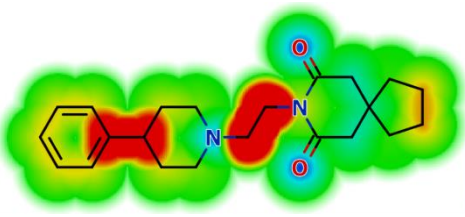
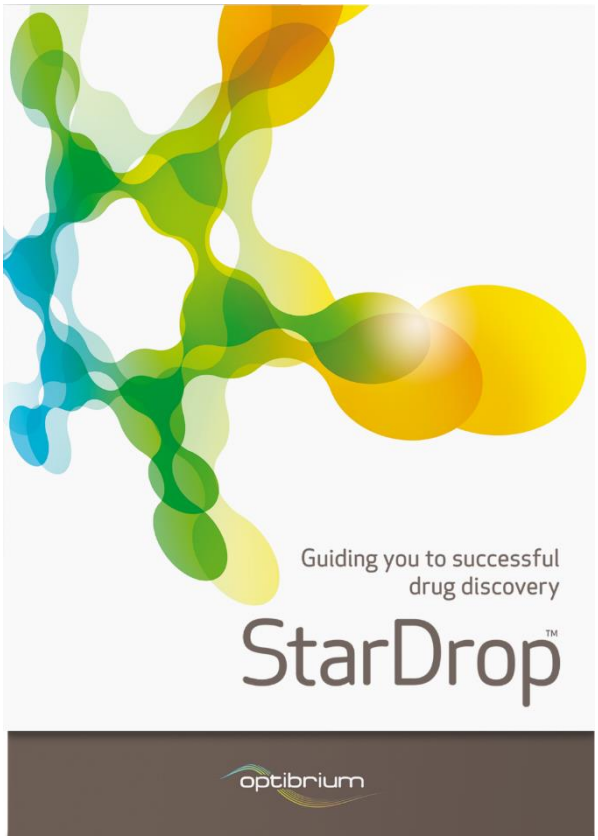



Innovative Lead Optimization and Candidate Selection by *in Silico* Synthesis and ADMET Prediction

December 2-3, 2015, Chungnam National University

Course Leaders: Professor Young Shin and Dr Matthew Segall

Hands On Examples



Contents

Introduction to StarDrop: Getting Started…

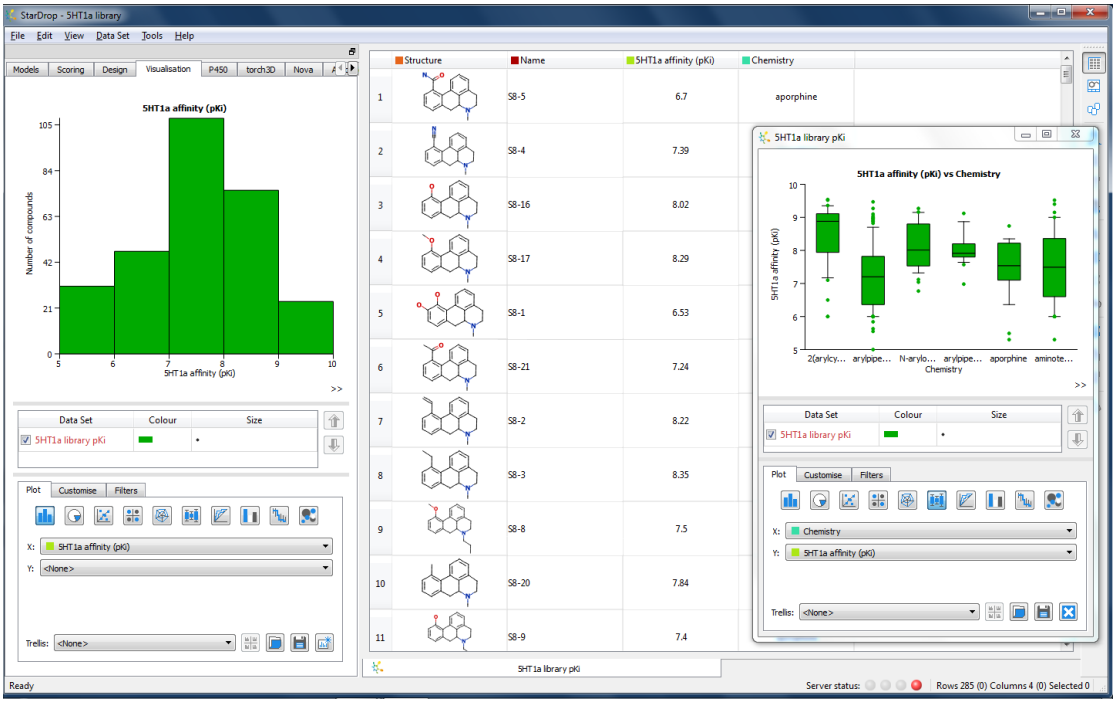
In drug discovery we are faced with many challenges as we look for compounds that have the right balance of properties in order for them to become successful drugs. In the early hit-to-lead stages we often wish to explore a wide range of chemistry in order to find those areas of chemical space which contain lead series with the greatest potential. As we move into lead optimisation we need to focus more closely to learn about the structure-activity relationships (SAR) and design new ideas around specific scaffolds. In this first section we will become familiar with the StarDrop interface and touch on a number of these themes.

Objectives

- Gaining familiarity with the StarDrop interface
- Importing compound data
- Calculating simple properties for compounds
- Visualising data
- Using Card View™

Exercise

- Open the project file **5HT1A library.sdproj**.



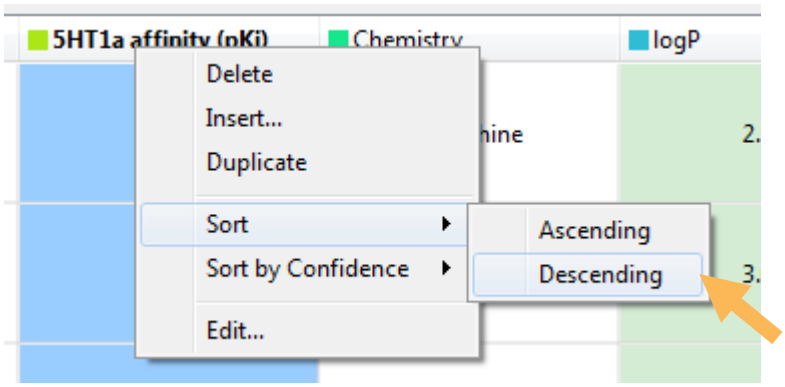
1. Which chemistry has the highest average potency?

   Answer:_____

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

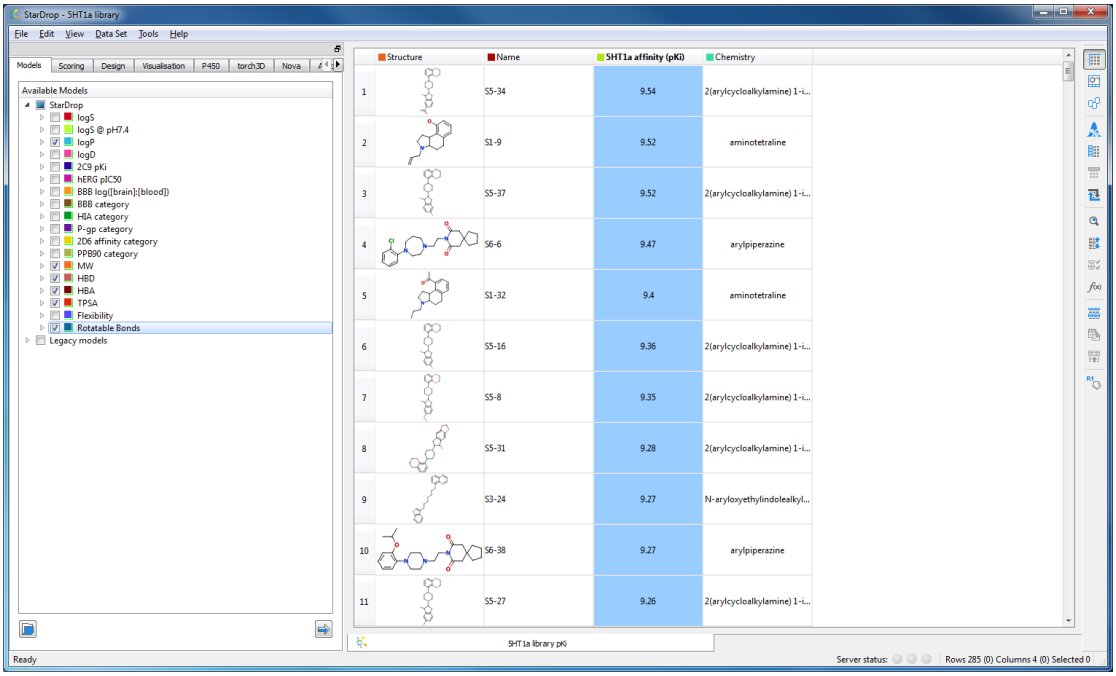
   Answer:_____

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

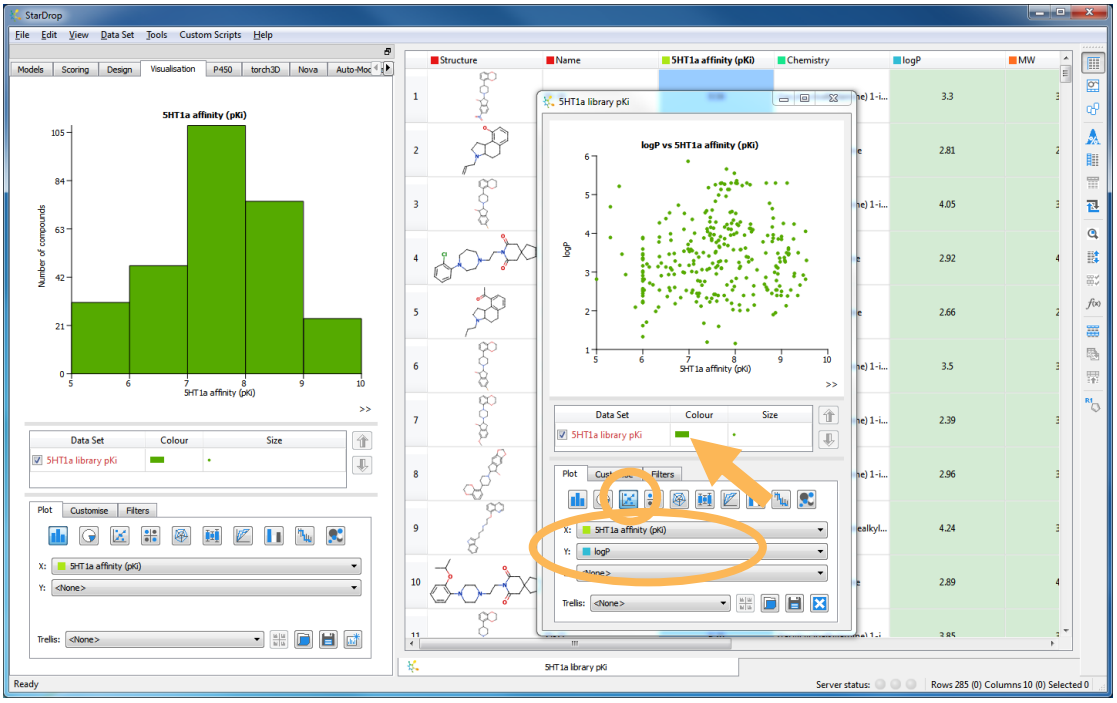


Initially, we will explore the relationship between the measured potency (pK_i) and simple 'drug like' properties of the compounds in this library. We are going to calculate some properties and then use the visualisation tools to find a chemistry that meets a number of property criteria. To do this, use the following steps:

- Select the following properties in the **Models** tab: logP, MW, HBD, HBA, TPSA and Rotatable Bonds.

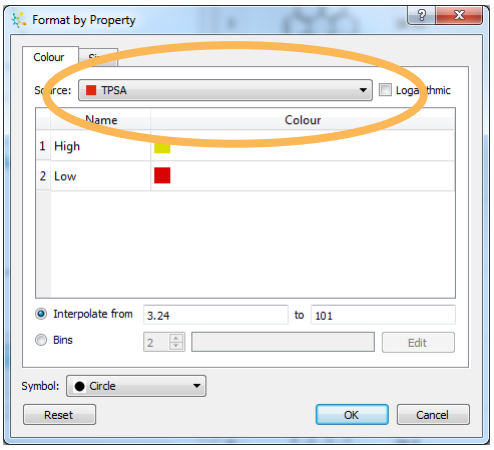


- Calculate these properties for all of the compounds in the data set by clicking the  button.
- In the **Visualisation** tab, click the  button to open another plot window.
- In the new window, plot a 2D scatter plot of **5HT1a affinity (pK_i)** against **logP** by clicking the  button and selecting 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)

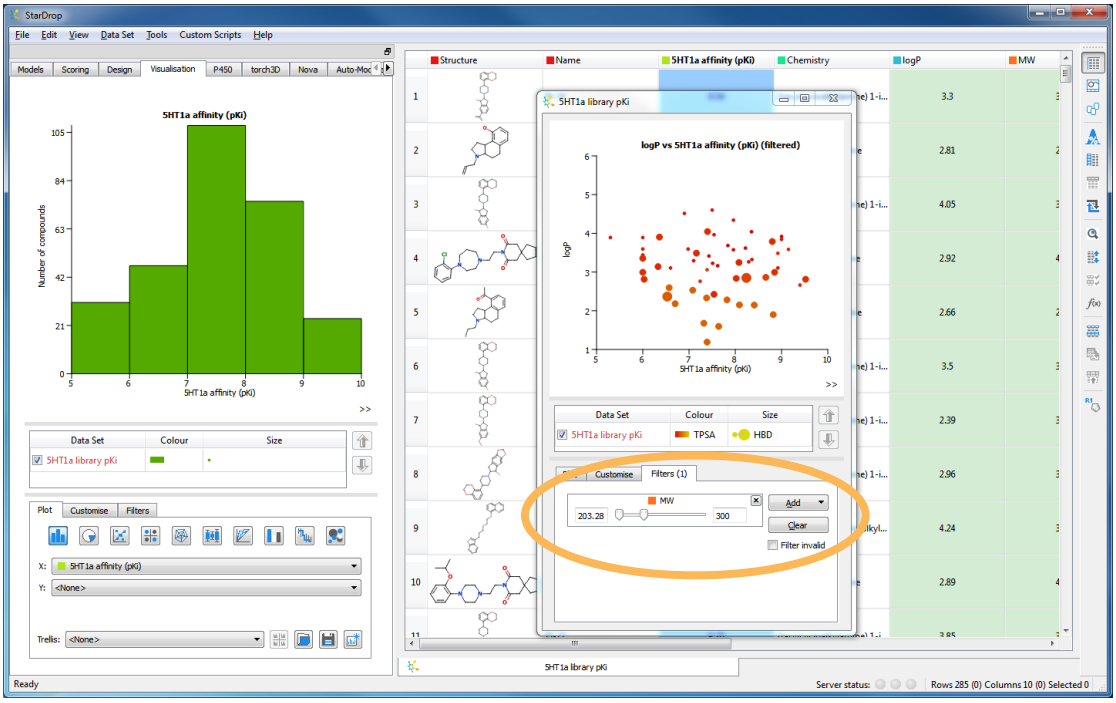


- Now colour the points by **TPSA** and size the points by **HBD**.

  Hint: Click on the colour block in the key (indicated by the arrow in the screenshot above) to display the **Format by Property** dialogue box and choose **TPSA** from the **Source:** menu, as shown right. Click on the **Size** tab in the dialogue box to size by **HBD** in a similar way.



- Finally, change to the **Filters** tab in the plot window to remove points from the plot which have **MW** greater than 300. Click the **Add** button and choose **MW** to create a new filter and then adjust the range as shown below:



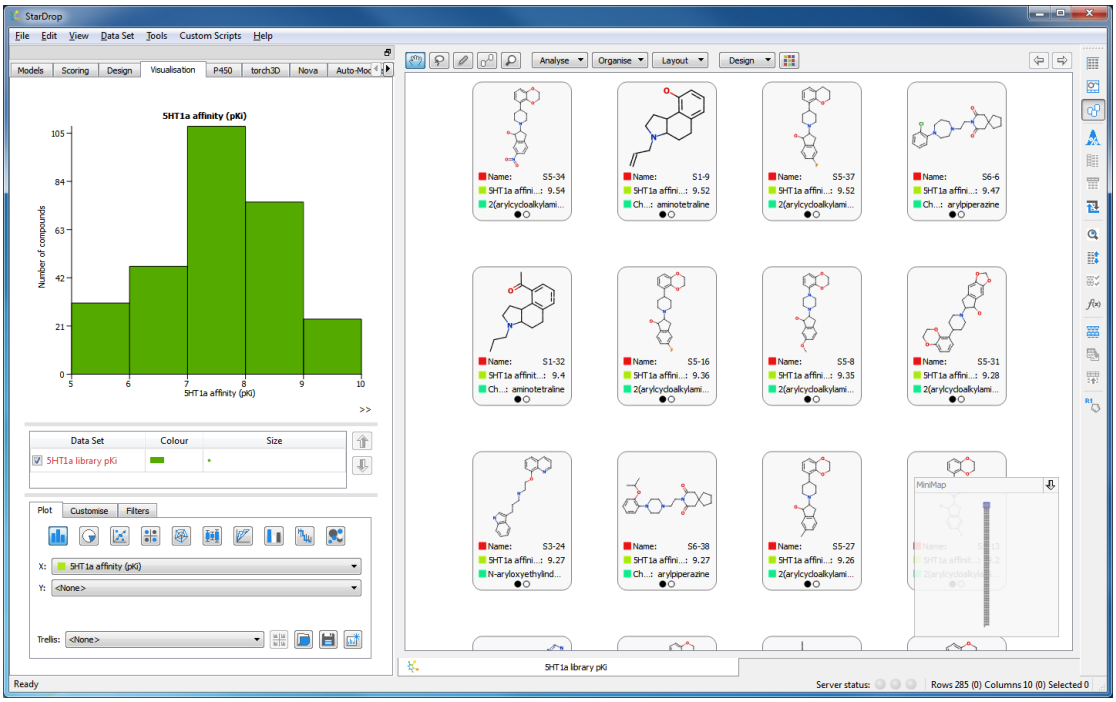
- Select a region in the bottom right corner of the plot.

