Methods S1. Cell adhesion molecule network differential gene expression analysis. Related to Figures 7, S8, Table S5.

Network differential gene expression analysis

Decoding the transcriptional programming of neural connectivity is one of the central questions in molecular neuroscience. The introduction of single-cell sequencing technologies and the ability to accurately measure wiring diagrams have made this problem tractable. Here we generalize existing statistical tools designed to examine the transcriptional diversity of neurons and introduce a novel technique termed network differential gene expression analysis to decode the transcriptional patterns correlated with connectivity.

Standard differential gene expression (DGE) analysis (Soneson and Delorenzi, 2013; Wang et al., 2019) is useful for delineating univariate transcriptional differences between groups of neurons and characterizing their diversity. This is typically done by segregating neurons into two groups and testing the hypothesis that gene expression in both groups comes from the same distribution. This hypothesis is rejected if sufficient evidence suggests the contrary resulting in differentially expressed genes in either group. Mathematically, DGE can be expressed as a regression equation:

$$\mathbf{G = X\,\beta_{DGE}^{T} + \epsilon (Eq.\,1)}$$

where $G \in R^{n \times d}$ represents the gene expression matrix of n neurons, and d genes, $X \in \{-1,1\}^{n}$ represents the "design matrix" that segregates the neurons to their respective groups (-1 vs. 1), $\beta_{DGE} \in R^{d}$ represents the regression coefficients that reflect whether particular genes are enriched or depleted in respective groups, and ϵ captures the model error, typically a Gaussian or a Laplacian variable depending on the choice of the error distribution. Different DGE packages such as DESeq2, edgeR, MAST (Finak et al., 2015; Love et al., 2014; Robinson et al., 2009) impose different distributional assumptions about gene expression or the modeling error, which is reflected in their choice of ϵ. The schematic diagram of standard DGE analysis can be seen in Figure I A, B.

While standard DGE analysis is useful for delineating univariate differences between groups of neurons, it is not designed to differentiate bivariate genetic effects such as those observed in synapses between groups of pairs of neurons. To address this limitation, here we introduce a generalization of DGE, termed "network" DGE (nDGE), to test the differential genetic effects observed in synaptic couplings. Unlike DGE, where gene expression levels of disjoint groups of neurons are compared, in nDGE, the multiplicative co-expression of genes between sets of pairs of neurons (representing edges in a network) is compared. Conveniently, the mathematical

