Worked Example:

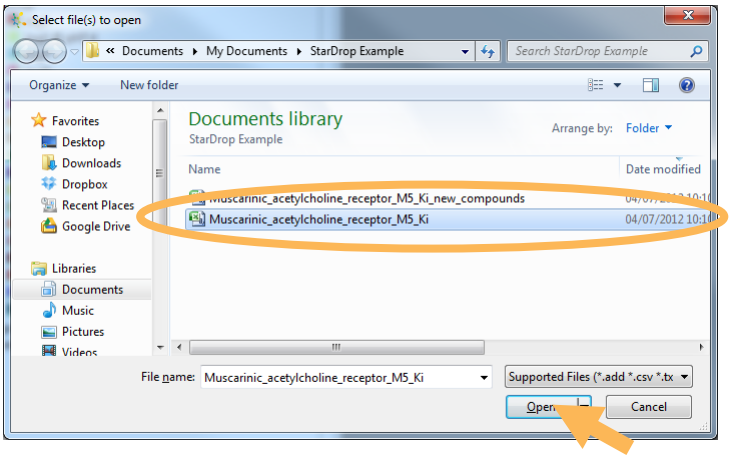
Automatic QSAR Model Building and Validation

In this example we will explore the application of StarDrop's Auto-Modeller to build a QSAR model of potency against the Muscarinic Acetylcholine M5 receptor, based on a set of public domain K_i data obtained from the ChEMBL database (https://www.ebi.ac.uk/chembl/). The resulting model will be applied to an additional set of compounds to predict their properties and visualise the structure activity relationship.

Step-by-step instructions for all the features you will need to use in StarDrop are provided, along with screenshots and examples of the output you are likely to generate. If you have any questions, please feel free to contact stardrop-support@optibrium.com.

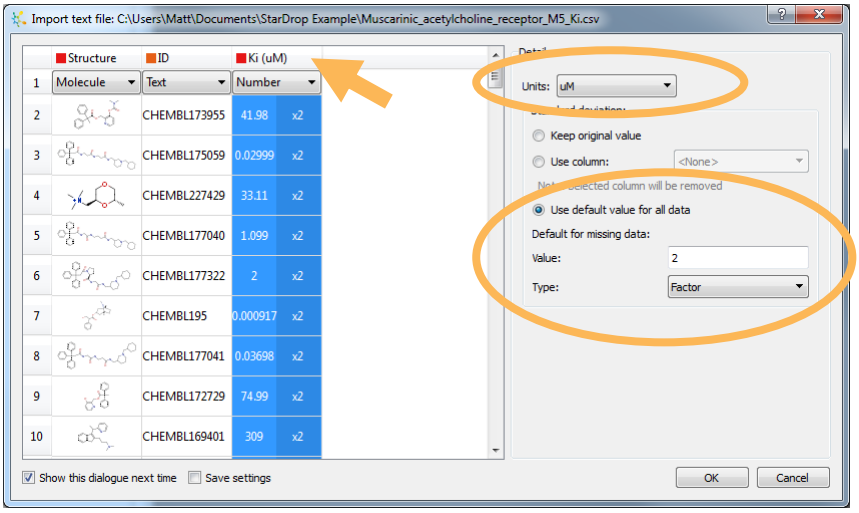
Exercise

- Start StarDrop from the Start menu
- Open the file **Muscarinic_acetylcholine_receptor_M5_Ki.csv** from the **File->Open** menu option. This is a comma-separated value file, such as can be saved from Excel.

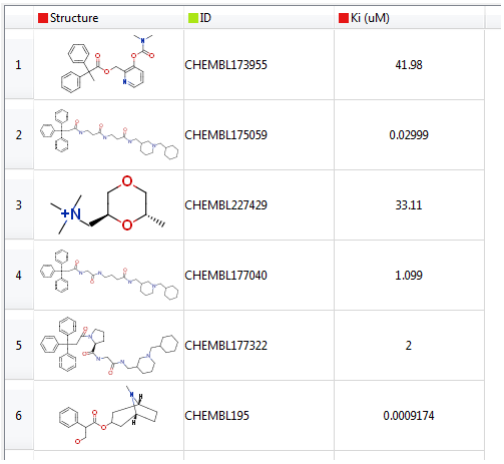


The file import dialogue box shows that the CSV file contains three columns: the first contains compound structures, the second, compound identifiers and the third, experimental data for the potency (K_i) against the Muscarinic Acetylcholine Receptor M5.

