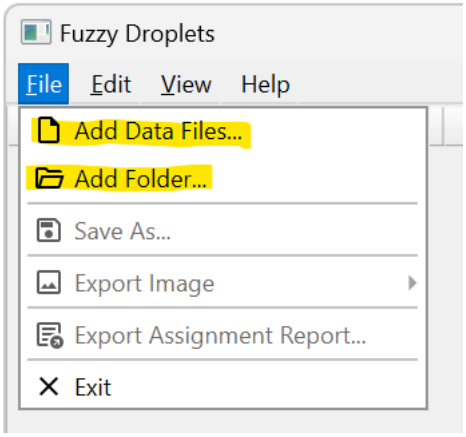
Fuzzy Droplets Tutorial, 1st March 2024

Step 1: Load your data

Load your data using the File menu:



With “Add Data Files…” you can select one or more individual files. With “Add Folder…” you can open all files within the specified folder.

Fuzzy Droplets expects data in .csv format. Each csv file contains the information for one biological sample (i.e., one well within a PCR plate). Each row of the .csv file contains information about a single droplet. The first two columns of each row are the amplitudes in channels 1 and 2, respectively. Hence, a minimal row could look like:

```
17254.93, 8002.4
```

If the droplet has been assigned an unfuzzy cluster membership, this can be represented by adding a third row containing a positive integer identifier for the cluster. Zero is reserved for unassigned data. Cluster identifiers should start at 1 and incrementally increase. Hence, an unfuzzy droplet with cluster membership could look like:

```
17254.93, 8002.4, 3
```

Upon reading this line, Fuzzy Droplets will infer that there are at least 3 clusters represented in the data, and the current droplet is a member of cluster 3.

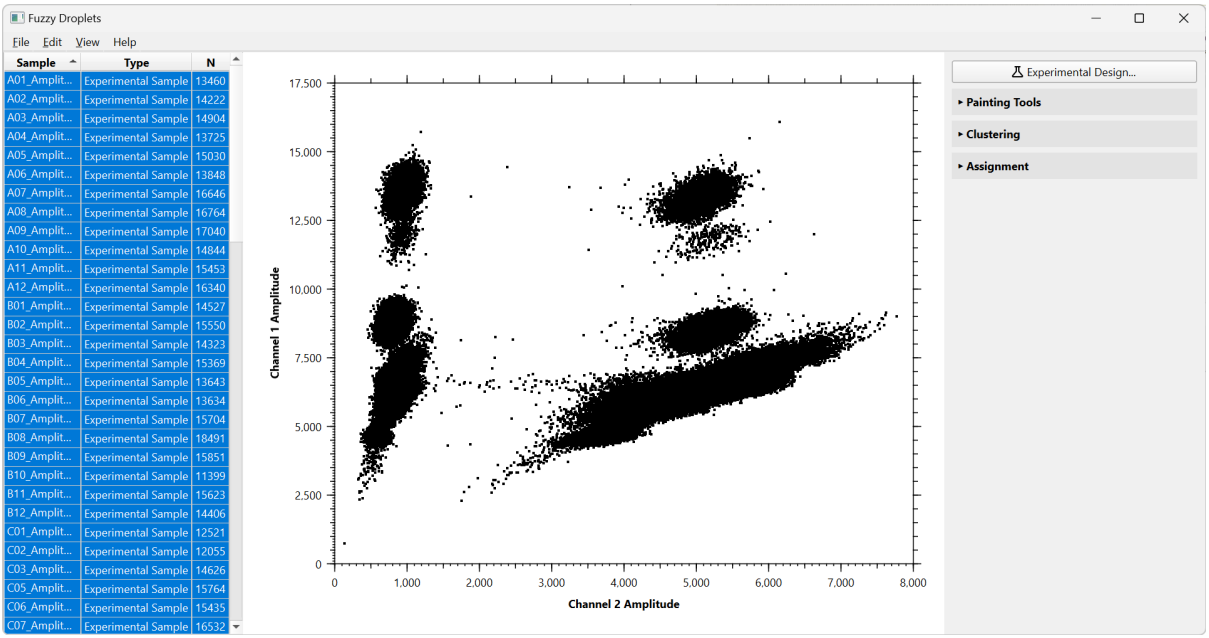
If the droplet has been assigned fuzzy cluster membership, the first two columns should be followed by *N* columns when there are *N* clusters. The *n*th column contains the probability that the droplet is a member of cluster *n*. For this reason, the sum of these values should be less than or equal to one (if less than one, the remaining value is the probability of being unassigned, i.e., a member of cluster zero). However, if they sum to more than one, Fuzzy Droplets will normalize them. Hence, a fuzzy droplet in an experiment generating eight clusters could look like:

```
17254.93, 8002.4, 0, 0, 0.3, 0.7, 0, 0, 0, 0
```

This is interpreted as the droplet having 30% likelihood of being part of cluster 3, and 70% likelihood of being part of cluster 4.

Within a single csv file, all rows should be in the same format.

The csv file may contain a header line, but this is currently ignored by Fuzzy Droplets. Here is some raw data loaded into the software:



Our data files are titled such that H10_Amplitude.csv contains information about the tenth well on plate H. The number of droplets in each well is listed under N on the left panel. Samples can be sorted by clicking on the “Sample”, “Type” and “N” headers. Here, all droplets are coloured black (or white, if you use a dark theme) indicating that it is unclassified data. In the figure above, all samples are selected. You can click on the list in the left panel to select individual samples, or control-click/shift-click to selected multiple samples and ranges.

Step 2: Specify sample types

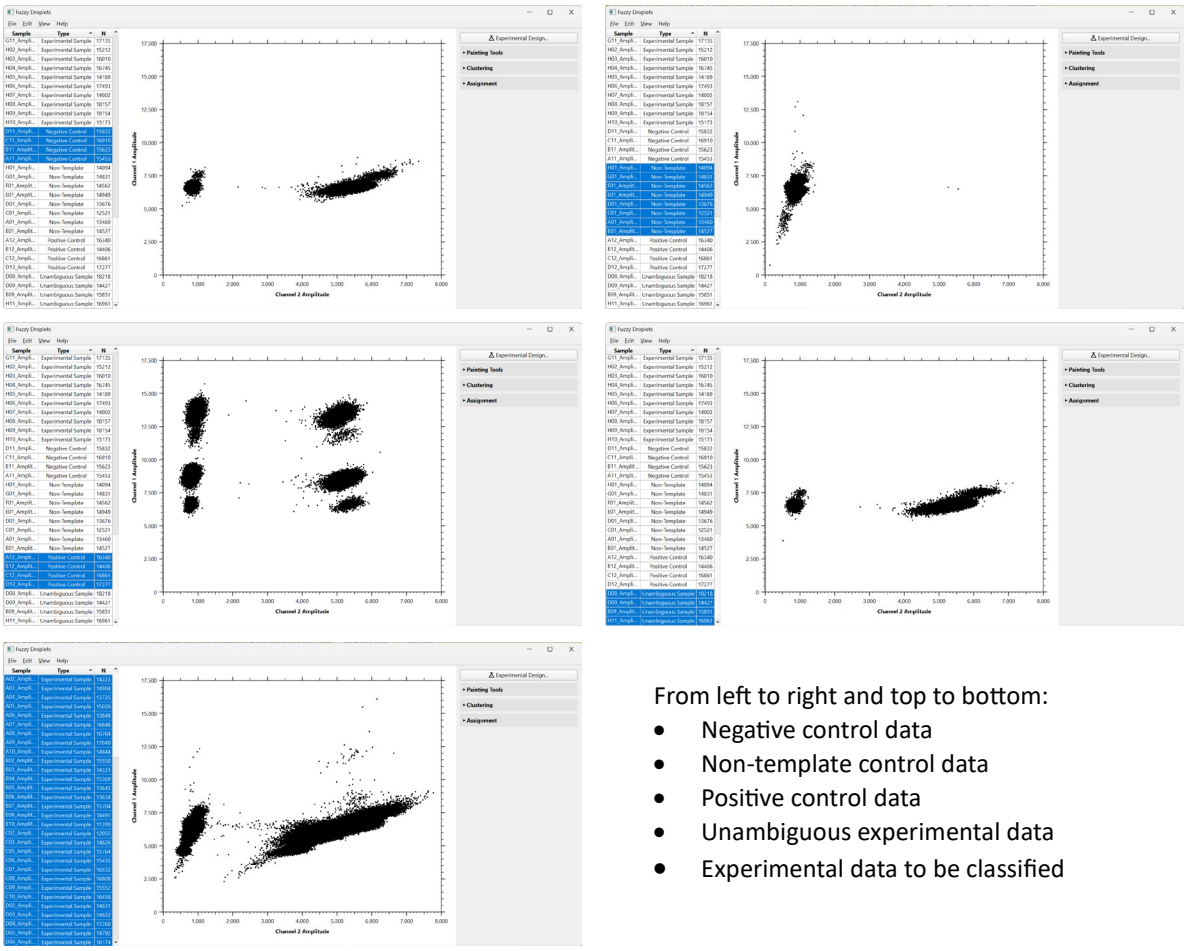
Your samples will consist of several different types:

- Positive controls: wells known to contain each of the targets
- Negative controls: wells known to contain nucleic acids but none of the targets
- Non-template controls: wells containing all of the reagents, but no biological material
- Experimental Samples: wells containing experimental data that you would like to classify
- Unambiguous Samples: wells containing experimental data that nevertheless can be unambiguously labelled and used as training data for the machine learning model

The goal of Fuzzy Droplets is to use the control data to train machine learning models to classify the sample data. So, we need to label each sample according to its sample type. This is accomplished by double clicking on a sample’s “Type” column in the left panel and selecting the appropriate sample type from the dropdown list, or by selecting multiple samples, right clicking on them, and setting their type using the context menu.

