**Neuronal contact predicts connectivity in the *C. elegans* brain**

Steven J. Cook1, Cristine A. Kalinski 1, Oliver Hobert 1

Department of Biological Sciences, Columbia University;

Howard Hughes Medical Institute, New York, NY

Correspondence:

Steven J. Cook (steven.cook@alleninstitute.org), Oliver Hobert (or38@columbia.edu)

**Abstract:**

How neurons choose appropriate synaptic targets among numerous potential partners remains a poorly understood problem. Here, we leverage the richness of a growing set of *C. elegans* wiring diagrams. We find that pairs of contacting neurons are more conserved across individuals than synaptic connections, with a ‘core’ set of adjacencies and connections present across individuals. We used a machine learning approach to generate a simplistic model (called Nematode Connectivity Classification Model or NCCM) to better understand the relationship between neuronal contact and connectivity. NCCM is highly accurate at predicting connectivity between neurons using aggregate neuronal adjacency and strata information. Our model also showed high performance when predicting connectivity within the pharyngeal nervous system. Encouraged by the information encoded within neuronal contacts, we lastly recontextualize how sexualized the wiring diagram is in the context of neuronal adjacencies and connectivity. The prevalence of sex-specific connections is commensurate with inter-individual variability and is rarely the product of sex-specific neuronal adjacencies. Our results suggest relative differences in neuronal adjacency may instruct synaptic connections between neurons.

**Introduction:**

How neurons choose appropriate synaptic targets among numerous potential partners remains a fundamental step in creating a nervous system. Multiple distinct mechanisms have been proposed to contribute to defining synaptic partnerships. In a ‘lock-and-key' mechanism, a neuron selectively chooses a synaptic partner among several different, adjacent target cells, based on a specific molecular recognition code[(Agi et al., 2020)](https://paperpile.com/c/8BrfnR/Yq5n). Alternatively, Peters’ rule posits that neurons indiscriminately form connections with other neuron types in their proximity; hence, neighborhood choice, dictated by initial neuronal process outgrowth and position, is the sole predictor for connectivity[(Braitenberg and Schüz 2013; Peters and Feldman 1976)](https://paperpile.com/c/8BrfnR/73v7+Co5L). Whether Peters’ rule plays an important role in synaptic wiring remains unresolved, and different invocations of the rule have been applied to understand synaptic organization[(Rees et al. 2017)](https://paperpile.com/c/8BrfnR/07A9). Here, we suggest that Peters’ rule, or the contact between two neuron pairs predicts connectivity, provides the main organizational principle of synaptic targeting choice in the brain of the nematode *C. elegans.*

The *C. elegans* is an appropriate model to study synaptic choice given its compact, largely invariant, and exquisitely mapped NS. Early ssEM analyses of *C. elegans* ultrastructure focused on finding stereotypy across individuals[(Albertson et al., 1976; Ward et al., 1975; White et al., 1976)](https://paperpile.com/c/8BrfnR/bcv2+suKi+pZEx), while recent advances in generating and analyzing nematode wiring diagrams have provided the opportunity to re-evaluate connectomic constancy (Fig 1A). This constancy can be leveraged by making a direct comparison of neuronal identity across sex[(Cook et al. 2019)](https://paperpile.com/c/8BrfnR/exvR), development[(Witvliet et al. 2021)](https://paperpile.com/c/8BrfnR/3rGA), and evolutionary time[(Bumbarger et al. 2013; Hong et al. 2019)](https://paperpile.com/c/8BrfnR/5DdP+AFti).

The brain of *C. elegans* is its nerve ring (NR), a bundle of axons where the majority of its neurons and synapses are localized(Fig 1B,C,D). Nearly all neurons within the NR exhibit axonal characteristics and make *en passant* synapses with a subset of their neighbors. The *C. elegans* nerve ring is not laminated, layered, or physically separated in such a way to prevent any two neurons from contacting and subsequently making synaptic connections (Fig 1C). Most neurons projecting into the NR do, however, computationally cluster into stratified groups with similar functions[(Brittin et al., 2021; Moyle et al., 2021)](https://paperpile.com/c/8BrfnR/2fpo+yc3u).

