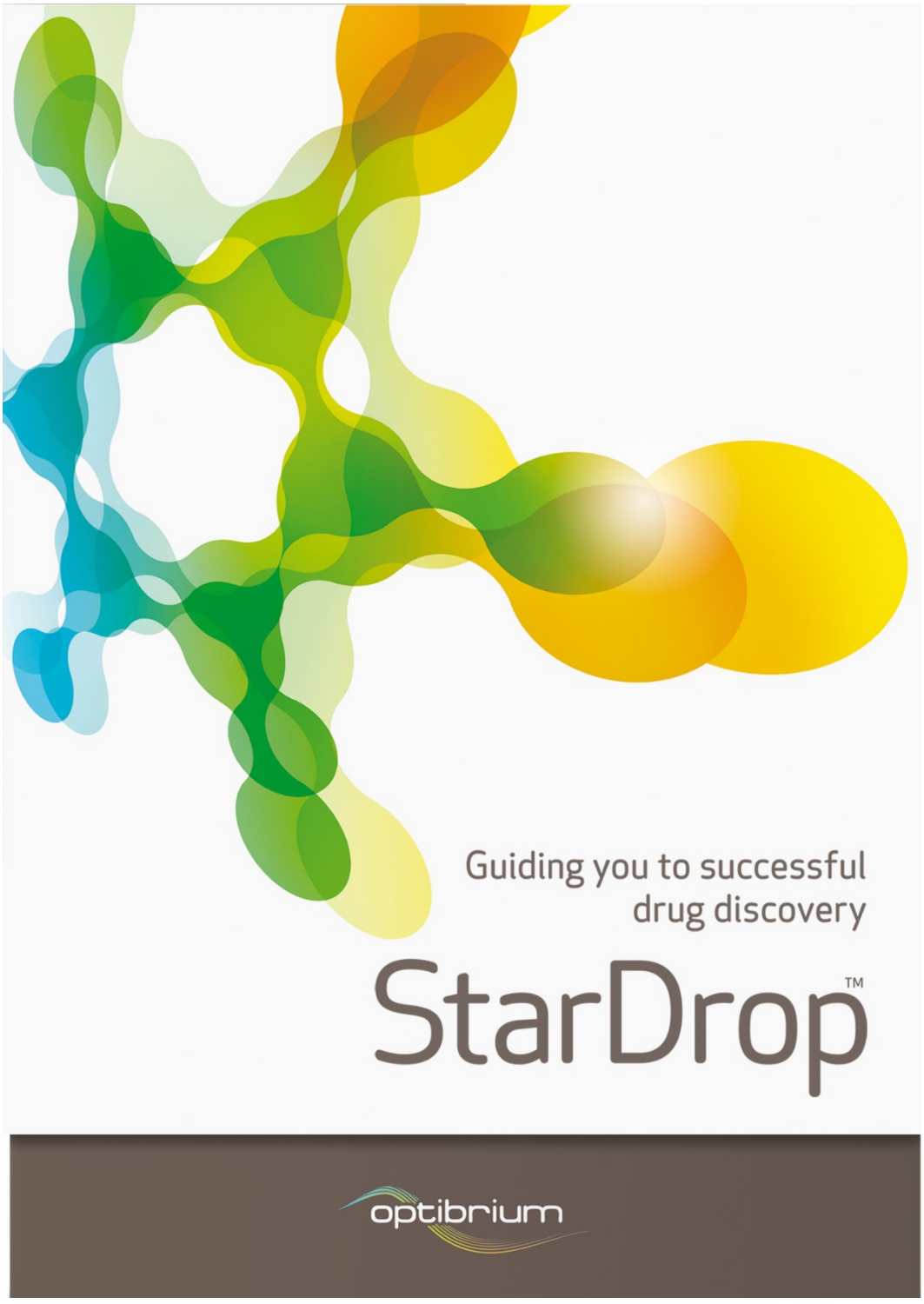
StarDrop™ User Guide

Version 6.4



Contents

1 Installation

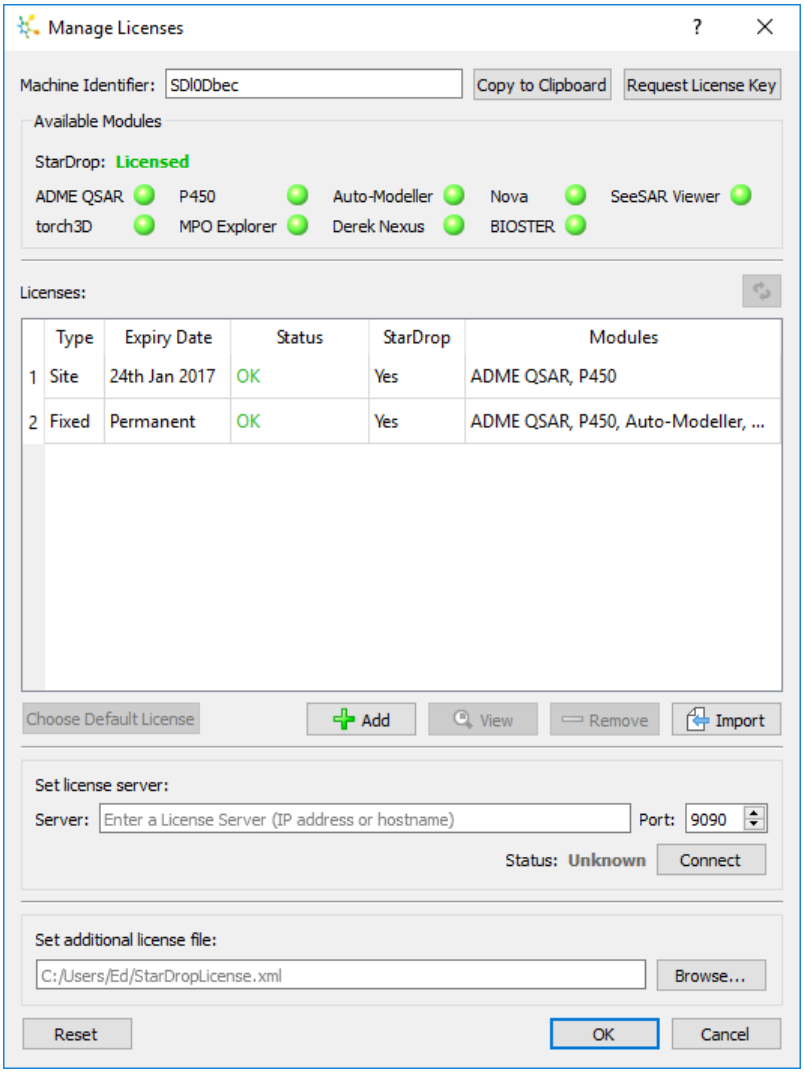
To install StarDrop on Windows® double-click the installation file **StarDrop-6.4-installer-32-bit.msi** for 32-bit Windows® installations or **StarDrop-6.4-installer-64-bit.msi** for 64-bit Windows® installations. A menu item for running StarDrop will be added to the chosen folder (by default this folder is called StarDrop). On a Mac®, double-click the image file **StarDrop-6.4.dmg** and drag the StarDrop app to the directory in which you want to install the application (typically the Applications folder).

1.1 License keys

The first time you run StarDrop, or when there isn't a valid license key, it will start in **Viewer mode.** While in viewer mode you will be able to view previously saved files but all major functionality will be disabled. Once a valid key has been installed, all functionality will be restored.

1.1.1 Installing keys

You can install, or update, your StarDrop license key by choosing **Manage Licenses...** from the **Help** menu. This opens the **Manage Licenses** dialogue which enables you to add your license key.



You can enter a key by clicking the **Add** button and copying and pasting the details. Alternatively, if you have a license file which contains one or more keys you can import this by clicking the **Import** button. You can also use a license file if it has been saved into a folder on your computer or a network drive (perhaps for sharing with other users) by indicating this in the **Set additional license file** box.

1.1.2 Fixed license

A fixed license key will work on one machine only. The **Manage Licenses** dialogue will display a **Machine Identifier** which is linked to the key. Once you have added a valid fixed key, the expiry date and the list of optional modules enabled by the license key will be displayed. The key will remain valid until it expires or the hardware configuration is changed.

1.1.3 Site license

A site license key will work on all machines at a single site. Once you have added a valid site key, along with its associated password, the expiry date and the list of optional modules enabled by the license key will be displayed.

1.1.4 Floating license

A floating license key requires StarDrop to contact a floating license server to request a key each time you start StarDrop. These keys are shared with other users but the number of concurrent users is managed by the server. Once you have added a valid floating license key, along with its associated password, the expiry date and the list of optional modules enabled by the license key will be displayed. Having done this, you must now specify details of your floating license server in the **Server** field below. If unsure of these details you should contact your network administrator.

If the maximum number of concurrent users has already been reached, StarDrop will start in **Viewer mode**. This will enable you to view previously saved files but all major functionality will be disabled. To start StarDrop in **Viewer mode** and avoid consuming a floating license, you can type '<installation directory>/stardrop.exe --viewer' at a command prompt (replacing <installation directory with the path on your PC to the directory where StarDrop is installed).

1.2 Servers

In order to use the P450 models and Auto-Modeller™ you will need to know the name and port numbers for the computer on which the P450 Server and Automatic Model Generation (AMG) Server have been installed. Running models on the StarDrop Model Server also requires equivalent information. If unsure of these details you should contact your network administrator. Once these details have been entered in the **Preferences** (see section 23) it will not be necessary to do so again. These details are stored in the file **serverconfig.xml** which is saved in a StarDrop folder in your home area. This file can be copied between users and may be provided by a network administrator.

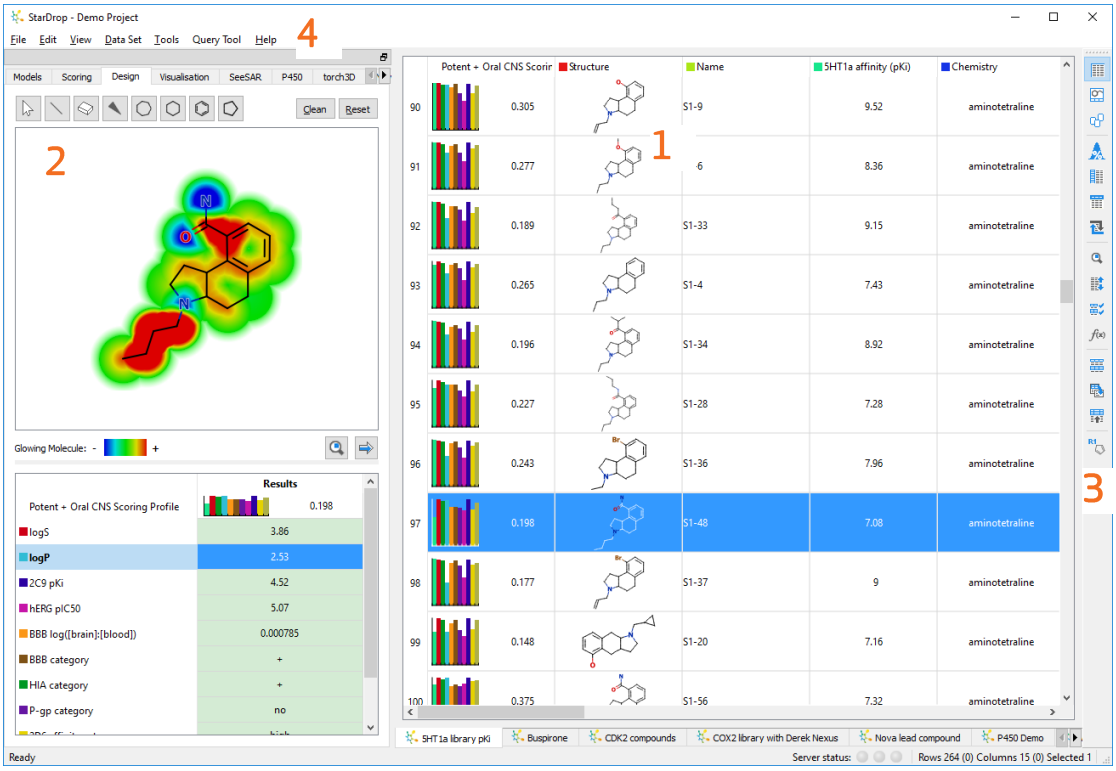
2 Introduction

StarDrop is designed to help you to quickly identify and design more effective compounds, enabling you to make decisions with confidence. Proprietary measured and predicted data can be imported and used alongside, or in place of, data generated within StarDrop.

Many of StarDrop's key features can be accessed through a different tab:

- **Models** – Generate property predictions using the StarDrop models or custom models
- **Scoring** - Score molecules against project objectives
- **Design** - Create and edit molecules to investigate the effect of structural changes on model predictions and scores
- **Visualisation** - Create plots and chemical space projections of one or more data sets
- **SeeSAR** – View 3D protein-ligand interactions
- **P450** – Predict sites of metabolism by the major Cytochrome P450 enzymes
- **torch3D™** - Compare the 3D fields of your molecules against a known active compound
- **Nova™** - Generate new compound ideas around existing chemistry or enumerate virtual libraries
- **Auto-Modeller™** - Build models of proprietary data

The main window of StarDrop is shown below with a guide to the key areas:



Key:

1. **Data sets** display your compounds and their data
2. **Tab area** contains all of StarDrop main functionality
3. **Tool bar** provides quick access to some of the key features
4. **Menu bar** provides general access to the features

2.1 Projects

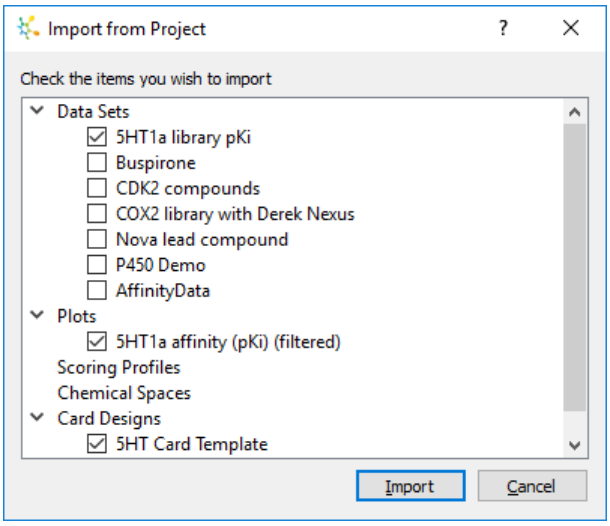
Everything that you are working with in StarDrop at any one time is considered to be part of a single project. When you save a project, all data sets, visualisations, scoring profiles, summary analyses,

functions and Card View templates will be saved with the project enabling you to save, restore and share complete StarDrop sessions.

You can only open one project at a time in StarDrop and will be prompted to save your current project before you open a new one.

When you open a data set, scoring profile, plot or template it will become included within the project in which you are currently working. If you have closed your project then when you open a data set it will become part of a new project.

You can import items into your project that have been saved as part of another project by selecting **Import From Project…** from the **File** menu. Select a project file from which to import and click **Open**.



This will display a list of items in the selected project. Tick the boxes next to those items which you would like to copy into our current project and click **Import**.

2.2 Data sets

Multiple data files can be opened at any one time within a single project and can be maximised or minimised within the data window.

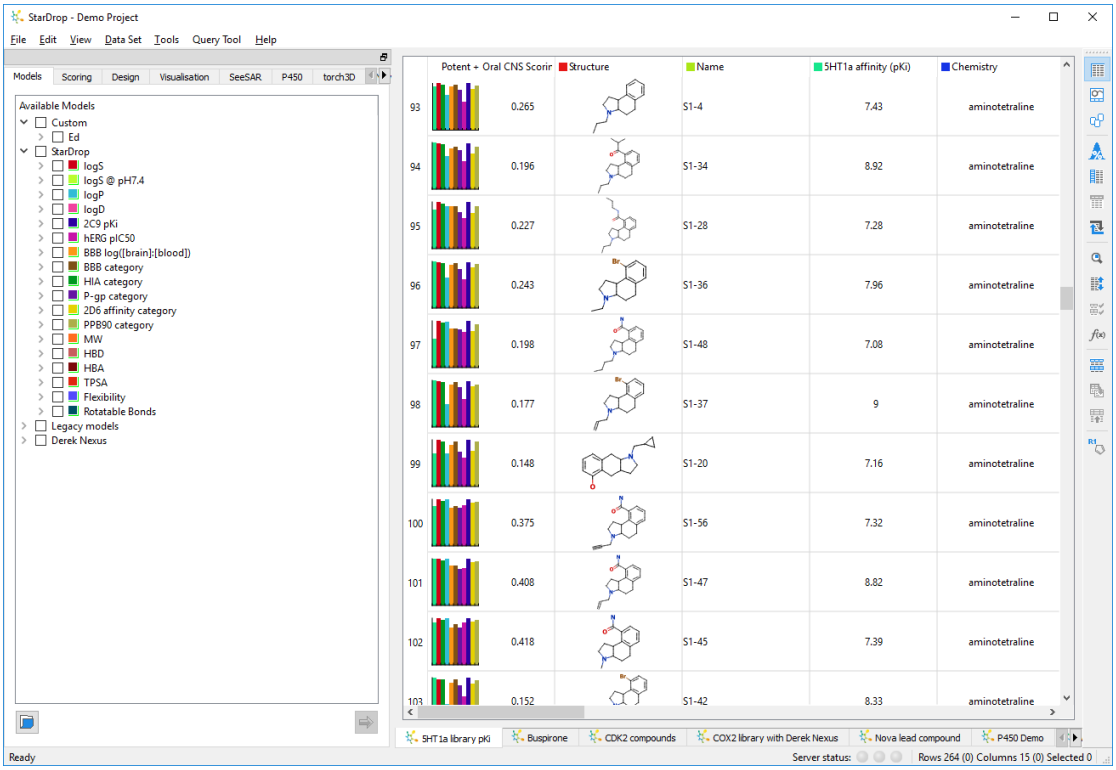
You can toggle between data sets by pressing **Ctrl+Tab**.

If multiple data sets are open then these can be selected, tiled or cascaded from the **Windows** menu.

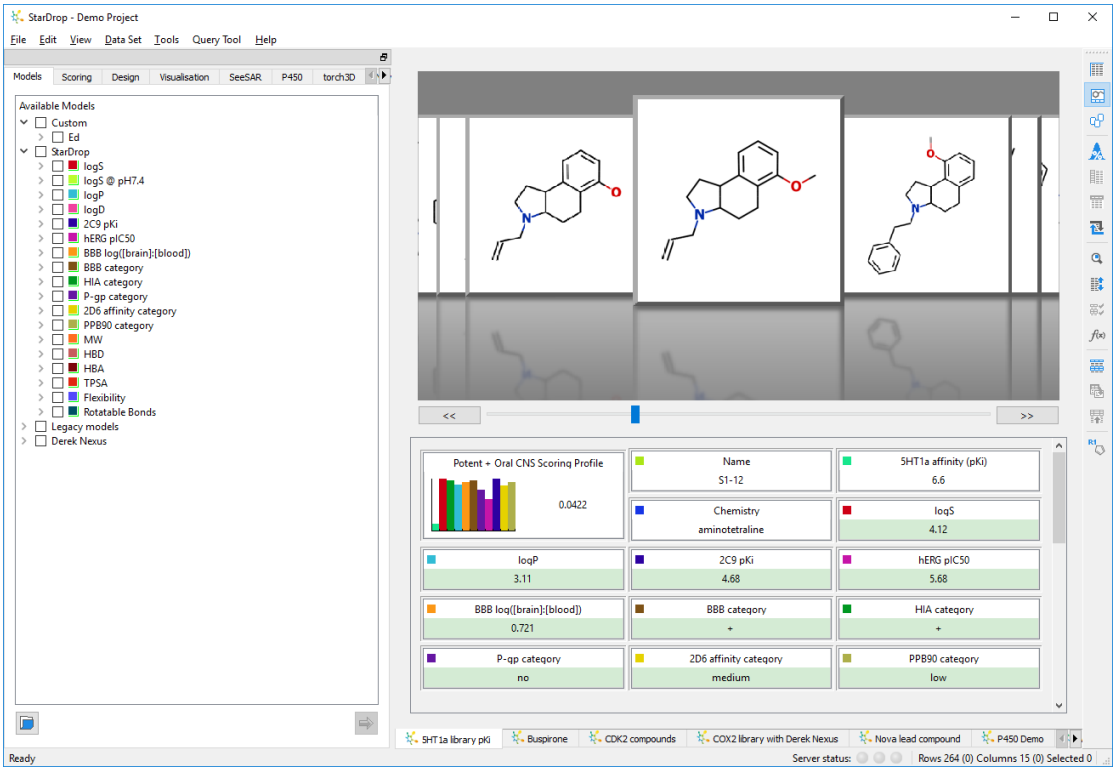
New data sets can be created by drawing molecules in the **Design** tab (see section 6). Additionally, you can create a new data set by selecting rows in an existing data set and then clicking the  button on the toolbar, or alternatively by selecting **Create From Selection…** from the **Data Set** menu. See section 15 for alternative ways of creating new data sets by selection.

2.2.1 Data set views

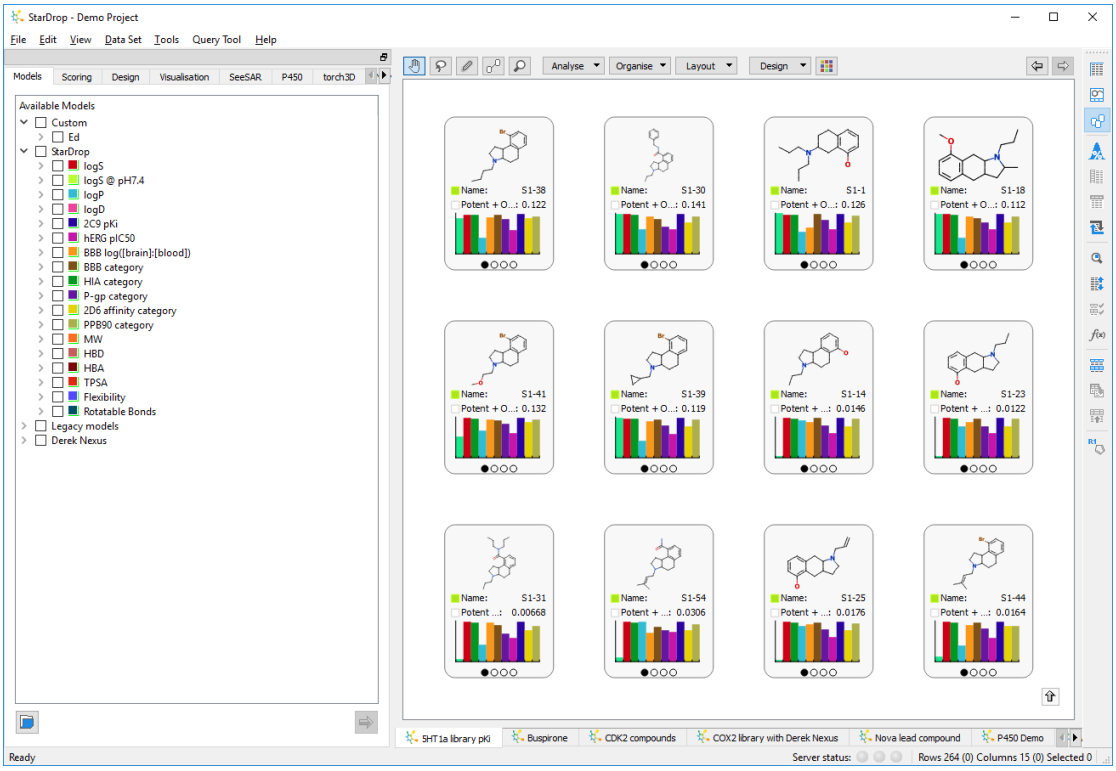
There are three different views that can be used for browsing the information in a data set. The table view shows all the information in a table where each row displays a different compound and each column displays a different data type.



The molecule view shows one compound at a time to make it easier to see all the properties for a single molecule. In molecule view you can scroll the molecules left and right. The data below will change to display values for the current molecule. You can drag and drop the data items in the grid below to arrange them as desired. Clicking on a molecule will select/deselect it. Clicking to either side of the molecule will scroll the view by one molecule in the chosen direction.



The Card View is an interactive way to view compounds and their relationships in the context of your discovery projects (see section 5).



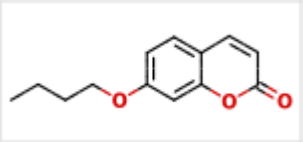
On the tool-bar are three buttons that can be used to switch between these views. The button switches to table view, the button switches to molecule view and the button switches to Card View.

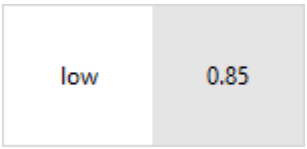
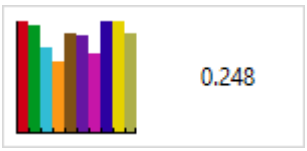
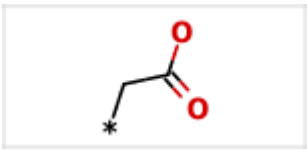
2.2.2 Data types

A data set can contain different types of information. There is one row for each molecule and each column contains a single type of data. There are a number of different types of data that a data set can hold, some of which can be converted from one type to another (this is necessary when data is imported from text files).

Note: For numerical and categorical data, standard deviations or probabilities will only be displayed if **Display standard deviations/probabilities** is ticked in **General** preferences (see section 24.1).

The types are listed below with an example of how they are displayed:

Data Type	Display	Details
Molecule		This shows a 2D representation of the molecular structure. If you select a column of model predictions for which **Glowing Molecule** (see section 8) values have been calculated the molecules will be highlighted with Glowing Molecule information. If there are multiple 3D conformers then the number of poses will be shown in the top right corner.
Number		This is a value from a model or assay that produces numerical results. Every number will have an associated standard deviation.

Category		This is a value from a model or assay that produces discrete results in the form of categories or classifications. Every category will have an associated probability.
Score		The score is the likelihood of the molecule meeting the criteria in a scoring profile (see section 6). Beside this is a histogram showing how each property has contributed to the overall score.
P450		The display shows a **Metabolic Landscape** from the P450 3A4 regioselectivity model along with the **Composite Site Lability** (CSL) value (see section 9). This may also show whether or not a P450 value is **Uncalculated**, **Running** or **Queued**.
R-group/Fragment		This is like a category but can display R-Groups/fragments.
Text	Compound 1	This is a string of text. The default type for any data that is not numerical, categorical or molecular.
Date	11/12/2015	This is a date which can formatted in a number of ways.
torch3D	0.753 0.1	This is a score calculated by torch3D. This may also display **Running** or **Queued** while results are being calculated.
Derek Nexus	Plausible	This is a Derek Nexus prediction. It is a categorical prediction, but coloured to highlight potential toxicities.
Glowing Molecule	3.31 0.435	Any values containing Glowing Molecule data will have a pale green background.
Invalid	?	Any values that do not contain all of the necessary information are considered invalid. These can include items for which model predictions have not yet been generated. If models are running on the server, the cell where the result will be displayed may be shown as an invalid value while the result is being calculated and will be updated when the result is returned.

3 How do I... Access my data?

To work with previously saved files StarDrop supports the following data file types:

- *.sdproj – StarDrop project file
- *.add – StarDrop data set file
- *.apd – StarDrop scoring profile file
- *.sdp – StarDrop visualisation plot file
- *.csp – StarDrop chemical space file
- *.aim – StarDrop model file

In addition, StarDrop also allows you to import data from following text file types:

- *.csv – Comma separated variable (CSV) file
- *.smi – SMILES file
- *.sd and *.sdf – V2000 or V3000 SD file
- *.mol – MDL Mol file
- *.mol2 – Tripos Mol2 file
- *.txt – Text file
- *.* - Any file which contains text delimited by tabs, spaces, commas, colons or semicolons

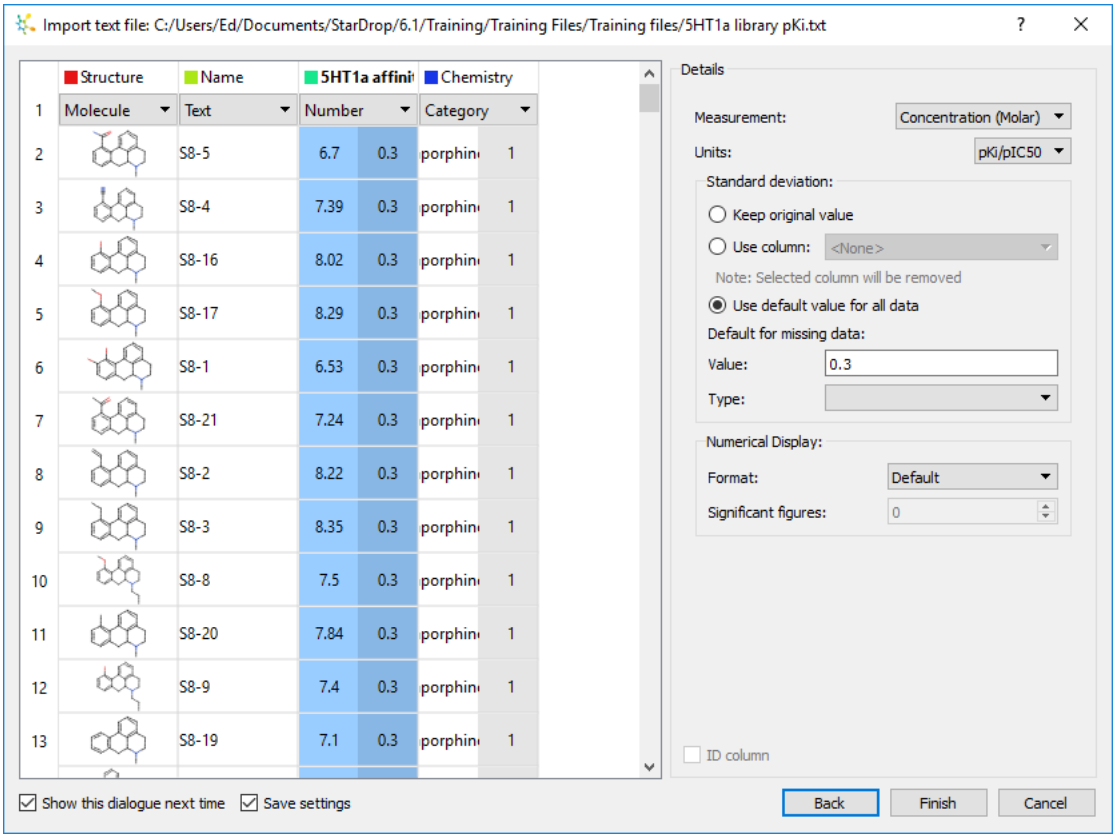
To open a file:

- From the **File** menu select **Open**. A dialogue will be displayed.
- Select the required file and click the **Open** button.
- StarDrop files will automatically be displayed as data sets.
- Text files will, by default, open in a preview dialogue (see 3.1) to enable you to configure the way the data is displayed. However if you have disabled this then they will be displayed as data sets.

Note: For text files of type *.csv or *.txt the first row is imported as a header row and will supply the column title, unless you specify otherwise in the preview dialogue.

3.1 Importing data from text files

When a text file is opened, unless you have specified otherwise in the preferences (see 24.1), the **Import text file** dialogue will be displayed enabling you to preview the data. If the file is of type *.txt then StarDrop will first ask you to confirm the delimiters used. Having done this and clicked **Next**, you can then check whether StarDrop has correctly determined the type of data in each column, and customise the way it is treated.



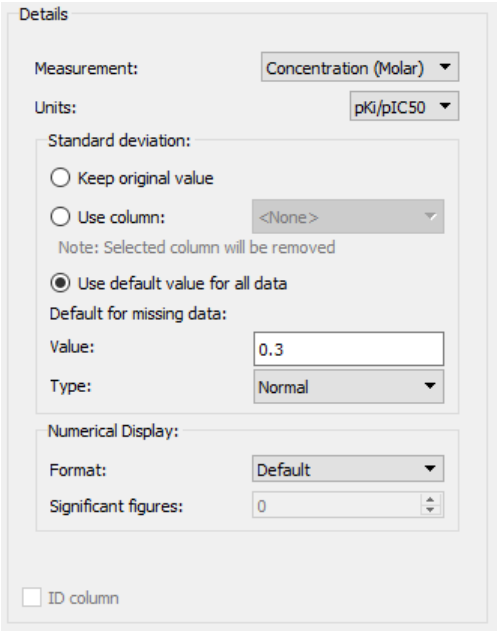
StarDrop will attempt to recognise the type of data in each column by analysing the first 500 rows (or all rows if the data set has fewer than this). If StarDrop has incorrectly chosen the type of data you can change this using the drop-down list at the top of each column.

For each column you can customise the way its data are displayed by selecting it and editing the values in the **Details** section on the right. See the options for different data types below.

Assuming the **Save Settings** option at the bottom is selected, StarDrop will remember the details for any columns where you have customised the way it is imported and automatically apply these again next time. You can see, and manage, the columns for which StarDrop has remembered details in the preferences (see 24.1).

3.1.1 Numbers

StarDrop will recognise as numbers any values that are numerical. If the numbers have modifiers then these will also be recognised. The following modifiers are accepted: >, <, >=, <=. In addition, numbers written as X-Y will be imported with the value displayed as the mean of the two numbers and the standard deviation as half the difference - e.g. 6-7 will be displayed as 6.5 with a standard deviation of 0.5.



StarDrop will show default units and standard deviation values based on any preferences you have set, however, these can be changed here.

You can specify the units of the data by choosing the type of units from the **Measurement** drop-down list and then selecting specific units from the **Units** drop-down list.

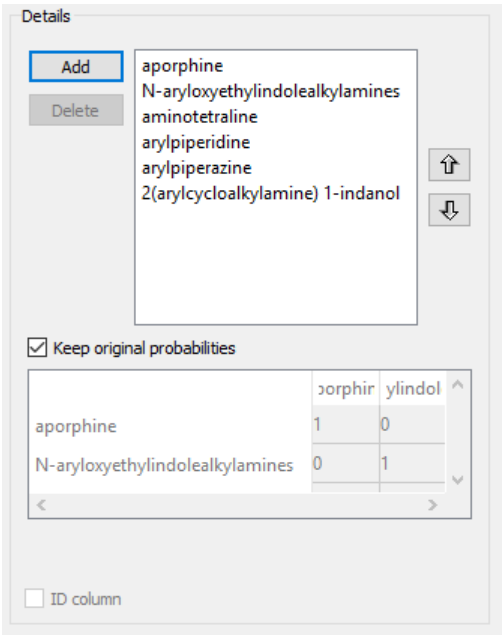
For the standard deviation you can specify that you would like to **Use default value for all data** to manually enter a **Value**. If you are entering a standard deviation then the type should be **Normal**. If, however, you are entering unlogged values where the errors are considered as factors then you can choose **Factor** the from the **Type** drop-down. Alternatively you can also choose to set a standard deviation that is a **Percentage** of the property value.

If you have imported a column of values which represent the standard deviations then you can instead choose **Use column** and then select that column from the drop-down list.

If you wish to specify a particular format in which to display the values for this column then you can modify the **Format** and **Significant figures** options in the **Numerical Display** section.

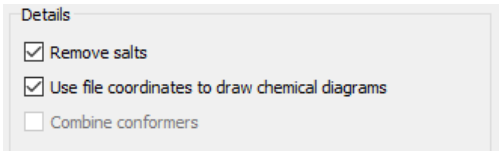
3.1.2 Categories

If StarDrop finds a column that contains text values that repeat throughout the column then it will treat these values as categorical data. You can edit these if necessary simply by double-clicking each individual category. You can use the **Add** and **Delete** buttons to modify the list contents and you can select categories and then use the up and down buttons to change the order (this will affect how the categories are displayed in visualisations).



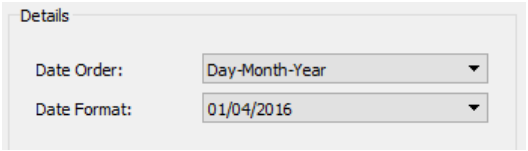
StarDrop will assign the probabilities to each category based upon the value set in the preferences assuming an equal likelihood for all incorrect categories. If you wish to assign probabilities to apply specifically when scoring data for missing values (i.e. if you know that one category is more likely to occur than another) then you can edit the probability matrix by unticking **Keep original probabilities** and double-clicking in each cell.

3.1.3 Molecules



Any text values that appear to be SMILES strings will be converted into molecules. You can indicate whether or not you wish StarDrop to **Remove salts** or mixtures from the molecules. If you have opened an SD file which contains 2D coordinates then you can choose to **Use file coordinates to draw chemical diagrams**. If your SD file has 3D coordinates then these will be preserved so that you can export them back into an SD file. If your data set appears to have multiple conformers then you can choose whether to **Combine conformers** into a single row in the data set. All data from the individual conformers will be preserved.

3.1.4 Dates



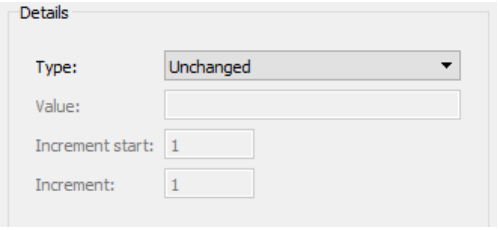
Any text values that appear to be formatted as dates will be displayed as dates. You can modify the **Date Order** and the following **Date Formats** will be recognised:

- 22/12/2001
- 22-Dec-2001
- 22-December-2001
- 12 22 2001
- Dec 22 01
- December 22 01
- 2001 Dec 01

- 2001 December 01
- 20011201
- 2001 12 01

3.1.5 Text

All other data are treated as text. If you wish to overwrite the text you have imported then you can change the Type and, if necessary, indicate that StarDrop should insert a new Value which, if desired can be incremented each row from a starting value.



In addition, you can also specify that a text columns be treated as the **ID column** (check box at bottom of dialogue not show in image). If there is an ID column specified then this can be displayed in visualisations and in Card View.

3.2 Updating data sets

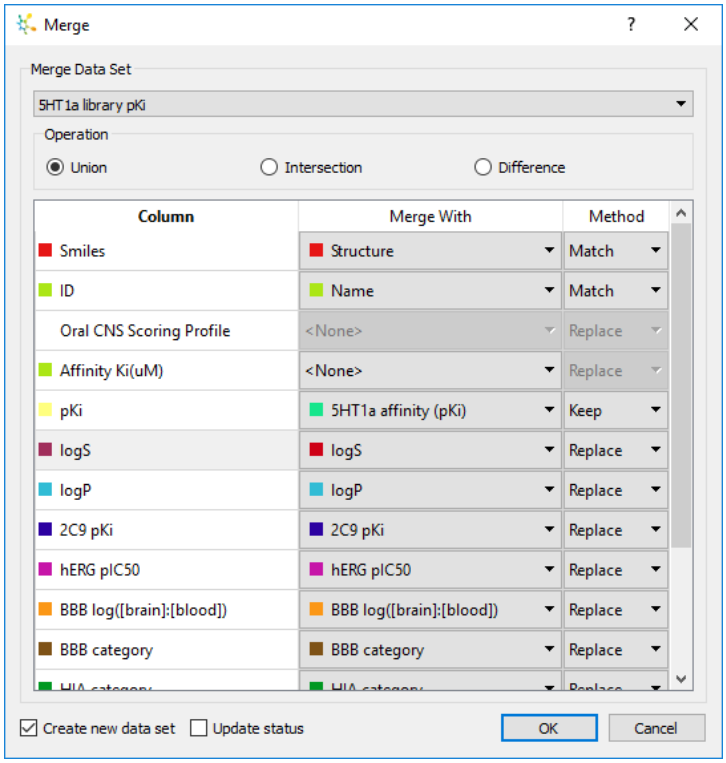
Data sets can be connected with sources of data, such as your in-house database, custom scripts, Pipeline Pilot protocols or imported text files and updated seamlessly when new compounds or results are available. If the underlying source of your data has been updated to contain new compounds or assay results then StarDrop's refresh functionality can be used to pull this new information into your data set. To update your data set menu, open the **Data Set** menu and select **Refresh**. If your data set was created by a custom script, database query or Pipeline Pilot protocol then refreshing it will result in the same query, script or protocol being run. If your data set was created by importing a text file then the same file will be reopened (on the assumption that it has been updated). In all cases, any new compounds or data returned that are not already present in your data set will be merged into your data set.

3.3 Merging data sets

The current data set can be merged with another, creating a new data set with columns from both the current data set and those of a selected set.

To merge two data sets:

- Open the two data sets to be merged.
- With one of the data sets being the active window (if necessary bring it to the front using the **Windows** menu), select menu **Data Set** -> **Merge…**.This opens the **Merge** dialogue.



- Select the other data set to be merged from the drop-down list (assuming the other data set is not already showing).
- Choose whether you would like the resulting data set to be the **Union**, **Intersection** or **Difference** of the two sets.
- The table shows the columns from the current data set under **Column** and the equivalent column from the other data set, with which each will be paired, under **Merge With**. By default, columns with the same name and type will be paired, but the drop-down lists enable you to make changes.
- To pair any other columns which have different names select the appropriate column name from the drop-down list.
- For each pair of columns you can indicate whether their values should **Match**. When comparing rows between the original and merging data set, if all the pairs of Match columns have identical values then those two rows are considered to be equivalent, and will therefore be merged together in the resulting data set. When this happens, for all the other pairs of columns you can indicate whether to **Keep** the value from the original data set or **Replace** it with the value from the merging data set.
- Tick the **Create new data set** option if you would like to create a new data asset containing the results, otherwise the original data set will be modified.
- Click the **OK** button.

Where values are missing in some of the rows, these will not be considered when matching and in the resulting data set the missing values will be replaced where possible.

4 How do I... Use Card View?

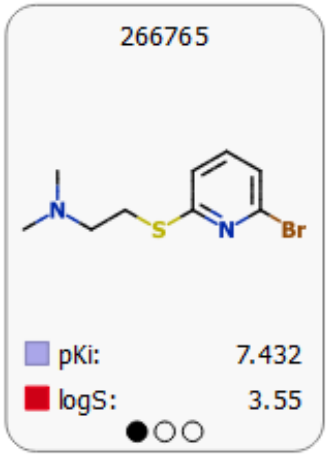
Card View is a ground-breaking and intuitive way to view compounds and their relationships in the context of your discovery projects. It provides a powerful alternative to StarDrop's table and molecule views.

To access Card View for the current data set, click the  button on the main toolbar. You can switch between Card View and other views of your data set without losing any of the Card View features that you have set up.

4.1 Definitions

4.1.1 Cards

A card is a representation of a single molecule in the data set.

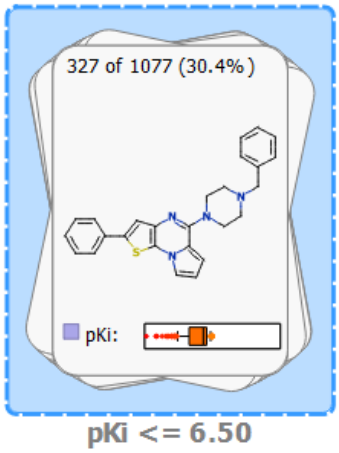


A card can show a structure and any property values you would like to see. Each card can have multiple pages and you can choose what is displayed on each page using the card designer (section 4.3). You can change the page by clicking on the circles at the bottom of the card. This will cause all cards to show the same page. Cards may be assigned a background colour (section 4.4) based on a property value and they can also be selected in the same way as rows in the traditional table view.

Cards can be moved freely around the desktop without affecting their position in the other views.

4.1.2 Stacks

A stack represents a collection of cards, which you can create manually by dragging cards on top of each other or onto existing stacks.

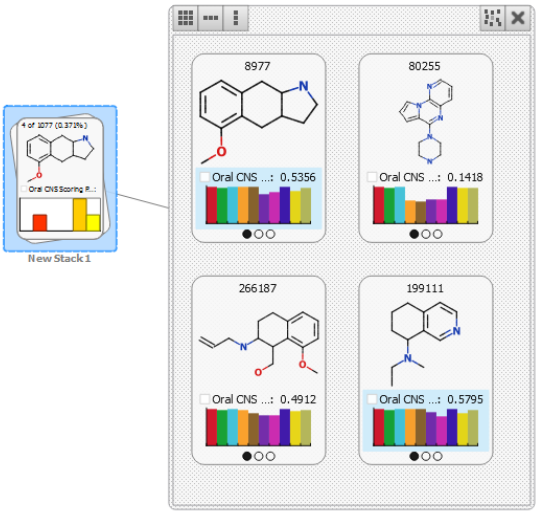


They can also be created automatically on the basis of a property or analysis (see section 4.5). In the same way that you can design cards, you can also determine which information you would like to see about a stack (section 4.3).

Selecting a stack will select all the molecules contained within it.

Stacks may be moved around the desktop in the same way as cards.

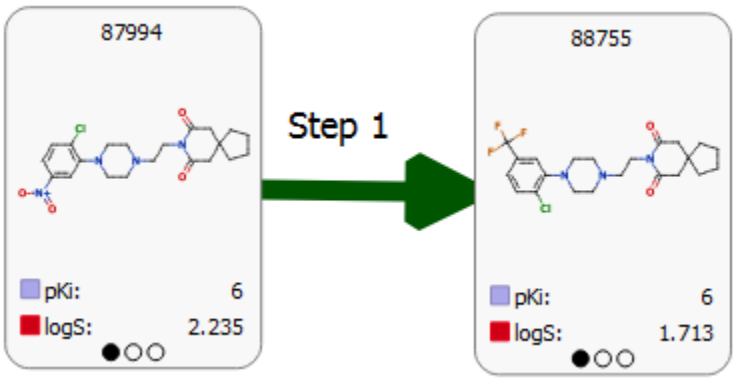
A newly created stack will be given a default name which you can edit by clicking on the name and typing.



You can view the cards within a stack by right-clicking and selecting the **Inspect** menu option. This will display a separate window in which all the cards in a stack can be viewed. Dragging a card from this window to the desktop will remove it from the stack.

4.1.3 Links

Links represent a connection between two cards. Just like stacks, these can be created manually using the link tool (section 4.2.4) or automatically by one of the analyses (section 4.5). Links can be annotated and coloured (section 4.4) and can, optionally, have a direction associated with them. Individual links can be edited at any time. Right-click over a link to add a label, edit the properties of the link or remove it from the view.



4.2 Tools

A number of tools are available to manipulate the Card View and the objects within it. You can choose a tool from the toolbar at the top of the Card View display.

4.2.1 Movement

This is used for basic card manipulation operations.

Clicking on any card will select it. If the Ctrl key is pressed, any existing selections will be retained and the newly selected card will be added to the selection. Otherwise, existing selections will be cleared. Clicking on the space between cards will clear the selections.

To drag a card, click on the card and hold the left-mouse button down while moving the mouse. Releasing the mouse button will drop the card at the new position. Stacks can be moved in exactly the same way. If multiple cards are selected, all the selected cards will be moved as the mouse moves.

To pan the view, click on a space between cards and move the mouse while holding the left-mouse button down. Turn the mouse wheel forward to zoom in to Card View, and turn backward to zoom out.

4.2.2 Selection

This behaves similarly to the movement tool, but allows for lasso selections to be made. To make a lasso selection, move the mouse while holding the left-mouse button down. On releasing the button, all cards totally or partially contained by the lasso will be selected.

To pan the view in this mode, move the mouse while holding the right-mouse button down.

4.2.3 Drawing

With this set of tools you can add your own labels and illustrations to Card View. Any additions that you make will be included in a printed image, and will be saved when you save the data set.

To add your own drawing to Card View, move the mouse while holding the left-mouse button down. To change the colour and thickness of the line, use the **Colour** and **Size** controls in the toolbar.

From this toolbar, you can also select an eraser tool, add a label to Card View, or remove all labels and annotations from the view by clicking the reset button on the right.

To pan the view in this mode, move the mouse while holding the right-mouse button down.

4.2.4 Link

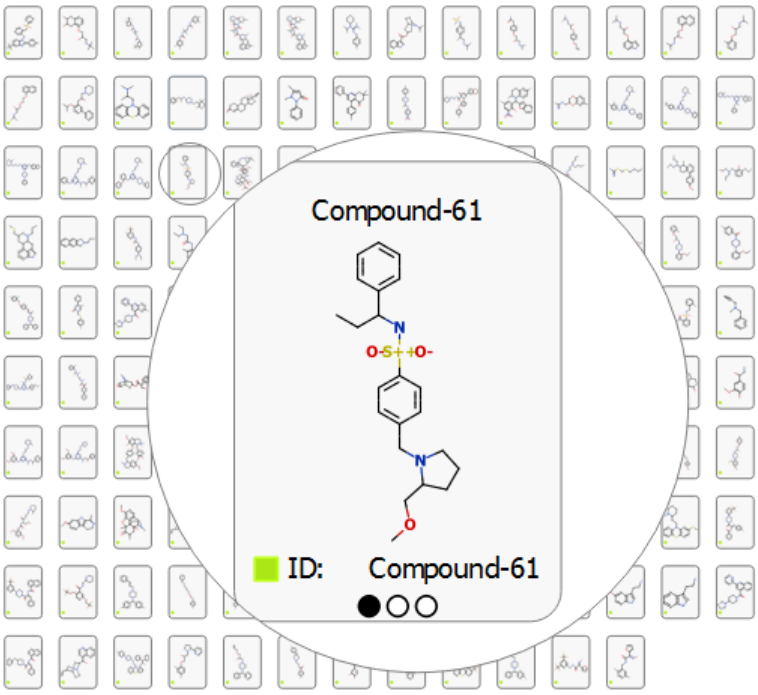
With the link tool you can add links between cards. To add a link, click on the card that you wish to be at the start of the link, and then click the card at the end of the link. A link will appear, and can be edited by right-clicking over it.

You can change the type of link to add by selecting from the **Type** control in the toolbar. Links may be either directed (in which case they appear with an arrow pointing to the second card clicked) or undirected;

From this toolbar, you can also remove all links from the view by clicking the button on the right.

4.2.5 Magnifier

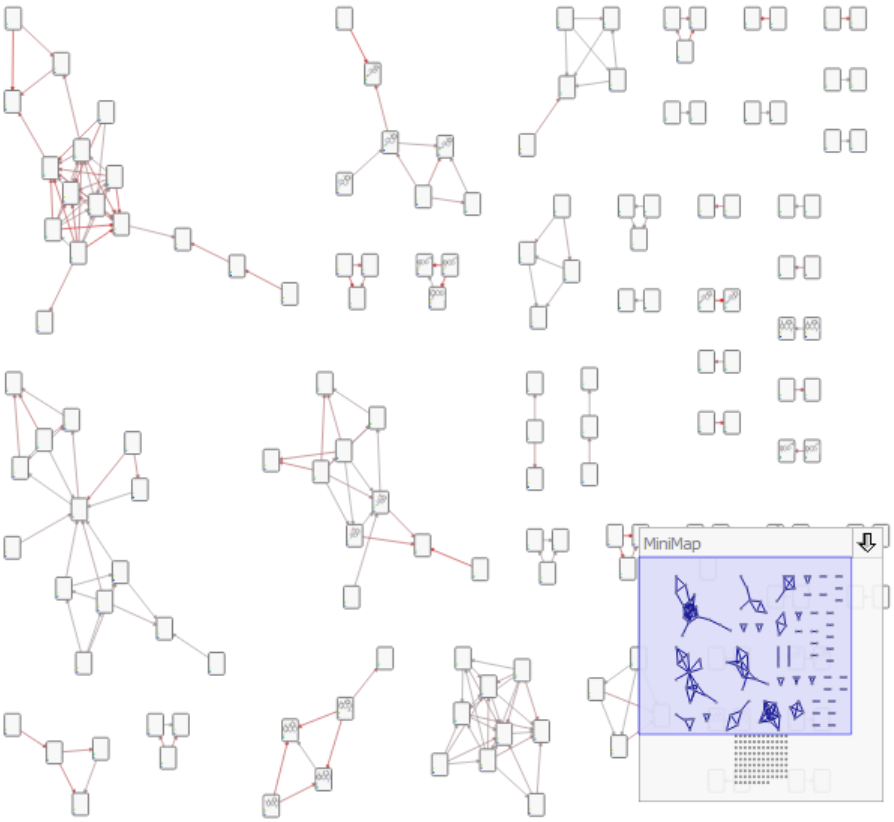
This tool enables you to magnify a selected portion of the view. It is useful to inspect individual cards when the view is zoomed out too far to see details on the cards. To see a magnified view, click anywhere in the view and hold the mouse button down. A magnified view of the part of the view under the mouse will appear, and will follow the mouse as you move around the view.



To pan the view in this mode, move the mouse while holding the right-mouse button down.

4.2.6 Mini-map

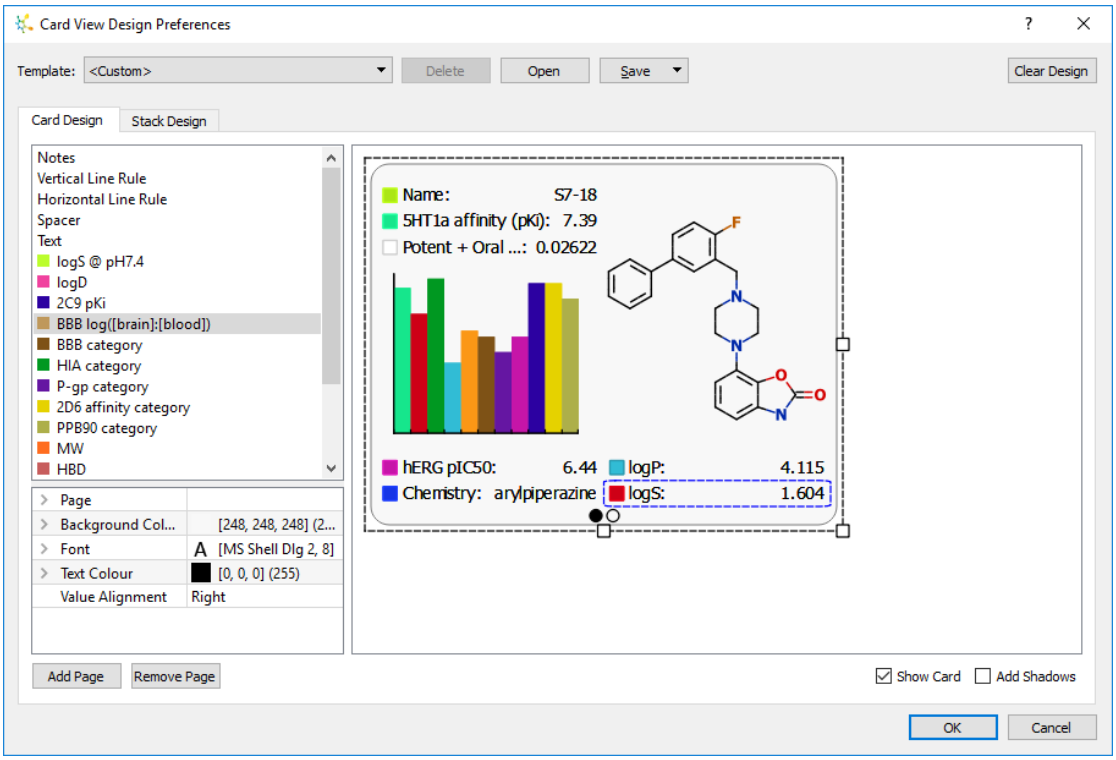
To aid navigation around the Card View desktop, a mini-map appears at the bottom right of the view. This will always show the whole view, with the currently visible portion highlighted. You can move the main view to a particular area by clicking on the mini-map.



To hide the mini-map, click the arrow above it. A small arrow icon will remain visible, and the mini-map can be restored by clicking this.

