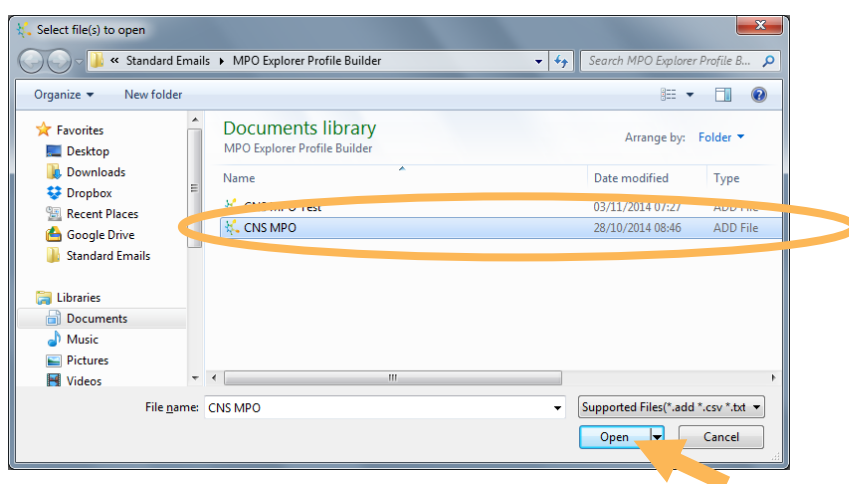# Worked Example:

Automatically Building a Scoring Profile with Rule Induction

In this example we will use the Profile Builder in StarDrop's MPO Explorer module to derive a multi-parameter scoring profile, based on a data set initially described by Wager *et al.* [ACS Chem. Neurosci. **1** p. 435 (2010)]. The authors used this data set to develop a multi-parameter optimisation method for selection of compounds intended for CNS indications. The 'CNS MPO score' derived by Wager *et al*. is calculated as the sum of the values of desirability functions for six physicochemical parameters, calculated logP (clogP), calculated logD at pH 7.4 (clogD), molecular weight (MW), topological polar surface area (TPSA), number of hydrogen bond donors (HBD) and the pKa of the most basic center (pKa), resulting in a value between 0 and 6. The authors compared the CNS MPO score for a set of 119 marketed drugs for CNS targets with 108 Pfizer CNS candidates and found that 74% of the marketed drugs achieved a desirability score of ≥ 4 compared with only 60% of the Pfizer candidates.

The scoring profile derived by MPO Explorer will contain one or more rules corresponding to combinations of properties that significantly increase the chances of identifying a drug and we will compare these with the results of the CNS MPO score.

Exercise

- Start StarDrop
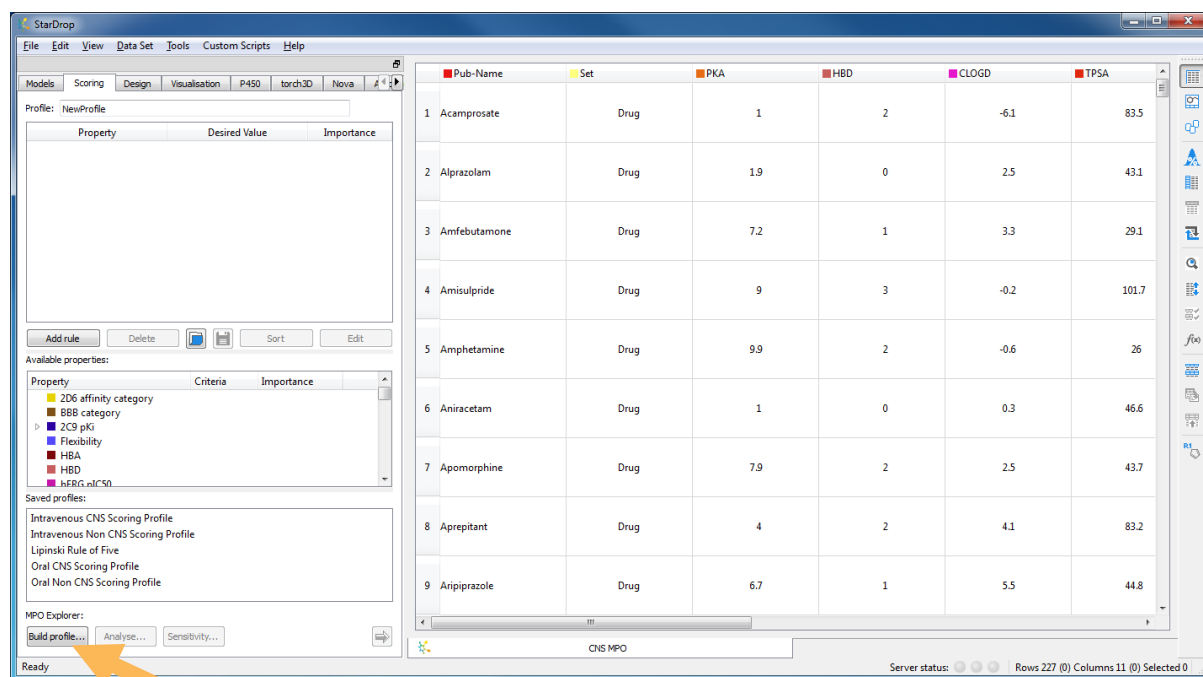- From the **File->Open** menu item, open the file **CNS MPO.add**.

