

Worked Example:

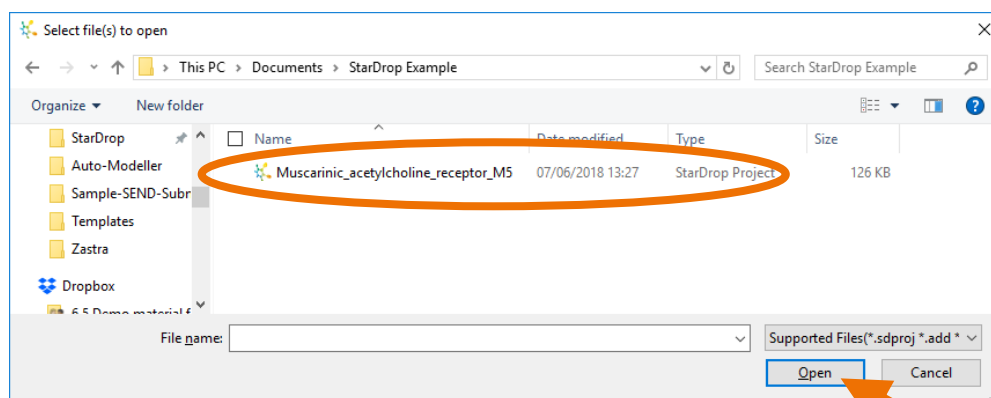
Automatic QSAR Model Building and Validation

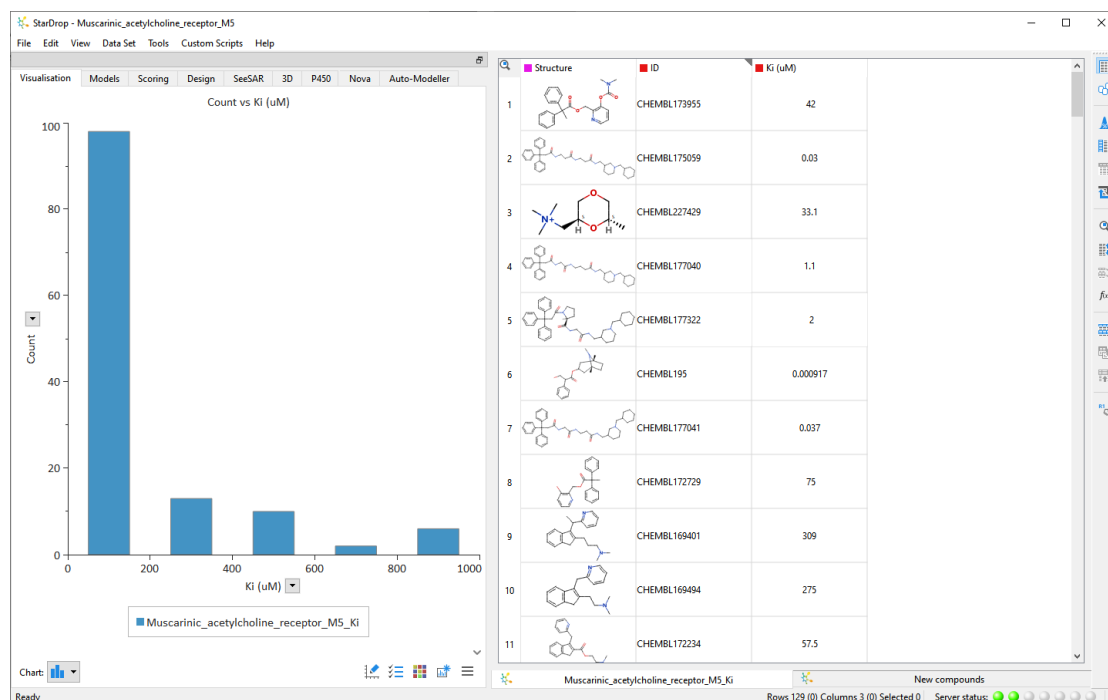
In this example, we will explore the application of StarDrop's Auto-Modeller to build a QSAR model of potency against the Muscarinic Acetylcholine M5 receptor, based on a set of public domain K_i data obtained from the ChEMBL database (<https://www.ebi.ac.uk/chembl/>). The resulting model will be applied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

