Circuit analysis of the *Drosophila* brain using connectivity-based neuronal classification reveals organization of key information processing pathways

Ketan Mehta[1], Rebecca F. Goldin[2], and Giorgio A. Ascoli[1*]

[1] Department of Bioengineering and Center for Neural Informatics, Structures & Plasticity, George Mason University, Fairfax, VA, USA

[2] Department of Mathematical Sciences and Center for Neural Informatics, Structures & Plasticity, George Mason University, Fairfax, VA, USA

[*] Corresponding author
Email: ascoli@gmu.edu (GA)

Short Title: Circuit analysis of *Drosophila* brain

Abstract

We present a functionally relevant, quantitative characterization of the neural circuitry of *Drosophila melanogaster* at the mesoscopic level of neuron types as classified exclusively based on network connectivity. Starting from a large neuron-to-neuron brain-wide connectome of the fruit fly, we use stochastic block modeling and spectral graph clustering to group neurons together into a common "cell class" if they connect to neurons of other classes according to the same probability distributions. Thus, nodes of the identified circuit represent cell classes, while the directed edges represent the connection probabilities between neurons in those respective classes. We then characterize the connectivity-based cell classes with standard neuronal biomarkers, including neurotransmitters, developmental birthtimes, morphological features, spatial embedding, and functional anatomy. Mutual information indicates that connectivity-based classification reveals aspects of neurons that are not adequately captured by traditional classification schemes. Next, using graph-theoretic and random walk analyses to identify neuron classes as hubs, sources, or destinations, we detect patterns of directional connectivity and information flow that potentially underpin specific functional interactions in the *Drosophila* brain. We uncover a core of highly interconnected dopaminergic cell classes functioning as the backbone information highway for multisensory integration. We also recognize glutamatergic-dominant connections between the motor and optic regions, possibly facilitating circadian rhythmic activity. Additional predicted pathways pertain to spatial orientation, fight-or-flight response, and olfactory learning. Mapping the developmental birthtimes of neurons also pinpoints the temporal change of the circuit blueprint over the fly lifespan, revealing distinct critical growth periods of individual pathways. Our analysis provides experimentally testable hypotheses critically deconstructing complex brain function from organized connectomic architecture.

Significance Statement

The potential synaptic circuitry of a neural system constitutes the fundamental architectural underpinning of its in vivo dynamics, plasticity, and functions. The fruit fly neural circuit presented here captures the latent stochastic patterns of network connectivity and provides a fundamental parts list for reverse-engineering brain computation. Mapping the interactions among connectivity-based neuronal classes to development,

morphology, physiology, and transcriptomics links connectomic structure with dynamic functional integration in the brain. These mappings result in testable hypotheses on the relationship between neural architecture and behavior.

Introduction

While it is well established that the brain is composed of distinct cell types, the extent of this cellular diversity remains a long standing open question in neuroscience [1]. A popular approach to studying brain computation is to model the nervous system as a giant interconnected network of these distinct cell types, each playing a specific role. The underlying assumption is that the intricate connectivity patterns of the neural network constitute the fundamental architectural underpinning of its in vivo dynamics and functions [2,3]. From this perspective, quantitatively characterizing the distinct cell types and their relation to synaptic circuitry is paramount to deconstructing brain computation.

Recent advancements in data acquisition and imaging techniques have enhanced our ability to construct very large scale maps of the neural circuitry in the form of connectomes. Macroscale connectomes are well suited to map the circuitry across the whole brain by parcellating it into distinct anatomical regions, but lack the ability to trace individual neurons. On the other hand, microscale connectomes enable the mapping of neural circuitry at the level of single neurons, capturing information pertaining to cell bodies, neurites, and individual synapses. These fine details are critical to determine the relation between structure and signal processing in the brain. However, the inherently massive scale of microscopic connectomes is not ideal for directly inferring brain-wide mechanisms and properties. A common practice is to group neurons in a microscale connectome into a common cell type by subsets of multifarious properties, including physiology, biochemistry, and morphology, and subsequently analyze the interactions between these groups. The expedient abundance of data has allowed the creation of increasingly unbiased descriptive taxonomies [4, 5] for the grouping of neurons. However, these experimentally accessible dimensions are only indirect proxies for the mechanistically more relevant features of network connectivity, developmental control, and experience-dependent plasticity [6, 7].

To address these challenges, we construct a mesoscopic level circuit of neuron types as classified exclusively

based on their synaptic connectivity patterns. We leverage a recently developed mathematical framework [8], which models a connectome as a directed stochastic block model (SBM) graph, to group neurons together into a common "cell class" if they connect to neurons of other classes according to the same probability distributions. Here we build and expand upon that approach applying it to a 19,902-neuron connectome of *Drosophila melanogaster*. The nodes of the identified circuit represent the derived connectivity-based neuronal classes, while the directed edges represent the connection probabilities between neurons in those respective classes. To the best of our knowledge, this is the first neuron type level circuit of its kind.

The *Drosophila* has a tractable brain size of approximately 150,000 neurons, making it an excellent model organism for connectomic analysis. Over the last several years, numerous *Drosophila* studies have contributed to an increased understanding of how certain anatomical regions and neuron types support specific function and behavior. However, the exact underlying mechanics of these functions remain largely unknown, especially when involving multiple modalities and cross-region integration.

To interpret the functional interactions captured by our derived circuit, we map the connectivity-based neuronal classes with traditional neuronal biomarkers, including neurotransmitters, developmental birthtimes, morphological features, spatial embedding, and functional anatomy. In conjunction with these mappings, we employ graph theoretic measures and random walks to analyze the patterns of information flow on the circuit, subsequently identifying pathways underpinning specific functions in the *Drosophila* brain. The predicted pathways pertain to multi-sensory integration, rhythmic circadian activity, spatial orientation, fight-or-flight response, and olfactory learning. These results demonstrate that the derived circuit captures latent patterns of interaction not revealed by traditional neuron type classification alone. Overall, our analysis provides experimentally testable hypotheses critically deconstructing complex brain function from organized connectomic architecture.

Materials and Methods

Data Source

We begin with a neuron-to-neuron brain-wide connectome of the *Drosophila* brain comprised of 19,902 single neurons, obtained from the recently released FlyCircuit v1.2 database [9]. The connectome consists

of a list of reconstructed neurons, each co-registered to a common brain atlas, and of their axonal-dendritic spatial overlaps identifying potential synaptic connections [10]. The neurons form a strongly connected network, in which every neuron can reach all others and can be reached by all others. For each neuron, the FlyCircuit database also provides its associated neurotransmitter, developmental birthtime, and functional-anatomical community.

The connectome is represented as a directed and weighted adjacency matrix A_c, with non-negative, integral entries $a_{c_{ij}}$ specifying the post-synaptic *connection strength* between the i-th and j-th neuron, respectively. More precisely, the connection strength $a_{c_{ij}}$ is the number of overlapping segment pairs between the axonal arbor of i-th neuron and the dendritic arbor of j-th neuron. We refer to the connectome as the strength connectome, and correspondingly to A_c as the strength connectivity matrix. A_c is a very sparse matrix with only 0.9% of its entries being non-zero, a maximum connection strength of 1040, and mean connection strength among non-zero entries of 7.8.

Experimental Design

We stochastically generate multiple binary (unweighted) connectomes from the strength (weighted) connectome. Our functional assumption here is that the structural connectome can be represented as a directed graph with binary edges, i.e., a synaptic connection is either present or absent. Because of the stochastic nature of our model, no two generated binary connectomes are identical and can be considered analogous to connectomes sampled from individual flys. Each generated binary connectome is modeled as a directed (stochastic block model) graph, and clustered using spectral graph clustering. By modeling each graph separately and then merging the individual clustering results via consensus clustering allows for a consistent inference. The consensus clusters are used to construct the final circuit. We elaborate on each of these steps below.

Stochastic Generation of Binary Connectomes. The strength connectivity matrix is used to derive a connection probability matrix A_p, whose entries $a_{p_{ij}}$ represent the probability that the i-th neuron forms a post-synaptic connection with the j-th neuron. Intuitively, the larger is the number of overlapping neurites between two neurons, the greater is the chance that they form a synaptic connection. Let p_{conn} denote the probability that an overlapping axon-dendritic segment-pair forms a synaptic connection. The connection

probability is then derived from the corresponding connection strength using a binomial distribution as follows

$$a_{p_{ij}} = 1 - (1 - p_{\text{conn}})^{a_{c_{ij}}}. \tag{1}$$

We then proceed to generate multiple binary adjacency matrices $A^{(\ell)}$, $l = 1, 2, \ldots, G$, each a stochastic realization of A_p. The entry $a_{ij}^{(\ell)}$ in matrix $A^{(\ell)}$ is an independent Bernoulli draw, being either 1 or 0, with corresponding probabilities $a_{p_{ij}}$ and $1 - a_{p_{ij}}$, respectively. Further, we trim $A^{(\ell)}$ to ensure that it does not have any all-zero rows or columns, i.e., discard any neuron that does not have both a dendritic and axonal connection.

Spectral Graph Clustering. Given a $n \times n$ binary adjacency matrix A, we partition the neurons based exclusively on their patterns of stochastic connectivity. In particular, we partition neurons into a common *class* if they connect to neurons in other classes according to the same probability distribution. We refer to these classes as **connectivity-based classes**. To this aim, we employ a recently developed mathematical framework [8] that models the connectome as a SBM graph, uses spectral graph clustering to predict the number of connectivity-based classes, and assign each neuron to a class.

A SBM graph is parameterized by (i) a block membership probability vector $\rho = (\rho_1, \ldots, \rho_k)$ in the unit simplex $\sum \rho_i = 1$, which partitions the n vertices into k disjoint subsets, and (ii) a $k \times k$ block connectivity probability matrix P, with entries $p_{ij} \in [0, 1]$. A SBM assumes that the probability that vertices (neurons) in the i-th class form an edge (synaptic connection) with vertices in the j-th class can be modeled as an independent Bernoulli trial with parameter p_{ij}.

The spectral graph clustering is a two-step process comprising of adjacency spectral embedding (ASE) in conjunction with Gaussian mixture model (GMM)-based modeling [8]. In the first step, the adjacency matrix is embedded into a lower dimensional space via spectral decomposition of the matrix into its latent vectors. In the second step, the latent vectors are modeled as a Gaussian mixture model (GMM) and clustered using Expectation-Maximization (EM) algorithm.

For any $d \leq \mathrm{rank}(A)$, one can approximate A by a rank d singular value decomposition

$$A \sim U_d D_d V_d^t = XY^t, \tag{2}$$

where $X := U_d\sqrt{D_d}$ and $Y := V_d\sqrt{D_d}$. X and Y are $n \times d$ matrices, and D_d is a $d \times d$ diagonal matrix with nonnegative entries called the singular values. For each A and choice of embedding dimension d, the vectors forming the columns in the augmented matrix $\mathbf{X} := [X|Y]^t$ provide a dot product embedding of A in a $2d$-dimensional space. The columns of the concatenated matrix $\mathbf{X}$ are called **latent vectors**.

