

Cellprofiler and Cellprofiler Analyst Practical

Cédric Hassen-Khodja, Volker Baeker, Clément Benedetti, Thibault Odor

1. Materials

You will analyze images of cells in a standard 384-well plate. However, you will focus on a subset of these images. These images are located in ./exercises/CP-CPA/images. A text file named **BBBC022_v1_image_subset.csv** is provided, indicating the location of the wells on the plate and the cell treatment. This file is located in ./exercises/CP-CPA/metadata.

2. Exercise: Using the CellProfiler software to identify features and obtain measurements from cellular images

1. Start **CellProfiler** by double-clicking the desktop icon 
2. Import the file named **CellPainting.cppipe** (located in ./exercises/CP-CPA/pipeline) by clicking on **Import > Pipeline from File....**

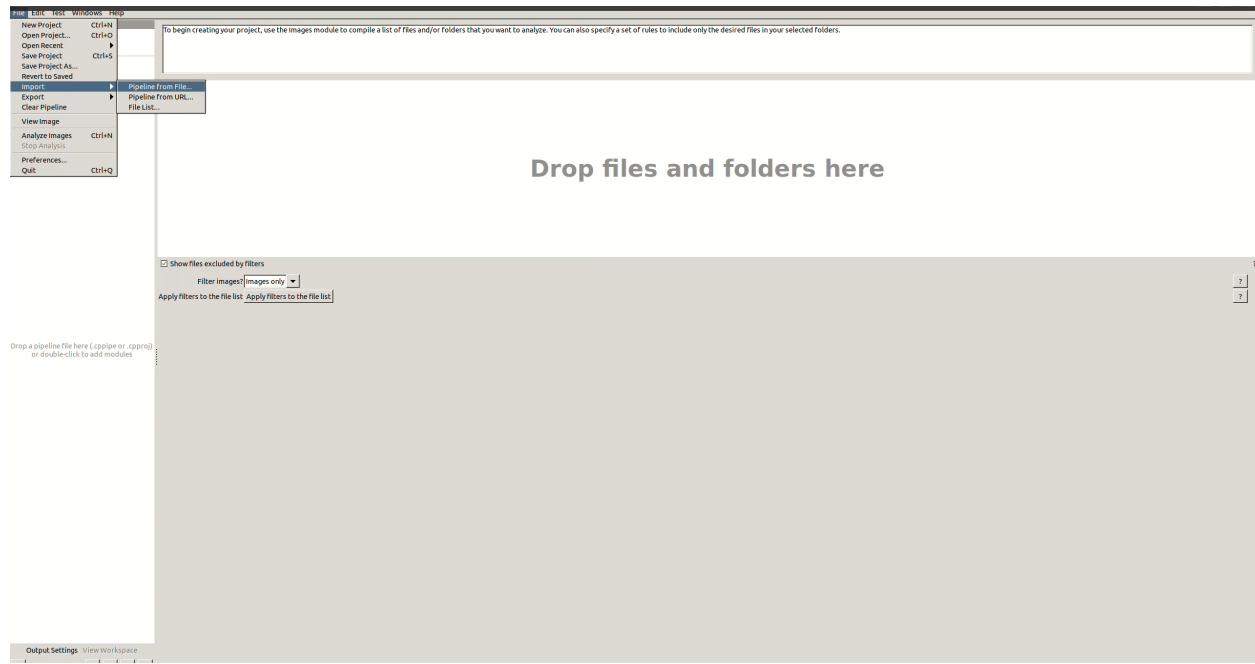


Figure 1: Import Pipeline from File

3. Within the interface, there's a “Drop files and folders here” section, part of the **Images** module’s main interface. Drag the *images* folder into this section. Your files from this folder will then be visible.

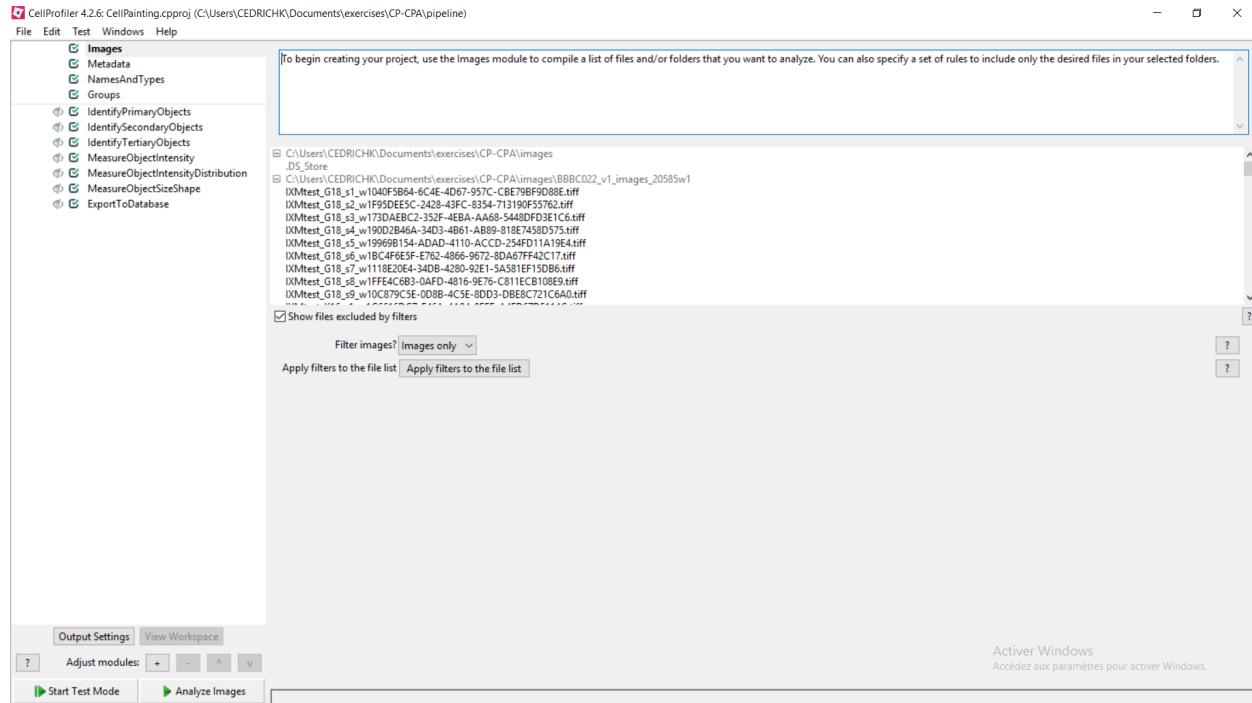


Figure 2: Drop images

4. Open the **Metadata** module, the second in the Input module panel. This module extracts details from image filenames. Click the magnifying glass icon  near *Regular expression*. A dialog will open for regular expression comparison. The regular expression is :