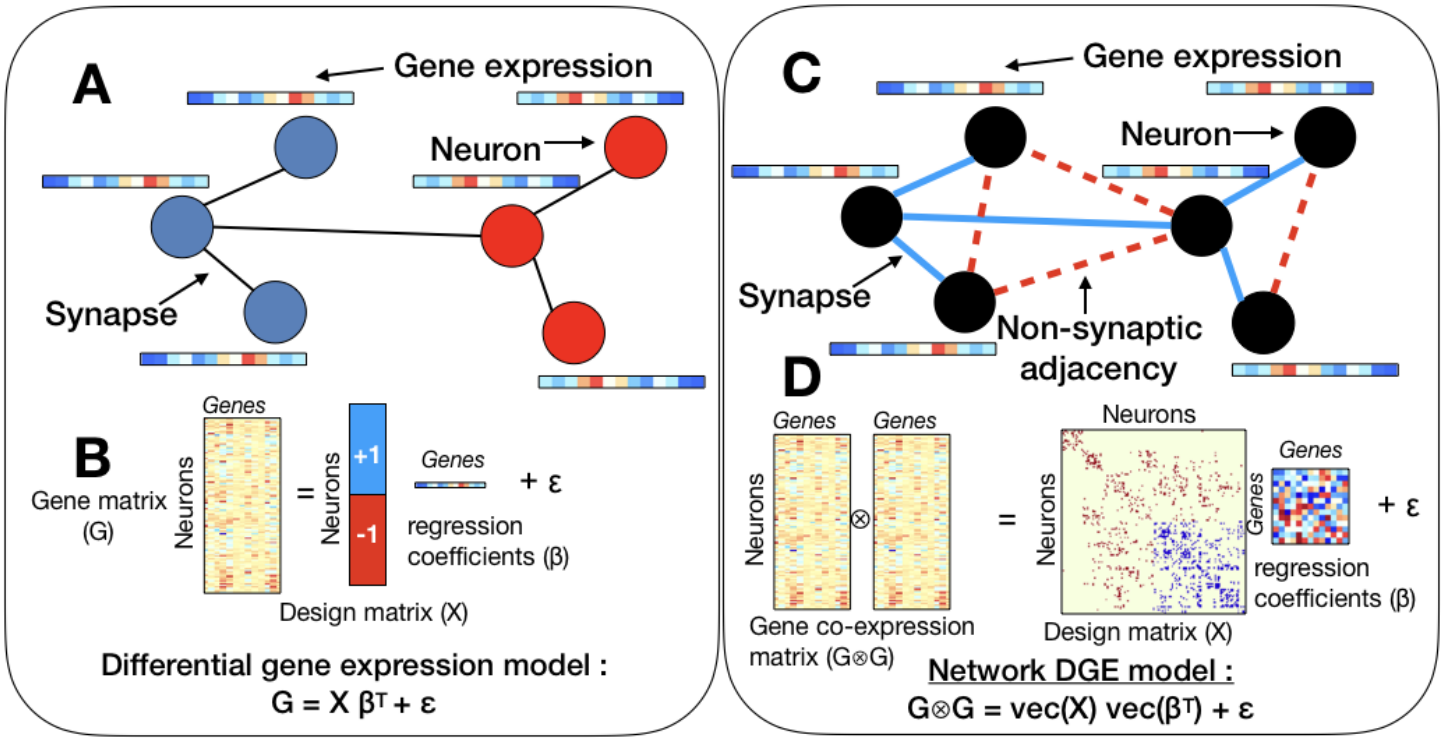

Figure I: Overview of standard differential gene expression analysis (DGE) vs. network differential gene expression analysis (nDGE)

formulation of nDGE can be derived as a generalization of DGE. Starting with the regression model of DGE in equation 1, we can represent gene expression matrix **G** with a gene co-expression matrix between all pairs of neurons. This can be compactly represented as a tensor product, $G \otimes G \in R^{nd \times nd}$ where $\otimes$ denotes the Kronecker product. Using this, the nDGE model aims to determine gene co-expression differences between pairs of connected neurons versus pairs of neurons that are disconnected. Mathematically, this can be cast as a regression of the gene co-expression patterns between all neurons as a function of differential groups of connectivity, e.g., connected pairs of neurons vs. non-connected neurons using the following model:

$$\mathbf{G} \otimes \mathbf{G} = \mathbf{X}\,\boldsymbol{\beta}_{\mathbf{NDGE}}^{\mathbf{T}} + \boldsymbol{\epsilon} \qquad (\mathbf{Eq.\,2})$$

Here $X \in \{-1,0,1\}^n$ represents the "design matrix" that segregates the neuron-pairs to their respective groups of whether they are connected (1), disconnected (-1) or excluded from analysis (0). In other words, if neuron i is pre-synaptically connected to neuron j, then $X_{ij} = 1$, if neuron i is disconnected from neuron j, $X_{ij} = -1$. If we want to exclude the synapse of neuron i to neuron j from the analysis, then we set $X_{ij} = 0$. The regression coefficients are represented by $\beta_{NDGE} \in R^{d \times d}$ and reflect whether particular gene-pair combinations are enriched or depleted in respective groups of connected or disconnected neurons. Thus, β_{pq} represents the effect of gene p in the pre-synaptic neural partner and gene q in the post-synaptic partner. If $\beta_{pq} > 0$ then the pth gene observed in the pre-synaptic neuron and qth gene observed in the post-synaptic neuron are enriched in connected neurons. If $\beta_{pq} < 0$, the opposite is true. Note that since the connectome is asymmetric, i.e. neuron i could be pre-synaptic to neuron j, but not vice versa, β_{pq} is not necessarily equal to β_{qp}, allowing us to gain two perspectives of the effect of gene co-expression on connectivity: one on the pre-synaptic side and the other on the post-synaptic side. Lastly, ϵ captures the model error. In this analysis, we assume Gaussian error given log-transformed gene counts in G. The visual representation of the nDGE statistical model can be seen in Figure I C-D.

Null distribution and p-value estimation

Like standard DGE, the nDGE test's output is a set of t-statistics and log-fold changes for gene associations with differential connectivity. However, unlike standard DGE, the t-statistics and log-fold changes in nDGE capture the effect of co-expression of pairs of genes, one corresponding to the gene observed in the pre-synaptic neuron partner and the other corresponding to the gene observed in the post-synaptic one. To deem a pair of genes significant under nDGE analysis, we also utilize the Bonferroni correction for p-values accounting for the number of pairs of genes in the multiple comparison setting.