The optimal choice of d is a known open problem in literature, with no consensus on a best strategy. Choosing a low d can result in discarding important information, while choosing a higher d than required not only increases computational cost but can adversely effect clustering performance due to the presence of extraneous variables which contribute towards noise in the subsequent statistical inference. We select d by using the method outlined in [11], which determines the cut-off 'elbow point' between relevant and non-relevant dimensions by maximizing a profile likelihood function of the singular values.

For sufficiently dense graphs, and large n, the adjacency spectral embedding (ASE) central limit theorem demonstrates that the n points in $\mathbb{R}^{2d}$ behave approximately as random samples from a k-component GMM [12]. The data is fitted to a GMM using the EM algorithm, which after an initialization of the parameters iteratively improves upon the estimates by maximizing the expected log-likelihood of the probability density function. The accuracy of the EM model fit is known to be sensitive to the initial choices of the parameters. Here, we initialize the EM using parameters obtained by applying model-based hierarchical agglomerative clustering (MBHAC) [13, 14] to all n data points. Finally, the number of components in the GMM are selected using the Bayesian Information Criterion (BIC), which penalizes the model based on the number of free parameters, i.e., model complexity. The EM model fitting is performed using the `mclust` R-package [15].

Consensus Clustering using Modified IVC. Let π_ℓ denote the clustering that results from applying the GMM$\circ$ASE framework to the corresponding binary adjacency matrix $A^{(\ell)}$, for $\ell = 1, \ldots, G$. The clustering π_ℓ maps the n vertices $x_1, \ldots, x_n$ into κ_ℓ clusters, such that there is an association of a cluster number $\pi_\ell(x_i) \in \{1, \ldots, \kappa_\ell\}$, for $i = 1, \ldots, n$. In other words, the clustering π_ℓ partitions the vertices into κ_ℓ

subsets,

$$V_k^{(\ell)} = \{x_i : \pi_\ell(x_i) = k\}, \text{ for } k = 1, \ldots, \kappa_\ell.$$

Note that in general each of the clusterings $\pi_1, \pi_2, \ldots, \pi_G$ has different number of clusters $\kappa_1, \ldots, \kappa_G$, respectively.

Consensus clustering refers to the process of merging the G clusterings into a single, final robust clustering. By combining several clusterings we eliminate any inconsistencies due to noise or outliers, in turn integrating the stochastic variations of individual solutions into a consistent inference. We implement a modified version of the iterative voting consensus (IVC) method introduced by [16]. The modification removes the reliance on a choice of reference clustering. The two-step process is summarized below.

Step 1: For each clustering π_ℓ we create a corresponding updated version π'_ℓ using IVC.

Let $V_k^{(\ell)}$ be a cluster of π_ℓ. For each $j = 1, 2, \ldots, \kappa_\ell$, let $\pi_j(V_k^{(\ell)})$ denote the most common label that π_j assigns to the elements of $V_k^{(\ell)}$. Following [16], define the center $C_k^{(\ell)}$ of $V_k^{(\ell)}$ as the sequence of labels associated with $V_k^{(\ell)}$ as we vary over the clusterings.

While the original method relies on a choice of reference clustering, we iterate sequentially over all G clusterings, obtaining a new clustering for each reference clustering. The procedure for obtaining the updated clusterings is detailed in Algorithm 1.

Step 2: The final clustering $\hat{\pi}$ is constructed as follows: we cluster vertices x and x' together when they occur in a chain of vertices that are pairwise clustered together at least τG times among the $\pi'_1, \ldots, \pi'_G$ clusterings, where τ is a value between 0 and 1, explained below. We then apply a cutoff threshold for a minimal cluster size.

The *similarity matrix* [16] for the clusterings $\pi'_1, \ldots, \pi'_G$ is an $n \times n$ matrix with the ij entry given by the proportion of times that x_i and x_j are clustered together among the clusterings $\pi'_1, \ldots, \pi'_G$. Explicitly, let

$$s_{ij} := \frac{1}{G} \sum_{\ell=1}^{G} \mathcal{I}\left(\pi'_\ell(x_i) = \pi'_\ell(x_j)\right), \tag{3}$$

where $\mathcal{I}$ is the indicator function that is 1 when the statement is true, and 0 otherwise.

Algorithm 1 IVC over all clusterings

Input: $\pi_1, \ldots, \pi_G$, with corresponding number of clusters $\kappa_1, \ldots, \kappa_G$, respectively.

loop $\ell \in \{1, 2, \ldots, G\}$
  $\pi^* = \pi_\ell$ with clusters $V_1^{(\ell)}, \ldots, V_{\kappa_\ell}^{(\ell)}$
  repeat
    Let $\pi_j(V_k^{(\ell)}) := \arg\max_s |\{x \in V_k^{(\ell)} : \pi_j(x) = s\}|$ for $j = 1, \ldots, G$ and $k = 1, \ldots, \kappa_\ell$
    Let $C_k^{(\ell)} := (\pi_1(V_k^{(\ell)}), \ldots, \pi_G(V_k^{(\ell)}))$ for $k = 1, \ldots, \kappa_\ell$.
    loop $i \in \{1, 2, \ldots n\}$
      Let $y_i := \langle \pi_1(x_i), \pi_2(x_i), \ldots, \pi_G(x_i) \rangle$.
      Calculate distance $d(y_i, C_k^{(\ell)}) := \sum_{j=1}^{C} \mathcal{I}(\pi_j(x_i) \neq \pi_j(V_k^{(\ell)}))$.
      Re-assign label $\pi^*(x_i)$ to the value k that minimizes the distance $d(y_i, C_k^{(\ell)})$.
    end loop
  until π^* converges.
  $\pi'_\ell := \pi^*$
end loop

Output: $\pi'_1, \ldots, \pi'_G$.

We define an equivalence relation that depends on a parameter τ among the vertices. We say $x \sim x'$ if there exists a sequence of vertices $x_{i_1}, x_{i_2}, \ldots, x_{i_m}$ where $x_{i_1} = x$ and $x_{i_m} = x'$, such that $s_{i_k, i_{k+1}} \geq \tau G$ for $k = 1, \ldots, m-1$. One may think of two equivalent vertices as having a chain of vertices between them, each pair along the chain being grouped together at least proportion of τ times, among the clusterings $\pi'_1, \ldots, \pi'_G$.

This equivalence relation defines a partition of the n vertices into disjoint subsets, each subset consisting of equivalent vertices. We arbitrarily number these subsets $V_1, \ldots, V_m$, noting that together these are all the vertices.

A low value of τ results in vertices classed together in the final clustering $\hat{\pi}$ even if they are in a chain of vertices that are somewhat infrequently clustered together, while a high value of τ indicates that vertices are in a chain of vertices along which adjacent pairs are almost always clustered together. Therefore, in general, a high value of τ results in consistent inference, but may create many classes each with too few vertices.

To guarantee a minimal size for all clusters in $\hat{\pi}$ we apply a cutoff threshold c_{size} and disregard those subsets of size smaller than this threshold. The final modified iterative voting consensus clustering $\hat{\pi}$ is given by $\{V_k : |V_k| \geq c_{size}\}$, where c_{size} is a natural number indicating the minimum cluster size. We relabel the

subsets so that $\hat{\pi}$ consists of $V_1, \ldots, V_{\hat{\kappa}}$ for some $\hat{\kappa}$. This approach does not require $\hat{\kappa}$ to be known apriori, nor does it require that $\hat{\kappa}$ be a member of $\{\kappa_1, \ldots, \kappa_G\}$.

Each subset V_k of the final clustering $\hat{\pi}$ is referred to as a *connectivity-based class*. As may be observed from the clustering process, not all vertices are classified using modified IVC; vertices not paired frequently enough with other vertices are not included in the final clustering, and are eliminated from further analysis.

Block Probability Estimates. Finally, we estimate the block connectivity probability matrix $\widehat{P}$ by using the clustering $\hat{\pi}$ and averaging across the G binary matrices. Recall that $A^{(\ell)} = (a_{rs}^{(\ell)})$ corresponds to the ℓth binary adjacency matrix.

The ij-th entry of $\widehat{P}$ is the average proportion of connected neurons between the i-th and j-th clusters of the final clustering $\hat{\pi}$, given by

$$\hat{p}_{ij} := \frac{1}{G} \sum_{\ell=1}^{G} \frac{|\{a_{rs}^{(\ell)} = 1 : \hat{\pi}(x_r) = i \text{ and } \hat{\pi}(x_s) = j\}|}{|\{x_r : \hat{\pi}(x_r) = i\}||\{x_s : \hat{\pi}(x_s) = j\}|} \tag{4}$$

for each $r, s \in \{1, 2, \ldots, n\}$, and $i, j = \{1, 2, \ldots, \hat{\kappa}\}$. The ratio in (4) defines a value from 0 to 1.

Statistical Analysis

Model Stability - Parameter Robustness. The model parameters we chose for the cluster analysis were $\mathbb{P} = \{p_{\text{conn}} = 0.15, d = 11, G = 100, \tau = 0.95, c_{\text{size}} = 100\}$. Since the final clustering is a function of these parameters, we must validate the analysis to ensure that the stochastic framework is reasonably robust with respect the the specific choice of values. Let $\mathbb{P} = \{p_{\text{conn}}, d, G, \tau, c_{\text{size}}\}$ denote the set of the five key model parameters that impact the cluster analysis, and $\widehat{\pi}$ denote the final clustering (resulting from IVC) which assigns a label to every vertex. For different values of the parameters $\mathbb{P}_1, \mathbb{P}_2, \ldots$, we obtain a corresponding clustering $\hat{\pi}(\mathbb{P}_1), \hat{\pi}(\mathbb{P}_2), \ldots$, respectively. Unfortunately, the relation between the parameters $\mathbb{P}$ and the clustering function $\hat{\pi}(\mathbb{P})$ is too complex to derive an analytical expression. Also, the total number of possible combinations of valid parameter choices is extremely large, and the clustering process is computationally expensive. Therefore, it is not possible to re-perform the clustering multiple times by randomizing the choice of parameters, and use brute force to examine the clustering statistics.

Instead of randomly varying the parameters, we specifically choose only those (meaningful) parameter val-

ues which are most likely to yield a successful clustering. The set of values considered for these parameters can be found in Table 2. We compare $\hat{\pi}(\mathbb{P}_1)$ and $\hat{\pi}(\mathbb{P}_2)$ for each pair of parameters $\mathbb{P}_1$ and $\mathbb{P}_2$ using the *adjusted Rand index* (ARI) [17]. ARI is a popular similarity score for comparing two partitioning schemes for the same data points. A higher value of ARI indicates high similarity, with a (maximum) value of one indicating that they are identical and an expected value of zero for two independent clusterings.