4.3 Card and stack appearance

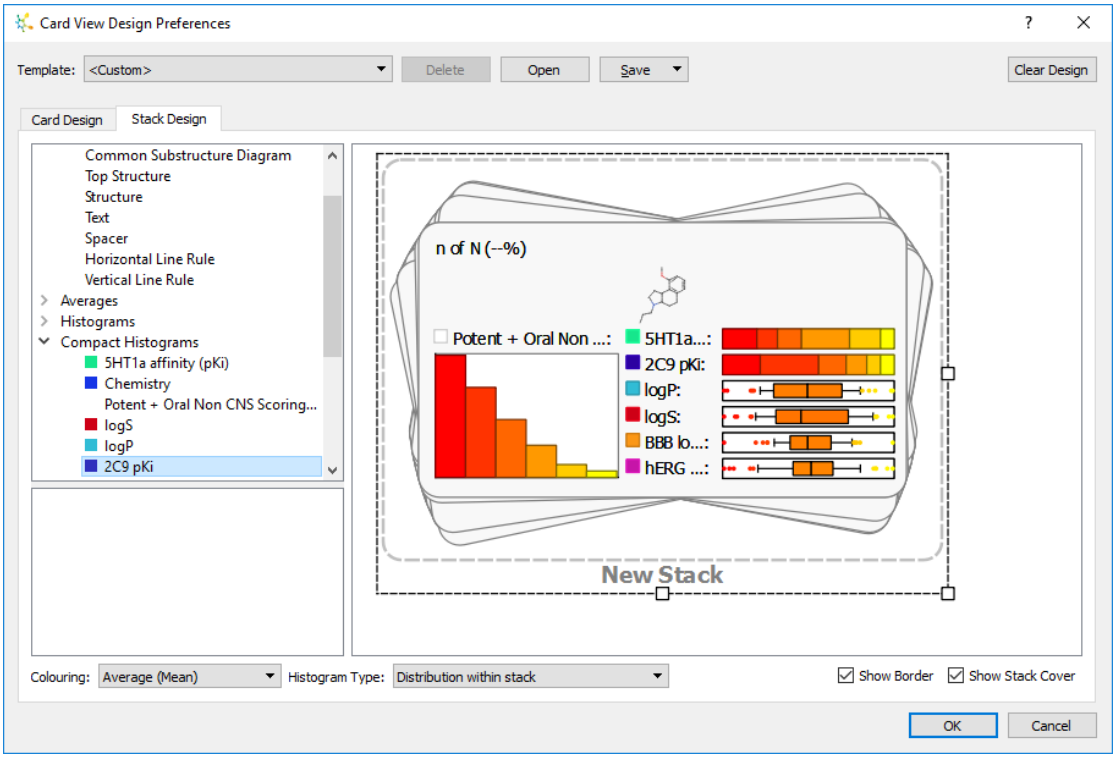
You can control the appearance of cards and stacks using the **Design** button. By default, StarDrop will create a single page design showing a structure diagram and one or two properties, but this is entirely customisable. From the **Design** button, you can specify a pre-defined design or select **Custom…** to launch the Card View Design Preferences.



From this dialogue you can choose any size of card by simply dragging its border. You can drag properties onto the card and a blue guide will show where the property can appear, enabling you to create grids or columns. You can use the Spacer, Line Rules and Text items to customise the layout and selecting any item on the card will enable you to modify its background colour, font and alignment. Large items are automatically sized, but again, by selecting them you can modify their relative sizes. The Add Page button can be used if you want to show information on separate pages.

When you have a design that you like, you can save it to a file or as part of your project by using the **Save** button. Any saved design templates can be opened using the **Open** button in future StarDrop sessions and reused. You can also indicate in the Preferences (section 24.2) a location from which StarDrop should automatically import saved card designs.

The dialogue also enables you to specify the appearance of stacks.



Stacks can only have one page, so the amount of information that can be displayed is more restricted. Stacks can show some summary information about the number of cards in the stack, a representative structure and small histograms or box plots showing the distribution of a property across the contents of the stack.

4.4 Colouring

Cards and links can be coloured according to a property, allowing you to create a 'heat map' effect in Card View. Click the  button to launch the **Colour By** dialog. Choose whether to colour the cards or the links and then select the desired property from the drop down menu. You can choose to colour using a continuous range, or by applying one or more thresholds. Any stacks you have created will also be coloured according to the values of the specified property within the stack.

When colouring links colours will be set according to the difference in a specified property between the two linked cards. If the links are undirected (i.e. no arrow) the colour will reflect the magnitude of the difference in property values. If the links are directed (i.e. they have an arrow) then the colour will reflect the magnitude of the property difference and also whether the change (in the direction of the arrow) is positive or negative.

To apply a colour to an individual card, right-click over the card and select the **Change Card Colour** option. If you have more than one card selected, all the selected cards will be given the selected colour.

4.5 Arranging cards and stacks

StarDrop provides a range of options for organising and laying out the cards in a data set. The **Layout** button enables to arrange the cards simply in a variety of layouts. The **Organise** button enables you to select an arrangement that reflects one or more properties of the compounds, in some cases creating or removing stacks. The **Analyse** button enables you to perform a range analyses on the data set, some of which may create or remove stacks and add or remove links to represent the results.

4.5.1 Layout

From the **Layout** button you can select the following options.

- **Grid** - Lays out all the cards in a grid. The grid flows from left to right and top to bottom, and the initial width is set by the width of the Card View window when the option is selected.
- **Row -** All cards are laid out in a single row
- **Column** - All cards are laid out in a single column
- **Network** - This will lay out any linked cards in a network view that maximises the visibility of the connections between them. Any cards not connected to any others are displayed below the network in a grid.
- **Hierarchy** - As above, but will take account of the directionality of any links when constructing the network.

4.5.2 Organise

From the **Organise** button you can perform the following actions which lay cards out based upon their properties. Any links or stacks you have created will remain within the arrangement.

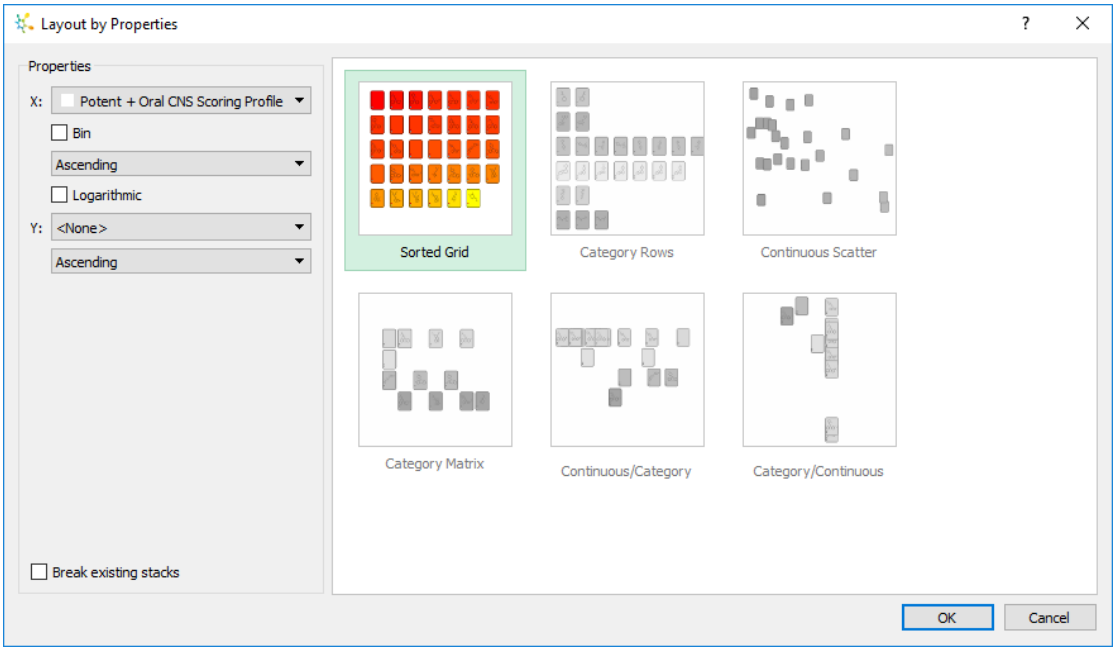
Stack

Selecting the **Stack** menu gives you several options:

- **Stack All -** Puts all the cards into a single stack
- **By Property** - Enables you to select a property to use to split the data set into stacks. If you choose a category property, StarDrop will create one stack for each category. For a continuous property, StarDrop will bin the data and create two stacks (although you can control this to set the number of bins and the thresholds to use).
- **Recursive Partitioning** - You can choose two or more properties to classify your data and StarDrop will create a stack for each branch of the tree generated. As above, you can specify category properties or binned continuous properties to create the tree.
- **Break -** Any stacks will be broken and displayed as individual cards
- **Save Stack Details to Data Set** You can add a new category property to the data set, capturing the current stacks. You can use this property in other StarDrop operations, including Visualisation and Scoring.

By Property

Selecting the **By Property** option enables you to arrange the cards in a 2-dimensional layout according to one or two property values.

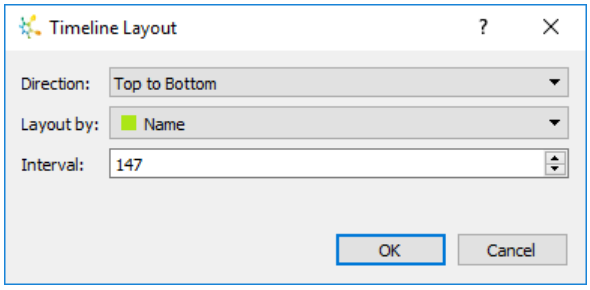


The types of layout available will vary according to whether you have one or two properties selected, and whether the properties are continuous or categorised. Checking the **Break existing stacks** box will

ensure that the layout is performed on all cards. If this is not checked, and stacks are present, they will be positioned according to the average values of the relevant properties within them.

Timeline

If you select the **Timeline** option, the cards will be laid out according to the directed (arrowed) links you have created between the compounds. When chosen you will see the **Timeline Layout** dialogue:

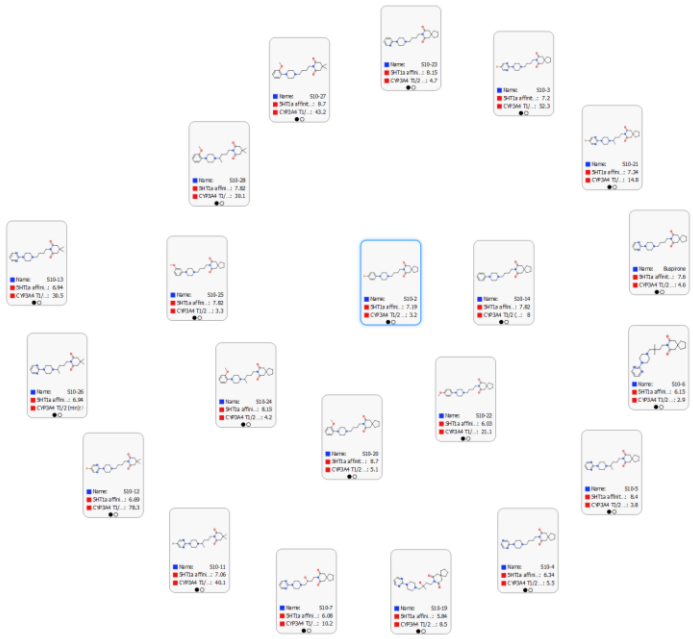


You can choose whether you would prefer a top-to-bottom or left-to-right layout. The **Layout by** option provides the opportunity to indicate a date or compound ID range of values over which the timeline occurs. This, in combination with the interval, then enables you to control the scale. For example, if you choose to lay the timeline out using a date column which spans a year, if the interval is a week then the timeline will have 52 rows (assuming it is top-to-bottom) where compounds within the same week are all on the same row.

Note: If you use compound ID then they must have some numerical component to them (e.g. ID123456) and not just be compound names.

Similarity

Select a single compound and then select the **Similarity** option. The compounds will be arranged in a spiral, with the selected compound at the centre and structurally similar compounds close to it.

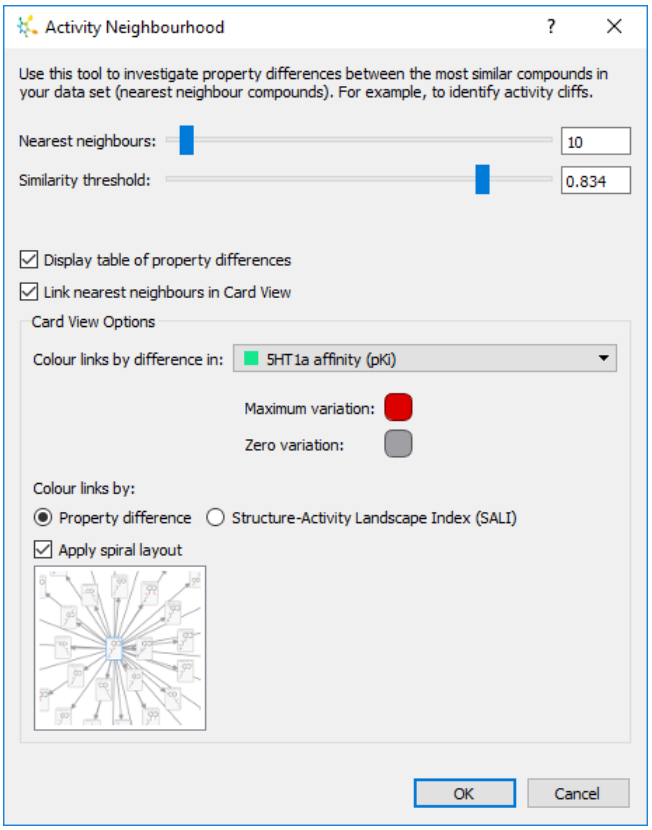


4.5.3 Analyse

The **Analyse** button in Card View provides direct access to a range of StarDrop tools that are particularly suited to use within Card View. **Note:** Links, stacks and card positions may be altered in order to represent the results of the analysis.

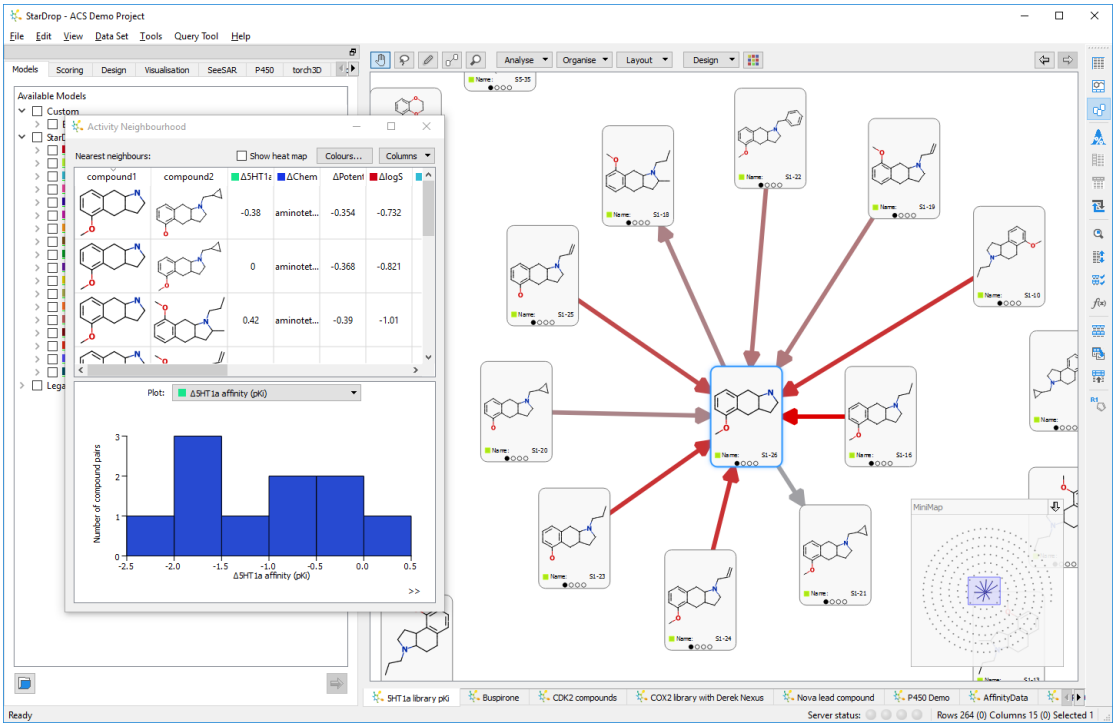
- **Find Matched Pairs** – The matched pairs tool enables you to arrange the cards within a network such that pairs of compounds which differ by just a single functional group change are linked. For more details about this tool see section 9.2.
- **Clustering** – The clustering tool enables to arrange the cards into clusters based upon similarities between their structures, similarities between the properties or common substructures. For more details about this tool see section 9.1.
- **Activity Landscape** – The activity landscape tool enables you to create a network where similar cards are linked and differences between their properties are indicated by coloured directed links. For more details about this tool see section 9.3.
- **Activity Neighbourhood** – The activity neighbourhood tool provides a cut-down version of the activity landscape tool where the focus is just on a single starting compound which you select before starting the analysis.

When you start this analysis you will see the following dialogue:



You can specify the number of nearest neighbours you would like to see in the results, or the similarity threshold you would like to use. These are linked so that you will be able to see the similarity threshold required to achieve a given number of nearest neighbours and vice versa.

The additional options enable you to choose how the results will be displayed; in a table, in Card View or both. If you choose to link nearest neighbours in Card view then you can choose a property for colouring the links. The links will be coloured to indicate either the magnitude of the difference in property values between the pairs or the Structure-Activity Landscape Index (SALI). An arrow will be shown on the link to indicate the direction of increase. You can also choose to organise the results within Card View in a network.



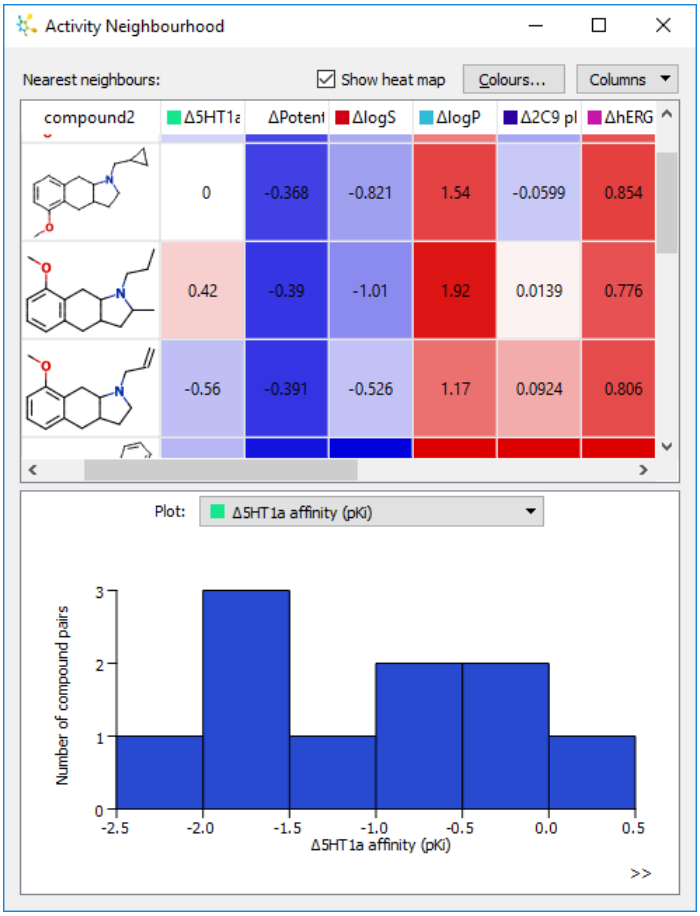
The **Activity Landscape** results show a table and a histogram and in Card View will show the selected compound at the centre of a spiral network with links indicating the magnitude and direction of change of the chosen property.

The **Nearest Neighbours** table lists all the nearest neighbours. The subsequent columns indicate the difference in property value between the chosen compound and that neighbour – these are calculated for all the properties in the data set.

The histogram at the bottom shows the distribution of property differences between the pairs of compounds for the property selected in the drop-down menu above it.

If you select a row in the **Nearest Neighbours** table then those compounds are selected in the data set (whichever view of the data set you are using). If your data set is displayed using Card View then the view will zoom and pan to show the selected compounds.

You can also choose to display a heat map in the tables by ticking the **Show heat map** option at the top.



You can edit the **Colours...** which are there to make it easy to spot which pairs of compounds have significant property differences.

You can choose which properties are displayed in the tables by clicking the **Columns...** button.

4.6 Saving and restoring

At any time you can use the forward and back arrows in the toolbar at the top of Card View to undo or redo changes made to the view. If you have a particular arrangement that you want to preserve, you can save it by selecting **Save** from the **Organise** menu. That layout will then be immediately available by selecting the **Load** option from that menu.

4.7 Printing and copying images

4.7.1 Card images

To save an image of an individual card, right-click on the card and select **Copy Image** or **Save Image** from the menu. The copy command copies the card image to the clipboard for pasting into other applications for use in presentations etc. The save command saves the image as a PNG file.

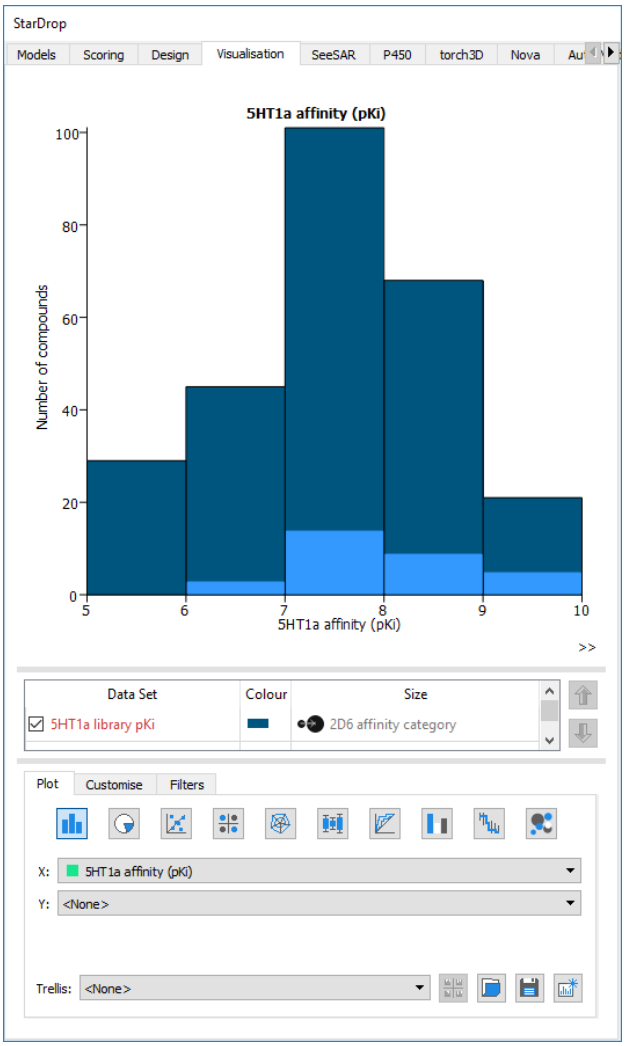
4.7.2 Card View image

To save an image of the whole view, right-click on an empty part of the view and select **Copy Image** or **Save Image** from the menu. The copy command copies the image to the clipboard for pasting into other applications for use in presentations etc. The save command saves the image as a PNG file.

From this menu, you can also choose to print the whole view.

5 How do I... Visualise my data?

It is possible to create a number of different plot types within StarDrop, all of which can be copied into other documents. Plots can be interactively created in the Visualisation tab.



5.1 General

A number of options are available for customising plots.

5.1.1 Labels

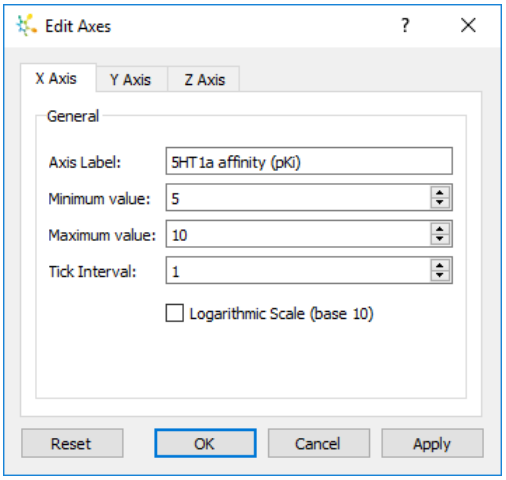
For all plot types the title and any axis labels can be edited by clicking on the text and typing a new label. Grabbing the edge of the label enables it to be moved by the mouse.

Right-clicking on the label displays a menu enabling the Font and Colour to be specified.

5.1.2 Axes

Any axis which is displaying numerical data (with the exception of axes in radar Plots or the Y axes on Histograms or Snake Plots) can be edited. To change the scale of the axes, grab one of the tick marks and move it left or right. Alternatively you can use the mouse-wheel to zoom into or out of a plot in a homogeneous way. You can also use the cursor keys to pan the plot left, right, up and down.

Alternatively, right-click on the axis and choose the menu item **Edit Axis...** from the axis menu. This will show the **Edit Axes** dialogue (also available by choosing the main plot **Edit Axes...** menu item) with the dialogue displaying the tab for the axis that was clicked.



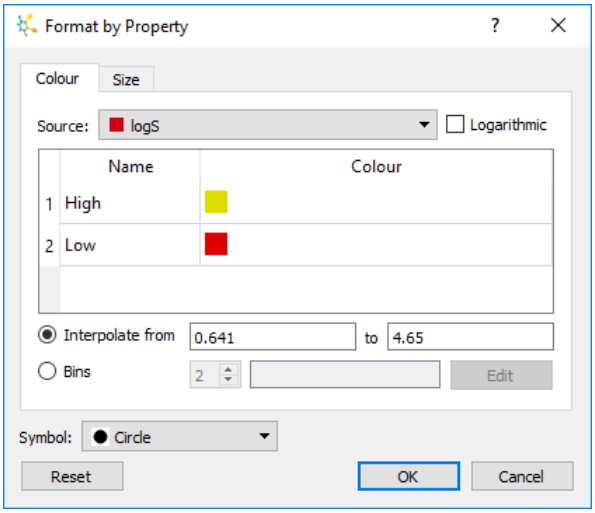
Specify your desired values and click the OK button to change the values.

For scatter plots and the Y-axis of 2D histograms and box plots you can also specify that you would like the axis to be on a log scale (base 10). If you choose this option then the tick interval will be disabled and appropriate ticks will be automatically determined.

The axes can be reset at any time by right-clicking on the plot area to display the plot menu and choosing **Reset Axes**.

5.1.3 Colour and Size

To edit the size and colour of plotted data points, click on the Colour or Symbol value in the list of data sets which is just below the plot area. You can set the colour and size of data points to a single colour and a single size. You can also colour by value and set the data point size by value. For example, you can colour data points by their logP value.



Changing the Source value allows you to specify the column on which to base the colours and you can set specific colours to use for the high and low ends of the range of values. Choosing the **Logarithmic** option will result in the data points colour and size being set based on a log scale (base 10).

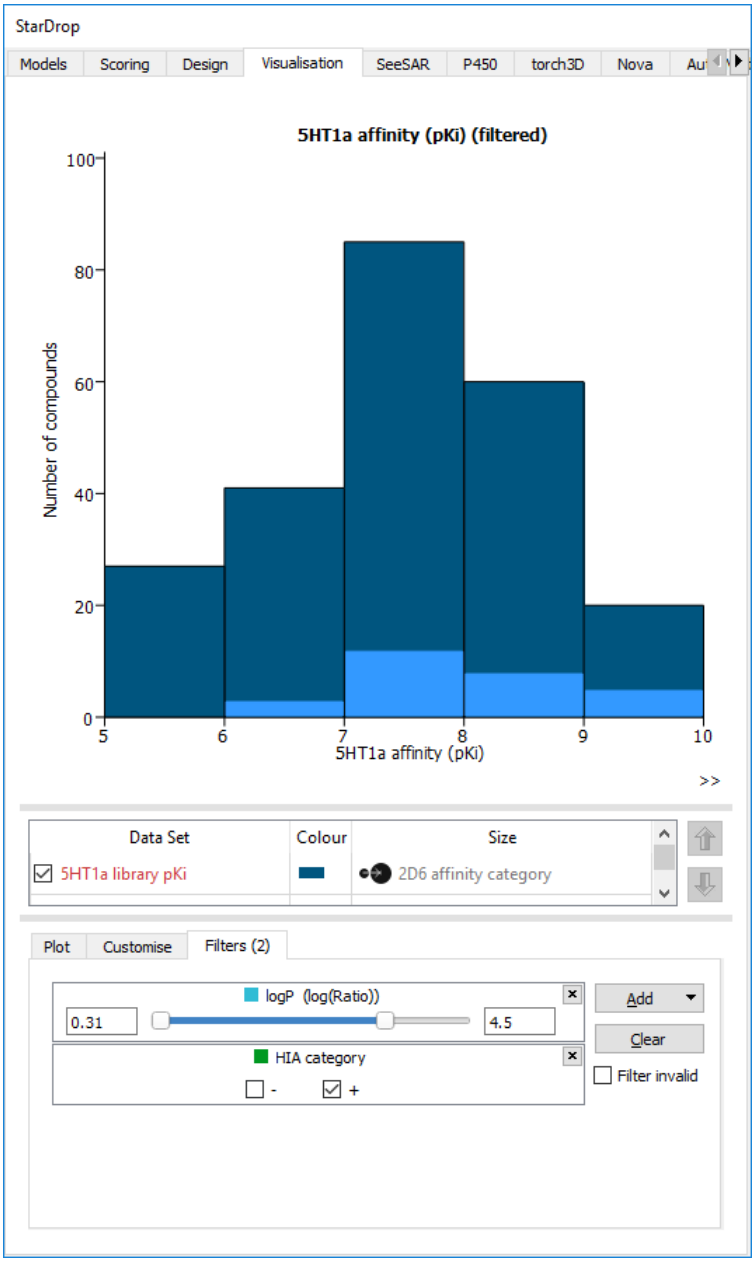
You can choose to either **Interpolate** or use **Bins** when specify colours and sizes. For any plots where data for multiple rows is represented by a single item in the plot (e.g. a histogram bar or a pie chart segment), choosing interpolate will result in the mean value of all the represented rows being used to set the colour or size.

The same process can be used on the Size tab.

At the bottom of the dialogue you can specify a specific symbol to use for the points in the data set.

5.1.4 Filtering

Click on the Filters tab below the plot if you wish to filter the data that is displayed on a plot based on one or more properties.



Clicking the **Add** button will enable you to choose a property on which to filter. Having added the property to the list you can either select/deselect categories if it is a categorical property, or move the sliders if it is numerical.

To remove all the filters click the **Clear** button.

5.1.5 Customising

A number of options are available for customising the plot area using the plot menu (right-click on the plot) or by clicking on the **Customise** tab below the plot. The following options are available:

- Setting the background colour
- Displaying a grid

- Displaying horizontal and/or vertical error bars (where appropriate)
- Displaying a regression line (where appropriate)
- Displaying an identity line (where appropriate)

5.1.6 Saving and Copying

The plot image can be copied using the plot menu (right-click on the plot). Images copied from the plot will be available on the clipboard and are created approximately 1500 pixels wide (with vertical scale set proportionally).

The plot can be saved using the plot menu (right-click on the plot) or by clicking the  button. Saved plots can be retrieved by clicking the  button. Saved plots can be removed from the list in the Visualisation preferences (see section 24.6).

5.1.7 Selections

You can select individual data points using the left mouse button. Multiple points can be selected by holding down the Ctrl key. (Note: when using a 3D plot you need to hold the Ctrl button for all selections). You can also select multiple points by holding down the left mouse button and drawing a boundary to lasso the points you wish to select. Selections made in a plot will be reflected in all open data sets and any other detached plots. Conversely, selections made in a data set will be reflected in any plots containing those points.

5.1.8 Key

A Key is available for all plots. Clicking the >> button in the bottom-left corner of the plot will display a key window. Right-clicking on the key window will display a menu enabling its contents to be copied of saved.

5.2 Graph types

There are a number of different types of graph that can be created:

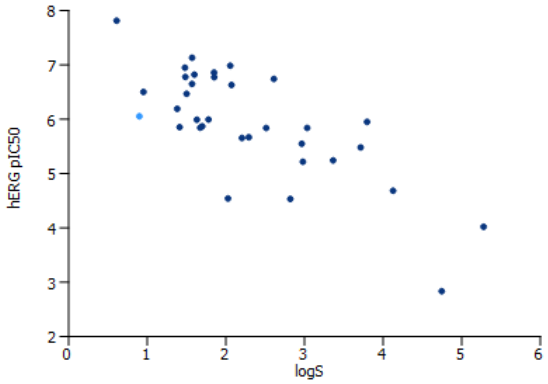
- Histogram
- Pie Chart
- Scatter Plot
- SAR Plot
- Radar Plot
- Box plot
- Criteria Histogram
- Snake Plot
- ROC Curve
- Chemical space

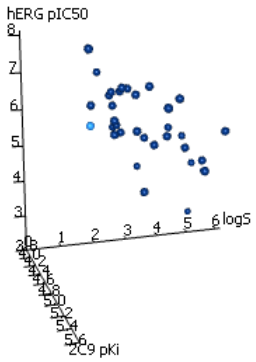
Multiple data sets can be plotted at the same time (assuming the necessary data is available in all data sets), either as combined totals or as individual data sets overlaid.

In addition, all plot types can be trellised to show multiple plots, each of which corresponds with a sub-set of the data that falls within a defined region or category of a specific property.

A graph that has been created can be detached by clicking the  button, enabling you to see multiple graphs at the same time.

5.2.1 Scatter Plot





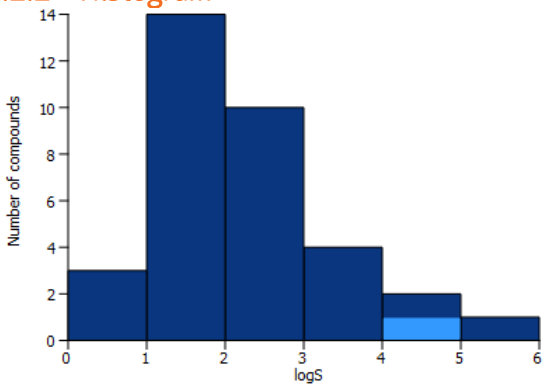
Scatter plots can created by clicking the  button and be used to plot any two or three columns of numerical or categorical data. If the Z axis is <None> then the plot will be displayed in 2D. If there are three columns then the plot will be displayed in 3D.

To select points in a Scatter Plot, click on them individually or draw around them to lasso multiple points. When displaying a 3D plot you need to hold the Ctrl button while lassoing to stop the plot rotating.

Hovering the mouse over a point will invoke a pop-up display showing the structure (assuming the data set contains structures).

If you have created a 3D plot you can spin the plot in any direction by clicking and dragging the mouse. If you let go of the mouse while doing this you can cause the plot to rotate smoothly in any direction. If you click on the plot again it will stop rotating.

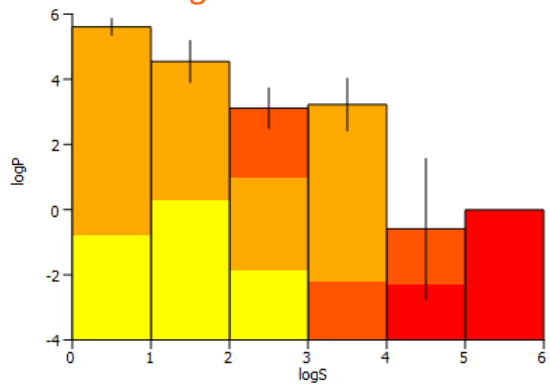
5.2.2 Histogram



Histograms can be created by clicking the  button and are used to display the distribution of values in a column of numerical or categorical data. When a histogram is displayed the customisation options enable you to choose whether to display multiple data sets in a stacked histogram.

To select data in a histogram, click on the bar. Holding down the **Ctrl key** while clicking allows you to select multiple bars.

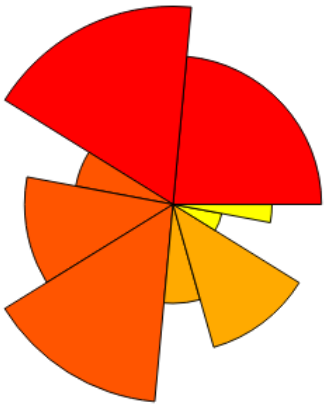
5.2.3 2D Histogram



If you also select a property for the Y-axis of a histogram a 2D histogram will display the distribution of values for that property within each of the bars.

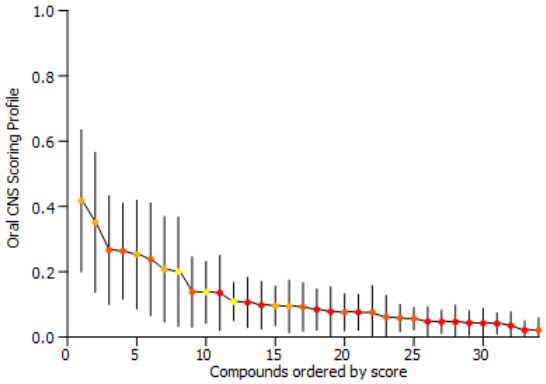
The height of each bar shows the mean value of the property on the Y axis within the given range of values on the X axis. The vertical error bars display the standard deviation of the data within this range.

5.2.4 Pie Chart



A pie chart can be created by clicking the  button. The **Colour** and **Size** options determine the segments that will be displayed within a pie chart for a single data set (see 5.1.3).

5.2.5 Snake Plot

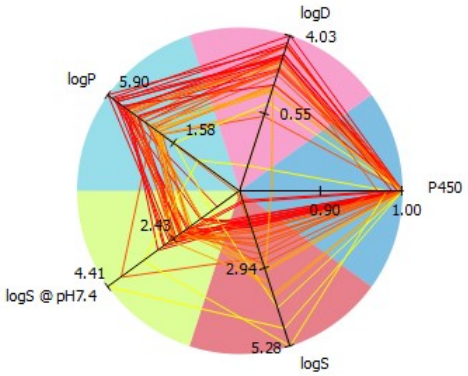


A Snake Plot can be created by clicking the  button and is used to display data from a column of scores. The compounds are displayed in order of highest to lowest score along the X axis with the score value displayed on the Y axis with the vertical error bars displayed.

To select points in a Snake Plot, click on them individually or draw around them to lasso multiple points.

Hovering the mouse over a point will invoke a pop-up display showing the structure (assuming the data set contains structures).

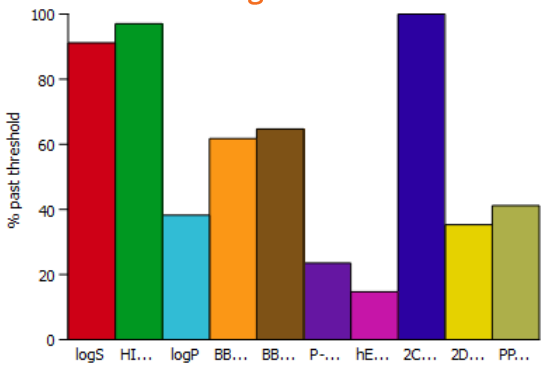
5.2.6 Radar



A Radar Plot can be created by clicking the  button and is used to display data from either a column of scores or from multiple properties. When displaying scores the grey region in the background indicates the "ideal" region from the scoring profile. Each compound in the data set is drawn as a series of lines between the axes indicating a shape corresponding to the properties of that compound.

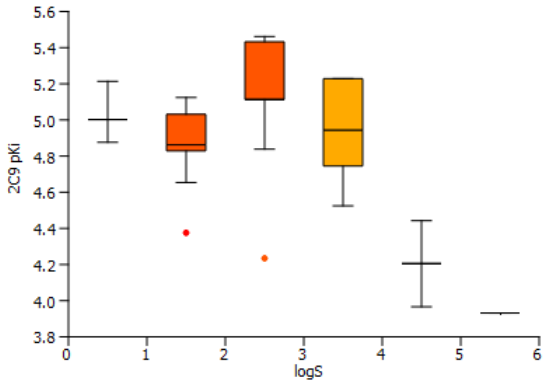
Hovering the mouse over an axis displays a tool-tip indicating the value at that position along the axis.

5.2.7 Criteria Histogram



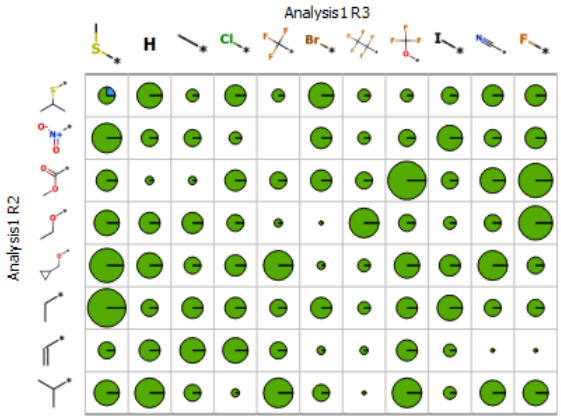
A Criteria Histogram can be created by clicking the  button and is used to display data from a column of scores. Each bar indicates the number of compounds within the data set which passed the criteria specified for that property in the scoring profile.

5.2.8 Box Plot



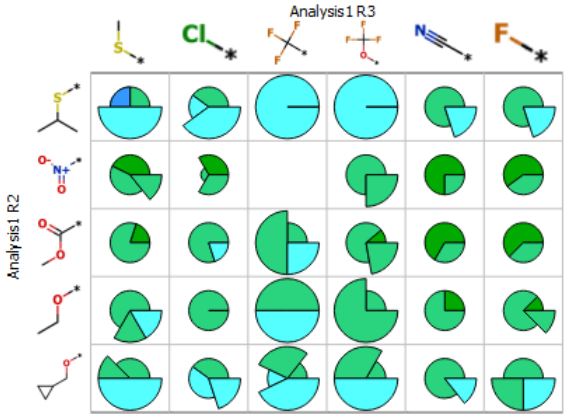
You can create a box plot by clicking the  button. A box plot enables you to look at the distribution of one properties value with respect to a second property. You can customise the whiskers displayed, choosing the percentile that they should indicate.

5.2.9 SAR Plot

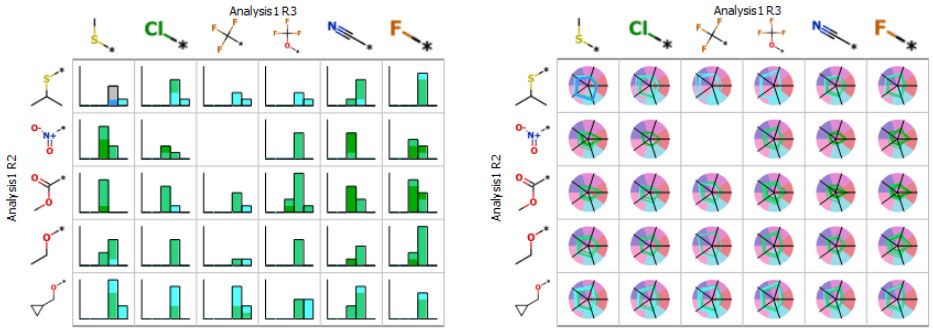


An SAR plot can be created by clicking the  button and is used to display the values from two columns of category or R-Group data (section 10.1). To select data, click on the segment. Holding down the **Ctrl key** while clicking allows you to select multiple segments. When plotted with R-Group data you can hover over the R-Groups to see a larger pop-up of the structure (see below).

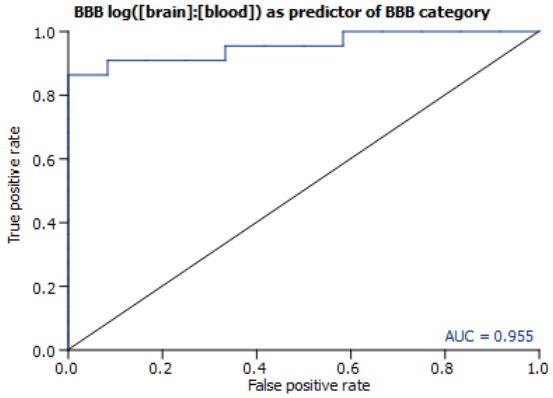
The default SAR plot will display basic pie charts in each cell, the size of which indicates the number of compounds represented by that cell. A with pie charts, changing the size and colour options can alter the displayed pie charts to represent other properties.



Alternatively you can choose to display histogram or radar plots within the cells of an SAR plot by changing the options in the Customise tab. Here you can also specify which properties to display within the plot type chosen.

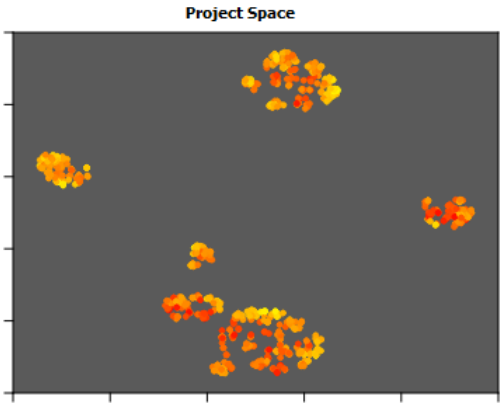


5.2.10 ROC Curves



A Receiver Operating Characteristic (ROC) curve can be created by clicking the  button. The curve(s) will display how well the **Classifier**, which can be a numerical or categorical property, can be used to classify the chosen **Desired result** of the **Property**.

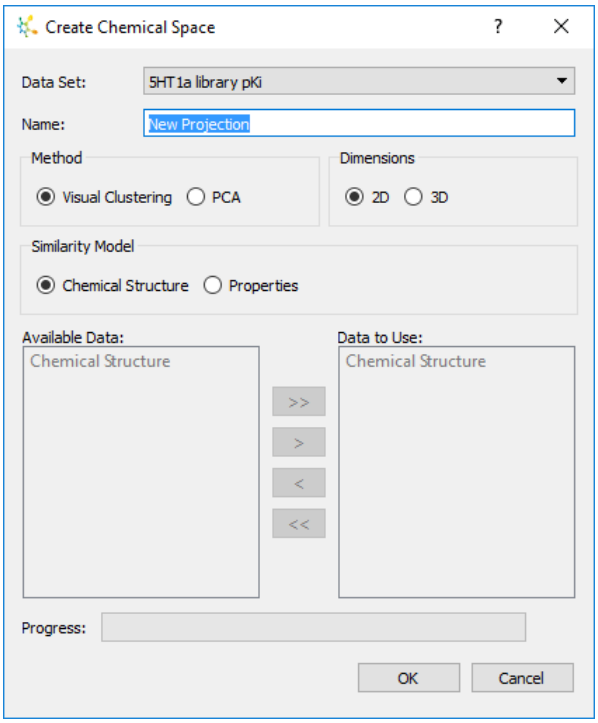
5.2.11 Chemical space



A Chemical Space plot can created by clicking the  button and be used to display a data set within a chemical space projection.

To open a saved chemical space click the **Import...** button or to create a new projection, click the **Create...** button to display the Select Chemical Space Dialogue.

If you have created a 3D chemical space you can spin the plot in any direction by clicking and dragging the mouse. If you let go of the mouse while doing this you can cause the chemical space to rotate smoothly in any direction. If you click on the chemical space again it will stop rotating.



5.2.12 Creating a new chemical space projection

An open data set is required in order to generate a chemical space projection.

Select one of the open data sets from the drop-down list and, optionally, type in a new name for the projection. Select a **Similarity Model** to use for the chemical space projection. The options are:

- Chemical Structure Only
- Properties Only

If in doubt, select **Chemical Structure Only**.

Select a **Method**. The options are:

- Visual Clustering
- PCA

Finally select whether to create a **2D** or a **3D** space.

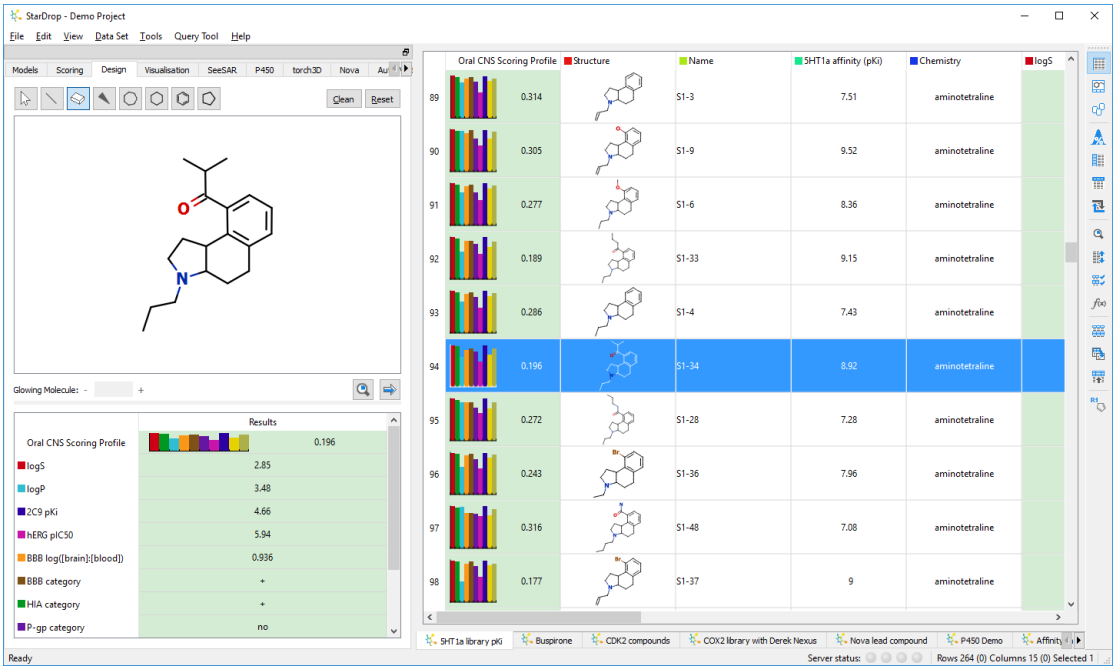
Each method will produce a different chemical space of the same data. You will probably find that for large, diverse data sets the Visual Clustering method produces more distinct clusters of compounds.

If **Properties Only** is selected, a list of the properties in the data set is displayed. Use the buttons to choose which properties from the data set to use when generating the projection.

Click the **OK** button to start generating the chemical space. An indicator bar will indicate progress giving the option to cancel the process.

6 How do I... Use the molecule designer?

The **Design** tab enables you to edit molecules and, if required, add new molecules to a data set. The top half of the tab consists of the molecule editor and the bottom half displays a summary of the calculated predictions.



A molecule highlighted in the data set will automatically be shown in the editor. The **Summary** will update to show the calculated properties for the displayed molecule. **Note:** Running the P450 models is computationally expensive, so these do not appear in the **Summary** display.

6.1 Drawing and editing

The molecule editor is a simple interface for editing and drawing chemical structures with many of the functions bound to specific keys.

General keys

Hot Key	Function
Ctrl-z	Undo
Ctrl-y	Redo
Ctrl-a	Select all

Select tool keys

Hot Key	Function
E	Rotate selection clockwise
Q	Rotate selection anticlockwise
Ctrl-x	Cut a selection
Ctrl-c	Copy a selection(clears current one)
Ctrl-v	Paste a selection

Template tool keys

Hot Key	Function
E	Rotate selection or template clockwise
Q	Rotate selection or template anticlockwise

Note: The items in the main **Edit** menu only apply to the data sets, not to the molecular editor.

6.1.1 Select tool

This is used for basic operations on structures and for editing atom types

To select an atom or bond click the desired item. When the mouse is near an atom or bond it will be highlighted to indicate which item will be selected.

To select multiple atoms or bonds, hold the **Ctrl key** and click on successive atoms or bonds. Alternatively, click and drag to lasso an area of interest.

You can move a selection by clicking and dragging and you can rotate a selection using the mouse wheel.
The standard cut/copy/paste operations are bound to keys **Ctrl-x**, **Ctrl-c**, and **Ctrl-v** respectively. Cutting a selection removes the selected atoms or bonds from the drawing area.

Note: Removing an atom will remove all connected bonds. Paste operations place a copy of a previously copied or cut atoms or bonds at the current mouse position.

Atom types can be changed by typing the specific atom label. By default, all atoms are carbon if not otherwise specified. For example, to change an atom from carbon to nitrogen, select the atom, or hover the select tool over the atom until it appears selected, and type **N**. To add charges simply type these in after the atom label using **+** or **-**.

6.1.2 Bond drawing tool

To draw a bond, move the mouse to the location where a bond is desired, press and hold down the left mouse button and drag the mouse to the end-point for the bond.

Bond types (single, double, triple) may be cycled by clicking on a specific bond.

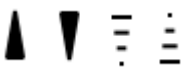
To edit an atom type, hover the bond tool over the atom until it is selected and then type in the atom label. To add charges simply type these in after the atom label using **+** or **-**.

6.1.3 Wedge bond drawing tool

To draw a wedge bond, move the mouse to the location where a wedge bond is desired, press and hold the left mouse button and drag the mouse to the end-point for the bond.

Clicking on an existing non wedge bond will convert the bond to a wedge bond.

The wedge bond type (dashed wedge or solid wedge) and the wedge bond direction may be cycled by clicking on a specific bond. Subsequent mouse clicks will cycle the wedge types in the following order:



6.1.4 Enhanced stereochemistry

To specify stereo-centres in the designer as members of the same enhanced stereo group, multi select the atoms then right click, and choose the group type from the **Enhanced Stereochemistry** menu. Enhanced stereo label choices include:

- **Absolute -** The absolute configuration is known and corresponds to that indicated by the wedge bond.
- **Mixture (And)** - The relative configuration is known. A mixture of the two enantiomeric relative configurations is present.
- **Relative (Or)** - The relative configuration is known. Only one of the two enantiomeric relative configurations is present. There is no assumption about which of the two configurations is present.
- **None** – This will remove any label.

Note: To view enhanced stereochemistry labels, please make sure that the **Show stereochemistry labels** option is checked in the Design preferences (see section 24.5).

6.1.5 Erase tool

To erase an atom or bond click on the item.

6.1.6 Template tools

These are used to draw simple chemical fragments.

When using a template tool, the template will be drawn under the mouse pointer. Click the left mouse button to add a template to the chemical drawing. Highlighted atoms, shown with a blue circle around them will be merged with the template when the left mouse button is clicked. Prior to clicking the left mouse button the template may be rotated to match atoms as described above.

6.1.7 Clean button

Click the **Clean** button to tidy the drawing, introducing standard bond lengths and angles and laying out the molecule in 2D keeping all atoms and bonds visible. Cleaning may also rearrange the wedge bonds. The following wedge bond modifications may be applied:

- Discard wedges involving atoms that cannot be stereo centres
- Discard wedges if the wedge tip is not on the stereo centre
- Transfer wedges in order to produce an 'improved' diagram. For example, transfer a wedge from a cyclic to an acyclic bond
- Discard wedges to retain only one wedge bond per stereo centre

6.1.8 Adding a molecule to a new data set

To add a molecule to a data set, click the  button. By default, new molecules are added at the position of the first selected row or at the end of a data set if no rows are selected.

6.1.9 Reset button

Click the **Reset** button to clear the drawing area.

6.1.10 Copy & paste images and molecules

Right-clicking over the drawing area displays a menu.

Copy Image copies an image of the structure to the clipboard allowing it to be pasted into other applications

Save Image saves the structure as a PNG image file.

Copy SMILES copies a SMILES string representing the structure to the clipboard which can be pasted into other structurally aware applications.

Paste Molecule enables you to paste a SMILES string from another application into the editor.

6.2 Searching for structures

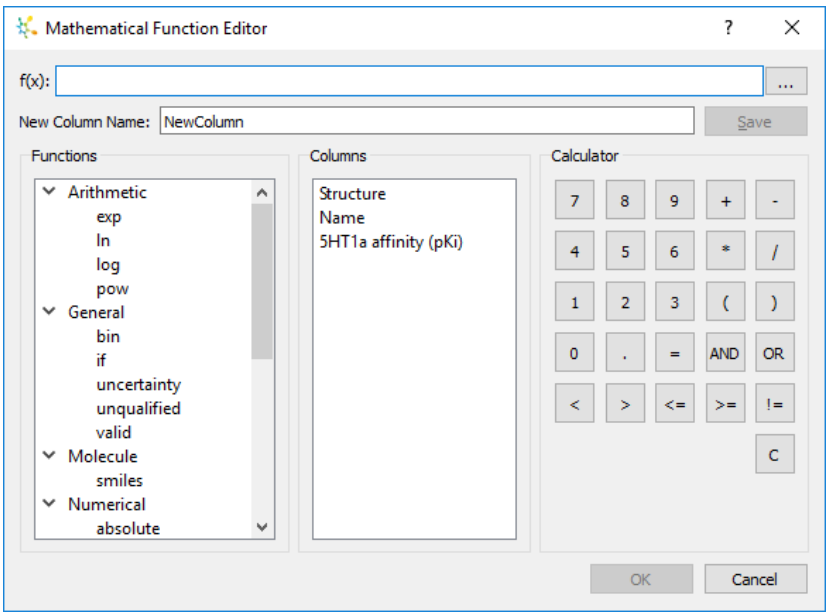
To search for molecules that contain the sub-structure displayed in the Design area, click the  button to display the Find dialogue. The structure copied into the search area. See section 8.1 for details on how to use the Find dialogue.

