Neuroinformatics (2021) 19:669–684
https://doi.org/10.1007/s12021-021-09513-y

ORIGINAL ARTICLE

Identification of Neuronal Polarity by Node-Based Machine Learning

Chen-Zhi Su[1,2] · Kuan-Ting Chou[1,3] · Hsuan-Pei Huang[4] · Chiau-Jou Li[1,3] · Ching-Che Charng[4] · Chung-Chuan Lo[1,4] · Daw-Wei Wang[2,3,5]

Accepted: 13 January 2021 / Published online: 5 March 2021

Abstract

Identifying the direction of signal flows in neural networks is important for understanding the intricate information dynamics of a living brain. Using a dataset of 213 projection neurons distributed in more than 15 neuropils of a *Drosophila* brain, we develop a powerful machine learning algorithm: node-based polarity identifier of neurons (NPIN). The proposed model is trained only by information specific to nodes, the branch points on the skeleton, and includes both Soma Features *(which contain spatial information from a given node to a soma)* and Local Features *(which contain morphological information of a given node)*. After including the spatial correlations between nodal polarities, our NPIN provided extremely high accuracy (>96.0%) for the classification of neuronal polarity, even for complex neurons with more than two dendrite/axon clusters. Finally, we further apply NPIN to classify the neuronal polarity of neurons in other species (Blowfly and Moth), which have much less neuronal data available. Our results demonstrate the potential of NPIN as a powerful tool to identify the neuronal polarity of insects and to map out the signal flows in the brain's neural networks if more training data become available in the future.

Keywords Neuronal polarity · Machine learning · Drosophila · Connectome · Axon · Dendrite

Introduction

Rapid technology advances in recent years have led to the development of several connectomic projects and large-scale databases for cellular-level neural images (Chiang et al. 2011; Kuan et al. 2015; Milyaev et al. 2012; Parekh and Ascoli 2013; Peng et al. 2015; Shinomiya et al. 2011; Xu et al., 2013; Xu et al. 2020). However, how to integrate and transform the data to address scientific questions (Lo and Chiang 2016) remains a central challenge. Overall, these projects aim to provide sufficient information for the analysis of information flows in the brain. This goal is difficult to achieve in the current stage, as many neural images do not provide information on polarity (axons and dendrites). The axon-dendrite polarity of a neuron can be identified by experimental methods (Craig and Banker 1994; Matus et al. 1981; Wang et al. 2004). However, these methods are not practical for large-scale neural image projects and for the image datasets that were already acquired. Morphology-based polarity identification at the post-imaging stage is possible, but this is particularly challenging for insects because of their highly diverse neuronal morphology (Cuntz et al. 2008; Lee et al. 2014).

To address this issue, the method of skeleton-based polarity identification of neurons (SPIN) has been developed using several classic machine-learning (ML) algorithms (Lee et al. 2014). Although SPIN reaches a decent performance in neuronal polarity identification for fruit flies, *Drosophila melanogaster*, with 84%–90% accuracy, the method suffers from the cluster-sorting

Chen-Zhi Su and Kuan-Ting Chou contributed equally to this work.

✉ Chung-Chuan Lo
cclo@mx.nthu.edu.tw

✉ Daw-Wei Wang
dwwang@phys.nthu.edu.tw

1 Brain Research Center, National Tsing Hua University, Hsinchu 30013, Taiwan

2 Physics Division, National Center for Theoretical Sciences, Hsinchu 30013, Taiwan

3 Department of Physics, National Tsing Hua University, Hsinchu 30013, Taiwan

4 Institute of Systems Neuroscience, National Tsing Hua University, Hsinchu 30013, Taiwan

5 Center for Quantum Technology, National Tsing Hua University, Hsinchu 30013, Taiwan

problem. Most projection neurons (i.e., neurons that innervate more than one neuropil) possess two or more clusters of neural processes. Each cluster can be either axon or dendrite, but not both. Using this observation, the SPIN method first identifies the clusters of processes in a neuron and then identifies the polarity of each cluster. The strategy is highly efficient, but incorrect sorting of clusters can lead to incorrect polarity classification of a large number of terminal points at once. This is a major source of errors in the SPIN method.

n the past decade, modern ML algorithms have been applied in many research fields and in daily life. The popularity of modern ML grows because of rapid developments in computational algorithms, high-speed processors, and big data available from various resources (LeCun et al. 1998; Krizhevsky et al. 2012; LeCun et al. 2015). Some widely successful algorithms —for example, deep neural networks (DNN) and extreme gradient boosting (XGB)— may recognize hidden patterns more efficiently than human knowledge/experience, after proper training on big data. Therefore, ML opens a new era when precise classification and/or prediction becomes possible even without full knowledge of the given data. As a result, many applications of ML have recently appeared in biological and medical research (Asri et al. 2016; Malta et al. 2018; Mohsen et al. 2018). It is reasonable to expect that one may apply modern ML for the identification of neuronal polarity solely using optical images of the fruit fly's brain. For neurons of this insect, several tenths of thousands of high-resolution optical images are already available, which is the largest dataset among all species.

In the present work, we develop a new classifier: node-based polarity identifier of neurons (NPIN). The proposed model achieves much higher accuracy (>96%) than SPIN or the human eye for the identification of neuronal polarity in the *Drosophila* brain. Our NPIN is developed using a node-based feature extraction method. Specifically, NPIN includes both Soma Features (spatial information between a soma and a given node) and Local Features (morphological information around a given node). Two state-of-the-art supervised learning algorithms—XGB and DNN—are used as two complementary classifiers, making the method applicable to complex neurons (which have more than two axon/dendrite clusters) with a competition between Soma Features and Local Features. We find that NPIN provides extremely good results for the classification of neuronal polarity, identifying important local features compared with the known soma features. We further apply NPIN to classify the neuronal polarity of other species of insects (in this case, Blowfly and Moth), which may have insufficient data for standard ML. These important achievements of NPIN are all important steps toward the understanding of signal flow dynamics in neural networks, and should speed up the connectomic projects for the whole brain when more data are available for training.

Method

Overview

The axon-dendrite polarity of a neuron is correlated with certain aspects of its morphology, such as the distance (or path length) from a terminal to the soma, the number of nodes involved in a domain/cluster, and the thickness of neurites (Craig and Banker 1994; Hanesch et al. 1989; Rolls 2011; Squire et al. 2008). However, so far, very few theoretical frameworks have systematically investigated the relationship between these features and neuronal polarity. These empirical conditions are loosely defined, with many exceptions for different types of neurons. Therefore, it is difficult to identify neuronal polarity by traditional rule-based computational programs. SPIN (Lee et al. 2014), which is developed using classical ML algorithms, can be improved in many aspects.

In order to significantly improve the previous methods, here we develop a new polarity identifier based on the morphologic features, which are extracted from neuronal nodes and handled by modern ML algorithms. Different from clusters, which are usually ill-defined from computational point of view, nodes are always well-defined by the bifurcation in a neuronal skeleton. The whole process of polarity identification, therefore, is composed by the following four major steps in our NPIN model. It is instructive to briefly describe them (see Fig. 1) before the further explanation in the rest of this paper:

Step I. **(Data Preparation and Reorganization)**: We invent a diagrammatic method to map a 3D neural skeleton structure of a given neuron onto 2D tree diagrams, called level trees and reduced trees. This effective representation makes it easy to extract representative features for ML.

Step II. **(Node-Based Feature Extraction)**: We determine the nodal polarity using the features of each node. Specifically, we identify and extract both Soma Features and Local Features for each node.

Step III. **(ML Models)**: In NPIN, we apply two powerful ML algorithms—XGB and DNN—together. They provide two different but complementary approaches for the classification of axons and dendrites.

Step IV. **(Implementation of Spatial Correlation)**: The spatial correlation of the nodal polarity in the nearby region is implemented by relabeling the nodal polarity suggested by ML models. This approach can significantly enhance the accuracy of the final output.

Typical ML methods concentrate on the algorithms in Step III. Instead, we put more emphasis on the other three steps in a

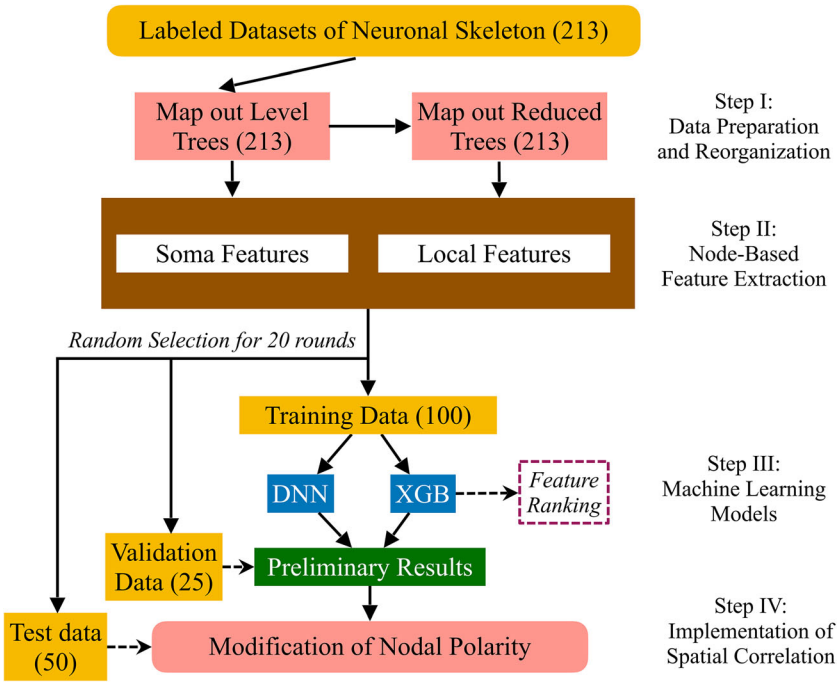

Fig. 1 Flowchart of the NPIN model. NPIN includes four major steps, as described in the text. The dataset contains 213 neurons with labeled polarity as the ground truth. We randomly choose 100/25/50 neurons from the datasets for training/validation/test sets. Every neuron in the training/validation sets is mapped to a level tree and a reduced tree. We then extract Soma Features and Local Features from these neuronal data for training. Preliminary results are obtained by XGB and DNN algorithms after validation. We then relabel the classification by including spatial correlations of nodal polarities before comparing them with the test data with known polarities. The whole process is repeated 20 times to cover all 213 neurons in the original dataset. As a result, each neuron could be selected to be a test sample and classified by a model trained on other neurons

way specifically useful for the determination of neuronal polarity. Figure 1 shows the flowchart of the whole calculations. We will explain these strategies in the rest of this section.

