Classifying GABAergic interneurons with semi-supervised projected model-based clustering

Bojan Mihaljević, Ruth Benavides-Piccione, Luis Guerra, Javier DeFelipe, Pedro Larrañaga, Concha Bielza

ABSTRACT

Objectives: A recently introduced pragmatic scheme promises to be a useful catalog of interneuron names. We sought to automatically classify digitally reconstructed interneuronal morphologies according to this scheme. Simultaneously, we sought to discover possible subtypes of these types that might emerge during automatic classification (clustering). We also investigated which morphometric properties were most relevant for this classification.

Materials and methods: A set of 118 digitally reconstructed interneuronal morphologies classified into the common basket (CB), horse-tail (HT), large basket (LB), and Martinotti (MA) interneuron types by 42 of the world's leading neuroscientists, quantified by five simple morphometric properties of the axon and four of the dendrites. We labeled each neuron with the type most commonly assigned to it by the experts. We then removed this class information for each type separately, and applied semi-supervised clustering to those cells (keeping the others' cluster membership fixed), to assess separation from other types and look for the formation of new groups (subtypes). We performed this same experiment unlabling the cells of two types at a time, and of half the cells of a single type at a time. The clustering model is a finite mixture of Gaussians which we adapted for the estimation of local (per-cluster) feature relevance. We performed the described experiments on three different subsets of the data, formed according to how many experts agreed on type membership: at least 18 experts (the full data set), at least 21 (73 neurons), and at least 26 (47 neurons).

Results: Interneurons with more reliable type labels were classified more accurately. We classified HT cells with 100% accuracy, MA cells with 73% accuracy, and CB and LB cells with 56% and 58% accuracy, respectively. We identified three subtypes of the MA type, one subtype of CB and LB types each, and no subtypes of HT (it was a single, homogeneous type). We got maximum (adapted) Silhouette width and ARI values of 1.03, 0.79, and 0.42, when unlabling the HT, CB, LB, and MA types, respectively, confirming the quality of the formed cluster solutions. The subtypes identified when unlabling a single type also emerged when unlabling two types at a time, confirming their validity. Axonal morphometric properties were more relevant that dendritic ones, with the axonal polar histogram length in the \([\pi, 2\pi]\) angle interval being particularly useful.

Conclusions: The applied semi-supervised clustering method can accurately discriminate among CB, HT, LB, and MA interneuron types while discovering potential subtypes, and is therefore useful for neuronal classification. The discovery of potential subtypes suggests that some of these types are more heterogeneous than previously thought. Finally, axonal variables seem to be more relevant than dendritic ones for distinguishing among the CB, HT, LB, and MA interneuron types.

1. Introduction

GABAergic interneurons of the cerebral cortex are key elements in many aspects of cortical function in both health and disease. Nevertheless, the classification of GABAergic interneurons is a difficult task and has been a topic of debate for a long time, since the
We extend the mentioned paper by refining some of the predictions by few experts. To cope with the semi-supervised projected model-based clustering algorithm [3], we sought to discover possible subtypes and see if the types agreed upon by at least three experts, with being a 'label reliability ability', i.e., such that the label of each neuron in the subset was labeled whereas others are not. Since all our neurons were labeled by the experts, we fitted the semi-supervised scenario by removing the labels of (a) one type at a time; (b) two types at a time; and (c) half the instances of each type, simultaneously. By doing this we sought to discover possible subtypes and see if the types could be automatically discriminated. We used an adaptation of the semi-supervised projected model-based clustering algorithm (SeSProc) introduced in [12]. This is a probabilistic clustering algorithm which estimates the number of clusters and the relevance of each predictive feature for each of the clusters. The estimation of feature relevance within model-based clustering was introduced in [13].

We quantified the neurons with nine simple axonal and dendritic morphological variables, such as the axonal length close to the soma, and labeled them according to the choices of the expert neuroscientists. In [3] each instance was given up to 42 labels—coming from the 42 experts that concluded the study. Following a common practice in supervised learning [14], we reduced this vector of 42 labels to its mode (i.e., the most common value), thus obtaining a single label per neuron. However, since experts frequently disagreed, such labels were often not reliable, i.e., they were backed by few experts. To cope with the label noise [15,16] that expert disagreements may be introducing, we analyzed three subsets of our neuron population, each with a different minimum of 'label reliability', i.e., such that the label of each neuron in the subset was agreed upon by at least three experts, with being a 'label reliability threshold'.

This paper is an extension of [17] and is the result of close collaboration between experts in neuroanatomy and machine learning. We extend the mentioned paper by refining some of the predictor variables, adapting the SeSProc algorithm, and considering two additional experimental settings. The remainder of this paper is organized as follows: Section 2 describes the materials and methods we used; Section 3 reports and discusses the obtained results; while Section 4 provides conclusions.

2. Materials and methods

2.1. Data

We used 237 three-dimensional (3D) reconstructions of interneurons from several areas and layers of the cerebral cortex of the mouse, rat, and monkey. These neurons were used in [3], and were originally extracted from NeuroMorpho.org [18]. From this population of neurons, we formed subsets by imposing minimums on the number of experts that agreed on the label of an included cell (i.e., a 'label reliability threshold'), considering that a higher threshold yields more confidence in the cells' labels. We used thresholds 18, 22 (half plus one out of the 42 experts), and 26 to build three databases: th18, th22, and th26, respectively. These data sets contained interneurons of four different types (classes): common basket (CB), horse-tail (HT), large basket (LB), and Martinotti (MA). Table 1 shows the distribution of different types at the three label reliability thresholds.

