Neuroscience
Computational and Systems Biology

# Deciphering the Genetic Code of Neuronal Type Connectivity: A Bilinear Modeling Approach

**Mu Qiao**

LinkedIn, Mountain View, CA, 94043



**Reviewed Preprint**

Revised by authors after peer review.

About eLife's process

**Reviewed preprint version 2**
April 11, 2024 (this version)

**Reviewed preprint version 1**
November 28, 2023

**Sent for peer review**
August 25, 2023

**Posted to preprint server**
August 4, 2023

## Abstract

Understanding how different neuronal types connect and communicate is critical to interpreting brain function and behavior. However, it has remained a formidable challenge to decipher the genetic underpinnings that dictate the specific connections formed between neuronal types. To address this, we propose a novel bilinear modeling approach that leverages the architecture similar to that of recommendation systems. Our model transforms the gene expressions of presynaptic and postsynaptic neuronal types, obtained from single-cell transcriptomics, into a crosscorrelation matrix. The objective is to construct this cross-correlation matrix that closely mirrors a connectivity matrix, derived from connectomic data, reflecting the known anatomical connections between these neuronal types. When tested on a dataset of *Caenorhabditis elegans*, our model achieved a performance comparable to, if slightly better than, the previously proposed spatial connectome model (SCM) in reconstructing electrical synaptic connectivity based on gene expressions. Through a comparative analysis, our model not only captured all genetic interactions identified by the SCM but also inferred additional ones. Applied to a mouse retinal neuronal dataset, the bilinear model successfully recapticulated recognized connectivity motifs between bipolar cells and retinal ganglion cells, and provided interpretable insights into genetic interactions shaping the connectivity. Specifically, it identified unique genetic signatures associated with different connectivity motifs, including genes important to cell-cell adhesion and synapse formation, highlighting their role in orchestrating specific synaptic connections between these neurons. Our work establishes an innovative computational strategy for decoding the genetic programming of neuronal type connectivity. It not only sets a new benchmark for single-cell transcriptomic analysis of synaptic connections but also paves the way for mechanistic studies of neural circuit assembly and genetic manipulation of circuit wiring.

**eLife assessment**

This is an **important** computational study that applies the machine learning method of bilinear modeling to the problem of relating gene expression to connectivity. Specifically, the author attempts to use transcriptomic data from mouse retinal neurons to predict their known connectivity with promising results. On revision, the approach was tested against a second data set from C. Elegans. A limited number of genes studied in this second dataset may have resulted in performance that matched but did not exceed prior models, however, taken together, the results were felt to provide **solid** evidence for the value of the approach.

# 1 Introduction

One of the fundamental objectives in neuroscience is understanding how diverse neuronal cell types establish connections to form functional circuits. This understanding serves as a cornerstone for decoding how the nervous system processes information and coordinates responses to stimuli [1]. Despite this, the genetic mechanisms determining the specific connections between distinct neuronal types, especially within complex brain structures, remains elusive [2, 3].

Recent advances in transcriptomics and connectomics provide opportunities to probe this. Single-cell transcriptomics enables high-resolution profiling of gene expressions across neuronal types [4, 5], while connectomic data offers detailed maps quantifying connections between neuronal cell types [6, 7, 8]. However, the challenge of linking gene expressions derived from single-cell transcriptomics to neuronal type connectivity evident from connectomic data to uncover the genetic underpinnings has yet to be fully addressed.

Drawing inspiration from the field of machine learning, particularly recommendation systems, we introduce a bilinear model to bridge this gap. This model, in the context of recommendation systems, has been successful in capturing intricate user-item interactions [9]. By treating the gene expressions of pre- and post-synaptic neurons and their connectivity akin to users, items, and their ratings, we adapt the architecture of recommendation systems to the neurobiological domain. We hypothesize that a similar model could capture the complex relationships between genetic patterns of presynaptic and postsynaptic neurons and their connectivity.

This bilinear modeling approach was first applied to a *Caenorhabditis elegans* (*C. elegans*) neuronal dataset, where it not only matched but slightly outperformed the spatial connectome model (SCM) in reconstructing the connectivity of electrical synapses or gap junctions from innexin gene expressions. Notably, it revealed additional genetic interactions beyond those uncovered by the SCM. When extended to mouse retinal neurons, we demonstrate that it could effectively reconstruct synaptic connectivity between bipolar cells (BCs) and retinal ganglion cells (RGCs) from their gene expressions. The model not only unveils connectivity motifs between BCs and RGCs but also provides biologically meaningful insights into candidate genes and the genetic interactions that orchestrate this connectivity. Furthermore, our model predicts potential BC partners for RGC transcriptomic types, with these predictions aligned substantially with functional descriptions of these cell types from previous studies. Collectively, this work significantly contributes to the ongoing exploration of the genetic code underlying neuronal connectivity and suggests a potential paradigm shift in the analysis of single-cell transcriptomic data in neuroscience.

# 2 Background and Related Work

## 2.1 Synaptic Specificity

The intricate neural networks that form the basis of our nervous system are a product of specific synaptic connections between different types of neurons. This specificity is not a mere coincidence but a meticulously orchestrated process that underpins the functionality of the entire network [3, 10]. Each neuron can form thousands of connections, or synapses, with other neurons, and the specificity of these connections determines the neuron's function and, by extension, the network's function as a whole.

Synaptic specificity encompasses both chemical synapses, which rely on neurotransmitter-mediated communication between pre- and post-synaptic neurons [3], and electrical synapses, where direct transmission of ions or small molecules occurs via gap junctions [10]. A classic example of chemical synaptic specificity is observed in the retina, where different types of BCs form specific synaptic connections with various types of RGCs [7, 11, 12]. These connections create parallel pathways that transform visual signals from photoreceptors to RGCs, which subsequently transmit the information to the brain [13, 14]. Meanwhile, specific gap junction connections, composed of connexins in vertebrates and innexins in invertebrates, has been observed between *C. elegans* neurons [15, 16, 17, 18, 19]. They function broadly in neural circuits of sensory processing and behavioral output [10, 20].

The genetic principles guiding the formation of these specific connections, particularly in complex brain structures, remains elusive. The brain's complexity, with its billions of neurons and trillions of synapses, poses significant challenges in identifying the specific genes and genetic mechanisms that guide the formation of these connections. Despite advances in genetic and neurobiological research, such as understanding the roles of certain recognition molecules and adhesion molecules in synaptic specificity, the genetic foundation of connectivity between neuronal types is still largely unknown [3, 21, 10].

Emerging tools and technologies offer unprecedented opportunities to unravel these mysteries. Among these, transcriptome and connectome are particularly promising [3, 22]. Transcriptome, the complete set of RNA transcripts produced by the genome, can provide valuable insights into the genes that are active in different types of neurons and at different stages of neuronal development. This can help identify candidate genes that may play a role in guiding neuronal connectivity. Connectome, on the other hand, provides a detailed map of the connections between neurons. By combining information from transcriptome and connectome, it is possible to link specific genes to specific connections, thereby shedding light on the genetic basis of synaptic connectivity.

## 2.2 Previous Approaches

Prior research has reported several methodologies to unravel the genetic underpinnings of neuronal connectivity. For instance, Kaufman et al. showed a correlation between gene expression of *C. elegans* neurons and their connectivity [23], and Varadan et al. developed an entropy minimization approach for understanding the molecular logic of synaptic connectivity in *C. elegans* [24]. These models, however, did not fully account for spatial constraints for synaptic formation.

In response, subsequent studies proposed methodologies that integrate gene expressions with neuronal connectivity, taking into consideration physical contacts between neurons [25, 26, 27]. Specifically, the Spatial Connectome Model (SCM) in Kovács et al. correlates the gene expression of neurons with their connectivity via a rule matrix. This model aims to minimize the

discrepancy between predicted connectivity based on gene expression, and observed connectivity. By restricting the analysis to neuron pairs that are in physical contact, the SCM transforms the original problem into a regression between the Kronecker product of the gene expression matrix and an edge list that captures neuronal connectivity [25].

Additionally, Taylor et al. introduced the network differential gene expression analysis (nDGE), a statistical method that expands upon traditional differential gene expression analysis by examining the co-expression of gene pairs between neuron pairs, comparing synaptic versus non-synaptic neuronal groups through t-tests. It incorporates physical contacts between neurons through the generation of "pseudoconnectomes" for null distribution estimation. Unlike multivariate methods such as the SCM, nDGE operates as a mass-univariate method, focusing on single gene pairs' contributions to synaptic formation without considering the complex interactions among multiple co-expressed genes. This makes nDGE's findings inherently conservative, ensuring strict control over type 1 errors but potentially underestimating the multifaceted nature of synaptic connectivity [27].

While the SCM and nDGE models have focused on the connectivity of individual neurons and were tested using *C. elegans* datasets, their generalization to neuronal cell types has not been explored. As we move from the invertebrate nervous systems to the neural architectures of vertebrates, such as those in mice or macaques, we need methodologies capable of unraveling the genetic basis of neuronal type connectivity [4, 28].

## 2.3 Collaborative Filtering

Our strategy draws inspiration from the concept of collaborative filtering using bilinear models, a technique fundamental to recommendation systems [29, 30]. These systems predict a user's preference for an item (e.g., a movie or product) based on user-item interaction data.

Bilinear models capture the interaction between users and items via low-dimensional latent features [9, 31]. Mathematically, for user i and item j, we denote their original features as $\boldsymbol{x}_i \in \mathbf{R}^{1\times p}$ and $\boldsymbol{y}_j \in \mathbf{R}^{1\times q}$, respectively. These features are then projected into a shared latent space with dimension d via transformations $\boldsymbol{x}_i\boldsymbol{A}$ (where $\boldsymbol{A} \in \mathbf{R}^{p\times d}$) and $\boldsymbol{y}_j\boldsymbol{B}$ (where $\boldsymbol{B} \in \mathbf{R}^{q\times d}$). The predicted rating of the user for the item is then formulated as:

$$r_{ij} = (\boldsymbol{x}_i\boldsymbol{A})(\boldsymbol{y}_j\boldsymbol{B})^T \tag{1}$$

In the context of collaborative filtering, the goal is to optimize the transformation matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ to align the predicted rating $r_{ij}$ with the ground-truth $z_{ij}$. This is expressed as the following optimization problem:

$$\min_{\boldsymbol{A},\boldsymbol{B}} \sum_{ij} (z_{ij} - (\boldsymbol{x}_i\boldsymbol{A})(\boldsymbol{y}_j\boldsymbol{B})^T)^2 \tag{2}$$

Or in the matrix form:

$$\min_{\boldsymbol{A},\boldsymbol{B}} ||\boldsymbol{Z} - (\boldsymbol{X}\boldsymbol{A})(\boldsymbol{Y}\boldsymbol{B})^T||_F^2 \tag{3}$$

Here, the objective is to minimize the Frobenius norm of the residual matrix $\boldsymbol{Z} - (\boldsymbol{X}\boldsymbol{A})(\boldsymbol{Y}\boldsymbol{B})^T$.

In our study, we interpret neuronal connectivity through the lens of recommendation systems, viewing presynaptic neurons as "users", postsynaptic neurons as "items", and the synapses formed between them as "ratings". Our chosen bilinear model extracts latent features of pre- and post-synaptic neurons from their respective gene expressions. One key advantage of the bilinear model is its capacity to assign different weights to the gene expressions of pre- and post-synaptic

neurons, enabling the model to capture not just homogeneous but also complex, heterogeneous interactions fundamental to understanding neuronal connectivity. Prior studies have highlighted such heterogeneous interactions, noting the formation of connections between pre- and post-synaptic neurons expressing different cadherins, indicative of a heterogeneous adhesion process [32, 33].

# 3 Bilinear Model for Neuronal Type Connectivity

We discuss the bilinear model for neuronal type connectivity in the following two scenarios: the first in which gene expression and connectivity of each cell are known simultaneously and the second where connectivity and gene expressions of neuronal types are from different sources. The bilinear models for these two situations are illustrated in **Figure 1**.

## 3.1 Gene Expression and Connectivity of Each Cell are Known Simultaneously

### 3.1.1 Objective Function

We begin with an ideal scenario where both the gene expression profiles and connectivity of individual cells are known concurrently. In this setting, we have a presynaptic neuronal types and b postsynaptic neuronal types, indexed by i and j, respectively. Each type contains a number of neurons, signified as $n_i$ for presynaptic and $n_j$ for postsynaptic types. The gene expression vector for the $k^{th}$ cell in the presynaptic type i is designated as $\boldsymbol{x}_{(ik)}$, where $k \in 1, 2, ..., n_i$, while for the $l^{th}$ cell in postsynaptic type j, it is $\boldsymbol{y}_{(jl)}$ with $l \in 1, 2, ..., n_j$. We depict the connectivity metric between a presynaptic neuron and a postsynaptic neuron as $z_{(ik)(jl)}$.

Drawing from the principles of collaborative filtering, we develop the following optimization objective:

$$\min_{\boldsymbol{A},\boldsymbol{B}} \sum_{i=1}^{a} \sum_{j=1}^{b} \left( \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} (z_{(ik)(jl)} - (\boldsymbol{x}_{(ik)}\boldsymbol{A})(\boldsymbol{y}_{(jl)}\boldsymbol{B})^T)^2 \right) \tag{4}$$

Here, $\boldsymbol{A}$ and $\boldsymbol{B}$ denote the transformation matrices we aim to learn. This formula can also be expressed in its matrix form as:

$$\min_{\boldsymbol{A},\boldsymbol{B}} ||\boldsymbol{W} \odot (\boldsymbol{Z} - (\boldsymbol{X}\boldsymbol{A})(\boldsymbol{Y}\boldsymbol{B})^T)||_F^2 \tag{5}$$

In this equation, $\boldsymbol{W}$ symbolizes a weight matrix where each element $w_{(ik)(jl)} = \frac{1}{\sqrt{n_i n_j}}$. As our study focuses on the genetic code of pre- and post-synaptic neuronal types rather than individual neurons, this weight matrix ensures that the model does not disproportionately favor neuronal types with a greater number of neurons over rarer types. Note that this formulation can be generalized to individual cell level analysis by treating each cell as a type and setting $n_i = n_j = 1$, thus allowing exploration of genetic underpinnings of connectivity at the single-cell resolution.

In the context of high dimensionality of gene expressions, the bilinear model may face a common issue in machine learning called multicollinearity, a condition where one or more predictor variables are highly correlated. To mitigate this, we can perform principal component analysis (PCA) on the gene expression vectors, transforming them into a new coordinate system and removing components with negligible eigenvalues to reduce redundant information. Alternatively, we can apply regularization techniques, such as L2 regularization (Ridge) or L1 regularization

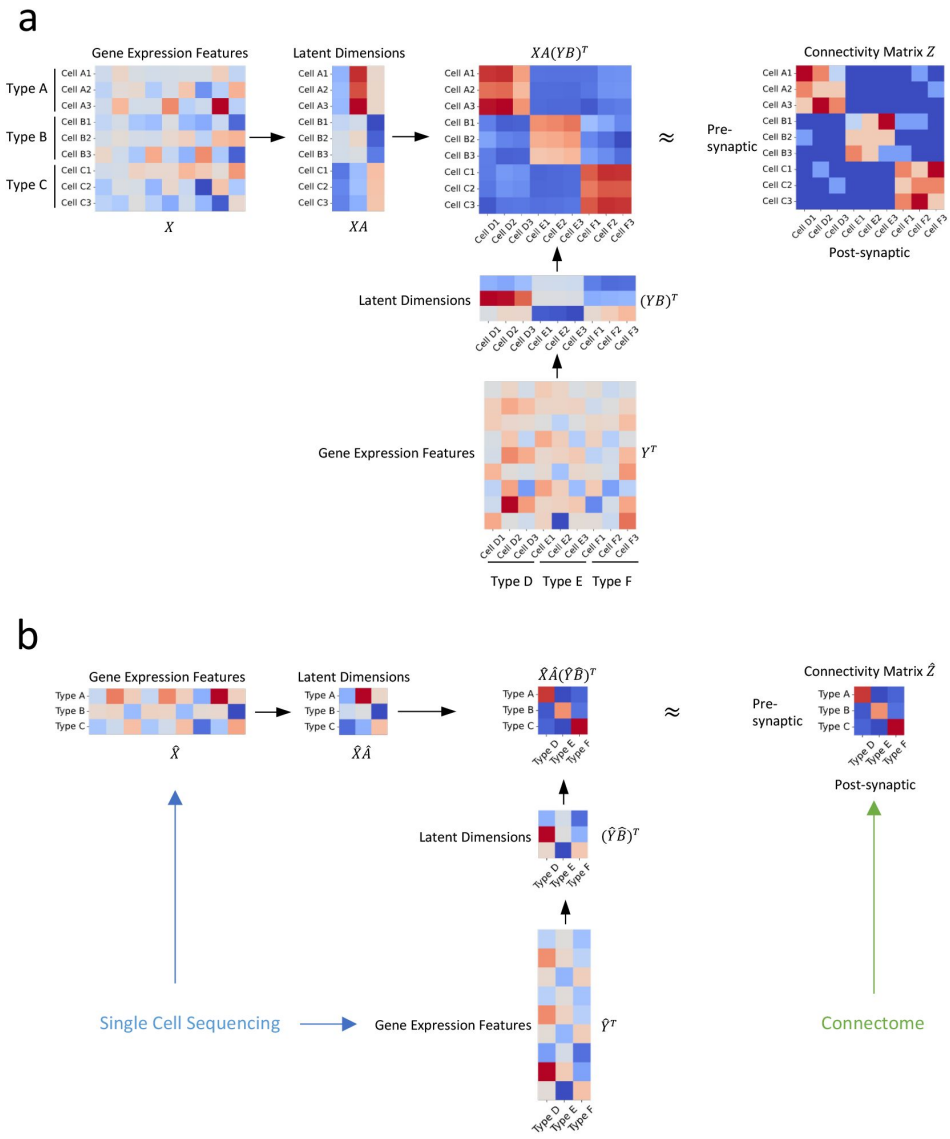



**Figure 1.**

Illustration of our approach. (a) In an ideal scenario where gene expression profiles and connectivity data of individual cells are available simultaneously, we establish the relationship between connectivity and gene expression profiles via two transformation matrices ***A*** and ***B*** (b) In practical situations where the gene expression profiles are derived from distinct sources, such as single-cell transcriptomic and connectomic data, we propose that the connectivity of individual cells and their latent gene expression features can be approximated by the averages of their corresponding cell types, and establish their relationship through transformation matrices Â and $\hat{\boldsymbol{B}}$.