Dataset

Our main dataset represents 213 neurons with experimental ground truth from the *Drosophila* brain, which are available from the FlyCircuit database (http://www.flycircuit.tw/) (Chiang et al. 2011). These 213 neurons are ALL projection neurons selected from various regions across the brain to represent the diversity of neuronal morphology as much as possible (Fig. 1(a)). Local neurons with axon/dendrite coexistence in the same branch/cluster are not included in our research. These projection neurons innervate 15 neuropils: AL, AOTU, CAL, CCP, DMP, EB, FB, IDFP, LH, LOB, MED, NO, PB, VLP, and VMP. Among these 213 neurons, 107 neurons have been included in the dataset used in the development of the previous model, SPIN, and we have 106 additional neurons for the present work. As we will show later, due to the improvement of feature extraction and the ML algorithm, our model, NPIN, substantially outperforms SPIN, not only in the overall precision and recall but also in the applicability in more brain regions as well as more types of complex structures. In Appendix E, we list these 213 neurons with information including the brain regions innervated by the dendrites and axons of each neuron, the numbers of axon/dendrite terminals, and precision/recall obtained by our model.

We divide the neurons in our dataset into two types: (i) simple neurons, which have two clusters of terminals (one dendrite and one axon); (ii) complex neurons, which have more than two clusters of terminals. In Figs. 2(b1)–(b3) and (c1)–(c4), we show some typical skeleton structures of these two types of neurons. In our dataset, we have 89 simple neurons and 124 complex neurons with previously reported polarity. Among complex neurons, most complex neurons have three clusters (two dendrites and one axon, or one dendrite and two axons). Only a few neurons have more than three clusters. The reason to classify these neurons is to investigate how the distance to soma and the number of clusters can influence the identification of neuronal polarity. Moreover, we can examine how well NPIN performs even when the polarity is difficult to be identified by the human eyes in the case of three or more terminal clusters. This is one of the most important criteria for a polarity identifier to be practically applicable for the determination of signal flow in neuronal networks of the insect brain. There are, of course, some other types of projection or local neurons, which may not be easily classified by the number of clusters or by their polarity distribution. We do not

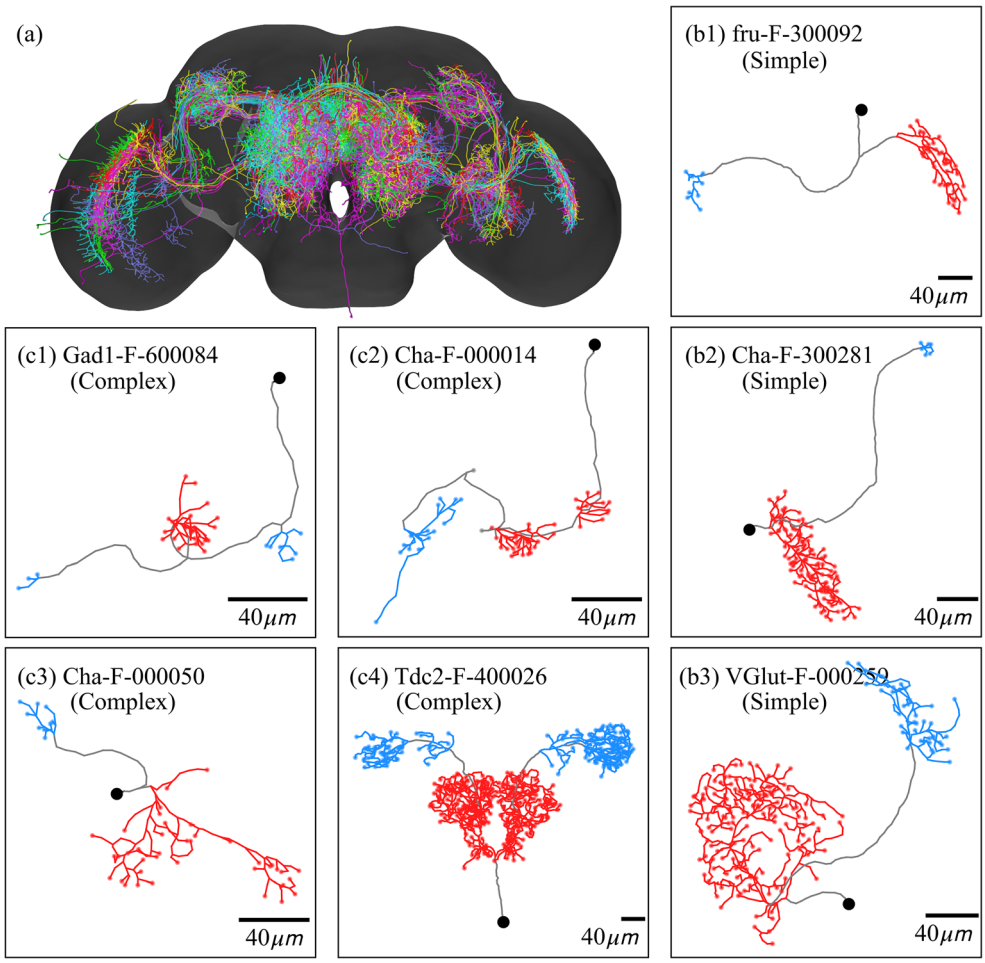

Fig. 2 ***Drosophila melanogaster*** **(fruit fly) neurons used in the present study.** (a) All 213 neurons in our dataset, shown in their actual locations in the standard fly brain. (b1)–(b3) Skeleton structures for several simple neurons. (c1)–(c4) Skeleton structures for several complex neurons. Black dots represent somas. Black lines are the main trunks of neurons. Blue or red lines indicate the axonal or dendritic clusters, respectively. Each neuron is labeled by its ID in the FlyCircuit database

include them in the dataset of this study because of a lack of data with confirmed polarity to be used for training. Our approach developed here, however, may still be applicable to these neurons when more data are available in the future.

Standardized Representation: Level Trees and Reduced Trees

To improve the accuracy of our ML model, we first need to define how to "standardize" the morphological information of these neurons, which are so different from each other in their original 3D structures. Figures 3(a1) and (b1) show two examples of a simple neuron and a complex neuron. First, we start with the 3D skeleton structures (see Figs. 3(a2) and (b2)) extracted from the raw images, where the width information of the trunks or branches are ignored temporarily in order to make our model more generally applicable. In our work, we further map the 3D skeleton structure onto a level tree (see Figs. 3(a3) and (b3)), which keeps all information on the position of each node (including soma, terminals, and cross points between branches) and the path length between them, but it ignores the trunk and branch information, such as width or shapes. To express this information in a 2D diagram, we introduce the level structure according to the generation of nodes: a soma is placed in the top-level (level 0), and the next two nodes are placed in the lower level (level 1), and so on for their offsprings, until all the ending nodes (terminals) are properly placed. We take the convention that the branches with more successive non-empty levels are placed in the left-hand side and the branches with less successive non-empty levels in the right-hand side (Figs. 3(a3) and (b3)). We believe that most morphological features of the neuronal cluster are still extractable from such standardized representation because the spatial positions of all nodes (including soma and terminals) are still available. The only missing information in the level tree (compared with the 3D skeleton image of neurons) is the shapes and widths of neuronal branches that connect neighboring nodes. As we will see below, this missing information seems not crucial for the determination of neuronal polarity in NPIN.