3. Which chemistry includes the majority of compounds with high potency, low MW and appropriate logP values?

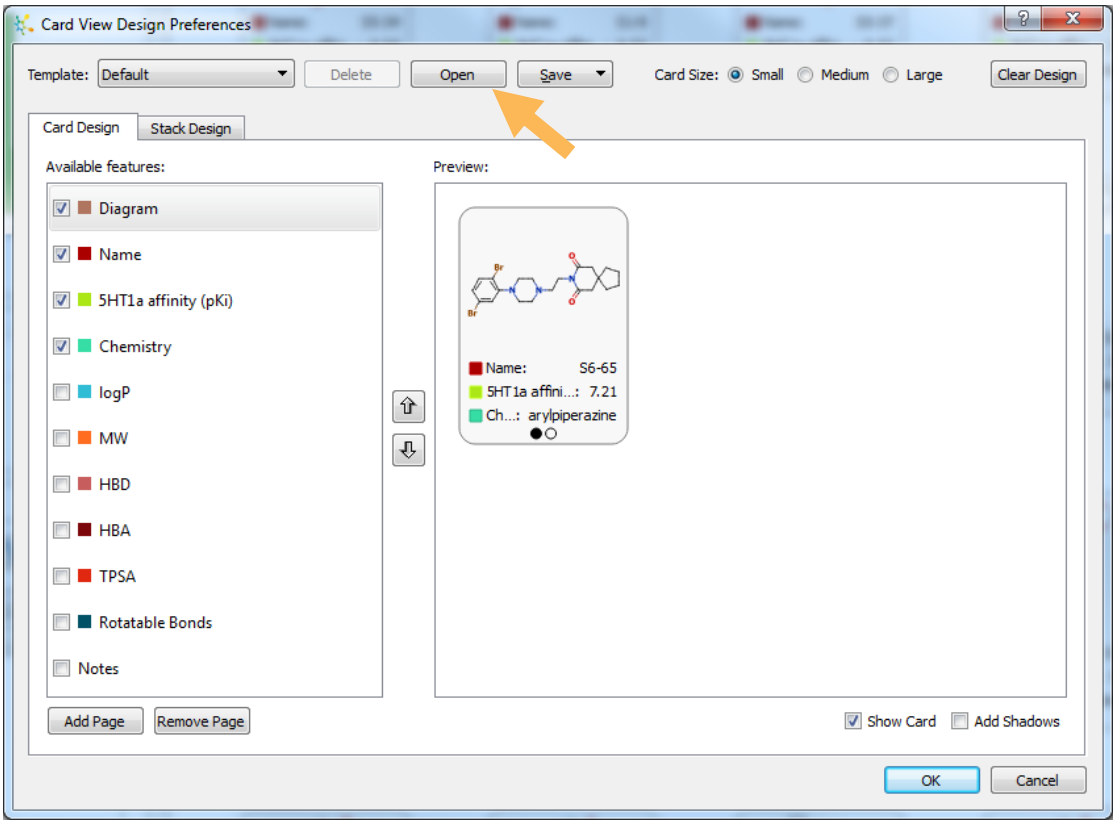
   Answer:_____

Let's take a closer look at the ranges of properties across the different chemical series in this data set.

- Click the  button to switch into Card View.

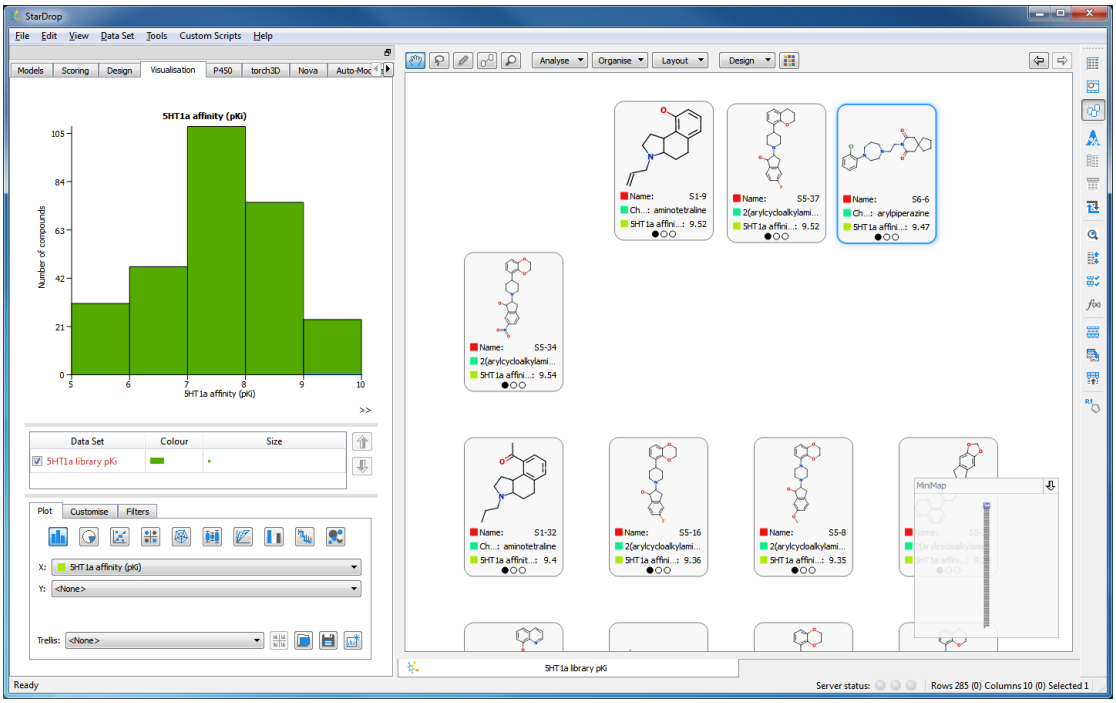


- Click the **Design** button and choose **Custom…**

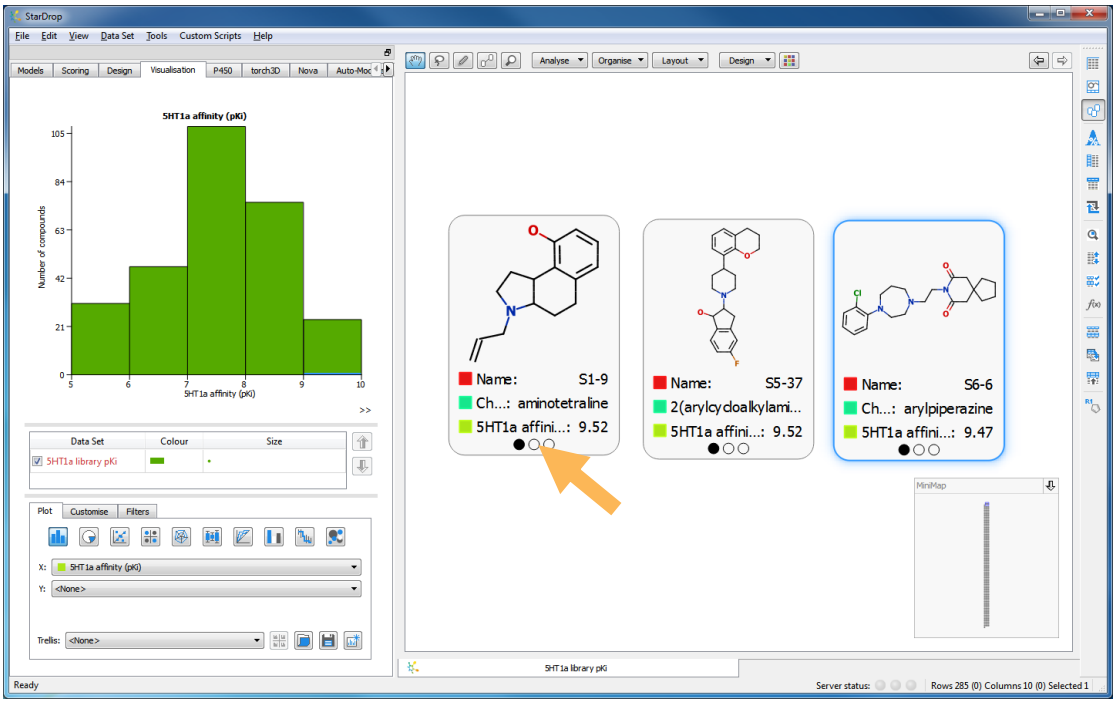


- Click the **Open** button to load the **5HT1a Card Template**. Now click **OK**.

- We can easily compare representative, potent compounds from three different series. Move the Card View table top down by clicking in the empty space between cards and dragging down. Now move the 2nd, 3rd and 4th cards up into the space above, next to each other.



- Zoom in (using the mouse-wheel or the **Ctrl** + **=** keys) to take a closer look at their properties.



- Toggle the property pages to compare their different properties by clicking the circles at the bottom of one of the cards.

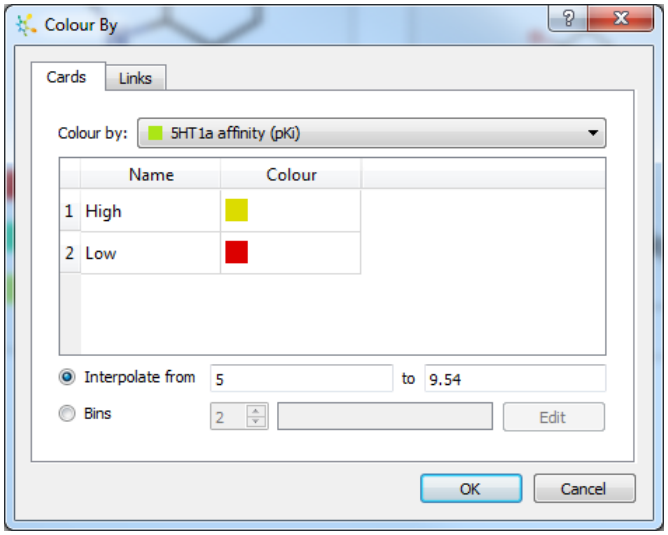
4. All three compounds are very potent so on the basis of these three examples which chemistry type looks promising as a potential lead?

   Answer:_____

- Change to the third page and add a note for compound S1-9 indicating that this compound has potential.

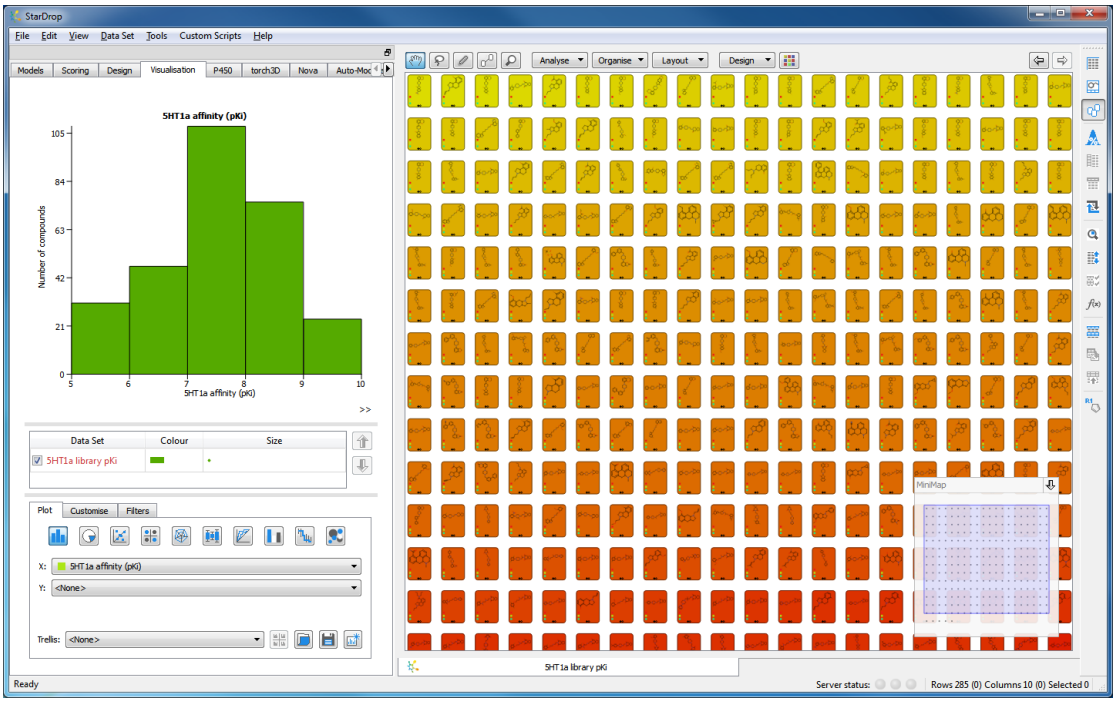
  Hint: Click on the third circle on the card and then double-click in the **Add Notes…** field.

- Click the  button to colour the cards. In the **Colour By** dialogue choose the property **5HT1a affinity (pK$_i$)**.



- Click **OK**.

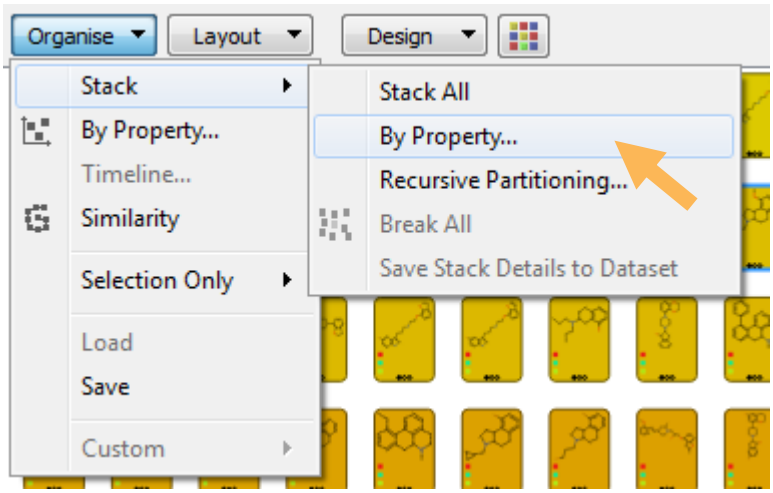
- Using the mouse-wheel (or the **Ctrl** and **–** keys) zoom out again and then from the **Layout** menu choose **Grid**.



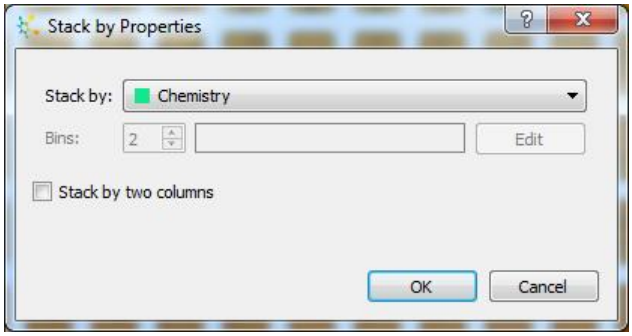
Our data set is already sorted with the most potent compounds at the top.

- Once again, select a region in the bottom right corner of the scatter plot to see the associated cards.

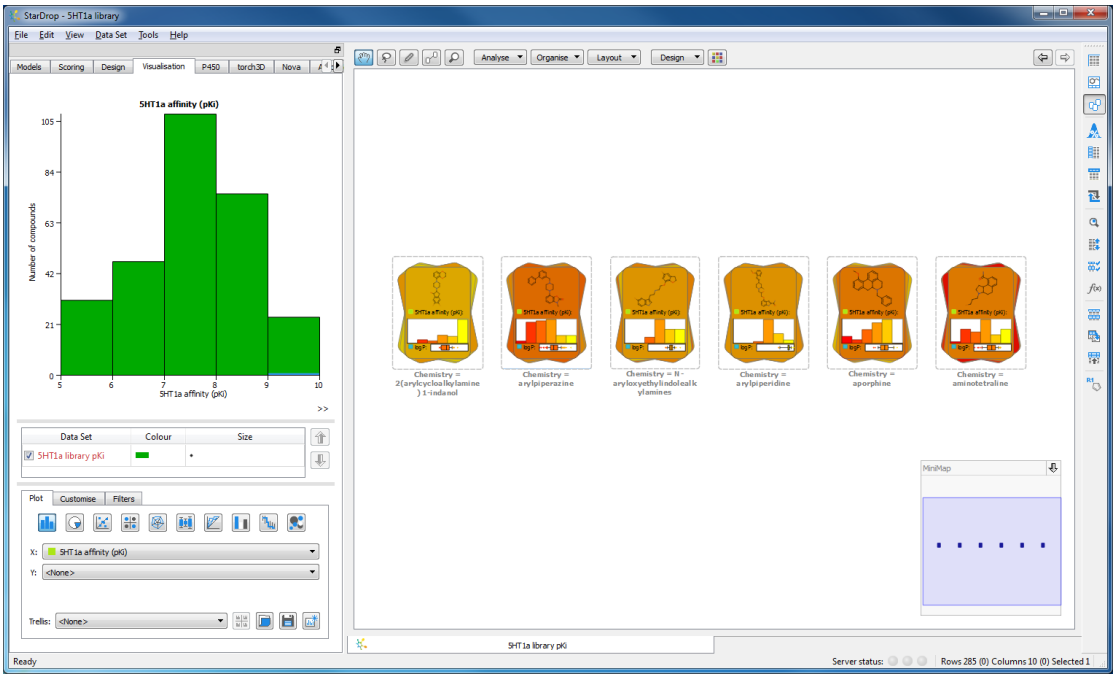
- We are now going to take a look at the properties of each chemical class (as indicated by the "Chemistry" column in the data set) so from the **Organise** menu choose **Stack** -> **By Property…**.



- In the **Stack by Properties** dialogue choose "Chemistry".



- Click **OK**.



5. Which stacks appear to contain compounds with the best ranges of potency and logP values?

   Answer: _____

Applying Probabilistic Scoring for Multi-Parameter Optimisation

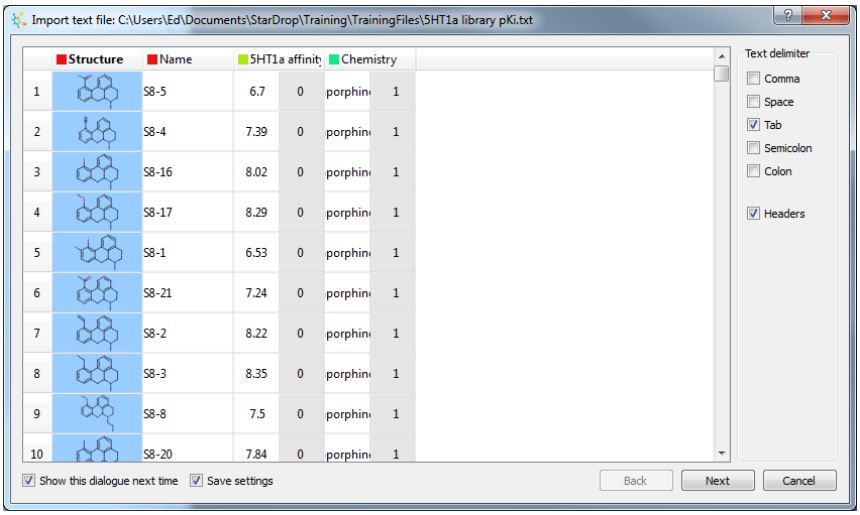
In hit-to-lead, it is important to quickly focus resources on the chemistries that are most likely to yield a high quality lead series. In this example we will explore how data from primary screening of a library for potency against the target 5HT1a can be combined with predictions for a range of ADME 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.

Objectives

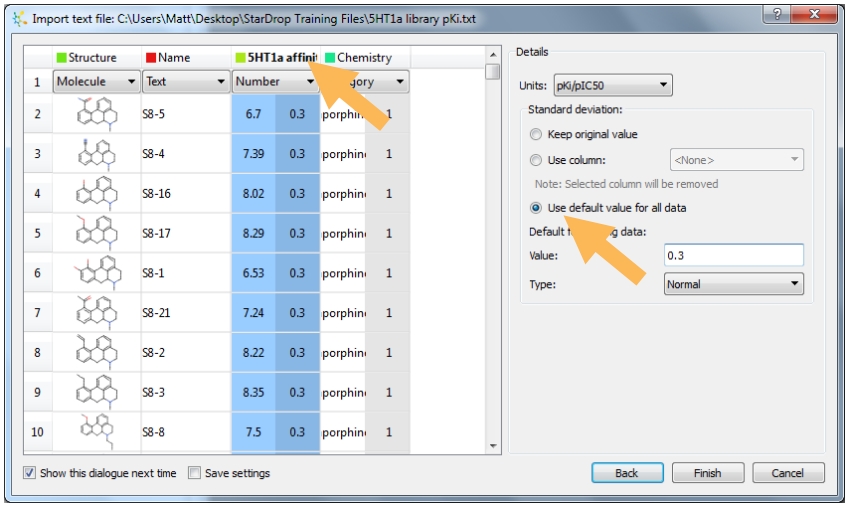
- Using predictive ADME QSAR models
- Creating and using scoring profiles
- Interpreting scores

Exercise

- From the **File** menu choose **Close Project** (**Discard**ing any changes) and open the file **5HT1A library pKi.txt**.