We characterized each neuron using nine features of axonal and dendritic morphology. While one may compute many morphological features (e.g., [3] used over 2000 features for classification), none are known, so far, as good predictors of interneuron type. Since detailed morphometric information on 3D reconstructed cortical interneurons is relatively scarce (a few hundred reconstructed neurons are available, comprising different types), it might be counterproductive to use many predictor variables. Therefore, we kept the number of variables low by defining variables which capture how, in our opinion, an expert classifies an interneuron upon visual examination.

We consider that an expert classifies an interneuron by estimating the distribution and the orientation of axonal and dendritic arborizations. We therefore measured the axonal and dendritic length according to the Sholl (5 features) and polar histogram (4 features) analyses from NeuroExplorer, the data analysis companion to Neurolucida [19]. Sholl analysis computes axonal and dendritic length at different distances from the soma whereas the polar histogram [20] describes the overall direction of dendritic growth; we only distinguished between two halves of the histogram, namely, the bifurcation angles falling in the [0, π] interval and those falling in the [π, 2π] interval. See Table 2 and Fig. 1 and for further details on predictor variables. We standardized all variables (transformed them so to have zero mean and unit standard deviation) prior to classification.

While an expert who classifies using a similar rationale can only roughly estimate these features, our classifier uses exact values, thus possibly being more objective. This is important as some of the features that we use, such as the length of the axonal arbor at a certain distance from the soma, are rather hard for an expert to estimate.

---

<table>
<thead>
<tr>
<th>Label Reliability Threshold (th)</th>
<th>CB</th>
<th>HT</th>
<th>LB</th>
<th>MA</th>
</tr>
</thead>
<tbody>
<tr>
<td>th18</td>
<td>25</td>
<td>9</td>
<td>21</td>
<td>24</td>
</tr>
<tr>
<td>th22</td>
<td>19</td>
<td>5</td>
<td>18</td>
<td>22</td>
</tr>
<tr>
<td>th26</td>
<td>12</td>
<td>4</td>
<td>19</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 1: Distribution of interneuron types with respect to label reliability threshold. Lowermost row shows total number of cells per dataset.
2.2 Semi-supervised projected model-based clustering

Table 2

<table>
<thead>
<tr>
<th>Variable</th>
<th>Arbor type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₁</td>
<td>Axon</td>
<td>Polar histogram length (µm) for the [0, π) radians interval</td>
</tr>
<tr>
<td>X₂</td>
<td>Axon</td>
<td>Polar histogram length (µm) for the [π, 2π) radians interval</td>
</tr>
<tr>
<td>X₃</td>
<td>Axon</td>
<td>Sholl analysis length (µm) at less than 150 µm from the soma</td>
</tr>
<tr>
<td>X₄</td>
<td>Axon</td>
<td>Sholl analysis length (µm) at more than 150 and less than 300 µm from soma</td>
</tr>
<tr>
<td>X₅</td>
<td>Axon</td>
<td>Sholl analysis length (µm) at more than 300 µm from soma</td>
</tr>
<tr>
<td>X₆</td>
<td>Dendrites</td>
<td>Polar histogram length (µm) for the [0, π) radians interval</td>
</tr>
<tr>
<td>X₇</td>
<td>Dendrites</td>
<td>Polar histogram length (µm) for the [π, 2π) radians interval</td>
</tr>
<tr>
<td>X₈</td>
<td>Dendrites</td>
<td>Sholl analysis length (µm) at less than 90 µm from the soma</td>
</tr>
<tr>
<td>X₉</td>
<td>Dendrites</td>
<td>Sholl analysis length (µm) at more than 90 µm from the soma</td>
</tr>
</tbody>
</table>

**Fig. 1.** A schematic representation of the nine morphological features of axonal and dendritic morphology that we used as predictor variables. (A) An example of a 3D reconstructed interneuron classified by expert neuroscientists in [3], showing its axonal (displayed in blue) and dendritic (red) arborizations. The gray vertical shadow indicates the extent of the cortical column (assumed to be 300 µm wide) whereas the dimensions of the squares are 100 µm x 100 µm. (B) Schematic representation of the overall direction of dendritic growth (polar histogram) for the dendrites (above; red) and the axon (below; blue). Features X₁ and X₆ encode axonal growth length in the angle intervals [0, π] and [π, 2π], respectively, whereas features X₃ and X₇ capture the dendritic growth length in the same angle intervals. (C) Schematic representation of the features encoding axonal and dendritic arborization lengths at different distances from the soma (Sholl analysis): features X₁, X₄, and X₅ encode, respectively, axonal lengths at 0–150 µm (marked with a dark blue circle), 150–300 µm (blue circle), and over 300 µm (outside of the blue circle) from the soma. Features X₆ and X₉ measure dendritic arborization at 0–90 µm (pink circle) and over 90 µm (outside of the pink circle) from the soma, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

2.2.2. A finite mixture model with local feature selection

Let \( X = \{X₁, \ldots, X₉\} \) be observed data, with \( xᵢ \in \mathbb{R}^F, \forall i \in \{1, \ldots, N\} \), where \( F \) denotes the number of features. In model-based clustering, we assume that the data are generated from a finite mixture of \( K \) components. The density function for an instance \( xᵢ \) is

\[
 f(xᵢ | \Theta) = \sum_{m=1}^{K} \pi_m f(xᵢ | \theta_m),
\]

with \( \pi_m \in [0, 1] \) and \( \sum_{m=1}^{K} \pi_m = 1 \), and \( \theta_m \) being the parameters of component \( m \), with \( \Theta = \{\Theta₁, \ldots, \Theta_K, \tau_₁, \ldots, \tau_K\} \).