The data set contains 227 compounds, 119 drugs for CNS targets and 108 Pfizer development candidates that did not reach the market (published by Wager *et al.*).
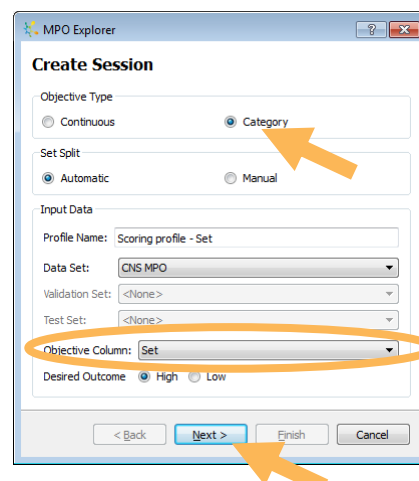
For each compound, 6 properties have been previously calculated, labelled MW, CLOGP, TPSA, CLOGD, HBD and PKA, as described in the introduction. For comparison, the CNS MPO score, as defined in Wager *et al.* is also included in the data set.

We will use the data in the data set to find a scoring profile, based on these simple properties, with which to distinguish marketed CNS drugs from unsuccessful candidates.

- Change to the **Scoring** tab in StarDrop and click the **Build Profile** button, under MPO Explorer, as shown above. The **MPO Explorer** wizard, shown right, will be displayed enabling us to control the profile building process.

- We would like to build a profile based on the categorical objective in the **Set** column. Therefore, choose **Category** under **Objective Type** and confirm that the **Objective Column** is defined as **Set,** as shown to the right.
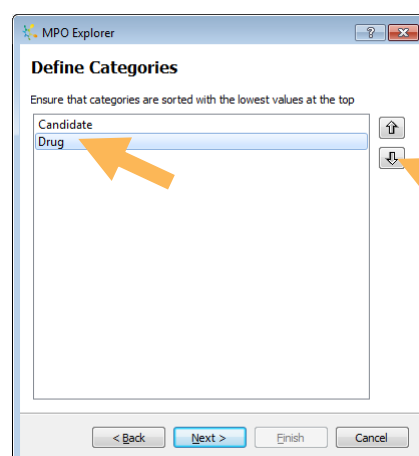
  Note that the **Desired Outcome** is **High,** i.e. we would like to maximise the value of the objective.

- Click **Next** to move onto the next page, **Define Categories** (shown right).

- Here we can define the order of the categories, from the lowest at the top, to the highest at the bottom. Remembering that we have chosen to maximise the objective, we need to order the categories, such that the **Drug** is at the bottom. Therefore, select the **Drug** category and click the down arrow.

- Click **Next** to move onto the **Set Selection** page.
- Here we can define the parameters for splitting the data set into training, validation and test sets. In this case, we will use the defaults, putting 70% of the compounds into the training set and 30% into the validation set. Ideally, we would like an external test set, but given the small size of the data set, it is not practical in this case. Click **Next**.

- On the **Select Properties** page (shown right) we can choose the properties that we would like to explore to identify a scoring profile. In this case, we would like to use the 6 simple compound properties, as described above. However, we should un-tick the **CNS MPO score** property, as shown right, to avoid using this.