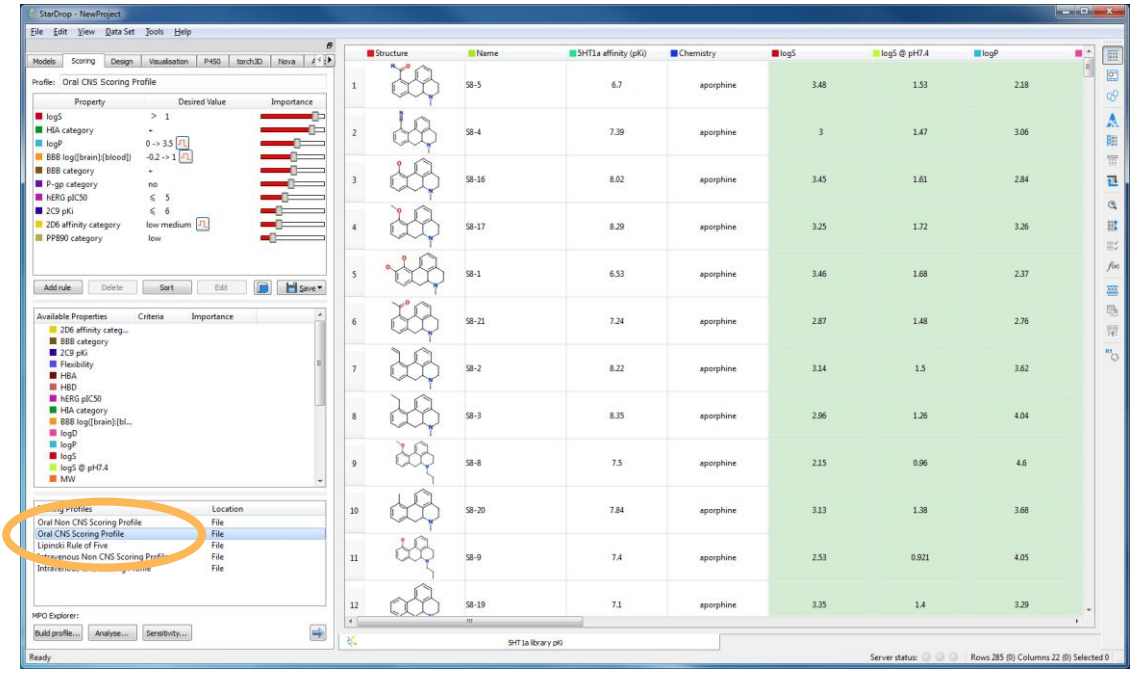
- Confirm the text delimiter that has been used to create this file (**Tab**) and click **Next**.
- As we import the data, we are going to provide some additional information about the experimental values we are importing. Select the **5HT1A affinity (pK$_i$)** column in the import tool. Set the units of the experimental pK$_i$ data to pK$_i$/pIC50 and the uncertainty to 0.3 log units (equivalent to a factor of two in the K$_i$).



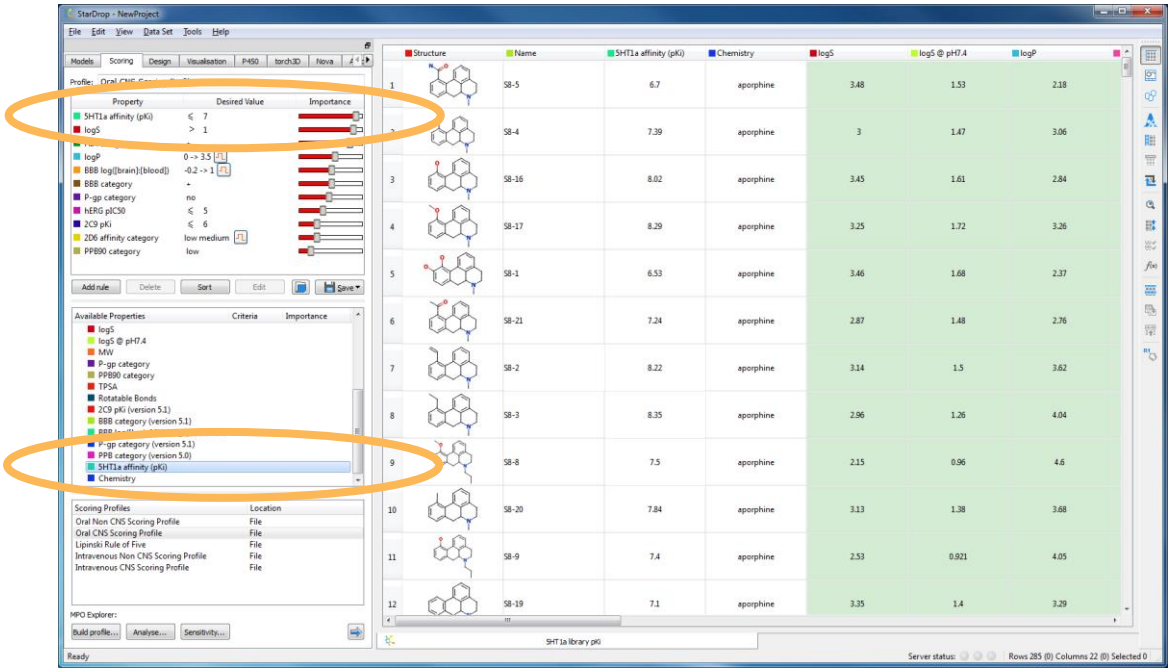
- Click **Finish** to complete the data import.

- On the **Models** tab select all the StarDrop models and click the  button to run them.

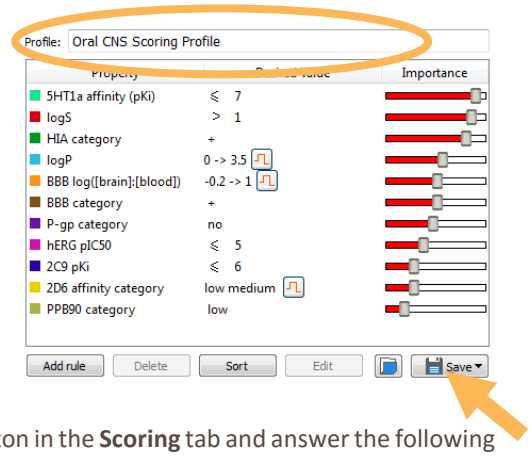
- Change to the **Scoring** tab and select the **Oral CNS Scoring Profile**, listed under **Saved profiles**, as shown below:



- Add **5HT1a affinity (pK$_i$)** 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 **Save** button below the scoring profile and choosing **Save to File…**. It will appear in the list of scoring profiles 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. 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.

7. Which compound has the highest score?

   Answer:_____

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

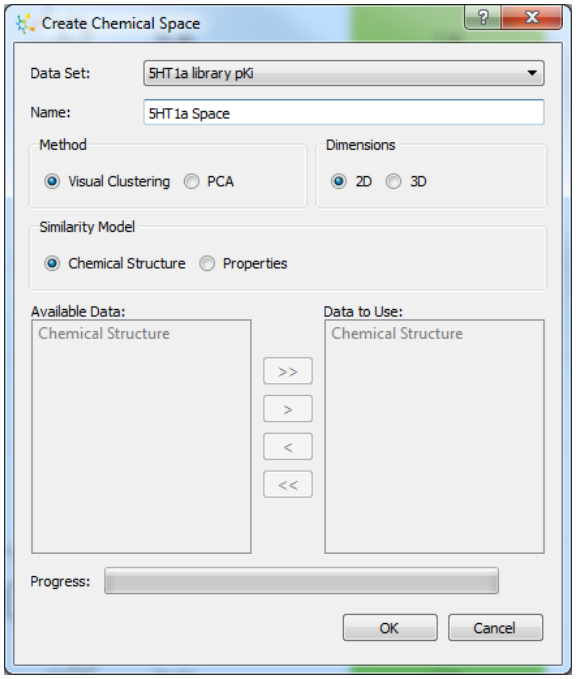
Exploring Chemical Space to Balancing Quality and Diversity

Objectives

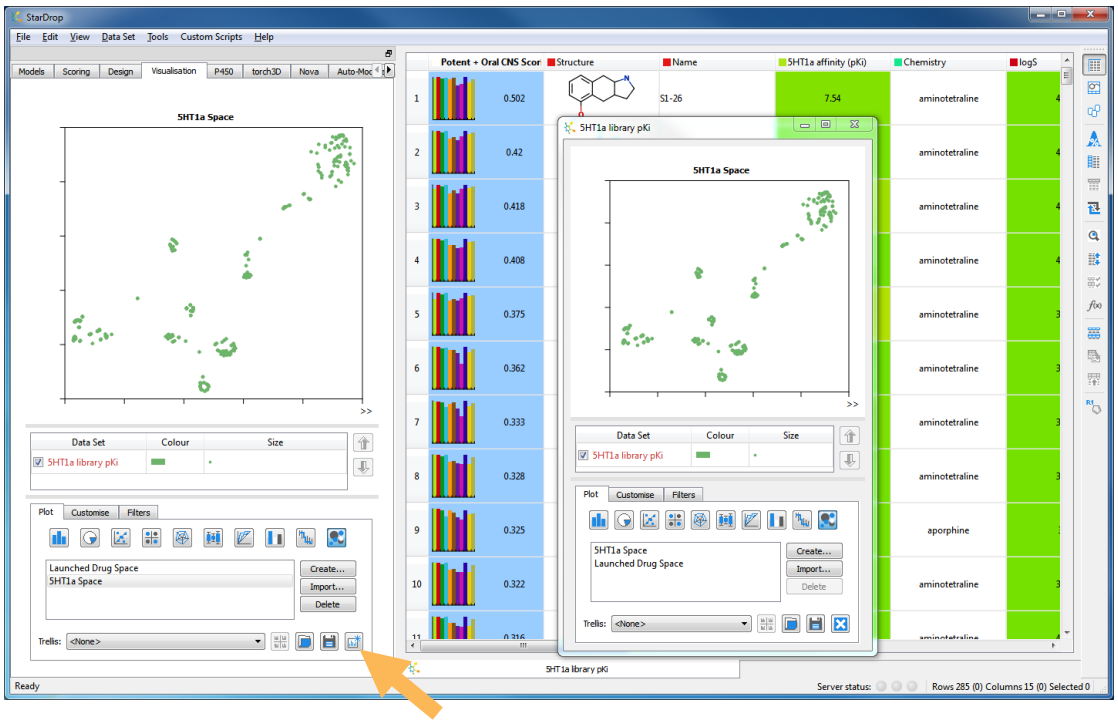
- Generating and visualising chemical spaces
- Considering uncertainty when selecting compounds
- Exploring an appropriate balance of quality and diversity

Exercise

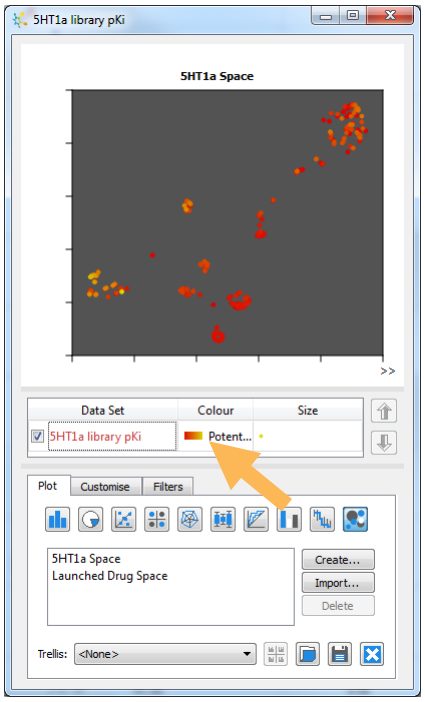
- Change to the **Visualisation** tab.
- Click the chemical space button  . Click **Create...** to create a new chemical space. In the **Create Chemical Space** dialogue, give the new projection a name of **5HT1a Space** and then use the default settings, as shown to the right, to generate a chemical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

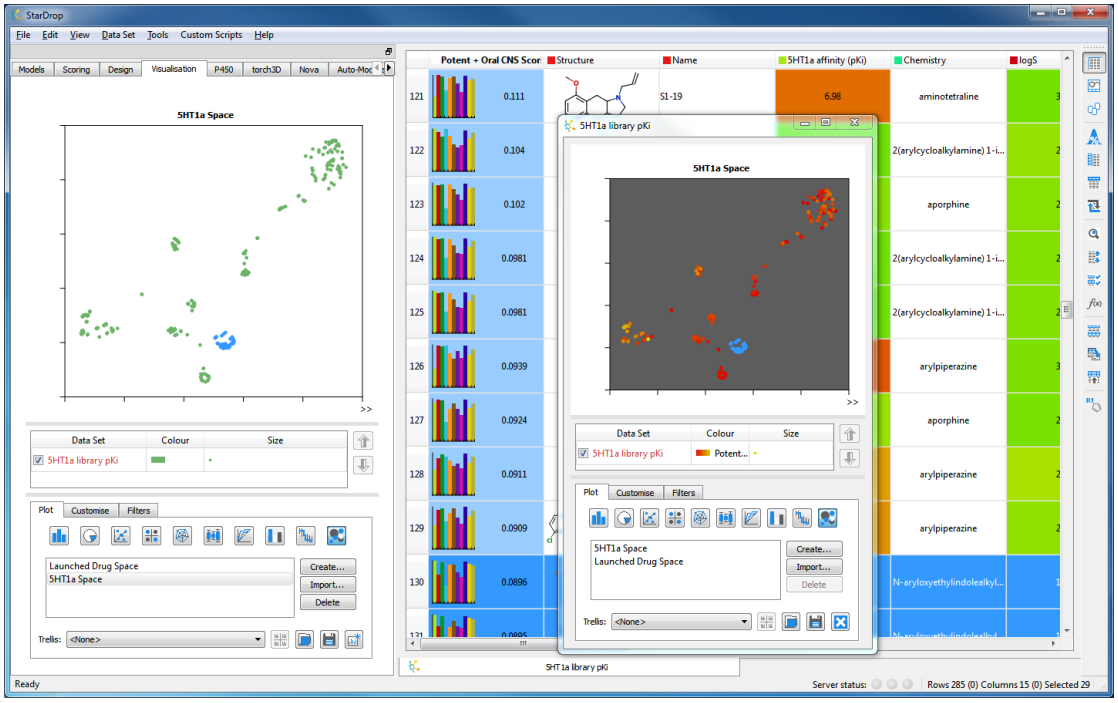


Hint: You can change the background colour on the plot by right-clicking on the plot and selecting **Change Background...**

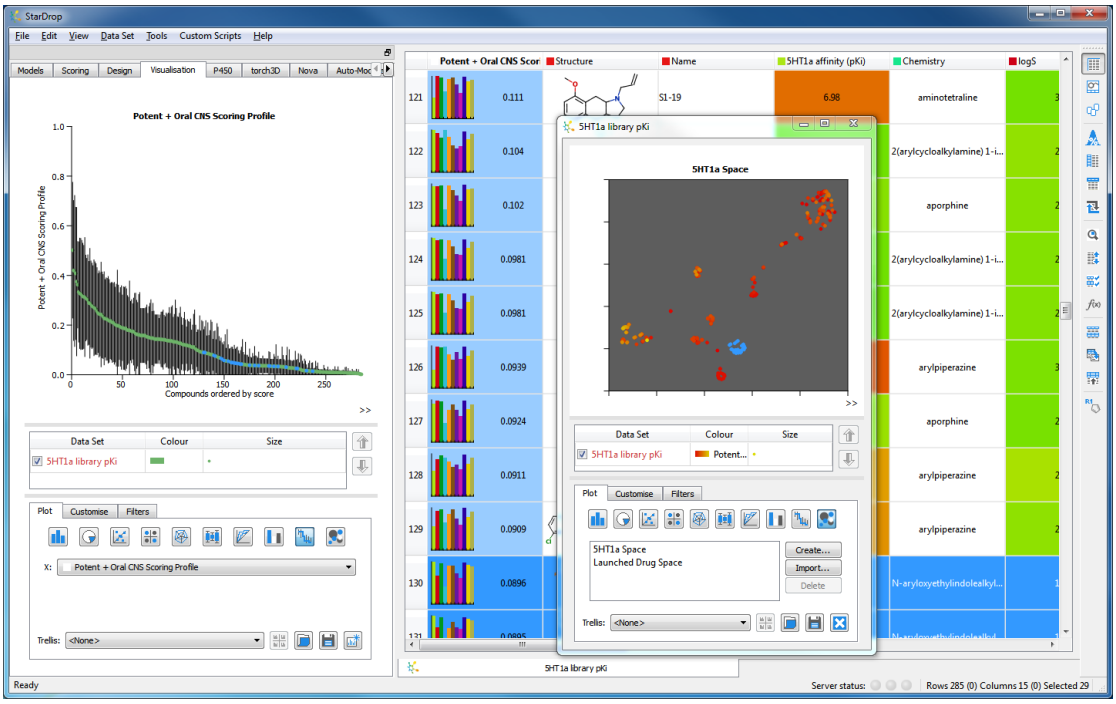
8. Which chemistry contains the majority of the top 10 compounds?