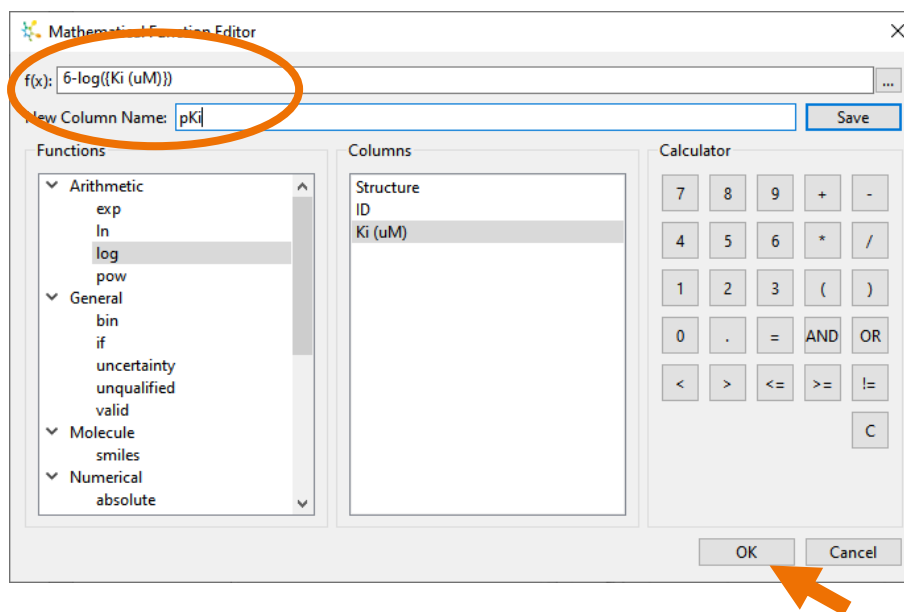
- Open the StarDrop Project file **Muscarinic_acetylcholine_receptor_M5.sdproj** by selecting **Open** from the **File** menu.



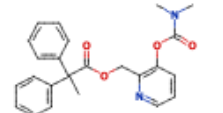

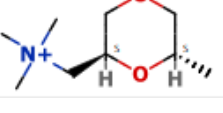


The first data set displayed contains 129 compounds with measured activity against the muscarinic acetylcholine receptor M5. These data are K_i values in μM . However, to build a good model, the data should be converted to logged units. Logged units provide a more even distribution of values to model and a better correlation with the compound descriptors used to build the model. Therefore, we will use the **Mathematical Function Editor** in StarDrop to generate $\text{p}K_i$ values.

- Select the $f(x)$ button on the right-toolbar to open the **Mathematical Function Editor**.
- In the **f(x)** field, enter the equation “ $6 - \log(\{K_i \text{ (uM)}\})$ ”. This can be easily achieved by pointing and clicking in the editor (or 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 these data.

	Structure	ID	Ki (uM)	pKi
1		CHEMBL173955	42	4.38
2		CHEMBL175059	0.03	7.52
3		CHEMBL227429	33.1	4.48

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

	Structure	ID	Ki (uM)	pKi
1	<chem>C1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3</chem>	CHEMBL173955	42	4.38
2	<chem>C1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3</chem>	CHEMBL175059	0.03	7.52
3	<chem>C1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3</chem>	CHEMBL227429	33.1	4.48
4	<chem>C1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3</chem>	CHEMBL177040	1.1	5.96
5	<chem>C1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3</chem>	CHEMBL177322	2	5.7
6	<chem>C1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3</chem>	CHEMBL195	0.000917	9.04
7	<chem>C1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3</chem>	CHEMBL177041	0.037	7.43
8	<chem>C1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3</chem>	CHEMBL172729	75	4.12
9	<chem>C1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3</chem>	CHEMBL169401	309	3.51
10	<chem>C1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3</chem>	CHEMBL169494	275	3.56
11	<chem>C1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3</chem>	CHEMBL172234	57.5	4.24

This will open the Auto-Modeller wizard.

The first page enables 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 provide these separate sets yourself. Finally, you can confirm the data set and the columns containing the structures and property values to be modelled. We will use most of the default settings in this case, but we need to select the correct property column to model.

StarDrop Auto-Modeller

Create Session

Model Type: Continuous Category

Set Split: Automatic Manual

Model Data:

Name: Muscarinic_acetylcholine_receptor_M5_Ki

Data Set: Muscarinic_acetylcholine_receptor_M5_Ki

Validation Set: <None>

Test Set: <None>

Value Column: pKi

Structure Column: Structure

< Back Next > Finish Cancel

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

The next page enables 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**.

