

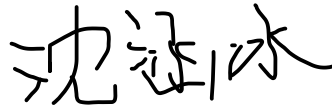
Unsupervised clustering for putative dopamine
neuron identification from tetrode recordings in
behaving mice

521

Statement of originality

I hereby declare that to the best of my knowledge, the content of this dissertation is my own work. I certify that the intellectual content of this dissertation is the product of my own work and has been done during the specified period of my research project. All the assistance received in preparing this dissertation and sources have been properly acknowledged. The content acquired outside the specified period of my research project has been properly indicated in the dissertation.

Student's name: Shen Hanbing

Handwritten signature of Shen Hanbing in Chinese characters.

Date: 2021/5/10

Acknowledgments

I would like to thank Naoshige Uchida (Harvard University), Sara Matias (Harvard University), HyungGoo Kim (Sungkyunkwan University) for their extraordinary support in this thesis process.

Abstract

Extracellular recordings from midbrain dopamine (DA) neurons in behaving animals provide a powerful tool to dissect the midbrain DA system. Nevertheless, one challenge is distinguishing putative DA (pDA) neurons from other neuron types by current methodologies, including simple electrophysiological criteria and opto-tagging technique. Using simple electrophysiological criteria to identify pDA neurons could be problematic due to the heterogeneity of midbrain DA neurons regarding their electrophysiological characteristics. opto-tagging is time-consuming, and it might miss real DA neurons. Therefore, this project aims to utilize unsupervised clustering methods to link electrophysiological properties' analysis with opto-tagging for unbiased and quantitative large-scale pDA neuron identification. The first goal is to extract the electrophysiological properties from recorded neurons. The second goal is to use unsupervised clustering methods to identify additional pDA neurons that might have been missed by the opto-tagging technique. K-means clustering with two clusters was evaluated as the best clustering model to distinguish pDA neurons from non-pDA neurons, achieving unbiased and relatively stable pDA identification with high accuracy. Feature extraction and unsupervised clustering methods introduced here could advance future research in the midbrain DA system. In the future, applying opto-tagging on other midbrain neurons is necessary for a more ground-truth pDA identification.

Keywords

Midbrain dopamine neurons, extracellular recordings in vivo, opto-tagging, unsupervised clustering, k-means

Table of contents

1.	Introduction.....	8
	1.1 Dopamine and midbrain dopamine neurons.....	8
	1.2 Extracellular recordings from midbrain dopamine neurons in behaving animals and their current limitations.....	8
	1.3 Unsupervised clustering for neuron classifications.....	9
2.	Aim and Hypothesis.....	9
3.	Materials and Methods.....	9
	3.1 Data source.....	9
	3.2 Extraction of spike waveform properties from single-units.....	10
	3.3 Extraction of firing properties from the time series activity of single-units.....	10
	3.4 Dimensionality reduction.....	11
	3.5 Unsupervised Classification Algorithms.....	11
	3.6 Evaluations of Unsupervised Classification Algorithms.....	12
	3.7 Predictive Analysis.....	13
4.	Results.....	13
	4.1 Representative feature extractions and dimensionality reduction.....	13
	4.2 Unsupervised clustering comparisons: k-means and GMM are superior regarding their accuracy.....	14
	4.3 K-means outperforms GMM.....	15
	4.4 K-means clustering classification of pDA and non-pDA neurons.....	15
	4.5 The clustering structure in K-means model could predict pDA neurons in the testing dataset	16
5.	Discussion.....	16
	5.1.K-means outperforms other unsupervised clustering methods on pDA identification.....	16
	5.2 K-means clustering classification of spike waveforms and firing rate in pDA and non-pDA clusters	18
	5.3 Limitations of PCA for dimensionality reduction.....	18
	5.4 Conduct optotagging on other midbrain neurons and explore semi-supervised classifier.....	18
6.	Conclusion.....	19
7.	References.....	20

List of figures and tables

Figure 1. Workflow of unsupervised clustering for putative dopamine neurons identification.....	24
Figure 2. Representative extractions of spike waveform properties and firing properties.....	26
Figure 3. Principal component analysis for dimensionality reduction and t-SNE.....	28
Figure 4. Unsupervised clustering of extracellularly recorded neurons in the VTA.....	29
Figure 5. Evaluation of k-means and GMM.....	31
Figure 6. k-means clustering of extracellular waveforms and firing patterns from pDA and non-pDA clusters.....	32
Figure 7. Prediction models using labels obtained from k-means clustering.....	33
Supplementary Figure 1. Extracted firing property features.....	34
Supplementary Figure 2. Parameter set estimations of DBSCAN.....	34
Supplementary Figure 3. Confusion matrix of the linear SVM model.....	35

List of abbreviations

ChR2	Channelrhodopsin-2
DA	Dopamine
DAT	Dopamine Transporter
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
GMM	Gaussian Mixture Model
ISI	Interspike Interval
M	Min Number of Neighbors
NLISI	Normalized Log-Transformed ISI
OCSVM	One-Class Support Vector Machine
PC	Principal Components
PCA	Principal Component Analysis
pDA	Putative Dopamine
R	Radius of Neighborhood around a Point
RGS	Robust Gaussian Surprise
SNC	Substantia Nigra Pars Compacta
SVMs	Support Vector Machines
TH	Tyrosine Hydroxylase
TP	True Positive Rate
t-SNE	<i>t</i> -distributed Stochastic Neighbor Embedding
VTA	Ventral Tegmental Area

1. Introduction

1.1. Dopamine and Midbrain Dopamine Neurons

Dopamine (DA), a neurotransmitter, has been reported to modulate motivation, reward processing, and learning (Wise, 2004). The majority of midbrain DA neurons are located in the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc), receiving inputs from the striatum and from motor, somatosensory, and automatic areas (Watabe-Uchida et al., 2012). The activity of midbrain DA neurons in reward-related behaviors has been paralleled to the reward prediction error signal used in classical and state-of-the-art Reinforcement Learning models (Schultz et al., 1997) (Cohen et al., 2012) (Dabney et al., 2020). Dysregulation of the midbrain DA system has been implicated in the pathophysiology of several psychiatric and neurologic disorders (Volkow et al., 2015). Nevertheless, additional studies are required to cope with multiple unverified hypotheses like ramping DA signals in the midbrain (Mohebi et al., 2019) (Kim et al., 2020), to reveal the role of midbrain DA neurons across circuit and behavioral levels.

1.2. Extracellular Recordings from Midbrain Dopamine Neurons in Behaving Animals and Their Current Limitations

However, in extracellular recordings, it is challenging to distinguish putative DA (pDA) neurons from other neuron types (e.g., GABAergic and glutamatergic neurons) in the midbrain with existent methodologies. A traditional criterion to recognize pDA neurons among recorded neurons was based on several electrophysiological characteristics, including slow baseline firing rate, spike waveform, and regularity of baseline firing pattern (Matsumoto and Hikosaka, 2009). Nevertheless, the electrophysiological characteristics DA neurons in the VTA are extremely heterogeneous (Morales and Margolis, 2017), and hence, no single electrophysiological characteristic can determine DA neurons (Margolis et al., 2006). For example, Lammel and his group (2008) have identified an atypical fast-firing mesocortical DA population with correspondingly low dopamine transporter (DAT) mRNA expression.

Later, a combination of Cre transgenic mouse lines and optogenetics ushers in a new phase in identifying DA neurons (Lammel et al., 2015). The opto-tagging (Lima et al., 2009) technique can directly photo-stimulate cells that express a light-sensitive opsin under genetic control, and either DAT or tyrosine hydroxylase (TH) promoters can be used to capture DA neurons (Cohen et al., 2012).

However, the emergency of silicon probes for high-density recordings such as Neuropixels (Jun et al., 2017) in the brain allows the simultaneous recordings of hundreds of neurons, which is difficult to achieve with classical tetrode recordings with opto-tagging. Opto-tagging is largely at the mercy of transgenic systems, which are time-consuming with a limited scope, and meanwhile, incomplete virus expression or a long distance to the probe might omit real DA neurons (Jia et al.,

2018). Incomplete virus expression was observed in one study, in which around 10% of TH-labeled cells did not co-express Channelrhodopsin-2 (ChR2) protein, the expression of which is mediated by the virus (Cohen et al., 2012). Therefore, the classical tetrode recordings with opto-tagging might miss a group of pDA neurons. Given that DA neurons are highly heterogeneous regarding their electrophysiological properties, humans cannot distinguish between DA neurons and non-DA neurons, and the methods used until now have substantial limitations, as described above. Therefore, we need a more automated methodology to link electrophysiological properties' analysis with opto-tagging to boost the sensitivity and maintain the specificity for pDA neuron identification.

1.3. Unsupervised Clustering for Neuron Classifications

Unsupervised learning can learn patterns and group data from unlabeled data (Jardine and Sibson, 1968). Supervised and unsupervised classification methods have been utilized to categorize neocortical neurons on the basis of electrophysiology properties (McGarry, 2010) (Guerra et al., 2011) or morphological (DeFelipe et al., 2013). Trainito and his group (2019) have identified four waveform-based cell classes with distinct functions in primate cortex. Prévost-Solié and his colleagues (2020) have conducted an unsupervised clustering analysis on DA neurons. Thus, it is promising to apply unsupervised clustering to link electrophysiological properties' analysis with opto-tagging for neuron identification.

2. Aim and Hypothesis

This project aims to utilize unsupervised clustering methods to link electrophysiological properties' analysis with opto-tagging for unbiased and quantitative large-scale pDA neuron identification. The first goal is to extract electrophysiological properties of recorded neurons that will be used in the clustering analysis. The second goal is to use unsupervised clustering methods to identify additional pDA neurons that the opto-tagging technique might have missed.

3. Materials and Methods

3.1. Data Source

The experiment arrangements are summarized in a workflow as shown in Figure 1 and all data analysis were conducted in MatLab R2019b. Single-unit electrophysiological data was obtained from previous recordings done in the Uchida Lab by Athar Malik (Kim et al., 2020). The VTA of male DAT-Cre mice were injected with Cre-dependent AAV virus expressing ChR2. After a virtual-reality reward delivery task, pulses of light were delivered to activate ChR2-expressing neurons (Figure 1A). This information was used to classify neurons as photo-identified DA neurons or unidentified neurons. 489 neurons were recorded in total, and 122 of them are photo-identified DA

neurons. 320 neurons were randomly selected as the training dataset, and the remaining 169 neurons is the testing dataset.