SeSProc encodes the clustering solution with a latent \( K \)-dimensional binary variable \( Z \), with \( zᵢm = 1 \) if \( xᵢ \) is assigned to component \( m \), and \( zᵢm = 0 \) otherwise. In a semi-supervised setting, we observe the class labels \( cᵢ, cᵢ \in \{1, \ldots, C\} \), of some instances \( i \in \{1, \ldots, L\} \), \( L < N \). We use the class labels to set the minimum number of components in the mixture, with one component per class label (i.e., \( K = C \)). We also use the labels to fix \( zᵢm = \|cᵢ = m\| \), where \( \| \cdot \| \) returns 1 if \( \cdot \) is true and 0 otherwise, for \( i < L \). This means that the algorithm does not change \( zᵢm \) for the labeled \( xᵢ \)—they always belong to clusters specified by their labels; the goal is to cluster the unlabeled instances \( xᵢ, L < i < N \).

Localized feature selection. Not all predictor variables are necessarily relevant for clustering. Following [13], we assume that a feature \( j \) is irrelevant if its distribution is independent on the components. Such a feature then follows some distribution \( f(· | λ_j) \), independent of any particular component, instead of following a distribution \( f(· | θ_j) \), dependent on a component \( m \). SeSProc then models the distribution of a feature \( X_j \) as a mixture of \( f(· | λ_j) \) and

\[\text{This differs from the original SeSProc method which follows the notion of irrelevance given in [21]. Also, our model differs from the one used in [13] in that it considers the per-component relevance of each feature.}\]
EM for mixture learning

If $Z$ and $Y$ were observed, the log-likelihood function would be

$$
\log \mathcal{L}(\Theta, X, Z, Y) = \sum_{i=1}^{N} \sum_{m=1}^{K} z_{im} \log \pi_m + \sum_{i=1}^{N} \sum_{m=1}^{K} \sum_{j=1}^{F} (z_{im} \log \rho_{mj} + \log f(x_{ij} | \theta_{mj})) + \sum_{i=1}^{N} \sum_{m=1}^{K} \sum_{j=1}^{F} \gamma_{mj} (1 - \gamma_{mj}) \log(1 - \rho_{mj}) + \log f(x_{ij} | \Theta_{mj})
$$

where $\Theta_{mj}$ are the parameters from iteration $t - 1$ and $\gamma()$ is the expectation function.

Assuming that both $f \cdot (\theta_{mj})$ and $f \cdot (\lambda_{j})$ are Gaussian distributions, the parameters are updated in the M-step by maximum likelihood, as follows:

$$
\pi_m = \frac{1}{N} \sum_{i=1}^{N} z_{im}, \quad \rho_{mj} = \frac{1}{N} \sum_{i=1}^{N} \gamma(z_{im}) \gamma(v_{mj}), \quad \mu_{\phi mj} = \frac{1}{N} \sum_{i=1}^{N} \gamma(z_{im}) \phi_{ij}, \quad \sigma_{\phi mj}^2 = \frac{1}{N} \sum_{i=1}^{N} \gamma(z_{im}) (\phi_{ij} - \mu_{\phi mj})^2, \quad \mu_{\lambda mj} = \frac{1}{N} \sum_{i=1}^{N} \gamma(z_{im}) \lambda_{ij}, \quad \sigma_{\lambda mj}^2 = \frac{1}{N} \sum_{i=1}^{N} \gamma(z_{im}) (\lambda_{ij} - \mu_{\lambda mj})^2
$$

for $m = 1, \ldots, M; j = 1, \ldots, F$.

### Estimating the number of components

Assuming that we have $C$ classes in the data, the initial mixture contains $C$ components. Its parameters are estimated with the EM procedure and its quality evaluated (see below) and recorded. Then, in the next iteration of SeSProC, this mixture is augmented with an additional component (see below), its parameters again estimated with EM and its quality computed and recorded. The process is repeated until a mixture $\mathcal{M}^K$ with $K$ components is better than a mixture $\mathcal{M}^{K-1}$, and yields $\mathcal{M}^{K}$ as the final model. We evaluate the quality of a model with the AIC score [23],

$$
AIC = -2 \log \mathcal{L} + 2R
$$

where $R$, the number of parameters in a model, is a function of the number of components $K$ and the number of features $F$:

$$
R = 2KF + (K - 1) + KF
$$

Additionally, we halt mixture augmentation if $\mathcal{M}^K$ contains a component $m$ with less than two instances (i.e., such that $m$ is the most likely component for less than two instances), returning $\mathcal{M}^{K-1}$, unless $K = C$, in which case $\mathcal{M}^K$ is returned.