`^.*_(?P<Well>[A-P][0-9]{2})_s(?P<Site>[0-9])_w(?P<ChannelNumber>[0-9]).*`

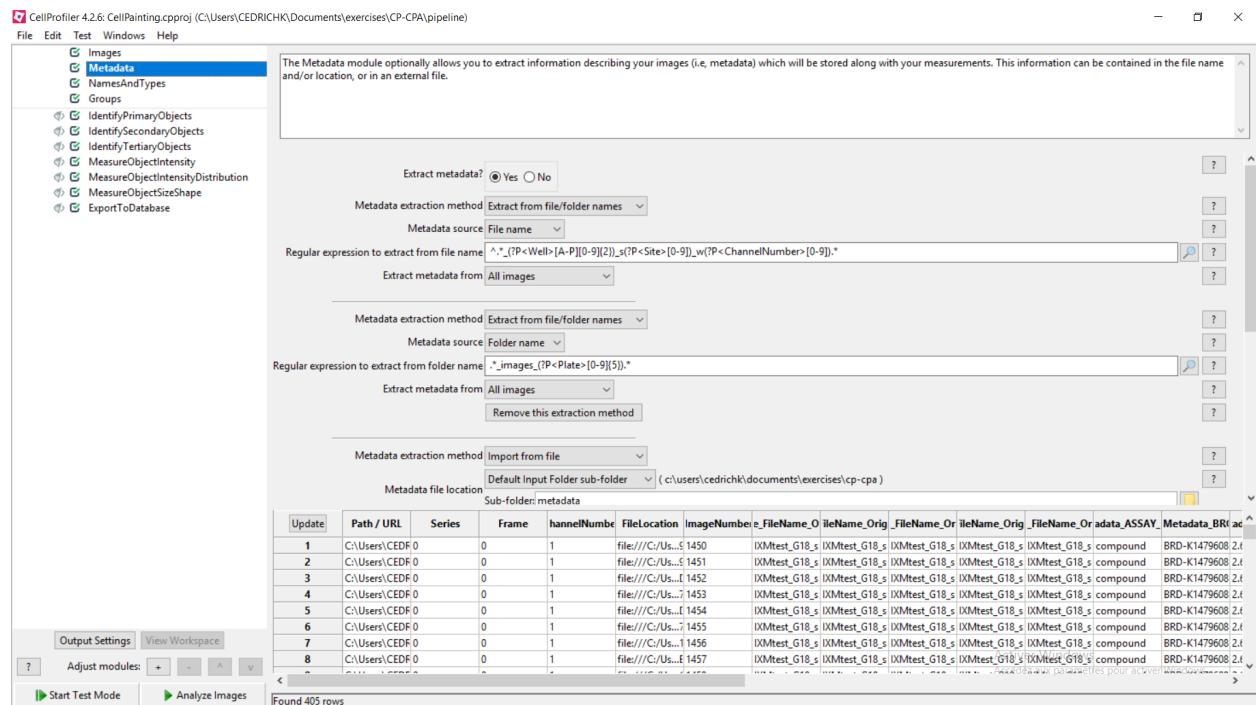
After setting up, click **Submit**.

Click on **Add another extraction method** to extract plate information with **Folder name** as *Metadata source*. The regular expression is:

`.*_images_(?P<Plate>[0-9]{5}).*`

For the third metadata extraction, navigate to the metadata folder and select the **BBBC022_v1_image_subset.csv** file.

Back in the **Metadata** module, click **Update**. You'll see columns in the Metadata window. Ensure that *Well*, *Plate* and *Site* matches in both the *CSV Metadata* and *Image Metadata* dropdowns.



Update	Path / URL	Series	Frame	channelNumber	FileLocation	ImageNumber	FileName_O	fileName_Orig	Metadata_BR
1	C:\Users\CEDRICHK\Documents\exercises\CP-CPA\image\BBBC022_v1_image_subset.csv	0	1	file:///C:/Users/.../1450	IXMtest_G18_s	IXMtest_G18_s	IXMtest_G18_s	IXMtest_G18_s	compound
2	C:\Users\CEDRICHK\Documents\exercises\CP-CPA\image\BBBC022_v1_image_subset.csv	0	1	file:///C:/Users/.../1451	IXMtest_G18_s	IXMtest_G18_s	IXMtest_G18_s	IXMtest_G18_s	compound
3	C:\Users\CEDRICHK\Documents\exercises\CP-CPA\image\BBBC022_v1_image_subset.csv	0	1	file:///C:/Users/.../1452	IXMtest_G18_s	IXMtest_G18_s	IXMtest_G18_s	IXMtest_G18_s	compound
4	C:\Users\CEDRICHK\Documents\exercises\CP-CPA\image\BBBC022_v1_image_subset.csv	0	1	file:///C:/Users/.../1453	IXMtest_G18_s	IXMtest_G18_s	IXMtest_G18_s	IXMtest_G18_s	compound
5	C:\Users\CEDRICHK\Documents\exercises\CP-CPA\image\BBBC022_v1_image_subset.csv	0	1	file:///C:/Users/.../1454	IXMtest_G18_s	IXMtest_G18_s	IXMtest_G18_s	IXMtest_G18_s	compound
6	C:\Users\CEDRICHK\Documents\exercises\CP-CPA\image\BBBC022_v1_image_subset.csv	0	1	file:///C:/Users/.../1455	IXMtest_G18_s	IXMtest_G18_s	IXMtest_G18_s	IXMtest_G18_s	compound
7	C:\Users\CEDRICHK\Documents\exercises\CP-CPA\image\BBBC022_v1_image_subset.csv	0	1	file:///C:/Users/.../1456	IXMtest_G18_s	IXMtest_G18_s	IXMtest_G18_s	IXMtest_G18_s	compound
8	C:\Users\CEDRICHK\Documents\exercises\CP-CPA\image\BBBC022_v1_image_subset.csv	0	1	file:///C:/Users/.../1457	IXMtest_G18_s	IXMtest_G18_s	IXMtest_G18_s	IXMtest_G18_s	compound
...

Figure 3: Metadata module

5. Click on the **NamesAndTypes** module, which is the third module in the Input module panel. This module allows you to assign a meaningful name to each image by which other modules will refer to it. Note how the images are assigned to channels: images containing “1” in their file name are assigned to the name “Hoechst”, those with “2” are assigned “ER”, those with “3” are assigned “Syto”, those with “4” are assigned “Ph_golgi” and those with “5” are assigned “Mito”.

Set Image set matching method to “Metadata”: Select **Plate** in all columns. Do the same for **Well** and **Site**.

Click the **Update** button below the divider to display a table that shows each channel pair matched up for the 81 wells in the assay.