- Select the header of the column labelled **Ki (uM)** and on the right specify the units of this data to be **uM** from the **Units** drop-down menu. Also, assign an uncertainty of a factor of two to the data by choosing the option **Use default value for all data**, setting the **Value** to 2 and choose **Factor** from the **Type** drop-down menu

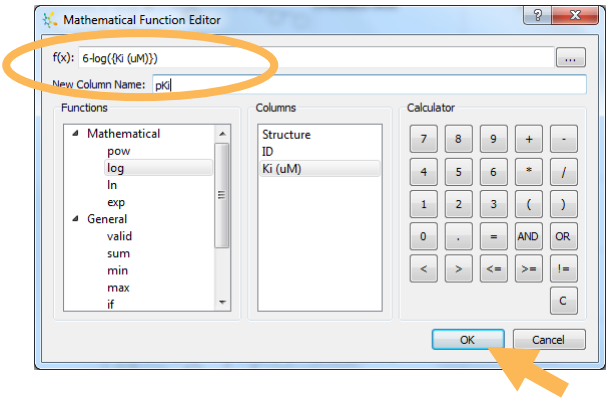


- Click **OK** to complete the import and a new data set will be created in StarDrop

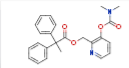
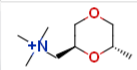
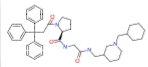
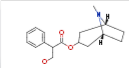


These data are K_i values in µM. However, to build a good model the data should be converted to logged units. Logged units provide a more even distribution of values to model and a better correlation with the compound descriptors used to build the model. Therefore, we will use the **Mathematical Function tool** in StarDrop to generate pK_i values.

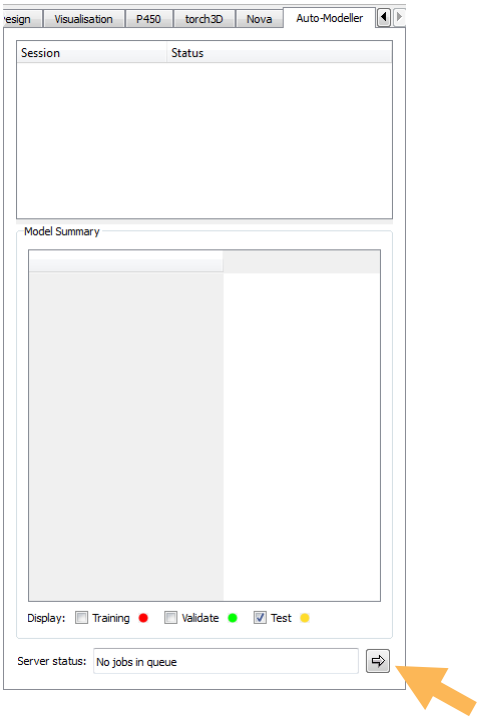
- Select the  tool from the toolbar to open the **Mathematical Function Editor**
- In the **f(x)** field, enter the equation “6-log({Ki (uM)})”. This can be easily achieved by pointing and clicking in the editor (or by copying and pasting without the quotes). Enter the name of the new column, **pKi**, in the **New Column Name** field and click **OK**.



The new column containing pK_i values will appear in the data set. Now we’re ready to build a model of this data.

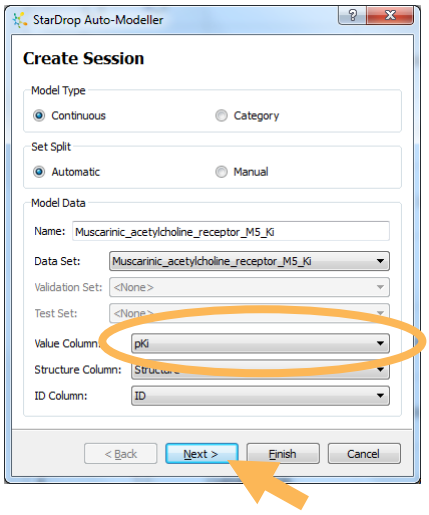
	Structure	ID	Ki (uM)	pKi
1		CHEMBL173955	41.98	4.377
2		CHEMBL175059	0.02999	7.523
3		CHEMBL227429	33.11	4.48
4		CHEMBL177040	1.099	5.959
5		CHEMBL177322	2	5.699
6		CHEMBL195	0.0009174	9.037

- Change to the **Auto-Modeller** tab on the left and click on the  button to begin a new modelling session



This will open the Auto-Modeller wizard.