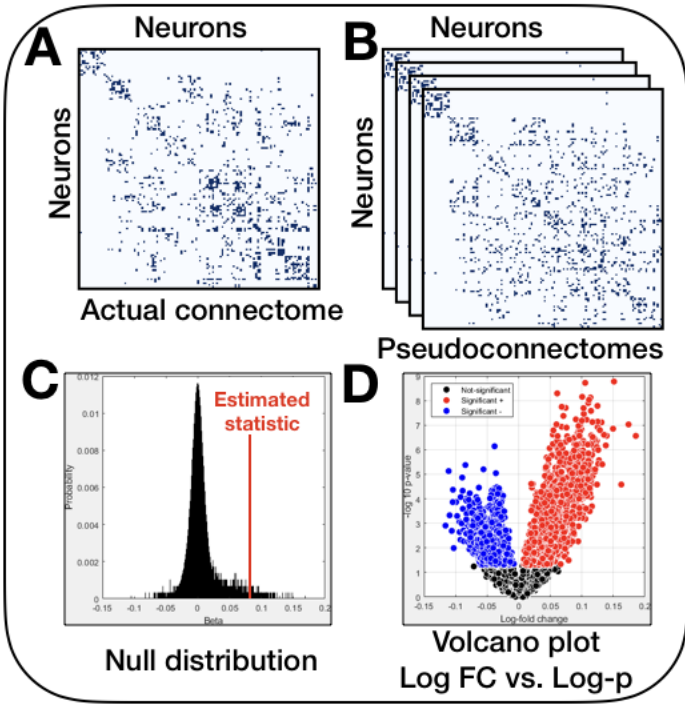

Figure II: Overview of null hypothesis generation using randomized topology preserving pseudoconnectomes

The main caveat in nDGE is the lack of independence of samples that are compared between groups. Since "samples" in nDGE are the co-expression of genes in pairs of neurons, the information from a particular neuron will inevitably be represented multiple times and possibly in different groups, e.g., the gene expression from neuron AIA is represented in multiple synaptic gene co-expression values for all synaptic partners of AIA as well as the non-synaptic adjacent partners of AIA (Figure 7B in the main text). This lack of independence in the test samples can falsely inflate/deflate the sample variance, which can introduce excess false positives and false negatives. To accurately estimate the null distribution of the nDGE test statistics, we generate randomized "pseudoconnectomes" that respect the topology of the original connectome. Specifically, the pseudoconnectomes preserve the same number of synaptic partners for each neuron and the shuffled synaptic partners are confined to be neurons that have membrane contact. We follow the degree

preserving graph randomization procedure (Milo et al., 2003). The latter constraint prevents infeasible pseudoconnectomes where synapses exist between neurons that do not share a membrane contact. We execute nDGE analysis with the design matrices corresponding to 1000 pseudoconnectomes and compute a t-statistic using the mean and variance of the resulting null distribution. The schematic of the null distribution estimation and pseudoconnectomes that are generated using the chemical connectome and membrane contact adjacency matrices are displayed in Figure II.