**Initialization of a new component $K$.** We have modified SeSProC to use the following heuristic for initializing new components. Starting from the previous mixture, $\mathcal{M}^{K-1}$, we consider the neighborhood of each unlabeled point as a potential new cluster. Thus, we take the nearest unlabeled neighbors (according to the Euclidean distance in full dimensionality), where $c$ is a parameter of the algorithm, to a point $x_i$ and assign them to a new cluster, by setting $z_{ik} = 1$ and $z_{jm} = 0$, $m \neq K$ for all $x_i$ in the neighborhood of $x_i$; we then update all parameters (this includes the $\theta_{kj}$) by maximum likelihood (i.e., with an M-step) and compute the likelihood of the thereby obtained model. The neighborhood that yields the most likely model is then used to initialize $\theta_{kj}$ in the new model $\mathcal{M}^K$. All assignments to $Z$ here described are then undone after $\theta_{kj}$ is initialized (i.e., these assignments were only temporary).

---

1 The first addend in the formula for $R$ corresponds to $\pi$'s, second to $\lambda$'s, third to $\mu$'s and fourth to $\sigma$'s.
2.3. Empirical setup

We tested SeSPROC's capacity to discriminate among interneuron classes and explored the existence of their subtypes in three experimental settings, corresponding to three rules for hiding class labels. First, we unlabeled all the cells of a single class and then ran the algorithm once for each class. Here, there was initially a cluster for each of the other (labeled) types and the desired result was to assign unlabeled instances to a (one of) newly formed cluster(s), allowing us to explore the potential subtypes of each class separately. Second, we simultaneously unlabeled all cells of each pair of classes, yielding six clustering scenarios; this allowed to assess whether cells of different classes would be clustered together and whether, and to what extent, would the subtypes identified in the previous setting re-appear. Finally, we simultaneously unlabeled a portion of cells of each of the four classes. This allowed unlabeled cells to be placed in their ‘true’ cluster, other classes (i.e., be misclassified), or assigned to a new cluster, providing insight into the homogeneity of each interneuron class. Here we unlabeled half the instances of each class (rounding down when necessary). We selected unlabeled cells by random sampling at each label reliability threshold (thus a cell might have been unlabeled at th26 but not at th18, for example), and repeated the sampling ten times. We used these three unlabeled settings for each label reliability threshold, i.e., for th18, th22, and th26.

For the first two settings, we defined per-class discrimination accuracy as acc_t = c_t / u_t, where u_t is the number of cells of the unlabeled class t and c_t the cardinality of the subset of u_t assigned to a (one of) newly formed clusters(s) (and therefore not assigned to one of the other classes). In the second setting, we averaged acc_t across the three ‘scenarios’ in which t was unlabeled (e.g., HB was unlabeled together with CB, LB, and MA). For the third setting, we defined two measures —‘error’ and ‘accuracy’—as follows: err_t = \sum c_t - p_t / u_t, where c_t is the number of unlabeled cells of class t assigned to class t and u_t the number of unlabeled cells of class t, and acc_t = a_t / u_t. ‘Accuracy’ considers the proportion of unlabeled cells of class t classified as t whereas ‘error’ does not penalize assignments to newly formed clusters.

When starting the EM procedure for a mixture model \( M^K \), one has to choose how to initialize the parameters. Furthermore, when \( K > C \) one can keep or adapt the parameters from the previous model, \( M^{K-1} \). We initialize the parameters with the following heuristics:

- For \( m \leq C \), estimate \( \theta_m \) from cells labeled as belonging to class m.
- Keep the \( \theta \) and \( \rho \) parameters from \( M^{K-1} \) for \( C < m < K \). That is, \( \theta^K = \theta^{K-1} \), \( \rho^K = \rho^{K-1} \).
- Estimate \( \theta_m \) from the instances selected as described in paragraph ‘Initializing a new component’ in Section 2.2.3.
- For components \( m \in \{1, \ldots, C\} \cup \{K\} \), i.e., those of fixed classes and the newly introduced one, make all features equally relevant and irrelevant: \( \rho_m = 0.5 \), \( \forall m \in \{1, \ldots, C\} \cup \{K\} \).
- Adapt \( \pi \) from \( M^{K-1} \) to give more weight to newly discovered components \( m > C \) than to class components \( m \leq C \): \( \pi^K_m = 2\pi^{K-1}_m \), \( \forall m > C \), \( \forall m \leq C \).
- Estimate \( \lambda \) as if \( p_{\lambda} = 0.5 \), \( \forall m, j \), i.e., as if each feature was equally relevant and irrelevant for every component, in order to fully ‘reset’ the \( \lambda \) estimates.

The EM procedure iterates until log-likelihood converges or up to 25 iterations. We set \( c_t = 5 \) for initializing new components, because this value produced the best results in preliminary experiments.

3. Results and discussion

3.1. Discriminating among classes

3.1.1. Unlabeling a single class

When unlabeling a single class, HB and MA cells were better distinguished from other classes when label reliability increased; the opposite happened for LB, while the discrimination accuracy for CB cells was rather unaffected (see Fig. 2a). At th26, HB and MA cells were identified with rather high accuracy.

HT was the most easily identified class: with perfect accuracy at th22 and th26, and high accuracy (0.89) at th18.

Although accurately identified at th22 and th26, MA cells were confused with all the other classes, particularly with HB at th18 and th22, and CB at th26. In this respect, the MA seemed to be the most heterogeneous interneuron type with respect to the used variables.

LB and CB cell types were often confused with each other but easily distinguished from other types (with the exception of CB at th18, where it was heavily confused with HB; see Table 3). This confusion is not surprising, as even expert neuroscientists often struggle to discern these two classes [3].