We deem the model to be 'stable' if two different clusterings resulting from the same or similar parameters have an adjusted rand index of 0.90 or higher. This conservative threshold for determining a stable set of parameters ensures minimal dependence on stochastic elements of our process.

Comparison with other Neuronal Classification Schemes. How much of a neuron's identity derived from network connectivity is explained by the knowledge of its neurotransmitter, or morphology (and vice versa)? To answer this question, we employ an information theoretic approach to compare and quantify the interdependency between different classification schemes.

Let random variable X, Y denote two different classifications of the same neuronal dataset, e.g., X represents the connectivity-based classification, and Y represents anatomical-functional community. We then calculate the joint and conditional distributions empirically, e.g., the posterior probability $p(Y = j \mid X = i)$ that the neuron is associated with the j-th functional-anatomical community, given that it belongs to the i-th connectivity-based class. The joint and conditional distributions are in turn used to estimate the entropy $H(X)$ [18], which quantifies the amount of information associated with the outcome of a random variable X. We also estimate the mutual information $I(X;Y)$, which is a measure of the mutual dependency between the two random variables. Mutual information intuitively measures how much information about Y is provided by knowledge of X alone, and vice-versa. If the two variables are independent then $I(X;Y) = 0$.

The proportion of information explained by X about Y can then be measured using normalized mutual information or NMI [19], often also referred to as the uncertainty coefficient [20],

$$NMI(X \rightarrow Y) = I(X;Y)/H(Y)\ , \qquad NMI(Y \rightarrow X) = I(X;Y)/H(X)\,. \tag{5}$$

Unlike mutual information which is a symmetric measure, NMI is a directional measure while also being normalized between $[0, 1]$.

Information Pathways - Absorption and Driftiness

We perform a series of random walks on the derived circuit to (i) identify key nodes in the circuit, and (ii) identify information pathways (the most important edges in the network that connect these key nodes). The advantage of using random walks is that it provides a *dynamic* measure, unlike neighborhood-based measures (e.g., degree centrality) or path length-based measures (e.g., betweenness centrality), which focus only on the intrinsic structure of the network. Specifically, a random walk describes a signal originating at a single node and propagating through the network via its structural connections. We refer to the origin node as the source, and the final destination node as the target. The random walk model we use to analyze information flow on our directed, and weighted graph is an extension of the *absorption* and *driftiness* measures introduced in [21].

Consider a SBM graph with k number of blocks $\mathcal{V} = \{V_1, V_2, \ldots, V_\kappa\}$, and parameterized by the block connection probability matrix $P = (p_{ij}) \in [0,1]^{\kappa\times\kappa}$. The total number of vertices in the i-th block is denoted by $|V_i|$, also referred to as the size of the block. The probability that a vertex in the i-th block forms an edge with a vertex in the j-th block, is given by an independent Bernoulli distributed with probability p_{ij}. The total number of edges between any two blocks i and j is then a binomially distributed random variable, with expected value $p_{ij}|V_i||V_j|$.

In our circuit representation of the SBM graph, each circuit node corresponds to a block, and the directed, weighted edges of the circuit correspond to the block connection probabilities $P(V_i, V_j) = p_{ij}$. A random walk on the circuit describes a sequence of n nodes that form a path $\mathcal{P} = (V_{j_1}, V_{j_2}, \ldots, V_{j_n}) \in \mathcal{V}^n$. For ease of notation we re-index by setting $W_i = V_{j_i}$, for $i = 1, 2, \ldots, n$. We consider only simple paths, i.e., the random walk visits each node only once, and ignore self-loops.

Let $|\widetilde{V}_i|$ denote the scaled block sizes,

$$|\widetilde{V}_i| := \frac{|V_i|}{\min\{|V_1|, |V_2|, \ldots, |V_\kappa|\}}.$$

The cost of the step from node W_i to W_{i+1} on the path is then defined as

$$c(W_i, W_{i+1}) := \frac{1}{P(W_i, W_{i+1})|\widetilde{W}_i||\widetilde{W}_{i+1}|}\,,$$

such that, the larger the expected number of edges between two blocks in the SBM graph, the lower is the cost required to traverse that particular path on the circuit.

The total path length of a random walk is obtained by summing up the cost of all steps (S1 Fig),

$$\ell(\mathcal{P}) \;:=\; \sum_{i}^{n-1} c(W_i, W_{i+1})\,.$$

Absorption is defined as the average path length between a source node V_s and target node V_t,

$$A(V_s, V_t) := \frac{1}{m} \sum_{\mathcal{P}\in\mathcal{P}(V_s,V_t)} \ell(\mathcal{P})\,, \quad \text{for } V_s, V_t \in \mathcal{V},$$

where m is the number of random walks between V_s and V_t, and $\mathcal{P}(V_s, V_t)$ denotes the set of all random walks from V_s to V_t. Ideally, we would want m to equal the total number of possible paths between a source and target node. For a large circuit with multiple edges, however, the total number of possible paths increases factorially, and calculating $A(V_s, V_t)$ is computationally prohibitive. In practice, by the law of large numbers, choosing a very large m is sufficient to guarantee convergence to the true average path length.

In addition to pairwise absorption values, we also calculate the average out-absorption, and average in-absorption for each of the nodes,

$$A_o(V_s) := \frac{1}{\kappa - 1} \sum_{k=1}^{\kappa} A(V_s, V_k)\,, \qquad A_i(V_t) := \frac{1}{\kappa - 1} \sum_{k=1}^{\kappa} A(V_k, V_t).$$

The average in- and out-absorption can be used to detect important nodes. A high out-absorption indicates that the node is located upstream relative to the propagation of information flow. These nodes would be more likely to have a signal originate and act as transmitters. A high out-absorption therefore would imply that the node has a high potential to activate other nodes. On the other hand, a high in-absorption would

indicate nodes that are downstream in the chain of signal propagation, acting as receivers where the signal terminates. A node with both high out- and in-absorption would most likely indicate a hub node serving as an intermediary connector for information exchange.

Driftiness is defined as the ratio between absorption and the shortest path length between the source and target node,

$$D(v_s, v_t) := \frac{A(V_s, V_t)}{\delta(V_s, V_t)}.$$

Correspondingly, the average out-driftiness and in-driftiness for each node are respectively given by

$$D_o(V_s) := \frac{1}{\kappa - 1} \sum_{k=1}^{\kappa} D(V_s, V_k)\,, \qquad D_i(V_t) := \frac{1}{\kappa - 1} \sum_{k=1}^{\kappa} D(V_k, V_t).$$

Driftiness can be used to detect important information paths. High out (or in) driftiness indicates presence of multiple outgoing (or incoming) paths, and can be thought of as a measure of robustness between outgoing (or incoming) signals. On the other hand, low values of driftiness indicates the presence of a key path, possibly stressing the importance of an individual information path. Nodes with low average driftiness are often local nodes.

Periods of Critical Growth

The circuit for each day is obtained by only using neurons that were born on that day or before. For example, the embryo (day 0) circuit consist only of embryo neurons, whose total is denoted by n_0. The day 1 circuit consists of a total of n_1 number of neurons whose birthtime was either embryo or day 1, i.e., $n_1 = n_0 +$ day 1 neurons. The circuit size therefore grows cumulatively, reaching on day 9 the original full circuit.

The circuit for the i-th day is a SBM parameterized by $\{P^{(i)}, \rho^{(i)}, n_i\}$, where $P^{(i)}$ is the $\kappa \times \kappa$ block probability matrix, and $\rho^{(i)}$ is the block membership probability vector $\rho^{(i)} := (\rho_1^{(i)} \dots \rho_\kappa^{(i)})$, respectively.

The cluster sizes on the i-th day are then given by the product $n_i(\rho_1^{(i)} \dots \rho_\kappa^{(i)})$. The size of the k-th cluster on the i-th day is therefore $n_i \rho_k^{(i)}$, $k = 1, 2, \dots, \kappa$. The expected reference cluster sizes are chosen as $n_i(\rho_1^{(9)} \dots \rho_\kappa^{(9)})$. These are the cluster sizes that we would expect on i-th day if the final circuit was simply

scaled down to n_i. This is therefore equivalent to comparing a circuit parameterized by $\{P^{(i)}, \rho^{(i)}, n_i\}$ against a reference circuit parameterized by $\{P^{(i)}, \rho^{(9)}, n_i\}$.

Cluster growth on the i-th day is defined as a percentage difference between actual daily cluster growth relative to uniform growth

$$g_k^{(i)} = \frac{\text{actual size} - \text{expected size}}{\text{expected size}} \times 100\% \quad = \quad \frac{n_i \rho_k^{(i)} - n_i \rho_k^{(9)}}{n_i \rho_k^{(9)}} \times 100\%. \tag{6}$$

The top 25 percentile of the days which showed the greatest rate of growth as compared to the day before $(g_k^{(i)} - g_k^{(i-1)})$, are determined to be the "critical growth period".

Results

Mesoscale Circuit Blueprint of the Fly Brain

To construct the mesoscale level circuit we clustered the neuron-to-neuron brain-wide connectome obtained from the FlyCircuit v1.2 database [9], using the stochastic framework described in "Experimental Design". This process identified a total of 54 connectivity-based classes. Out of the 19,902 neurons, 15,571 (78.24%) were assigned to a class, while the remaining neurons were discarded from the circuit analysis. Each connectivity-based class is represented as a circuit node, while the directed, weighted edges represent the connection probabilities. For any two connectivity-based classes in the circuit, each neuron within one class has the same probability of forming a synaptic connection (indicated by the incident edge) with each member of the other class of neurons.

Using posterior conditional probabilities (as described in "Comparison with other Neuronal Classification Schemes"), we identified the dominant functional-anatomical community type among the neurons in a connectivity-based class, and labeled the class accordingly (Fig. 1a). Specifically, if $p(Y = j \mid X = i) > 0.67$ then the i-th connectivity cluster was assigned the "label" corresponding to the j-th anatomical-functional community. Clusters with no dominant community meeting this 67% threshold criterion were not assigned any community label.