- Click **Next** to move onto the last page of the wizard.

- The **Profile Parameters** page enables us to specify some additional conditions. In this case, we will use the default minimum profile coverage of 20% (i.e. we will only consider rules that are applicable to ≥20% of the compounds in the data set).
- Set the **Minimum desirable category** to **Drug.** This is used to calculate the performance statistics for the rule's ability to distinguish desirable and undesirable outcomes. This parameter is most useful if the objective is a continuous property or has more than two categories.
- Finally, click **Finish** to begin the profile building session.

Once the first rule has been found, this will be displayed within the MPO Explorer profile analysis window, as shown below.
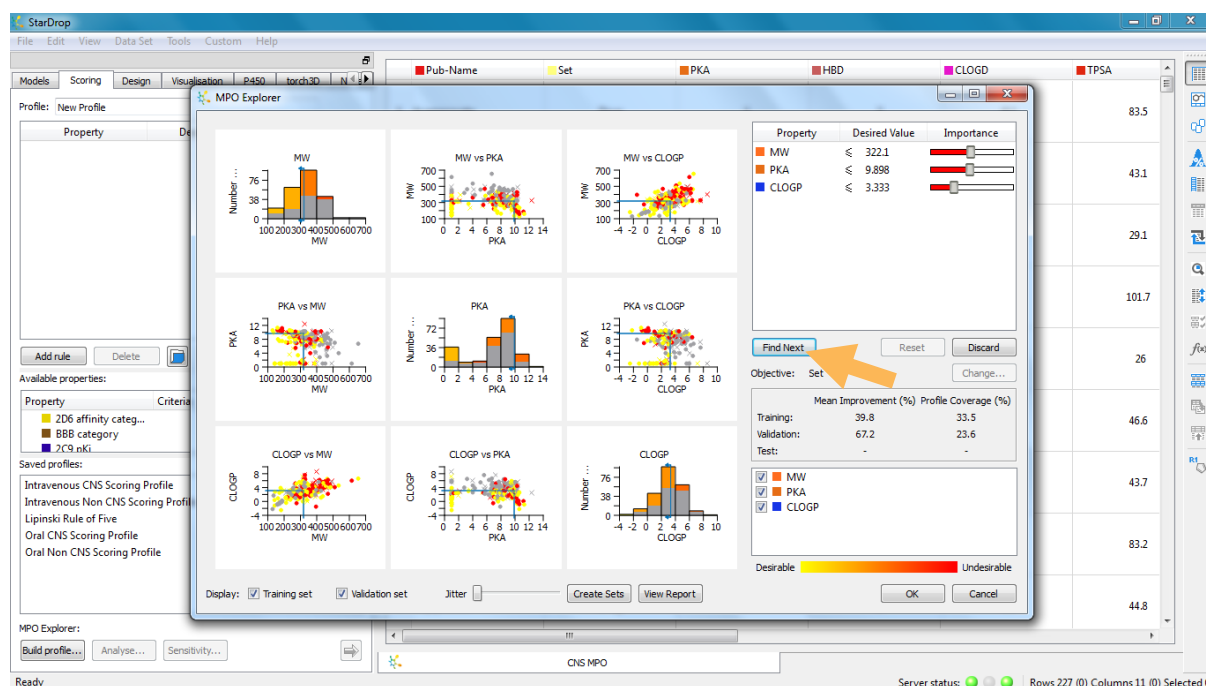
The rule is shown in the top-right. In this case, the rule suggests that compounds with a MW < 322 and a most basic PKA < 9.9 and a CLOGP < 3.33 will have an increased chance of success. The criteria for MW and PKA are slightly more important than CLOGP. (**Note:** the exact values you see will depend upon the precision you have specified for displaying numbers in the preferences)

Below this, the statistics for the corresponding multi-parameter rule can be seen; in this case the mean improvement for compounds obeying the rule is >67%, i.e. compounds that meet these property criteria have a 67% greater chance of being a drug than a compound selected randomly from the set. If you hover your mouse over this statistic a tool-tip will display additional information; in this case the p-value for this rule is 0.0002 suggesting that it is highly statistically significant and the odds ratio is 10.4, which means that compounds meeting all three property criteria have a ~10× higher chance of success than compounds that do not. A detailed report on the statistics can be generated by clicking the **View Report** button.

