

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.

👯 Select file(s) to open			x
Correction - Libraries - C	Documents 🔸 StarDrop Example	👻 🍫 Search	StarDrop Example
Organize 🔻 New folder			≣ - 🔟 🔞
★ Favorites ■ Desktop	Documents library StarDrop Example		Arrange by: Folder 🔻
Downloads	Name	Data modified	Туре
Stropbox	B Muscarinic_acetylcholine_receptor_M5_Ki	04/07/2012 10:10	Microsoft (cel Comma Se
👍 Google Drive	Muscarinic_acetylcholine_receptor_M5_Ki_new_comp	04/07/2012 10-10	incrosoft Excel Comma Se
🎉 Standard Emails			
Documents			
J Music			
Pictures			
Videos 🔫			•
File <u>n</u> ame	Muscarinic_acetylcholine_receptor_M5_Ki	✓ Support	ted Files(*.add *.csv *.bt Cancel



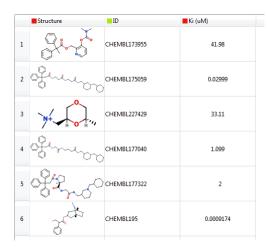
Optibrium[™], StarDrop[™], Nova[™] Glowing Molecule[™] and Auto-Modeller[™] are trademarks of Optibrium Ltd.

The **Import text file** dialogue 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 choosing Factor from the Type drop-down menu.

	Structure	ID	Ki (uM)
	Molecule 🔻	Text 🔹	Number -
	Surg	CHEMBL173955	41.98
<	Junion	CHEMBL175059	0.02999
	¥,	CHEMBL227429	33.11
	ofinin	CHEMBL177040	1.099
5	of the co	CHEMBL177322	2
	3 ch	CHEMBL195	0.0009174
3	ofino	CHEMBL177041	0.03698
,	्रद्व	CHEMBL172729	74.99
0	and	CHEMBL169401	309

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



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.

f(x): 6-log({Ki (uM)})		
New Column Name: pKi		
Functions	Columns	Calculator
 Mathematical pow 	Structure ID	7 8 9 + -
log In	Ki (uM)	4 5 6 * /
exp ▲ General		1 2 3 ()
valid sum		
min max	-	< > <= >= != C
if		
		OK Cancel

	Structure	ID ID	Ki (uM)	_ рКі
1		CHEMBL173955	41.98	4.377
2	ofineos	CHEMBL175059	0.02999	7.523
3	N+ H 0 H	CHEMBL227429	33.11	4.48
4	ofinios	CHEMBL177040	1.099	5.959
5	Jan o	CHEMBL177322	2	5.699
6	3 ch	CHEMBL195	0.0009174	9.037
5	of the wood	CHEMBL177322	2	5.699

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

tarDrop					
Edit View Data Set Tools Custom Help					
sign Visualisation P450 torch3D Nova Auto-Modeller	Structure	ID	Ki (uM)	<mark>p</mark> Ki	
ision Status	$_{1}$ Q_{1}	CHEMBL173955	41.98	4.377	
		D .			
	2 often	CHEMBL175059	0.02999	7.523	
del Summary	3	CHEMBL227429	33.11	4.48	
	· ofin	CHEMBL177040	1.099	5.959	
	· 42	CHEMBL177322	2	5.699	
	6 5	CHEMBL195	0.0009174	9.037	
	1 ofin	CHEMBL177041	0.03698	7.432	
	8	CHEMBL172729	74.99	4.125	
splay: 🛄 Training set ● 📃 Validation set ● 📝 Test set ●	•	CHEMBL169401	309	3.51	
ver status: No jobs in queue		Muscarinic_acetylcholine_receptor_	M5 Ki	1	
v			-	Server st	atus: 🥥 💿 🥥 🛛 Rows 129 (0) Columns 4 (0) Se

This will open the Auto-Modeller wizard.

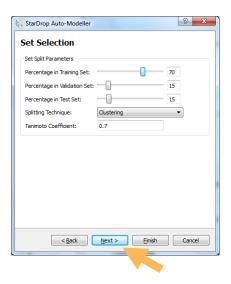
The first page enables you to choose the type of model to build: continuous (i.e. numerical) or category (i.e. classification). You can also choose whether to allow StarDrop's Auto-Modeller to split the data into Training, Validation and Test sets automatically or to provide these separate sets yourself. 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.

reate 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> v</none>
Test Set:	<none></none>
reacaeta	
Value Column.	ркі 👻
Value Column	
Value Column. Structure Colu	mn: Structure

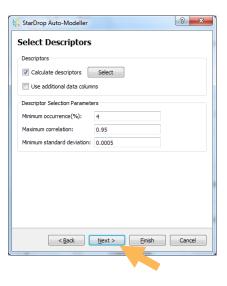
Choose the pKi column from the Value Column drop-down menu and click Next

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



The next page enables 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 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; in this case all of the methods will be selected by default.