We also created similar mappings for the neurotransmitter distribution (Fig. 1b), and developmental birth-

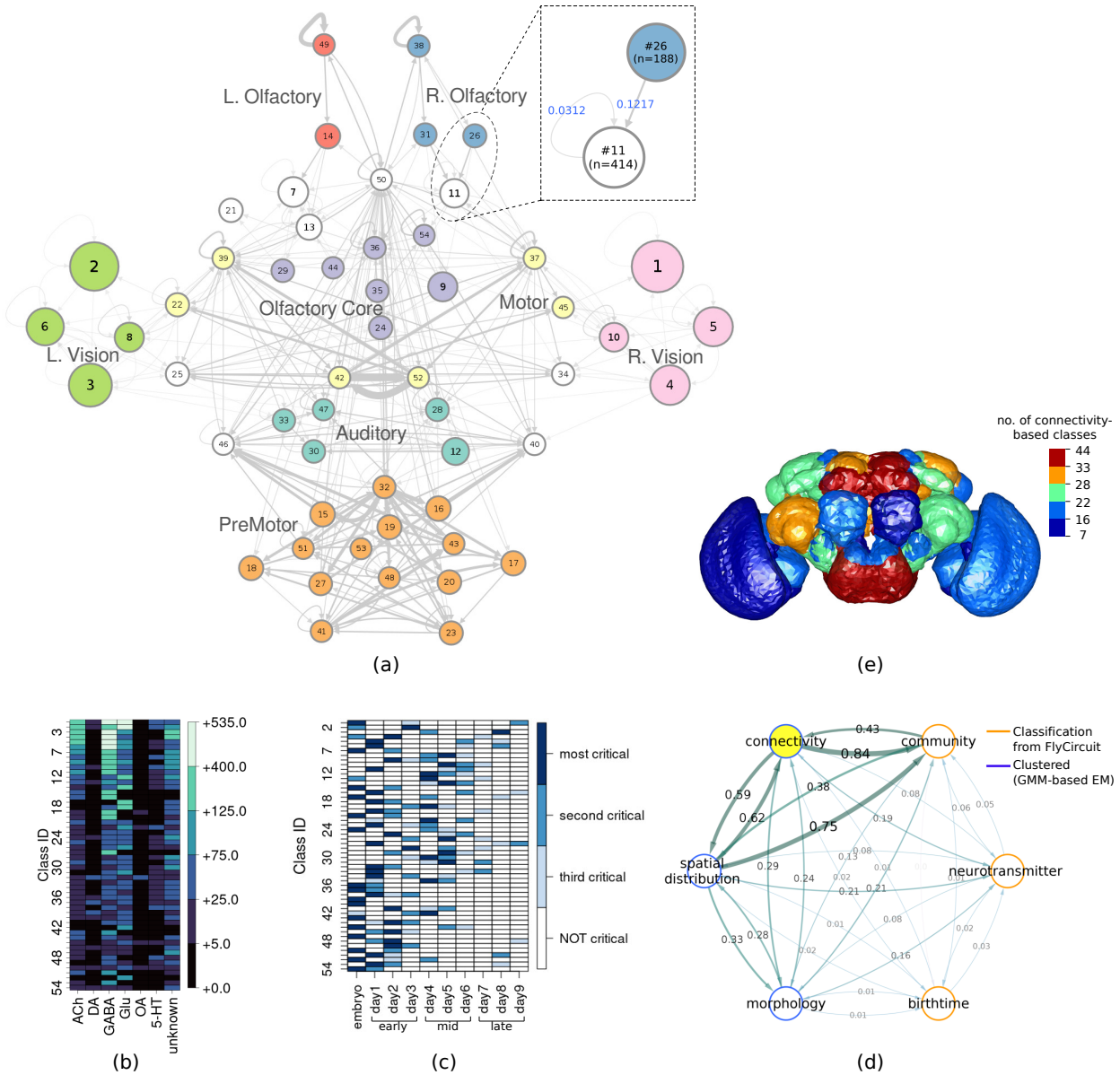

Figure 1: (a) Circuit diagram of 54 connectivity-based cell classes. The size of each node is proportional to the number of neurons in that cell class, while the edge thickness is proportional to the directional connection probabilities. Classes are color-coded by their dominant anatomical-functional community as reported in FlyCircuit.tw [9]. (b) Distribution of neurons in the 54 connectivity-classes by neurotransmitter. ACh: acetylcholine; DA: dopamine; GABA: gamma-amino-butyric acid; Glu: glutamate; OA: octopamine; 5-HT: serotonin. (c) Mapping of connectivity-based classes by periods of critical of growth (embryo, early, mid, late), identified by the developmental birthtimes of constituent neurons. (d) Normalized mutual information among six classification schemes. A value of 1 indicates the two classifications are identical, while 0 indicates that they are independent. (e) Spatial distribution showing the number of connectivity-based classes innervating each of the 75 neuropil regions.

times (Fig. 1c). These mappings enable us to characterize and interpret the circuit in relation to the traditional neuronal biomarkers. For example, connectivity-based class 1 was identified as a part of right vision, contains no octopaminergic cells, and shows critical growth in late development. A detailed list of all descriptor labels assigned to each connectivity-based class is provided in S1 File. Further, we compiled the evolution of the circuit over the span of 10 days into a movie clip identifying critical growth periods (S1 and S2 Movies).

To quantify how much knowledge of the traditional biomarkers is explained by neuronal connectivity and vice versa, we quantified the inter-dependency between each pair of classification schemes using normalized mutual information. The results (Fig. 1d) reveal that connectivity-based classification, neurotransmitter, and birth time are practically independent of each other ($\approx 0 - 0.20$ NMI). While connectivity-based classification largely explains (0.84 NMI) functional-anatomical communities, the reverse is not true (0.43 NMI), suggesting that connectivity-based classification constitutes a high-resolution refinement of functional-anatomical communities [9]. As expected, connectivity-based classification reflects certain aspects of the spatial neuropil distribution (≈ 0.60 NMI). However, the classification obtained using only the spatial embedding of the neurons is substantially different from our connectivity-based classification, which also groups together spatially nonadjacent neurons that innervate distinct neuropils (Fig. 1e). We additionally performed a morphology-based classification by combining 15 morphological features and branch patterns as quantified by persistence homology vectors [22], obtained from NeuroMorpho.Org, but observed very low dependency between our connectivity-based classification and morphology (< 0.30 NMI). Altogether, these results indicate that the connectivity-based classification captures latent network characteristics that are otherwise not revealed by known neuronal biomarkers.

Parameter Choice and Model Stability

Binomial Probability of Synaptic Connection. The binomial connection probability parameter p_{conn} was chosen to obtain a mean connection probability among potentially connected vertices of 0.5. Recall that p_{conn} determines the connection probability matrix A_p, whose entries are $a_{p_{ij}}$ are derived from the post-synaptic connection strengths $a_{c_{ij}}$ using a binomial distribution (1).

The mean connection probability $\bar{a}_p$ is the mean value of non-zero entries in A_p. We pick p_{conn} to be the

value that minimizes $|\overline{a}_p - 0.5|$. Empirically, we found $p_{\text{conn}} = 0.15$.

To validate our choice of p_{conn}, we compare our stochastically generated connectomic matrix obtained using FlyCircuit data, versus the recently released Janelia hemibrain connectome [23]. The edge strength of a stochastically generated FlyCircuit connectomic matrix is the number of overlapping axonal-dendritic segments that successfully form a synapse if modeled using a binomial distribution with success probability 0.15. The edge strength in the Janelia connectomic matrix is the synapse count between two neurons, experimentally determined using electron microscopy (EM). The vertex degree is the number of incoming and outgoing synapses of a neuron in the connectome (Table 1).

Connectome	# Neurons	Sparsity	Avg. Vertex Deg	Avg. Edge Str
EM Experimental Data (Janelia)	21,739	0.75%	326.64	4.04
Stochastically Generated from FlyCircuit Data	19,902	$0.46\% \pm 0.01$	181.79 ± 0.07	2.32 ± 0.01
Truncated Connectomes: Hemibrain Only				
EM Experimental Data (Janelia)	11,461	$0.75\% \pm 0.01$	172.73 ± 2.23	4.01 ± 0.09
Stochastically Generated from FlyCircuit Data	11,461	$0.76\% \pm 0.01$	173.36 ± 0.09	2.58 ± 0.02

Table 1: Comparing the EM-derived Janelia hemibrain connectome [23] against our stochastically generated connectome using $p_{\text{conn}} = 0.15$. Both connectomic datasets were appropriately resized so that they span the same neuropils in the *Drosophila* hemibrain. The mean ($\pm$ standard deviation) values for the sparsity, degree, and strength, were calculated using a total of 100 randomly sampled connectomes.

Since the FlyCircuit dataset has neurons sampled from the entire brain, we appropriately truncate the stochastically generated connectome by discarding all vertices that correspond to neurons that were not embedded in the neuropil regions investigated in the Janelia hemibrain dataset. We then also truncated the Janelia hemibrain connectome to match the size of the truncated FlyCircuit connectome, by randomly (uniformly) sampling neurons from the entire Janelia dataset. The sparsity and degree statistics of the resized connectomes are almost identical (Table 1), supporting our choice for $p_{\text{conn}} = 0.15$.

Given the difference in average edge strength, we also compare the normalized histograms distributions of the particular variable (Fig. 2) for the two truncated hemibrain connectomes. Let $f(x)$ and $g(x)$ denote the probability density function (PDF) of the edge strengths for the the Janelia hemibrain connectome and the stochastically generated FlyCircuit connectome, respectively. To measure the similarity between the two

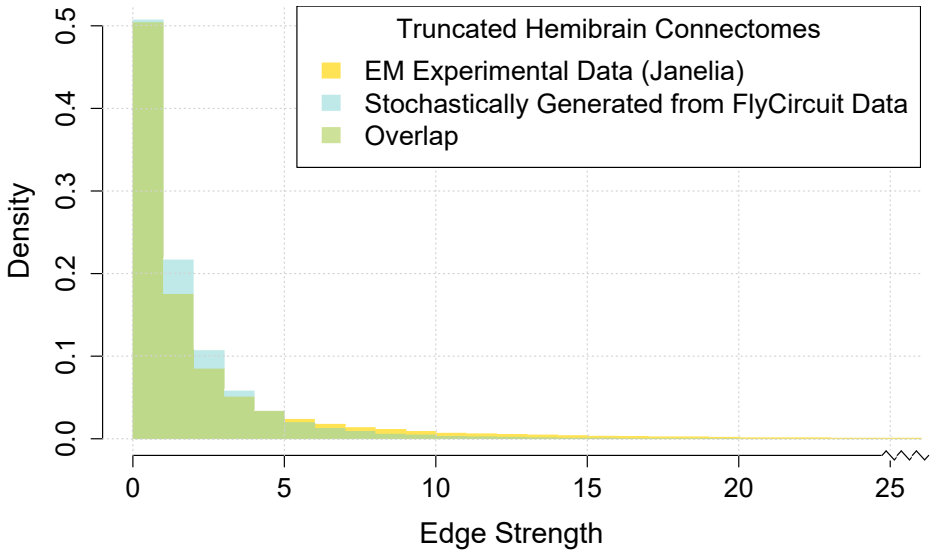

Figure 2: Normalized histogram distributions (probability density functions) of the edge strengths, comparing the experimentally determined Janelia hemibrain connectome [23] against our stochastically generated connectome using $p_{\text{conn}} = 0.15$. Both connectomic datasets were appropriately resized so that they span the exact same neuropils in the *Drosophila* hemibrain.

distributions we calculate the area over which the two PDFs intersect [24]:

$$\text{Int}(f, g) = \int \min\big(f(x), g(x)\big) dx \,,$$

where an intersection value $\text{Int}(f, g) = 1$ indicates prefect overlap. The intersection between the histogram distributions of edge strengths (Fig 2) is 0.964 revealing that, despite a difference in average edge strength due to long-tail effects, the two distributions are nearly identical. The high measures of similarity between the graph-theoretical statistics of the stochastically generated connectome when compared to the EM experimental data corroborate our choice of p_{conn}.

