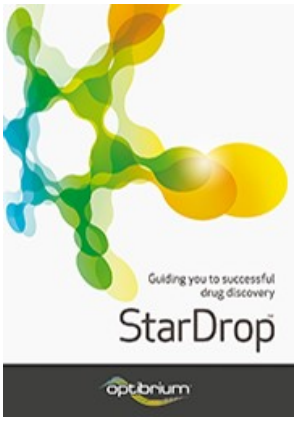


Worked Example:

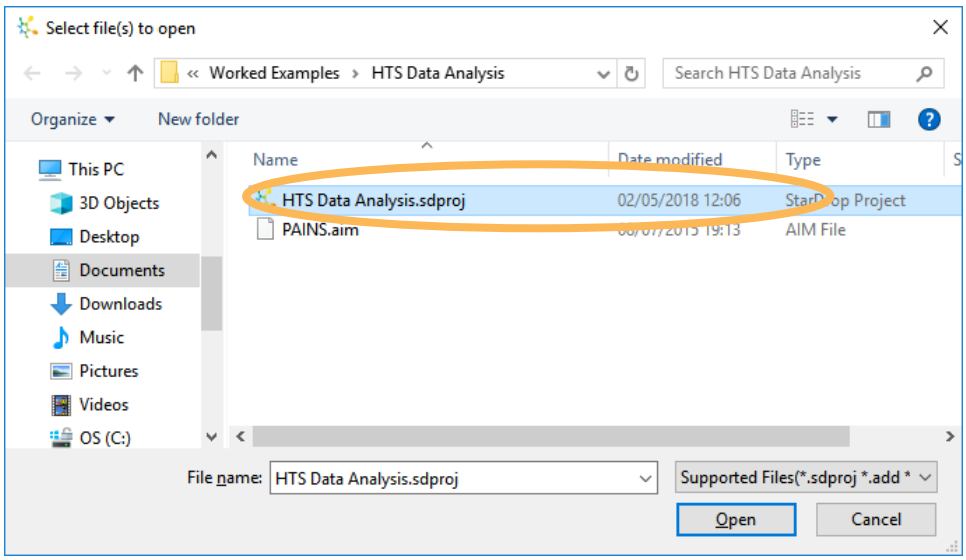
Analysis of High-throughput Screening Data

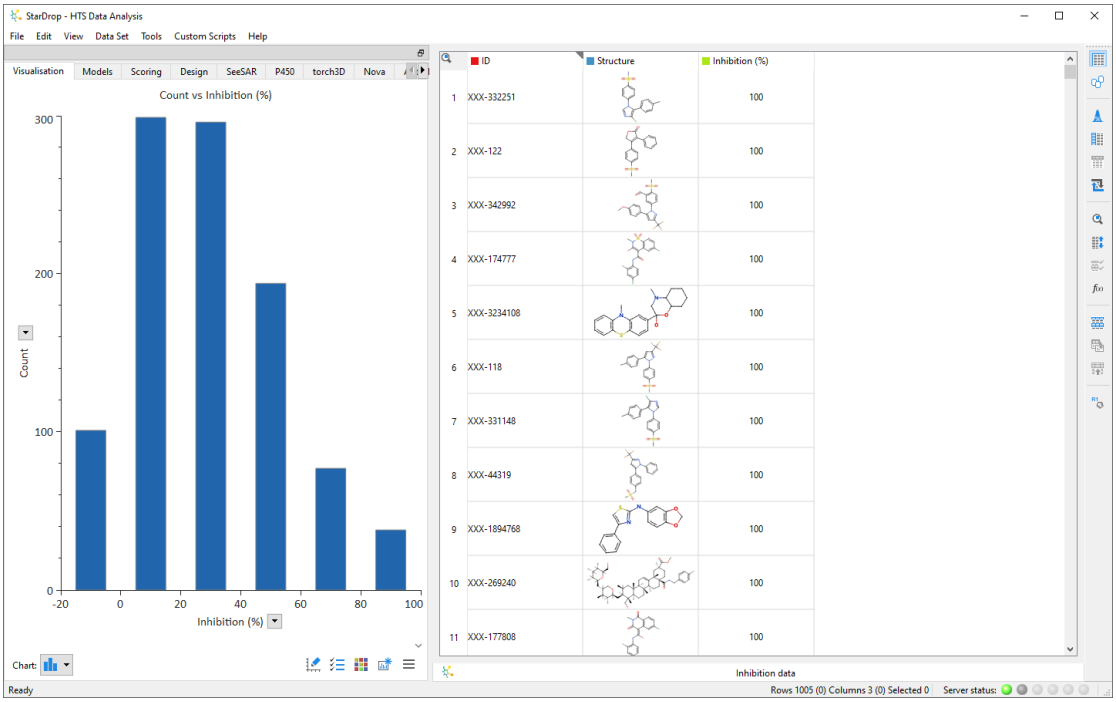
The following example uses a publicly available set of over 1000 compounds with data from a COX-2 inhibition screen. The example demonstrates different ways to analyse inhibition data from an HTS campaign to identify high-quality chemotypes for optimisation. Whilst many HTS campaigns provide much larger data sets than this, the principles demonstrated can be applied to much larger data sets.

If you have any questions, please feel free to contact stardrop-support@optibrium.com.

Exercise

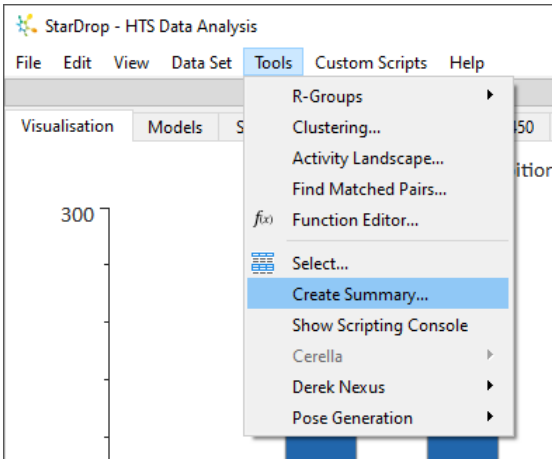
- In StarDrop, open the file **HTS Data Analysis.sdproj** by selecting **Open** from the **File** menu.





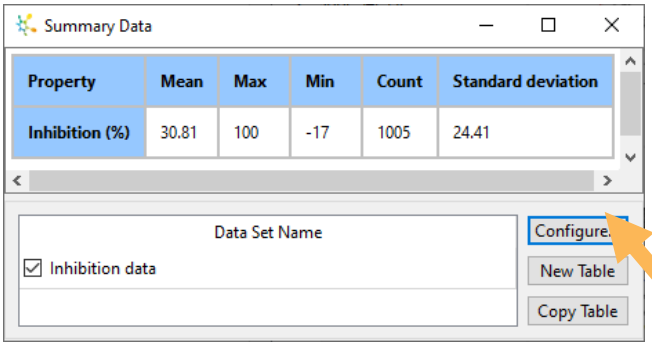
The project contains a data set of just over 1000 compounds with % inhibition data against COX-2 and a histogram showing their distribution.

To generate a more quantitative summary of these data, we're going to create a summary table.



- Select **Create Summary** from the **Tools** menu.

This will create a table of summary statistics for each property in the data set. You can configure the columns, statistics, classifiers and display of the table by clicking on the **Configure** button underneath the table (for more information you can watch a quick Summary Analysis video which gives some hints and tips).



In this case, the statistics we require are included in the table. The average inhibition measured is approximately 31%, and the standard deviation is approximately 24%. A reasonable cut-off for selection of hits might be

two standard deviations above the mean, which is approximately 80% inhibition for this set.

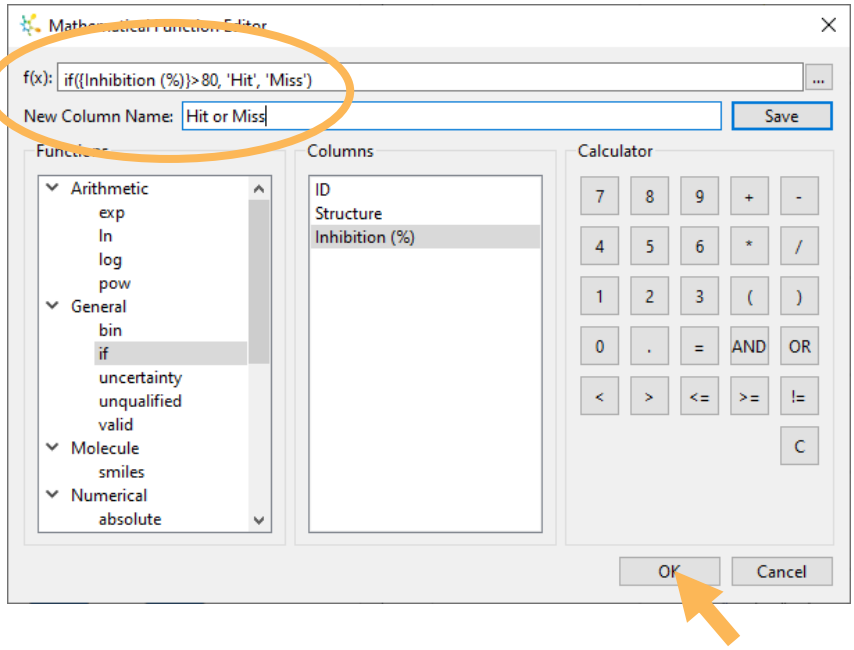
Hint: You can close or minimise the summary table to save space on the screen. We won't use it again in this exercise.

Using this, we can create a new column classifying compounds as either a 'Hit' or a 'Miss'.

- Click the **Function Editor** button $f(x)$ on the right-hand toolbar (you can also open the Function Editor by selecting **Function Editor** from the **Tools** menu).
- In the Function Editor, enter "Hit or Miss" as the **New Column Name** and type or paste in the following equation:

  ```
  if({Inhibition (%)}>80, 'Hit', 'Miss')
  ```

  and click the **OK** button.

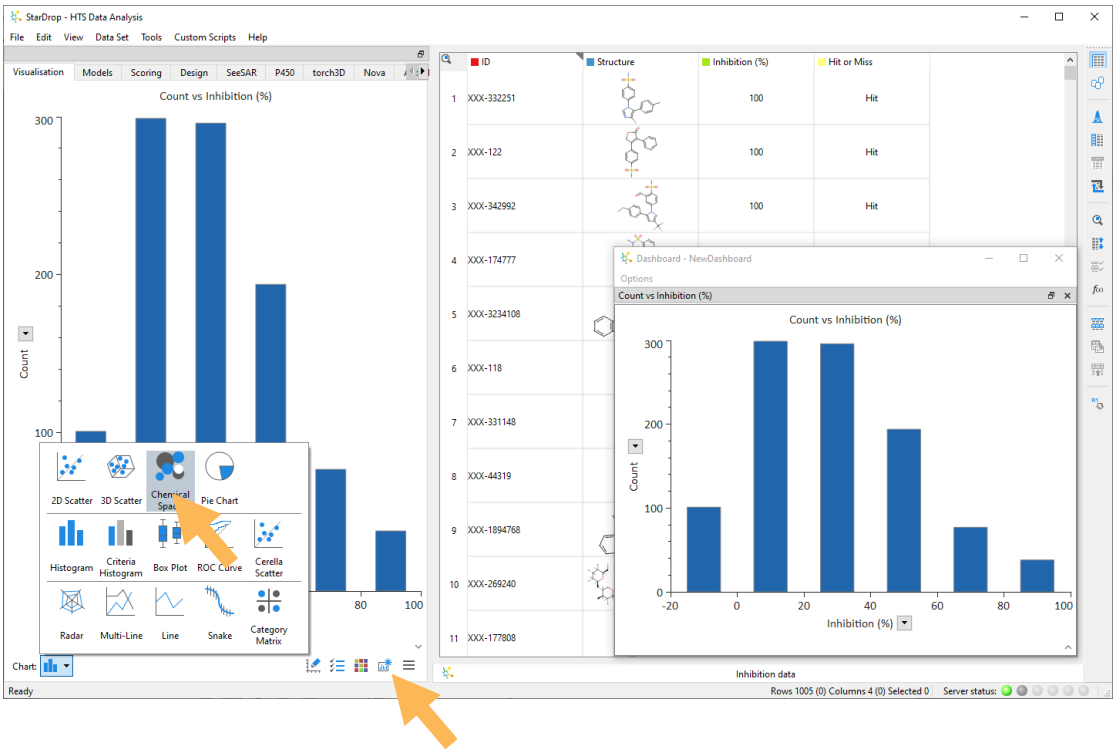


Note: you can enter most of the equation by selecting functions and columns from the lists below, but ensure that you put single quotes around 'Hit' and 'Miss' in the equation to specify these as categories rather than text.

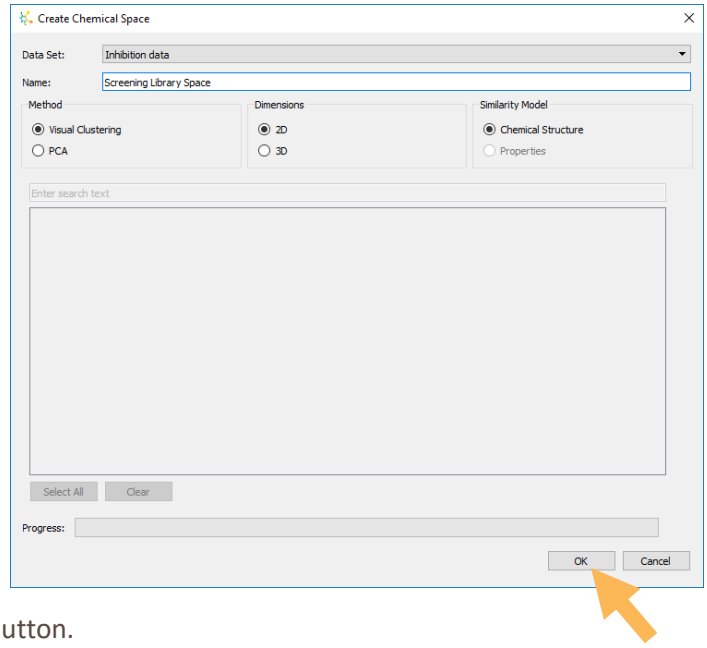
A new column called **Hit or Miss** will be added to the data set indicating those compounds that meet the cut-off of 80% that we are considering. Now we'll explore the distribution of these hits across the chemical diversity of the screening library.