3.2. Extraction of Spike Waveform Properties from Single-Units

The average waveform from the best-unreferenced channel (i.e. the channel in which the spike waveform shows the largest amplitude) was used to extract waveform features. Spline interpolation that calculates the cubic polynomials between two adjacent data points was used for curve fitting. After curve fitting, two types of total duration (Gale and Perkel, 2016), half-width (Markham and Stoddard, 2015), peak-to-trough duration, and repolarization time (Trainito et al., 2019) were extracted from the spike trains in each neuron (Figure 1B). The first total duration is the distance between the baseline and the peak. The second total duration is the duration between the first point that reaches ten percent of the amplitude of the global peak and the point that returns to ten percent of the amplitude before this peak (Gale and Perkel, 2016). The distance at half-amplitude is defined as the half-width, also called spike width (Markham and Stoddard, 2015). The peak-to-trough duration ranges from the global peak to the following local minimum if the primary peak is positive-going or from the global minimum to the next local maximum if the primary peak is negative-going. The repolarization time is the distance between the primary peak and the following inflection point (Trainito et al., 2019). In sum, five spike waveform features were extracted.

3.3. Extraction of Firing Properties from the Time Series Activity of Single-Units

For the firing patterns, the neuron's response consists of a set of spikes separated by long or short intervals (Figure 1C). The arrangement of differential interspike intervals (ISIs) defines the distinct neuronal firing pattern. A burst is a group of spikes with a series of successive short ISIs, while a pause is a group of spikes with a series of successive long ISIs (Figure 1C).

Robust Gaussian Surprise (RGS) method was performed for pause and burst quantification (Ko et al., 2012). The conventional template method lacks statistical significance since an ISI below 80 msec defines a burst while an ISI above 160 msec defines a pause (Grace and Bunney, 1984). The RGS method first normalizes the log-transformed ISI in each spike train (NLISI), which generates a cumulative Gaussian probability distribution. Burst threshold and pause threshold are defined based on the 0.5 percentile and 99.5 percentile of this central distribution, respectively (Ko et al., 2012). The burst threshold is estimated to be 2.58 times lower than the median average deviation, while the pause threshold is estimated to be 2.58 times higher than the median average deviation. Finally, NLISI in front or behind the threshold-defined seed set will be added to form a set of continuous strings of NLISIs, until the string reaches the minimum P-value (Ko et al., 2012). After the pause and burst quantification, the whole spike train was divided into bursts, pauses, and tonics. Electrophysiological features were extracted from the first 5 min of each recording session. The total duration of each firing pattern was extracted at first (Supplementary Figure 1). After duration,

intrapattern frequency, number of spikes, within pattern ISIs, and instantaneous firing frequency being extracted, their distributions were analyzed by calculating the following five values: mean, median, variation, kurtosis, and skewness (Supplementary Figure 1).

3.4. Dimensionality Reduction

After z-score standardization, principal component analysis (PCA) was used for dimensionality reduction (Jolliffe, 1986), and it aims to compute the new principal components (PCs) that are used to preserve most of the information in the original dataset.

t-SNE (*t*-distributed Stochastic Neighbor Embedding) was used to embed points with high dimensionality in low dimensions (Maaten et al., 2008), which is suitable to visualize natural clusters in high-dimensional data after PCA.

3.5. Unsupervised Classification Algorithms

Unsupervised learning can learn patterns and group data from unlabeled data. Five unsupervised classification algorithms were run 100 times to first examine their performance regarding pDA identification.

K-means. K-means is an iterative method, which aims to partition n observations into a given number of clusters (Lloyd, 1982). The algorithm starts with an initial guess for the centroids, as the cluster centers. Then, each observation will be placed into the closest cluster (Pihur et al., 2009). The centroids are then updated, and the whole process is repeated until the centroids no longer move. Given a data set $\{x_i\}_{i=1}^n$ and a given number of k clusters, this algorithm aims to minimize the sum of the squared Euclidean distance between data points and centroids as the following:

$$J(r, c) = \sum_{j=1}^k \sum_{i=1}^n r_{ij} \|x_i - c_j\|_2^2 \quad (1)$$

where c_j is the j_{th} cluster and $r_{ij} \in \{0,1\}$. If r_{ij} is equal to 1, x_i is assigned to c_j .

K-Medoids. Like K-means, k-medoids aims to minimize the sum of the dissimilarities between data points and cluster centers. However, the cluster center in K-medoids clustering is not the mean of data points in each cluster but an actual point in this cluster (Aggarwal et al., 1999). Thus, K-medoids is less sensitive than k-means to the outliers.

The Gaussian mixture model (GMM). Similar to K-Means, GMM is required to specify the number of clusters before fitting the model. In contrast to K-mean, which gives exactly how each data point is assigned to a specific cluster, GMM gives the probabilities of each data point belonging to each of the possible clusters (Reynolds, 2015). A data set $\{x_i\}_{i=1}^n$ is assumed to be generated from a mixture

of a given number k of Gaussian distributions, $\{N(x_i | \mu_j, \Sigma_j)\}_{j=1}^k$, where μ_j and Σ_j are the mean and the covariance of j_{th} Gaussian. Given the probability of observing x_i from j_{th} Gaussian is π_j , the probability of observing x_i from a mixture of Gaussian distributions is as the following:

$$P(x_i) = \sum_{j=1}^k \pi_j N(x_i | \mu_j, \Sigma_j) \quad (2)$$

Generally, it aims to maximize $\sum_{i=1}^n \ln(P(x_i))$, the GMM likelihood, by using the iterative Expectation-Maximization algorithm (Reynolds, 2015).

Density-based spatial clustering of applications with noise (DBSCAN). DBSCAN is a common method for class identification by locating high density regions (Ester et al., 1996). It is mainly based on two manually defined parameters, R (radius of the neighborhood around a point) and M (min number of neighbors). For one thing, if an R includes enough points within it, this area is called a dense area. For another, M determines the minimum number of neighbors required to define a cluster. A point is a core point if it has at least M neighbors around it, while a data point is a border point if it does not have enough points in the neighborhoods. An outlier is a point that is not reachable from any other point. After all data points are labeled, this algorithm connects core points that are neighbors and puts them in the same cluster.

One-class support vector machine (OCSVM). Support Vector Machines (SVMs) are most frequently used for classification (Christianini and Shawe-Taylor, 2000). OCSVM is similar to regular SVM except that it analyzes training data with only one class. OCSVM first embeds data in a feature space with high dimensionality by using the radial basis function kernel, the default kernel function within OCSVM (Chen et al., 1996). In this feature space, it aims to find a max-margin hyperplane to separate the origin and the data points.

3.6. Evaluations of Unsupervised Clustering Algorithms

Five unsupervised clustering algorithms were run 100 times. First, to evaluate the mean accuracy of each model, the phototagged DA neurons, which serve as externally provided class labels, were used to calculate the mean and standard deviation (SD) of the true positive rate (TP) and the proportion of pDA neurons among recorded neurons (%). Second, the best performance was selected based on a higher TP with a relatively low pDA (%) across 100 runs. To visualize the clustering results of each model in its best performance, 3D t-SNE plot was used for visualization (Figure 1D).

Third, to use the internal information to evaluate the goodness of a clustering structure, the Calinski-Harabasz index (Calinski and Harabasz, 1974) was calculated for GMM and K-means with different numbers of clusters. The Calinski-Harabasz index is the ratio of the sum of inter-clusters

variance to within-cluster variance for all clusters. With n number of observations in the cluster i , the inter-cluster variance is as the following:

$$V_B = \sum_{i=1}^k n_i ||m_i - m||^2 \quad (3)$$

, where m_i and m are the cluster centroid and the overall mean of the dataset, respectively.

With a data point x belonging to a given cluster, c_i , the within-cluster variance is as the following:

$$V_w = \sum_{i=1}^k \sum_{x \in c_i} ||x - m_i||^2 \quad (4)$$

The Calinski-Harabasz index is then calculated as the following:

$$V_{CH} = \frac{V_B}{V_w} \times \frac{(N-k)}{(k-1)} \quad (5)$$

A higher index score infers a better performance.

Last, to compare the stability of GMM and K-means model, K-means and GMM were run with a series of sample sizes from 19 to 489 for 100 times. Given each sample size, cells were randomly selected. The mean and SD of TP (%) and pDA (%) were calculated.

3.7. Predictive Analysis

To test the predictive efficiency of k-means clustering, the first strategy is to utilize the centroids in the training dataset from the k-means model. By calculating the Euclidean distance between data points from the testing dataset ($n=169$) and centroid in each cluster, each data point from the testing dataset is assigned to pDA or non-pDA cluster.

The second strategy is to train a linear SVM classifier on the training dataset with the labels generated by k-means model. This SVM classifier was evaluated using 5-fold cross-validation. 5-fold cross-validation, which is a resampling method, randomly divides the dataset into 5 groups, and the first group will be treated as a validation set (Anguita et al., 2005).

4. Results

4.1. Representative Feature Extractions and Dimensionality Reduction

54 features were extracted from extracellularly recorded neurons in total. Spike waveform properties, including total duration 1, total duration 2, half-width, peak-to-trough duration and repolarization time, were successfully extracted as shown in representative extractions (Figure 2A -

C). Pauses and bursts were quantitated by the RGS method ($p < 0.05$; Figure 2D). After that, the whole spike train was divided into bursts, pauses, and tonics. Features of the firing properties (Supplementary Figure 1) were extracted (Prévost-Solié et al., 2020).

320 neurons were randomly selected as the training dataset, and the remaining 169 neurons would be the testing dataset. The extracted 54 features from either the training dataset or the testing dataset were z-scored on mean and SD of the training dataset. PCA was performed on these 54 features, and 38 PCs were chosen to explain 99.3% variances of the training dataset (Figure 3A and B). t-SNE with PCA initialization shows a clear cluster of phototagged DA neurons mixed with additional non-phototagged neurons (Figure 3C), indicating that these mixed neurons could form the pDA cluster.