Varying the Parameters. To validate the model stability relative to the values of all parameters in our framework (Table 2), we kept $p_{\text{conn}} = 0.15$ constant and re-performed the clustering for alternative parameter values. The possible choices for the embedding dimensionality $d \in \{11, 15\}$ were determined by identifying the first and second elbow-point [11], respectively, on the scree plot (S2 Fig) of singular values. The values for the other parameters were similarly picked to yield optimal clustering (Table 2).

Parameter	Description	Value
p_{conn}	Binomial probability of synaptic connection	**0.15**
d	Dimensionality of embedding	**11**, 15
G	Number of graphs	**100**, 200, 400
τ	Times two neurons were placed in same cluster during IVC	0.99, **0.95**, 0.90
c_{size}	Minimum cluster size	150, 125, **100**, 75, 50

Table 2: Parameter choices (values used for creating the final circuit are bold)

	100	200	400
100	0.9514 (71.94%)	0.9818 (72.93%)	0.9692 (72.97%)
200	-	0.9649 (71.92%)	0.9737 (73.35%)
400	-	-	0.9878 (74.89%)

Table 3: Pairwise ARI values for varying the number of graphs $G \in \{100, 200, 400\}$ while keeping all other parameters constant $\{p_{\text{conn}} = 0.15, d = 11, \tau = 0.95, c_{\text{size}} = 100\}$. The value in parenthesis indicates the percentage of neurons classified (not discarded).

For example, to investigate whether increasing the number of the stochastically generated binary graphs impacted clustering, we successively doubled only the parameter G (100, 200, 400) while keeping all other parameters constant (at the bold values in Table 2). The very high pairwise adjusted rand index (ARI) values (Table 3) indicate that the resulting clusterings are all extremely similar, and that the model is stable over this choice of parameter values. Therefore, increasing the number of graphs beyond 100 does not substantially alter the final clustering, but only raises computational complexity. The ARI values along the diagonal entries of the table were obtained by using identical parameters and number of graphs for each clustering, but repeating the stochastic generation process with different random seeds. The diagonal ARI values are therefore an indication of the dependence of the clusterings on the stochastic elements of our process, and how well we can replicate the results.

Similar ARI results for other different combinations of the parameters in Table 2 are detailed in the S2 File. Overall, the ARI remained high ($\gtrapprox 0.90$) across several different combinations, demonstrating model stability over a wide range of parameter values.

Dopaminergic Hubs form a Backbone Information Highway for Multisensory Integration

To investigate the dopaminergic pathway (Fig. 3a), we specifically consider only those neurons in each connectivity-based class whose neurotransmitter is dopamine, and appropriately recalculate the connection probabilities between classes. Classes without dopamine neurons are therefore devoid of incoming or outgoing edges. Six out of 54 classes (32,50,37,39,40,46) alone account for 57% (143/252) of dopaminergic neurons. These same classes also show the highest weighted degree and betweenness-centrality among all classes (Fig. 3b), acting as information hubs in the circuit.

The inferred pattern of information flow (Fig. 3c) indicates that these six interconnected dopaminergic classes work together to facilitate integration across the sensory and motor communities, forming a backbone information highway. The tightly interconnected loop between the hub classes observed at the circuit level was also identified at the level of individual neurons (Fig. 3d). Interestingly, these hub classes also exhibit the most extensive neuronal morphologies among all classes, as characterized by highest total length and number of bifurcations. Developmental birthtimes reveal that the backbone information highway is one of the earliest formed pathways in the entire circuit with 90% of its connections (27/30 edges) established in embryo. The three sensory-innervated hub classes show critical growth during the embryo period, while the three premotor-innervated hub classes show critical growth on day 1.

Overall, the analysis suggests that the identified pathway is critical for cross-modal integration of stimuli. The mechanics of how a fly combines input from multiple sensory modalities to guide behavior in its natural environment are not fully understood, e.g., when combining mechanosensory signals about wind direction along with visual cues to track an odor. We predict that selectively suppressing activity of neurons in the identified hub classes using the TH-Gal4 driver line will disrupt the fly's ability to perform such complex integrative tasks.

Glutamate-dominant Pathway bridges the Motor System inter-laterally and interlinks it with Vision to facilitate Rhythmic Activity

To identify the glutamate-dominant pathway, we restricted attention to glutamatergic neurons and appropriately recalculated the connection probabilities between classes. We retained the top 20% of the edges by weight (Fig. 4a) and eliminated any edges below this threshold. This pathway shows critical growth during

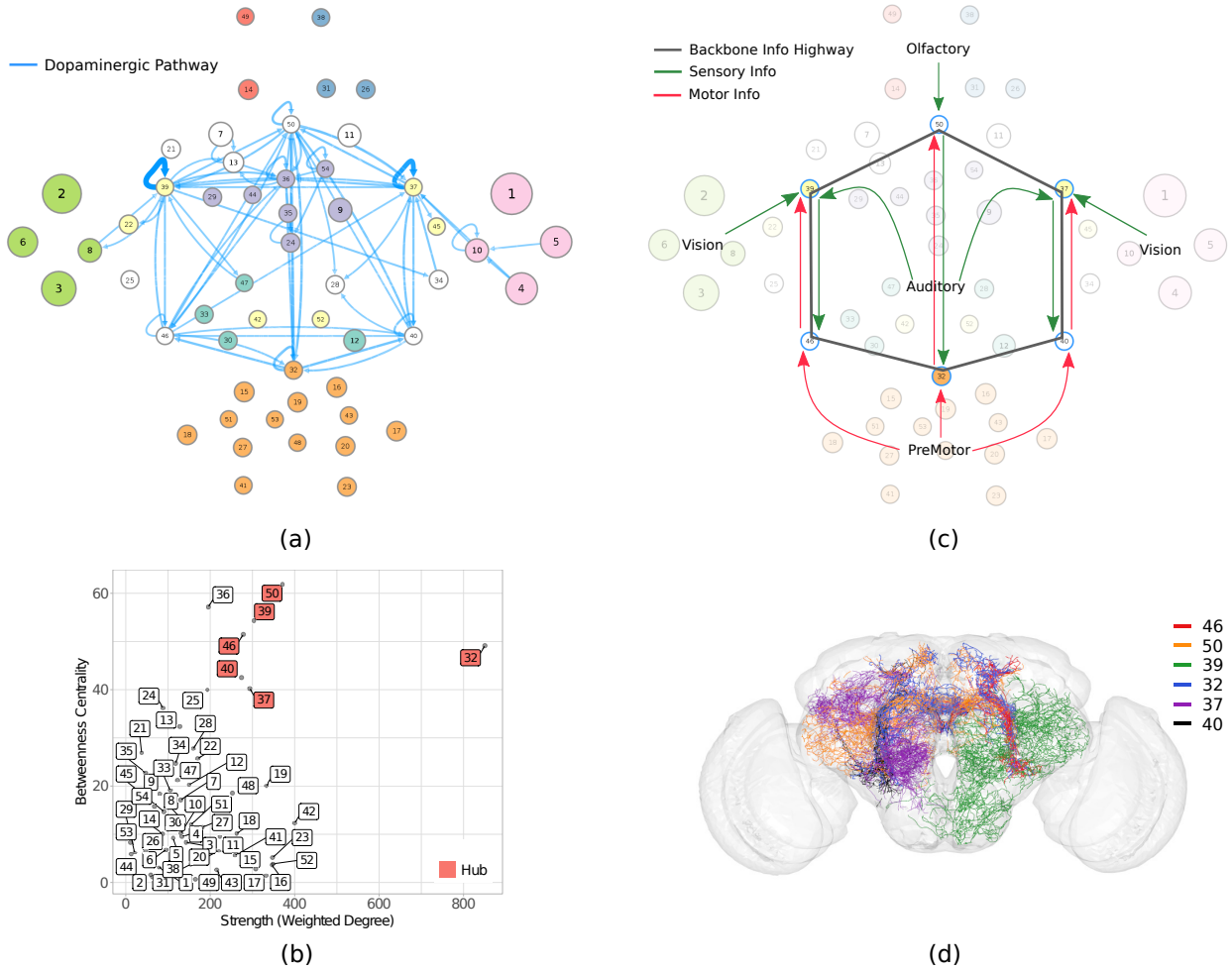

Figure 3: (a) Dopaminergic pathway. (b) Graph theoretic connectivity measures calculated for the circuit show that six classes (32, 50, 37, 39, 40, 46) have both high weighted degree and betweenness centrality, and act as information hubs. (c) Pattern of directional information flow identified using random walks on the circuit reveal a backbone information highway for multisensory integration. (d) Twenty-five representative dopaminergic neurons from the six hub classes that form a single continuous path, looping over all 30 edges of the backbone information highway. Numerous such subsets of neurons exist within these identified classes that reflect the circuit-level pathway.

early development, with classes 42 and 52 especially populated in embryo. These two classes also have extremely high weighted degree (but low betweenness centrality, Fig. 4b) and appear to act as a central bridge for coordinating motor activity between hemispheres.