- First, click the **Detach** button at the bottom of the Visualisation area to add the histogram to a new dashboard.

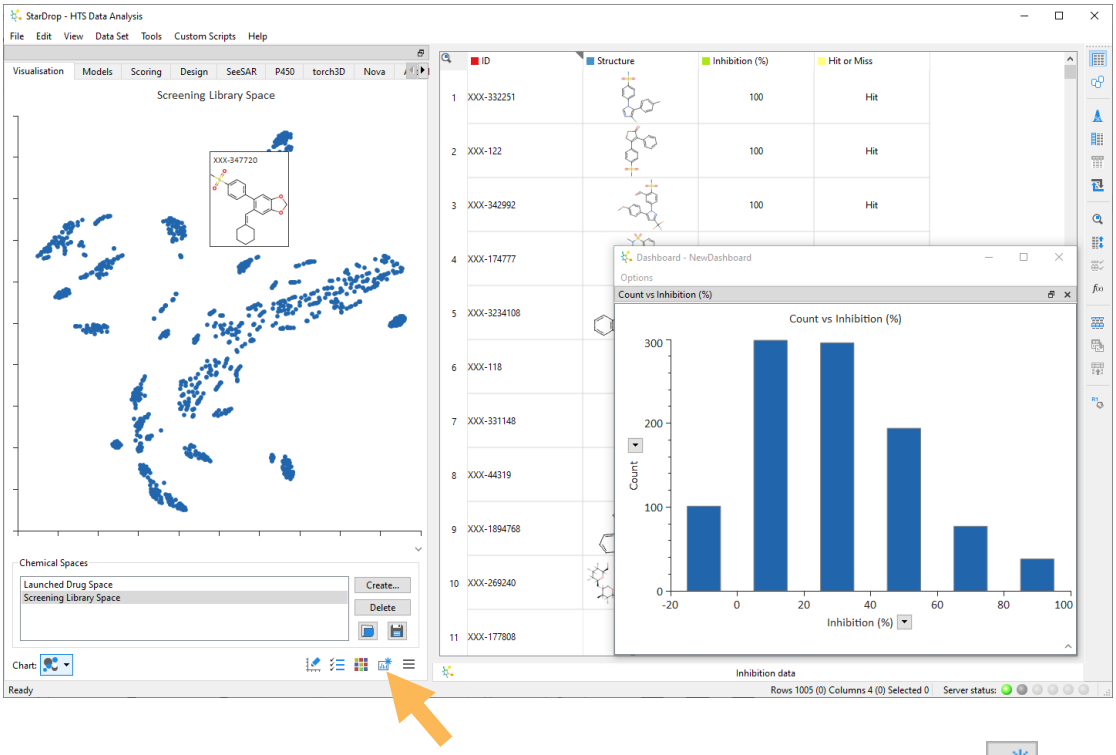
This enables us to keep it available when we create a new chemical space visualisation.



- Select **Chemical Space**  from the **Chart** menu and click the **Create** button to generate a new chemical space representing the library.
- In the **Create Chemical Space** dialogue, give the new projection a name (e.g. "Screening Library Space") and click the **OK** button.

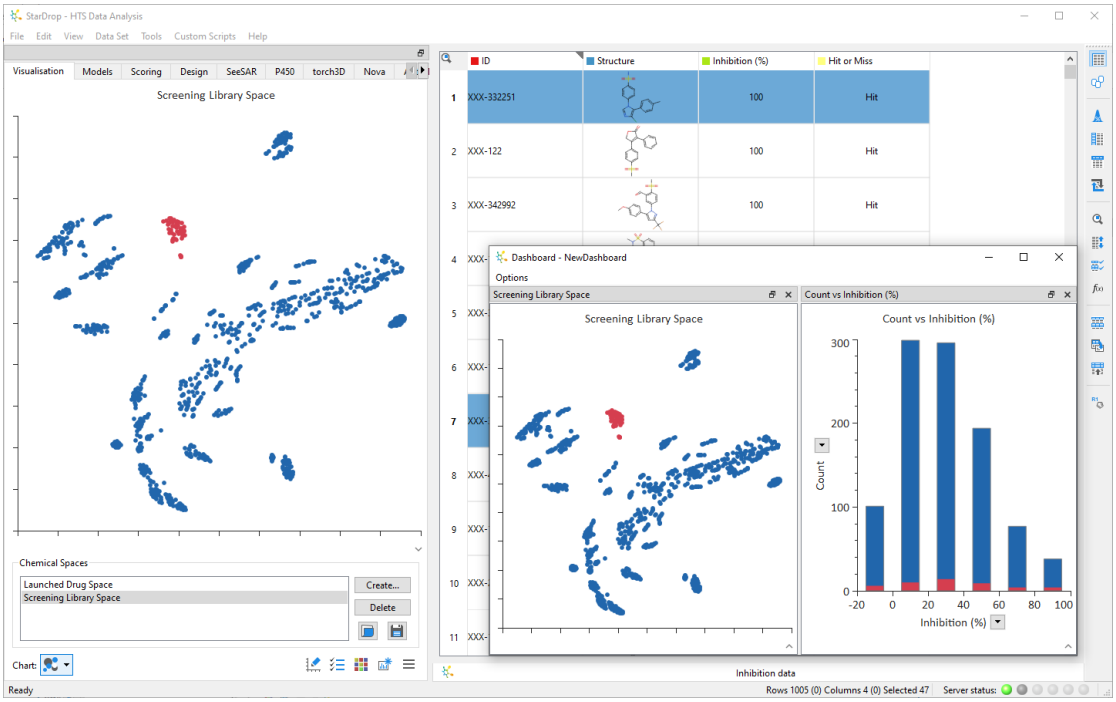


It may take about 30 seconds to generate a chemical space illustrating the diversity of the screening library. In this visualisation, each point represents a compound and structurally similar compounds are clustered together. You can see the structure corresponding to a point by hovering the mouse over it.



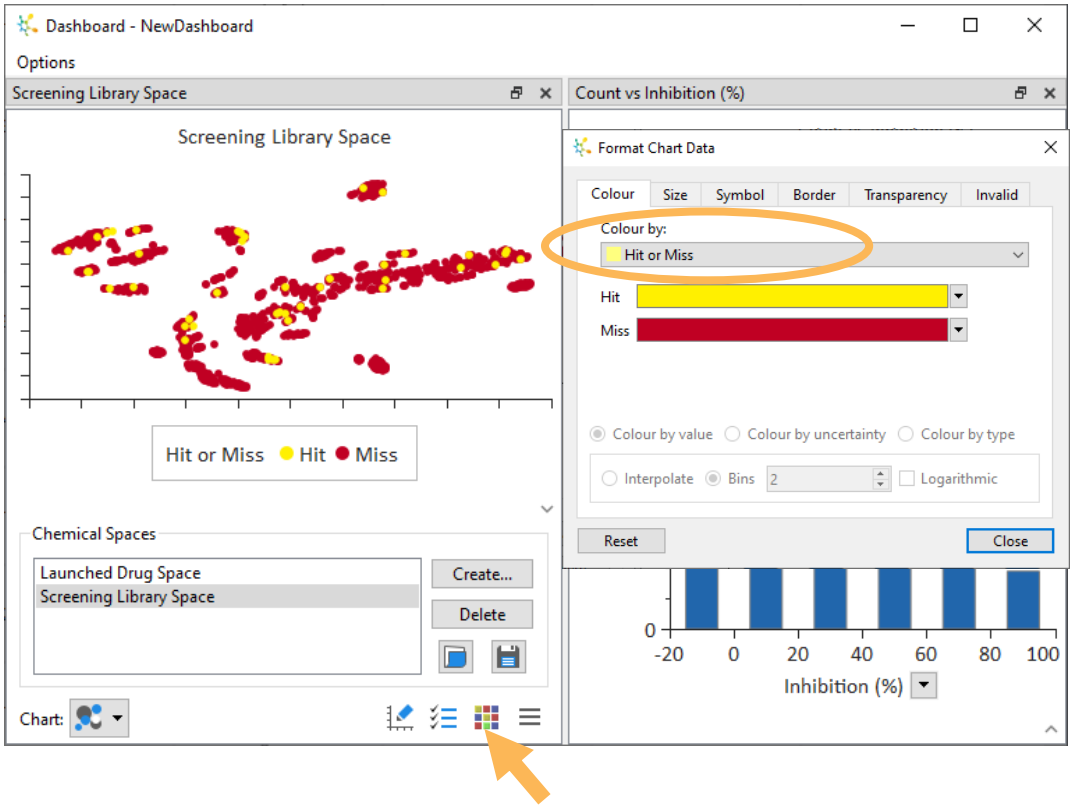
- Add the chemical space to the dashboard by clicking on the **Detach** button  in the bottom-right of the **Visualisation** area.

You can add as many charts as you like to a dashboard and drag the charts and the spacers between them to choose how you want to see them. **Note:** Making a selection in any of your charts, or the data set, enables you to see those compounds selected everywhere else.



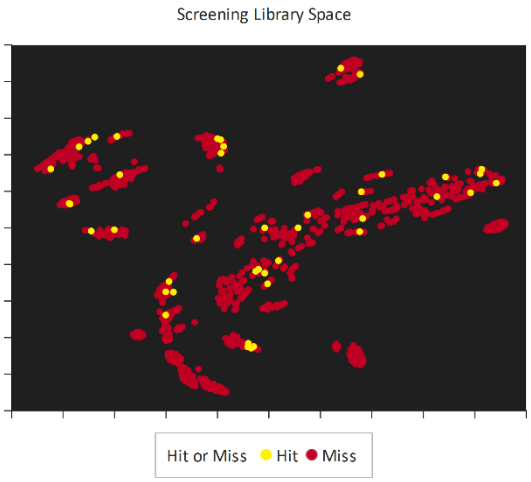
We can colour the points in the chemical space to highlight the hits.

- In the dashboard, click the arrow at the bottom of the chemical space to display the controls and click on the **Format** button  .



- In the **Format by Property** dialogue, choose **Hit or Miss** from the **Colour by** list. In this example, we have coloured the “Hits” yellow and the ”Misses” red.
- Right-click on the chart itself and select **Change Background** from the menu to set an alternative background colour (we have chosen dark grey).

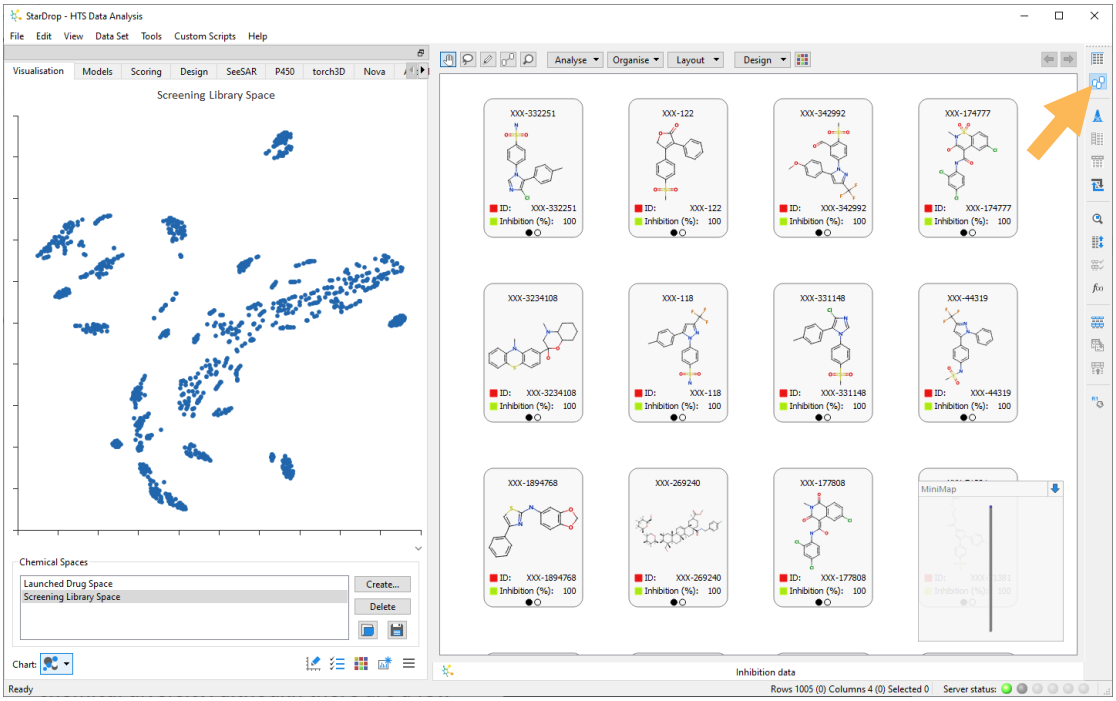
With the hits highlighted we can see that they are distributed over a wide range of chemical diversity, although there are a few clusters of compounds which contain no hits.



