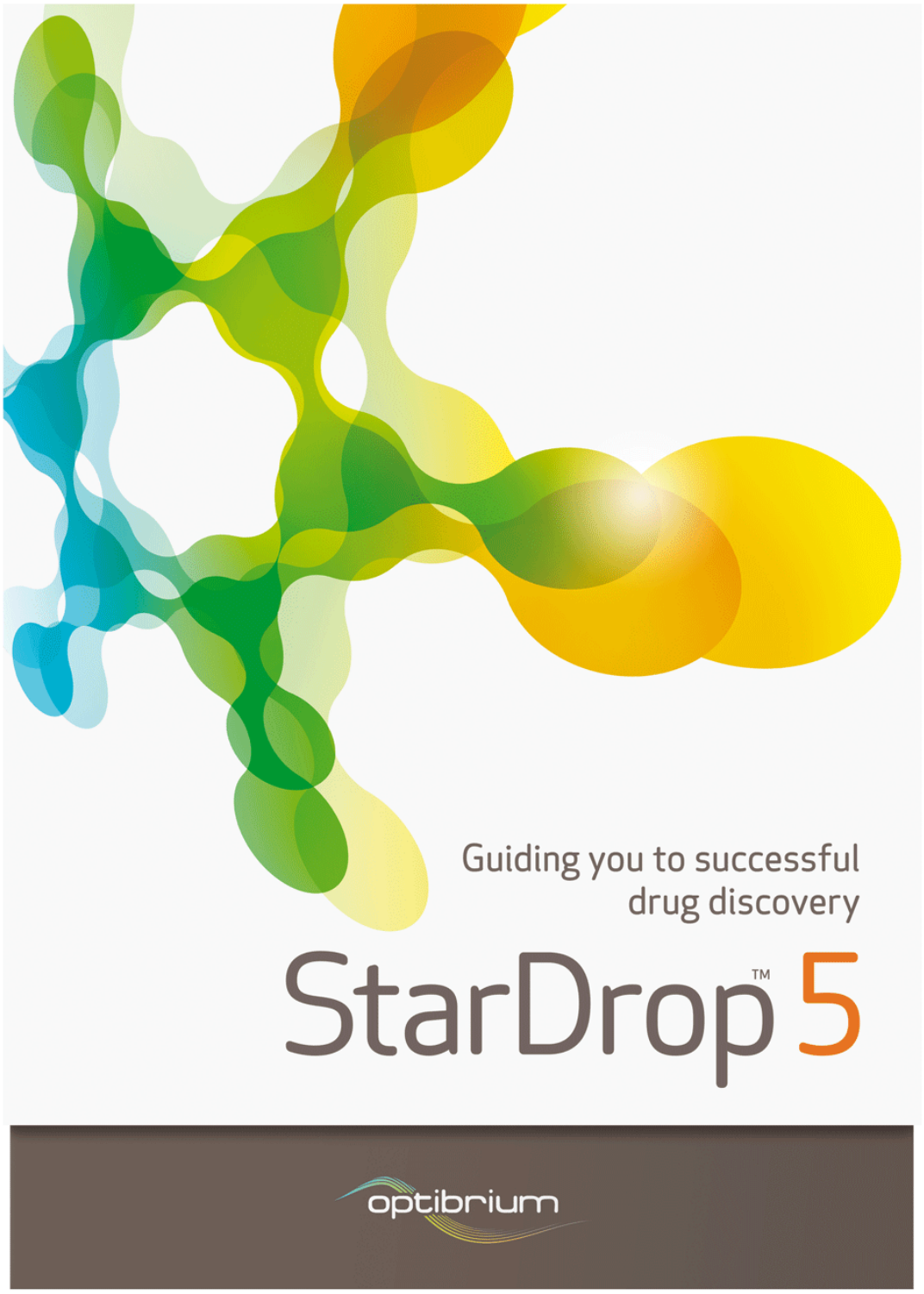
StarDrop™ Training Manual

Version 5.5



Contents

Hit-to-Lead: Targeting High Quality Lead Series

In hit-to-lead, it is important to quickly focus resources on the chemistries that are most likely to yield a high quality lead series. In this example we will explore how data from primary screening of a library for potency against the target 5HT1a can be combined with predictions for a range of ADME and physicochemical properties to identify chemistries with a good balance of properties. At the same time, given the uncertainty in the underlying data due to experimental variability and statistical error, it is important that we do not reject compounds inappropriately and risk missing valuable opportunities.

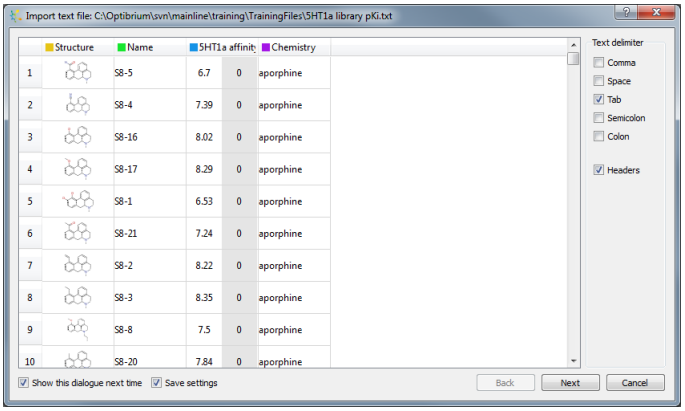
Importing Your Data and Predicting Compound Properties

Objectives

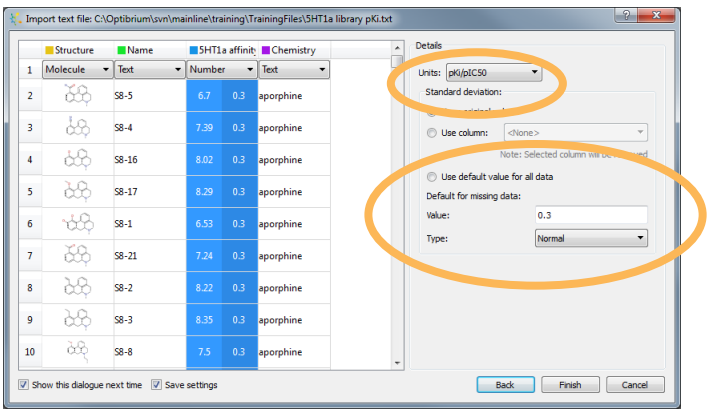
- Familiarity with the StarDrop interface
- Importing compound data
- Running models
- Exploring your data set
- Simple data visualisation

Exercise

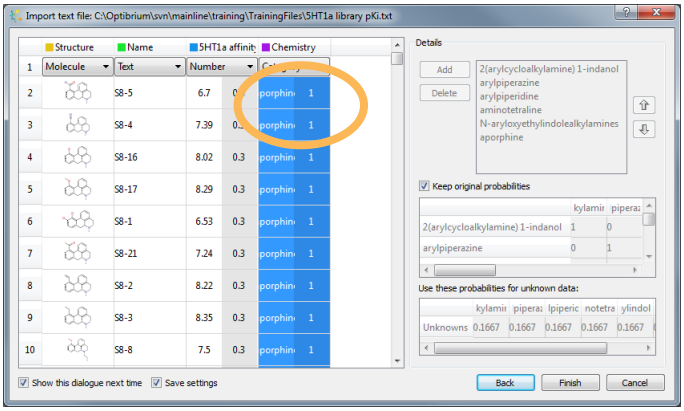
- Open the file **5HT1A library pKi.txt**. Confirm the text delimiter that has been used to create this file and click **Next**.



- Set the units of the experimental pKi data to **pKi/pIC50** and the uncertainty to **0.3 log units** (equivalent to a factor of two in the pK_i).



- Set the type of the **Chemistry** column to **Category**, as shown below:



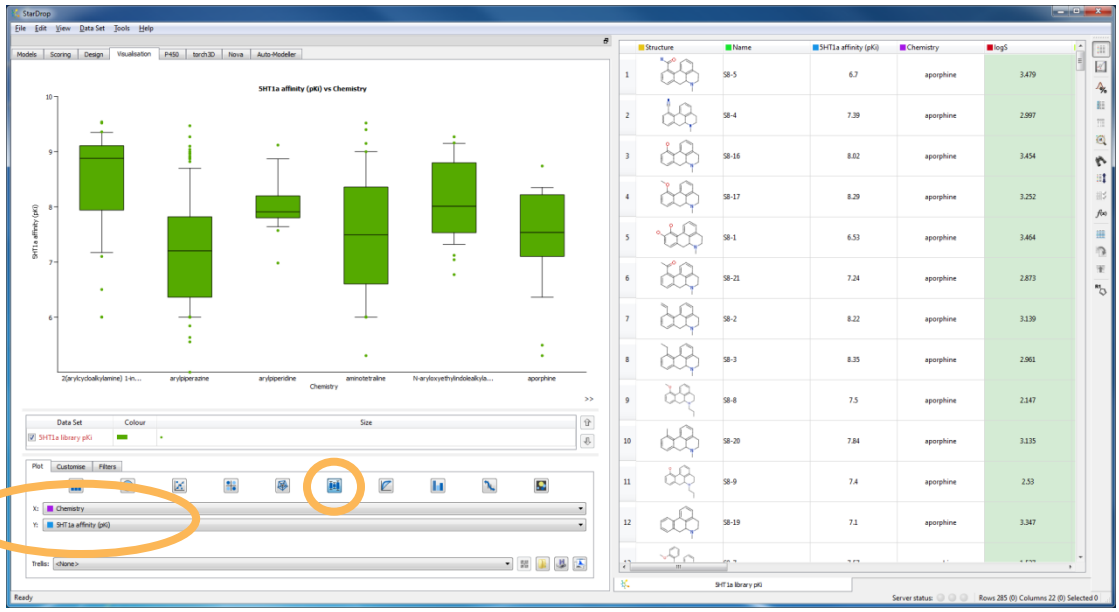
- Run all of the StarDrop ADME QSAR models and calculated properties from the **Models** tab in StarDrop by ticking them all and clicking the  button .

- Answer the following questions using the data set management and visualisation tools in StarDrop:

1. Which chemistry has the highest average potency?

   Answer:_____

   Hint: Change to the **Visualisation** tab in StarDrop and create a **box plot** by clicking the  button, with the **Chemistry** column on the x-axis and **5HT1a affinity (pK_i)** on the y-axis, as shown below:



2. Which chemistry has the highest average predicted solubility?

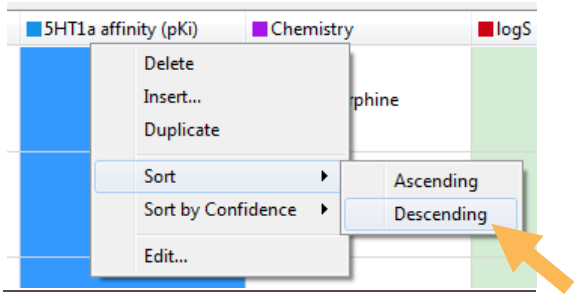
   Answer:_____

   Hint: Plot a box plot with the **Chemistry** column on the x-axis and **logS** on the y-axis.

3. What is the identifier of the most potent compound?

Answer:_____

Hint: Sort the data set by pK_i in descending order by right-clicking on the **5HT1a affinity (pKi)** column, as shown below:



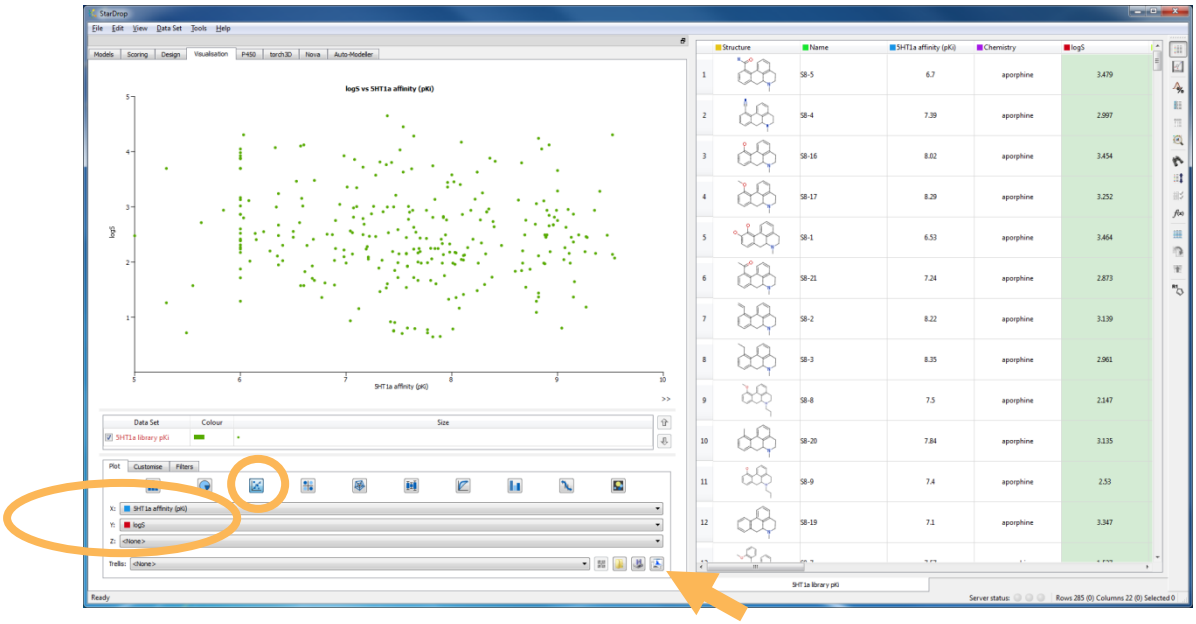
4. How do this compound's properties compare with those of compound S1-26?

Answer:_____
_____
_____

Hint: Freeze the row corresponding to the most potent compound by selecting the row and clicking on the **Freeze row** button on the toolbar ( ). Then search for compound ID "S1-26" using the **Find** tool (the button on the toolbar). When you're finished you can unfreeze the row by clicking the **Freeze row** button again.

Next we are going to use the visualisation tools to find a chemistry that meets a number of property criteria. To do this, use the following steps:

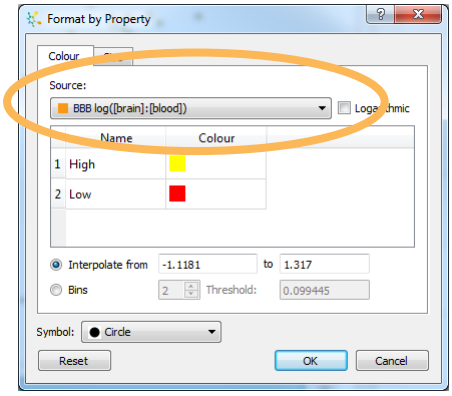
- Change to the **Visualisation** tab and plot a 2D scatter plot of **5HT1a affinity (pKi)** against **logS**, by clicking the button and selecting the properties from the **X:** and **Y:** menus, as shown below: (Alternatively you can hold down the Ctrl key while selecting these two columns in the data set)



- Click the detach button on the Visualisation tab to create a separate window containing the plot.