Here we have assigned sample types and sorted samples based on type by clicking on the “Type” column header. By selecting each sample type in turn (you can do this manually or by the menu using Edit > Selection) we can examine our training data, and the data that we want to classify. As you can see, in our particular data (see Step 3 below for more information) our positive controls contain eight

clusters, our negative controls contain two clusters, and our non-template controls contain one cluster.



From left to right and top to bottom:

- Negative control data
- Non-template control data
- Positive control data
- Unambiguous experimental data
- Experimental data to be classified

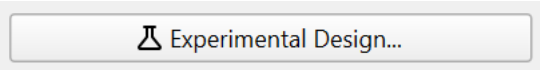
Step 3: Set up experimental design

We should provide information about the number, names and approximate locations of nucleotide targets in our data. This information is used to improve some of the clustering algorithms and used to properly label the clusters in the output.

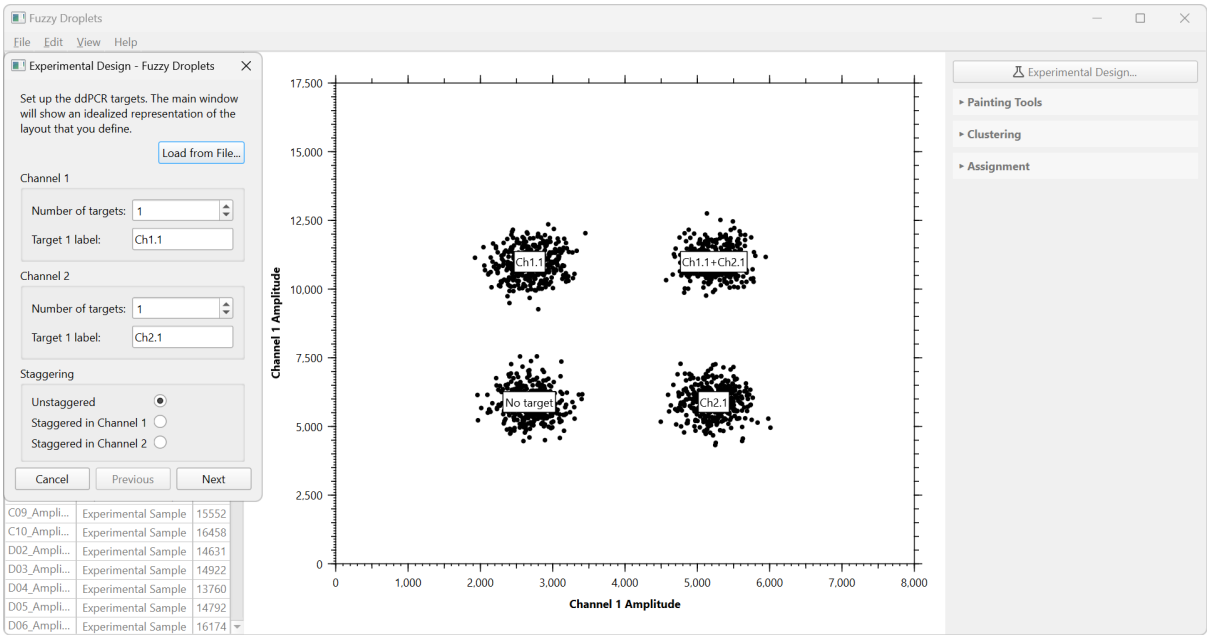
In the data illustrated in this tutorial, we have two targets on channel 1 (GFP and SRY) and one target on channel 2 (TERT). This gives rise to eight possible classes of droplets with respect to the targets for which they are positive, reflected in the eight clusters in our positive control data:

- **No targets**
- **GFP**
- **SRY**
- **GFP+SRY**
- **TERT**
- **TERT+GFP**
- **TERT+SRY**
- **TERT+GFP+SRY**

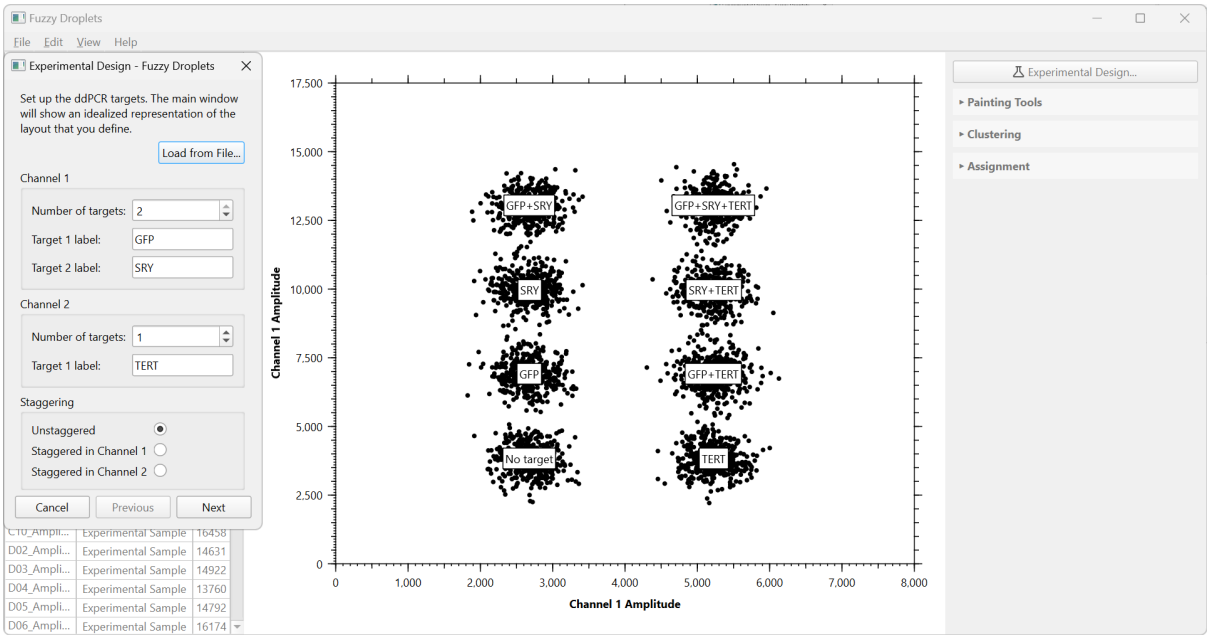
In order to provide this information, first select the positive control data, then press the “Experimental Design…” button or access it via the Edit menu.



This will launch the experimental design setup wizard.



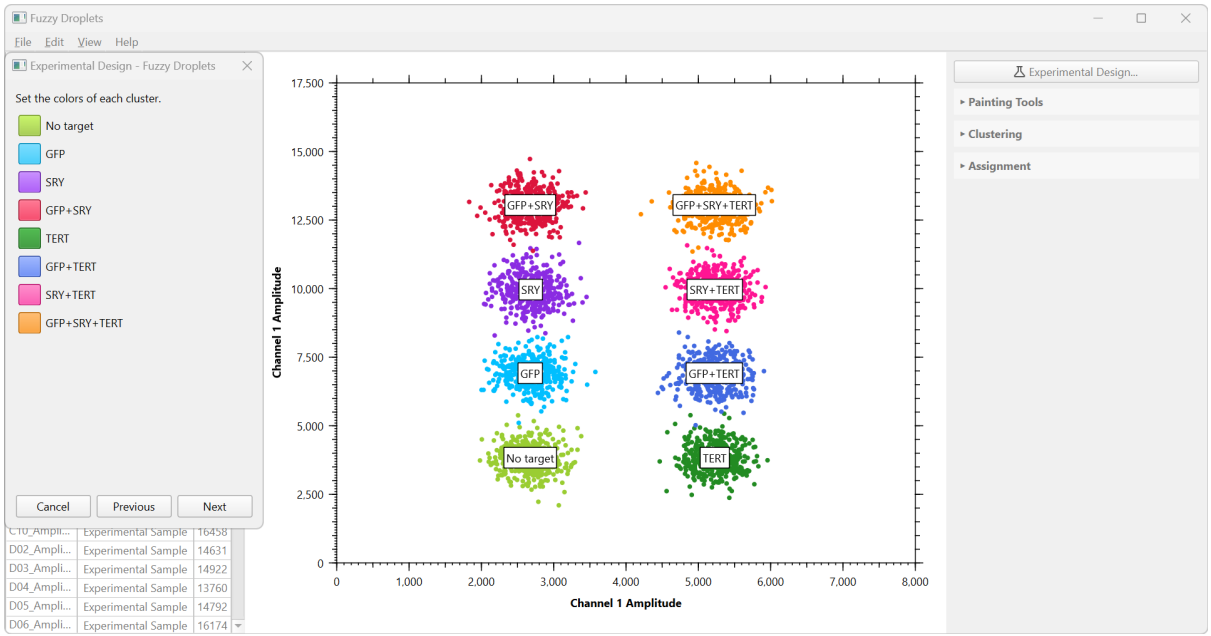
By default, there is one target on each channel. In the graph, we can see an idealized representation of the layout of clusters associated with this number of targets, along with the targets for which these clusters are positive. Modify the number of targets and their labels and the idealized representation will be updated appropriately. You can also setup a staggered experiment using the radio buttons at the bottom of the wizard. Here is the situation after our data has been entered:



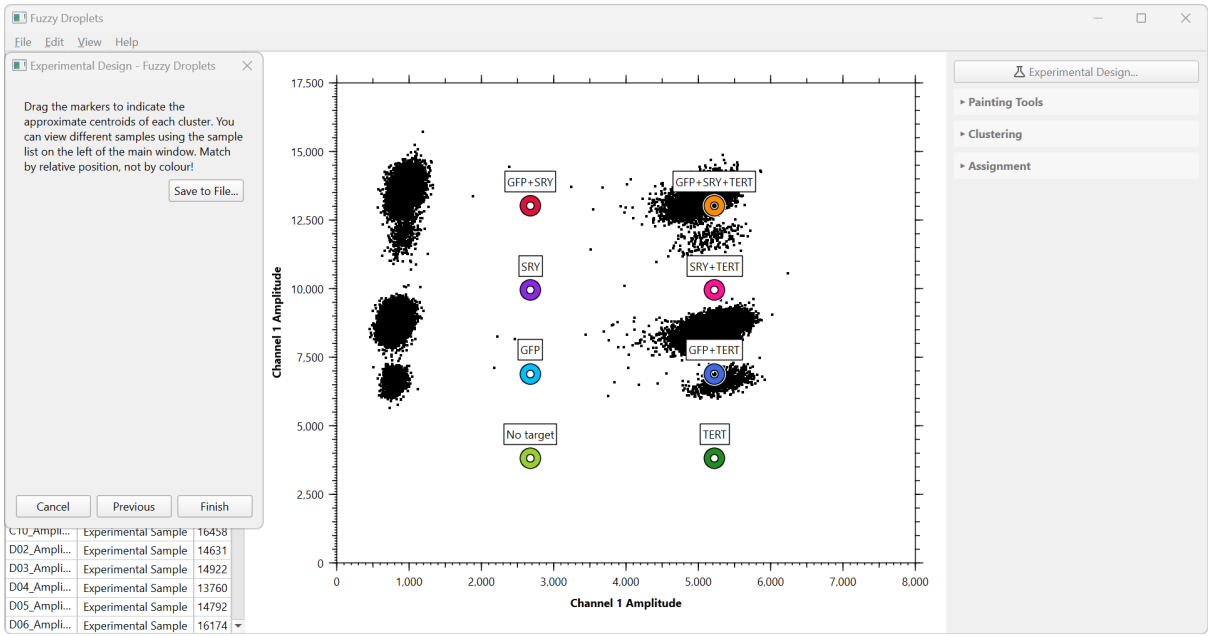
When you are happy, click next. **You must ensure that the labels are set correctly because this will be used in the output!**

If you have previously set up an experimental design, you can load it from file in this first page of the wizard. The option to save your design is available on the last page of the wizard.

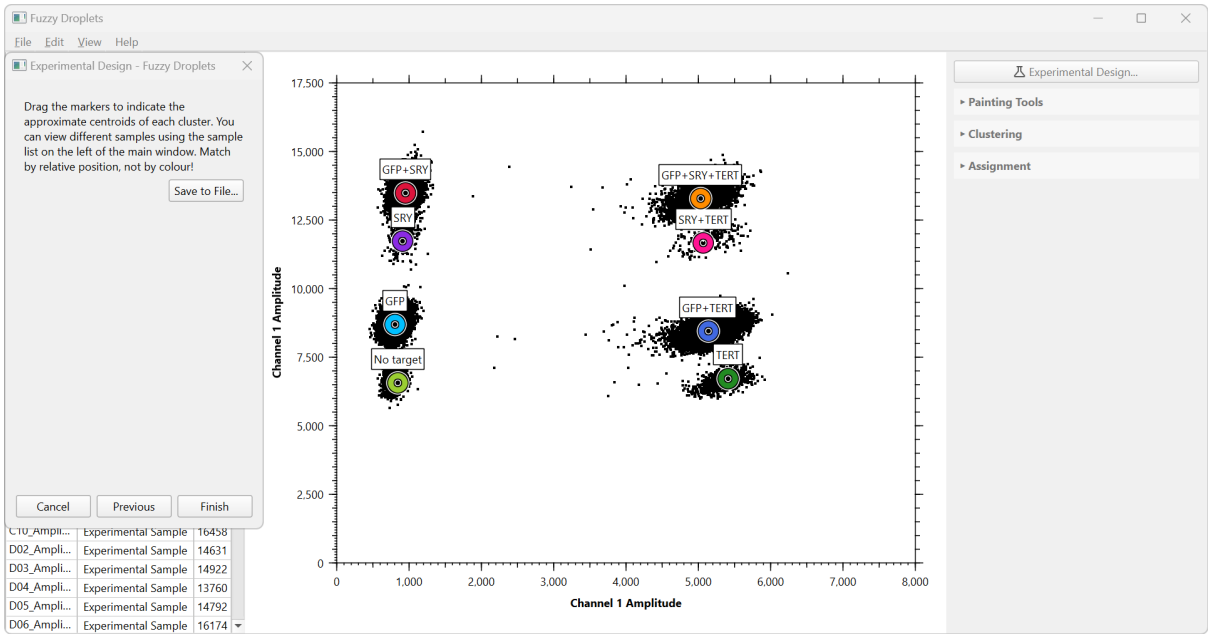
The next page of the wizard allows us to choose the colours we would like to use to represent each cluster. Click on the square with a colour you wish to change, and a colour chooser will pop up. Black and white are reserved for unclassified data.



When you are happy, click Next.



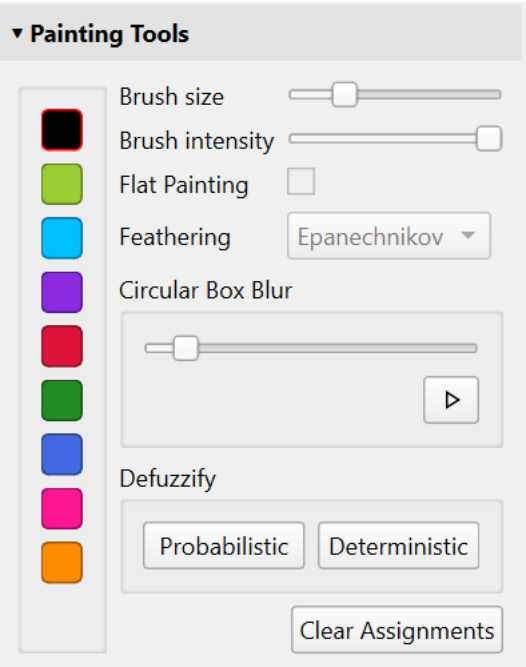
In the final page of the wizard, we should indicate the approximate centre of each cluster. You will see coloured rings representing each cluster, each with the appropriate label. Drag each ring to the approximate centre of its cluster. Remember not to mix up the relative position of the centroids. Here is our result:



When satisfied, click Finish. If your original data already had colours assigned, these will be adjusted to correspond to the colours you have entered in the wizard.

Step 4: Label the training data

The easiest samples to label are the non-template controls. These by definition have only one cluster, so we merely need to apply the appropriate colour (in our case, light green) to the data. We will do this using the Painting Tools, which can opened by clicking on the shelf in the right hand panel.



On the left is the palette, containing the eight colours we specified in the experimental design wizard, along with black to represent unlabelled data. If your computer uses a dark desktop theme, this will be white instead, so that it remains visible.

To paint on the graph, simply select the colour from the palette and stroke your mouse over the droplets whose colour you wish to change. You can modify the brush size from very small to very large. When you have finished, leave painting mode by unselecting the colour (click on it again).

Changing the brush intensity allows fuzzy painting (it allows intermediate colours to be formed by blending, which are used to represent uncertainty about the assignment of droplets to any particular cluster).

When the brush intensity is less than 100%, you are allowed to check “Flat Painting”. This applies a fixed amount of the selected colour to your data. For example, if set to 60%, any droplets you paint will be set to 60% of the selected colour, with their previous colours scaled such that the total amounts to 100%. This allows you to set precise levels of ambiguity (i.e. 50% red and 50% blue).

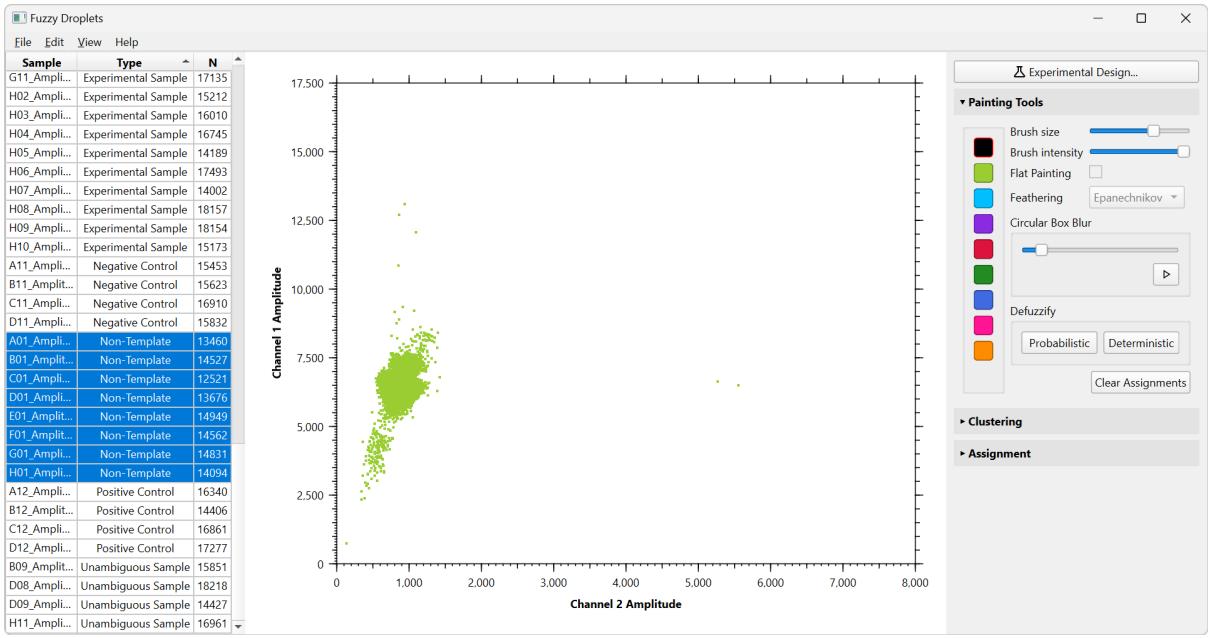
Feathering changes the amount of softness around the edge of the brush (though these options do not make much difference in practice).

