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

Mu Qiao
LinkedIn, Mountain View, CA, 94043
muqiao0626@gmail.com

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 pre- and post-synaptic 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 mouse bipolar cells (presynaptic) and retinal ganglion cells (postsynaptic), obtained from single-cell transcriptomics, into a covariance matrix. The objective is to construct this covariance matrix that closely mirrors a connectivity matrix, derived from connectomic data, reflecting the known anatomical connections between these neuronal types. Our model successfully recaptiulates recognized connectivity motifs and provides interpretable insights into genetic interactions that shape the connectivity. Specifically, it identifies 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.

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 precise neuronal connectivity patterns evident from connectomic data 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

Preprint.

that a similar model could capture the complex relationships between genetic patterns of presynaptic and postsynaptic neurons and their connectivity.

Applying this model to single-cell transcriptomic and connectomic data from mouse retinal neurons, we demonstrate that it can effectively learn connectivity patterns between bipolar cells (BCs, presynaptic) and retinal ganglion cells (RGCs, postsynaptic). 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: Synaptic Specificity between Neuronal Types

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]. 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. A classic example of this is seen in the retina, where different types of BCs form specific synaptic connections with various types of RGCs [7, 10, 11]. These connections create parallel pathways that transform visual signals from photoreceptors to RGCs, which subsequently transmit the information to the brain [12, 13].

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 [14, 3, 15].

Emerging tools and technologies offer unprecedented opportunities to unravel these mysteries. Among these, the transcriptome and connectome are particularly promising [3, 16]. The 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. The connectome, on the other hand, provides a detailed map of the connections between neurons. By combining information from the transcriptome and connectome, it is possible to link specific genes to specific connections, thereby shedding light on the genetic basis of neuronal connectivity.

3 Related Work: Collaborative Filtering

Our strategy draws inspiration from the concept of collaborative filtering using bilinear models, a technique fundamental to recommendation systems [17, 18]. 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, 19]. 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 [20, 21].

4 Bilinear Model for Neuronal Type Connectivity

4.1 Objective Functions

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.

4.1.1 Gene Expression and Connectivity of Each Cell are Known Simultaneously

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 kth cell in the presynaptic type i is designated as $\boldsymbol{x}_{(ik)}$, where $k \in 1, 2, ..., n_i$, while for the lth 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 $w_{(ik)(jl)} = \frac{1}{n_i n_j}$. 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.

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

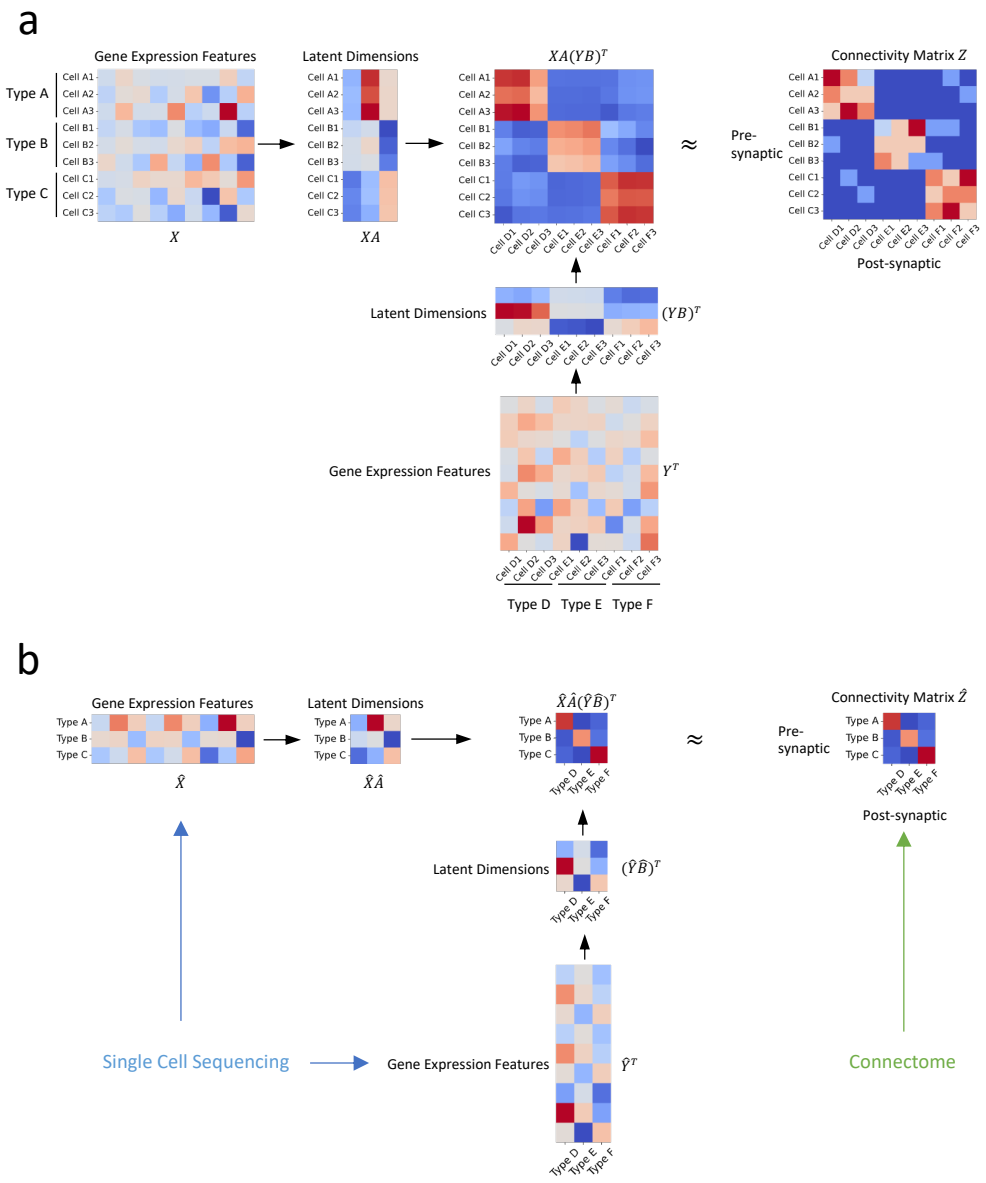

Figure 1: Illustration of our approach. (a) In an ideal scenario where gene expression profiles and connectivity data of individual cells are available simultaneously, we establish the relationship between connectivity and gene expression profiles via two transformation matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ (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 $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$.