The question of wiring specificity [(White et al., 1983)](https://paperpile.com/c/8BrfnR/U9Bf) of the *C. elegans* connectome predates its formal publication[(White et al., 1986)](https://paperpile.com/c/8BrfnR/C00S). Through an approximation of adjacent neuronal processes, White estimated that neurons make synaptic connections with 45% of their adjacent neighbors[(White et al., 1983)](https://paperpile.com/c/8BrfnR/U9Bf). While this suggests that selective fasciculation plays a large role in forming neuronal neighborhood structure, some neurons break this trend and can abruptly switch their neighbors and synaptic partners[(White et al. 1986; Brittin et al. 2021; Moyle et al. 2021)](https://paperpile.com/c/8BrfnR/C00S+2fpo+yc3u).

Here, we directly test an invocation of Peters’ rule in the context of an entire brain (Fig 1E). We first describe the properties of neuronal adjacencies and connectivity across individuals, and then apply a machine learning approach to accurately classify whether neurons connect based on their relative adjacencies and strata locations. We then apply this model to the pharyngeal NS, where our model also achieves high accuracy. Lastly, we use the information encoded within adjacency to re-evalaute sexual dimorphism of the NS. Our study provides new evidence for the importance of neuronal adjacency to constrain synaptic specificity.

**Results:**

*C. elegans* is a gold standard for connectomics research. The compact nervous system of *C. elegans* was the first to be mapped by ssEM, and recent works have established it as the currently best model to study inter-individual differences in nanoscale neuronal wiring. To capture as many nanoscale wiring configurations as possible, we used recently published wiring diagrams of the *C. elegans* brain[(Brittin et al. 2021; Witvliet et al. 2021)](https://paperpile.com/c/8BrfnR/2fpo+3rGA). These 10 datasets represent the largest and most complete sampling of any brain’s wiring diagram to date (Supplemental data 1). To create a composite dataset which accurately compares differences within and across individual EM samples, we made several assumptions. We evaluated only embryonically-born and sex-shared (i.e. present in both sexes) neurons. The majority of *C. elegans* neurons are left-right symmetric, and we therefore grouped and aggregated data by neuronal class (nodes). Adjacencies and chemical synaptic connections (edges) were treated as undirected edges unless noted otherwise. To adjust for size differences across individuals, we scaled individual adjacencies within each dataset. As a proxy for synaptic strength, we counted the total number of individual synaptic connections between neurons. Gap junctions are excluded from our analysis due to their ambiguous ultrastructural characteristics. Our combined dataset of 2683 adjacencies and 1404 connections represents our goal of making unbiased comparisons between wiring diagrams.

Despite its compact and conserved structure, the *C. elegans* nervous system is highly interconnected. Calculating from our aggregate dataset, neurons in the *C. elegans* brain have a high probability of coming into contact with each other. By neuronal class, the median contact of a neuron is 79.5% (range 45.8% to 100%) of all other neurons (Fig 2A). Neurons are more selective with their choice of synaptic partners, making connections with an average of 38.6% (range 13.3% to 68.7%) of all other neurons (Fig 2A). We next calculated the filling fraction, or proportion of adjacencies that also have a synaptic connection for each sample. The aggregate filling fraction was 0.523, while the mean of individual filling fractions was 0.303 (Fig 2B). These data update previous estimates of filling fraction(cite white 1983 and witvliet), suggesting the *C. elegans* has a wider range of possible neuronal adjacency and connectivity configurations than previously reported.

How repeatable are adjacencies and connectivity across individual samples? We calculated a conservation score for each connection and adjacency in our dataset, representing the number of samples where that edge is present. The most frequently observed connectivity conservation level is 0, where two neurons are adjacent but asynaptic (Fig 2C). The second-most observed conservation in 1, which indicates our sensitivity to ‘rare’ synaptic connections. However, the third most common connectivity conservation level was 10, with 14.52% of all connections observed. The distribution of adjacency is heavily skewed, where the most common observed adjacency is present in every sample (Fig 2D). These distributions have two implications 1) adjacency is more conserved than connectivity 2) a previously predicted (Brittin et al., 2021) ‘core-circuit’ is likely present, albeit embedded within variability.

Because both neuronal adjacency and connectivity are variable we asked whether these two metrics share any commonalities. We plotted both adjacency scores and found that while adjacency conservation can set an upper bound for connectivity, it does not strongly relate to connectivity conservation. Only completely conserved adjacencies were larger, on average, than any other conservation level (Fig 2E)(DO THE ANOVA HERE). Next, we plotted our weight metric for synaptic connections, and similarly observed that the only statistically significant difference in synaptic size occurs when an adjacency is perfectly conserved (Fig 2F)(DO THE ANOVA HERE). The influence of completely conserved adjacency on edge size is further obvious when plotted as a cumulative density function of mean scaled adjacency (Fig 2G) or median connection weight (Fig 2H).