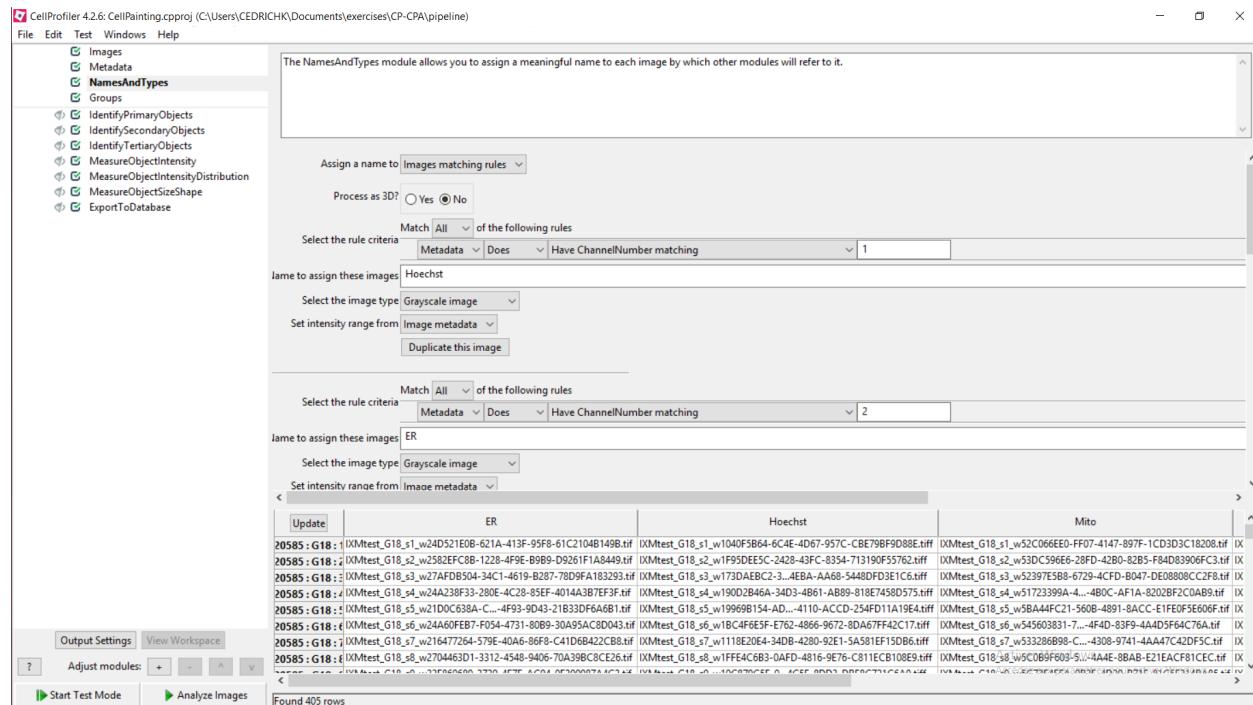


Figure 4: NamesAndTypes module

6. Click on the **ExportToDatabase** module to save measurements to a database. This allows CellProfiler Analyst to retrieve the data. Check the following parameters:

- Set the *Database type* to “SQLite”.
- Enable “Yes” for *Create a CellProfiler Analyst properties file*.
- Choose “Nuclei” for *Which objects should be used for locations?*.
- Select “384” for *Select the plate type*.
- For *plate and well metadata*, choose “Plate” and “Well” respectively.
- Set *Output file location* to “Default Output Folder.”
- Enable *Write image thumbnails directly to database?* and select “ER”, “Hoechst”, “Mito”, “Ph_golgi” and “Syto” using Ctrl-click (Windows) or Command-click (Mac).

Leave other settings at their default values.

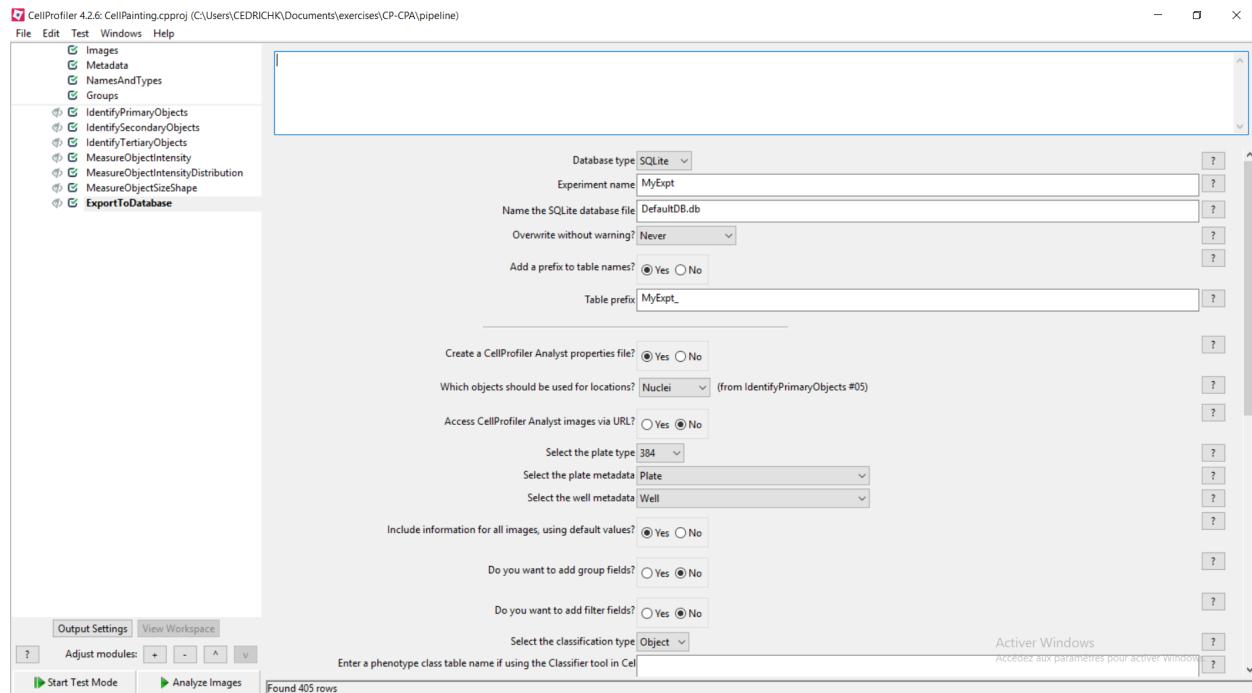


Figure 5: ExportToDatabase module

The settings for the CellProfiler pipeline have been pre-selected to optimize the analysis of the given images, making the pipeline ready for processing the full set of images.