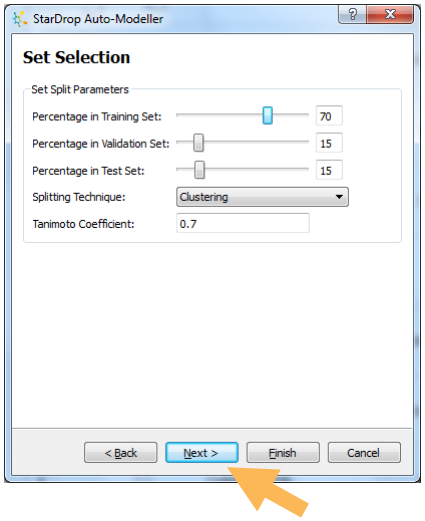
The first page allows you to choose the type of model to build: continuous (i.e. numerical) or category (i.e. classification). You can also choose whether to allow StarDrop's Auto-Modeller to split the data into Training, Validation and Test sets automatically or to provide these separate sets automatically. Finally, you can confirm the data set to model and the columns containing the property values to model, the compound structures and compound identifiers. In this case, we will use the default settings, but we need to select the correct column to model.



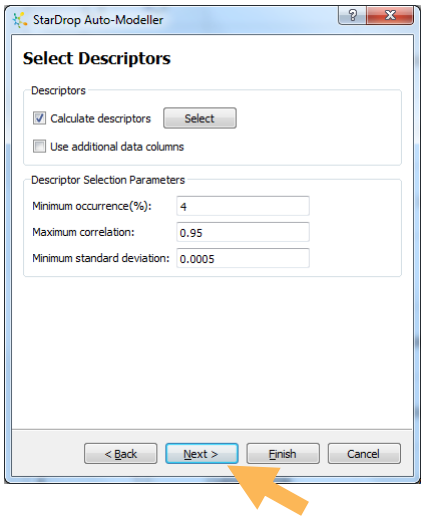
- Choose the **pKi** column from the **Value Column** drop-down menu and click **Next**

The next page allows you to configure the parameters for the automatic selection of Training, Validation and Test sets. In this case we will use the default set split parameters.

- Click **Next**

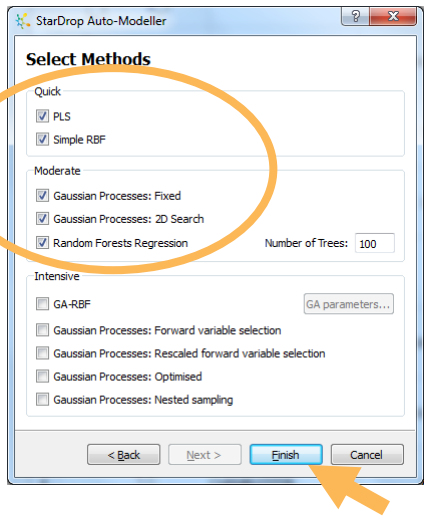


The next page allows you to select the descriptors to use. The built-in library of whole molecule and 2D descriptors will be used by default. More details are available by clicking the **Select** button and new descriptors can be imported as SMARTS. Also, additional columns in the data set can be used as descriptors. Finally, the parameters for selection of descriptors can be defined. Again, we will use the default settings.

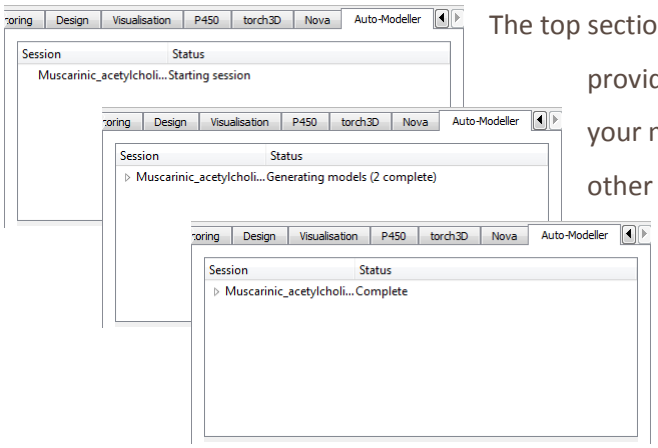


- Click **Next**

The final page of the Auto-Modeller wizard allows you to select the modelling methods to apply. The methods are categorised by their computational cost and the methods selected by default will depend on the size of the data set; in this case all of the methods will be selected by default.