In addition to the level tree representation for a neuronal structure, in this work, we further define a reduced tree for each neuron. The reduced tree aims to retain the major branches of the skeleton structure to identify an axon or

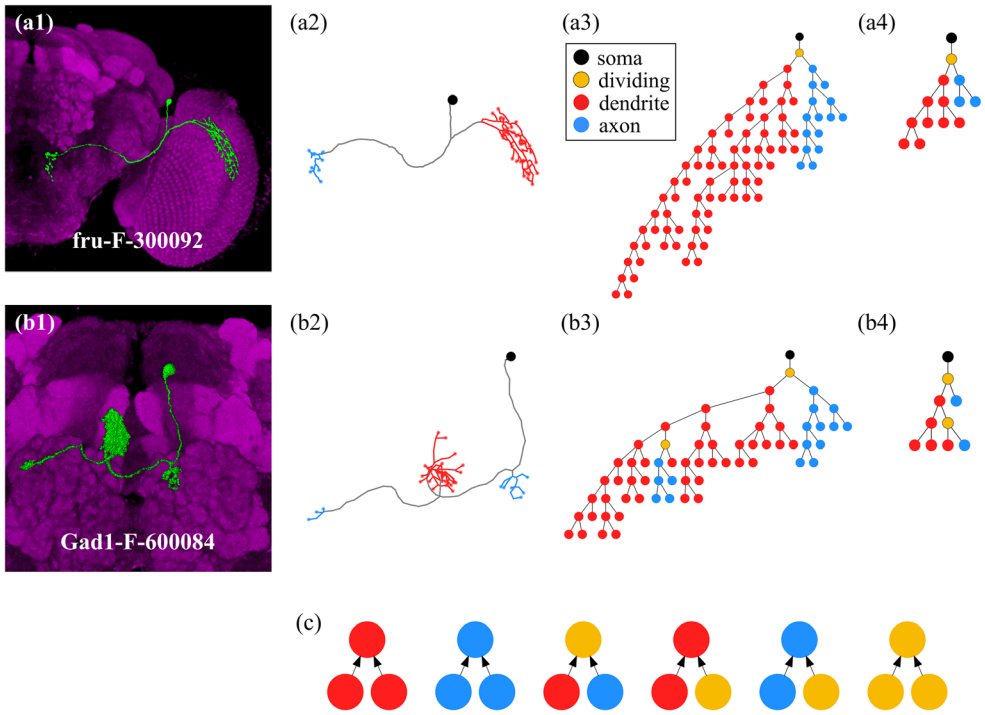

Fig. 3 Encoding 3D optical images of neurons into level trees and reduced trees. First, the volume image of a neuron (a1) is converted into the skeleton (a2), and then a level tree (a3), which is a 2D plot with a standardized method to label most features of the original neurons. Red, blue, and yellow dots represent dendrites, axons, and dividing nodes (including terminals), respectively. (a4) represents the reduced tree of the same neuron cell. (b1)–(b4) show the same reduction for a complex neuron. Because a complex neuron has more than two clusters, there can be more than one dividing node that separates axon clusters from dendrites. In (c), we graphically show the rules to define the nodal polarity based on the polarity of terminals in the level tree (see the text). Upward arrows indicate that the nodal polarity in the upper level is defined by the nodal polarities of the two nodes/terminals in the lower level

dendrite cluster. This information is important for the determination of cluster curvature and aspect ratio for nodal features within each cluster (explained below). The reduced tree of a neuron can be obtained by repeatedly removing the ending nodes with the branches shorter than a characteristic length determined by the branch distribution, until it stops automatically or only five levels are left (see Figs. 3(a4) and (b4)). The basic assumption behind this procedure is that the major branch of a neuron skeleton structure is contained in the "inner" (closer to the soma) and "longer" branches. Shorter and outsider branches are minor or unimportant for determining the clusters. See Appendix A for the detailed procedure of producing the reduced tree from a level tree.

Nodal Polarity

The polarity of the neurons in our dataset are all predetermined using the presynaptic (Syt::HA) or postsynaptic (Dscam17.1::GFP) markers (C.-Y. Lin et al. 2013) or using the morphological features described in previous studies (Fischbach and Dittrich 1989; Hanesch et al. 1989; Wu et al. 2016). There are 7142 terminals identified as dendrites and 2310 as axons. However, because the axon-dendrite polarity of these terminals is highly correlated to the morphological structure of their neurons, in this study, we extend the definition of polarity from terminals to nodes, and we use this information to extract features in NPIN. In other words, we use a bottom up method to assign the polarity for a node to be the axon/dendrite class, if its offspring branches are connected to pure axon/dendrite terminals or nodes. See below for more detail.

We emphasize that using features extracted from nodes has several important advantages over using features extracted from terminals or clusters for the training process of ML. First, the number of nodes is much larger than the number of clusters in each neuron. Therefore, the polarity identification has significantly higher accuracy due to the larger training data. Second, nodes are well-defined in the skeleton structure (compared with clusters) and could include more morphological features (compared with terminals). Finally, these nodes can also be systematically labeled in the skeleton structure or in our level tree diagram, making it easy to include their correlated features in the spatial distribution. This node-based feature extraction is crucial in NPIN, making an accurate identification of neuronal polarity possible.

To extend the polarity definition from terminals, as provided in the dataset, to nodes on the skeleton of a neuron, we apply the following series of rules to define the nodal polarity

according to the polarity of terminals (Fig. 3. (c)): (1) If two child nodes (or terminals) are both axons (or dendrites), their parent node (the node that directly connects to them in the upper level) is also defined as an axon (or dendrite). (2) If one of the child nodes (or terminals) is an axon, and the other is a dendrite, their parent node is defined as a "dividing node." (3) If one of the child nodes is an axon (or dendrite), and the other is a dividing node, their parent node is defined as an axon (or dendrite). Finally, (4) if two child nodes are both dividing nodes, their parent node is also defined as a dividing node (however, we do not have such a case in our dataset). The definition of diving nodes is just for the convenience and consistency of nodal polarity. These dividing nodes are very few (mostly none or at most two in each neuron of our dataset) and therefore not included in our training data. If not defining dividing points in such a way, we could not properly identify the polarity of a node connecting to both dendrite and axon nodes.

After applying these rules, we can label the polarity of all nodes of any neuron using the polarity information of their terminals. Note that this expansion of nodal polarity should not be misunderstood as introducing any artifacts or unconfirmed polarity labeling, because the morphological features of terminals should be directly related to the nearby nodes by definition. The introduced nodal polarity is just for the convenience of feature extraction and for data augmentation in machine learning language, and will not be shown in the evaluation of NPIN performance. In other words, the precision of NPIN is still calculated based on terminal polarity rather than nodal polarity, and will show (see below) a significant enhancement of prediction accuracy compared to the results using terminal information only. Finally, we note that the dividing node is defined to mark the position to separate axon and dendrite clusters, and it should be important in the nerve cell development. Since the number of dividing points is much less (one or at most two points in each neuron) than the number of axon or dendrite nodes, we do not include them in the training and testing processes. Figures 3(a3) and (b3) show some representative level trees, where all nodes are properly labeled.

Feature Extraction for Nodal Polarity

In principle, the level tree representation defined above contains all information of a 3D neuron and can be used for the identification of neuronal polarity. We test more than a dozen of features, including (1) path length to its parent node, (2) normalized path length to its parent node, (3) path length to soma, (4) normalized path length to soma, (5) direct distance to soma, (6) normalized direct distance to soma, (7) Strahler number, (8) angle between branches to the children nodes, (9) ratio of path lengths to the children nodes, (10) layer number in the cluster, (11) number of terminals in the clusters, (12) eigenvalues of moment of inertia of the cluster, (13) curvature (varicosities) of the cluster, (14) aspect ratio of the cluster, (15) volume of the cluster, etc. We do not include arbor thickness because not all neurons have such information in their optical images.

After systematic studies and comparison of the prediction results, we eventually find out the nine most relevant features, which can be classified into 2 groups: Soma Features (SF) and Local Features (LF). Soma Features contain spatial information from a given node to a soma, including the path length along the neuronal branches and the direct distance in 3D space. Local Features contain certain information on the local morphology of a given node, including the curvature and aspect ratio of the cluster it belongs to. Hence, Local Features do not include any information about the soma, while Soma Features do not include any information about the local morphology. Let i be the index of a given node. Soma Features of node i can be expressed as a four-component vector: $SF = [l_{si}, nl_{si}, d_{si}, nd_{si}]$, which are the path length to soma, normalized path length to soma, distance to soma and normalized distance to soma, respectively. Local Features of node i can be expressed as a five-component vector: $LF = [l_{pi}, nl_{pi}, c_i, ar_i, rl_i]$, which are the path length to the parent node, the normalized path length to the parent node, curvature of the cluster, aspect ratio of the cluster, and the ratio of path lengths to the children nodes, respectively. If a children node does not exist, its features are replaced by the number, −1. We then train different ML models on various combinations of features to identify their roles in the identification of neuronal polarity. In Appendix B, we explain how to identify and calculate soma features and local features (from the level trees and reduced trees defined above).

Machine Learning Models

We train our model by supervised learning using the training data extracted from the dataset. We implement several ML algorithms: random forest, gradient boosting decision tree, XGB, support vector machines, and DNN. We find that, in general, XGB and DNN provide the best and complementary results from the features we selected. Therefore, we use them in our NPIN. In Appendix C, we explain the details of how to implement these two algorithms in the present study.

In addition to the algorithms, an ML model also depends on the features used during the training process. To investigate the effects of different morphological features on the identification of nodal polarity, we develop three models by using three types of features in NPIN: Model I (using both Soma Features and Local Features), Model II (using Soma Features only), and Model III (using Local Features only). As we will see later, we can gain insight into the relationship between morphological features and polarity by systematically

comparing the polarity identification results between different models and different types of neurons.

Implementation of Spatial Correlation of Nodal Polarity

In the standard application of supervised learning for classification, one usually obtains the results from the output probabilities directly when the model is well-trained on the training data. The training aims to minimize the cross-entropy between the output results and the known answers by backpropagation. However, this ML process does not guarantee reasonable results all the time without violating some necessary conditions, which could not be included in the input features of training data. For the task of nodal polarity identification in our present work, for example, the polarities of nodes are highly dependent on its neighboring nodes: nodes in the same cluster (and, therefore, close in space) are usually of the same type (a dendrite or axon), but such loosely defined necessary condition cannot be implemented in the loss function if the polarity of each node is identified individually. Therefore, we have to include such a spatial correlation of polarity by adding other methods in the ML model.

In this work, spatial correlations between nodal polarities can be included by the modification of the polarity provided by XGB or DNN, if the probability for axon or dendrite is below a certain threshold. More precisely, such a modification process contains three steps: (1) we perform the ML process for the test data and obtain the polarity and its probability for each node. (2) Next, we accept the result of a given node if the probability is higher than a threshold, and we reject the result otherwise by changing it to be unidentified. (3) Finally, we relabel these rejected/unidentified nodes according to the polarity of its neighboring nodes. As a result, we identify spatial correlations between nodal polarities. More details of such polarity modification and its effects on the NPIN performance are described in Appendix D.