7. For a more efficient run, hide all module display windows by choosing “Hide All Windows on run” from the *Windows* menu. The “eyeball” icons next to the modules will switch from open () to closed (). This speeds up the process as CellProfiler won’t need to generate and display each module’s window. Save Your pipeline by selecting **File > Save Project As...**, name the pipeline, and save it for future reference. Start the analysis of all images by clicking the **Analyze images** button. If using Windows, you may need to allow network access for CellProfiler.exe. Once started, the analysis of all images will take a few minutes to complete.

3. Exercise: Using the CellProfiler Analyst software to visualize the data and classify cells

You're ready to use CellProfiler Analyst (CPA) to delve into the data collected from the cell images.



1. To launch CPA, double-click its desktop icon. Upon starting, CPA will prompt you to select a properties file. Choose the **DefaultDB.properties** file from the Default Output Folder, which the **ExportToDatabase** module from CellProfiler created.
- This properties file is crucial as it contains the configurations for CPA to interface with the database created by CellProfiler, which includes measurements from the images and their locations on your hard drive.
- Remember, if you move the database file, you must update the properties file to reflect the new location of the database.

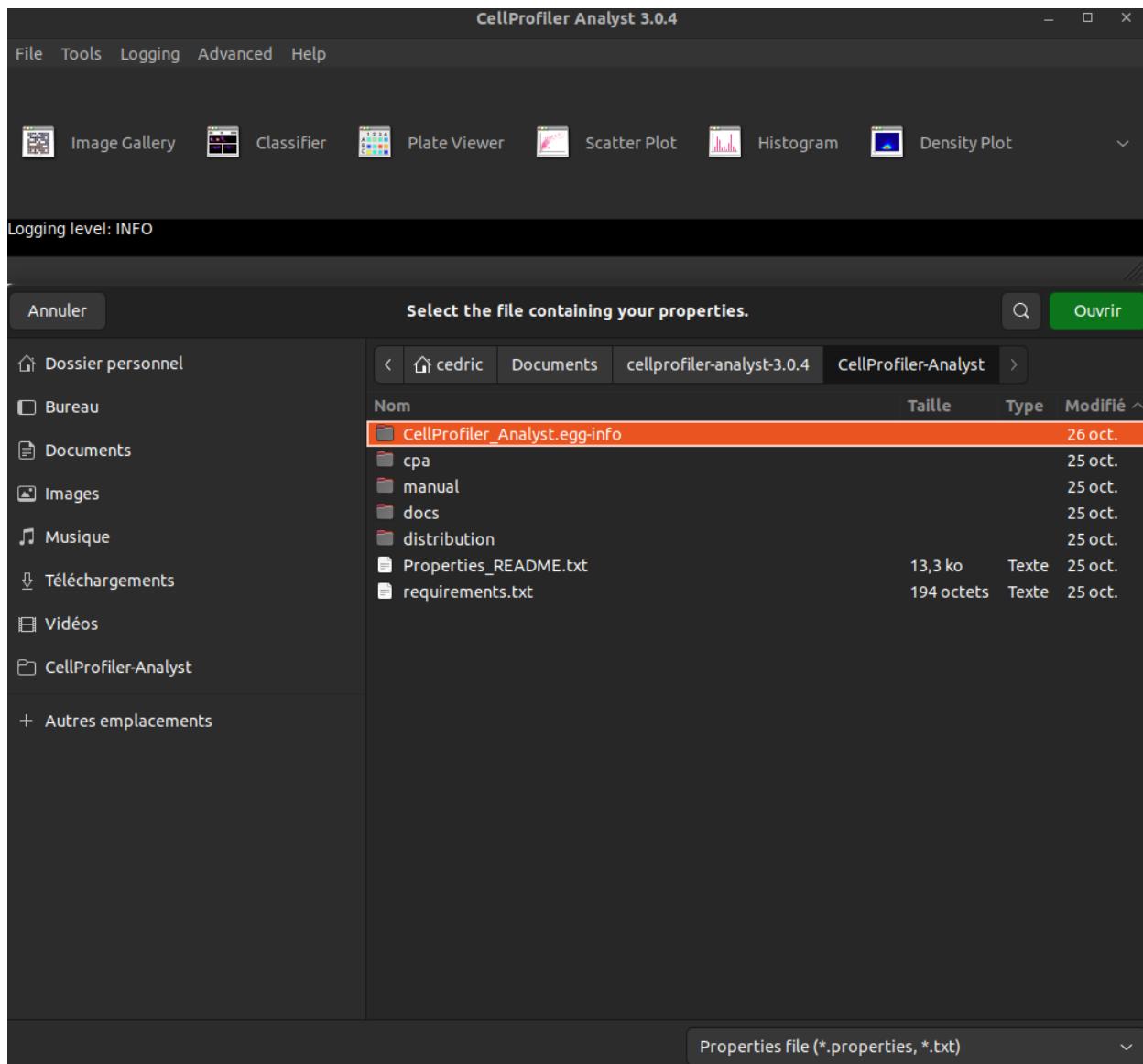


Figure 6: Select properties file

2. Visualizing the measurements in a 384-well plate layout view.

CPA offers a variety of data display tools, including a plate layout view for multi-well plate data, like the 384-well plate used in this assay.

- To access this view, click the **Plate Viewer** icon  in CPA's main window. This displays a 384-well plate layout where colored squares represent wells with data, and crossed-out wells lack data. In this case, 3 wells have data. Hover over the wells to see tooltips with per-well values.
- Change **Number of plates** to 3, since we have used three plates in this experiment and want to visualize them at the same time.
- Initially, the color coding shows the image index, which isn't relevant for our analysis. Change the measurement view to “Image_Count_Cells” to visualize a specific measurement: The wells have the following treatments:
 - K16: **negative controls (no drug) or mock treated**.
 - G18: **Treated with berberine chloride**.
 - P20: **Treated with rapamycin**.