Another goal in the analysis of HTS data is to identify a hit series with good structure-activity relationships (SAR); this can give us confidence that the results are genuine and not the result of assay interference or impurities. Consistent SAR may also indicate opportunities for further optimisation, so next, we’re going to explore the activity landscape around potential hit series.

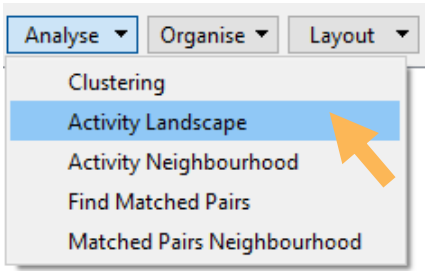
Note: You can minimise the dashboard to give more room for other visualisations, but don't close it because we'll return to it again later.

- To help us visualise this, change to StarDrop's **Card View™** by clicking on the **Card View** button on the right-hand toolbar.



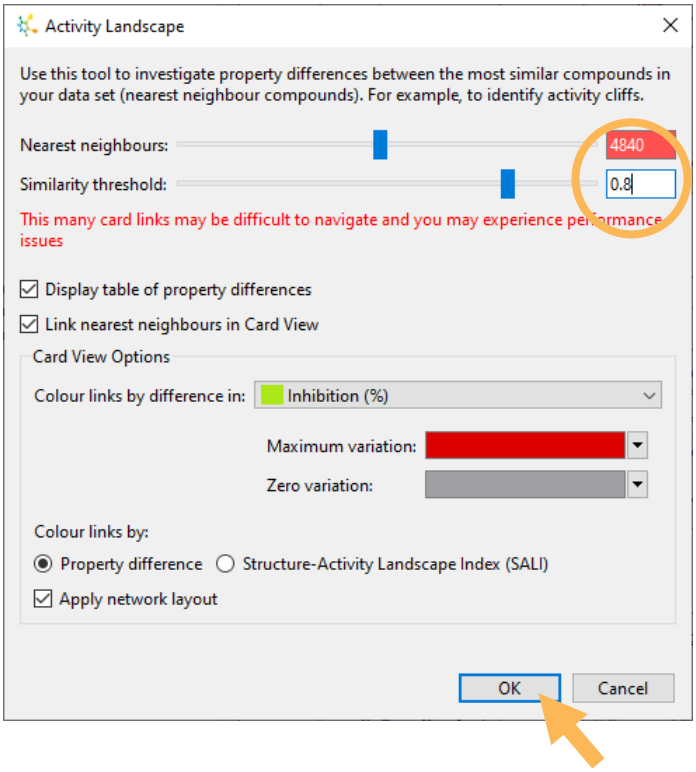
In Card View, each compound is represented by a card and, by default, these show the compound structure and the first few properties from the data set; in this case the Identifier, Inhibition (%) and whether it is a Hit or Miss. You can change the data shown on a card using the **Design** menu at the top of Card View. For more information on Card view, please watch the series of short videos starting with Getting Started in Card View.

- From the **Analyse** menu at the top of Card View, select **Activity Landscape**.



- In the **Activity Landscape** dialogue, set the **Similarity threshold** to **0.8**. This defines the threshold above which pairs of compounds are considered similar in the analysis.

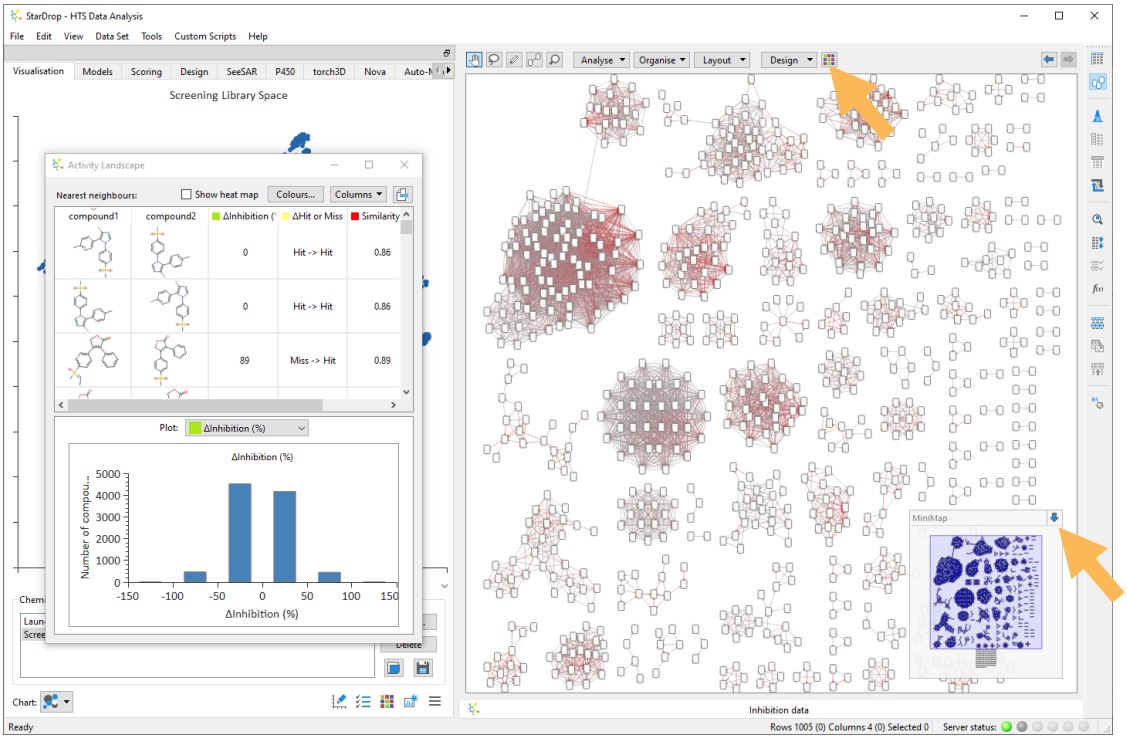
  Note: you will see a warning about the number of card links that will be generated, but you can safely ignore it in this case.



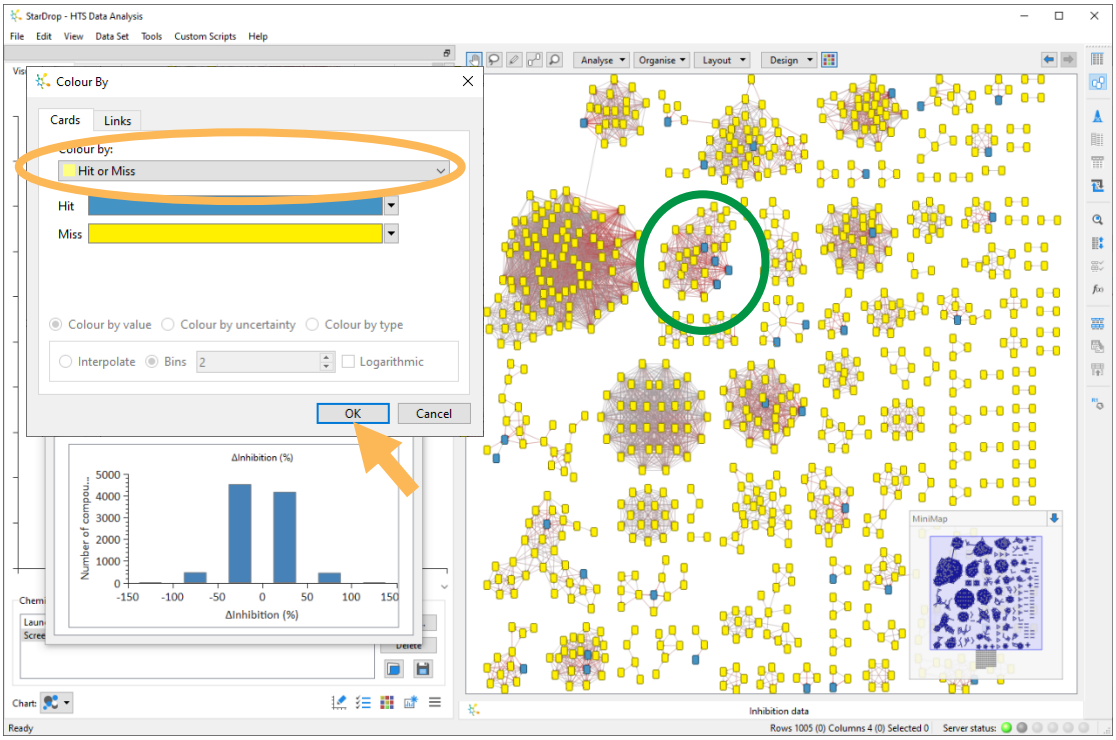
- We don't need to change any other options, so click the **OK** button to run the analysis.

The result is a number of networks that each represent a 'neighbourhood' of similar compounds. Compounds with a link have a similarity greater than the threshold (0.8), the arrow on the link indicates the direction in which activity increases and the colour indicates the size of the increase from red (high) to grey (zero), A table of nearest neighbours is also shown.

Note: You may wish to hide the mini-map which is in front of some of the network sections by clicking the arrow in its top-right corner.



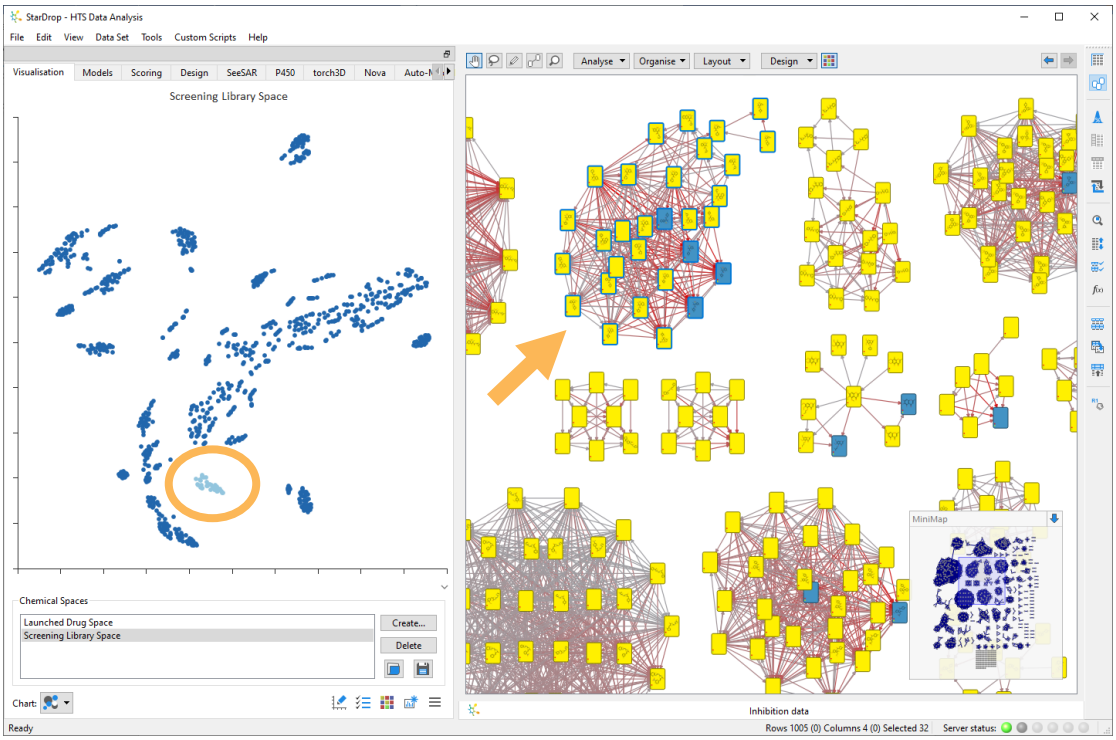
- Colour the cards using the categories from the **Hit or Miss** column by clicking the **Format** button  at the top of Card View and selecting **Hit or Miss** from the **Colour by** list. You can choose any colour scale you like, but here we have set the hits to be blue and the misses yellow. Click the **OK** button.



We are going to focus on one section of the network, highlighted by the green circle above.

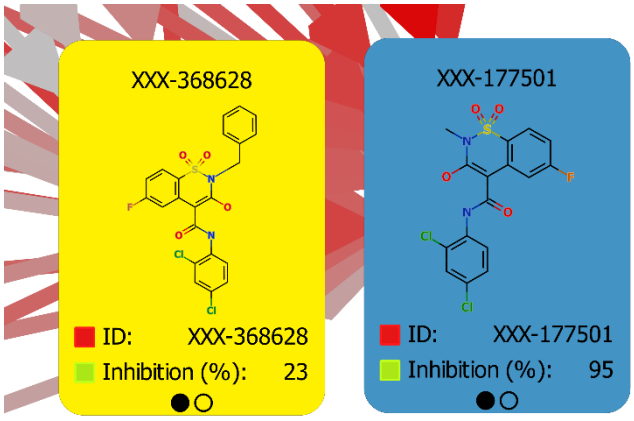
Note: You can zoom in to any region of the layout by pointing the mouse at it and using the mouse wheel, pinching or zooming on your trackpad or touchscreen, or using the **Ctrl** and **-/=** keys.

If in doubt, select the cluster in chemical space, as shown below, to identify the correct network section.

