

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 (<u>https://www.ebi.ac.uk/chembl/</u>). 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.

K. Select file(s) to open	the second se	x
OO - 🚺 « Docu	ments 🕨 My Documents 🔸 StarDrop Example 🔹 🍫 Search StarDrop Exam	nple 🔎
Organize 👻 New f	older 🛛 🔠 🔻	
☆ Favorites ■ Desktop	Documents library StarDrop Example	Folder 🔻
Downloads	E Name	Date modified
Secent Places	wuscarinic_acetylcholine_receptor_M5_Ki_new_compounds	04/07/2010 10:10
Google Drive	Muscarinic_acetylcholine_receptor_M5_Ki	04/07/2012 10:1
🕞 Libraries		
🌙 Music		
Pictures		
Videos	▼	•
Fi	le name: Muscarinic_acetylcholine_receptor_M5_Ki	*.csv *.tx ▼ Cancel



Optibrium™, StarDrop™, Nova™ Glowing Molecule™ and Auto-Modeller™ are trademarks of Optibrium Ltd.

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

	Structure	ID	Ki (uN	1)
1	Molecule 🔹	Text 🔹	Numbe	•
2	Strog	CHEMBL173955	41.98	
3	guero	CHEMBL175059	0.02999	
4	XQ.	CHEMBL227429	33.11	
5	ginion	CHEMBL177040	1.099	
6	of 2	CHEMBL177322	2	
7	30th	CHEMBL195	0.000917	
8	ofins	CHEMBL177041	0.03698	
9	53	CHEMBL172729	74.99	
10	аŝ,	CHEMBL169401	309	

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

	Structure	ID	Ki (uM)
1	J. J.	CHEMBL173955	41.98
2	ofuneror	CHEMBL175059	0.02999
3	+	CHEMBL227429	33.11
4	ofinero	CHEMBL177040	1.099
5	072,00	CHEMBL177322	2
6	git	CHEMBL195	0.0009174

These data are K_i values in μ M. 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 $f^{(x)}$ 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)	рКі
1	J. J.	CHEMBL173955	41.98	4.377
2	officer	CHEMBL175059	0.02999	7.523
3		CHEMBL227429	33.11	4.48
4	offiningo	CHEMBL177040	1.099	5.959
5	07-2-00	CHEMBL177322	2	5.699
6	g.A	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

1	Visualisa	ation	P450	torch3D		Nova	Auto	Modeller		
ess	ion			Status						
Mod	el Summa	ry								
Disp	olay: 📃	Training	• [🔲 Validate	•	V Tes	t 💻			
		a								
erve	r status:	No job	s in quei	ue					5	
		140 JUD:	s in que	uc.					-	

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.

Create Se	ssion
Model Type	
Continuous	Category
Set Split	
Automatic	Manual
Model Data	
Name: Musca	rinic_acetylcholine_receptor_M5_Ki
Data Set:	Muscarinic acetylcholine receptor M5 Ki
Validation Set:	<none></none>
	<none></none>
Test Set:	
Test Set: Value Column	ркі
Test Set: Value Column Structure Colu	pKi
Test Set: Value Column Structure Colu ID Column:	t pří v Struur v ID v
Test Set: Value Column: Structure Colu ID Column:	při v Struce v ID v

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

 Percentage in Training Set:
 0

 Percentage in Validation Set:
 15

 Percentage in Test Set:
 15

 Splitting Technique:
 Lustering

 Tanimoto Coefficient:
 0.7

StarDrop Auto-Modeller

Set Split Parameters

? <mark>X</mark>

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.

👯 StarDrop Auto-Modeller		? ×
Select Descriptors		
Descriptors Calculate descriptors Use additional data column	Select	
Descriptor Selection Paramete	rs	
Minimum occurrence(%):	4	
Maximum correlation:	0.95	
Minimum standard deviation:	0.0005	
< <u>B</u> ack	Next > Einish	Cancel

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

👯 StarDrop Auto-Modeller	8 ×
Select Methods	
Quick	
V PLS	
Simple RBF	
Moderate	
Gaussian Processes: Fixed	
Gaussian Processes: 2D Search	
Random Forests Regression Number of Tr	ees: 100
Intensive	
GA RBF GA p	parameters
Gaussian Processes: Forward variable selection	
Gaussian Processes: Rescaled forward variable selection	
Gaussian Processes: Optimised	
Gaussian Processes: Nested sampling	
	Crarel
<u> vext</u> <u>Pext</u>	

toring	Design	Visualis	sation	P450	torch3D	Nova	Aut	:o-Modelle	r ()	The	e top
Se	ssion		State	us							
	Muscarinic	acetylch	oliStarti	ing sess	ion						
Session Stat Muscarinic_acetylcholiStart Session Design Session > Muscarinic		Visualisation P450 torch3D Nova				va Auto	Auto-Modeller				
	Session		on		Sta	tus					
		⊳M	luscarinic	_acetyl	choli Ger	ierating	models	(2 compl	ete)		
				oring	Design	Visuali	sation	P450	torch3D	Nova	Auto-Mo
				Sess	ion		St	atus			
		⊳ N	1uscarinic _.	_acetylcl	noli Co	mplete					