The Clear Assignments button will remove all labels from the currently selected data.

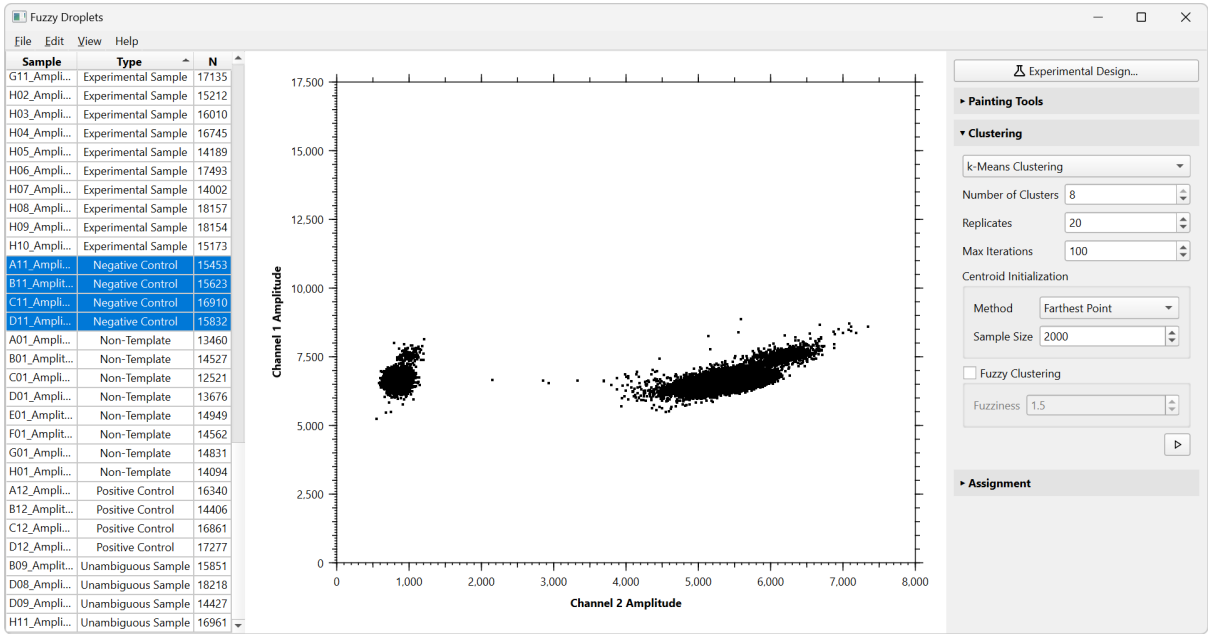
Fuzziness can be removed at any time from the currently selected samples by using the defuzzify. First, the data can be unfuzzified deterministically. In this case, each fuzzy droplet will be assigned unfuzzy member of the cluster for which that droplet has the highest probability of membership. Second, the data can be unfuzzified probabilistically. In this case, each fuzzy droplet is assigned randomly to a cluster depending on the probability of cluster membership. For example, if a droplet is 20% red and 80% blue, it will be assigned fully to red with a probability of 0.2 and to blue with a probability of 0.8.

Fuzziness can also be added any time using the circular box blur tool. This makes each droplet the weighted average colour of its neighbours within a certain radius specified by the slider. The size of the radius is shown when moving the slider as a grey circle (or oval if your axes are of different scales). Hence, you have control over the amount of fuzziness in the data you export.

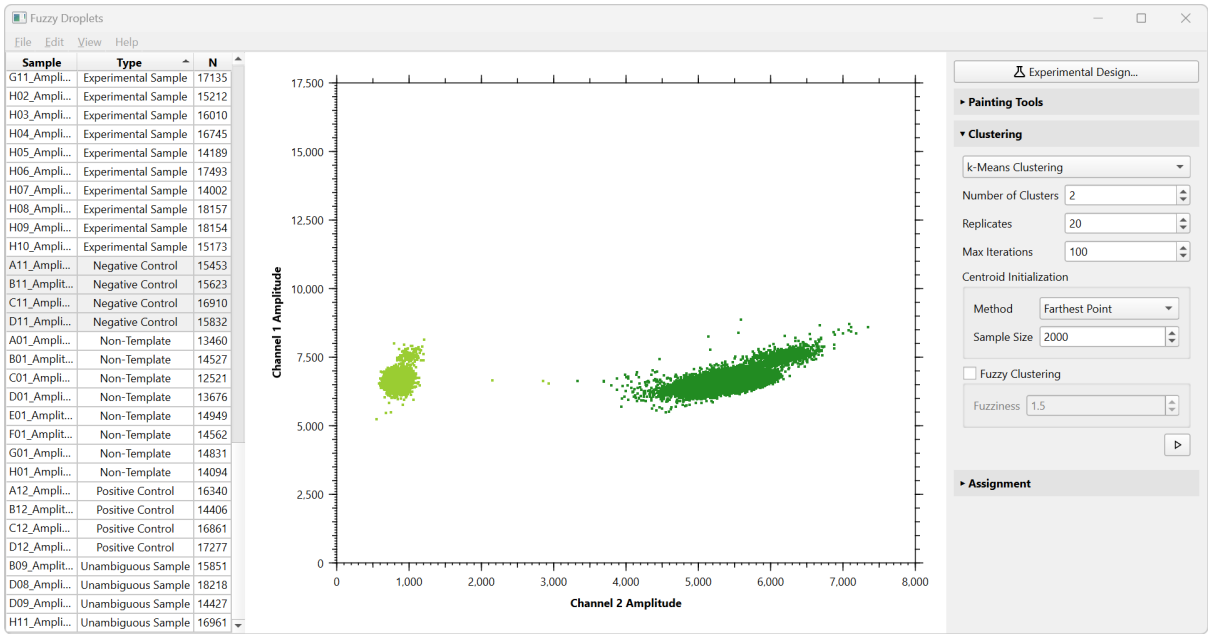
Here, we have selected all of the non-template controls and painted them the appropriate colour based on the colours we assigned to clusters earlier:



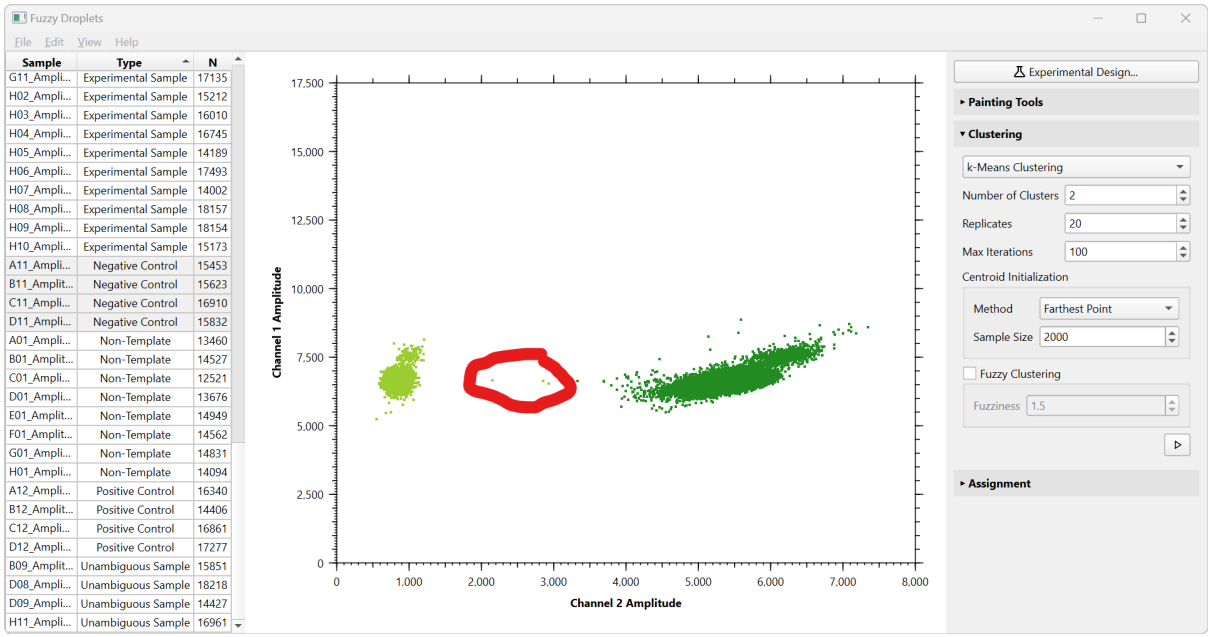
Next, we will label the negative control data. Close painting tools, select all of the negative control samples, and open the Clustering tools.