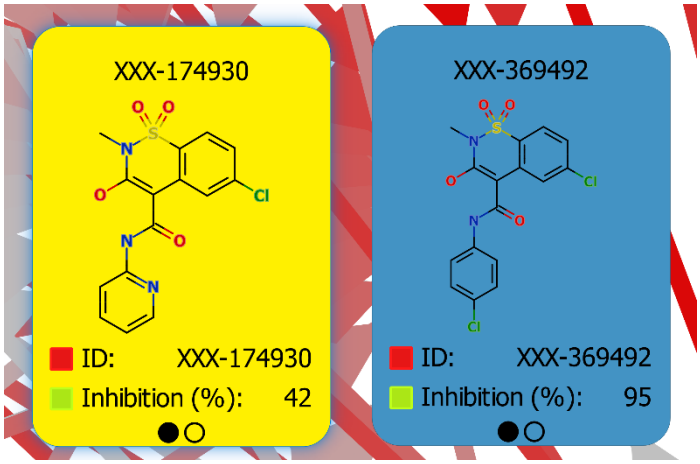


This series may be relevant because it includes multiple hits (blue cards) and it exhibits some interesting SAR. Large changes in activity resulting from small changes in structure are indicated by red arrows. Pointing at a link causes the linked cards to 'pop up' so that you can compare them side-by-side. You can also move the cards around for a clearer view by dragging them.

For example, we can see that the addition of bulky groups on the 1,2-benzothiazine nitrogen is not tolerated:

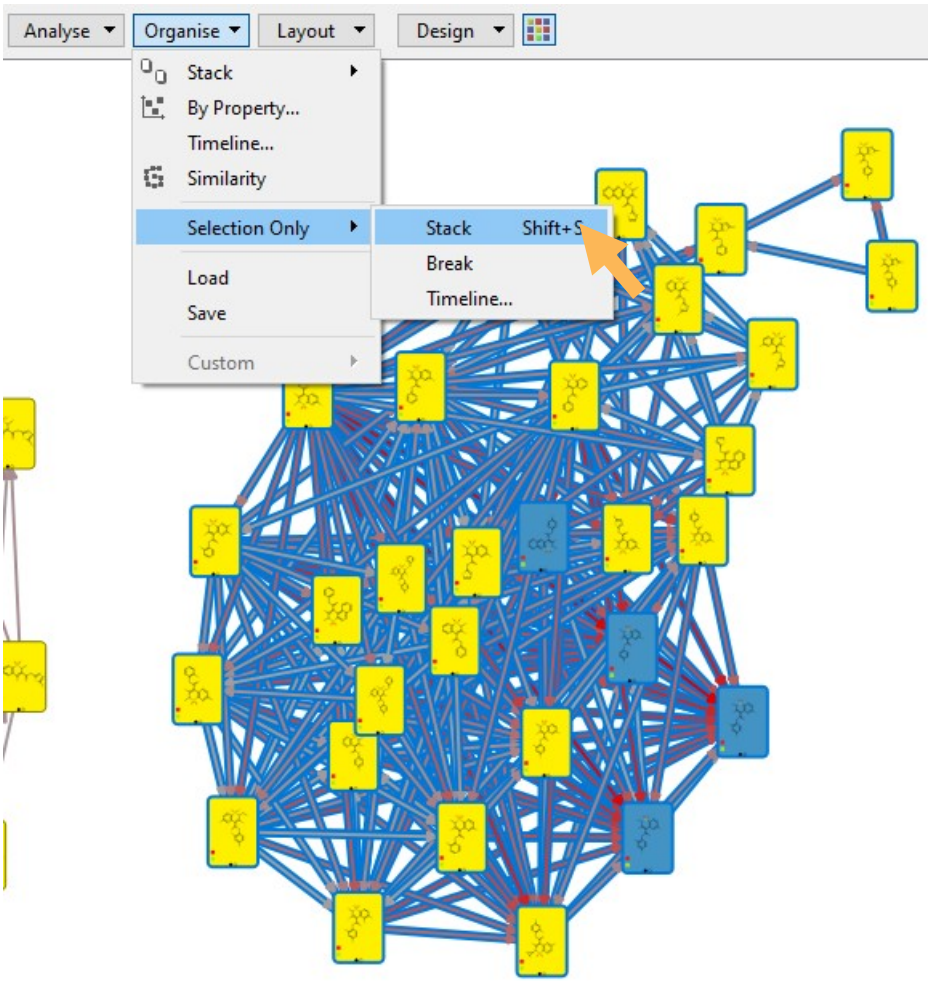


And, lipophilic aryl groups may be preferred over polar groups substituted on the amide.



We can tag the compounds in this series for further follow-up.

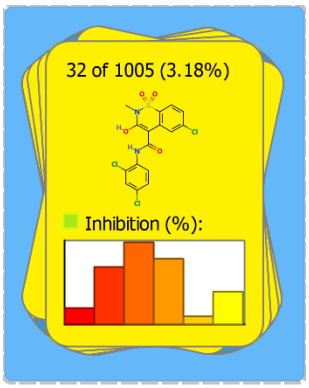
- Select the **lasso** tool at the top of Card View and draw around the network section to select all the cards therein.
- From the **Organise** menu at the top of Card View, select **Selection Only** and then **Stack**.



This will create a stack containing 32 cards, on which a representative structure and the distribution of inhibition values are displayed.

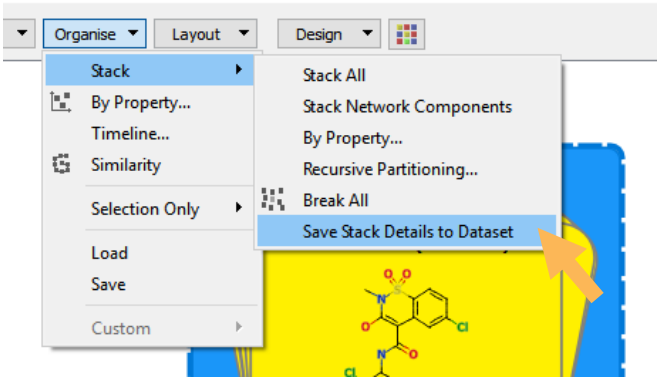
You can choose the data to display on each stack using the card and stack designer by choosing **Custom** from the **Design** menu. For more information, watch the short video Introduction to Card View Designer.

- Click on the label **New Stack 1** below the stack to provide a more appropriate name, in this case, we'll use "Follow up".

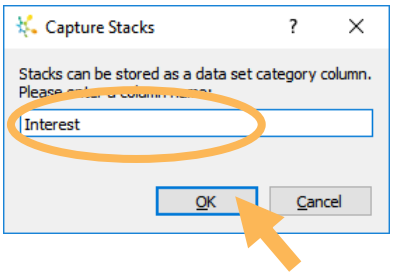


Follow up

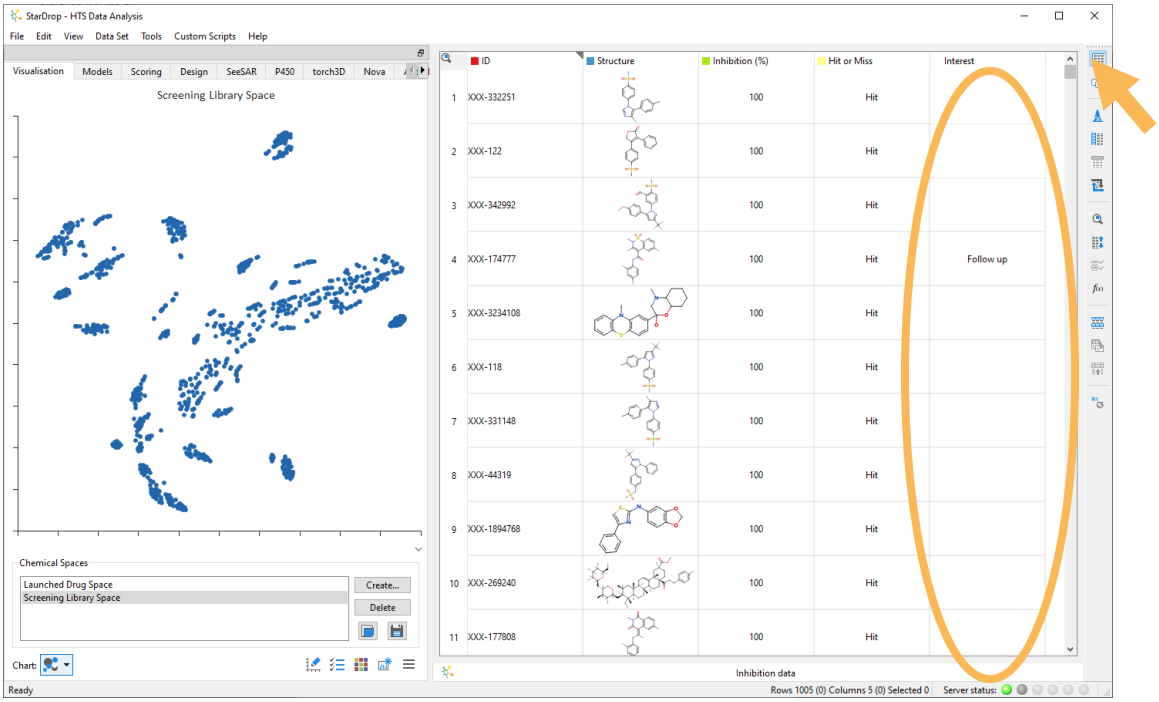
- We can tag these compounds in our data set by selecting **Stack** and then **Save Stack Details to Dataset** from the **Organise** menu at the top of Card View.



- Enter a name for the column in which the labels will be stored, for example, "Interest", and click the **OK** button.

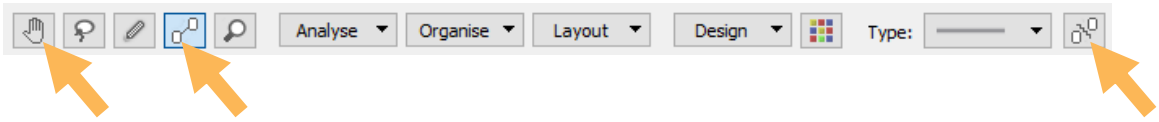


- This will create a new column in the data set in which the stacked compounds will be labelled as "Follow up". Change back to **Table View** by clicking on the **Table View** button at the top of the right-hand toolbar and look to the right to see this new column.



Another approach to identifying chemical series with good activity is by using clustering.

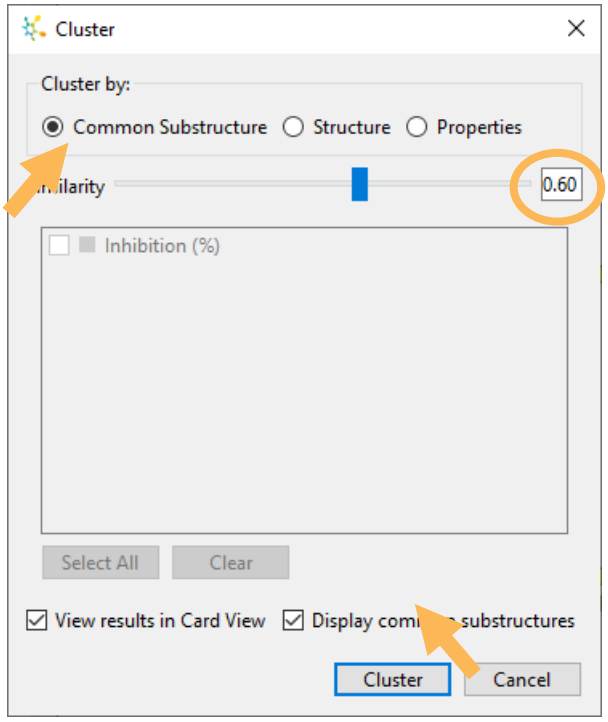
- Return to **Card View** by clicking on the **Card View** button on the right-hand toolbar and remove the links between cards by clicking on the **link** tool at the top, clicking the **clear links** button and finally selecting the **move** tool.



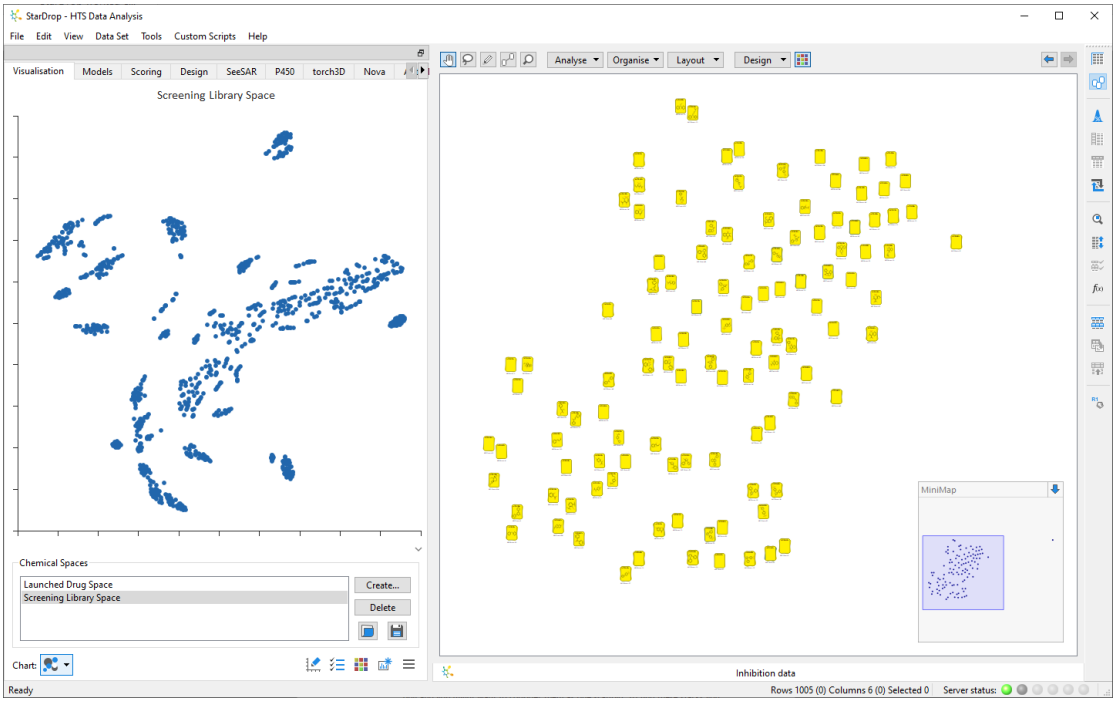
- From the **Analyse** menu at the top of Card View, select **Clustering**.