(Lasso) to effectively manage the multicollinearity. These regularization methods work by imposing a penalty on the size of the linear coefficients in the model, thereby shrinking the coefficients and stabilizing their estimates.

### 3.1.2 Optimization Algorithm

Incorporating L2 regularization, we minimize the following loss function with regularization hyperparameters $\lambda_A$ and $\lambda_B$:

$$L(\boldsymbol{A}, \boldsymbol{B}) = ||\boldsymbol{W} \odot (\boldsymbol{Z} - (\boldsymbol{X}\boldsymbol{A})(\boldsymbol{Y}\boldsymbol{B})^T)||_F^2 + \frac{\lambda_A}{2}||\boldsymbol{A}||_F^2 + \frac{\lambda_B}{2}||\boldsymbol{B}|_F^2 \tag{6}$$

To optimize this function, we propose an alternative gradient descent algorithm. This algorithm alternates between updating the transformation matrices ***A*** and ***B***, using the gradient descent optimization method.

The algorithm begins by initializing transformation matrices ***A*** and ***B*** using random values drawn from a standard normal distribution. The central aspect of the algorithm is an iterative loop that alternates the updates of ***A*** and ***B***. During each iteration, the algorithm first computes the predicted connectivity metric ***Z*** using the current estimates of ***A*** and ***B***. Subsequently, the gradient of the loss function with respect to the transformation matrices is calculated, and the matrices are updated by moving in the negative gradient's direction. This iterative process is repeated until the transformation matrices ***A*** and ***B*** converge to a steady solution. Upon completion, the algorithm yields the optimized transformation matrices.

This gradient descent-based algorithm provides a computationally efficient solution to the bilinear mapping problem between gene expression profiles and connectivity metrics. As a result, it produces associations between gene expression profiles of cell types and their connectivity.

**Algorithm 1 Alternative Gradient Descent (AGD) for 4.1**
**Algorithm 1 Alternative Gradient Descent (AGD) for 4.1**

**procedure** AGD($\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{Y}, d, r, \lambda_A, \lambda_B$) ▷ $d$: latent space dimension; $r$: learning rate
  $q \leftarrow$ second dimension of $\boldsymbol{X}$
  $p \leftarrow$ second dimension of $\boldsymbol{Y}$
  Initialize $\boldsymbol{A}$ with random values of size $(q, d)$
  Initialize $\boldsymbol{B}$ with random values of size $(p, d)$
  **while** not converged **do**
    $\hat{\boldsymbol{Z}} \leftarrow \boldsymbol{X}\boldsymbol{A}(\boldsymbol{Y}\boldsymbol{B})^T$ ▷ $\hat{\boldsymbol{Z}}$: prediction of $\bar{\boldsymbol{Z}}$
    Compute $\boldsymbol{A}_{\text{grad}} \leftarrow 2\boldsymbol{X}^T(\boldsymbol{W} \odot (\hat{\boldsymbol{Z}} - \boldsymbol{Z}))\boldsymbol{Y}\boldsymbol{B} + \lambda_A\boldsymbol{A}$
    Update $\boldsymbol{A} \leftarrow \boldsymbol{A} - r * \boldsymbol{A}_{\text{grad}}$
    Compute $\boldsymbol{B}_{\text{grad}} \leftarrow 2\boldsymbol{Y}^T(\boldsymbol{W} \odot (\hat{\boldsymbol{Z}} - \boldsymbol{Z}))^T\boldsymbol{X}\boldsymbol{A} + \lambda_B\boldsymbol{B}$
    Update $\boldsymbol{B} \leftarrow \boldsymbol{B} - r * \boldsymbol{B}_{\text{grad}}$
  **end while**
  **return** $\boldsymbol{A}, \boldsymbol{B}$
**end procedure**

## 3.2 Connectivity and Gene Expressions of Neuronal Types are from Different Sources

### 3.2.1 Objective Function

In real scenarios, gene expression profiles and connectivity information are often derived from separate sources, such as single-cell sequencing [34, 35] and connectome data [7, 36, 37]. Bridging these datasets requires classifying neurons into cell types based on their gene expression profiles and morphological characteristics. These cell types from different sources are subsequently aligned according to established biological knowledge (e.g., specific gene markers are known to be expressed in certain morphologically-defined cell types [38]).

The primary challenge in this scenario is that, while we can align cell types (denoted by indices $i$ and $j$ in equation 4), we are unable to associate individual cells (represented by indices $k$ and $l$ in equation 4). To tackle this issue, we adopt a simplifying assumption that the connectivity and latent gene expression features of individual cells can be approximated by the averages of their corresponding cell types. This premise hinges on the notion that the connectivity metrics and latent gene expression features of individual cells are close enough to the mean value of their corresponding cell types.

As a result, our optimization objective in equation 4 becomes:

$$\min_{\boldsymbol{A},\boldsymbol{B}} \sum_{i=1}^{a}\sum_{j=1}^{b}(z_{(i.)(j.)} - (\boldsymbol{x}_{(i.)}\boldsymbol{A})(\boldsymbol{y}_{(j.)}\boldsymbol{B})^T)^2 \tag{7}$$

In this equation, $z_{(i.)(j.)}$ denotes the mean connectivity metric between presynaptic cell type $i$ and postsynaptic cell type $j$. Meanwhile, $\boldsymbol{x}_{(i.)}$ and $\boldsymbol{y}_{(j.)}$ represent the average gene expressions of cell types $i$ and $j$ respectively.

While optimizing the transformation matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we impose constraints on these matrices to ensure that the variance of latent gene expression features within each neuronal type is minimized. Specifically, we define ϵ as a small enough value and impose the following constraints on $\boldsymbol{A}$:

$$||\boldsymbol{A}^T\boldsymbol{\Sigma}_x\boldsymbol{A}||_F^2 \leq \epsilon \tag{8}$$

Where

$$\boldsymbol{\Sigma}_x = \sum_{i=1}^{a}(\frac{1}{n_i}\sum_{k=1}^{n_i}(\boldsymbol{x}_{(ik)} - \boldsymbol{x}_{(i.)})^T(\boldsymbol{x}_{(ik)} - \boldsymbol{x}_{(i.)})) \tag{9}$$

and $\boldsymbol{B}$:

$$||\boldsymbol{B}^T\boldsymbol{\Sigma}_y\boldsymbol{B}||_F^2 \leq \epsilon \tag{10}$$

Where

$$\boldsymbol{\Sigma}_y = \sum_{j=1}^{b}(\frac{1}{n_j}\sum_{l=1}^{n_j}(\boldsymbol{y}_{(jl)} - \boldsymbol{y}_{(j.)})^T(\boldsymbol{y}_{(jl)} - \boldsymbol{y}_{(j.)})) \tag{11}$$

These conditions assure that the latent gene expression features of individual cells are proximate enough to the average value within their respective cell types. With these constraints in mind, we formulate the optimization problem as follows:

$$\min_{\boldsymbol{A},\boldsymbol{B}} ||\bar{\boldsymbol{Z}} - \bar{\boldsymbol{X}}\boldsymbol{A}(\bar{\boldsymbol{Y}}\boldsymbol{B})^T||_F^2, \quad s.t. ||\boldsymbol{A}^T\boldsymbol{\Sigma}_x\boldsymbol{A}||_F^2 \leq \epsilon, ||\boldsymbol{B}^T\boldsymbol{\Sigma}_y\boldsymbol{B}||_F^2 \leq \epsilon \tag{12}$$

In this equation, $\bar{\boldsymbol{X}} \in \mathbf{R}^{a\times p}$ denotes the average gene expressions of the a presynaptic cell types, wherein each element $\bar{x}_{im}$ is indicative of the average gene expression feature m within cell type *i*. Likewise, $\bar{\boldsymbol{Y}} \in \mathbf{R}^{b\times q}$ represents the average gene expressions of the b postsynaptic cell types, with each element $\bar{y}_{jm}$ signifying the average gene expression feature *m* in cell type *j*.

In practical application, we approximate $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$ with their diagonal estimates $diag(\hat{\sigma}^2_{x_1}, \hat{\sigma}^2_{x_2}, ..., \hat{\sigma}^2_{x_p})$ and $diag(\hat{\sigma}^2_{y_1}, \hat{\sigma}^2_{y_2}, ..., \hat{\sigma}^2_{y_q})$ [39, 40]. We then transform the initial optimization problem into the following:

$$\min_{\hat{\boldsymbol{A}},\hat{\boldsymbol{B}}} ||\bar{\boldsymbol{Z}} - \hat{\boldsymbol{X}}\hat{\boldsymbol{A}}(\hat{\boldsymbol{Y}}\hat{\boldsymbol{B}})^T||_F^2, \quad s.t. ||\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}}||_F^2 \leq \epsilon, ||\hat{\boldsymbol{B}}^T\hat{\boldsymbol{B}}||_F^2 \leq \epsilon \tag{13}$$

Here, elements in $\hat{\boldsymbol{X}} \in \mathbf{R}^{a\times p}$ are defined as $\hat{x}_{im} = \frac{\bar{x}_{im}}{\hat{\sigma}_{x_m}}$ and elements in $\hat{\mathrm{Y}} \in \mathbf{R}^{b\times q}$ are given by $\hat{y}_{im} = \frac{\bar{y}_{im}}{\hat{\sigma}_{y_m}}$. The optimization of this formulation tends to be computationally more tractable.

In summary, our methodology adapts when gene expression profiles and the connectivity matrix originate from distinct sources. Instead of aligning at the level of individual cells, we focus on the alignment of neuronal types. We achieve this by mapping gene expressions into a latent space via transformation matrices $\hat{\mathrm{A}}$ and $\hat{\boldsymbol{B}}$, with the optimization process aiming to minimize the discrepancies between these two sources of information while maintaining consistency of the gene expression features within individual neuronal types.

### 3.2.2 Optimization Algorithm

To solve the optimization problem as outlined in equation 13, we construct the following loss function:

$$L(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}) = ||\bar{\boldsymbol{Z}} - \hat{\boldsymbol{X}}\hat{\boldsymbol{A}}(\hat{\boldsymbol{Y}}\hat{\boldsymbol{B}})^T||_F^2 + \frac{\lambda_A}{2}||\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}}||_F^2 + \frac{\lambda_B}{2}||\hat{\boldsymbol{B}}^T\hat{\boldsymbol{B}}||_F^2 \tag{14}$$

where $\lambda_A$ and $\lambda_B$ are hyperparameters whose optimal values are determined through a grid search.

To optimize this loss function, we employ an alternative gradient descent algorithm analogous to that described in section 3.1.2, by iteratively updating the transformation matrices $\hat{\mathrm{A}}$ and $\hat{\boldsymbol{B}}$.

**Algorithm 2 Alternative Gradient Descent (AGD) for 4.2**
**Algorithm 2 Alternative Gradient Descent (AGD) for 4.2**

**procedure** AGD($\bar{\boldsymbol{Z}}, \hat{\boldsymbol{X}}, \hat{\boldsymbol{Y}}, d, r, \lambda_A, \lambda_B$) ▷ $d$: latent space dimension; $r$: learning rate
  $q \leftarrow$ second dimension of $\hat{\boldsymbol{X}}$
  $p \leftarrow$ second dimension of $\hat{\boldsymbol{Y}}$
  Initialize $\hat{\boldsymbol{A}}$ with random values of size $(q, d)$
  Initialize $\hat{\boldsymbol{B}}$ with random values of size $(p, d)$
  **while** not converge **do**
    $\hat{\boldsymbol{Z}} \leftarrow \hat{\boldsymbol{X}}\hat{\boldsymbol{A}}(\hat{\boldsymbol{Y}}\hat{\boldsymbol{B}})^T$ ▷ $\hat{\boldsymbol{Z}}$: prediction of $\bar{\boldsymbol{Z}}$
    Compute $\hat{\boldsymbol{A}}_{grad} \leftarrow \hat{\boldsymbol{X}}^T(\hat{\boldsymbol{Z}} - \bar{\boldsymbol{Z}})\hat{\boldsymbol{Y}}\hat{\boldsymbol{B}} + \lambda_A\hat{\boldsymbol{A}}(\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}})$
    Update $\hat{\boldsymbol{A}} \leftarrow \hat{\boldsymbol{A}} - r * \hat{\boldsymbol{A}}_{grad}$
    Compute $\hat{\boldsymbol{B}}_{grad} \leftarrow \hat{\boldsymbol{Y}}^T(\hat{\boldsymbol{Z}} - \bar{\boldsymbol{Z}})\hat{\boldsymbol{X}}\hat{\boldsymbol{A}} + \lambda_B\hat{\boldsymbol{B}}(\hat{\boldsymbol{B}}^T\hat{\boldsymbol{B}})$
    Update $\hat{\boldsymbol{B}} \leftarrow \hat{\boldsymbol{B}} - r * \hat{\boldsymbol{B}}_{grad}$
  **end while**
  **return** $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{B}}$
**end procedure**

# 4 Datasets and Pre-processing

To validate and assess the efficacy of our bilinear model, we utilized two distinct datasets available from previous studies:

## 4.1 Gap Junction Connectivity and Innexin Expression Data of *C. elegans* Neurons

We first used a dataset of gap junction connectivity and innexin expressions of individual *C. elegans* neurons. Derived from the work of Cook et al. [41] and subsequently analyzed by Kovács et al. [25], this dataset included expression profiles of 18 innexin genes across 184 neurons, alongside detailed gap junction connectivity between these neurons. We followed the same procedure outlined by Kovács et al. to obtain the innexin expression matrix $\boldsymbol{X}$ and $\boldsymbol{Y}$ (in this case $\boldsymbol{X} = \boldsymbol{Y}$ with the dimensions of 184 × 18), and the connectivity matrix between individual *C. elegans* neurons $\boldsymbol{Z}$.

To incorporate spatial constraints by considering only neuron pairs in physical contact, we extracted a contact matrix from the dataset. This was transcribed into the weight matrix $\boldsymbol{W}$ in our model, with values set to 0 for neuron pairs without physical contact and 1 for those with contact. This enabled our bilinear model to focus on the 5,592 neuron pairs that exhibit physical contacts, restricting the analysis to biologically plausible connections.

The utilization of this dataset serves a dual purpose. It not only provides a validation for our bilinear model but also enables a direct comparison with the model employed by Kovács et al., offering a comprehensive evaluation of the bilinear model in the context of established connectomic research.

## 4.2 Single-cell Transcriptomic and Connectomic Data of Mouse Retinal Neurons

The second dataset encompassed data of mouse retinal neurons, integrating single-cell transcriptomic data from various studies with connectomic data obtained from the EyeWire project. The data provide us with connectivity information and gene expression profiles of mouse

BCs and RGCs, and are important for applying our proposed bilinear model and testing its effectiveness in a more complex neuronal environment compared to the *C. elegans* dataset.

### 4.2.1 Single-cell Transcriptomic Data

The single-cell transcriptomic data include the gene expression profiles for two classes of mouse retinal neurons - presynaptic BCs as reported by Shekhar et al. [34], and postsynaptic RGCs as reported by Tran et al. [35].

Preprocessing of this data adhered to previously documented procedures [34, 35, 42]. The transcript counts within each cell were first normalized to align with the median number of the transcripts per cell, followed by a log-transformation of the normalized counts. High variable genes (HVGs) were then selected using an approach based on establishing a relationship between mean expression level and the coefficient of variance [43, 44, 45]. We focused on those cells whose types correspond with the neuronal types outlined in the connectomic data, as delineated later in **Table S1**, **Table S2**, and **Table S3**. This yielded two matrices, $\boldsymbol{X}$ and $\boldsymbol{Y}$, representing presynaptic BCs and postsynaptic RGCs, where each row pertains to a cell and each column represents an HVG. The dimensions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are 22453 × 17144 and 3779 × 12926, respectively.

Next, we performed a principal component analysis (PCA) on these matrices to transform the gene expression data into the principal component (PC) space. We retained only the PCs that account for a cumulative 95% of explained variance. Consequently, the gene expression of the BCs in $\boldsymbol{X}$ and the RGCs in $\boldsymbol{Y}$ were featurized by their respective PCs, resulting in matrices of dimensions 22453 × 11323 and 3779 × 3142, respectively.

Based on each cell's neuronal type, we computed the variance of gene expression features within these types. Mathematically, the variance of gene expression feature m within the BC types and the RGC types are expressed as:

$$\hat{\sigma}_{x_{im}} = \sum_{i=1}^{a}\left(\frac{1}{n_i}\sum_{k=1}^{n_i}(x_{(ik)m} - x_{(i.)m})^2\right) \tag{15}$$

$$\hat{\sigma}_{y_{jm}} = \sum_{j=1}^{b}\left(\frac{1}{n_j}\sum_{l=1}^{n_j}(y_{(jl)m} - y_{(j.)m})^2\right) \tag{16}$$

Taking $\bar{x}_{im}$ and $\bar{y}_{jm}$ to represent the average gene expression feature *m* of the BC type i and the RGC type j, we were able construct matrices, $\hat{\boldsymbol{X}}$ and $\hat{\mathrm{Y}}$, in which $\hat{x}_{im} = \frac{\bar{x}_{im}}{\hat{\sigma}_{x_m}}$ and $\hat{y}_{im} = \frac{\bar{y}_{im}}{\hat{\sigma}_{y_m}}$. In these matrics, each row represents a cell type, with the dimensions of $\hat{\boldsymbol{X}}$ being 25 × 11323 and $\hat{\mathrm{Y}}$ being 12 × 3142. These matrices serve to bridge the gene expression of BC types and RGC types with the connectivity matrix of these neuronal types derived from the connectomic data.