Results

Our dataset includes 213 neurons with verified polarities as the ground truth. In our training procedures (Fig. 1), we randomly select 100 neurons from the dataset for training, 25 for validation, and 50 for testing. This process is repeated for 20 rounds, so that each neuron can be tested (by different models trained by other neurons) for 4–5 times on average. We then average these probabilities for their nodal polarity and make the final comparison with the ground truth. Using this method, the obtained results for the nodal polarity of each neuron can be much more stable because the fluctuations due to the dataset selection are reduced. In our training data and in comparison with the ground truth, the dividing points are not included because their numbers are too few to be statistically relevant. In the testing neurons, they could be recovered using the predicted polarities of other nodes (see Appendix D).

In the following sections, we will first present the distribution of nodal features, including both soma features and local features, obtained from all neurons in our dataset. This provides a deep understanding of neuronal morphology and its relationship with other results. Next, we show the results of polarity identification provided by Model I (with both Soma Features and Local Features) for our whole neuron dataset, followed by results using Model II (with Soma Features only). We then focus on the results obtained by using complex neurons as training data for comparison. As an example of application in other species, we apply NPIN to test the blowfly. Finally, we summarize these calculation results and our findings.

Feature Distribution and Importance Ranking

Before presenting the results of neuronal polarity by NPIN, we investigate the distribution of different features (Soma Features or Local Features) for different types of neurons (simple neurons or complex neurons). This provides a better picture which helps to understand and explain the results of the present algorithm. In Fig. 4, we show the distribution of axon nodes and dendrite nodes (including terminals) of all neurons as a function of the normalized path length (relative to the largest length to the soma). Results of simple neurons (a1) and complex neurons (a2) are shown together for comparison. As expected, most axons have a longer path length to soma compared with most dendrites in simple neurons, but the distribution of dendrite is certainly wider than the distribution of axons. A wider distribution pattern for dendrites in simple neurons directly implies that it is easier to correctly classify a node to be a dendrite, while it is more difficult to include all dendrite nodes by the same classifier. Hence, this explains why the precision is higher (or lower) than the recall for dendrites (or axons) of simple neurons (Fig. 5(a1) and (b1)). On the other hand, in Fig. 4(a2), axon nodes have a wider distribution than dendrite nodes in complex neurons, explaining why the precision is lower (or higher) than the recall for dendrites (or axons) of complex neurons (see Fig. 5(a2) and (b2)).

In addition to the path length to the soma, we have also included the direct distance from a node to a soma as a feature (Appendix B and Fig. S2(b)). Besides, the ratio of direct distance to the path length reflects a global morphological feature of a given node: if the distance to a soma is close to the path length to a soma, the neuron branches are more straight in the real space. The path is more curved if this ratio is much smaller than one. This implied that the node is close to the soma in space with a long and curved neuronal branch in between. In Figs. 4(b1) and (b2), we show the distribution of axon and dendrite nodes in the space of normalized length to the soma

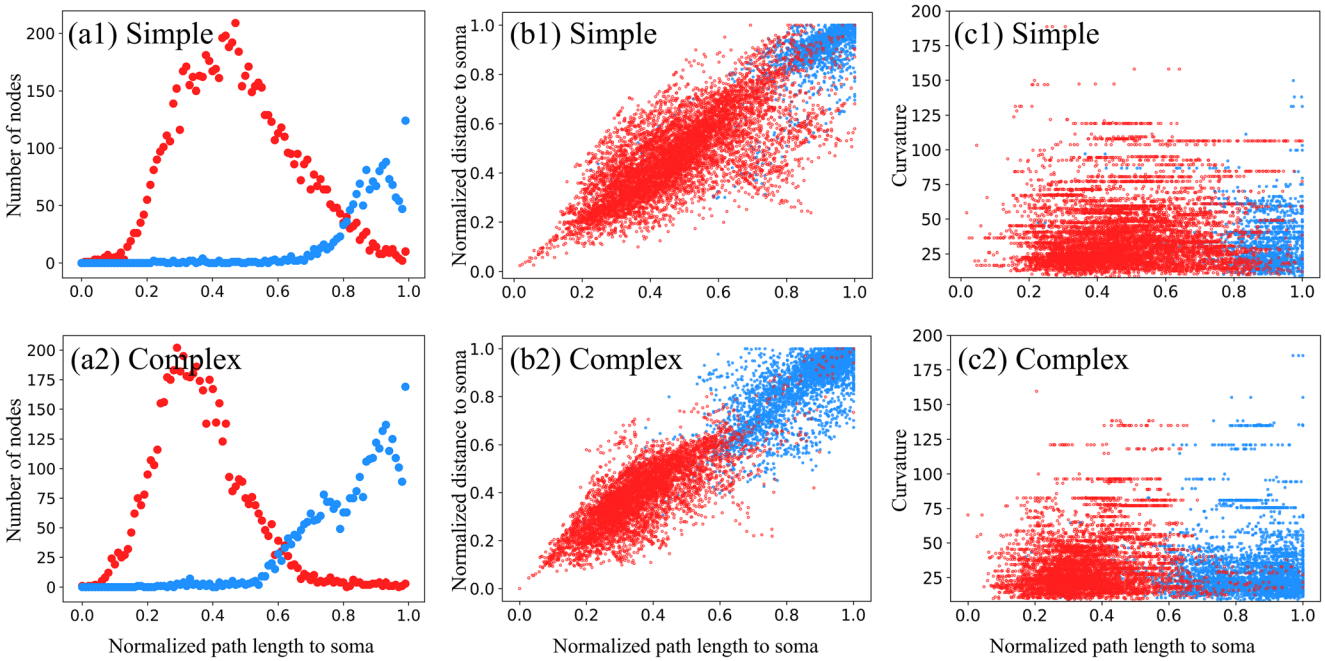

Fig. 4 Feature distributions of axons and dendrites for all neurons in our dataset. (a1) and (a2) show the distribution of axon and dendrite nodes along the normalized path length to soma, for simple and complex neurons, respectively. (b1) and (b2) display the nodal distribution in terms of the normalized path length and the normalized distance to the soma. (c1) and (c3) show the nodal distribution in terms of the normalized path length to the soma and the curvature of the associated cluster. Blue and red dots represent axon and dendrite nodes, respectively. Details of curvature calculations are described in Appendix B

and normalized direct distance to the soma. The distribution clearly indicates that most nodes are well-separated in such 2D space. In fact, feature ranking by XGB also reveals these two features as the most important features for the identification of nodal polarity.

Apart from the two soma features mentioned above, in Fig. 4(c1) and (c2), we also present the distribution of nodal polarity in the space of normalized length to the soma and the cluster curvature near a given node. We suggest that the polarity classification can be effectively enhanced by including

(a1) Simple, Model I, XGB

96.5%	axon	dend
axon	738	45
dend	125	3909
	prc	rcl
axon	85.5%	94.3%
dend	98.9%	96.9%

(a2) Complex, Model I, XGB

95.3%	axon	dend
axon	1480	201
dend	15	2939
	prc	rcl
axon	99.0%	88.0%
dend	93.6%	99.5%

(a3) All, Model I, XGB

95.9%	axon	dend
axon	2218	246
dend	140	6848
	prc	rcl
axon	94.1%	90.0%
dend	96.5%	98.0%

(b1) Simple, Model I, DNN

96.5%	axon	dend
axon	745	38
dend	133	3901
	prc	rcl
axon	84.9%	95.2%
dend	99.0%	96.7%

(b2) Complex, Model I, DNN

95.7%	axon	dend
axon	1501	180
dend	22	2932
	prc	rcl
axon	98.6%	89.3%
dend	94.2%	99.3%

(b3) All, Model I, DNN

96.1%	axon	dend
axon	2246	218
dend	155	6833
	prc	rcl
axon	93.5%	91.1%
dend	96.9%	97.8%

(c)

Accuracy		Predict polarity	
		axon	dend
Actual polarity	axon	AA	AD
	dend	DA	DD
		prc	rcl
axon		PA	RA
dend		PD	RD

PA(precision of axon)=AA/(AA+DA)
PD(precision of dendrite)=DD/(AD+DD)
RA(recall of axon)=AA/(AA+AD)
RD(recall of dendrite)=DD/(DA+DD)
Accuracy=(AA+DD)/(AA+AD+DA+DD)

Fig. 5 Performance of NPIN with Model I, where both Soma Features and Local Features are used. (a1)–(a3) are the confusion matrix and precision/recall table of the terminal polarity, based on the XGB algorithm for simple, complex, and all neurons, respectively. (b1)–(b3) are the same as in (a1)–(a3) but calculated by the DNN algorithm. (c) defines the confusion matrices shown in this figure. In the upper part of the table, each row indicates the actual polarity, and each column indicates the polarity predicted by NPIN. The lower part of the table displays the precision and recall of axonal and dendritic terminals. Precision and recall are defined in the equations below (c)

curvature as one of the local features because visual inspection reveals that typically more dendrites (compared to the axon nodes) can be found in the regime of larger curvatures. Such effects look more significant in simple neurons than in complex neurons. However, if we use curvature or other local features alone, the performance of polarity classification cannot be as good as using the path length to the soma.