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.

odel Summary				
	Val RSqr	Val RMSE	Test RSqr	Test RMSE
GPFixed	0.842017	0.573166	0.780085	0.482378
RBF Model	0.829932	0.594684		
GP2DSearch	0.827723	0.598533		
Random Forest Regression Model	0.787084	0.665394		
PLS Model	0.740853	0.734088		

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.
- Scoring Design Visualisation P450 torch3D Nova Auto-Modeller Status Session Muscarinic_acetylcholi...Complete RBF Model PLS Model Random Forest Re. GP GP2DS€ Generate Models... Change Default Settings. Delete All Sessions Model Summa Create Folder Save Model. View Model 9٦

Units: pKi/pIC50

•

- 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.



K. Save Ma

Long Name

Description

Name: Muscarinic_Ach_M5_Ki

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



⇔

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.

StarDrop						<u>ار می</u>	5 - X	3
<u>File Edit View Data Set Tools H</u> elp								
8	Structure	ID	Ki (uM)	Muscarinic_Ach_M5_Ki	logS	logS @ pH7.4	ī	
Available Models	1	CHEMBL353500	0.05794	6.766	1.111	0.324		⊴ ∧
▷ ♥ ■ logS ▷ IogS ▷ IogS □ <td< td=""><td>2</td><td>CHEMBL352808</td><td>28.97</td><td>4.196</td><td>0.8481</td><td>0.8481</td><td></td><td></td></td<>	2	CHEMBL352808	28.97	4.196	0.8481	0.8481		
	3	CHEMBL352723	309	3.433	2.134	2.016	4	0 *
▷ ✓ ●	4	CHEMBL343182	10.99	5.858	5.202	2.717	E	f(x)
 ▷ ♥ ■ HBD ▷ ♥ ■ HBA ▷ ♥ ■ TPSA ▷ ♥ ■ TPSA ▷ ♥ ■ Flexibility 	5	CHEMBL339719	0.001187	7.628	1.767	1.767	-	0
V Catable Bonds Catable Bonds Catable Bonds V Legacy models O catable Nexus V Temporary	6 🚴	CHEMBL338801	0.003019	7.888	1.21	1.21	R	* ' 0
V Custom V StarDrop Example W Muscarinic_Ach_M5_Ki	7	CHEMBL283320	18.62	4.427	3.085	2.372		
	8	CHEMBL275104	0.5105	6.013	4.988	2.879		
	9	CHEMBL105593	204.2	3.832	1.763	0.6673	Ŧ	
	K. Muscarinic_acetylcholine	_receptor_M5_Ki	Muscarinic_acetylcholine_recep	otor_M5_Ki_new_compounds			<u> </u>	

Choosing the **Design** and selecting a row in the data set will allow you to explore

optimisation strategies, guided by the Glowing Molecule.

StarDrop							I X
<u>Eile Edit V</u> iew <u>D</u> ata Set <u>T</u> ools <u>H</u> elp							
Models Scoring Design Visualisation P450 torch30 4 (*)	Structure	ID	Ki (uM)	Muscarinic_Ach_M5_Ki	■logS	logS @ pH7.4	1
	2 365	CHEMBL352808	28.97	4.196	0.8481	0.8481	
	3	CHEMBL352723	309	3,433	2.134	2.016	
	4	CHEMBL343182	10.99	5.858	5.202	2.717	<u>8</u>
	5	CHEMBL339719	0.001187	7.628	1.767	1.767	iii iii: fix
	6	CHEMBL338801	0.003019	7.888	1.21	1.21	
Glowing Molecule: -	, <i>a</i> f	CHEMBL283320	18.62	4.427	3.085	2.372	RI
Summary Results Results Summary	8	CHEMBL275104	0.5105	6.013	4.988	2.879	
log\$ 1.763 log\$ pH7.4	9						
logP 4,878 logD 3,606	10	CHEMBL105148	323.6	3.452	2.146	0.9273	-
■2C9 pKi 5.402 ▼	Muscarinic_acetylcholine	e_receptor_M5_Ki	Muscarinic_acetylcholine_rece	ptor_M5_Ki_new_compounds			<u> </u>

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 <u>stardrop-support@optibrium.com</u>.