To inform our modeling approach, we next investigated the direct relationship of each edge in or dataset. By plotting mean scaled adjacency against median connection weight, we observe a moderate linear correlation (pearson’s r=0.682). However, given the abundance of asynaptic adjacencies (Fig 2C), we were motivated to investigate how the edge weights of our individual samples correlate. In addition to comparing across samples, we used the strata classification as defined by (Brittin et al., 2021) and (Moyle et al., 2021) and binarized whether the two adjacent neuronal classes were in the same or different strata. The scaled adjacency weights found in each individual sample were strongly correlated with each other and weakly correlated with the strata binary (Figure 3B). Similarly, the scaled connection weights found in each individual sample were strongly correlated with each other and weakly correlated with strata (Figure 3B). The consistent relationship between our samples suggested the appropriateness of a model to understand the relationship between adjacency and connectivity.

We approached our problem of inferring connectivity from adjacency as a machine learning based classification problem. Our model includes three predictors and one target variable. For full model selection criteria, please see methods. In brief, the predictors are mean scaled adjacency, strata classification (same) and strata classification (different)(cite brittin). The target variable was a binarized connectivity value of median connection weight > 0 vs median connection weight = 0. We adjusted for imbalance in our classifier, and reported multiple performance metrics for each model. After splitting the data into a training and test set, we cross-validated the performance of multiple classifiers on our training set (Table 1). Although the classifiers performed similarly, the logistic regression classifier was chosen for its performance, interpretability, and simplicity(Figure 3D,E). We named this model Nematode Connectivity Classification Model (NCCM) and refer to it by this name hereafter. Because of our imbalanced synaptic classifier, we do not rely on accuracy alone to assess our model. Our final logistic model on the test set generated ROC-AUC (Ability to discriminate between positive class and negative class), PR-AUC (precision score per recall threshold), and weighted F1 (harmonic mean of precision and recall) scores of 0.940, 0.76, and 0.88 respectively (see Methods) (Figure 3F,G). This suggests that NCCM is both precise and robust at predicting connectivity from adjacency and brain strata information. Intuitively, a logistic regression model approximates a threshold of physical contact which must be met to yield a synapse between neurons.

Encouraged by our NCCM results, we tested its ability to predict connectivity from adjacency in a different region of the *C. elegans* NS. Separated from its brain, *C. elegans* has a small, self-contained pharyngeal nervous system which directly connects to its brain through a single neuron (Fig 4A). The pharyngeal NS is composed of 20 neurons that fall into 14 classes, making several hundred synaptic connections. Previous work suggests the pharyngeal NS is more promiscuous than its somatic counterpart – its filling fraction is larger in both *C. elegans* and the related nematode *P. pacificus*. We applied NCCM to the densest region of the pharyngeal NS, treating the pharyngeal nerve ring as an unstratified structure (Supplemental dataset 2). Similar to the brain region, we also found a moderately strong correlation between our adjacency and connectivity weights in the pharynx (Fig 4B). The performance criteria for our pharyngeal NCCM were strong, with ROC-AUC, PR-AUC, and F1 scores of 0.762, 0.881, and 0.58, respectively. These results suggest that NCCM was more precise but less robust when applied to the pharynx compared to the NR. Together, our results suggest NCCM is robust to different brain regions, and different regions of the *C. elegans* NS may share fundamental rules of synaptogenesis.