The importance of each feature can also be obtained from the feature ranking calculation of XGB (however, this function is not available in DNN). This can also be obtained by comparing the overall accuracy after systematically removing certain features during the training. Our result suggests the top six features for the determination of nodal polarity: (1) unnormalized path length to the soma, (2) normalized path length to the soma, (3) unnormalized distance to the soma, (4) normalized distance to the soma, (5) curvature of the associated cluster, and (6) aspect ratio of the associated cluster. Other features are less important but can still contribute to the overall performance of NPIN. These results also confirm that local features are secondary factors for the determination of nodal polarity.

Identification Results of Model I: Using both Soma Features and Local Features

To present the results of polarity identification by NPIN, we start from Model I by using both soma features and local features for the whole dataset (with both simple and complex neurons). Figures 5(a1)–(a3) show the confusion matrix of Model I based on XGB and the associated precision/recall table for the polarity of terminals in simple neurons, complex neurons, and all neurons, respectively. Figures 5(b1)–(b3) show the results of Model I but based on DNN for comparison. Figure 5(c) presents the definition of the confusion matrix and explains how the precision and recall are calculated for axons and dendrites. Because the final result is a binary classification of terminal polarity, the dividing nodes are not included either in the training data or in the test data.

From results shown in Figs. 5(a3) and 5(b3), we discover that NPIN is a very powerful classifier with an overall accuracy of 96%. This is achieved by including both Soma Features and Local Features. The model is trained and applied on both simple and complex neurons. According to our results, in general, the precision and recall for the polarity identification of dendritic terminals are better than those of axons by 3%–8%. One possible reason is that the total number of dendrite terminals is approximately three times more than the number of axon terminals, providing more training data that may increase the precision.

Comparing the confusion matrices for simple neurons (Figs. 5(a1) and 5(b1)) and complex neurons (Figs. 5(a2) and 5(b2)), we can observe similar performance of XGB and DNN on simple and complex neurons: The accuracy for simple neurons is higher than that of complex neurons by 1.2% for XGB (compare Figs. 5(a1) and 5(a2)), while it becomes 0.8% if calculated by DNN (compare Figs. 5(b1) and 5(b2)).

However, such similar accuracy of polarity identification for simple and complex neurons is surprising, because complex neurons have more than two clusters. Therefore, the polarity of middle clusters cannot be easily identified according to its relative distance to soma. There are also various kinds of complex neurons (see Fig. 2, for example), which may also have an axon cluster close to the soma. As a result, a naïve comparison of the path length to the soma should not work well for a complex neuron. Hence, it is reasonable to believe that, in our NPIN, the contribution of soma features to simple and complex neurons should be different from the contribution of local features. To understand how this result is related to the feature selection in various types of neurons, in the following section, we demonstrate the performance of NPIN with different feature selections.

Identification Results of Model II: Using Soma Features Only

To clarify the role of Soma Features and Local Features in the identification of neuronal polarity, we additionally use Model II, which is trained by using Soma Features only. The model is trained using the same protocol as in the previous section. The results are shown in Fig. 6.

According to Fig. 6, when using Soma Features only, we find that the overall accuracy drops to 95.5% (94.7%) for simple neurons, and 93.1% (90.0%) for complex neurons, respectively, if using XGB(DNN) algorithms. The performance on all neurons, as shown in Fig. 6(a3) and (b3), is between those of the simple and complex neurons, as expected.

Several important conclusions can be made. First, the overall accuracy of Model II is lower than for Model I (compare Fig. 6(a3) with Fig. 5(a3) for XGB and compare Fig. 6(b3) with Fig. 5(b3) for DNN). However, the difference is only 1.6% for XGB, while it is 3.5% for DNN. This means that the contribution of local features, which exists in Model I but not in Model II, is more significant for DNN than XGB. Second, if we compare the results for simple and complex neurons, we can see that the influence of local features is much more significant for complex neurons than for simple neurons. For example, for XGB, we find that the accuracy decreases by 1% only in simple neurons (compare Fig. 5(a1) and Fig. 6(a1)), while it decreases by 2.3% for complex neurons (compare Fig. 5(a2) and Fig. 6(a2)). These two values become 2.2% and 5.7%, respectively, for DNN. This clearly implies that Local Features which are included in Model I are more important for complex neurons compared to simple neurons. The most obvious reason is that complex neurons have more

(a1) Simple, Model II, XGB

95.5%	axon	dend
axon	727	56
dend	162	3872
	prc	rcl
axon	81.8%	92.8%
dend	98.6%	96.0%

(a2) Complex, Model II, XGB

93.1%	axon	dend
axon	1413	268
dend	50	2904
	prc	rcl
axon	96.6%	84.1%
dend	91.6%	98.3%

(a3) All, Model II, XGB

94.3%	axon	dend
axon	2140	324
dend	212	6776
	prc	rcl
axon	91.0%	86.9%
dend	95.4%	97.0%

(b1) Simple, Model II, DNN

94.7%	axon	dend
axon	720	63
dend	192	3842
	prc	rcl
axon	79.0%	92.0%
dend	98.4%	95.2%

(b2) Complex, Model II, DNN

90%	axon	dend
axon	1450	231
dend	231	2723
	prc	rcl
axon	86.3%	86.3%
dend	92.2%	92.2%

(b3) All, Model II, DNN

92.6%	axon	dend
axon	2170	294
dend	423	6565
	prc	rcl
axon	83.7%	88.1%
dend	95.7%	94.0%

(c)

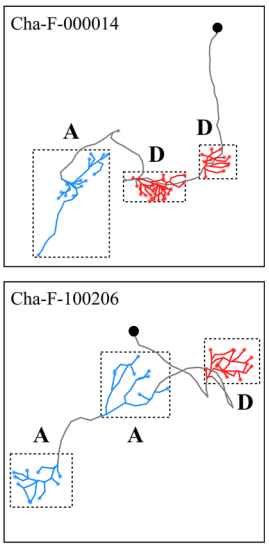

Fig. 6 Performance of NPIN using Model II, where only Soma Features are included. (a1)–(a3) show the results for simple neurons, complex neurons, and all neurons, respectively, using the XGB algorithm. (b1)-(b3) are the same as (a1)–(a3) but for the DNN algorithm. (c) shows two similar complex neurons, where middle clusters have opposite polarities. The cluster labeled by A/D is axons/dendrites

than two clusters and, therefore, the simple application of soma features could not provide enough information for the identification of polarity. As an example, Fig. 6(c) shows two types of complex neurons, where the middle clusters have different polarities. These middle clusters are difficult to classify by Soma Features only. As a result, we conclude that local features are crucial for the polarity identification of the middle clusters in complex neurons and the DNN algorithm may be more sensitive to these differences than XGB.

Comparison of Models I, II, and III for Complex Neurons

In this experiment, to investigate how NPIN works with complex neurons and to examine its relationship with local features, we focus on complex neurons only: no simple neurons are included in either training data or test data. Three models are used for comparison: Model I (with both Soma Features and Local Features), Model II (with Soma Features only), and Model III (with Local Features only). Because the influence of Local Features is more significant in DNN than in XGB (see above), here we will apply the DNN algorithm only for simplicity.

According to the results shown in Fig. 7(a1)–(a3), the accuracy of classification is the best for Model I and slightly reduces for Model II, but it significantly drops for Model III (which uses Local Features only). This result indicates that, without any information on its relative distance to the soma, Local Features alone for a given node perform poorly in polarity identification but are not completely useless (with 71% accuracy, see Fig. 7(a3)). Indeed, we find that the inclusion of local features plays a complementary role in polarity identification, especially for the middle clusters of complex neurons. More precisely, by comparing Fig. 7(a2) to Fig. 7(a1), we find that local features can significantly reduce the number of incorrect identification for axons (upper right corner of the confusion table, from 150 to 66); hence, the number of correctly identified axons is increased.

In Fig. 7 (b1)–(b3), (c1)–(c3), and (d1)–(d3), we show three representative complex neurons with three or more clusters of terminals. Figure 7 (b1), (b2), and (b3) show the same neuron with polarities identified by Model I, Model II, and Model III, respectively. Figure 7 (c1)–(c3) and (d1)–(d3) show similar information but for another two neurons. The results obtained by using Local Features alone (Model III) are not satisfactory: some axon clusters with larger curvatures may be incorrectly classified as dendrites (see, for example, two axon clusters in Fig. 7(c3)). Moreover, some dendrite clusters with divergent branches may be incorrectly classified as axons (see, for example, the dendrite cluster in Fig. 7(d3)). Using Soma Features only (Model II), on the other hand, provides a much better result (with an accuracy of 95.8%), because clusters that are closest to or farthest from the soma are identified as dendrites or axons, respectively. However, as we see in Fig. 7(c2), (d2), and (e2), the middle clusters (defined from their distance to soma) of these complex neurons cannot be identified easily by Model II (with Soma Features only), because their relative distance to the soma is not well-defined compared to the other clusters.