4.2 Unsupervised Clustering Comparisons: K-Means and GMM are Superior Regarding their Accuracy

To identify the pDA cluster in an unsupervised way, five unsupervised clustering methods were performed and compared. The accuracy of these models was first evaluated by calculating the proportion of phototagged DA neurons assigned to the putative DA group, named TP (%). Meanwhile, the proportion of putative DA neurons among all neurons was calculated to be weighted with a reference value of 70% (Ungless and Grace, 2012), serving as another measure of the model accuracy. Each unsupervised clustering method was run 100 times. The best performance of each model was shown in Figure 4A-E, in which pDA and non-pDA clusters are visualized in 3D t-SNE plot with green and red circular borders representing k-means cluster labels.

k-means. The k-means performed exceptionally well in identifying photo-tagged DA neurons (Figure 4 A and F), and its best performance reached a TP of 100%, suggesting that it can assign all photo-tagged DA neurons into the pDA cluster (Figure 4A). Its best performance shows a pDA of 75.94%, which is very close to the reference value 70%, strengthening its accuracy. Besides, it yielded a mean TP of 97.93 ± 5.56 % (mean \pm S.D.) and a mean pDA of 83.79 ± 10.26 %, indicating a certain level of stability.

k-medoids. k-medoids was completely inaccurate because of its inability to distinguish pDA from non-pDA (pDA in best performance: 99.69%, mean pDA = 99.69 ± 14.9 %; Figure 4C and G).

DBSCAN. The performance of DBSCAN varied with different manually defined parameter sets. A parameter set, in which $R=3$ and $M=12$, was chosen because of its good performance among 247 parameter sets (Supplementary Figure 2). The best performance of DBSCAN (TP = 82.5%, pDA = 89.38%; Figure 4B) is worse than the best performance in k-means due to DBSCAN's lower TP and higher pDA. Furthermore, the mean TP of DBSCAN (82.32 ± 3.01 %) is lower than k-means (97.93 ± 5.56 %). Therefore, k-means is superior to DBSCAN.

GMM. GMM plays relatively well. The GMM with a shared covariance matrix performed best in identifying true phototagged DA neurons (mean TP = $100 \pm 0\%$), but it did not perform well in distinguishing pDA from non-pDA (pDA = $98 \pm 0\%$). In contrast, a GMM without a shared covariance matrix yielded extremely good performance with TP of 92.5% and pDA of 80.31% (Figure 4D), but its convergence was not guaranteed. Therefore, a GMM without a shared covariance matrix could be a promising supervised clustering method for the pDA assignment.

OCSVM. OCSVM was utterly inaccurate, exhibiting a mean TP of $56.77 \pm 0.11\%$ and a mean pDA of $48.43 \pm 0.06\%$. The TP in the best performance, 62.50%, is also very low (Figure 4E). Thus, this algorithm is unable to identify most phototagged DA neurons.

Because of the relatively good performance of k-means and GMM in terms of accuracy, these two models would subsequently be evaluated.

4.3 K-Means Outperforms GMM

First, for real-world applications, a great model is expected to manage different sample sizes. The stability of these two models with different sample sizes was tested as shown in Figure 5A and B. TP (%) and pDA (%) in k-means changes slowly across different sample sizes, whereas these two values vary a lot across different sample sizes in GMM (Figure 5A and B). It suggests that k-means is more stable than GMM with varying sample sizes. Second, the Calinski-Harabasz index was used to evaluate the goodness of a clustering structure by using the internal information. This index is the ratio of the sum of inter-clusters variance to within-cluster variance for all clusters. This index is higher in k-means model than that in GMM regardless of the number of clusters, indicating that k-means can generate a better clustering structure (Figure 5C and D). Besides, in k-means, the number of 2 clusters has the best performance, validating the presence of two clusters (Figure 5C). When taking the accuracy, the Calinski-Harabasz index, and the stability into consideration, the k-means outperforms the GMM and becomes the best unsupervised clustering method for pDA identification.

4.4 K-Means Clustering Classification of pDA and Non-pDA Neurons

From k-means clustering, spikes within the pDA cluster have comparatively wider durations (Figure 6A and F). Consistent with previous research in VTA DA neurons (Ungless et al., 2004), neurons classified as pDA neurons by k-means had a higher firing rate (Figure 6B-F) and a wider half-width, also named spike width (Figure 6C). Interestingly, wider peak-to-trough duration was observed in pDA neurons, while the repolarization time of neurons from the pDA cluster and the non-pDA cluster almost overlaps (Figure 6D and E).

4.5 The Clustering Structure in K-means Model could Predict pDA Neurons in the Testing Dataset

To examine whether the clustering structure in the training dataset generated from the k-means model could predict pDA neurons when analyzing the testing dataset ($n = 169$), two strategies were used. The first strategy is to use the centroids in the training dataset from the k-means model. By calculating the Euclidean distance between data points from the testing dataset and centroid in each cluster, each data point from the testing dataset would be assigned to pDA or non-pDA cluster. As a result, the performance is well, exhibiting a TP of 97.67% and a pDA of 75.54% (Figure 7A).

The second strategy is to train a linear SVM classifier to classify a dataset into two classes by using a straight line. The training dataset trained the linear SVM classifier with the labels generated by the k-means model. This SVM classifier was evaluated using 5-fold cross-validation, yielding an extremely high precision ($> 99\%$) calculated by the Confusion Matrix (Supplementary Figure 3). Nevertheless, the performance of this classifier is worse than the Euclidean distance-based method (TP = 95.24%, pDA = 78.11%; Figure 7B).

Therefore, both strategies can predict pDA neurons in the testing dataset, and Euclidean distance clustering predictions were superior to the linear SVM.

5. Discussion

In this report, five unsupervised clustering methods were employed and compared for pDA neuron identification from tetrode recordings in behaving mice. k-means clustering with two clusters was evaluated as the best clustering model to distinguish pDA neurons from non-pDA neurons, allowing unbiased and relatively stable pDA identification with high accuracy. It is highly inspiring that unsupervised clustering methods could link electrophysiological properties' analysis with opto-tagging for neuron classification and identification. Feature extraction methods and the unsupervised clustering methods introduced here could advance future research in the midbrain DA system, particularly with respect to dissecting DA neurons *in vivo* with Neuropixels (Steinmetz et al., 2021).

5.1. K-means Outperforms Other Unsupervised Clustering Methods on pDA Identification

It is surprising that k-means performed the best among the five models in terms of accuracy (Figure 4) and outperformed the GMM due to higher stability and higher Calinski-Harabasz index with two clusters (Figure 5). K-means is not sensitive to different sample sizes, indicating that its performance largely is independent of how big the dataset is. Meanwhile, the clustering structure obtained from the k-means model can also predict new neurons from the testing dataset with a high accuracy (Figure 7). However, despite k-means's insensitivity across different sample sizes, the SD of its performance is a little high, suggesting a certain level of instability given specific sample size.

One possible reason for such instability is that the initial guess of centroids from k-means in a dataset is random. Consequently, the clustering structure is highly dependent on centroids initializations, and this algorithm does not guarantee a convergence to a global optimum. To improve k-means, several optimized k-means algorithms can be used to achieve a better centroids initialization, such as Particle Swarm Optimization (Pednekar, 2019) and another algorithm that makes the initial centroids spread as far as possible (Barakbah and Helen, 2005).

In contrast to K-mean, which gives exactly how each data point is assigned to a certain cluster, GMM is more flexible, which gives the probabilities of the data point belonging to each of the possible clusters. GMM has been applied in DA neurons classification (Prévost-Solié et al., 2020), but its efficiency and stability have not yet been compared with other unsupervised clustering methods in this context. In this project, k-means outperformed GMM (without shared covariance matrix) in terms of accuracy (Figure 4A and D). The specific structure of the pDA neurons cluster in high dimensions within the dataset could be a potential reason, and such structure may be more fitted to the boundaries in k-means than that in GMM. This is because each cluster in GMM is fitted as a Gaussian distribution, and in contrast to circular boundaries within k-means, the boundaries in GMM could be elliptical due to the presence of a covariance matrix.

DBSCAN did not perform better than GMM and k-means (Figure 4A, B, D, F, and G). Although DBSCAN is robust to discover arbitrarily shaped clusters, it usually fails to deal with a sparse dataset and to identify a cluster with varying density, since the R-M combination cannot be appropriately chosen for all clusters. The dataset is highly sparse due to the heterogeneity of both pDA and non-pDA neurons, which could explain the DBSCAN's disappointing performance. Meanwhile, since the performance of DBSCAN is highly sensitive to different R-M combinations, an improper R-M set could lead to poor performance. Even though this project has selected the best set among 247 R-M combinations, it is necessary to conduct hill-climbing, an iterative algorithm for local search, to find a more appropriate set (Johnson and Jacobson, 2002).

Although both k-means and k-medoids belong to partitional clustering methods. K-medoids failed to distinguish pDA neurons from non-pDA neurons (Figure 4C and G). In k-medoids, its medoid but not its mean is the center of a cluster, and the medoid is the cluster member, which minimizes the dissimilarities between itself and other cluster points. Thus, k-medoids is less sensitive to outliers than k-means, serving as a more median-type approach. However, since the distribution of neurons across different dimensions is highly sparse with varying density, k-medoids might regard some true DA neurons as outliers and introduce errors in the entire clustering structure.

As for OCSVM, it can be used for classifying binary samples with a highly imbalanced class distribution in which very few samples are in the minority (Shravan Kumar and Ravi, 2017). Therefore, a potential explanation for its poor performance in our dataset could be the cluster-distribution regularity (Figure 4E).

5.2.K-means Clustering Classification of Spike Waveforms and Firing Rate in pDA and Non-pDA Clusters

Classified pDA neurons from k-means clustering had higher firing rate and wider spike width (Figure 6C). This is consistent with previous research in VTA DA neurons from which the action potential waveform during cell-attached recordings shows a good separation with a 99% confidence interval between GFP neurons and GAD-GFP neurons (presumed GABAergic) (Ungless et al., 2004). Patch-clamp recordings from the mouse VTA from other studies also support it, showing no overlap in spike width from TH-positive DA neurons and TH-negative non-DA neurons (Ford et al., 2006) (Chieng et al., 2011). Surprisingly, there is an overlapping region of the spike width between pDA neurons and non-pDA neurons (Figure 6C).