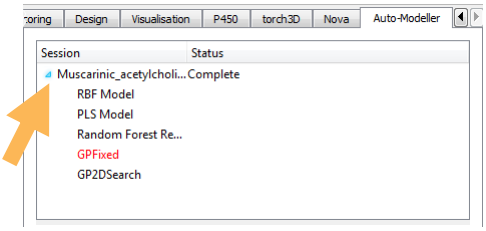


- We would recommend unselecting the **Intensive** methods to provide a quick example
- Click **Finish** to begin the modelling session



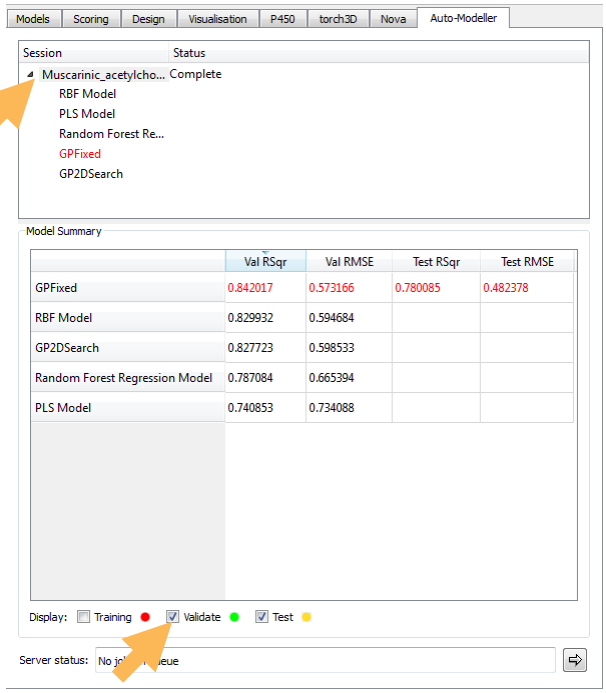
The top section of the **Auto-Modeller** tab will provide a running update of the progress of your modelling session. If there are no other modelling sessions ahead of yours in the queue on the server, this should only take a few minutes (you can check the status of the server at the bottom of the **Auto-Modeller** tab).

- When the modelling session is complete, click the arrow or plus next to the session in the **Auto-Modeller** tab to see a list of models generated.



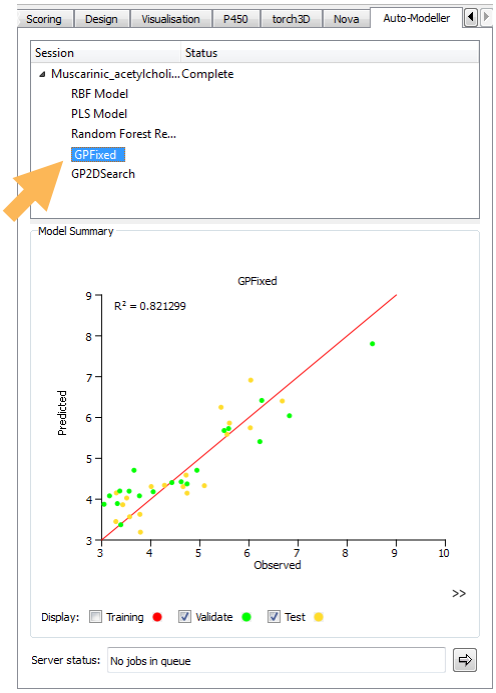
The model with the best result on the Validation set will be highlighted in red. Please note that the specific results you see may be slightly different to the examples shown here due to a random element in the assignment of compounds to the Training, Validation and Test sets.

- Select the modelling session to see a table summarising the results of the different models. This will show the result of the best model on the Test set.
- Tick the **Validate** box at the bottom of the tab to see the results for all of the models on the Validation set.