We're going to cluster together compounds that share a significant common substructure.

- In the **Cluster** dialogue, select the **Common Substructure** option, set the **Similarity** to **0.6** and then click **Cluster**.

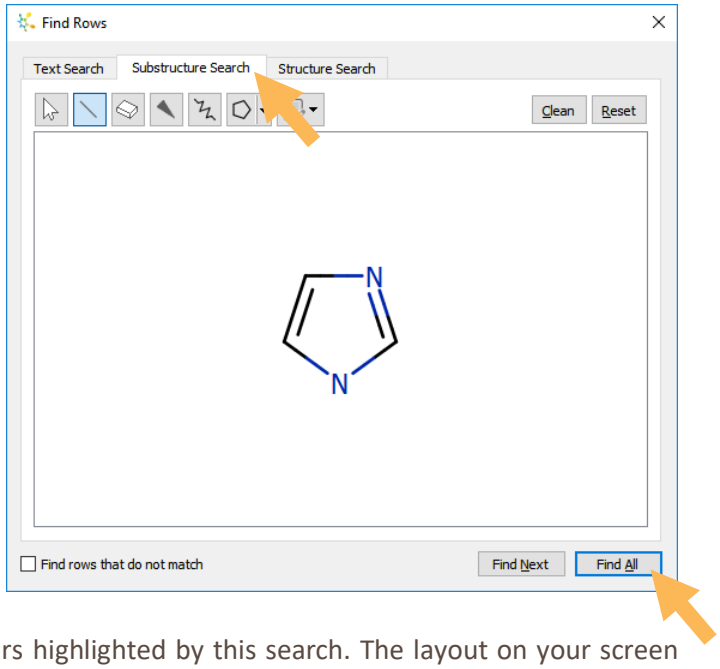


The resulting clusters are represented by stacks of cards in Card View. On the top of the stacks, the number of compounds is shown along with the substructure that all the compounds in the cluster have in common.



The cards are positioned such that stacks representing clusters with similar common substructures are close to one another.

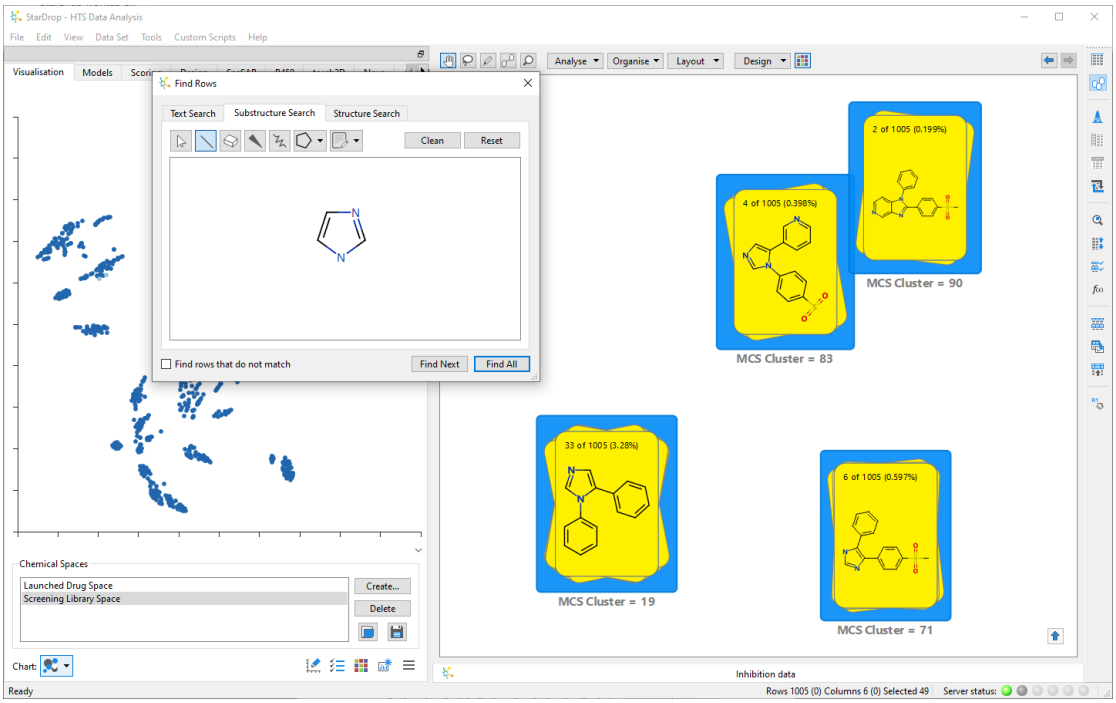
- A number of stacks are grouped around various substitution patterns on an imidazole ring, and you might want to consider them as one scaffold. To find these stacks, you can search for the imidazole substructure. Click the **Find** button on the right-hand toolbar and select the **Substructure Search** tab. Draw an imidazole as below and click the **Find All** button.



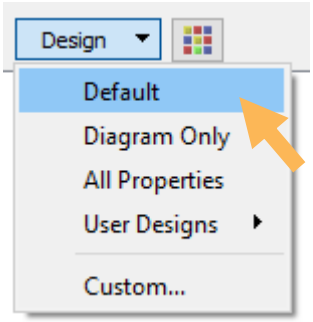
The substructure search tool provides enormous flexibility for matching variable atom and bond types, as well as linkers. For more details, watch the short video Flexible Substructure Searching.

Note: You can close the **Find Rows** dialogue if you wish to save screen space.

- Let's zoom in on the clusters highlighted by this search. The layout on your screen may be slightly different, depending on the dimensions of your screen.

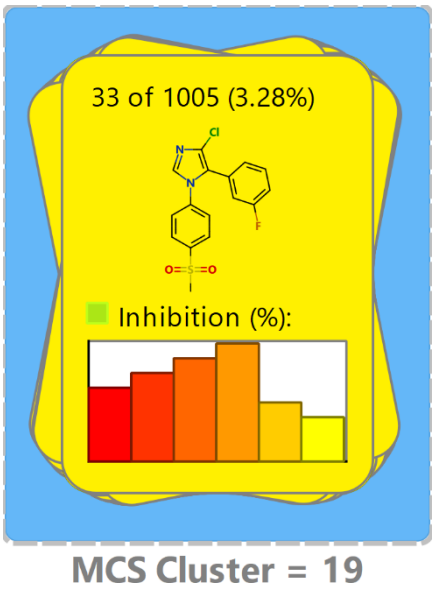


- To restore the default information displayed on the stack, select **Default** from the **Design** menu at the top of Card View.



The stacks will change to display the default information, in this case, the number of compounds, a representative structure and a histogram showing the distribution of the **Inhibition (%)** data for the compounds within the stack.

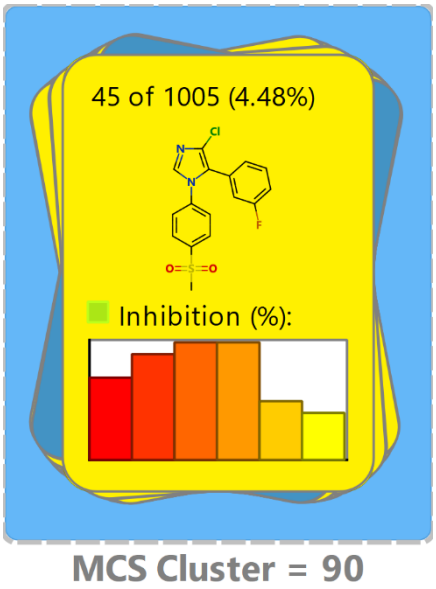
For example, this stack represents a series of 33 imidazoles with a good distribution of activity.



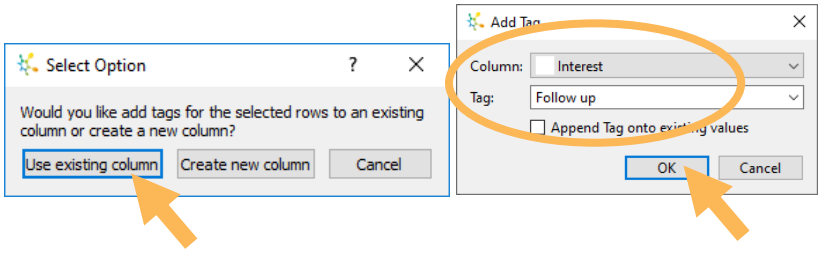
- Combine the four stacks by dragging them on top of one another.

You should end up with a stack containing 45 compounds.

We'll also tag this series for follow-up.



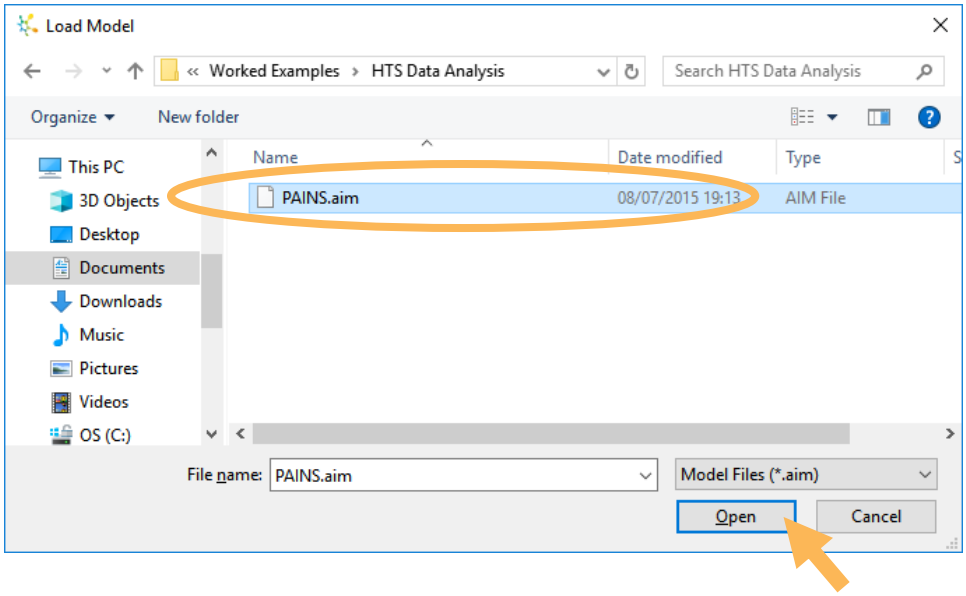
- Select the merged stack by clicking on it.
- Click the **Tag Selected Items** button on the right-hand toolbar  .
- Choose the **Use Existing Column** option and select the **Interest** column from the drop-down. Enter "Follow up" as the **New tag** and click the **OK** button.



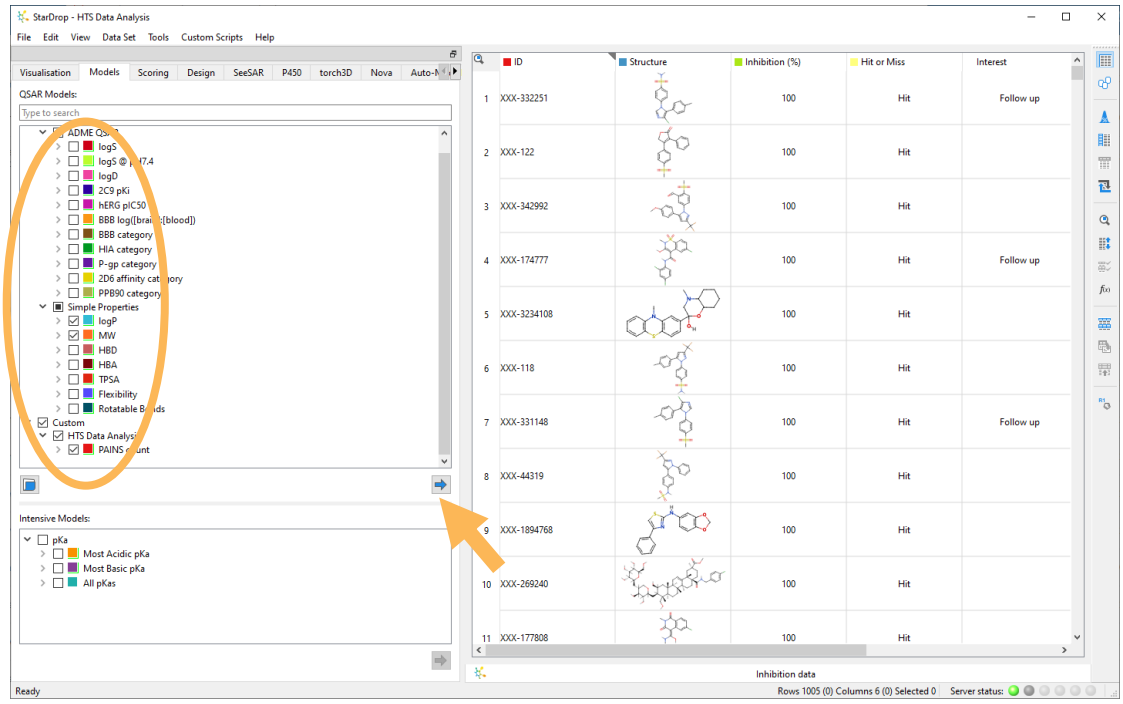
Target inhibition is not the only requirement for a high-quality hit series, so we'll now consider some other simple properties that may be relevant to the selection of compounds for further follow-up: lipophilicity (logP), molecular weight (MW) and the number of structural alerts corresponding to pan-assay interference compounds (PAINS) [Baell and Holloway, J. Med. Chem. 2010 53(7) pp. 2719-2740].