Interestingly, wider peak-to-trough duration was also observed in pDA neurons (Figure 6D), advancing our understandings of electrophysiological features of VTA DA neurons. In addition, the repolarization time between pDA and non-DA neurons is almost overlapping with each other (Figure 6E), indicating that this feature may be redundant during the unsupervised clustering.

5.3. Limitations of PCA For Dimensionality Reduction

Neuron classification is challenging because of the difficulty of choosing the best features to define different neuron types. In this report, as a feature transformation technique, PCA was used to reduce the dimensionality of extracted 54 features (Figure 3A and B). Although PCA performs well in revealing latent structure in the dataset, since it still preserves the relative distance between different samples, its effectiveness will be primarily reduced when there are multiple irrelevant features that would hide clusters in noisy data. A single-channel recording research suggested that some features that are not necessarily dependent on neuron type could bring ambiguity to the clustering structure (Parikh et al., 2018). Thus, the importance of each feature should be estimated. Feature selection can alleviate this issue through eliminating irrelevant and redundant dimensions in the whole dataset (Blum and Langley, 1997). A more advanced method, subspace clustering, as an extension of feature selection, can find clusters that exist in different subspaces of the same dataset (Parsons et al., 2004). Thus, a combination of subspace clustering and PCA could be a better solution for dimensionality reduction.

5.4. Conduct Opto-Tagging On Other Midbrain Neurons and Explore Semi-Supervised Classifier

One limitation of this project is that the dataset only consists of true-positive data, and hence, unsupervised methods but not supervised methods are preferred. One study has applied opto-tagging to identify glutamatergic and GABAergic neurons in the cerebellar nucleus of *VGLUT2-cre* and *GAD-cre* mice (Özcan et al., 2020). In the future, opto-tagging can be applied to other neuron types in the midbrain, and as a result, a group of midbrain neurons can be labeled by ground-truth

neuron types. Then, a semi-supervised classifier can be trained. A semi-supervised model can deal with a dataset with both the labeled and unlabeled data (Zhu and Goldberg, 2009), providing the benefits of both unsupervised and supervised learning. Supervised classification methods were found to outperform unsupervised algorithms in several studies regarding neuron type classifications (Guerra et al., 2010) (Vasques et al., 2016), but these methods are limited in the requirement of a large amount of labeled data. In contrast, a semi-supervised classifier can be trained by a small number of labeled neurons to predict whether a new neuron is pDA or not.

6. Conclusion

In sum, after feature extraction and dimensionality reduction, k-means clustering with two clusters was evaluated as the best clustering model to distinguish pDA neurons from non-pDA neurons, achieving an unbiased pDA identification with high accuracy. It is highly inspiring that unsupervised clustering methods could link electrophysiological properties' analysis with opto-tagging for neuron classification and identification. However, applying opto-tagging on other midbrain neurons is essential to achieve a more ground-truth pDA identification.

7. Reference

- Aggarwal, C., Wolf, J., Yu, P., Procopiuc, C. and Park, J., 1999. Fast algorithms for projected clustering. *ACM SIGMOD Record*, 28(2), pp.61-72.
- Anguita, D., Ridella, S. and Riveccio, F., 2005. K-fold generalization capability assessment for support vector classifiers. *Proceedings. 2005 IEEE International Joint Conference on Neural Networks*, vol. 2, pp. 855-858.
- Barakbah, A.R., and Afrida H., 2005. Optimized K-means: an algorithm of initial centroids optimization for K-means. *Proc. Seminar on Soft Computing, Intelligent System, and Information Technology (SIIT), Surabaya*.
- Blum, A. and Langley, P., 1997. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2), pp.245-271.
- Boser, B., Guyon, I. and Vapnik, V., 1992. A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*,.
- Calinski, T. and Harabasz, J., 1974. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1), pp.1-27.
- Chen, S., Chng, E. and Alkadhimi, K., 1996. Regularized orthogonal least squares algorithm for constructing radial basis function networks. *International Journal of Control*, 64(5), pp.829-837.
- Chieng, B., Azriel, Y., Mohammadi, S. and Christie, M., 2011. Distinct cellular properties of identified dopaminergic and GABAergic neurons in the mouse ventral tegmental area. *The Journal of Physiology*, 589(15), pp.3775-3787.
- Cohen, J., Haesler, S., Vong, L., Lowell, B. and Uchida, N., 2012. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature*, 482(7383), pp.85-88.
- Coomans, D. and Massart, D., 1982. Alternative k-nearest neighbour rules in supervised pattern recognition. *Analytica Chimica Acta*, 136, pp.15-27.
- Cover, T. and Hart, P., 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), pp.21-27.
- Cristianini, N. and Shawe-Taylor, J., 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511801389.
- Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C., Hassabis, D., Munos, R. and Botvinick, M., 2020. A distributional code for value in dopamine-based reinforcement learning. *Nature*, 577(7792), pp.671-675.
- DeFelipe, J., López-Cruz, P., Benavides-Piccione, R., Bielza, C., Larrañaga, P., Anderson, S., Burkhalter, A., Cauli, B., Fairén, A., Feldmeyer, D., Fishell, G., Fitzpatrick, D., Freund, T., González-Burgos, G., Hestrin, S., Hill, S., Hof, P., Huang, J., Jones, E., Kawaguchi, Y., Kisvárdy, Z., Kubota, Y., Lewis, D., Marín, O., Markram, H., McBain, C., Meyer, H., Monyer, H., Nelson, S., Rockland, K., Rossier, J., Rubenstein, J., Rudy, B., Scanziani, M., Shepherd, G., Sherwood, C., Staiger, J., Tamás, G., Thomson, A., Wang, Y., Yuste, R. and Ascoli, G., 2013. New insights into the classification and nomenclature of cortical GABAergic interneurons. *Nature Reviews Neuroscience*, 14(3), pp.202-216.
- Ester, M.; Kriegel, H.-P.; Sander, J. & Xu, X. (1996), A Density-Based Algorithm for Discovering

Clusters in Large Spatial Databases with Noise, *in* 'Proc. of 2nd International Conference on Knowledge Discovery and', pp. 226-231.

Ford, C., 2006. Properties and Opioid Inhibition of Mesolimbic Dopamine Neurons Vary according to Target Location. *Journal of Neuroscience*, 26(10), pp.2788-2797.

Gale SD and Perkel DJ (2006) Physiological properties of zebra finch ventral tegmental area and substantia nigra pars compacta neurons. *J Neurophysiol*, **96**:2295–2306.

Grace AA, Bunney BS. The control of firing pattern in nigral dopamine neurons: burst firing. *J Neurosci*. 1984 Nov;4(11):2877-90. doi: 10.1523/JNEUROSCI.04-11-02877.1984. PMID: 6150071; PMCID: PMC6564720.

Guerra, L., McGarry, L., Robles, V., Bielza, C., Larrañaga, P. and Yuste, R., 2010. Comparison between supervised and unsupervised classifications of neuronal cell types: A case study. *Developmental Neurobiology*, 71(1), pp.71-82.

Jardine, N. and Sibson, R., 1968. The Construction of Hierarchic and Non-Hierarchic Classifications. *The Computer Journal*, 11(2), pp.177-184.

Jia, X., Siegle, J., Bennett, C., Gale, S., Denman, D., Koch, C. and Olsen, S., 2019. High-density extracellular probes reveal dendritic backpropagation and facilitate neuron classification. *Journal of Neurophysiology*, 121(5), pp.1831-1847.

Johnson, A. and Jacobson, S., 2002. On the convergence of generalized hill climbing algorithms. *Discrete Applied Mathematics*, 119(1-2), pp.37-57.

Jolliffe, I., 1986. Principal Component Analysis and Factor Analysis. *Principal Component Analysis*, pp.115-128.

Jun, J., Steinmetz, N., Siegle, J., Denman, D., Bauza, M., Barbarits, B., Lee, A., Anastassiou, C., Andrei, A., Aydın, Ç., Barbic, M., Blanche, T., Bonin, V., Couto, J., Dutta, B., Gratiy, S., Gutnisky, D., Häusser, M., Karsh, B., Ledochowitsch, P., Lopez, C., Mitelut, C., Musa, S., Okun, M., Pachitariu, M., Putzeys, J., Rich, P., Rossant, C., Sun, W., Svoboda, K., Carandini, M., Harris, K., Koch, C., O'Keefe, J. and Harris, T., 2017. Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679), pp.232-236.

Jung, Y., 2017. Multiple predictingK-fold cross-validation for model selection. *Journal of Nonparametric Statistics*, 30(1), pp.197-215.

Kim, H., Malik, A., Mikhael, J., Bech, P., Tsutsui-Kimura, I., Sun, F., Zhang, Y., Li, Y., Watabe-Uchida, M., Gershman, S. and Uchida, N., 2020. A Unified Framework for Dopamine Signals across Timescales. *Cell*, 183(6), pp.1600-1616.e25.

Ko, D., Wilson, C., Lobb, C. and Paladini, C., 2012. Detection of bursts and pauses in spike trains. *Journal of Neuroscience Methods*, 211(1), pp.145-158.

Lammel, S., Hetzel, A., Häckel, O., Jones, I., Liss, B. and Roeper, J., 2008. Unique Properties of Mesoprefrontal Neurons within a Dual Mesocorticolimbic Dopamine System. *Neuron*, 57(5), pp.760-773.

Lammel, S., Steinberg, E., Földy, C., Wall, N., Beier, K., Luo, L. and Malenka, R., 2015. Diversity of Transgenic Mouse Models for Selective Targeting of Midbrain Dopamine Neurons. *Neuron*, 85(2), pp.429-438.