As a summary, we find that the accuracy to classify the polarity of middle clusters in a complex neuron can be significantly enhanced after combining Soma Features and Local Features in

(a1) Complex, Model I, DNN

97.7%	axon	dend
axon	1615	66
dend	43	2911
	prc	rcl
axon	97.4%	96.1%
dend	97.8%	98.5%

(a2) Complex, Model II, DNN

95.8%	axon	dend
axon	1531	150
dend	44	2910
	prc	rcl
axon	97.2%	91.1%
dend	95.1%	98.5%

(a3) Complex, Model III, DNN

71%	axon	dend
axon	1058	623
dend	721	2233
	prc	rcl
axon	59.5%	62.9%
dend	78.2%	75.6%

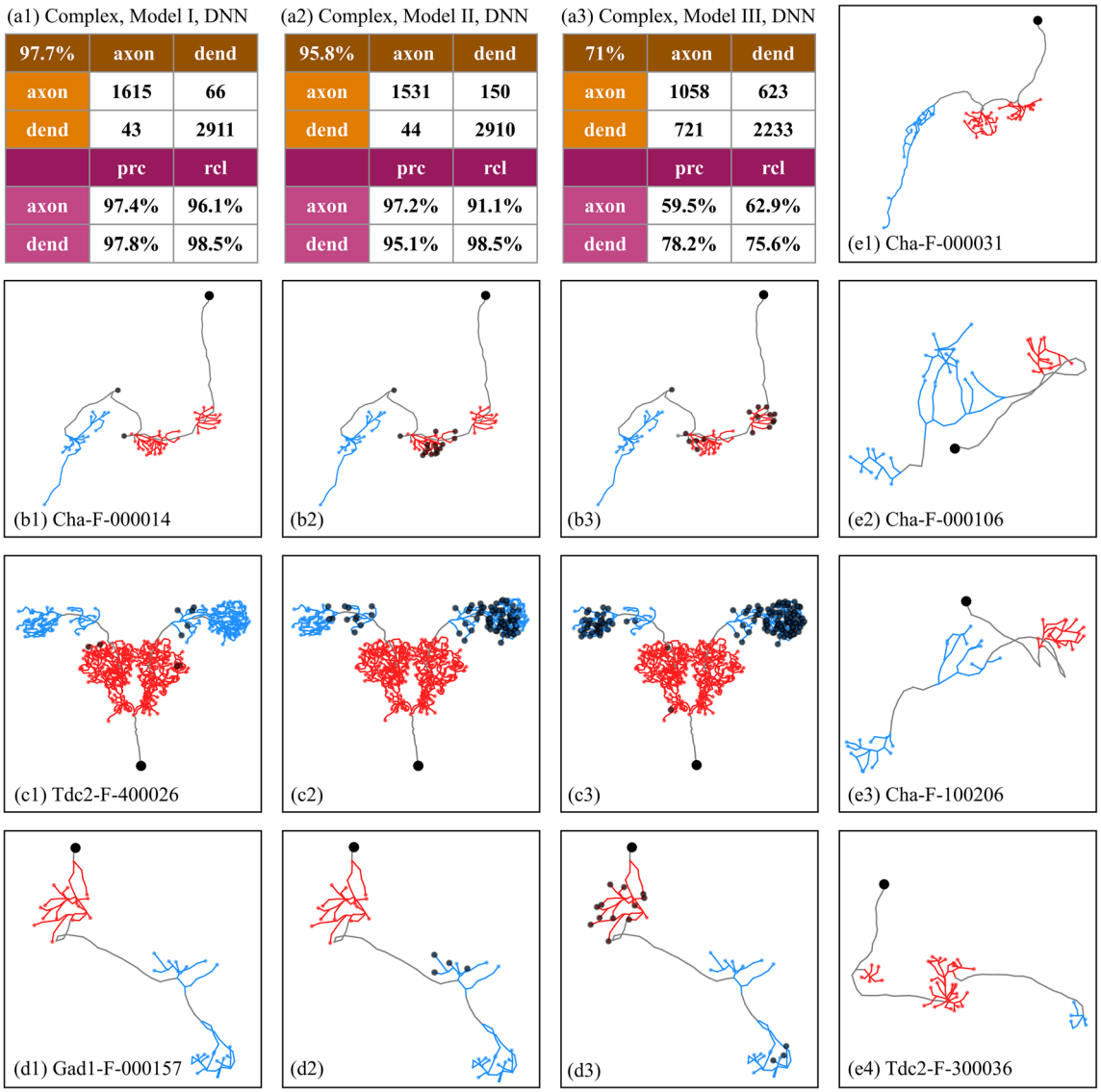

Fig. 7 Performance of NPIN with DNN algorithm for complex neurons in three different models. (a1)–(a3) are the confusion matrix and precision-recall table for the terminal polarity for Model I (with both Soma Features and Local Features), Model II (with Soma Features only), and Model III (with Local Features only), respectively. (b1)–(b3) display the same complex neuron with polarity classification using Model I, Model II, and Model III, respectively. Filled gray circles indicate the terminals of incorrect classification. (c1)–(c3) and (d1)–(d3) are the same as in (b1)–(b3) but with two different complex neurons. (e1)–(e4) are four different complex neurons, where polarities are classified by Model I with 100% accuracy by DNN algorithm

Model I. More examples of complex neurons with correct polarity identification by Model I are shown in Fig. 7(e1)–(e4).

Application to Other Species of Insects: Blowfly and Moth

In principle, our NPIN, trained on the *Drosophila* brain neurons, can also be applied to the polarity identification of other species, if the training data is replaced by the neurons of that species. However, the number of publicly available neuronal data samples of other species with identified polarity is much less than that of *Drosophila*. Therefore, such an application may not be practical. However, it is still instructive to see how our NPIN, trained by *Drosophila* neurons, can be directly used for other species of insects, which should have similar morphological features as *Drosophila*. Here, we take the neuron images of the blowfly and the moth from the Neuromorpho database (http://neuromorpho.org/) as an example. The database lists 19 blowfly neurons and 3 moth neurons with labeled polarity.[1] These data were generated by different labs

[1] The IDs of 19 Blowfly neurons are HSE-fluoro05, HSE-fluoro11, HSE-fluoro15, HSN-cobalt, HSN-fluoro04, HSN-fluoro06, HSN-fluoro08, HSS-cobalt, VS1-cobalt, VS2-fluoro01, VS2-fluoro03, VS2-fluoro10, VS3-cobalt, VS4-cobalt, VS4-fluoro02, VS4-fluoro07, VS4-fluoro09, VS5-cobalt, VS9-cobalt. The IDs of 3 Moth neurons are Nevron-komplett-08-02-28-2a, Nevron-komplett-08-03-13-2a, Nevron-komplett-08-08-28-1a-A.

using different reconstruction methods from that of our *Drosophila* dataset. To save space, below we just present NPIN results of the blowfly in detail and mention the results of the moth data in brief.

Figure 8 shows the results of polarity identification for 19 blowfly neurons obtained by Model I, Model II, and Model III of NPIN, which is trained on 213 *Drosophila* neurons in our dataset with the DNN model. We find that Model I, using both soma features and local features, still provides a decent level of accuracy (83.4%). The main error stems from the pretty low precision and recall of the axons, which have much fewer terminal numbers than dendrites (dendrites: axons ratio = 22.8:1). Similar results are also observed for Model II, as shown in Fig. 8(a2).

However, a surprising result is obtained when using Model III, where only local features are included for the training on *Drosophila* neurons. The overall accuracy, as well as the precision and recall for both dendrites and axons, are very high (accuracy = 98.98%). This result is even better than that obtained by using the blowfly data for the training process (Fig. 8(b)). The results clearly indicate that, unlike *Drosophila*, where Local Features are only secondary factors compared with Soma Features, Local Features are the primary factors for the identification of neuronal polarity for blowfly neurons that we tested in the present study. This can also be observed from the skeleton structure of dendrite clusters in Fig. 8(c1)–(c4). Therefore, to apply NPIN (trained on *Drosophila* neurons) to neurons of other insects, it is necessary to provide not only Model I, but also Model II and Model III, to maximize the range of applications. However, we have to emphasize that since the 19 blowfly neurons are all collected from the visual system, and therefore we could not exclude the possibility that the success of NPIN may be due to their special morphology. More analysis on other types of neurons in the blowfly should be carried out in the future when more neurons with known polarity are available.

In addition to the blowfly, we also collected 3 moth neurons with known polarities from the Neuromorpho database. Among the 194 dendrite terminals and 358 axon terminals, the overall accuracy of the polarity identification by NPIN (trained by the 213 *Drosophila* neurons with the DNN model) is 98.2%, 99.0% and 65.6% for Model I, II, and III respectively. This result reflects the fact that the polarity of these three neurons could be much easily identified by Soma Features only. This could be a complementary example of the blowfly and show the importance of including both Soma Features and Local Features for a general application of NPIN.

Summary of Results

We summarize the results of the present study in Fig. 9 by showing the accuracy of NPIN in all test conditions including three models (Model I: all features, Model II: Soma Features only, Model III: Local Features only) and three types of test data (simple neurons, complex neurons, and all neurons). For simplicity, we only display the results using the DNN algorithm.

As explained above, the overall accuracy cannot reflect the complete information on model performance, especially when the numbers of dendrites and axons are highly imbalanced. To generate a reliable ML model, we suggest that the precision and recall for both axons and dendrites have to be larger than 50%, or, in other words, we have more correctly identified terminals than incorrect ones. We put stars "*" in Fig. 9 to mark those results that do not meet these criteria.

Discussion

Comparison of NPIN and SPIN

A previously developed machine-learning-based method, SPIN (described in the introduction), has identified the polarity of insect's neurons with an overall accuracy 84%–90% (Lee et al. 2014). SPIN starts by identifying clusters of neuronal arbors in each neuron and then classifies the polarity of each cluster according to its geometric structure and distance to the soma. As a result, terminals in a cluster are all classified as having the same polarity. However, this approach has two challenges. First, a cluster might not be easily identified for neurons with complex morphology, and incorrect clustering could lead to a large number of incorrectly classified terminals, for example, 14 of 213 neurons used in the present study were not processable by SPIN. Second, the number of available clusters may not be sufficient to achieve good training results because each neuron has only a few clusters. Due to these issues, SPIN often failed to classify part of or even all terminals of a neuron if its arbors were not clustered correctly.