- In the new plot window, colour the points by **BBB log([brain]:[blood])** and size the points by **logP**.

Hint: Click on the colour block in the key to bring up the **Format by Property** dialogue box and choose **BBB log([brain]:[blood])** from the **Source:** menu, as shown right. Click on the size symbol in the key to size by **logP** in a similar way.



- Finally, change to the **Filters** tab in the plot window to filter the points in the scatter plot and remove compounds with predicted **hERG pIC50** of greater than 5, as shown below:

5. Which chemistry includes compounds with 5HT1a pK_i >7, logS > 1, log BBB penetration > -0.2, logP < 3.5 and hERG pIC50 < 5?

Answer:_____

Hint: You can find a compound in the data set by clicking on the point in the scatter plot. Alternatively, select multiple points by drawing around the points while holding down the left mouse button. To create a new set containing the selected compounds, click the  button on the toolbar.

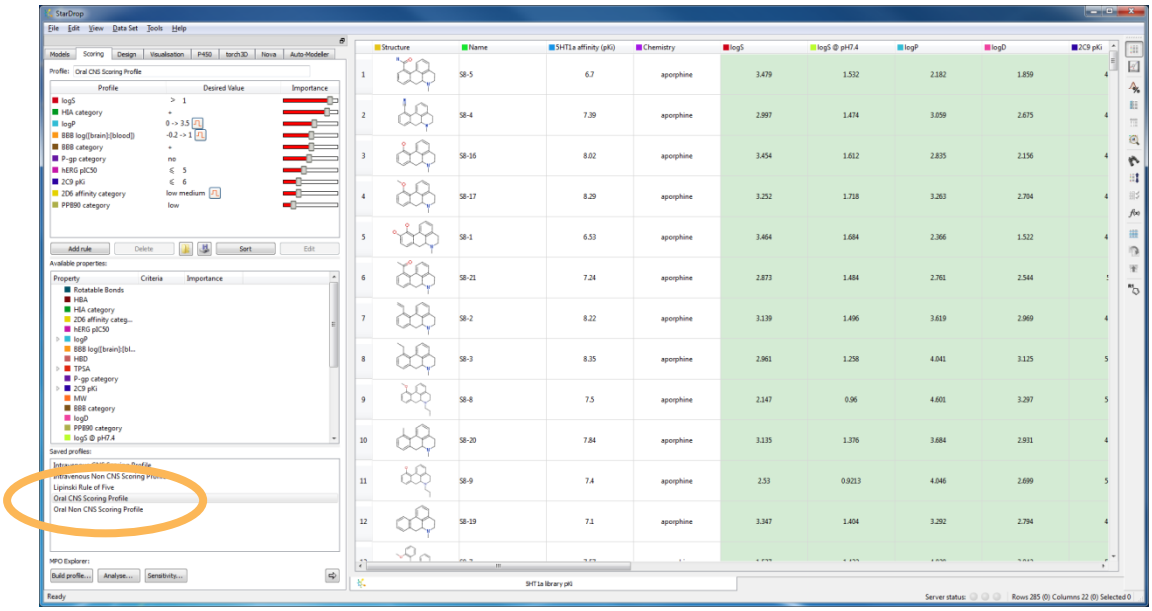
Probabilistic Scoring: Prioritising Compounds with a Balance of Properties

Objectives

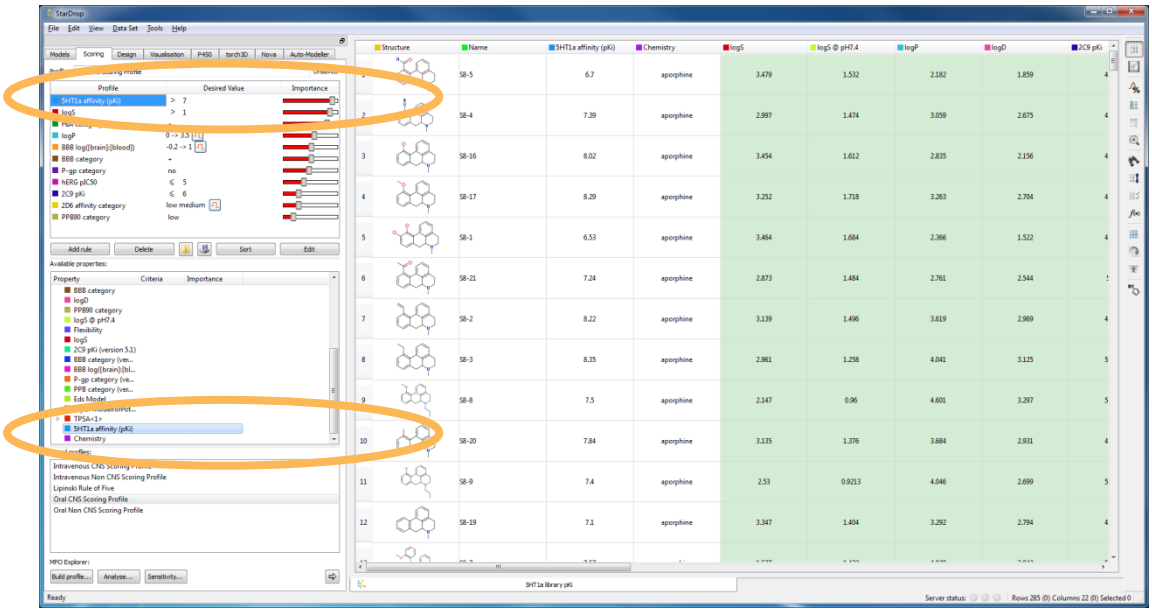
- Editing scoring profiles
- Running scoring profiles
- Interpreting scores

Exercise

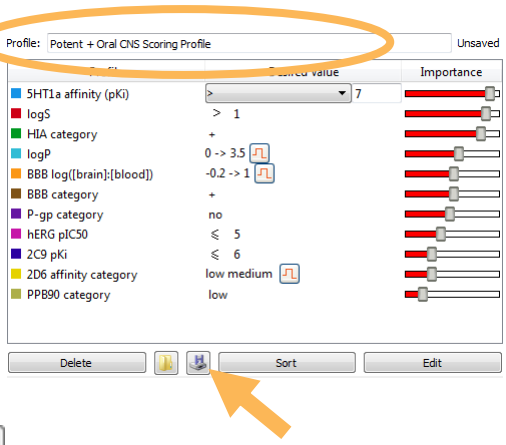
- Change to the **Scoring** tab in StarDrop and select the **Oral CNS Scoring Profile**, listed under **Saved profiles**, as shown below:



- Add **5HT1a affinity (pKi)** from the list of **Available properties** to the scoring profile by dragging the property into the profile editor. Set a **Desired Value** of >7 and an **Importance** of 0.95 for this property, as shown below:



- Give the resulting profile a name by entering it in the **Profile:** box above the scoring profile and save it in a convenient place by clicking on the  button below the scoring profile, as shown to the right. It will appear in the **Saved profiles:** list at the bottom of the **Scoring** tab, so that you can retrieve it easily.



- Run this scoring profile by clicking on the  button in the **Scoring** tab and answer the following questions:

6. Which compound has the highest score?

   Answer:_____

   Hint: Sort the data set by score in descending order.

7. What are the most critical issues that should be addressed to significantly improve the chance of success of compound S3-23?

   Answer:_____

   Hint: Use the Find tool (the  button on the toolbar) to locate compound ID S3-23.

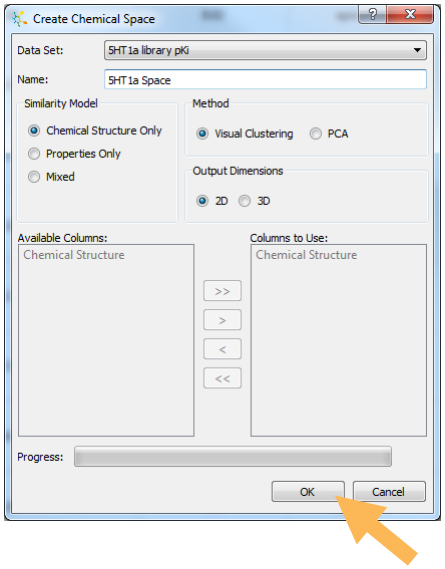
Exploring Chemical Space: Balancing Quality and Diversity

Objectives

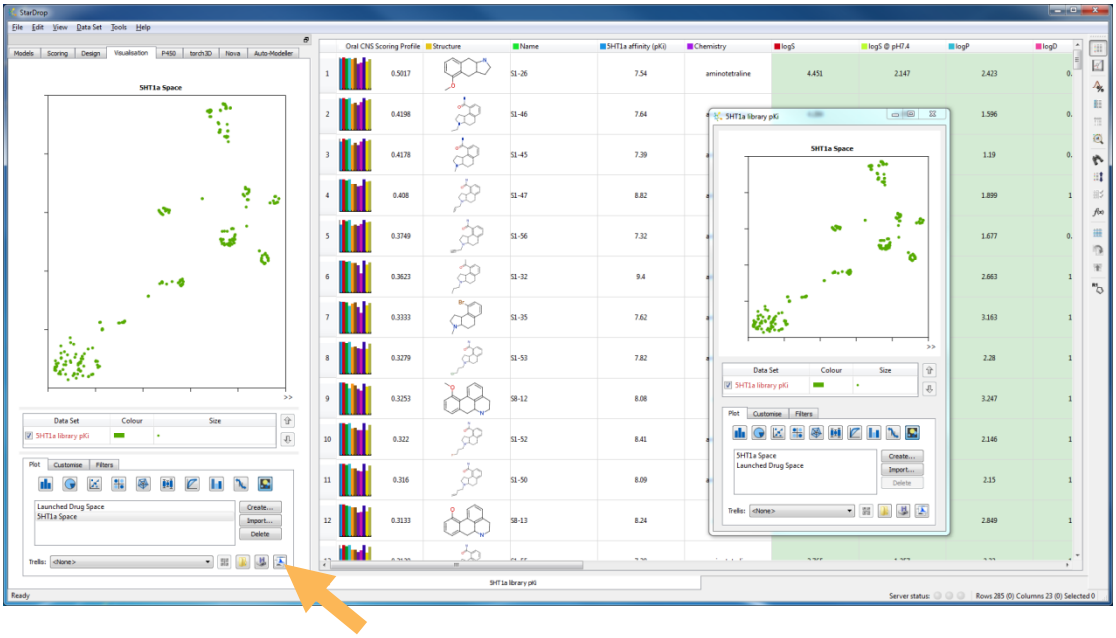
- Generating and visualising chemical spaces
- Considering uncertainty when selecting compounds
- Exploring an appropriate balance of quality and diversity

Exercise

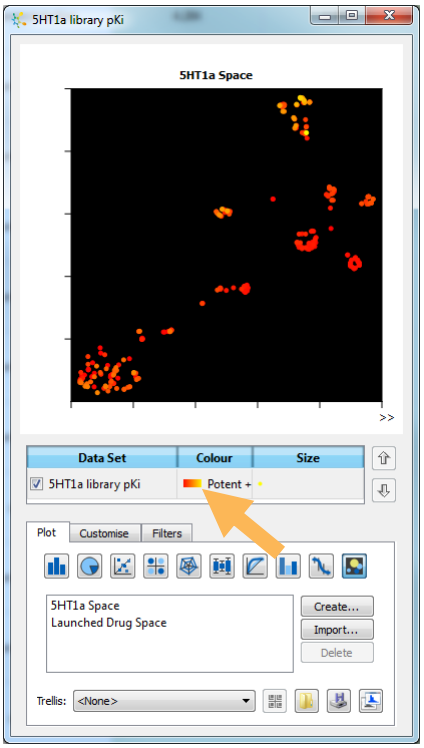
- Change to the **Visualisation** tab and select the chemical space button  . Click **Create...** to create a new chemical space. In the **Create Chemical Space** dialogue, give the new projection a name of **5HT1a Space** and then use the default settings, as shown to the right, to generate a chemical space based on chemical structure alone. Click **OK** to generate the chemical space plot.



- Click the detach button  on the Visualisation tab to create a separate window containing the chemical space plot, as shown below:



- By clicking on the colour block in the key of the detached plot, colour the points by the overall score, as shown below:

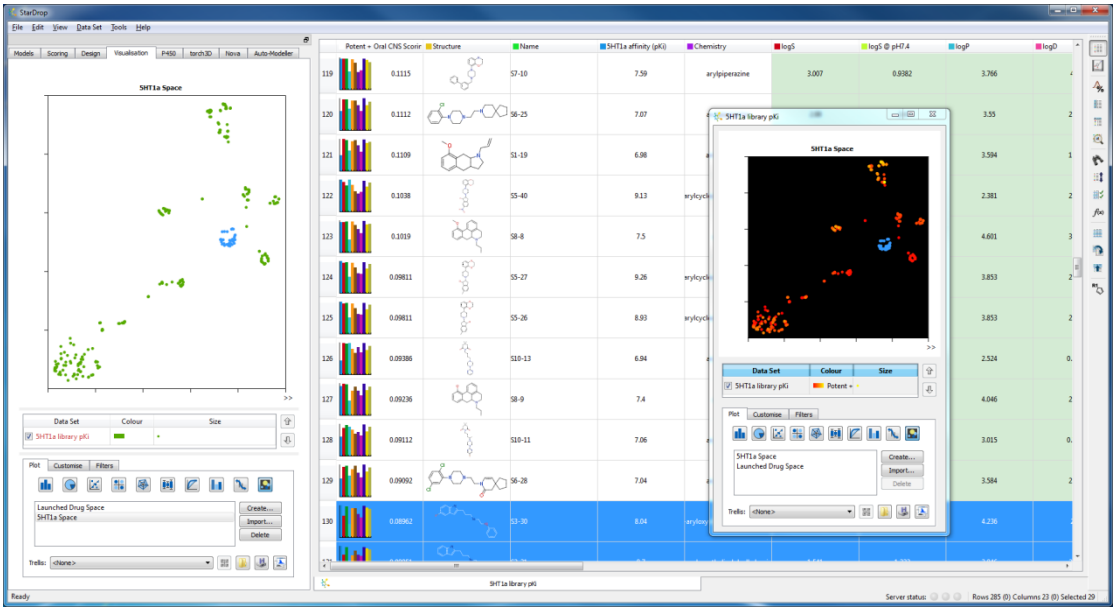


Hint: You can change the background colour on the plot by right-clicking on the plot and selecting **Change Background...**