3.1.2. Unlabeling two classes

When hiding two classes at a time, discrimination accuracy generally increased with label reliability (see Fig. 3a). CB cells were better discriminated than when hiding a single class, HT cells equally well, whereas LB and MA cells better on some label reliability thresholds and worse on others.

At th26, which contained the most reliably labeled cells, all classes were identified more accurately than when unlabeling a single class (see Figs. 2a and 3a). Furthermore, the classes were well separated in the formed clusters: only six out of the 20 clusters formed at th26 (columns ‘A’, ‘B’, etc., in Table 4a–f) were ‘mixed’, i.e., containing instances of more than one class (shown in black). Thus, even though MA cells were misclassified as CB at th26 (when unlabeling the MA type), these two types were neatly separated when unlabeled simultaneously (see Table 4a). HT cells were almost never placed in clusters containing other cell types (only a single CB cell was assigned to a HT cluster; see Table 4b). LB cells were least separated in the formed clusters: they were mixed with both CB and MA cells in two clusters (see Table 4e and f).

3.1.3. Partially unlabeling all classes

In this setting, discrimination accuracy generally improved with label reliability (see Fig. 4a), except for MA and CB at th22, where it decreased due to more instances being assigned to new clusters (indicated by their low error at th22; Fig 4). Likewise, discrimination error generally decreased, except for MA and LB at th26.

At th26, the most accurately classified and most homogeneous types were HB and CB, as they had lowest error and highest accuracy. HB and CB, on the other hand, displayed low accuracies but relatively low errors, suggesting they were more heterogeneous than HB and CB at th26.

3.1.4. Summary

Summarizing the previous three sections, we can note that, despite some fluctuations—possibly due to small data samples—discrimination accuracy tended to increase with label reliability in all three experimental settings (e.g., accuracy was almost universally higher and error lower at th26 than th18, see, e.g., Figs. 2a, 3a, 4a and 4b). This might suggest that the degree of label
noise decreased with label reliability threshold, and that therefore, the most reliable results were obtained at \( th_{26} \).

### 3.2. Potential subtypes

The number of subtypes generally decreased with label reliability (see Figs. 2b and 3b). This may indicate that there was more heterogeneity among less reliably labeled cells; nonetheless, this heterogeneity may simply be due to the higher number of instances at lower thresholds (especially for the CB, LB, and HT types). HT appeared as the most compact class as all of its cells were clustered together (in a single cluster) at \( th_{22} \) and \( th_{26} \), in both the first and the second labeling scenario (see Figs. 2b and 3b and Table 4b-d). MA, on the other hand, seemed to be the most heterogeneous—at

### Table 3

Clustering of unlabeled instances when hiding a single class. Each row corresponds to one 'labeling scenario', e.g., CB is the hidden class in the first row, whereas columns represent the classes to which the unlabeled cells were assigned. The cells assigned to newly formed clusters are considered as correctly classified and are thus displayed in blue color on the diagonal whereas the cells assigned to other classes are shown in red, outside the diagonal. Thus, for example, 16 MA cells were correctly classified, four were misclassified as CB, one as HT, and one as LB at \( th_{26} \) (row four, rightmost table). Note that the number of formed subtypes is not shown—the 16 MA cells were placed in three clusters; the formed clusters are discussed in Section 3.2. Zeros were omitted.

<table>
<thead>
<tr>
<th></th>
<th>CB</th>
<th>HT</th>
<th>LB</th>
<th>MA</th>
</tr>
</thead>
<tbody>
<tr>
<td>th_{18}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CB</td>
<td>29</td>
<td>7</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>HT</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LB</td>
<td>3</td>
<td>22</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>MA</td>
<td>1</td>
<td>13</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>th_{22}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CB</td>
<td>14</td>
<td>9</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>HT</td>
<td></td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LB</td>
<td>6</td>
<td>12</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>MA</td>
<td>5</td>
<td>2</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>th_{26}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CB</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>HT</td>
<td></td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LB</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>MA</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>16</td>
</tr>
</tbody>
</table>

### Table 4

Clustering of unlabeled instances when hiding two classes at a time at \( th_{26} \). Rows represent the hidden classes whereas columns denote the clusters the instances were ascribed to. ‘N’, ‘N’, etc., denote newly formed clusters. Misclassified instances are shown in red, instances in ‘pure’ (not-mixed) clusters are shown in blue, and instances in mixed clusters in black. Zeros were omitted.

#### (a) Unlabeling MA and CB

<table>
<thead>
<tr>
<th></th>
<th>CB</th>
<th>HT</th>
<th>LB</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CB</td>
<td>3</td>
<td>1</td>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### (b) CB and HT

<table>
<thead>
<tr>
<th></th>
<th>CB</th>
<th>HT</th>
<th>MA</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB</td>
<td>1</td>
<td></td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HT</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### (c) HT and LB

<table>
<thead>
<tr>
<th></th>
<th>CB</th>
<th>MA</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LB</td>
<td>4</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### (d) HT and MA

<table>
<thead>
<tr>
<th></th>
<th>CB</th>
<th>LB</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT</td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MA</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

#### (e) LB and MA

<table>
<thead>
<tr>
<th></th>
<th>CB</th>
<th>HT</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HT</td>
<td>8</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

#### (f) CB and LB

<table>
<thead>
<tr>
<th></th>
<th>CB</th>
<th>MA</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LB</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
th26 MA cells were clustered into at least three subtypes (see Fig. 2b and Tables 4a, d, and e). Thus, we focused on th26 for analyzing the formed subtypes as it contained the most reliably labeled interneurons.