- Change back to **Table View** by clicking the **Table View** button at the top of the right-hand toolbar.
- Change to the **Models** area in StarDrop to calculate these properties by clicking on the **Models** tab.

- The PAINS substructure alerts are not a standard model in StarDrop, so we'll load this additional model. Click the button in the **Models** area, select the **PAINS.aim** model file and click the **Open** button.



- In the Models area, tick the boxes next to **logP**, **MW** and **PAINS count** (**Note:** the PAINS count model will appear in a branch with the name of the directory from which it was loaded) and click the  button. The calculated properties will be added to the data set.

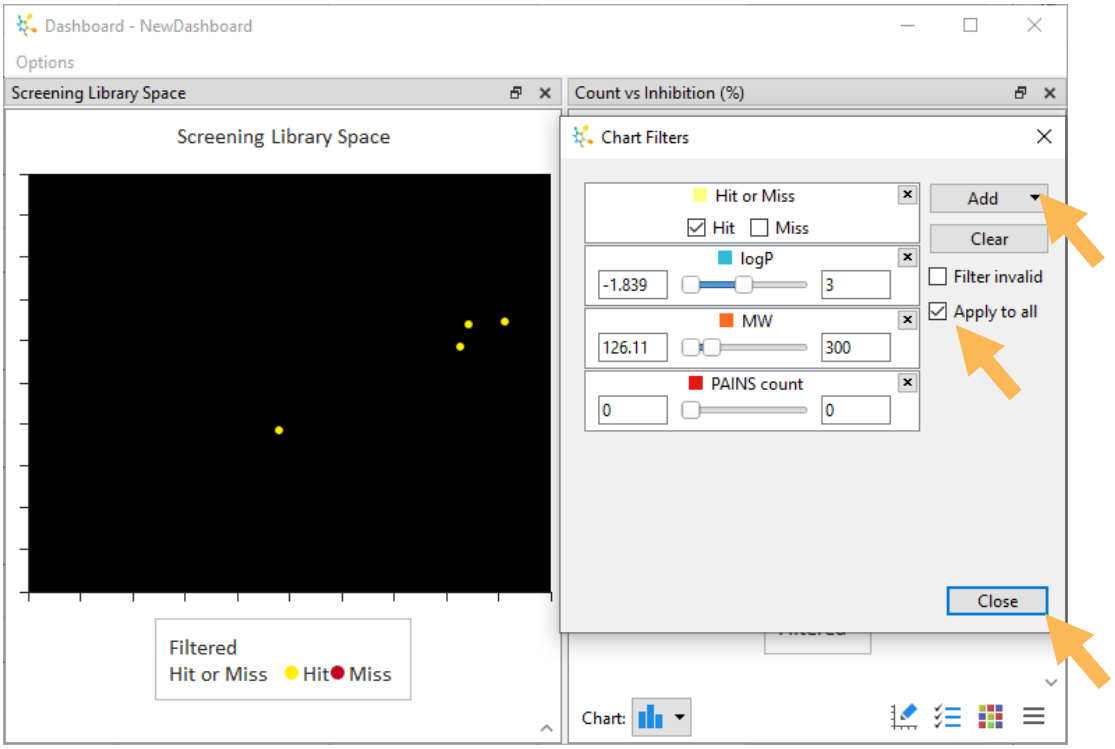


A common approach is to filter the hits according to 'lead-like' property criteria, for example:

- logP < 3
- MW < 300
- No PAINS hits (PAINS count ≤0)

These can easily be applied using the Filter tool (available if you select **Filter** from the **Data Set** menu). We can also achieve this by applying dynamic filters to our visualisations, which we'll try here.

- Open the dashboard showing the chemical space and distribution of inhibition (%) data.
- Click the arrow beneath one of the charts to show the chart controls.
- Click on the **Chart menu** button ☰ and choose **Filter**.
- Click the **Add** button and add a filter for **Hit or Miss** (uncheck Miss), **logP** (set the upper threshold to 3), **MW** (set the upper threshold to 300) and **PAINS count** (set the upper bound to 0).
- Tick the **Apply to all** box to ensure the filters are applied to all the charts in the dashboard.



You will see that as you add the filters, they are applied to all the charts immediately.

Note: You can use the numerical filters to specify ranges and if you wish to invert the range, simply click on the slider bar. You can shift the range by dragging it.

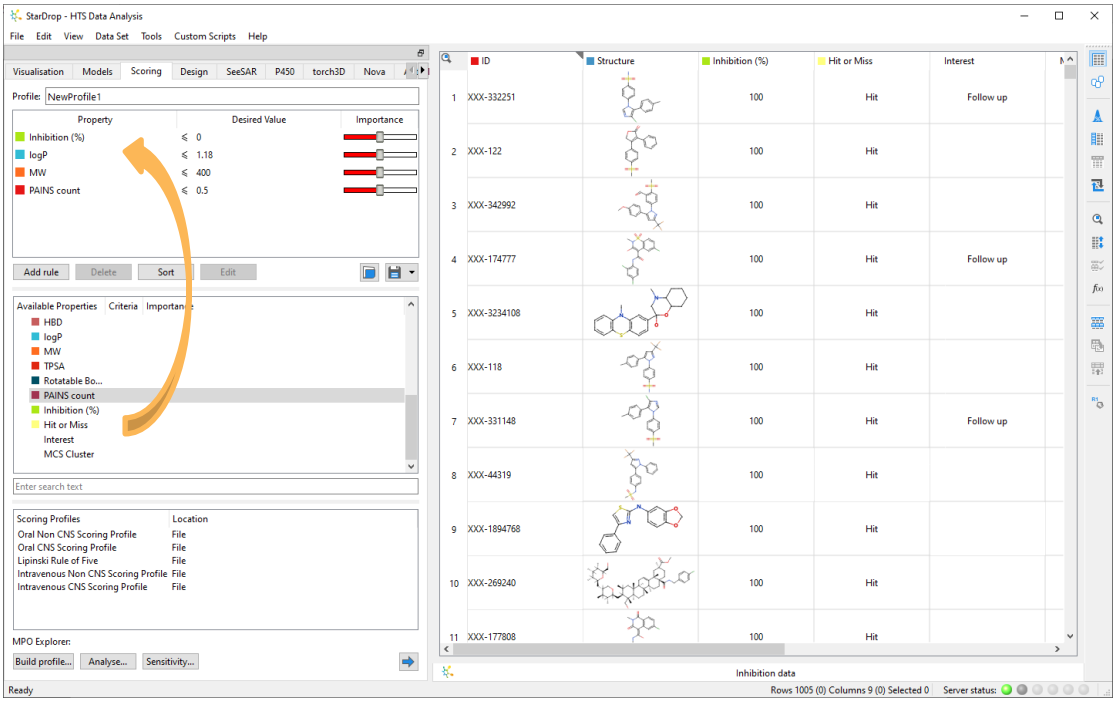
Only four of the hits remain that pass all these criteria. You can find these in the data set by drawing around the points in the chemical space to select them.

- Clear all the filters by clicking on the **Clear** button in the **Chart Filters** dialogue and then click the **Close** button.

Applying these hard filters dramatically reduces the number of potential hits to follow up, and may artificially restrict the choice of direction for the project (in this example, some of the 4 remaining hits may be considered questionable) or inappropriately reject good compounds. We have chosen a cut-off of >80% inhibition to define 'hits' but, given the variability in the assay results, we cannot confidently reject compounds with measured values close to this cut-off. Furthermore, the 'lead-like' criteria are not hard-and-fast rules; for example, a compound with MW of 299 Da does not represent a significantly better starting point than one with MW of 301 Da. Similarly, the PAINS filters, although popular, have been shown to correlate poorly with promiscuity of binding and many approved drugs contain PAINS alerts [Capuzzi *et al.* JCIM DOI: 10.1021/acs.jcim.6b00465].

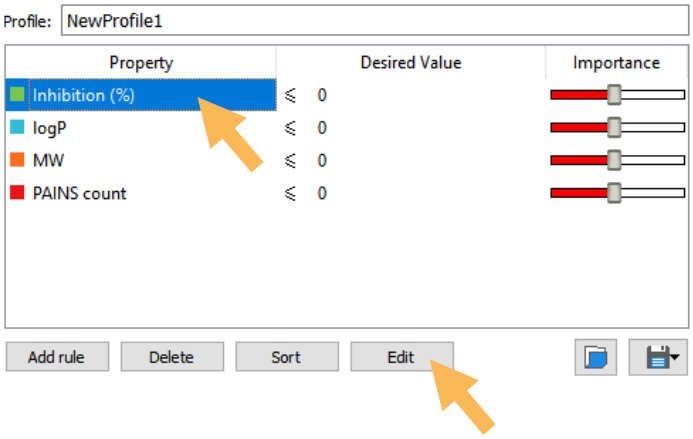
Therefore, a better approach to the prioritisation of potential hits for follow-up is to apply a multi-parameter optimisation method, where appropriate weights can be given to the experimental and calculated results. To explore this, we will use StarDrop's Probabilistic Scoring.

- Minimise the dashboard again and change to the **Scoring** area in StarDrop by clicking on the **Scoring** tab.
- From the list of **Available Properties** drag the **Inhibition (%)**, **logP**, **MW** and **PAINS count** properties into the scoring profile near the top of the area (it may be necessary to scroll down the list of available properties to find them).



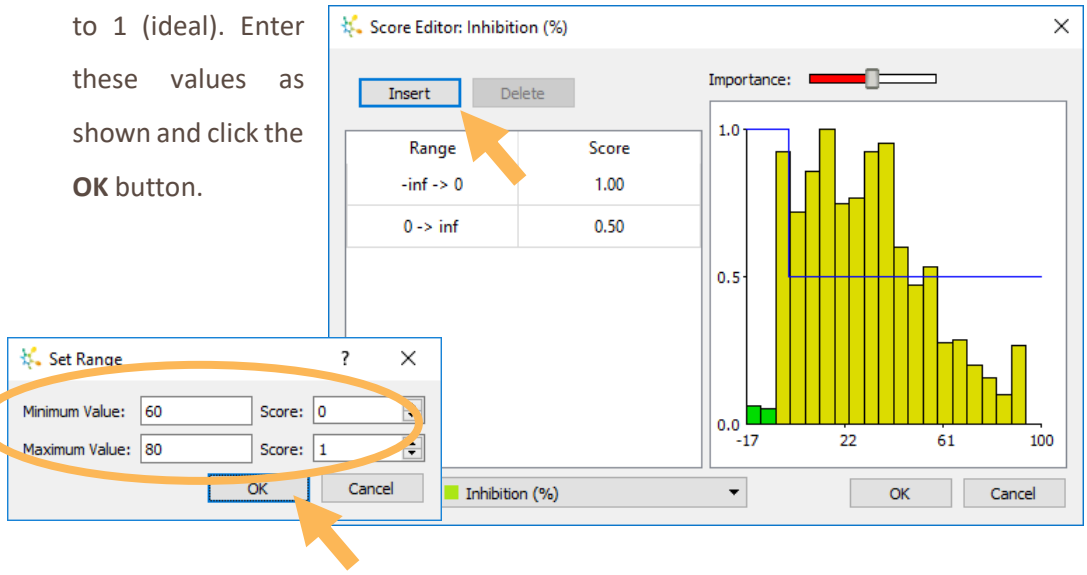
For each property, we are going to define a function which describes an ideal value range and its relative importance.

- Select the **Inhibition (%)** property in the scoring profile and click the **Edit** button below the scoring profile (alternatively, double-click Inhibition (%) in the list).

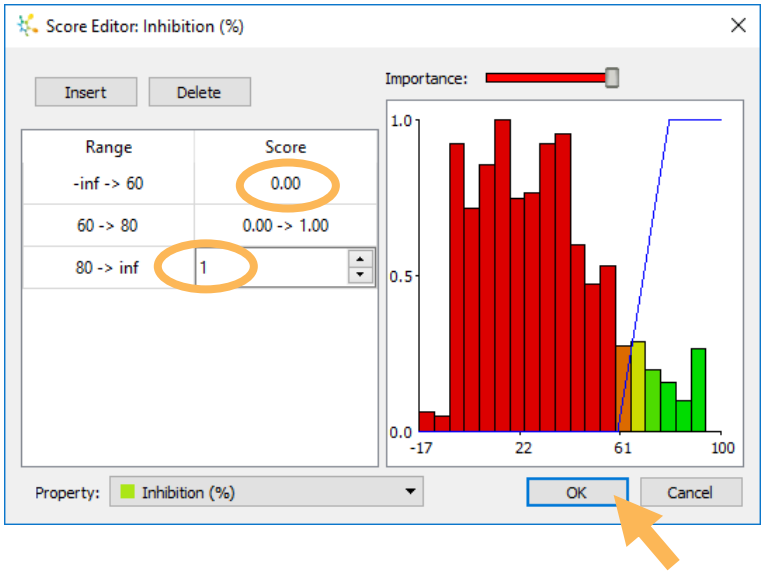


As we found previously, a reasonable value for percentage inhibition is >80 % to identify a hit, but we might be willing to accept slightly lower if all other properties were good. So, a hard cut-off is not necessarily appropriate.