- Lima, S., Hromádka, T., Znamenskiy, P. and Zador, A., 2009. PINP: A New Method of Tagging Neuronal Populations for Identification during In Vivo Electrophysiological Recording. *PLoS ONE*, 4(7), p.e6099.
- Lloyd, S., 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), pp.129-137.
- Ludwig, K., Miriani, R., Langhals, N., Marzullo, T. and Kipke, D., 2011. Use of a Bayesian maximum-likelihood classifier to generate training data for brain-machine interfaces. *Journal of Neural Engineering*, 8(4), p.046009.
- Margolis, E., Lock, H., Hjelmstad, G. and Fields, H., 2006. The ventral tegmental area revisited: is there an electrophysiological marker for dopaminergic neurons?. *The Journal of Physiology*, 577(3), pp.907-924.
- Markham MR, Stoddard PK. Adrenocorticotrophic hormone enhances the masculinity of an electric communication signal by modulating the waveform and timing of action potentials within individual cells. *J Neurosci*. 2005;25:8746–8754.
- Matsumoto, M. and Hikosaka, O., 2009. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, 459(7248), pp.837-841.
- McGarry, 2010. Quantitative classification of somatostatin-positive neocortical interneurons identifies three interneuron subtypes. *Frontiers in Neural Circuits*,.
- Mensi, S., Naud, R., Pozzorini, C., Avermann, M., Petersen, C. and Gerstner, W., 2012. Parameter extraction and classification of three cortical neuron types reveals two distinct adaptation mechanisms. *Journal of Neurophysiology*, 107(6), pp.1756-1775.
- Mohebi, A., Pettibone, J., Hamid, A., Wong, J., Vinson, L., Patriarchi, T., Tian, L., Kennedy, R. and Berke, J., 2019. Dissociable dopamine dynamics for learning and motivation. *Nature*, 570(7759), pp.65-70.
- Morales, M. and Margolis, E., 2017. Ventral tegmental area: cellular heterogeneity, connectivity and behaviour. *Nature Reviews Neuroscience*, 18(2), pp.73-85.
- Morales, M. and Margolis, E., 2017. Ventral tegmental area: cellular heterogeneity, connectivity and behaviour. *Nature Reviews Neuroscience*, 18(2), pp.73-85.
- Noble, W., 2006. What is a support vector machine?. *Nature Biotechnology*, 24(12), pp.1565-1567.
- Özcan, O., Wang, X., Binda, F., Dorgans, K., De Zeeuw, C., Gao, Z., Aertsen, A., Kumar, A. and Isope, P., 2019. Differential Coding Strategies in Glutamatergic and GABAergic Neurons in the Medial Cerebellar Nucleus. *The Journal of Neuroscience*, 40(1), pp.159-170.
- Parikh, R., 2018. Large-scale neuron cell classification of single-channel and multi-channel extracellular recordings in the anterior lateral motor cortex.
- Parsons, L., Haque, E. and Liu, H., 2004. Subspace clustering for high dimensional data. *ACM SIGKDD Explorations Newsletter*, 6(1), pp.90-105.
- Pednekar, A.M., 2018. 'Ramond-Ramond cohomology and O(D,D) T-duality'. [Preprint]. Available at: arXiv:1904.09098 (Accessed: 1 May 2021).
- Pihur, V., Brock, G. and Datta, S., 2009. Cluster Validation for Microarray Data: An Appraisal. *Advances in Multivariate Statistical Methods*, pp.79-94.

- Prévost-Solié, C., Girard, B., Righetti, B., Tapparel, M. and Bellone, C., 2020. Dopamine neurons of the VTA encode active conspecific interaction and promote social learning through social reward prediction error.
- Reynolds, D., 2015. Gaussian Mixture Models. *Encyclopedia of Biometrics*, pp.827-832.
- Schultz, W., Dayan, P. and Montague, P., 1997. A Neural Substrate of Prediction and Reward. *Science*, 275(5306), pp.1593-1599.
- Shravan Kumar, B. and Ravi, V., 2017. Text Document Classification with PCA and One-Class SVM. *Advances in Intelligent Systems and Computing*, pp.107-115.
- Trainito, C., von Nicolai, C., Miller, E. and Siegel, M., 2019. Extracellular Spike Waveform Dissociates Four Functionally Distinct Cell Classes in Primate Cortex. *Current Biology*, 29(18), pp.2973-2982.e5.
- Ungless, M. and Grace, A., 2012. Are you or aren't you? Challenges associated with physiologically identifying dopamine neurons. *Trends in Neurosciences*, 35(7), pp.422-430.
- Ungless, M., 2004. Uniform Inhibition of Dopamine Neurons in the Ventral Tegmental Area by Aversive Stimuli. *Science*, 303(5666), pp.2040-2042.
- Van der Maaten, L. J. P., and Hinton, G. E., Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605)
- Vapnik, V., 2000. *The Nature of Statistical Learning Theory*. New York, NY: Springer New York.
- Vasques, X., Vanel, L., Villette, G. and Cif, L., 2016. Morphological Neuron Classification Using Machine Learning. *Frontiers in Neuroanatomy*, 10.
- Volkow, N. and Morales, M., 2015. The Brain on Drugs: From Reward to Addiction. *Cell*, 162(4), pp.712-725.
- Watabe-Uchida, M., Zhu, L., Ogawa, S., Vamanrao, A. and Uchida, N., 2012. Whole-Brain Mapping of Direct Inputs to Midbrain Dopamine Neurons. *Neuron*, 74(5), pp.858-873.
- Wise, R., 2004. Dopamine, learning and motivation. *Nature Reviews Neuroscience*, 5(6), pp.483-494.
- Zhu, X. and Goldberg, A., 2009. Introduction to Semi-Supervised Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1), pp.1-130.

Figures, diagrams or tables

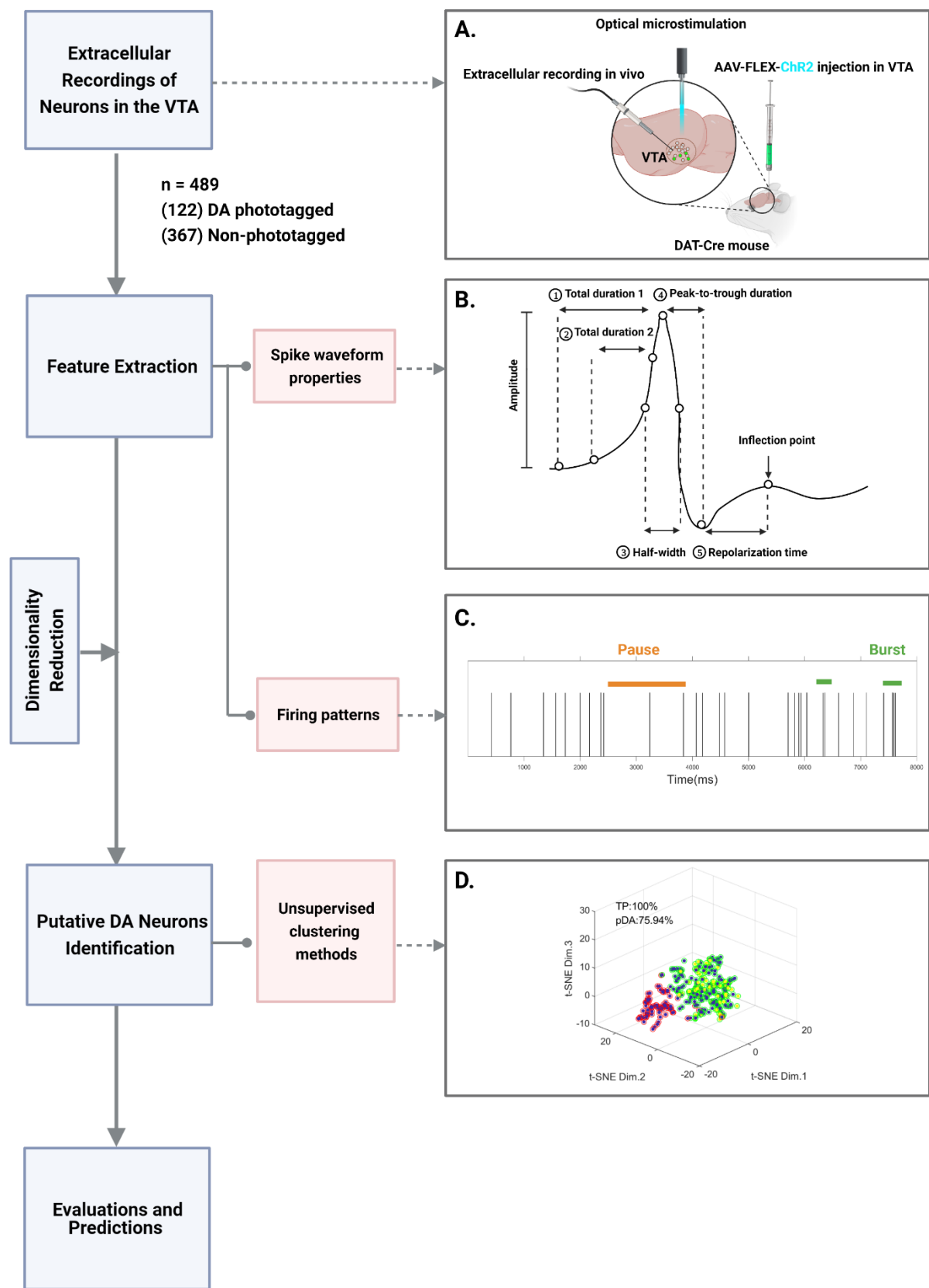
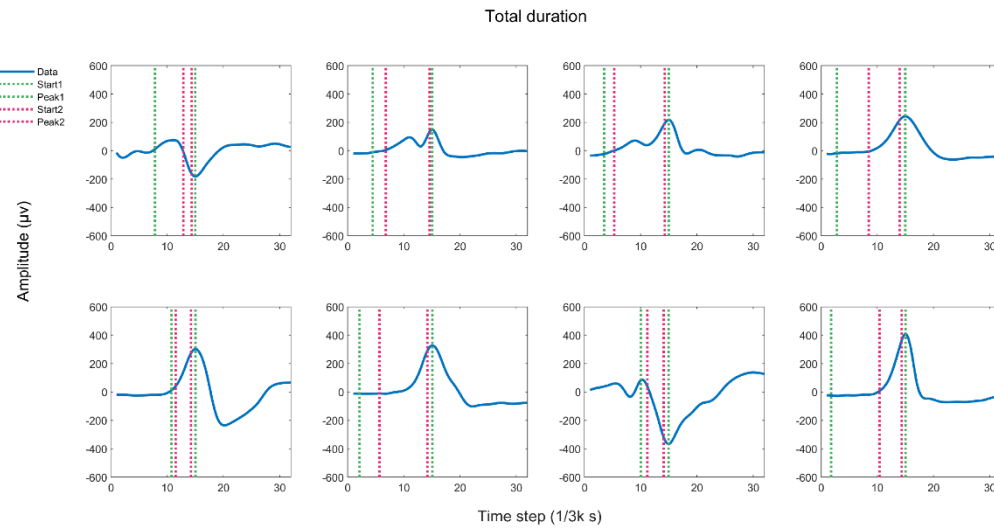