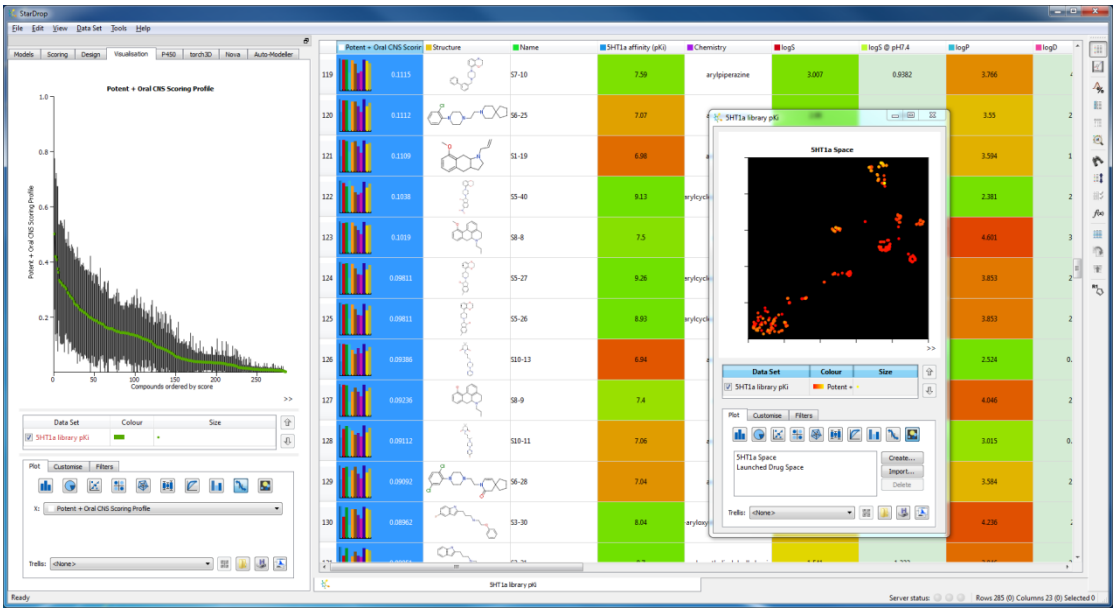
Which chemistry contains the majority of the top-10 compounds?

Answer:_____

Hint: Selecting points from the chemical space will select the corresponding compounds in the data set and vice-versa, as illustrated below (Please note, the selection shown below is not the top-10!):



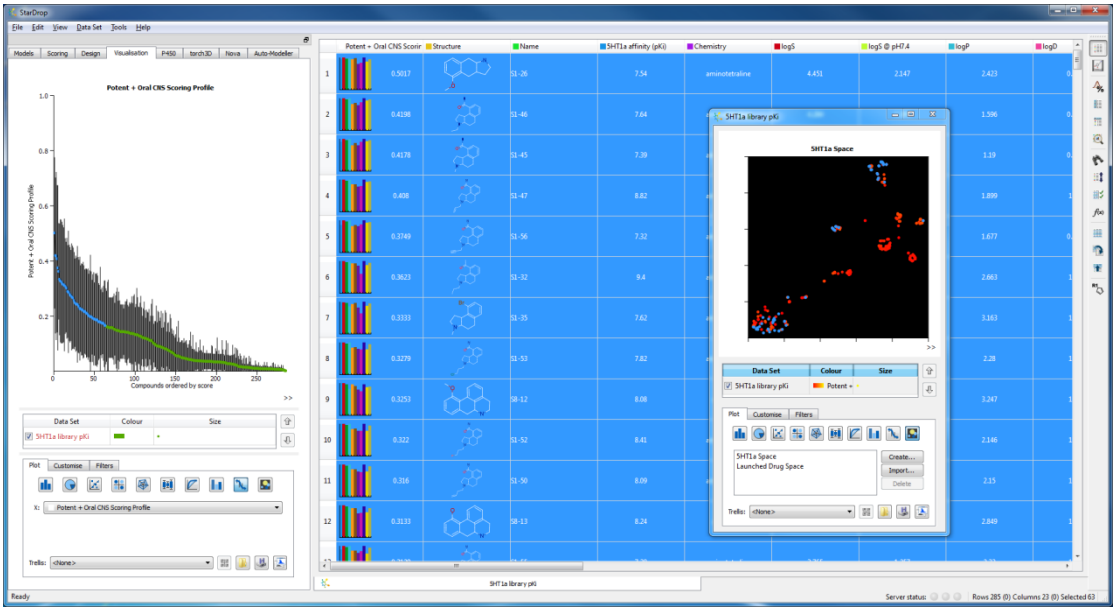
- Select the scoring column to generate a **Snake Plot** for the compound scores in this library, as shown below:



- Points in the **Snake Plot** for which the error bar overlaps with that of the first compound cannot be confidently distinguished from the highest scoring compound, based on the selected scoring profile and the uncertainty in the available data. By selecting compounds with an appropriate range of scores from the Snake Plot, as illustrated below, answer the following question:

8. Which other chemistries should we consider in the search for a high quality lead series?

Answer:_____
_____



Answers

1. Which chemistry has the highest average potency?

   Answer: 2(arylcycloalkylamine) 1-indanol

2. Which chemistry has the highest average predicted solubility?

   Answer: Aminotetraline

3. What is the identifier of the most potent compound?

   Answer: S5-34

4. How do this compound's properties compare with those of compound S1-26?

   Answer: Other than its lower potency, compound S1-26 appears to have generally more favourable properties: higher solubility, lower affinity for P450 enzymes and hERG, lower logP, not a P-gp substrate etc.

5. Which chemistry includes compounds with 5HT1a pK_i >7, logS > 1, log BBB penetration > -0.2, logP < 3.5 and hERG pIC50 < 5?

   Answer: Aminotetraline

6. Which compound has the highest score?

   Answer: S1-26

7. What are the most critical issues that should be addressed to significantly improve the chance of success of compound S3-23?

   Answer: Low solubility and high logP.

8. Which chemistry contains the majority of the top-10 compounds?

   Answer: Aminotetraline

9. Which other chemistries should we consider in the search for a high quality lead series?

   Answer: Both the aporphines and arylpiperazines have a number of promising compounds

Lead Optimisation: Guiding the Design of Balanced Compounds

One of the chemical series chosen for progression from the hit-to-lead project, explored in the previous section, was a series of Arylpiperazines (series S10). Due to their similarity with the drug Buspirone, which is known to have issues due to rapid metabolism by Cytochrome P450 CYP3A4, the compounds were initially tested for stability against this P450 isoform.

In the following example, we will explore approaches to investigate the structure-activity relationships (SAR) for these properties along with additional predicted properties. Furthermore, we will see how the Glowing Molecule visualisation can help to guide the design of compounds to overcome potential liabilities, while monitoring other properties to ensure that improvements to one property do not have a negative impact on other important factors.

SAR Analysis

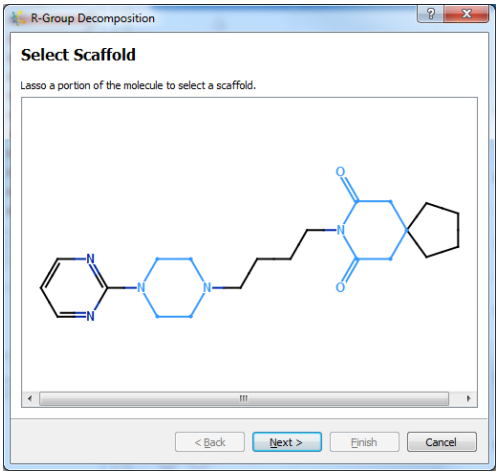
Objectives

- Generating R-group decompositions
- Visualising Structure-Activity Relationship (SAR) tables
- Multi-parameter optimisation within a chemical series

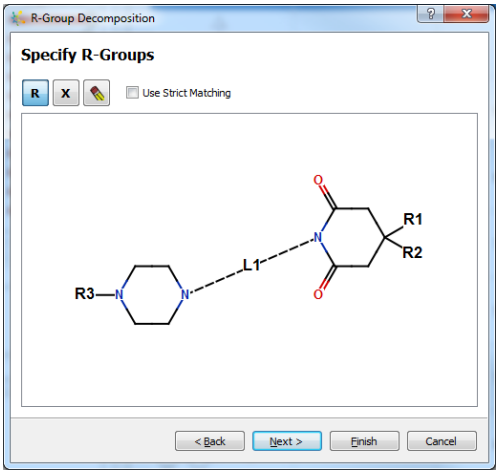
Exercise

- Open the file **Arylpiperazine series S10.add** containing 21 compounds from series S10, for which both potency (pK_i) against the 5HT1a target and half-life for metabolism by CYP3A4 have been experimentally measured.
- Select the first compound in the chemical series (Buspirone) and click the **Run R-group Decomposition** button on the toolbar (

).
- First, we will define the regions representing the common scaffold within the chemical series. Select the two ring functionalities, highlighted in blue below, by drawing around them to lasso the area and then holding down the **Shift** or **Ctrl** button to select a second region.

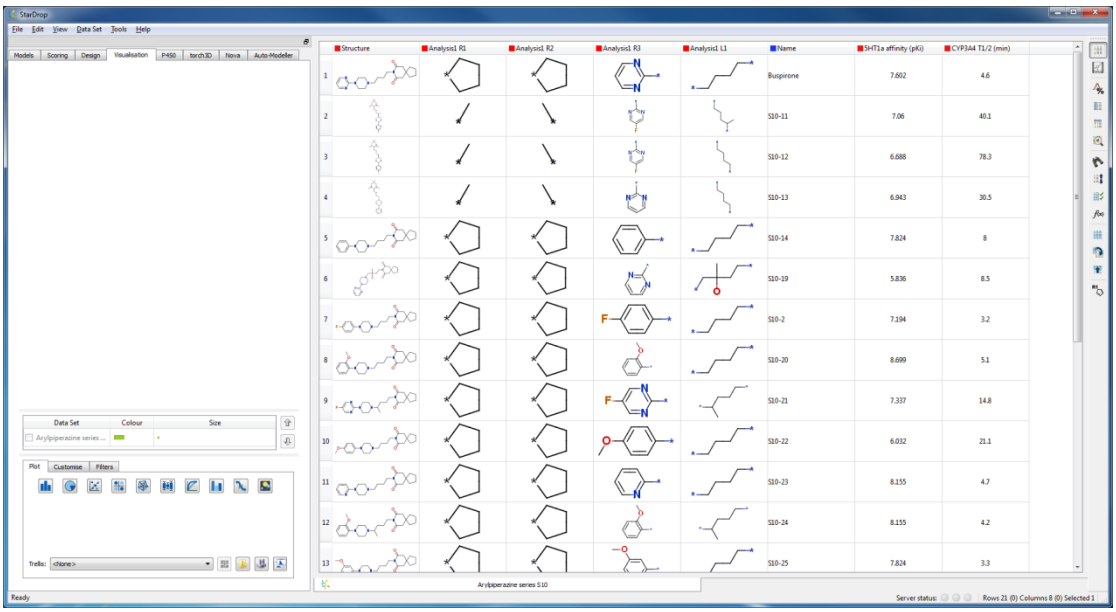


- Click **Next** to confirm the R-Groups that will be analysed. In this example, the scaffold represents two separate regions, so the connecting group will be considered to be a variable linker.

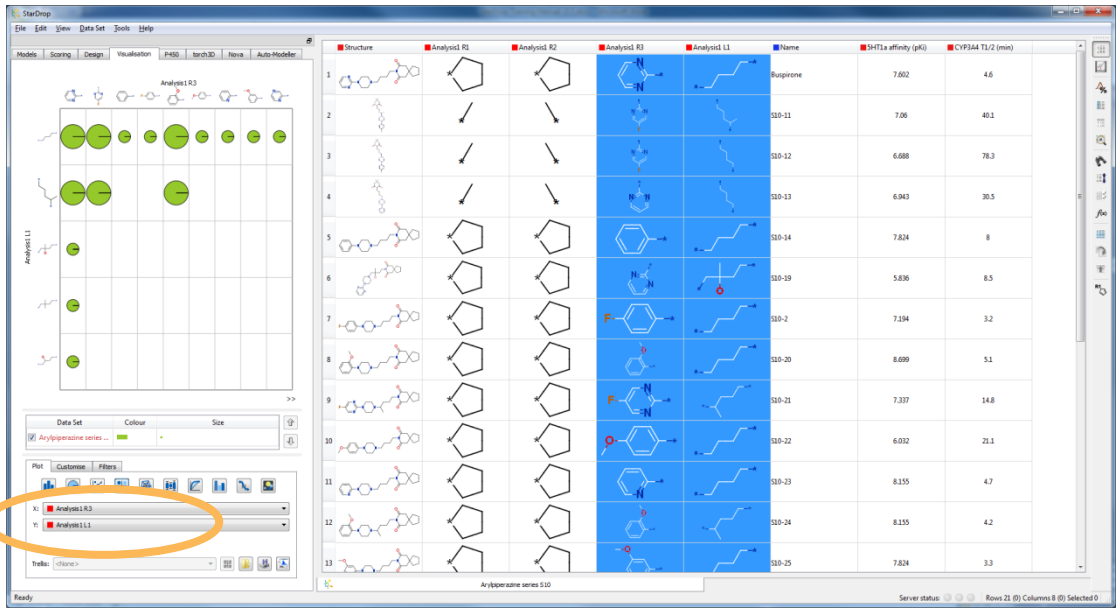


Hint: Although it is not necessary in this case, you can indicate additional R-Groups by selecting the R button and then clicking at the point on the scaffold where they should be included. If you wish to specify variable atoms or fragments within the scaffold, click the X button and then select atoms on the scaffold which may vary between the compounds in the chemical series.