On the left, plots show the rule in property space corresponding to the property criteria. The properties that are shown are controlled by the tick boxes in the bottom right of the analysis tool. In each plot, the blue lines indicate the boundaries implied by the rules. These boundaries can be dragged to modify the criteria and the statistics will be updated instantly. The compounds in the training set are represented by circles and those in the validation set by '×'. Desirable compounds (i.e. drugs) are shown in yellow and undesirable compounds (i.e. unsuccessful candidates) are shown in red. Grey points indicate compounds which have been filtered out by criteria other than those represented in the plot.

- In this case, we will accept the rule that has been generated automatically. To search for a second rule, click **Find Next**.

The second rule, shown above, depends only on CLOGD and PKA. In this case the mean improvement is only 21% and the corresponding p-value is 0.15, suggesting that this rule is not statistically significant. Therefore, we will reject this rule by clicking **Discard**.

- The final profile, which in this case contains just one rule, will be displayed and we can accept this by clicking **OK**.

The scoring profile will be displayed in the **Scoring** tab in StarDrop. As with any scoring profile, we can modify, rename or save the profile. Clicking **Analyse** under **MPO Explorer** will return to the MPO Explorer profile analysis tool. (**Note:** you can analyse any scoring profile, not only those built using MPO Explorer's Profile Builder).

- Run the new scoring profile on the full data set by clicking the ⇨ button on the Scoring tab. (**Note:** you can ignore the warning about data with zero uncertainty).
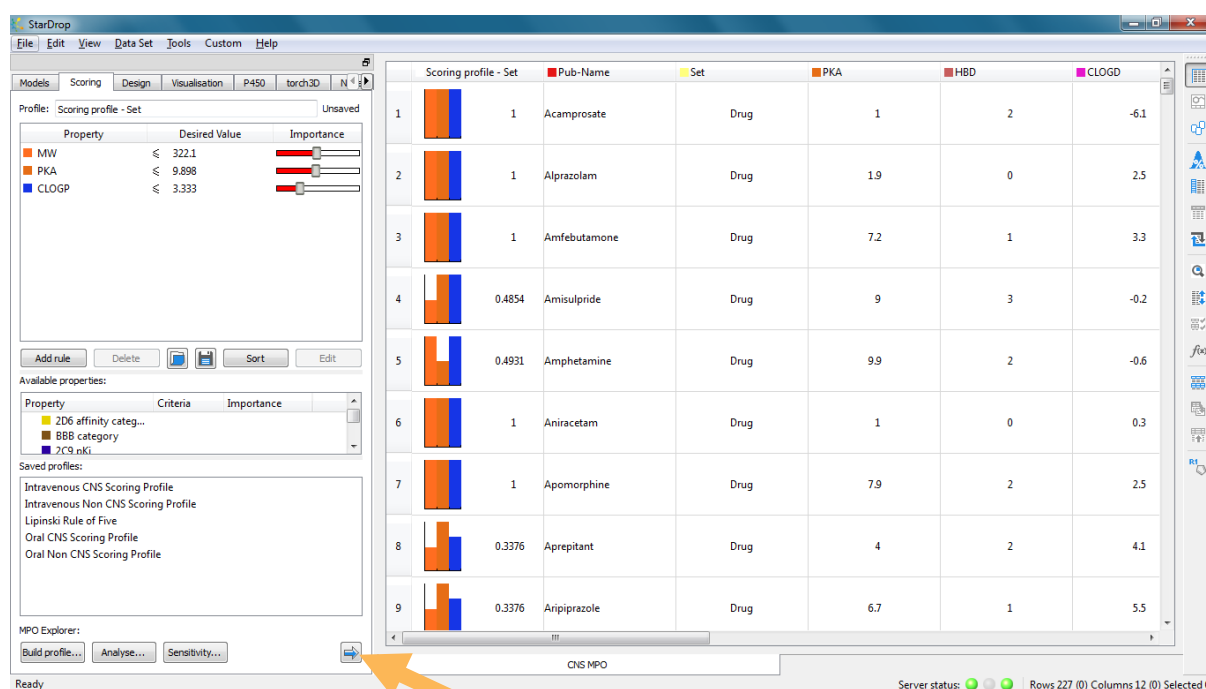
We can compare the performance of this scoring profile with the CNS MPO score by plotting a Receiver Operating Characteristic (ROC) plot.