7 How do I... Organise my data?

A number of features are available in StarDrop to allow you to easily manipulate a data set.

7.1 Using mathematical functions

You can use the function editor to create new columns which are functions calculated using the values in other columns. Click the  button on the toolbar to open the **Mathematical Function Editor**.



Clicking on a function in the list displays a template in the editor with the first item that needs to be filled in already selected and ready for input.

To add column names, choose them from the list or type them in. Note: Column names should be enclosed in curly brackets e.g. {logP}.

If you wish to compare a value in a column with some text then the text should be enclosed within double quotes.

If you wish to generate categories in the results then the desired category names should be enclosed within single quotes.

Example:

To categorise logP values into high and low, simply use the function:

if({logP}>3,'high','low')

Clicking the **OK** button will close the dialogue and add a new column to the data set which has the name specified in the New Column Name.

If there is an error in the function, a message will appear indicating that StarDrop cannot calculate the result until the error is corrected.

Note: If values in columns which are used as part of the function are changed after the function column has been created, the values in the function column will update to reflect the changed values.

7.1.1 Function types

The following table lists the different types of function which are grouped together in the editor

Group	Description
Arithmetic	Functions which can be applied to numerical data
General	General functions
Numerical	Functions which can be applied to multiple numerical values. If one column is provided then the function is applied over the whole column. If multiple columns are provided then the function is applied across each row using the values from those columns
Molecule	Functions which can be applied to chemical structures
Text	Functions which can be applied to text or categorical data

7.1.2 Functions

The following table lists the function available along with a description

Function	Description
absolute	Return the absolute value for each value in a column
average	Calculate the mean of all values in a column. If multiple columns are given then calculate the mean across those columns in each row
bin	Create category bins to represent the entries in a column by specifying thresholds E.g. bin({logP},1.5,3,5)
contains	Return true or false indicating whether the first value contains the second E.g. contains({idcolumn},"MyID")
count	Count the number of valid values in a column. If multiple columns are given then count the number of valid entries across those columns in each row
endswith	Return true or false indicating whether the first value ends with the second E.g. endswith({idcolumn},"MyID")
exp	Calculate e raised to the power of each value in a column
if	Partition values into two categories based on a test. Use nested if statements to partition data into more than two categories
indexof	Return a number indicating the position at which the second value is found in the first, returning -1 if not found E.g. indexOf({idcolumn}, "MyID")
ln	Calculate the natural logarithm of each value in a column
log	Calculate the logarithm (base 10) of each value in a column
max	Return the maximum value in a column. If multiple columns are given then return the maximum value across those columns in each row
min	Return the minimum value in a column. If multiple columns are given then return the minimum value across those columns in each row

pow	Calculate the first value raised to the power of the second. If either value is a column then the column values will be used
replace	Return a copy of the text in which the first occurrence of the first value has been replaced by the second. E.g. replace({idcolumn}, "MyID")
smiles	Return the SMILES representation for each molecule in a column
startswith	Return true or false indicating whether the first value starts with the second E.g. startswith({idcolumn}, "MyID")
stddev	Calculate the standard deviation across all values in a column. If multiple columns are given then calculate the standard deviation across those columns in each row
substring	Return a substring of the text value starting at the character of the first value and of length given by the second E.g. substring({idcolumn},4,6)
sum	Return the sum of the values in a column. If multiple columns are given then return the sum across those columns in each row
uncertainty	Return the uncertainty for a value. For numbers this is a standard deviation, for categories this is a probability
unqualified	Return a copy of an entry without any qualifier
valid	Return true or false indicating whether the value is valid

7.1.3 Qualifiers

When data are qualified (e.g. >, >=, ~, <=, <), the qualification will be taken into consideration when the function is applied and the result qualified accordingly (e.g. >5 + >3 = >8). When qualifiers are not compatible and there is no way to combine them in the result then the result will be invalid. If you have qualified data and would like the function editor to ignore the qualifier then you can use the **unqualified** function, e.g.:

The function “{pIC50} - {logP}”, where the pIC50 is “<6” and the logP is “>2”, gives an invalid result.

Whereas:

The function ”{pIC50} - unqualified({logP})” gives a result of “<4”

Note: You could also apply unqualified to {pIC50} to give a result of “4”

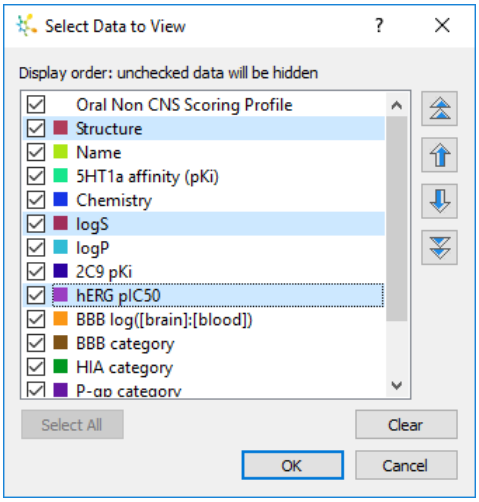
7.1.4 Propagation of error

When functions are applied, standard deviations and probabilities are calculated for the results based upon those in the input data. Please be aware that error estimates for non-linear functions are therefore only approximations, the accuracies of which are partially dependent on the mathematical function and partially on the scale of the numbers being used.

As a result, if data values are calculated externally and then imported with explicit errors specified as factors, scores generated using these numbers may differ marginally from scores generated from the equivalent data values calculated by functions within StarDrop.

7.2 Re-ordering columns/properties

To change the order in which properties are displayed click the  button on the toolbar to bring up the **Select Data to View** dialogue.



Drag and drop properties (or highlight them and click the  or  buttons) to change the order in which properties are displayed. The  and  buttons move selected items to the top or bottom of the list, respectively.

Note: The order will be changed in both the table view and the molecule view.

Uncheck and recheck the properties to hide and unhide them respectively.

7.3 Sorting data

A data set can be sorted by values within one or more specific columns or by structural similarity to a specific molecule.

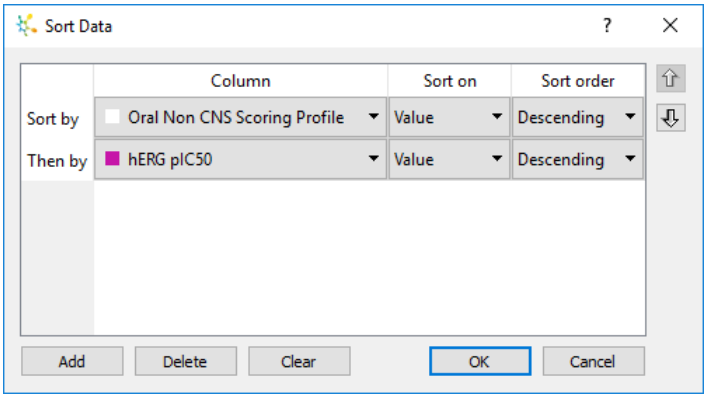
7.3.1 Sorting by a single column value

Click the right mouse button on the column title to display the menu and select **Sort** and either **Ascending** or **Descending**. The data within the data set will be sorted based upon the data in the selected column.

Alternatively you can sort the data based on their standard deviation or probability by clicking the right mouse button on the column title to display the menu and selecting **Sort by confidence** followed by either **Ascending** of **Descending**.

7.3.2 Sorting by multiple column values

Click the  button on the toolbar to display the **Sort Data** dialogue



Select the column you wish to sort on and then specify whether you wish to sort by the values or by the confidence (standard deviation or probability depending on the data type) and whether you wish to sort in ascending or descending order.

Click the **Add** button to sort by a second column. Select a row and click the **Delete** button if you wish to remove a column from the list. You can start again by clicking the **Clear** button.

To sort the data click the OK button.

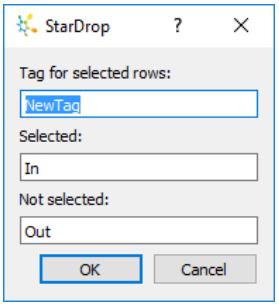
7.3.3 Sorting by rows

The data set can be sorted based upon the structural similarity to a molecule in a specific row. Click the right mouse button at the beginning of the row containing the comparator molecule to display the menu. Select **Sort by structural similarity** followed by **Ascending** or **Descending**:

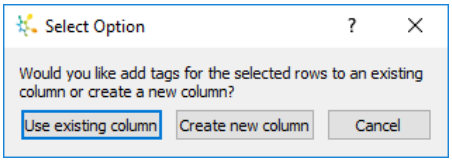
- If **Ascending** is selected, then the row chosen as a comparator will end up at the bottom of the data set.
- If **Descending** is selected, then the row chosen as a comparator will appear at the top.

7.4 Tagging data

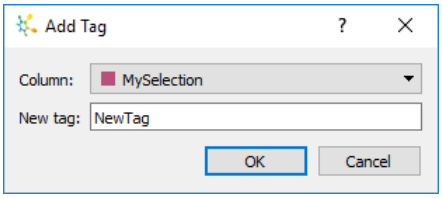
To remember any selected rows click the  button on the toolbar. This will display a dialogue enabling you to specify the name of a new column and a flag to add in each cell to remember whether or not that row was selected.



If you already have one or more category columns in your data set then you be given the option to add the tags to an existing column.



If you choose to use an existing column, then you can type in the tag you would like to use.



To tag rows quickly, use the keyboard shortcuts **Ctrl+t,I** and **Ctrl+t,o** to tag a row as being "In" or "Out" respectively. When using these shortcuts, if you have a category column selected then this will automatically be filled in with the tag value. If no column is selected then you will be prompted to choose a column, as described above.

7.5 Editing data

7.5.1 Editing rows

To edit one or more rows of data, select them and then right-click on the row label to bring up the menu.

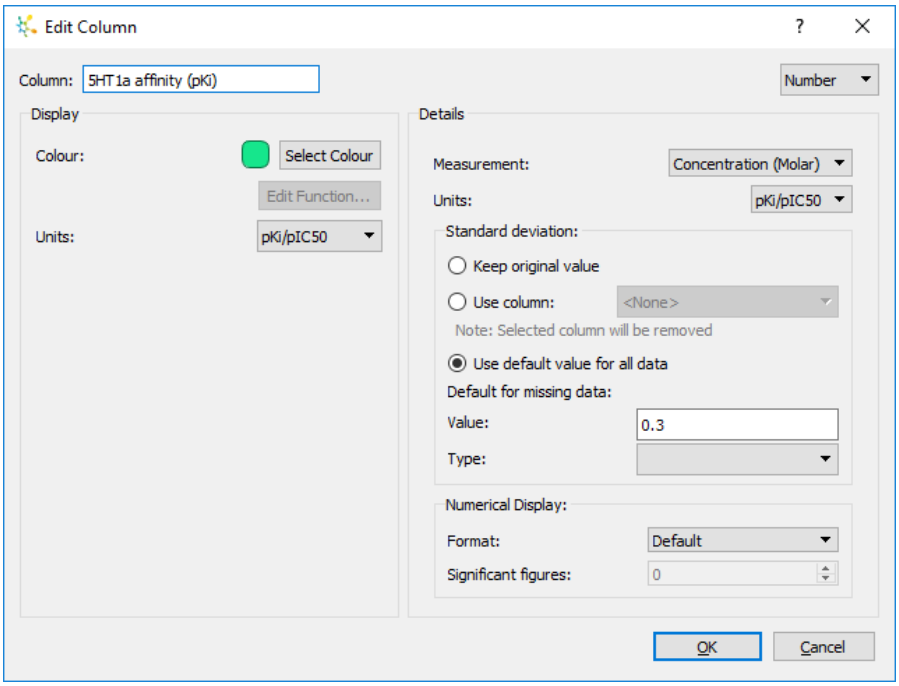
To **delete, cut or copy** the selected rows select the appropriate menu item. To **paste** data from the clipboard select the **Paste** menu item. Cut, copy and paste are also available on the **Edit** menu and via the shortcuts **Ctrl-x** (cut), **Ctrl-c** (copy) and **Ctrl-v** (paste).

7.5.2 Editing columns

To edit one or more columns of data right-click on the column name to display the menu:

To **delete** the selected columns select the **Delete** menu item. To **insert** a new column select the **Insert...** menu item and then enter a name for the new column in the dialogue.

To edit the **properties** and **data** of a column select the **Edit...** menu item. The **Edit Column** dialogue will be displayed.



This dialogue enables you to change different characteristics of the column, depending on the type of column selected. If the data was imported, then the Details section will show the same options that were available in the Import File dialogue (see section 3.1)

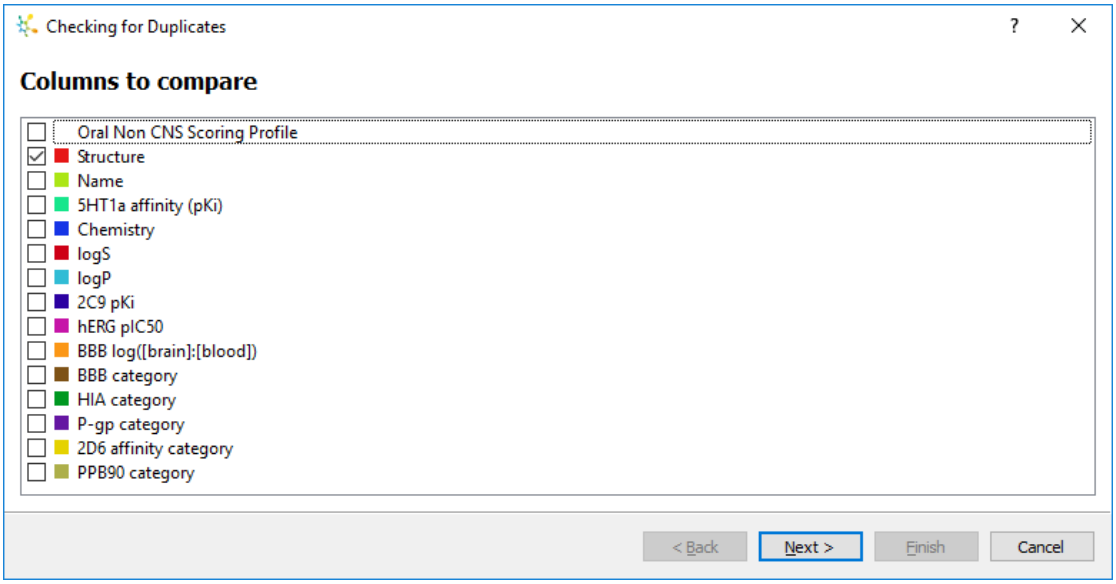
7.5.3 Editing cells

To edit the value in an individual cell, double-click the cell. This is only possible for numerical, category, date and text data.

Edit the values displayed in the dialogue that appears. Clicking OK will close the dialogue. Pressing the Enter key will put the value(s) into the cell and then display the data from the cell below to enable you to continue editing data down a column.

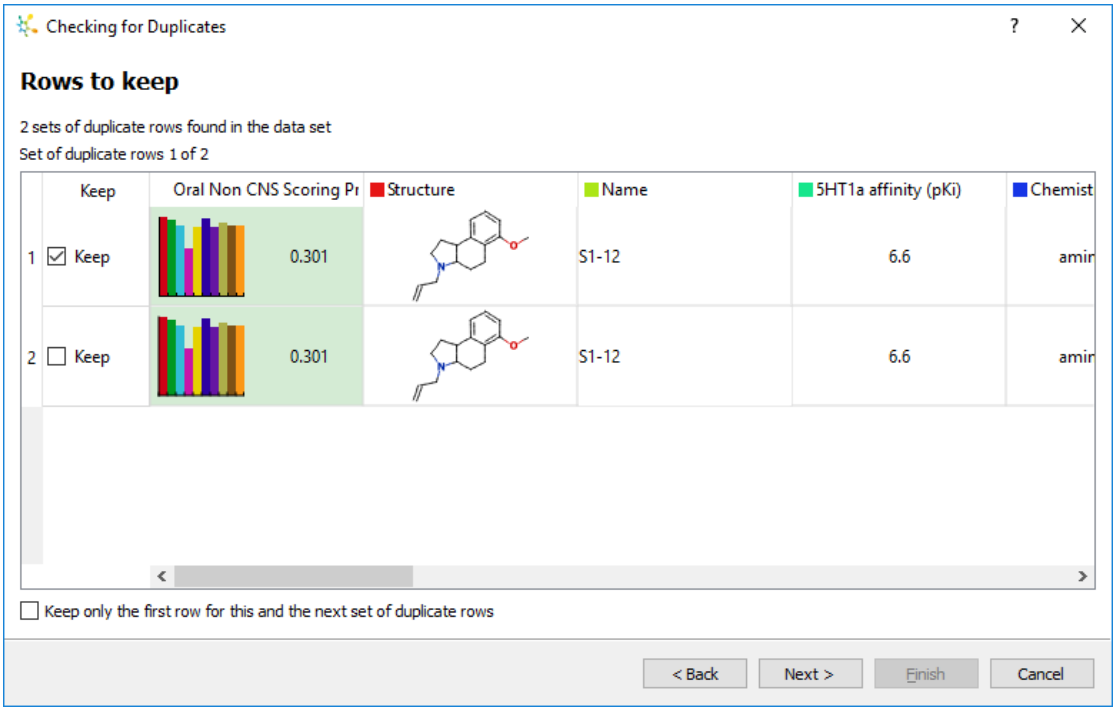
7.6 Checking for duplicates

To check and remove rows in your data set that that have duplicate information, select **Check For Duplicates...** from the **Data Set** menu.



Firstly, select the columns to be compared. If you select more than one column then the corresponding values in all of the columns must be identical for two rows to be considered duplicates.

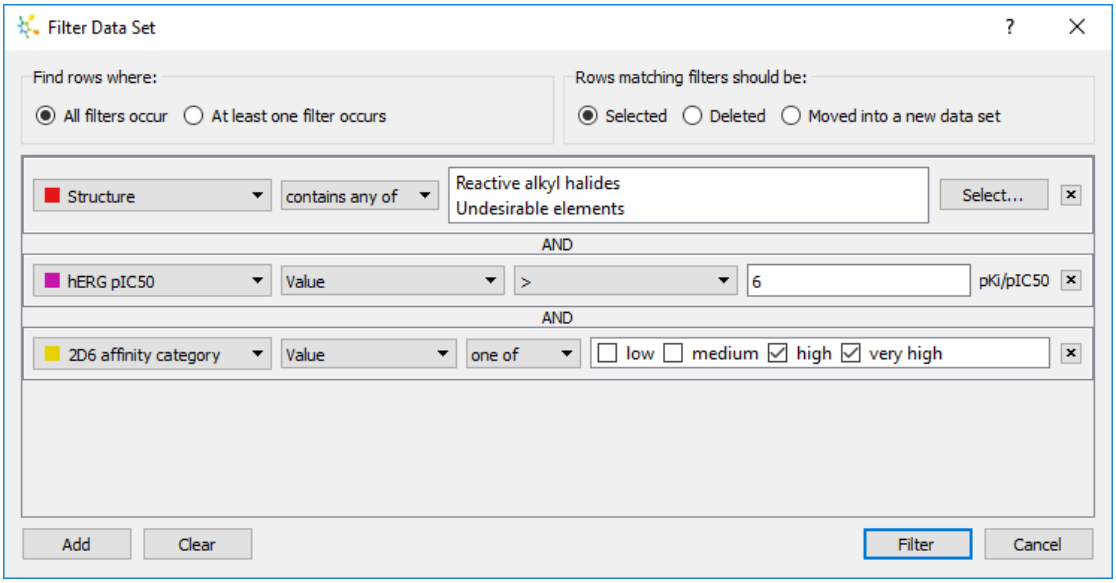
Once the data set has been analysed, the different sets of duplicate rows are shown, enabling you to confirm which rows should be kept and removed.



The final step of the process enables you to choose whether the duplicate rows should be removed from the actual data set or whether a new data should be created which just contains the duplicate-free rows.

7.7 Filtering data

To filter out data set rows that satisfy one or more criteria, select **Filter...** from **the Data Set** menu.



The dialogue that appears enables you to define as many filters as you need by clicking the **Add** button and then choosing the property and criteria. Each of the filters contains the name of the column to check, whether to filter by value or confidence, and the actual condition to be met, which will depend on the column type. The option to filter invalid values is also available.

The rows that match the combination of filters can be selected, deleted from the data set or moved out of the current data set into a new data set.

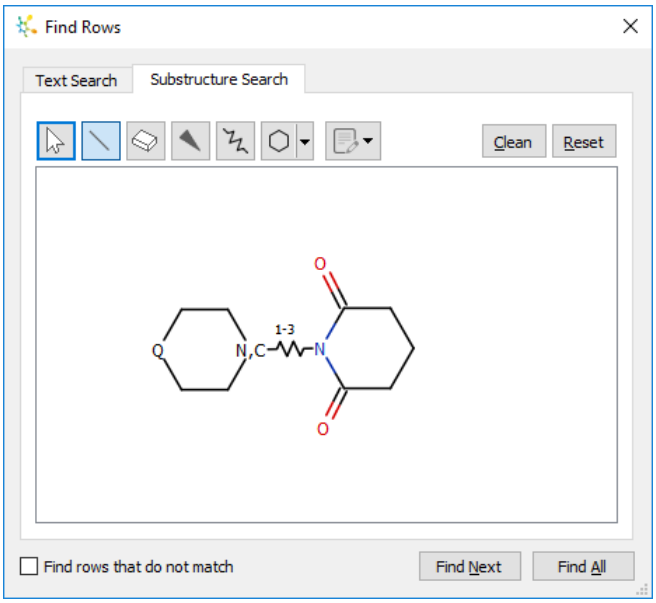
8 How do I... Find compounds or data?

StarDrop's Find tool enables you to search within a data set for substructures or within one or more selected columns for data. To start the Find tool, click the  button on the toolbar or press **Ctrl+f**.

Whether you are searching for data or a substructure, when you click the **Find Next** button StarDrop will search for the next row in the data set which matches the search pattern. If you click Find All then all rows which match the search pattern will be selected. If you tick the **Find rows that do not match** checkbox, then an inverse search will be performed.

8.1 Finding substructures

To search for a substructure, select the **Substructure Search** tab and use the drawing tools.



The editor is very similar to that which is in the main Design tab, however it enables you to define query structures which can represent multiple compounds rather than individual molecules.

As well as the standard drawing tools, described below, each atom, bond and linker drawn can be modified with different search constraints enabling them to represent multiple possibilities that could be matched when the search is performed.

Note: When sketching substructure search queries there are no implicit hydrogens. Hydrogen atoms can be sketched and hydrogen count constraints can be applied in the atom constraints.

8.1.1 Atom constraints

The following constraints can be specified for any atom in a query structure:

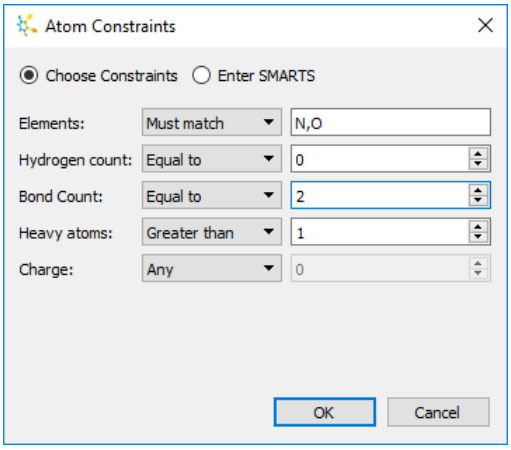
- Elements: one or more elements that must/must not match
- Hydrogen count: exactly/not equal to/greater than/less than a specified number of attached hydrogens
- Bond count: exactly/not equal to/greater than/less than a specified number of bonds
- Heavy atoms: exactly/not equal to/greater than/less than a specified number of bonds to heavy atoms
- Charge: exactly/not equal to/greater than/less than a specified charge

To specify elements for an atom you can simply select an atom and type the element you wish to match. If you type in multiple elements separated by commas, then any of these can match. The symbol "X" can be used to represent any halogen. The symbol "Q" can be used to match any

heteroatom. You can also use the symbol "!" before an element to indicate that this element must not match.

To add charges simply type these in after the atom label using **+** or **-**.

To edit atom constraints, select one or more atoms and click the  menu button and choose **Edit Atom Constraints…** to display the **Atom Constraints** dialogue.



If you choose the **Enter SMARTS** option, then you can paste a SMARTS pattern to represent the constraints that will be applied to the selected atoms.

Note: While some features of the substructure search go beyond what is possible with SMARTS patterns, there is also a small number of SMARTS features that are not supported.

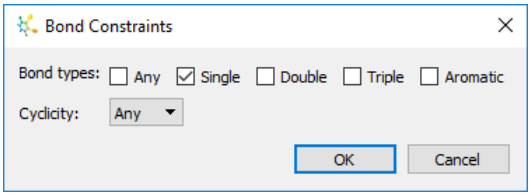
You can also modify atom constraints by selecting one or more atoms and then right-clicking to bring up a menu enabling you to specify each atom constraint individually.

8.1.2 Bond constraints

The following constraints can be specified for any bond in a query structure:

- Bond types: can be any, single, double, triple or aromatic
- Cyclicity: can be any, chain or ring
- Stereochemistry: wedge bonds can also match unspecified stereochemistry

To edit bond constraints, select one or more bonds and click the  menu button and choose **Edit Bond Constraints…** to display the **Bond Constraints** dialogue.



If the bond displays a wedge, then there is also a **Match Unspecified** option available.

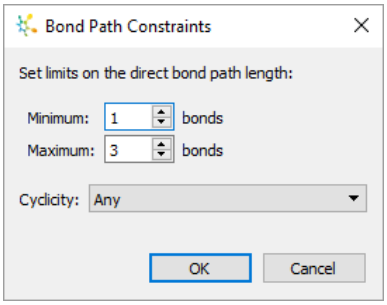
You can also modify bond constraints by selecting one or more bonds and then right-clicking to bring up a menu enabling you to specify each bond constraint individually.

8.1.3 Bond path constraints

The following constraints can be specified for any bond path in a query structure:

- Length: must be between a minimum and maximum specified number of bonds
- Cyclicity: must allow/not allow cyclic bonds in the direct path

To edit bond path constraints, select one or more variable paths and click the  menu button and choose **Edit Bond Path Constraints...** to display the **Bond Path Constraints** dialogue.



You can modify bond path constraints by selecting one or more bonds and then right-clicking to bring up a menu enabling you to specify each bond constraint individually.

8.1.4 Select tool

This is used for basic operations on structures and for editing atom types

To select an atom or bond click the desired item. When the mouse is near an atom or bond it will be highlighted to indicate which item will be selected.

To select multiple atoms or bonds, hold the **Ctrl key** and click on successive atoms or bonds. Alternatively, click and drag to lasso an area of interest.

You can move a selection by clicking and dragging and you can rotate a selection using the mouse wheel.
The standard cut/copy/paste operations are bound to keys **Ctrl-x**, **Ctrl-c**, and **Ctrl-v** respectively. Cutting a selection removes the selected atoms or bonds from the drawing area.

Note: Removing an atom will remove all connected bonds. Paste operations place a copy of a previously copied or cut atoms or bonds at the current mouse position.

8.1.5 Bond drawing tool

To draw a bond, move the mouse to the location where a bond is desired, press and hold down the left mouse button and drag the mouse to the end-point for the bond.

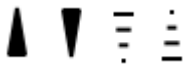
Bond types (single, double, triple) may be cycled by clicking on a specific bond.

8.1.6 Wedge bond drawing tool

To draw a wedge bond, move the mouse to the location where a wedge bond is desired, press and hold the left mouse button and drag the mouse to the end-point for the bond.

Clicking on an existing non wedge bond will convert the bond to a wedge bond.

The wedge bond type (dashed wedge or solid wedge) and the wedge bond direction may be cycled by clicking on a specific bond. Subsequent mouse clicks will cycle the wedge types in the following order:



If you

8.1.7 Erase tool

To erase an atom or bond click on the item.

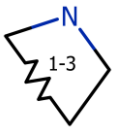
8.1.8 Template tools

These are used to draw simple chemical fragments.

When using a template tool, the template will be drawn under the mouse pointer. Click the left mouse button to add a template to the chemical drawing. Highlighted atoms, shown with a blue circle around them will be merged with the template when the left mouse button is clicked. Prior to clicking the left mouse button the template may be rotated to match atoms as described above.

8.1.9 Linker tool

The linker tool allows you to set limits on the length of the direct, or shortest, bond path between two atoms. This is useful for searching for linkers and rings of variable size.



8.1.10 Clean button

Click the **Clean** button to tidy the drawing, introducing standard bond lengths and angles and laying out the molecule in 2D keeping all atoms and bonds visible.

8.1.11 Reset button

Click the **Reset** button to clear the drawing area.

8.1.12 Copy & paste images and queries

Right-clicking over the drawing area displays a menu.

Copy Image copies an image of the query structure to the clipboard allowing it to be pasted into other applications

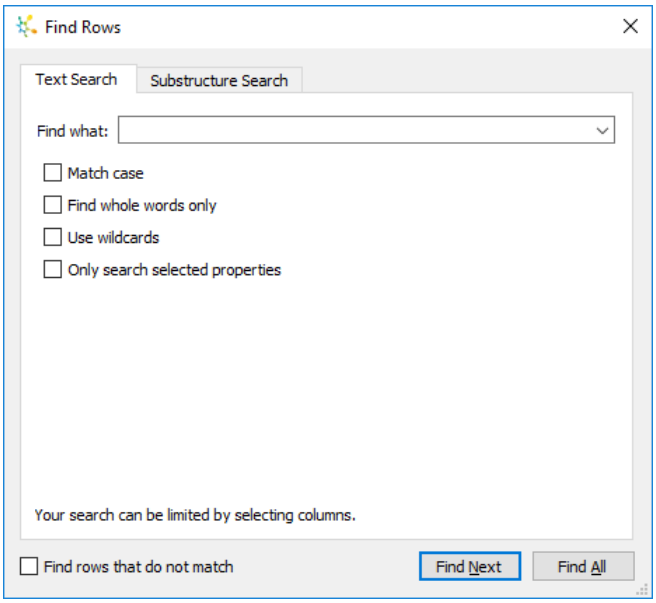
Save Image saves the query structure as a PNG image file.

Copy SMARTS copies a SMARTS pattern representing the query structure to the clipboard which can be pasted into other structurally aware applications.

Paste SMARTS enables you to paste a SMARTS pattern from another application into the editor. If the pattern can be interpreted as SMILES then the corresponding molecule is pasted. For example, benzene in kekule form, C1=CC=CC=C1, is pasted as a molecule with aromatic bonds.

8.2 Finding data

To search for a data, select the **Text Search** tab.

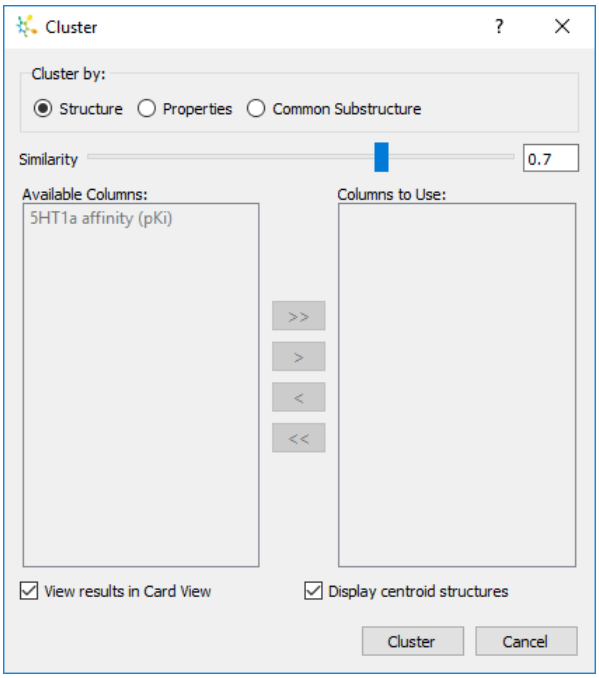


The drop-down list will show recent searches and you can further control the search that is performed using the **Match case**, **Find whole word only**, **Use wildcards** (inserting a * in place of characters) and **Only search selected properties** options.

9 How do I... Analyse my data?

9.1 Clustering

The clustering tool enables you to analyse your data by clustering it based on chemical structures, property values or common substructures. To cluster a data set, select **Clustering...** from the **Tools** menu. The clustering tool can also be accessed from within the **Analyses** drop-down menu in Card View.



Here you can indicate how you would like to cluster your compounds and you can indicate a minimum level of similarity which must be achieved for two rows to be clustered together - the default value is 0.7.

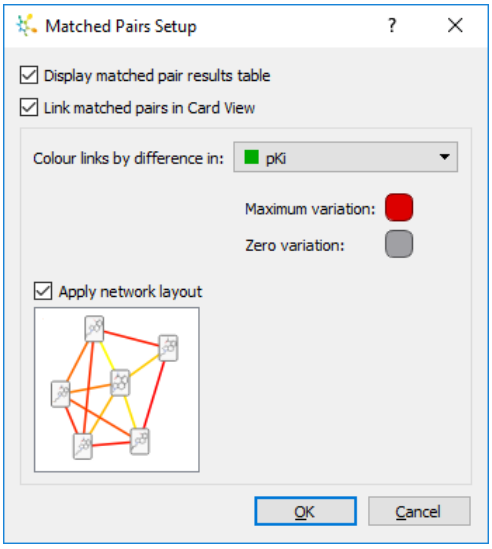
When clustering by structure, comparisons between rows are made by using the Tanimoto index to compare the molecules. When clustering by properties, comparisons are made by calculating the Euclidean distance between compounds using the properties that you have chosen.

The algorithm describing the process of clustering by common substructure is described in the StarDrop Reference Guide.

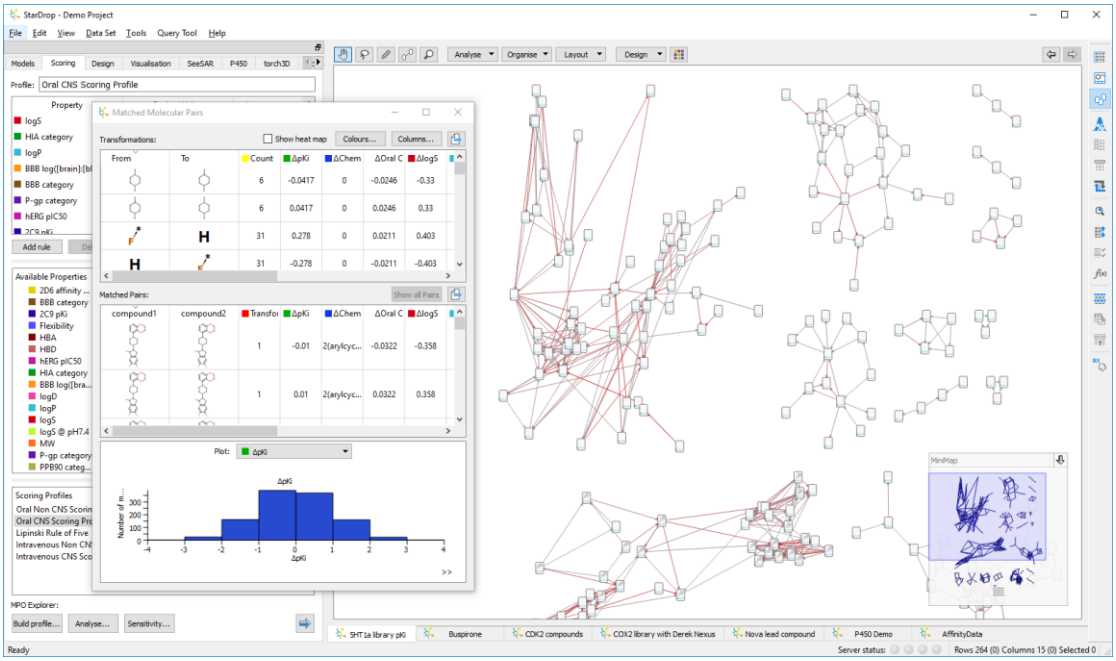
When the process is complete, two new columns are added to the data set indicating the cluster to which each row belongs and whether the row contains a cluster centroid or a cluster member. All the rows which do not belong to any cluster (singletons) are assigned to group 0.

9.2 Matched Pairs

The Matched Pairs tool is a generalised approach for finding pairs of compounds that differ by just a single functional group. For more information about the algorithm used to find the matched pairs please take a look at the StarDrop Reference Guide. To find matched pairs within a data set select **Find Matched Pairs** from the **Tools** menu. The matched pairs tool can also be accessed from within the **Analyses** drop-down menu in Card View.



The options enable you to choose how the results will be displayed; in a table, in Card View or both. If you choose to link matched pairs in Card view then you can choose a property for colouring the links. The links will be coloured to indicate the magnitude of the difference in property values between the pairs. An arrow will be shown on the link to indicate the direction of property increase. You can also choose to organise the results within Card View in a network.



The **Matched Molecular Pairs** results show two tables and a histogram.

The table at the top, **Transformations**, lists all of the functional group transformations which occur within the matched pairs found in the data set. The **Count** column indicates how often this transformation has occurred. The subsequent columns indicate the average change in property value when that transformation is seen – these are calculated for all the properties in the data set.

The table in the middle, **Matched Pairs**, shows all the pairs of compounds and the associated property columns show the change in property values between those two compounds.

The histogram at the bottom shows the distribution of property differences for the property selected in the drop-down menu above it.

If you select a row in the **Transformations** table then only compounds which relate to that transformation are shown in the **Matched Pairs** table and the histogram. At the same time, those compounds are selected in the data set (whichever view of the data set you are using). If your data set is displayed using Card View then the view will zoom and pan to show the selected compounds.

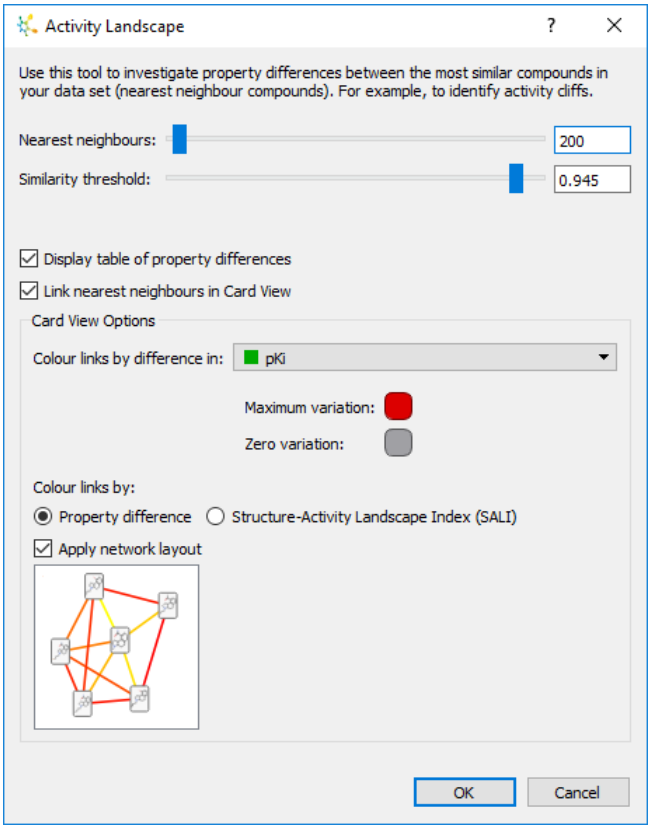
You can also choose to display a heat map in the tables by ticking the **Show heat map** option at the top.

You can edit the **Colours...** which are there to make it easy to spot which transformations significantly alter the property values and also to highlight whether there are trends where a transformation has consistently changed a particular property.

You can choose which properties are displayed in the tables by clicking the **Columns...** button.

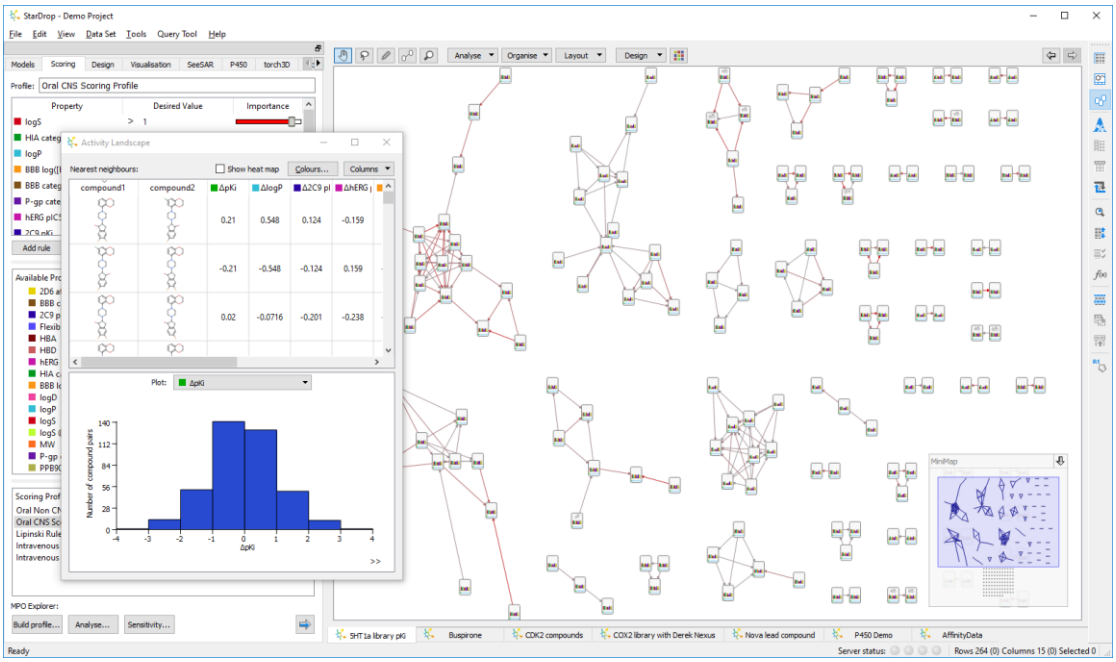
9.3 Activity Landscape

The Activity Landscape tool enables you to investigate property differences between similar compounds in your data set and can be accessed by selecting **Activity Landscape** from the **Tools** menu. The activity landscape tool can also be accessed from within the **Analyses** drop-down menu in Card View.



You can specify the number of nearest neighbours you are interested in finding, or the similarity threshold you would like to use. These are linked so that you will be able to see the similarity threshold required to achieve a given number of nearest neighbours and vice versa.

The additional options enable you to choose how the results will be displayed; in a table, in Card View or both. If you choose to link nearest neighbours in Card view then you can choose a property for colouring the links. The links will be coloured to indicate either the magnitude of the difference in property values between the pairs or the Structure-Activity Landscape Index (SALI). An arrow will be shown on the link to indicate the direction of increase. You can also choose to organise the results within Card View in a network.



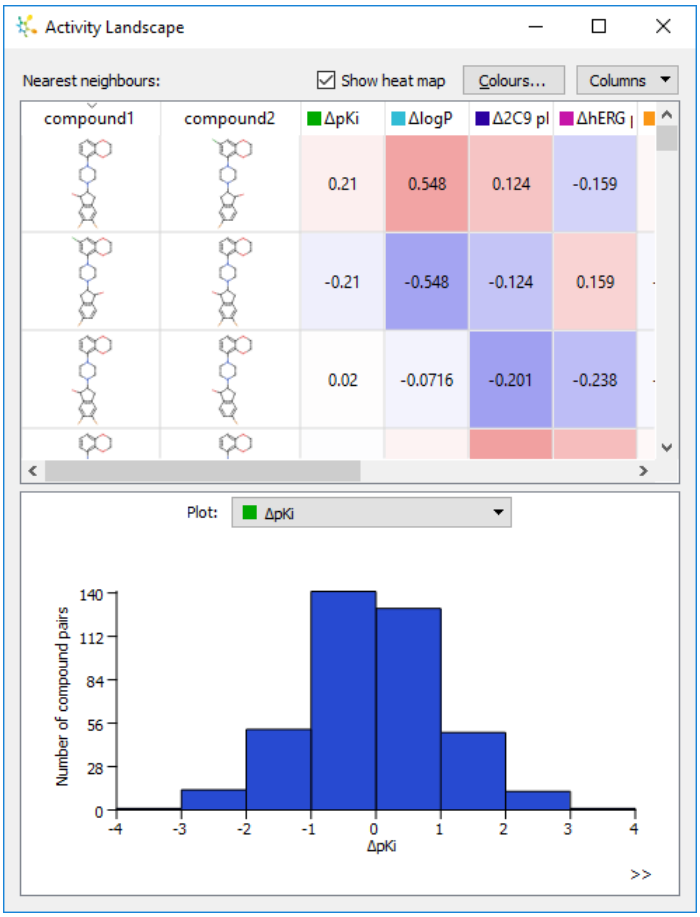
The **Activity Landscape** results shows a table and a histogram.

The **Nearest Neighbours** table lists all the pairs of nearest neighbours. The subsequent columns indicate the difference in property value between the first and the second compound in the pair – these are calculated for all the properties in the data set.

The histogram at the bottom shows the distribution of property differences between the pairs of compounds for the property selected in the drop-down menu above it.

If you select a row in the **Nearest Neighbours** table then those compounds are selected in the data set (whichever view of the data set you are using). If your data set is displayed using Card View then the view will zoom and pan to show the selected compounds.

You can also choose to display a heat map in the tables by ticking the **Show heat map** option at the top.

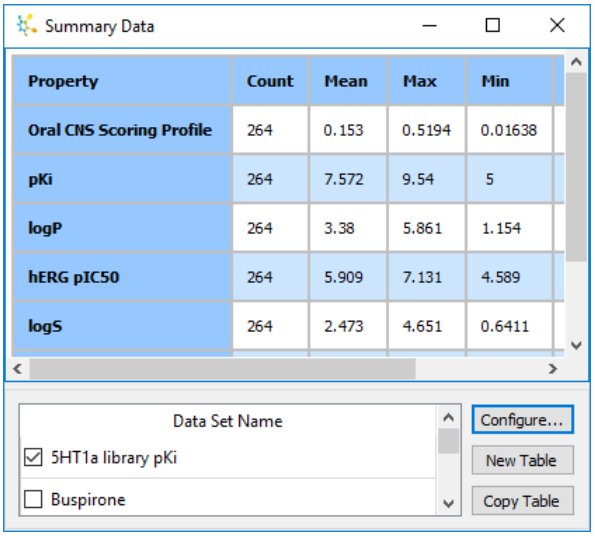


You can edit the **Colours...** which are there to make it easy to spot which pairs of compounds have significant property differences.

You can choose which properties are displayed in the tables by clicking the **Columns...** button.

9.4 Summary Analysis

You can create a summary of the data in one or more data sets by choosing the **Tools**, **Create Summary...** menu.



The window displays a summary of all of the properties in your data set. You can choose which data sets are include within the summary by ticking them below the table.

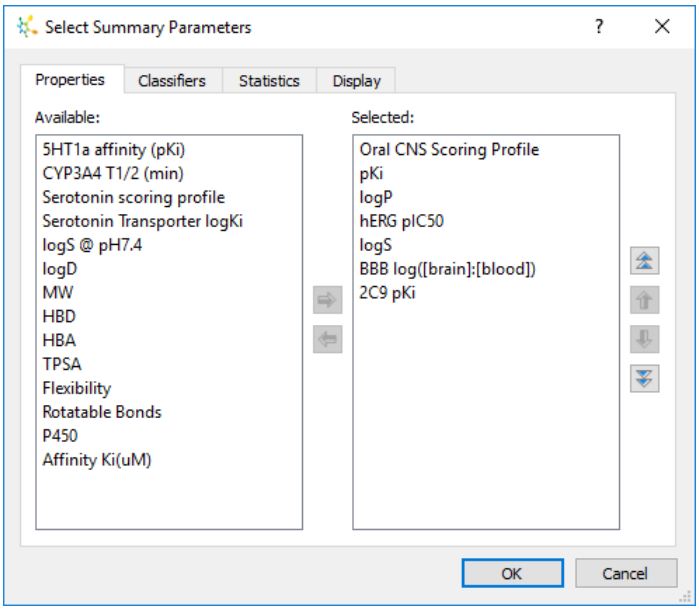
If you click the **New Table** button, a new window will appear with a copy of your current summary analysis which you can configure to show different statistics.

If you click the **Copy Table** button, the summary will be copied to the clipboard so that you can paste it into documents, presentations or other applications.

To configure the properties, classifiers, statistics and display, click the **Configure...** button.

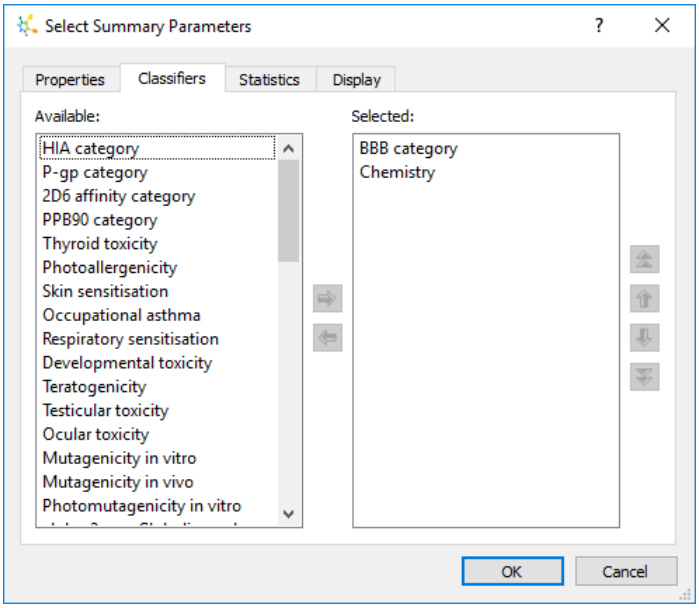
9.4.1 Properties

On the properties tab you can choose which properties to summarise in the table.



9.4.2 Classifiers

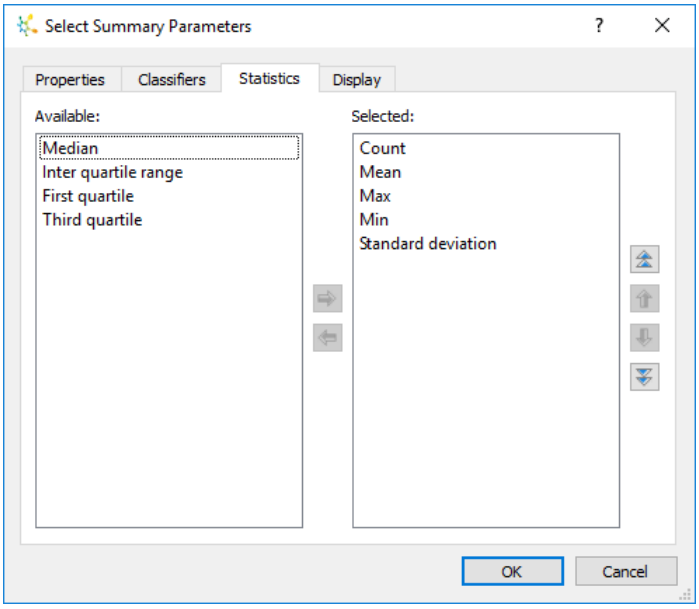
When you select a classifier (a categorical property) the summary table will be split to show the statistics for each of the different categories.



Selecting multiple classifiers will see each section of the table sub-divided further.

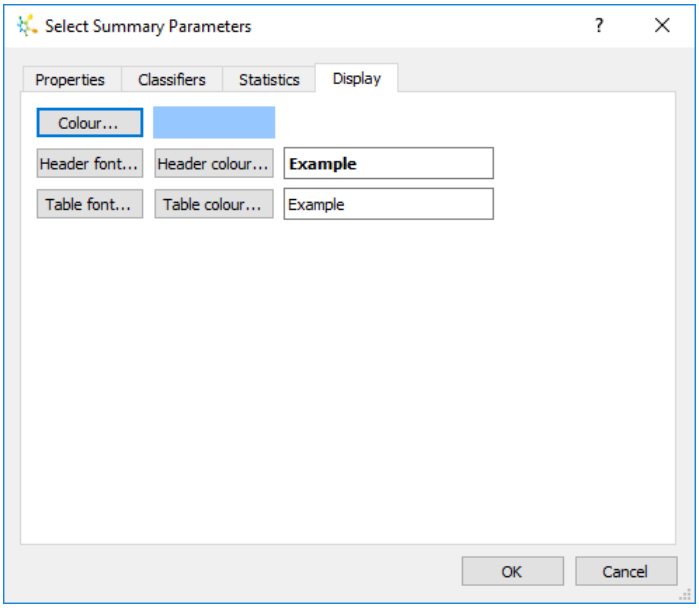
9.4.3 Statistics

You can choose which statistics to include within the table.



9.4.4 Display

The display options enable you to configure the way the table looks.

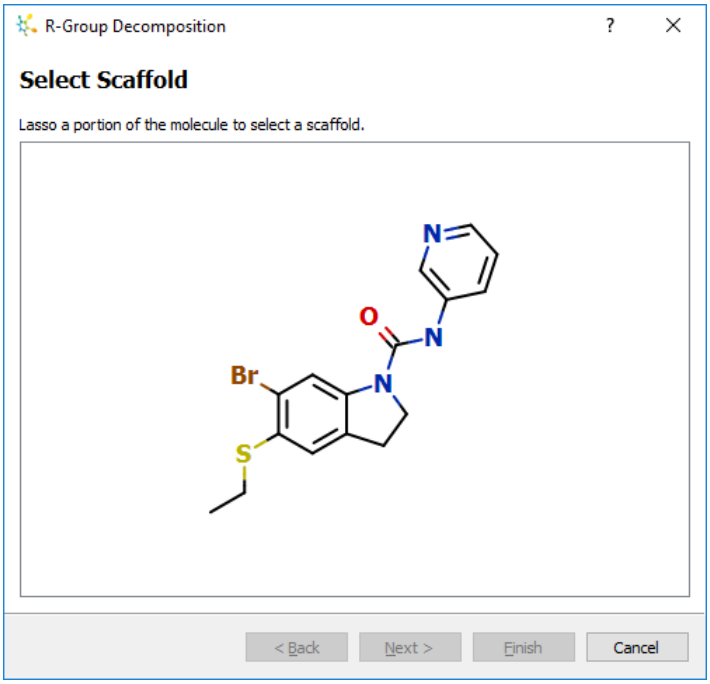


These can also be set in the preferences (see section 24.1) so that you do not need to configure them every time.

10 How do I... Carry out an R-Group decomposition?

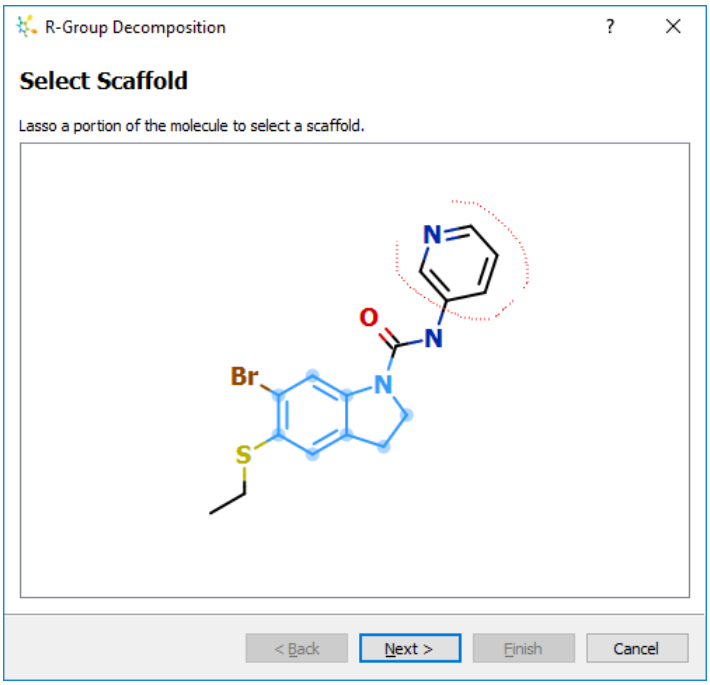
StarDrop's R-Group decomposition tool enables you to analyse a chemical series to investigate the impact of variations to R-Groups, linkers, atoms or fragments on compound properties.

To start the R-Group decomposition wizard, select an example molecule for each scaffold that exists in your data set and click the  button on the tool bar. Alternatively, select **R-Group Decomposition...** from the **Tools**, **R-Groups** menu.