### 3.2.1. With a single hidden class

MA cells were clustered in three groups (see Fig. 5 for representative examples), with six, five, and five cells each, whereas a single distinct subtype was identified for the CB and LB classes, counting five and seven cells each, respectively. Since all HT cells were placed in a single cluster (see Fig. 2b), no potential subtypes of HT were identified.

Overall, MA subtypes showed relatively sparse axonal arbors, as indicated by their low or medium values for $X_1$ and $X_5$ (see Figs. 5a–c), with MA-A cells being less sparse than MA-B and MA-C ones. MA-A cells had plenty of axonal arborization far from soma (high $X_5$ values in Fig. 5a) and dendritic polar histogram length in the $[\pi, 2\pi]$ interval ($X_7$). MA-B cells exhibited medium values for all variables (Fig. 5b) whereas MA-C had the sparsest axons (low values for $X_1$ and $X_5$ in Fig. 5c) and, like MA-A, plenty of axon far from soma ($X_4$) and dendrites in the $[0, \pi]$ polar histogram interval ($X_7$). CB-A cells displayed relatively sparse axons (medium $X_1$ and low $X_5$ values in Fig. 5d), with little axon far from the soma (low $X_5$ values) and little dendritic arborization far from soma and in the $[\pi, 2\pi]$ polar histogram interval (low $X_1$ and $X_5$ values). LB-A cells exhibited the most dense axonal and dendritic arbors, with high or medium values for $X_1$, $X_3$, $X_6$, and $X_7$. $X_4$ values (axonal length at medium distance from soma) was especially high (see Fig. 5e).

#### Validating the produced clusters.

While the discovered subtypes are relatively small—their sizes ranging from five to seven cells, they may nonetheless be relevant in the domain of neuronal classification, where 3D neuronal reconstructions, and reliably classified reconstructions in particular, are scarce. Cluster quality indices [24, 25] may thus help assess the goodness of the clustering solution.

One type of indices compares the obtained (crisp) clustering partition with the original one, given by class labels. We performed such an analysis in Section 3.1 when we computed accuracy, and now we report a measure more commonly used for this end, the Adjusted Rand Index (ARI; [26, 27]),

$$
\text{ARI} = \frac{\sum_{c=1}^{C} \sum_{m=1}^{K} \frac{N_{cm}}{2} - \frac{1}{2} \left[ \sum_{c=1}^{C} \frac{N_c}{2} \right] \sum_{m=1}^{K} \frac{N_m}{2} } { \frac{1}{2} \left[ \sum_{c=1}^{C} \frac{N_c}{2} \right]^2 + \sum_{m=1}^{K} \frac{N_m}{2} - \frac{1}{2} \left[ \sum_{m=1}^{K} \frac{N_m}{2} \right] \sum_{c=1}^{C} \frac{N_c}{2} } - \frac{1}{2} \left[ \sum_{c=1}^{C} \frac{N_c}{2} \right] \sum_{m=1}^{K} \frac{N_m}{2} }.
$$

where $N_{cm}$ is the number of instances of class $c$ assigned to cluster $m$. ARI rewards two types of agreements: (a) clustering a pair of instances together (i.e., as members of a same cluster) in both partitions; and (b) clustering a pair of instances separately (i.e., as members of different clusters) in both partitions. It reaches its maximum value, 1, when the two partitions agree perfectly, and this occurs when unlabeling HT cells (see Table 5). We also achieved high ARI values when unlabeling CB and LB cells whereas we obtained a low one when unlabeling MA cells; this is due to the discovery of three MA subtypes in the setting, which was not favored by ARI because it increased the difference among the two partitions.

We cannot easily compute a second type of clustering quality indices, based on assessing properties such intra-cluster compactness and inter-cluster separation, because that would require computing distances among data points, something which is unclear how to do when points are located in different feature subspaces. When then assessed our clustering in terms of probabilistic concordance, considering that a clustering is good if cluster membership probabilities, $p(z_{ij})$, are similar among the members of a same cluster and different among members of different clusters. We measured the similarities among cluster membership probabilities of two data instances with Jensen–Shannon divergence,

$$
d_{JS}(p(z_i), p(z_j)) = \frac{1}{2}(d_{KL}(p(z_i), r) + d_{KL}(p(z_j), r)),
$$

where $r = \frac{1}{2}(p(z_i) + p(z_j))$ and $d_{KL}(p(z_i), p(z_j))$ is the Kullback–Leibler divergence [28] between $p(z_i)$ and $p(z_j),

$$
d_{KL}(p(z_i), p(z_j)) = \sum_{m=1}^{K} p(z_{im}) \log \left( \frac{p(z_{im})}{p(z_{jm})} \right).
$$

Note that $p(z_{im})$ is simply $\gamma(z_{im})$ computed in the final step of the EM algorithm. Unlike Kullback–Leibler divergence, Jensen–Shannon divergence is symmetric, does not require absolute continuity (i.e., that $p(z_{im}) = 0 \Rightarrow p(z_{jm}) = 0$), its square root is a metric, and it is bounded: $0 \leq d_{JS} \leq 1$ [29].

Using Jensen–Shannon divergence as the measure of distance among data instances, we computed the Silhouette width [30] clustering index, thus measuring intra-cluster compactness and inter-cluster separation, whereas rows correspond to different metrics.