### 4.2.2 Connectivity Data

The connectivity matrix of neuronal types is derived from connectomic data acquired through the process of serial electron microscopy (EM)-based reconstruction of brain tissues [6, 7, 8]. From these reconstructed tissues, connectivity measurements are usually expressed as either the contact area or the number of synapses between neurons [7, 46]. When normalized to the total contact area or total number of synapses of each neuron, the resulting metric, ranging from 0 to 1, signifies the percentage of contact area or synapses formed between neurons. This normalized metric provides a quantitative connectivity measure, where 0 indicates no connectivity and 1 implies complete connectivity between two neurons.

Our analysis utilized the neural reconstruction data of mouse retinal neurons, courtesy of the EyeWire project, a crowd-sourced initiative that generates 3D reconstructions of neurons from serial section EM images [47]. This extensive dataset facilitated the derivation of a comprehensive connectivity matrix between two classes of mouse retinal neurons - BCs [37] and RGCs [36]. The data were sourced from the EyeWire Museum (*https://museum.eyewire.org/*), which offers detailed information for each cell in a JSON file, including attributes like "cell id", "cell type", "cell class", and "stratification". The stratification profile describes the linear density of voxel volume as a function of the inner plexiform layer (IPL) depth [47, 37, 36].

We approximated the connectivity metric between a BC and a RGC using the cosine similarity of their stratification profiles. Let $\boldsymbol{v}_{ik}$ and $\boldsymbol{v}_{jl}$ denote the stratification profiles of the $k^{th}$ cell in BC type i and the $l^{th}$ cell in RGC type *j*, respectively. The connectivity metric $z_{(ik)(jl)}$ between these two neurons can be expressed as:

$$z_{(ik)(jl)} = \frac{\boldsymbol{u}_{ik}\boldsymbol{v}_{jl}}{|\boldsymbol{v}_{ik}||\boldsymbol{v}_{jl}|} \tag{17}$$

This equation represents the degree of overlap in their voxel volume profile within the IPL, resulting in the connectivity matrix $\boldsymbol{Z}$ between mouse BCs and RGCs. To allow for both positive and negative values within the matrix, we standardized $\bar{\boldsymbol{Z}}$ by subtracting the mean of $\bar{\boldsymbol{Z}}$ and then dividing by its standard deviation. Subsequently, the connectivity matrix $\bar{\boldsymbol{Z}}$ between mouse BC and RGC neuronal types was calculated, with each element $\bar{z}_{ij} = z(i.)(j.)$ representing the average of the connectivity metrics between cells of BC type *i* and cells of RGC type *j*.

### 4.2.3 Correspondence of Mouse Retinal Cell Types

Aligning neuronal types as annotated in the single-cell transcriptomic data and those identified in the connectomic data was informed by findings from previous studies. Notably, a one-to-one correspondence exists between BC cell types classified by Shekhar et al. [34] and Greene et al. [37]. This correspondence is presented in **Table S1**.

Regarding RGC types, alignment between cell types annotated in Tran et al. [35] and Bae et al. [36] was established primarily based on the findings from Goetz et al. [38]. This study presents a unified classification of mouse RGC types, based on their functional, morphological, and gene expression features. The corresponding RGC types were mainly obtained from **Supplementary Table S3** of Goetz et al. (**Table S2**), with additions derived from **Supplementary Table S1** of Tran et al., based on the expressions of genetic markers of these RGC types (**Table S3**).

## 5 Model Training, Validation, and Comparison

Our approach of training and validating the bilinear model involved an iterative optimization of transformation matrices using the AGD algorithm, as outlined in **Section 3**. The primary goal was to minimize the defined loss function. With the matrices initially generated from a standard normal distribution, the optimization process continued until the loss change was less than a threshold of $10^{-6}$, or a maximum of $10^6$ iterations were completed.

During optimization, we focused on two key hyperparameters: the regularization parameters, $\lambda_A$ and $\lambda_B$, and the latent feature space dimensionality. Preliminary tests indicated that a lower loss was achieved when both regularization parameters were set equally, leading us to consolidate them into a single parameter, $\lambda$.

## 5.1 *C. elegans* Neuronal Dataset

For the *C. elegans* dataset, which provides simultaneous gene expression and connectivity data for individual cells, we employed the model configuration described in **section 3.1**. The model's hyperparameters, λ and the latent feature space dimensionality, were fine-tuned through 5-fold cross-validation, exploring a range of values for λ and different dimensions for the latent feature space. The optimal hyperparameters were identified based on the lowest validation loss observed during cross-validation (**Figure S1**).

Given the prior utilization of this dataset in validating the SCM proposed by Kovács et al. [25], our bilinear model was positioned for a direct comparison with the SCM. The SCM introduced a rule matrix $\boldsymbol{O}$ with the aim to minimize the discrepancy between the observed connectivity and the gene expression-based predicted connectivity $\boldsymbol{XOX}^T$, employing L2 regularization on $\boldsymbol{O}$. Our bilinear model echoes this approach, where we seek to minimize the divergence between the connectivity matrix and the bilinearly predicted connectivity $\boldsymbol{XA}(\boldsymbol{XB})^T$, with L2 regularization imposed on matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. In essence, the bilinear form decomposes the rule matrix into two lower-dimensional matrices, which represent projections onto latent dimensions.

To quantitatively compare the bilinear model's transformation matrix product $\hat{\mathbf{O}} = \boldsymbol{AB}^T$ with the SCM's rule matrix $\boldsymbol{O}$, and to systematically identify the genetic interaction each model uniquely captured, we introduced the discrepancy score (DS). For each pair of corresponding entries in the matrices at indices *i* and *j*, the DS is calculated as follows:

$$DS_{ij} = \frac{|\hat{o}_{ij} - o_{ij}|}{|\hat{o}_{ij}| + |o_{ij}|} \tag{18}$$

This metric, ranging from 0 to 1, quantifies the relative discrepancy between the two matrices, normalizing it in relation to their magnitudes. A score close to 1 indicates a large discrepancy, while a score near 0 suggests a negligible difference between the entries. Through this lens, we can further scrutinize the corresponding entries with the score above a certain threshold to reveal specific genetic interactions captured by one model but potentially missed by the other.

## 5.2 Mouse Retinal Neuronal Dataset

The model's application to the mouse retina dataset, which involves gene expression and connectivity data from disparate sources, was facilitated by the approach outlined in section 3.2. Optimal hyperparameters were determined through 5-fold cross-validation, adjusting λ and exploring various dimensionalities for the latent feature space (**Figure S2**). Notably, the lowest validation loss was achieved with the dimensionality of two. Given the chosen hyperparameters, we performed the final round of training on the entire dataset to yield the definitive transformation matrices Â and $\hat{\boldsymbol{B}}$.

To assess the consistency of our model under PCA pre-processing across different replicates, we repeated the optimization procedure five times, each time adhering to the previously identified optimal hyperparameters. In the context of our solution, where $\hat{\mathbf{A}} = [\boldsymbol{u}_1\ \boldsymbol{u}_2]$ and $\hat{\boldsymbol{B}} = [\boldsymbol{v}_1 \quad \boldsymbol{v}_2]$, with vectors $\boldsymbol{u}_1, \boldsymbol{v}_1$ representing coefficients for the first latent dimension and $\boldsymbol{u}_2, \boldsymbol{v}_2$ for the second, we noted that negating the coefficients of any latent dimension in both matrices (for instance, $\hat{\mathbf{A}} = [-\boldsymbol{u}_1\ \boldsymbol{u}_2]$ and $\hat{\boldsymbol{B}} = [-\boldsymbol{v}_1 \quad \boldsymbol{v}_2]$ results in an equivalent solution. Therefore, to compare solutions across different repetitions, we calculated the absolute value of cosine similarity for each latent dimension's coefficient vectors, and reported the similarity between solutions as the average of the values across the two latent dimensions. Moreover, we recognized that swapping the positions of the coefficient vectors (yielding $\hat{\mathbf{A}} = [\boldsymbol{u}_2\ \boldsymbol{u}_1]$ and $\hat{\boldsymbol{B}} = [\boldsymbol{v}_2 \quad \boldsymbol{v}_1]$) also leads to an equivalent solution. To accommodate this, we evaluated both the original and swapped vector pairings for

each repetition. The final measure of consistency was determined by taking the maximum of the two average absolute cosine similarities, ensuring a comprehensive and robust assessment of solution consistency across multiple runs.

We observed a high degree of consistency across multiple repetitions of the solutions under PCA pre-processing (**Figure S3**). The majority of the average absolute cosine similarity scores are close to 1, and even the minimum observed similarities are well above 0.75, suggesting that the optimization yields stable solutions.

# 6 Results

## 6.1 Comparative Analysis using *C. elegans* Neuronal Data

### 6.1.1 Reconstruction of *C. elegans* Gap Junction Connectivity from Innexin Expressions

Utilizing the *C. elegans* neuronal dataset, we first tried to reconstruct the gap junction connectivity network based solely on the expression profiles of innexin genes. Using $\boldsymbol{A}$ and $\boldsymbol{B}$ generated by the bilinear model, we processed the innexin expression data to predict gap junction connectivity between neuron pairs as $\boldsymbol{XA}(\boldsymbol{Y}\,\boldsymbol{B})^T$ (**Figure 2a**). This approach was then compared to the SCM proposed by Kovács et al. [25], which used a rule matrix $\boldsymbol{O}$ to correlate gene expression with observed connectivity in the form of $\boldsymbol{XOX}^T$ (**Figure 2b**).

The effectiveness of both models was evaluated against the observed gap junction connectivity matrix of *C. elegans* neurons (**Figure 2c**). Given the binary nature of the ground truth matrix (where 1 denotes a connection and 0 indicates its absence) and the continuous nature of reconstructed connectivity matrices from both models, we conducted Receiver Operating Characteristic (ROC) analysis. This involves varying a threshold to binarize the continuous predictions, under which the true positive rate is plotted against the false positive rate for each possible cutoff. This process yields the ROC curve, which is a graphical representation of the trade-off between sensitivity and specificity at various thresholds (**Figure 2d**).

Subsequently, we calculated the Area Under the ROC Curve (AUC), providing a singular value summarizing the overall predictive performance of the model across all thresholds. The ROC-AUC metric is particularly informative as it aggregates the model's effectiveness over all possible thresholds, with a score of 1 indicating perfect prediction and 0.5 denoting a performance no better than random chance. From the calculation, the bilinear model achieved a ROC-AUC score of 0.6435, slightly surpassing the SCM's score of 0.6428. While both scores are reasonably close, the slight edge of the bilinear model indicates its nuanced efficiency in mapping gene expressions to connectivity. However, it is noteworthy that both scores, while above 0.5, are substantially distant from the ideal score of 1. This observation suggests that relying exclusively on innexin expression data might be insufficient for fully capturing the detailed gap junction connectivity in *C. elegans*.

### 6.1.2 Comparison of Rule Matrix from SCM and Bilinear Transformation Matrices

In light of the challenge in fully capturing the *C. elegans* gap junction connectivity based on innexin expression data alone, instead of analyzing connectivity motifs between *C. elegans* neurons, our focus pivoted towards exploring and comparing the genetic rules inferred by both the bilinear model and the SCM, which was also the key discussion presented in Kovács et al. [25]. As mentioned in Section 5.1 and discussed in Section 7, the product of the bilinear

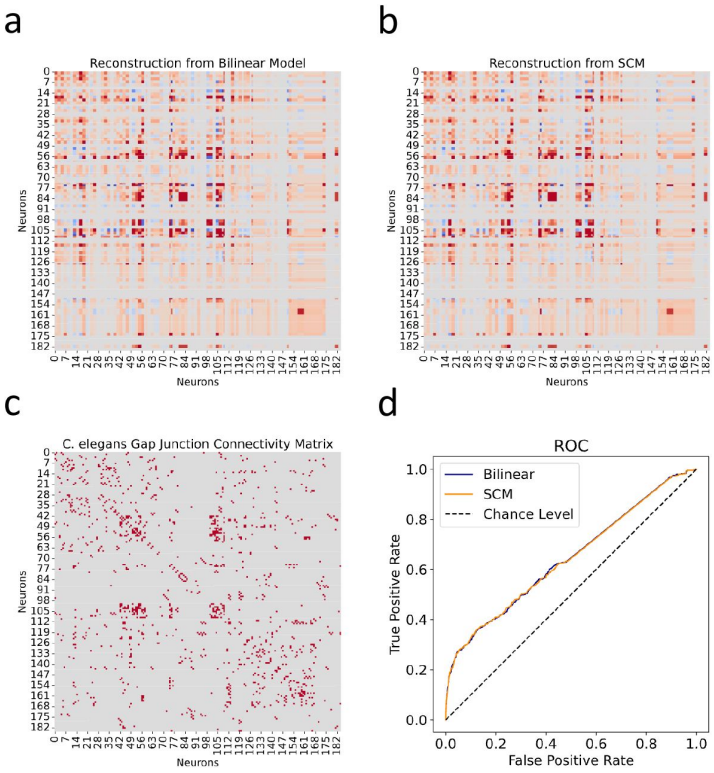



**Figure 2.**

Reconstructed gap junction connectivity from innexin expression data. (a) Connectivity matrix predicted by the bilinear model. (b) Connectivity matrix modeled from Kovács et al.'s SCM. (c) Observed gap junction connectivity matrix, serving as ground truth. The color spectrum from red to gray denotes the spectrum from strong connections to weak or no connections. (d) ROC Curves from both the bilinear model and the SCM. Dashed line indicates the chance level.

transformation matrices, $\hat{\mathbf{O}} = \boldsymbol{AB}^T$, can be interpreted as a lower-dimensional reconstruction of the rule matrix $\boldsymbol{O}$ used in the SCM. This perspective steered us to a meticulous comparative analysis between the two matrices.

The rule matrix solved from the SCM establishes a baseline for the comparison (**Figure 3b**). Against that, we compared the product of the bilinear transformation matrices (**Figure 3a**). Visualization of the two matrices suggests a high degree of similarity between them, which is quantitatively supported by a Pearson correlation coefficient of 0.90 ($p < 0.001$), underscoring a strong alignment.

To discern specific genetic interactions uniquely characterized by each model, we applied the DS metric to corresponding matrix entries (**Figure S4a**). This metric, ranging from 0 (no discrepancy) to 1 (maximum discrepancy), was thresholded at 0.5 to highlight entries with substantial differences. Further, to account for the regularization effect that pushes less important coefficients toward zero, we filtered out entry pairs where both values were less than 0.1 (**Figure S4b,c**). The remaining pairs are highlighted in black boxes in both matrices (**Figure 3**).

Comparing the values of highlighted entry pairs, we found that the bilinear model not only captured all genetic interactions identified by the SCM but also inferred additional ones: certain innexins (inx-11, inx-8, inx-5, and inx-2) were implicated in co-expression patterns within connected neurons, while another set (inx-11, inx-9, inx-3, inx-5, inx-7) was associated with an avoidance pattern, suggesting a lack of co-expression in neuron pairs forming gap junctions. These findings provide extra candidates to be tested in future experiments.

## 6.2 Application of Bilinear Model to Mouse Retinal Neuronal Data

### 6.2.1 Bilinear Model Reconstructs Neuronal Type-Specific Connectivity Map from Gene Expression Profiles

In our application of the bilinear model to the mouse retinal neuronal data, upon completion of the final training process, our optimized bilinear model produced transformation matrices, $\hat{\mathbf{A}}$ and $\hat{B}$. We used these matrices to project the normalized single-cell transcriptomic data, $\hat{X}$ and $\hat{\mathbf{Y}}$, into a shared latent feature space. Consequently, we obtained projected representations for BC and RGC types, $\hat{X}\hat{A}$ and $\hat{Y}\hat{B}$, respectively. With these latent representations, we were able to reconstruct the cell-type-specific connectivity matrix: $\hat{\boldsymbol{X}}\hat{\boldsymbol{A}}(\hat{\boldsymbol{Y}}\hat{\boldsymbol{B}})^T$ (**Figure 4a**).

To evaluate our model, we compared the reconstructed connectivity matrix with the one derived from connectomic data (**Figure 4b**). We calculated the Pearson correlation coefficient between entries of the two matrices to assess their agreement. The resulting correlation of 0.83 ($p < 0.001$) demonstrated a robust association between the transformed gene expression features and the connectomic data. This result attests to our model's capability in capturing the relationship between these two distinct types of biological information.

To gain insights into our model's reconstruction accuracy, we employed the DS metric to identify entries with substantial deviations between the reconstructed and the actual connectivity matrices (**Figure S5a**). This examination specifically quantified the extent of connections in the target matrix (positive entries) that were not captured in the model's reconstruction (negative entries) (**Figure S5b,c**). Notably, the analysis revealed that only a small fraction, specifically 9 out of 115 connections, were not represented in the reconstructed matrix.

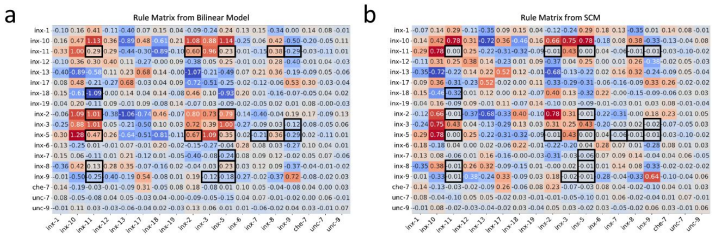



**Figure 3.**

Genetic rules from the bilinear model and the SCM. (a) The rule matrix $\boldsymbol{AB}^T$ derived from the bilinear model. (b) The rule matrix $\boldsymbol{O}$ from the SCM. Black boxes highlight entries with substantial differences.