- Select the **Visualisation** tab in StarDrop.

- Click on the [icon] button to generate a ROC plot. Choose **Set** as the property and **Drug** as the desired result and use the new score **Scoring profile – Set** as the classifier, as shown below.
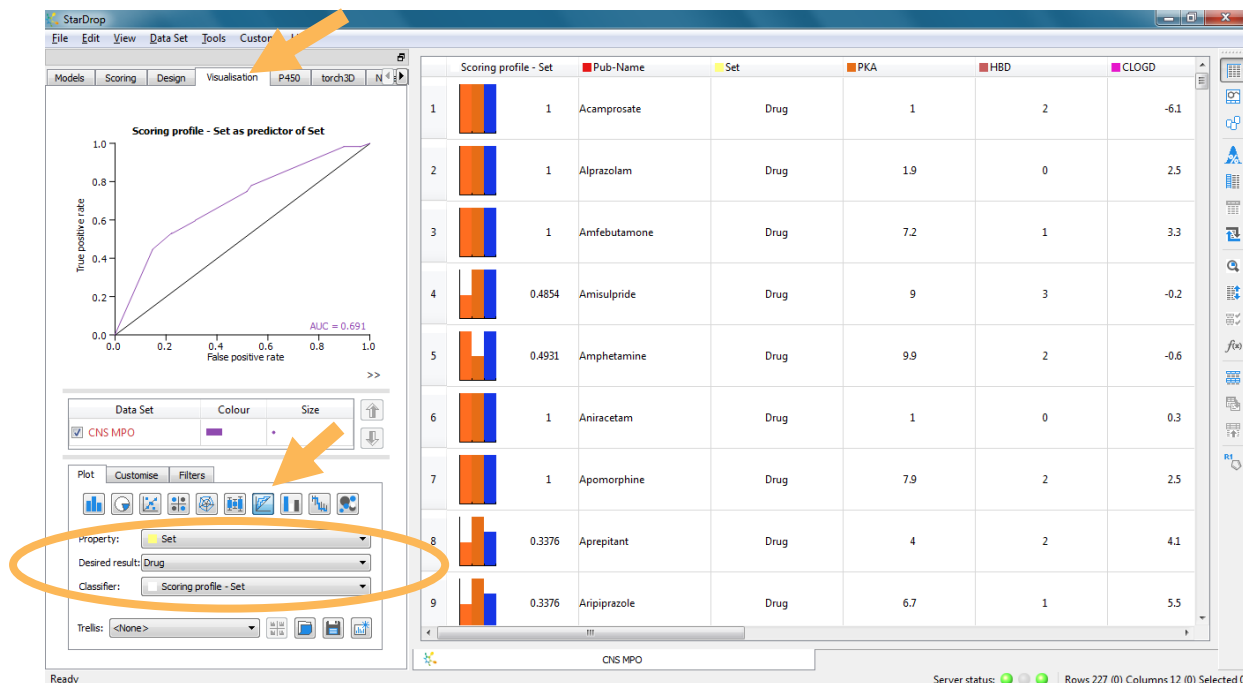
We would like to see a ROC curve above the black identify line (which corresponds to the performance of a random selection) and ideally as close to the top-left corner as possible. A higher area under the curve (AUC) corresponds to better predictive performance. For more details of ROC plots, see http://en.wikipedia.org/wiki/Receiver_operating_characteristic.

Here we can see that the AUC for the scoring profile we have generated is 0.69. We can compare this performance with that of the CNS MPO score.