If NCCM is an appropriate model, this implies that biological differences in connectivity, such as sexually dimorphic connectivity, could be attributed to relative neuronal adjacency between sexes. *C. elegans* has sexually dimorphic circuitry arising from both sex-specific and sex-shared neurons. Nearly every NR neuron is sex-shared, with the exceptions of HSN in the adult hermaphrodite, and CEM, EF, and MCM in the male. Previous work to classify sex-specificity of sex-shared synaptic connections relied on a statistical analysis to determine whether a connection’s strength was outside of expected variability[(Cook et al., 2019)](https://paperpile.com/c/8BrfnR/exvR). Although limited by no available contactome data for the male, we were compelled to re-evaluate previous classifications of sexual dimorphism in *C. elegans* within the context of hermaphrodite adjacency data (Supplemental datasets 3 and 4). To contextualize sexually dimorphic connectivity, we compared wiring of the sexes through the lens of individuality, i.e. connections found in only one sample but in no others. We compared connectivity both as a directed and undirected graph, and found that on average XXX%(Range XXX%-XXX%) of connections were unique to a specific sample (Fig 5A). The male sample contains the second-highest number of individual connections, consistent with a trend of older animals containing both a greater number of and a greater proportion of unique connections(cite witvliet). We next looked at individuality from the perspective of sex and classified synapes as being 1) present in both sexes 2) never found in the male sample (Herm. Only) or 3) only found in the male sample. We found that the majority of connections were found only in the hermaphrodite sample, followed by those in both sexes, and a small percentage found only in the male (Fig 5B). Our power to determine ‘male only’ synapses is much greater than ‘hermaphrodite’ only due to unbalanced sample sizes. We also noted a consistent trend when comparing directed edges to undirected edges such that there was no proportional increase in sex-specific connections. These proportions of sex-specific synapses are much lower than previously reported, suggesting that individuality clouds inference regarding classification of sex-specific synapses.

Having re-classified sex-specific connectivity, we next asked whether the contactome predicts sexual dimorphism. We stratified undirected connections by their sex-specific classification and plotted this against their mean scaled adjacency weight. We found that connections found in both sexes have larger mean scaled adjacency, on average (Fig 5C)(ANOVA STATS). We also found that sex-specific connections were less likely to be between neurons with strong adjacencies. Only XXX male-specific connections were never found to be adjacent in the hermaphrodite, suggesting 1) contact alone is insufficient to yield sex-dimorphism 2) we likely have a near-complete sample of possible contactomic configurations in our dataset.

Our dataset contains a developmental timeline, allowing us to determine the temporality of sexually dimorphic connectivity. Previous work classified hermaphrodite synapses by their strength, establishment, and maintenance, allowing us to ask whether hermaphrodite-specific and sex-shared connections arise early or late. We found that hermaphrodite-specific connections were most likely to be born post-embryonically and are variable in nature. The stable connections had the highest proportion of hermaphrodite-specific connections, suggesting our dataset captures the ‘core-circuit’ of hermaphrodite connectivity well. Finally, we compared our classifications to that of Cook et al. 2019. We found that the hermaphrodite-only classification did not yield a robust prediction, with approximately half of the connections found in males. There are very few male only connections, and some of the connections that didn’t pass a significance threshold for sexual dimorphism previously could be easily re-classified by this study. Together, our individuality and sexual dimorphism evaluation shows the contactome is capable of instructing the core-circuit of both sexes, with sex-specific synapses represented by smaller adjacnecies, on average.

**Discussion:**

Here, we provide evidence for how neuronal adjacency informs connectivity. Despite variability in the set of adjacencies and connections between individual *C. elegans*, when aggregated we were able to discern constancies and model synaptic specificity. Our NCCM is simplistic and mirrors Peters’ rule: a neuron is directed to specific brain strata after which a connection is formed if a threshold of physical contact is met. If our model is generalizable, biological differences such as sexual dimorphism could be approximated by neuronal adjacency alone. Despite extensive genetic screens searching for molecules that confer synaptic specificity, few molecules have been described. Contrastingly, the pursuit of axon guidance and fasciculation molecules has been more productive. Our results may suggest that the roles of synaptic specificity in *C. elegans* may be secondary to neuronal adjacency or possibly redundant with these processes.

While we are confident in our ability to describe the differences between individual EM samples, the underlying causes of variability are less clear. While neuronal adjacencies are computationally inferred, nematode synaptic connectivity can be uniquely scored by different trained individuals[(Cook et al., 2019; Xu et al., 2013)](https://paperpile.com/c/8BrfnR/Sng9+exvR). The difficulty in discerning synaptic features is both biological and technical. Ambiguity is introduced by inconsistent electron-dense features in *C. elegans* chemical postsynapses, which is further compounded by sample preparation technique differences. These issues are even more apparent when scoring gap junctions. Computational methods that describe the topology of the nervous system can also yield similar, yet non-identical results [(Brittin et al., 2021; Moyle et al., 2021)](https://paperpile.com/c/8BrfnR/2fpo+yc3u). Conclusions drawn from such analyses should neither be viewed as deterministic nor subjective, but rather as an encouragement to generate new wiring diagrams to observe which neuroanatomical features are most conserved.

