functional changes in neuron
subtypes between conditions

Given MESSI’s ability to accurately predict gene expression based on
signaling genes, we used MESSI to explore cell type specific functional
changes in signaling networks related to behavioral changes. For this, we
compared MESSI models learned from naive female animals, with models
learned from virgin females and mothers subjected to behavior stimuli as
described in Moffitt et al. (2018). Following Moffitt et al. (2018), we refer
to these samples as virgin parenting and parenting samples.

We first compared the performance of models learned using the naive
or behavioral training samples on the task of predicting expression values

<table>
<thead>
<tr>
<th>Model</th>
<th>Behavior</th>
<th>Raw Behavior</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>MESSI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XGBoost</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Summary of classification performance when using the reduced dataset
from 0.7K to 6K compared to 19K to 55K for the full set. Results for
the naive model. Behavior raw – predictions based on learning from the raw values in the
corresponding behavioral samples. Behavior change - predictions based on learning from
the raw values in the corresponding behavioral samples subtracted by the predictions from
the naive model. See Supplementary Methods for details. ***: p-value below 1e-3; **: p-
value below 1e-2; *: p-value below 5e-2; ---: p-value larger than 5e-2; $: more than half
of the CV groups show non-significant improvement when using behavioral models.

Fig. 5. CV results of the MERFISH U-2 OS cell line and the STARmap data. Left: Results for
the analysis of the MERFISH U-2 OS cells data. Right: Results for the analysis of the
STARmap data. ***: p-value below 1e-3; $: more than half of the CV groups show non-
significant improvement when using MESSI; MLP single: MLP with single-node output.

3.3 Testing MESSI on additional datasets

To test the generality of our method, we further tested MESSI on two
other spatial cell expression datasets, both with a smaller sample
size. The first is another dataset profiled by MERFISH that only focused
on a single cell type (U-2 OS cell line) (Xia et al. 2019). This dataset
included about 13K cells though many more genes (about 10K) were
profiled for each cell. The second was from a method called STARmap,
which profiled mouse medial prefrontal cortex (mPFC) cells (Wang et al.
2018). Again, this dataset was much smaller than the first MERFISH
dataset, with about 2K cells profiled from the biggest cell type: excitatory
neurons. STARmap provides information on the expression of 166 genes
per cell. Since we use the prediction accuracy of response genes to validate
the signaling genes and subtypes MESSI learned, we have also assessed
MESSI’s performance when using different sizes of response gene sets.
For this, we used the smaller MERFISH data and varied the number of top
DE genes selected as response genes between 38 and 693. The nested CV
resulted in the selection of a single expert (essentially equivalent to the
MROTS method) for all responses sets. Same as for the first MERFISH
data, MESSI performed best for this data as well, as shown in Figure 5
when using 38 responses and Figure S7 for all settings (see Supplementary
Results 2.1 for details). MLP single-node output with the number of hidden
nodes equal to 2 performed much worse than the other models and are not
shown in these plots (see Supplementary Results Figure S10). For the
Fig. 7. MESSI reveals changes in key signaling molecules and relevant signaling networks activated upon experience. Top a): Coefficients for top signaling molecules in a subset of the MESSI experts of excitatory cells for different conditions (X axis) for several response genes (Y axis). Note the large increase in Oxt coefficients between naive and parenting or virgin parenting models. See Supporting Methods for selection of top features. Bottom: Cells assigned to specific MESSI experts. b), c), e): spatial location of the cells on an example bregma; d): the proportion of interacting partners from each expert as indicated by the expression level of a selected neighboring feature.

for unseen behavioral samples. For cell types where the performance of these models is roughly the same, we can conclude that no major changes occur between the two conditions (since models trained on one perform well on the other). However, if we observe large differences in prediction abilities, then for these cell types, signaling molecules’ impact on the expression levels of the response genes are likely different. When using this approach for the MERFISH hypothalamus data, we observe a large improvement in the prediction ability for models learned using behavioral samples for inhibitory and excitatory cell types, while no obvious change in improvement in the prediction ability for models learned using behavioral samples. For cell types where the performance of these models is roughly the same, we can conclude that no major changes occur between the two conditions (since models trained on one perform well on the other).

MESSI indicates that these are likely oxytocin secreting magnocellular neurosecretory neurons located in the paraventricular nucleus (PV A) (Figure 7 b)).