Answer: _____

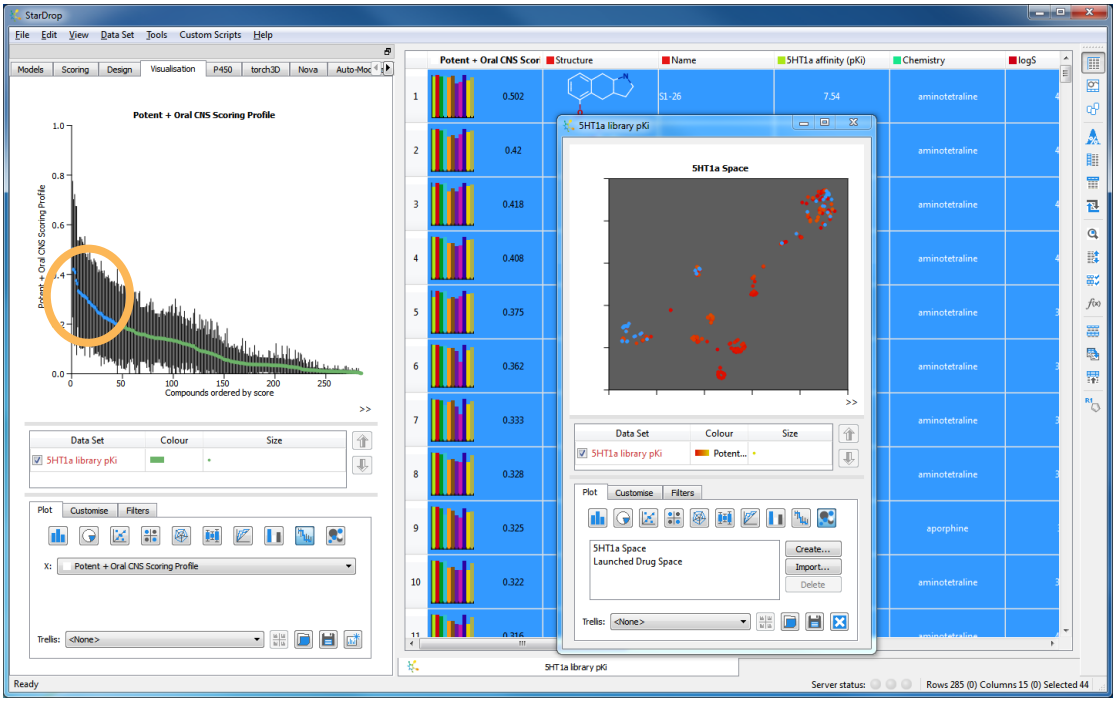
Hint: Selecting points from the chemical space will select the corresponding compounds in the data set and vice-versa, as illustrated below (Please note, the selection shown below is not the top 10!):



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

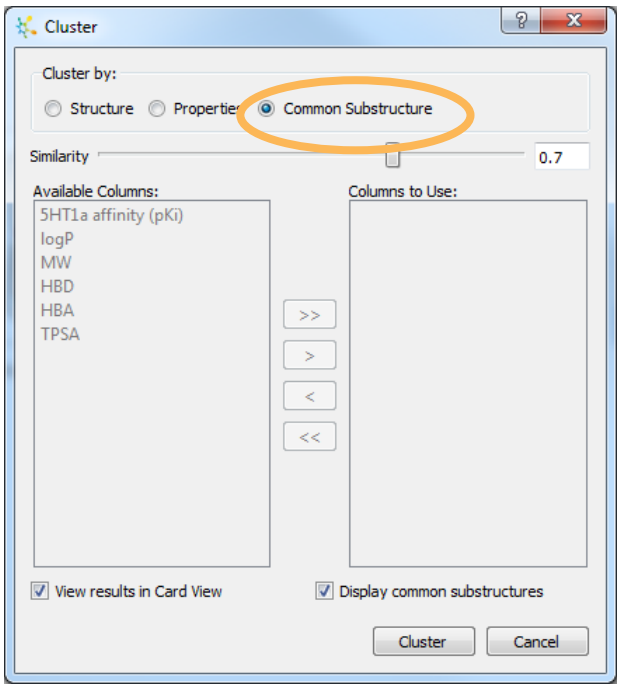


- Points in the **Snake Plot** for which the error bar overlaps with that of the first compound cannot be confidently distinguished from the highest scoring compound, based on the selected scoring profile and the uncertainty in the available data. By selecting compounds with an appropriate range of scores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

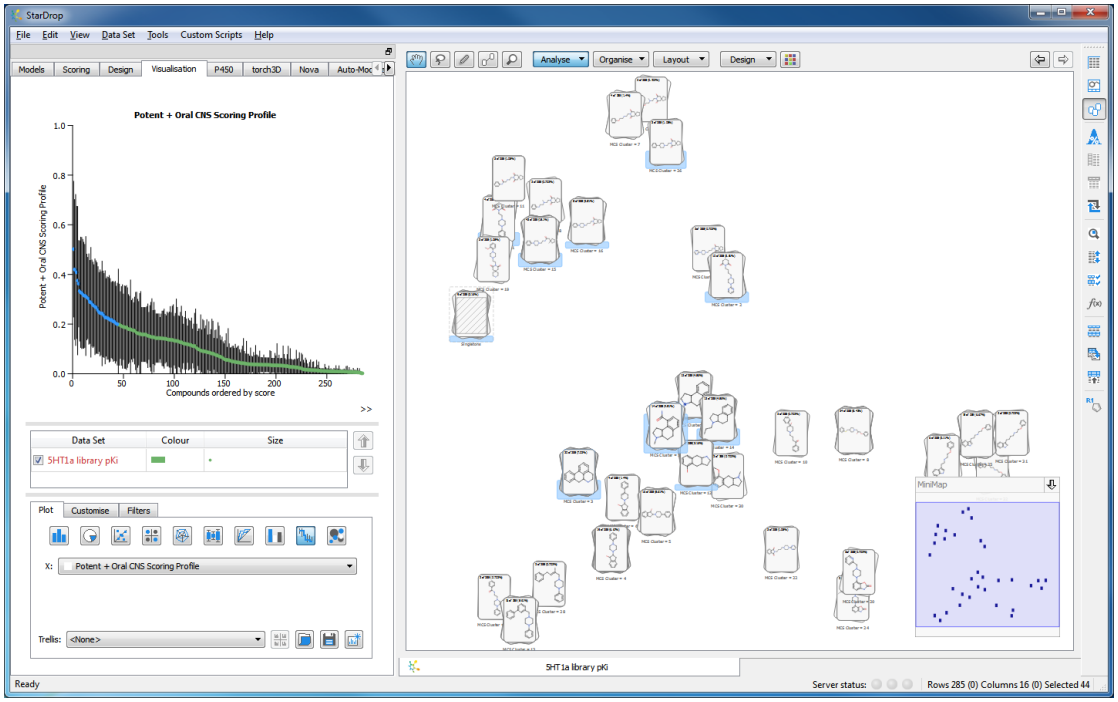


As an alternative approach to explore the distribution of high scoring compounds within this library, we can use StarDrop's analysis tools to cluster our compounds based upon their structures or properties.

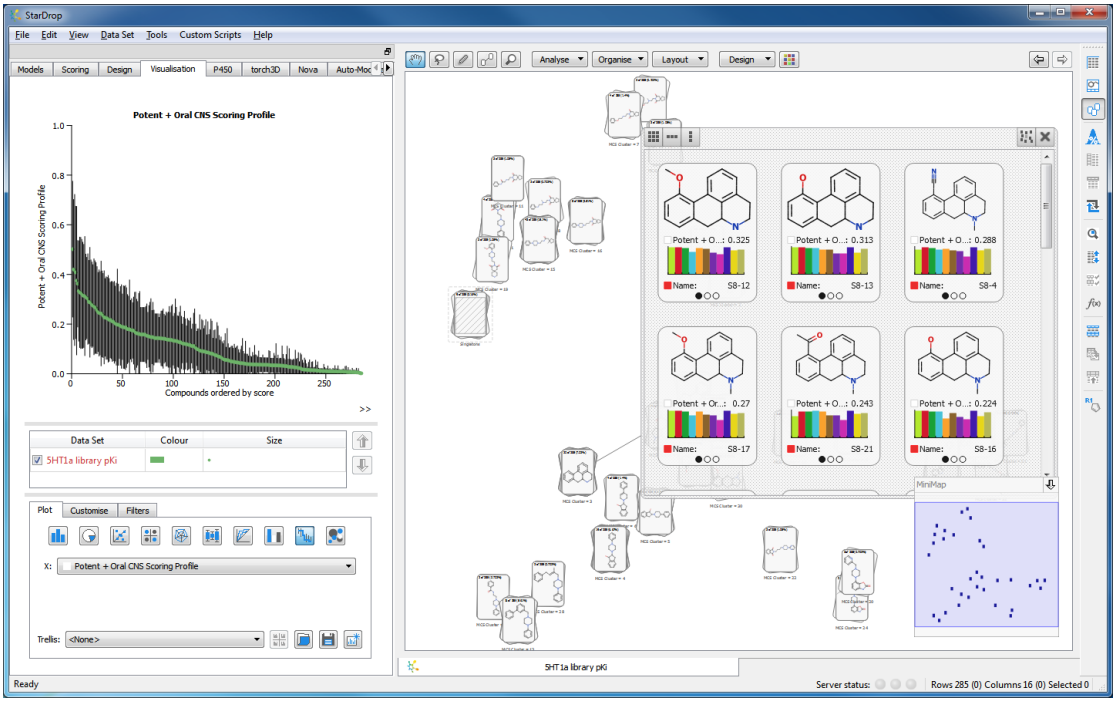
- Switch back into Card View by clicking the  button.
- From the **Analyse** menu choose **Clustering**.



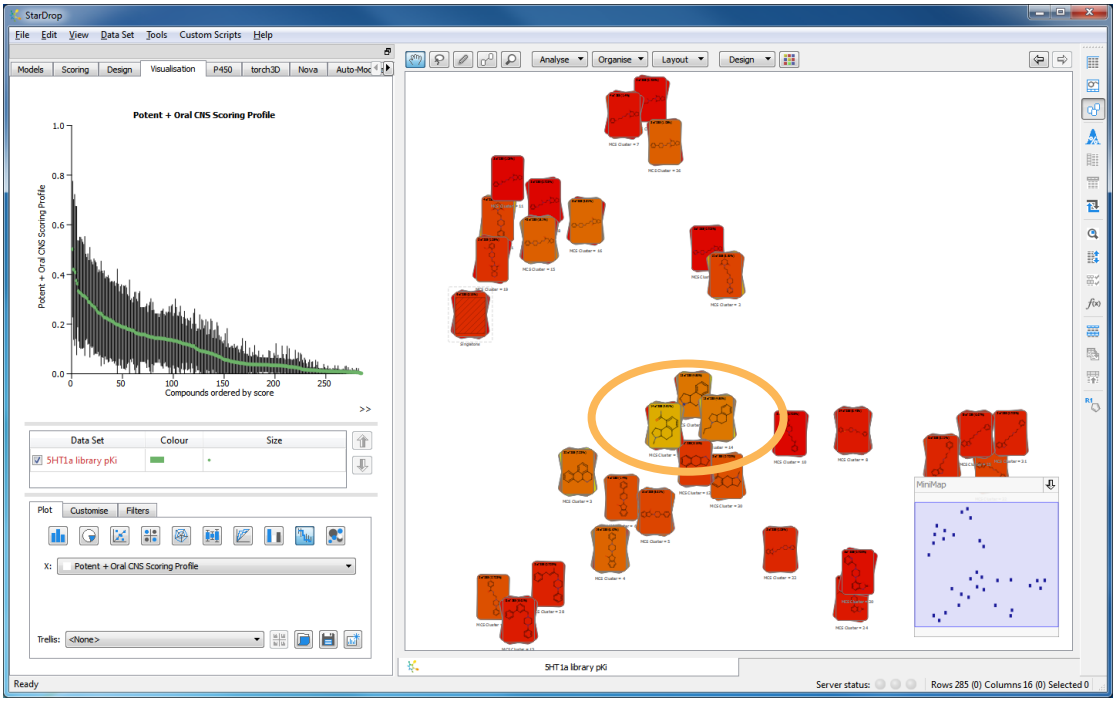
- Select **Common Substructure** and then click the **Cluster** button.



- Right-click on a stack and choose **Inspect** from the menu. This enables us to browse the cards in this stack and, if necessary, remove cards, simply by dragging them out of the window.



- Click the  button and colour the cards by the overall score – interpolating from 0 to 0.3.

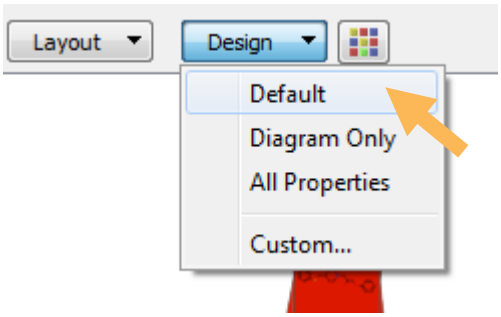


- We can see that the Aminotetralines generally have the best overall scores, but we'd like to refine the clustering results to combine the three stacks which contain Aminotetralines with common sub-structures (highlighted above) by dragging one on top of the other.

  Note: On some computers this display may be rotated but the Aminotetralines we'd like to combine are highlighted above.

The newly combined stack will show the common sub-structure of all the compounds in the three stacks that we have combined. This flexibility enables us to 'fine tune' the layout based on our understanding of the chemistry.

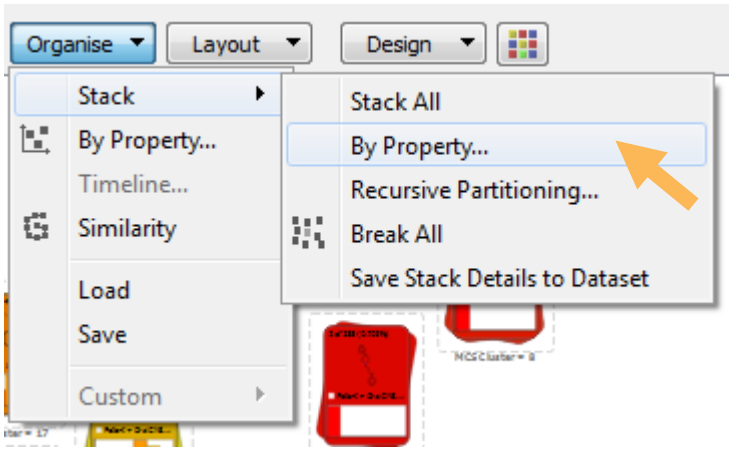
We can easily compare the distributions of scores within the clusters by showing these on the stacks:



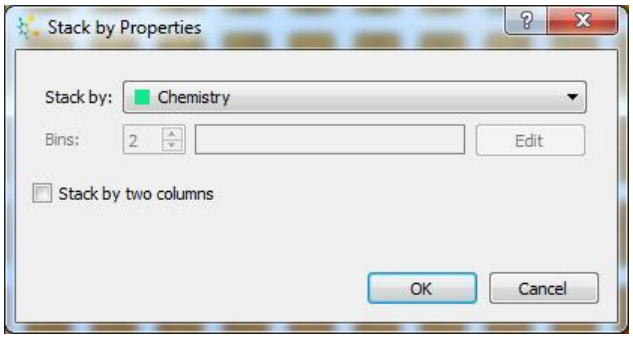
- From the **Design** menu in Card View, choose the **Default** option.

To save time in combining the remaining clusters into chemical series, we can use the pre-defined "Chemistry" column to create a stack for each chemical series:

- From the **Organise** menu in Card View, choose **Stack** -> **By Property…**.



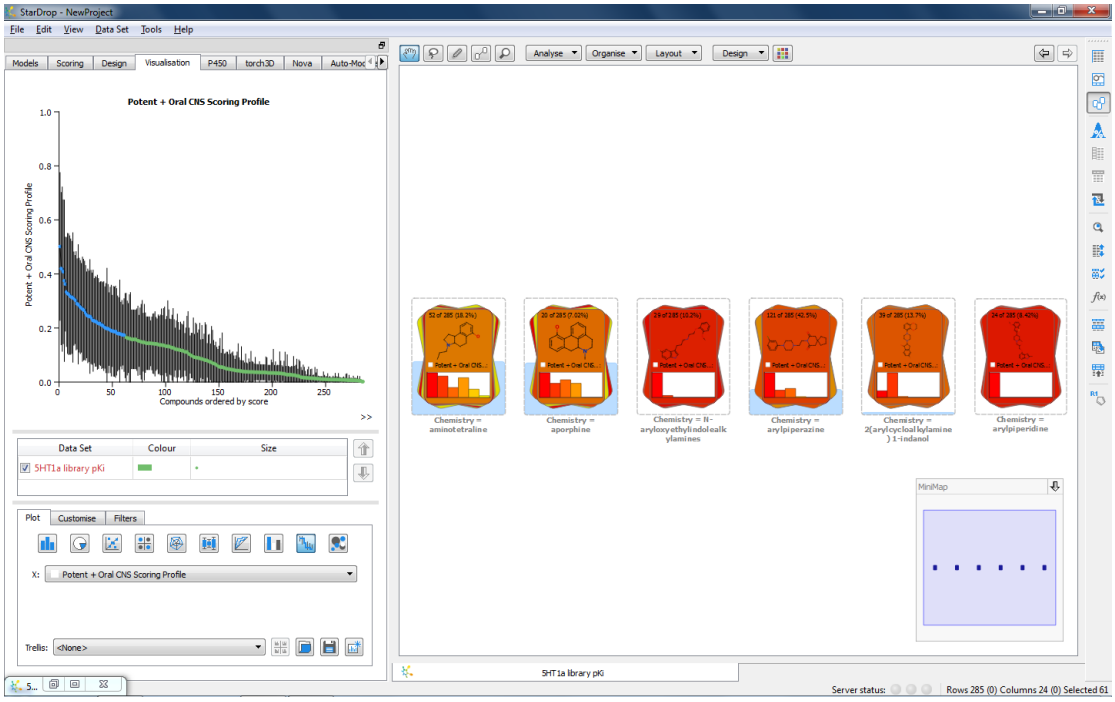
- In the **Stack by Properties** dialogue choose “Chemistry”.



As we saw previously, the Aminotetraline series contains the majority of the high scoring compounds. However, we can now easily see which other chemical series should also be considered:

- In the **Snake Plot** in the **Visualisation Tab** select the compounds that are not confidently distinguishable from the top-scoring compound.

  Hint: If the Snake Plot is no-longer visible, click the  button on the **Visualisation** tab to create this plot again.



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

   Answer:_____
   _____

Interactive Design and the Glowing Molecule

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).

In the following example, we will explore 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.

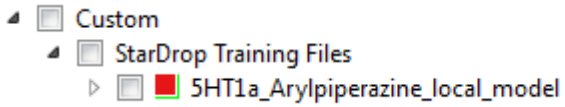
One of the potential issues identified for this chemical series is inhibition of the hERG ion channel, indicating a risk of QT prolongation and cardiotoxicity. Therefore, we will use the interactive designer, guided by the Glowing Molecule, to explore potential strategies to reduce the predicted hERG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

Objectives

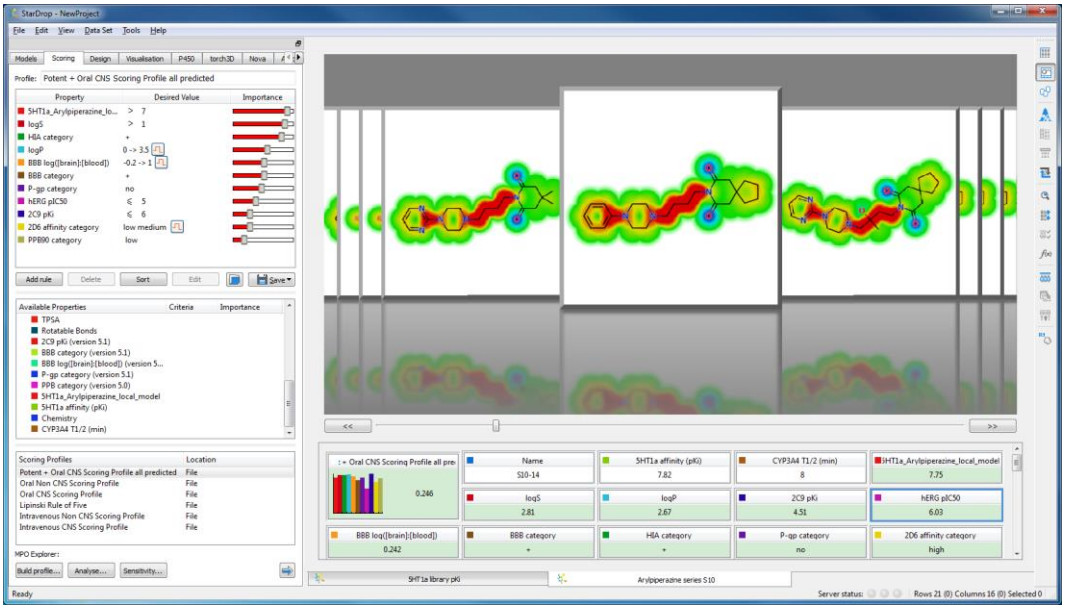
- Loading additional predictive models and scoring profiles
- Using the interactive designer
- Interpreting the Glowing Molecule

Exercise

- 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.
- 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:

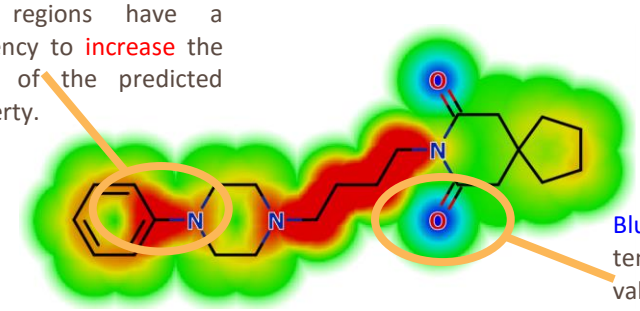


- Now we will load a new scoring profile that uses this model. Change to the **Scoring** tab, click the  button and select the scoring profile **Potent + Oral CNS Scoring Profile all predicted.apd**. This is the same as the profile we used earlier, using predicted values of the 5HT1a pK_i in place of 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 clicking the  button on the toolbar. Find compound S10-14, which is the highest-scoring compound (**Hint:** use the **Find** tool  on the toolbar). Click on the **hERG pIC50** model result to display the Glowing Molecule for this property, as shown below.