Pacemaker neurons are glutamatergic in Drosophila, implicating this neurotransmitter in behavioral rhythmicity, such as generation of locomotion and modulation of circadian photoreception [25–27]. Although

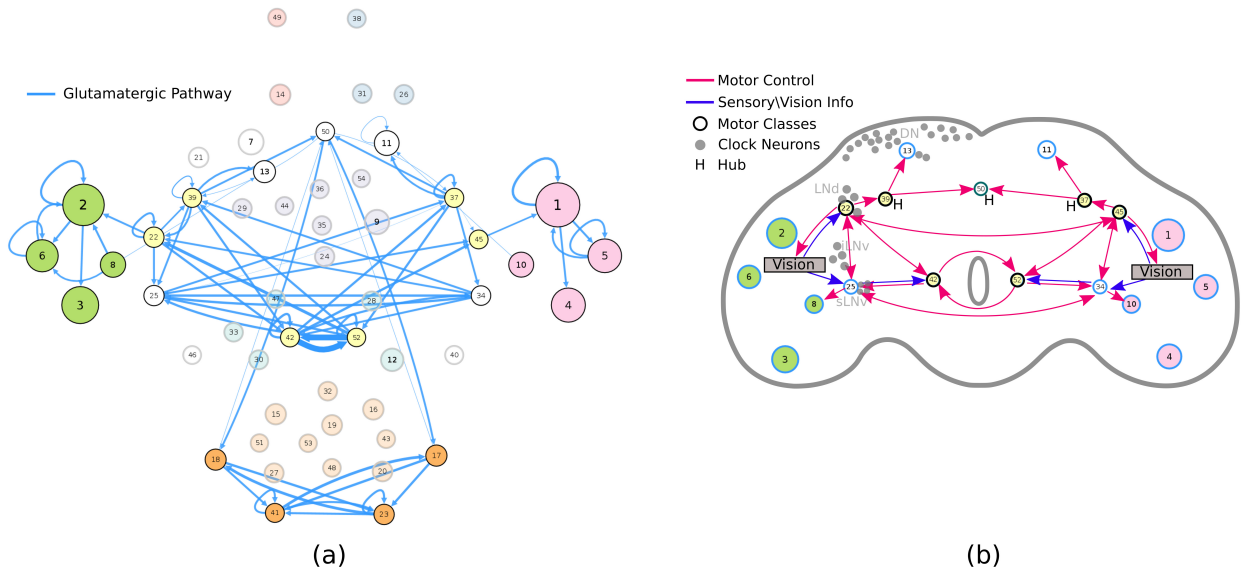

Figure 4: (a) Dominant glutamatergic pathway. (b) Inferred pattern of information flow associated with rhythmic motor activity identified using random walks on the circuit. The representation of information flow is superimposed on a schematic of the fly brain illustrating the approximate location of the dominant glutamatergic classes in relation to clock neurons.

the anatomical locations of clock neurons in the Drosophila brain are known, their connectivity pathways remain largely unexplored. Circadian properties may be an emergent function of the the underlying circuitry and not the property of clock neurons alone [26, 28].

Random walk analysis over the circuit (Fig. 4b) revealed that classes 22 and 25 in the left hemisphere, and 34 and 45 in the right hemisphere, serve as the primary gateways between the vision and motor systems. Specifically, these classes receive signals from the visual system and pass on this information to the downstream motor regions, while also transmitting motor signals back to vision – possibly control signals to modulate input. Further, the inferred patterns of information flow feature multiple inter-connecting loops in the motor region, resembling feedback control, with strong interaction between left and right hemispheres. We predict that suppressing the activity of glutamatergic neurons in these identified classes, but not elsewhere, by VGlut-GAL4 driver line targeting will disrupt motor patterning and circadian rhythms.

Mechanosensory Pathways rely on Parallel Serotonergic and Octopaminergic Neurotransmission

We identified connectivity-based classes associated with sensory-motor processing through mappings to functional-anatomical communities and spatial distribution. Specifically, we examined those classes whose neurons heavily innervate the central complex and are identified to be part of the premotor, auditory, or motor communities. The central complex has been linked to spatial representation, spatial orientation, visual place learning, and navigation [29]. These combined pathways are also believed to be associated with other mechanosensory functions such as gravity sensation, wind sensation, and air current feedback during flight [30].

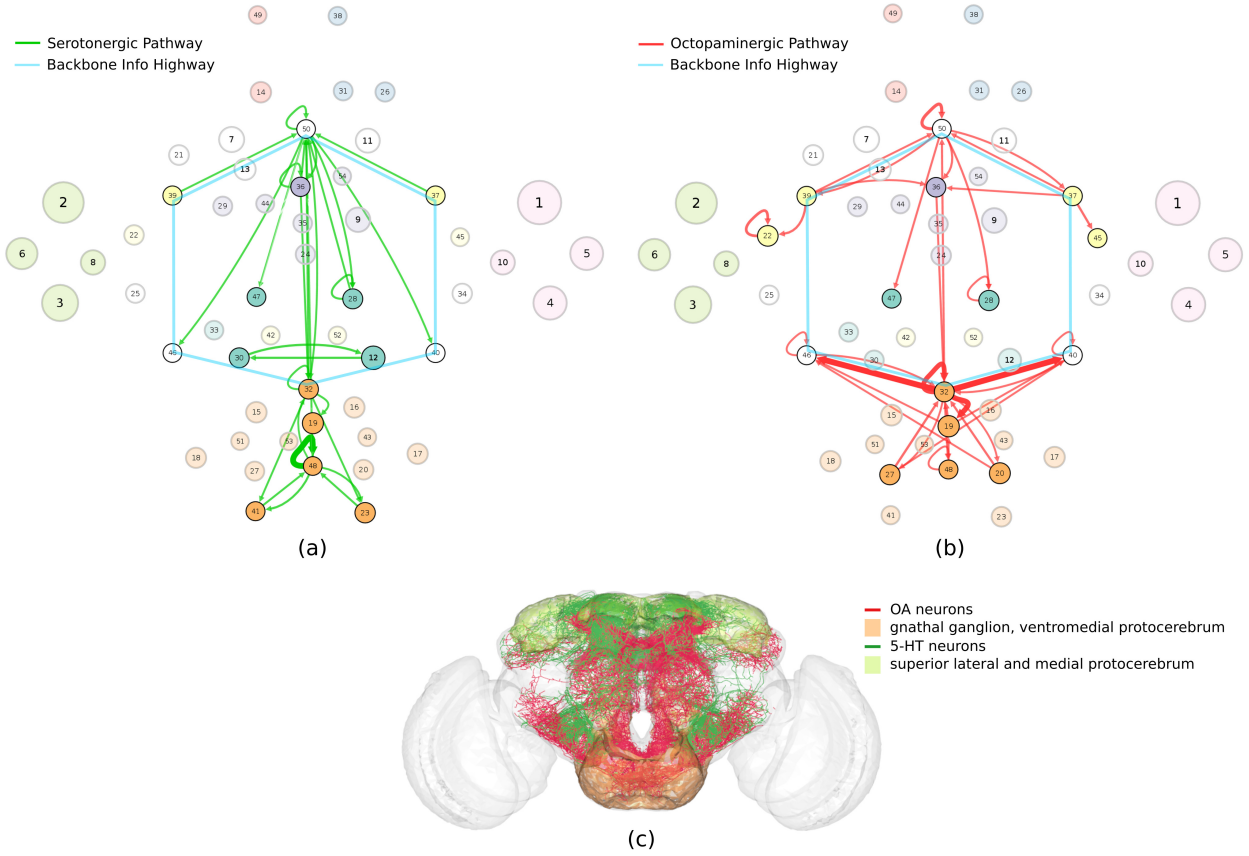

Figure 5: (a) Serotonergic pathway associated with spatial orientation. (b) Octopaminergic pathway associated with fight-or-flight response. (c) The mechanosensory pathways diverge along the dorsal-ventral axis with serotonergic and octopaminergic neurons innervating distinct neuropils.

The serotonergic pathways were identified by restricting attention to the serotonergic neurons in the classes

identified above, and appropriately recalculating the connection probabilities (Fig. 5a). The serotonergic system is a critical component of place learning [31]. The octopaminergic pathways were similarly identified by considering only the octopaminergic neurons in the same classes as above (Fig. 5b). While there are only 140 octopaminergic neurons in our connectome, octopamine is critical for essentially all sensory inputs and is known to play a role in fight-or-flight response [32], stress-related enhancement of motor activity [33], and aggression [34]. Compared to the premotor region which shows critical growth in the early period, the auditory classes show growth predominantly during mid-development, with the octopaminergic classes (19, 20, and 27) showing continued growth even in the latest stages. Interestingly, while running in parallel throughout most of the mechanosensory system, these two pathways diverge along the dorsal-ventral axis (Fig. 5c), with octopaminergic neurons predominantly innervating the gnathal ganglion and ventromedial protocerebrum, and serotonergic neurons innervating the superior lateral protocerebrum and superior medial protocerebrum.

The analysis suggests that these parallel pathways aid in navigation and place learning by working in conjunction with the backbone information highway to gather information using sensory cues during flight. We predict that inhibiting the serotonergic neurons in this pathway using the Trh-Gal4 driver would interfere with the fly's ability to orient itself in relation to its surrounding. In addition, we hypothesize that selectively inhibiting octopaminergic neurons in this pathway using the Tdc2-Gal4 driver would prevent the fly from responding appropriately to potential threats such as moving away from a predator.

Olfactory Learning depends on non-redundant GABAergic and Glutamatergic Inhibition, while Odor Tracking relies on fast Serotonergic Pathways.

To identify the circuit pathways involved in olfaction, we examined those classes which are associated with the olfactory communities and whose neurons heavily innervate the antennal lobes, lateral horn, and mushroom body. The mushroom body plays a central role in associative learning and memory, particularly of olfactory information [35]. Within the above identified classes, we considered the neurons associated with the three neurotransmitters involved in olfactory processing, namely GABA [35], glutamate [36], and serotonin [37], and recalculated the connection probabilities accordingly (Fig. 6). The identified pathways are consistent with findings that GABA and glutamate function in parallel to gate olfactory input from the

antennal lobe to the mushroom body [36].

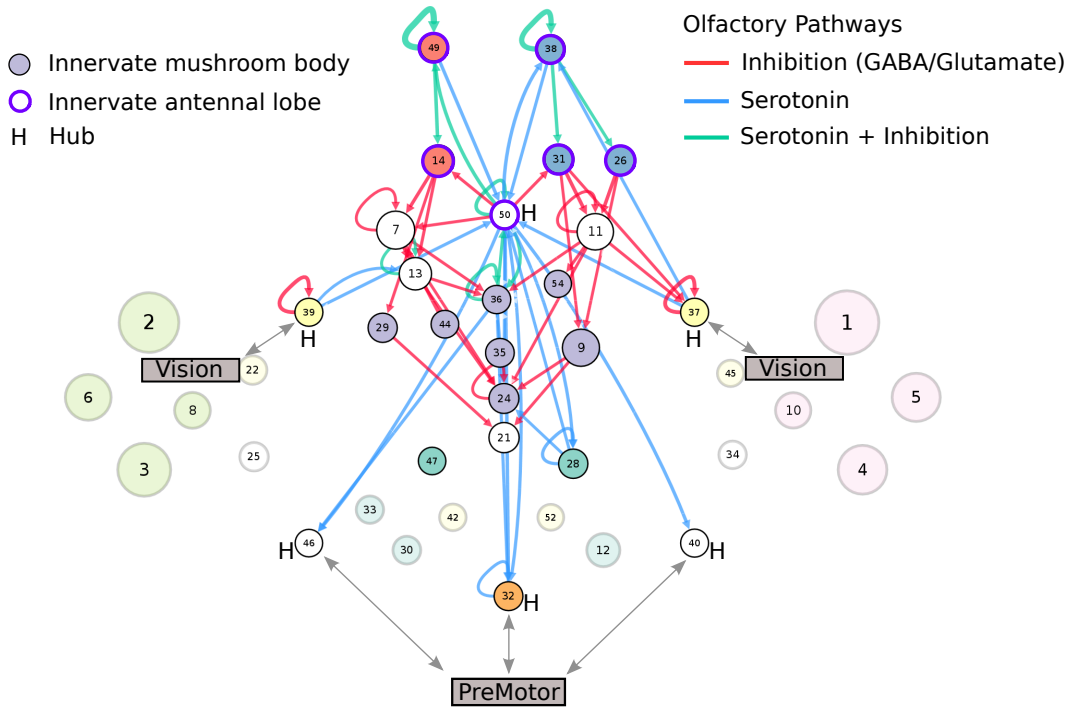

Figure 6: Predicted olfactory pathways. GABA and glutamate function in parallel to enable olfactory learning, while serotonergic connections enable cross-modal integration to drive odor tracking.