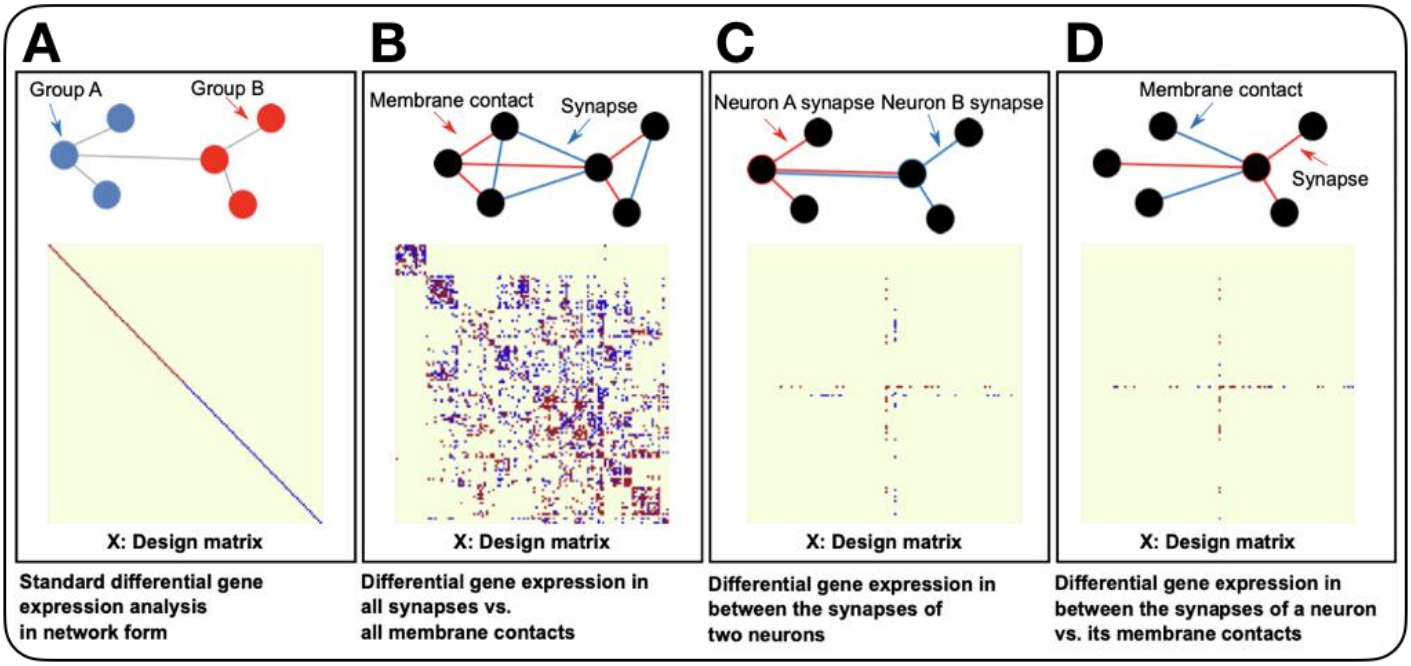

Figure III: Different types of design matrices in nDGE enable the testing of various connectivity hypotheses

Different design matrices enable testing various connectivity hypotheses

Since nDGE is a generalization of standard DGE, it enables testing a variety of hypotheses in addition to what is testable in standard DGE. The types of hypotheses that are tested are encoded in the design matrix of nDGE of which several examples are displayed in Figure III. Figure IIIA shows how standard DGE can be executed through nDGE, by placing 1s and -1s in the diagonal of the design matrix corresponding to the neuron groups. Three other types of hypotheses that can be tested are whether particular gene pairs have global effects of synaptic formation across all the neurons (Figure IIIB), whether there are differential gene co-expression differences in the synapses of two different neurons (Figure IIIC), or which gene co-expression patterns are implicated in the synapses of an individual neuron (Figure IIID). In these scenarios, the design matrix has 1s where there is a synapse and a -1 where there is membrane contact, but no synapse, restricted to the sets of neurons of interest (all, pair, or one, respectively).

Limitations

nDGE model can only account for a single co-expressed gene in either of the two synaptic terminals (pre/post). For this reason, the nDGE model will tend to underestimate the effects of trimeric (or higher-order) protein complexes in the formation and maintenance of synapses. Therefore, it is imperative to keep in mind that lack of significant hits for a particular neuron might not mean that there are no genes implicated in the formation of synapses for that neuron, but rather that higher-order gene interactions might be at play. Conceptually, it is straightforward to extend the model to higher-order gene interactions, but the prohibitive number of combinatorial gene co-expression enumeration is a computational bottleneck.

Comparison with another connectivity analysis technique

A feature of nDGE is that it is a mass-univariate method, which does not consider the possibility of interaction of different co-expressed genes in forming or inhibiting synapses. Therefore, the significance results output by nDGE tends to be very conservative with strict control of type 1 errors. This contrasts with multivariate methods

for explaining the genetic bases of connectivity (Kovacs et al., 2020). Due to the relatively high dimensionality of the gene expression data compared to the number of synapses in the chemical connectome, multivariate models tend to overfit and introduce type 1 errors.

Differential expression of cell adhesion molecules among neurons and their synaptic partners

We utilize nDGE to profile the differential usage of cell adhesion molecules (CAM) in specifying synaptic partners. To accomplish this, we use 84 individualized design matrices for every single neuron class in the nerve ring. For each neuron class, we denote its pre and post-synaptic partners as +1 in the design matrix in their respective rows and columns, with the convention that rows denote pre-synaptic partners and columns denote post-synaptic partners. We also record each neurons' neighbors with membrane contacts but no synaptic links as -1 in their respective rows and columns. See Figure IVB for an illustration of the design matrices. For each of these 84 design matrices, we obtain the nDGE statistic by solving the linear regression problem in equation 2. The null distribution is obtained by generating