Figure 1. Workflow of unsupervised clustering for putative dopamine neurons identification

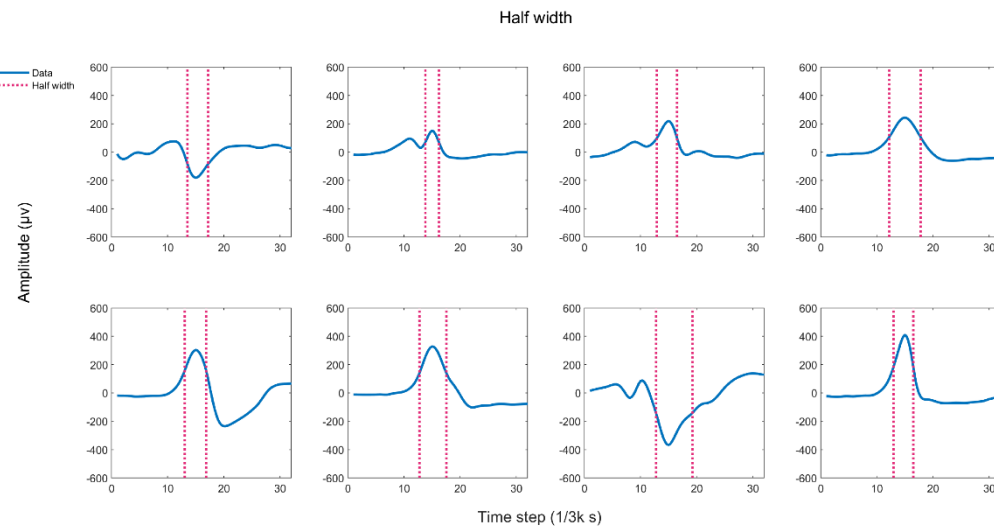
The sketch of the workflow contains five main steps: extracellular recordings of neurons in the VTA, feature extraction, dimensionality reduction, putative DA (pDA) neurons identification by unsupervised clustering methods, and model evaluations and predictions.

- (A) The VTA of male DAT-Cre mice were injected with Cre-dependent AAV virus expressing ChR2. Pulses of light were delivered to activate ChR2-expressing neurons after a virtual-reality reward delivery task.
- (B) Five spike waveform properties.
- (C) The spike trains from an example neuron. Orange and green horizontal bars represent pause and burst respectively.
- (D) An example 3D t-SNE scatter plot for extracellularly recorded neurons in the VTA. For unsupervised clustering under best performance, set circular borders on them according to their membership (non-pDA, red borders; pDA, green borders) determined by a specific unsupervised clustering.

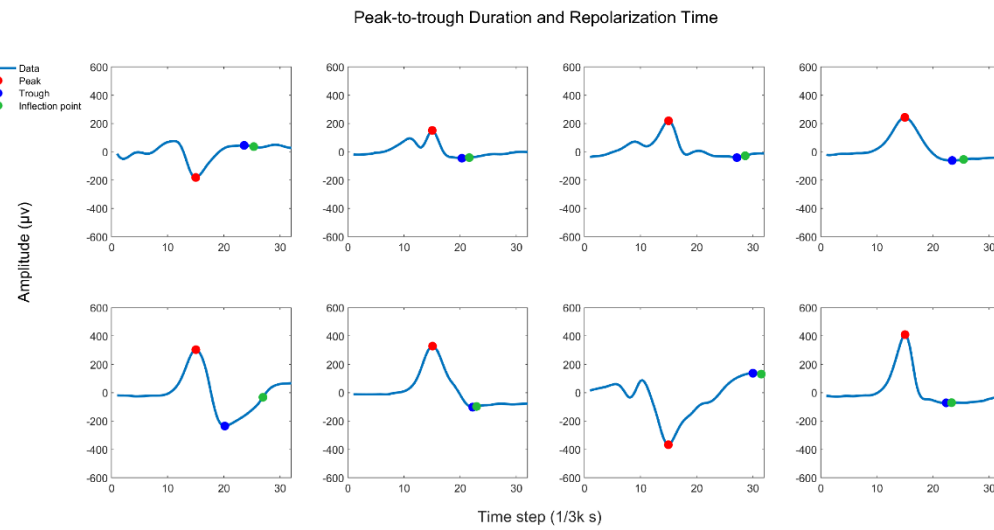
A



B



C



D

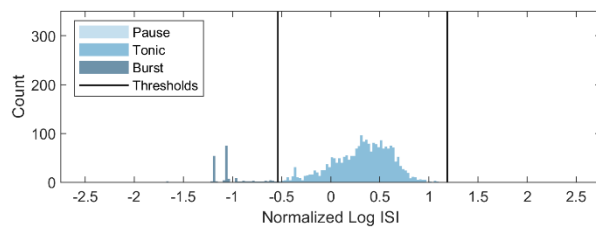


Figure 2. Representative extractions of spike waveform properties and firing properties

- (A) Total duration 1 and Total duration 2 extracted from 8 example neurons. The time courses of spike waveform amplitude from each neuron (blue line). Total duration 1 is the distance between green dotted lines, while Total duration 2 is the distance between pink dotted lines.
- (B) Half-width extracted from 8 example neurons. Half-width is the duration between pink dotted lines.
- (C) Peak-to-trough duration and repolarization time extracted from 8 example neurons. The red, blue, and green points in each subplot correspond to peak, trough, and inflection points respectively. Peak-to-trough duration represents the distance between peak and trough, while repolarization time is the distance between trough and inflection point.
- (D) Normalized log inter-spike intervals (ISIs) of a representative neuron in the VTA. Two dark vertical lines define the burst and pause thresholds. Different shades of blue represent pause, tonic and burst respectively.

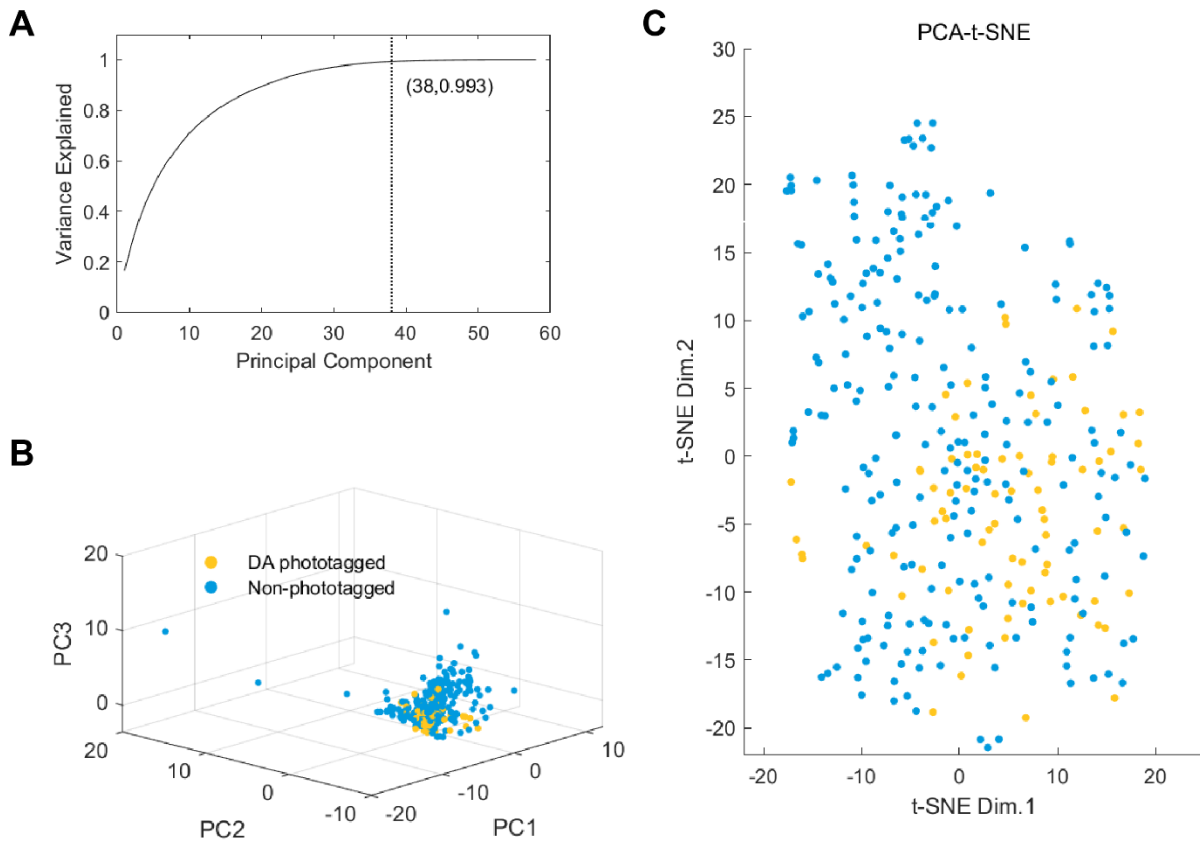


Figure 3. Principal component analysis for dimensionality reduction and t-SNE

(A) Variance explained by principal components (PCs). 38 PCs explain 99.3% variances of the training dataset.

(B) 3D scatter plot of the first three PCs showing the clustering of phototagged DA neurons (yellow points; $n = 80$) and non-phototagged neurons (blue points; $n = 240$).

(C) 2D t-SNE scatter plot with PCA initialization.