StarDrop Auto-Modeller

Set Selection

Set Split Parameters

Percentage in Training Set: 70

Percentage in Validation Set: 15

Percentage in Test Set: 15

Splitting Technique: Clustering

Tanimoto Coefficient: 0.7

< Back Next > Finish Cancel

The third page enables you to select the descriptors to use. The built-in library of whole-molecule and 2D descriptors are selected by default. More details are available by clicking **Select**. It is possible to add your own descriptors as SMARTS patterns or to use additional columns in the data set as descriptors. Finally, the parameters for the selection of descriptors can be defined. Again, we will use the default settings.

- Click **Next**.

StarDrop Auto-Modeller

Select Descriptors

Descriptors

Calculate descriptors **Select**

Use additional data columns

Descriptor Selection Parameters

Minimum occurrence(%): 4

Maximum correlation: 0.95

Minimum standard deviation: 0.0005

< Back Next > Finish Cancel

The final page of the Auto-Modeller wizard enables 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.

- As shown to the right, we would recommend deselecting the Intensive methods for this quick example.
- Click **Finish** to begin the modelling session.

StarDrop Auto-Modeller

Select Methods

Quick

PLS

Simple RBF

Moderate

Gaussian Processes: Fixed

Gaussian Processes: 2D Search

Random Forests Regression Number of Trees: 100

Intensive

GA-RBF GA parameters...

Gaussian Processes: Forward variable selection

Gaussian Processes: Rescaled forward variable selection

Gaussian Processes: Optimised

Gaussian Processes: Nested sampling

< Back Next > Finish Cancel

Visualisation | Models | Scoring | Design | SeeSAR | 3D | P450 | Nova | Auto-Modeller

Session	Status
Muscarinic_acetylcholine_receptor_M5_Ki	Starting session

Visualisation | Models | Scoring | Design | SeeSAR | 3D | P450 | Nova | Auto-Modeller

Session	Status
> Muscarinic_acetylcholine_receptor_M5_Ki	Generating models (4 complete)

Visualisation | Models | Scoring | Design | SeeSAR | 3D | P450 | Nova | Auto-Modeller

Session	Status
> Muscarinic_acetylcholine_receptor_M5_Ki	Complete

The top section of the **Auto-Modeller** area provides a running update on 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 server's status at the

bottom of the **Auto-Modeller** area).

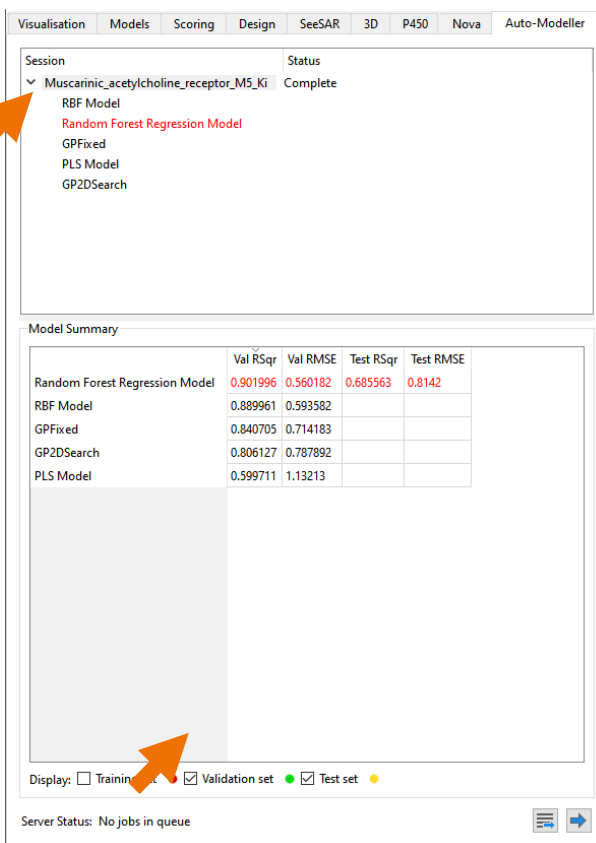
- When the modelling session is complete, click the arrow next to the session to open up the list of models that have been generated.