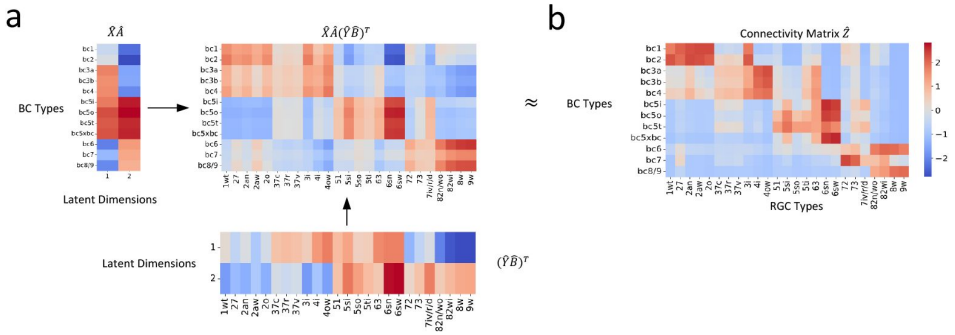



**Figure 4.**

Reconstruction of connectivity map from gene expression profiles. (a) The reconstructed connectivity matrix, derived from the shared latent feature space projections. (b) The connectivity matrix obtained from connectomic data. Differences in color intensity represent the strength of connections, with dark red indicating strong connections and dark blue indicating weak or no connections.

### 6.2.2 Bilinear Model Recapitulates Recognized Connectivity Motifs

Our cross-validation procedure indicated that the optimal number of latent dimensions was two (**Figure S2**). This finding suggested that these two dimensions capture the essential connectivity motifs between BC and RGC types. This led us to further investigate what are these motifs and how they are different from each other.

We first reconstructed connectivity using only the first latent dimension. The first dimension appeared to emphasize connectivity patterns between BCs and RGCs that laminate within the IPL's central region, as well as those that laminate within the marginal region (**Figure 5a,d,g**). We then reconstructed connectivity using only the second latent dimension. Notably, the spotlight shifted to connections between BCs and RGCs that laminate within the outer and inner regions of the IPL, respectively (**Figure 5b,e,i**).

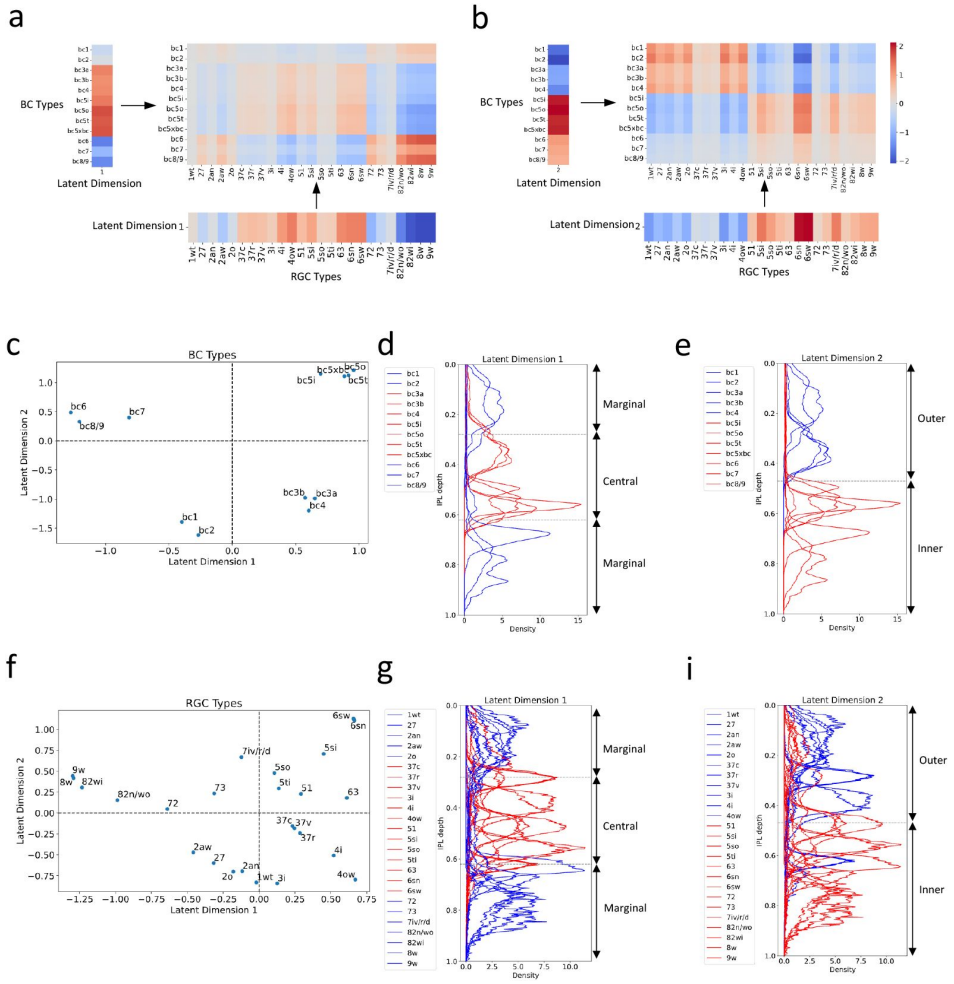
To confirm these observations, we further visualized BC and RGC types within the two-dimensional latent feature space (**Figure 5c,f**). Grouping BC and RGC types based on whether they fell within the positive or negative halves of the latent dimensions, we color-coded their stratification profiles within the IPL by group. BCs and RGCs that fell within the positive half of latent dimension 1 tend to stratify within the IPL's central region, delineated by the boundaries formed by the ON and OFF starburst amacrine cells (SACs) (**Figure 5d,g**). Conversely, those falling within the negative half of this dimension tend to stratify in the marginal region of the IPL. As for the second latent dimension, BCs and RGCs that fell within the positive half predominantly stratify in the inner region of the IPL (**Figure 5e,i**), while those within the negative half primarily stratify in the IPL's outer region.

Interestingly, these distinct connectivity motifs align with two widely recognized properties of retinal neurons: kinetic attributes that reflect the temporal dynamics (transient versus sustained responses) of a neuron responding to visual stimuli, and polarity (ON versus OFF responses) reflecting whether a neuron responds to the initiation or cessation of a stimulus [48, 11, 12, 49]. This correlation implies that our bilinear model has successfully captured key aspects of retinal circuitry from gene expression data.

### 6.2.3 Bilinear Model Reveals Interpretable Insights into Gene Signatures Associated with Different Connectivity Motifs

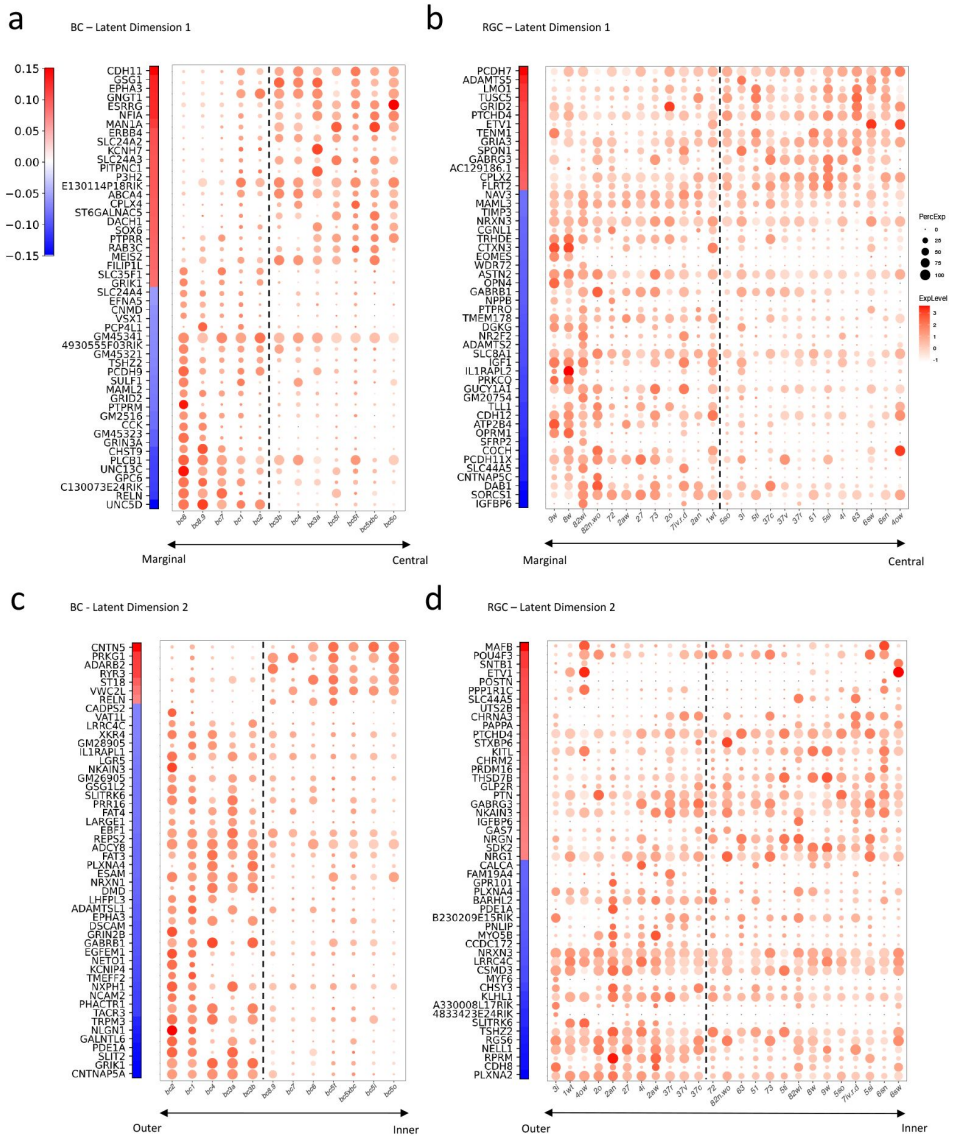
The inherent linearity of our bilinear model affords a significant advantage: it enables the direct interpretation of gene expressions by examining their associated weights in the model. These weights signify the importance of each gene in determining the connectivity motifs between the BC and RGC types. We identified the top 50 genes with the largest positive or negative weights for BCs and RGCs across both latent dimensions. We plotted their weights alongside their expression profiles in the respective cell types (**Figure 6**).

Our analysis unveiled distinct gene signatures associated with the connectivity motifs revealed by the two latent dimensions. In the first latent dimension, genes like CDH11 and EPHA3, involved in cell adhesion and axon guidance, carried high weights for BCs forming synapses in the IPL's central region. In contrast, for BCs synapsing in the marginal region, we observed high weights in the cell adhesion molecule PCDH9 and the axon guidance cue UNC5D (**Figure 6a**). This pattern was echoed in RGCs but involved a slightly different set of molecules. For example, in RGCs forming synapses in the IPL's central region, the cell adhesion molecule PCDH7 carried high weights, whereas for RGCs synapsing in the marginal region, cell adhesion molecules PCDH11X and CDH12 were associated with high weights (**Figure 6b**).




## Figure 5.

Distinct connectivity motifs revealed by the two latent dimensions. (a, b) The reconstructed connectivity using only latent dimension 1 or 2, respectively. Differences in color intensity represent the strength of connections. (c) BC types plotted in the latent feature space, with each point representing a specific BC type. Dashed lines indicate zero values for latent dimensions 1 and 2. (d, e) Stratification profiles of BC types in IPL, color-coded based on their positions along the first (d) or second (e) latent dimension. Red indicates BC types on the positive half, while blue indicates BC types on the negative half. (f) RGC types plotted in the latent feature space, with each point representing a specific RGC type. (g, h) Stratification profiles of RGC types in IPL, color-coded based on their positions along the first (g) or second (h) latent dimension. Dashed lines in (d) and (g) mark the positions of ON and OFF SACs [36]. BCs and RGCs stratifying between them tend to exhibit more transient responses, and those stratifying outside them exhibit more sustained responses. Dashed lines in (e) and (h) denote the boundary of the outer and inner IPL [36]. Synapses between BCs and RGCs in the outer retina mediate OFF responses, while those in the inner retina mediate ON responses.




## Figure 6.

Gene signatures associated with the two latent dimensions. (a, b) Weight vectors of the top 50 genes for latent dimension 1, along with their expression patterns in BC types (a) and RGC types (b). The weight value is indicated in the color bar, with the sign represented by color (red: positive and blue: negative), and the magnitude by saturation. The expression pattern is represented by the size of each dot (indicating the percentage of cells expressing the gene) and the color saturation (representing the gene expression level). BC and RGC types are sorted by their positions along latent dimension 1, as shown in Figure 5c,f, with the dashed line separating the positive category from the negative category. (c, d) Weight vectors of the top 50 genes for latent dimension 2, and their expression patterns in BC types (c) and RGC types (d), depicted in the same manner as in (a) and (b). BC and RGC types are sorted by their positions along latent dimension 2.

The second latent dimension revealed a comparable pattern, albeit with different gene signatures. For BCs laminating in the IPL's outer region, high weights were assigned with guidance cues such as SLIT2, NLGN1, EPHA3 and PLXNA4, as well as the adhesion molecule DSCAM. For BCs in the inner region, the adhesion molecule CNTN5 was associated with a high weight (**Figure 6c**). In RGCs, we noticed that guidance molecules such as PLXNA2, SLITRK6 and PLXNA4 along with adhesion modules CDH8 and LRRC4C were associated with high weights for cells forming synapses in the IPL's outer region. In contrast, the adhesion molecule SDK2 was among the top genes for RGCs laminating and forming synapses in the IPL's inner region (**Figure 6d**). Some of these genes or gene families, such as Plexins (PLXNA2, PLXNA4), Contactin5 (CNTN5), Sidekick2 (SDK2), and Cadherins (CDH8,11,12), are known to play crucial roles in establishing specific synaptic connections [50, 51, 52, 53, 32, 33, 54]. Others, particularly delta-protocadherins (PCDH7,9,11x), emerged as new candidates potentially mediating specific synaptic connections [3].

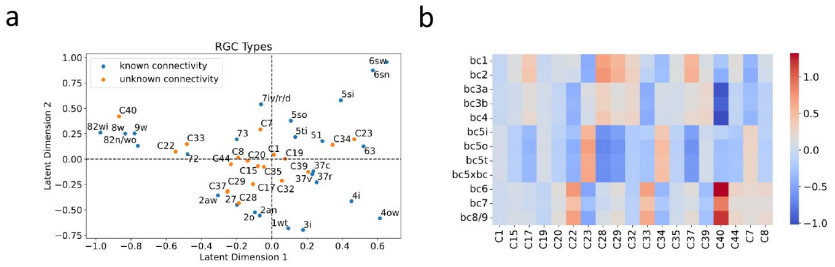
To elucidate the biological implications of these identified gene sets, we further conducted Gene Ontology (GO) enrichment analysis on the top genes through g:Profiler, a public web server for GO enrichment analysis [55, 56]. This tool allowed us to delve into the molecular functions, cellular pathways, and biological processes associated with these genes. Intriguingly, when we listed the top 10 significant GO terms for each latent dimension based on their adjusted p-values, we found two common themes: neuronal development and synaptic organization (**Table S4**). **Table S4** also highlights the number of the top genes associated with each GO term, revealing that overall about 47% of these genes are involved in neural development and synaptic organization. Such findings underscore the potential roles of these genes in forming and shaping the specific connections between BC and RGC types.

### 6.2.4 Bilinear Model Predicts Connectivity Partners of Transcriptomically-Defined RGC Types

The success of recommendation systems in accurately predicting the preferences of new users inspired us to leverage the bilinear model for predicting the connectivity partners of RGC types whose interconnections with BC types remain uncharted. There are some RGC types defined from single-cell transcriptomic data [35], which lack clear correspondence with those identified through connectomics studies [36]. This discrepancy leaves the connectivity patterns of these transcriptionally-defined RGC types unknown, providing an opportunity for our model to predict their BC partners.

To accomplish this, we first projected these RGC types into the same latent space as those used to train the model (**Figure 7a**). We then employed this projection to construct a connectivity matrix between these RGC types and BC types (**Figure 7b**), facilitating educated estimates about their connectivity partners. For each transcriptionally-defined RGC type, we identified the top three BC types as potential partners, determined by the highest values present in the connectivity matrix. These three BC types could provide insight into the potential synaptic input to each RGC type. Detailed predictions are presented in **Table S5**.

Although the ground truth connectivity of these RGC types remains unknown due to the absence of matching types in connectomic data, Goetz et al. [38], via Patch-seq, attempted to match some transcriptomic types with functionally defined RGC types. These functional descriptions may hint at the BC partners of these RGC types. For instance, an RGC exhibiting OFF sustained responses is likely to be synaptically linked with BC types bc1-2, known to mediate OFF sustained pathways. Conversely, an RGC that displays ON sustained responses likely receives synaptic inputs from BC types bc6-9, which oversee ON sustained pathways. We summarized these functional descriptions in **Table S5**, referencing **Figure 5A** from Goetz et al. [38], and highlighted whether our




## Figure 7.

BC partner prediction of transcriptionally-defined RGC types. (a) Projection of transcriptionally-defined RGC types with unknown connectivity into the same latent space as those with known connectivity. (b) The resulting predicted connectivity matrix between these RGC types and BC types. Transcriptionally-defined RGC types are named according to Tran et al. [35]

predictions were consistent with these functional annotations. Among the ten predictions made, eight aligned with these functional descriptions, lending support to the predictive power of our model.