In *C. elegans,* synaptic target choice appears to be driven by relative axodendritic placement and, hence, initial axodendritic outgrowth and selective fasciculation are the primary determinants of synaptic connectivity. We speculate that promiscuous synaptic connectivity asserted by Peters’ rule represents an ancestral mechanism of nervous system wiring. Increases in brain complexity, manifested by increases in the number of neurons, distinct neuron types, and potential adhesive molecular codes, may yield additional mechanisms for making selective choices among potential partners.



**Figure 1. Structure of the *C. elegans* nervous system and models of synaptic specificity.** A) Graphical depiction of animals sampled for this study. Animals from Cook et al 2019 are in red, Brittin et al 2020 in Brown, and Witvliet et al 2021 in Blue. B) Cartoon of the *C. elegans* anterior, with the region of our study presented in a dashed-line box, the nerve ring is multi-colored, neuronal cell bodies are small shaded ovals, and the pharynx is in dark gray. C) 3D reconstruction of pharyngeal isthmus and three nerve ring strata D) Example electron micrograph showing a dense reconstruction of the nematode nerve ring. E) Simple depiction of how synaptic specificity may be achieved between adjacent neurons. Either neurons may selectively choose among adjacent partners, or indisciminitably synapse with all adjacent partners.

****

**Figure 2. Presence and sources of variability in the *C. elegans* brain**. A) Boxplots of percent adjacency and connections by neuron class. B) Dot plot of filling fraction, or percent of adjacencies that result in a connection, of each EM series evaluated with mean and aggregate in red and black, respectively. C) Histogram of connectivity conservation across EM series. 14.52% of undirected connections are found in each EM series. D) Histogram of adjacency conservation across EM series. 40.92% of adjacencies are found in each EM series. E) Boxplot of adjacency conservation vs synaptic conservation. F) Boxplot of adjacency conservation vs median connection weight. G) Cumulative distributions of mean scaled adjacency vs proportion of adjacencies, colored by conservation across EM series. H) Cumulative distributions of median connectivity weight vs proportion of connections, colored by conservation across EM series.



**Figure 3. Modeling adjacency as a predictor of connectivity**. A) Blue scatter plot of mean scaled adjacency vs median connection weight with linear regression line shown in red, pearson’s r=0.682. B) Heatmap of correlations between scaled adjacency values of all samples and the binarized strata value. C) Heatmap of correlations between scaled connection weights of all samples and the binarized strata value. D) ROC curves of cross-validated model (kfolds=10), black line is the mean ROC-AUC=0.93 +/ 0.02. E) PR curves of cross-validated model (kfolds=10), black line is the mean PR-AUC=0.80+/0.06. F) ROC curve of final model performance, ROC-AUC=0.95 G) PR curve of final model performance, PR-AUC=0.764.

****

**Figure 4. NCPM performance extends to the pharyngeal nervous system.** A) (top) Micrograph of all neurons in the *C. elegans* anterior with the pharyngeal NR outlined by a red dashed box (box) subway line depiction of the pharyngeal nervous system with the pharyngeal NR highlighted in a black dashed box B) Blue scatter plot of mean scaled adjacency vs median connection weight with linear regression line shown in red, pearson’s r= 0.734. C) ROC curve of NCPM performance, ROC-AUC=0.762 D) PR curve of NCPM performance, PR-AUC=0.881.



**Figure 5. Individuality contextualizes sexualization of the connectome.** A) Dot plot of percent individuality, defined as a connection present in only a single dataset. B) Stacked barplot of the number of connections present in the undirected and directed datasets. Connections found in both sexes, hermaphrodites only, or male only are in red, blue, and orange, respectively C) Boxplot of mean scaled adjacency for connections found in both sexes (red), hermaphrodites only (blue), and males only (orange). D) Stacked barplot of the proportion of Witvliet et al. classified connections found in both sexes (red) and in hermaphrodites only (blue). E) Barplot of the proportion of Cook et al. classified connections found in both sexes (red), in hermaphrodites only (blue), and in males only (orange).