The proposed NPIN avoids these issues by adopting node-based rather than cluster-based classification. To compare the performance of SPIN and NPIN, we examine the results of the polarity identification by SPIN on the same 213 neurons we used here (Huang et al. 2019). We find that, among these 213 neurons, only 79 neurons are fully identified (i.e., without any "non-classified" terminals), 120 neurons are partially predicted (i.e., some clusters cannot be identified), and 14 neurons cannot be predicted. Among 9452 terminals of these 213 neurons, there are 1207 unclassified terminals and 8247 classified terminals. Within the SPIN-classified terminals, 8038 terminals are correctly identified for their polarities. Therefore, the overall accuracy of SPIN is 85.04% only if we consider all terminals in the dataset, while it could be 97.49% if we considered only classified terminals.

We emphasize that, in the present study, we develop a completely different approach by identifying the polarity of

(a1) Blowfly, Model I, DNN

88.7%	axon	dend
axon	112	103
dend	475	4424
	prc	rcl
axon	19.1%	52.1%
dend	97.7%	90.3%

(a2) Blowfly, Model II, DNN

87.6%	axon	dend
axon	85	130
dend	833	4066
	prc	rcl
axon	9.3%	39.5%
dend	96.9%	83.0%

(a3) Blowfly, Model III, DNN

98.98%	axon	dend
axon	210	5
dend	47	4852
	prc	rcl
axon	81.7%	97.7%
dend	99.9%	99.0%

(b) Blowfly, Model I*, DNN

95.3%	axon	dend
axon	83	132
dend	109	4790
	prc	rcl
axon	43.2%	38.6%
dend	97.3%	97.8%

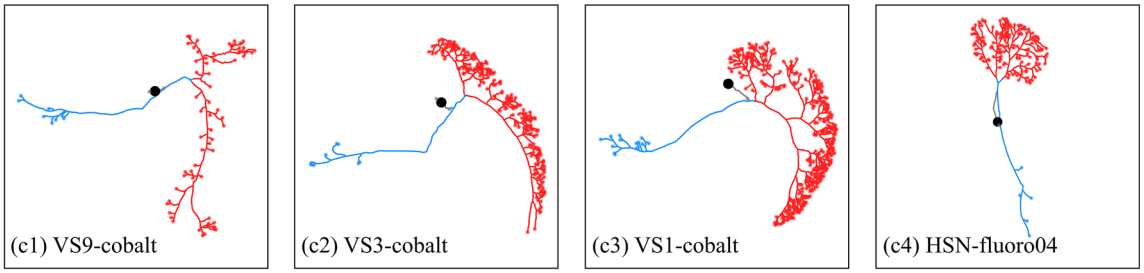

Fig. 8 Performance of NPIN on blowfly brain neurons. (a1)–(a3) are the confusion matrices and precision-recall tables for Model I, Model II, and Model III, respectively. The models are trained on 213 fruit-fly neurons in our dataset. (b) is the result for Model I but trained on blowfly neurons directly. (c1)–(c4) display four example skeleton structures of the blowfly neurons used in this test

each node, which can be unambiguously defined in the skeleton structure of each neuron, with a nodal polarity also well-defined through the polarity of terminals (see Fig. 3). Such node-based feature extraction, therefore, takes advantage of the fact that the number of nodes is much larger than the number of clusters in each neuron. It can achieve a much higher accuracy (>96%) for the whole dataset (213 neurons and 9452 terminals) after including the spatial correlation. Therefore, we conclude that NPIN outperforms SPIN in the polarity identification, showing an important step toward the reconstruction of the connectome. We expect to analyze the information flow of the brain in much finer scales in the near future, revealing more detailed functional relationships between subregions of the *Drosophila* brain.

Neurons Not in the Dataset

As described in the flowchart of NPIN in Fig. 1, the neuronal polarity predicted by NPIN for the 213 *Drosophila* neurons is obtained by randomly selecting 100 neurons for training, 25 for validation and 50 for test. In other words, each neuron shown in Tables S1 and S2 of Appendix E is tested by models, trained on other neurons in the dataset, and therefore there is no overlap between the test data and training data in all the results presented above.

However, in order to show how well NPIN can be applied for neurons not in the same dataset, we find another 22 neurons with distinct connection types (mostly from AOTU to BU, and from MED to VMP) from those in the NPIN dataset for the test. The polarities of these neurons are determined by our experimental collaborators and therefore are not published before (and also not in the original dataset either). 12 of them have dendrites located in AOTU, 9 in MED and 1 complex neuron in MB, see Table S3 of Appendix E. The predicted results by NPIN (trained by the 213 neurons together) show that NPIN could still provide very high accuracy. More precisely, for these 12 neurons in AOTU, 11 of them are 100% correct and only one is of 75% accuracy. For the 9 neurons

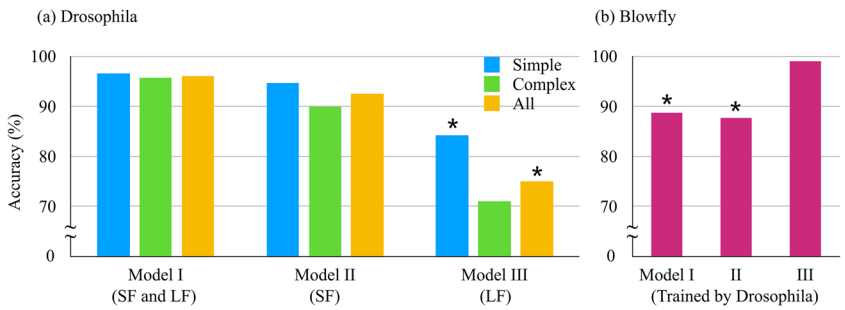

Fig. 9 Summary of NPIN accuracies in all test conditions using the DNN algorithm. (a) shows the results for Model I (with both Soma Features and Local Features), Model II (with Soma Features only), and Model III (with Local Features only), for three types of test data: simple neurons, complex neurons, and all neurons, respectively. (b) shows the results for the same models but with the blowfly neurons (trained by our *Drosophila* dataset). Results with precision or recall of less than 50% are indicated by "*" (see the text)

with dendrites in MED, 7 of them are 100% correct and the other two are of 94.3% and 80.6% accuracy. The complex neuron with dendrites in MB (also not shown in the dataset of NPIN) is predicted with 100% accuracy. These results show that our NPIN should be applicable to neurons in other brain regions. Although like all machine learning algorithms, NPIN is trained by labeled data with similar features (neuronal morphology) to those of unlabeled data, we found that NPIN is still able to successfully classify polarity for neurons that are morphologically distinct from the training neurons.

Nevertheless, we have to acknowledge that the number of neurons in our dataset is far less than the total number of neurons (approximately 135 K) in the *Drosophila* brain. There must be other types of neurons with polarity-specific morphological features, which can be very different from what we have addressed in this study. For example, the dendrites and axons of some local neurons are co-localized in the same cluster of arbors, and some projection neurons develop axonal clusters that are closer to the soma than the dendritic ones. We will include more morphologically distinct neurons into the training set, once their experimentally verified polarities are available. Therefore, although more training data are necessary when applying our NPIN for the polarity identification of the whole *Drosophila* brain, the present work at least demonstrates the possibilities to have a high precision identification through the node-based feature extraction in NPIN. We believe that the future versions of NPIN, after including more types of neurons in the training data, will provide a much wider range of applications.

Neurons with Low Accuracy

To examine the performance of our NPIN, we investigate those neurons not identified well in their polarity. As described in the Results section, we could obtain this information by randomly selecting 150 neurons (100 for training, 25 for validation, and 50 for testing) out of the 213 neurons in the dataset for each training/test process and then repeating it for 20 rounds. As a result, each neuron can be tested (by different models trained by other neurons) for 4–5 times on average, and their polarity identification results can be obtained by averaging their probabilities before relabeling. The final results calculated by the DNN model are shown in Appendix E. Within these 213 neurons, the terminal polarity of 166 neurons is identified with 100% accuracy. Only 14 simple neurons and 33 complex neurons are not fully identified. Concentrating on those neurons with a lower accuracy (say below 85%), we find only 5 simple neurons and 24 complex neurons left.

When looking into the skeleton structures of these neurons with a lower accuracy, we find the following features of these neurons: Simple neurons have a very similar distance for axon clusters and dendrite clusters to the soma, and the number of dendrite terminals is much larger than the number of axon terminals. The former makes it difficult to distinguish axons from dendrites, while the latter could confuse NPIN by mispredicting all terminals to be dendrites (as a result, the precision and recall of axons are both small). For complex neurons, the incorrectly identified terminals usually appear in the middle clusters, as one may expect. However, the most complex neurons have been correctly predicted by NPIN with a very high accuracy (91 of the 124 complex neurons are identified with 100% accuracy). In our node-based feature extraction, it is challenging to correctly identify the clusters of fewer terminals or nodes, because their local features are less representative of their local morphology. Therefore, finding a better way to define local features (less dependent on the number of terminals in the same clusters) could enhance polarity identification in future work.

Comparison with Results of Electronic-Microscopy Images

Finally, a large set of electronic-microscopy images (the EM dataset) of the *Drosophila* brain has recently been released (C. S. Xu et al. 2020). This dataset includes identified polarities, and hence can be potentially used as the training data for NPIN or be compared with the results predicted by NPIN on the fluorescence images. However, after careful examination of that dataset, we discovered two major differences in the morphological characteristics between the two datasets: (1) the neuronal skeletons in the EM dataset exhibit much more details, e.g., a larger number of short terminal branches than what have been found in the fluorescent images in the present study. (2) Some neurons in the EM dataset have incomplete tracing or discontinuous branches. These issues prevent us from directly using the EM dataset. For future work, we suggest that heavy preprocessing, containing the reconstruction of the connectome and the algorithm of matching terminals of the same neuron from two image types, is required, before NPIN can utilize the EM dataset. Moreover, we have to emphasize that the current hemi-brain EM database is from ONE fly only, but neural images from light-microscopy based databases are often accumulated from multiple individuals. Although this database serves as crucial reference data for the fly community, it is unlikely that full-brain or hemi-brain EM data from many more flies or from flies with different genetic manipulation will become available in the near future. By contrast, data from optical images are continuously generated by a large number of labs in the world. We therefore believe that our NPIN will have its impact and be widely used by many fly labs in the future.