### Table 5

Cluster validation indices when hiding a single class at th26. Columns denote the unlabeled class whereas rows correspond to different metrics.

<table>
<thead>
<tr>
<th>Class</th>
<th>HT</th>
<th>MA</th>
<th>CB</th>
<th>LB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjusted Rand index</td>
<td>1.00</td>
<td>0.42</td>
<td>0.83</td>
<td>0.79</td>
</tr>
<tr>
<td>Silhouette</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
inter-cluster separation in terms of this distance. The Silhouette width is given by

\[ SW = \frac{1}{N} \sum_{i=1}^{N} \frac{b_i - a_i}{\max(b_i, a_i)}, \]

where \( a_i \) is the average distance between \( x_i \) and other points in its cluster, while \( b_i \) is its average distance to the points in the closest cluster (defined as that yielding the lowest \( b_i \)). We achieved maximum Silhouette values (its values range from -1 to 1) for all labeling scenarios (see Table 5), showing that the cluster membership probabilities had converged reasonably.

A partial cause for high Silhouette widths was that we only unlabeled a subset of instances in each setting, and it was only those instances' \( \hat{p}(z_i) \) that were estimated by our algorithm, while the rest instances' \( \hat{p}(z_i) \) were fixed to degenerate distributions (probability
1 for the class labels' cluster; 0 for all other clusters) which were different among the clusters. Thus, e.g., when unlabeling the HT cells, we estimated the $p(z)$ of only four instances, the overall Silhouette width therefore necessarily being high, due to the inter-cluster differences among the fixed $p(z)$. Yet, as Table 6 shows, Silhouette widths were also high for the newly discovered clusters, not only for those corresponding to the known types (and thus containing many cells with fixed degenerate $p(z)$).

Finally, in the next section we validated the discovered subtypes with a different experimental setting—the hiding of two classes, thus evaluating their ‘stability’, i.e., their robustness to different labeling scenarios.

### 3.2.2. With two hidden classes

The validity of the above-described subtypes of the MA, CB, and LB types, identified when hiding a single class, was confirmed when hiding two classes simultaneously. That is, in almost every labeling scenario in the second setting (i.e., for every pair of hidden classes), there was a cluster that greatly resembled the corresponding subtypes. So, for example, when hiding CB and MA, cluster E (Table 4a) was identical to subtype CB-A. Furthermore, four cells from the CB-A subtype were clustered together in every labeling scenario, that is, the intersection of CB-A and cluster E in Table 4a, cluster B in Table 4b, and cluster B in Table 4f consisted of four cells. This subset of CB-A cells emerged as its ‘core’ of highly similar instances, showing the robustness of this subtype (which totalled five cells).

A ‘core’ of the MA-C subtype (i.e., the MA-C cells that were always clustered together) emerged, consisting of four cells which formed the intersection of the pure MA clusters C, A, and C in Table 4a, d, and e, respectively (the latter two clusters being identical). Regarding MA-A, a ‘core’ of three cells emerged, defined by the intersection of clusters A, C and A in Table 4a, d, and e, respectively (the latter being a mixed cluster). Finally, clusters C and D in Table 4a and d contained four and three MA-B cells, respectively.

### 3.2.3. Feature relevance

We focused on the first setting (i.e., hiding a single class) and th26 to analyze the estimated relevance of predictor variables. Overall, all predictor variables seemed useful, as each one was very likely relevant (around 100% chance of being relevant; dark green boxes in Fig. 6) for at least two subtypes/classes identified at th26. Feature $X_{1}$—axonal polar histogram in the $(\pi, 2\pi)$ interval—appeared to be the most useful as it was very likely relevant for all the subtypes/classes (see Fig. 6). While its relevance for the HT class and MA subtypes is clear—an MA’s axon grows predominantly upwards from the soma whereas the opposite holds for HT—it is interesting that it was relevant for the CB and LB subtypes as well. Features $X_{4}$ and $X_{5}$, which capture the length of axonal arborization at 150–300 μm and over 300 μm from soma, were relevant for five (out of six) subtypes. On the other hand, feature $X_{6}$—length of dendritic arborization at over 90 μm from soma—appeared as least useful as it was very likely relevant for only two subtypes (CB-A and MA-B). In general, axonal features ($X_{1}$ to $X_{5}$) were more likely to be relevant than dendritic ones ($X_{6}$). For example, an axonal feature was, on average, very likely relevant for 4.6 subtypes/classes in Fig. 6 whereas a dendritic one was for 2.5. Likewise, a dendritic feature was, on average, probably irrelevant (below 50% chance of being relevant; red and brown boxes in Fig. 6) for more subtypes than an axon feature.

Moreover, as shown by Fig. 6, the relevance of axonal features differed among the subtypes/classes. All axonal features were very likely relevant for all MA subtypes (except $X_{1}$ for MA-A: MA-A, unlike MA-B and MA-C, did not stand out regarding $X_{1}$—see the注 green box for $X_{1}$ in Fig. 5) and for the HT class. This suggests that the axonal arborizations of the MA subtypes and the HT class were distinct among themselves and from the CB-A and LB-A subtypes according to all axonal features. On the other hand, two and three (out of five) axonal features were relevant for the CB-A and LB-A subtypes, respectively, and, in particular, polar histogram in the $[0, \pi)$ interval and close to soma axonal arborization length ($X_{1}$ and $X_{3}$, respectively) were relevant with only 50% chance for both of those subtypes. This suggests that CB-A and LB-A were not particularly distinct according to those variables.