- Click **Next** to enter a name for the analysis and then the **Finish** button. New columns will be added to the data set indicating the R-Groups and linkers that have been found for each compound, as shown below:

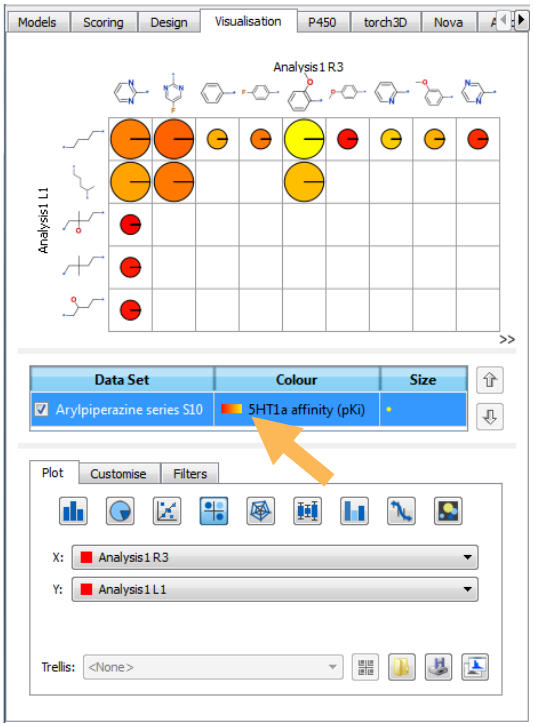


- With the **Visualisation** tab selected, plot **R3** of the SAR analysis against the linker **L1** to generate a SAR table, as shown below:



Hint: You can do this by clicking the SAR plot button  selecting R3 from the **X:** menu and L1 from the **Y:** menu in the **Visualisation** tab or simply selecting the headers of the corresponding columns in the data set (holding the **Ctrl** key to select multiple columns)

- Colour the circles in the SAR plot by the potency, **5HT1a affinity (pKi)**, by clicking on the colour bar in the key next to the **Arylpiperazine series S10** data set, as shown right . Size the circles by the stability, **CYP3A4 T1/2 (min)**.



10. Which groups at position R3 and linker L1 give the highest potency?

    Answer: _____

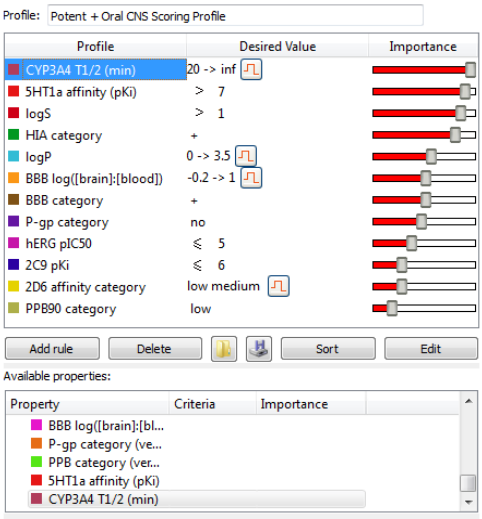
11. Which group at position R1 and linker combination give the highest stability against CYP3A4 metabolism?

    Answer: _____

- Click the detach button  to create a separate window containing the SAR plot for future comparison.

- We would like a balance of potency and metabolic stability, so we'll use Probabilistic Scoring to help us to identify the substituents that are most likely to achieve a good balance of these experimental properties (and other predicted properties).

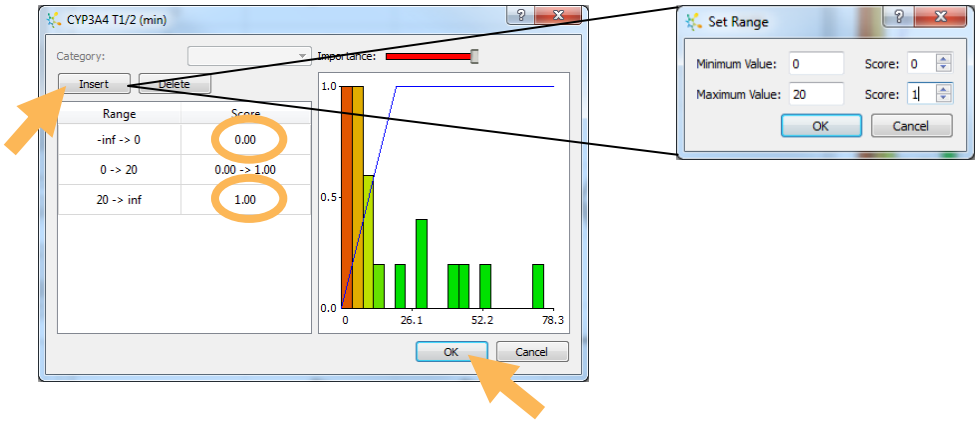
- Change to the **Scoring** tab and select the scoring profile we used earlier, in the Hit-to-Lead example. If you do not have the scoring profile in the list of **Saved Profiles:,** click the  button and open the scoring profile file called **Potent + Oral CNS Scoring profile.apd**



- Add the **CYP3A4 T1/2 (min)** property to the scoring profile by dragging it from the **Available properties:** list into the profile editor, as show to the right.

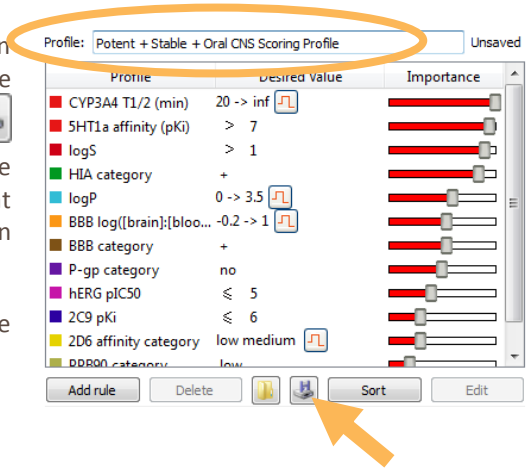
- Ideally, we would like a half-life of > 20 minutes, but we may be willing to accept a lower half-life if other properties are ideal, so we will set up scoring function with a trend for this property. Select the **CYP3A4 T1/2 (min)** property in the scoring profile and click **Edit**.

- In the scoring function editor that appears, click **Insert** to insert a range between 0 and 20, with a score of 0 for $T_{½}$ = 0 and a score of 1 for $T_{½}$ = 20, as shown below. Ensure that the score for $T_{½}$ below 0 is set to 0 and for $T_{½}$ above 20 is set to 1, as shown below. Click **OK** when you have finished.



- Give the resulting profile a name by entering it in the **Profile:** box above the scoring profile and save it in a convenient place by clicking on the  button below the scoring profile, as shown to the right. It will appear in the **Saved profiles:** list at the bottom of the **Scoring** tab, so that you can retrieve it easily.

- Run the new scoring profile by clicking on the  button.



12. Which compound is most likely to have a good balance of potency, metabolic stability and the other properties required by this project?

Answer:_____

Interactive Design and the Glowing Molecule

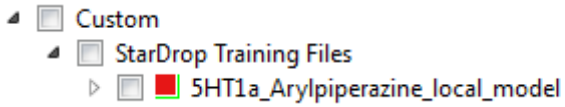
One of the potential issues identified for this chemical series is inhibition of the hERG ion channel, indicating a risk of QT prolongation and cardiotoxicity. Therefore, we will use the interactive designer, guided by the Glowing Molecule, to explore potential strategies to reduce the predicted hERG pIC_{50}.

Objectives

- Loading additional predictive models and scoring profiles
- Using the interactive designer
- Interpreting the Glowing Molecule

Exercise

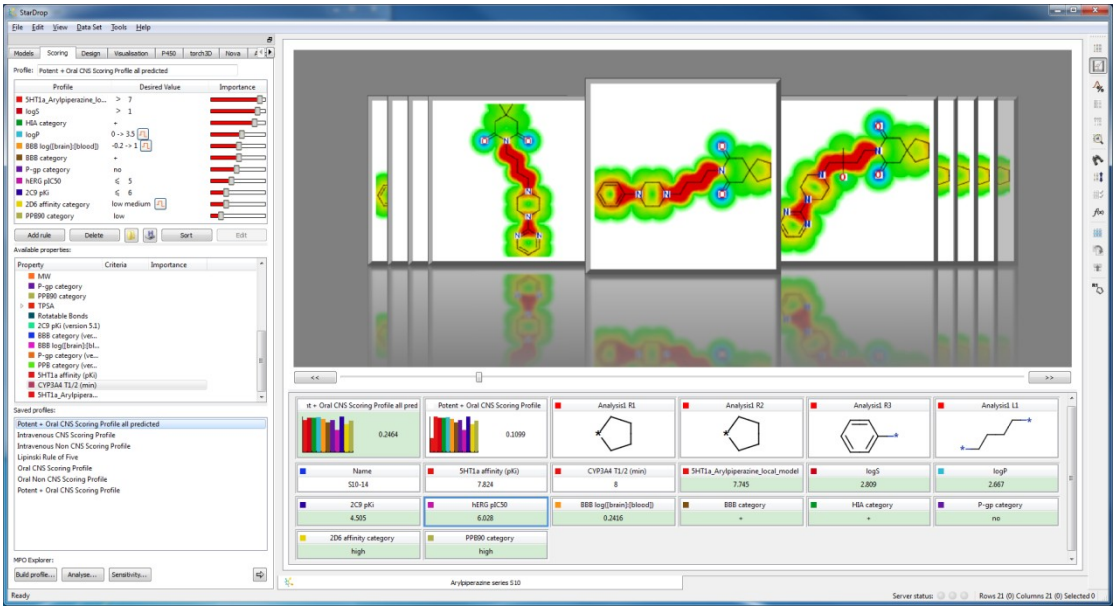
- Change to the **Models** tab, click the  button and select the model **5HT1a_pKi_Arylpiperazine_local_model.aim**. This will appear in the list of models under the branch called **StarDrop Training Files**, as shown below:



- Now we will load a new scoring profile that uses this model. Change to the **Scoring** tab, click the  button and select the scoring profile **Potent + Oral CNS Scoring Profile all predicted.apd**. This is the same as the profile we used earlier, using predicted values of the 5HT1a pK_i in place of experimental pK_i (but not CYP3A4 stability, because a model is not available for this property). We will explore strategies for optimisation of stability with respect to P450 metabolism in the P450 exercise later (if applicable).

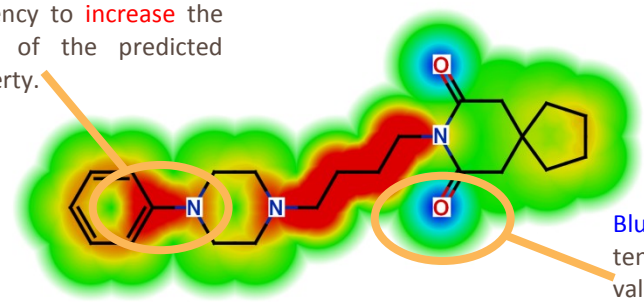
- Score the compounds using the new profile by clicking the  button.

- To help us to get an overview of each compound's properties change to the **Molecule View** by selecting the  button from the toolbar. Find compound S10-14, which is the highest-scoring compound (**Hint:** use the **Find** tool,  , on the tool bar). Click on the **hERG pIC50** model result to display the Glowing Molecule for this property, as shown below.



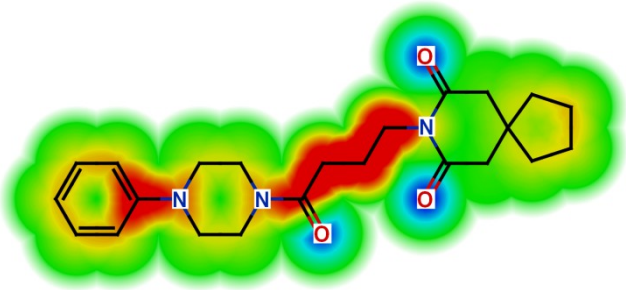
- Change to the **Design** tab to explore strategies to reduce the potential for hERG inhibition. Select the displayed structure in the molecule view and it will be displayed in the editor. When the **hERG pIC50** prediction is selected, the Glowing Molecule will also be displayed, as shown below:

Red regions have a tendency to increase the value of the predicted property.



Blue regions have a tendency to decrease the value of the predicted property.

- Edit the structure, as shown below:



13. What effect does this have on the predicted **hERG pIC50**?

    Answer: _____

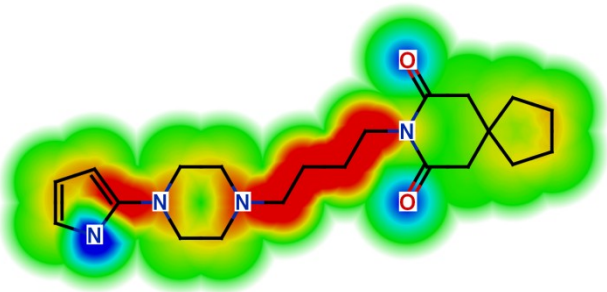
14. What effect does this have on the overall predicted score?

    Answer: _____

15. Why?

    Answer: _____

- Try the structure below instead, replacing the phenyl of compound S10-14 with a pyrole:



Hint: To quickly delete the phenyl ring in the previous compound, draw around the phenyl ring with the  tool and type **Ctrl-X**.

16. What effect does this have on the predicted **hERG pIC50**?

    Answer:_____

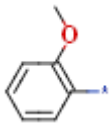
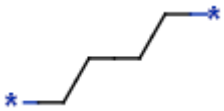
17. What effect does this have on the overall predicted score?

    Answer:_____

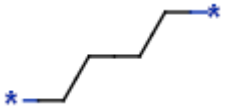
- Add this compound to the data set by clicking the  button below the editor. If you wish, you can give the compound a name by double-clicking in the **Name** cell.
- Feel free to explore some additional ideas for how to reduce the predicted hERG pIC50 without having a detrimental effect on the overall balance of properties. In a later exercise (if applicable), we will explore how the Nova module in StarDrop can help to automatically explore a large number of ideas to identify those most likely to give a good balance of properties.

Answers

10. Which groups at position R3 and linker L1 give the highest potency?

Answer:  at R3 and  as the linker give the highest potency

11. Which group at position R1 and linker combination give the highest stability against CYP3A4 metabolism?

Answer:  at R1 and  as the linker have the greatest stability

12. Which compound is most likely to have a good balance of potency, metabolic stability and the other properties required by this project?