Conclusion

In this study, we have developed NPIN, a completely new ML model to identify the polarity of projection neurons in a

Drosophila brain with high precision (>96%). This result was achieved due to three major contributions: node-based feature extraction, separation of Local Features from Soma Features, and implementation of spatial correlations between nodal polarities. In the experiments, we systematically compare the results of different models for various types of neurons. We demonstrate that, apart from Soma Features, Local Features are the secondary factors to determine the neuronal polarity. Local Features can significantly improve the polarity identification, especially for the middle clusters of complex neurons, which cannot be well-identified by using Soma Features only. Besides the *Drosophila* neurons, we show that NPIN can also be applied to identify the neuronal polarity of other insects, such as the blowfly. As a result, we believe that the development of NPIN and its applications is an important step toward the determination of signal flows in complex neural networks.

Information Sharing Statement

The NPIN software package contains data of sample neurons with skeletal data available from the FlyCircuit database (http://www.flycircuit.tw). We also provide two online versions of NPIN to be used or tested by other research groups at the following address: website (https://npin-for-drosophila.herokuapp.com) and Gitlab code (https://gitlab.com/czsu32/npin).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12021-021-09513-y.

Acknowledgments This work is supported by the Ministry of Science and Technology grant (MOST 107-2112-M-007-019-MY3) and by the Higher Education Sprout Project funded by the Ministry of Science and Technology and the Ministry of Education in Taiwan. We thank the National Center for Theoretical Sciences and Brain Research Center in the National Tsing Hua University for providing full support in the interdisciplinary collaboration. We thank the National Center for High-Performance Computing for providing the FlyCircuit database. We appreciate Prof. Che-Rung Lee for providing computational resources, Yi-Ning Juan for helping with the website, Dr. Yu-Chi Huang and Prof. An-Shi Chiang for helpful discussion about the neuronal datasets.

Funding This work is supported by the Ministry of Science and Technology grant (MOST 107–2112-M-007-019-MY3) and by the Higher Education Sprout Project funded by the Ministry of Science and Technology and the Ministry of Education in Taiwan.

Data Availability The FlyCircuit database (http://www.flycircuit.tw/) is provided by the National Center for High-Performance Computing.

Code Availability We provide an online version and a source code of NPIN (XGB version) at the following websites:
Test Page (https://npin-for-drosophila.herokuapp.com/).
Gitlab code (https://gitlab.com/czsu32/npin).

Declarations

Conflict of Interests The authors declare that they have no conflict of interests.

References

Asri, H., Mousannif, H., Moatassime, H. A., & Noel, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science, 83*, 1064–1069. https://doi.org/10.1016/j.procs.2016.04.224.

Chiang, A.-S., Lin, C.-Y., Chuang, C.-C., Chang, H.-M., Hsieh, C.-H., Yeh, C.-W., et al. (2011). Three-Dimensional Reconstruction of Brain-wide Wiring Networks in Drosophila at Single-Cell Resolution. *Current Biology, 21*(1), 1–11. https://doi.org/10.1016/j.cub.2010.11.056.

Craig, A. M., & Banker, G. (1994). Neuronal Polarity. *Annual Review of Neuroscience, 17*(1), 267–310. https://doi.org/10.1146/annurev.ne.17.030194.001411.

Cuntz, H., Forstner, F., Haag, J., & Borst, A. (2008). The Morphological Identity of Insect Dendrites. *PLoS Comput Biol, 4*(12), e1000251. https://doi.org/10.1371/journal.pcbi.1000251.

Fischbach, K.-F., & Dittrich, A. P. M. (1989). The optic lobe of Drosophila melanogaster. I. A Golgi analysis of wild-type structure. *Cell and Tissue Research, 258*(3), 441–475. https://doi.org/10.1007/BF00218858.

Hanesch, U., Fischbach, K.-F., & Heisenberg, M. (1989). Neuronal architecture of the central complex in Drosophila melanogaster. *Cell and Tissue Research, 257*(2), 343–366. https://doi.org/10.1007/BF00261838.

Huang, Y.-C., Wang, C.-T., Su, T.-S., Kao, K.-W., Lin, Y.-J., Chuang, C.-C., et al. (2019). A Single-Cell Level and Connectome-Derived Computational Model of the Drosophila Brain. *Frontiers in Neuroinformatics, 12*. https://doi.org/10.3389/fninf.2018.00099.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25*(pp. 1097–1105). Curran Associates, Inc. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf. Accessed 13 April 2020

Kuan, L., Li, Y., Lau, C., Feng, D., Bernard, A., Sunkin, S. M., et al. (2015). Neuroinformatics of the Allen Mouse Brain Connectivity Atlas. *Methods, 73*, 4–17. https://doi.org/10.1016/j.ymeth.2014.12.013.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. Presented at the Proceedings of the IEEE. https://doi.org/10.1109/5.726791.

Lee, Y.-H., Lin, Y.-N., Chuang, C.-C., & Lo, C.-C. (2014). SPIN: A Method of Skeleton-Based Polarity Identification for Neurons. *Neuroinformatics, 12*(3), 487–507. https://doi.org/10.1007/s12021-014-9225-6.

Lin, C.-Y., Chuang, C.-C., Hua, T.-E., Chen, C.-C., Dickson, B. J., Greenspan, R. J., & Chiang, A.-S. (2013). A Comprehensive Wiring Diagram of the Protocerebral Bridge for Visual Information Processing in the Drosophila Brain. *Cell Reports, 3*(5), 1739–1753. https://doi.org/10.1016/j.celrep.2013.04.022.

Lo, C.-C., & Chiang, A.-S. (2016). Toward Whole-Body Connectomics. *Journal of Neuroscience, 36*(45), 11375–11383. https://doi.org/10.1523/JNEUROSCI.2930-16.2016.

Malta, T. M., Sokolov, A., Gentles, A. J., Burzykowski, T., Poisson, L., Weinstein, J. N., et al. (2018). Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell, 173*(2), 338–354.e15. https://doi.org/10.1016/j.cell.2018.03.034.

Matus, A., Bernhardt, R., & Hugh-Jones, T. (1981). High molecular weight microtubule-associated proteins are preferentially associated with dendritic microtubules in brain. *Proceedings of the National Academy of Sciences of the United States of America, 78*(5), 3010–3014.

Milyaev, N., Osumi-Sutherland, D., Reeve, S., Burton, N., Baldock, R. A., & Armstrong, J. D. (2012). The Virtual Fly Brain browser and query interface. *Bioinformatics, 28*(3), 411–415. https://doi.org/10.1093/bioinformatics/btr677.

Mohsen, H., El-Dahshan, E.-S. A., El-Horbaty, E.-S. M., & Salem, A.-B. M. (2018). Classification using deep learning neural networks for brain tumors. *Future Computing and Informatics Journal, 3*(1), 68–71. https://doi.org/10.1016/j.fcij.2017.12.001.

Parekh, R., & Ascoli, G. A. (2013). Neuronal Morphology Goes Digital: A Research Hub for Cellular and System Neuroscience. *Neuron, 77*(6), 1017–1038. https://doi.org/10.1016/j.neuron.2013.03.008.

Peng, H., Hawrylycz, M., Roskams, J., Hill, S., Spruston, N., Meijering, E., & Ascoli, G. A. (2015). BigNeuron: Large-Scale 3D Neuron Reconstruction from Optical Microscopy Images. *Neuron, 87*(2), 252–256. https://doi.org/10.1016/j.neuron.2015.06.036.

Rolls, M. M. (2011). Neuronal polarity in Drosophila: Sorting out axons and dendrites. *Developmental Neurobiology, 71*(6), 419–429. https://doi.org/10.1002/dneu.20836.

Shinomiya, K., Matsuda, K., Oishi, T., Otsuna, H., & Ito, K. (2011). Flybrain neuron database: A comprehensive database system of the Drosophila brain neurons. *The Journal of Comparative Neurology, 519*(5), 807–833. https://doi.org/10.1002/cne.22540.

Squire, L. R., Berg, D., Bloom, F., Lac, S. du, & Ghosh, A. (Eds.). (2008). *Fundamental Neuroscience, Third Edition* (3rd ed.). Academic Press.

Wang, J., Ma, X., Yang, J. S., Zheng, X., Zugates, C. T., Lee, C.-H. J., & Lee, T. (2004). Transmembrane/Juxtamembrane Domain-Dependent Dscam Distribution and Function during Mushroom Body Neuronal Morphogenesis. *Neuron, 43*(5), 663–672. https://doi.org/10.1016/j.neuron.2004.06.033.

Wu, M., Nern, A., Williamson, W. R., Morimoto, M. M., Reiser, M. B., Card, G. M., & Rubin, G. M. (2016). Visual projection neurons in the Drosophila lobula link feature detection to distinct behavioral programs. *eLife, 5*, e21022. https://doi.org/10.7554/eLife.21022.

Xu, C. S., Januszewski, M., Lu, Z., Takemura, S., Hayworth, K. J., Huang, G., et al. (2020). A Connectome of the Adult Drosophila Central Brain. *bioRxiv*, 2020.01.21.911859. 10.1101/2020.01.21.911859.

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/10.1371/journal.pone.0054050.