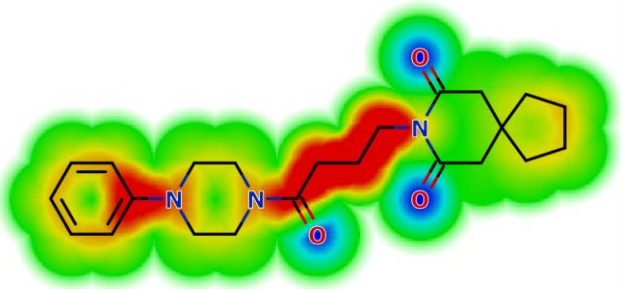
- Change to the **Design** tab to explore strategies to reduce the potential for hERG inhibition. Select the displayed structure in the molecule view and it will be displayed in the editor. When the **hERG pIC50** prediction is selected, the Glowing Molecule will also be displayed, as shown below:

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:



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

    Answer: _____

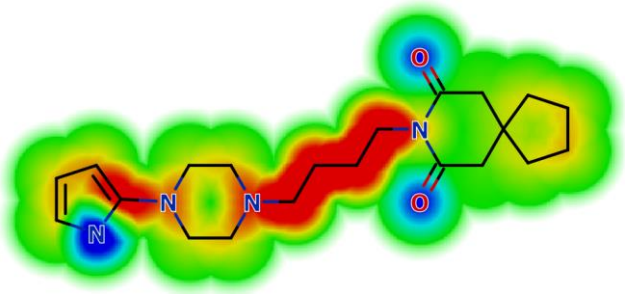
11. What effect does this have on the overall predicted score?

    Answer: _____

12. 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**.

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

    Answer:_____

14. 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.

Auto-Modeller Exercise

Objectives

- Using the Auto-Modeller wizard to build models
- Building continuous 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.

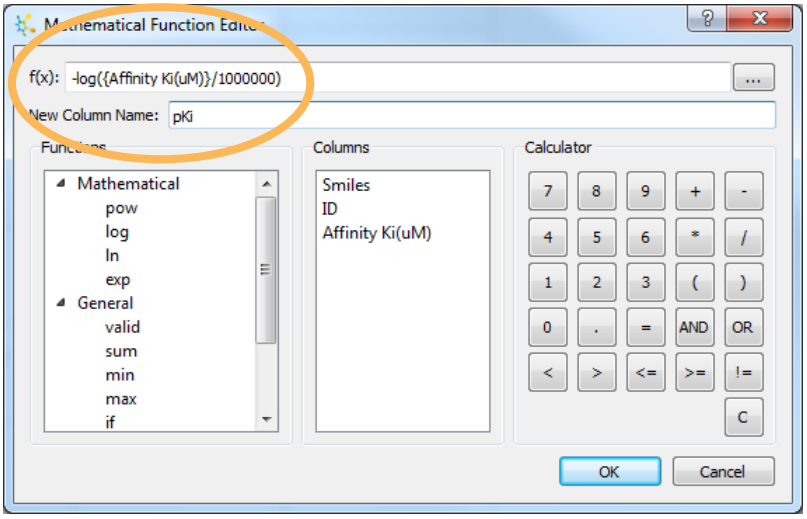
Exercise

- Open the file **Affinity Data.sdproj**.

This project contains two data sets AffinityData and AffinityData2. We will use the first of these data sets to build a QSAR model of the target potency. First, we need to convert the target K_i data to appropriate units.

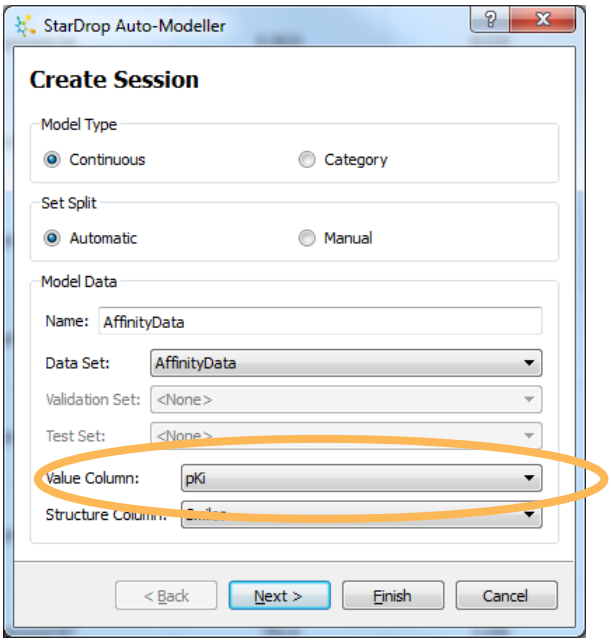
- Open the **Mathematical Function Editor** by clicking the  button and create a function to convert the K_i values into pK_is in a new column called **pK_i**. The Ki values are µM and so the function is:

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



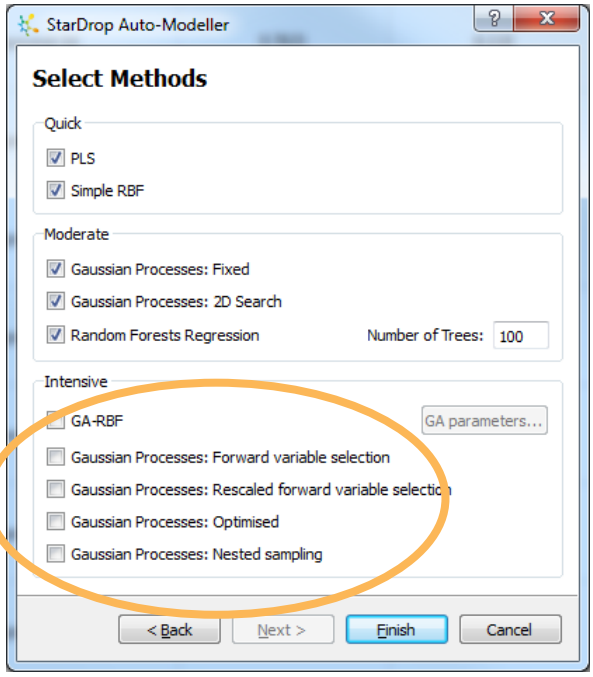
- On the **Auto-Modeller** tab start the Auto-Modeller wizard by clicking the  button.

- On the first wizard page, accept all the default settings but change the **Value Column:** to be **pKi**.



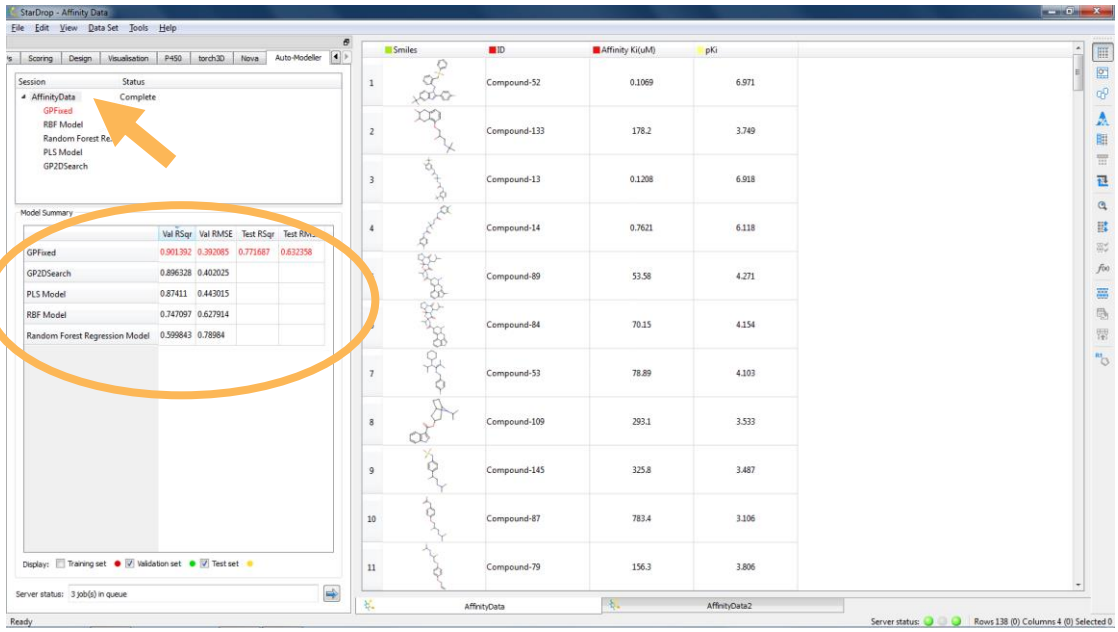
- Click **Next** until you reach the **Select Methods** page. Here, un-tick all the **Intensive** methods.

  (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.

15. 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 right-click on the modelling session and select the **View Session Details** menu option.

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

Answer:_____

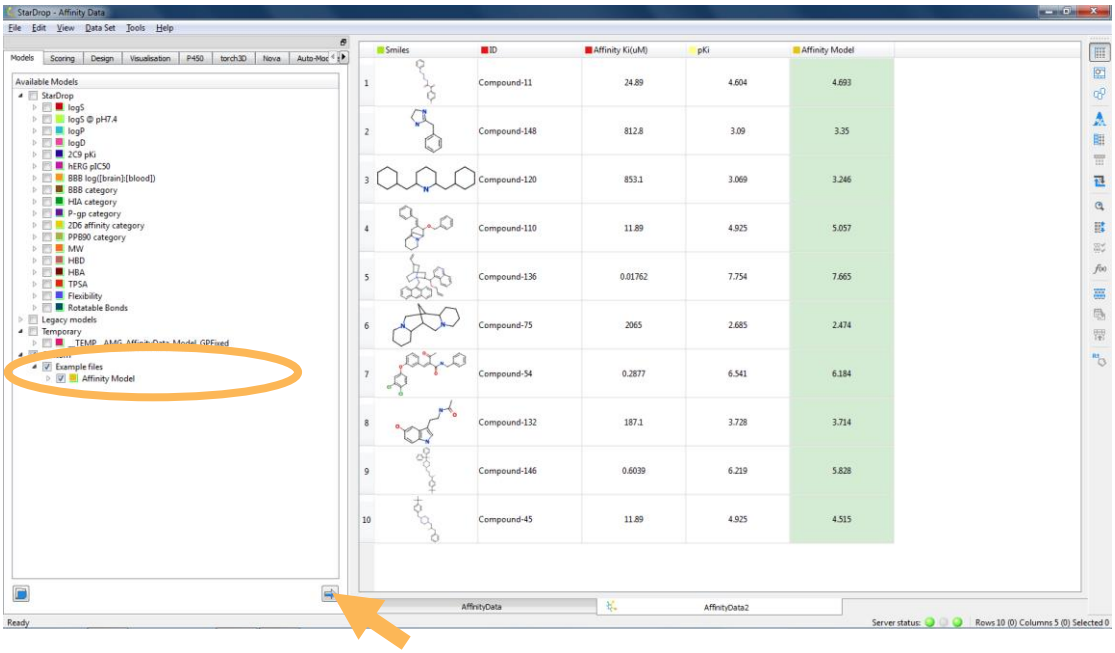
_____

_____

_____

- Save the best model by right-clicking on them and choosing **Save Model...** from the menu.
- Change to the **AffinityData2** data set.
- From the Models tab, run the best continuous model against this set of compounds.

  Hint: the saved models will appear in the Custom section.



17. Comment on how well this model performed and which might be most useful when making decisions about which compounds to synthesise in the future (For convenience, the K_i values have already been converted into pK_is to help you to compare them with the models):

Comments:_____

_____

_____

_____

_____

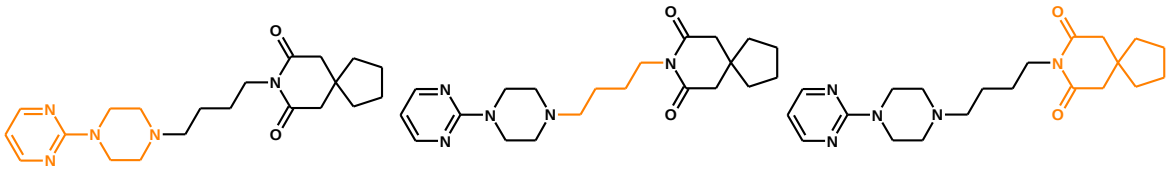
Predicting Metabolism by Cytochrome P450 to Guide Optimization of Metabolic Stability

Objectives

Background

In this example we will explore the feasibility of pursuing a fast-follower for Buspirone, a 5-HT_{1A} ligand used as an anti-anxiolytic therapeutic. Buspirone has a known liability due to rapid metabolism by CYP3A4, leading to low oral bioavailability and a short half-life in man. The project wished to efficiently identify analogues of Buspirone with an *in vitro* CYP3A4 half-life 3-times longer than Buspirone and a minimum loss of receptor affinity.

The structure of Buspirone can be broken down into three regions:



Arylpiperazine

- Protonatable recognition element, receptor affinity
- Metabolism: Hydroxylation at pyrimidine C5

Tetramethylene linker

- Metabolism: N-dealkylation α to piperazine N4

Piperidinedione moiety

- Metabolism: oxidation of spirocyclopentane ring

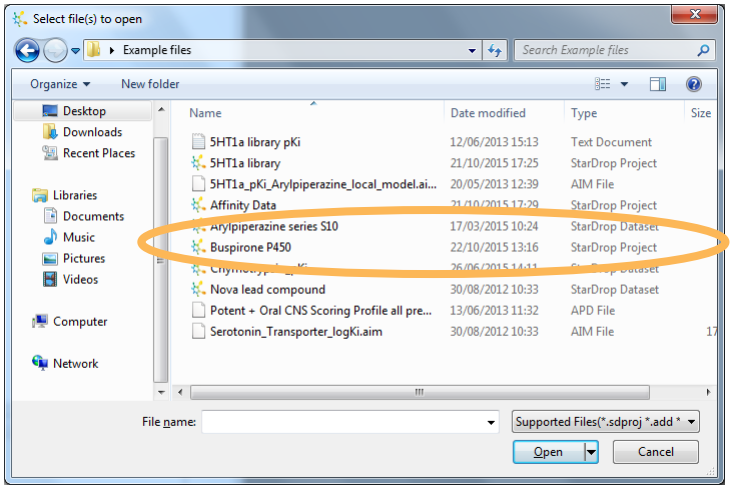
This example illustrates the use of the P450 metabolism models to explore structural modifications in each of these regions in order to identify those most likely to significantly improve the stability with respect to CYP3A4 metabolism.

Exercise

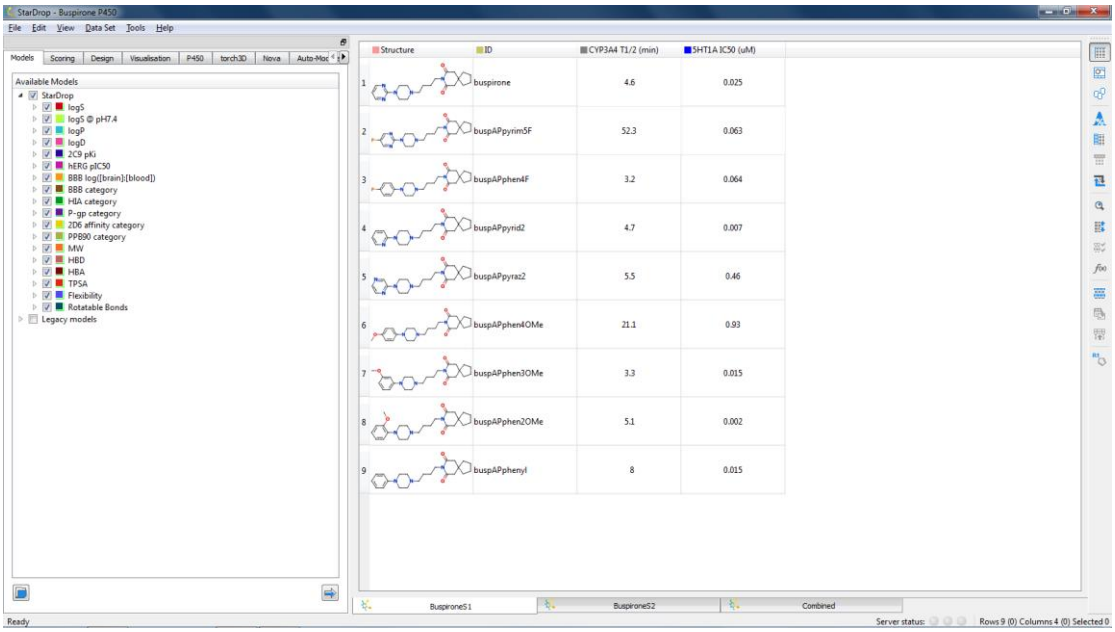
Our objective is to identify structural modifications that reduce the vulnerability of key sites of metabolism, as indicated by decreasing the **site lability** and, ultimately, identify molecules that are likely to meet the project goal of increased half-life with respect to metabolism by CYP3A4 by reducing the **composite site lability** (CSL).

We will explore modifications to the different regions of Buspirone identified above using two different series: Series 1 will explore alternative aryl substitutions on the piperazine; Series 2 will explore modifications to the tetramethylene linker and piperidinedione moiety. These series were designed to maintain potency against 5-HT_{1A} as well as improve metabolic stability.

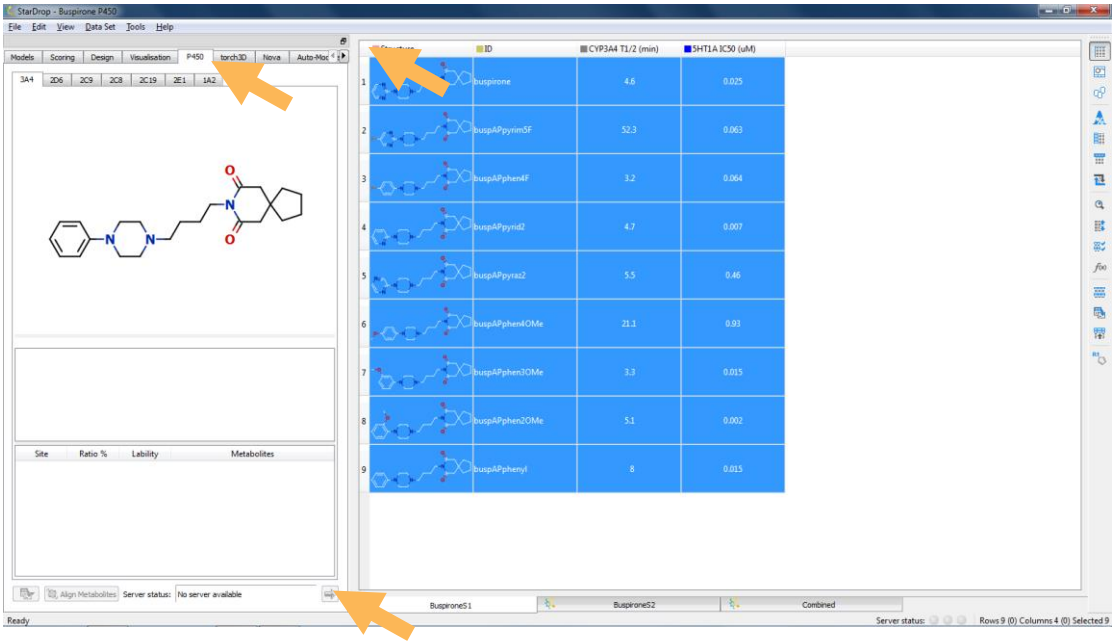
- Open the file **Buspirone P450.sdproj** by using the **File -> Open** menu option.