    Answer: S10-27 has the best overall balance of properties

13. What effect does this have on the predicted **hERG pIC50**?

    Answer: The predicted hERG pIC50 decreases to 5.382

14. What effect does this have on the overall predicted score?

    Answer: Despite improving the hERG pIC50 the overall score has decreased to 0.1524

15. Why?

    Answer: The predicted blood-brain barrier penetration is now lower and this property is more important in terms of the overall score than the hERG pIC50

16. What effect does this have on the predicted **hERG pIC50**?

    Answer: The predicted hERG pIC50 decreases to 5.366

17. What effect does this have on the overall predicted score?

    Answer: The overall score increases slightly to 0.2525

P450 Exercise

Objectives

- Using the P450 models
- Interpreting P450 results

Background

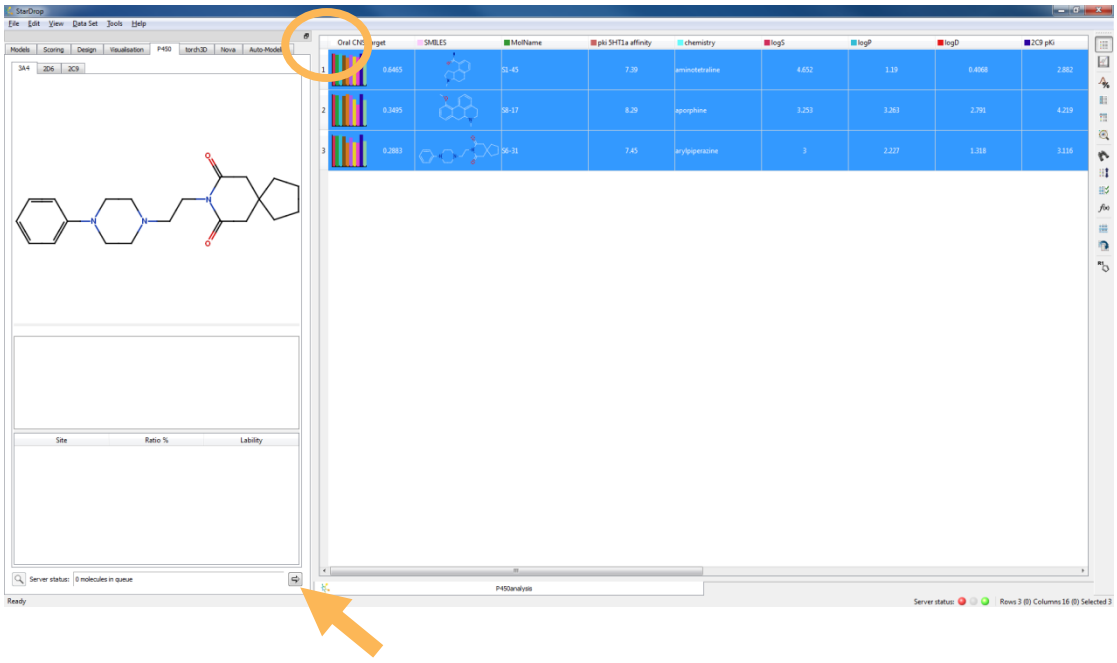
The **P450** tab provides models for the prediction of the regioselectivity and site lability of molecules to metabolism by Cytochrome P450 enzymes, CYP3A4, CYP2D6 and CYP2C9. These models provide guidance for the redesign of molecules to overcome metabolic liabilities and also to identify molecules with a high risk of rapid metabolism by Cytochrome P450 enzymes.

(**Note:** The results of these models are only relevant if the compound is a substrate for the specific isoform - these models do not give any information about whether or not compounds are substrates)

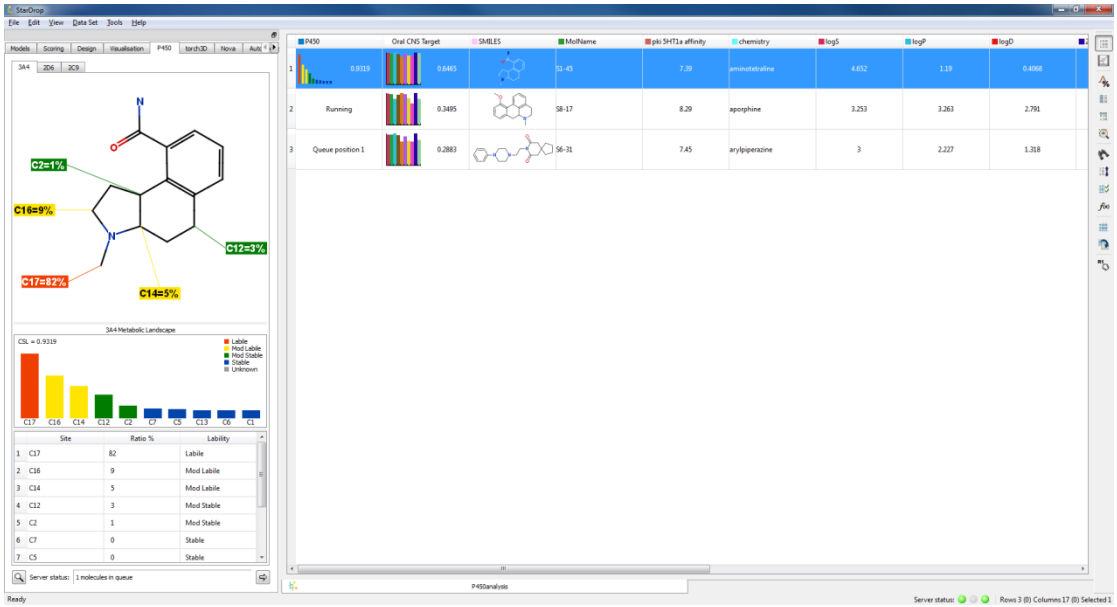
The exercise below is designed as an introduction to these tools and illustrates the effect of chemical modification within a chemical series.

Exercise

- Open the file **P450analysis.add** containing compounds S1-45, S8-17 and S6-31
- From the **P450** tab, run the P450 models on these compounds by selecting all three rows and clicking the  button. You can select a complete data set by clicking the button at the top left of the table or by pressing keys **Ctrl-a**.



- While the calculations are running an indicator will be displayed in the P450 column showing the queue position for each molecule or whether it is being calculated. Once a result is complete it will be returned and displayed in the row. Clicking on that row enables you to see the results in the P450 tab.

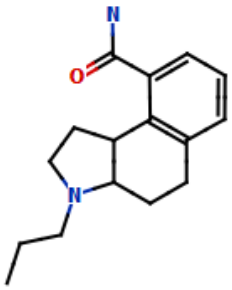
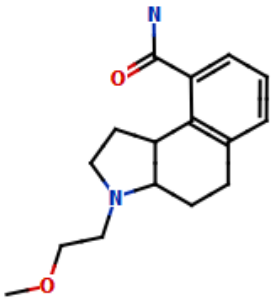
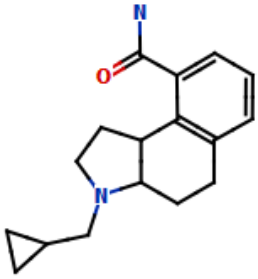


1. Complete the table for each compound inserting the CSL and the site(s) most likely to form at least 50% of products due to metabolism by CY3A4?

Compound	CSL	Likely Site(s) of Metabolism
S1-45		
S8-17		
S6-31		

- Select compound S1-45, modify it to form the following three compounds and add them to the data set:

  (**Hint:** Go to the **Design** tab, click on the row containing S1-45 and edit it to create each analogue in turn. To add the edited molecule to the data set click the  button)

Compound	Structure
Analogue 1	
Analogue 2	
Analogue 3	

2. The P450 models predict the vulnerability of each site to metabolism by CYP3A4. Which of the above compounds is most likely to show an improvement in metabolic stability over S1-45?

Answer: _____

(**Note:** If you select a compound that's already queued, the  button will change to  indicating that clicking the button again will remove the compound from the queue.)

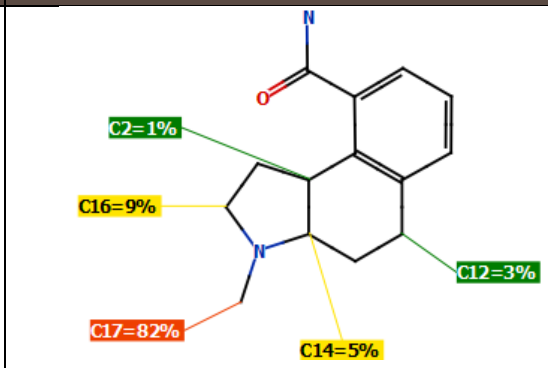
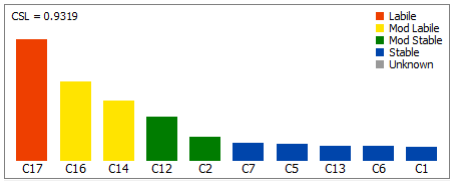
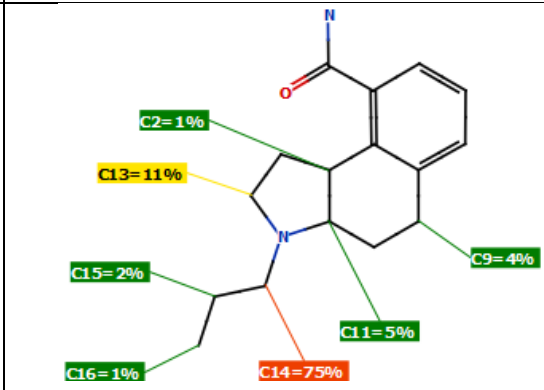
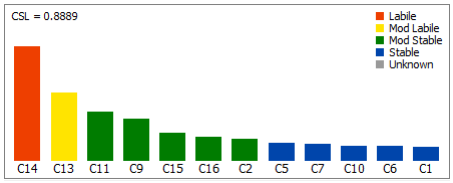
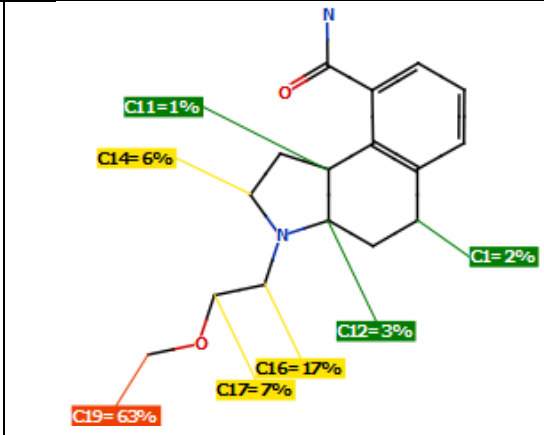
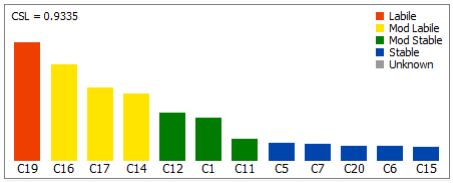
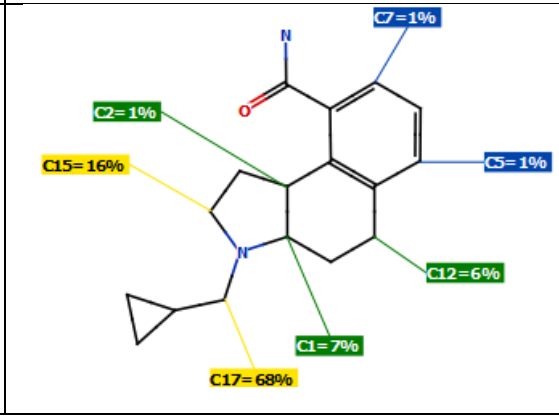
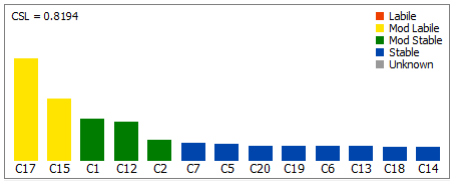
Answers

1. Complete the table for each compound inserting the CSL and the site(s) most likely to form at least 50% of products due to metabolism by CY3A4:

Compound	CSL	Likely Site(s) of Metabolism
S1-45	0.93	C17 (82%)
S8-17	0.95	C18 (38%), C16 and C8 (22%)
S6-31	0.83	C7 (35%), C21 and C22 (16%)

2. Which of the above compounds is most likely to show an improvement in metabolic stability over S1-45?

   Answer: The CYP3A4 regioselectivity and metabolic landscapes for S1-45 and modified analogues are as follows:

Compound	3A4 Regioselectivity	Metabolic Landscape
S-145		
Analogue 1		
Analogue 2		
Analogue 3		

This suggests that Analogue 3 is likely to be metabolised the least if it is a substrate for 3A4. It is the only compound with no labile sites and has a lower CSL.

Auto-Modeller Exercise

Objectives

- Using the Auto-Modeller wizard to build models
- Changing the Auto-Modeller settings
- Building continuous models
- Building category models
- Viewing, saving and using new models

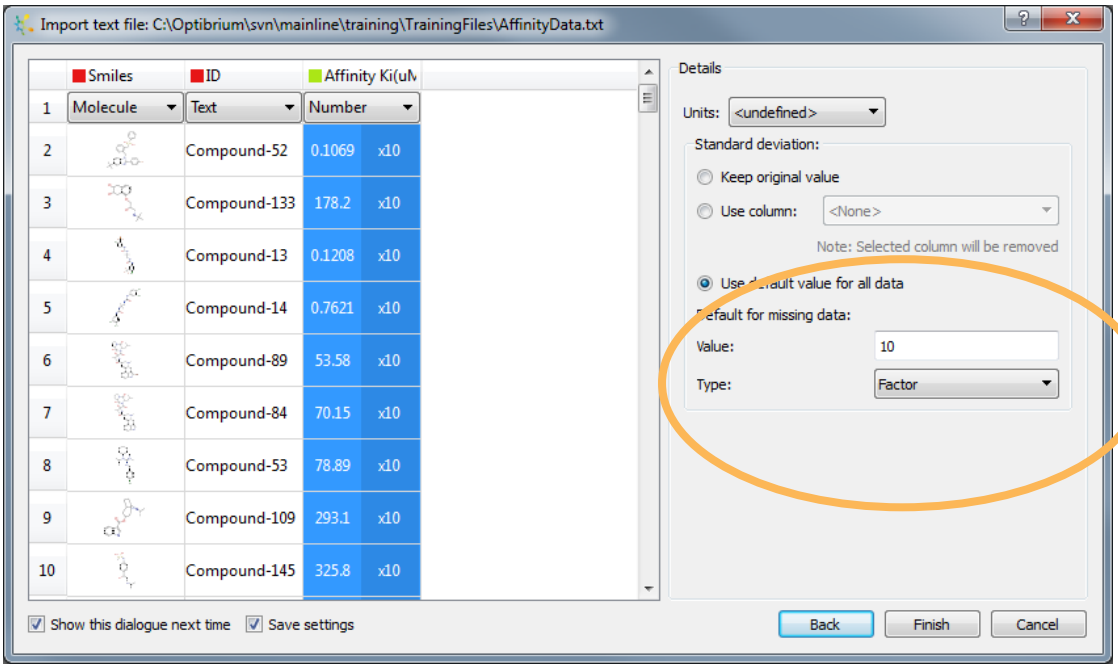
Background

Your project has measured some affinity data for the target you are working on. The compounds already synthesised and measured show a spread of affinities across a number of different chemotypes. In order to use this for decision making about new compounds to be synthesised it is necessary to build a model of this data that can be used alongside the ADME models. Your project has recently measured the affinity for the target for a small number of further compounds and these will be available to test the models built with the original set of data.