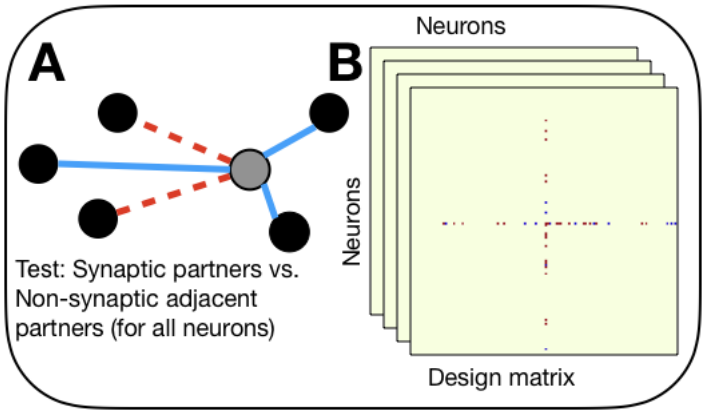

Figure IV: Using neuron specific connectivity design matrices to elucidate the diversity of CAM utilization in synaptic partners in the nerve ring.

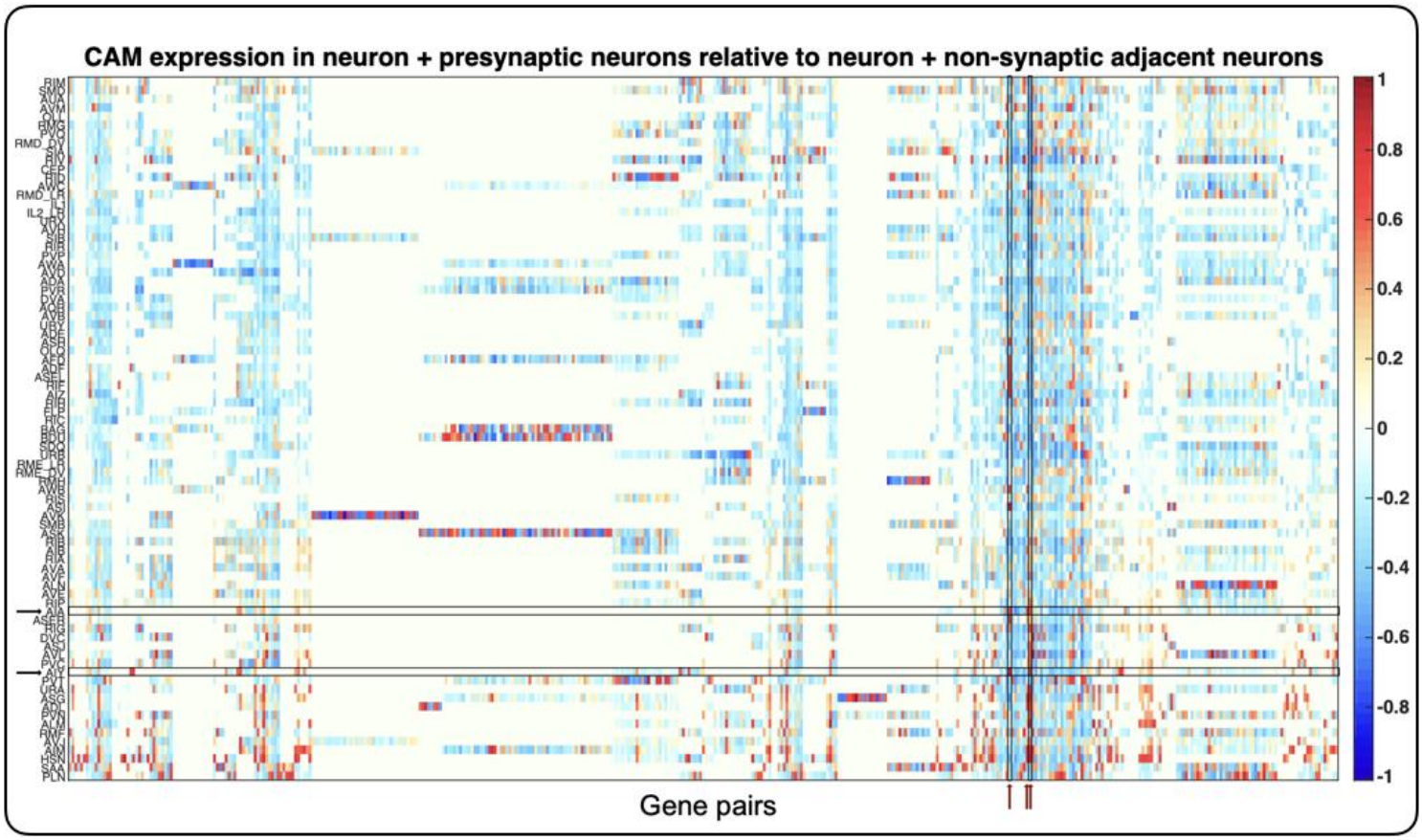

Figure V: Heatmap of CAM usage in presynaptic connections for each neuron in the nerve ring. Neurons are rows, gene pairs are columns. Colors indicate the log of the t-statistic comparing gene pair expression in each neuron + presynaptic partners vs each neuron + non-synaptic adjacent neurons. Positive values (orange and red) indicate higher expression in synaptically-connected neurons relative to adjacent neurons. Negative values (blue) indicate higher expression in adjacent neurons relative to synaptically-connected neurons. Only gene pairs with a log fold change in expression > 0.2 in either direction are displayed. Black arrows indicate AIA and AIY, and red arrows indicate three gene pairs with high log fold changes in presynaptic partners for AIA and AIY.

a pseudo-connectome using the procedure illustrated in Figure II and re-calculating the nDGE statistic using equation 2 with the new design matrix for each neuron class that results from the pseudoconnectome. This procedure is repeated 1000 times to estimate a t-statistic for nDGE by subtracting the null nDGE statistic from the actual statistic and dividing by the standard deviation.