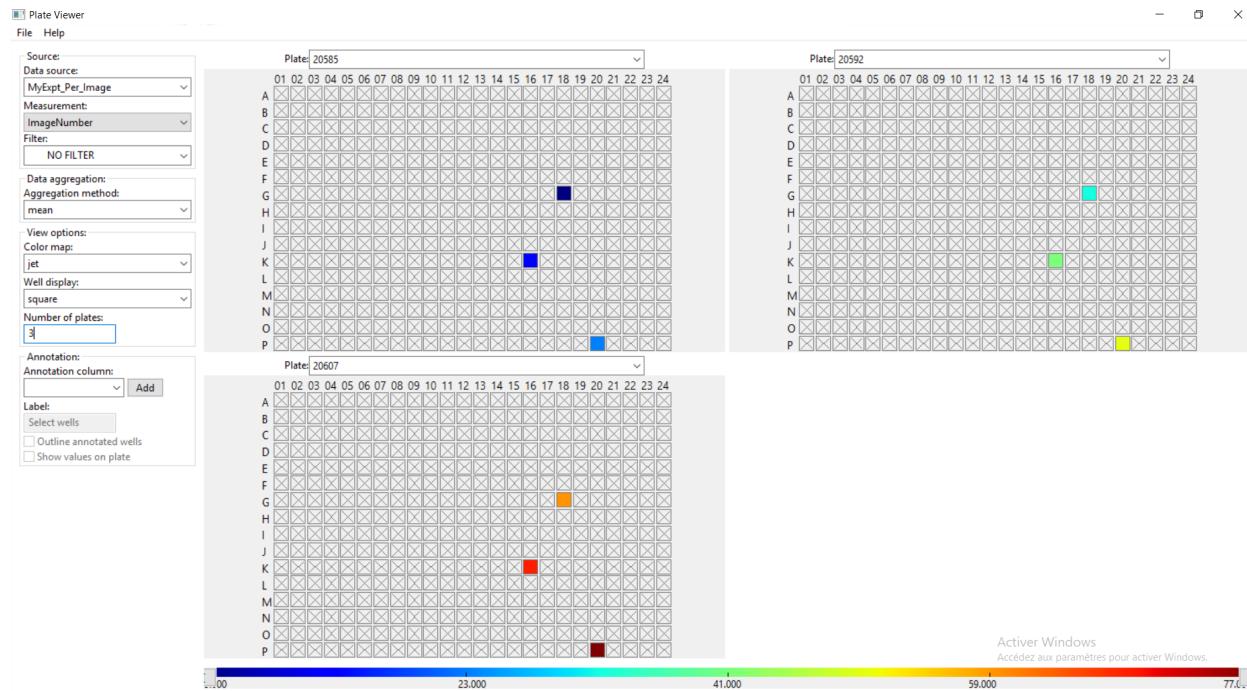


Figure 7: Plate Viewer

- For *per-object measurements*, set the Data Source to “Per-object” and choose a measurement like “Nuclei_Intensity_IntegratedIntensity_Syto”.
 - Do we have homogeneous values or heterogeneous values for this measure?
 - If heterogeneous what are the wells that have the lowest values?
- The Plate Viewer will show an **aggregate statistic** for these per-object measurements in each well, which you can adjust by selecting a different “Aggregation method.”
- To view **image thumbnails** in the Plate Viewer, select “thumbnail” under *Well display*. This replaces the colored squares with thumbnails of the original images.

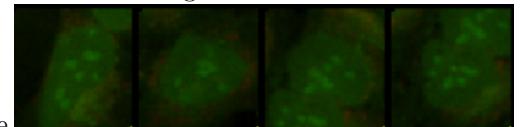
Keep the Plate Viewer open as you'll need it for further steps in the exercise.

3. Classifying the cells' phenotypes with CPA's Classifier

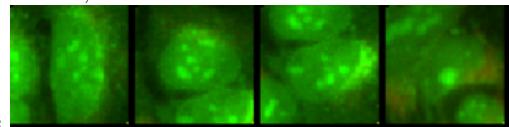
CellProfiler Analyst's Classifier is a machine-learning tool used to automatically distinguish different cell phenotypes. In this exercise, you'll train the Classifier to recognize phenotypic changes induced by the two drug treatments.



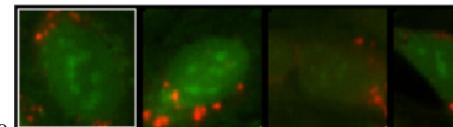
- Click the **Classifier** icon in CPA to open the interface.
- we want to explore three phenotypes: **negative**, **reduced_nucleolar_size**, and **redistribution_of_mitochondria**. Right click in the **positive** window and select *Rename class*. Write **reduced_nucleolar_size** (no spaces) as the new class name. Then press the **Add new class** button in the bottom right corner and write **redistribution_of_mitochondria** as the new class name.
- Press the **Fetch!** button to load a default set of cell images (usually 20) for classification. Thumbnails of cells will populate the “unclassified” panel.
- In the top menu bar, it is possible to select which color channel (or None) to use for the display of different images. For this exercise, the most important images are **Syto** and **Mito**. Select your preferred colors for those two channels, e.g., Green and Red, and select *None* for the other images.



- Drag and drop cells into the “reduced_nucleolar_size” bin, as shown here



- Drag cells into the “negative” bin, as shown here



- Similarly, drag cells into the “redistribution_of_mitochondria” bin, as shown here

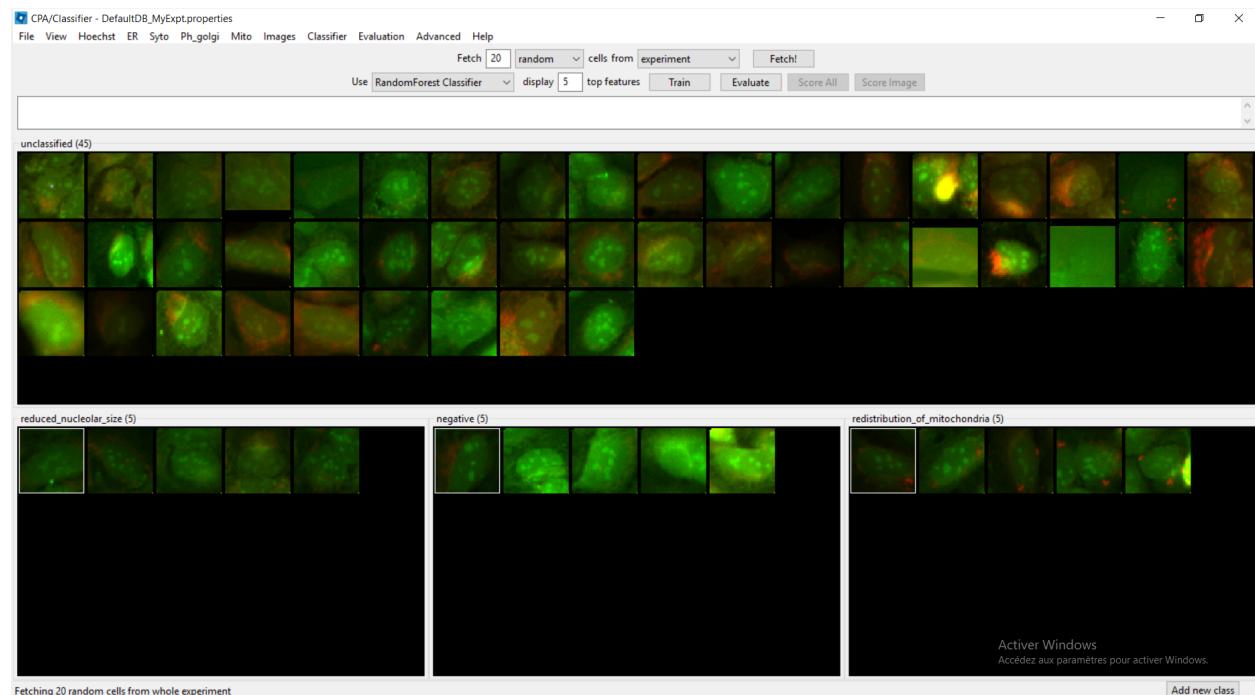


Figure 8: The Classifier interface

- Once you have at least 5 cells in each bin, click **Train Classifier**.
 - If the initial batch doesn't have 5 clearly definable cells in each class, use "Fetch!" again until you do.
 - The training set you create by sorting these cells will vary since the images are randomly sampled. Consequently, the training and resulting classifications will differ among users. This set of sorted cells is known as the **training set**.
- 4. Reviewing the rules that CPA established (based on your training set) to classify positive and negative cells**

In this exercise, you'll evaluate the criteria that CellProfiler Analyst (CPA) uses to distinguish different phenotypes, derived from the training set you created.