| Metric | Multi-layer Perceptron | Random Forest | XGBoost | Decision Tree | Logistic Regression |
| --- | --- | --- | --- | --- | --- |
| ROC\_AUC | 0.937600475 | 0.904718849 | 0.915540625 | 0.923959412 | 0.93730694 |
| PR\_AUC | 0.791002832 | 0.698312366 | 0.653441347 | 0.703207654 | 0.79090089 |
| F1 | 0.706096694 | 0.677465615 | 0.537985882 | 0.709319337 | 0.730452724 |

**Table 1.** **Performance of multiple ML models**. ROC\_AUC, PR\_AUC, and F1 are reported for all models tested.

**Supplemental data 1: NR modeling*.*** Aggregate dataset of adjacency and connectivity for the hermaphrodite datasets.

**Supplemental data 2: Pharyngeal modeling*.*** Aggregate dataset of adjacency and connectivity for the pharyngeal datasets.

**Supplemental data 3: Undirected male modeling*.*** Aggregate undirected edge dataset of adjacency and connectivity for all datasets, including the male.

**Supplemental data 4: directed male modeling*.*** Aggregate directed edge dataset of connectivity for all datasets, including the male.

**ACKNOWLEDGEMENTS:**

We thank David Miller III, John G. White, Eviatar Yemini, and Kang Shen for the discussion of this manuscript. This work was funded by the Howard Hughes Medical Institute.

**AUTHOR CONTRIBUTIONS**:

SJC – conceptualization, methodology, investigation, formal analysis, visualization, writing

CAK – investigation, formal analysis, visualization, writing

OH – writing, funding acquisition

**DECLARATION OF INTERESTS**:

The authors declare no conflicts of interest.

**Methods:**

***Data sources***

Nerve ring connectivity data are derived from[(Cook et al., 2019)](https://paperpile.com/c/8BrfnR/exvR), and [(Witvliet et al., 2021)](https://paperpile.com/c/8BrfnR/3rGA). Nerve ring adjacency data for nerve ring data are derived from [(Brittin et al., 2021)](https://paperpile.com/c/8BrfnR/2fpo) and [(Witvliet et al., 2021)](https://paperpile.com/c/8BrfnR/3rGA). Neuronal strata classifications are derived from [(Brittin et al., 2021)](https://paperpile.com/c/8BrfnR/2fpo) and [(Moyle et al., 2021)](https://paperpile.com/c/8BrfnR/yc3u). Pharyngeal connectivity and adjacency data are from [(Cook et al., 2020)](https://paperpile.com/c/8BrfnR/76Er).

***Code availability***

For peer review purposes, all code used to perform this analysis and generate figures can be found a<https://drive.google.com/drive/folders/1GHmW5LH0SU7N9x03XniAVPElcAG7b_ZC?usp=sharing>. Upon publication, code will be publicly shared on the Hobert Lab’s github page.

***Experiment Methodology***

Implementation of data preprocessing methods and classification algorithms were performed using Python, including the scikit-learn machine learning library[(“Scikit-Learn: Machine Learning in Python,” n.d.)](https://paperpile.com/c/8BrfnR/N8Hw), imbalanced-learn library[(Lemaître et al., 2017)](https://paperpile.com/c/8BrfnR/wh7p), and the XGBoostgradient boosting library[(Chen et al., 2015)](https://paperpile.com/c/8BrfnR/eEVF).

***Data Preprocessing***

We limited our analysis to adjacencies and chemical synapses between neuronal classes (as compared to individual neurons) that innervate the NR embryonically. Adjacency edge weights are undirected sums of individually adjacent membrane profiles[(Brittin et al., 2021)](https://paperpile.com/c/8BrfnR/2fpo). Connection edge weights are undirected reciprocal sums of connectivity. To account for the differences in adjacency sizes and scoring across samples, we scaled the 9 samples with an available volumetric reconstruction individually using a standard scaler which removes the mean and scales to unit variance. We then created a new feature, ‘ave\_scaled\_adjacency’ by taking the mean of the scaled adjacencies across these 9 series. This permits future implementation of our model irrespective of volumetric reconstruction sample size. For each neuron we used the strata classification as defined by[(Brittin et al., 2021)](https://paperpile.com/c/8BrfnR/2fpo) and[(Moyle et al., 2021)](https://paperpile.com/c/8BrfnR/yc3u). We created boolean variables to assign whether each row had neurons in the same or different strata (brittin/moyle\_bool\_1 and brittin/moyle\_bool\_2). We generated our target variable (dummy\_size) by taking the median of the connection weights across all 10 samples, and subsequently binarizing this value (>0 connection weight = 1, 0 connection weight = 0).