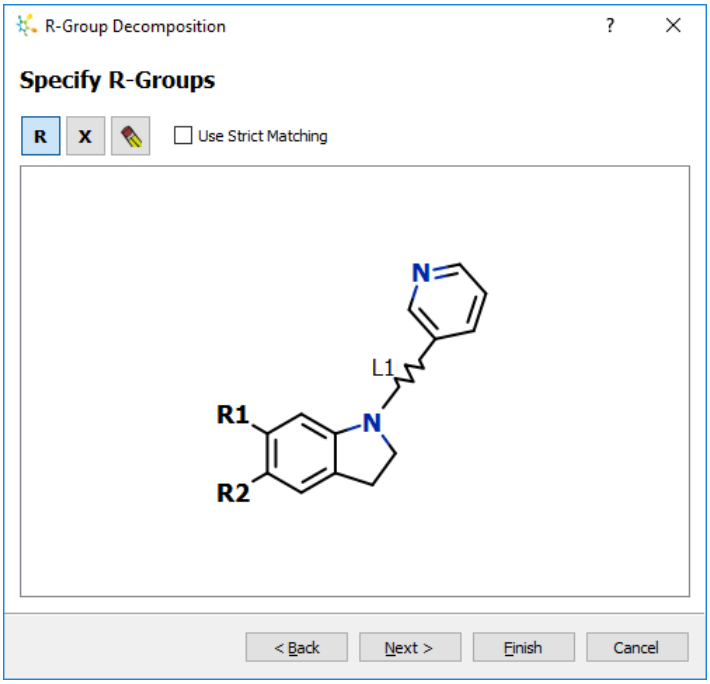


If no molecule is selected, the first one in the data set will be used.

Hold down the left mouse button and draw round the scaffold structure to select it. If there is more than one scaffold region, hold the SHIFT or CTRL key down while selecting the second region.



Click the **Next** button to confirm the R positions.



The selected parts of the molecule will remain visible, but the other regions will be displayed as R-Groups. If multiple regions were selected then the region joining them will be displayed as a linker.

To specify additional R-Group positions select the 

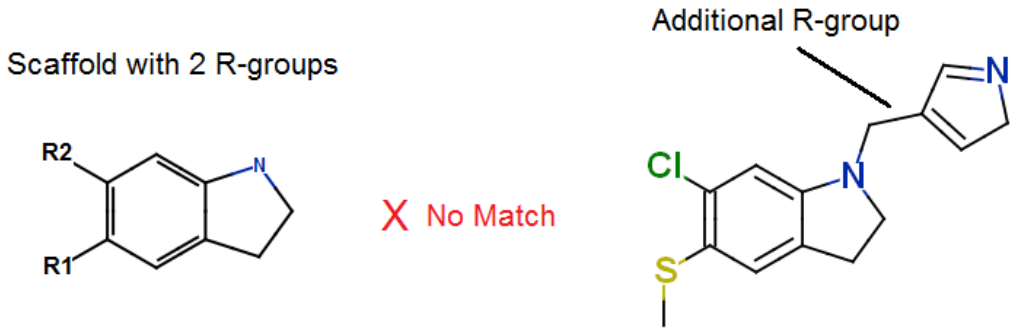
button and click at the positions on the molecule where these may occur. To change the name of an R-Group, select it and type in the new name.

To specify variable atoms, select the 

button and then click the atom positions which may vary within the data set.

To remove an R-Group or variable atom, select the  button and then click on the R-Group you wish to remove.

If you choose to **Use Strict Matching** then hydrogen atoms will be included in the scaffold definition.. You can use strict matching to avoid matching scaffolds that have *additional* R-Groups. For example, the following figure shows a scaffold that has two R-Groups. In this case, with strict matching on, the molecule on the right is *not matched* because it has an additional third R-Group. You may wish to avoid matching such molecules because any additional R-Group will not be present in the subsequent R-Group table.

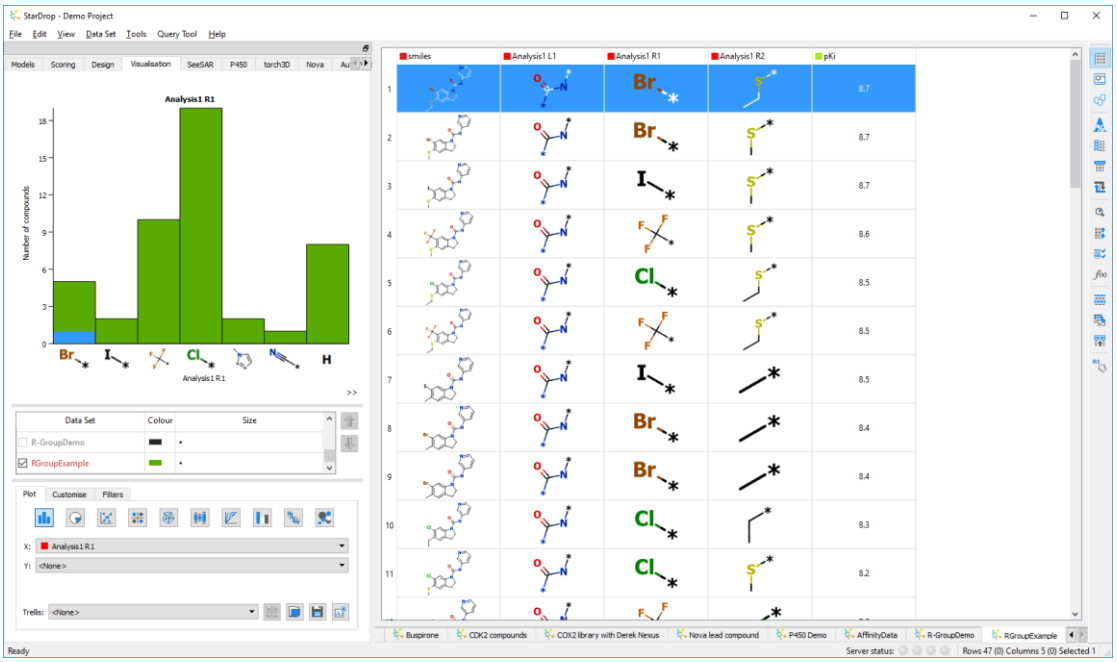


If you selected more than one starting molecule then the **Select Scaffold** and **Specify R-Groups** pages will be shown again for each additional molecule so that you can define each of the scaffolds and its corresponding R-Group positions in turn.

Once these have been defined, click the **Next** button to give the analysis a name and then click **Finish**.

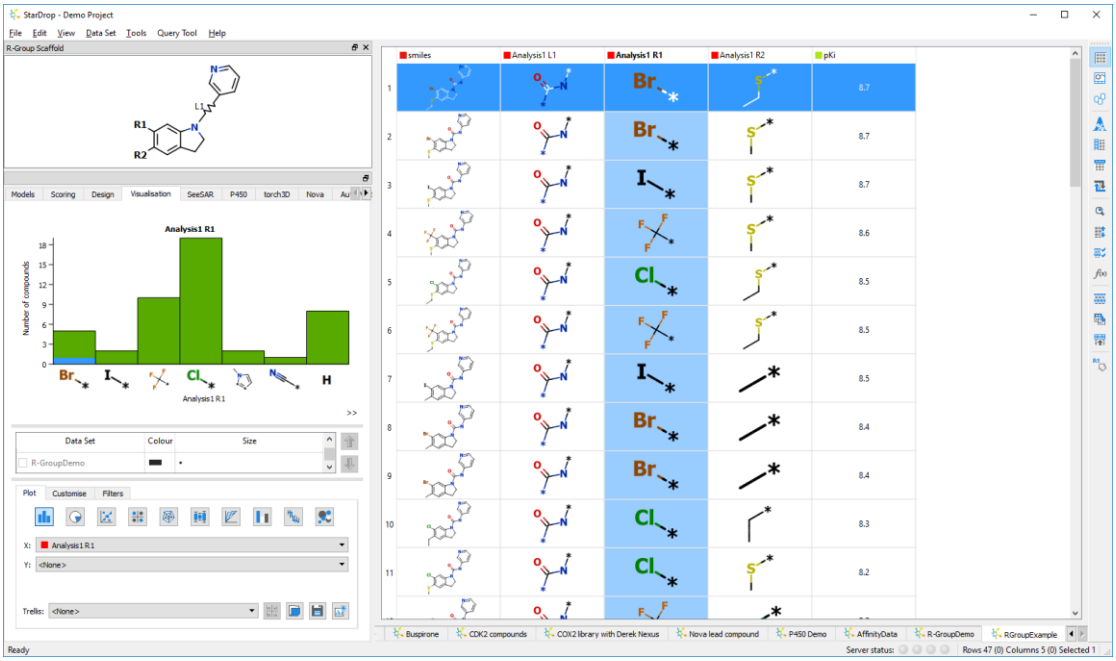
10.1 R-Group decomposition results

Once the wizard has finished, the R-Group decomposition will be carried out and the results will be added to the data set. Each R-Group position, variable atom or linker will be given its own column in the data set.



When calculating the R-Groups, where there is a choice over R group column assignment the following properties are used, in sequence, to determine the order of R-Group assignment: aromatic bond count, polar atom count and molecular weight.

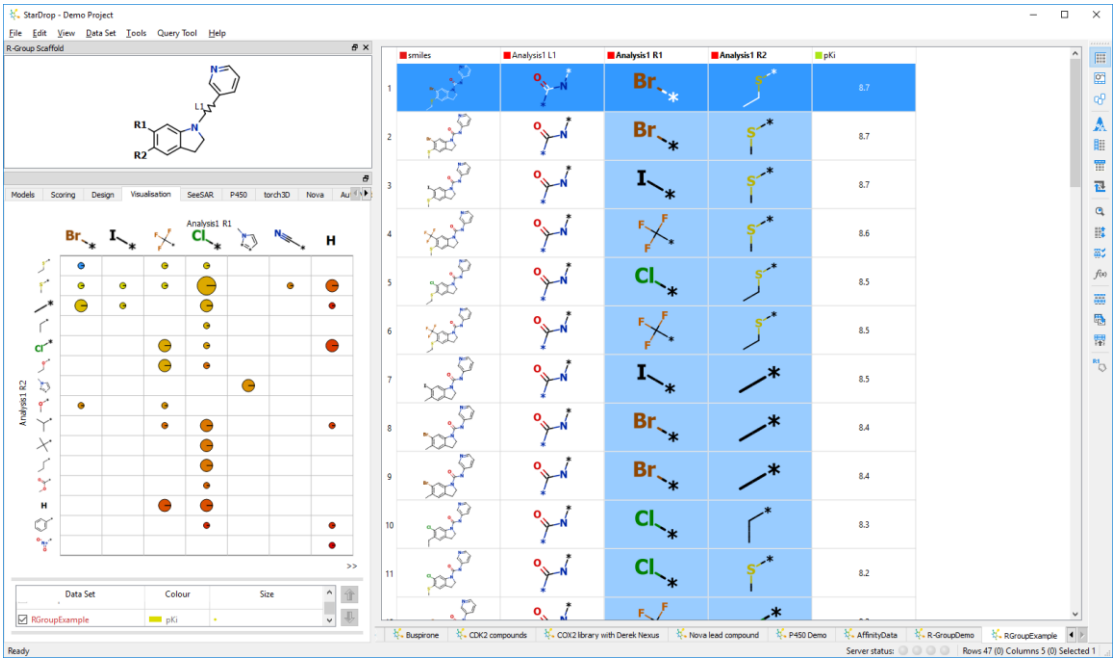
To see the scaffold structure at any time, hover the mouse over any of the R-Groups. Alternatively, right-click on the header of one of the R-Group columns and select menu option **View Scaffold**. Alternatively, select **View Scaffold** from the **Tools**, **R-Groups** menu. This will create a new window which contains the scaffold. This window can be docked if dragged to the tabbed area of the main window.



If you have carried out multiple R-Group analyses within your data set then the displayed scaffold will always be the one which relates to the R-Group analysis column that you have selected at any given time.

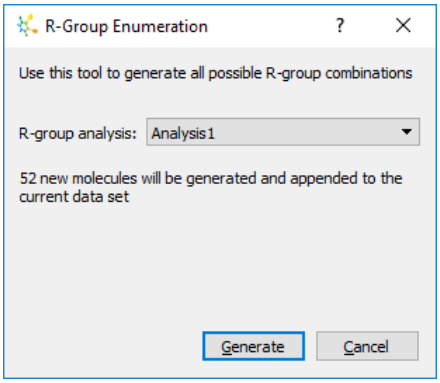
The results from the R-Group column can be visualised in the usual way by selecting those columns while viewing the Visualisation tab.

Selecting two R-Group columns and then colouring the plot based upon a property enables you to explore relationships between structural characteristics and a property.



10.2 Enumerating the possibilities

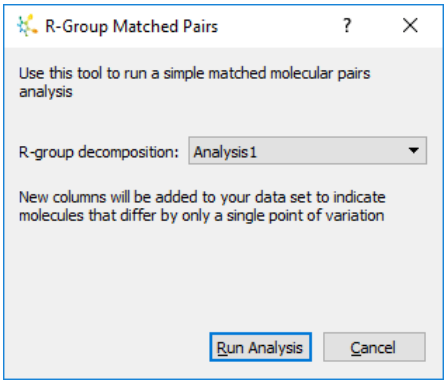
StarDrop can also generate molecules that represent all the missing combinations of R-Groups. This can be done by right-clicking on an R-Group column and selecting **R-Group Enumeration...**. Alternatively, this can be run by selecting **R-Group Enumeration...** from the **Tools**, **R-Groups** menu.



The **R-Group Enumeration** dialogue confirms which R-Group decomposition the enumeration should be based on (there's nothing to stop you carrying out multiple decompositions within a single data set) and indicates the approximate number of molecule that will be added to the data set when you click the **Generate** button.

10.3 R-Group Matched Pairs analysis

StarDrop can also perform an R-Group Matched Pairs analysis to identify pairs of molecules that have only a single point of variation which is based upon an R-Group enumeration. This is distinct from the more general tool for finding Matched Pairs which is described in section 9.2. To run an R-Group Matched Pairs analysis you must first perform an R-Group decomposition. Once you have a set of R-Group decomposition results you can right-click on an R-Group column and select **R-Group Matched Pairs…**. Alternatively, select **R-Group Matched Pairs…** from the **Tools**, **R-Groups menu**.



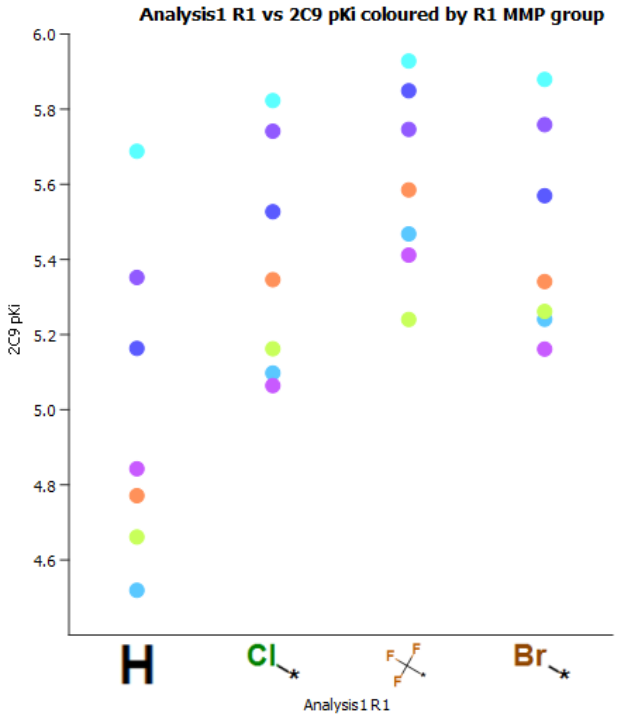
The **R-Group Matched Pairs** dialogue confirms the R-Group decomposition on which the matched pairs analysis is to be based.

Running the analysis will add new columns to the data set. One MMP column is added for each R-Group position.

	smiles	Analysis1 R1	Analysis1 R2	MMPAnalysis1 R1	MMPAnalysis1 R2	pKi
1				Group1	Group3	7.6
2				Group1	Group1	7.7
3				Group2	Group3	7.5
4				Group2	Group1	8

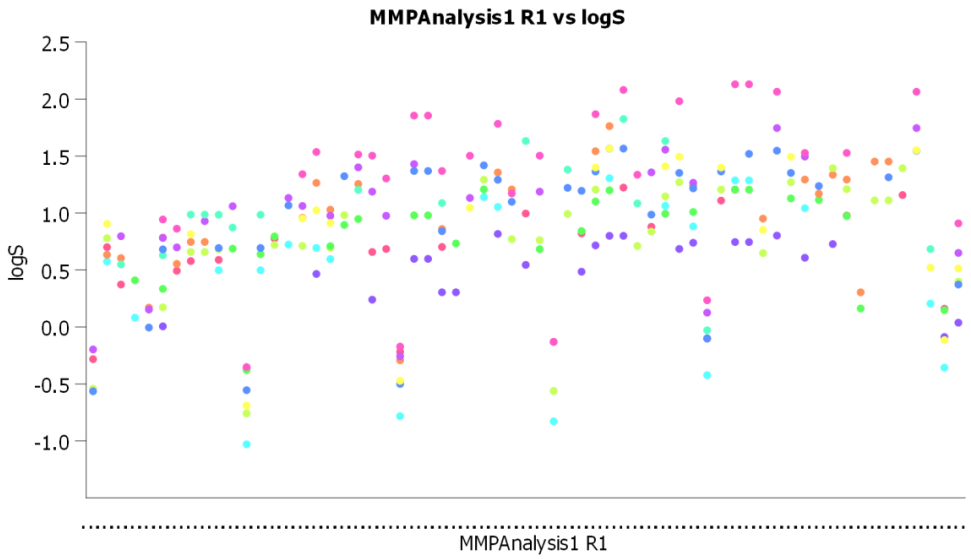
In this simple example, the 'MMPAnalysis1 R1' column identifies matched molecular pairs that differ only at the R1 position and the 'MMPAnlaysis1 R2' column identifies matched molecular pairs that differ only at the R2 position. Molecules that have matched pairs are assigned an 'MMP group'. Any two molecules that are in the same MMP group (for that column) are a matched pair. Also, all molecules that are in a group have all other substituents in common.

You may wish to sort the data set using an MMP column. You may also wish to use the MMP columns to create plots. One suggestion is to create a scatter plot of a property of interest, such as pKi, against an R-Group position such as R1 to see how particular substituent changes can affect a property. You can then colour the plot by the MMP analysis column for R1 to highlight trends that may exist between matched molecular pairs.



In this plot of pKi versus R1 the plot has been coloured by the MMP analysis results column for position R1. We can see that, on average, the change H → Cl, may have an effect to increase pKi. Colouring the plot in this way allows us to focus on cases where the change H → Cl is the *only* change made because we can identify trends by looking at the differences in pKi values between points of the same colour. The plot shows that for every case where H was replaced by Cl at R1, and all other substituents were constant, the effect was to increase the value of pKi.

An alternative plot that you may find useful is to plot the property of interest against an MMP column, as shown in the plot below. This is coloured by the substituent at the corresponding position.



In this case, a substituent that corresponds to a consistently high, or low, property value for multiple MMP groups may indicate a strong influence of that substituent on the property. The colours of the points may help to identify such a group. However with many different R-Groups, this can be difficult to see. In this case, you may find it useful to use a second plot of substituents at R1 to select combinations of R-Groups and use the option to plot **only selected data** (available in the **Customise** options of the Visualisation tab). This allows smaller numbers of R-Groups to be compared interactively.

11 How do I... Save or export my data?

After working with one or more data sets in a project you may wish to save your results for future analysis. A number of options are available.

11.1 Saving a project

This is the easiest way to save and reload data into the application for future analysis. Saving a project will preserve all data sets, visualisations, scoring profiles, summary analyses and functions.

From the **File** menu choose **Save Project**. If the project has been saved before, this will update the saved version. If the project has not been saved before, the **Save As** dialogue will appear. Specify the file name where prompted and click the **Save** button. The file will be saved with the suffix **.sdproj** to indicate that it is a StarDrop project file.

To save the data set as a separate file, select **Save Project As...** from the **File** menu. The same dialogue will appear giving you the option to choose a new file name for the project.

11.2 Saving a data set

If you wish to save the contents of a single data set there are a number of formats you can choose.

11.2.1 As a StarDrop file

From the **File** menu choose **Save Data Set As...** and the **Save As** dialogue will appear. Specify the file name where prompted and click the **Save** button. The file will be saved with the suffix **.add** to indicate that it is a StarDrop data set file.

11.2.2 As a SMILES file

Unless the data set only contains 2D structures and an identifier, saving a data set in this format is not recommended as it does not save all the information contained in the data set. The format of a SMILES file is such that only the structure and identifier information will be saved.

Note: If the column immediately to the right of the structure column contains text it will be used as the identifier for the SMILES string.

To save the data set as a SMILES file, select **Save Data Set As...** from the **File** menu, select SMILES Files from the **Save as type** drop-down list and specify the file name. The file will be saved with the suffix **.smi** to denote a SMILES file.

11.2.3 As an SD file

To export an SD File, select **Save Data Set As...** from the **File** menu, select either **V2000 SD Files** or **V3000 SD Files** from the **Save as type** drop-down list and specify the file name. The file will be saved with the suffix **.sdf** to denote an SD file.

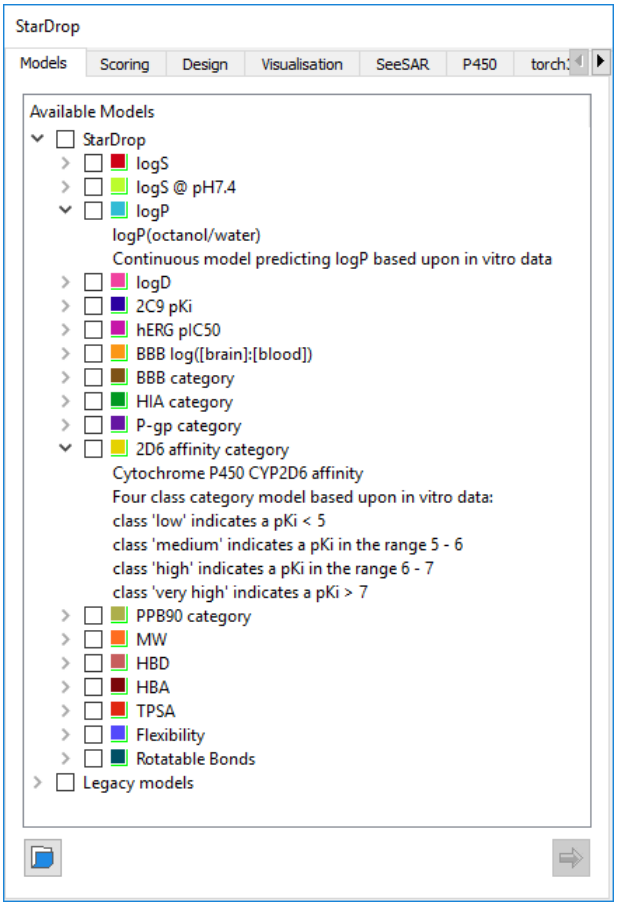
Noe: If your data set contains enhanced stereochemistry labels then please ensure that you choose the V3000 SD Files option.

11.2.4 As a Comma Separated Variable (CSV) file

To export a CSV File, select **Save Data Set As...** from the **File** menu, select Comma-Separated Variable Files from the **Save as type** drop-down list and specify the file name. The file will be saved with the suffix **.csv** to denote a CSV file.

12 How do I... Use models?

StarDrop contains a set of predictive models, and for users with the ADME QSAR module this will contain a range of ADME and physicochemical models. Click on the indicator next to a model name to display a brief description of that model:



Some additional information is also listed below:

2C9pK_i

The model output is pK_i (-$\log_{10}K_i$). A conversion table is shown below:

pK_i	K_i (µM)
3	1000
4	100
5	10
6	1
7	0.1

BBB log([brain]:[blood])
A conversion table for log ([brain]:[blood]) output is shown below:

logBB	([brain]:[blood])
-1	1:10
-0.5	1:3
-0.2	2:3
0	1:1
1	10:1

hERG pIC_{50}
A conversion table for the model output ($-\log_{10}IC_{50}$) is shown below:

pIC_{50}	IC_{50} (µM)
3	1000
4	100
5	10
6	1
7	0.1

logD
The model output is the logarithm of the distribution coefficient between a buffer at pH7.4 and octan-1-ol. Log $D_{7.4}$ differs from logP in that ionised species are considered as well as the neutral form of the molecule.

logS and logS@pH7.4
A conversion table of the logS output is shown below:

logS (µM)	Solubility (µM)
-1	0.1
0	1
1	10
2	100
3	1000

2D6 affinity category
A conversion table for the model output is shown below:

Class	pK_i Range	K_i Range (µM)
Low	<5	>10
Medium	5-6	1-10
High	6-7	0.1-1
Very High	>7	<0.1

BBB category
The classification boundary is nominally set at a logBB of -0.5 (brain:blood ratio of 1:3).

HIA category
The model predicts absorption from the human intestine into the hepatic portal vein. First pass metabolism effects are not considered, so this is not a direct indication of oral bioavailability. The data

used to generate the model relate primarily to passively absorbed compounds so active uptake and efflux are not considered.

Whole molecule properties

StarDrop also provides a number of whole molecule properties. These properties are often used inside the models and unlike the models are exact calculations. The following properties are available:

- Molecular weight - **MW**
- Proportion of total bonds that are rotatable **- Flexibility**
- **Rotatable bonds**
- Hydrogen bond donor count - **HBD**
- Hydrogen bond acceptor count - **HBA**
- Topological polar surface area (nitrogen, oxygen) - **TPSA**

Additional models

You can also load additional models that have been built using the Auto-Modeller or from external sources (see the StarDrop Scripting and Customisation Guide for further details of how to do this).

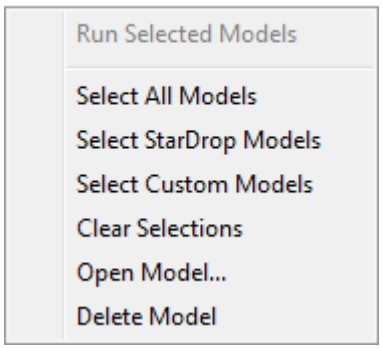
12.1 Running models

To run the ADME QSAR models, select the models to run using the checkboxes and click the  button. (Alternatively, click the right mouse button and select **Run Selected Models**.) A progress bar will be displayed while the models are run. Once the results are calculated, a new column will be added to the data set for each calculated property.

Note: Model calculations run on the server will incrementally populate the data set. For a large dataset, this process may take several minutes.

12.1.1 Models right-click menu

Right-clicking on the models tab will display a menu:



From this you can select or de-select models. Alternatively, you can choose **Open Model…** to add a new model to the list (perhaps one created by another user) or **Delete Model** to remove a model from the list.

12.2 Auto-Modeller models

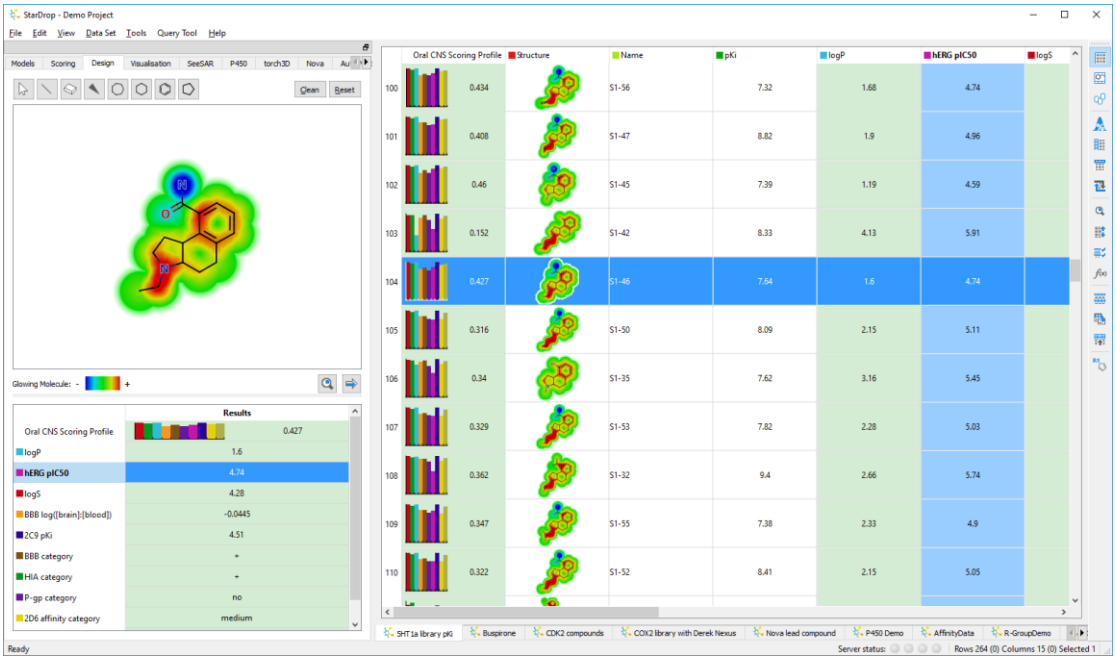
Models saved from the Auto-Modeller will also be available in the Models tab as Custom models. In addition, models created by the Auto-Modeller that have not yet been saved may also be temporarily available (see section 20.3).

13 How do I… Use Glowing Molecule™?

A **Glowing Molecule** is a representation of the model result indicating the parts of the molecule that have the greatest influence on the prediction. A Glowing Molecule is displayed as a heat map in which the regions of the molecule increasing the prediction are coloured red, the regions decreasing the prediction are coloured blue, and regions having no overall influence are coloured green. The colours are interpolated between these extremes to indicate a greater or lesser effect and the default colours can be changed by double-clicking on the button **Colours…** in the **General Preferences** (see section 24.1). **Note:** For some properties an increase might be good, whereas for others it may be bad.

To use Glowing Molecules you must have access to models capable of producing Glowing Molecule results. All the StarDrop ADME QSAR models and any models created by the Auto-Modeller (with the exception of those generated using external descriptors) can generate Glowing Molecules. Models that can generate Glowing Molecules are displayed in the **Models** tab with a green edge to the colour for that model.

To display Glowing Molecule results in a data set, select a column for which Glowing Molecule results have been calculated. This will result in the column of molecules being displayed with the corresponding Glowing Molecule representations:



Additionally, if the **Design** tab is visible, this will automatically select the corresponding row in the summary table so that any molecule displayed in the designer is also displayed as a Glowing Molecule for that property.

Note: The designer will only display Glowing Molecules if Enable Glowing Molecule is checked in the preferences (see secton 24.1).

To view the Glowing Molecule for any other property or score in the **Design** tab click on the corresponding row in the summary table. **Note**: This will not change the selected column in the data set. However changing the selected column in the data set will change the selected row in the summary table.

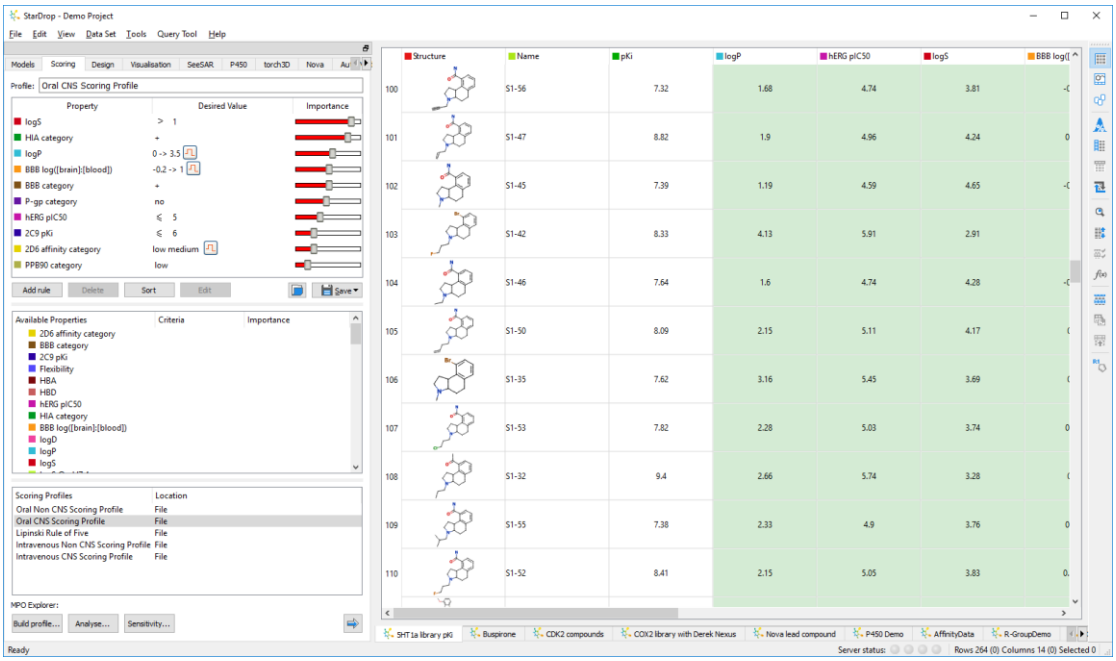
14 How do I... Score my compounds?

A standard set of **Scoring Profiles** are installed with StarDrop in the Examples folder. These profiles may be used directly or edited to produce scoring profiles specific to your target product profile. These will automatically be displayed in the **Scoring** tab.

Scoring profiles enable you to generate a score for every molecule in a data set indicating its likelihood of meeting all the project's requirements. The proprietary algorithm used not only utilises the values returned from models or experimental data, but also the uncertainty associated with the prediction or measurement.

14.1 Loading a scoring profile

Existing scoring profiles can be loaded into StarDrop in two ways. The default StarDrop profile and previously saved profiles in the directories specified in the preferences (see section 24.2) are displayed in the **Saved profiles** section of the scoring tab, and can be displayed simply by clicking on the appropriate entry.

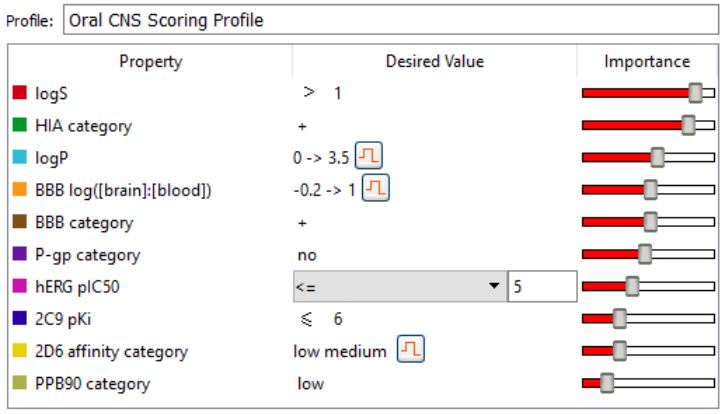


Alternatively, saved profiles can be loaded by clicking the  button. This will display the **Load Profile** dialogue from which you can browse for and select a scoring profile.

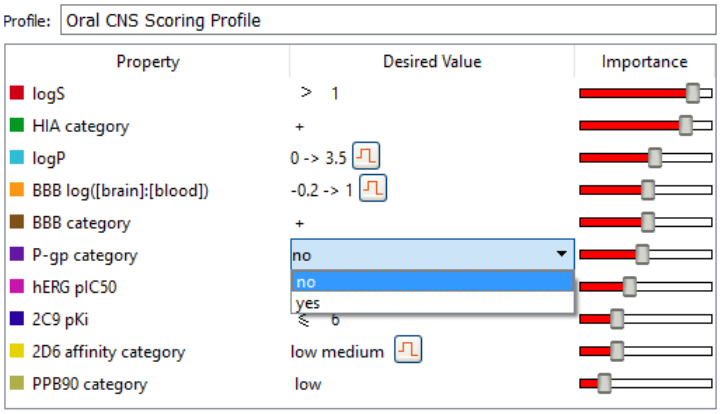
14.2 Editing a scoring profile

Profiles can be edited in the Scoring tab:

To edit a value, click on the entry and the display will change to show two controls that allow you to both change the threshold and indicate whether the desired value is greater or less than the threshold.



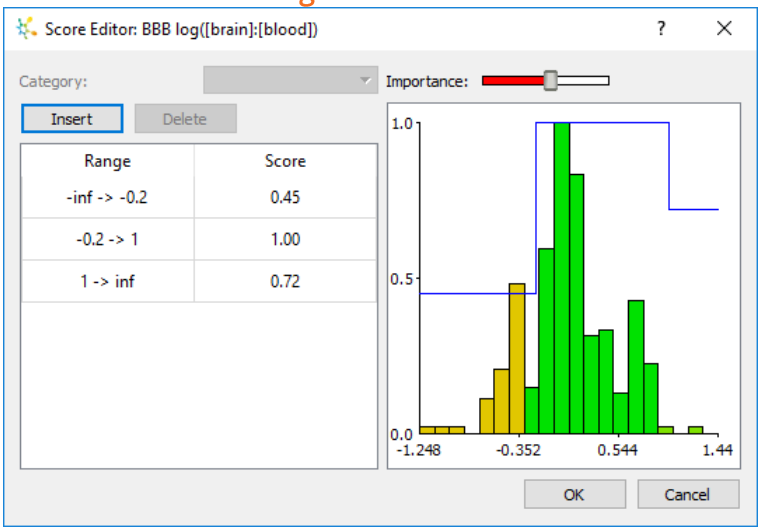
To change the preferred category for a classification model, click on the **Desired Value** as before. In this case the display will show a drop-down menu listing the available categories.



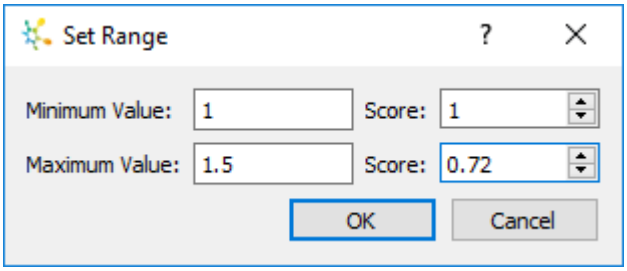
To change the importance of a property in the profile, use the slider bar in the **Importance** column. Moving the slider to the right makes the property more important. The greater the importance, the higher the penalty that will be applied for compounds that fail to meet the desired criteria. Saved profiles are always displayed with the most important property at the top, but you can sort a profile manually while editing by clicking the **Sort** button or by selecting **Sort** from the menu.

The importance slider represents an exact number which can be seen by hovering the mouse over the slider bar. More precise control of these details is available in the **Edit Score** dialogue (see section 14.2.1), which can be displayed by double-clicking a property in the profile or by clicking the **Edit** button when a property is highlighted.

14.2.1 Edit Score Dialogue



Clicking the Insert button displays a small dialogue enabling you to add a range to the list and indicate the score to assign for that range.



This will then be added into the list and the previous ranges altered to accommodate the newly specified values.

When criteria containing more than two ranges have been defined these cannot be easily edited using the drop-down lists in the Scoring tab. Any such criteria show an indicator  .

To add a property to a profile drag the item from the **Available properties** box into the **Profile** window. The **Available properties** box displays all the properties that can be included in your scoring profile. If you wish to save a specific scoring function for a property, drag that item from the **Profile** back into the **Available properties** box. The item will be added as a sub-item to the entry with the same name enabling you to save multiple functions for the same property.

To remove a property from a profile, select it and press the **Delete** key.

To remove all properties and begin a new profile, right-click over one of the properties in the profile and select the **Clear** option from the menu.

14.3 Saving a scoring profile

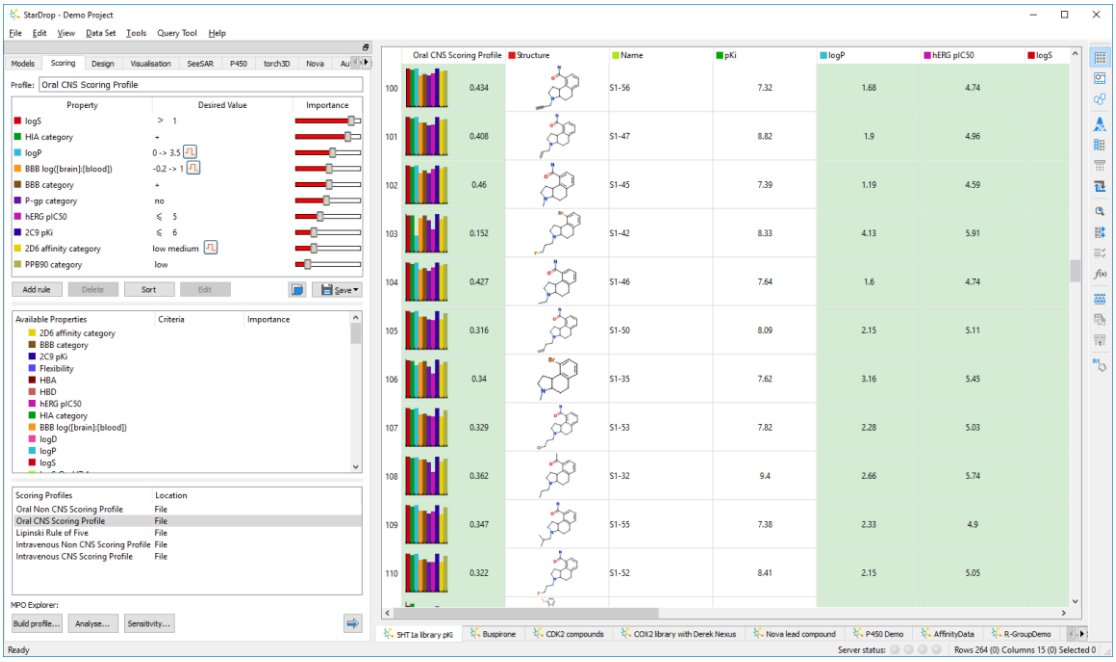
Scoring profiles can be saved for future use by clicking the **Save** button menu which enables you to choose whether to save the profile into the current project or as a file. Alternatively you can right-click over the scoring profile to bring up the menu. Specify a location and file name click the **Save** button. If the profile is saved as a separate file, then file will be saved with a '.apd' suffix indicating that it is a StarDrop scoring profile file.

If the profile was originally loaded from a file, the **Save** menu item will also be enabled. Selecting this will cause the original file to be overwritten with the new profile.

14.4 Using a scoring profile

Any scoring profile displayed in the **Profile** box can be used to score the current data set. To generate scores for a data set, click the  button or right-click to bring up the menu and click **Run**.

If all the properties specified in the profile are present in the data set, then the profile will run immediately, and the results will be displayed in an additional column. The column title will be the name of the scoring profile used. (For an explanation of the results column, see section 14.5).



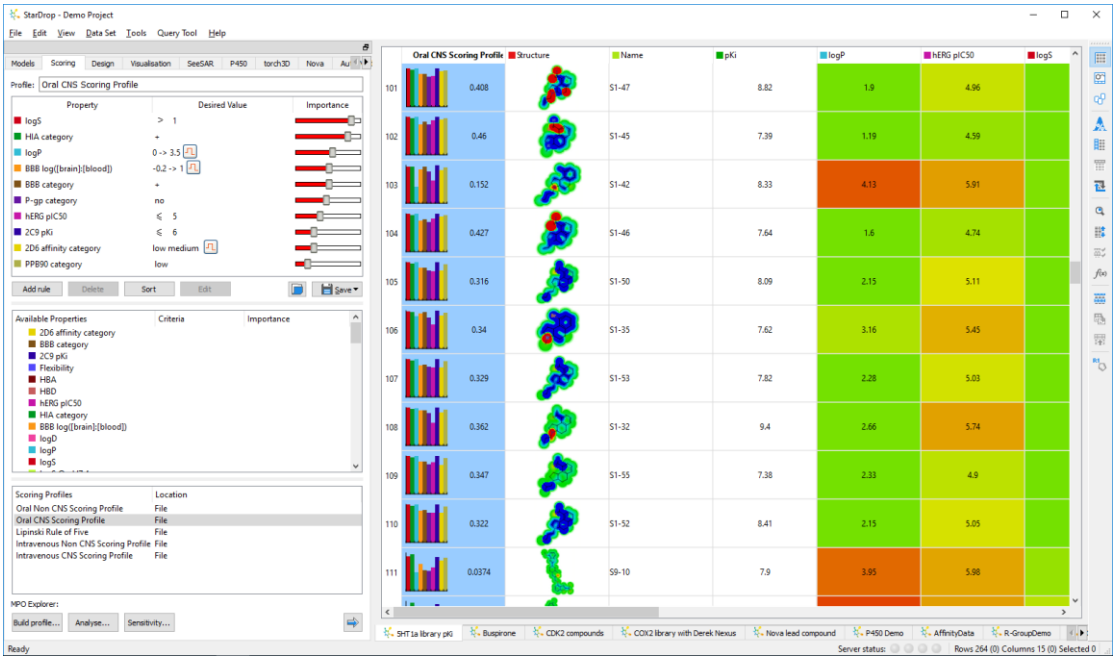
Alternatively, if the profile contains properties that are not present in the data set but can be calculated by StarDrop, you will be prompted to indicate whether you would like these values to be added to the data set when StarDrop calculates the scores.

If the profile contains properties that are not present in the data set which cannot be calculated by StarDrop you will be prompted to indicate whether you wish to continue generating scores. In the scoring preferences (section 24.4) you can indicate whether or not you wish StarDrop to calculate scores when there are missing data. If StarDrop does calculate a score with missing data, the histogram that is displayed will be faded to indicate that this has happened.

14.5 Explanation of scoring profile results

Scores range from 0 to 1. They have an associated uncertainty that can be displayed in the same way as for model results. The higher the value the more likely it is that a molecule meets the criteria in the scoring profile and thus the target product profile. The histogram displays the relative contributions of assay or model data to the overall score.

If you select the column of scores, the scoring profile used will automatically be displayed in the Scoring tab. In addition, the properties that have contributed to the overall score will be highlighted in the data set with a colour indicating how well that property contributed to the score.



By default, properties that scored well will be coloured green, properties that scored badly will be red, and properties that scored somewhere in between will be shaded to reflect this. These colours can be defined by clicking the **Colours…** button in the **General Preferences** (see section 24.1).

14.6 Creating a new scoring profile

To create a new scoring profile, right-click over the Profile box and select **Clear** to remove any currently displayed entries. Then add the desired properties to the profile by selecting them in the **Available properties** box and dragging them into the **Profile** box. Each property can then be edited as previously described in order to customize the profile. When the profile is complete, add a name by entering it in the text window at the top. The profile can now be run against the current data set, or saved for future use.

Creating good scoring profiles can be challenging, so listed below are some tips to assist you when creating Scoring Profiles:

2C9pK_i

If metabolism by CYP2C9 is a concern, we would normally recommend a pK_i less than 5. In the absence of a metabolism issue, we would recommend a pK_i less than 6 to avoid drug-drug interactions.

logBB([Brain]:[Blood])

A logBB below -0.5 almost always indicates that there will be no BBB penetration. With a logBB above -0.2 BBB penetration is usually present unless the compound is a substrate for active efflux transporters such as P-gp. Between these values, there is a less obvious range where the CNS activity will be dependent on the potency of the compound. With higher potency, less BBB penetration is acceptable.

hERG pIC_{50}

The criteria for hERG inhibition will ultimately depend on the relative affinity for this channel against potency for the therapeutic target. Typically, a factor of at least 100 is recommended. However, in the absence of known potency, a pIC_{50}>5 would suggest the need for experimental confirmation and pIC_{50}>6 would be a significant cause for concern.

logP

The logP of compounds is already taken into account in the other StarDrop models, so this should not be used as a criterion for properties such as solubility or absorption. However, we find logP to be a useful indicator of the risk of metabolism, particularly by CYP3A4. We would recommend a logP of <3.5 as this significantly decreases the risk of metabolism by this abundant enzyme. Of course, for many projects this may not be achievable due to the lipophilicity required for target affinity. In these cases other approaches to minimizing metabolism, such as reducing site lability may be necessary

logD

The logD of compounds is already taken into account in the other StarDrop models, so this should not be used as a criterion for properties such as solubility or absorption. However, we find logD to be a useful indicator of the risk of metabolism, particularly by CYP3A4.

logS / logS @ pH 7.4

Ideally, compounds should have logS greater than 2, indicating that compounds will be comfortably in solution at typical assay conditions, in the gut and in circulation. In practice this is not achievable for many projects and effort will be needed to achieve a suitable formulation for compounds with lower solubility. Therefore, we would recommend a minimum threshold for logS of 1.

2D6 Affinity Category

The primary concern with CYP2D6 inhibition is the potential for drug-drug interactions. Therefore, we would recommend a classification of Low or Medium to avoid this. If metabolism by CYP2D6 is a concern, then a classification of Low is recommended.

P-gp Category

P-gp is expressed in the gut wall and hence can limit oral bioavailability. However, potentially high drug concentrations in the gut means that well absorbed compounds typically saturate P-gp efflux and absorption is not significantly affected. P-gp transport is of greater concern for compounds intended for a CNS target, as P-gp is highly expressed in the blood-brain barrier (BBB) and circulating concentrations are typically lower than those found in the gut. Thus, P-gp transport can significantly limit BBB penetration.

PPB Category

Concerns over the effect of protein binding on efficacy or pharmacokinetics are not normally expressed unless plasma protein binding is very high (>95%). However, this model is useful to indicate potential issues when *in vitro* testing includes both isolated enzyme/receptor and cell-based assays.

BBB Category

A classification of '+' indicates that BBB penetration is usually present unless the compound is a substrate for active efflux transporters such as P-gp. A classification of '-' indicates that there is likely to be little or no BBB penetration.

A note on BBB values

If using both BBB models at the same time when scoring, the importance for each of these must be set appropriately to allow for this duplication. Because the calculation of the overall score is multiplicative, using both models could result in a compound being penalised excessively for an undesirable BBB value. As such, the importance of each property must be reduced so that, when combined, the necessary overall importance is appropriate. The following example shows how two models would be used individually or combined while maintaining the same overall importance.

Example: If the desired logBB value for a CNS target is set at -0.5 and the required score for compounds with importance of 0.8, when using both models together the importance of each should be set to 0.55.

logBB only:

Model	Desired values	Importance
logBB	>-0.3	0.8

BBB Category only:

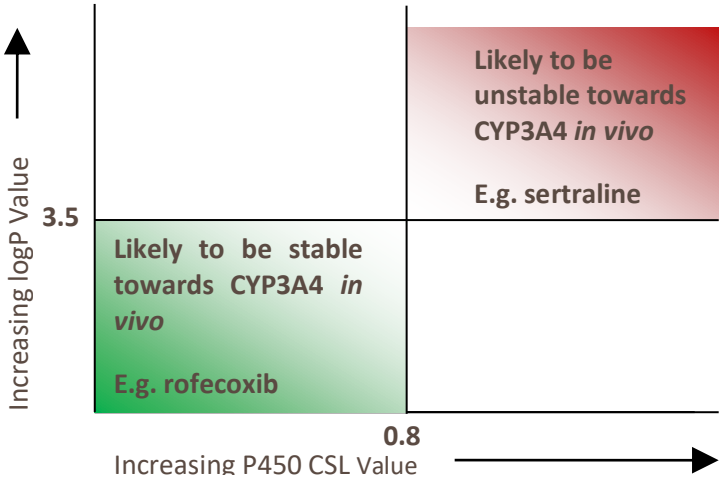
Model	Desired category	Importance
BBB Category	+	0.8

Both BBB models:

Model	Desired values/category	Importance
logBB	> -0.3	0.55
BBB Category	+	0.55

P450 CSL values

CSL values are only generated for the CYP3A4 P450 model within StarDrop and they give an estimate of the efficiency of metabolism for the molecule by CYP3A4 (see section 9). However, metabolism will only occur if the molecule is a substrate for CYP3A4 and this is more likely when the molecule has a high logP value. In combination these two models (CSL and logP) give an indication that the molecule may be unstable due to metabolism by CYP3A4.

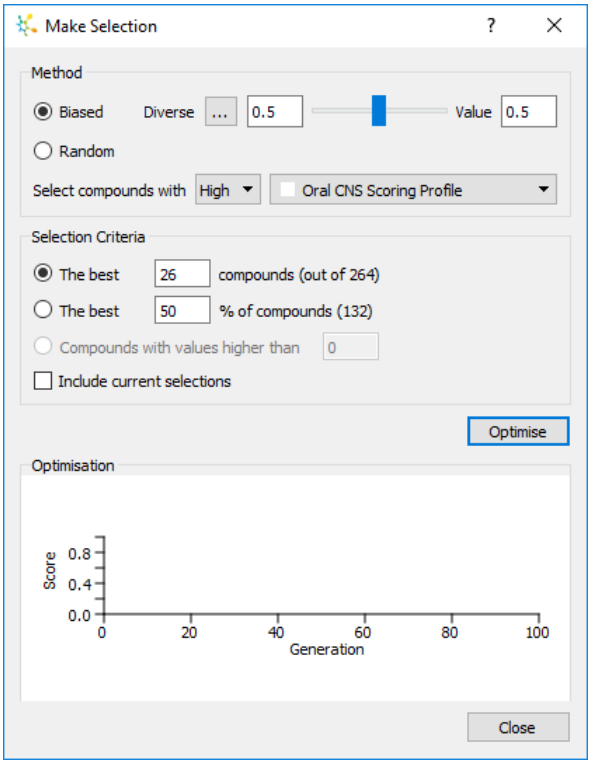


The following table is an example of how you might set the importance using these two models:

Model	Desired values	Importance
logP	< 3.5	0.6
CSL	< 0.8	0.6

15 How do I... Choose which compounds to select?

The **Selection** tool enables you to make selections based on chemical diversity, score or a combination of the two. You can also make random selections. The selection tool is available from the toolbar by clicking the  button.

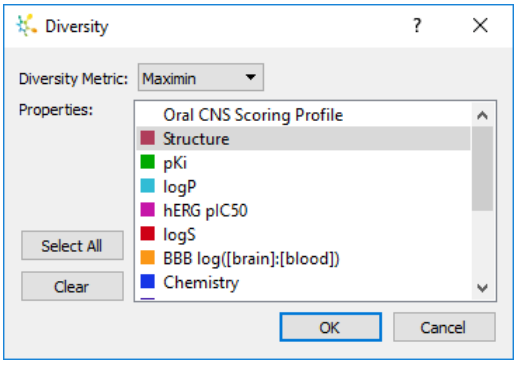


15.1 Source data

You can use this tool to select molecules that have the best possible balance between structural diversity (**Diversity**) and another property (**Value)**. If a data set contains both chemical structures and data, then **Biased** selections can be made. If a data set has either properties or structures, but not both, then the selection can be based upon the information available. For any data set **Random** selections can be made.

15.1.1 Diversity

When making a biased selection which involves diversity you can choose the properties you would like to be used to determine diversity, along with the diversity metric, by clicking the  button.



Select the properties you would like to use when determining diversity and choose a metric from the list. For descriptions of the diversity metrics please refer to the StarDrop Reference guide.

When choosing a property from the drop-down list of available options, you can also specify whether you would rather bias towards high or low values for this property.

Note: The drop down lists will only be populated if there is appropriate data in the data set.

15.1.2 Balanced score versus diversity selections

If there are both data and structure columns in your data set, you can make a selection choosing how much to bias the selection between a property and chemical diversity.

Slide the bar to control the balance you would like to achieve between a given property and the chemical diversity. The numbers next to **Diverse** and **Value** display the ratio desired. For example, if the selection is to be based 20% on structure and 80% on the property the **Diverse** value will read 0.20 and the **Value** value will read 0.80.

You can specify the size of the set to select either as a fixed number, as a percentage, or ensuring that all the compounds which meet some criteria get selected.

Click on the **Optimise** button to start the process. The bottom of the **Optimisation** area will show the **Generation** and the **Fitness** of the algorithm. The closer the value is to 1.00 the more reliable the result. **Note:** It is usually not possible for the **Fitness** value to equal 1.00.

Click the Stop button when the **Fitness** has levelled out. This will select the chosen rows in the data set.