- Click the **Insert** button in the **Score Editor** to insert a range between 60 and 80% inhibition. In this range, the score will increase from a low value of 0 (unacceptable) to 1 (ideal). Enter these values as shown and click the **OK** button.

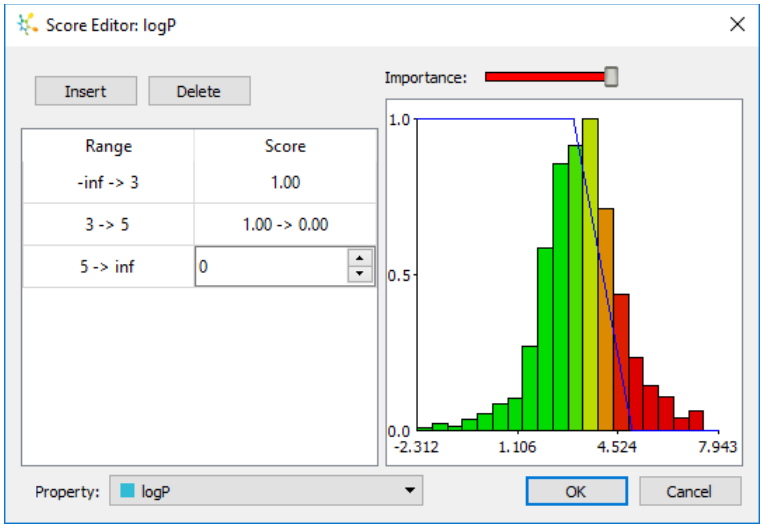


- Below 60% inhibition, the score should be 0, and above 80% the score should be 1, so enter these values in the Score Editor, then click the **OK** button.

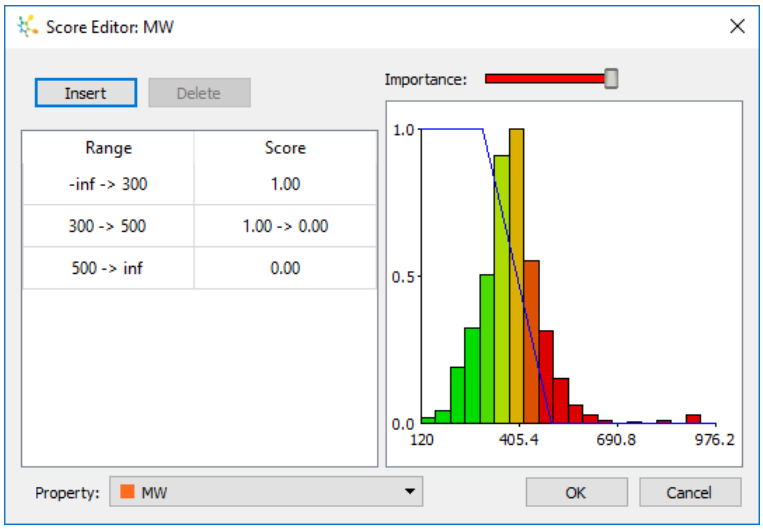


- Repeat this process to define scoring functions for logP, MW and PAINS count as shown on the next page.

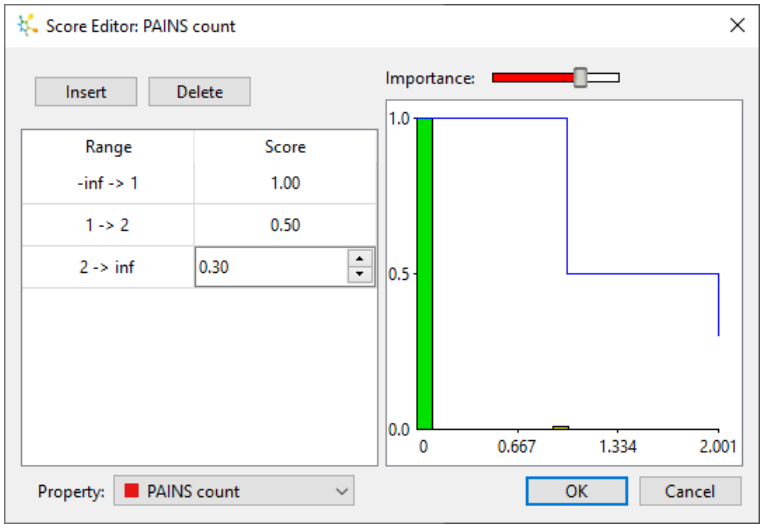
- logP



- MW

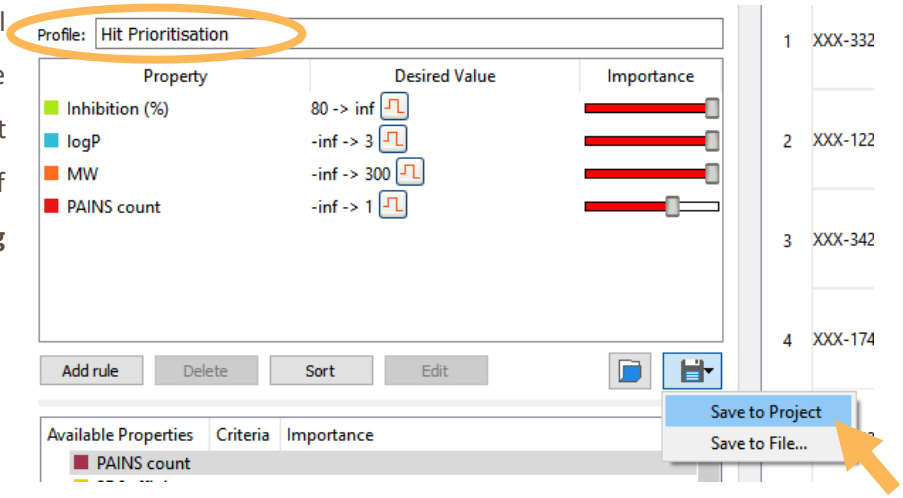


- PAINS count



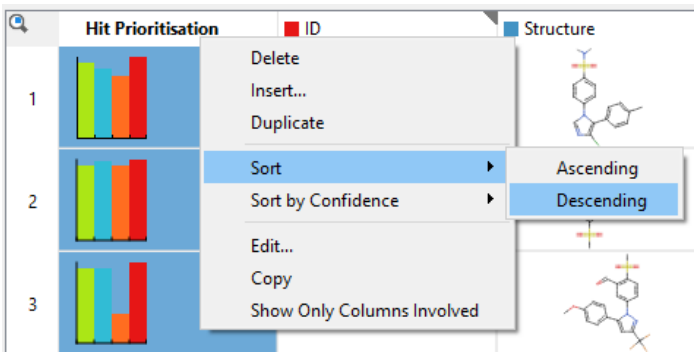
- Give the profile a name by editing the **Profile** text above the scoring profile (we have used “Hit Prioritisation”) and save the profile to the project for future reference by selecting **Save to Project** from the **Save** drop-down below the profile.

The profile will appear in the list of profiles at the bottom of the **Scoring** area.



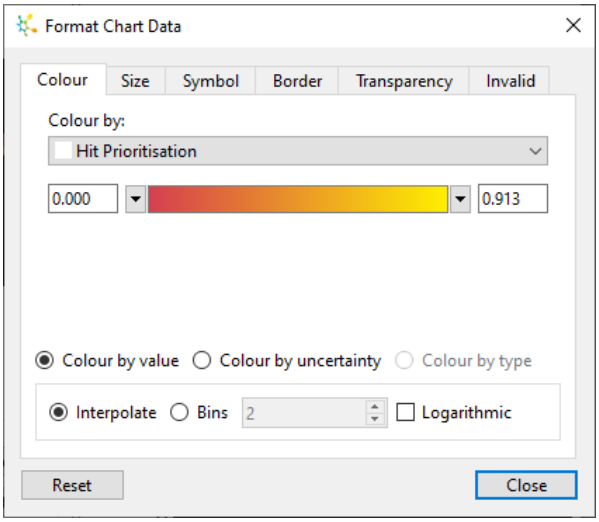
- Click the  button at the bottom of the **Scoring** area to run the profile, ignoring the warning about zero uncertainty in the MW and PAINS count columns; this is correct for these properties, so you can click the **OK** button to continue.

- Sort the compounds by score by right-clicking on the score column and selecting **Descending** from the **Sort** menu. **Note**: You can choose to apply the new order also to Card View.

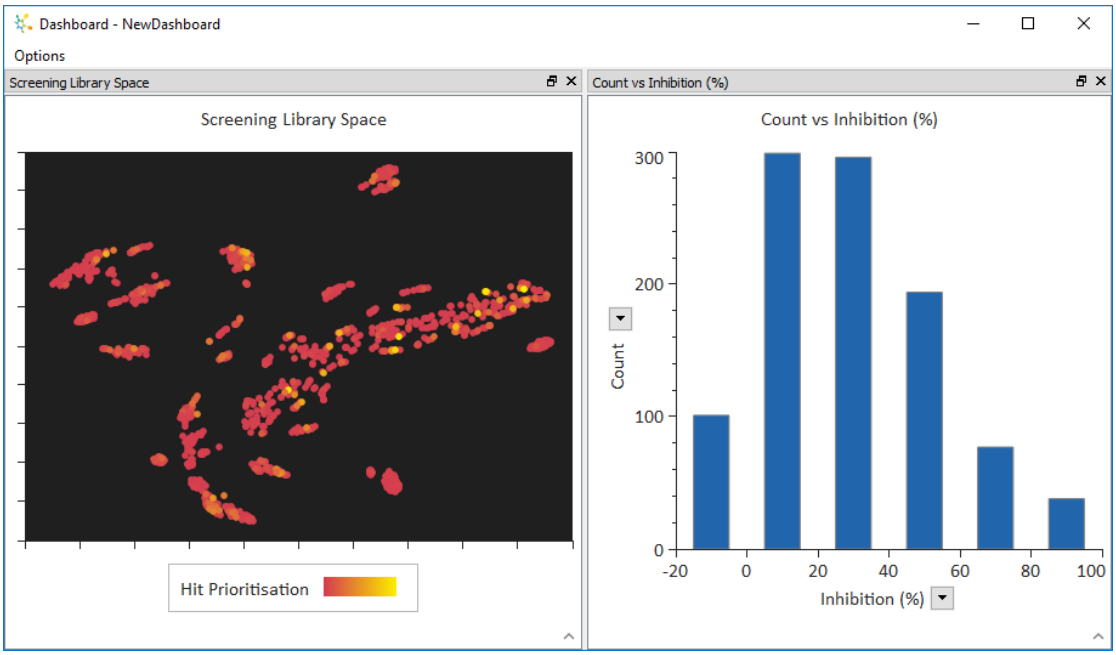


We can now explore how the compound scores are distributed across the chemical space of the screening library by changing the colour formatting.

- In the dashboard, click the arrow at the bottom of the chemical space window, click the **Format** button  and select "Hit Prioritisation" (the scoring profile name) from the **Colour by** drop-down in the **Format by Property** dialogue.

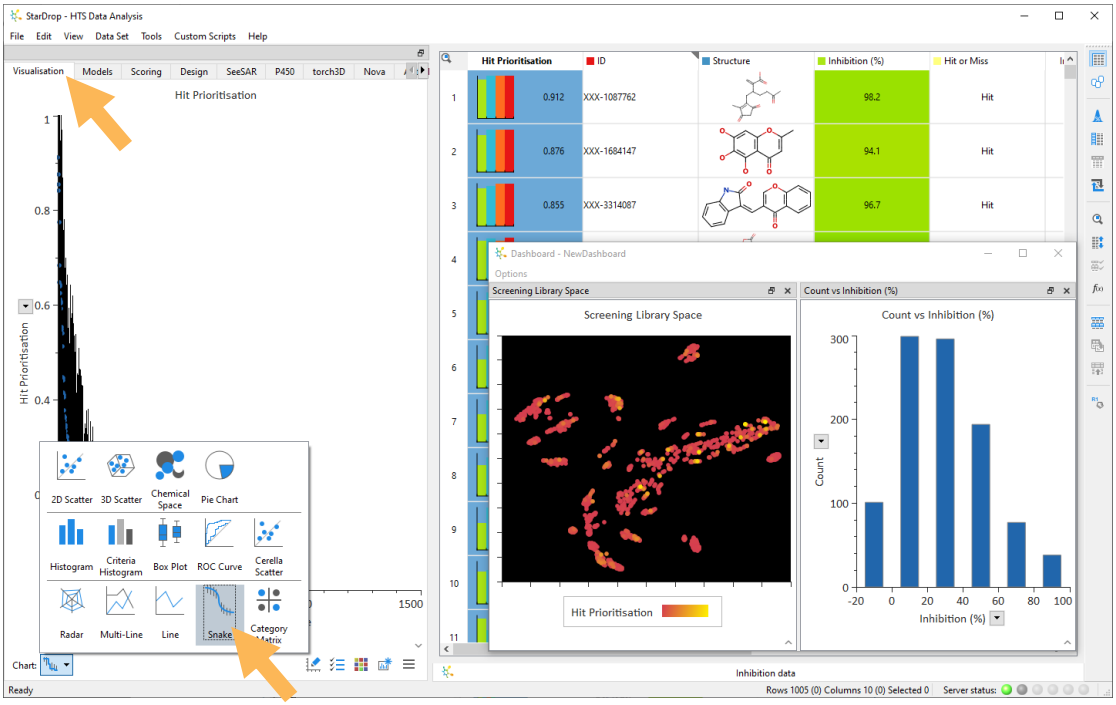


In the resulting chemical space, we can see that there are high-scoring compounds in several clusters, representing a broader diversity than the four compounds that passed all four filters earlier.