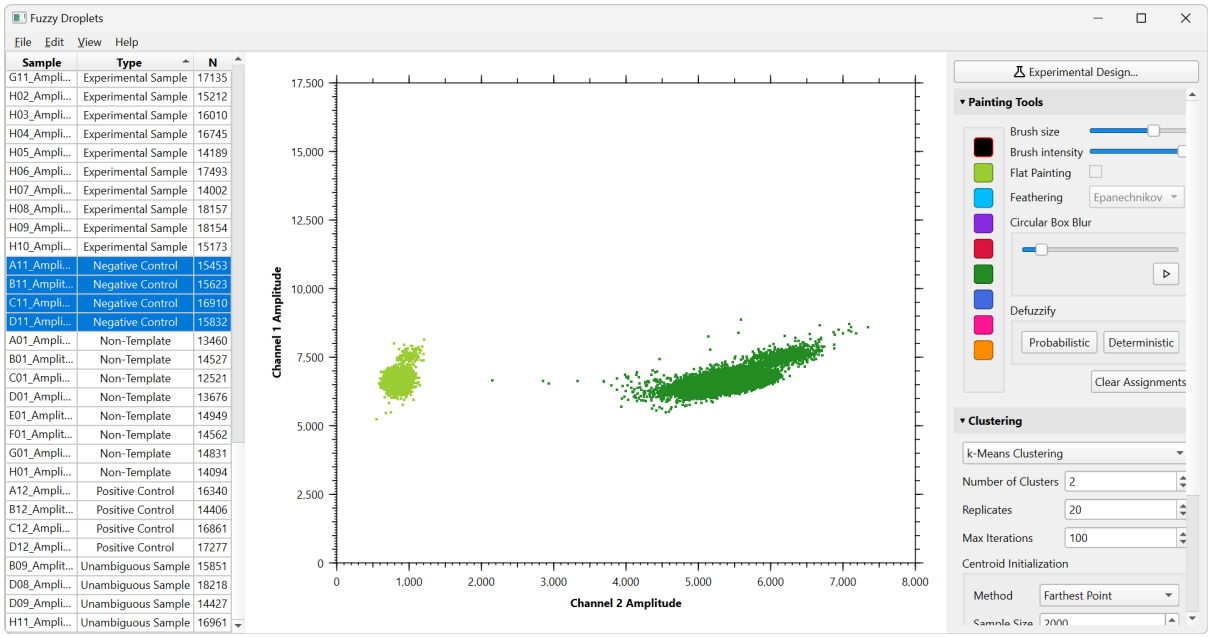
We could actually use the painting tools to label this data as well, but we will here demonstrate clustering. Several methods are provided and can be chosen using the drop-down list at the top of the clustering panel. If the method you use allows the number of clusters to be set, it should be set appropriately. Here we will demonstrate k-means clustering. We set the number of clusters to two and leave all of the other options as default. Run the clustering algorithm by pressing the button in the bottom right.



As you can see, k-means has worked well here because the clusters are well separated. Fuzzy Droplets automatically assigns the correct colour to each cluster, based on the information you specified in the experimental design wizard. If you look closely, you will see some “rain” droplets between the dark green and light green cluster. Rain typically falls in a downward direction, so these should probably be coloured dark green, but k-means has labelled them light green.

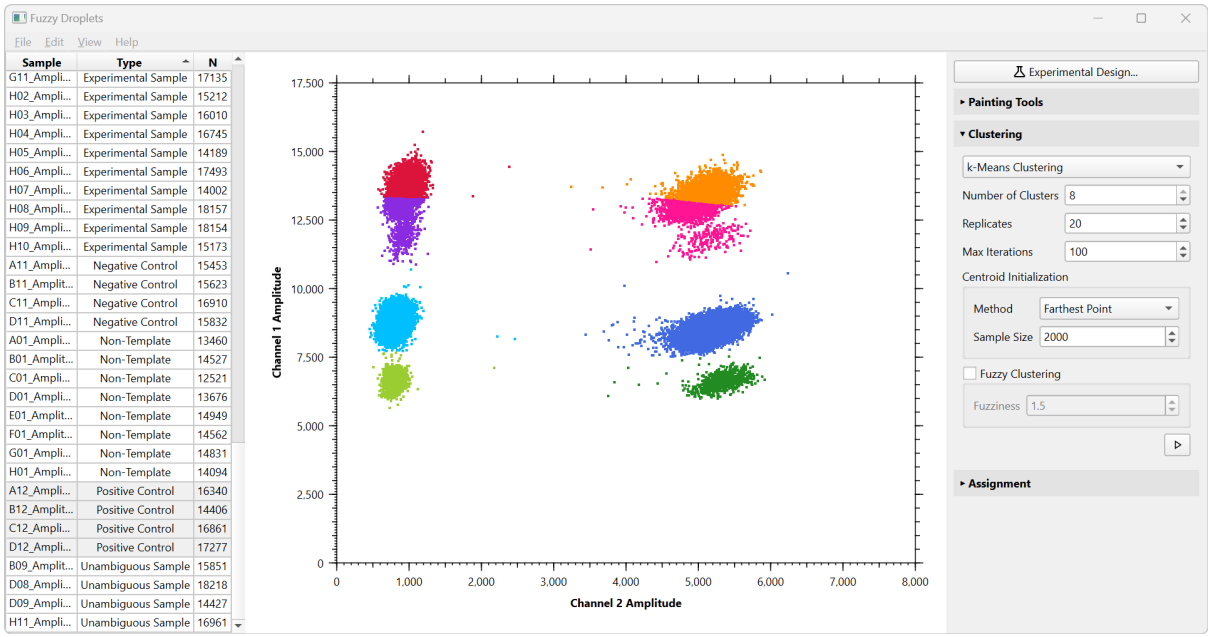


We can use the painting tools to correct this error:

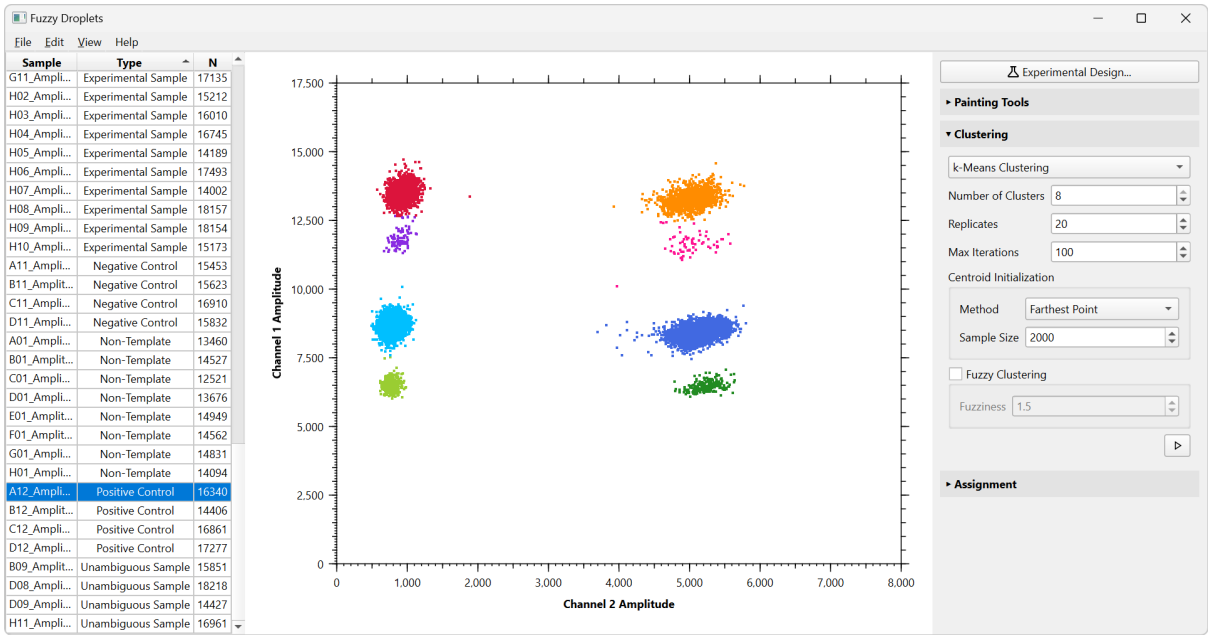


We can label our unambiguous samples in a similar way.

Now, let's turn to the positive control data with all eight clusters. We set the number of clusters to eight, and run k-means as above:

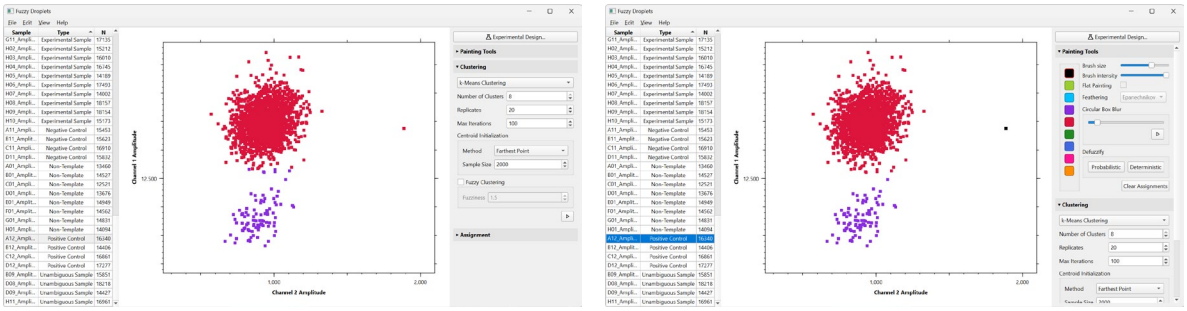


In this case, k-means clustering performs poorly. This is because the clusters vary a lot in size and density. To improve clustering we will label each positive control sample in turn. We select the first sample and apply k-means as above. It works quite well:

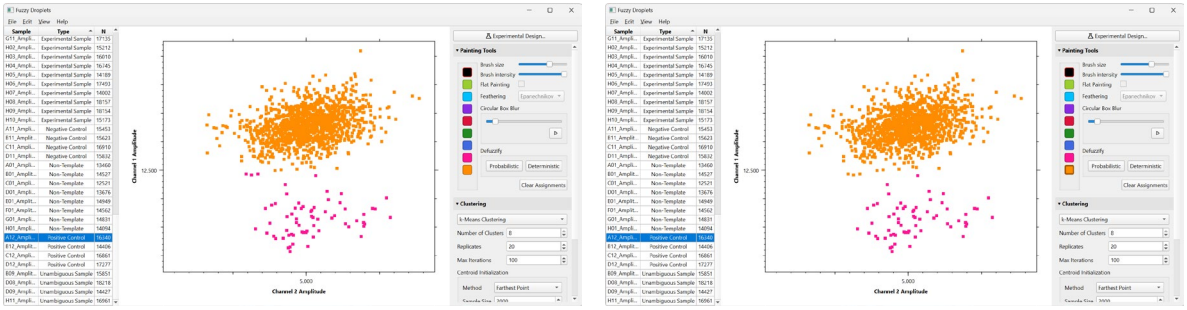


There are some apparent errors in the upper half of the graph, where the clusters are very close to each other. We will use the painting tools to adjust this. We will also zoom and pan into the regions of interest. Zooming is accomplished depending on your platform and hardware. Zooming could involve “pinch-to-zoom” on a touchpad, control + mouse wheel, or control + swipe. Zooming can also be accomplished by the actions in the View menu. Panning could involve mouse wheel to pan vertically and shift + mouse wheel to pan horizontally, or swiping on a touchpad. You can always return to view the whole graph using View -> Reset Zoom. For the purposes of creating figures here we also adjust marker size using the options in the View menu.

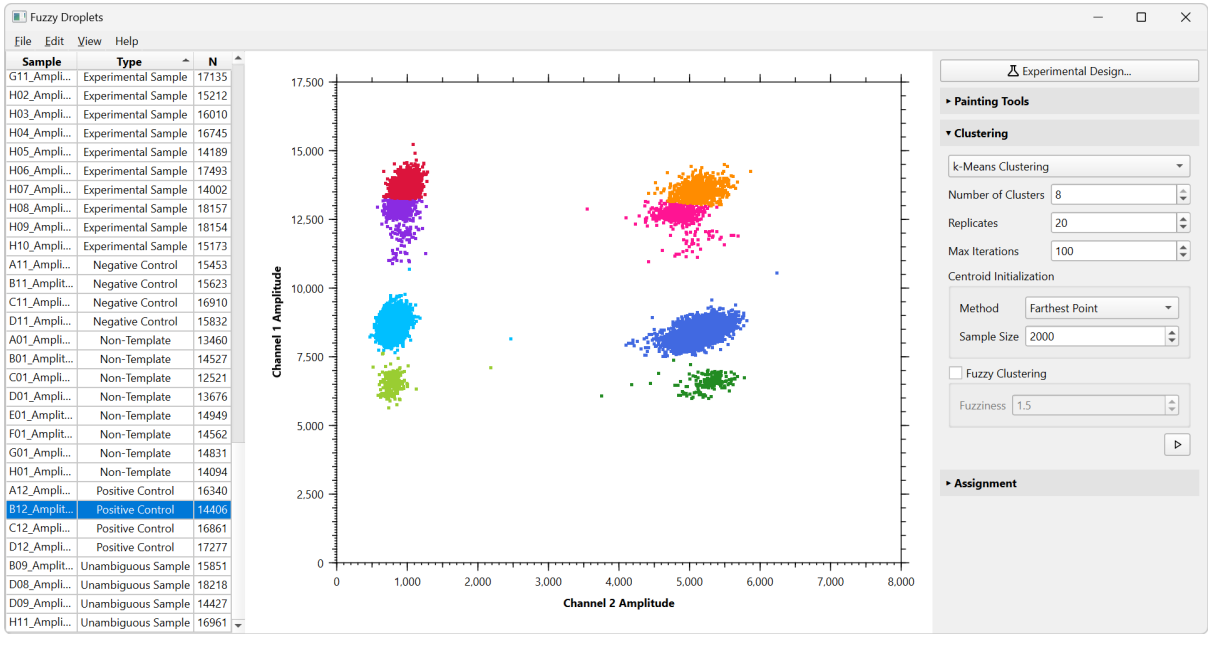
Below shows the before (left) and after (right). You will notice we have also painted a wayward droplet black, to show that we are fully uncertain about what colour it really should be.



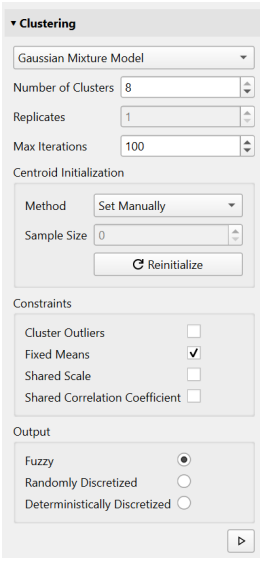
Similarly, we make adjustments on the right-hand side:



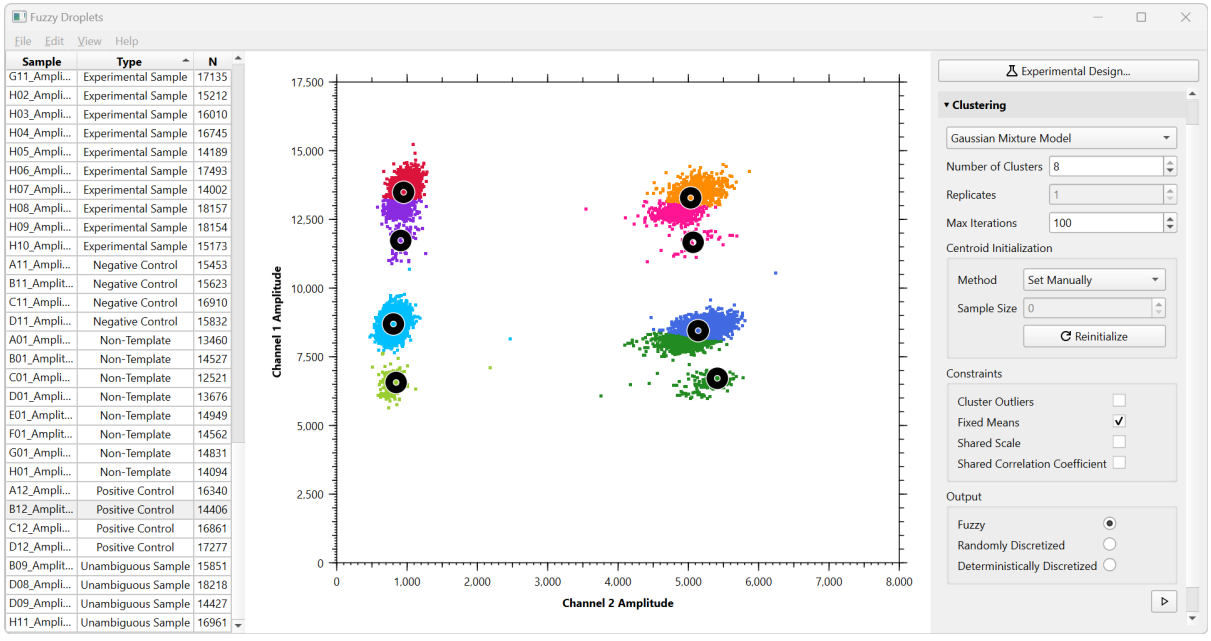
K-means clustering fails on the second positive control sample because of the closeness of the upper clusters. It seems that information about the location and size of the clusters is unlikely to be automatically deduced from the data that is available:



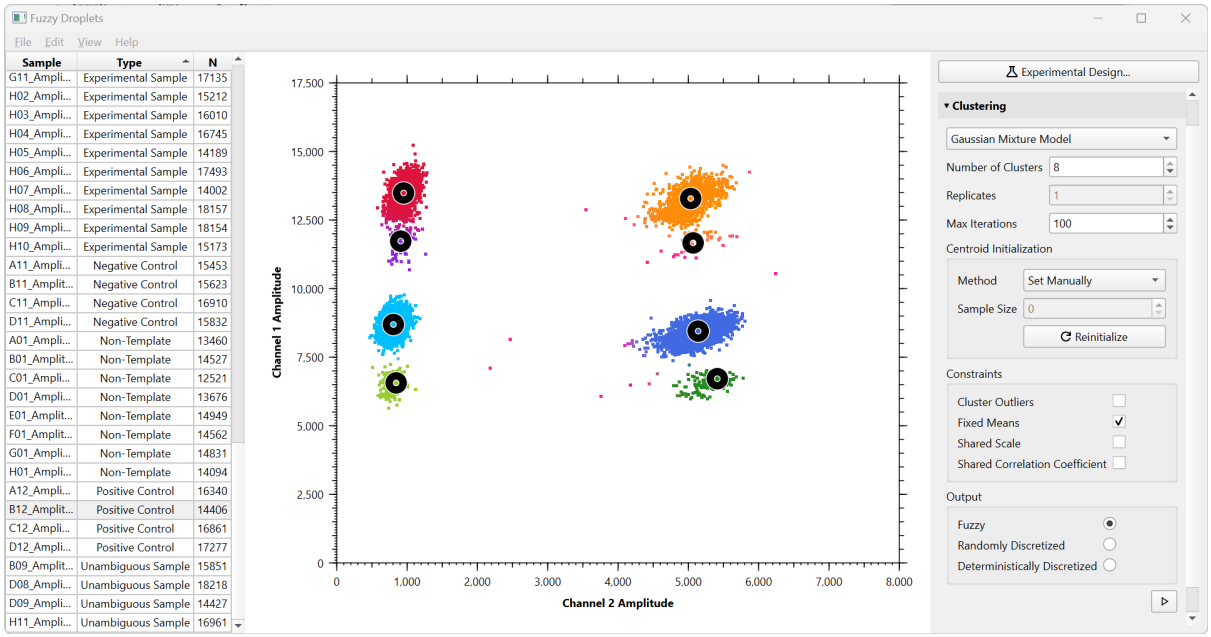
In this case we will use a highly constrained Gaussian mixture model. This consists of a set of bivariate normal distributions fitted to the data. We will fix the centroids of these distributions so that only their scale and correlation coefficients are estimated during model fitting. Change k-means to Gaussian mixture model in the drop-down box and set the appropriate constraints:



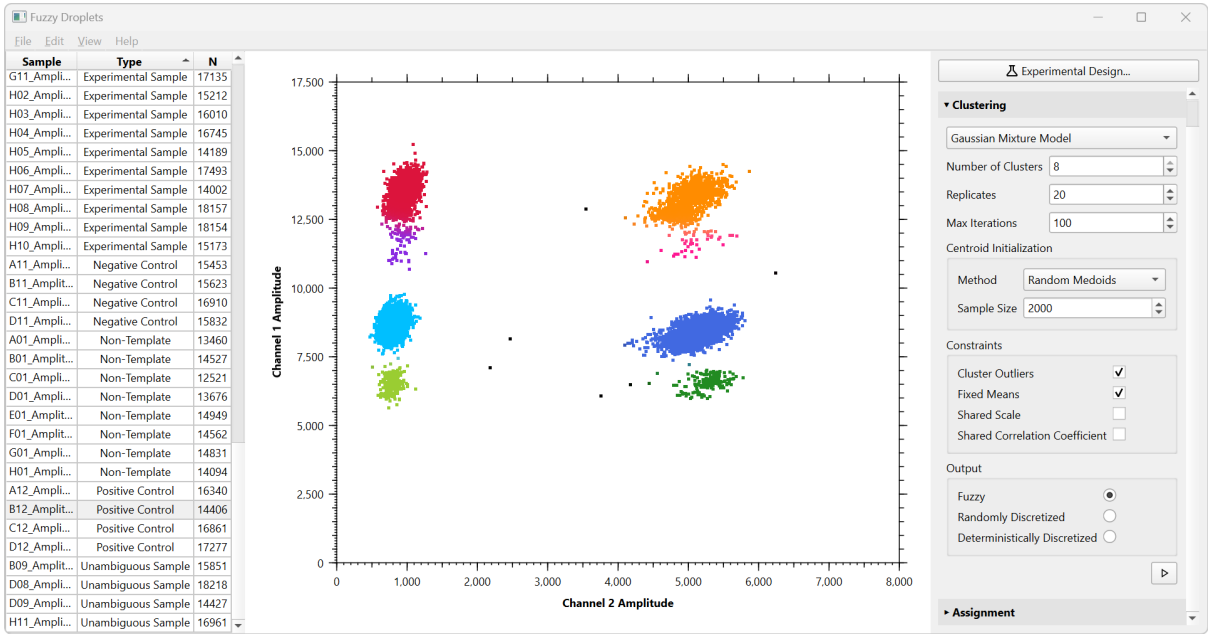
Since we will use fixed means, we need a way of specifying. In the Centroid Initialization box, I have chosen “Set Manually”. This makes the eight centroids appear on the graph as shown below. The centroids initial position is based on what we entered during experimental design, but they can be dragged around if you need to reposition them. To me, they look ok.



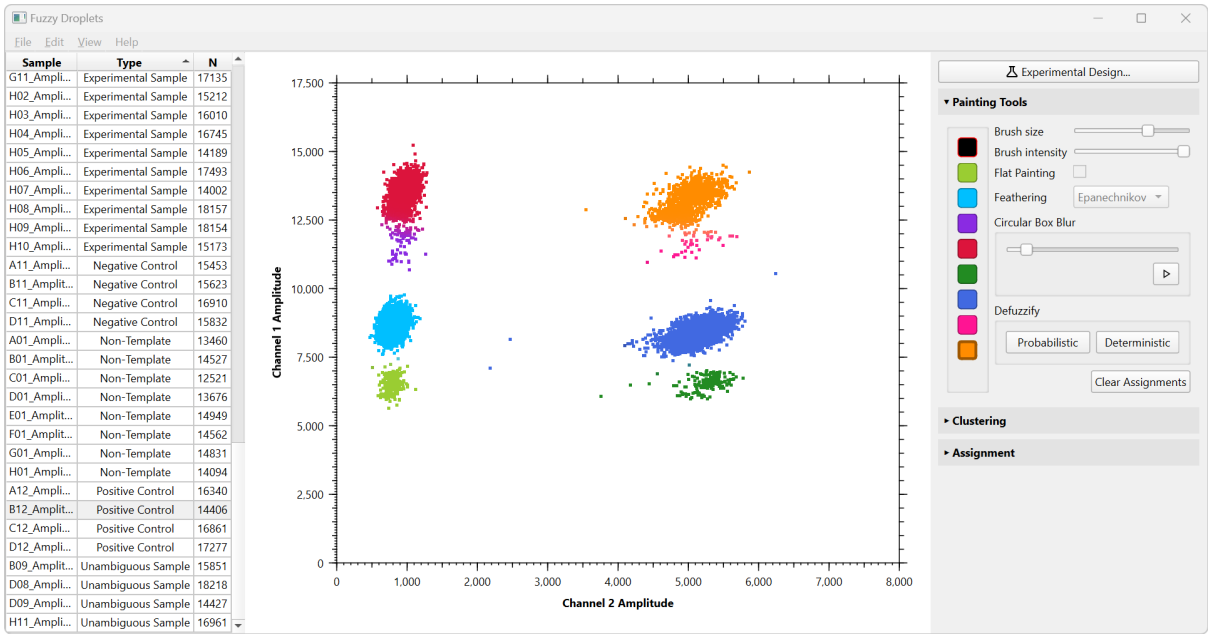
After running the GMM, we find that this has worked better than k-means, but some of the rainy outliers are unexpected colours. Specifically, in order to maximize the likelihood of the fit, the pink cluster has been given a very large scale so that it can scoop up all of the outliers, allowing the other denser clusters to have smaller scales and higher likelihoods:



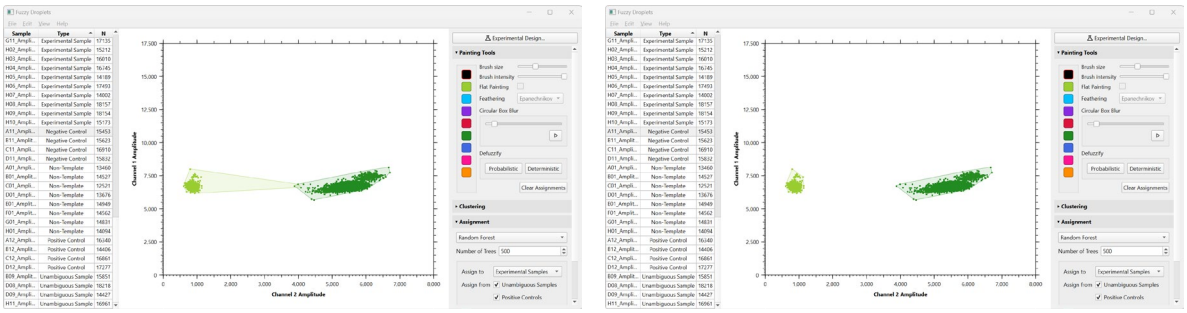
We can fix this problem in various ways. First, we could force the GMM to also have shared scales. This means that each cluster has exactly the same size. This doesn't reflect the reality of our data very well. Instead, we can explicitly "cluster outliers". This allows outlying droplets to be assigned to the "unclassified" category (i.e. they will be painted in black). With outliers explicitly clustered, the results look better. Note that the GMM has introduced appropriate fuzziness into the top left, where red grades into purple.



I use the Painting Tools to assign the black droplets to their correct clusters:

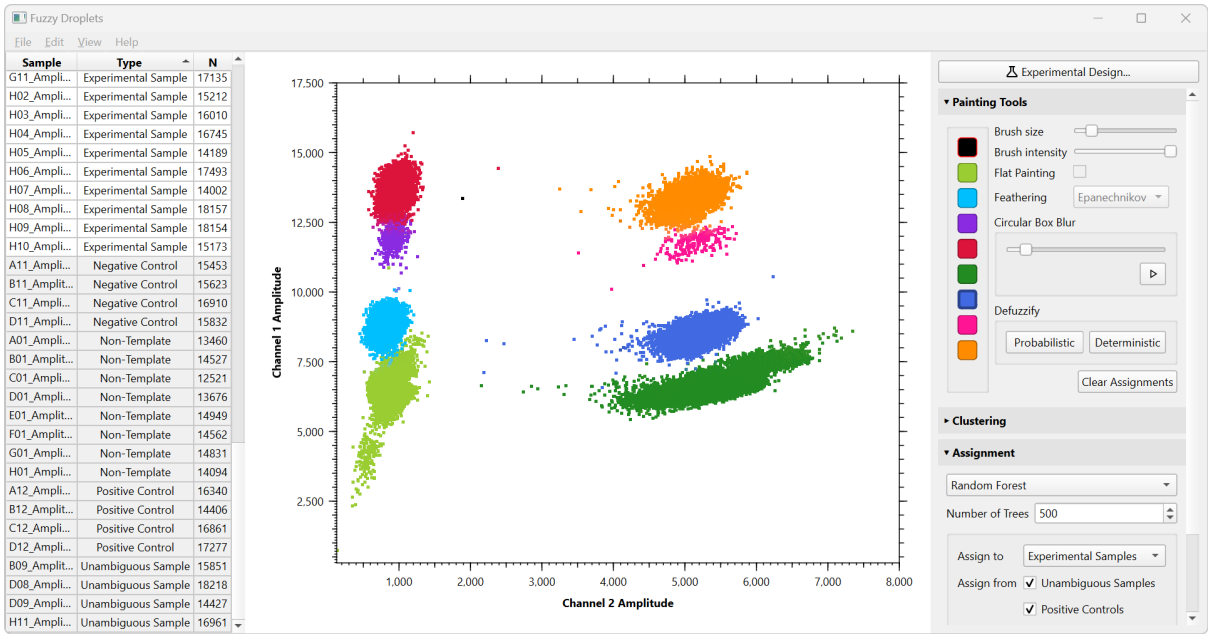


Proceeding this way through all of the positive control data, we finish labelling the training data. It is important to carefully check the quality of the labels that we have specified. One aid provided by Fuzzy Droplets is that it can draw convex hulls around clusters (View -> Convex Hulls, or right click for the context menu, or Ctrl+H). These can clearly indicate wayward droplets that are misclassified. Here, I have fixed them using the Painting Tools:



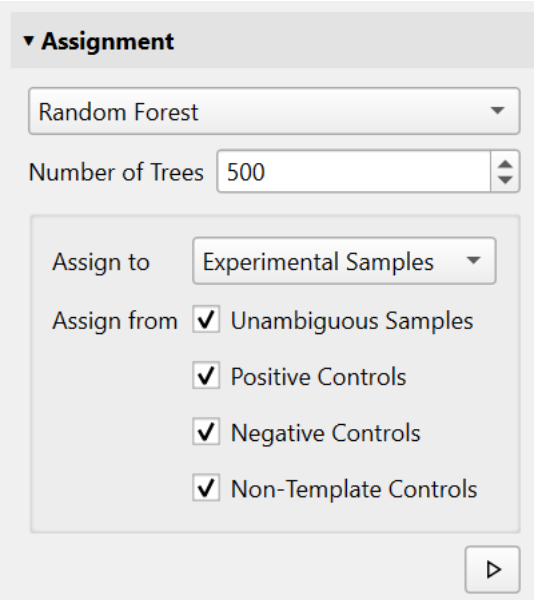
Fuzzy Droplets also allows you to adjust the z-order of the different colours, by right clicking on a coloured droplet and choose from the options in the context menu. This can help to reveal wayward droplets by bringing them to the front of the graph where they are more visible.

Finally, here is the combined training data for our experiment:

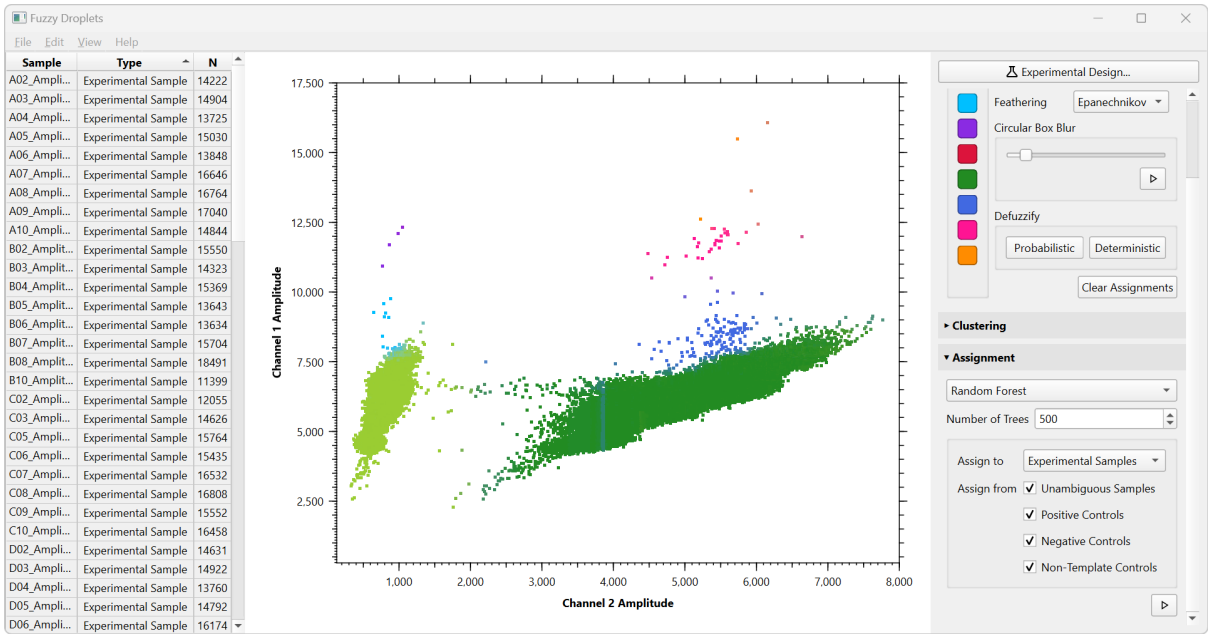


Step 5: Classify the experimental data

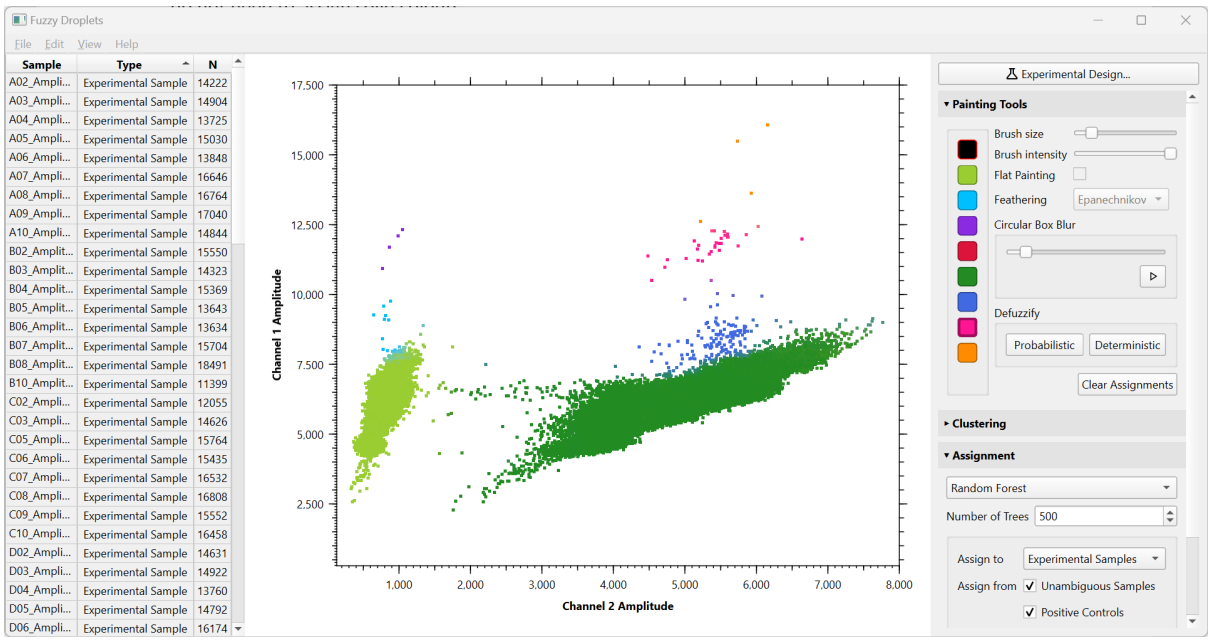
We will use the control data to train a machine learning model which will then classify the experimental samples. Open the Assignment section of the right panel. Here I will use a simple and efficient method, “Random Forest”.



This panel allows you to specify the training data (Assign from) and the data to be classified (Assign to). After running the classification, the experimental samples have been assigned to clusters in a fuzzy manner:



As you can see, there has been a little difficulty in assigning cluster memberships to droplets in the far left region of the dark green cluster. This is because the experimental data exhibits substantially more scatter in this dimension than the training data, and the classifier essentially has no local information about what colour a droplet in this region should be. One important lesson here is that the training data should exhibit scatter similar to what is seen in the experimental data. But, the messy data we are classifying in this example is probably unusual in this regard. There is also a faint blue vertical line in the left side of the dark green cluster, perhaps due to poor labelling of the training data or imperfections in the underlying random forest algorithm. We can fix the classification using fuzzy painting tools. I set the brush intensity to 20% and gently stroke the affected region to improve the classification. To the extent that we believe this represents real ambiguity, we do not need to assign solid colours:



Note that elsewhere, for example the area between light green and light blue, the classifier has introduced an appropriate degree of uncertainty which genuinely reflects the underlying data.

Step 6: Export data

The labelled data can be exported into a folder by choosing File -> Save As. Fuzzy Droplets will also generate a table of raw count data for each cluster by choosing File -> Export Assignment Report. The count data will be integers if the data is unfuzzy, but continuous numbers if the data is fuzzy. In the latter case, the count for a cluster is the sum of probabilities of being members of that cluster across all droplets. The sum of these counts across all clusters in a sample will be equal to the total number of droplets, so it is analogous to integer counts.

Fuzzy Droplets can output high quality images ready for publication. Use File > Export Image to create either jpeg or tiff versions of the current view on the graph.