A subtype of oxytocin secreting cells located in the same region was also identified by MESSI for another experimental condition, virgin parenting (subtype 7 in Figure 7 a) and c)). However, while we identified Oxt for both behaviors, the response genes influenced most from Oxt for the virgin parenting model are different from those predicted to be impacted in the parenting animals. This may imply differences in mechanisms for Oxt secretion regulation. Indeed, for virgin parenting, we found that a major impact for Oxt arises from neighboring cells, whereas for parenting, its main impact is internal (Figure 7 a)). Unlike many other neuropeptides released primarily from axons, oxytocin is known to be exclusively released within the PV A from dendrites or somata and to act on nearby neighboring cells (Pow and Morris (1989)) further corroborating our hypothesis about the identity of this subtype. The difference in the sign for the coefficients of neighboring oxytocin and intracellular oxytocin may result from a negative feedback loop that impacts the production of oxytocin in nearby neurons.

Compared to the results from a clustering-based approach applied in the original paper (Moffitt et al. (2018)), these two groups of cells (subtype 5 in parenting and subtype 7 in virgin parenting) corresponds to cluster E-9 (80% overlapped). In line with their differential genes analysis results, we also found the genes Ntng1, Cck as top features in the classifiers determining if a cell belongs to these two subtypes (Supplementary Figures S1 & S2).
This likely explains why Oxt was not identified in the original paper and why the GnRH cluster contains only very few cells (Moffitt et al. (2018)). In contrast, by looking at their impact on other genes in the network, MESSI was able to identify both as important signaling factors for the parenting/virgin parenting animals. Indeed, neuropeptides are known to be activated post-translationally, which can explain the relatively small changes in expression levels observed (Varrò (2007)). Another possible reason is that oxytocin is secreted in pulses for lactation and parturition (Fink et al. (2012)), and the dendritic release within PVA facilitates pulsatile secretion (Bealer et al. (2010)), and so it is not universally high in all excitatory cells. GnRH is another classic example of a pulsatile secretion hormone whose transcription level oscillates in an episodic pattern (Choe et al. (2013)). Thus, it is possible that methods that do not allow the analysis of gene-gene interactions in subtypes of cells (for example, those using clustering) miss such events, whereas MESSI can identify them using the different experts and their learned regression models. While MESSI can accurately predict gene expression and is able to reveal the activity of key signaling genes, there are a number of ways in which it can be further improved. First, we noticed that some signaling genes are highly correlated. Such collinearity results in reduced importance assigned to each of the correlated genes such that the model may miss identifying some of the key signaling genes. A possible way to mitigate this problem is through the use of a more stringent regularization penalty.