- We would recommend unselecting the Intensive methods for this quick example
- Click Finish to begin the modelling session

K StarDrop Auto-Modeller	×
Select Methods	
Quick V PLS Simple RBF	
Moderate Image: Constraint of the second s	00
Intensive GA-RBF GA paramet	
Gaussian Processes: Forward variable selection Gaussian Processes: Rescaled forward variable selection Gaussian Processes: Optimised Gaussian Processes: Nested sampling	
	ncel

toring	Design	Visual	isation	P450	torch3D	Nova	Aut	to-Modeller		The	e top
Sess	ion		Stat	us							
P	Auscarinic	_acetylc	holiStarti	ing ses	sion						
		toring	Design	Visu	alisation	P450	torch3	3D Nov	/a Auto	-Modeller	
		Ses	sion		Sta	itus					
		⊳∣	Muscarinic	_acetyl	choliGer	nerating r	nodels	(2 comple	ete)		
				oring	Design	Visualis	sation	P450	torch3D	Nova	Auto-Mo
				Sess	ion		St	atus			
				⊳ N	Auscarinic	_acetylch	oliCo	mplete			

section of the Auto-Modeller tab will provide 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 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 open up the list of models generated.

oring De	sign	Visualisation	P450	torch3D	Nova	Auto-Modeller	
Session			Status				
R P R G	BF Mo LS Mo	del 1 Forest Re	Complete				

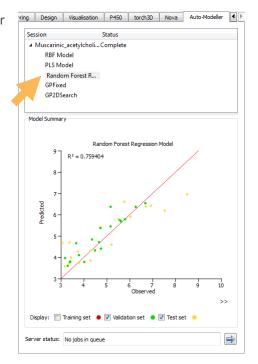
The model with the best result on the Validation set will be highlighted in red. Please note that the specific results you see may be slightly different to the examples shown here due to a random element in the assignment of compounds to the Training, Validation and Test sets.

- 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 Validation set box at the bottom of the tab to see the results for all of the models on the Validation set.

Muscarinic, acetylcholi Complete RBF Model PLS Model Random Forest Re GPFixed GP2DSearch	•			
1odel Summary	Val RSqr	Val RMSE	Test RSqr	Test RMSE
Random Forest Regression Model	0.762848	0.496961	0.755099	0.79306
RBF Model	0.708223	0.551232		
GP2DSearch	0.526001	0.702583		
GPFixed	0.495782	0.724634		
PLS Model	0.398659	0.791352		
Display: 🔲 Training set 🔍 Valida	tion set 🌘 🔽	Test set 😑		

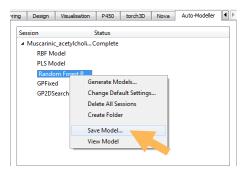
You can see that the Auto-Modeller has built a model with each of the modelling methods and compared the performance of these for the Validation set to identify the most predictive model. This best model is then further validated using the external Test set. A robust model should have good performance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

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



Units: pKi/pIC50

? X

•

- In the Save Model dialogue enter an appropriate name, set the units to pKi/pIC50 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.

ı boxes).				<u> </u>	Cancel
		nts 🕨 StarDrop Example	• [+j]	Search StarDrop Ex	
Organize 👻 New	w folder				≣≕ ▼ (
☆ Favorites ■ Desktop		Documents library StarDrop Example		Arrange by:	Folder 🔻
〕 Downloads 😵 Dropbox	≡ 1	Name		Date modified	Туре
Recent Places Google Drive Standard Email	ls		No items match your search		
📜 Libraries					
Documents					
Pictures		III			
File <u>n</u> ame:	Muscarinic_Ach_M	15_pKi			
Save as <u>t</u> ype:	Model Files (*.aim)				
Hide Folders			ſ	Save	Cancel

👯 Save Model

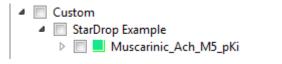
Long

Description:

Name: Muscarinic Ach M5 pKi

Note: 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

Organize 👻 New folde	r	8==	- 🔟 🕜
🔆 Favorites 📃 Desktop	Documents library StarDrop Example	Arrange by:	Folder 🔻
🐌 Downloads	Name	Date modified	Туре
Dropbox Recent Places	Eacetylcholine_receptor_M5_Ki	04/07/2012 10:10	Missosoft Exce
Google Drive	Muscarinic_acetylcholine_receptor_M5_Ki_new_compo	04/07/2012 10:10	Microsoft D
Standard Emails	Museprinic Ach M5 pKi.aim	02/11/2014 04:10	rile
Libraries Documents			
 Documents Music Pictures 			
Documents			
 Documents Music Pictures 	• • m		

- Import this new data using the procedure described previously
- 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 主 button.