- First choose the **BuspironeS1** data set, containing the compounds in Series 1.You will see a spreadsheet containing structures, identifiers and their measured half-life with respect to metabolism by CYP3A4. The data set contains 9 compounds, the first of which is Buspirone.



- Change to the P450 tab, select all of the compounds in the data set by clicking in the top left corner of the spreadsheet and submit these to the P450 models by clicking on the  button.



When the calculations are complete, the results will be returned from the server and a summary will be displayed in the spreadsheet. Each molecule will take roughly 2-3 minutes to calculate; however, if the results for a molecule have previously been calculated on your server, the results will be returned instantly.

- Selecting a row in the spreadsheet will display the detailed results for a single molecule in the P450 tab (**Hint:** the regions of the P450 tab can be resized to enlarge the regioselectivity and site lability views). The results for Buspirone are shown below:

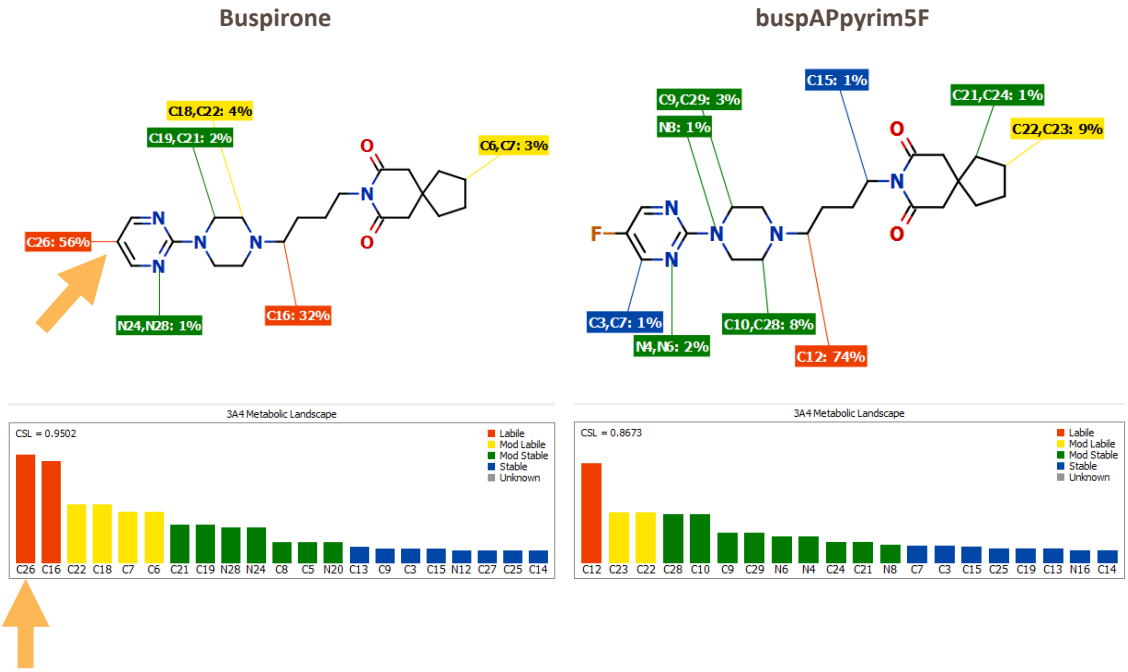
18. Which are the major sites of metabolism on Buspirone predicted for CYP3A4?

Answer:_____

_____

_____

- Examine each of the remaining compounds in turn to identify modifications to the aryl group that improve the vulnerability of this region of the molecule to metabolism by CYP3A4. This will be indicated by lower **site lability** bars shown for the corresponding sites in the **Metabolic Landscape** view. An example of such a modification is shown below:



19. Which are the most promising modifications of the aryl ring to reduce lability in this region of the molecule?

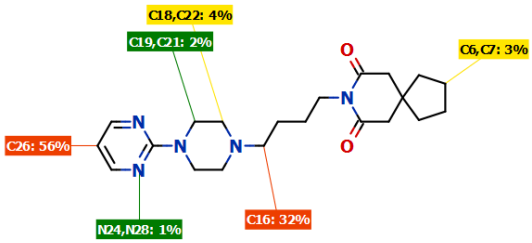
Answer:_____
_____
_____

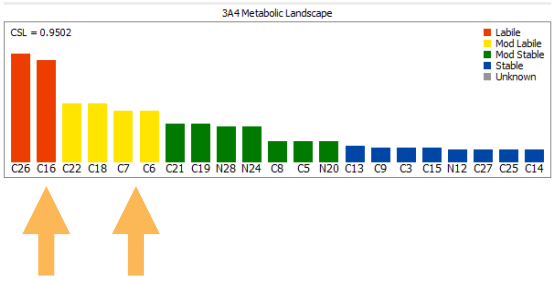
The **composite site lability** (CSL) for a molecule is a measure of the efficiency of the product formation step in the catalytic cycle of CYP3A4. Thus, a lower CSL value indicates greater stability. In this case, as we are modifying only one region of the molecule, other moderately labile sites remain, so there may be only a small change in the overall CSL, even for a beneficial modification. Also, other factors influence the overall rate of metabolism (in particular logP and pKa) therefore we do not necessarily expect a direct correlation between the small changes to CSL and the CYP3A4 half-life at this stage.

- Change to the **BuspironeS2 data** set.. This contains the compounds in Series 2 that explore modifications to the tetramethylene linker and piperidinedione moiety.

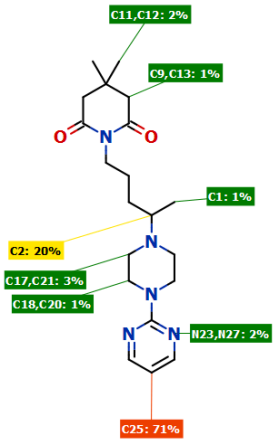
- Run the P450 calculations for Series 2, as described above for Series 1, and explore the resulting Metabolic Landscapes. An example is shown below:

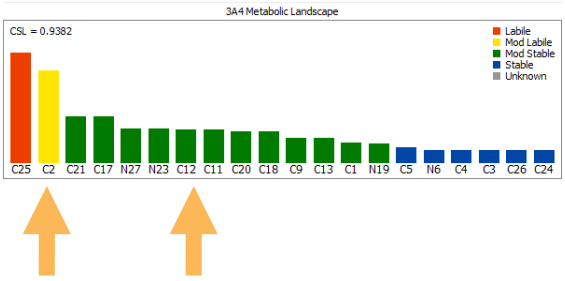
Buspirone





buspLSamegemdimet

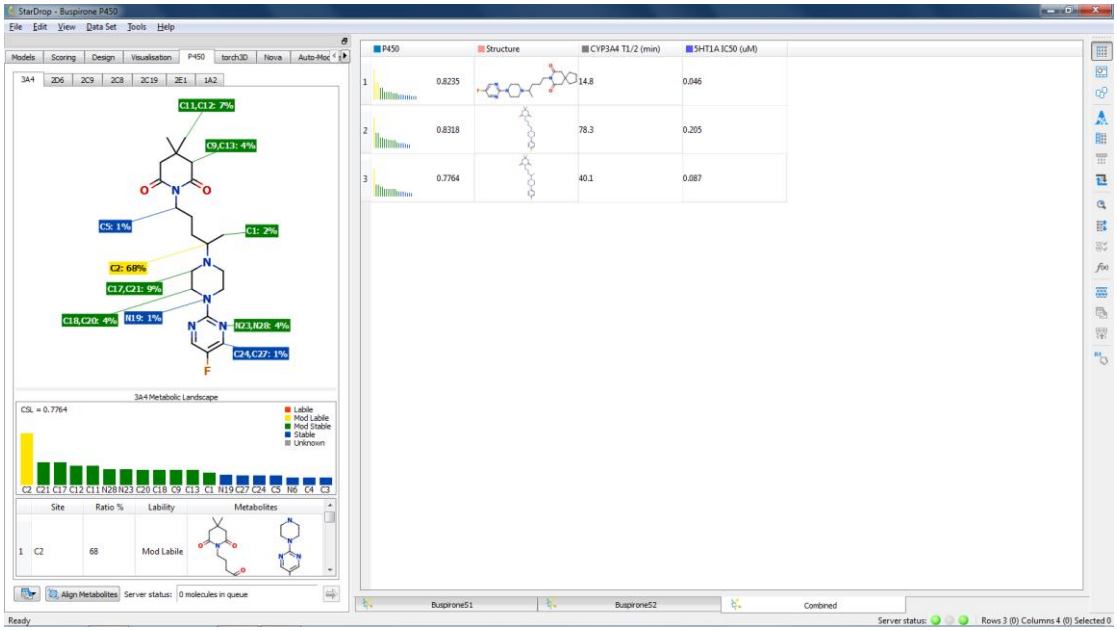




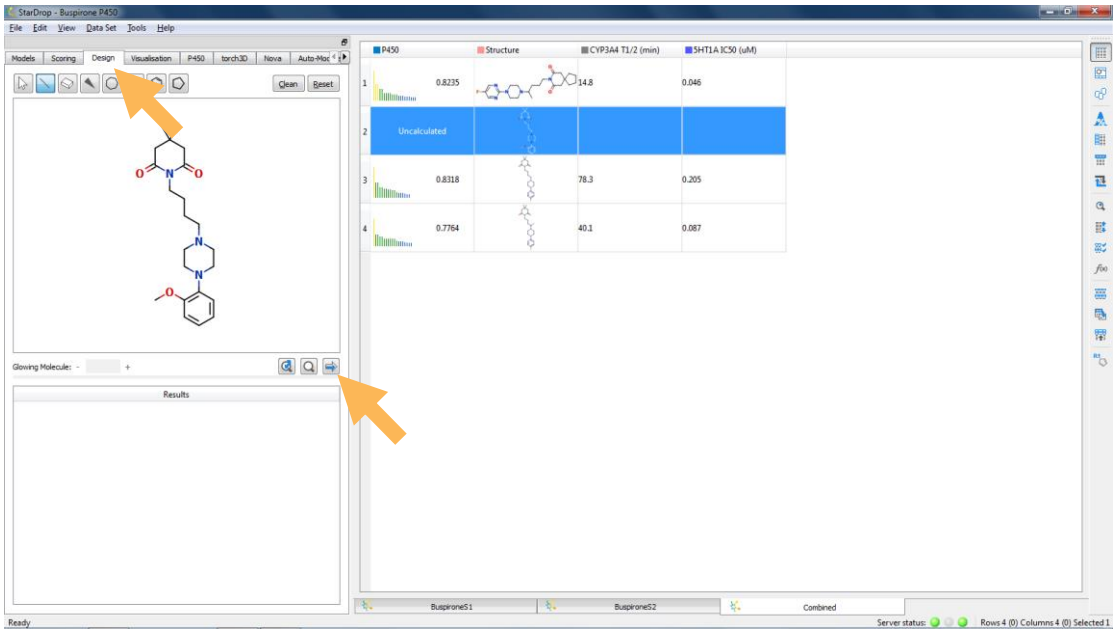
20. Which are the most promising modifications of the tetramethylene linker and piperidinedione moiety to reduce lability in these regions of the molecule?

Answer:_____
_____
_____

- Finally, we would like to explore combinations of the modifications we have identified to find compounds with improved overall stability, while avoiding those changes that caused a large decrease in potency. Change to the **Combined** data set to load three such examples and run the P450 models as described above.



- Note that all of these compounds have significantly better (lower) CSL values than Buspirone and meet the objective of greater than 3-times the half-life of Buspirone. Furthermore, in two cases, IC_{50} values against 5-HT_{1A} of less than 0.1 µM have been retained.
- Further modifications can be explored by drawing new molecules in the **Design** tab. Add these to the dataset using the  button before switching to the **P450** tab and submitting the molecules to the P450 models.



Further details of the chemistry, assays and results in this study can be found in Tandon *et al.* The design and preparation of metabolically protected new arylpiperazine 5-HT1A ligands. Bioorg. Med. Chem. Lett. 2004 **14**(7) pp. 1709-12.

Applying Matched Series Analysis to Improve Target Activity

Objectives

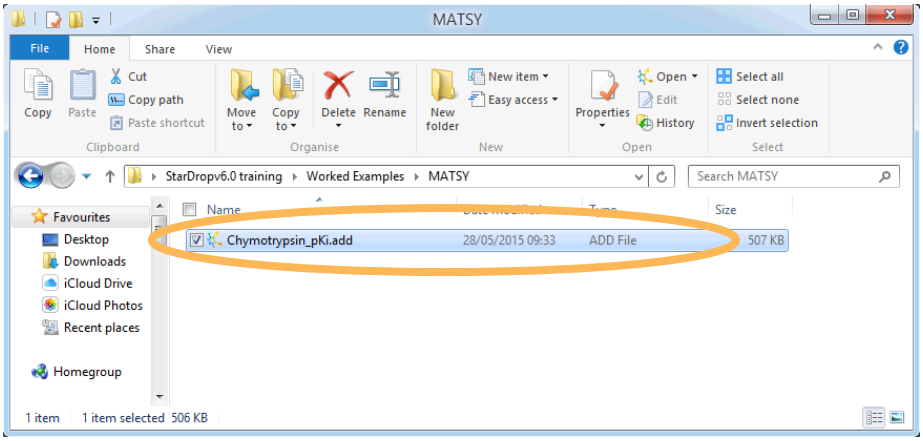
- Using Nova's Matched Series Analysis to generate new compound ideas
- Controlling the different approaches for generating suggestions for matched series
- Exploring Matched Series Analysis results

Background

This example uses a publically available set of Human Chymotrypsin K_i data and searches the ChEMBL pIC_{50} knowledge base to find matched series that indicate new substitutions with a high likelihood of improving the binding at Chymotrypsin.

Exercise

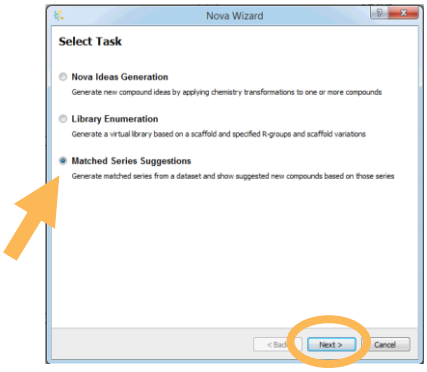
- Open the file **Chymotrypsin_pKi.add** by using the **File -> Open** menu.



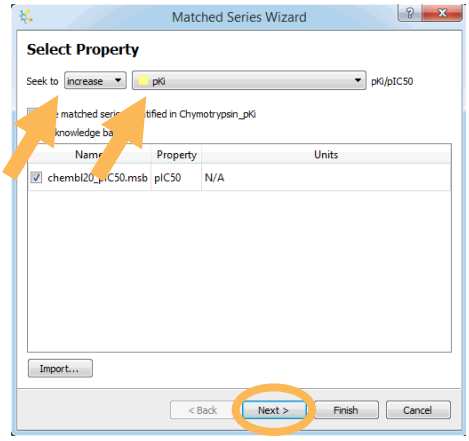
You will see a spreadsheet containing 115 structures and their measured affinities for Human

`Chymotrypsin C (in the column labelled pKi).

- Change to the **Nova** tab and then the arrow button  at the bottom of the tab.



- Select the **Matched Series Suggestions** option and click **Next**

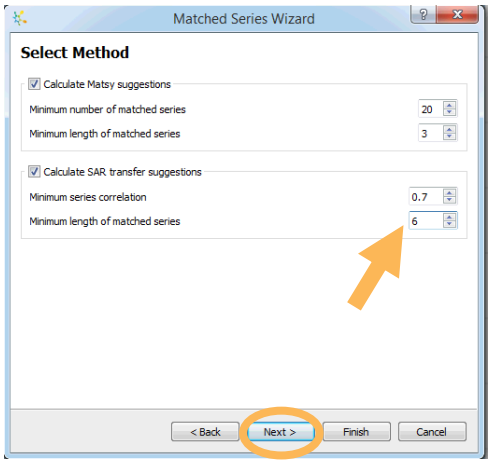


In the dialogue box that appears, you can specify the column containing the property you wish to improve. In this case, the column we are interested in, **pKi**, is already chosen and we want to find suggestions that **increase** this value so this default option is also correct.

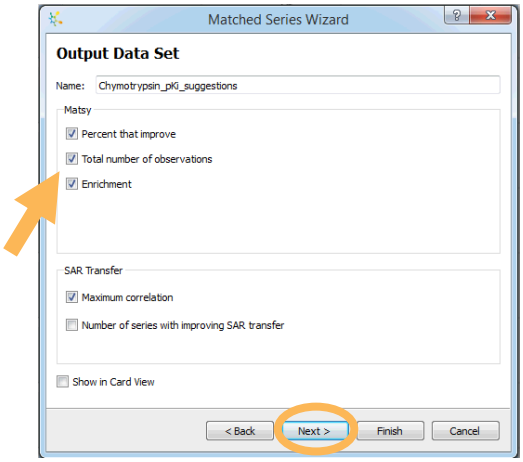
- Select **Next** to continue.

At this point you can change the limitations placed on the suggestions returned. In Matsy™ the support for a suggestion comes from the number of times it has been seen; the more frequent the occurrence of the order of the input series, the more likely it is that the suggestion will be an improvement. Hence, to find many examples in the ChEMBL knowledge base, the compared series are generally short.

With SAR transfer, the support for a suggestion comes from a long series of derivatives that shows a consistent trend with that seen in the input data set. This example data set is too small to have matched series with the default minimum number of derivatives (8), so for this example we will decrease this limit.

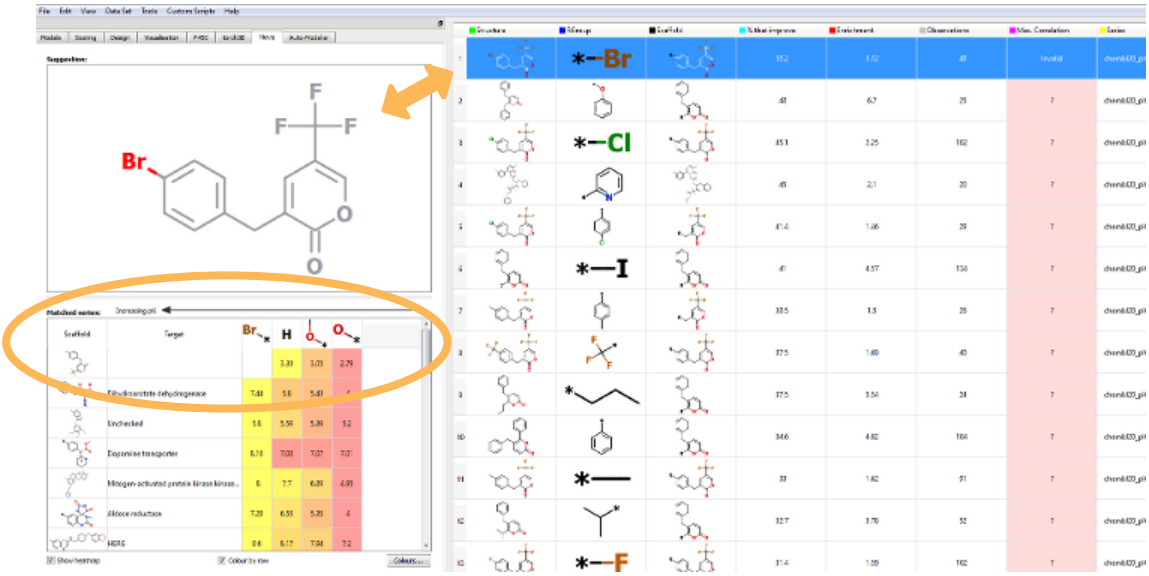


- Click on the **Minimum length of matched series** box in the **SAR transfer** section and change the value to **6** as shown in the image above.
- Click the **Next** button to continue.