- By default, it displays the top 5 features.
 - Write down the top 5 features.
- If you do not have interesting features, you can add more images to each bin and start the training again.

5. Reviewing the accuracy of the classification with the confusion matrix

you'll assess the precision of your classifier by using a **confusion matrix**. This matrix helps verify how well the classifier's rules, which were developed during training, can predict the actual classification of each cell in your training set. CPA accomplishes this by applying the rules to each cell's data to predict if it's positive or negative and then checks these predictions against your initial categorization.

- To visualize the classifier's performance, click the **Evaluate** button. This action will produce a confusion matrix that plots the true classifications you assigned against the predicted ones made by CPA.
 - How accurate is your classification after adding only a few cells to your training set?

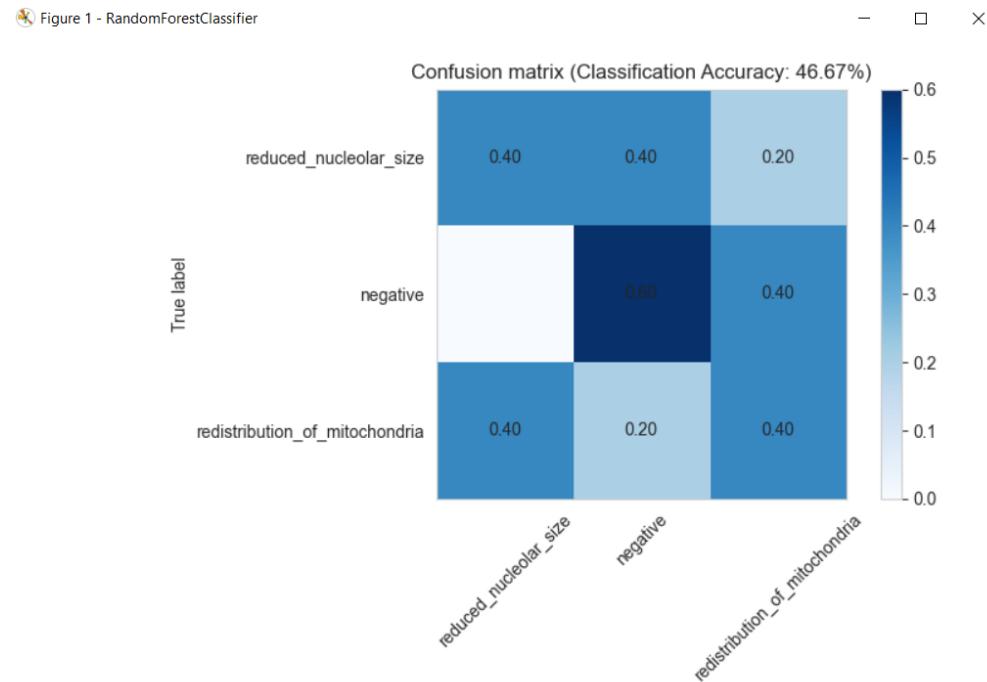


Figure 9: The Confusion matrix

- Remember, the confusion matrix reflects the classifier's accuracy on the training set, not the entire dataset. Even if the matrix shows a high level of accuracy, it's essential to test the classifier on a broader range of data to ensure it's reliable for scoring your entire experiment.

6. Refining the training set by obtaining samples from specific wells

if you know which images are likely to contain the phenotype, you can manually select cells from those images.

- Launch the **Plate Viewer** and navigate to well *K16*, which contains a negative control. Open an image from this well by double-clicking on it.
- In the image, identify a cell that lacks the phenotype (negative) and drag it into the negative bin. Do this for *five such cells*.
- Follow the same process for *P20* (a well containing cells treated with rapamycin), where you'll find "reduced_nucleolar_size" phenotypes. Do the same for *G18* (a well containing cells treated with berberine chloride), dropping the cells into the "redistribution_of_mitochondria" bin.
- After you've selected five cells for each category, click the **Train classifier** button to update your classifier with these new examples.

7. Refining the training set by correcting misclassified cells in an image

- Open an image by double-clicking on a **thumbnail** of a cell in the bins.
- Once the image is open, go to the menu and select **Classify** followed by **Classify Image** to initiate the classification process.
- The cells in the image will be **color-coded** based on their classification according to the current rules.
 - For Windows users, to understand what each color represents, click on the **Show controls >>** button at the bottom of the window to display the color legend.
 - For Mac users, go to the image menu, select **View** then choose **View cell classes as numbers** for a numerical representation. To interpret what each number signifies, click the **Show controls >>** button to bring up the numbered class list.
- Identify and select up to **five cells** that have been *incorrectly classified*. Click and drag each one into the correct bin.
- After reclassifying any mislabeled cells, click the **Train classifier** button to update the classifier with these corrections.