Random walk analysis revealed that these pathways have predominantly short local connections with important individual information paths and uncharacteristically few alternative communication routes. Thus, we predict that selectively suppressing glutamatergic (via VGlut-Gal4 driver) and GABAergic (via Gad1-Gal4) transmission in this pathway would impact olfactory learning tasks such as discriminating among similar odors or varying odor concentrations. Flies are known to track odor by integrating cross-modal mechanosensory and visual input in flight [38], yet little is known about these descending neural pathways. While the inferred pattern of information flow reveals strong olfaction-premotor and olfaction-vision link via the hub classes, we discovered that there are very few serotonergic connections to or from the mushroom body (olfactory core) classes themselves. Birthtime analysis shows that the critical growth period of the olfaction-vision connections occurs in early development, while the olfaction-premotor connections are already present in embryo. We hypothesize that this pathway enables fast connections from a stimulus (antennal lobe) to a behavioral response (premotor), and that selective blockage of those serotonergic neurons via Trh-Gal4 driver line targeting would interfere with odor localization.

Discussion

Over the last two decades, *Drosophila* has been a popular model organism for studying how structural connections in the brain give rise to functional interactions. Numerous studies have previously focused on the functional dissection of *regional* neural circuitry in the *Drosophila*, including that of vision [39, 40], olfaction [35], mushroom body [41], and central complex [29], to name a few. With advancements in data acquisition and processing, recent large-scale projects [23, 42] have been successful in generating detailed connectomes of the *brain-wide* neural circuitry, that go well beyond a single region. While a complete brain-wide wiring diagram is a necessary prerequisite, unraveling the underlying mechanistic pathways to interpolate function and behavior, is a problem that cannot be solved by gathering more data alone [23]. One of the fundamental obstacles is the absence of a quantitative specification of neuron types and how they relate to neural circuitry.

The prevalent approach to analyzing brain-wide microscopic connectomes is to group neurons by a subset of their anatomy and/or morphology, e.g., local processing units [42] or compartments [43], and then attempt to characterize the connectivity between these groups. Therefore what is effectively mapped is the anatomical connectivity between brain regions, and not the true latent connectivity between neurons. While there is undoubtedly a strong spatial component to synaptic connectivity, anatomical organization is by itself insufficient to deconstruct brain computation.

In contrast, we group neurons in the microscopic connectome based exclusively on their stochastic patterns of connectivity. The derived mesoscopic circuit is therefore a direct representation of the latent connectivity between neurons. The information captured by connectivity-based classification is complementary, yet synergistic, to that provided by traditional neuronal biomarkers - together enabling functional interpretation not previously possible. By inferring the information flow on the derived circuit and simultaneously characterizing the circuit with respect to multiple biomarkers, we have identified functional pathways supporting multiple sensory modalities and cross-region integration.

A key pathway identified by our circuit is the backbone information highway comprising of six interconnected dopaminergic-centric hubs working together to facilitate integration of the sensory and motor pathways. The neural circuitry for vision, olfaction, and hearing are well characterized, but the pathways for

how these modalities combine to guide behavior, e.g., during navigation, remain elusive. Interestingly, two out of these six hubs are the only classes overall that consist of all three neuromodulators — DA, OA, and 5-HT. Further, since the hubs innervated multiple neuropils, this pathway would not have been revealed by a classification purely based on anatomy, neurotransmitter, or morphology. Dopamine is a known neuromodulator central to reinforcement signaling [44], and responding to salient stimuli related to olfaction and vision. Also, while dopamine affects visual tracking [45], its role in place learning or spatial memory remains to be elucidated [31].

Another prominent pathway revealed in our analysis is the glutamate-dominant pathway for circadian activity. The clock neuron subsets in the *Drosophila* brain are well described including their spatial location, morphology, neurotransmitter, and role in behavioral rhythmicity. Nevertheless, the inputs from other neurons onto clock neurons, and the exact neuronal pathways underlying the entrainment of the clock for synchronization with the environment are poorly known [46]. The identified pathways link the motor and vision regions inter-laterally, suggesting involvement in coordination of rhythmic activity. Unfortunately, however, it is unknown which, if any, of the neurons in the 19,902-neuron FlyCircuit dataset are actually clock neurons.

Over the last several years, optogenetics [47] have been invaluable in studying how dynamical interactions on the neural circuit give rise to specific *Drosophila* behavior. These include circuits relating to rhythmic motor function [48], olfaction [49–51], central complex function [52], and learning and memory [53]. Our derived circuit is amenable to optogenetic manipulation, thus allowing for experimental verification of the predicted functional behavior of our identified pathways. Specifically, for each of the characterized circuit pathways, we identify GAL4-UAS driver lines [54] that could potentially be used for the selective neuronal targeting. We also predict the associated loss (or gain) of function that would accompany the selective deactivation (or activation) of the corresponding neurons in these pathways.

Our connectomic dataset consists of 19,902 neurons reconstructed from multiple specimen using light microscopy, approximately corresponding to 12% of the total number of neurons in a *Drosophila* brain. Given that that these neurons may not necessarily be uniformly sampled from the entire brain, it is unlikely that our analysis adequately captured all essential features of the brain-wide circuitry. The constructed circuit does

however show excellent left-right symmetry, indicating bilaterally uniform sampling of the neurons. Moreover, our modeling framework scales extremely well with the size of the connectome; thus, this analysis can be progressively refined as more neurons continue to be morphologically reconstructed, registered to a common atlas, and shared with the community [55]. Our model could also be applied to dense connectomic reconstructions from electron microscopy, which is considered the gold standard in the field. However, we view the knowledge derived from sparse light microscopy collected from multiple brains valuable in its own merit, as it provides complementary information on the shared circuit elements across individuals [56]. In this regard, it is useful to note that our probabilistic model, currently restricted to Bernoulli adjacency matrices, can be generalized to cluster non-binary connectomes using a weighted SBM [57, 58], which would allow accounting for the number of synapses between connected neuron pairs [59] in the identification of connectivity-based classes.

Acknowledgements

We thank our colleagues at the Center for Neural Informatics, Structures, and Plasticity (CN3) for many insightful discussions. We are especially grateful to Dr. Sumit Nanda for feedback on the neuropil distribution of the FlyCircuit reconstructions and to Dr. Carolina Tecuatl for help with the Janelia hemibrain connectome tracing files. This research was supported in part by NIH grants R01NS39600, R01NS86082, U01MH114829, and RF1MH128693.

Code Accessibility

The source code for all R scripts and Supporting Information including supplementary figures, movie, tables, and a collation of all biomarker labels are available at github.com/k3t3n/FlyConn.

Supporting Information

S1 Fig. Random walk path length.

S2 Fig. Scree plot.

S3 Fig. Neurotransmitter pathways.

S4 Fig. 3D spatial embedding of each class

S1 File. Class labels and data.

S2 File. ARI results.

S1 Movie. Critical growth per day.

S2 Movie. Critical growth time periods.

References

[1] Zeng H, Sanes JR. Neuronal cell-type classification: challenges, opportunities and the path forward. Nature Reviews Neuroscience. 2017;18(9):530–546. doi:10.1038/nrn.2017.85. 3

[2] Abbott LF, Bock DD, Callaway EM, Denk W, Dulac C, Fairhall AL, et al. The mind of a mouse. Cell. 2020;182(6):1372–1376. doi:10.1016/j.cell.2020.08.010. 3

[3] Jonas E, Kording K. Automatic Discovery of Cell Types and Microcircuitry from Neural Connectomics. eLife. 2015;4:e04250. doi:10.7554/eLife.04250. 3

[4] DeFelipe J, López-Cruz PL, Benavides-Piccione R, Bielza C, Larrañaga P, Anderson S, et al. New insights into the classification and nomenclature of cortical GABAergic interneurons. Nature Reviews Neuroscience. 2013;14(3):202–216. doi:10.1038/nrn3444. 3

[5] Yuste R, Hawrylycz M, Aalling N, Aguilar-Valles A, Arendt D, Armañanzas R, et al. A community-based transcriptomics classification and nomenclature of neocortical cell types. Nature neuroscience. 2020;23(12):1456–1468. doi:10.1038/s41593-020-0685-8. 3

[6] Armañanzas R, Ascoli GA. Towards the automatic classification of neurons. Trends in Neurosciences. 2015;38(5):307–318. doi:10.1016/j.tins.2015.02.004. 3

[7] Shepherd GM, Marenco L, Hines ML, Migliore M, McDougal RA, Carnevale NT, et al. Neuron names: A gene-and property-based name format, with special reference to cortical neurons. Frontiers in Neuroanatomy. 2019;13:1–25. doi:10.3389/fnana.2019.00025. 3

[8] Mehta K, Goldin RF, Marchette D, Vogelstein JT, Priebe CE, Ascoli GA. Neuronal Classification from Network Connectivity via Adjacency Spectral Embedding. Network Neuroscience. 2021;5(3):689–710. doi:10.1162/netn_a_00195. 4, 6

[9] Shih CT, Lin YJ, Wang CT, Wang TY, Chen CC, Su TS, et al. Diverse Community Structures in the Neuronal-Level Connectome of the Drosophila Brain. Neuroinformatics. 2020;18(2):267–281. doi:10.1007/s12021-019-09443-w. 4, 15, 16, 17

[10] Rees CL, Moradi K, Ascoli GA. Weighing the evidence in Peters' rule: does neuronal morphology predict connectivity? Trends in Neurosciences. 2017;40(2):63–71. doi:10.1016/j.tins.2016.11.007. 5