- Detach the ROC plot for the scoring profile by clicking the button in the bottom right of the **Visualisation** tab.
- On the **Visualisation** tab, change the **Classifier** to **CNS MPO score** to create a ROC curve showing the performance of CNS MPO score, as shown below:
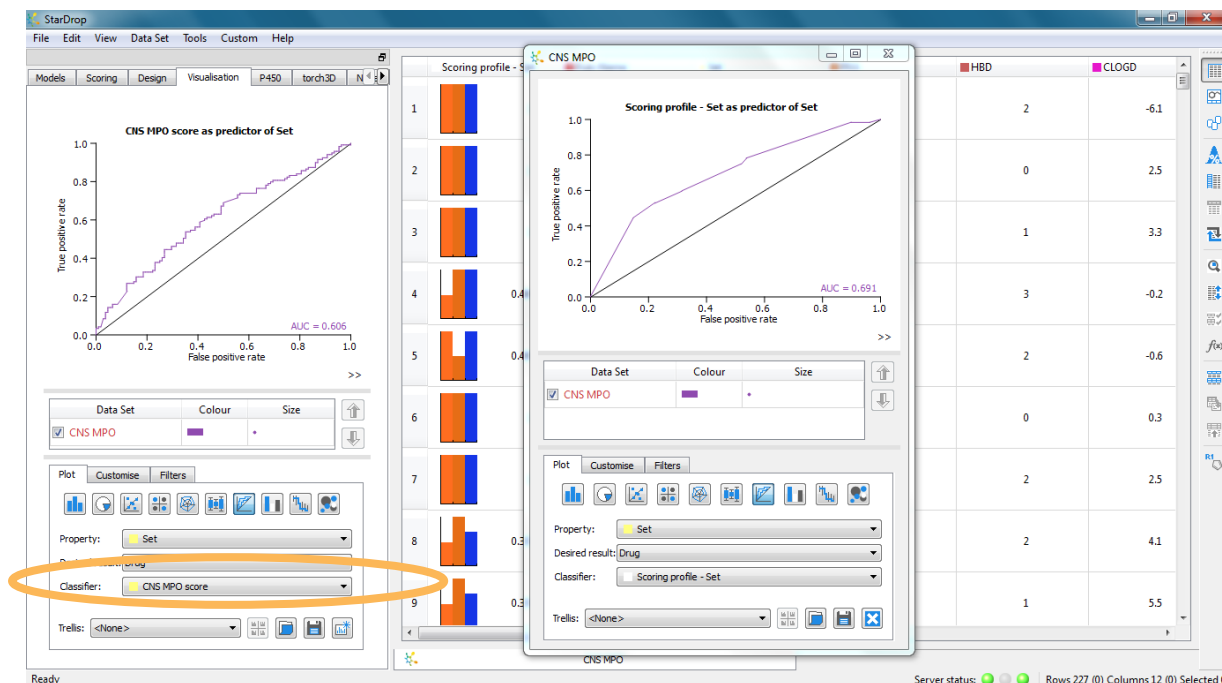

From this, we can see that the perfomance of CNS MPO score is not as good as the scoring profile we have generated because the ROC curve is closer to the identity line and the AUC is lower.

This is not a fair test of CNS MPO score and the new scoring profile that we have derived for two reasons:

- Compounds nominated as clinical candidates will generally have reasonable properties, so we would expect it to be quite challenging to distinguish candidates from successful drugs based on these simple properties. A more realistic test would be to distinguish drugs from early 'lead' compounds from drug discovery projects.
- We have assessed the performance of the scoring profile on the same set used to train and validate the corresponding rule. Therefore, the measure of performance may be artificially high.

To address these concerns, we can apply these scoring approaches to an independent test set.

- From the **File->Open** menu item, load the file **CNS MPO Test.add.**
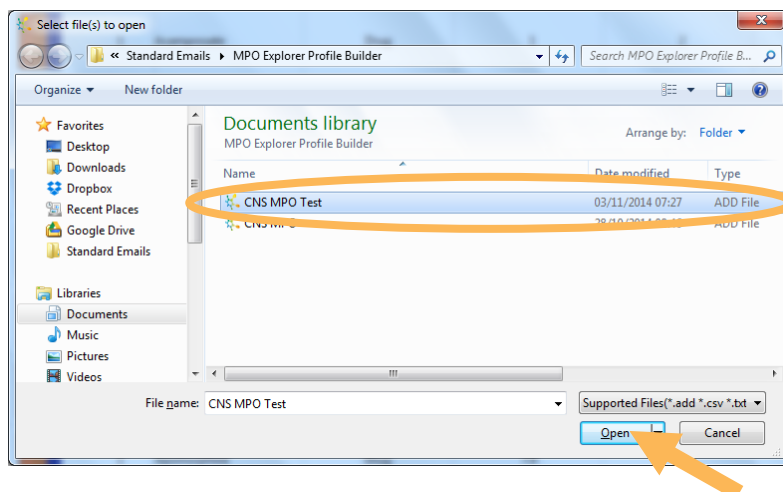
This data set contains 118 drugs (different from those used to find the rule) and 1000 compounds randomly selected from compounds in the ChEMBL database (https://www.ebi.ac.uk/chembl/) with a pK_i/pIC_{50} of greater than 6 (i.e. a K_i/IC_{50} less than 1 µM) against a CNS target. The target and measured pK_i/pIC_{50} of each compound from ChEMBL is included in the data set.