are highly correlated. To overcome this, we can perform principal component analysis (PCA) on the gene expression vectors, transforming them into a new coordinate system. By excluding components whose negligible eigenvalues, we effectively get rid of redundant information, thus mitigating the effects of multicollinearity. This transformed gene expression data can then be used in our bilinear model, leading to more stable and reliable estimates of connectivity. In mathematical terms, applying PCA to the zero-centered and unit-variance adjusted matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ results in the approximations:

$$\boldsymbol{X} \approx \boldsymbol{X}\boldsymbol{U}\boldsymbol{U}^T \tag{6}$$

and

$$\boldsymbol{Y} \approx \boldsymbol{Y}\boldsymbol{V}\boldsymbol{V}^T \tag{7}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ indicate the PCA transformation of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively. The original optimization problem now becomes:

$$\min_{\boldsymbol{A}^{'},\boldsymbol{B}^{'}} ||\boldsymbol{W} \odot (\boldsymbol{Z} - (\boldsymbol{X}^{'}\boldsymbol{A}^{'})(\boldsymbol{Y}^{'}\boldsymbol{B}^{'})^T)||_F^2 \tag{8}$$

where $\boldsymbol{X}^{'} = \boldsymbol{X}\boldsymbol{U}$, $\boldsymbol{Y}^{'} = \boldsymbol{Y}\boldsymbol{V}$, $\boldsymbol{A}^{'} = \boldsymbol{U}^T\boldsymbol{A}$, $\boldsymbol{B}^{'} = \boldsymbol{V}^T\boldsymbol{B}$. It's worth noting that if the optimization problem is solved in the PCA-transformed space, we need to transform $\boldsymbol{A}^{'}$ and $\boldsymbol{B}^{'}$ back $\boldsymbol{A}$ and $\boldsymbol{B}$ in order to retrieve the corresponding weight of each gene.

4.1.2 Connectivity and Gene Expressions of Neuronal Types are from Different Sources

In real scenarios, gene expression profiles and connectivity information are often derived from separate sources, such as single-cell sequencing [22, 23] and connectome data [7, 24, 25]. 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 [26]).

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:

$$\min_{\boldsymbol{A},\boldsymbol{B}} \sum_{i=1}^{a}\sum_{j=1}^{b}(z_{(i.)(j.)} - (\boldsymbol{x}_{(i.)}\boldsymbol{A})(\boldsymbol{y}_{(j.)}\boldsymbol{B})^T)^2 \tag{9}$$

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 $\boldsymbol{A}$:

$$||\boldsymbol{A}^T\boldsymbol{\Sigma}_x\boldsymbol{A}||_F^2 \leq \epsilon \tag{10}$$

where

$$\boldsymbol{\Sigma}_x = \sum_{i=1}^{a}(\frac{1}{n_i}\sum_{k=1}^{n_i}(\boldsymbol{x}_{(ik)} - \boldsymbol{x}_{(i.)})^T(\boldsymbol{x}_{(ik)} - \boldsymbol{x}_{(i.)})) \tag{11}$$

and $\boldsymbol{B}$:

$$||\boldsymbol{B}^T\boldsymbol{\Sigma}_y\boldsymbol{B}||_F^2 \leq \epsilon \tag{12}$$

where

$$\boldsymbol{\Sigma}_y = \sum_{j=1}^{b}(\frac{1}{n_j}\sum_{l=1}^{n_j}(\boldsymbol{y}_{(jl)} - \boldsymbol{y}_{(j.)})^T(\boldsymbol{y}_{(jl)} - \boldsymbol{y}_{(j.)})) \tag{13}$$

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:

$$\min_{\boldsymbol{A},\boldsymbol{B}} ||\bar{\boldsymbol{Z}} - \bar{\boldsymbol{X}}\boldsymbol{A}(\bar{\boldsymbol{Y}}\boldsymbol{B})^T||_F^2, \quad s.t. ||\boldsymbol{A}^T\boldsymbol{\Sigma}_x\boldsymbol{A}||_F^2 \leq \epsilon, ||\boldsymbol{B}^T\boldsymbol{\Sigma}_y\boldsymbol{B}||_F^2 \leq \epsilon \tag{14}$$

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 $diag(\hat{\sigma}_{x_1}^2, \hat{\sigma}_{x_2}^2, ..., \hat{\sigma}_{x_p}^2)$ and $diag(\hat{\sigma}_{y_1}^2, \hat{\sigma}_{y_2}^2, ..., \hat{\sigma}_{y_q}^2)$ [27, 28]. We then transform the initial optimization problem into the following:

$$\min_{\hat{\boldsymbol{A}},\hat{\boldsymbol{B}}} ||\bar{\boldsymbol{Z}} - \hat{\boldsymbol{X}}\hat{\boldsymbol{A}}(\hat{\boldsymbol{Y}}\hat{\boldsymbol{B}})^T||_F^2, \quad s.t. ||\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}}||_F^2 \leq \epsilon, ||\hat{\boldsymbol{B}}^T\hat{\boldsymbol{B}}||_F^2 \leq \epsilon \tag{15}$$

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 $\hat{\boldsymbol{Y}} \in \mathbf{R}^{b\times q}$ 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 $\hat{\boldsymbol{A}}$ 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.

4.2 Optimization Algorithm

To solve the optimization problem as outlined in equation 15, we construct the following loss function:

$$L(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}) = ||\bar{\boldsymbol{Z}} - \hat{\boldsymbol{X}}\hat{\boldsymbol{A}}(\hat{\boldsymbol{Y}}\hat{\boldsymbol{B}})^T||_F^2 + \lambda_A||\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}}||_F^2 + \lambda_B||\hat{\boldsymbol{B}}^T\hat{\boldsymbol{B}}||_F^2 \tag{16}$$

This function can be regarded as the Lagrangian of the initial problem. Here, λ_A and λ_B are treated as hyperparameters, and their optimal values are determined through a grid search.

Given this loss function, we propose an alternative gradient descent algorithm to find the solutions. This algorithm alternates between updating the transformation matrices $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$, using the gradient descent optimization method.