This exercise uses the **Auto-Modeller** tab and the **Mathematical Function Editor** in StarDrop.

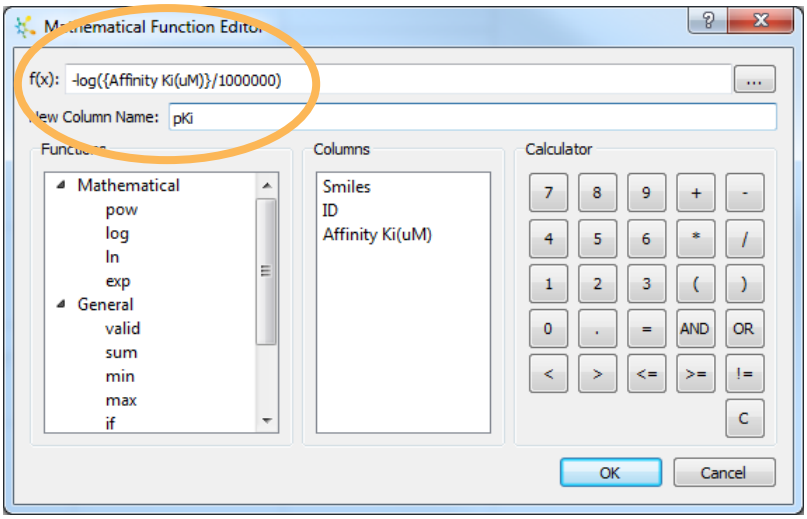
Exercise

- Open the file **AffinityData.txt**, click **Next** to confirm that it is a tab-delimited file.
- Select the Affinity Ki(uM) column, tick the **Use default value for all data** option and specify the uncertainty value to be a **factor** of **10**.

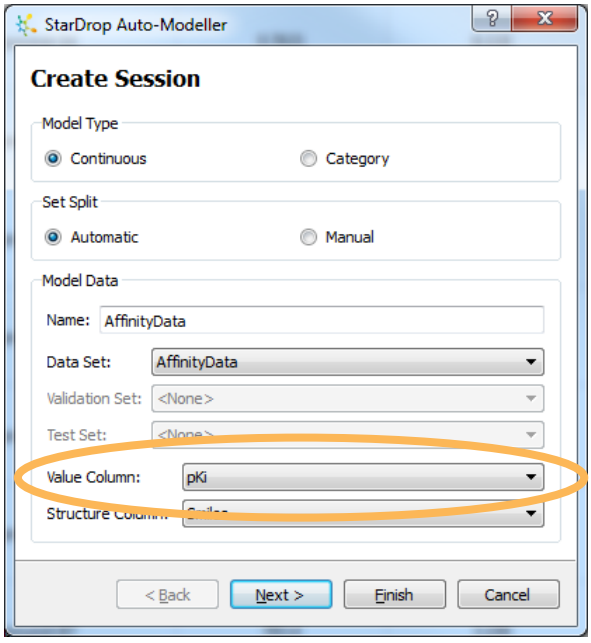


- Open the **Mathematical Function Editor** by clicking the  button and create a function to convert the Ki values into pKis in a new column called **pKi**. The Ki values are uM and so the function is:

  -log({Affinity Ki(uM)}/1000000)

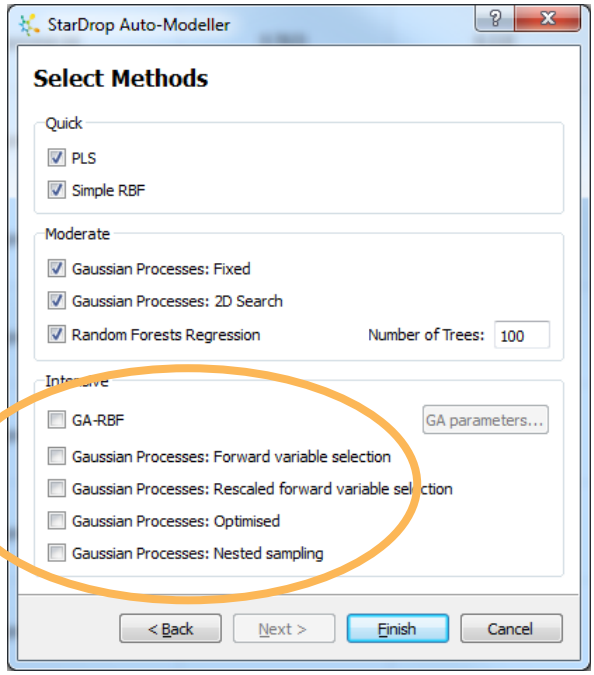


- Start the Auto-Modeller wizard by clicking the  button.
- On the first wizard page, accept all the default settings but change the **Value Column:** to be **pKi**.

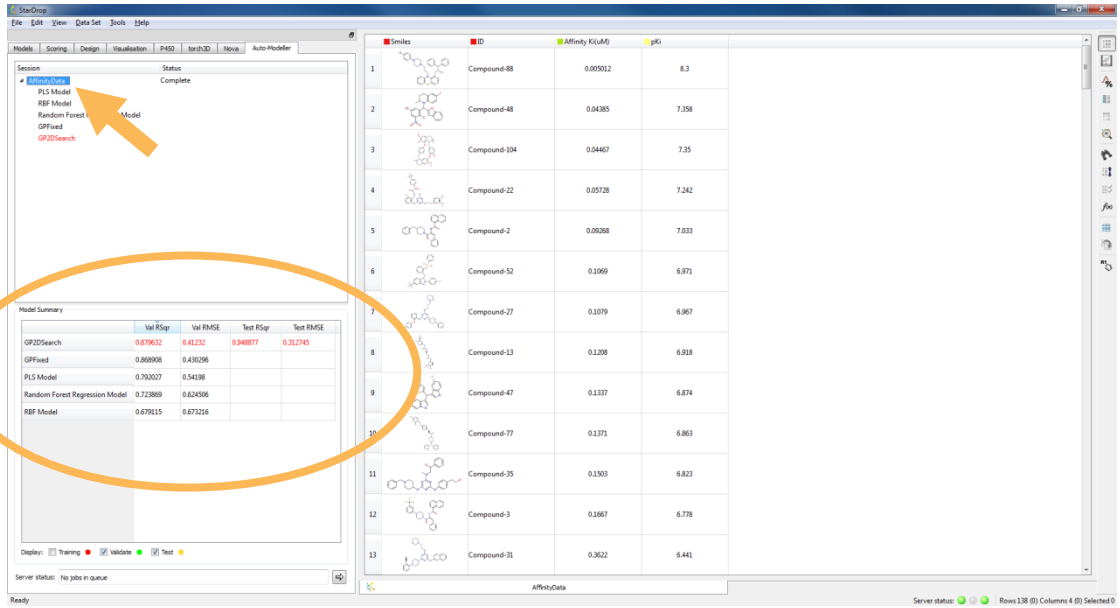


- Click **Next** until you reach the **Select Methods** page. Here, untick all the Intensive methods.

  (In practise you would normally use all these methods for a data set of this size, but for this example we will just use the quicker methods).



- Click **Finish** to start the model building process.
- Once the process has completed, select the session to see statistics for all of the models.



Hint: To see a graph of the model results for an individual model, select the model in the list.

1. Complete the following table and then comment on the differences:

Model	Training		Validation		Test	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
PLS						
GP Fixed						
GP 2D Search						

RBF						
Random Forests						

Comments:_____

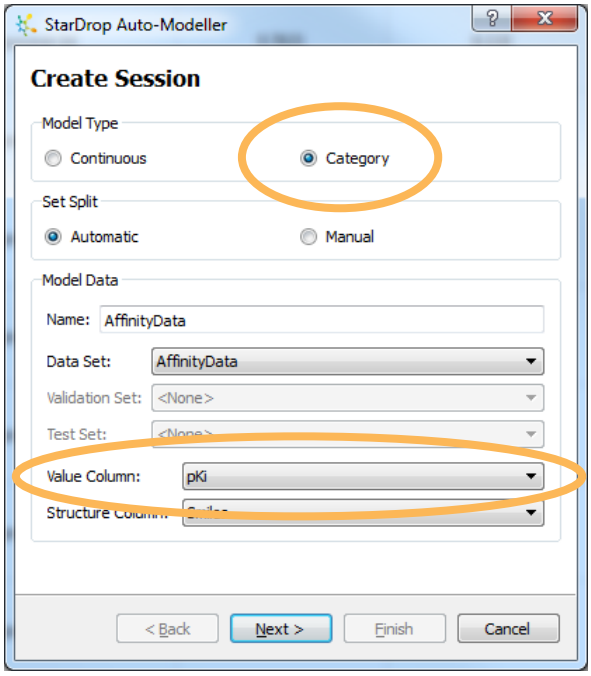
_____

_____

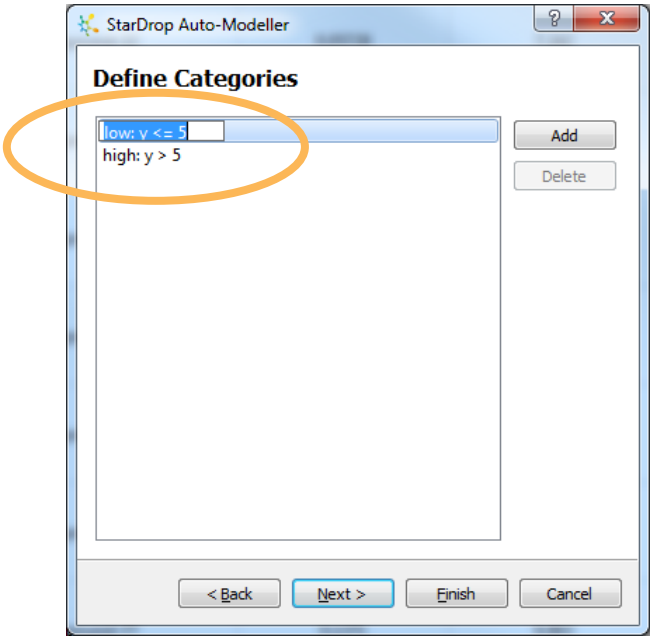
_____

(**Hint:** To see the Test results for all models ensure **Test best model only** is not ticked in the **Auto-Modeller preferences**)

- Start the Auto-Modeller wizard again. Change the **Model Type** to **Category** and once again change the **Value Column:** to be **pKi**.



- The input data (pKi) is numerical and so this time when you click Next, you'll be prompted to define categories. The default valuies will show the average value as the cutoff but if you double-click on this you can change the value to 5.



- Click **Finish** to start the model building process.

2. Which is the best model and how does it compare with the others?

Answer:_____

_____

_____

_____

- Save the best continuous and classification models by right-clicking on them and choosing Save Model... from the menu.
- Open the file **AffinityData2.txt**
- From the Models tab, run the best continuous and the best category model against this set of compounds. (**Hint:** the saved models will appear in the Custom section).

Comment on how well these models performed and which might be most useful when making decisions about which compounds to synthesise in the future: (Don't forget to apply the same function to convert the K_i values into pK_is before you compare them with the models.)

Comments:_____

_____

_____

_____

_____

Summary & Answers

1. Complete the table and then comment on the differences between the models:

Comments: The set selection process introduces an element of randomness when splitting the original data into training, validation and test sets. However, looking at the statistics for the three models, all are likely to be reasonable and produce similar results.

2. Which is the best model and how does it compare with the others?

   Answer: Again, without using pre-split data, the model results can vary. However it is likely that one or two of the models correctly categorize the majority of the data. It is probable that some of the models produce identical results and as such would be equally useful.

3. Comment on how well these models performed and which might be most useful when making decisions about which compounds to synthesize in the future:

   Comments: Both the best continuous and the best classification model produce good results and give good predictions for the additional data. For use within a project, it would be advisable to use the continuous model because this model is accurate enough to give a better indicator of affinity than the classification model. Therefore, if included in a scoring profile alongside ADME parameters the continuous model will be more helpful in differentiating between compounds when prioritising and selecting.

Nova Exercise

Objectives

- Using Nova to generate new compound ideas
- Choosing preferences to control the way new molecules are generated
- Exploring Nova results
- Loading custom models

Background

Company X has found a lead compound which they would like to try and evolve into a candidate. The compound has a good profile of ADME properties but insufficient inhibition of the target (Serotonin transporter). Your task is to see if you can generate some new ideas for compounds which can improve the potency while maintaining the balance of other properties.

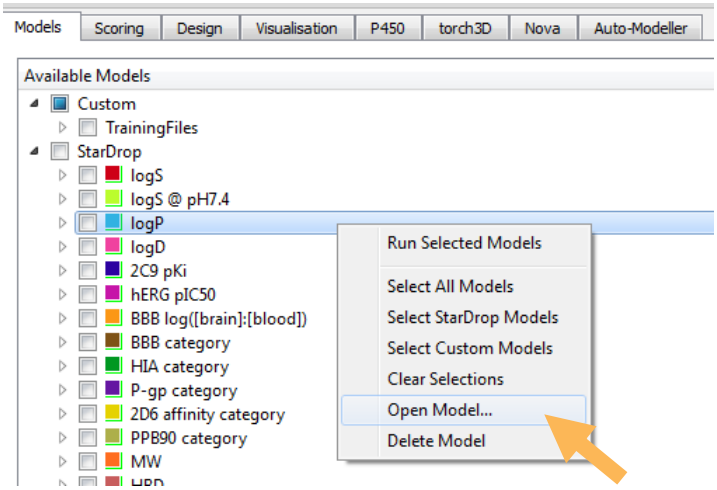
We will use a model for Serotonin transporter inhibition built with public domain data from the ChEMBL database using StarDrop's Auto-Modeller to monitor potency during the exploration.

Nova is capable of generating data sets of many hundreds of thousands of compounds if left to run for many generations and so for this exercise we're going to take a look at how we can manage this.

This exercise uses the **Models** and **Nova** tab in StarDrop.