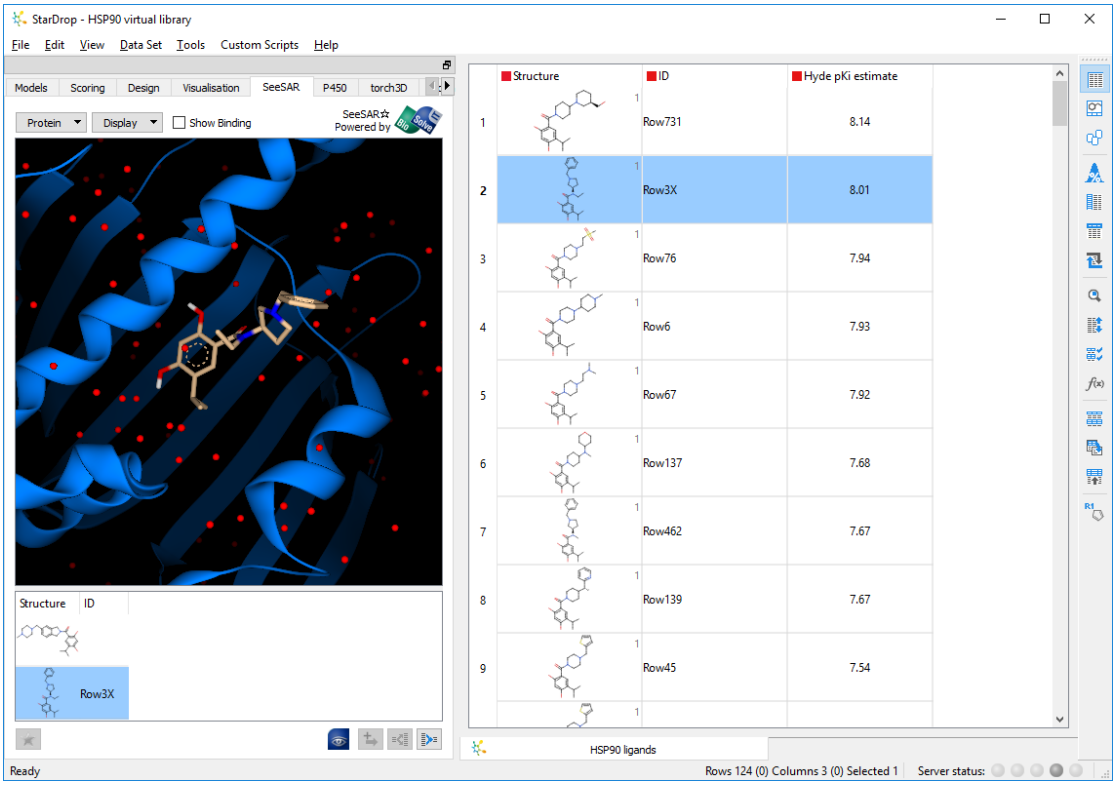
If you choose to **Include current selections** then all the rows which are selected when you start the process will be included in the final selection, with the algorithm choosing the best compounds to complement these in order to best meet the selection criteria.

15.2 Copying selections

At any time within StarDrop you can copy selected rows into a new data set. To do this, click the  button on the toolbar and provide a name for the new data set.

16 How do I... Use SeeSAR™?

In the SeeSAR tab you can view your 3D structure-based design docking results to help understand a compound's binding to its therapeutic target. The viewer enables you to load a protein file and then display this along with any compounds for which you have 3D coordinates. While it is not a requirement, it is recommended that you view proteins and ligands which have been created in the same coordinate space.



16.1 Loading proteins

To load a protein into the viewer, click on the **Protein** menu and choose **Open...** or **Download...**. The **Open...** menu option enables you to load a PDB file. The **Download...** menu option enables you to download a protein from the RCSB Protein Data Bank, simply by entering the 4-character PDB ID. If you have previously loaded a protein, then the **Recent** menu will show you up to the last 10 proteins you have viewed.

Once you have loaded a protein it will be displayed. Any ligands that are loaded from the PDB file will also be displayed and will be listed in the table below the viewer. To remove these from the view, simply deselect them in the table below.

To add a ligand that was imported with the protein to your main data set, select it and click the  button

To remove a protein from the view, choose **Remove** from the **Protein** menu.

To save a protein file, select **Save As...** from the Protein menu.

16.2 Managing conformers

When you load an SD file or Mol2 file containing ligands with 3D coordinates, these will be displayed in the StarDrop data set. Molecules that are represented by more than one conformer will, by default, only appear as a single row in your data set, although this is an option you can change during the

importing process (see section 3.1.3). When you select one or more compounds in the data set, assuming that they have 3D coordinates, they will be added to the table below SeeSAR and will be displayed in 3D. When you select a compound in your data set which has multiple conformers, the individual conformers will be displayed below the viewer, enabling you to view them individually.

When there are multiple conformers of a compound, the table below the viewer will show columns containing the properties where there is at least one different value for a conformer. One of the conformers will have a blue star next to it  which indicates that this is the primary conformer. When the conformers are represented by a single row in the main data set, the properties of the primary conformer will be shown in the cells for that row.

To change the primary conformer, select the conformer you wish to be the primary conformer in the table below the viewer and click the  button .

If your data set shows all the individual conformers and you wish to collapse these to be represented by a single row, click the  button.

If your data set shows all the conformers in your main data set represented by a single row and you wish to expand these to show all the individual conformers, click the  button.

16.3 3D view controls

The 3D view is controlled by your mouse:

- To zoom into the display, roll the mouse-wheel forward
- To zoom out, roll the mouse wheel backwards
- To rotate the view, hold down the left mouse button and move the mouse - the point around which the view is rotating will be indicated
- To reposition the view, hold down the right mouse button and move the mouse

16.4 Display options

The options controlling the different ways that the protein and ligands are displayed are available under the **Display** menu.

The **Protein** sub-menu enables you choose between displaying the protein as **Stick**, **Wireframe** or **Secondary Structure**. The Protein menu also provides the option to display a **Surface**. When a surface is displayed, you can choose an appropriate level of transparency and whether to colour the surface based on logP or atom type.

The **Ligand** and **Waters** sub-menus enable you to choose between displaying these as **Stick**, **Wireframe**, **Ball and Stick** or **Spacefill**. The Ligand sub-menu enables you to choose whether to display a **Surface**. The Waters sub-menu enables you to choose whether to **Show Waters**.

The **Label & Measure** sub-menu enables you to choose what happens when you click on atoms in the view. If **Label Atoms** is selected, then when you click on an atom a label will be added, identifying the atom. If **Measure Distance** is selected, then when you click on an atom it will be highlighted pink. A connector will then follow the mouse and when you hover the mouse over another atom the distance will be displayed until you move the mouse. If you click on the second atom, then the connector will be drawn between the atoms and the distance will remain displayed as a label.

To remove any labels and distances, select the **Remove Labels** option from the Label & Measure sub-menu.

The **Change Background** menu option enables you to choose a different colour for the background.

The **Show Binding** option provides a quick way to highlight any binding interaction that may exist between the ligand and the protein. When this option is checked, the view will be limited to the region in which both the protein and the ligand exist in close proximity and hydrogen bonds will be displayed between specific points of interaction. Note: If you load a protein and display a ligand which is in different coordinate space there may be no obvious interaction and the **Show Binding** option may result in a blank view.

16.5 Transferring to the full SeeSAR application

If you also have a copy BioSolveIT's SeeSAR software installed, then you can transfer any selected ligands from your data set, along with the protein you have loaded directly into SeeSAR by clicking the SeeSAR button.

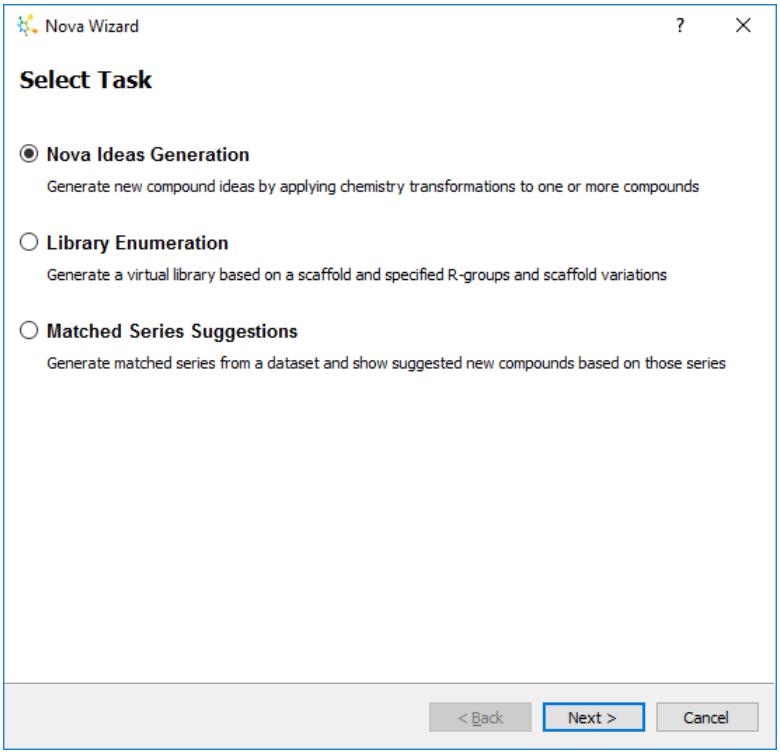
SeeSAR provides equivalent functionality for sending the same data back into StarDrop enabling you to edit ligands in 3D and then continue your analyses in StarDrop.

17 How do I… Use Nova™?

Nova enables you to generate new chemistry ideas either by applying transformations to a compound, finding suggestions based upon matched series analysis or by enumerating a virtual library.

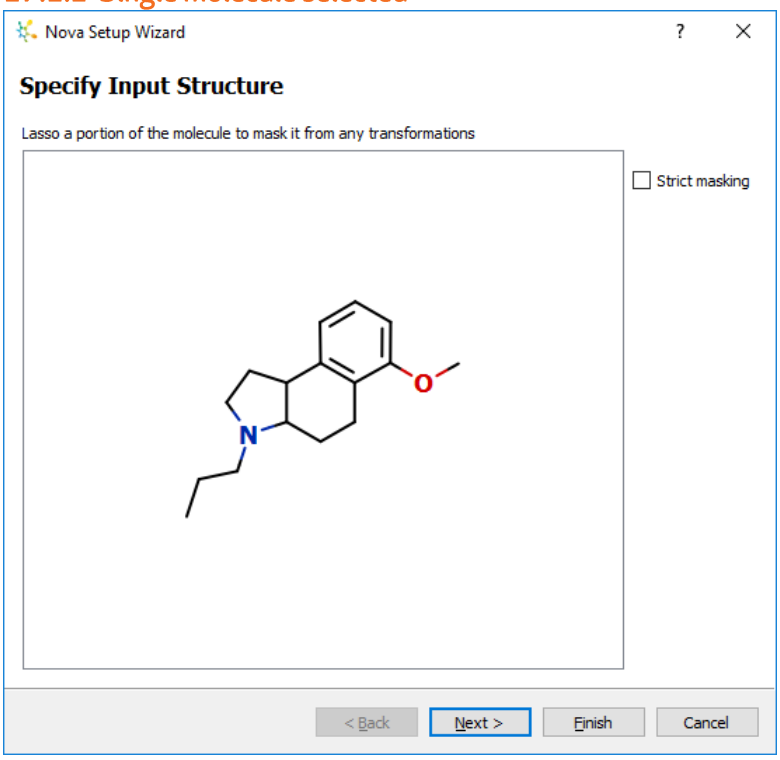
Optionally select a row in the data set which contains the molecule you wish to use as a starting point (for idea generation or library enumeration) and click the button to start the Nova wizard.

When you start the Nova wizard you must choose whether or not you would like to generate new ideas (see section 17.1), carry out a matched series analysis (see section 17.2) or enumerate a virtual library (see section 17.3).

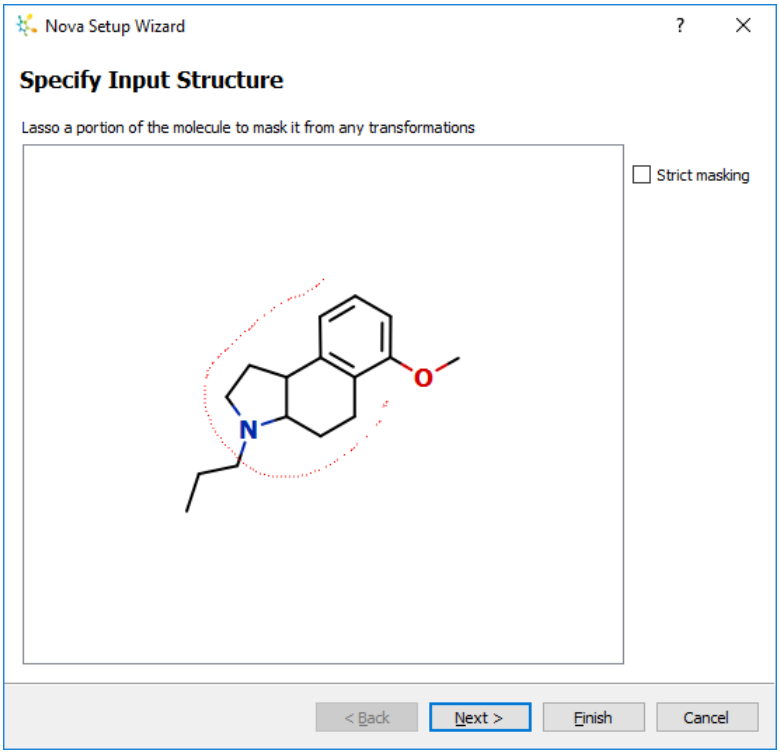


17.1 Nova - Idea generation

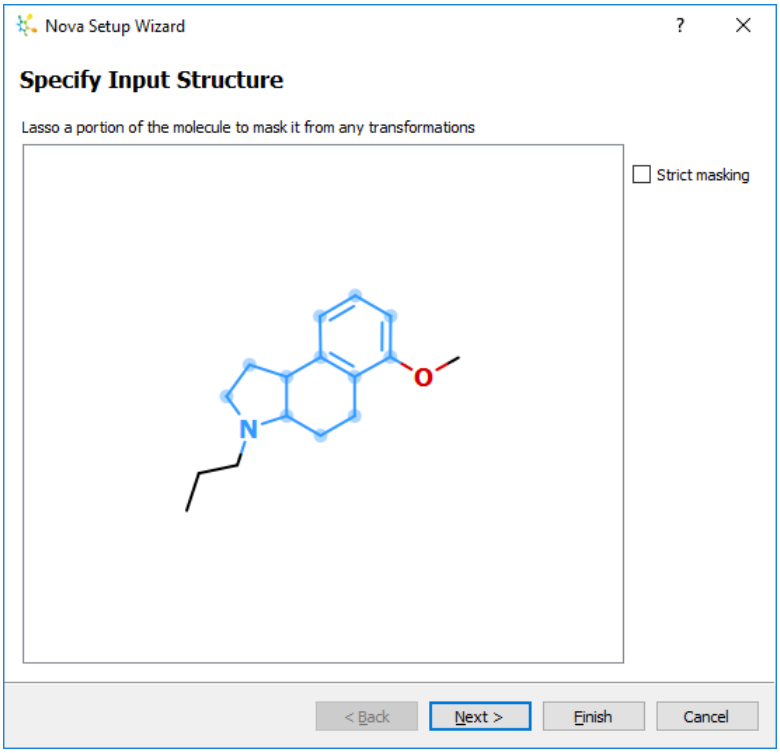
17.1.1 Single molecule selected



The selected structure is displayed. If you wish to ensure that any portion of the structure is 'masked' i.e. not changed by any of the transformations that are applied, then lasso that region of the molecule.



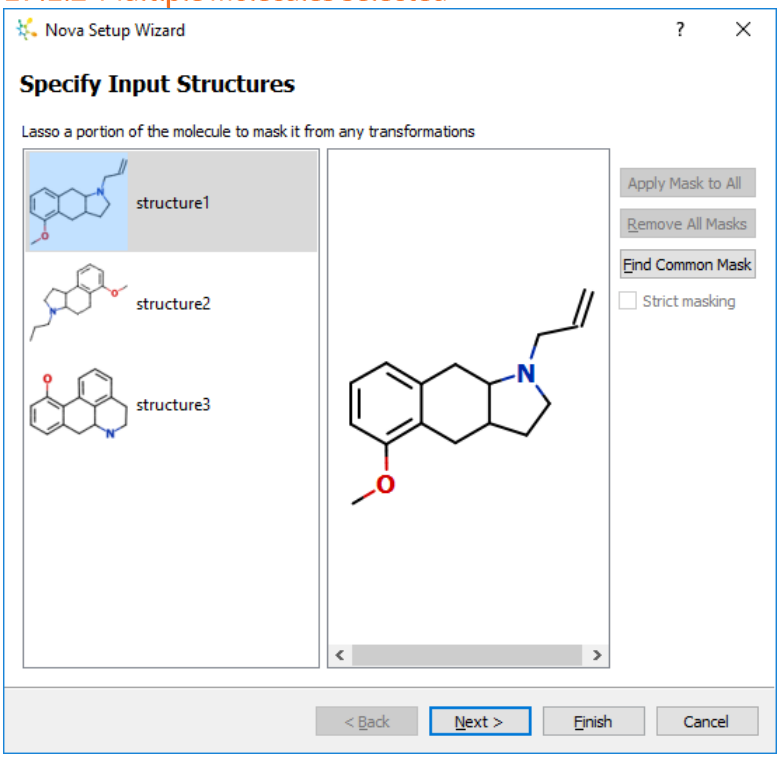
The selected region will be displayed and will remain selected unless you click on the structure to reset the selection or select a different region.



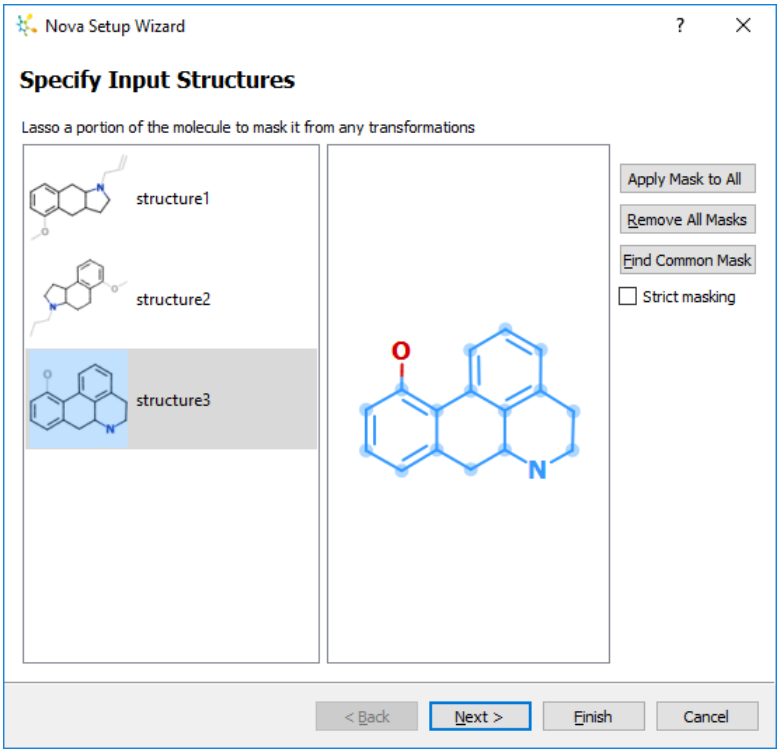
Clicking the strict masking button will not only ensure that the selected region is present in all the generated molecules, but that no atoms or functional groups will be added to this substructure either.

Click the **Next** button to select which transformations to apply.

17.1.2 Multiple molecules selected



The selected structures are displayed. If you wish to ensure that any portion of any of the structures are 'masked' i.e. not changed by any of the transformations that are applied, then lasso that region of each molecule.



The selected region will be displayed and remain selected unless you click on the structure to reset the selection or select a different region.

If you wish the same region to be masked for all the structures then you can click the **Apply Mask to All** button and where possible the mask will be applied to all the molecules.

You can check or set the mask individually for each compound by selecting them from the list on the left. Alternatively you can clear the masks set on all the compounds by clicking the **Remove All Masks** button.

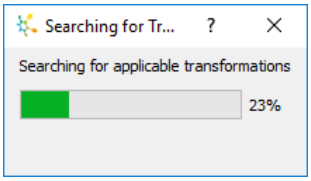
If you click the **Find Common Mask** button StarDrop will calculate and select the largest common mask that can be applied to all the compounds in the list.

Clicking the strict masking button will not only ensure that the selected region is present in all the generated molecules, but that no atoms or functional groups will be added to this substructure either.

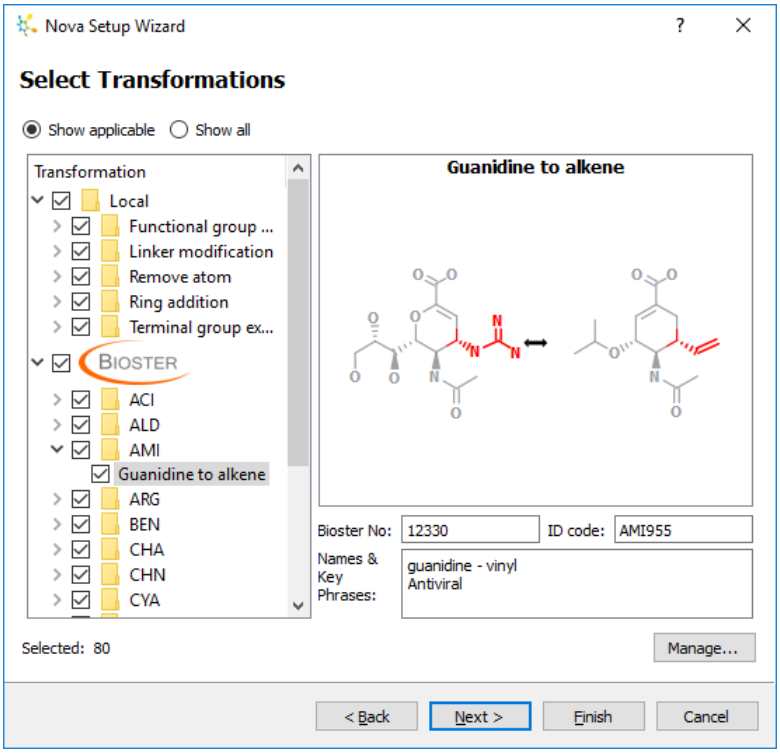
Click the **Next** button to select which transformations to apply.

17.1.3 Select Transformations

Having defined one or more input structures, Nova will determine which of the available transformations are applicable. If you are using the BIOSTER™ database then it may take a few seconds to search and you will see a progress indicator.



The list of applicable transformations will then be displayed from which you can select those you would like to apply.

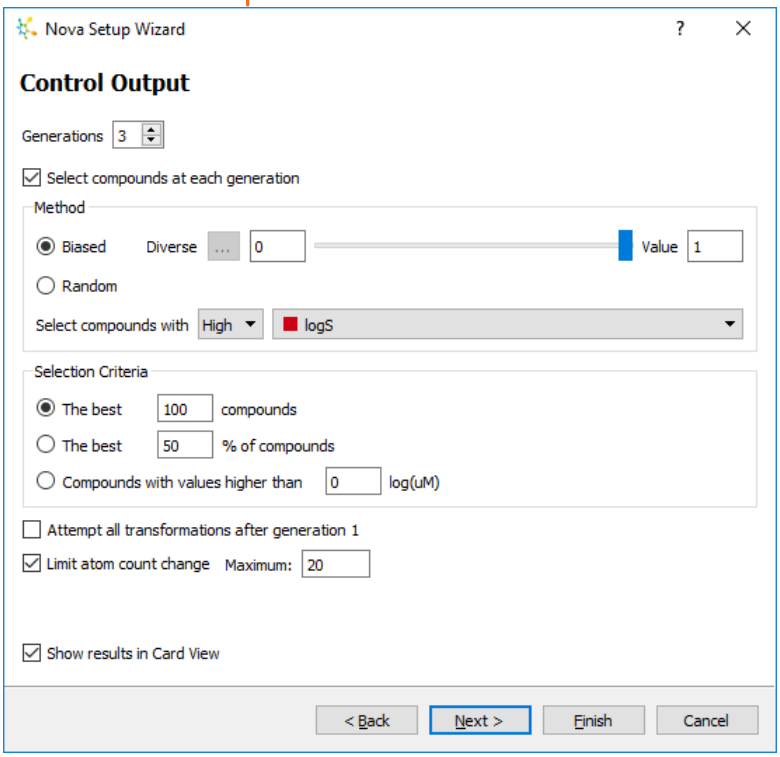


The list will display only those transformations that are applicable to your input structure(s) but choosing **Show all** will display the complete list.

To open the Transformation manager enabling you to see further details of BIOSTER transformations, click the **Manage...** button.

Click the **Next** button to specify options for controlling the number of generations of compounds produced.

17.1.4 Control Output



The values displayed here will be those specified in the Nova preferences (see section 0).

If you select more than one generation then in each subsequent generation after the first, the transformations will be applied to all the molecules generated in the previous generation. This may result in exponential growth of compound numbers and so you can define a process of **Compound Selection** to limit the number of molecules passed on to the next generation.

To do this, you can choose whether to select compounds based upon properties, scores or chemical diversity. When biasing, either partially or fully, towards a property or score you can select this from the drop-down list. These values will be calculated for all generated molecules. You can then indicate whether you are interested in compounds which have properties or scores better than a specified threshold or you can indicate a number or percentage of the set to select based upon looking for high or low values.

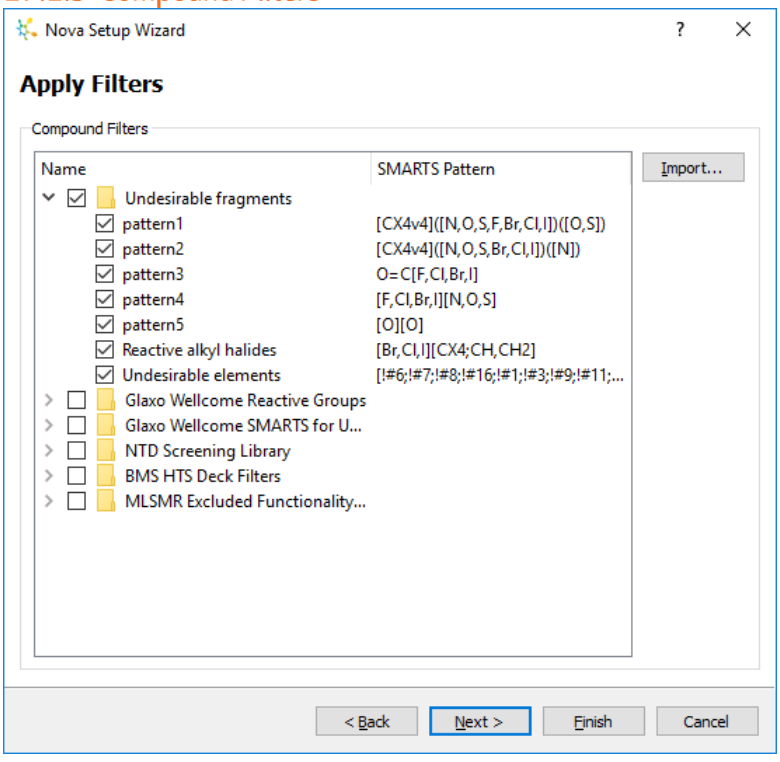
Tick the **Attempt all transformations after the first generation** checkbox if you would like Nova to look for additional transformations that will be applicable to compounds within subsequent generations other than those previously selected.

The **Limit atom count change** can be used to ensure that generated molecules do not differ in size from the parent compound by more than a certain amount.

If you choose to **Show results in Card View** then the structures generated will be displayed as a network within Card View.

Click the **Next** button to specify compound filters.

17.1.5 Compound Filters



Nova may generate compounds that contain undesirable fragments. Such compounds can be filtered using a set of SMARTS pattern filters that are selected by checking the associated checkbox. Filters will be applied as a final step in the generation process *after* all generations are complete.

You can also define your own filters. To do this first create a text file containing smarts patterns that represent your filters and include associated names. The SMARTS patterns must not contain any spaces and there should be a space to separate the pattern from the name:

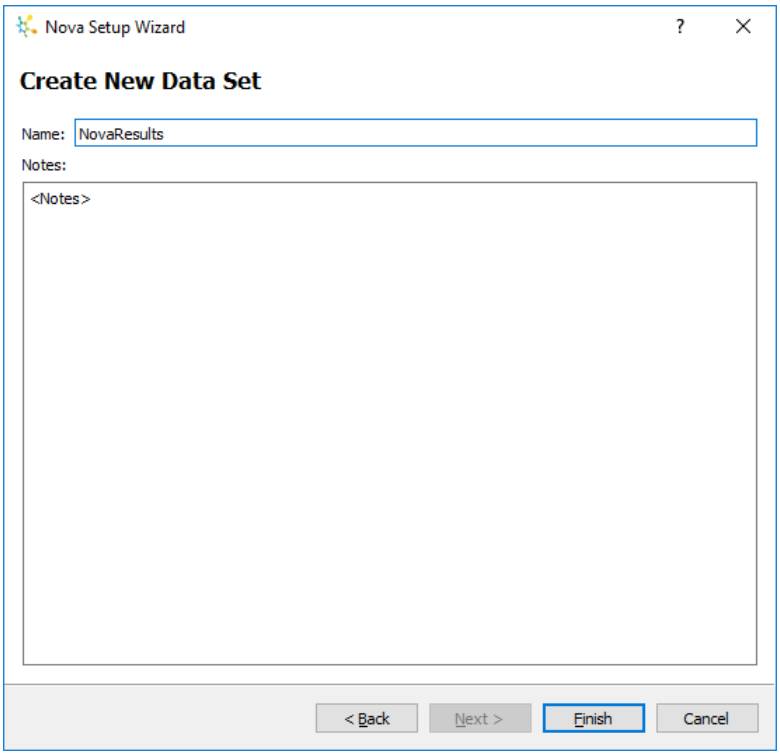
```
[S,C](=[O,S])[F,Br,Cl,I] acid halide
[Cl]C([C&R0])=N chloramidine
[P,S][F,Cl,Br,I] P/S halide
```

Then, import the SMARTS file using the import button.

The filters that are selected (checked) by default will be those specified in the Nova preferences (see section 0).

Click the **Next** button to specify any details about the session that you would like to save.

17.1.6 Create New Data Set



Once you have given the session a name and added any notes, click the **Finish** button to start the process. A box listing all the sessions currently running will be displayed at the bottom of the tab. This will indicate what stage the process is at and how many molecules have been generated.

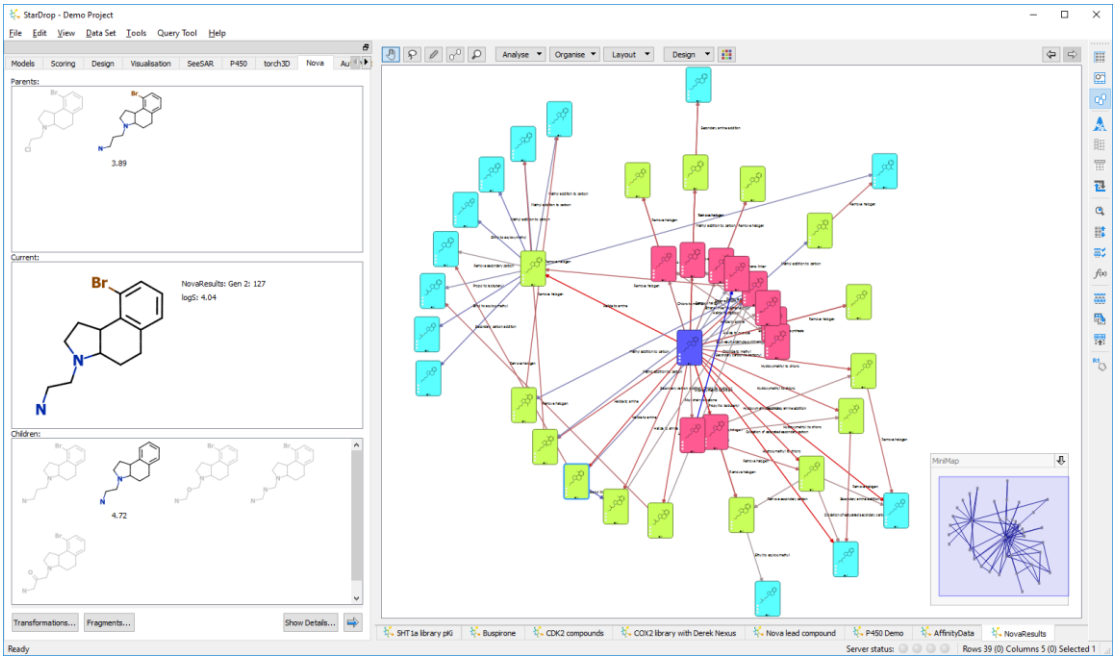
Note: The number of molecules indicated will only increase when novel compounds are generated (using multiple generations it is often possible to generate the same compounds using different combinations of transformations) and so the total number will increase more slowly as the process continues.

17.1.7 Nova Results

When the process has completed a new data set will be displayed containing all the molecules.

The data set will contain a category column indicating in which generation the molecule was generated. Another column will display the last transformation that was applied to generate the compound.

Selecting a row in the data set will display information about the molecule, its parents and any children generated from it.



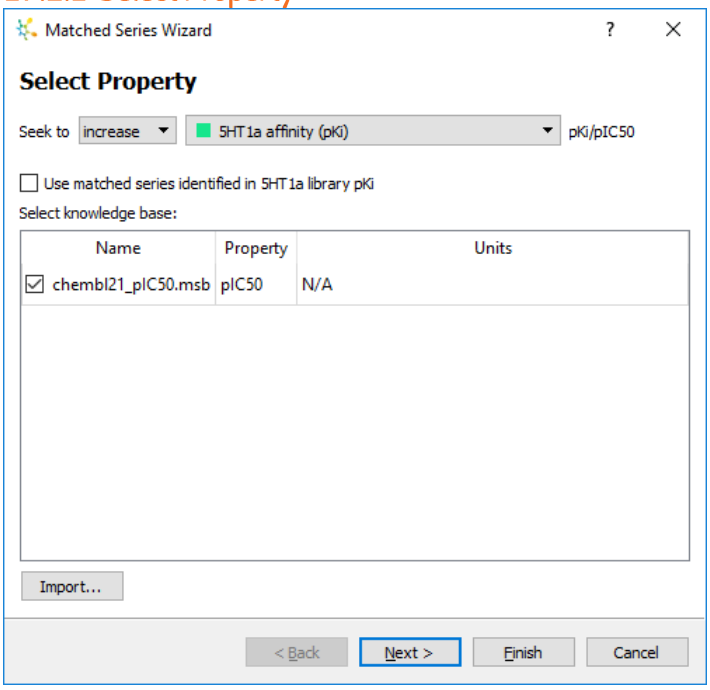
Selecting molecules in the Nova tab will select them in the data set. Double-clicking molecules in the Nova tab will promote them to be the current molecule of interest with their children and parents displayed.

If you display the results in Card View then each ring in the network will indicate a generation. The arrows between cards indicate transformations that have taken place with the colours indicating whether the transformation resulted in an increase or a decrease in the desired property.

17.2 Nova - Matched Series Analysis

To carry out a matched series analysis you must start with a data set containing at least one series of compounds for which you have measured potency data.

17.2.1 Select Property



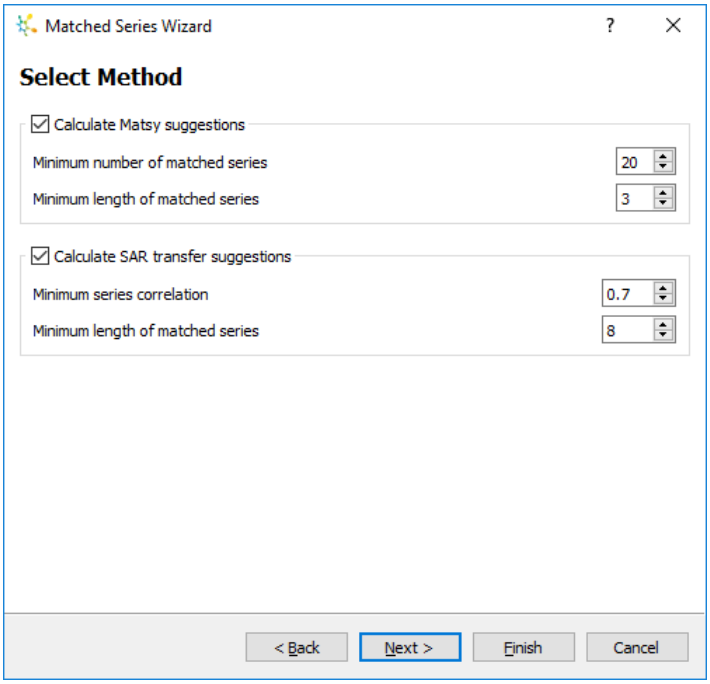
A list of all the properties in your data set will be displayed and you must choose the one you would like to improve. Depending upon the units of the chosen property you should indicate whether you would like to generate suggestions for compounds which increase or decrease this.

The default knowledge base generated from ChEMBL contains potency data and can be used to generate suggestions for compounds which improve potency. To generate suggestions which improve other properties you can use matched series from within your data set or must generate a new knowledge base (this is something which is available as an additional service – for more information please contact stardrop-support@optibrium.com).

If you have access to additional knowledge bases you can access these by clicking **Import…**

Click **Next** to choose which matched series methods to apply.

17.2.2 Select Method



Two different approaches can be used to generate suggestions, Matsy and SAR transfer. For more details on these methods see the StarDrop Reference Guide.

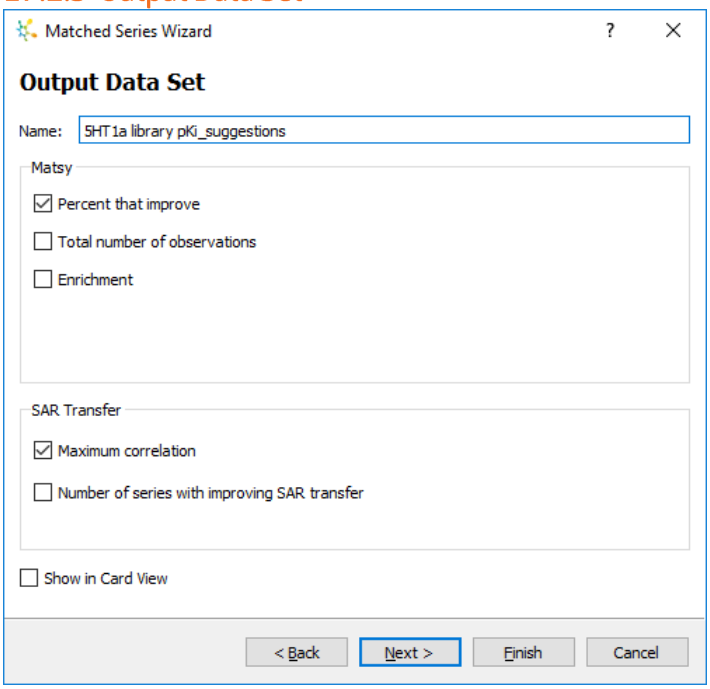
For each method you can configure the **minimum length of matched series** to consider – the default values give an indication of suggested values.

For suggestions based on Matsy you can also specify the **minimum number of matched series** – suggestions will only be given where the series occurs this number of times or more.

For suggestions based on SAR transfer you can also specify a **minimum series correlation**.

Click **Next** to choose the columns that you would like to see in the result data set.

17.2.3 Output Data Set



In the resulting data set you can choose which supporting information is added.

For Matsy:

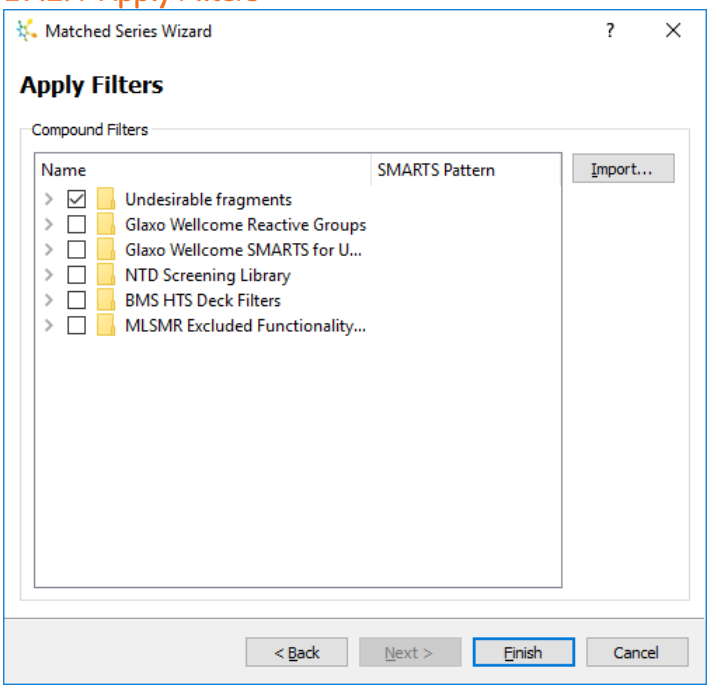
- **Percent that improve** indicates the proportion of times that the suggestion improves the property value in the supporting data
- **Total number of observations** provides a count of the number of times that the series was observed
- **Enrichment** indicates the ratio of the observed frequency of a particular series to that expected by chance, assuming all N! orders of the series are equally likely

For SAR transfer:

- **Maximum correlation** shows the value highest correlation between the series in your data set and the matching series found in the supporting data
- **Number of series with improving SAR transfer** indicates the number of series found that support a given suggestion

Click **Next** to specify any filters you would like to use.

17.2.4 Apply Filters



Matched series analysis may, on occasion, generate compounds that contain undesirable fragments. Such compounds can be filtered using a set of SMARTS pattern filters that are selected by checking the associated checkbox. Filters will be applied as a final step in the generation process *after* all suggestions have been found.

You can also define your own filters. To do this first create a text file containing smarts patterns that represent your filters and include associated names. The SMARTS patterns must not contain any spaces and there should be a space to separate the pattern from the name:

```
[S,C](=[O,S])[F,Br,Cl,I] acid halide
[Cl]C([C&R0])=N chloramidine
[P,S][F,Cl,Br,I] P/S halide
```

Then, import the SMARTS file using the import button.

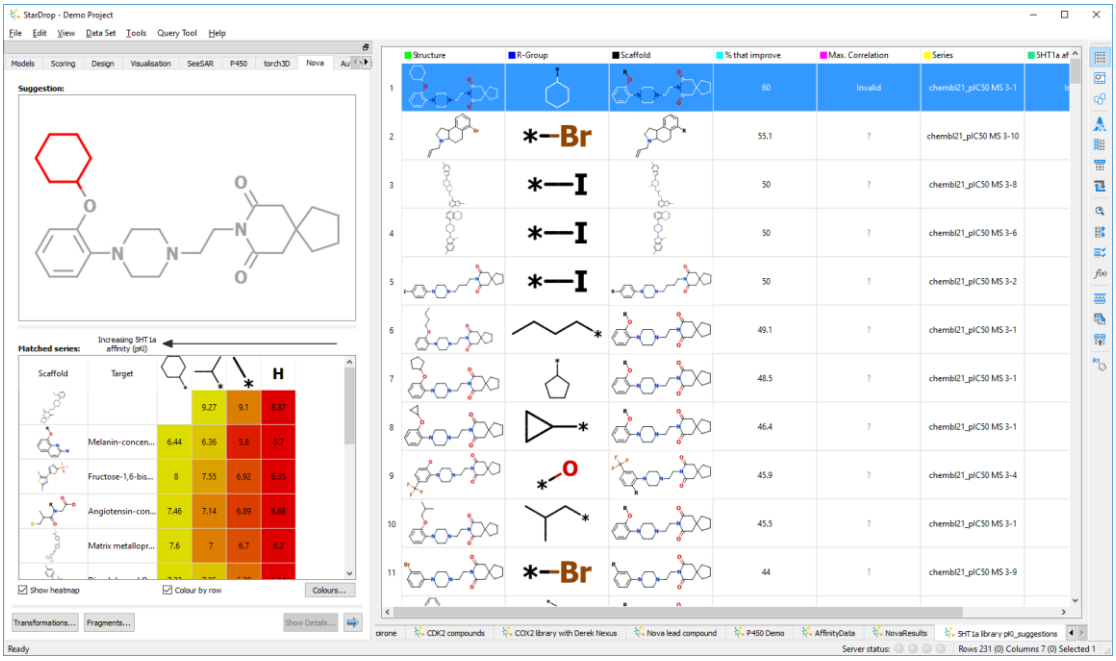
The filters that are selected (checked) by default will be those specified in the Nova preferences (see section 0).

Click **Finish** to run the matched series analysis.

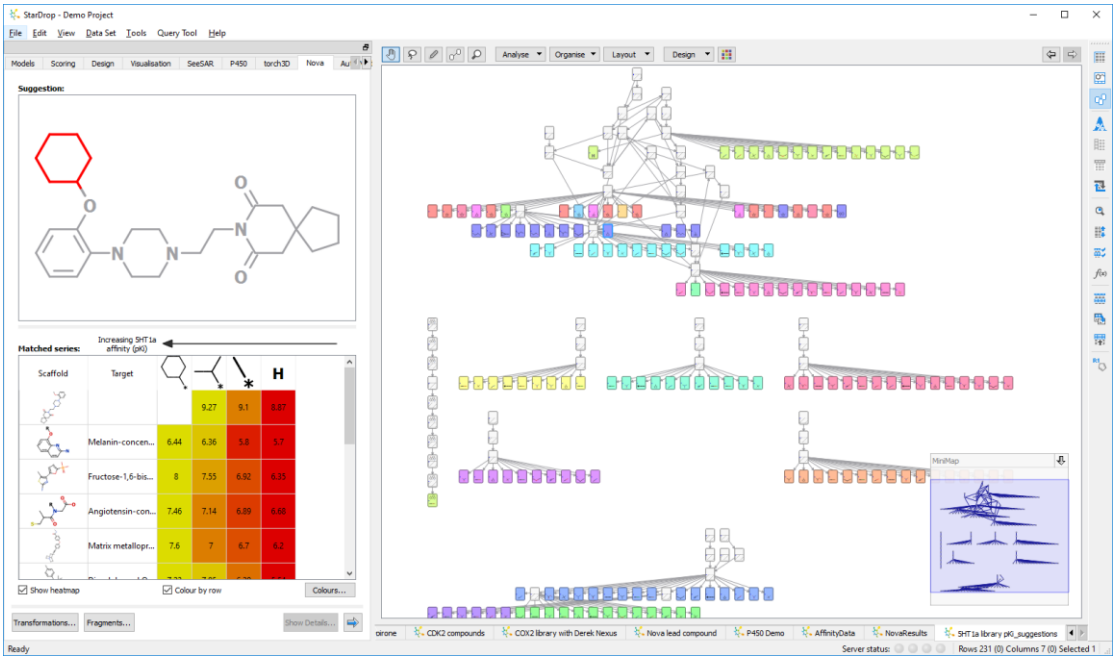
17.2.5 Matched Series Results

While the analysis is running a status will be displayed in the Nova tab.

Once complete a new data set will be displayed which shows the suggested compounds. Each row represents one suggestion, shown in the **Structure** column. The **Scaffold** column shows the common scaffold which was found in your original data set and the **R-group** column shows the suggested substituent. The **Series** column shows a name for each series (e.g. “chembl21_pIC50 MS 3-30”) where the first part of the name indicates the database from which the suggestion was derived and the second part indicates the approach used to generate the series (MS – Matsy, SAR – SAR transfer). The numbers (e.g. 3-30) indicate the length of the series, “3”, and a unique id, “30”, to distinguish all series of this length derived from the same database using the same method. The columns chosen in the wizard are also shown indicating the level of support for each of the suggestions.



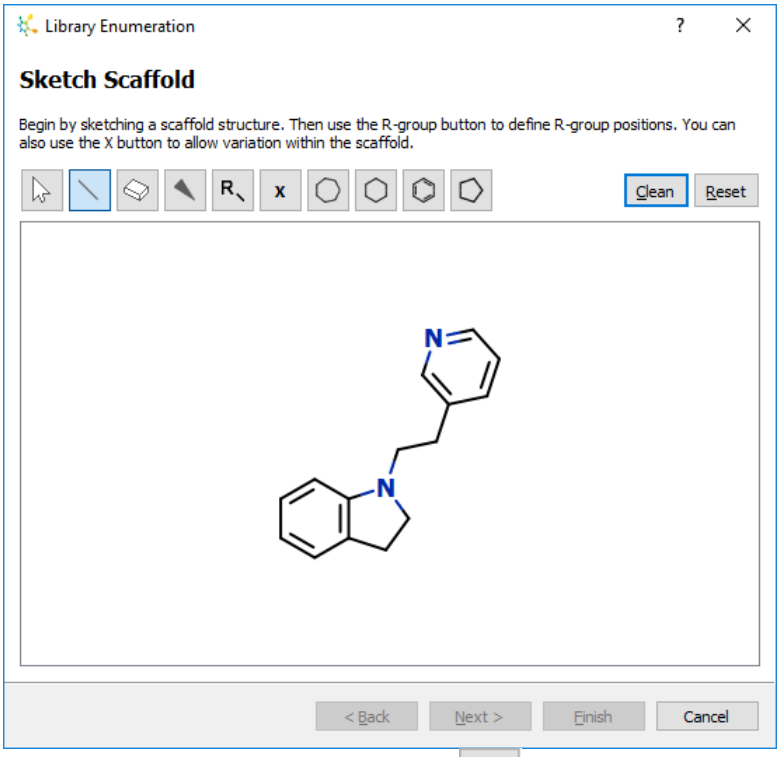
In the Nova tab, the suggestion is displayed along with the supporting evidence. The table below the structure shows the examples found where the same matched series occurs. If you are using the ChEMBL knowledge base and have an internet connection then clicking on any of targets will open a web page showing the data from ChEMBL. Equally, clicking in any of the cells (in any of the rows other than the top one which shows your own data) will show you that compound



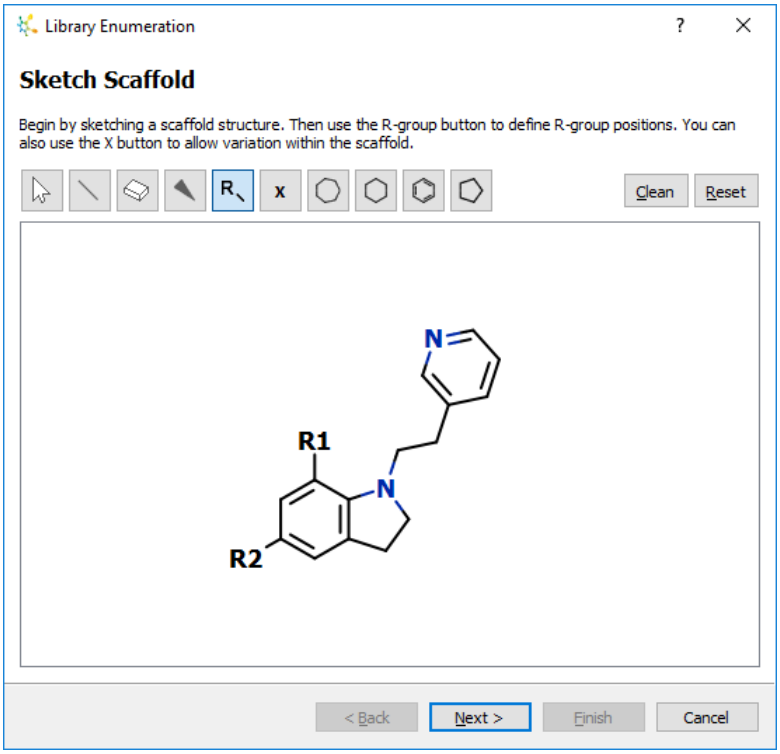
17.3 Nova - Library Enumeration

17.3.1 Sketch Scaffold

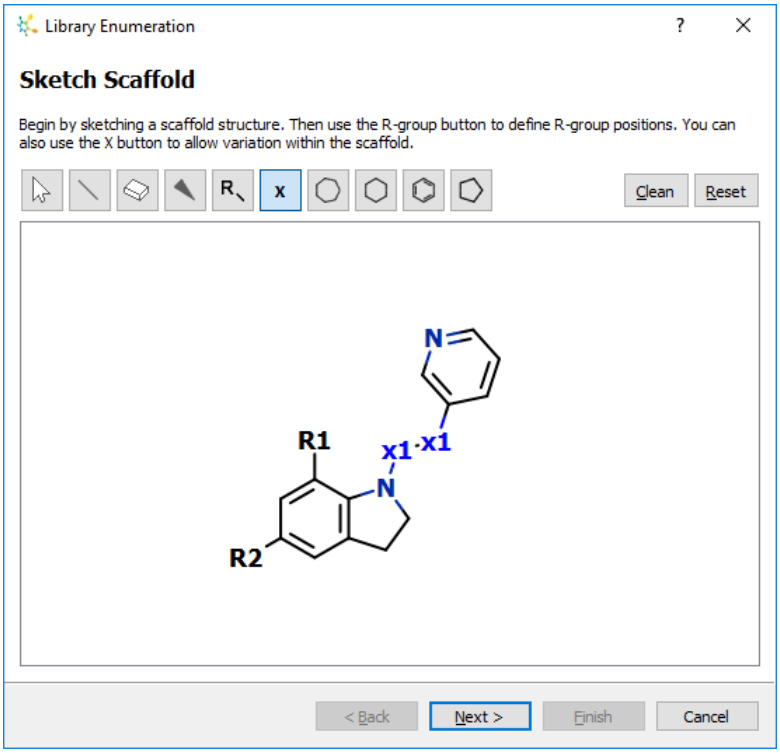
First, draw the structure you wish to use as a scaffold. If you have a data set open with a compound selected then this molecule will be displayed, which you can clear (to start again) or edit.



To specify R-Group positions click on the  button and then click to draw R-Group indicators at the required positions on the scaffold.



To specify variable atoms or linkers click on the  button and then select the atoms or linkers on the scaffold.

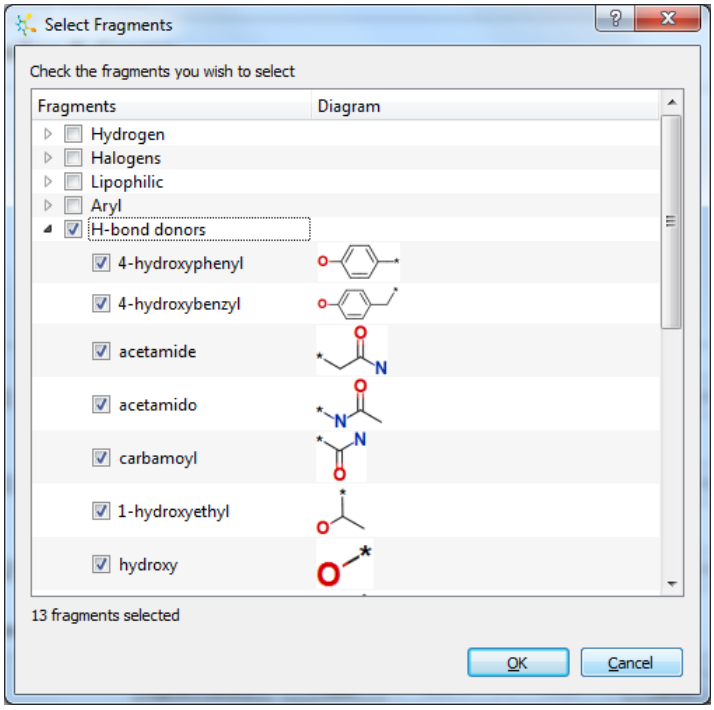


Click the **Next** button.

17.3.2 Define R-Groups

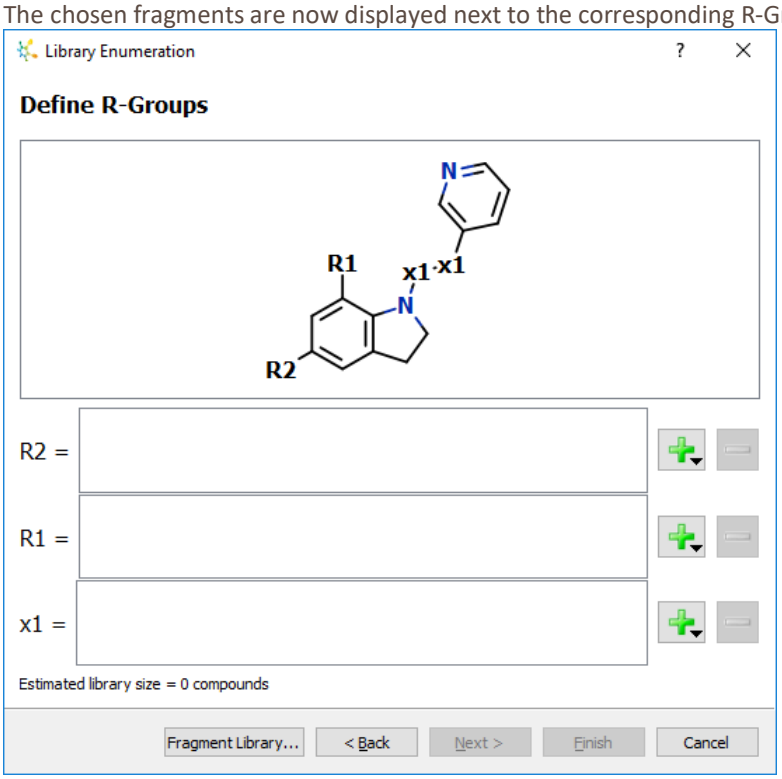
On this page (shown above) we can define the R-Groups and linkers with which to decorate our scaffold.