The algorithm begins by initializing transformation matrices $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ using random values drawn from a standard normal distribution. The central aspect of the algorithm is an iterative loop that alternates the updates of transformation matrices $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$.

Algorithm 1 Alternative Gradient Descent (AGD) for Bilinear Mapping

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 A, B
end procedure

During each iteration, the algorithm computes the predicted connectivity metric $\hat{\boldsymbol{Z}}$ using the current estimates of $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{X}}$, and $\hat{\boldsymbol{Y}}$. 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 $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ 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 while adhering to the constraints unique to our problem design. As a result, it produces associations between gene expression profiles of cell types and their connectivity.

5 Datasets and Pre-processing

Our study hinges on two distinct data collections of mouse retina neurons: single-cell transcriptomic data from previous studies and connectomic data from the EyeWire project. Together, these datasets provide us with connectivity information and gene expression profiles, which constitute the key ingredients for our proposed bilinear model.

5.1 Single-cell Transcriptomic Data

The single-cell transcriptomic data used in our study include the gene expression profiles for two classes of mouse retina neurons - presynaptic BCs as reported by Shekhar et al. [22], and postsynaptic RGCs as reported by Tran et al. [23].

Preprocessing of this data adhered to previously documented procedures [22, 23, 29]. 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 [30, 31, 32]. We focused on those cells whose types correspond with the neuronal types outlined in the connectomic data, as delineated later in Table 1, Table 2, and Table 3. 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 $\times$ 17144 and 3779 $\times$ 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 $\times$ 11323 and 3779 $\times$ 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:

$$\hat{\sigma}_{x_{im}} = \sum_{i=1}^{a}(\frac{1}{n_i}\sum_{k=1}^{n_i}(x_{(ik)m} - x_{(i.)m})^2) \tag{17}$$

$$\hat{\sigma}_{y_{jm}} = \sum_{j=1}^{b}(\frac{1}{n_j}\sum_{l=1}^{n_j}(y_{(jl)m} - y_{(j.)m})^2) \tag{18}$$

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 $\hat{\boldsymbol{Y}}$, 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 $\hat{\boldsymbol{Y}}$ 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.

5.2 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, 33]. 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 [34]. This extensive dataset facilitated the derivation of a comprehensive connectivity matrix between two classes of mouse retina neurons - BCs [25] and RGCs [24]. 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 [34, 25, 24].

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 kth cell in BC type i and the lth cell in RGC type j, respectively. The connectivity metric $z_{(ik)(jl)}$ between these two neurons can be expressed as:

$$z_{(ik)(jl)} = \frac{\boldsymbol{u}_{ik}\boldsymbol{v}_{jl}}{|\boldsymbol{v}_{ik}||\boldsymbol{v}_{jl}|} \tag{19}$$

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.

5.3 Correspondence of Cell Types between Datasets

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. [22] and Greene et al. [25]. This correspondence is presented in Table 1.

Table 1: Correspondence of Mouse BC types [25, 22]

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

Regarding RGC types, alignment between cell types annotated in Tran et al. [23] and Bae et al. [24] was established primarily based on the findings from Goetz et al. [26]. 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 2):

Table 2: Correspondence of Mouse RGC types [24, 23, 26]

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

with the following additions derived from Supplementary Table S1 of Tran et al., based on the expressions of genetic markers of the following RGC types (Table 3):

Table 3: Correspondence of Mouse RGC types [24, 23, 26]

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

We carried out subsequent analyses based on the alignment of these neuronal types.

6 Model Training and Validation

With the bilinear model outlined in Section 4, we iteratively optimized transformation matrics, denoted as $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$, using the ADG optimization algorithm. The goal was to minimize the prescribed loss function. The initial states of these transformation matrices were randomly generated from a standard normal distribution, and updates were continued until either the change in loss fell below a predetermined threshold of 10^{-6} or a maximum iteration count of 10^6 was reached.

We examined two sets of hyperparameters during the optimization process: the regularization parameters, λ_A and λ_B, and the dimensionality of the latent feature space. Preliminary tests suggested that a lower loss was achieved when λ_A and λ_B were set to equivalent values. Consequently, both were unified under a single parameter, λ.

We utilized 5-fold cross-validation to identify the optimal hyperparameters. This procedure partitioned the connectivity matrix entries into five unique subsets or "folds". The model was then trained using a combination of four folds, with the remaining one serving as a validation set. This procedure was repeated five times, with each iteration reserving a different fold for validation. Throughout this cross-validation process, we varied λ across the range [0.1, 1, 10, 100] and the latent feature space's dimensionality across the range [1, 2, 3, 4, 8].

Figure 2a presents a heatmap of the logarithm (base 10) of the validation loss, highlighting variations of λ and dimensionality. It can be observed that the lowest validation loss is achieved when λ equals 1 and the dimensionality equals 2, as shown in Figure 2b,c. These specific values were thus chosen as the optimal hyperparameters. With these optimal hyperparameters, we performed the final round of training on the entire dataset to yield the definitive transformation matrices, $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$, using a learning rate of 5×10^{-7}.

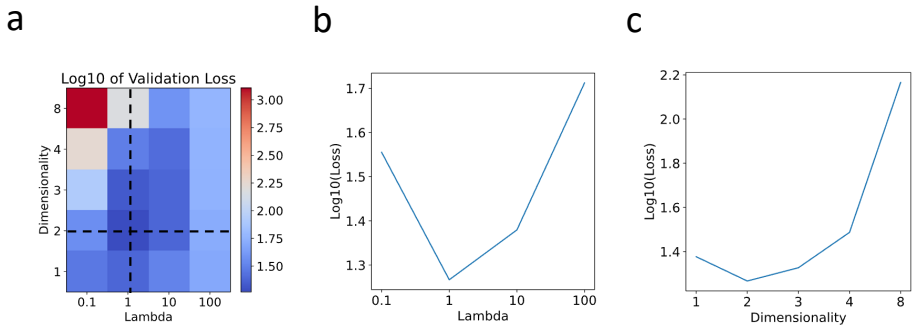

Figure 2: Hyperparameter selection through cross-validation. (a) Heatmap plot of the logarithm (base 10) of the validation loss, showing variations with respect to λ across [0.1, 1, 10, 100] and dimensionality across [1, 2, 3, 4, 8]. (b) Plot showing the logarithm (base 10) of the validation loss against 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].