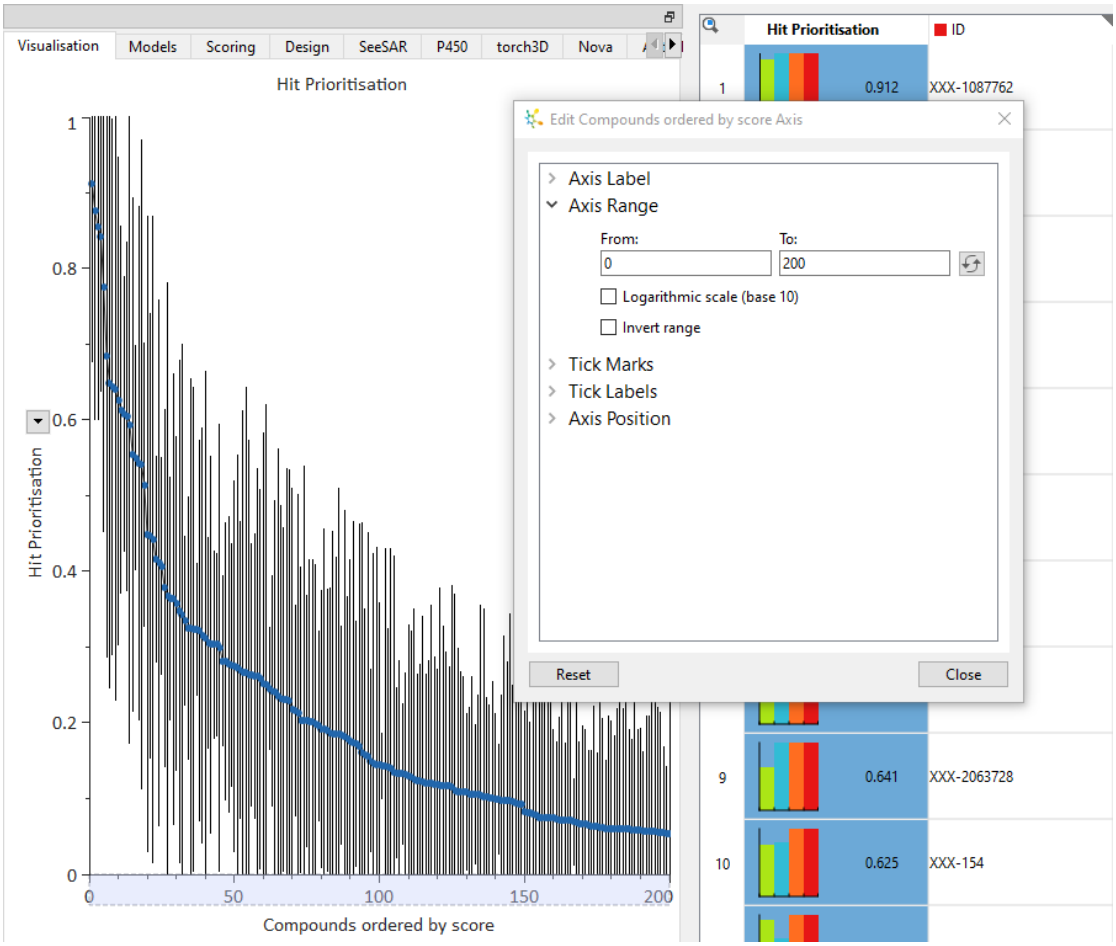
Now we should consider how many of these compounds should be evaluated for follow-up.

- Change to the **Visualisation** area and select **Snake** from the **Chart** menu.



The resulting visualisation illustrates the distribution of scores within the library. The compounds are ordered by score along the x-axis, from the highest on the left to the lowest on the right. The compound scores are plotted on the y-axis and error bars indicate the uncertainties in the overall scores, given the uncertainties in the experimental inhibition data and the predicted logP.

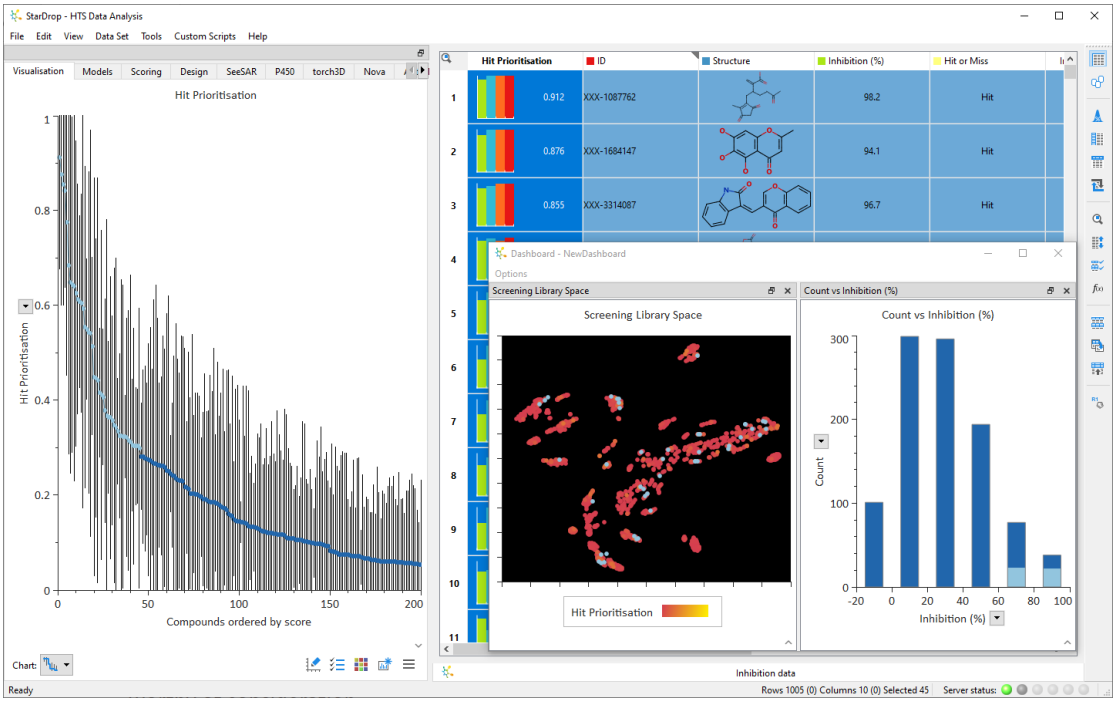
- Zoom into the highest-scoring compounds by right-clicking on the x-axis and choosing **Edit** from the menu. Modify the **Axis Range** to be **0** to **200**.



Note: As with formatting and other chart customisations, as you adjust axes, the chart will update to show you the results of your modifications.

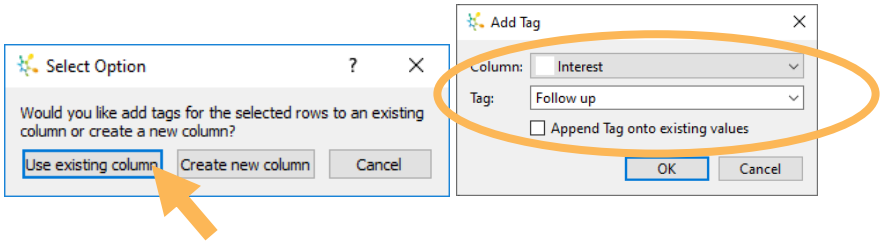
From this, we can see that roughly the top 30 compounds are not confidently distinguishable from the highest-scoring (their error bars overlap with the first compound). Alternatively, approximately 45 of the top-scoring compounds have a score that is statistically better than zero (the error bar does not meet the x-axis). So, it would be reasonable to further consider the top 30 to 45 compounds.

- Select these compounds by drawing around the points on the snake plot. The corresponding points will also be highlighted in the dashboard charts.



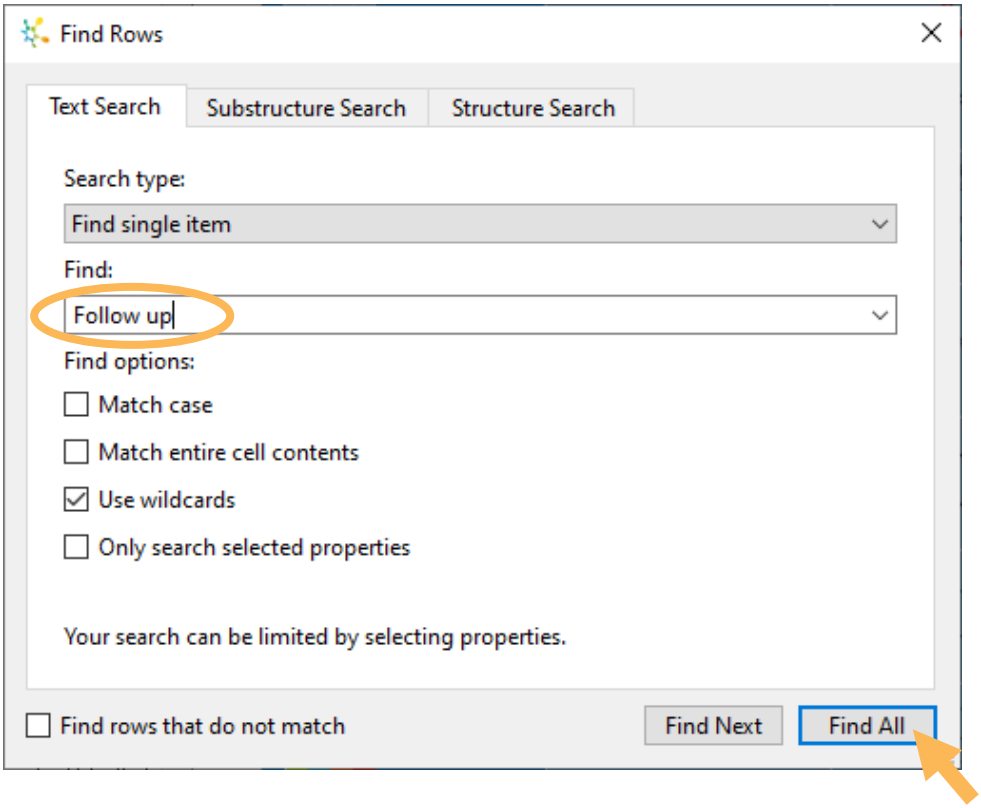
Here we can see that the selected compounds highlight several chemical series worthy of consideration.

- As before, tag the compounds in the data set for follow-up by clicking the **Tag Selected Items** button on the right-hand toolbar.
- Click the **Use Existing Column** button and select the **Interest** column from the drop-down, specifying “Follow up” as the **New tag**, as before. Click the **OK** button.



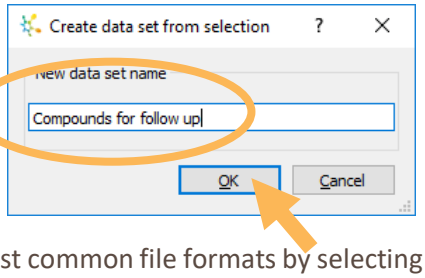
We are now going to copy all the compounds that we have flagged for follow-up into a new data set.

- Click the **Find** button on the right-hand toolbar.
- In the **Find Rows** dialogue, select the **Text Search** tab, enter “Follow up” and click **Find All** to select all the data set rows where this text is found.

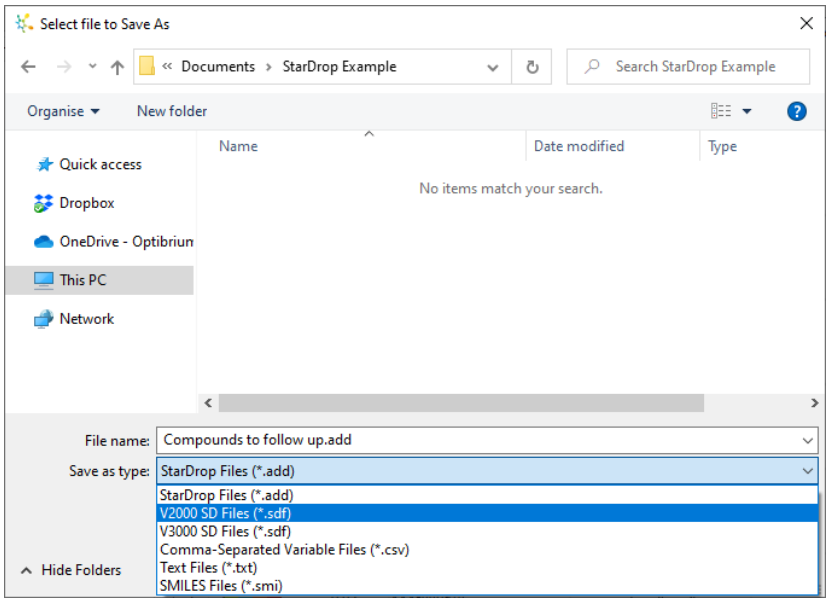


- Create a new data set containing only the selected compounds by clicking the **Create Data Set From Selection** button on the right-hand toolbar.

  Give the new data set a name and click the **OK** button.



- The resulting data set can then be exported in most common file formats by selecting **Save Data Set As** from the **File** menu.



This example has illustrated several ways in which HTS results can be analysed within StarDrop to identify high-quality hits. We have used a relatively small data set from the public domain in this example for speed; however, these techniques can be applied to data sets of several tens of thousands of compounds, depending on the memory and performance of your computer. Please note that the Card View approach to visualisation is, realistically, limited to ~20,000 compounds due to the complexity of representing compounds as individual cards.

If you have any questions, please contact stardrop-support@optibrium.com and explore our online community at www.optibrium.com/community for more tutorials and videos.