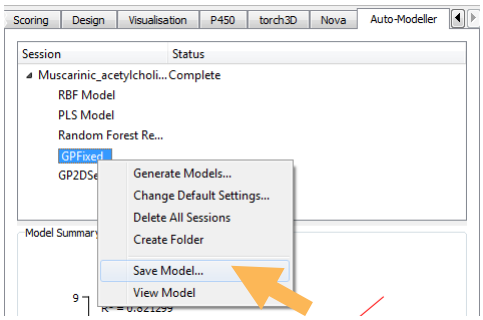
You can see that the Auto-Modeller has built a model with each of the modelling methods and compared the performance of these on the Validation set to identify the most predictive model. This best model is then further validated using the external test set. A robust model should have good performance on both the Validation and Test sets.

- Select a model to see a plot of the results for the Validation and Test sets and confirm that the model is producing reliable predictions.
- Hover the mouse pointer over a data point to see the corresponding structure.

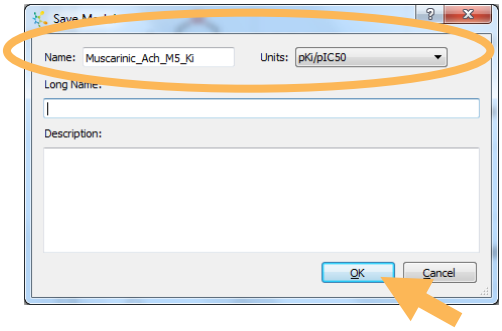


If you are happy with the results for a model, you can save it so that you can apply it to new compounds to predict the affinity and visualise the structure-activity relationship.

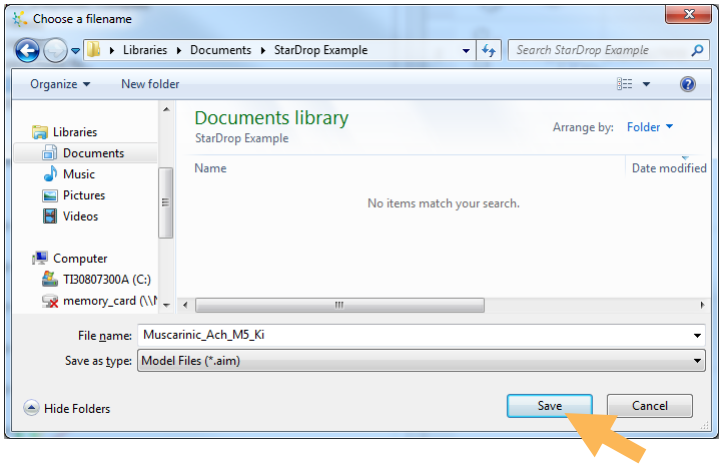
- Right click on the best model in the **Auto-Modeller** tab and select the **Save Model** option .



- Enter an appropriate name and set the units to **pKi/pIC50** in the **Save Model** dialogue box that appears and click **OK** (if you wish you can also enter more information in the **Long Name** and **Description** boxes).

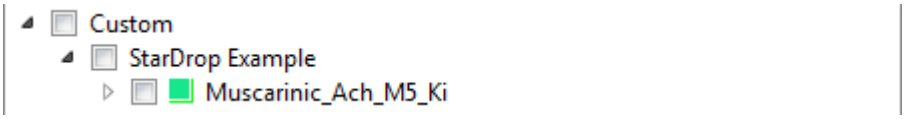


- Finally, navigate to a convenient directory and click **Save** to save the model.



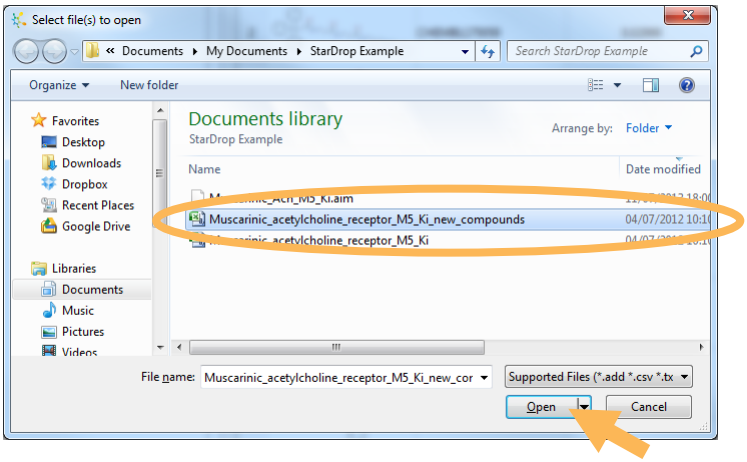
N.B. The resulting file can be shared with any other StarDrop user who can load and use the model in their copy of StarDrop.