7 Results

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

Upon completion of the final training process, our optimized bilinear model produced transformation matrices, $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$. We used these matrices to project the normalized single-cell transcriptomic data,

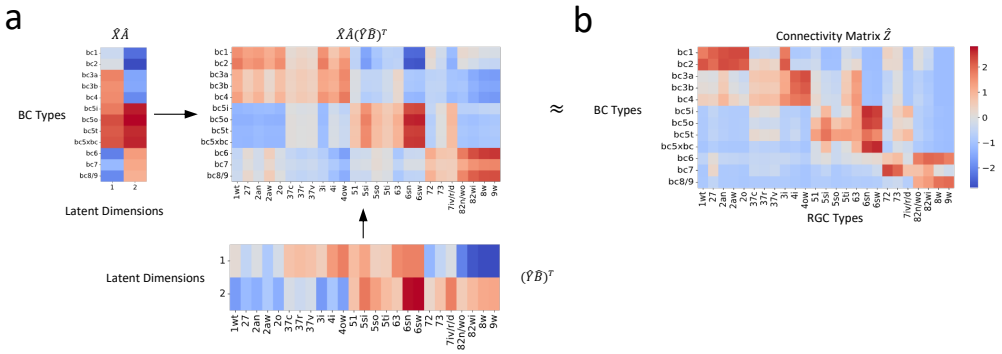

Figure 3: Reconstruction of connectivity map from gene expression profiles. (a) The reconstructed connectivity matrix, derived from the shared latent feature space projections. (b) The connectivity matrix obtained from connectomic data. Differences in color intensity represent the strength of connections, with dark red indicating strong connections and dark blue indicating weak or no connections.

$\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{Y}}$, into a shared latent feature space. Consequently, we obtained projected representations for BC and RGC types, $\hat{\boldsymbol{X}}\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{Y}}\hat{\boldsymbol{B}}$, 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 3a).

To evaluate our model, we compared the reconstructed connectivity matrix with the one derived from connectomic data (Figure 3b). 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.

7.2 Bilinear Model Recapitulates Recognized Connectivity Motifs

Our cross-validation procedure indicated that the optimal number of latent dimensions was two. 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 4a,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 4b,e,i).

To confirm these observations, we further visualized BC and RGC types within the two-dimensional latent feature space (Figure 4c,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 4d,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 4e,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 [35, 10, 11, 36]. This correlation implies

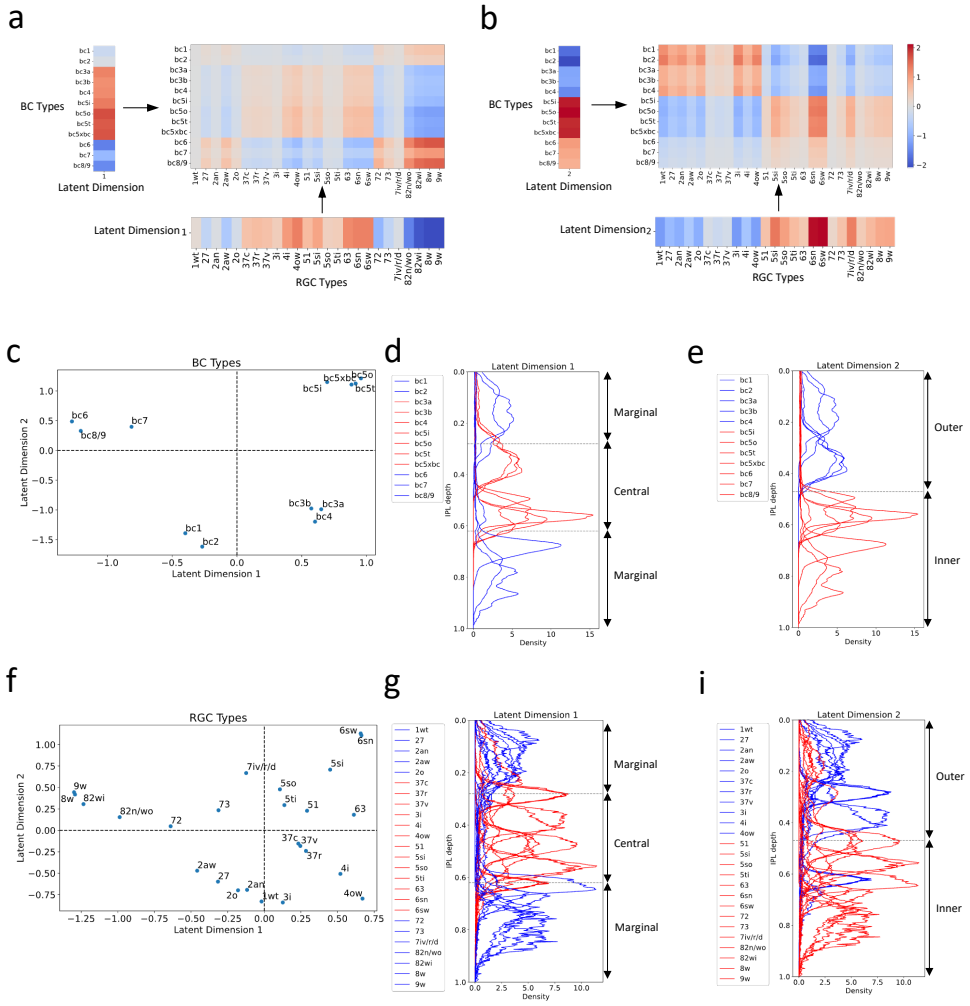

Figure 4: Distinct connectivity motifs revealed by the two latent dimensions. (a, b) The reconstructed connectivity using only latent dimension 1 or 2, respectively. Differences in color intensity represent the strength of connections. (c) BC types plotted in the latent feature space, with each point 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 [24]. 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 [24]. Synapses between BCs and RGCs in the outer retina mediate OFF responses, while those in the inner retina mediate ON responses.

that our bilinear model has successfully captured key aspects of retinal circuitry from gene expression data.

7.3 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 5a-d).

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 5a). 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 5b).