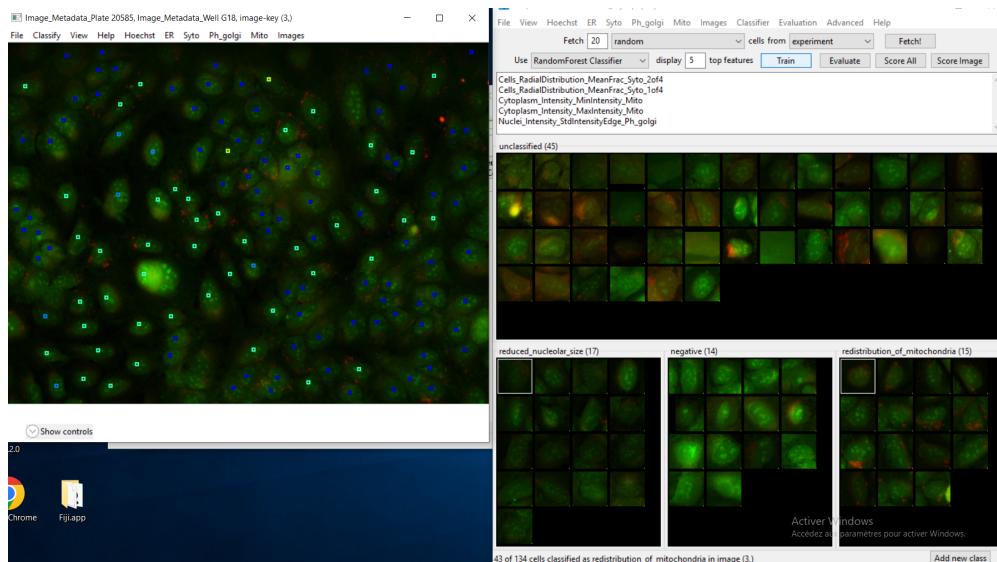


Figure 10: Classification of objects into image

8. Refining the training set by fetching positive and negative cells

- Adjust the number of cells to fetch by changing the value next to **Fetch** from **20** to **5**.
- In the fetch controls, click the **random** drop-down box and select *positive*.
- Press the **Fetch!** button to collect samples that the computer identifies as positive cells, based on the existing rules.
- Review the fetched cells:
 - For correctly identified positive cells (true positives), move them to the positive bin.
 - For incorrectly identified negative cells (false positives), move them to the negative bin.
- If uncertain about a cell's classification, select it and press **Delete** to exclude it from your training set.
- Continue this process until you have a minimum of **20 cells** in each bin.
- Click the **Train classifier** button to update the classifier.
 - Has the accuracy of the model increased ? If not, go back to the first step and repeat, until the classifier displays the desired level of accuracy.

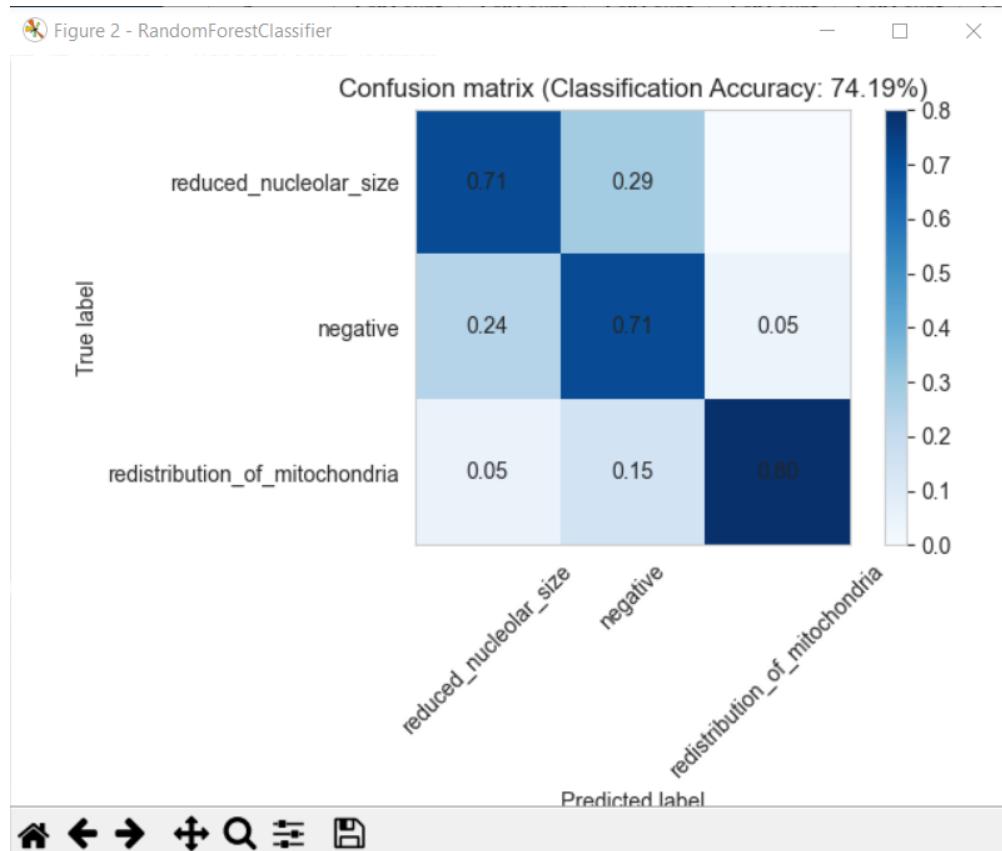


Figure 11: The Confusion matrix after refining the training

9. Applying the classifier to the entire dataset:

- Once satisfied with the classifier's accuracy, initiate the **scoring** of the entire dataset by clicking the **Score all** button.
- In the dialog box that appears, confirm the default settings by clicking **OK** to start the scoring process. The classifier will evaluate each cell in every image as either positive or negative.
- A summary window, known as the **Hit table**, will appear, showing the scores for each image. It includes the total count of cells and the number of cells classified in the different classes, along with the **enrichment score** for each.
- To organize the data, click on the **Enriched Score positive** column header. Adjust the window size if necessary to view this column.
- Sort the scores in *descending order* to have the highest enrichment scores at the top of the list.
- To view the image with the highest enrichment score, double-click the asterisk (*) next to the "ImageNumber" in the first row.

This process allows you to view the images corresponding to the most enriched scores, which can provide insights into the accuracy and effectiveness of your classifier across the dataset.