- Change to the Scoring tab and run the scoring profile by clicking the ⬛ button.

- Select the **Visualisation** tab in StarDrop and, as before, click on the ⬛ button to generate a ROC plot. Choose **Set** as the property, **Drug** as the desired result and use the new score **Scoring profile – Set** as the classifier. Check that the **CNS MPO Test** set is selected in the key below the plot.
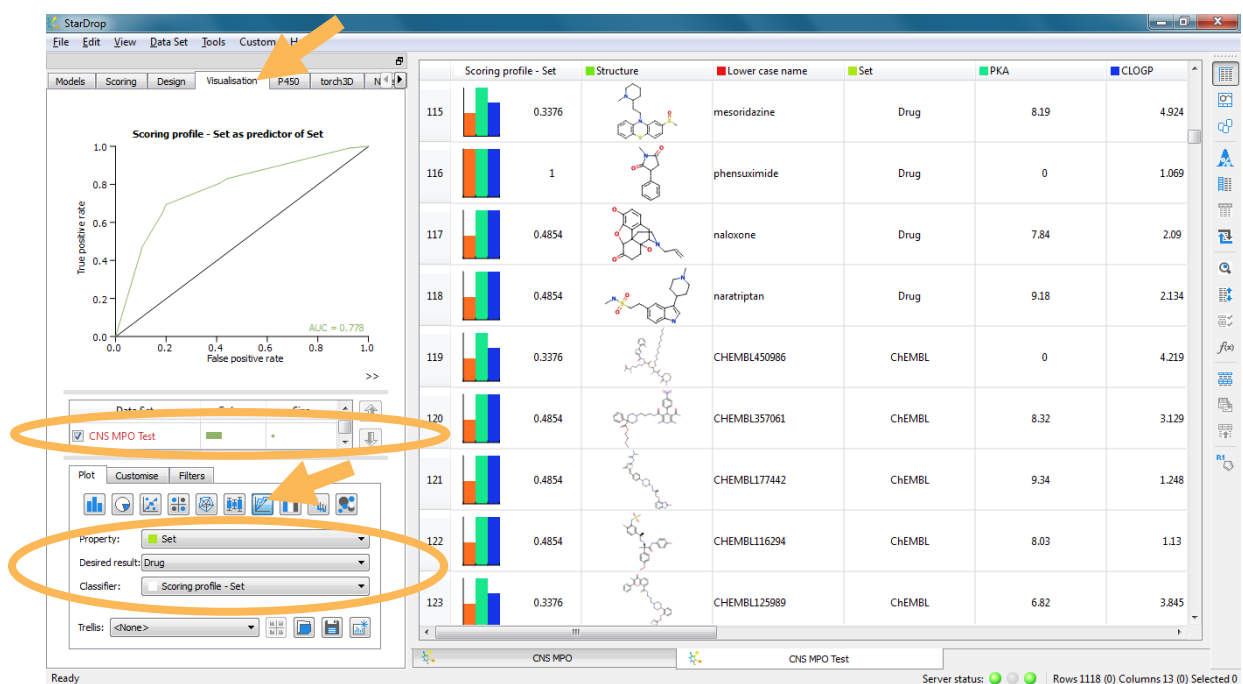
- Detach the ROC plot for the scoring profile by clicking the button in the bottom right of the **Visualisation** tab.
- On the **Visualisation** tab, change the **Classifier** to **CNS MPO score** to create a ROC curve showing the performance of CNS MPO score, as shown below:
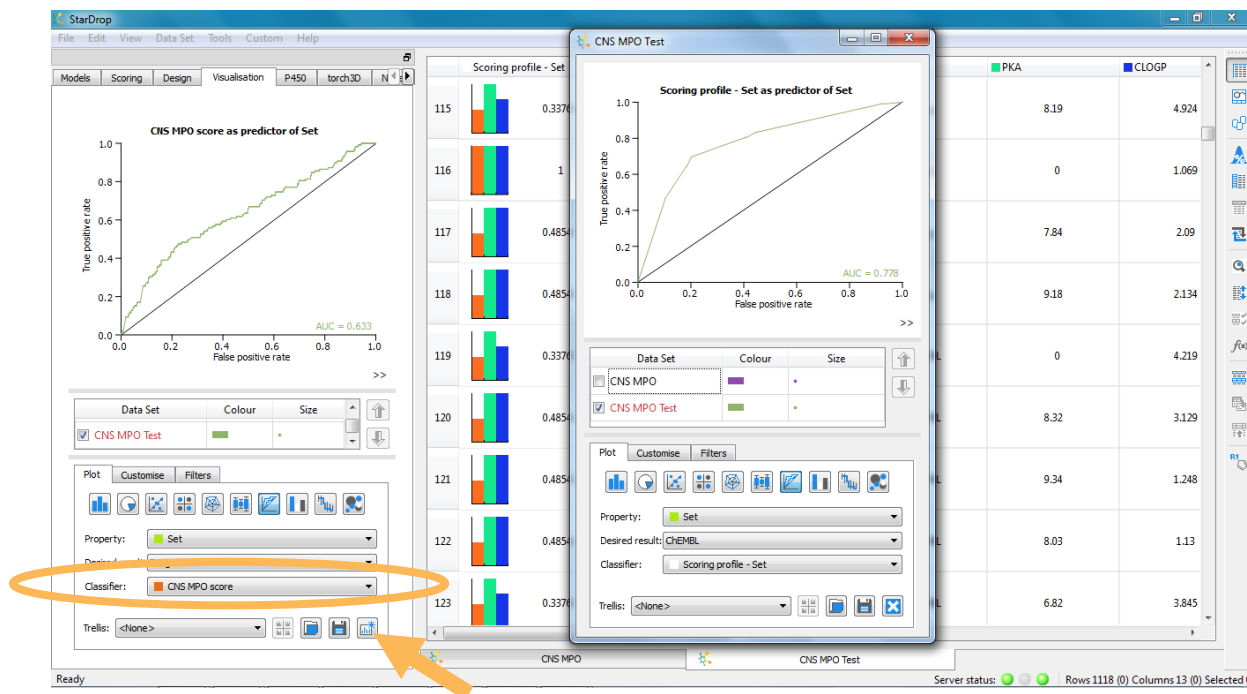
From this we can see that the performance of the scoring profile is better on this set, achieving an AUC of 0.78. However, AUC for the CNS MPO score has improved only marginally to 0.63, indicating that it cannot confidently distinguish between 'lead' compounds and drugs.