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 5c). 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 5d). 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 [37, 38, 39, 40, 20, 21, 41]. 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 [42, 43]. This tool allowed us to delve into the molecular functions, cellular pathways, and biological processes associated with these genes. Intriguingly, when we plotted the top 10 significant GO terms for BCs and RGCs in latent dimension 1 or 2 (Figure 5e-h), we found two common themes associated with the top genes: neuronal development and synaptic organization. This observation underlines the potential role of the top genes in forming and shaping the specific connections between BC and RGC types.

7.4 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 [23], which lack clear correspondence with those identified through connectomics studies [24]. 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 6a). We then employed this projection to construct a connectivity matrix between these RGC types and BC types (Figure 6b), 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 4.

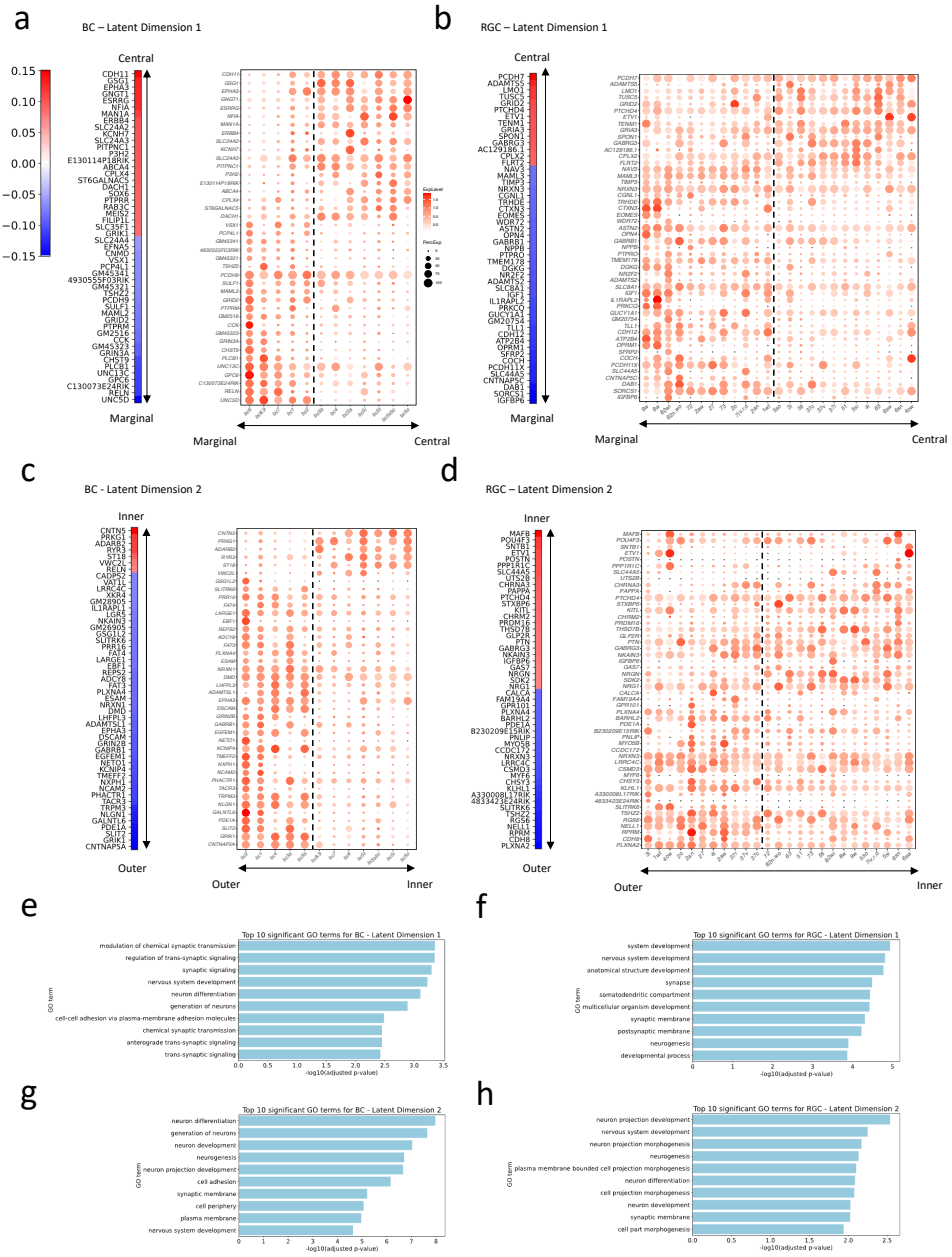

Figure 5: Gene signatures associated with the two latent dimensions. (a, b) Weight vectors of the top 50 genes for latent dimension 1, along with their expression patterns in BC types (a) and RGC types (b). The weight value is indicated in the color bar, with the sign represented by color (red: positive 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 4c,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. (e, f) The top 10 significant Gene Ontology (GO) terms extracted from the top 50 genes for BC types (e) and RGC types (f) in latent dimension 1. (g, h) The top 10 significant Gene Ontology (GO) terms extracted from the top 50 genes for BC types (g) and RGC types (h) in latent dimension 2. The significance of each GO term is indicated by its adjusted p-value.

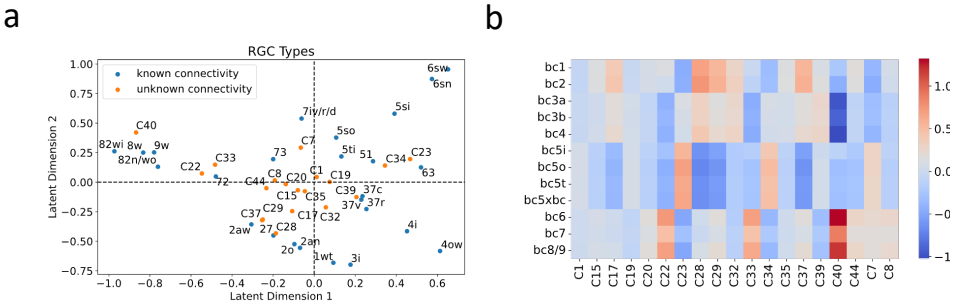

Figure 6: BC partner prediction of transcriptionally-defined RGC types. (a) Projection of transcriptionally-defined RGC types with unknown connectivity into the same latent space as those with known connectivity. (b) The resulting predicted connectivity matrix between these RGC types and BC types. Transcriptionally-defined RGC types are named according to Tran et al. [23]

Table 4: Predicted BC Partners of Transciptionally-defined RGC Types

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

Although the ground truth connectivity of these RGC types remains unknown due to the absence of matching types in connectomic data, Goetz et al. [26], 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 4, referencing Figure 5A from Goetz et al. [26], and highlighted whether our 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.