	File	View	Help	Hit table (grouped by Image) (DefaultDB_MyExpt.properties)														
ImageNumber	je_Metadata_1	je_Metadata_2	Total Cell Count	iced_nucleolar Cell Count	Negative Cell Count	ution_of_mite Cell Count	p(Enriched)	p(Enriched)	p(Enriched)	enriched Score	enriched Score	enriched Score	Training Total	Reduced_nucleolar Cell Count	aining Cell Count	Negative Cell Count	distribution_of_ Cell Count	
* 17	20585	K16	116	28	87	1	0.12936611742	0.08250739781	0.03638778634	0.8280150204	1.74948140957	-1.42294665994	0	4	0	0		
* 72	20607	K16	115	59	55	1	0.63900422670	0.73881092248	0.32682025442	0.24800161379	0.45157829385	-0.1716205562	2	1	1	0		
* 73	20607	P20	70	52	17	1	0.93433029424	0.27160296617	0.07052634651	1.15113536006	-1.1198856972	0	0	0	0	0		
* 53	20592	P20	68	39	28	1	0.7352811757	0.06987032031	0.0311638917	0.44355024445	0.19402847467	-1.1030104672	0	0	0	0		
* 18	20585	K16	66	8	57	1	0.03436957231	0.99598624617	0.07586471926	-1.4486368142	2.38206744687	-0.1056965995	5	0	0	0		
* 64	20607	K16	108	58	47	3	0.68323969485	0.66111992743	0.08776228642	0.33384737873	0.29038803271	-1.018000936	0	0	0	0		
* 11	20585	K16	97	27	67	3	0.1872505759	0.05798616385	0.10072214886	-0.675333090	1.3796698928	-0.9507689943	0	0	0	0		
* 12	20585	K16	97	29	65	3	0.2204885007	0.04668633904	0.10072214886	0.5484365632	1.2508904571	-0.9507689943	1	0	0	0		
* 74	20607	P20	59	40	17	2	0.87012526336	0.36489137532	0.13317687442	0.82605709833	-0.2406844120	-0.8135016710	0	0	0	0		
* 67	20607	K16	132	86	40	6	0.86405596051	0.38404953797	0.13376703649	0.80318169988	-0.205158383	-0.8112856005	0	0	0	0		
* 25	20585	P20	107	62	40	5	0.75655248295	0.53831384838	0.14340204344	0.4924337147	0.06668807144	-0.7730480137	0	0	0	0		
* 16	20585	K16	121	59	56	6	0.58929714222	0.7132819414	0.14926304463	0.15800661020	0.3958430026	0	0	0	0	0		
* 14	20585	K16	104	22	77	5	0.05998879990	0.078622264210	0.14945411697	0.0739532992	1.6606612317	-0.75188638	0	0	0	0		
* 13	20585	K16	119	36	77	6	0.22162127675	0.3611514623	0.15239239578	-0.5455795036	1.16619359489	-0.7452315475	0	0	0	0		
* 15	20585	K16	68	6	59	3	0.01824061628	0.09635794956	0.15180767970	0.0277655359	2.43706943484	-0.7359258952	1	0	0	0		
* 70	20607	K16	147	110	29	8	0.95339301670	0.16782759187	0.15931347415	1.31277972420	-0.6953499490	-0.7223815783	0	0	0	0		
* 50	20592	P20	49	26	21	2	0.65436204565	0.63385580444	0.16409843306	0.27719404349	0.23833831797	-0.7070507052	0	0	0	0		
* 24	20585	P20	77	36	37	4	0.54611599699	0.73345573292	0.17301932978	0.08034002176	0.43960456150	-0.6794007331	0	0	0	0		
* 79	20607	P20	46	21	23	2	0.52247357364	0.745761205	0.17566830283	0.0390669154	0.4673463192	-0.674861650	0	0	0	0		
* 66	20607	K16	93	37	50	6	0.40629563716	0.82221398340	0.20505563444	-0.1447280814	0.6650872591	-0.5884649750	1	0	0	0		
* 21	20585	P20	52	27	22	3	0.63636684740	0.62557565313	0.20927773316	0.2430407508	0.22291575711	-0.5773009462	0	0	0	0		
* 80	20607	P20	69	43	21	5	0.0895499626	0.39469924023	0.2407492809	0.6246261354	-0.1857049286	-0.498820230	0	0	0	0		
* 26	20585	P20	127	91	26	10	0.92871194019	0.18415695782	0.24261163172	1.11486423416	-0.6464183396	-0.4944070123	0	0	0	0		
* 52	20592	P20	40	13	24	3	0.29446348180	0.85995657095	0.27307797505	-0.3794881877	0.78821378312	-0.4252011539	0	0	0	0		
* 23	20585	P20	88	47	33	8	0.67421345451	0.5399051268	0.2908886752	0.31586427039	0.06947248042	-0.3869661484	1	0	0	0		
* 78	20607	P20	27	15	10	2	0.67109314962	0.52394545145	0.29355923966	0.3097097491	0.04163005187	-0.381379951	0	0	0	0		
* 81	20607	P20	76	41	28	7	0.6807903645	0.52584799360	0.29854508068	0.3289373205	0.04494262855	-0.3709882510	0	0	0	0		
* 71	20607	K16	128	88	27	13	0.90289223178	0.19524235211	0.31770885094	0.96838194344	-0.6150910815	-0.3319424236	2	0	0	0		
* 75	20607	P20	33	16	14	3	0.5680496602	0.61602932280	0.32592873880	0.1189524971	0.2053032480	-0.315580663	0	0	0	0		
* 65	20607	K16	75	25	42	8	0.28700558677	0.84364420555	0.34384297646	-0.395195762	0.73204534742	-0.2804767242	0	0	0	0		
* 61	20607	G18	60	23	30	7	0.3842949121	0.75593440464	0.247112512	0.49097754829	-0.2156587665	0	0	0	0			
* 10	20585	K16	112	46	52	14	0.43091404629	0.7145227524	0.39440363836	-0.207872155	0.39844456340	-0.1864233446	0	0	0	0		
* 40	20592	K16	127	91	20	16	0.92871194019	0.10803414228	0.39630655241	1.11486423416	-0.9167872026	-0.1827852092	1	0	0	0		
* 20	20585	P20	63	22	33	8	0.23025227619	0.79156557106	0.40733427701	-0.3268553686	0.57951743966	-0.1628588530	0	0	0	0		
* 22	20585	P20	37	13	19	5	0.34077119073	0.75455273767	0.43721575197	-0.2865719411	0.48773140673	-0.1096429560	1	0	1	0		
* 54	20592	P20	64	29	26	9	0.51730755436	0.59855298893	0.4458203115	0.3007831859	0.17347437506	-0.0944877303	0	0	0	0		
* 46	20592	P20	50	27	16	7	0.67073016847	0.43059116605	0.44627930580	0.3089958927	-0.121359049	-0.0936840056	0	0	0	0		
* 39	20592	K16	107	54	37	16	0.62191348821	0.47862323691	0.46871785602	0.21613879179	-0.031758918	-0.0544137213	1	0	0	0		

Figure 12: Hit table

- You can also save your training set and/or classifier model for future reference or to make changes later; do so by going to **File > Save Training Set** or **File > Save Classifier Model**.

10. Storing and Visualizing Classification Scores

After scoring your experiment, save the results to the measurement database for future visualization with CellProfiler Analyst's (CPA) tools.

- In the “Hit table” window, go to the menu and select **File**, then **Save table to database**. When asked for a table name, type in **HitTable** and choose to *save it permanently*.
- Access the Plate Viewer in CPA and set the “Data source” to “*OTHER TABLE*.”
- When prompted, select the “HitTable” from the list of tables.
- For the table type, choose “per-well” and then link the hit table to the image measurements table by matching the columns:
 - Match “Image_Metadata_Plate” with itself.
 - Match “Image_Metadata_Well” with itself.
- Open a new Plate Viewer and from the Measurement dropdown, select “pEnriched_reduced_nucleolar_size” to display the enrichment scores in the context of the plate layout.
 - What does this correspondence (or lack thereof) tell you about the classifier?
- A strong match may indicate a successful classification, while discrepancies could suggest areas for improvement or further investigation.