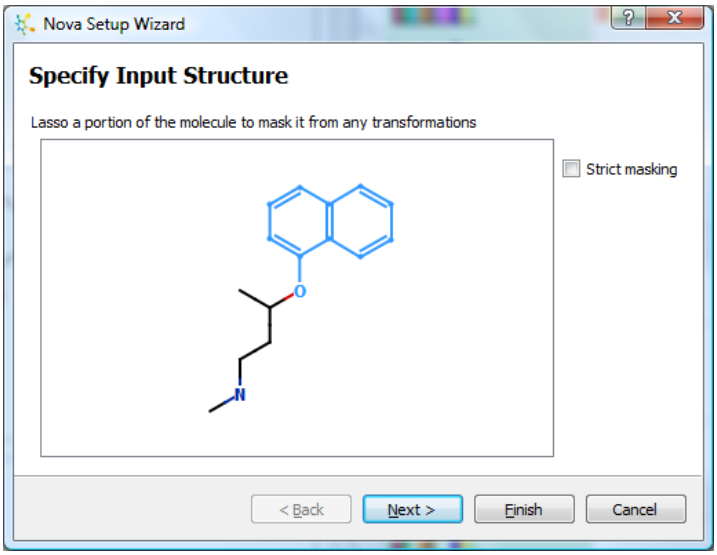
Exercise

- On the **Models** tab, right-click over the models and select **Open Model…** from the menu and then open the model file **Serotonin_Transporter_logKi.aim**

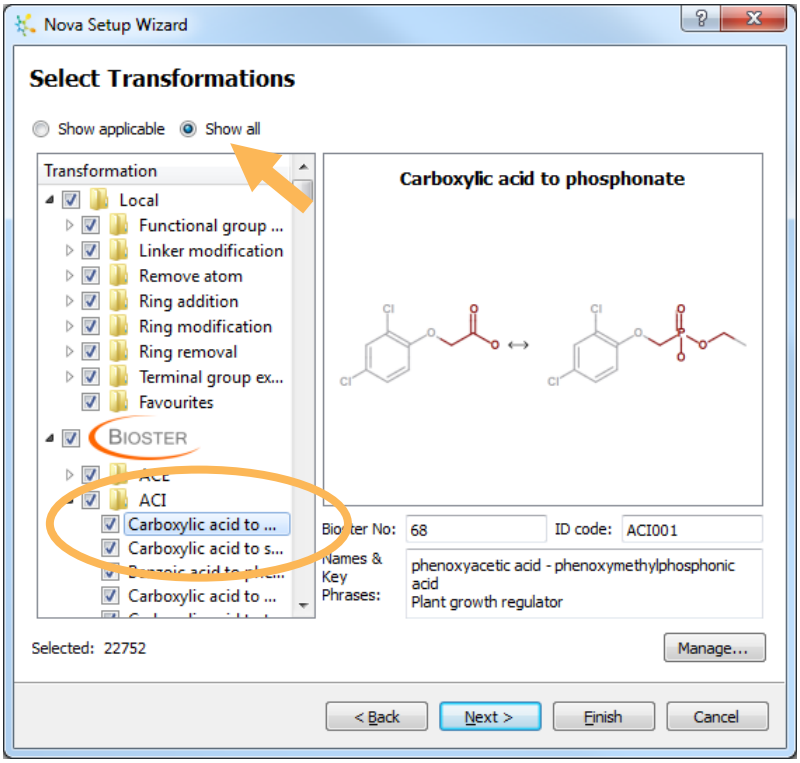


- The model will appear in the **Custom** model section.
- Open the data file **Nova lead compound.add**
- Select the only row in the data set and on the **Nova** tab, click the  button to start the Nova wizard. (**Hint:** like the P450 models, when you start Nova it only applies to those rows which are selected).
- The Nova wizard can be used for idea generation and library enumeration. In this example we are going to select **Nova Ideas Generation**, the click **Next**.

- On the **Specify Input Structure** page of the wizard, select the napthol group, by drawing around it, to ensure that this is not modified during the process.

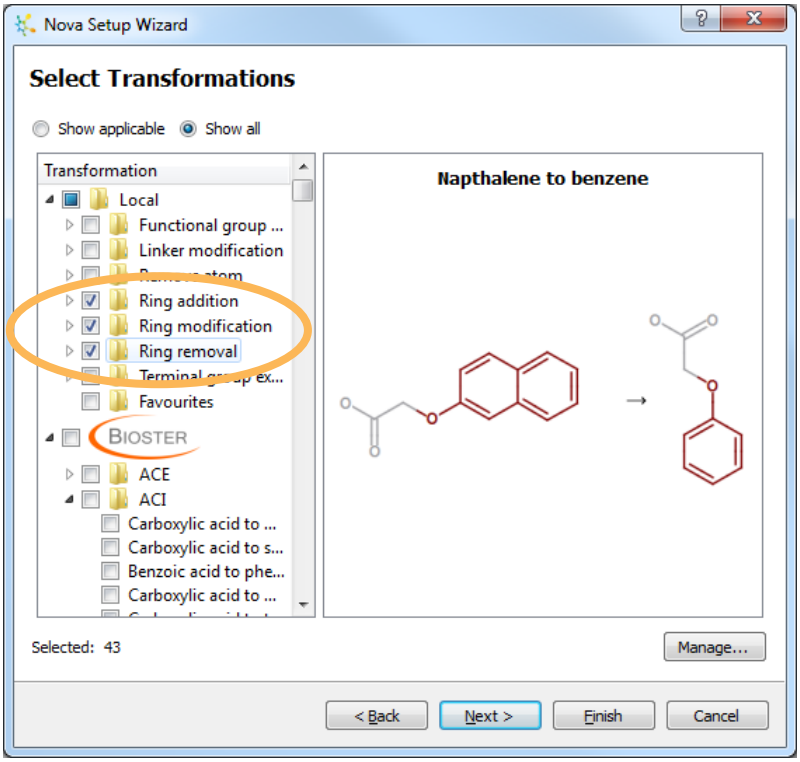


- Click **Next** to go to the **Select Transformations** page. Nova will search for, and display, only those transformations which are applicable to the input structure. However, if you are going to allow Nova to run for multiple generations (as we will in this example) it is sometimes useful to select additional transformations. These can not be applied to the first molecule, but may be applicable in subsequent generations. Click **Show all** to display the complete list of transformations



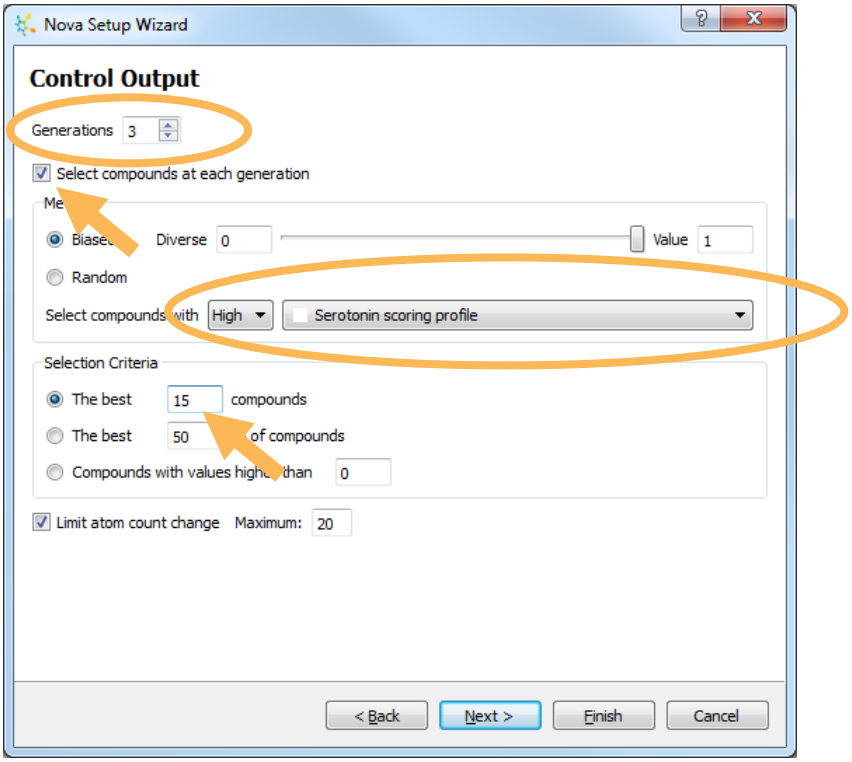
(**Hint:** selecting an individual transformation will enable you to see an example and any other available details).

- For this example, we are going to limit Nova to a small number of transformations. Select just the following groups: **Ring addition**, **Ring modification** and **Ring removal**.

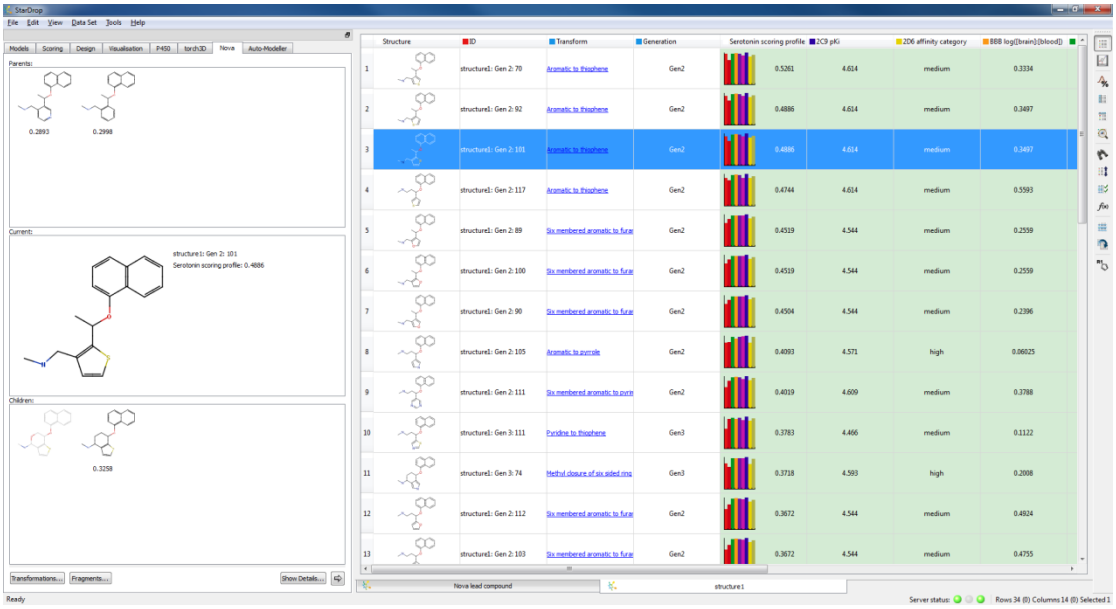


- Click Next to go to the **Control Output** page. Change the number of **Generations** to **3** and tick the **Compound Selection** box. Choose to select compounds with **High Serotonin scoring profile**. Choose a **biased** selection with a weight of 1 on Value (in this example we will not search for diverse solutions). From each generation select **The best 15 compounds**.

  (**Hint:** You don't have to specify any criteria when running Nova. However, beware that with over 200 possible transformations, running it for three generations without any limiting criteria can produce data sets with over 1,000,000 compounds – which could take a little while.)



- Start the process by clicking **Finish**. The Nova job will take a couple of minutes to run…
- While in progress an indicator will show the highest score that has been achieved so far. Once complete, a new data set will be displayed.



- Selecting a row will result in the compound being displayed within the Nova tab, along with its parent compounds and any compounds generated from it. Any compounds which are filtered out will be displayed faded-out to indicate that they were created but are not present in the final data set. (**Hint:** the penultimate wizard page, not used in this example, enables you to choose which filters, if any, to apply).

1. What are the main differences between the compound with the highest score and the initial lead compound?

   Answer:_____

   _____

   _____

   _____

   _____

2. There is a major drug in the list of newly generated compounds. Can you spot it?

Guidance
Summary & Answers

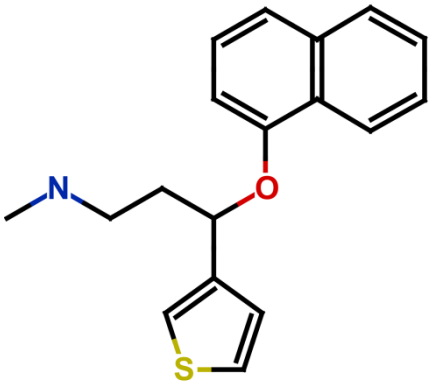
- Nova will automatically generate new compound ideas from an initial starting molecule
- Lassoing a portion of a molecule will mean that this sub-structure will be present in all results. **Note:** this relates only to the heavy atoms.
- You can choose which transformations to apply, creating favourite lists for easy access or importing and reorganising the transformations to match your chemistry
- You can limit the number of results generated and bias it towards individual properties or scores (including custom models)

1. What are the main differences between the compound with the highest score and the initial lead compound?

Answer: The Glowing Molecule shows that the additional heteroatom present in the thiophene and pyrrole groups in the top scoring compounds has a positive effect on the Ki, significantly increasing the overall score.

2. There is a major drug in the list of newly generated compounds. Can you spot it?

   Answer: The fourth best molecule produced is actually the drug Duloxetine.



torch3D Exercise

Objectives

- Comparing the 3D fields generated by compounds to identify key interactions
- Screening new compounds against a known bioactive to look for novel potent compounds and identify opportunities for optimisation

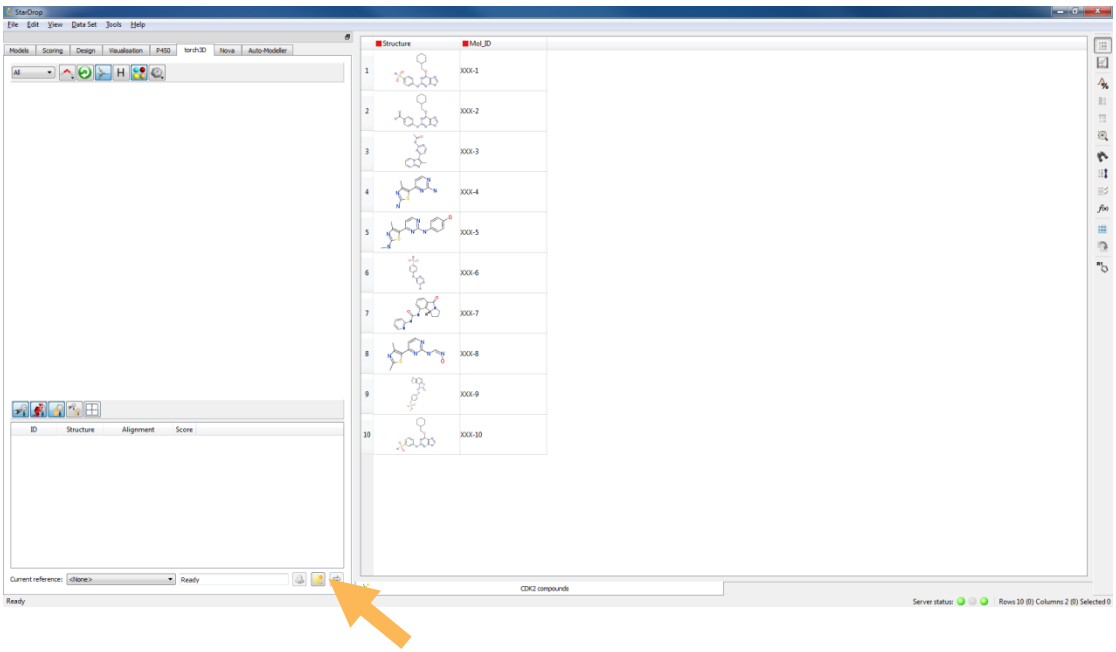
Background

You have a 3D structure of a bioactive compound from your project, derived from a crystal structure of the ligand bound to CDK2, and a number of other active compounds. You would like to compare these active compounds to understand the underlying SAR and identify opportunities for further optimisation.