8 Discussion

This study showcased an innovative computational strategy that integrates transcriptomic and connectomic data to elucidate genetic principles dictating neural circuit wiring. Leveraging a bilinear model, we reconstructed a neuronal type-specific connectivity map from gene expression profiles. We have demonstrated the model's capability to recapitulate recognized connectivity motifs of the retinal circuit and provide interpretable insights into the gene signatures associated with these motifs. Moreover, our model can predict potential connectivity partners for transcriptomically-defined neuronal types whose connectomes are yet to be fully characterized. These make our bilinear model a valuable tool for investigating gene-regulatory mechanisms involved in neural circuit wiring.

The inspiration for our bilinear model stemmed from recommendation systems – a machine learning domain focused on capturing intricate interactions between users and items and predicting user preferences in commercial settings. This analogy served as a useful framework in our study, where the roles of users and items in the recommendation systems were 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.

Our bililinear model successfully 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. These motifs align well with recognized properties of retinal neurons: kinetic attributes (transient versus sustained responses) and polarity (ON versus OFF responses). 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. 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) [37, 38]. 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) [39]. 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 [40]. Similarly, Cadherins (CDH8,11,12), whose combinations have been implicated in synaptic specificity within retinal circuits [20, 21], 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 [41].

In additional 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 [44, 45, 46]. Future experimental studies are needed to validate these findings and further unravel the roles of these genes in neural circuit formation and function.

While our approach was illustrated using the mouse retina, it holds vast potential to decipher the genetic programming of neuronal connectivity in other areas of the nervous system, given sufficient data on both gene expression profiles and synaptic connections. For instance, single-cell sequencing across multiple cortex regions and the hippocampus in mice has defined transcriptomic cell types [47, 48, 49]. Concurrently, considerable progress has been made in connectomic studies of mouse cortices, particularly the visual cortex [50, 51, 33, 52]. Integrating these data sources enables a

comprehensive examination of the genetic underpinnings of connectivities in these brain regions. Our approach serves as a pioneering step in investigating how gene expression patterns contribute to the diversity of neuronal circuits across different brain regions, thereby setting the stage for a holistic understanding of the genetic blueprint governing neuronal circuit wiring throughout the entire brain.

9 Future Directions

This study introduces a novel and powerful methodology for inferring connectivity maps from gene expression data, bridging the gap between gene expression and synaptic connectivity in neural circuits. It offers a data-driven, systematic approach to decipher the genetic regulation of neural circuit formation. Nevertheless, the study is not without its limitations, which, in turn, highlight potential avenues for future research.

Firstly, the inherent linearity of the model facilitates the identification of key genes influencing neuronal connectivity but reduces the intricate, multi-stage process of synapse formation to a linear transformation. This simplification, while useful for gaining insights, does not fully capture the complexity of synapse formation. However, the bilinear model provides a solid theoretical foundation for developing more sophisticated, non-linear models. For instance, we could extend the bilinear model by incorporating kernels, possibly in the form of neural networks, to capture non-linear interactions between gene expressions and connectivity patterns (Figure 7). A manifestation of this non-linear approach is the "two-tower model", in which each "tower" represents a deep neural network that learns a non-linear transformation of the input features [53, 54]. This model, widely used in contemporary recommendation systems, has demonstrated potency in capturing complicated user-item interactions.

Secondly, our model operates on the assumption that the transcriptomic profile of a neuron type largely determines its connectivity profile. However, other factors, such as neuronal activity, also influence synaptic connectivity, and our model does not account for these elements. Future models could incorporate activity-dependent variables and additional omics data, to capture a broader range of factors shaping neural circuits.

References

[1] Sebastian Seung. *Connectome: How the Brain's Wiring Makes Us Who We Are*. Houghton Mifflin Harcourt, 2011.

[2] Franck Polleux and William Snider. Establishment of axon-dendrite polarity in developing neurons. *Annual review of neuroscience*, 36:467–488, 2013.

[3] Joshua R. Sanes and S. Lawrence Zipursky. Synaptic Specificity, Recognition Molecules, and Assembly of Neural Circuits. *Cell*, 181(3):536–556, 2020.

[4] Hongkui Zeng and Joshua R Sanes. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nature Reviews Neuroscience*, 18(9):530–546, 2017.

[5] Antonio Scialdone, Valentine Svensson, Anja Wilbrey-Clark, Valentina Proserpio, and Sarah A Teichmann. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 19(3):133–145, 2018.

[6] Winfried Denk and Heinz Horstmann. Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure. *PLOS biology*, 2(11), 2004.

[7] Moritz Helmstaedter, Kevin L. Briggman, Srinivas C. Turaga, Viren Jain, H. Sebastian Seung, and Winfried Denk. Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*, 500(7461):168–174, 2013.

[8] Juan Carlos Tapia, Narayanan Kasthuri, Kenneth J Hayworth, Richard Schalek, Jeff W Lichtman, Stephen J Smith, and JoAnn Buchanan. High-contrast en bloc staining of neuronal tissue for field emission scanning electron microscopy. *Nature Protocols*, 7(2):193–206, 2012.

[9] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. In *Computer*, volume 42, pages 30–37, 2009.

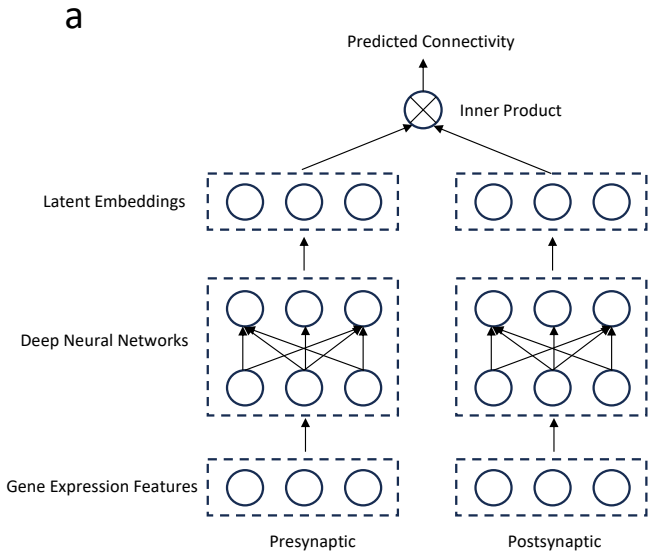

Figure 7: Future direction: A two-tower model. (a) Gene expression profiles of pre- and post-synaptic neurons are transformed into latent embedding representations via deep neural networks. The connectivity metric between the pre- and post-synaptic neurons is predicted by taking the inner product of their respective latent embeddings.