Visualisation | Models | Scoring | Design | SeeSAR | 3D | P450 | Nova | Auto-Modeller

Session	Status
▼ Muscarinic_acetylcholine_receptor_M5_Ki RBF Model Random Forest Regression Model GPFixed PLS Model GP2DSearch	Complete

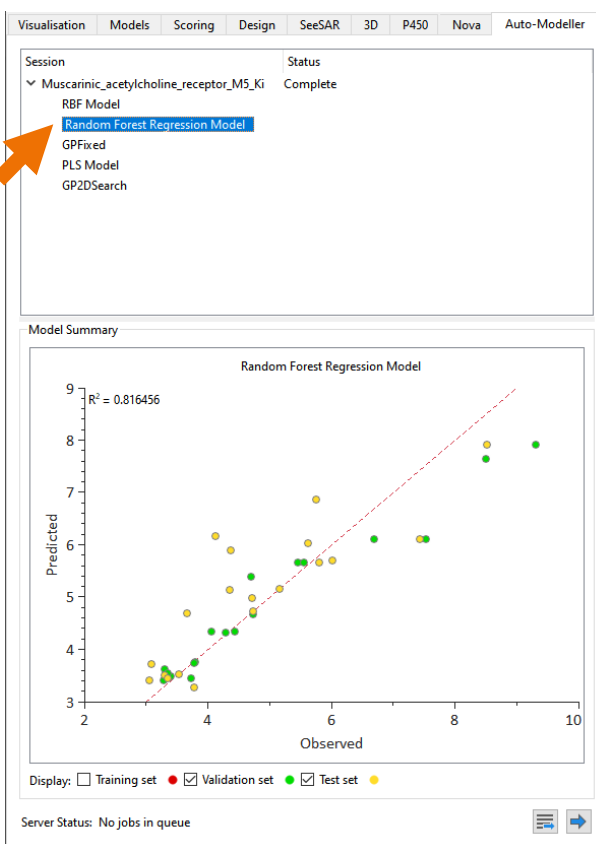
The model with the best result on the validation set will be highlighted in red. Please note that the specific results you see may differ from the examples shown here due to a random element in the assignment of compounds to the training, validation and test sets.

- Select the modelling session to see a summary of the different models. This will show the result of the best model on the test set.
- Tick the **Validation set** option at the bottom to see results for all of the models on the validation set.

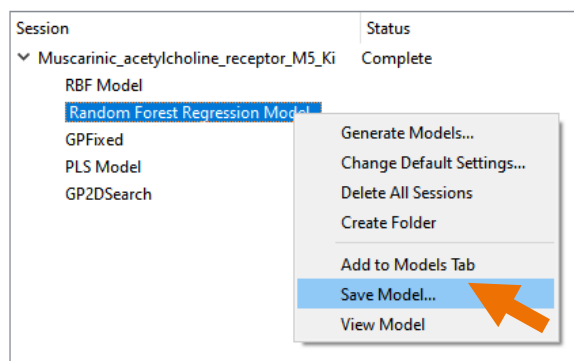


You can see that the Auto-Modeller has built a model with each of the modelling methods and compared the performance of these for the validation set to identify the most predictive model. This best model is then further validated using the external test set. A robust model should perform well for both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

- Select a model to see a plot of the validation and test sets and confirm that the model is producing reliable predictions.
- Hover the mouse pointer over a data point to see its corresponding structure.

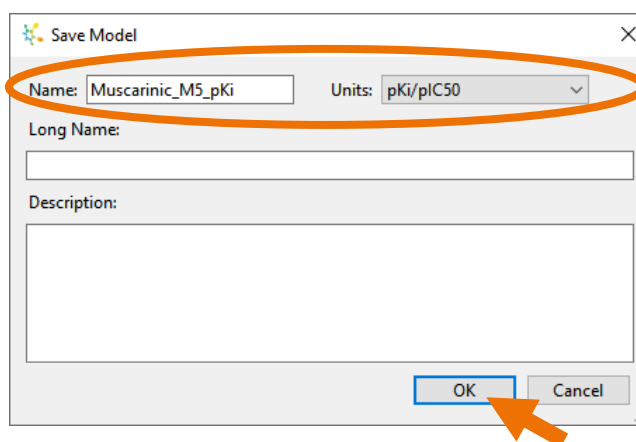


If you are happy with the results for a model, you can save it to apply it to new compounds to predict the potency and visualise the structure-activity relationships. In this case, we will save the random forest regression model.