- Switch to the **Models** tab and you will see that the model has appeared under the corresponding directory name, ready to run on new compounds.

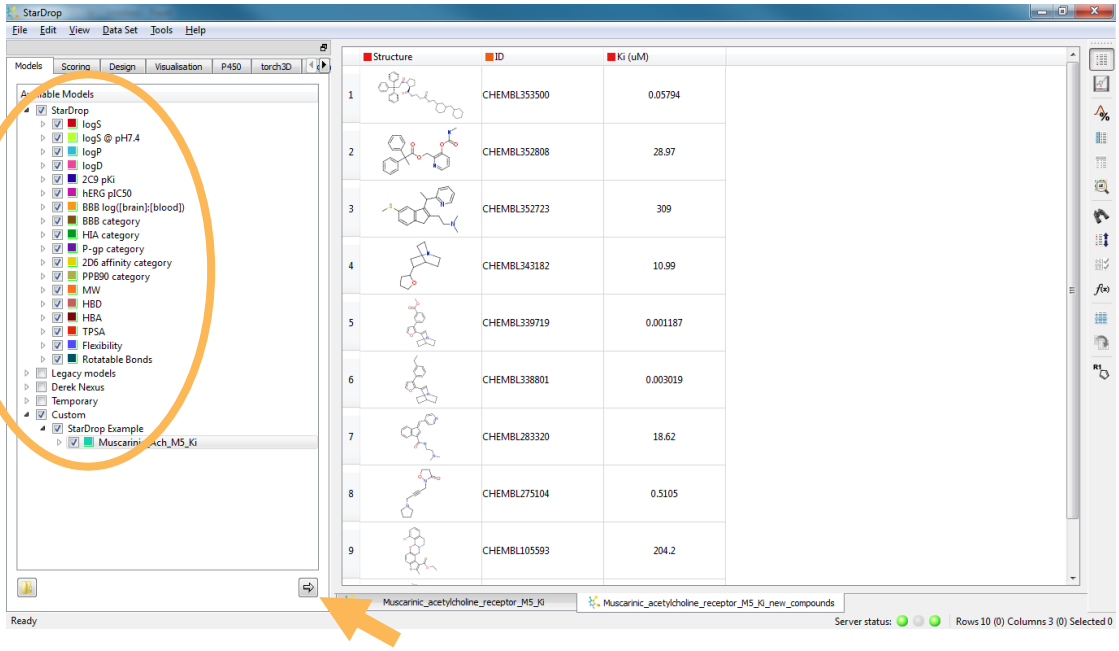


To illustrate the application of this model to new compounds, we're going to load another data set containing an additional 10 compounds.