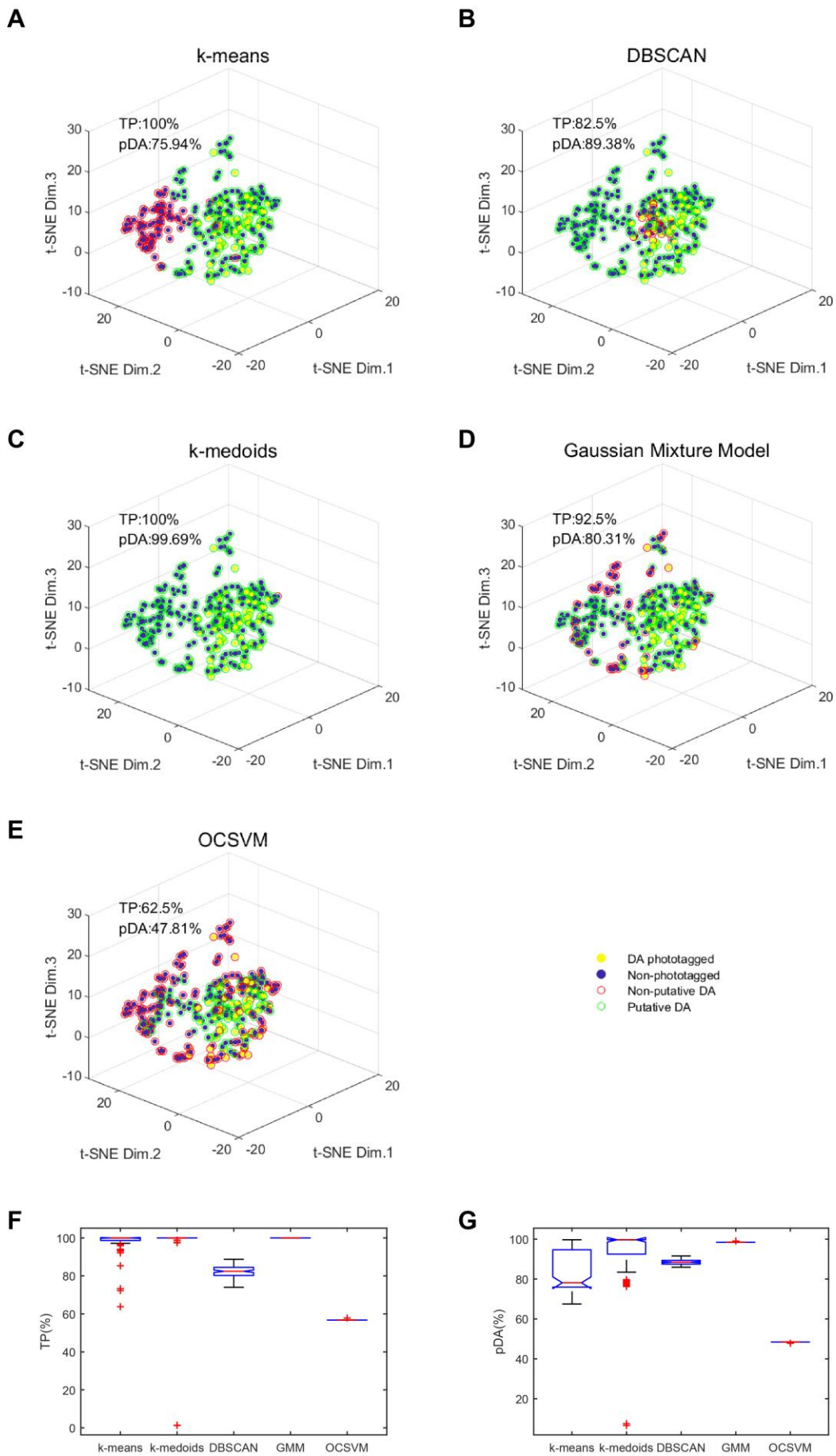


Figure 4. Unsupervised clustering of extracellularly recorded neurons in the VTA

(A - E) 3D t-SNE scatter plot for extracellularly recorded neurons in the VTA (phototagged DA

neurons, yellow, n=80; non-phototagged neurons, dark blue, n=240). For unsupervised clustering under best performance, set circular borders on them according to their membership (non-pDA, red borders; pDA, green borders) determined by k-means, DBSCAN, k-medoids, GMM, and OCSVM. True positive rate (TP) (%) and the proportion of putative DA neurons among all recorded neurons (%) were calculated in each clustering model.

(F) TP (%) of various unsupervised clustering methods. Each method for 100 times.

(G) Proportion of putative DA neurons of various unsupervised clustering methods. Each method for 100 times.

Box plots indicate the median (red center line), first quartiles (the bottom and top edges), minimum/maximum values (whiskers), and outliers (red +).

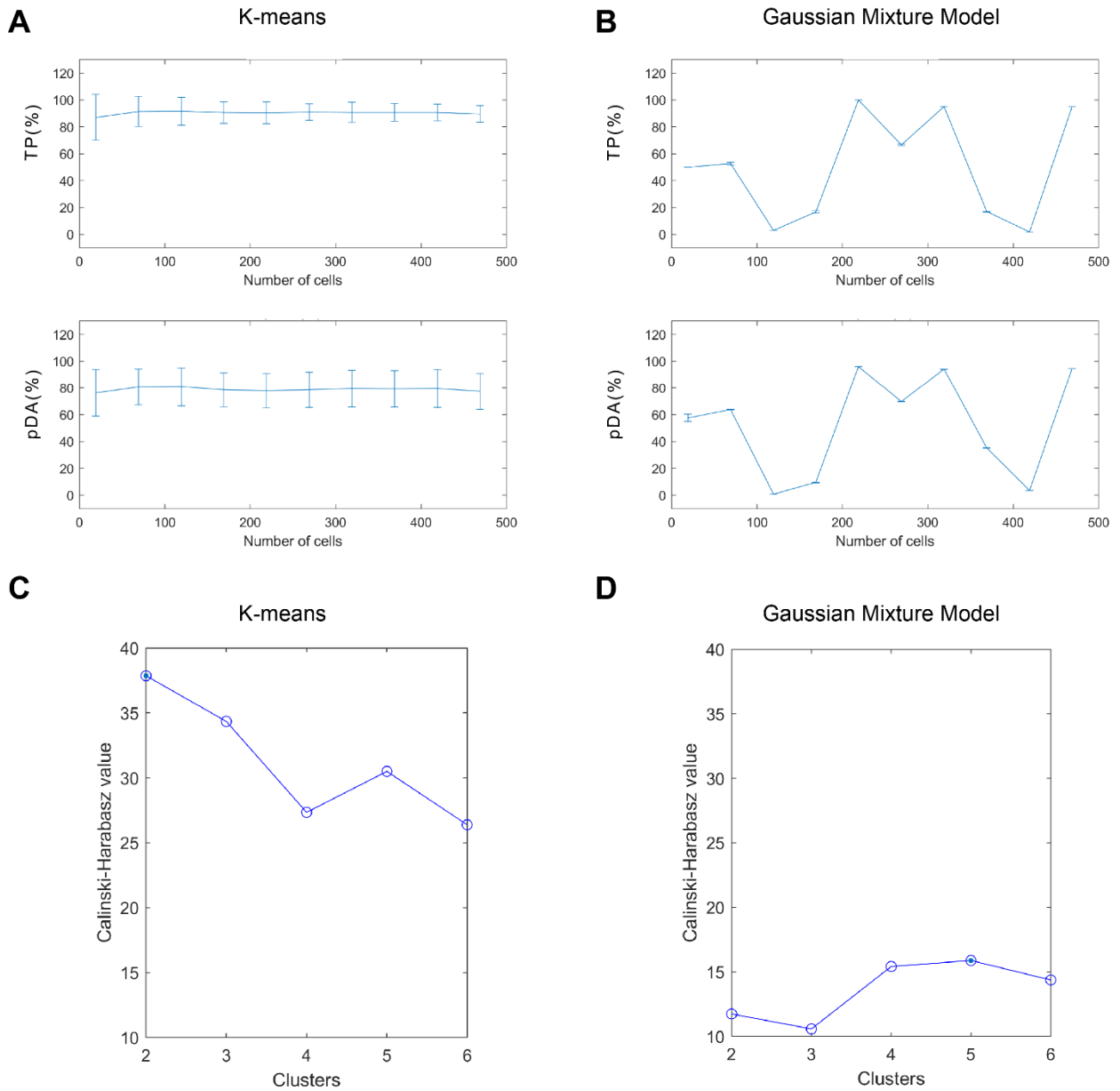


Figure 5. Evaluation of k-means and GMM

(A) TP (%) and pDA (%) across different sample sizes ranging from 19 to 489 based on k-means clustering. Plotted in mean \pm S.D.

(B) TP (%) and pDA (%) across different sample sizes ranging from 19 to 489 based on k-means clustering. Plotted in mean \pm S.D.

(C) The Calinski-Harabasz index with varying clusters in k-means. The best number of cluster is marked by a filled blue point.

(D) The Calinski-Harabasz index with varying clusters in GMM. The best number of cluster is marked by a filled blue point.