[10] Thomas Euler, Silke Haverkamp, Timm Schubert, and Tom Baden. Retinal bipolar cells: elementary building blocks of vision. *Nat Rev Neurosci*, 15(8):507–519, 2014.

[11] Joshua R Sanes and Richard H Masland. The types of retinal ganglion cells: Current status and implications for neuronal classification. *Annual Review of Neuroscience*, 38:221–246, 2015.

[12] Tim Gollisch and Markus Meister. Eye smarter than scientists believed: Neural computations in circuits of the retina. *Neuron*, 65(2):150–164, 2010.

[13] Rava Azeredo da Silveira and Botond Roska. Cell types, circuits, computation. *Current Opinion in Neurobiology*, 21(5):664–671, 2011.

[14] Joris de Wit and Anirvan Ghosh. Specification of synaptic connectivity by cell surface interactions. *Nature Reviews Neuroscience*, 17:4, 2016.

[15] Adam J Isabella, Eduardo Leyva-Díaz, Takuya Kaneko, Scott J Gratz, Cecilia B Moens, Oliver Hobert, Kate O'Connor-Giles, Rajan Thakur, and HaoSheng Sun. The field of neurogenetics: where it stands and where it is going. *Genetics*, 218(4):iyab085, 2021.

[16] Alex Fornito, Aurina Arnatkevičiūtė, and Ben D. Fulcher. Bridging the Gap between Connectome and Transcriptome. *Trends in Cognitive Sciences*, 23(1):34–50, 2019.

[17] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B Kantor. Introduction to recommender systems handbook. *Recommender systems handbook*, 1:1–35, 2011.

[18] Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009, 2009.

[19] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 452–461, 2009.

[20] Xin Duan, Arjun Krishnaswamy, Irina De la Huerta, and Joshua R. Sanes. Type ii cadherins guide assembly of a direction-selective retinal circuit. *Cell*, 158(4):793–807, 2014. Free article.

[21] Xin Duan, Arjun Krishnaswamy, Irina De la Huerta, and Joshua R Sanes. Cadherin combinations recruit dendrites of distinct retinal neurons to a shared interneuronal scaffold. *Neuron*, 99(6):1145–1154.e6, 2018.

[22] Karthik Shekhar, Sylvain W. Lapan, Irene E. Whitney, Nicholas M. Tran, Evan Z. Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z. Levin, James Nemesh, Melissa Goldman, Steven A. McCarroll, Constance L. Cepko, Aviv Regev, and Joshua R. Sanes. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, 166(5):1308–1323.e30, August 2016.

[23] Nicholas M. Tran, Karthik Shekhar, Irene E. Whitney, Anne Jacobi, Inbal Benhar, Guosong Hong, Wenjun Yan, Xian Adiconis, McKinzie E. Arnold, Jung Min Lee, Joshua Z. Levin, Dingchang Lin, Chen Wang, Charles M. Lieber, Aviv Regev, Zhigang He, and Joshua R. Sanes. Single-cell profiles of retinal neurons differing in resilience to injury reveal neuroprotective genes. *bioRxiv*, page 711762, July 2019.

[24] J. Alexander Bae, Shang Mu, Jinseop S. Kim, Nicholas L. Turner, Ignacio Tartavull, Nico Kemnitz, Chris S. Jordan, Alex D. Norton, William M. Silversmith, Rachel Prentki, Marissa Sorek, Celia David, Devon L. Jones, Doug Bland, Amy L. R. Sterling, Jungman Park, Kevin L. Briggman, H. Sebastian Seung, and the EyeWirers. Digital museum of retinal ganglion cells with dense anatomy and physiology. *Cell*, 173(5):1293–1306.e19, 2018.

[25] M. J. Greene, J. S. Kim, H. S. Seung, and the EyeWirers. Analogous convergence of sustained and transient inputs in parallel on and off pathways for retinal motion computation. *Cell Reports*, 14:1–9, 2016.

[26] Jillian Goetz, Zachary F Jessen, Anne Jacobi, Adam Mani, Sam Cooler, Devon Greer, Sabah Kadri, Jeremy Segal, Karthik Shekhar, Joshua R Sanes, and Gregory W Schwartz. Unified classification of mouse retinal ganglion cells using function, morphology, and gene expression. *Cell Reports*, 40(2):111040, 2022.

[27] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, May 2018.

[28] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902.e21, June 2019.

[29] Mu Qiao. Factorized linear discriminant analysis and its application in computational biology. *arXiv preprint arXiv:2010.02171*, 2020.

[30] Hung-I. Harry Chen, Yufang Jin, Yufei Huang, and Yidong Chen. Detection of high variability in gene expression from single-cell RNA-seq profiling. *BMC Genomics*, 17 Suppl 7:508, August 2016.

[31] Shristi Pandey, Karthik Shekhar, Aviv Regev, and Alexander F. Schier. Comprehensive Identification and Spatial Mapping of Habenular Neuronal Types Using Single-Cell RNA-Seq. *Curr. Biol.*, 28(7):1052–1065.e7, April 2018.

[32] Yerbol Z Kurmangaliyev, Juyoun Yoo, Samuel A LoCascio, and S Lawrence Zipursky. Modular transcriptional programs separately define axon and dendrite connectivity. *eLife*, 8:e50822, November 2019.