Clicking on the  button next to each R-Group enables you to choose between sketching a fragment, selecting one or more fragments from a list or loading a file of fragments. You can also pick from a list of common fragments. For information on sketching or loading fragments take a look at section 24.9.5. If you choose to **Select...** then you will see the **Select Fragments** dialogue.

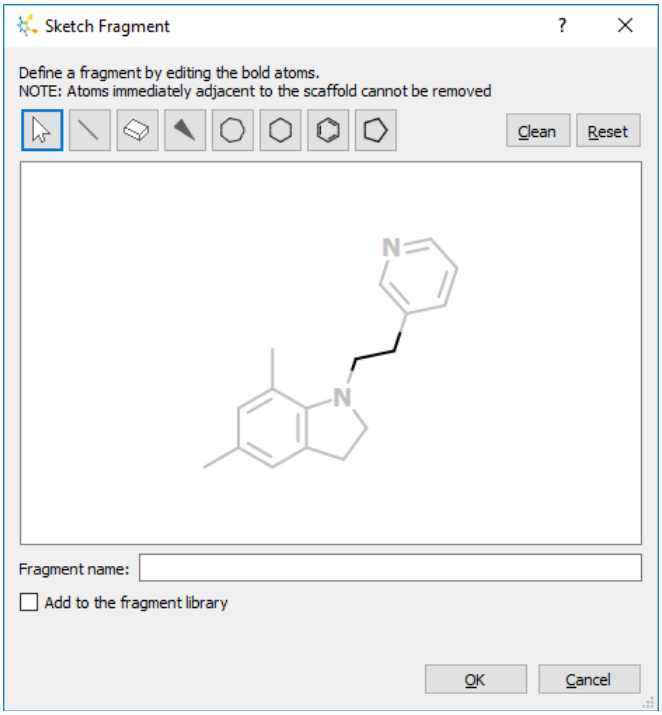


Tick the groups, or individual fragments, that would like to include at the given position and click **OK**.

The chosen fragments are now displayed next to the corresponding R-Group.



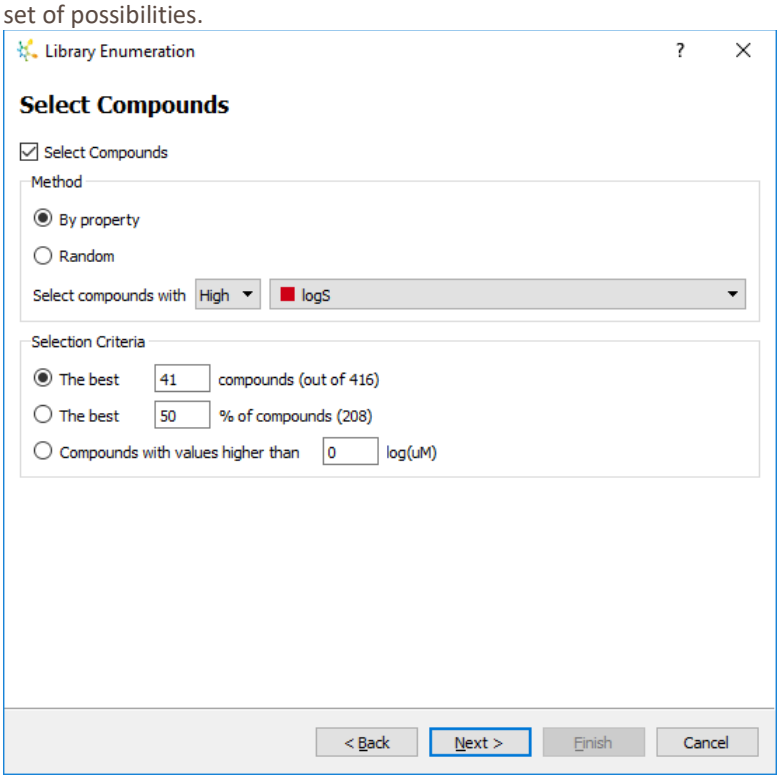
If you choose to **Sketch...** a linker then the **Sketch Fragment** dialogue will display a template molecule into which you can draw the linker.



Note: to draw a longer linker chain, use the erase tool  to remove the bond and the  tool to draw in a longer chain connecting the attachment atoms. Once you have added R-Groups for each position click the **Next** button.

17.3.3 Select Compounds

When enumerating a large virtual library, you can choose whether to create just a subset of the full set of possibilities.



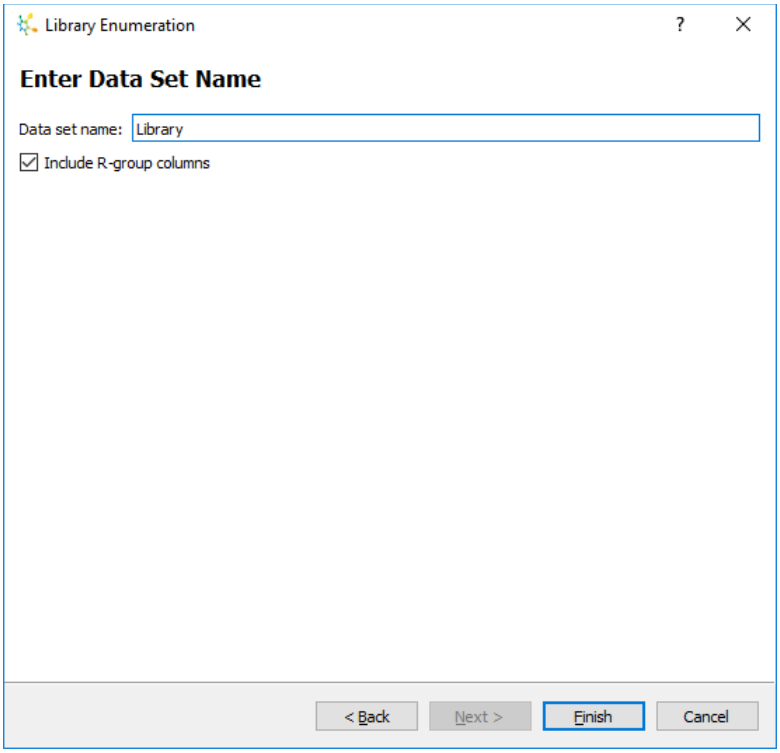
To select compounds tick the **Select Compounds** checkbox. Choose a **Method** and, if **choosing By property**, specify the desired property to optimise and whether or not high or low values are desirable.

For the chosen property you can choose to select compounds either by selecting:

- a given number of the best compounds
- a given percentage of the data set
- all compounds which exceed some criteria

Once you have specified the criteria click **Next**.

17.3.4 Enter Data Set Name



Finally, enter a **Data set name** for the new library and indicate whether you would like to **Include R-Group columns** in the data set which show the fragments that have been used.

Click **Finish** to complete the process.

While the library is being enumerated an indicator will be displayed in the Nova tab. Once the library is ready the new data set will be displayed.

18 How do I... Use the P450 models?

To run the P450 regioselectivity and site lability models, select the **P450** tab. There are models for seven P450 isoforms: CYP3A4, CYP2D6, CYP2C9, CYP1A2, CYP2C19, CYP2C8 and CYP2E1. Each model will generate a list of the sites across the molecule that are susceptible to metabolism and display these as annotations.

Note: Regioselectivity indicates the distribution of metabolites **if** the molecule is metabolised by the given isoform.

It is assumed that in running the model a user believes that the molecule may be a substrate for one of the P450 isoforms. In addition, the CYP3A4 Model also produces a **metabolic landscape**. The vertical bars indicate the degree of lability of each site on the molecule. The category assigned to each site is indicated by the colour of the bar from red (**labile**) to blue (**stable**). The top left of the landscape display shows the predicted **Composite Site Lability** (CSL) value. This is an indication of the efficiency of metabolism. A value close to 1 indicates that the metabolism is likely to be extremely efficient. It is important to note that CSL is **not** a prediction of the **rate** of metabolism.

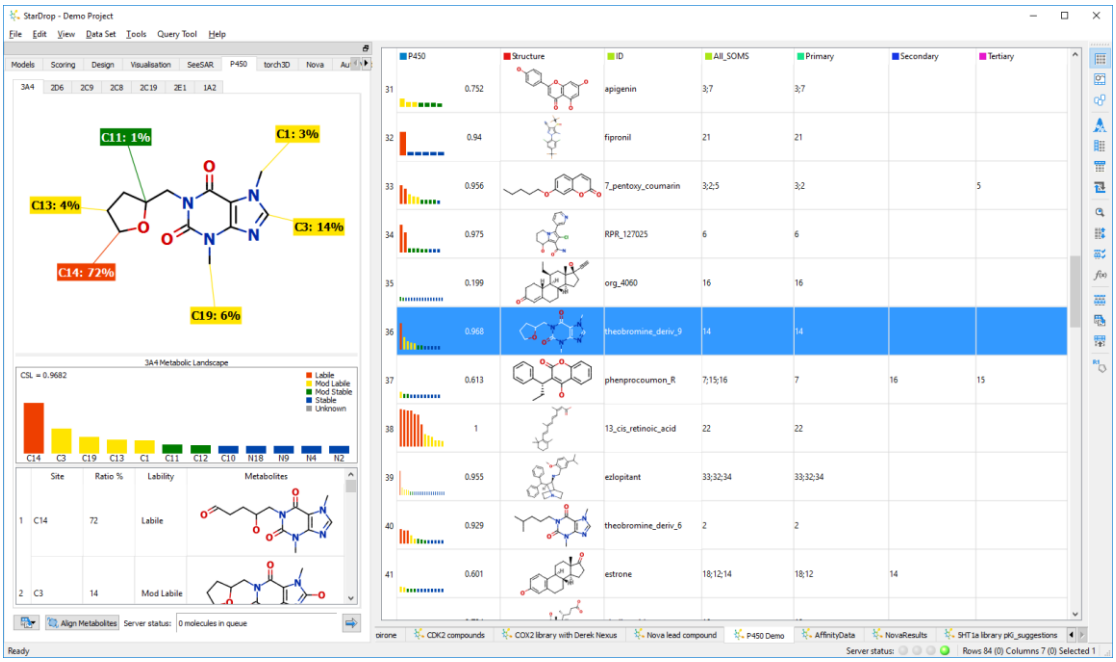
18.1 Running P450 models

The P450 models will only be run on rows that are selected. To select all molecules click the square in the top left hand corner of the data set. Alternatively, choose **Select All** from the **Edit** menu. A specific set of molecules can be selected by holding the **Ctrl** key while clicking on selected rows within a data set. Blocks of molecules can be selected by holding the **Shift** key while clicking on rows within a data set.

Click the  button to run the models.

Note: If more than 10 compounds are selected, a prompt will be displayed to warn you that the models can be slow.

Once the models are running an additional column will appear in the data set called **P450**. While the molecules are being processed this column shows their status.



If required, the molecules that have a **Queue** position displayed in the P450 column can be removed from the queue. It is not possible to remove molecules that are already **Running**. To remove a molecule, highlight the required molecule(s) in the data set and click the  button.

18.2 Recalculating P450 model predictions

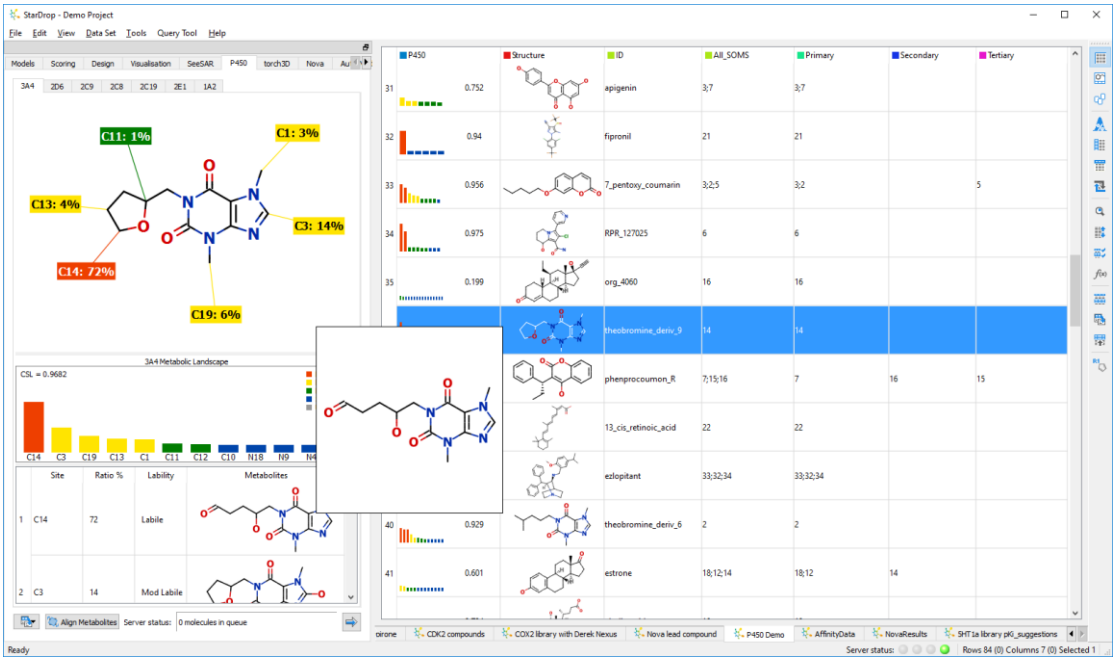
P450 models can sometimes take a considerable amount of time to run. Saved data sets can be closed during this process and the server will continue to process the molecules and the results will appear in the data set the next time it is opened in StarDrop. Example results are shown below.

Note: For this reason, the calculated results are cached on the P450 server and are recalled if the same molecule is resubmitted.

To force a molecule to be recalculated, open the P450 preferences and tick the **Force re-run of submitted molecules** checkbox.

18.3 Metabolites

To view the metabolite resulting from metabolism at a given site, hover the mouse over the annotation to see a pop-up.



At the bottom of the tab a table displays the list of all the sites and for each one indicates its regioselectivity and the metabolite that will be produced if metabolism occurs at that position.

To align the displayed metabolites with the parent molecule click the **Align Metabolites** button.

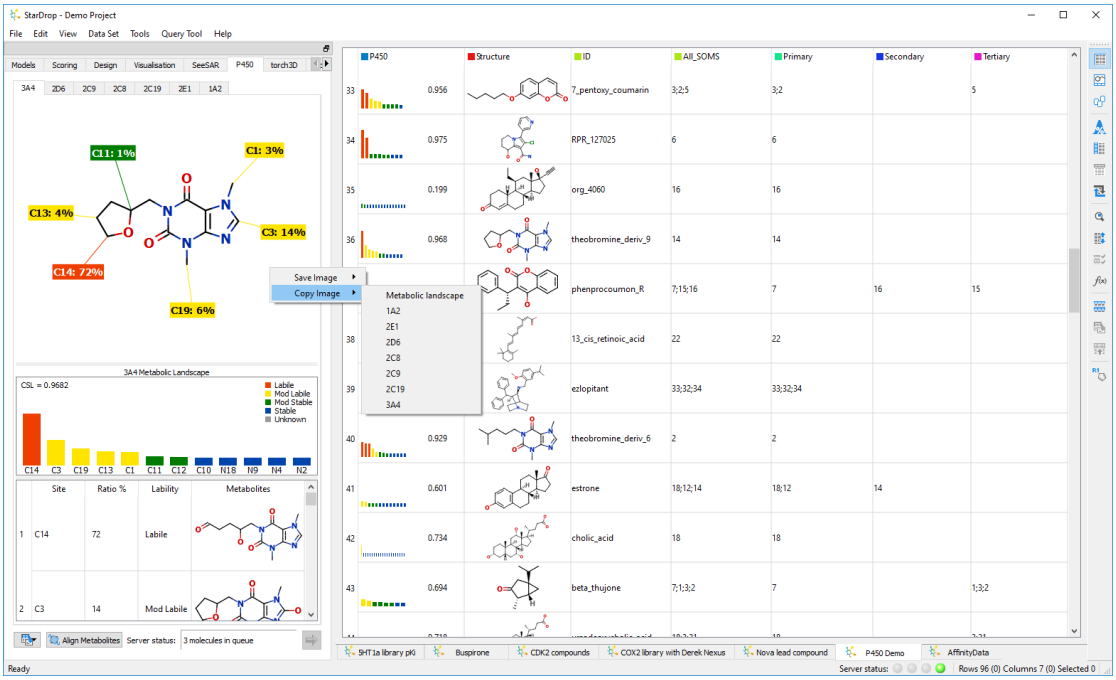
To export the metabolites into their own data set, click the  button and choose whether to export metabolites for the current molecule of for all selected molecules.

The resulting data set will show each metabolite on a separate row along with the site, parent molecule, exact mass, regioselectivity with respect to each of the seven P450 isoforms and information about the steric and orientation factors considered within the prediction.

18.4 Saving and copying the P450 images

The diagrams showing the P450 regioselectivity for each isoform can be saved as images.

Click the right mouse button on the diagram to display the menu.



Select the image to save and this will display the **Save As** dialogue. Choose an appropriate name and click the **Save** button.

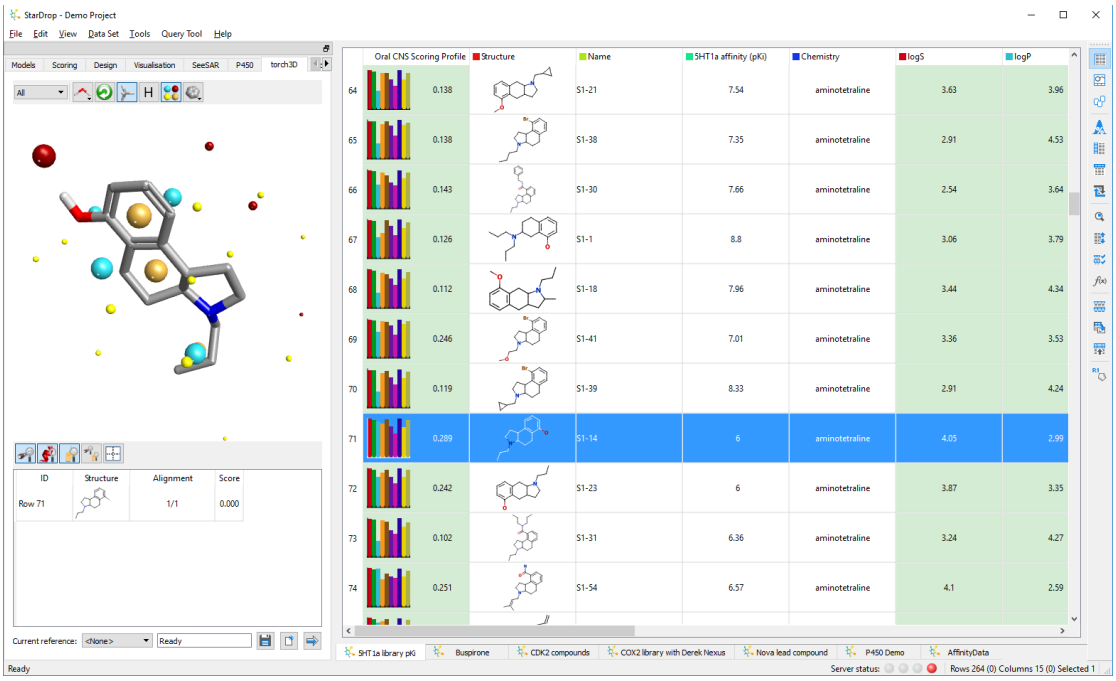
Note: Normally only those sites that have more than 1% of total metabolites resulting from metabolism are annotated. To show all potential sites of metabolism, open the P450 preferences and tick the **Display annotations for all sites** checkbox.

You can also copy the image directly to the clipboard from the **Copy Image** menu.

19 How do I... Use torch3D™?

torch3D enables you to compare the fields of active compounds to identify key interactions with their protein target and further optimise potency. New compounds can be scored according to their field similarity with a known active to identify novel series with greater structural diversity or design focused libraries for synthesis and screening.

To take a look at the field patterns around a molecule in 3D, select the molecule of interest and click the  button.



To export an SD file containing the 3D coordinates, click the  button and choose an appropriate name and destination for the file.

19.1 torch3D wizard

To generate a torch3D reference against which to compare your compounds:

Click the  button to start the torch3D wizard

On the first page of the wizard, type a reference name. This will be the name given to the column of results that appears for this reference.

Click **Next** and then load an SD file of the bioactive compound against which you would like to compare your compounds. Ideally, this should be an SD file which contains the 3D coordinates of the known active conformation. You also have the option to extract a reference molecule from a PDB file (see section 19.3).

Click **Next** to (optionally) load an excluded volume – this could be the protein target.

Click **Next** to specify the speed/quality of the calculations. As a guide, often using the **Fast** method will be a good first step – if the results look interesting then the best compounds could be run against the slower higher-quality method for confirmation.

Click **Finish** to complete the process.

A new column will be added into the StarDrop data set. In addition, the reference will be added to the drop-down list at the bottom of the torch3D tab. Selecting the reference column, or selecting an item from this list, will change the reference that is on display within the tab.

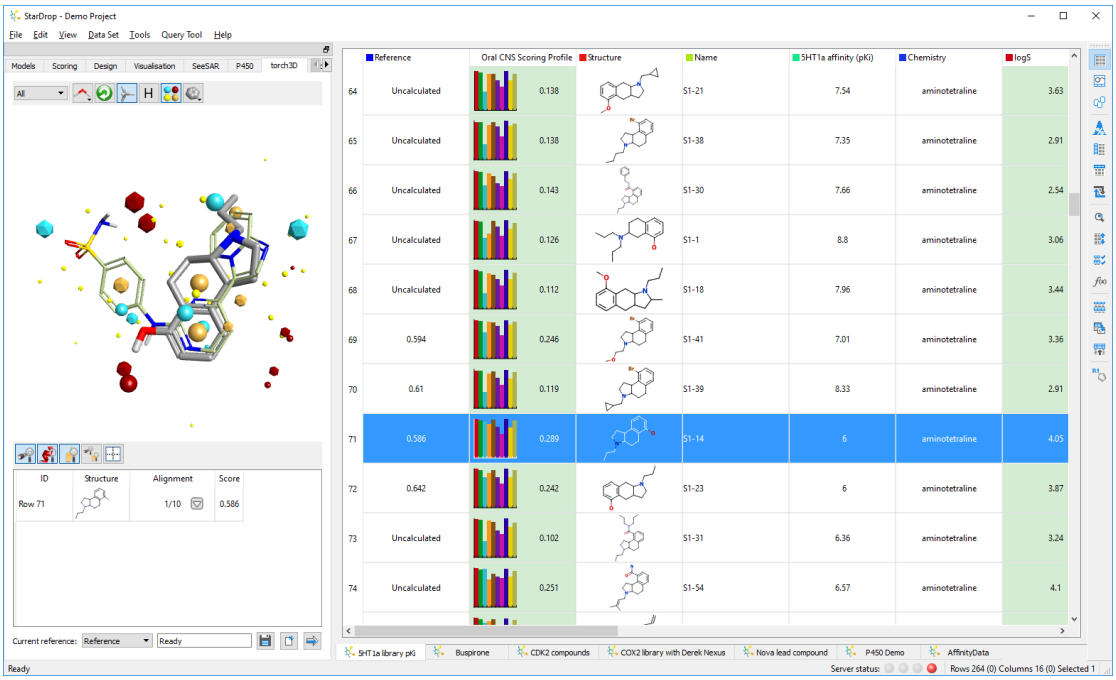
19.2 torch3D results

Once a torch3D reference has been set up, results for individual molecules can be calculated

To calculate a torch3D score for a compound, select that row in the data set and then click the  button. If multiple rows are selected then these will be queued with the status of each calculation displayed in the cell until the result has been calculated.
Note: Due to size constraints, a maximum of 500 molecules can be run against any given reference.

At the bottom of the torch3D tab, an indicator will show how many molecules are currently queued. If you wish to un-queue or cancel any calculations, simply select those rows in the data set and click the  button.

To view a result and see how the fields around a compound compare with the reference molecule, select that row in the data set.



The selected compound will be displayed superimposed upon the reference compound in the 3D display. Below the 3D display is a table showing which of the ten best alignments is currently displayed. To choose which of the alignments to view, click the up and down arrow buttons to change the displayed alignment. The score for the chosen alignment will be the one displayed within the data set. Selecting multiple rows will enable you to view them at the same time, either superimposed or in a grid.

19.2.1 Toolbar buttons

You can use the buttons above the 3D display as follows:

	Style-Surface Chooser The Style-Surface Chooser sets the domain of applicability or relevance of the

	display and surface toolbars. Possible values for the Relevance toolbar are: • All • References • Aligned For example, setting the Style-surface chooser to **Aligned** and clicking the change molecule colour button on the Toolbar causes the new colour to be applied to just the aligned molecules.
	Show structure as: • Lines • Thin sticks • Sticks • Ball and stick • van der Waals
	Reset the display
	Show/Hide the structure (Fields remain visible)
	Show/Hide hydrogens
	Show/Hide Field points
	**Surface Chooser** The Surface chooser is used to show, hide and remove molecular and Field surfaces to any molecule. Changes are only applied to the molecules specified by the Style-Surface Chooser. Field surfaces (positive, negative, shape, and hydrophobic) are shown at the contour level given in the spin box. • Remove all surfaces • Show/Hide solvent surface • Show/Hide positive Fields • Show/Hide negative Fields • Show/Hide shape Fields • Show/Hide hydrophobic Fields • Change Field contour level

The buttons below the 3D display enable you to:

	Show/Hide all molecules marked as search molecules
	Show/Hide all molecules marked as protein molecules
	Show/Hide selected molecules
	View reference and aligned molecules side-by-side
	Show molecules in a grid display rather than overlaid

19.2.2 3D Window

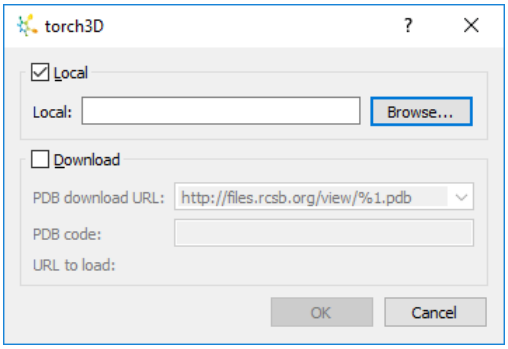
The display in the 3D window is controlled by the buttons described above. As a general rule, visual inspection of the 3D alignment is the best way to choose both successful alignments and good molecule correspondences.

The 3D display can be controlled by the mouse as follows:

Left mouse button	Rotate view around x/y axes
Right mouse button	Translate view
Middle mouse button	Zoom in/out
ALT key + Left mouse button	Zoom in/out
CTRL + Left mouse button	Rotate view around z axis
Mouse wheel up/down	Zoom in/out
'<' and '>'	Zoom in/out
SHIFT + Left mouse button	Clip in the Z-plane (up or down). The current clip value is shown in the lower right corner of the 3D window.

19.3 Extracting a reference molecule from a protein

To extract a reference molecule from a protein, click the **Extract from Protein** button on the **Load reference molecule** page of the torch3D wizard.

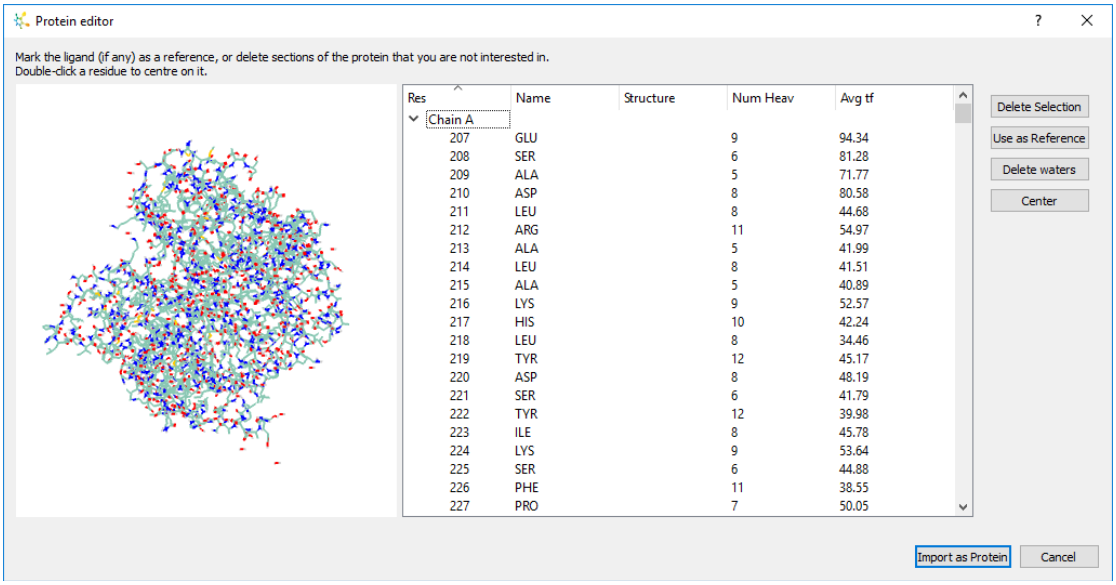


If you choose **Local** you can **Browse** for a PDB or mol2 file.

If you choose **Download** you need to specify a URL from which StarDrop can download the PDB file. The template provided needs to be update due to recent changes to the RCSB website, so in the **PDB download URL** box enter "http://files.rcsb.org/view/%1.pdb" (without the quotes). Having done this, specify the **PDB code** for the protein of interest (e.g. 1oit).

Click **OK** to open the file.

Note: The download URL will be retained so on subsequent uses you will only need to specify the PDB code.



If you are working on a target that has a protein crystal structure then it is possible to use the protein information surrounding your ligands as “excluded volumes”. As the name implies, the protein information is used to mark areas that ligands should not enter and pharmacophoric information (e.g. presence/absence of N, O, etc.) is not used.

You can use the various actions to process the protein and, once ready the **Import as Protein** button closes the editor and saves the remaining residues into the main application as a protein.

The table view of the protein contains a selectable list of protein residues that can be sorted using the column headers. The 2D structure of any non-protein residues will be displayed in the structure column. Usually this column contains the structure of the ligand(s) and hence clicking on the **Structure** header will cause the ligands to be sorted to the top or bottom of the list. Sometimes PDB files list anions and cations or other groups as HET atoms which can cause the ligand to be difficult to find. Often this can be solved by sorting the list on the number of heavy atoms that are present, bringing the interesting ligands to the top or bottom of the list.

Selecting a ligand (or other residue) in the table or the 3D window then clicking the **Use as Reference** button will make the selected residue the current reference molecule for the project. The molecule will be displayed in the 3D window in “Stick” form to indicate that it is now a reference molecule and no longer part of the protein.

If you double click a residue in the table, then the 3D window will centre on this residue. The window can then be zoomed (middle mouse or wheel) to view the residue in its 3D context.

The average temperature factor, **Avg tf**, column should be used to identify any highly flexible loops in the protein. If flexible loops exist close to the active site of the protein then it is advisable to remove them from protein excluded volume.

Any residue can be selected in the table and then deleted by clicking the **Delete** button or using the **Delete** key on the keyboard. Waters can be removed directly without preselecting them using the **Delete Waters** button.

A typical workflow for processing a PDB file is as follows:

1) Download the pdb file
   Choose **Download** and specify the **PDB code** (e.g. 1oit)
2) Identify the ligand

Click the **Structure** column header (to sort it in descending order) twice and click on the picture of the ligand

3) Label the ligand as a reference molecule

Click the **Use as reference** button

4) Delete the crystalized water

Click **Delete water** button

5) Optionally identify and remove flexible loops

Click on **Avg tf** column twice (to sort it in descending order), highlight residues with high values and inspect the protein complex in the 3D window to decide if they need removing then delete (using **Delete Selection** button) or modify selection as appropriate. High values are typically more than double the average across the whole structure.

6) Click **Import as Protein** to return to the main application

20 How do I... Use the Auto-Modeller™?

The Auto-Modeller enables you to build models of your own data sets. The process has been automated as far as possible to eliminate the requirement for you to be familiar with all the practices and mathematical techniques commonly applied when building QSAR models. However, if you are familiar with these then you have the option to control the parameters and techniques employed. The process requires you to provide an appropriate data set which can be automatically split into training, validation and test data sets before a set of descriptors is selected and a mathematical technique applied to model the data. By default a number of techniques will be applied and you will be able to see the results of all the models built and select and save the model(s) you wish to keep for future use.

At the first time of use this tab will be empty; however, if previous sessions have been run then these will be displayed.

To build a model you must open a data set containing a column of molecules, a column of unique ids and a column of Y values (data to be modelled), which may be numerical or categorical data. In some circumstances, it is possible to generate a model without having a chemical structure column (see section 20.5.5).

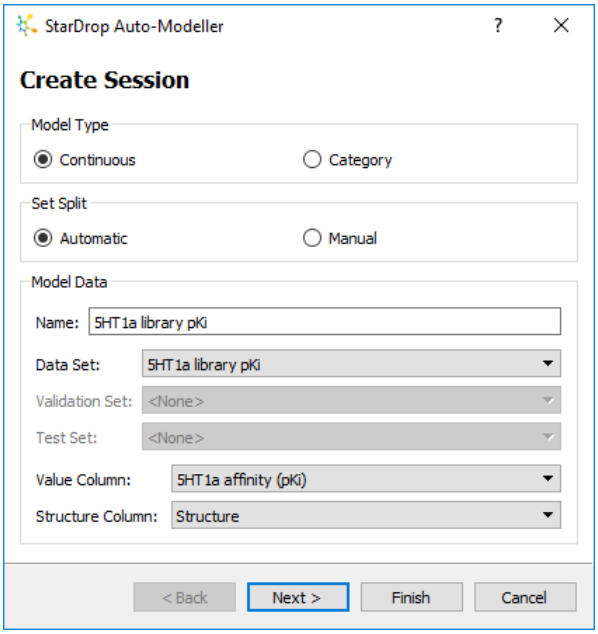
Click the  button to open the Auto-Modeller wizard dialogue (see section 20.1).

Note: This will only be possible if a server has been configured in the preferences (see section 24.3).

20.1 Auto-Modeller wizard

20.1.1 Create Session

In every step of the wizard, default values will be filled in, however, during the first step it will be necessary to confirm that the correct data set and model types are selected. After this you can click the **Finish** button at any time to start the session without making further choices.

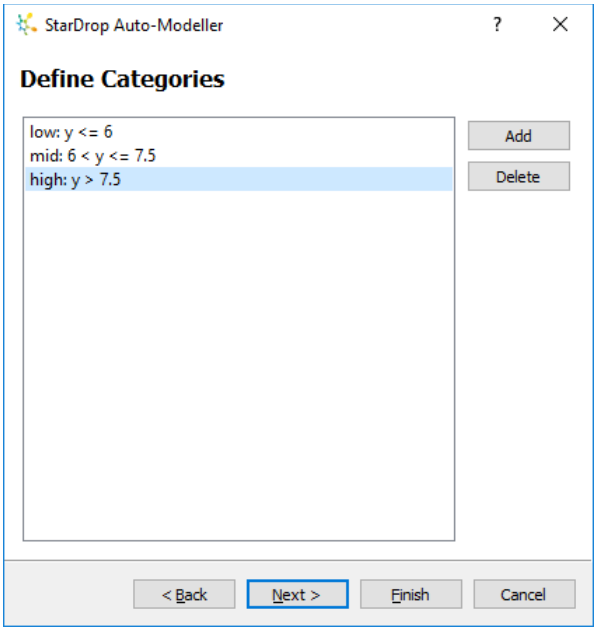


- The **Model Type** will be based upon the type of data in the current data set. If you choose **Category** the next step in the wizard will ask you to define categories to use.
- The **Set Split** will default to **Automatic**.
- The **Name** will default to the name of the current data set.

- The **Training Data Set** will default to the current data set if this is a valid set for generating models but all other open (and valid) data sets will be available in the drop-down list.
- The **Validation Data Set** and **Test Data Set** options will only be available if the **Set Split** is set to **Manual**. (see section 13.5.1)
- The **Value Column** defaults to the first column in the data set of the appropriate type but the drop-down list will contain all other appropriate columns.
- The **Structure Column** will default to the first column containing structures in the data set.
- The **ID Column** will default to the first column containing text in the data set but the drop-down list will contain any other columns that also contain text.

20.1.2 Define Categories

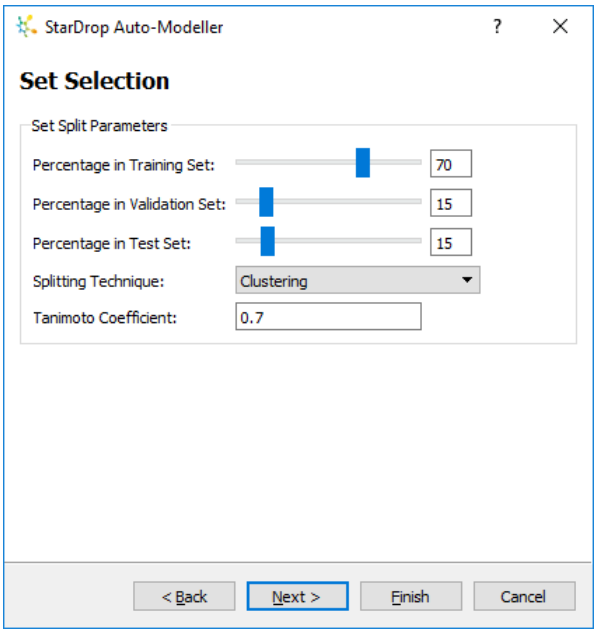
If you have chosen to build a **Category** model from numerical data then the following step will appear next in the wizard:



If the data are already categorical then these will appear in the list.

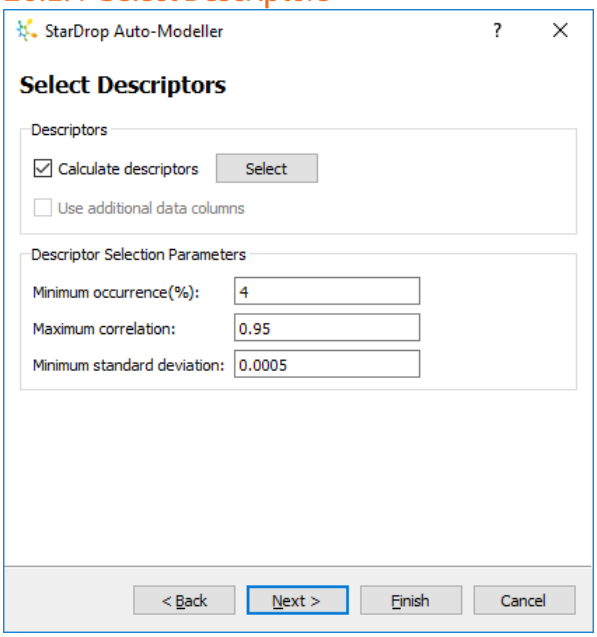
Click the **Add** and **Delete** buttons to add or remove categories respectively. Click on the category name to edit it. If the data are numerical then you can click on the range values to edit them.

20.1.3 Set Selection



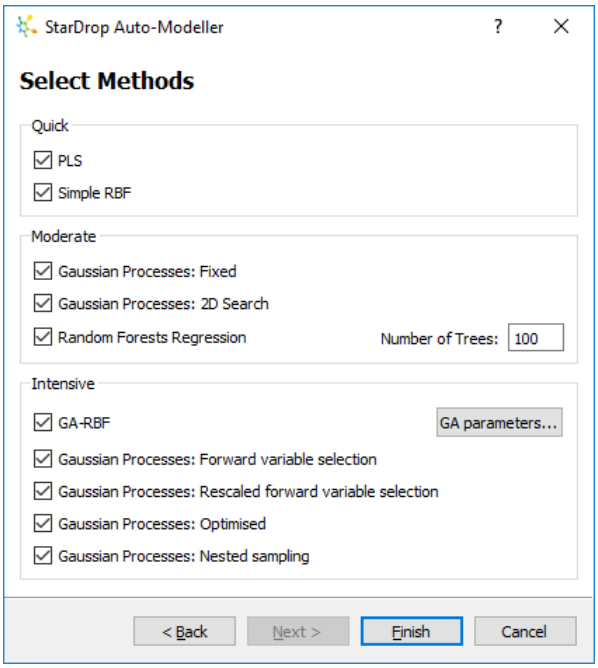
Click **Next** to set any parameters for set selection – default values will already be filled in (see section 20.5.1)

20.1.4 Select Descriptors



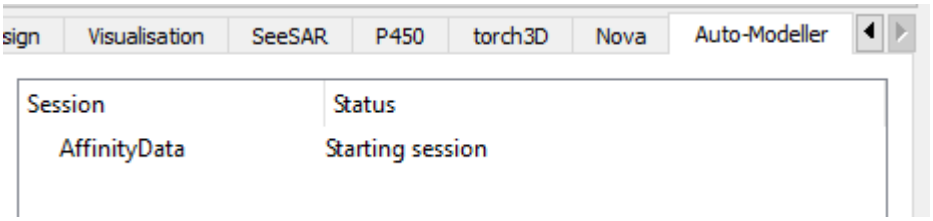
Click **Next** to set parameters for descriptor selection. Default values will already be filled in (see section 20.5.2). By default, **Calculate descriptors** will be checked. The descriptors to be calculated can be customised by clicking **Select** (see section 20.5.4). **Use additional data columns** will be available if the data set contains more than just the three mandatory columns because such columns can optionally be used as descriptors (see section 20.5.5).

20.1.5 Select Methods

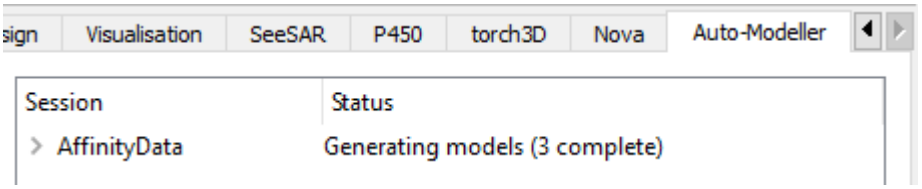


Click **Next** to choose modelling techniques to apply – a default selection will already be made (see section 20.5.3).

Click the **Finish** button to start the session. A model generating session will be started on the Auto-Modeller server and at the same time this will be displayed at the top of the **Auto-Modeller** tab.

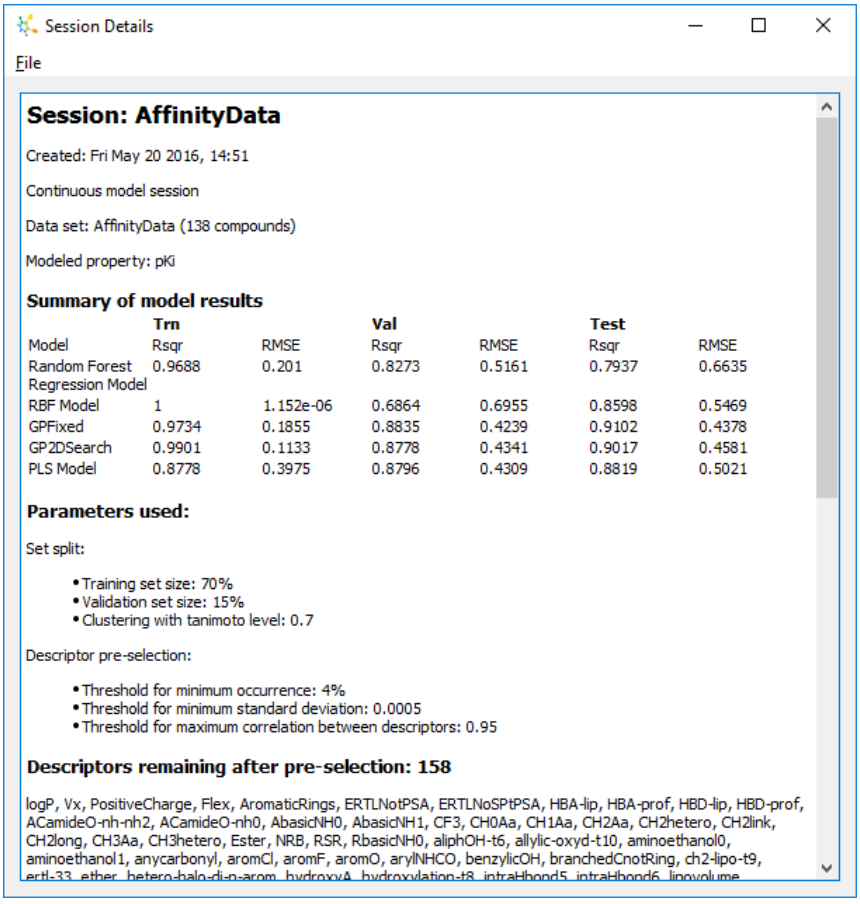


StarDrop will now receive updates from the Auto-Modeller server indicating the status of the session.



It is not necessary to wait until all models have been generated to start analysing them (see section 20.2). If you wish to stop a session before all models have been generated, this can be done by right-clicking over the session and selecting **Halt Session** from the menu.

Any time after a session has been started, you can obtain a report of the session details by right-clicking on the session and selecting **View Session Details** from the menu. This will display a window containing information about the parameters used to run the session, the results of the set splitting and descriptor selection, and statistics for the models generated when these are available:

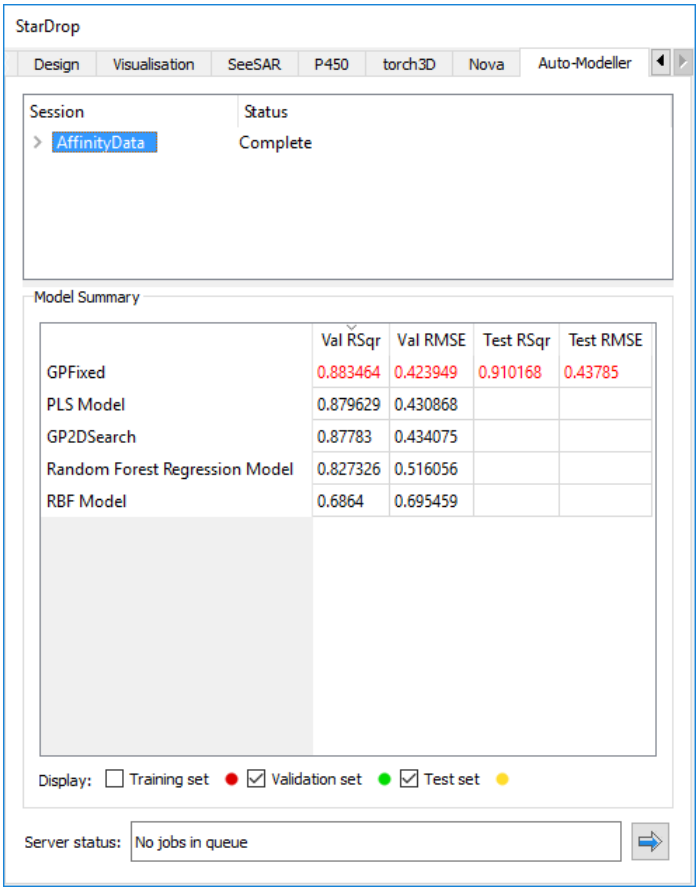


This report can be printed or saved as a PDF or HTML document, using the **File** menu.

Once all the models have been generated the status will say **Complete** and the best model will be highlighted in red.

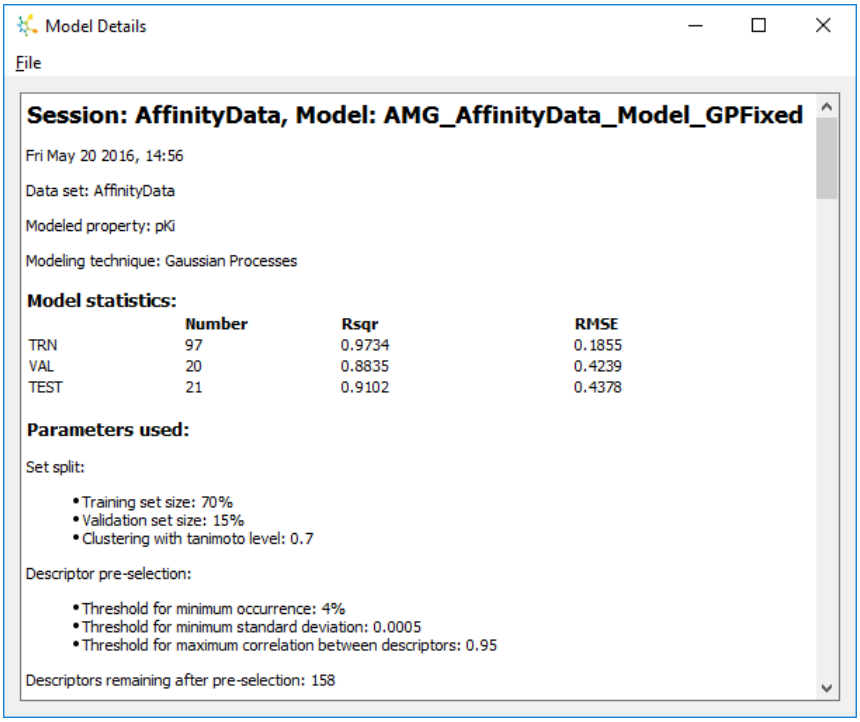
20.2 Analysing models

To see a summary of the statistics for the all the models generated (at any stage once at least one model has been returned), highlight the session at the top of the **Auto-Modeller** tab. A table will then be displayed in the bottom half of the **Auto-Modeller** tab showing the statistics (R-squared and Root Mean Square Error (RMSE) for continuous models, and the kappa statistic and accuracy for category models) for all the model building data sets. By default this will just be the test set:



By selecting the check-boxes at the bottom of the **Auto-Modeller** tab it is possible to show/hide all of the data sets used in the model generation process. By default the statistics for the test set will only be displayed for the best model. This is because the model was chosen as the best based upon the validation set statistics and the test set is there to provide a final assessment of the predictive power of the final chosen model.

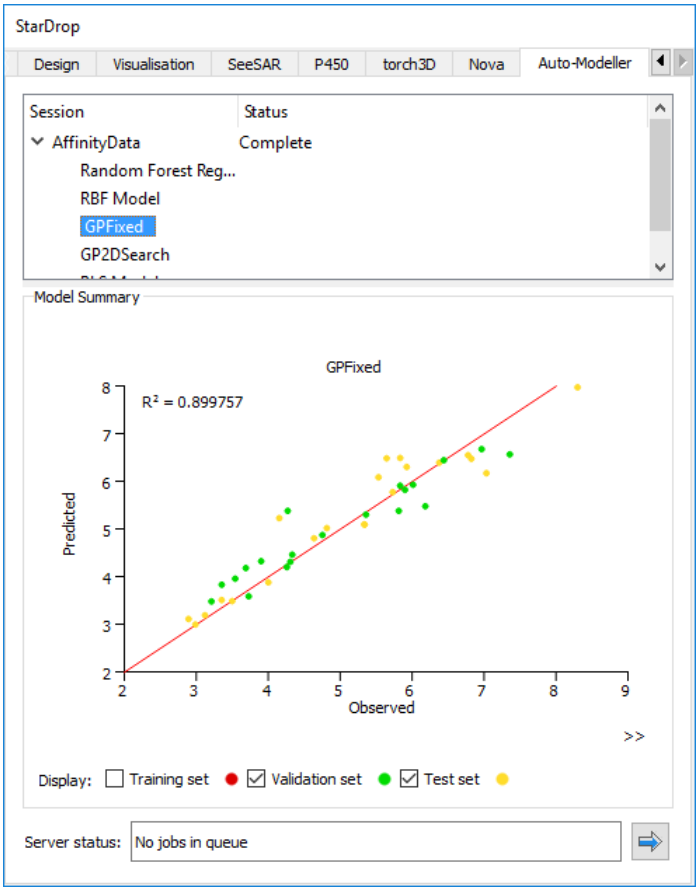
To view details of a model, right-click over the model to bring up the menu and select the menu option **View Model**. This will bring up a window giving model statistics, details about the settings used for descriptor selection, descriptors used and any other model technique specific information:



This can be printed or saved as a PDF or HTML document from the **File** menu at the top of the window.

20.2.1 Continuous models

To view an XY scatter graph displaying the observed versus predicted data for a continuous model select that model in the list.



To see the molecule associated with any point on the graph, hover the mouse over that point and a pop-up structure will be displayed.

As before, when looking at the statistics for each of the data sets, to show or hide points from each of the three model building sets (training, validation and test) check or uncheck the boxes below the graph. The training set points are coloured red, the validation set points green and the test set points yellow.

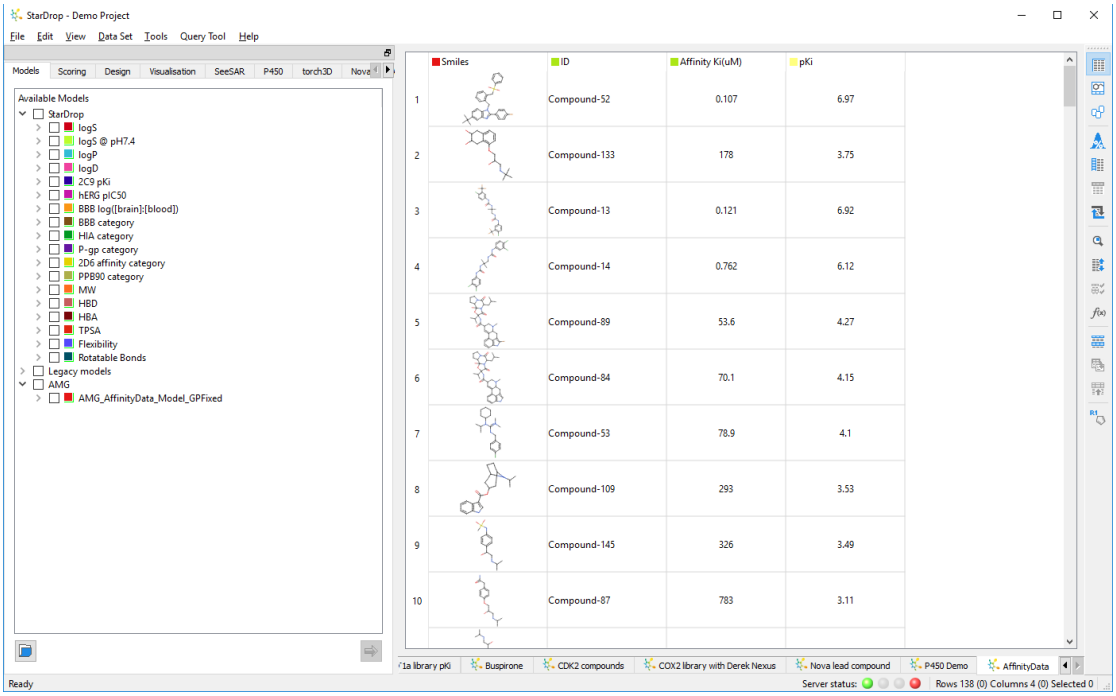
Right-clicking over the graph will bring up a menu enabling you to change the background colour, save the graph as a Portable Network Graphics (PNG) image file or copy it to the clipboard.

20.2.2 Category models

To view a confusion matrix of the results for a category model select that model from the list. Checking and un-checking the boxes below the matrix will change the model building data sets included in the matrix.

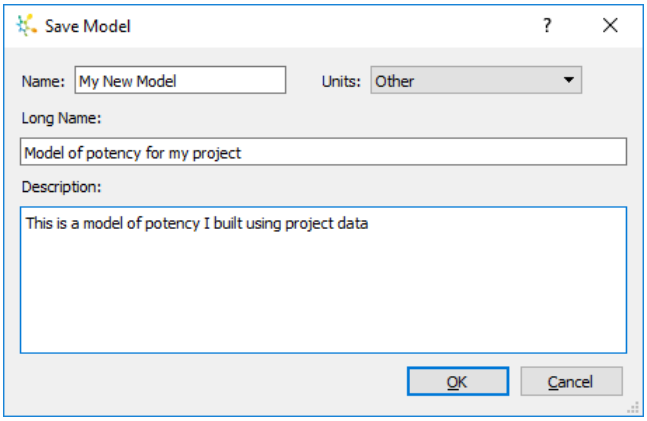
20.3 Using new models

Once models have been generated, it is possible to start using them to make predictions immediately. All the models that have been viewed will appear in the **Models** tab as **AMG** models.