We limit this analysis to the gene co-expression profiles of 138 CAM genes, resulting in 19,044 dimeric interactions for CAM pairs expressed in pre-and post-synaptic partners for each of the 84 neuron classes in the nerve ring. The resulting 19,044 t-statistics for each of these dimeric interactions for each neuron class then is visualized as a heatmap in Figure V and VI. In Figure V, the significance of a dimeric interaction in correlating to a neuron being pre-synaptic to its neuron partners is shown and in Figure VI, the propensity of a dimeric interaction to be correlated with post-synaptic connections to the neuron classes is shown. The heatmaps of CAM usage in the synaptic links of neurons in the nerve ring showcase the diversity of connectional encoding in the nerve ring and shows that different combinations of CAMs might be responsible for the connections of different subsets of neurons.

Case study of the dimeric interactions of post-synaptic partners of AVM and ALM and correlation to nerve ring strata

To closely examine the results of nDGE analysis, we display the differential gene co-expression patterns in the post-synaptic partners vs. adjacent but non-synaptic partners of the neurons AVM and ALM in Figure VII. We also evaluate the organization of neurons in the nerve ring by measuring how correlated each neurons' CAM usage is in post-synaptic connections. This analysis reveals that neurons that utilize similar CAMs in forming post-synaptic connections are not necessarily organized by the four strata observed in the nerve ring (Moyle et al., 2021).