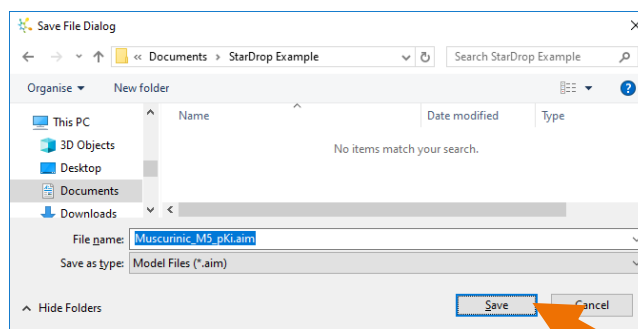


- Right-click on the **Random Forest Regression Model** under the modelling session in the **Auto-Modeller** area and select the **Save Model** option.

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

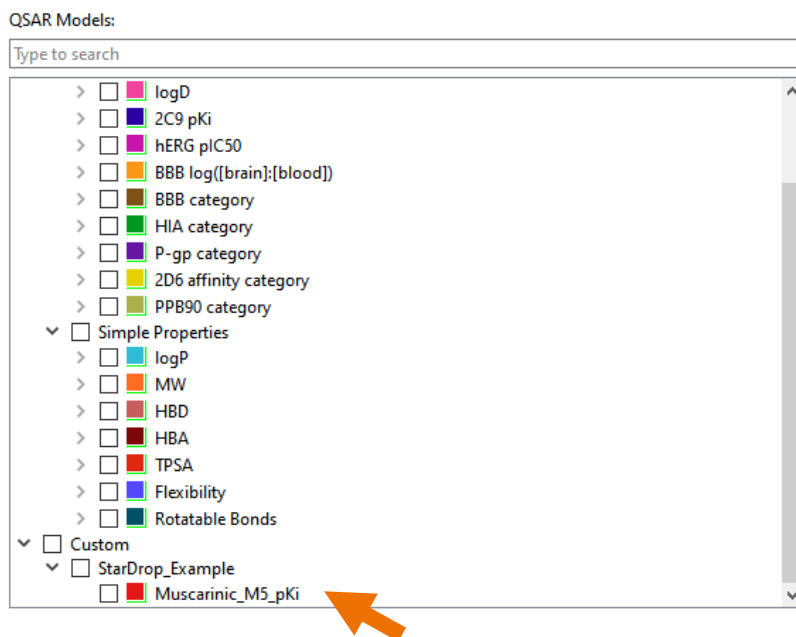


- Finally, navigate to a convenient directory and click the **Save** button to save the model.



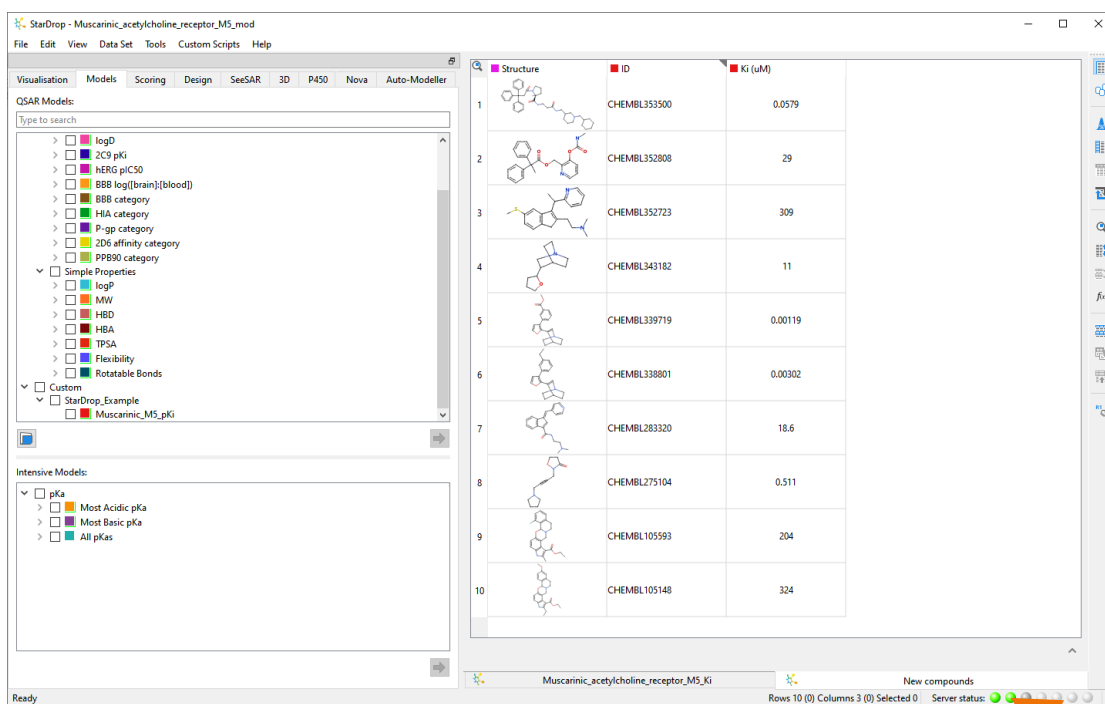
Note: The resulting model can be shared with any other StarDrop user to load and use in their copy of StarDrop.