These models are temporary because they are only available during the current session. If you wish to keep the model you can choose to save it. To save a model, right-click on the model in the **Auto-Modeller** tab to bring up the menu and select **Save Model**.

A dialogue box will be displayed prompting you to provide some details about the model. You can provide an appropriate set of details describing the model, including the units if appropriate. For a category model the description should include some definition of the categories used.



You can then choose a filename and save the model. Once saved the model will appear in the **Models** tab as a **Custom** model.

Each time StarDrop is started this model will be available as a custom model and can be run in the usual way. If you give other users copies of, or access to, the model file saved then they too can use this model.

20.4 Deleting sessions

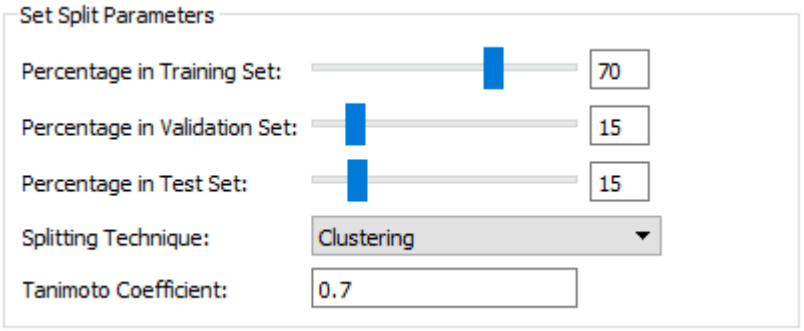
When a session has completed it will remain available on the server until deleted. To delete a session, right-click on the session and select **Delete Session** from the menu. Once a session has been deleted, it is no longer possible to access any of the information or models associated with it, unless they have been saved as described above. The same menu enables you to remove all current sessions by selecting **Delete All Sessions**.

20.5 Advanced features

There are a number of options to allow you to customise the way that the Auto-Modeller runs. This section gives an overview of how to change some of these settings but for further information on the underlying techniques and the implications of making changes to these settings please refer to the StarDrop Reference Guide. These options can be saved in the Auto-Modeller preferences (see section 2.5) or set in the Auto-Modeller wizard (see section 13.1).

20.5.1 Set splitting

You can make changes to the way that a data set is automatically split into training, validation and test data sets for the model generation process. By default, 70% of the data will be put into the training set with the rest split evenly between the validation and test sets. There are three techniques available for performing this split:



Clustering

The data set is clustered using fingerprints of the molecule structures with a specified Tanimoto coefficient (by default this is 0.7). All the cluster centres and all the entries which do not fall into any cluster are put into the training set. The remaining data in each cluster is then sorted on the property

value and split between the training, validation and test sets to ensure that the required number end up in the training set (by default this is 70% of the total). The rest are split between the validation and test sets (by default, 15% in each). If, after clustering, there are too few compounds in clusters to make up the validation and test sets then the split is performed using the Y-based procedure below.

Y-based

The data set is sorted on the property value and then randomly picked from bins of similar values to go into the training, validation and test sets such that each set will have a similar spread of property values and each will be the appropriate size.

Random

The data set is split randomly into the three sets in the correct proportions.

It is possible to specify a size of zero for the test set, although this is not recommended. However, both training and validation sets must have data in order to build models.

If **Clustering** is selected as the split technique, the Tanimoto coefficient may also be specified. It must always be between 0 and 1. If it is set at either 0 or 1 then no clustering will take place and the data will be split on the basis of property values alone.

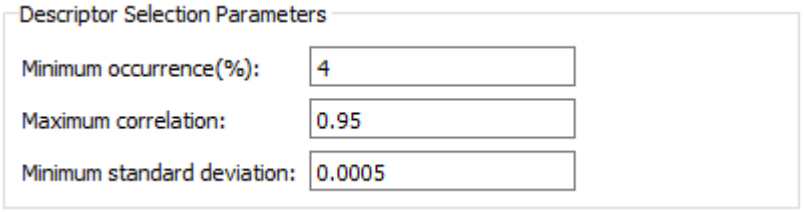
Manual splitting

Alternatively you can create your own training, validation and test data sets. To do this, create three data sets, each with exactly the same columns before starting the Auto-Modeller wizard where you must select the **Manual** option in the **Set Split** section when generating models.

Note: The IDs must be unique across all three data sets.

20.5.2 Descriptor selection

When selecting descriptors with which to build models, StarDrop first performs a number of initial checks to remove unnecessary descriptors.



Minimum Occurrence

All descriptors that have a value of zero too often are removed. This option determines the minimum percentage of the data set for which the descriptor must not be zero (by default this is 4%).

Maximum Correlation

Any descriptor that is too highly correlated with another descriptor will be removed. This option determines the maximum allowable correlation between two descriptors before one of them is removed. This value must be between 0 and 1 inclusive (by default this is 0.95).

Minimum Standard Deviation

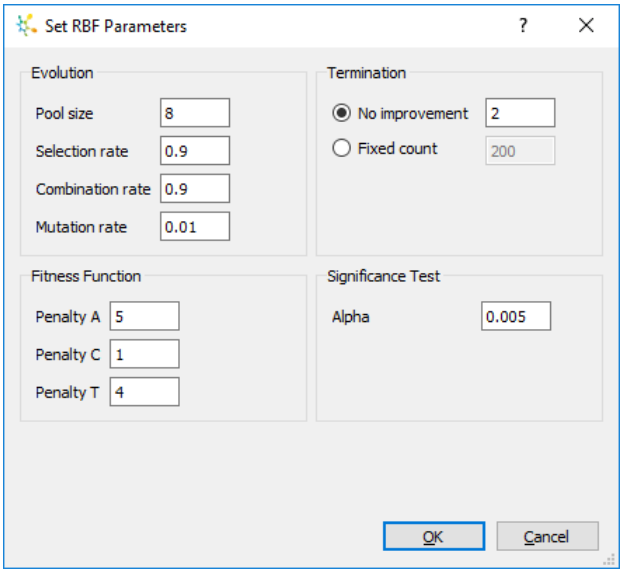
Any descriptor that has too little variation across the data set will be removed. This option determines the minimum acceptable standard deviation across the data set for any descriptor (by default this value is 0.0005).

20.5.3 Modelling techniques

When building continuous models, StarDrop uses a variety of mathematical techniques. By default, all are enabled.

For the Random Forests technique, setting the number of trees is available as an option.

Additionally, the GA-RBF technique uses a genetic algorithm to choose descriptors to ensure an appropriate ratio between the number of descriptors in the final model and the number of molecules in the data set, before applying a Radial Basis Function method. To change these parameters click the **GA Parameters…** button to display the **Set RBF Parameters** dialogue.



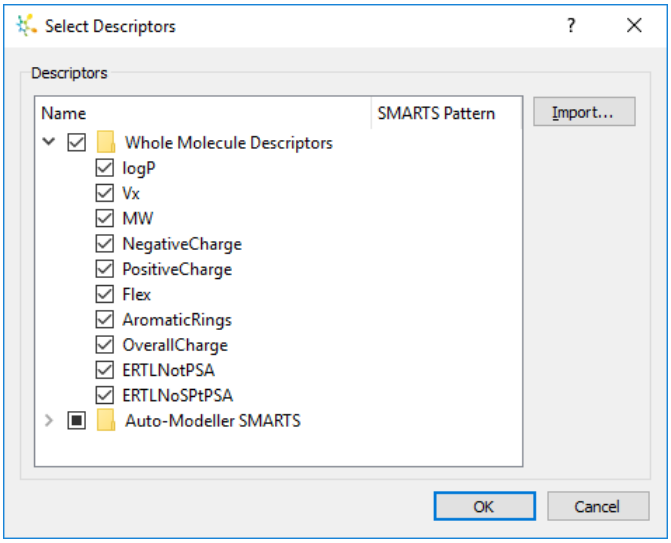
The genetic algorithm works by successively evolving generations of chromosomes, each of which describes a list of descriptors to be used, and builds a model for each one. In each generation, a number of chromosomes are generated and compared. A chromosome that produces the best model is considered to be the best chromosome. Successive generations of chromosomes are then evolved by selecting from the set of chromosomes for that generation. In some cases chromosomes are combined to produce a new chromosome with a bias towards combining those that produced the best models. This continues until the necessary stopping conditions have been satisfied. The whole process is carried out 250 times to determine the descriptors that consistently occur in the best models. The parameters that control the genetic algorithm are as follows:

- **Mutation Rate** determines the proportion of times (between 0 for never and 1 for every time) that a chromosome evolved for the next generation mutates.
- **Combination Rate** determines the proportion of times (between 0 for never and 1 for every time) that a selected chromosome is combined with a second chromosome to create two new chromosomes for the next generation rather than being used itself.
- **Selection Rate** determines the proportion of times (between 0 for never and 1 for every time) that the best chromosome is selected instead of the choice being random.
- **Pool Size** (an integer between 1 and 20) determines the number of chromosomes put into a selection pool to be used when selecting a chromosome.
- **Use Stopping Criteria** means that the genetic algorithm will be tested against a penalty function and the algorithm will continue until the score against the penalty function doesn't improve for a set number of generations. When this option is checked it is possible to define the number of generations for which the penalty function result must remain unchanged in order to end the process. When this option is unchecked the algorithm is run for a fixed number of generations.
- **Stop Count** is only enabled when there are no other stopping criteria. This value must be 1 or greater and determines the number of generations for which the genetic algorithm will run.
- **Stopping Criteria** is only enabled when **Use Stopping Criteria** is checked. This variable sets the number of times the best result against the penalty function must remain unchanged for the algorithm to stop.
- **Penalty A**, **Penalty C** and **Penalty T** determine the shape of the penalty function to use for comparing the results from successive generations.

- **Alpha** determines the threshold used for determining the most important descriptors at the end of the genetic algorithm. The lower this value (which must be between 0 and 1) the more frequently a descriptor must have occurred to make the final selection.

20.5.4 Customising descriptors

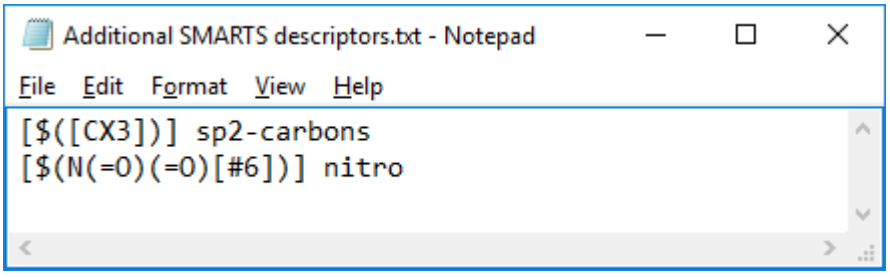
By default, StarDrop will calculate a set of descriptors to use in the model building process. These consist of some whole molecule descriptors, including molecular weight, logP and TPSA, and some SMARTS patterns (for more information on SMARTS patterns see the StarDrop Reference Guide). To customise this list for a particular model building session, click **Select** in the **Descriptors** section of the **Generate Models** dialogue. This will display the **Select Descriptors** dialogue:



By checking and un-checking the boxes, you can choose which descriptors are used.

Clicking the **Import...** button enables you to import your own filters to add to the list. To define your own you must create a text file containing SMARTS and their associated names. The SMARTS patterns must not contain any spaces and there should be a space to separate the pattern from the name, with one pattern and name on each line of the file.

Example:



Clicking the **Manage...** button enables you to choose which collections are available. StarDrop uses SMARTS patterns in different ways so it is not always useful or appropriate to use all the available patterns as descriptors.

20.5.5 Additional descriptors

You can use additional descriptors either as well as, or in place of, those already in StarDrop. These can either be provided as additional columns in the data set(s) or as a file of SMARTS patterns. If you do not have a column of molecular structures, then you can only use additional columns as descriptors.

When using extra descriptors you must then check the box labelled **Use additional data columns** in the Auto-Modeller wizard (see section 20.1.4). Having done this, you will then be prompted to confirm which of the additional columns should be used as descriptors:

If you have chosen to manually split the data set then each of the training, validation and test data sets must have the same additional columns.

Note: Models that use additional descriptors can only be run against structures in a data set if that data set already contains columns of the additional descriptors. It is not possible to generate Glowing Molecules from these models.

20.5.6 Test set statistics

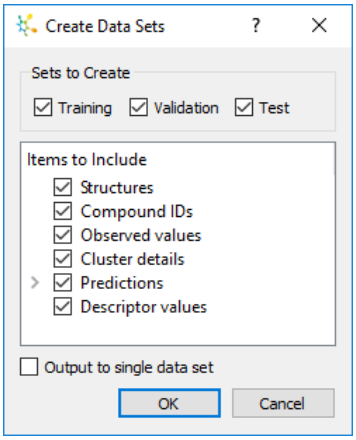
The test set is designed to provide a final independent assessment of the quality of the model chosen from all those generated in a session. Although it is possible to see test set statistics it is not recommended they be used for model selection, even though the test set results may not always be as good as the validation set results.

By default, test set statistics are only displayed in the **Auto-Modeller** tab for the best model in any model generation session. However, to view the test set statistics for all the models generated, ensure that the option **Test best model only** is unchecked in the Auto-Modeller preferences (see section 24.3).

20.5.7 Creating split data sets

Once a model building session has completed, you can retrieve the data set splits for further analysis. As well as structures, IDs and observed values, the information available can include clustering information, predictions for all models created, and descriptor values. For decision tree models you can also generate information indicating which compound belongs to which leaf/rule.

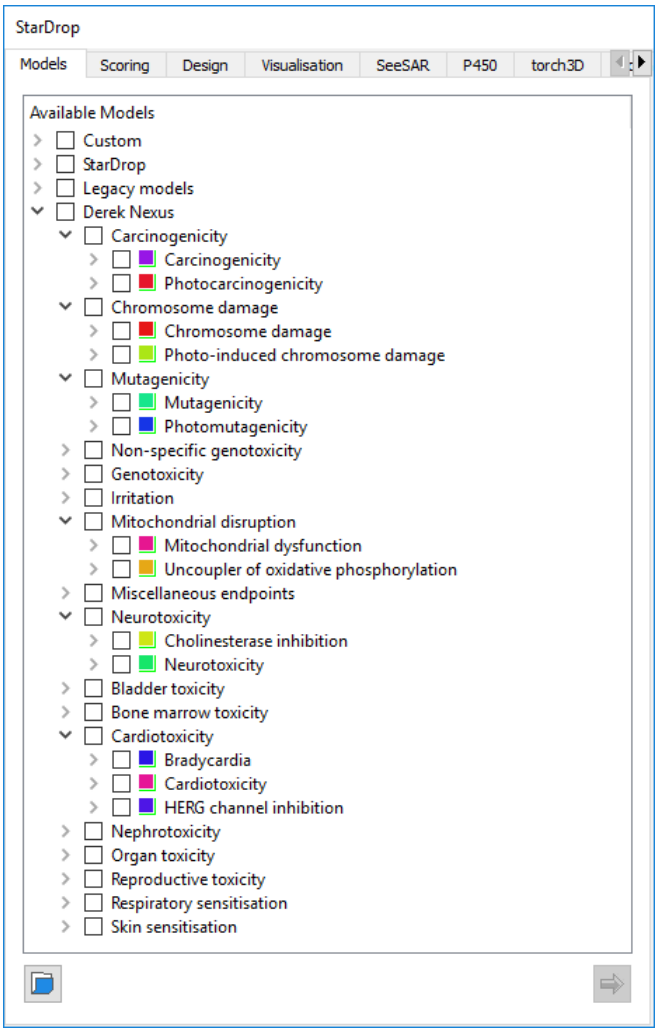
To do this, right-click over the session name and select **Create Split Data Sets…**. This will display the **Create Data Sets** dialogue enabling you to choose which sets to create and what data to include. Tick the **Output to single data set** option if you would like the splitting information in a single data set with a category assigned to each row indicating the set to which the row was assigned.



The data sets will then be added to the main StarDrop window.

21 How do I... Use Derek Nexus™ models?

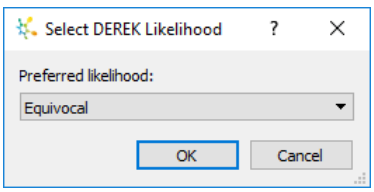
When StarDrop is started, if you have a license enabling the Derek Nexus models and have configured a connection to your Derek Nexus server (see section 0) then these will appear in the Models tab.



There are over 40 endpoints available and each model can be run in exactly the same way as StarDrop's other models (see section 12.1) by selecting the models of interest and clicking the  button.

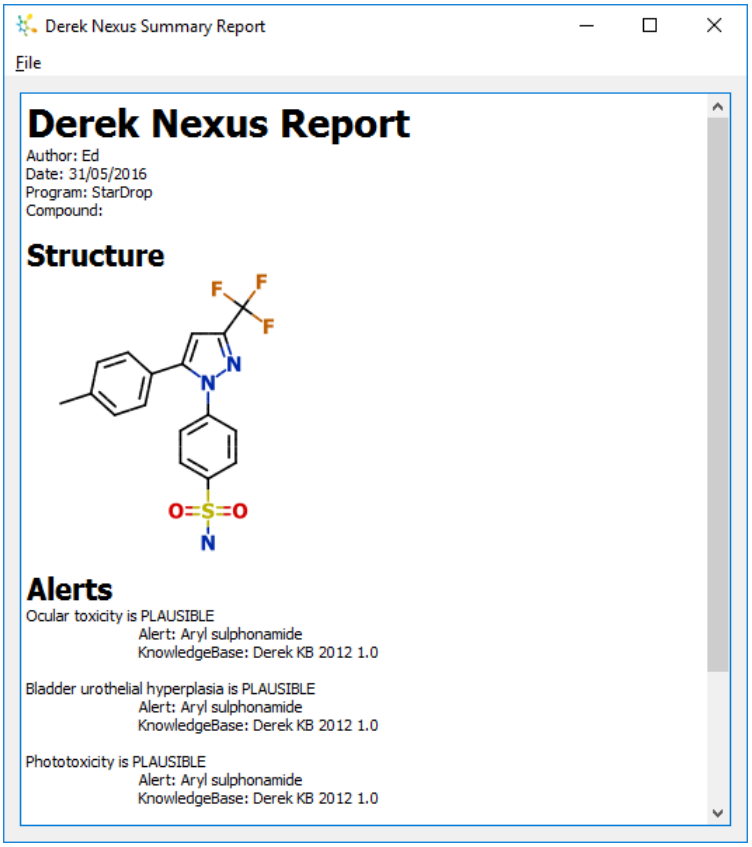
For any predictions that are returned with a result other than 'No report', hovering the mouse over the value will display a tool tip showing the species involved and a description of the alert that was triggered.

If you have run multiple Derek Nexus models you can create a summary of results using the **Tools->Derek Nexus->Add summary column...** menu item. This will present you with a dialogue enabling you to choose what level of likelihood you consider to be an alert.



This then creates a single column containing ALERT/NO ALERT categories based upon combining all the alerts from the predicted endpoints.

If you select a single row in the data set you can select the **Tools->Derek Nexus->Create a report** menu item. This will display a report of any predicted toxicities which can be printed or saved as a PDF document.



As with StarDrop's other models, the Glowing Molecule can be used to see which parts of the molecule have caused the alert.

22 How do I... Use MPO Explorer™?

MPO Explorer can be used to generate scoring profiles from your own data sets and analyse the robustness of your decisions to the selection criteria you have chosen. Essentially, it reverses the scoring process to determine whether or not there are easily interpretable rules encoded within the data which could be used to increase the likelihood of finding future compounds which meet your objectives.

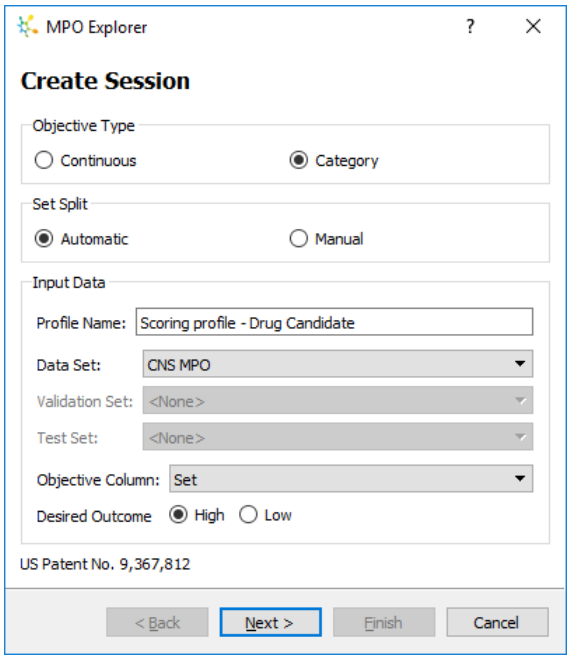
To build a profile you must open a data set containing a number of columns of property values and a column of objective values (data to be modelled). The objective and properties can be either numerical or categorical. MPO Explorer can handle missing data for the properties, but any row with a missing objective value will be ignored.

22.1 Profile Builder wizard

To build a new scoring profile, click the **Build Profile...** button on the **Scoring** tab to open the **Profile Builder wizard** dialogue.

22.1.1 Create Session

The **Objective Type** will be based upon the type of data in the current data set. If you choose **Category** the next step in the wizard will ask you to define categories to use.

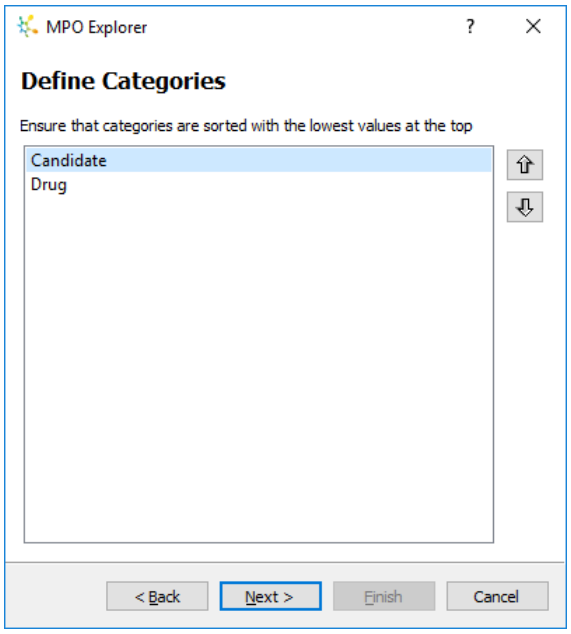


- The **Set Split** will default to **Automatic**.
- The **Profile Name** will default to "Scoring profile - " followed by the name of the chosen objective.
- The **Data Set** will default to the current data set if this is a valid set for building profiles but all other open (and valid) data sets will also be available in the drop-down list.
- The **Training Set**, **Validation Set**, and **Test Set** options will only be available if the **Set Split** is set to **Manual** (see section 22.1.4).
- The **Objective Column** defaults to the first column in the data set of the appropriate type but the drop-down list will contain all other appropriate columns.
- The **Desired Outcome** option lets you choose whether you want MPO Explorer to search for high or low values of the objective. (For categorical objectives, the next wizard page allows you to specify an ordering for the categories from "low" to "high".)

22.1.2 Define Categories – Categorical Objective

If you have specified a **Category** objective type for a *categorical* objective column, then the following step will appear next in the wizard:

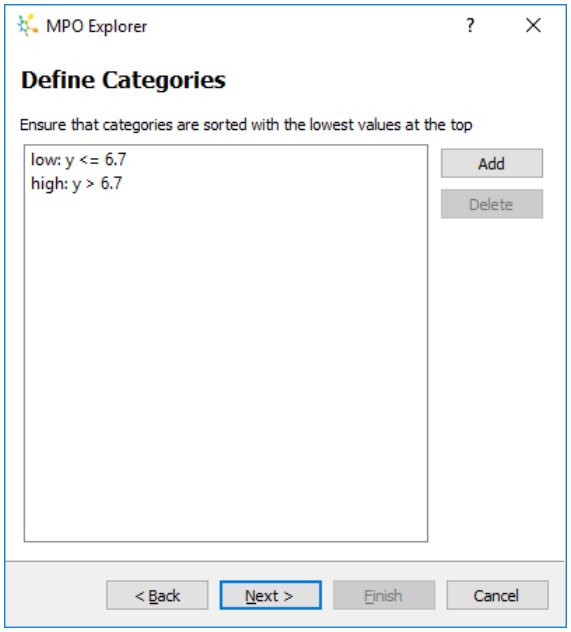
The categories used in the objective column will be automatically detected and added to the list. Because the Profile Builder works by searching for regions of property space where the mean objective value is relatively high (or low), you must specify an ordering for the categories from the "lowest" category to the "highest". Click on the arrows on the right to rearrange the categories in ascending order.



22.1.3 Define Categories – Numerical Objective

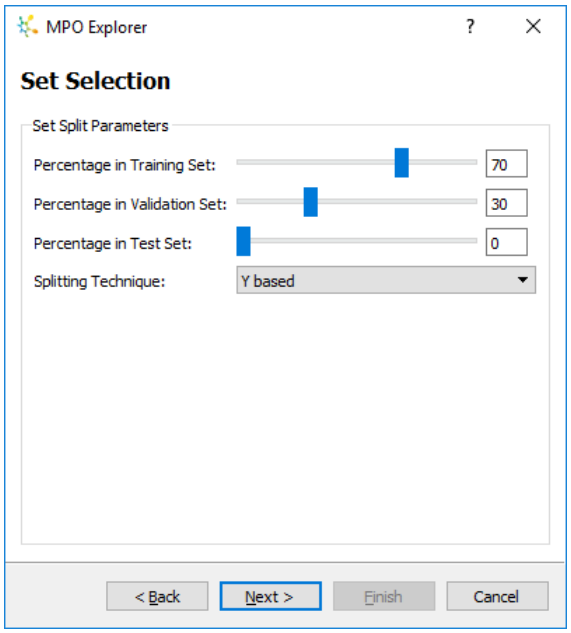
If you have specified a **Category** objective type for a *numerical* objective column then the following step will appear next in the wizard:

Click the **Add** and **Delete** buttons to add or remove categories respectively. Click on the category name to edit it. You can also click on the range values to edit them.



22.1.4 Set Selection

Click **Next** to set any parameters for set selection – default values will already be filled in. You can make changes to the way that a data set is automatically split into training, validation and test data sets for the model generation process. By default, 70% of the data will be put into the training set with the rest split evenly between the validation and test sets. There are two techniques available for performing this split:



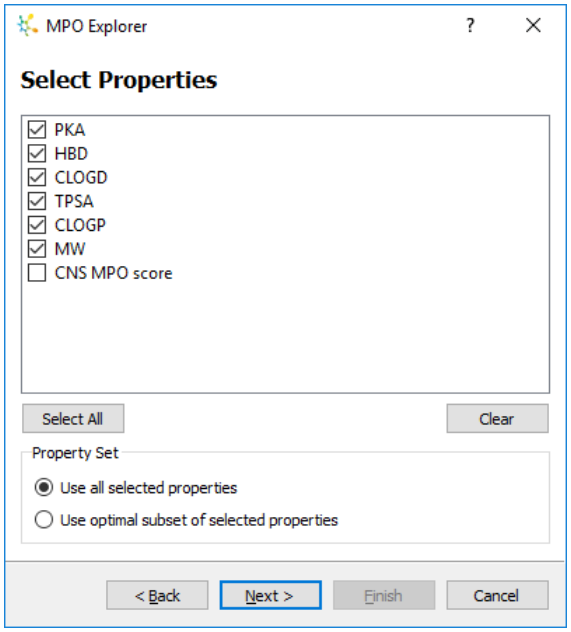
- **Y-based**
  The data set is sorted on the property value and then randomly picked from bins of similar values to go into the training, validation and test sets such that each set will have a similar spread of property values and each will be the appropriate size.

- **Random**
  The data set is split randomly into the three sets in the correct proportions.

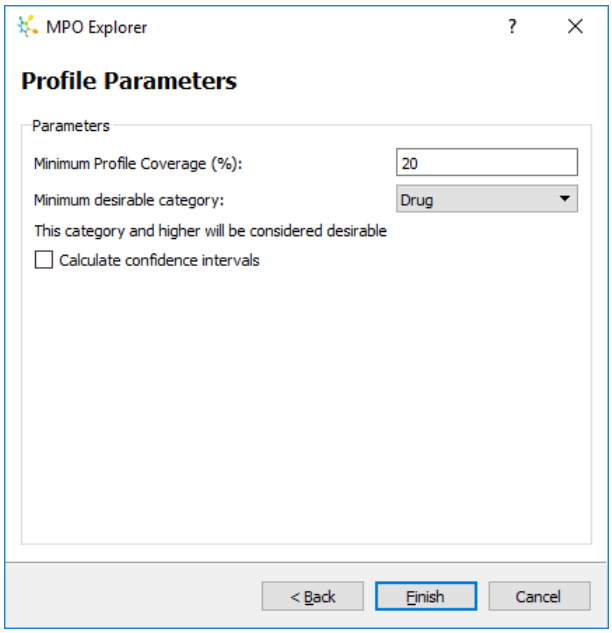
22.1.5 Select Properties

In this step, you can specify which of the available properties the Profile Builder should use. By default, all columns in the data set that can be used as properties will be selected.

You can also specify whether the Profile Builder should use *all* the selected properties for training, or whether it should first search for the most predictive subset of the selected properties and then use only these properties for training. The latter option is particularly useful when the set of properties is prohibitively large (e.g. over 15), however this step can take a lot longer to complete.



22.1.6 Profile Parameters



The **Minimum Profile Coverage** option indicates the minimum proportion of the data set which must meet the criteria of a rule for it to be considered acceptable; the Profile Builder will always seek to find rules that apply to at least this proportion of the data set. By default, this is set to 20%.

If you have specified a categorical objective, the **Minimum desirable category** option tells the Profile Builder which compounds should be considered "desirable", i.e. it is used to specify what objective values you are searching for; in particular when your categorical objective has more than two possible outcomes.

Similarly, if you have specified a continuous objective, the **True/False Threshold** tells the Profile Builder what the cut-off point is for compounds to be considered "desirable"; again, it is used to specify what objective values you are searching for.

Select the **Calculate confidence intervals** option to calculate "soft" boundaries for continuous properties instead of hard cut-offs, based on the amount of data used to generate the boundaries and any potential instability in the box thresholds. This option is deselected by default as it will cause the Profile Builder to take a lot longer to complete.

Click the **Finish** button to run the Profile Builder using the data set and parameters you have selected.

22.2Profile Builder view

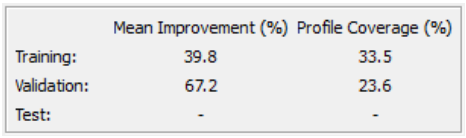
If the Profile Builder is able to find a rule, it will be displayed in an interactive "grid view" as shown above. This view enables you perform multi-parameter optimisation in a highly visual and interactive way; it allows you to easily see, manipulate, and understand the effect of the rule that was found.

Shown on the top right is the rule itself displayed as a standard StarDrop scoring profile, i.e. a list of properties together with their desired values and importances. You can adjust the desired values here and the graphs and statistics will update in real time; to delete a property criterion, click on it and press the **Delete** key.

If you have made changes to the rule, at any point you can click the **Reset** button to undo these and revert to the original rule found by the Profile Builder. Once you are satisfied with the rule, you can click the **Find Next** button to continue searching for another rule to add to the profile.

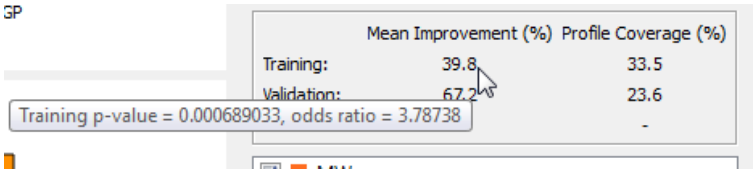
Note: As new rules are found and added to the profile you can choose which rule's plots are displayed by selecting it within the profile. However, you can only make changes to the most recent rule found. To make changes to other rules you must complete the process to generate the new profile and then click the **Analyse...** button on the Scoring tab.

You may also press **Discard** if you do not wish to use the current rule.



	Mean Improvement (%)	Profile Coverage (%)
Training:	39.8	33.5
Validation:	67.2	23.6
Test:	-	-

Below the profile on the right are a number of statistics showing how the current rule performs against the training, validation, and test sets. As you make changes to the rule, you can keep track of how the mean improvement and profile coverage are affected to ensure that your changes are having the desired outcome.



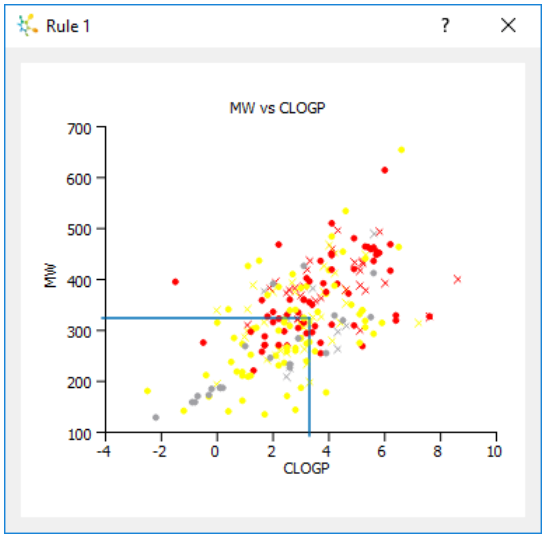
If you move your mouse cursor over the mean improvement statistics for each set, you can also see the *p*-value and odds ratio for the rule over each of the training, validation, and test sets.

Click on the colour bar on the bottom right of the Profile Builder to change the colour key used to depict desirable and undesirable compounds.

Clicking on the **Create Sets** button will generate the individual training, validation, and test sets used by the Profile Builder (in the background behind the window). These sets can then be saved and manipulated just as any other data sets within StarDrop.

By default, the main 'grid view' shows each property in the rule plotted against each other property for the full data set, i.e. the combined training, validation, and test sets (if present), each with a different symbol. You can choose whether or not to display the training, validation, and test sets by selecting or deselecting the **Training set**, **Validation set**, and **Test set** options on the bottom left of the main window. Each property criterion is indicated by the blue lines or boxes.

You can zoom into any particular plot by right-clicking on it and selecting **Detach**. To distinguish between overlapping compounds, you can also drag the **Jitter** slider in the main window to randomly perturb the position of each compound by a small distance.

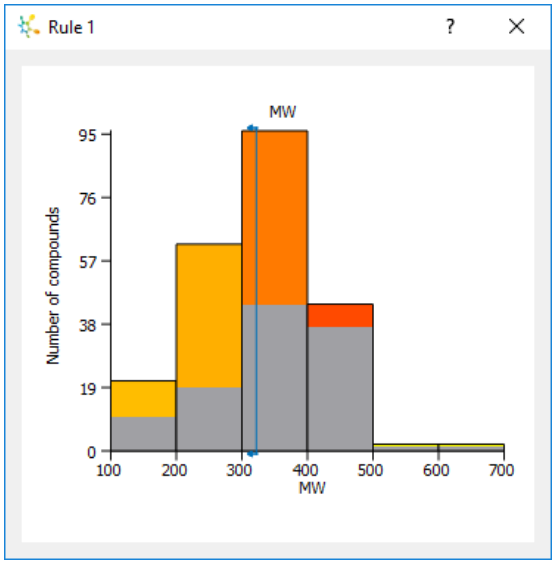


The two relevant property criteria from the rule are displayed as a blue box. You can adjust the boundaries of this box by dragging its edges; the goal is to place as many good (in this case yellow) compounds inside the box as possible while excluding as many bad (in this case red) compounds as you can.

As you change the boundaries, all the other plots and statistics will update to show the effect your changes are having on the other criteria.

Note that a number of compounds are also coloured in grey. A grey compound is one that has been filtered out by one of the property criteria *not* represented in the current plot, so including or excluding it in the blue box will have no effect on the rule's performance.

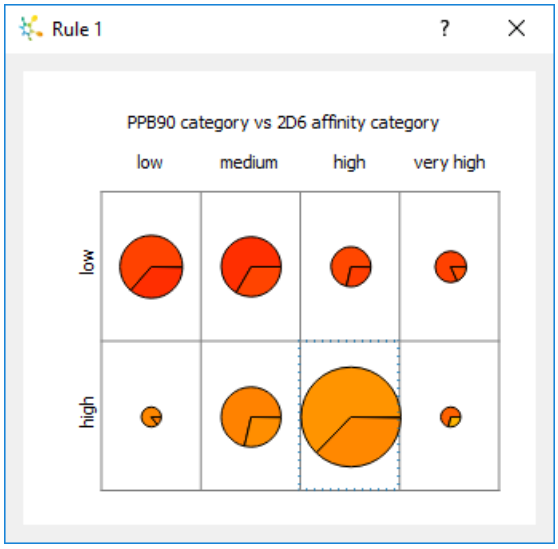
Having a large number of grey compounds in a plot implies that you should first focus on other property criteria, because adjusting the criteria represented in the plot will have a relatively small effect on the rule's performance.



Along the diagonal you will see histogram plots for each property. In the histogram above, you can see how the MW property is distributed across the full set of compounds. The property criterion is displayed as a blue vertical line, and dragging it will change the criterion.

The coloured portion of each bar represents the mean objective value within the property bin; thus 'yellower' bars are more desirable. The grey portion of each bar indicates how many compounds within the property bin are filtered out by some other property criterion.

As before, large grey regions in the plot imply that you should first focus on the effect other property criteria are having on the rule.



When categorical criteria are used the size of each pie represents the number of compounds in the pie. The property criteria are highlighted in blue; the dotted lines indicate that these criteria cannot be adjusted from the plot itself, but you can change them in the profile view on the right.

The coloured portion of each pie represents the mean objective value within the pie; thus 'yellower' pies are more desirable. The grey portion of each pie indicates how many compounds within the pie are filtered out by some other property criterion.

As before, large grey regions in the pie charts imply that you should first focus on the effect other property criteria are having on the rule.

You can select which properties are plotted within the grid view by selecting them from this list on the right-hand side of the Profile Builder.

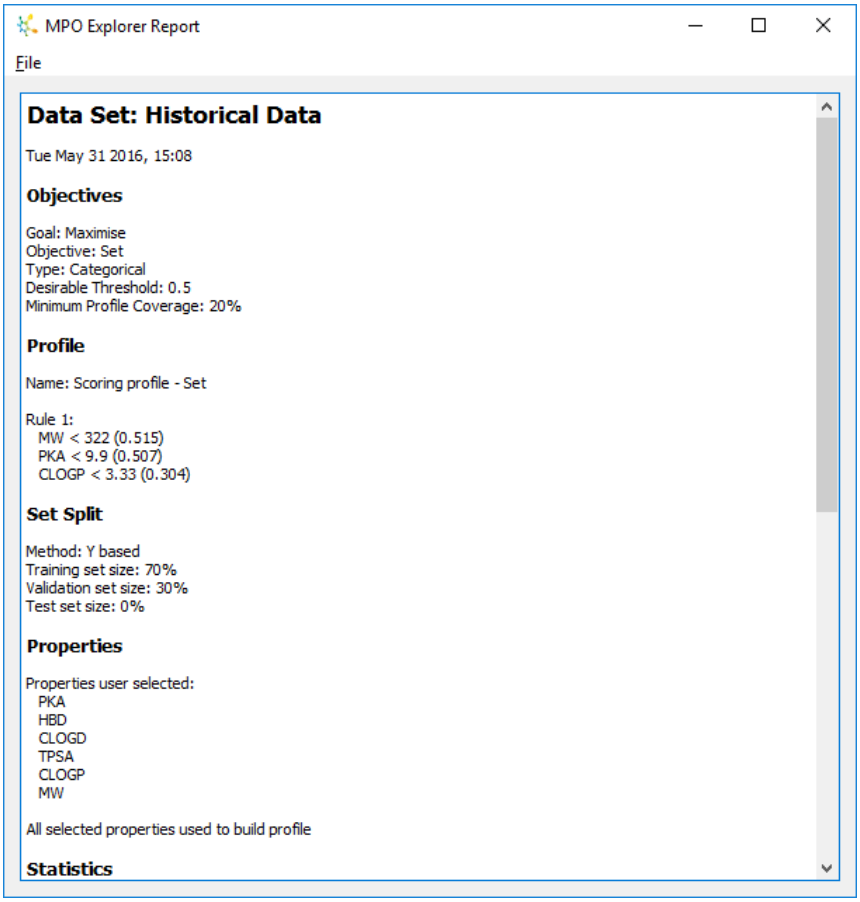
If you instruct the Profile Builder to generate confidence intervals for each continuous property threshold, they will be displayed as shaded regions within the grid view.

Once you are satisfied with the current profile, click the **OK** button to return to the Scoring tab where the new scoring profile will be displayed. Here, you can run, modify, and save the profile just as any other scoring profile within StarDrop.

Alternately, you can click on **Cancel** if you do not wish to use the profile you have generated.

22.3 Profile Builder report

To see a detailed report on the profiles created so far together with the values of the parameters specified in the MPO Explorer wizard, click on the **View Report** button in the main MPO Explorer window to bring up MPO Explorer Report window.

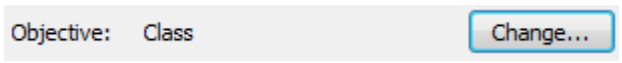


From the **File** menu you can choose to **Save** the report as a PDF file or to **Print** the report.

22.4Profile analysis

In the Scoring tab, you can also use the Profile Builder grid view to analyse generated profiles *post hoc* or to analyse profiles you have created by hand. To invoke this view, click on the **Analyse...** button. The **Analyse...** button will only be enabled when you have a data set open which contains the necessary columns of data for the profile.

The Analysis window is almost identical to the Profile Builder view, except that statistics are now only reported for the current data set and not separately for the training, validation, and test sets. In addition, you can now make changes to any of the rules in the profile.

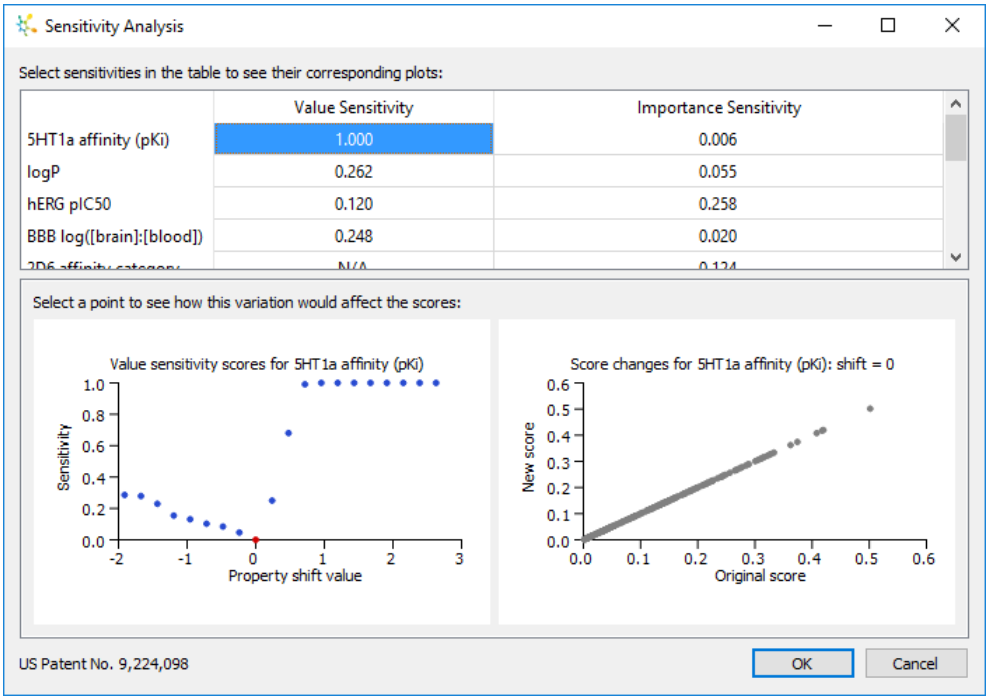


If you wish to see how the profile will perform against different objectives you can click the **Change...** button to choose a different one.

22.5 Sensitivity Analysis tool

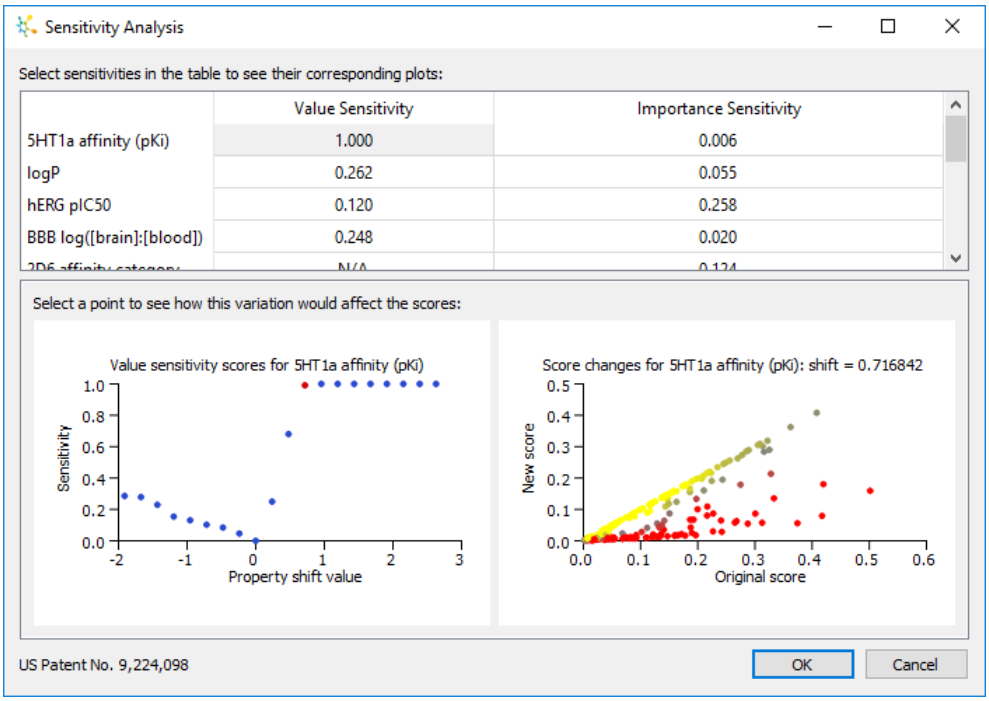
MPO Explorer can also be used to analyse the robustness of your decisions to the selection criteria you have chosen. The Sensitivity Analysis tool enables you to determine if adjusting any of a scoring profile's property criteria or their importances will have a substantial impact on the compounds you might select from a given data set. To achieve this, the tool assumes that the compounds you will be choosing are those with the highest scores.

To analyse the sensitivity of a particular data set to the criteria in a scoring profile, select the scoring profile, open the data set and click the **Sensitivity...** button.



The Sensitivity Analysis dialogue shows each of the criteria from the scoring profile in a table, listed with the most sensitive properties at the top. A *sensitivity score* between 0 and 1 is reported for each 'parameter' in the scoring profile, where a parameter is defined to be either a property's desired value range or its importance. A high sensitivity score means that adjusting the parameter will have a significant effect on the choice of top-scoring compounds from the data set. The sensitivity score also takes into account the effect of uncertainty in the data set. Any sensitivity scores of 0.7 or greater are highlighted in red because these are considered significant.

When a sensitivity score is selected in the table, the graph on the left shows how the sensitivity score varies as the property's original scoring function is adjusted. Each individual point in this graph represents a single scoring function, with the x-coordinate giving the amount that the original criteria or importance has been shifted and the y-coordinate giving the sensitivity score.

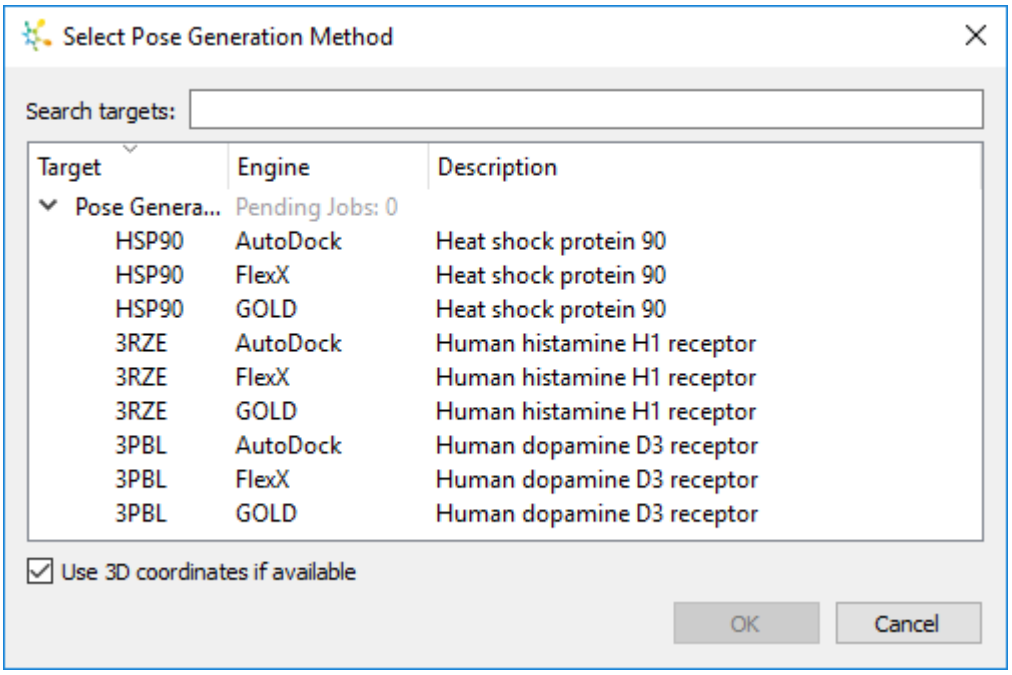


Selecting any point in this graph will bring up a second graph (on the right) with the 'new' scores of every compound in the data set, which relate to the shifted scoring function, plotted against the original scores. These points are coloured according to how much their corresponding compounds' ranks have changed: very red compounds have greatly decreased their rank, very yellow compounds have greatly increased their rank, and greyish compounds have not significantly changed their rank. Clicking on any point in this graph will highlight the corresponding compound in the data set itself. When this plot shows a large number of differences between the original and the 'new' scores it indicates that such a shift in the scoring function would have a significant impact upon the selection of compounds that you might make.

23 How do I... Use Pose Generation?

StarDrop's Pose Generation Interface enables you to send compounds to, and receive results directly from, your docking and alignments tools. Once you have configured access to any Pose Generation server(s) that have been set-up (see section 24.13) you will be able to access the published docking and alignment models.

To use the models, select the rows in your data set for which you would like to generate docking or alignment results and select **Run** from the **Tools**, **Pose Generation** menu.



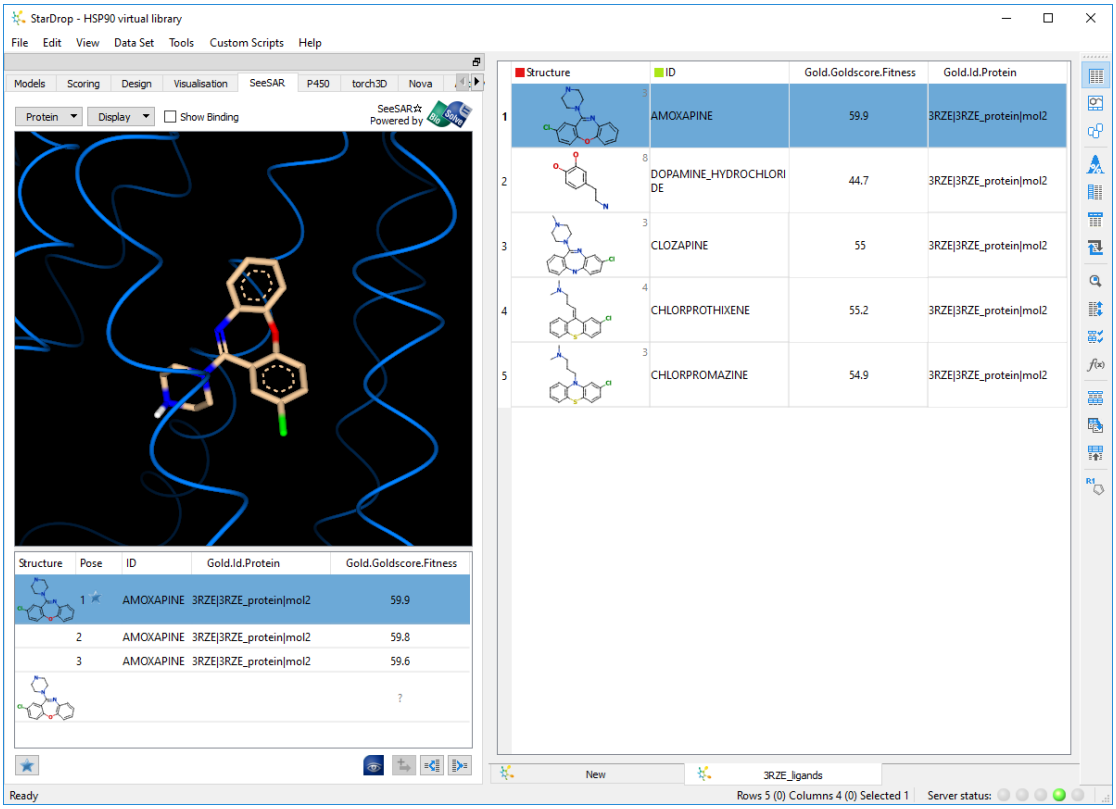
The **Select Pose Generation Method** dialogue enables you to choose a model from the list of those available. Please note that the models for each server to which you have connected will be displayed in separate branches of the tree. To find a model in the list, start typing in the **Search targets** box and the list will be refined as you type to show only those models whose name, engine or description contain the text.

The **Use 3D coordinates if available** option will be the same as you have defined in the Pose Generation Preferences (see section 24.13) and ensures that any compounds you have imported in 3D will be passed with the same coordinates to the Pose Generation server.

To run a model, select it from the list and click **OK**.

While a result is being calculated a * will be shown in the data set next to the compound structure and columns will be added to the data set in which the docking or alignment results will be displayed when the calculation is complete.

When the result has been calculated it will be added to the data set. If you save your data set and close StarDrop while it is calculating, then the result will be collected when you next open the project. The number displayed next to the structure in the data set indicates how many conformations were returned by the model.



If you have access to the SeeSAR module then you can use this to view the different conformations. To view a result, select that row in the data set. If protein information is associated with the result then this will automatically be downloaded into the SeeSAR viewer and displayed with the ligand.

For more details on using the SeeSAR viewer see section 16.

If the docking or alignment fails then the * will disappear and no results will be added to the data set. Information describing any problems that have occurred will be saved in a log file. On Windows this is called C:\Users\<Username>\AppData\Roaming\StarDrop\PoseGenerationInformation.log and on Mac OS this is /Users/<Username>/StarDrop/PoseGenerationInformation.log.

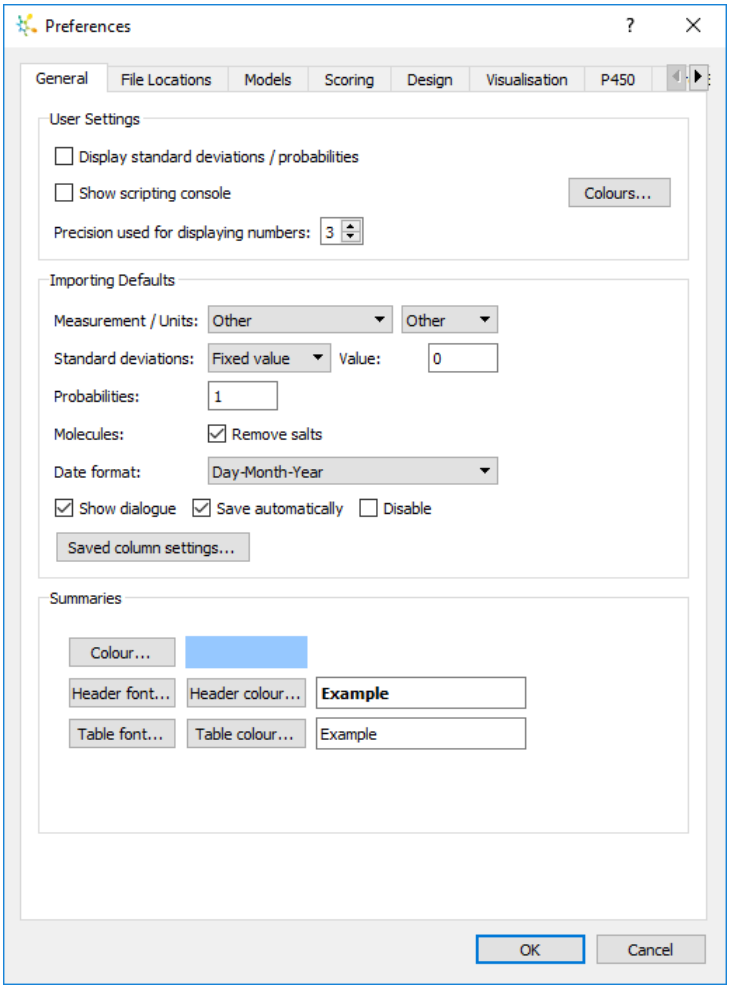
24 Preferences

Most user settings are stored in a StarDrop folder within your home area in a file called **clientconfig.xml**.