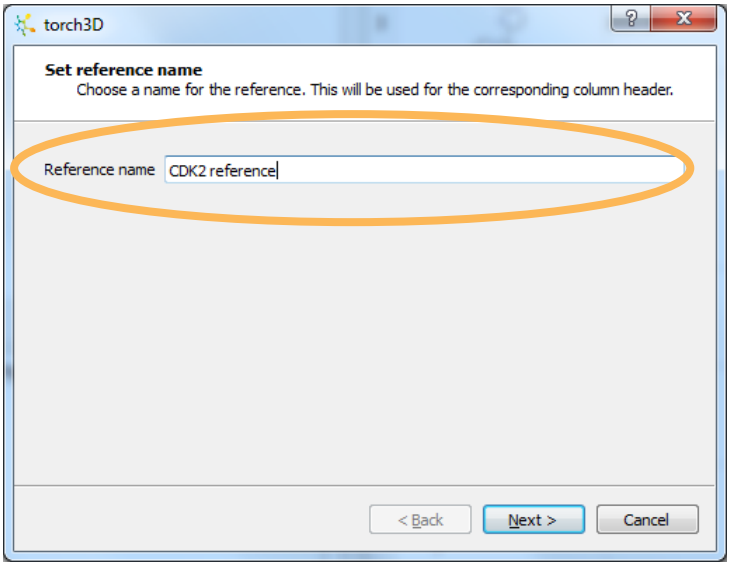
In this example we will use the torch3D molecule to study 3D Field alignments with the known bioactive structure. torch3D generates multiple energy-minimised conformations and compares their 3D shape and field patterns to find the best alignments to the reference structure. For each molecule the top-10 scoring alignments are calculated for detailed analysis and visualisation. A high Field Score indicates a better match of an alignment with the reference structure. It is not a guarantee that it will be potent, but helps to prioritise those compounds for more detailed investigation.

Exercise

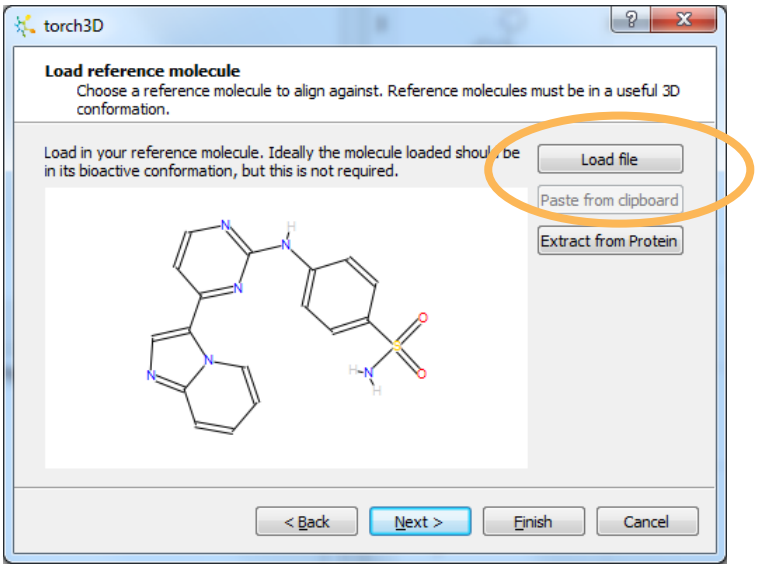
- Open the data set **CDK2 compounds.add**
- On the **torch3D** tab select the  button to start the torch3D wizard and define a new reference against which the new compounds will be compared.



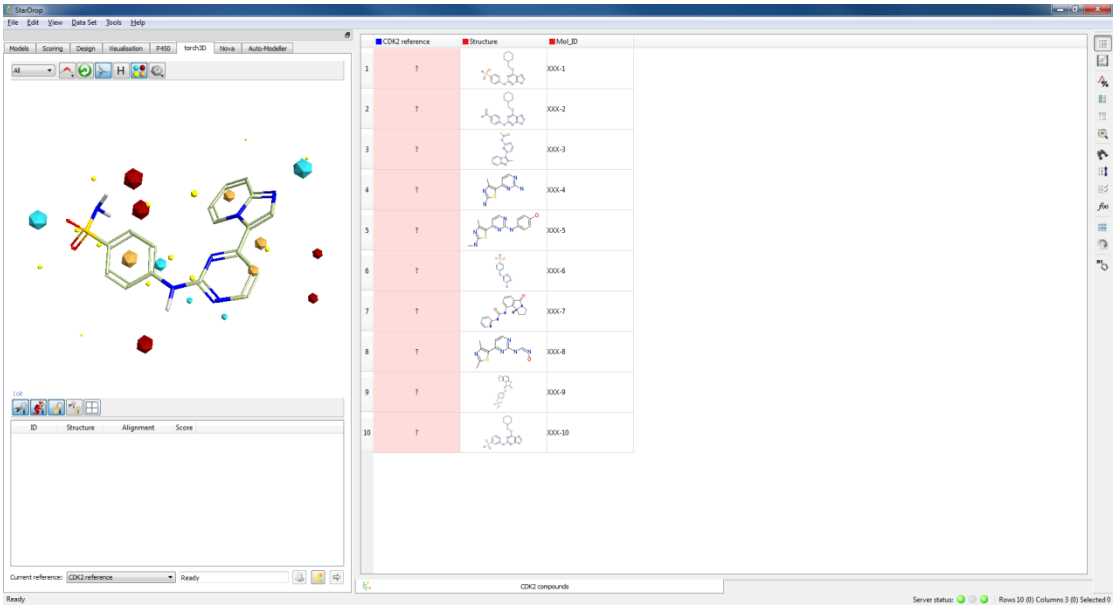
- On the first page of the Wizard, type in the name **CDK2 Reference** (this will be used as the column name in the data set where the results are displayed).



- Click the **Next** button and then load the reference molecule by clicking the **Load file** button and selecting the attached file **CDK2 bioactive.sdf**. The loaded structure will be displayed.

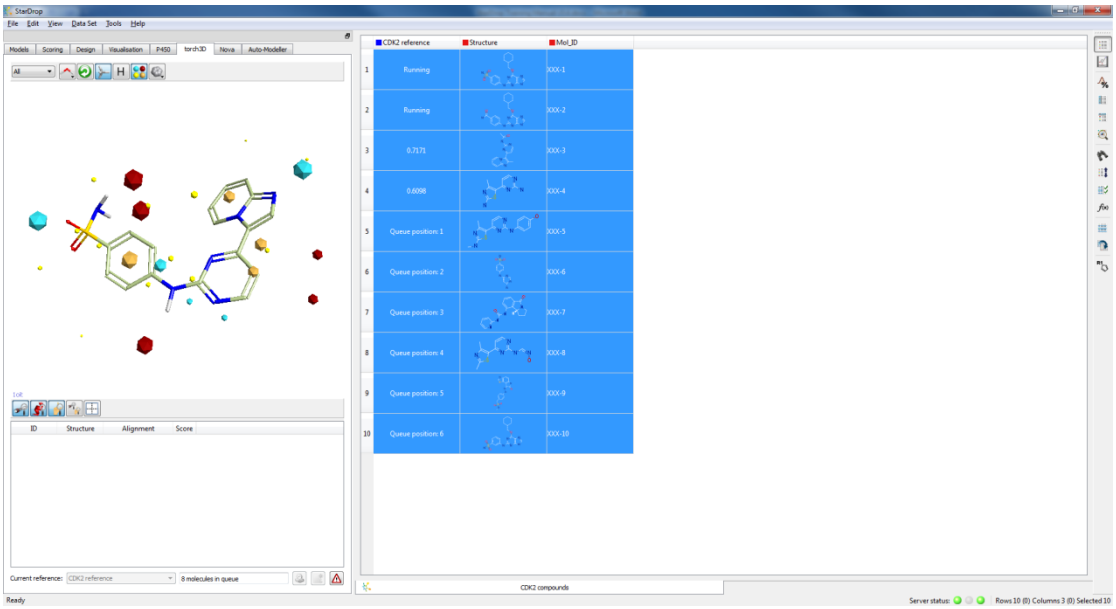


- Click **Next** to (optionally) load a protein structure to define excluded volume - in this example we will omit this step.
- Click **Next** to specify the speed of the calculations – in this case, in the interest of time, select **Fast** from the drop down menu and then click the **Finish** button
- The reference molecule will be loaded in 3D showing its field points and a new column will be added to the data set.



Now the reference has been set up we can carry out a field comparison against any compounds in the data set

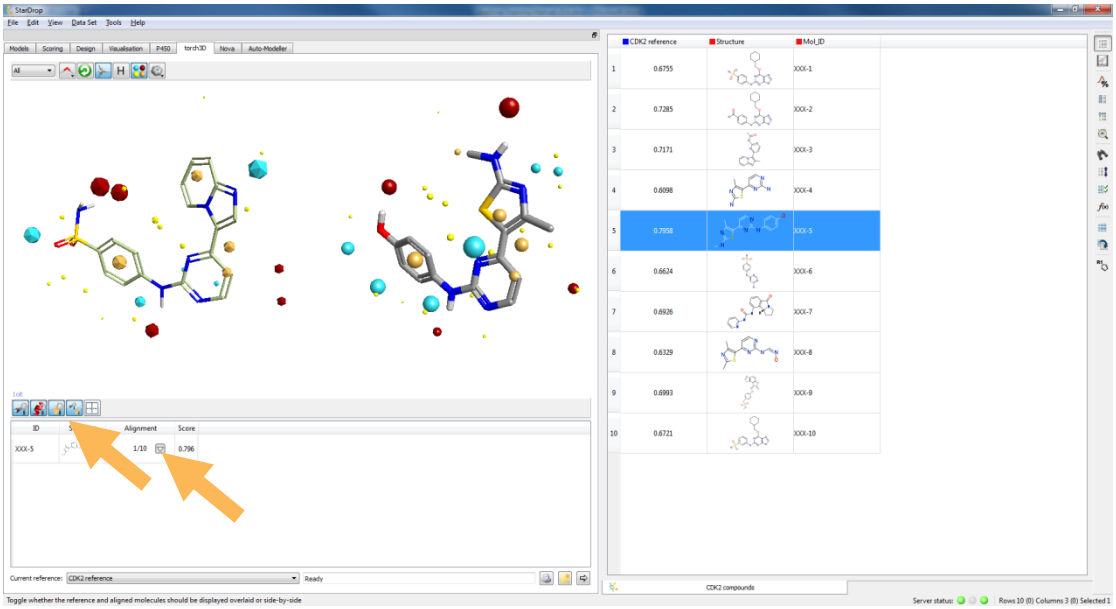
- Select all of the rows in the data set and then click the  button to start the calculations. The torch3D calculations will run in the background and results will appear in the table as they are completed. The field points of the reference are shown as polygons, those of the aligned compounds as spheres.



(**Note:** If you select a compound that's already queued, the  button will change to  indicating that clicking the button again will remove the compound from the queue.)

- Once a score has been returned for a compound, select that row to see the compound and its fields, superimposed on the reference molecule. Using the mouse you can zoom into and rotate the 3D molecules in order to see how their fields compare. You can explore the top 10 alternative alignments identified by clicking on  or  buttons in the "Alignment"

column. Click the  button to see the molecule and the reference compound side-by-side.



1. Many of the compounds are structurally similar to the reference compound. However, compound XXX-7 is from a different chemical class. What similarities may explain the similar activities of this compound and the reference?

   Answer:_____

   _____

   _____

   _____

   _____

2. Compound XXX-4 is smaller than the majority of the active compounds and the top-scoring alignment is similar to the structurally related compounds. Explore the alternative alignments proposed by torch3D. Can you identify any possible alterative binding modes that may suggest opportunities for further optimisation?

   Answer:_____

   _____

   _____

   _____

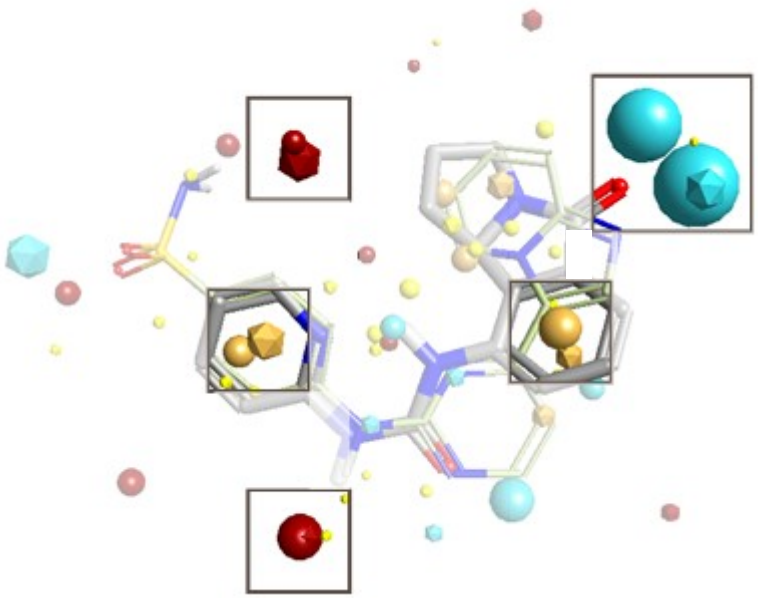
   _____

Answers

1. Many of the compounds are structurally similar to the reference compound. However, compound XXX-7 is from a different chemical class. What similarities may explain the similar activities of this compound and the reference?

   Answer: Despite the difference in chemical structure torch3D identifies a plausible alignment to explain the similarity in activity between XXX-7 and the reference, based on the similar field pattern (and hence interactions) generated by these structures (some are highlighted in the

figure below). Given the commonality between these points, they may indicate important features to give potency against the target.



2. Compound XXX-4 is smaller than the majority of the active compounds and the top-scoring alignment is similar to the structurally related compounds. Explore the alternative alignments proposed by torch3D. Can you identify any possible alterative binding modes that may suggest possible opportunities for further optimisation?

   Answer: An alternative alignment shown below indicates a potential alternative binding mode that may provide an opportunity to extend this compound from the piperidine to form additional interactions and further optimise potency.