StarDrop <u>Fi</u> le <u>E</u> dit <u>V</u> iew <u>D</u> ata Set <u>T</u> ools Custom <u>H</u> elp					
Fue Fau Tien Supper Joon envent Heb	8				
Model Design Visualisation P450 tord		Structure	ID ID	Ki (uM)	A
		~14p			
Available Models	1	"onen	CHEMBL353500	0.05794	
✓ StarDrop ▷ ♥ ■ logS		00			
IogS @ pH7.4		n. 👗			
 ▷ ♥ ■ logP ▷ ♥ ■ logD 	2	and i	CHEMBL352808	28.97	
V V 2C9 pKi					
▷ ☑ FrBS(pclS0) ▷ ☑ BB8 (sqlftrain)Ellocad ▷ ☑ HA2 (stegory) ▷ ☑ P-gp (stegory)					
	3	TO-	CHEMBL352723	309	
 P-gp category 2D6 affinity category 	4	1 A	CHEMBL343182	10.99	
PPB90 category		5			
▷ ♥ ■ MW ▷ ♥ ■ HBD		~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~			E
⊳ 🗹 📕 HBA	5	d_	CHEMBL339719	0.001187	
 TPSA Flexibility 		A			
Rotatable Bonds		5			
Legacy models Derek Nexus	6	Ľ	CHEMBL338801	0.003019	
Temporary		Å			
 ✓ Custom ✓ StarDrop Example 		~			
🕞 🗹 📕 M scarinic_Ach_M5_pKi	7	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	CHEMBL283320	18.62	
		<u> </u>			
	8	<u>``</u>	CHEMBL275104	0.5105	
		5	CI IEMBEZ/5104	0.5105	
		â			
	9	- B	CHEMBL105593	204.2	
		ye			
		× * * * * * * * *		1.	
Ready		Muscarinic_acetylcholi	ne_receptor_M5_Ki	K Muscarinic_acetylcholine_recept	Rows 10 (0) Columns 3 (0) Se

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.

dels Scoring Design Visualisation P450 torch3D	Structure	D ID	Ki (uM)	Muscarinic_Ach_M5_pK	logS	logS @ pH7.4	^
Available Models Image: Control of the second	1	CHEMBL353500	0.05794	6.318	1.111	0.324	
	2	CHEMBL35280B	28.97	4.67	0.8481	0.8481	
	3	CHEMBL352723	309	3.762	2.134	2.016	
	4	CHEMBL343182	10.99	5.507	5.202	2.717	Е
	5	CHEMBL339719	0.001187	6.916	1.767	1.767	
	6	CHEMBL338801	0.003019	8.028	1.21	1.21	
	7	CHEMBL283320	18.62	4.711	3.085	2.372	
	8	CHEMBL275104	HEMBL275104 0.5105	5.774	4.988	2.879	
	9	CHEMBL105593	204.2	3.674	1.763	0.6673	

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

le <u>E</u> dit <u>V</u> iew <u>D</u> ata Set <u>T</u> ools									
fodels Scoring Design Visua	alisation P450 torch3D 4 (+)		Structure	ID	Ki (uM)	Muscarinic_Ach_M5_pKi logS		logS @ pH7.4	*
	C C Clean Reset	2	Sent.	CHEMBL352808	28.97	4.67	0.8481	0.8481	
		з	-010	CHEMBL352723	309	3.762	2.134	2.016	
		4	se a la l	CHEMBL343182	10.99	5.507	5.202	2.717	
	0	5	a.	CHEMBL339719	0.001187	6.916	1.767	1.767	
		6	Ł	CHEMBL338801	0.003019	8.028	1.21	1.21	
		7	- R				3.085		E
ilowing Molecule: - +++++++++++++++++++++++++++++++++++	Results	8	8	CHEMBL275104	0.5105	5.774	4.988	2.879	
Muscarinic_Ach_M5_pKi	4.711 E								
logS	3.085	9		CHEMBL105593	204.2	3.674	1.763	0.6673	
logS @ pH7.4	2.372								
logP	2.404	10	2	CHEMBL105148	323.6	3.768	2.146	0.9273	
■ logD	0.8386		1						-
2C9 pKi	4.875	•	III						•

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 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>.