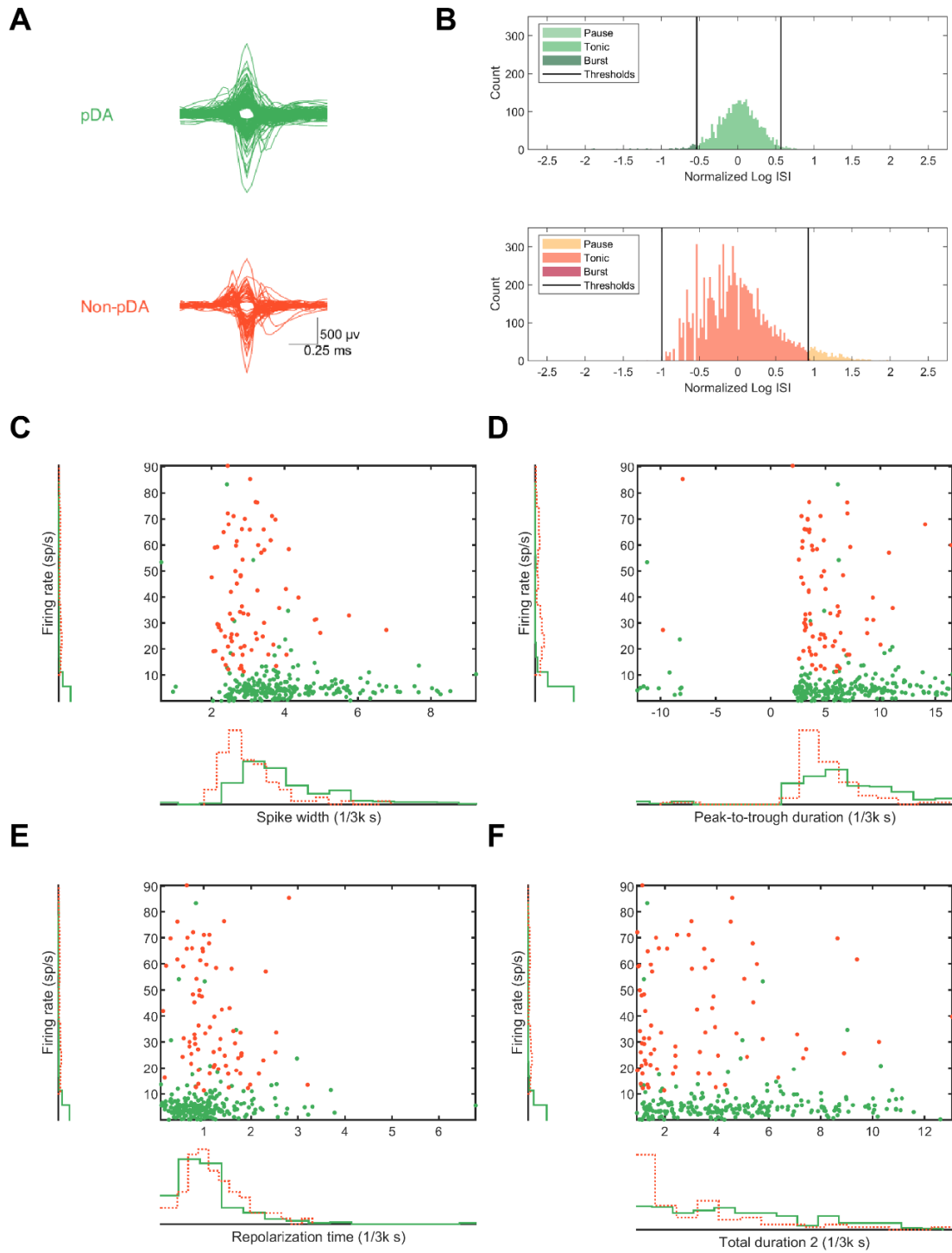


Figure 6. k-means clustering of extracellular waveforms and firing patterns from pDA and non-pDA clusters

(A) Normalized waveforms for units from pDA (green) and non-pDA (red) clusters (n=243 in the pDA cluster; n = 77 in the non-pDA cluster).

(B) Normalized log inter-spike intervals (ISIs) of representative neurons from the pDA cluster and the non-pDA cluster. Two thresholds define the burst and pause thresholds.

(C-F) Scatter plot with histograms for firing rates versus spike waveform properties of clustered neurons. Each data point represents a neuron. The baseline firing rate is plotted in the ordinate, spike width, peak-to-trough duration, repolarization time and total duration are plotted in the abscissa.

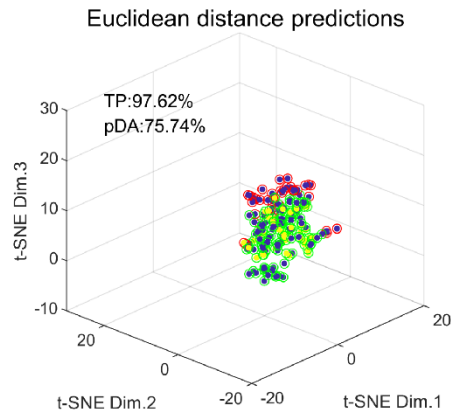
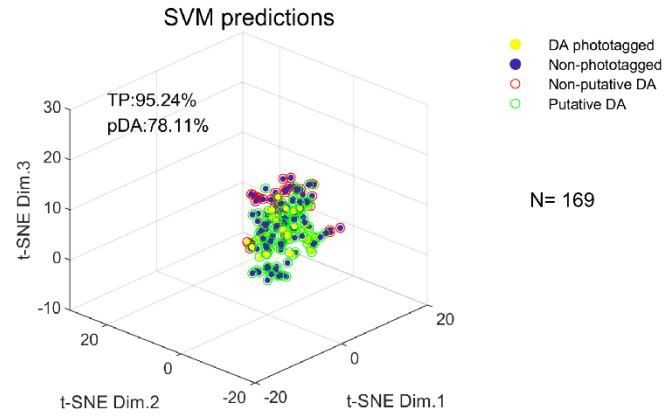
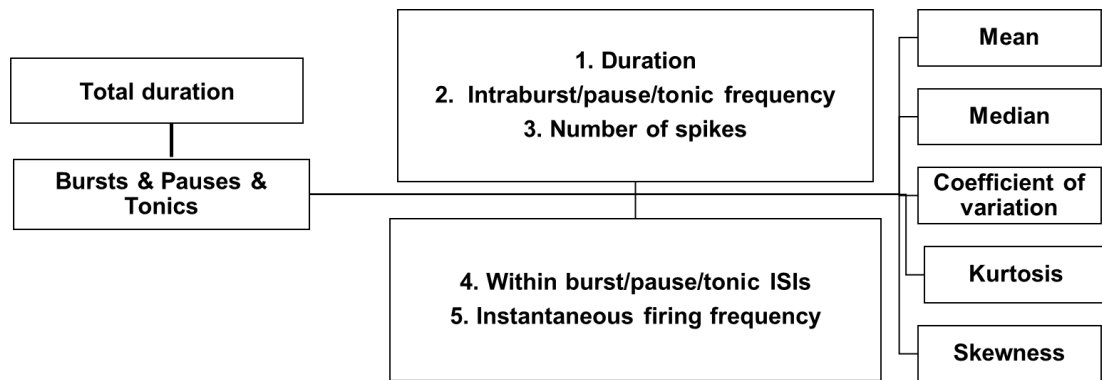
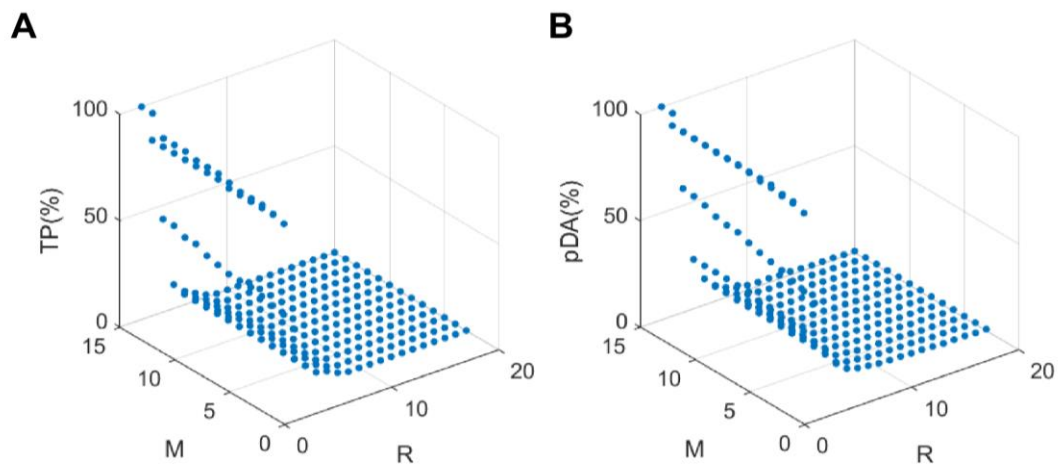
A**B**

Figure 7. Prediction models using labels obtained from k-means clustering

- (A) Euclidean distance clustering predictions for recorded neurons from the testing dataset (n=169).
- (B) Clustering predictions generated from a linear SVM for recorded neurons from the testing dataset (n=169).

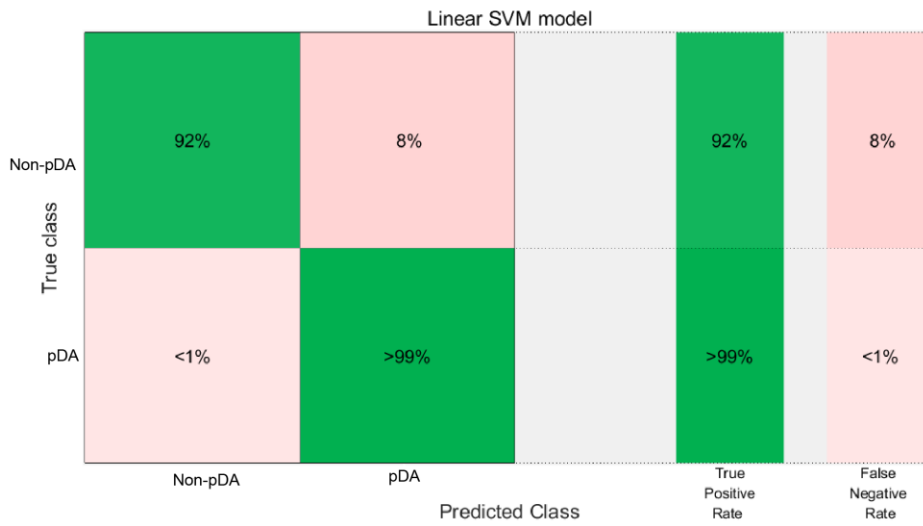


Supplementary Figure 1. Extracted firing property features



Supplementary Figure 2. Parameter set estimations of DBSCAN

DBSCAN with different parameter sets, R and M, showing the variation of TP (%) (a) and pDA (%) (b). R from 2 to 20 and M from 3 to 15.



Supplementary Figure 3. Confusion matrix of the linear SVM model

The linear SVM model was trained on the training dataset with the labels generated by the k-means model, and it was evaluated by 5-fold cross-validation.