- Check all the boxes for the **Matsy** output as shown above and click **Next.**
- Here you can choose structural filters to exclude certain chemical groups. In this case we will use the defaults, so click **Finish** to begin the matched series analysis.

The suggestions are returned in a table with the Matsy based suggestions first, followed by the SAR transfer suggestions. The Matsy suggestions are ordered by the **% that improve** column and the SAR transfer are ordered by the **Max Correlation** column. When a row is selected in the data set the suggestion is displayed in the main Nova tab and the supporting evidence is shown in a table below it.



SAR data from the input data set is in the first row (which is why the target and first substituent columns are empty) and the SAR data is ordered with the least active/desirable on the right to the most active/desirable on the left (as indicated by the colour coding in the table cells).

The first row of the data set shows one of the suggestions that is most likely to improve the pKi which is the creation of the Bromine derivative. This suggestion is based on the order of activity seen for the hydroxyl, methoxy, and unsubstituted derivatives, at that position on the displayed scaffold, seen in the input data set. The Bromine substitution on this scaffold has the greatest weight of evidence suggesting that this might be worth investigating next.

21. Which other variations on this scaffold might also be worth considering?

Answer:_____
_____
_____

Hint: Select the first row and the sort the data set by the scaffold column to show all the other suggestions for this scaffold ordered by the percentage of times that an improvement has been seen.

22. Which of the suggestions resulting from the SAR transfer method has the strongest evidence that it may improve activity?

Answer:_____
_____
_____

Hint: To see the SAR transfer suggestions, **right click** on the **Max correlation** column in the main data table and choose the **Sort->descending** menu item.

Applying Medicinal Chemistry Transformation Rules to Guide Optimisation

Objectives

- Using Nova to generate new compound ideas
- Controlling the way new molecules are generated
- Exploring Nova results

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.

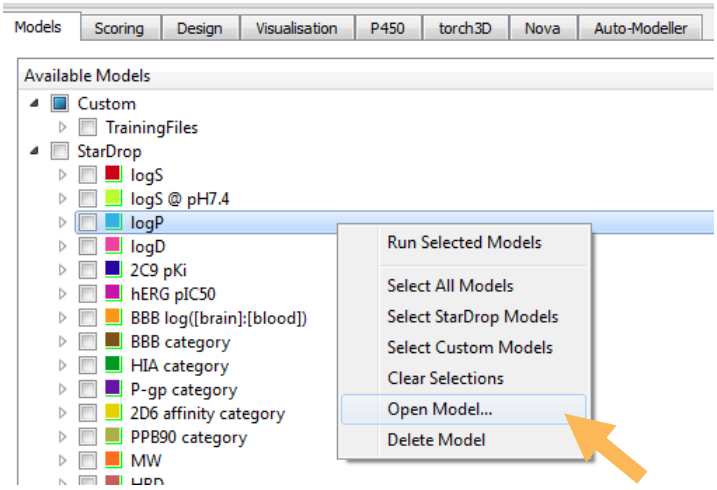
We will use a model for Serotonin transporter inhibition built with public domain data from the ChEMBL database using StarDrop's Auto-Modeller to monitor potency during the exploration.

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.

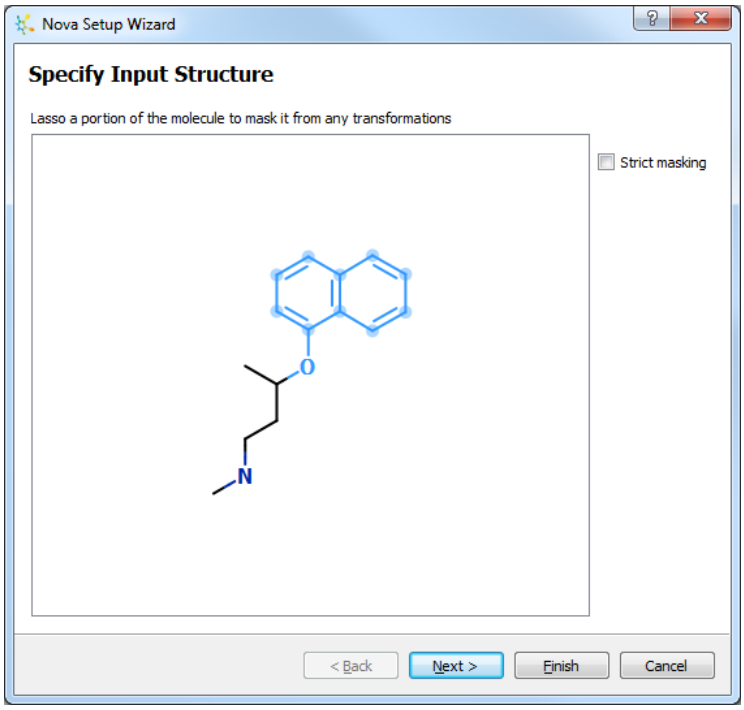
Exercise

- 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**

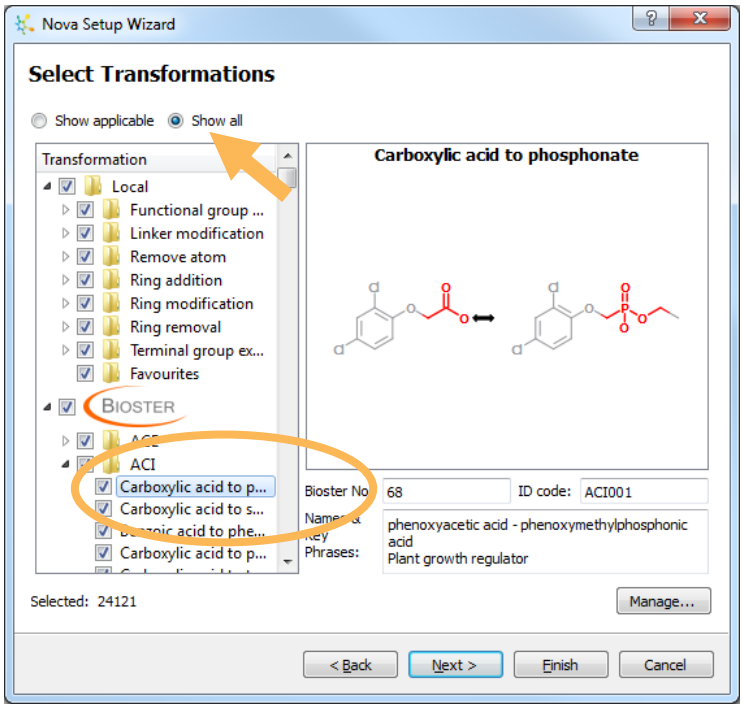


- The model will appear in the **Custom** model section.
- Open the data file **Nova lead compound.add**
- Select the only row in the data set and on the **Nova** tab, click the  button to start the Nova wizard. (**Hint:** like the P450 models, when you start Nova it only applies to those rows which 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.** Click **Next**.

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

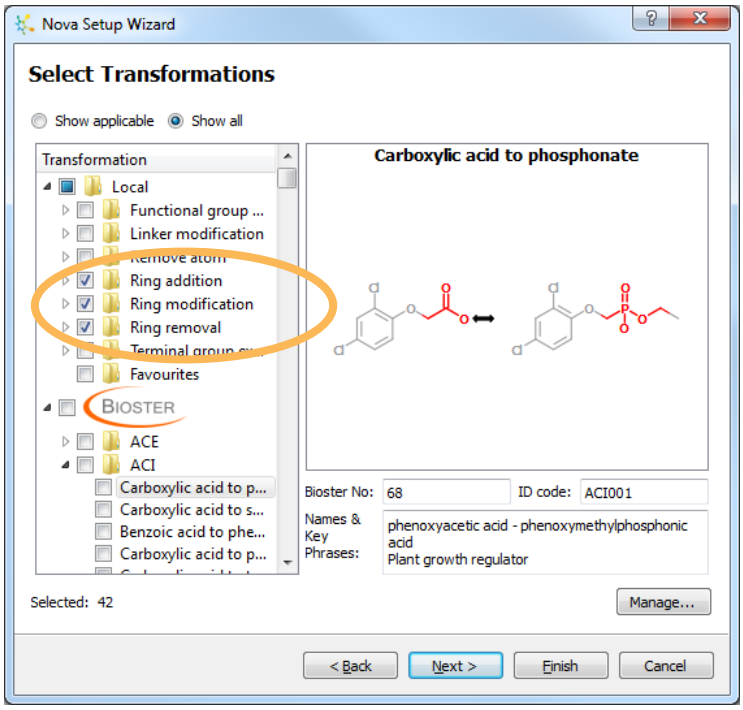


- Click **Next** to go to the **Select Transformations** page. Nova will search for, and display, only those transformations which are applicable to the input structure. However, if you are going to allow Nova to run for multiple generations (as we will in this example) it is sometimes useful to select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



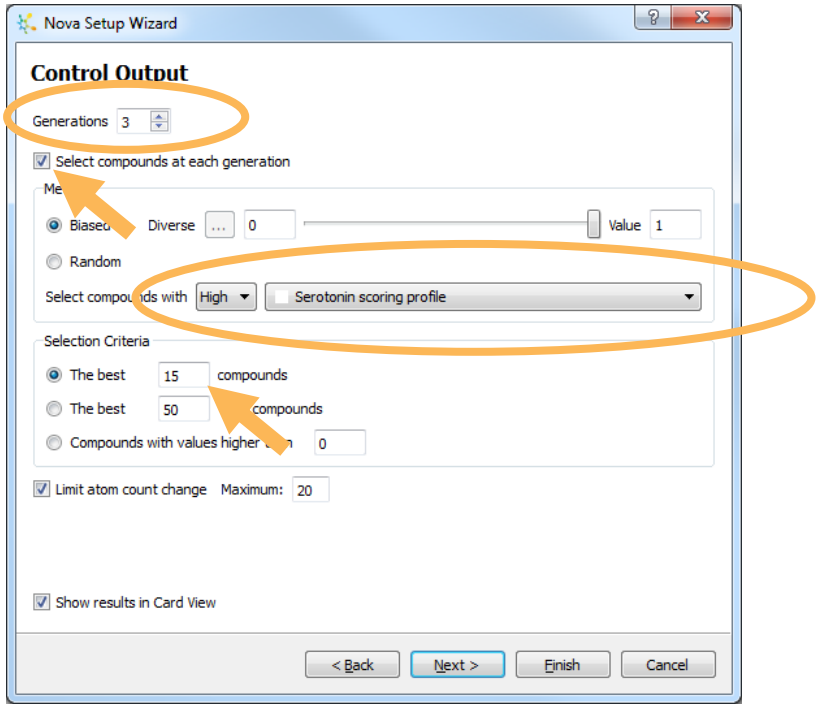
Hint: selecting an individual transformation will enable you to see an example along with any other available details.

- 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 just 200 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.

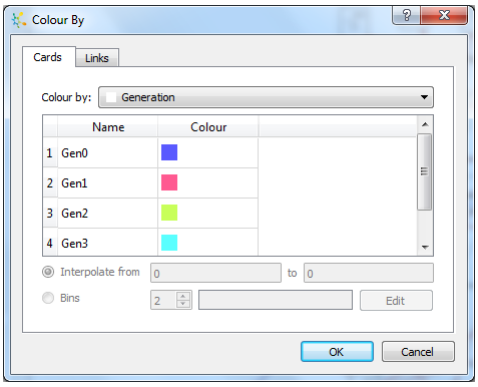


- Start the process by clicking **Finish**. The Nova job will take a couple of minutes to run…

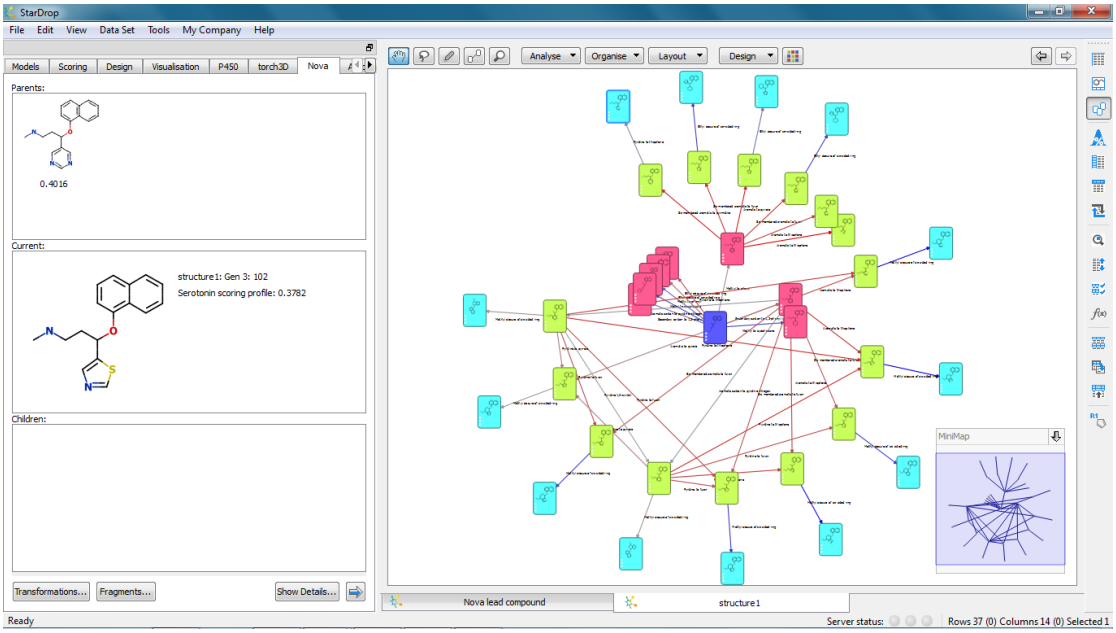
- While in progress an indicator will show the highest score that has been achieved so far. Once complete, a new data set will be displayed using Card View.

A radial network will be displayed with the initial compound at the centre and each subsequent generation in a ring around this.

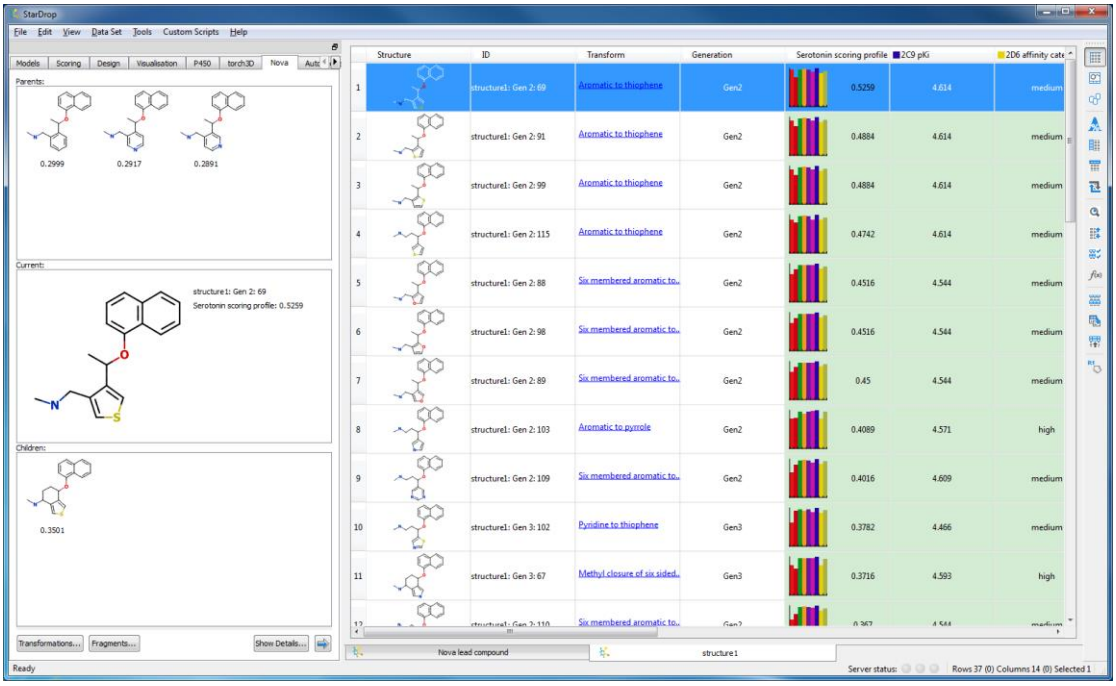
- You can easily see this by clicking on the colour palette () and colouring the cards by **Generation**.



The links between the cards indicate that a transformation was applied, with the arrow indicating the resulting molecule. The links are coloured such that red indicates a transformation which increases the score and blue indicates a transformation which decreases the score. We can see that in this network the highest scoring compounds were found in the second generation because most of the links to the last (third) generation of ideas are blue. Where there are multiple links to a compound, or links between compounds within the same generation, this shows that the same molecule was created via more than one combination of transformations.



- Switch to table view by clicking the  button



In table view the rows are sorted with the highest scoring idea at the top. 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).

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

Answer:_____

_____

_____

_____

_____

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

Answer:_____

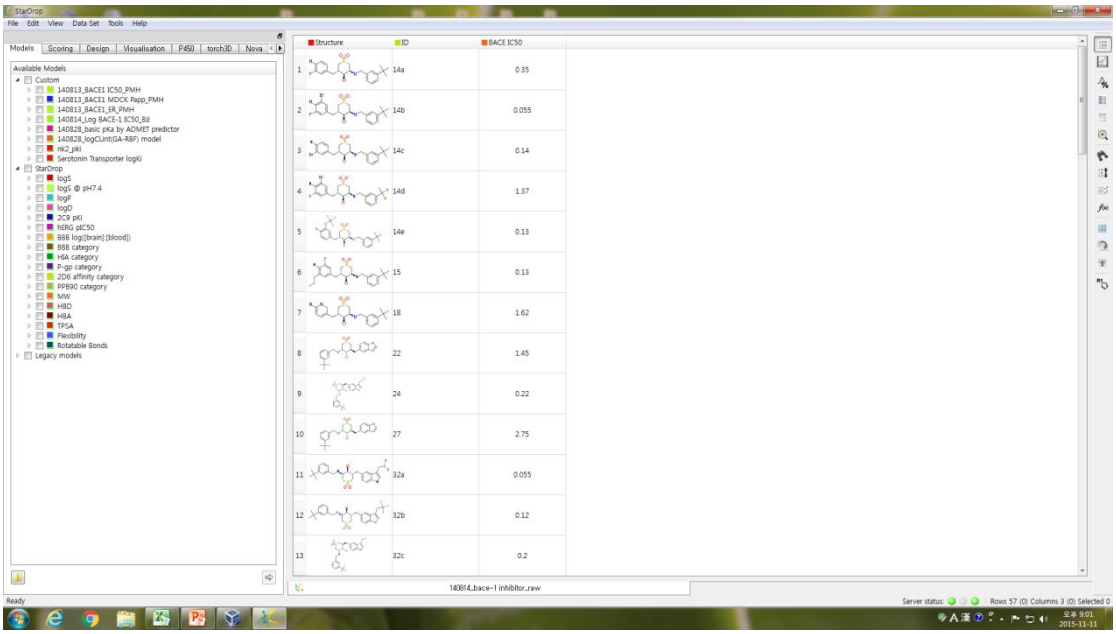
Prioritising Compounds by a Combination of Potency (IC_{50}), *in Vitro* Cl Prediction and ADME Properties by Building Predictive Models.