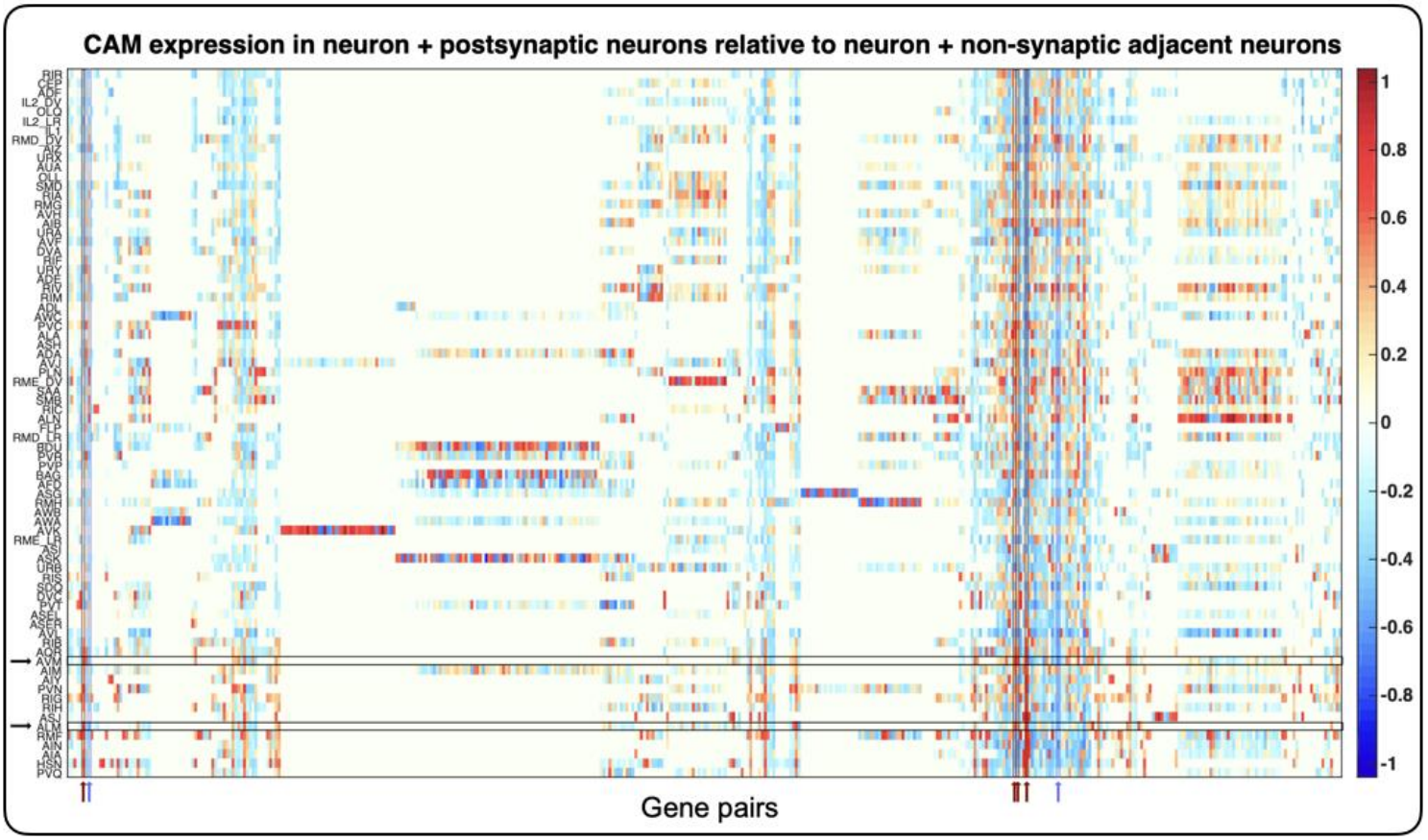

Figure VI: Heatmap as in Figure V, but for CAM usage in postsynaptic connections for each neuron in the nerve ring. Black arrows indicate AVM and ALM, red arrows indicate gene pairs with high log fold change in postsynaptic connections relative to adjacent neurons for both AVM and ALM, and blue arrows indicate gene pairs with high log fold change in adjacent neurons compared to synaptically connected neurons for both AVM and ALM.