# 7 Discussion

## 7.1 Summary of Study

This study showcased a novel application of the bilinear modeling approach within the realm of gene expression analysis of neuronal type connectivity, drawing inspiration from recommendation systems - a machine learning domain focused on capturing intricate interactions between users and items and predicting user preferences. This analogy served as a useful framework in our study, where the roles of users and items in the recommendation systems are mirrored by presynaptic and postsynaptic neurons, respectively. Likewise, the user-item preference matrix corresponds to the synaptic connection matrix in neural circuits. The recommendation systems are based on the assumption that user preferences and item attributes can be represented by latent factors; similarly, our model assumes that synaptic connectivity between various neuron types is determined by a shared latent feature space derived from gene expression profiles.

The applicability and effectiveness of our bilinear model were validated using two different datasets. Applying it to the *C. elegans* neuronal dataset, which include data of gap junction connectivity and innexin expression at the individual neuron level, we showed that the model could be generalized to single-cell level connectivity by treating each neuronal type as an individual cell (Section 3.1), and incorporate spatial constraints such as physical contact between neurons into the weight matrix (Section 4.1). In a more complex scenario where the transcriptomic and connectomic data are from different sources and aligned at the neuronal-type level, we demonstrated the model's capability in decoding the genetic underlying of the connectivity between neuronal types (Section 3.2), using the mouse retinal neuronal dataset (Section 4.2). This emphasizes the model's potential in offering insights into the genetic mechanisms that orchestrate synaptic connections across various nervous systems.

## 7.2 Insights from Analysis of *C. elegans* Dataset and Comparison with SCM

Using the *C. elegans* neuronal dataset, we conducted a comparative analysis between our bilinear model and the SCM, which correlates neuronal innexin expression with gap junction connectivity via a rule matrix [25, 26]. The SCM incorporates spatial constraints, such as physical contact between neurons, and represents the connectome as an edge list for regression against the Kronecker product of the gene expression matrix. Our model is closely related to the SCM, as it can be seen as factorization of the rule matrix into the product of two lower-dimensional transformation matrices. This factorization not only yielded a performance comparable to, if slightly better than, the SCM in reconstructing the gap junction connectivity matrix, but also revealing potential new innexin interactions for experimental exploration (**Figure 2**; **Figure 3**).

Beyond these, a crucial advantage of our bilinear model lies in in its computational efficiency, an attribute of significance when scaling to larger datasets, where the number of genes and the number of neurons or neuronal types escalates to the order of thousands, such as those of the mouse or macaque cortex [57, 58]. In such situation, the computational complexity of the SCM is substantial, given its reliance on the Kronecker product's dimensions and subsequent matrix inversion. In contrast, the computational demands of our bilinear model, driven primarily by matrix multiplication during gradient descent, are considerably more manageable, offering

scalability and feasibility even as dataset sizes increase. Furthermore, the requirement to calculate the Kronecker product in SCM significantly heightens memory usage, critical when the data scale is large but memory resources are constrained. These advantages ensure our bilinear model a scalable solution when applied to other organisms and brain regions.

In assessing the bilinear model's and the SCM's performance to reconstruct *C. elegans* gap junction connectivity, the resulting modest ROC-AUC scores (approximately 0.64, much lower than the ideal 1.0) underscore the challenges in predicting electrical synapse specificity using innexion expressions alone. This suggests that additional molecular mechanisms, beyond innexin interactions, play crucial roles in forming specific electrical synaptic connections. Indeed, in the realm of chemical synapses, it's increasingly recognized that synaptic specificity is significantly influenced by factors such as cell-cell adhesion and recognition molecules, rather than just the pre- or post-synaptic machinery [3]. Recent studies support this viewpoint. For instance, research on the *C. elegans* motor circuit has revealed how a developmental program fine-tunes cAMP signaling to guide neuron-specific assembly of electrical synapses [59]. Furthermore, the observed coexistence of electrical and chemical synapses in close proximity intimates potential shared mechanisms underlying their specificity [60].

## 7.3 Insights from Application to Mouse Retinal Neuronal Dataset

Applying to the mouse retinal neuronal dataset, our bilinear model successfully reconstructed a neuronal type-specific connectivity map from gene expression profiles and recapitulated two core connectivity motifs of the retinal circuit, representing synapses formed in central or marginal parts of the IPL, and synapses formed in outer or inner regions (**Figure 4**; **Figure 5**). These motifs align well with recognized properties of retinal neurons: kinetic attributes (transient versus sustained responses) and polarity (ON versus OFF responses) [48, 11, 12, 49]. Significantly, these motifs aren't predefined or explicitly encoded into the model; instead, they emerge naturally from the model, further attesting to the model's power to capture key aspects of retinal circuitry.

The bilinear model also revealed unique insights into the gene signatures associated with the connectivity motifs. The weight vectors in the transformation matrices provide a means to assess the relative importance of individual genes. This direct interpretability is a significant advantage of the linear model, allowing for a more intuitive understanding of the gene-to-connectivity transformation process. Our analysis discovered distinct gene signatures associated with different connectivity motifs (**Figure 6**). Among these genes, some have been previously implicated in mediating specific synaptic connections, thererby validating our approach. For instance, Plexins A4 and A2 (PLXNA4, PLXNA2), predicted to be crucial for RGCs' synapsing in the outer IPL, have been shown to be necessary for forming specific lamina of the IPL in the mouse retina, interacting with the guidance molecule Semaphorin 6A (SEM6A) [50, 51]. Contactin5 (CNTN5), which our model predicted as vital for BCs forming synapses in the inner IPL, has been shown to be essential for synapses between ON BCs and the ON lamina of ON-OFF direction-selective ganglion cells (ooDSGCs) [52]. Sidekick2 (SDK2), predicted to be critical for RGCs' synapses in the inner IPL, has been shown to guide the formation of a retinal circuit that detects differential motion [53]. Similarly, Cadherins (CDH8,11,12), whose combinations have been implicated in synaptic specificity within retinal circuits [32, 33], were highlighted for multiple connectivity motifs. In particular, Cadherin8 (CDH8), which our model predicted to be crucial for RGC's synaptic connections in the outer IPL, has been shown to be guided by the transciptional factor Tbr1 for laminar patterning of J-RGCs, a type of OFF direction-selective RGCs [54]. In addition to these validated gene signatures, our analysis identified promising candidate genes that may mediate specific synaptic connections. Particularly, delta-protocadherins (PCDH7,9,11x) appeared as potential new candidates. While their roles in synaptic connectivity aren't fully understood [3], mutations in delta-protocadherins in mice and humans have been linked with various neurological phenotypes, including axon growth and guidance impairments and changes in

synaptic plasticity and stability [61, 62, 63, 64]. Future experimental studies are needed to validate these findings and further unravel the roles of these genes in neural circuit formation and function in the mouse retina.

The bilinear model's utility extends beyond the identification of gene signatures, emerging as a potent tool for hypothesis generation, particularly in predicting connectivity for transcriptionallydefined neuronal types whose synaptic partners remain uncharted (**Figure 7**). Trained on data from a specific neural region, the bilinear model can facilitate the anticipation of synaptic partners for newly characterized transcriptional types within that region, thereby generating hypotheses on their functional roles within neural circuits. Furthermore, this model opens avenues for inferring neural wiring alterations resulting from genetic manipulations. For instance, by altering the genetic profile of certain neuronal types to create new transcriptionally defined types, we can use the model to predict changes in their synaptic partners, offering insights into the consequent reconfiguration of neural networks. This could be further extended to hypothesize the rewiring of the brain under psychological disorders, such as autism, where significant connectome changes suggest shifts in synaptic partner choices [65, 66]. With recent availability of neuronal gene expression data of autism [67, 68, 69], our model stands poised to predict the implications of such genetic profiles on neural circuitry, guiding the research of understanding and treating this psychological disorder.

While our bilinear model offers valuable insights into the connectivity motifs of retinal circuits and the associated gene signatures, with many findings aligning with existing literature, it is important to acknowledge certain limitations of this study. Firstly, the model's connectivity matrix was deduced from stratification profiles derived from EM reconstruction. Although prior research has indicated stratification as a meaningful indicator of connectivity within the mouse retina, as certain BC types preferentially connect with specific RGC types stratified in the same lamina [32, 53, 33], this metric may not capture the entire complexity of synaptic connections [70]. The incorporation of additional experimental data, such as electrophysiological measurements, could enhance both the accuracy and the reliability of the connectivity metrics. Secondly, the model, despite its overall success in reconstructing the connectivity matrix, missed several connections, notably among specific BC-RGC pairs such as those between RGC types 51, 5ti and BC types 3a, 3b, and 4 (**Figure S5**). This highlights the potential for a more complex approach, such as deep learning models, to capture the subtleties of synaptic connections. Finally, the list of top genes identified by our model is enriched with genes directly mediating synapse formation and maintenance, such as adhesion molecules (**Figure 6**; **Table S4**), yet overlooks transcription factors like Tbr1 known to affect synaptic specificity [54]. These factors, impacting various neuronal functionalities, might not be captured by a linear model that inherently favors predictor variables that strongly correlate with the target variable.

# 8 Future Directions

## 8.1 Experiment Validation of Candidate Genes

The bilinear model enables the predictions of possible changes in synaptic connections resulting from changes in expressions of the candidate genes. Emerging genome editing technologies, particularly CRISPR/Cas9 [71, 72], offers a precise and efficient way to validate these predictions through experiments. By leveraging CRISPR/Cas9, targeted genetic manipulations, such as gene silencing or modification, can be conducted to assess their impact on synaptic connectivity. In the context of the mouse retina, the delivery of CRISPR/Cas9 components into BCs or RGCs can be achieved through electroporation or adeno-associated virus (AAV) vectors, respectively, allowing for targeted gene intervention [73, 74].

The finding of delta-protocadherins (PCDH7, 9, 11x) as potential mediators of synaptic specificity in the mouse retina presents an exciting opportunity for experimental exploration. We propose to design CRISPR/Cas9 systems targeting these delta-protocadherins (PCDH7,9,11x), similar to those detailed in a recent study [75]. Delivered to the mouse retina using AAV vectors, we expect to knockdown delta-protocadherin expressions in RGCs [74]. With PCDH7 identified as a key factor in synapse formation within the central regions of the IPL, a focal point of our investigation will be RGC types like W3B RGCs, which are known to stratify in these central layers [76]. The consequences of PCDH7 downregulation on the connectivity of W3B RGCs can be examined through multiple approaches [53]: immunohistochemical techniques or the use of transgenic markers can reveal morphological changes indicative of altered connectivity; electrophysiological assessments, such as targeted recordings from postsynaptic neurons while optogenetically stimulating presynaptic partners, offer a functional probe into the synaptic alterations. Similarly, as PCDH9 and PCDH11x are implicated in synaptic connections within the marginal regions of the IPL, candidate RGCs for examination could include ON and OFF sustained alpha RGCs, known for their peripheral stratifications [77].

This experimental paradigm is not confined to the mouse retina but extends to a broad range of neuronal circuits, thanks to the flexibility and wide applicability of genome editing tools like CRISPR/Cas9 [78, 79, 80]. The capacity to induce targeted gene knockouts or modifications will empower researchers to validate our bilinear model's predictions and explore the underlying genetic mechanisms for synaptic formation and maintenance. This endeavor opens new avenues for deciphering the complex interplay between genetics and neural circuit wiring, furthering our comprehension of the molecular mechanisms driving synaptic specificity.

## 8.2 Application to Other Neural Systems

Our bilinear model, while illustrated using the *C. elegans* and mouse retina datasets, holds significant potential for elucidating the genetic underpinnings of neuronal connectivity across various species and brain regions, contingent upon the availability of comprehensive gene expression profiles and synaptic connection data. For instance, the advent of a comprehensive single-cell transcriptome atlas for the adult fruit fly brain, alongside the recent establishment of its complete connectome, offers a fertile ground for extending our model to decipher the complex neural circuits of *Drosophila* [81, 82].

In the context of the mouse brain, the depth and breadth of single-cell sequencing efforts have unveiled a rich tapestry of transcriptomic cell types across cortex regions and the hippocampus [83, 84, 85, 57]. These efforts, in tandem with connectomic studies that meticulously map neuronal connections, lay a foundation for integrating transcriptomic and connectomic data [86, 87, 46, 88]. Such integration, especially across diverse brain regions, presents an exciting avenue to uncover both neuronal connection mechanisms that are shared by neuronal types across different regions and those unique to specific regions. The scalability of our bilinear model, akin to collaborative filtering's effectiveness in commercial domains, supports the prospect of its cross-regional application. This approach positions our model at the forefront of efforts to explore how gene expression patterns contribute to the diversity of neuronal circuits across brain areas, moving us closer to a holistic understanding of the genetic blueprint of neuronal connectivity throughout the entire brain.

Nevertheless, we recognize the challenge that such well-aligned connectomic and transcriptomic data may not always be readily available. To address this, future research endeavors will also explore adaptations of our model to other available datasets, such as those that combine single-cell transcriptomic profiling with long-range neuronal projection mapping [89, 90]. Furthermore, our model is amenable to integration with trans-synaptic tracer-based sequencing methods [91, 92], expanding its utility in studies where detailed connectomic information is limited. Pursuing

these avenues is pivotal in broadening the model's utility and ensuring its relevance across a wider spectrum of brain connectivity research, making it an invaluable tool in the quest to unravel the complexities of neural circuitry.

## 8.3 Model Advancements

To enhance the model's fidelity and applicability, we propose several advancements. First, we advocate for the integration of auxiliary data types, including electrophysiological data, neuron tracing data, and an array of omics data such as proteomics and epigenetics data, to augment and enrich the model's training base [49, 91, 92, 93, 94]. These data modalities offer complementary insights into neuronal function and connectivity, providing valuable context that can inform and refine the model's predictions.

Second, we envision extending the bilinear model to incorporate non-linear interactions, capturing the intricate dynamics between gene expressions and synaptic connections. A potential pathway for this is through kernel methods or the integration of neural networks, specifically adopting the "two-tower model" framework renowned in modern recommendation systems (**Figure 8**). In this model, each "tower" is a deep neural network that undertakes the non-linear transformation of input features [95, 96]. This architecture has proven effective in capturing complex user-item interactions and could significantly enhance our model's ability to decipher the nuanced relationships between genetics and neural connectivity.

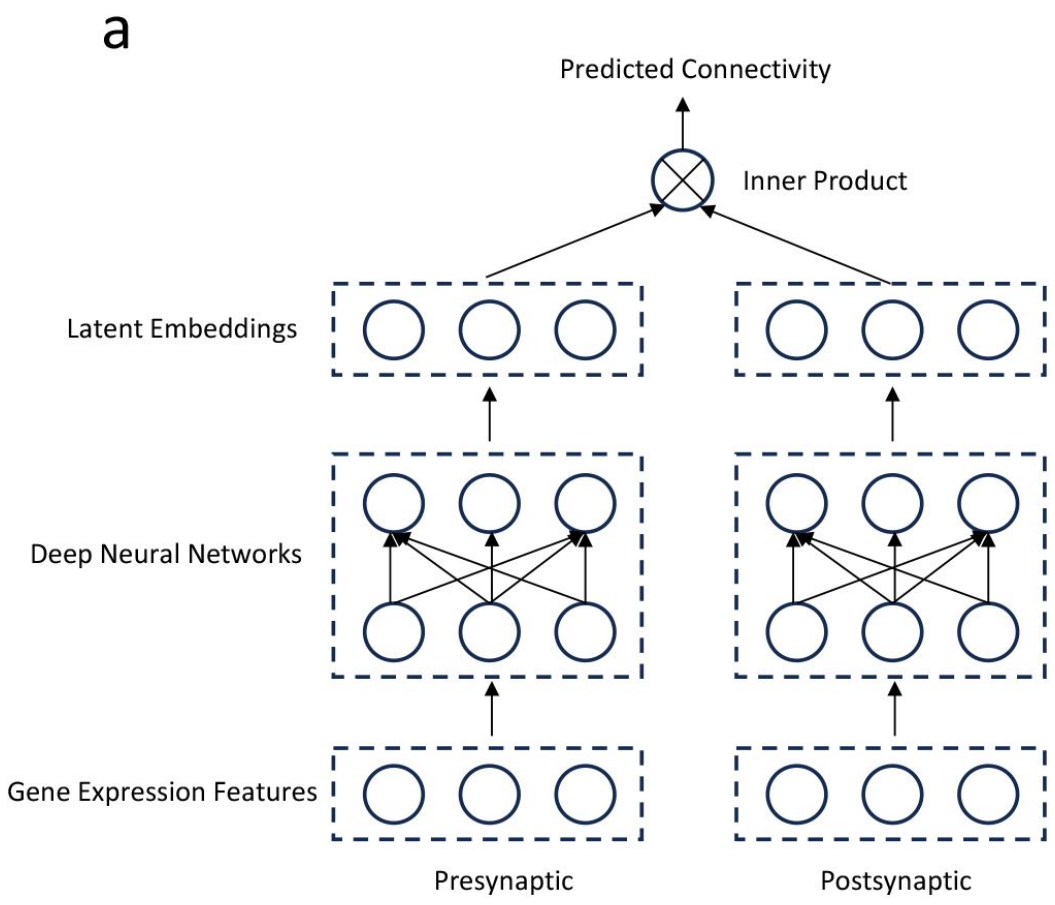


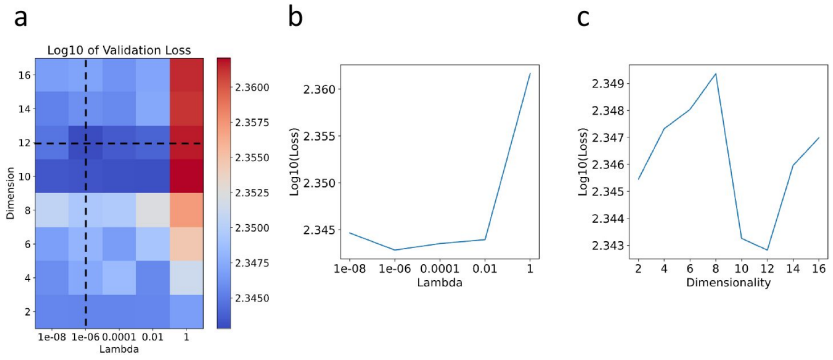

**Figure 8.**

Future direction: A two-tower deep learning model. (a) Gene expression profiles of pre- and post-synaptic neurons are transformed into latent embedding representations via deep neural networks. The connectivity metric between the pre- and post-synaptic neurons is predicted by taking the inner product of their respective latent embeddings.