In summary, the results clearly show that the length of axonal polar histogram in the $[\pi, 2\pi)$ interval ($X_{1}$) is particularly relevant, followed by axonal length relatively near to the cell body ($X_{3}$) and the extent of the axon far from the cell body ($X_{5}$). In addition, the dendritic arborization was highly relevant for the

### Table 6

<table>
<thead>
<tr>
<th>Subtype</th>
<th>HT</th>
<th>LB</th>
<th>MA</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB</td>
<td>1.000</td>
<td>1.000</td>
<td>0.9996</td>
</tr>
<tr>
<td>HT</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>LB</td>
<td>0.994</td>
<td>1.000</td>
<td>0.996</td>
</tr>
<tr>
<td>MA</td>
<td>0.990</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Relevance ($\rho$) of features ($X_{1}$–$X_{6}$) for the HT, LB, CB, and MA subtypes identified at th26 when hiding a single type. Each row corresponds to one subtype, e.g., the last three rows correspond to the three subtypes of MA. The displayed subtypes do not contain all the cells contained at th26, as some were classified (e.g., besides CB-A, CB cells at th26 were also assigned to the LB and MA classes; the latter cells are not contained in any subtype represented in this graph). For interpretation of the references to color in the text, the reader is referred to the web version of the article.

A LB–A core of five cells emerged, contained in clusters A, B, and A in Table 4c, e and f, respectively. The latter two contained a larger LB–A core of six cells.
characterization of CB cells and MA cells. These findings may be useful for generating an accurate automatic classifier of 3D reconstructed interneurons.

4. Conclusions

Automatic classification of interneurons is a hard problem because data are scarce, the cells are morphologically, molecularly, and physiologically variable, and experts disagree on the definitions of features that distinguish them. Thus, we introduced a semi-supervised approach to this problem. Besides discriminating between types, this approach leads to the discovery of new types of neurons. It uses the available knowledge—in the shape of class labels—and also attempts to identify the relevance of each feature for each type and subtype.

We presented results on the classification of common basket, large basket, horse-tail, and Martinotti cells. We quantified the neurons with simple morphological features which describe the distribution and the orientation of axonal and dendritic arbors, seeking to mimic the way in which an expert visually classifies a neuron. The algorithm accurately discriminated among the different types when one and two types were unlabeled at a time and when half the instances of all types were unlabeled. Importantly, it identified potential subtypes of common basket, large basket, and Martinotti cells, suggesting that these types are more heterogeneous than previously thought. Although the identified subtypes are small, they may be indicative of the characteristics that differentiate cells belonging to the same interneuron type.

The defined morphological variables seemed useful for discriminating among the types. Axonal features seemed more useful than dendritic ones, with the axonal polar histogram length in the $[\pi, 2\pi]$ interval appearing to be the most useful feature. The defined variables might then be considered in future studies of interneuron classification. Still, it is possible that more complex variables, such as those considering both the distance and position with respect to the soma (e.g., over 300 $\mu$m from soma and above it) could further improve discrimination accuracy.

Overall, the results suggest that a semi-supervised approach may be helpful in neuronal classification and characterization. Further studies, with different morphological features and more neurons, would be needed in order to obtain more conclusive results. Although we focused on a subpopulation of GABAergic interneurons, the presented methodology could be applied to other types of cells as well.

The used algorithm, SeSProc, is open to further improvements, like the inclusion of uncertainty into labels. Instead of a single label—the class most voted by the experts—per data instance, a probability distribution could be learned from the experts’ votes, thus taking all votes into account. Alternatively, an approach similar to [31] might be applied: a separate model could be learned for each set of experts with similar opinions, and these models then combined into a final, consensus model. Regarding the data, although it is generally thought that the same morphological types of neurons are found in all species, we cannot discard the possibility of interspecies variability. Thus, we plan to further analyze the data taking into account the different species in order to find types of neurons that may be representative of particular species.

Finally, it has been shown that GABAergic interneurons are affected in several brain diseases but not all of these interneurons are equally affected. For example, alterations of chandelier cells expressing the calcium binding protein parvalbumin have been associated with certain forms of epilepsy and schizophrenia. In addition, possible alterations of interneurons expressing somatostatin, neuropeptide Y and vasoactive intestinal peptide, probably including Martinotti and horse-tail (or double bouquet) interneurons, have been reported in schizophrenia (e.g., [32-35]). Since the algorithm used in the present study accurately discriminated between different interneuron types and new groups can be discovered, its application to study brain diseases may shed light on interneuron pathology in these diseases, providing a tool to more accurately determine which subtypes may be affected.

Acknowledgements

The authors thank anonymous reviewers whose comments have greatly helped to improve this paper. This work was supported by grants from the following entities: the Spanish Ministry of Economy and Competitiveness (grants TIN2013-41592-P to B.M., P.L., and C.B.; BU2012-34963 to J.D.F.), CIBERNED CB06/05006 to J.D.F., the Cajal Blue Brain Project (Spanish partner of the Blue Brain Project initiative from EPFL) to B.M., J.D.F., P.L. and C.B.; by the Regional Government of Madrid through the S2013/ICE-2845-CASI-CAMCM project to B.M., P.L. and C.B.; and the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 604102 (Human Brain Project) to J.D.F., P.L. and C.B.-P. was supported by the Spanish Ministry of Economy and Competitiveness (CSIC).

References