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



We'll illustrate the application of this model to an additional 10 compounds that were not included in the data set used to build and validate the model.

- Change to the **New compounds** data set by clicking on the tab at the bottom of the data set.



- In the **Models** area, select the new model and any others you would like to run by ticking the box next to the model and clicking the  button.

StarDrop - Muscarinic_acetylcholine_receptor_M5_mod

File Edit View Data Set Tools Custom Scripts Help

Visualisation Models Scoring Design SeeSAR 3D P450 Nova Auto-Modeller

QSAR Models

Type search

- logD
- 2C9 pKi
- HERG pIC50
- BBB log([brain]/[blood])
- BBB category
- HIA category
- P-gp category
- 2D6 affinity category
- PPB90 category
- Simple Properties
 - logP
 - MW
 - HBD
 - HBA
 - TPSA
 - Flexibility
 - Rotatable Bonds
- Custom
 - StarDrop_Example
 - Muscarinic_M5_pKi

Intensive Models

- pKa
 - Most Acidic pKa
 - Most Basic pKa
 - All pKas

Structure	ID	Ki (uM)
	CHEMBL353500	0.0579
	CHEMBL352808	29
	CHEMBL352723	309
	CHEMBL343182	11
	CHEMBL339719	0.00119
	CHEMBL338801	0.00302
	CHEMBL283320	18.6
	CHEMBL275104	0.511
	CHEMBL105593	204
	CHEMBL105148	324

Muscarinic_acetylcholine_receptor_M5_Ki

New compounds

Rows 10 (0) Columns 3 (0) Selected 0 Server status: ● ● ● ● ● ● ● ● ● ●

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.

Changing to the **Design** area and selecting a row in the data set will enable you to explore optimisation strategies guided by the Glowing Molecule.

StarDrop - Muscarinic_acetylcholine_receptor_M5_mod

File Edit View Data Set Tools Custom Scripts Help

Visualisation Models Scoring Design SeeSAR 3D P450 Nova Auto-Modeller

Clean Reset

Property: No Property

Results

Property	Results
2C9 aminty category	low
PPB90 category	4.75
logP	279
MW	0
HBD	2
HBA	16.4
TPSA	0.125
Flexibility	3
Rotatable Bonds	7.53
Muscarinic_M5_pKi	

Structure	ID	Ki (uM)	logS	logS @ pH7.4
	CHEMBL353500	0.0579	1.08	0.274
	CHEMBL352808	29	0.856	1.22
	CHEMBL352723	309	2.13	2.02
	CHEMBL343182	11	5.2	2.72
	CHEMBL339719	0.00119	1.77	2.29
	CHEMBL338801	0.00302	1.21	1.65
	CHEMBL283320	18.6	3.09	2.37
	CHEMBL275104	0.511	4.99	2.88
	CHEMBL105593	204	1.76	0.667
	CHEMBL105148	324	2.15	0.927

Muscarinic_acetylcholine_receptor_M5_Ki

New compounds

Rows 10 (0) Columns 22 (0) Selected 1 Server status: ● ● ● ● ● ● ● ● ● ●

This has been a quick example of the application of StarDrop's Auto-Modeller. There are, of course, additional features enabling expert modellers to control the parameters of the model

building process and explore the detailed results for each model. For more information or to arrange a comprehensive demo, please get in touch with stardrop-support@optibrium.com.