Conclusion

This example has shown how we can use the Profile Builder in MPO Explorer to generate scoring profiles with which to select compounds with a higher chance of success against our objective; in this case distinguishing CNS drugs from unsuccessful candidates.

In this example, we have only used the simple functionality of the Profile Builder. Other capabilities enable the automatic selection of properties from a large number of possibilities and the derivation of 'soft' criteria to take into account the sparseness of data, helping to avoid 'hard' cut offs that draw artificially harsh distinctions between compounds close to a property criterion.

It is notable that the simple scoring profile, using only three properties (logP, pK_a of the most basic site and MW) can outperform CNS MPO score, which uses six properties. This illustrates the fact that there is significant correlation between the properties used in CNS MPO score, for example logP and logD are strongly correlated (R^2=0.6 in this set). The inclusion of correlated properties can result in 'over counting' of the same factor, inappropriately biasing the selection of compounds. The Profile Builder will select only property criteria that contribute significantly to the selection of high quality compounds, avoiding the selection of multiple, highly correlated properties.

This example has used the simple 'drug like' properties included in the CNS MPO score in order to draw a direct comparison. However, the Profile Builder can be applied to any data, including predicted or experimental biological or physicochemical properties that are more directly related to the *in vivo* disposition or efficacy required in a successful compound.

Additional, practical examples can be found in Yusof and Segall, Drug Discov. Today **19**(5) pp. 680-687, a preprint of which can be downloaded from http://www.optibrium.com/community/publications/multi-parameter-optimisation/215-preprint-finding-the-rules-for-successful-drug-optimization.