## 9 Data and Code Availability

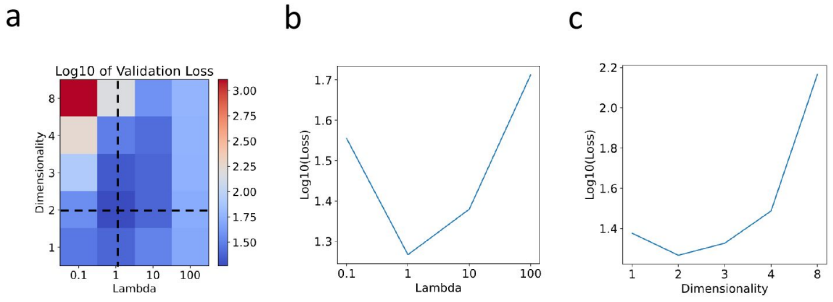
Pointers to the data used in this study and the source code of the bilinear model are available at *https://github.com/muqiao0626/Bilinear_Model*.

## 10 Supplementary Materials



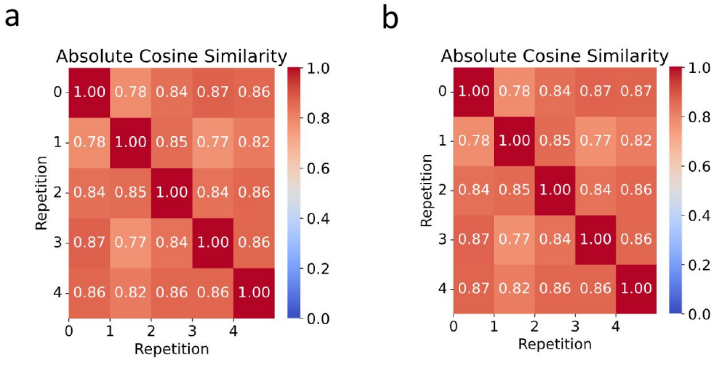

## Supplementary Figure S1

Hyperparameter selection through cross-validation for the *C. elegans* neuronal dataset. (a) Heatmap plot of the logarithm (base 10) of the validation loss, showing variations with respect to λ across [$10^{-8}$, $10^{-6}$, 0.0001, 0.01, 1] and dimensionality across [2, 4, 6, 8, 10, 12, 14, 16]. (b) Plot showing the logarithm (base 10) of the validation loss against λ over the range [$10^{-8}$, $10^{-6}$, 0.0001, 0.01, 1]. (c) Plot displaying the logarithm (base 10) of the validation loss against dimensionality over the range [2, 4, 6, 8, 10, 12, 14, 16].



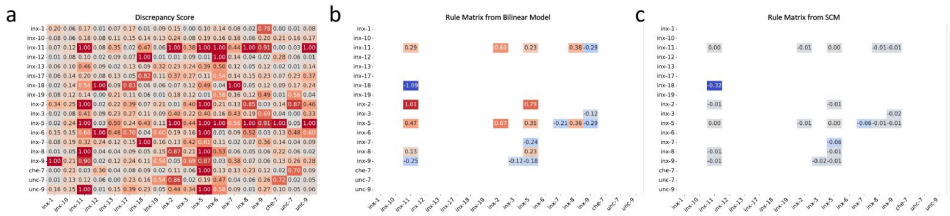

## Supplementary Figure S2

Hyperparameter selection through cross-validation for the mouse retinal neuronal dataset. (a) Heatmap plot of the logarithm (base 10) of the validation loss, showing variations with respect to λ across [0.1, 1, 10, 100] and dimensionality across [1, 2, 3, 4, 8]. (b) Plot showing the logarithm (base 10) of the validation loss against λ over the range [0.1, 1, 10, 100]. (c) Plot displaying the logarithm (base 10) of the validation loss against dimensionality over the range [1, 2, 3, 4, 8].



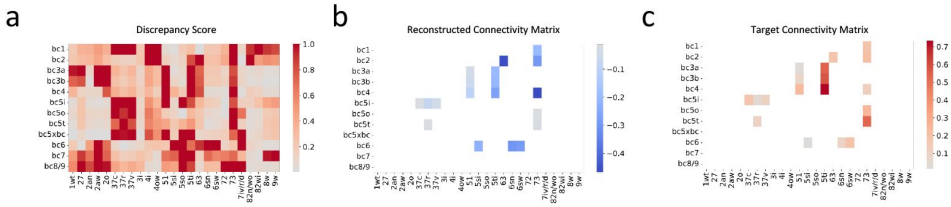

## Supplementary Figure S3

Heatmaps showcasing the average absolute cosine similarities across five optimization repetitions for (a) $\hat{\mathbf{A}}$ and (b) $\hat{B}$. The color scale reflects value of the metric.




## Supplementary Figure S4

Detailed discrepancy analysis between the bilinear model and SCM genetic rules. (a) Discrepancy scores (DS) identifying divergences between the models' rule matrices. (b, c) Significant entries from the bilinear model's rule matrix (b) and the SCM's rule matrix (c), respectively, with DS exceeding 0.5 and matrix entries no less than 0.1.




## Supplementary Figure S5

Detailed discrepancy analysis between the reconstructed and the target connectivity matrices. (a) Discrepancy scores (DS) identifying divergences between the two matrices. (b, c) Specific connections present in the target matrix (c) that were not captured in the reconstructed matrix (b), with DS exceeding 0.5, indicating notable deviations.

| Shekhar et al. 2016 | Greene et al. 2016 |
|---|---|
| BC1a, BC1b | bc1 |
| BC2 | bc2 |
| BC3a | bc3a |
| BC4 | bc4 |
| BC5a | bc5i |
| BC5b | bc5o |
| BC5c | bc5t |
| BC5d | (bc5)xbc |
| BC6 | bc6 |
| BC7 | bc7 |
| BC8, BC9 | bc8/9 |

## Supplementary Table S1

**Correspondence of Mouse BC types [37, 34]**

| Goetz et al. (2022) Type | Bae et al. (2018) Eyewire type | Tran et al. (2019) Cluster |
|---|---|---|
| ON sus alpha | 8w | C43 |
| OFFhOS | 2aw | C9 |
| OFFvOS | 2aw | C5 |
| ON tr SmRF | 6sn | C30 |
| OFF tr alpha | 4ow | C45 |
| OFF tr SmRF | 4i | C21 |
| ONhOS SmRF | 82wi | C27 |
| ONhOS LgRF | 82n/wo | C36 |
| ONvOS SmRF | 72 | C38 |
| ON DS sus DN | 7iv | C10 |
| ON DS sus T | 7ir | C10 |
| ON DS sus V | 7id | C10 |
| OODS D | 37v | C16 |
| OODS T | 37r | C24 |
| HD1 | 5si | C13 |
| HD2 | 5so | C6 |
| UHD | 5ti | C2 |
| LED | 51 | C11 |
| F-mini-ON | 63 | C3 |
| ON delayed | 73 | C14 |
| ON bursty | 3i | C18 |
| bSbC | 2o | C25 |
| sSbC EW27 | 27 | C26 |

## Supplementary Table S2

**Correspondence of Mouse RGC types [36, 35, 38]**

| Goetz et al. (2022) Type | Bae et al. (2018) Eyewire type | Tran et al. (2019) Cluster |
|---|---|---|
| M2 | 9w | C31 |
| OFF sus alpha | 1wt | C42 |
| OODS N | 37c | C12 |
| F-mini-OFF | 2an | C4 |
| ON tr alpha | 6sw | C41 |

## Supplementary Table S3

**Correspondence of Mouse RGC types [36, 35, 38]**

| Term Name | Term ID | Adjusted P-value | Number of Genes |
|---|---|---|---|
| BC Latent Dimension 1 | | | |
| modulation of chemical synaptic transmission | GO:0050804 | 0.000440326 | 9 |
| regulation of trans-synaptic signaling | GO:0099177 | 0.000445672 | 9 |
| synaptic signaling | GO:0099536 | 0.000504129 | 10 |
| nervous system development | GO:0007399 | 0.000591482 | 15 |
| neuron differentiation | GO:0030182 | 0.000776891 | 12 |
| generation of neurons | GO:0048699 | 0.001293397 | 12 |
| cell-cell adhesion via plasma-membrane adhesion molecules | GO:0098742 | 0.003255938 | 6 |
| chemical synaptic transmission | GO:0007268 | 0.003529887 | 9 |
| anterograde trans-synaptic signaling | GO:0098916 | 0.003529887 | 9 |
| trans-synaptic signaling | GO:0099537 | 0.003763337 | 9 |
| RGC Latent Dimension 1 | | | |
| system development | GO:0048731 | 1.14E-05 | 21 |
| nervous system development | GO:0007399 | 1.53E-05 | 17 |
| anatomical structure development | GO:0048856 | 1.67E-05 | 25 |
| synapse | GO:0045202 | 3.22E-05 | 13 |
| somatodendritic compartment | GO:0036477 | 3.61E-05 | 11 |
| multicellular organism development | GO:0007275 | 3.76E-05 | 22 |
| synaptic membrane | GO:0097060 | 4.90E-05 | 8 |
| postsynaptic membrane | GO:0045211 | 5.92E-05 | 7 |
| neurogenesis | GO:0022008 | 0.000126449 | 14 |
| developmental process | GO:0032502 | 0.000135356 | 25 |
| BC Latent Dimension 2 | | | |
| neuron differentiation | GO:0030182 | 1.07E-08 | 17 |
| generation of neurons | GO:0048699 | 2.28E-08 | 17 |
| neuron development | GO:0048666 | 9.30E-08 | 15 |
| neurogenesis | GO:0022008 | 1.98E-07 | 17 |
| neuron projection development | GO:0031175 | 2.18E-07 | 14 |
| cell adhesion | GO:0007155 | 7.00E-07 | 15 |
| synaptic membrane | GO:0097060 | 6.20E-06 | 9 |
| cell periphery | GO:0071944 | 8.75E-06 | 27 |
| plasma membrane | GO:0005886 | 1.07E-05 | 26 |
| nervous system development | GO:0007399 | 2.36E-05 | 17 |
| RGC Latent Dimension 2 | | | |
| neuron projection development | GO:0031175 | 0.00289625 | 10 |
| nervous system development | GO:0007399 | 0.00563224 | 14 |
| neuron projection morphogenesis | GO:0048812 | 0.0067497 | 8 |
| neurogenesis | GO:0022008 | 0.00734347 | 12 |
| plasma membrane bounded cell projection morphogenesis | GO:0120039 | 0.00789142 | 8 |
| neuron differentiation | GO:0030182 | 0.00817177 | 11 |
| cell projection morphogenesis | GO:0048858 | 0.008392 | 8 |
| neuron development | GO:0048666 | 0.00937721 | 10 |
| synaptic membrane | GO:0097060 | 0.00944134 | 6 |
| cell part morphogenesis | GO:0032990 | 0.01143062 | 8 |

## Supplementary Table S4

**Gene Ontology (GO) Terms Associated with Latent Dimensions in BCs and RGCs**

| Tran et al. (2019) | Predicted BC Partners | Based on Goetz et al. (2022) Figure 5A | Consistent |
|---|---|---|---|
| C1: W3L1 | bc5o, bc5xbc, bc5t | Match to 6 functional types | - |
| C15 | bc2, bc1, bc8/9 | - | - |
| C17 | bc2, bc1, bc4 | - | - |
| C19 | bc5o, bc5xbc, bc5t | "HD2" like: Between SAC bands | Yes |
| C20 | bc6, bc8/9, bc7 | - | - |
| C22: MX | bc6, bc8/9, bc7 | "M2" like: ON, Sustained | Yes |
| C23: W3D2 | bc5o, bc5xbc, bc5t | "OFF tr alpha" like: OFF, Transient | No |
| C28: F-midi-OFF | bc2, bc1, bc4 | Match to 3 functional types | - |
| C29 | bc2, bc1, bc4 | "OFF sus alpha" like: OFF, Sustained | Yes |
| C32 | bc2, bc4, bc1 | Match to 2 functional types | - |
| C33: M1a | bc6, bc8/9, bc7 | "M2" like: ON, Sustained | Yes |
| C34 | bc5o, bc5xbc, bc5t | "ON tr MeRF" like: ON, Transient | Yes |
| C35 | bc2, bc1, bc4 | Match to 4 functional types | - |
| C37 | bc2, bc1, bc4 | "OFF tr MeRF" like: OFF, Transient | Yes |
| C39 | bc4, bc3a, bc3b | - | - |
| C40: M1b | bc6, bc8/9, bc7 | "M1": ON, Sustained | Yes |
| C44 | bc6, bc8/9, bc7 | "LED" like: Between SAC bands | No |
| C7 | bc5o, bc5i, bc5xbc | - | - |
| C8 | bc6, bc8/9, bc7 | "PixON" like: ON, Sustained | Yes |

## Supplementary Table S5

**Predicted BC Partners of Transciptionally-defined RGC Types**

# References


[1] Seung Sebastian (2012) **Connectome: How the Brain's Wiring Makes Us Who We Are**

[2] Polleux F., Snider William (2010) **Initiating and growing an axon** *Cold Spring Harb Perspect Biol* **2**

[3] Sanes Joshua R., Zipursky S. Lawrence (2020) **Synaptic Specificity, Recognition Molecules, and Assembly of Neural Circuits** *Cell* **181**:1434–1435

[4] Zeng Hongkui, Sanes Joshua R. (2017) **Neuronal cell-type classification: Challenges, opportunities and the path forward** *Nat Rev Neurosci* **18**:530–546

[5] Stegle Oliver, Teichmann Sarah A., Marioni John C. (2015) **Computational and analytical challenges in single-cell transcriptomics** *Nat Rev Genet* **16**:133–145

[6] Denk Winfried, Horstmann Heinz (2004) **Serial Block-Face Scanning Electron Microscopy to Reconstruct Three-Dimensional Tissue Nanostructure** *PLOS Biology* **2**

[7] Helmstaedter Moritz, Briggman Kevin L., Turaga Srinivas C., Jain Viren, Seung H. Sebastian, Denk Winfried (2013) **Connectomic reconstruction of the inner plexiform layer in the mouse retina** *Nature* **500**:168–174

[8] Tapia Juan Carlos, Kasthuri Narayanan, Hayworth Kenneth J., Schalek Richard, Lichtman Jeff W., Smith Stephen J., Buchanan JoAnn (2012) **High-contrast en bloc staining of neuronal tissue for field emission scanning electron microscopy** *Nat Protoc* **7**:193–206

[9] Koren Yehuda, Bell Robert, Volinsky Chris (2009) **Matrix Factorization Techniques for Recommender Systems** *Computer* **42**:30–37

[10] Martin E. Anne, Lasseigne Abagael M., Miller Adam C. (2020) **Understanding the Molecular and Cell Biological Mechanisms of Electrical Synapse Formation** *Front Neuroanat* **14**

[11] Euler Thomas, Haverkamp Silke, Schubert Timm, Baden Tom (2014) **Retinal bipolar cells: Elementary building blocks of vision** *Nat Rev Neurosci* **15**:507–519

[12] Sanes Joshua R., Masland Richard H. (2015) **The Types of Retinal Ganglion Cells: Current Status and Implications for Neuronal Classification** *Annual Review of Neuroscience* **38**:221–246

[13] Gollisch Tim, Meister Markus (2010) **Eye smarter than scientists believed: Neural computations in circuits of the retina** *Neuron* **65**:150–164

[14] da Silveira Rava Azeredo, Roska Botond (2011) **Cell types, circuits, computation** *Curr Opin Neurobiol* **21**:664–671

[15] Kumar Nalin M., Gilula Norton B. (1996) **The Gap Junction Communication Channel** *Cell* **84**:381–388

[16] Phelan Pauline, Bacon Jonathan P., Davies Jane A., Stebbings Lucy A., Todman Martin G. (1998) **Innexins: A family of invertebrate gap-junction proteins** *Trends in Genetics* **14**:348–349

[17] Rabinowitch Ithai, Schafer William R (2015) **Engineering new synaptic connections in the C. elegans connectome** *Worm* **4**

[18] Marcus Gary, Marblestone Adam, Dean Thomas (2014) **The atoms of neural computation** *Science* **346**:551–552

[19] Südhof Thomas C. (2017) **Synaptic Neurexin Complexes: A Molecular Code for the Logic of Neural Circuits** *Cell* **171**:745–769

[20] Hall David H. (2017) **Gap junctions in C. elegans: Their roles in behavior and development** *Dev Neurobiol* **77**:587–596

[21] de Wit Joris, Ghosh Anirvan (2016) **Specification of synaptic connectivity by cell surface interactions** *Nat Rev Neurosci* **17**:22–35

[22] Fornito Alex, Arnatkevičiūtė Aurina, Fulcher Ben D. (2019) **Bridging the Gap between Connectome and Transcriptome** *Trends Cogn Sci* **23**:34–50

[23] Kaufman Alon, Dror Gideon, Meilijson Isaac, Ruppin Eytan (2006) **Gene Expression of Caenorhabditis elegans Neurons Carries Information on Their Synaptic Connectivity** *PLoS Comput Biol* **2**