[33] Nicholas L Turner, Thomas Macrina, J Alexander Bae, Runzhe Yang, Alyssa M Wilson, Casey Schneider-Mizell, Kisuk Lee, Ran Lu, Jingpeng Wu, Agnes L Bodor, Adam A Bleckert, Derrick Brittain, Emmanouil Froudarakis, Sven Dorkenwald, Forrest Collman, Nico Kemnitz, Dodam Ih, William M Silversmith, Jonathan Zung, Aleksandar Zlateski, Ignacio Tartavull, Szi-Chieh Yu, Sergiy Popovych, Shang Mu, William Wong, Chris S Jordan, Manuel Castro, JoAnn Buchanan, Daniel J Bumbarger, Marc Takeno, Russel Torres, Gayathri Mahalingam,

Leila Elabbady, Yang Li, Erick Cobos, Pengcheng Zhou, Shelby Suckow, Lynne Becker, Liam Paninski, Franck Polleux, Jacob Reimer, Andreas S Tolias, R Clay Reid, Nuno Maçarico da Costa, and H Sebastian Seung. Reconstruction of neocortex: Organelles, compartments, cells, circuits, and activity. *Cell*, 185(6):1082–1100.e24, 2022.

[34] Jinseop S Kim, Matthew J Greene, Aleksandar Zlateski, Kisuk Lee, Mark Richardson, Srinivas C Turaga, Michael Purcaro, Matthew Balkam, Amy Robinson, Bardia F Behabadi, et al. Space–time wiring specificity supports direction selectivity in the retina. *Nature*, 509(7500):331–336, 2014.

[35] Richard H. Masland. The neuronal organization of the retina. *Neuron*, 76(2):266–280, 2012.

[36] Tom Baden, Philipp Berens, Katrin Franke, Miroslav Román Rosón, Matthias Bethge, and Thomas Euler. The functional diversity of retinal ganglion cells in the mouse. *Nature*, 529(7586):345–350, 2016.

[37] Ryota L. Matsuoka, Kim T. Nguyen-Ba-Charvet, Aijaz Parray, Tudor C. Badea, Alain Chédotal, and Alex L. Kolodkin. Transmembrane semaphorin signaling controls laminar stratification in the mammalian retina. *Nature*, 470(7333):259–263, 2011.

[38] Liqun O. Sun, Zhuoyi Jiang, Merav Rivlin-Etzion, Rachel Hand, Claire M. Brady, Reina L. Matsuoka, King-Wai Yau, Marla B. Feller, and Alex L. Kolodkin. On and off retinal circuit assembly by divergent molecular mechanisms. *Science*, 342(6158):1241974, 2013.

[39] Yi-Rong Peng, Nicholas M Tran, Arjun Krishnaswamy, Dimitar Kostadinov, Emily M Martersteck, and Joshua R Sanes. Satb1 regulates contactin 5 to pattern dendrites of a mammalian retinal ganglion cell. *Neuron*, 95(4):869–883.e6, 2017.

[40] Arjun Krishnaswamy, Masahito Yamagata, Xin Duan, Y. Kate Hong, and Joshua R. Sanes. Sidekick 2 directs formation of a retinal circuit that detects differential motion. *Nature*, 524(7566):466–470, 2015.

[41] Jinyue Liu, Jasmine D. S. Reggiani, Mallory A. Laboulaye, Shristi Pandey, Bin Chen, John L. R. Rubenstein, Arjun Krishnaswamy, and Joshua R. Sanes. Tbr1 instructs laminar patterning of retinal ganglion cell dendrites. *Nature Neuroscience*, 21(5):659–670, 2018.

[42] Jüri Reimand, Meelis Kull, Hedi Peterson, Jaanus Hansen, and Jaak Vilo. g:Profiler — a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, 2007. [PDF].

[43] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 2019. [PDF].

[44] Irene Kahr, Karl Vandepoele, and Frans van Roy. Delta-protocadherins in health and disease. *Progress in Molecular Biology and Translational Science*, 116:169–192, 2013.

[45] Sarah E.W. Light and James D. Jontes. Delta-protocadherins: organizers of neural circuit assembly. *Seminars in Cell & Developmental Biology*, 69:83–90, 2017. PMCID: PMC5582989.

[46] Stacey Peek, Kar Men Mah, and Joshua A. Weiner. Regulation of neural circuit formation by protocadherins. *Cellular and Molecular Life Sciences*, 74(22):4133–4157, 2017. PMCID: PMC5643215.

[47] Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tomasz Jarsky, Zizhen Yao, Boaz Levi, Louie T Gray, Stacey A Sorensen, Tim Dolbeare, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience*, 19(2):335–346, 2016.

[48] Bosiljka Tasic, Zizhen Yao, Lucas T Graybuck, Kimberly A Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Ellen Garren, Michael N Economo, Sarada Viswanathan, et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, 2018.

[49] Zizhen Yao, Cindy T.J. van Velthoven, Thuc Nghi Nguyen, Jeff Goldy, Adriana E. Sedeno-Cortes, Fahimeh Baftizadeh, Darren Bertagnolli, Tamara Casper, Megan Chiang, Kirsten Crichton, et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, 184(12):3222–3241.e26, 2021.

[50] David D. Bock, Wei-Chung Allen Lee, Alexander M. Kerlin, Marga L. Andermann, Gregory Hood, Ann W. Wetzel, Stanislav Yurgenson, Elisha R. Soucy, Hyun Seok Kim, and R. Clay Reid. Network anatomy and in vivo physiology of visual cortical neurons. *Nature*, 471:177–182, 2011.

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

[52] Shenqin Yao, Quanxin Wang, Karla E. Hirokawa, Benjamin Ouellette, Ruweida Ahmed, Jasmin Bomben, Krissy Brouner, Linzy Casal, Shiella Caldejon, Andy Cho, Nadezhda I. Dotson, Tanya L. Daigle, Tom Egdorf, Rachel Enstrom, Amanda Gary, Emily Gelfand, Melissa Gorham, Fiona Griffin, Hong Gu, Nicole Hancock, Robert Howard, Leonard Kuan, Sophie Lambert, Eric Kenji Lee, and Hongkui Zeng. A whole-brain monosynaptic input connectome to neuron classes in mouse visual cortex. *Nature Neuroscience*, 26:350–364, 2023. Article.

[53] Tian Wang, Yuri M. Brovman, and Sriganesh Madhvanath. Personalized embedding-based e-commerce recommendations at ebay. *arXiv preprint arXiv:2102.06156*, 2021.

[54] Yantao Yu, Weipeng Wang, Zhoutian Feng, Daiyue Xue, Meituan, and Beijing. A dual augmented two-tower model for online large-scale recommendation. *KDD*, 2021.