Objectives

In this example we will explore a strategy for prioritising compounds by a combination of potency (IC_{50}), *in vitro* clearance prediction and ADME properties by building predictive models with a user's own data set.

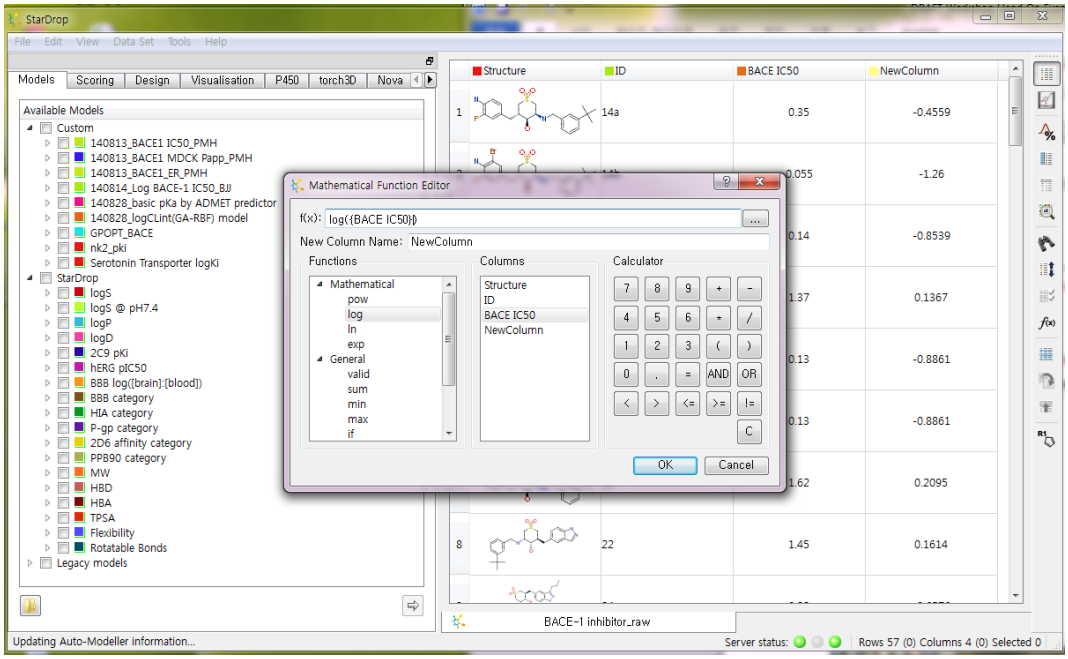
Exercise

- Open the file **1.BACE-1 inhibitor_raw.add** by using the **File -> Open** menu option.

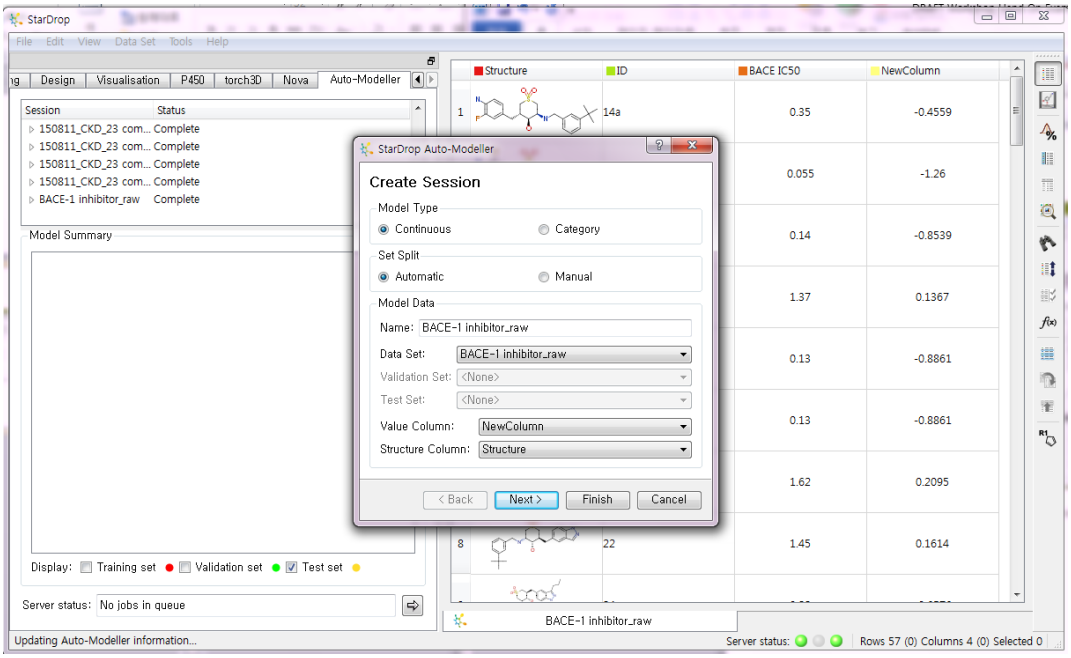


The IC_{50} values are from the experimental results in the manuscript.

- First convert the IC_{50} values to log(IC_{50}).



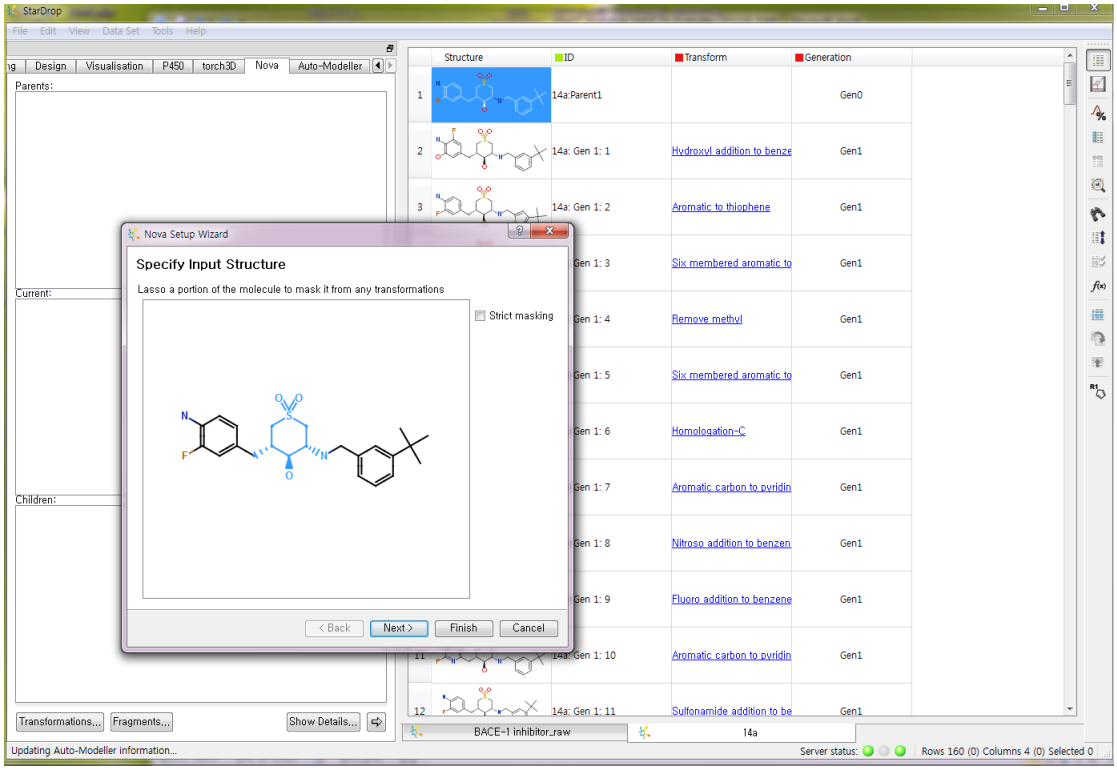
- After conversion, create your own predictive log(IC_{50}) models using the Auto-Modeller.



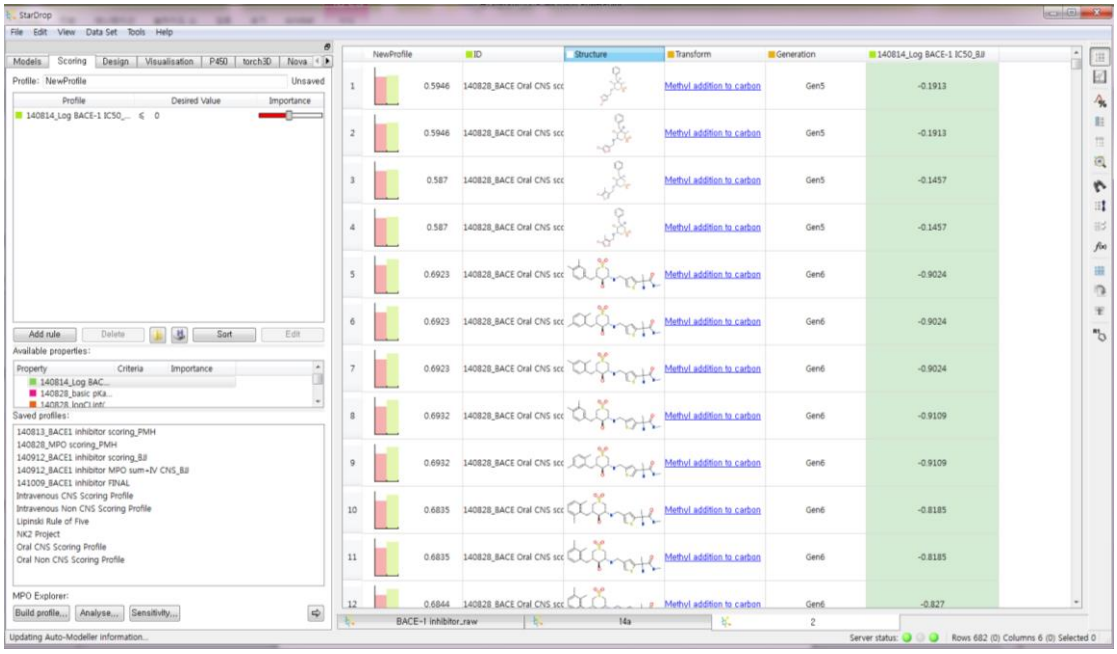
- Compare the R^2 on the validation set and training sets to choose the most appropriate model for your log(IC_{50}) predictions. Check that the performance is also good on the independent test set.

It is also possible to create a similar predictive model of intrinsic clearance (CL) model using your own *in vitro* CL data set, but we will not do that in this case.

- Create a virtual library using Nova. Select the pharmacophore in the parent structure.



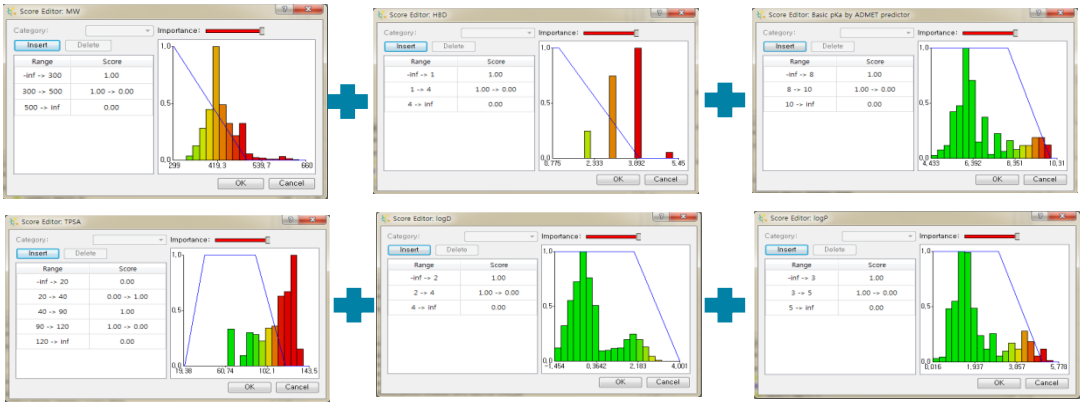
- After library generation, apply the mode of log(IC_{50}) to the library you generate.



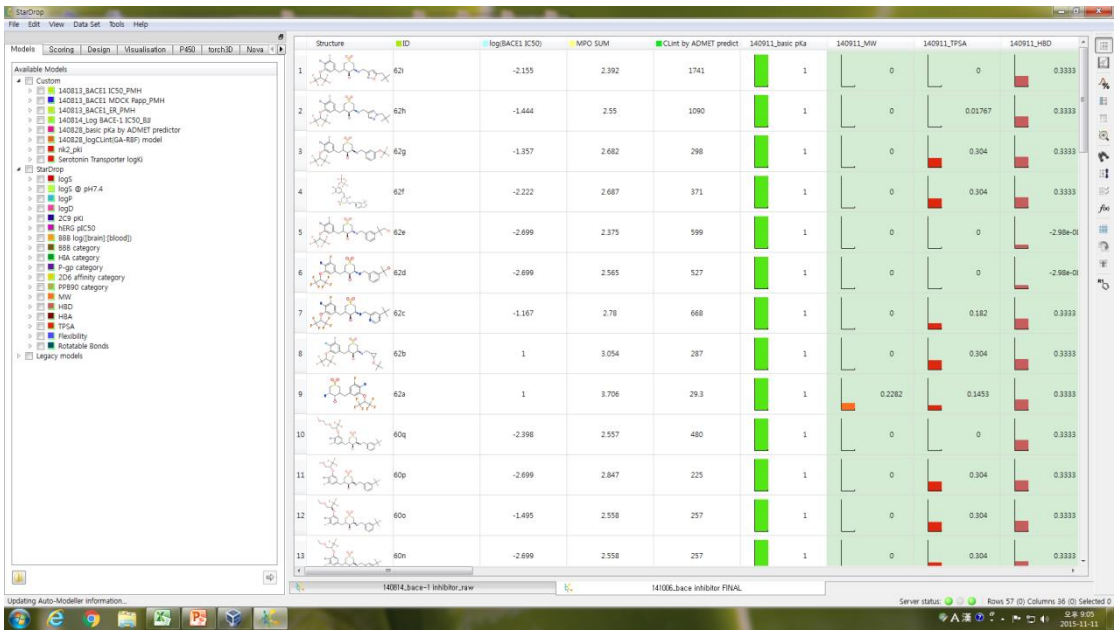
- Create MPO score profile.

.

$$\mathbf{MPO} = \sum \text{Score (clogP + clogD + PSA + MW + HBD + pKa)}$$



- Select log(BACE1 IC50), MPO SUM and log(Clint by ADMET predictor) to create a scoring profile with which to prioritise the compounds. You can also add other built-in ADME properties for rank order



- Sort all compounds virtually synthesized by Nova, using the score based on the three parameters selected. Try to explore other ADME properties for these compounds

Answers

1. Which chemistry has the highest average potency?

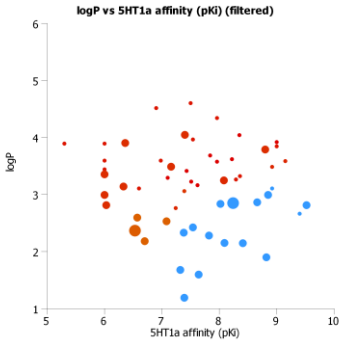
   Answer: 2(arylcycloalkylamine) 1-indanol

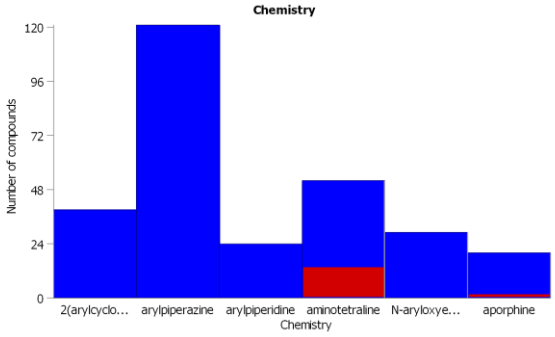
2. What is the identifier of the most potent compound?

   Answer: S5-34

3. Which chemistry includes the majority of compounds with high potency, low MW and appropriate logP values?

   Answer: Aminotetraline





4. All three compounds are very potent so on the basis of these three examples which chemistry type looks promising as a potential lead?

   Answer: Confirming our previous analysis, the aminotetraline S1-9 appears to offer a good starting point for lead optimisation, with high potency, low logP and low molecular weight.

5. Which stacks appear to contain compounds with the best potency and range of logP values?

   Answer: Based on potency and logP alone, both the 2(arylcycloalkylamine), 1-indanol, and aminotetraline stacks contain a reasonable proportion of potent compounds and a good range of logP values.

6. 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.

7. Which compound has the highest score?

   Answer: S1-26

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

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

    Answer: The predicted hERG pIC50 decreases to 5.382

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

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

12. 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

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

    Answer: The predicted hERG pIC50 decreases to 5.366

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

    Answer: The overall score increases slightly to 0.2525

15. 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.

16. 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.

17. 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.

18. Which are the major sites of metabolism on Buspirone predicted for CYP3A4?

    Answer: The major sites of metabolism are predicted to be the para position on the pyrimidine (C26) and the carbon at the alpha position to the piperazine nitrogen on the tetramethylene linker (C16). Other moderately labile sites are on the spirocyclopentane (C6,C7) and on the piperazine ring (C18,C22).

19. Which are the most promising modifications of the aryl ring to reduce lability in this region of the molecule?

    Answer: The 5-fluoro-pyrimidine and 4-fluoro-phenyl substitutions appear to give the greatest reduction in lability on the aryl ring.

20. Which are the most promising modifications of the tetramethylene linker and piperidinedione moiety to reduce lability in these regions of the molecule?

Answer: Introduction of steric bulk or polarity near to the alpha carbon appears to reduce the lability at this position. Replacement of the spirocyclopentane with a gem-dimethyl substitution is predicted to reduce the lability in that region.

21. Which other variations on this scaffold might also be worth considering?

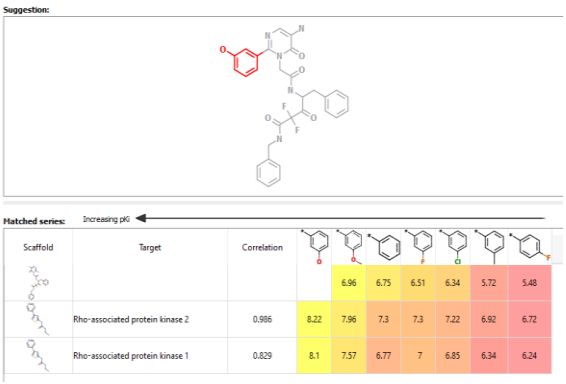
Answer: Whilst one might be surprised to find that Bromine is a suggestion, given the fact that the SAR shows that the smallest group is the most active in the series, this effect has been seen many times (over 56% of the 48 observations of this series) and may well be due to this group binding to a flexible hydrophobic area of a protein that can move to accommodate the larger bromine. The increased lipophilicity of the bromo-derivative may also help drive the increase in binding by hydrophobic collapse of the ligand with the protein in water.

The chloro, tri-fluoro, methyl and fluoro substitutions are also recommended having being observed 40 times or more (chloro and fluoro more than 100 times). In over 30 percent of cases all of these have resulted in an improvement in activity for this series.

22. Which of the suggestions resulting from the SAR transfer method has the strongest evidence that it may improve activity?

Answer: In the top suggestion you can see that the series of aryl derivatives from the input data correlates very well with the activities of derivatives at Rho Kinase.

Given the improvement in activity seen in the input data set, in moving from the more hydrophobic aryls to the anisole, the suggested phenol follows that trend for more polarity.



23. 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.

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

Answer: The molecule generated in row 19 is actually the drug Duloxetine.