[24] Varadan Vinay, Miller David M., Anastassiou Dimitris (2006) **Computational inference of the molecular logic for synaptic connectivity in C. elegans** *Bioinformatics* **22**:e497–506

[25] Kovács István A., Barabási Dániel L., Barabási Albert-László (2020) **Uncovering the genetic blueprint of the C. elegans nervous system** *Proceedings of the National Academy of Sciences* **117**:33570–33577

[26] Barabási Dániel L., Barabási Albert-László (2020) **A Genetic Model of the Connectome** *Neuron* **105**:435–445

[27] Taylor Seth R. *et al.* (2021) **Molecular topography of an entire nervous system** *Cell* **184**:4329–4347

[28] Zeng Hongkui (2022) **What is a cell type and how to define it?** *Cell* **185**:2739–2755

[29] Ricci Francesco, Rokach Lior, Shapira Bracha, Ricci Francesco, Rokach Lior, Shapira Bracha, Kantor Paul B. (2011) **Introduction to Recommender Systems Handbook** *Recommender Systems Handbook* :1–35

[30] Su Xiaoyuan, Khoshgoftaar Taghi M. (2009) **A Survey of Collaborative Filtering Techniques** *Advances in Artificial Intelligence* **2009**

[31] Rendle Steffen, Freudenthaler Christoph, Gantner Zeno, Schmidt-Thieme Lars (2012) **BPR: Bayesian Personalized Ranking from Implicit Feedback** *Comment: Appears in Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI2009)*

[32] Duan Xin, Krishnaswamy Arjun, Huerta Irina De la, Sanes Joshua R. (2014) **Type II cadherins guide assembly of a direction-selective retinal circuit** *Cell* **158**:793–807

[33] Duan Xin, Krishnaswamy Arjun, Laboulaye Mallory A., Liu Jinyue, Peng Yi-Rong, Yamagata Masahito, Toma Kenichi, Sanes Joshua R. (2018) **Cadherin Combinations Recruit Dendrites of Distinct Retinal Neurons to a Shared Interneuronal Scaffold** *Neuron* **99**:1145–1154

[34] Shekhar Karthik *et al.* (2016) **COMPREHENSIVE CLASSIFICATION OF RETINAL BIPOLAR NEURONS BY SINGLE-CELL TRANSCRIPTOMICS** *Cell* **166**:1308–1323

[35] Tran Nicholas M. *et al.* (2019) **Single-cell profiles of retinal neurons differing in resilience to injury reveal neuroprotective genes** *bioRxiv*

[36] Bae J. Alexander *et al.* (2018) **Digital Museum of Retinal Ganglion Cells with Dense Anatomy and Physiology** *Cell* **173**:1293–1306

[37] Greene Matthew J., Kim Jinseop S., Seung H. Sebastian, EyeWirers (2016) **Analogous Convergence of Sustained and Transient Inputs in Parallel On and Off Pathways for Retinal Motion Computation** *Cell Rep* **14**:1892–1900

[38] Goetz Jillian *et al.* (2022) **Unified classification of mouse retinal ganglion cells using function, morphology, and gene expression** *Cell Rep* **40**

[39] Butler Andrew, Hoffman Paul, Smibert Peter, Papalexi Efthymia, Satija Rahul (2018) **Integrating single-cell transcriptomic data across different conditions, technologies, and species** *Nat Biotechnol* **36**:411–420

[40] Stuart Tim, Butler Andrew, Hoffman Paul, Hafemeister Christoph, Papalexi Efthymia, Mauck William M., Hao Yuhan, Stoeckius Marlon, Smibert Peter, Satija Rahul (2019) **Comprehensive Integration of Single-Cell Data** *Cell* **177**:1888–1902

[41] Cook Steven J. *et al.* (2019) **Whole-animal connectomes of both Caenorhabditis elegans sexes** *Nature* **571**:63–71

[42] Qiao Mu (2023) **Factorized Discriminant Analysis for Genetic Signatures of Neuronal Phenotypes** *arXiv preprint*

[43] Chen Hung-I. Harry, Jin Yufang, Huang Yufei, Chen Yidong (2016) **Detection of high variability in gene expression from single-cell RNA-seq profiling** *BMC Genomics* **17**

[44] Pandey Shristi, Shekhar Karthik, Regev Aviv, Schier Alexander F. (2018) **Comprehensive Identification and Spatial Mapping of Habenular Neuronal Types Using Single-Cell RNA-Seq** *Curr. Biol* **28**:1052–1065

[45] Kurmangaliyev Yerbol Z, Yoo Juyoun, LoCascio Samuel A, Lawrence Zipursky S (2019) **Modular transcriptional programs separately define axon and dendrite connectivity** *eLife* **8**

[46] Turner Nicholas L. *et al.* (2022) **Reconstruction of neocortex: Organelles, compartments, cells, circuits, and activity** *Cell* **185**:1082–1100

[47] Kim Jinseop S. *et al.* (2014) **Space–time wiring specificity supports direction selectivity in the retina** *Nature* **509**:331–336

[48] Masland Richard H. (2012) **The Neuronal Organization of the Retina** *Neuron* **76**:266–280

[49] Baden Tom, Berens Philipp, Franke Katrin, Rosón Miroslav Román, Bethge Matthias, Euler Thomas (2016) **The functional diversity of retinal ganglion cells in the mouse** *Nature* **529**:345–350

[50] Matsuoka Ryota L., Nguyen-Ba-Charvet Kim T., Parray Aijaz, Badea Tudor C., Chédotal Alain, Kolodkin Alex L. (2011) **Transmembrane semaphorin signalling controls laminar stratification in the mammalian retina** *Nature* **470**:259–263

[51] Sun Lu O., Jiang Zheng, Rivlin-Etzion Michal, Hand Randal, Brady Colleen M., Matsuoka Ryota L., Yau King-Wai, Feller Marla B., Kolodkin Alex L. (2013) **On and off retinal circuit assembly by divergent molecular mechanisms** *Science* **342**

[52] Peng Yi-Rong, Tran Nicholas M., Krishnaswamy Arjun, Kostadinov Dimitar, Martersteck Emily M., Sanes Joshua R. (2017) **Satb1 Regulates Contactin 5 to Pattern Dendrites of a Mammalian Retinal Ganglion Cell** *Neuron* **95**:869–883

[53] Krishnaswamy Arjun, Yamagata Masahito, Duan Xin, Hong Y. Kate, Sanes Joshua R. (2015) **Sidekick 2 directs formation of a retinal circuit that detects differential motion** *Nature* **524**:466–470

[54] Liu Jinyue, Reggiani Jasmine D. S., Laboulaye Mallory A., Pandey Shristi, Chen Bin, Rubenstein John L. R., Krishnaswamy Arjun, Sanes Joshua R. (2018) **Tbr1 instructs laminar patterning of retinal ganglion cell dendrites** *Nat Neurosci* **21**:659–670

[55] Reimand Jüri, Kull Meelis, Peterson Hedi, Hansen Jaanus, Vilo Jaak (2007) **G:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments** *Nucleic Acids Res* **35**:W193–200

[56] Raudvere Uku, Kolberg Liis, Kuzmin Ivan, Arak Tambet, Adler Priit, Peterson Hedi, Vilo Jaak (2019) **G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update)** *Nucleic Acids Res* **47**:W191–W198

[57] Yao Zizhen *et al.* (2023) **A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain** *Nature* **624**:317–332

[58] Chen Ao *et al.* (2023) **Single-cell spatial transcriptome reveals cell-type organization in the macaque cortex** *Cell* **186**:3726–3743

[59] Palumbos Sierra Danielle (2021) **Molecular Determinants of Electrical Synaptic Specificity. Thesis**

[60] Lasseigne Abagael M, Echeverry Fabio A, Ijaz Sundas, Michel Jennifer Carlisle, Anne Martin E, Marsh Audrey J, Trujillo Elisa, Marsden Kurt C, Pereda Alberto E, Miller Adam C (2021) **Electrical synaptic transmission requires a postsynaptic scaffolding protein** *eLife* **10**

[61] Kahr Irene, Vandepoele Karl, van Roy Frans (2013) **Delta-protocadherins in health and disease** *Prog Mol Biol Transl Sci* **116**:169–192

[62] Light Sarah E.W., Jontes James D. (2017) **δ-Protocadherins: Organizers of neural circuit assembly** *Semin Cell Dev Biol* **69**:83–90

[63] Peek Stacey, Mah Kar Men, Weiner Joshua A. (2017) **Regulation of Neural Circuit Formation by Protocadherins** *Cell Mol Life Sci* **74**:4133–4157

[64] Bisogni Adam J, Ghazanfar Shila, Williams Eric O, Marsh Heather M, Yang Jean YH, Lin David M (2018) **Tuning of delta-protocadherin adhesion through combinatorial diversity** *eLife* **7**

[65] Roine Ulrika, Roine Timo, Salmi Juha, Nieminen-von Wendt Taina, Tani Pekka, Leppämäki Sami, Rintahaka Pertti, Caeyenberghs Karen, Leemans Alexander, Sams Mikko (2015) **Abnormal wiring of the connectome in adults with high-functioning autism spectrum disorder** *Molecular Autism* **6**

[66] Hong Seok-Jun *et al.* (2019) **Atypical functional connectome hierarchy in autism** *Nat Commun* **10**

[67] Velmeshev Dmitry, Schirmer Lucas, Jung Diane, Haeussler Maximilian, Perez Yonatan, Mayer Simone, Bhaduri Aparna, Goyal Nitasha, Rowitch David H., Kriegstein Arnold R. (2019) **Single-cell genomics identifies cell type–specific molecular changes in autism** *Science* **364**:685–689

[68] Nassir Nasna *et al.* (2021) **Single-cell transcriptome identifies molecular subtype of autism spectrum disorder impacted by de novo loss-of-function variants regulating glial cells** *Human Genomics* **15**

[69] Li Chong *et al.* (2023) **Single-cell brain organoid screening identifies developmental defects in autism** *Nature* **621**:373–380

[70] Dunn Felice A, Wong Rachel O L (2014) **Wiring patterns in the mouse retina: Collecting evidence across the connectome, physiology and light microscopy** *J Physiol* **592**:4809–4823

[71] Le Cong F. Ann Ran, Cox David, Lin Shuailiang, Barretto Robert, Habib Naomi, Hsu Patrick D., Wu Xuebing, Jiang Wenyan, Marraffini Luciano A., Zhang Feng (2013) **Multiplex genome engineering using CRISPR/Cas systems** *Science* **339**:819–823

[72] Mali Prashant, Yang Luhan, Esvelt Kevin M., Aach John, Guell Marc, DiCarlo James E., Norville Julie E., Church George M. (2013) **RNA-guided human genome engineering via Cas9** *Science* **339**:823–826

[73] Sarin Sumeet *et al.* (2018) **Role for Wnt Signaling in Retinal Neuropil Development: Analysis via RNA-Seq and In Vivo Somatic CRISPR Mutagenesis** *Neuron* **98**:109–126

[74] Tian Feng *et al.* (2022) **Core transcription programs controlling injury-induced neurodegeneration of retinal ganglion cells** *Neuron* **110**:2607–2624

[75] Biswas Sayantanee, Emond Michelle R., Chenoweth Kurtis P., Jontes James D. (2021) **δ-Protocadherins regulate neural progenitor cell division by antagonizing Ryk and Wnt/β-catenin signaling** *iScience* **24**

[76] Zhang Yifeng, Kim In-Jung, Sanes Joshua R., Meister Markus (2012) **The most numerous ganglion cell type of the mouse retina is a selective feature detector** *Proc Natl Acad Sci U S A* **109**:E2391–E2398

[77] Krieger Brenna, Qiao Mu, Rousso David L., Sanes Joshua R., Meister Markus (2017) **Four alpha ganglion cell types in mouse retina: Function, structure, and molecular signatures** *PLOS ONE* **12**

[78] Dickinson Daniel J., Goldstein Bob (2016) **CRISPR-Based Methods for Caenorhabditis elegans Genome Engineering** *Genetics* **202**:885–901

[79] Gratz Scott J., Rubinstein C. Dustin, Harrison Melissa M., Wildonger Jill, O'Connor-Giles Kate M. (2015) **CRISPR-Cas9 genome editing in Drosophila** *Curr Protoc Mol Biol* **111**

[80] Li Mingyu, Zhao Liyuan, Page-McCaw Patrick, Chen Wenbiao (2016) **Zebrafish genome engineering using the CRISPR-Cas9 system** *Trends Genet* **32**:815–827

[81] Davie Kristofer *et al.* (2018) **A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain** *Cell* **174**:982–998

[82] Dorkenwald Sven *et al.* (2023) **Neuronal wiring diagram of an adult brain** *bioRxiv*

[83] Tasic Bosiljka *et al.* (2016) **Adult mouse cortical cell taxonomy revealed by single cell transcriptomics** *Nat. Neurosci* **19**:335–346

[84] Tasic Bosiljka *et al.* (2018) **Shared and distinct transcriptomic cell types across neocortical areas** *Nature* **563**:72–78

[85] Yao Zizhen *et al.* (2021) **A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation** *Cell* **184**:3222–3241

[86] Bock Davi D., Lee Wei-Chung Allen, Kerlin Aaron M., Andermann Mark L., Hood Greg, Wetzel Arthur W., Yurgenson Sergey, Soucy Edward R., Kim Hyon Suk, Reid R. Clay (2011) **Network anatomy and in vivo physiology of visual cortical neurons** *Nature* **471**:177–182

[87] Lee Wei-Chung Allen, Bonin Vincent, Reed Michael, Graham Brett J., Hood Greg, Glattfelder Katie, Reid R. Clay (2016) **Anatomy and function of an excitatory network in the visual cortex** *Nature* **532**:370–374

[88] Yao Shenqin *et al.* (2023) **A whole-brain monosynaptic input connectome to neuron classes in mouse visual cortex** *Nat Neurosci* **26**:350–364

[89] Chen Xiaoyin, Sun Yu-Chi, Zhan Huiqing, Kebschull Justus M., Fischer Stephan, Matho Katherine, Huang Z. Josh, Gillis Jesse, Zador Anthony M. (2019) **High-Throughput Mapping of Long-Range Neuronal Projection Using In Situ Sequencing** *Cell* **179**:772–786

[90] Sun Yu-Chi, Chen Xiaoyin, Fischer Stephan, Lu Shaina, Zhan Huiqing, Gillis Jesse, Zador Anthony M. (2021) **Integrating barcoded neuroanatomy with spatial transcriptional profiling enables identification of gene correlates of projections** *Nat Neurosci* **24**:873–885

[91] Tsai Nicole Y. *et al.* (2022) **Trans-Seq maps a selective mammalian retinotectal synapse instructed by Nephronectin** *Nat Neurosci* **25**:659–674

[92] Zhang Aixin *et al.* (2023) **Rabies virus-based barcoded neuroanatomy resolved by single-cell RNA and in situ sequencing** *bioRxiv*

[93] Mazan-Mamczarz Krystyna, Ha Jisu, De Supriyo, Sen Payel (2022) **Single-Cell Analysis of the Transcriptome and Epigenome** *Methods Mol Biol* **2399**:21–60

[94] Bennett Hayley M., Stephenson William, Rose Christopher M., Darmanis Spyros (2023) **Single-cell proteomics enabled by next-generation sequencing or mass spectrometry** *Nat Methods* **20**:363–374

[95] Wang Tian, Brovman Yuri M., Madhvanath Sriganesh (2021) **Personalized Embedding-based e-Commerce Recommendations at eBay** *arXiv preprint*

[96] Yu Yantao, Wang Weipeng, Feng Zhoutian, Xue Daiyue (2021) **A dual augmented two-tower model for online large-scale recommendation** *KDD*

## Article and author information

**Mu Qiao**
LinkedIn, Mountain View, CA, 94043
**For correspondence:** muqiao0626@gmail.com

## Copyright



## Editors

Reviewing Editor
**Sacha Nelson**
Brandeis University, Waltham, United States of America

Senior Editor
**Sacha Nelson**
Brandeis University, Waltham, United States of America

**Reviewer #1 (Public Review):**

Summary:

In this study, the author aimed to develop a method for estimating neuronal-type connectivity from transcriptomic gene expression data. They sought to develop an interpretable model that could be used to characterize the underlying genetic mechanisms of circuit assembly and connectivity in various neuronal systems.

Strengths:

Many of the proposed suggestions were addressed by the author from the initial review. In general the claims made by the author are more strongly supported by the data and better situated in the literature. A major improvement includes the application of the model to the C. elegans gap junction neuronal system. Despite several key differences in the dataset as compared to the mouse retina data, the proposed model performs comparably to the SCM model currently considered state of the art in the literature (the author should remain cautious about claiming better performance given extremely marginal differences). In section 7.2, the author clearly outlines additional advantages of the proposed model including superior time and space complexity. The overall model performance remains modest, but it learns the same rules as the SCM model as well as other candidate patterns.

As in the initial submission, the bilinear model recapitulates key connectivity motifs for the mouse dataset. The algorithm is shown to converge across several runs affirming its stability/replicability. The model is also extended to predict connectivity on unknown RGC-BC cell type pairs. Without ground truth, the author posits how it should perform based on known functional properties of the RGC type. The hypotheses are confirmed for 8/10 neuronal types with unknown connectivity. The author more clearly describes how this

model can be used experimentally for hypothesis testing and presents a more comprehensive future roadmap regarding validation, avenues for improving the model, and incorporation of growing datasets.

Weaknesses:

While the C Elegans dataset is useful because it enables benchmarking to existing models, the dataset is quite different. The gene expression dimensionality is 18 genes as opposed to over 3000 genes in the mouse dataset. It is a strength that the model still works as intended, but a weakness that the bilinear model could not be tested on a similar mouse dataset. This distinction matters because it remains an open question if the PCA methodology would hold up in a dataset with varied distributions of gene expression. Variations of the PCA methodology could be evaluated further with the present dataset to make the generalizability of the model more convincing.