There are 13 tabs within **Preferences**, each of which refers to a different section of the application.

24.1 General preferences

The **General** preferences tab enables you to configure a number of details which are generally applicable across StarDrop. In this tab you can set the precision used to display decimal numbers and whether to display standard deviations or probabilities alongside the data. Additionally, you can choose whether to display the scripting console.



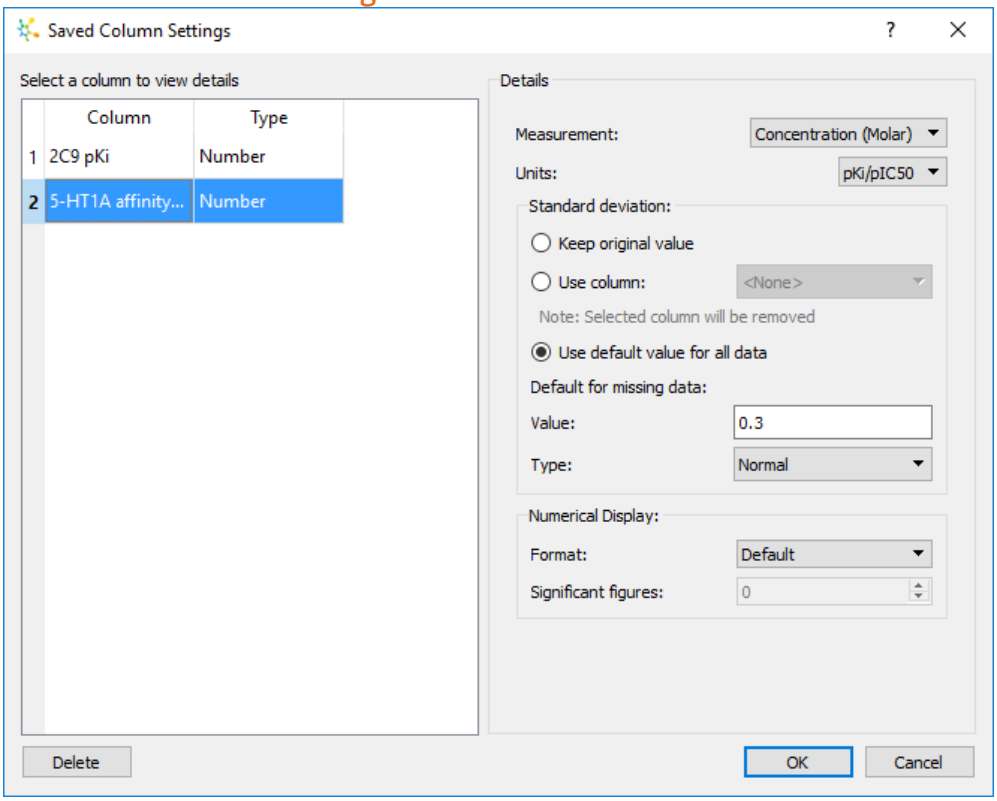
The **Importing Defaults** section enables you to specify any default values you wish StarDrop to use when importing data from non-StarDrop files (e.g. Comma-Separated Variable, Text, SD files). These values will only be used when the equivalent information is not specified in the file.

Clicking the **Colours...** button displays the **Colour settings** dialogue (see section 24.1.2) in which you can define the default colours to use for graphs, Card View, score backgrounds and Glowing Molecules.

When you import data, unless you choose otherwise, StarDrop will save details of any columns of data where you have specified something other than the default. To see the saved details, click the **Saved Column Settings...** button (see section 24.1.1).

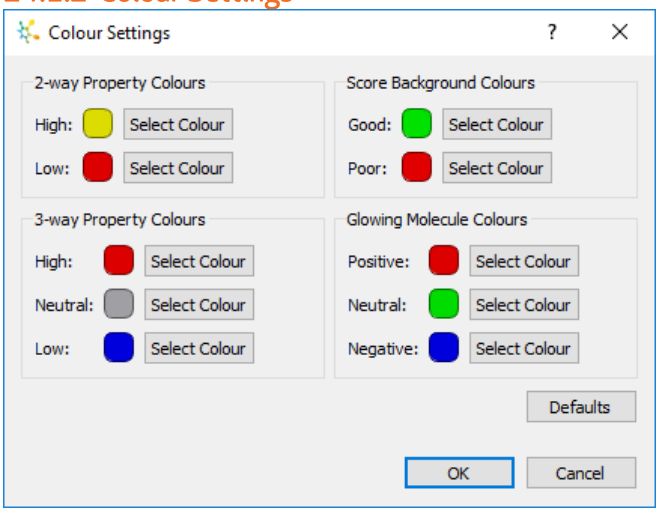
The **Summaries** section enables you to specify the way that summary analyses are displayed (see section 9.4).

24.1.1 Saved Column Settings



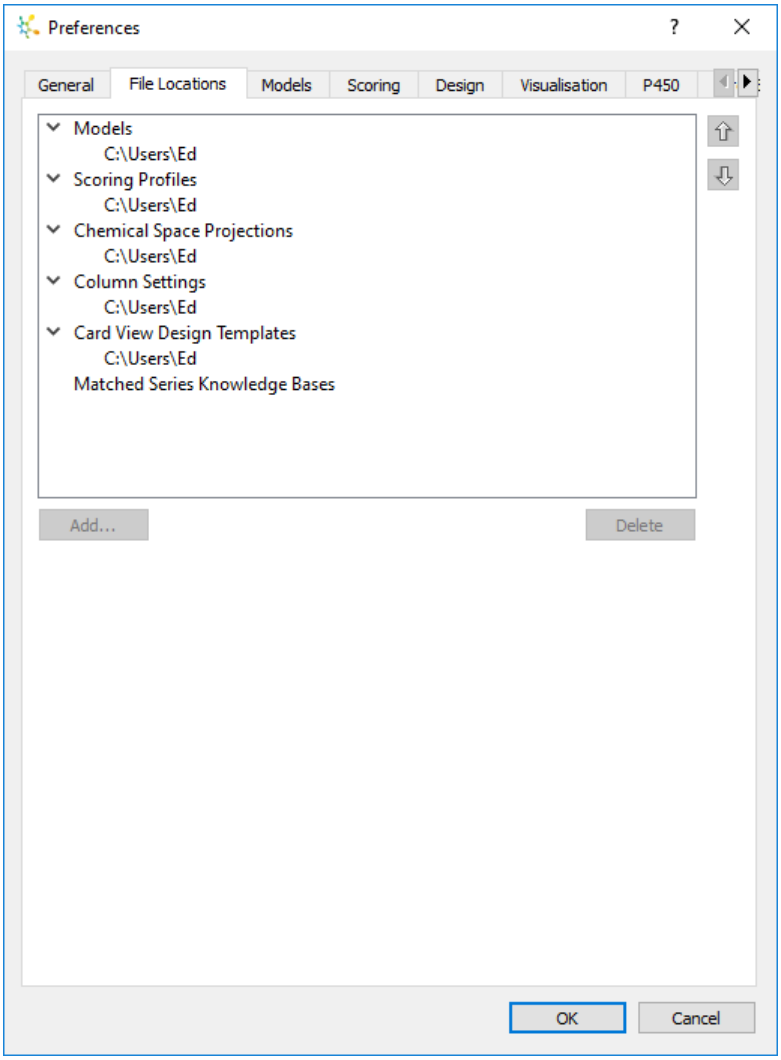
Selecting a column on the left-hand side will enable you to see the details that will be used when a column of the same name (containing the same kind of data) is imported. Click the **Delete** button to delete any selected. These settings are stored in a StarDrop folder within your home area in a file called **ColumnSettings.xml**.

24.1.2 Colour Settings



In this dialogue you can specify a default set of colours for StarDrop to use with its various features by clicking on the colour and choosing from the displayed dialogue.

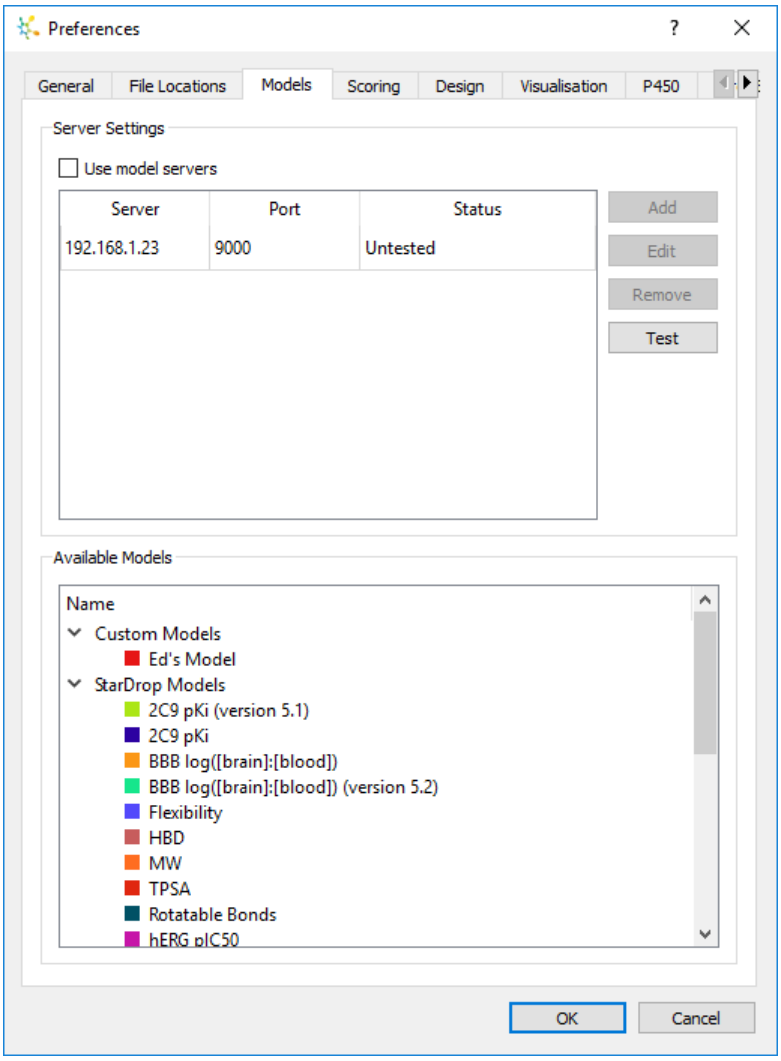
24.2 File locations



Here you can define directories which StarDrop will scan to look for models, scoring profiles, chemical space projections, column settings files, Card View design templates or matched series knowledge bases. Any files of the appropriate type located within these directories will be loaded each time StarDrop is started.

24.3 Models preferences

The **Models** preferences tab is where you can connect to a server to run models:

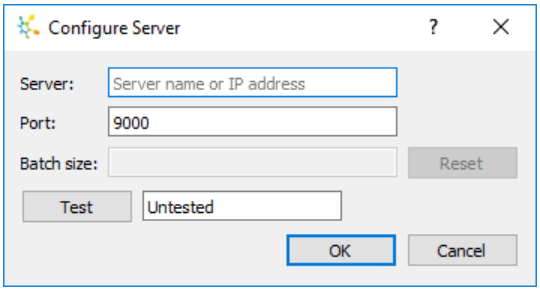


The StarDrop models can be used without access to a server. However, if a StarDrop Model Server is available on the network it will be possible to check the box to **Use model servers** to run models. You can add a server by clicking the **Add** button. The **Edit** and **Remove** buttons will be enabled when you select a server in the list making it possible to modify or remove a server.

Click the **Test** button to confirm that StarDrop is connected to the server. If this is successful, the **Available Models** area will be populated with the names of the models available on the server. StarDrop will always run models from the server if possible.

24.3.1 Adding a Model Server

Type in the name or network address of the model server in the **Server** box and the port in the **Port** box. Click the **Test** button to check the connection. If unsure of these details, contact your network administrator.

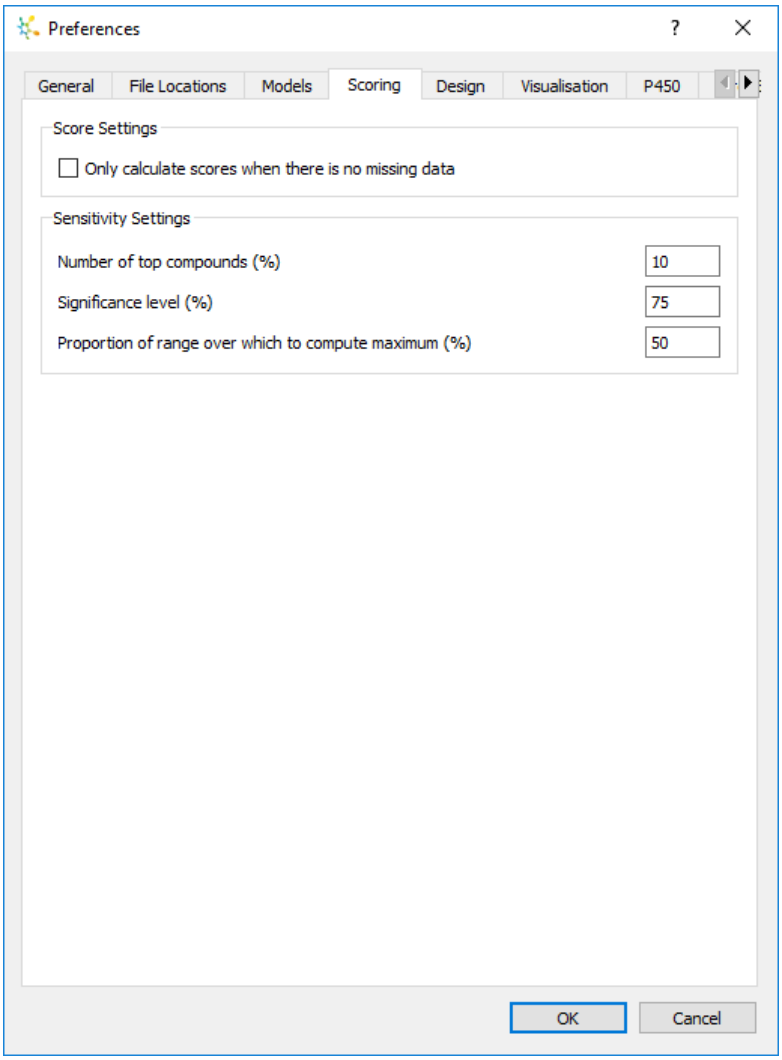


24.4 Scoring preferences

Within the scoring preferences you can indicate whether or not StarDrop should generate scores for compounds when there are missing data.

If the option is ticked then StarDrop will not generate scores when there are data missing. If this option is not ticked then StarDrop will generate scores for all compounds, but, where there are missing data, the score histogram will be faded to give a visual indication that this has happened.

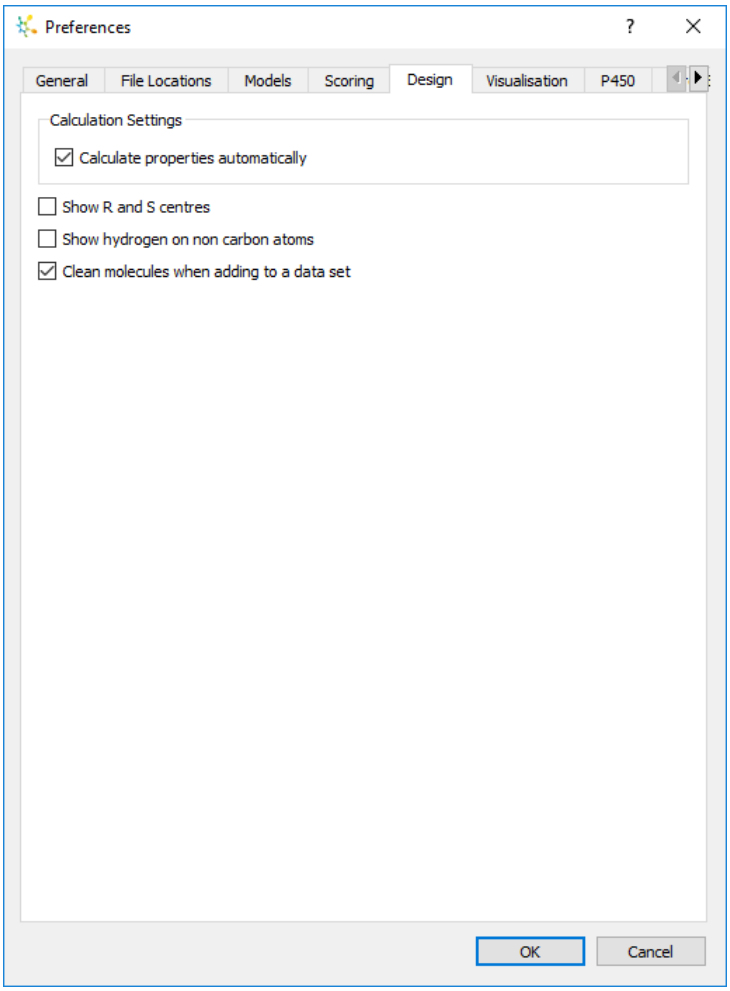
The **Sensitivity Settings** are available for the MPO Explorer module. The **Number of top compounds (%)** indicates the proportion of the data set (the proportion with the highest scores) whose ranks are analysed while determining sensitivity to one of the scoring criterion. The **Significance level (%)** indicates the threshold which must be achieved for there to be sensitivity. The **Proportion of the range over which to computer maximum (%)** indicates a range around a property criterion in which to search for sensitivity. For more detail on these parameters please refer to the StarDrop Reference Guide.



24.5 Design preferences

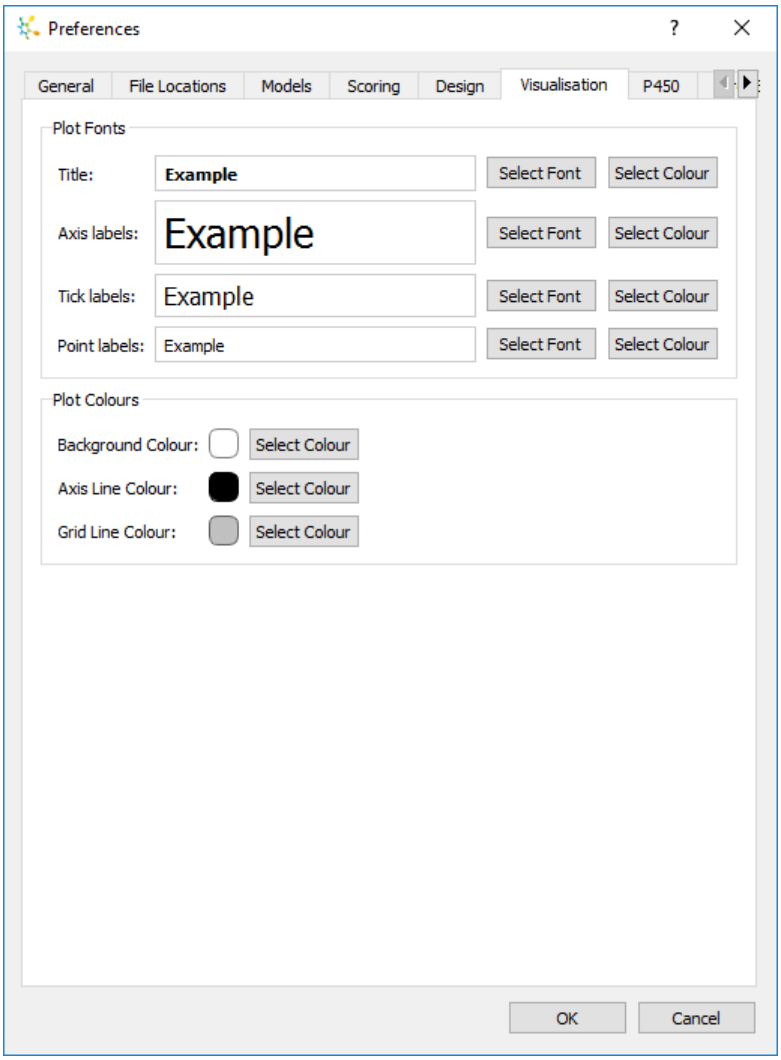
The **Design** preferences tab is where you can choose whether you would like StarDrop to **Calculate properties automatically** as you sketch or modify molecules.

You can also indicate whether you would like to see R and S centres with any enhanced stereochemistry flags or whether hydrogens should be displayed on non-carbon atoms.



24.6 Visualisation preferences

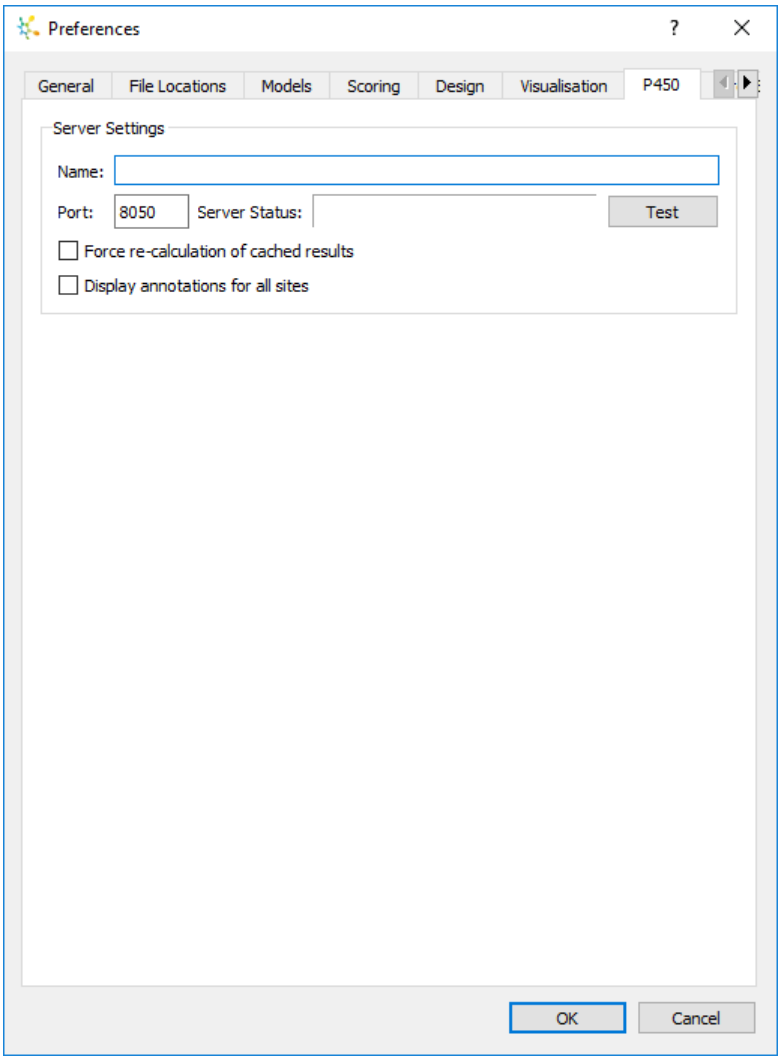
The **Visualisation** preferences enable you to specify default settings for displaying plots.



The colours and fonts specified here will be used each time a new plot is created.

24.7 P450 preferences

To run the StarDrop P450 Models a StarDrop P450 server must be connected.

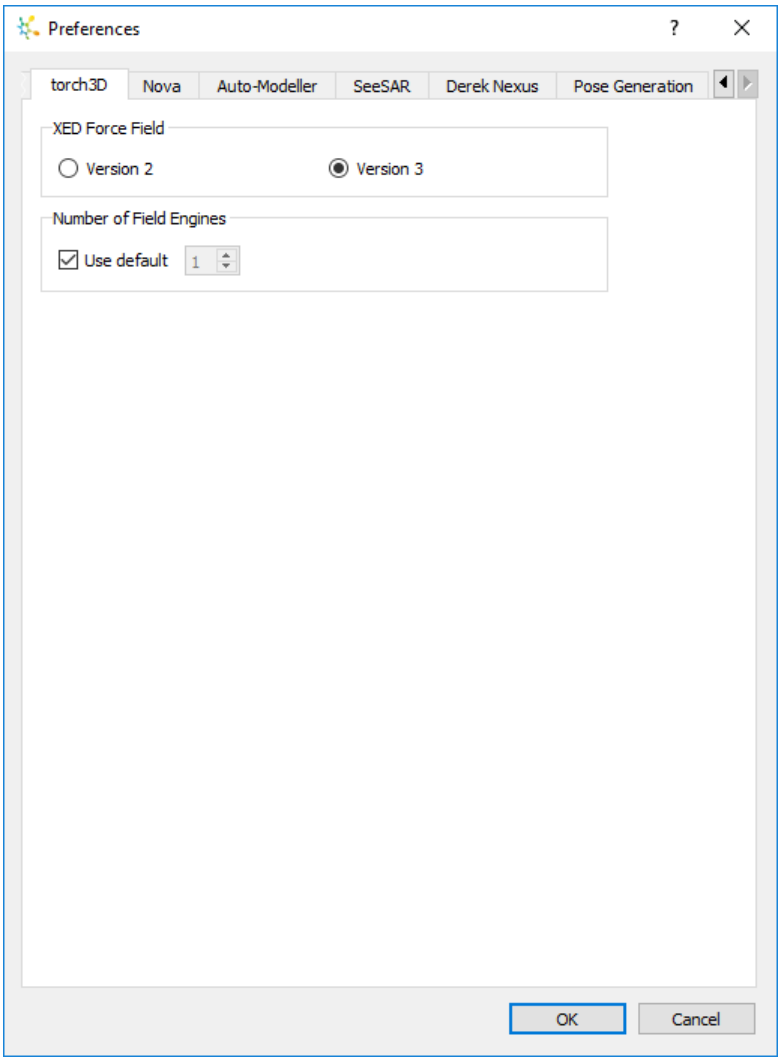


In the section called **Server Settings** type the **Name** and **Port** number of the P450 server. If unsure of these details, contact your network administrator. Click the **Test** button to confirm that StarDrop is connected to the P450 server.

See the **P450** section (see section 18) for use of the other options within this tab.

24.8 torch3D™ preferences

Here you can choose which version of the XED force field to use. You can also control the **Number of Field Engines** that will be used to calculate results if you do not wish all the available CPUs to be in use while torch3D results are being calculated.



24.9 Nova™ preferences

The Nova preferences tab enables you to configure some of the default values which will appear in the wizard when you run Nova (see section 17).

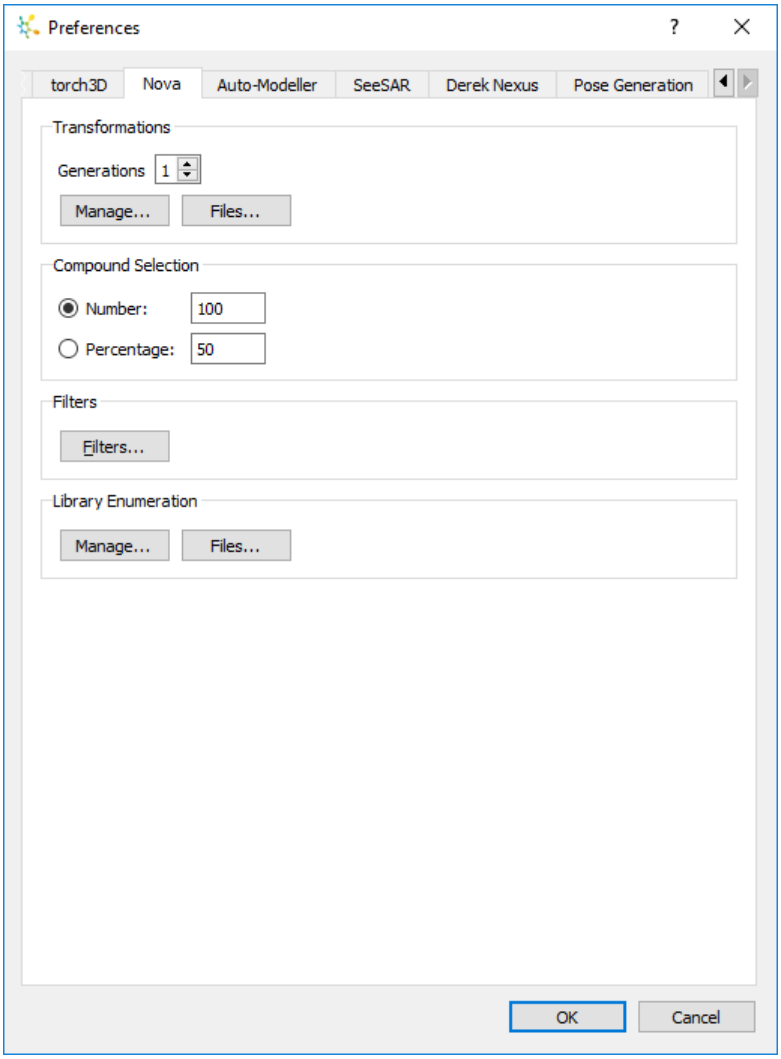
Within the **Transformations** section, clicking the **Manage**… button opens the Transformation Manager dialogue (see section 24.9.1) enabling you to manage the transformations which StarDrop will use when running Nova. Clicking the **Files…** button opens the Select Files dialogue enabling you to specify other files of transformations (see section 24.9.3) that you wish StarDrop to include when running Nova.

The **Compound Selection** offers you the opportunity to specify whether, when making selections during the Nova process, you wish to specify the number of results to return as a fixed **Number** or as a **Percentage**, and for each, a default value.

Clicking the **Filters...** button opens a dialogue enabling you to manage the filters which you can apply to your compounds (see section 24.9.4).

Within the **Library Enumeration** section, clicking the **Manage**… button opens the Fragment Library dialogue (see section 24.9.5) enabling you to manage the fragments which StarDrop will use when enumerating libraries. Clicking the **Files…** button opens the Select Files dialogue enabling you to specify other files of fragments (see section 24.9.6) that you wish StarDrop to include when enumerating libraries.

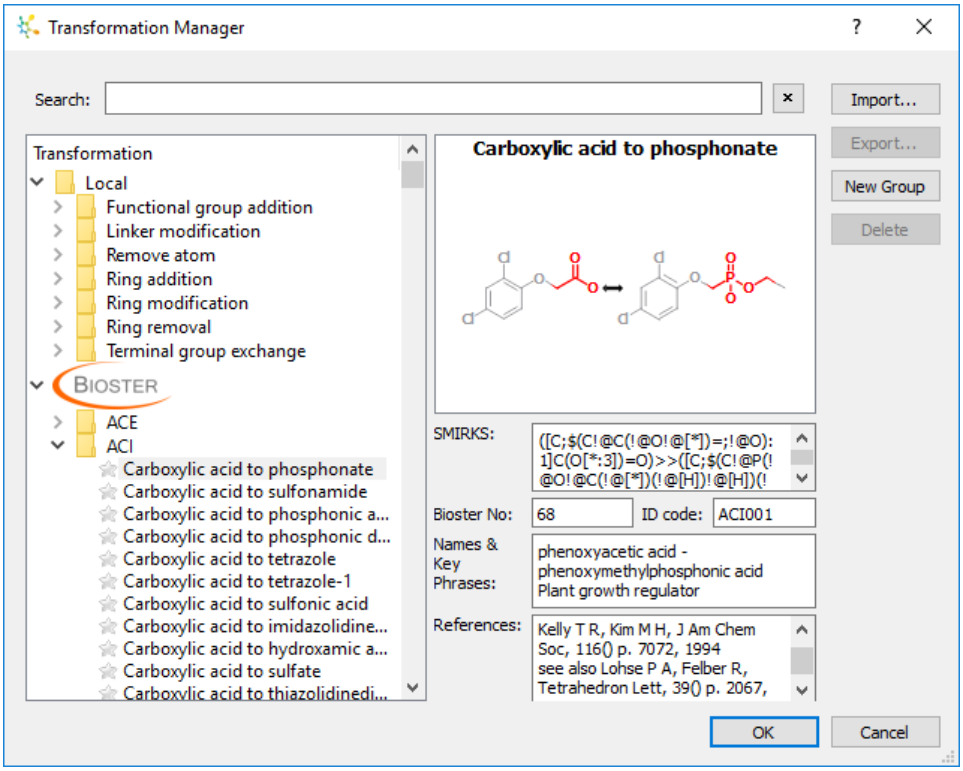
Note: Regardless of which transformations or fragments are imported and available, you can choose the exact set that will be used each time you run the Nova wizard.



24.9.1 Transformation Manager

The Transformation Manager enables you to import, export and organise sets of transformations. Clicking the star will toggle the 'favourite' flag. Transformations that are marked as favourites will be available for selection in their own group when you run Nova.

Typing in the **Search** bar will automatically reduce the displayed list of available transformations to those which contain the search term within the name or list of key phrases.



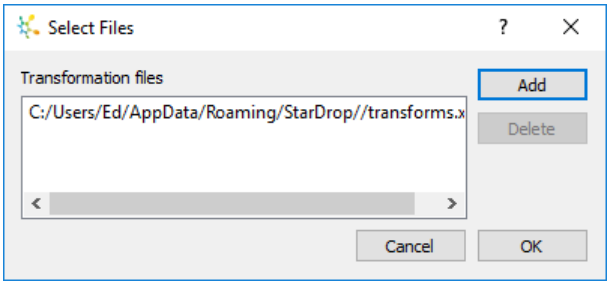
24.9.2 Importing transformations

To import a set of transformations, click the **Import...** button. This will enable you to browse for a file containing your own transformations, encoded as SMIRKS patterns. The name of the file will be used to provide the name of the new group created in the tree. The file must be a text file and each line should contain either two or three tab delimited entries. The first entry must be the SMIRKS pattern and the second entry should be a name of this pattern (this will be displayed in the tree). The third entry is an optional reference for this pattern.

Example:

```
[C:1][CH2][C:2]>>[C:1]O[C:2]      Secondary carbon to ether  Reference1
[C:1][CH2][C:2]>>[C:1]N[C:2]      Secondary carbon to amine  Reference2
[C:1][CH2][C;!R:2]>>[C:1][C;!R:2] Remove secondary carbon
```

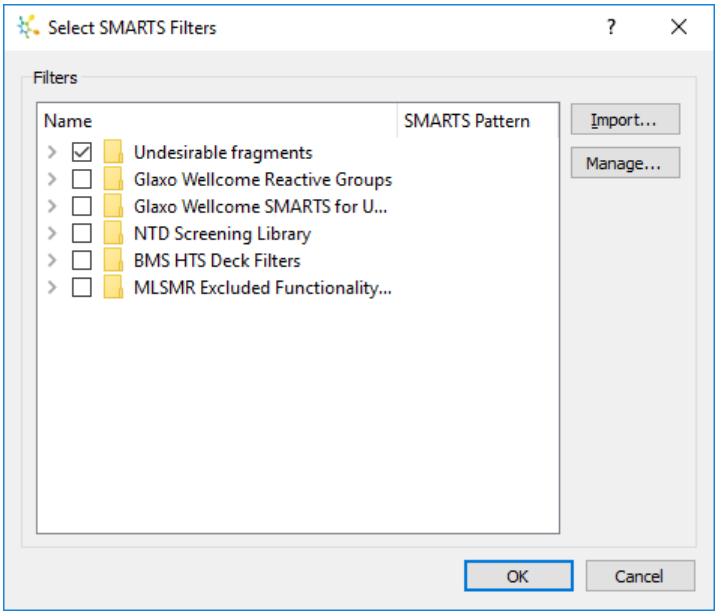
24.9.3 Select Transformation Files



This enables you to access a file of transformations created by another user in StarDrop in addition to your own local set. Therefore if you create a set of transformations within StarDrop and copy the resulting xml file to a shared folder, it can be used by other members of the team. Your local set of transformations is stored in a file called **transforms.xml** in the StarDrop folder in your home area.

24.9.4 Filters

The **Select Smarts Filters** dialogue enables you to choose which filters are selected by default when you run Nova. You can select whole groups or individual filters from within any of the groups.



Clicking the **Import...** button enables you to import your own filters to add to the list. To define your own you must create a text file containing SMARTS and their associated names. The SMARTS patterns must not contain any spaces and there should be a space to separate the pattern from the name, with one pattern and name on each line of the file.

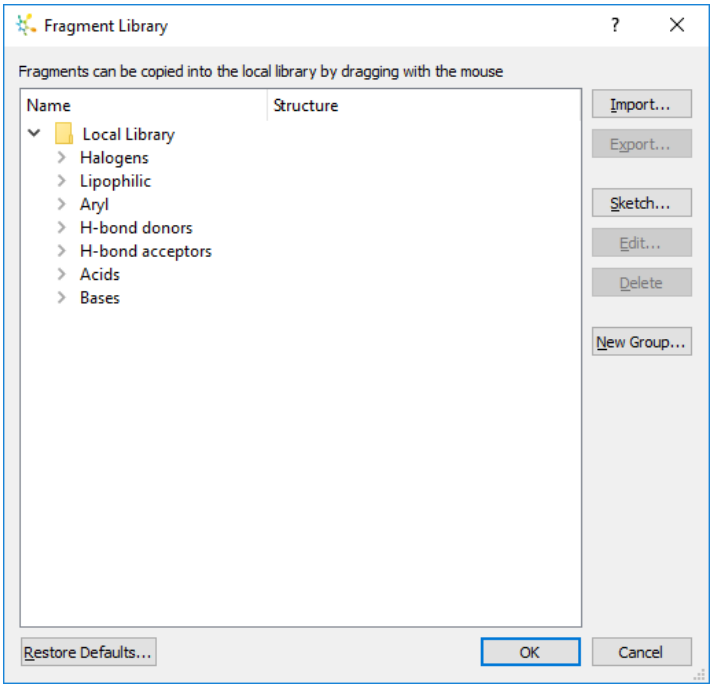
Example:

```
[S,C](=[O,S])[F,Br,Cl,I] acid halide
[Cl]C([C&R0])=N chloramidine
[P,S][F,Cl,Br,I] P/S halide
```

Clicking the **Manage...** button enables you to choose which collections are available. StarDrop uses SMARTS patterns in different ways so it is not always useful or appropriate to use all the available patterns as filters.

24.9.5 Fragment Library

The **Fragment Library** enables you to view and edit fragments.



Click the **Import...** button to import your own fragments in SD file format. The SD file format is designed to define a complete structure, whereas a fragment is specifically a sub-structure that can be connected. As such, appropriate data elements must be provided to indicate which atoms and bonds describe the actual fragment that will be used and which atoms and bonds are placeholders for the attachment point.

The <FragmentAtomIds> tag must be used to contain a semi-colon delimited list of indices of the atoms in the fragment.

The <FragmentBondIds> tag must be used to contain a semi-colon delimited list of indices of the bonds in the fragment.

The <CollectionName> tag is optional. Where used it provides the name of the group in which this fragment will be displayed within the StarDrop client. Where this is not provided, the filename will be used as the name of the group.

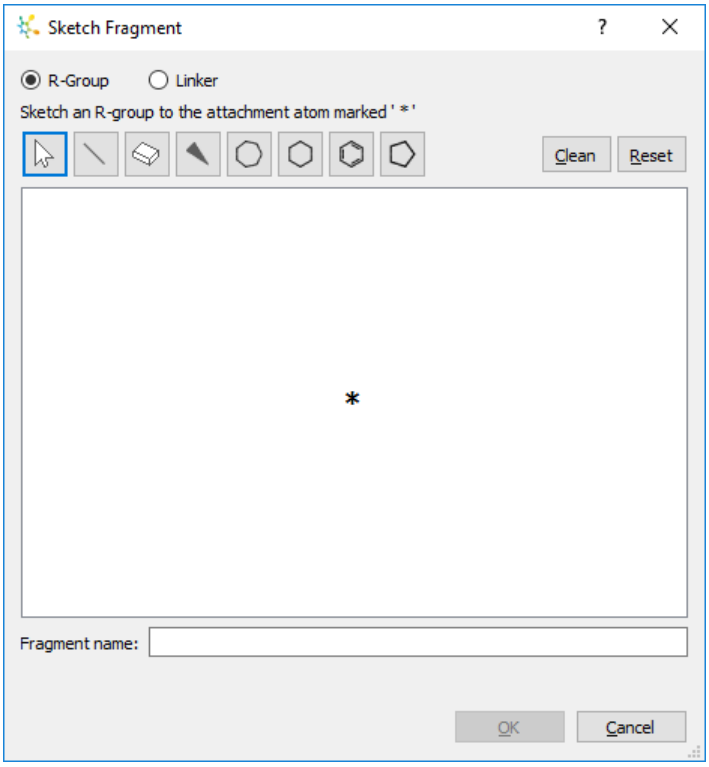
Example:

```
trifluoro
StarDropFragmentManager_1

  8 7 0 0 0 0 0 0 0 0999 V2000
     3.732 0.5 0 C 0 0 0 0 0 0 0 0 0 0 0 0
     2.866 0 0 C 0 0 0 0 0 0 0 0 0 0 0 0
     2.366 0.866 0 F 0 0 0 0 0 0 0 0 0 0 0 0
         2 -0.5 0 F 0 0 0 0 0 0 0 0 0 0 0 0
     3.366 -0.866 0 F 0 0 0 0 0 0 0 0 0 0 0 0
     4.042 -0.03694 0 H 0 0 0 0 0 0 0 0 0 0 0 0
     4.269 0.81 0 H 0 0 0 0 0 0 0 0 0 0 0 0
     3.422 1.037 0 H 0 0 0 0 0 0 0 0 0 0 0 0
  1 2 1 0 0 0 0
  2 3 1 0 0 0 0
  2 4 1 0 0 0 0
  2 5 1 0 0 0 0
  1 6 1 0 0 0 0
  1 7 1 0 0 0 0
  1 8 1 0 0 0 0
M END
> <FragmentAtomIds>
2;3;4;5;
>
> <FragmentBondIds>
2;3;4;
>
> <CollectionName>
Halogens
>
$$$$
```

To export a set of fragments click the **Export...** button and choose a file name. The exported fragments will be in the SD format defined above.

Click the **Sketch...** button to create a new fragment by drawing it. The **Sketch Fragment** dialogue will be displayed.



If the fragment is an **R-Group**, just one attachment point * will be displayed onto which you can draw the fragment (see description of editor tools in section 6.1). If you choose to draw a **Linker** then two attachment points will be available which you must join when drawing the linker. The **Fragment name** you provide will be displayed in the list of fragments.

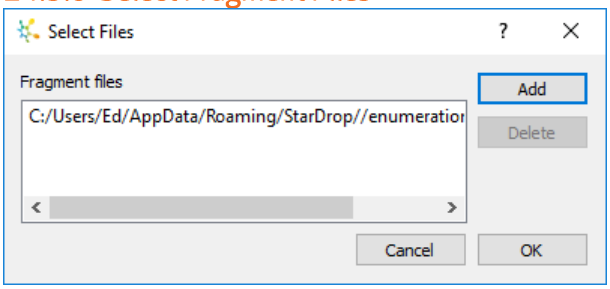
Click the **Edit...** button to edit a fragment. The **Sketch Fragment** dialogue, described above, will be displayed.

Click the **Delete...** button to delete a fragment or group.

Click the **New Group...** button to add a new group to the list.

Clicking the **Restore Defaults...** button enables you to reset the list of fragments back to the original set provided with StarDrop.

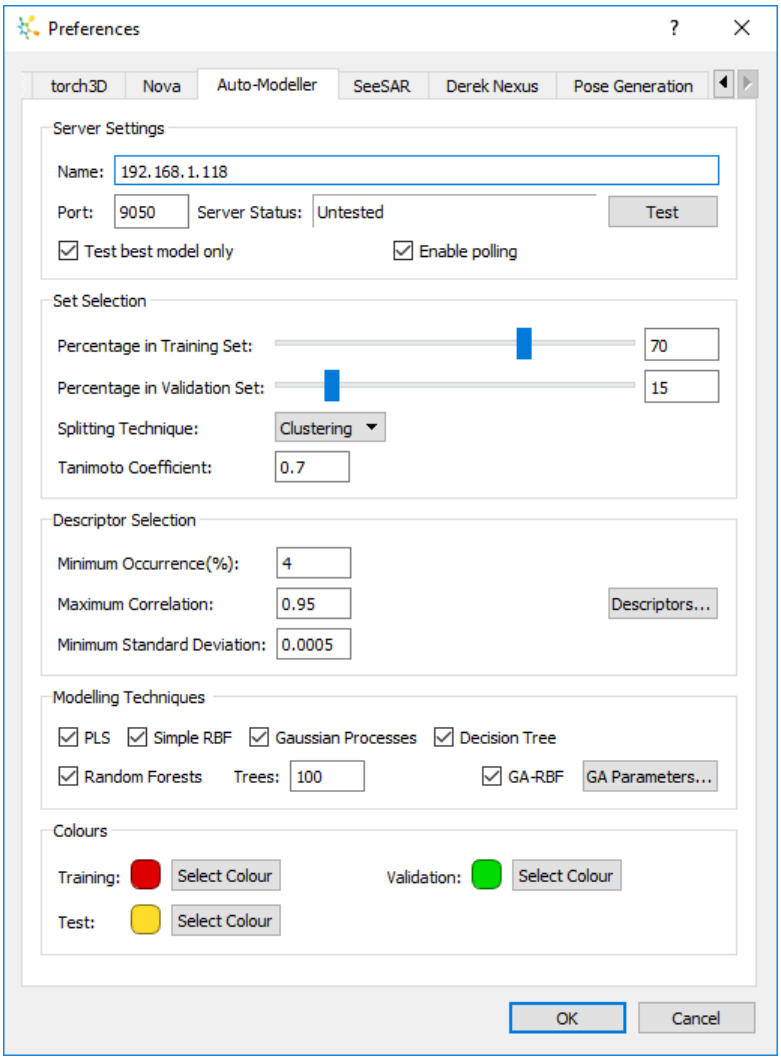
24.9.6 Select Fragment Files



This enables you to access a file of fragments created by another user in StarDrop in addition to your own local set. Therefore if you create a set of fragments within StarDrop and copy the resulting SD file to a shared folder, it can be used by other members of the team. Your local set of fragments is stored in a file called **enumerationfragments.sdf** in the StarDrop folder in your home area.

24.10 Auto-Modeller™ preferences

To run the StarDrop Auto-Modeller and generate new models the StarDrop Automatic Model Generation (AMG) Server must be connected.

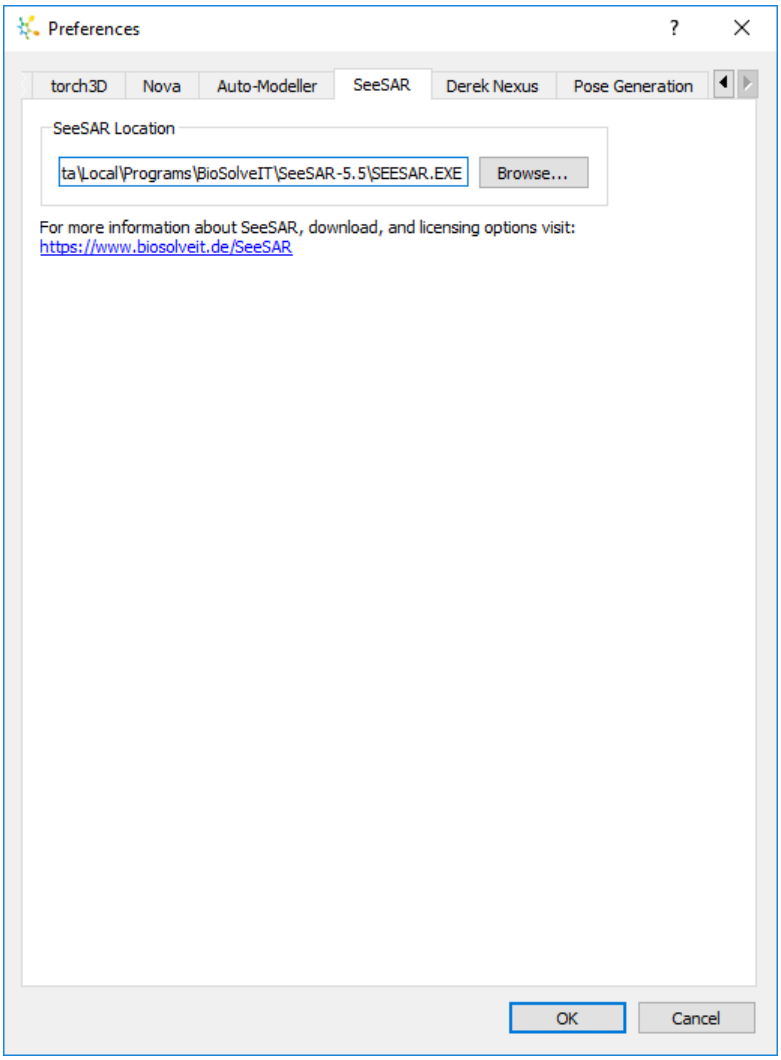


In the section called **Server Settings** type the **Name** and **Port** number of the server of the AMG server. If unsure of these details contact your network administrator. You can click the **Test** button to confirm that StarDrop is connected to the AMG server.

See the **Auto-Modeller** section (section 18) for use of the other options within this tab.

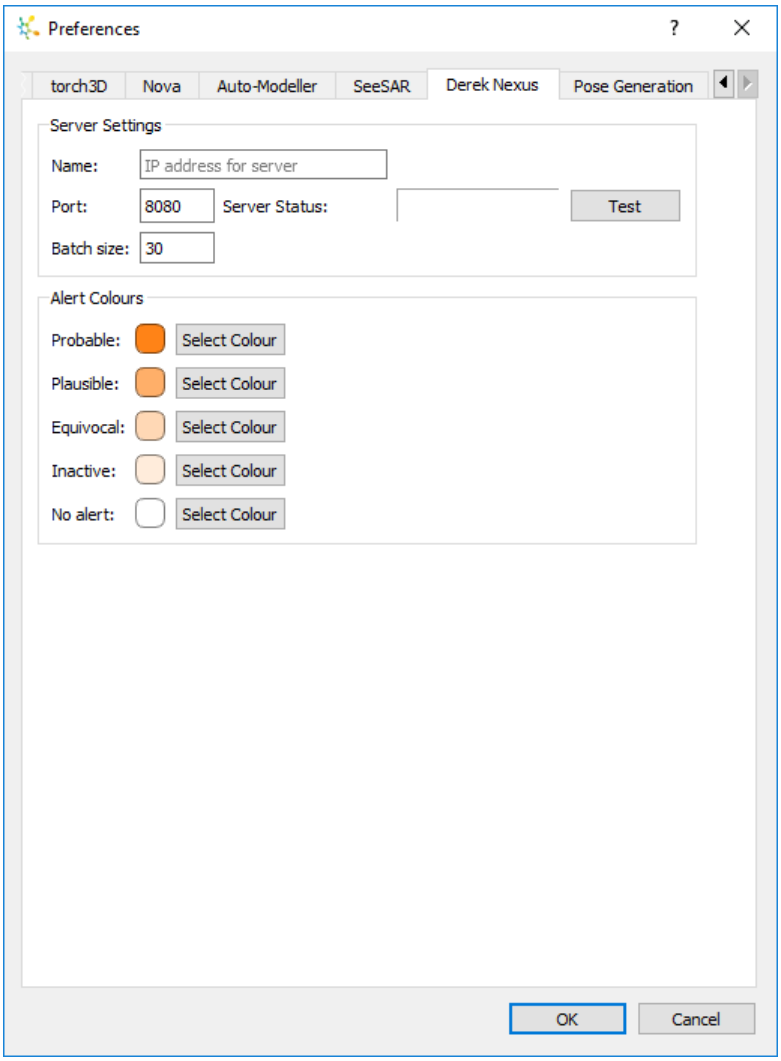
24.11 SeeSAR preferences

Here you can specify the location of you SeeSAR installation if you wish to transfer ligands and proteins directly into the full SeeSAR application.



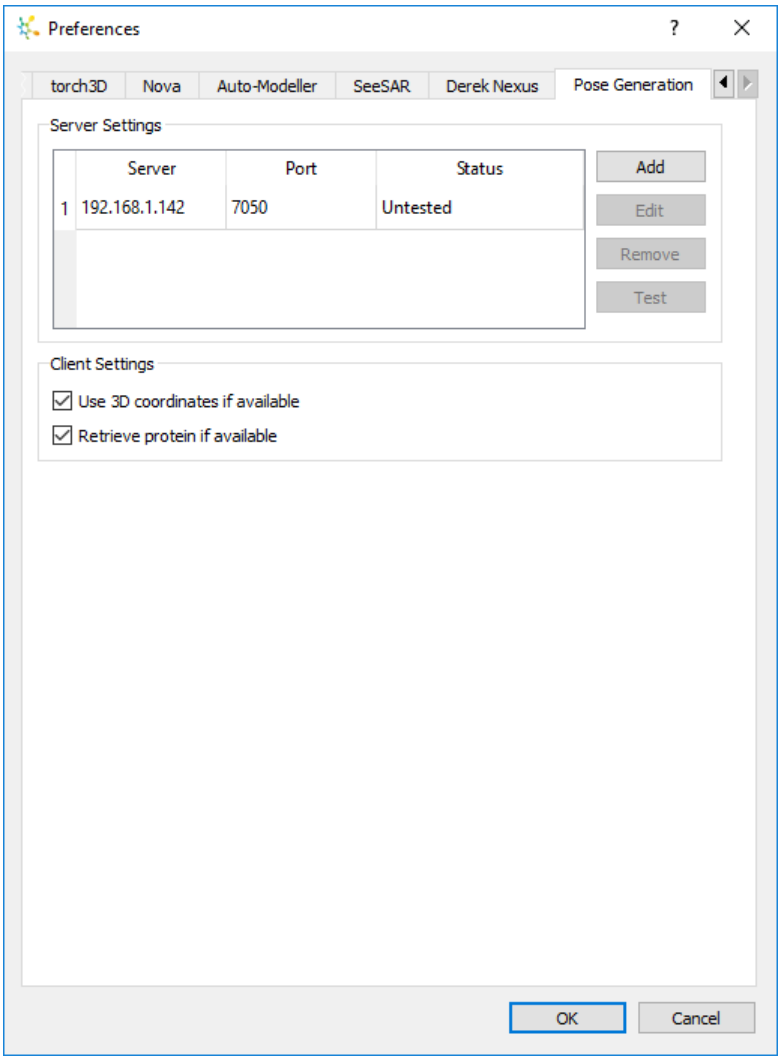
24.12 Derek Nexus™ preferences

Here you can indicate the **Name** and **Port** of your Derek Nexus server. Click the **Test** button to make sure that you are able to connect. You can also specify the colours to use to highlight possible toxicities.

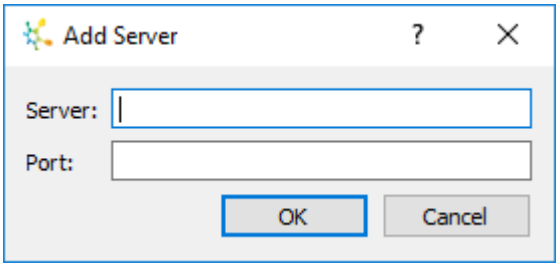


24.13 Pose Generation preferences

Here you provide details of can Pose Generation servers that you have configured for running 3D docking or alignment software.



Click the **Add** button to display the **Add Server** dialogue where you specify the name or IP address of the server, as well as the port number that has been configured.



You can also **Edit** the server details, **Remove** the server from the list or **Test** the connection.

The **Use 3D coordinates if available** option ensures that any compounds you have imported in 3D will be passed with the same coordinates to the Pose Generation server. The **Retrieve protein if available** option ensures that proteins are automatically downloaded from the pose generation server.