efremova, m., vento-tormo, m., teichmann, s. a., and vento-tormo, r. (2020).
li, w., yin, y., quan, x., and zhang, h. (2019). gene expression value prediction
jordan, m. i. and jacobs, r. a. (1994). hierarchical mixtures of experts and the em
kanasaki, h., mutiara, s., oride, a., purwana, i. n., and miyazaki, k. (2009).
pulse frequency-dependent gonadotropin gene expression by adenylyl cyclase-
activating polypeptide 1 in perfused mouse pituitary gonadotroph beta 2 cells.
biochemistry, 81(3), 465–472.
kumar, m. p., du, j., lapoudos, g., jiao, y., sawyer, a., drummond, d. c.,
lauffenburger, d. a., and rater, a. (2018). analysis of single-cell rna-seq
identifies cell-cell communication associated with tumour characteristics.
cell reports, 28(6), 1458–1468.
lein, e., born, l. e., and limansson, s. (2017). the promise of spatial
transcriptomics for neuroscience in the era of molecular cell typing.
scie, 358(6359), 64–69.
l, w., yin, y., quan, x., and zhang, h. (2019). gene expression value prediction
based on tsgod algorithm.
frontiers in genetics, 10, 1077.
differentially activates neurons expressing either fos or cocaine-and amphetamine-
regulated transcript in lactating, maternal rodents.
nature neuroscience, 8(5), 315–328.
moffitt, j. r., bambah-mukku, d., eichhorn, s. w., vaughn, e., shekhar, k., perez,
j. d., rabinstein, n. d., hao, j., reges, a., dalac, c., et al. (2010). molecular,
spatial, and functional single-cell profiling of the hypothalamic preoptic region.
science, 326(5955), 661–664.
pow, d. and morris, j. (1989). dendrites of hypothalamic magnocellular neurons
release neurohypophysial peptides by exocytosis.
nature neuroscience, 32(2), 435–439.
rai, p., kumar, a., and daume, h. (2012). simultaneously leveraging output and task
structures for multiple-output regression.
research, 7(1), 8815.
esources for multiple-output regression.
research, 25, pages 1855–1913.
curran associates, inc.
rakowski, j. a., goldberg, t., harbsburger, j., kloppmann, e., lizio, m.,
A draft network of ligand-receptor-mediated multicellular signalling in human.
nature communications, 6(1), 7866.
Rilling, J. K. and Young, L. J. (2014). the biology of mammalian parenting and its
effect on offspring social development.
science, 345(6198), 771–776.
straus, j. f. and barbieri, r. l. (2013). vos & jeff’s reproductive endocrinology
E-book: physiology, pathophysiology, and clinical management.
Elsevier health sciences.
for spatially resolved transcriptomic studies.
svensson, v., teichmann, s. a., and stingl, d. (2018). spatial: identification of
spatially variable genes.
tany, a. and reger, a. (2017). scaling single-cell genomics from phenomenology
to mechanism.
tosyuzaki, K., ishi, m., and nikaido, j. (2019). uncovering hypergraphs of cell-cell
interaction from single cell rna-sequencing data.
bioRxiv.
Transmitters.
American Cancer Society.
wang, X., allen, w. E., wright, M. A., Sylvester, E. L., samuisk, N., Vosina, S.,
intact-tissue sequencing of single-cell transcriptional states.
Science, 364(6450), eaat5691.
wang, Y., wang, R., zhang, S., Song, S., Jiang, C., Han, G., wang, M., ajani, J.,
Friel, a., and wang, L. (2019). tsk: an r package to characterize and illustrate
intracellular communication.
bioRxiv.
profiling by merfish reveals subcellular rna compartmentalization and cell cycle-
dependent gene expression.
proceedings of the national academy of sciences, 116(39), 19490–19499.
Extracting intracellular signaling network of cancer tissues using ligand-receptor
expression patterns from whole-tumor and single-cell transcriptomes.
Science reports, 7(1), 8815.
zhao, Q., shah, S., dries, R., cai, L., and yuan, G.-C. (2018). Identification of
spatially associated subpopulations by combining scRNAseq and sequential
fluorescence in situ hybridization data.

339–342.
Cellphone R: inferring cell-cell communication from combined expression of
multi-subunit ligand-receptor complexes.
Nature protocols, 15, 1844–1506.
Academic press.
algorithm.
Neural Computation, 6(2), 181–214.
Pulse frequency-dependent gonadotropin gene expression by adenylyl cyclase-
activating polypeptide 1 in perfused mouse pituitary gonadotroph beta 2 cells.
Kumar, M. P., Du, J., Lapoudos, G., Jiao, Y., Sawyer, A., Drummond, D. C.,
identifies cell-cell communication associated with tumor characteristics.
Cell reports, 28(6), 1458–1468.
transcriptomics for neuroscience in the era of molecular cell typing.
Science, 358(6359), 64–69.
Li, W., Yin, Y., Quan, X., and Zhang, H. (2019). Gene expression value prediction
based on tsgod algorithm.
Frontiers in Genetics, 10, 1077.
differentially activates neurons expressing either fos or cocaine-and amphetamine-
regulated transcript in lactating, maternal rodents.
Nature neuroscience, 8(5), 315–328.
Moffitt, J. R., Bambah-Mukku, D., Eichhorn, S. W., Vaughn, E., Shekhar, K., Perez,
J. D., Rabinstein, N. D., Hao, J., Reges, A., Dalac, C., et al. (2010). Molecular,
Spatial, and functional single-cell profiling of the hypothalamic preoptic region.
Science, 326(6241), eaau5324.
release neurohypophysial peptides by exocytosis.
structures for multi-output regression.
Research, 7(1), 8815.
Rakowski, J. A., Goldberg, T., Harbsburger, J., Kloppmann, E., Lizio, M.,
A draft network of ligand-receptor-mediated multicellular signaling in human.
Nature Communications, 6(1), 7866.