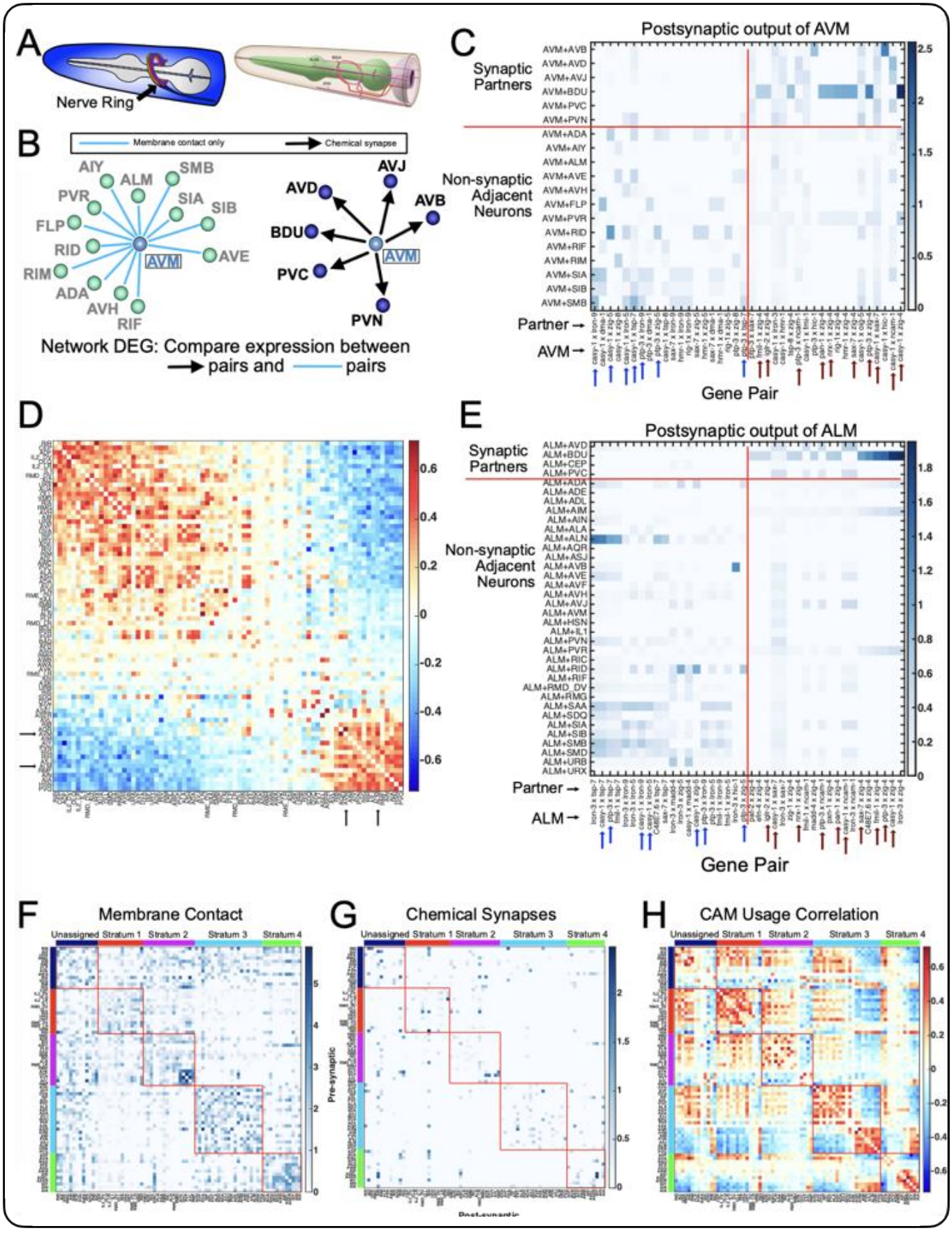

Figure VII: Case study of the nDGE analysis for the post-synaptic output of ALM and AVM (A,B,C,E), the overall diversity of post-synaptic usage of CAMs within the nerve ring (D) and correlation to strata organization (F,G,H). Caption continued on next page.

Figure VII (continued): A) Left: Cartoon representation of the *C. elegans* nerve ring. Right: Subset of nerve ring neurites from the touch neurons AVM, ALM and their synaptic partner BDU. Courtesy of WormAtlas. B) Diagram of neurons receiving chemical postsynaptic output from AVM (right) and neurons with membrane contact but no synapses with AVM (left). Differential cell adhesion molecule (CAM) gene expression was determined between AVM + synaptic partners and AVM + non-synaptic adjacent neurons. C) Heatmap of the 20 gene pairs with highest log fold change in AVM and its postsynaptic outputs compared to AVM and adjacent neurons (right of vertical red line). The 20 gene pairs with highest log fold change in AVM and adjacent neurons compared to AVM and synaptic partners are shown to the left of vertical red line. Blue and dark red arrows indicate gene pairs common among analyses of AVM and ALM (panel D). The first gene listed is expressed in AVM, the second gene is expressed in the synaptic or adjacent partner. D) Correlation matrix showing the relationship of CAM usage across all neurons in the nerve ring (84 neuron types). Note the similarities in CAM usage among the neurons in the lower right-hand corner. The black arrows indicate AVM and ALM (correlation 0.607). E) Heatmap as in C, for ALM. 10 of the top 20 gene pairs (dark red arrows) higher in ALM and its synaptic partners compared to ALM and adjacent neurons are common with the AVM analysis in B. On the left, 7 of the top 20 genes pairs higher in ALM/adjacent are common with AVM (blue arrows). F) The membrane adjacency matrix was grouped according to nerve ring strata as defined in Moyle, et al. Within each stratum, neurons were ordered according to their CAM usage correlations (see panel H). Membrane contact is denser within strata than between strata, but not for the unassigned neurons which generally contact neurons in other strata. Colored bars indicate neuron strata assignments. Each stratum is outlined by a thin red box. G) The same ordering as in F was imposed upon the chemical connectome revealing that most synapses are detected between neurons within the same stratum. H) The CAM usage correlation matrix (as in D) was grouped by strata, then sorted by similarity (using multidimensional scaling) within each stratum. Neurons within stratum 1 are similar to each other. Strata 2, 3, and 4 all show two distinct groups of neurons based on CAM usage at postsynaptic outputs.

Software implementation and brief user guide

The MATLAB implementation of nDGE can be found in Github at https://github.com/cengenproject/connectivity_analysis

Within this package, there is the option to interrogate the differential gene co-expression enrichment in the synaptic partners of user input neurons versus the non-synaptic partners of the user input neuron. Furthermore, there is the additional option to interrogate the differential gene co-expression enrichment between the synaptic partners of two neurons. This latter analysis enables testing whether two neurons utilize the same set of genes to form connections.

Further clarifications and guidance can be obtained by contacting the corresponding author of this software (Erdem Varol) at ev2430@columbia.edu.