***Experimental Setup***

The dataset used in our model (NR-modeling.csv) uses four features with ‘ave\_scaled\_adjacency’, ‘brittin\_bool\_1’, and ‘brittin\_bool\_2’ used as predictors and ‘dummy\_size’ used as the target variable. The data were split randomly into either the training set or test set, where 75% of the instances were placed in the training set, and 25% were placed in the test set. The training set was further split into a second training set and a validation set, where 75% of the instances were placed in the training set, and 25% were placed in the validation set. To address the class imbalance in our data, we used Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTENC) on our training set.

***Classification Algorithms***

The algorithms we evaluated were MLPClassifier, DecisionTreeClassifier, RandomForestClassifier, and LogisticRegression from the scikit-learn library, as well as the XGBClassifier from the XGBoost library[(Chen et al., 2015)](https://paperpile.com/c/8BrfnR/eEVF).

***Classifier Evaluation***

To evaluate the performance of the classifiers, we used a stratified k-fold cross validation (k=10) and took the mean of a variety of metrics including area under the curve (AUC) the receiver operating characteristic (ROC) curve (ROC-AUC), AUC of the Precision-Recall (PR) Curve (PR-AUC), and F1 score. We used these metrics because our target variable was imbalanced leading us to be more concerned about false positives. We tested performance of the classifiers using strata classifications defined by either [(Brittin et al., 2021)](https://paperpile.com/c/8BrfnR/2fpo) or [(Moyle et al., 2021)](https://paperpile.com/c/8BrfnR/yc3u) to see which resulted in a feature of higher relative importance.

***Model Selection and Performance***

Because the classifiers performed similarly across the three metrics, we chose to use logistic regression for our model due to its interpretability. We used a grid search to optimize the parameters of the logistic regression model. The Brittin et al. 2021 classifications were used due to their higher relative importance in the classifiers [(Brittin et al., 2021)](https://paperpile.com/c/8BrfnR/2fpo). We used the training set from the original training and test split, performed SMOTENC, and fit the model.

***Analysis of pharyngeal nervous system***

We further tested our model using an identical training set but with the pharyngeal dataset as the test set. The pharyngeal dataset (pharynx\_modeling.csv) was created the same way as the NR dataset except that only two samples were used, and all neurons were automatically classified as being in the same strata.

[Agi, E., Kulkarni, A., & Hiesinger, P. R. (2020). Neuronal strategies for meeting the right partner during brain wiring. *Current Opinion in Neurobiology*, *63*, 1–8. https://doi.org/](http://paperpile.com/b/8BrfnR/Yq5n)[10.1016/j.conb.2020.01.002](http://dx.doi.org/10.1016/j.conb.2020.01.002)

[Albertson, D. G., Thompson, J. N., & Brenner, S. (1976). The pharynx of Caenorhabditis elegans. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *275*(938), 299–325. https://doi.org/](http://paperpile.com/b/8BrfnR/bcv2)[10.1098/rstb.1976.0085](http://dx.doi.org/10.1098/rstb.1976.0085)

[Brittin, C. A., Cook, S. J., Hall, D. H., Emmons, S. W., & Cohen, N. (2021). A multi-scale brain map derived from whole-brain volumetric reconstructions. *Nature*, *591*(7848), 105–110. https://doi.org/](http://paperpile.com/b/8BrfnR/2fpo)[10.1038/s41586-021-03284-x](http://dx.doi.org/10.1038/s41586-021-03284-x)

[Chen, He, Benesty, & Khotilovich. (2015). Xgboost: extreme gradient boosting. *R Package Version*.](http://paperpile.com/b/8BrfnR/eEVF) <https://cran.microsoft.com/snapshot/2017-12-11/web/packages/xgboost/vignettes/xgboost.pdf>

[Cook, S. J., Crouse, C. M., Yemini, E., Hall, D. H., Emmons, S. W., & Hobert, O. (2020). The connectome of the Caenorhabditis elegans pharynx. *The Journal of Comparative Neurology*, *528*(16), 2767–2784. https://doi.org/](http://paperpile.com/b/8BrfnR/76Er)[10.1002/cne.24932](http://dx.doi.org/10.1002/cne.24932)