[11] Zhu M, Ghodsi A. Automatic Dimensionality Selection from the Scree Plot via the Use of Profile Likelihood. Computational Statistics & Data Analysis. 2006;51(2):918–930. doi:10.1016/j.csda.2005.09.010. 7, 19

[12] Athreya A, Priebe CE, Tang M, Lyzinski V, Marchette DJ, Sussman DL. A limit theorem for scaled eigenvectors of random dot product graphs. Sankhya A. 2016;78(1):1–18. doi:10.1007/s13171-015-0071-x. 7

[13] Fraley C. Algorithms for model-based Gaussian hierarchical clustering. SIAM Journal on Scientific Computing. 1998;20(1):270–281. doi:10.1137/S1064827596311451. 7

[14] Scrucca L, Raftery AE. Improved initialisation of model-based clustering using Gaussian hierarchical partitions. Advances in data analysis and classification. 2015;9(4):447–460. doi:10.1007/s11634-015-0220-z. 7

[15] Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal. 2016;8(1):289–317. doi:10.32614/RJ-2016-021. 7

[16] Nguyen N, Caruana R. Consensus Clusterings. In: Seventh IEEE International Conference on Data Mining (ICDM 2007). Omaha, NE, USA; 2007. p. 607–612. 8

[17] Hubert L, Arabie P. Comparing partitions. Journal of Classification. 1985;2(1):193–218. doi:10.1007/BF01908075. 11

[18] Cover TM. Elements of information theory. John Wiley & Sons; 1999. 11

[19] Särndal CE. A comparative study of association measures. Psychometrika. 1974;39(2):165–187. 11

[20] Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes 3rd edition: The art of scientific computing. Cambridge University Press; 2007. 11

[21] Costa LdF, Batista JL, Ascoli GA. Communication structure of cortical networks. Frontiers in Computational Neuroscience. 2011;5:6. doi:10.3389/fncom.2011.00006. 12

[22] Li Y, Wang D, Ascoli GA, Mitra P, Wang Y. Metrics for Comparing Neuronal Tree Shapes Based on Persistent Homology. PLOS ONE. 2017;12(8):e0182184. doi:10.1371/journal.pone.0182184. 17

[23] Scheffer LK, Meinertzhagen IA. A connectome is not enough–what is still needed to understand the brain of Drosophila? Journal of Experimental Biology. 2021;224(21):jeb242740. doi:10.1242/jeb.242740. 18, 19, 27

[24] Cha SH. Comprehensive survey on distance/similarity measures between probability density functions. International Journal of Mathematical Models and Methods in Applied Sciences. 2007;1(4):300–307. doi:10.1.1.154.8446. 19

[25] Guo F, Yu J, Jung HJ, Abruzzi KC, Luo W, Griffith LC, et al. Circadian Neuron Feedback Controls the Drosophila Sleep–Activity Profile. Nature. 2016;536(7616):292–297. doi:10.1038/nature19097. 22

[26] Azevedo RVDMd, Hansen C, Chen KF, Rosato E, Kyriacou CP. Disrupted Glutamate Signaling in Drosophila Generates Locomotor Rhythms in Constant Light. Frontiers in Physiology. 2020;11. doi:10.3389/fphys.2020.00145. 22, 23

[27] Zimmerman JE, Chan MT, Lenz OT, Keenan BT, Maislin G, Pack AI. Glutamate is a wake-active neurotransmitter in Drosophila Melanogaster. Sleep. 2017;40(2). doi:10.1093/sleep/zsw046. 22

[28] Dissel S, Hansen CN, Özkaya Ö, Hemsley M, Kyriacou CP, Rosato E. The Logic of Circadian Organization in Drosophila. Current Biology. 2014;24(19):2257–2266. doi:10.1016/j.cub.2014.08.023. 23

[29] Turner-Evans DB, Jayaraman V. The Insect Central Complex. Current Biology. 2016;26(11):R453–R457. doi:10.1016/j.cub.2016.04.006. 24, 27

[30] Boekhoff-Falk G, Eberl DF. The Drosophila Auditory System. WIREs Developmental Biology. 2014;3(2):179–191. doi:10.1002/wdev.128. 24

[31] Zars T. Spatial Orientation in Drosophila. Journal of Neurogenetics. 2009;23(1-2):104–110. doi:10.1080/01677060802441364. 25, 28

[32] Sotnikova TD, Gainetdinov RR. Octopamine and Other Monoamines in Invertebrates. In: Encyclopedia of Neuroscience. Academic Press; 2009. p. 9–15. 25

[33] Sujkowski A, Ramesh D, Brockmann A, Wessells R. Octopamine Drives Endurance Exercise Adaptations in Drosophila. Cell reports. 2017;21(7):1809–1823. doi:10.1016/j.celrep.2017.10.065. 25

[34] Hoyer SC, Eckart A, Herrel A, Zars T, Fischer SA, Hardie SL, et al. Octopamine in Male Aggression of Drosophila. Current Biology. 2008;18(3):159–167. doi:10.1016/j.cub.2007.12.052. 25

[35] Busto GU, Cervantes-Sandoval I, Davis RL. Olfactory Learning in Drosophila. Physiology. 2010;25(6):338–346. doi:10.1152/physiol.00026.2010. 25, 27

[36] Liu WW, Wilson RI. Glutamate Is an Inhibitory Neurotransmitter in the Drosophila Olfactory System. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(25):10294–10299. doi:10.1073/pnas.1220560110. 25, 26

[37] Johnson O, Becnel J, Nichols CD. Serotonin Receptor Activity Is Necessary for Olfactory Learning and Memory in Drosophila Melanogaster. Neuroscience. 2011;192:372–381. doi:10.1016/j.neuroscience.2011.06.058. 25

[38] Duistermars B, Frye M. Multisensory Integration for Odor Tracking by Flying Drosophila. Communicative & Integrative Biology. 2010;3:60–3. doi:10.4161/cib.3.1.10076. 26

[39] Zhu Y. The Drosophila Visual System: From Neural Circuits to Behavior. Cell Adhesion & Migration. 2013;7(4):333–344. doi:10.4161/cam.25521. 27

[40] Borst A, Haag J, Mauss AS. How Fly Neurons Compute the Direction of Visual Motion. Journal of Comparative Physiology A. 2020;206(2):109–124. doi:10.1007/s00359-019-01375-9. 27

[41] Li F, Lindsey JW, Marin EC, Otto N, Dreher M, Dempsey G, et al. The Connectome of the Adult Drosophila Mushroom Body Provides Insights into Function. eLife. 2020;9:e62576. doi:10.7554/eLife.62576. 27

[42] Shih CT, Sporns O, Yuan SL, Su TS, Lin YJ, Chuang CC, et al. Connectomics-Based Analysis of Information Flow in the Drosophila Brain. Current Biology. 2015;25(10):1249–1258. doi:10.1016/j.cub.2015.03.021. 27

[43] Scheffer LK, Xu CS, Januszewski M, Lu Z, Takemura Sy, Hayworth KJ, et al. A Connectome and Analysis of the Adult Drosophila Central Brain. eLife. 2020;9. doi:10.7554/eLife.57443. 27

[44] Waddell S. Reinforcement Signaling in Drosophila; Dopamine Does It All after All. Current Opinion in Neurobiology. 2013;23(3). doi:10.1016/j.conb.2013.01.005. 28

[45] Riemensperger T, Isabel G, Coulom H, Neuser K, Seugnet L, Kume K, et al. Behavioral Consequences of Dopamine Deficiency in the Drosophila Central Nervous System. Proceedings of the National Academy of Sciences. 2011;108(2):834–839. doi:10.1073/pnas.1010930108. 28

[46] Hamasaka Y, Wegener C, Nässel DR. GABA Modulates Drosophila Circadian Clock Neurons via GABAB Receptors and Decreases in Calcium. Journal of Neurobiology. 2005;65(3):225–240. doi:10.1002/neu.20184. 28

[47] Kohsaka H, Nose A. Optogenetics in Drosophila. In: Yawo H, Kandori H, Koizumi A, Kageyama R, editors. Optogenetics: Light-Sensing Proteins and Their Applications in Neuroscience and Beyond. Advances in Experimental Medicine and Biology. Springer; 2021. p. 309–320. 28

[48] Ravbar P, Zhang N, Simpson JH. Behavioral evidence for nested central pattern generator control of Drosophila grooming. Elife. 2021;10:e71508. doi:10.7554/eLife.71508. 28

[49] Bellmann D, Richardt A, Freyberger R, Nuwal N, Schwärzel M, Fiala A, et al. Optogenetically induced olfactory stimulation in Drosophila larvae reveales the neuronal basis of odor-aversion behavior. Frontiers in behavioral neuroscience. 2010;4:27. doi:10.3389/fnbeh.2010.00027. 28

[50] Owald D, Felsenberg J, Talbot CB, Das G, Perisse E, Huetteroth W, et al. Activity of defined mushroom body output neurons underlies learned olfactory behavior in Drosophila. Neuron. 2015;86(2):417–427. doi:10.1016/j.neuron.2015.03.025. 28

[51] Hige T, Aso Y, Modi MN, Rubin GM, Turner GC. Heterosynaptic plasticity underlies aversive olfactory learning in Drosophila. Neuron. 2015;88(5):985–998. doi:10.1016/j.neuron.2015.11.003. 28

[52] Franconville R, Beron C, Jayaraman V. Building a functional connectome of the Drosophila central complex. Elife. 2018;7:e37017. doi:10.7554/eLife.37017. 28

[53] Simpson JH, Looger LL. Functional Imaging and Optogenetics in *Drosophila*. Genetics. 2018;208(4):1291–1309. doi:10.1534/genetics.117.300228. 28

[54] Brand AH, Perrimon N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. Development. 1993;118(2):401–415. doi:10.1242/dev.118.2.401. 28

[55] Ascoli GA, Maraver P, Nanda S, Polavaram S, Armañanzas R. Win–win data sharing in neuroscience. Nature methods. 2017;14(2):112–116. doi:10.1038/nmeth.4152. 29

[56] Ascoli GA. Trees of the brain, roots of the mind. MIT Press; 2015. 29

[57] Ng TLJ, Murphy TB. Weighted stochastic block model. Statistical Methods & Applications. 2021; p. 1–34. doi:10.1007/s10260-021-00590-6. 29

[58] Aicher C, Jacobs AZ, Clauset A. Learning latent block structure in weighted networks. Journal of Complex Networks. 2015;3(2):221–248. doi:10.1093/comnet/cnu026. 29

[59] Tecuatl C, Wheeler DW, Ascoli GA. A method for estimating the potential synaptic connections between axons and dendrites from 2D neuronal images. Bio-protocol. 2021;11(13):e4073–e4073. doi:10.21769/BioProtoc.4073. 29