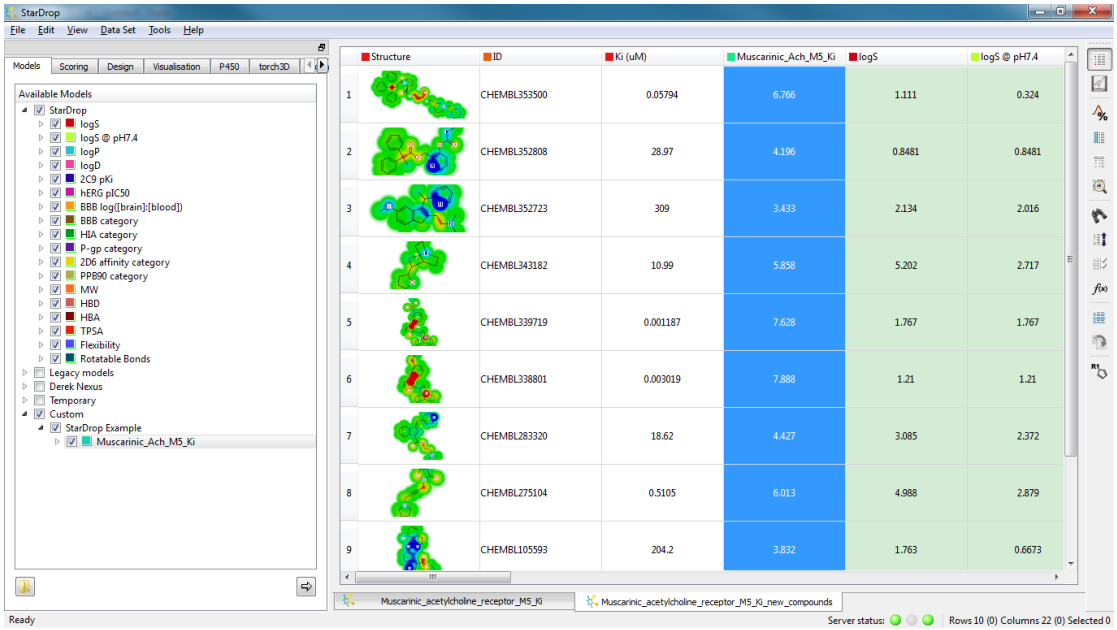
- Choose the **File->Open** menu option again and load the file **Muscarinic_acetylcholine_receptor_M5_Ki_new_compounds.csv**



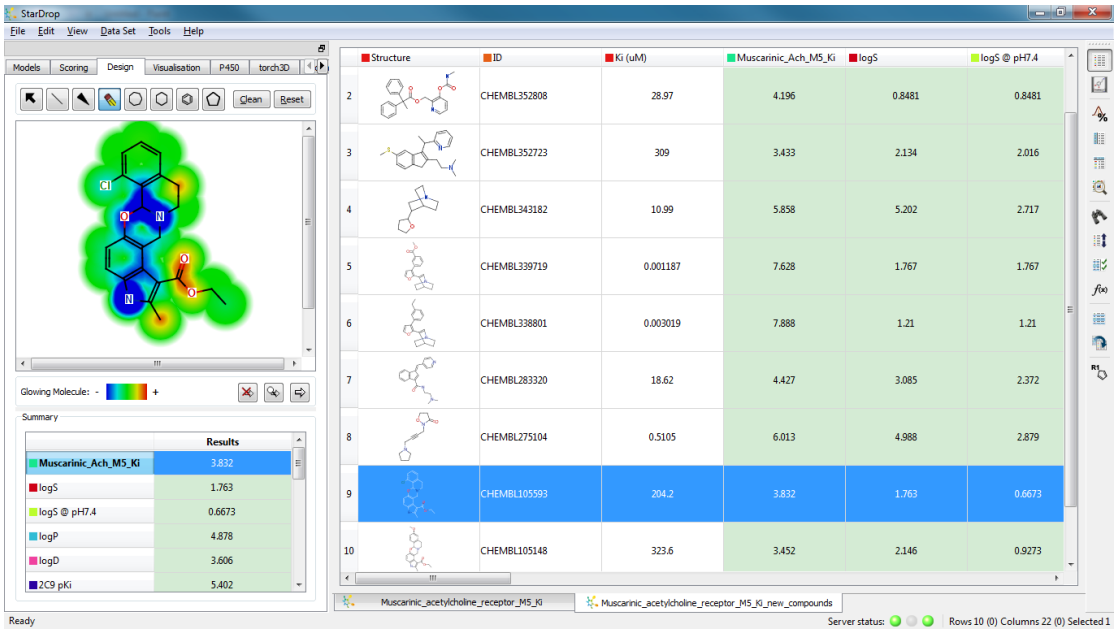
- Import this new data using the procedure described above
- In the **Models** tab, select the new model we have built along with any other models you would like to run by ticking the boxes next to the models and click the  button.



The new model can be used in the same way as any other model in StarDrop. For example, selecting the column header will display the Glowing Molecule visualisation for each compound, showing the structure-activity relationship captured by the model we have built.



Choosing the **Design** and selecting a row in the data set will allow you to explore optimisation strategies, guided by the Glowing Molecule.



This has been a quick example of the application of StarDrop's Auto-Modeller. There are, of course, additional features allowing expert modellers to control the detailed parameters of the model building process and explore the detailed results for each model. For more information or to arrange a comprehensive demo, please contact stardrop-support@optibrium.com.