[Cook, S. J., Jarrell, T. A., Brittin, C. A., Wang, Y., Bloniarz, A. E., Yakovlev, M. A., Nguyen, K. C. Q., Tang, L. T.-H., Bayer, E. A., Duerr, J. S., Bülow, H. E., Hobert, O., Hall, D. H., & Emmons, S. W. (2019). Whole-animal connectomes of both Caenorhabditis elegans sexes. *Nature*, *571*(7763), 63–71. https://doi.org/](http://paperpile.com/b/8BrfnR/exvR)[10.1038/s41586-019-1352-7](http://dx.doi.org/10.1038/s41586-019-1352-7)

[Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research: JMLR*, *18*(1), 559–563.](http://paperpile.com/b/8BrfnR/wh7p) <https://dl.acm.org/doi/10.5555/3122009.3122026>

[Moyle, M. W., Barnes, K. M., Kuchroo, M., Gonopolskiy, A., Duncan, L. H., Sengupta, T., Shao, L., Guo, M., Santella, A., Christensen, R., Kumar, A., Wu, Y., Moon, K. R., Wolf, G., Krishnaswamy, S., Bao, Z., Shroff, H., Mohler, W. A., & Colón-Ramos, D. A. (2021). Structural and developmental principles of neuropil assembly in C. elegans. *Nature*, *591*(7848), 99–104. https://doi.org/](http://paperpile.com/b/8BrfnR/yc3u)[10.1038/s41586-020-03169-5](http://dx.doi.org/10.1038/s41586-020-03169-5)

[Scikit-learn: Machine Learning in Python. (n.d.). *Https://www.jmlr.org › volume12 › pedregosa11ahttps://www.jmlr.org › volume12 › pedregosa11a*.](http://paperpile.com/b/8BrfnR/N8Hw) <https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>

[Ward, S., Thomson, N., White, J. G., & Brenner, S. (1975). Electron microscopical reconstruction of the anterior sensory anatomy of the nematode Caenorhabditis elegans. *The Journal of Comparative Neurology*, *160*(3), 313–337. https://doi.org/](http://paperpile.com/b/8BrfnR/pZEx)[10.1002/cne.901600305](http://dx.doi.org/10.1002/cne.901600305)

[White, J. G., Southgate, E., Thomson, J. N., & Brenner, S. (1976). The structure of the ventral nerve cord of Caenorhabditis elegans. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *275*(938), 327–348. https://doi.org/](http://paperpile.com/b/8BrfnR/suKi)[10.1098/rstb.1976.0086](http://dx.doi.org/10.1098/rstb.1976.0086)

[White, J. G., Southgate, E., Thomson, J. N., & Brenner, S. (1983). Factors that determine connectivity in the nervous system of Caenorhabditis elegans. *Cold Spring Harbor Symposia on Quantitative Biology*, *48 Pt 2*, 633–640. https://doi.org/](http://paperpile.com/b/8BrfnR/U9Bf)[10.1101/sqb.1983.048.01.067](http://dx.doi.org/10.1101/sqb.1983.048.01.067)

[White, J. G., Southgate, E., Thomson, J. N., & Brenner, S. (1986). The structure of the nervous system of the nematode Caenorhabditis elegans. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *314*(1165), 1–340. https://doi.org/](http://paperpile.com/b/8BrfnR/C00S)[10.1098/rstb.1986.0056](http://dx.doi.org/10.1098/rstb.1986.0056)

[Witvliet, D., Mulcahy, B., Mitchell, J. K., Meirovitch, Y., Berger, D. R., Wu, Y., Liu, Y., Koh, W. X., Parvathala, R., Holmyard, D., Schalek, R. L., Shavit, N., Chisholm, A. D., Lichtman, J. W., Samuel, A. D. T., & Zhen, M. (2021). Connectomes across development reveal principles of brain maturation. *Nature*, *596*(7871), 257–261. https://doi.org/](http://paperpile.com/b/8BrfnR/3rGA)[10.1038/s41586-021-03778-8](http://dx.doi.org/10.1038/s41586-021-03778-8)

[Xu, M., Jarrell, T. A., Wang, Y., Cook, S. J., Hall, D. H., & Emmons, S. W. (2013). Computer assisted assembly of connectomes from electron micrographs: application to Caenorhabditis elegans. *PloS One*, *8*(1), e54050. https://doi.org/](http://paperpile.com/b/8BrfnR/Sng9)[10.1371/journal.pone.0054050](http://dx.doi.org/10.1371/journal.pone.0054050)