The Gene Ontology analysis requires more methodological explanation. The author claims, " (the linear nature of the model) enables the direct interpretation of gene expressions by examining their associated weights in the model. These weights signify the importance of each gene in determining the connectivity motifs between the BC and RGC types." If I am correctly understanding the methods, the model weights in each dimension are indexing the importance of a gene expression feature as opposed to the importance of a single gene alone, "the gene expression of the BCs in X and the RGCs in Y were featurized by their respective PCs, resulting in matrices of dimensions 22453 × 11323 and 3779 × 3142, respectively." It would be helpful to explain how gene weights are extracted from a gene expression feature once highlighted.

There could be a more rigorous analysis of the predictive capacity of the model even with the current data. The model recapitulates connectivity patterns from the full dataset and a prediction is demonstrated for unknown data. The model is thus championed as a useful tool for predicting how genetic modifications will influence connectivity, but this is not empirically evaluated.

Appraisal of whether the author achieved their aims, and whether results support their conclusions:

In line with the aims of the paper, the author proposed an interpretable bilinear model to learn a shared latent feature space derived from gene expression profiles to predict synaptic connectivity between various neuron types. The model was shown to generalize to two distinct neuronal systems with varying levels of genomic and cellular resolution. While the performance remains modest, the model performs comparably to the existing state of the art despite improved computational complexity.

Discussion of likely impact of the work on the field, and utility of methods and data to the community:

The author has elaborated substantially on the impact of this work, particularly how it could be leveraged in experimental settings. The clear methodology could be implemented by other researchers to test the model on new datasets and for benchmarking novel methods.

https://doi.org/10.7554/eLife.91532.2.sa1

**Reviewer #2 (Public Review):**

Summary:

In this study, Mu Qiao employs a bilinear modelling approach, commonly utilised in the recommendation systems, to explore the intricate neural connections between different pre- and post-synaptic neuronal types. This approach involves projecting single-cell Transcriptomic datasets of pre- and post-synaptic neuronal types into a latent space through transformation matrices. Subsequently, the cross-correlation between these projected latent spaces is employed to estimate neuronal connectivity. To facilitate the model training, Connectomic data is used to estimate the ground-truth connectivity map. This work introduces a promising model for the exploration of neuronal connectivity and its associated molecular determinants. In the revised version of the manuscript, the author has applied and validated the model in both C. elegans gap junction connectivity and the retina neuron connectivity conditions.

Strengths:

This study introduces a succinct yet promising computational model for investigating connections between neuronal types. The model, while straightforward, effectively integrates single-cell transcriptomic and connectomic data to produce a reasonably accurate connectivity map, particularly within the context of retinal connectivity. Furthermore, it successfully recapitulates connectivity patterns and helps uncover the genetic factors that underlie these connections.

Weaknesses:

(1) When compared with the previous method - SCM, the new model shows a similar performance level. This may be due to the limitation of the dataset itself, as it only has the innexin expression data. Is it possible to apply the SCM model to the more complete retina dataset and compare the performance with the proposed bilinear modelling approach?

Minor Weakness:

(1) The study lacks experimental validation of the model's prediction results.

https://doi.org/10.7554/eLife.91532.2.sa0

**Author response:**

The following is the authors' response to the original reviews.

> ***eLife assessment***
>
> *This is a valuable computational study that applies the machine learning method of bilinear modeling to the problem of relating gene expression to connectivity. Specifically, the author attempts to use transcriptomic data from mouse retinal neurons to predict their known connectivity. The results are promising, although the reviewers felt that demonstration of the general applicability of the approach required testing it against a second data set. Hence the present results were felt to provide borderline incomplete support for a key premise of the paper.*

We thank the reviewers for their insightful and constructive feedback. In response to the reviews, we have undertaken a comprehensive revision of our manuscript, incorporating changes and improvements as outlined below:

(1) New results have been included showcasing the application of our bilinear model to a seconddataset focusing on C. elegans gap junction connectivity. This extension validates our

model with a biological context other than mouse retina and facilitates a direct comparison with the spatial connectome model (SCM).

(2) A new section titled "Previous Approaches" has been added to background, situating our studywithin the broader landscape of existing modeling methodologies.

(3) The discussion sections have been expanded to fully incorporate the suggestions and insightsoffered by the reviewers. This includes a deeper exploration of the implications of our findings, potential applications of our model, and a more thorough consideration of its limitations and future directions.

(4) To streamline the main text and ensure that the core narrative remains focused and accessible, select figures and tables have been relocated to the "Supplementary Materials" section.

> ***Reviewer 1 (Public Review):***
>
> *Summary of what the author was trying to achieve: In this study, the author aimed to develop a method for estimating neuronal-type connectivity from transcriptomic gene expression data, specifically from mouse retinal neurons. They sought to develop an interpretable model that could be used to characterize the underlying genetic mechanisms of circuit assembly and connectivity.*
>
> *Strengths:*
>
> *The proposed bilinear model draws inspiration from commonly implemented recommendation systems in the field of machine learning. The author presents the model clearly and addresses critical statistical limitations that may weaken the validity of the model such as multicollinearity and outliers. The author presents two formulations of the model for separate scenarios in which varying levels of data resolution are available. The author effectively references key work in the field when establishing assumptions that affect the underlying model and subsequent results. For example, correspondence between gene expression cell types and connectivity cell types from different references are clearly outlined in Tables 1-3. The model training and validation are sufficient and yield a relatively high correlation with the ground truth connectivity matrix. Seemingly valid biological assumptions are made throughout, however, some assumptions may reduce resolution (such as averaging over cell types), thus missing potentially important single-cell gene expression interactions.*

Thank you for recognizing the strengths of our work, particularly the clarity of the model presentation and its foundation in recommendation systems. In the revised manuscript we have also extended the model's capabilities to analyze gene interactions for neural connectivity at single-cell resolution, when gene expression and connectivity of each cell are known simultaneously.

> *Weaknesses:*
>
> *The main results of the study could benefit from replication in another dataset beyond mouse retinal neurons, to validate the proposed method. Dimensionality reduction significantly reduces the resolution of the model and the PCA methodology employed is largely non-deterministic. This may reduce the resolution and reproducibility of the model. It may be worth exploring how the PCA methodology of the model may affect results when replicating. Figure 5, 'Gene signatures associated with the two latent dimensions', lacks some readability and related results could be outlined more clearly in the results section. There should be more discussion on weaknesses of the results e.g.*

> *quantification of what connectivity motifs were not captured and what gene signatures might have been missed.*

We acknowledge the significance of validating our method across different datasets. In line with this, our revised manuscript now includes an expanded analysis utilizing a C. elegans gap junction connectivity dataset, which not only broadens the method's demonstrated applicability but also underscores its versatility across varied neuronal systems.

To address the concern of resolution and reproducibility associated with PCA preprocessing, we have conducted a comparative analysis from five replicates of the bilinear model, presenting the results in the revised manuscript (Figure S3). This analysis confirms the consistency of the solutions, as evidenced by the similarity metrics. Furthermore, we discussed alternative methodologies, such as L1 or L2 regularization, to tackle multicollinearity, offering flexibility in preprocessing choices.

In response to feedback on the original Figure 5's clarity, we have replaced the original Figure 5e-h with Table S4, which summarizes the gene ontology (GO) enrichment results and quantifies the number of genes associated with aspects of neural development and synaptic organization. This revision aims to improve the interpretability and accessibility of the results, ensuring a clearer presentation of the model's insights.

Finally, we have expanded our discussion to address the study's limitations more comprehensively. This includes exploration of potentially missed connections and gene signatures, such as transcription factors, which might not be captured by a linear model due to its inherent preference for predictors with strong correlations to the target variable.

> *The main weakness is the lack of comparison against other similar methods, e.g. methods presented in Barabási, Dániel L., and Albert-László Barabási. "A genetic model of the connectome." Neuron 105.3 (2020): 435-445. Kovács, István A., Dániel L. Barabási, and Albert-László Barabási. "Uncovering the genetic blueprint of the C. elegans nervous system." Proceedings of the National Academy of Sciences 117.52 (2020): 33570-33577. Taylor, Seth R., et al. "Molecular topography of an entire nervous system." Cell 184.16 (2021): 4329-4347.*

We value your suggestion to compare our model with established methods. The revised manuscript now includes a comparative analysis with the spatial connectome model (SCM) using the same C. elegans dataset. In addition, a section reviewing previous approaches has been included in the background part, and the discussion part has been extended for the comparison.

> *Appraisal of whether the author achieved their aims, and whether results support their conclusions: The author achieved their aims by recapitulating key connectivity motifs from single-cell gene expression data in the mouse retina. Furthermore, the model setup allowed for insight into gene signatures and interactions, however could have benefited from a deeper evaluation of the accuracy of these signatures. The author claims the method sets a new benchmark for single-cell transcriptomic analysis of synaptic connections. This should be more rigorously proven. (I'm not sure I can speak on the novelty of the method)*

In the revised manuscript. we emphasized the bilinear model's innovative application in the context of neuronal connectivity analysis, inspired by collaborative filtering in recommendation systems. We present quantitative performance metrics, such as the ROC-AUC score and Pearson correlation coefficient, as well as its comparison with the SCM, to benchmark our model's efficacy in reconstructing connectivity matrices. We also quantified the overlap of the genetic interactions revealed by the bilinear model and the SCM (using the

C. elegans dataset), and reported the percentage of the top genes associated with neural development and synaptic organization (using the mouse retina dataset). These numbers set a precedent for future methodological comparisons.

> *Discussion of the likely impact of the work on the field, and the utility of methods and data to the community : This study provides an understandable bilinear model for decoding the genetic programming of neuronal type connectivity. The proposed model leaves the door open for further testing and comparison with alternative linear and/or non-linear models, such as neural networkbased models. In addition to more complex models, this model can be built on to include higher resolution data such as more gene expression dimensions, different types of connectivity measures, and additional omics data.*

We are grateful for your recognition of the study's potential impact. The bilinear model indeed offers a foundation for future explorations, allowing for integration with more complex models, higher-resolution data, and diverse connectivity measures.

> ***Reviewer 1 (Recommendations For The Authors):***
>
> *The inclusion of predicted connectivity (Figure 6) of unknown BC neurons is useful as it shows that this is a strong hypothesis generation tool. This utility should potentially be showcased more as it is also brought up in the abstract, "genetic manipulation of circuit wiring", with an explanation of how the model could be leveraged as such. The discussion may benefit from a summarizing sentence regarding which key gene signatures were identified and are in line with the literature, which key gene signatures/connectivity motifs may have been missed, and which gene signatures are novel.*

Thank you for the insightful recommendation on emphasizing the model's utility in generating hypotheses, particularly regarding predicting connectivity. In the revised manuscript, we have expanded the discussion on how our model can be leveraged to guide genetic manipulations at altering circuit wiring and highlighted its potential impact in the field.

We have discussed key gene signatures identified from our model that are in line with existing literature, such as plexins and cadherins, which have been previously recognized for their involvement in synaptic connection formation and maintenance. We have also introduced potential new candidates, such as delta-protocadherins. In the revised manuscript, we summarized potentially missed gene signatures or synaptic connections, to provide a comprehensive view of our findings.

> ***Reviewer 2 (Public Review):***
>
> *Summary:*
>
> *In this study, Mu Qiao employs a bilinear modeling approach, commonly utilized in recommendation systems, to explore the intricate neural connections between different pre- and post-synaptic neuronal types. This approach involves projecting single-cell transcriptomic datasets of pre- and post-synaptic neuronal types into a latent space through transformation matrices. Subsequently, the cross-correlation between these projected latent spaces is employed to estimate neuronal connectivity. To facilitate the model training, connectomic data is used to estimate the ground-truth connectivity map. This work introduces a promising model for the exploration of neuronal connectivity and its associated molecular determinants. However, it is important to note that the current model has only been tested with Bipolar Cell and Retinal Ganglion Cell data, and its*

> *applicability in more general neuronal connectivity scenarios remains to be demonstrated.*
>
> *Strengths:*
>
> *This study introduces a succinct yet promising computational model for investigating connections between neuronal types. The model, while straightforward, effectively integrates singlecell transcriptomic and connectomic data to produce a reasonably accurate connectivity map, particularly within the context of retinal connectivity. Furthermore, it successfully recapitulates connectivity patterns and helps uncover the genetic factors that underlie these connections.*

Thank you for your positive assessment of the paper.

> *Weaknesses:*
>
> *(1) The study lacks experimental validation of the model's prediction results.*

We recognize the importance of experimental validation in substantiating the predictions made by computational models. While the primary focus of this study remains computational, we have dedicated a section in the revised manuscript, titled "Experimental Validation of Candidate Genes", to outline proposed methodologies for the empirical verification of our model's predictions. This section specifically discusses the experimental exploration of novel candidate genes, such as deltaprotocadherins, within the mouse retina using AAV-mediated CRISPR/Cas9 genetic manipulation. We plan to collaborate with experimental laboratories to facilitate the validation. Given the extensive nature of experimental work, both in terms of time and resources, it is more pragmatic to present a comprehensive experimental investigation in a follow-up study.

> *(2) The model's applicability in other neuronal connectivity settings has not been thoroughly explored.*

The question of the model's broader applicability is well-taken. In response, we have expanded our analysis to include additional neuronal data and connectivity settings. Specifically, the revised manuscript includes results where we apply the model to a dataset of C. elegans gap junction connectivity, demonstrating its potential in different neuronal systems. This extension serves to illustrate the model's adaptability and potential applicability to a broader range of neuronal connectivity studies.

> *(3) The proposed method relies on the availability of neuronal connectomic data for model training,which may be limited or absent in certain brain connectivity settings.*

We acknowledge the limitations posed by the model's dependency on comprehensive connectomic data, which may not be readily available across all research contexts. To address this, we have discussed in the revised manuscript several alternative strategies to adapt our model to the available data. This includes exploring the potential of applying the model to available data such as projectome, and integrating other data modalities such as electrophysiological measurements. These initiatives aim to enhance the model's applicability and ensure its utility in a broader spectrum of brain connectivity studies, especially in scenarios where detailed connectomic data are not available.

> ***Reviewer 2 (Recommendations For The Authors):***
>
> *Q1. In this work, the author has mainly been studying the retina neuronal type connectivity, it will be interesting to see whether the model works for other brain regions*

> *or other neuronal type connectivity as well.*

We value your interest in the model's applicability to other brain regions and neuronal types. To address this, we have extended our analysis in the revised manuscript to include a study on gap junction connectivity between C. elegans neurons. This extension demonstrates the model's versatility and its potential applicability across various nervous systems and connectivity types.

> *Q2. Whether the authors can use the same transformation matrices trained from the retina data to predict neuronal connectivity in other brain regions? Or an easier case, the connectivity between RGC types to the neuronal types in SC, dLGN, or other post-RGC-synaptic brain regions. As the neuronal connection mechanisms are conserved and widely shared between different neuronal types, one would expect the same transformation matrices may work in predicting other neuronal type connectivity as well (at least to some extent).*

The idea to use the same transformation matrices for predicting connectivity in other brain regions is intriguing. While direct application of these matrices to different regions remains challenging, we discussed the potential scalability of our model to other brain areas. By applying the model to combined datasets from various regions, we could uncover conserved neuronal connection mechanisms. This approach is theoretically feasible and is supported by the demonstrated scalability of the bilinear model and its deep learning variants in industrial applications.

> *Q3. Section 5.2 Connectivity metric generation: in this work, the author uses the stratification profiles of the neurons to estimate the connectivity metric, how reliable this method is? There will be a scenario where though two neuronal types project to a similar inner plexiform layer, they may not have any connection. Have the authors considered combining other experimental data (like electrophysiology data or neuron tracing data)?*

We discussed the reliability of using stratification profiles for estimating connectivity metrics, acknowledging potential limitations. In the revised manuscript, we added discussion on how the integration of additional experimental data, such as electrophysiological and neuron tracing data, could enhance the accuracy of the connectivity metrics.

> *Q4. Section 6 Model training and validation: does the author have a potential hypothesis as to why 2 dimensions are the best latent feature spaces dimensionality? One would imagine with more dimensionality, the model will give better results. Could it be that the connectivity data that is used to train the model is only considering the two-dimensional space of the neuronal stratification?*

The selection of two dimensions for the latent feature space was informed by 5-fold cross-validation, aimed at optimizing model generalization to unseen data. Here while increasing dimensionality improves performance on the training set, it does not necessarily enhance generalization to the validation set. Thus, the choice of two dimensions ensures good performance without overfitting to the training data.

> *Q5. Could the author provide the source code for the analysis? Or could the author make it a python/R package so that non-computational biologists can easily apply the method to their own data?*

We have included a "Data and Code Availability" section in the revised manuscript. This section provides a link to the source code with pointers to datasets used in our study, facilitating the application of our methods by researchers from various backgrounds.

> *Q6. I know it may be difficult for the author to do, but is it possible to design and perform some experiments to validate the model prediction results, either connectivity partners of transcriptomicallydefined RGC types or the function of the key genetic molecules (which hasn't been discovered before)? The author may consider collaborating with some experimental labs. The author may even consider predicting the connectivity between RGC with some of its post-synaptic neurons in the brain regions, like SC or dLGN, as recently there are a lot of single-cell sequencing data as well as connectivity data.*

We appreciate your suggestion regarding experimental validation. As a future direction, we have discussed potential experimental approaches to validate the model's predictions in the "Experimental Validation of Candidate Genes" section. Specifically, we propose an experimental design involving the manipulation of delta-protocadherins using AAV-mediated CRISPR/Cas9 and subsequent examination of connectivity phenotypes. We are also open to collaborating with experimental labs to further explore the model's predictions, particularly in predicting connectivity between RGCs and their post-synaptic neurons in other brain regions.