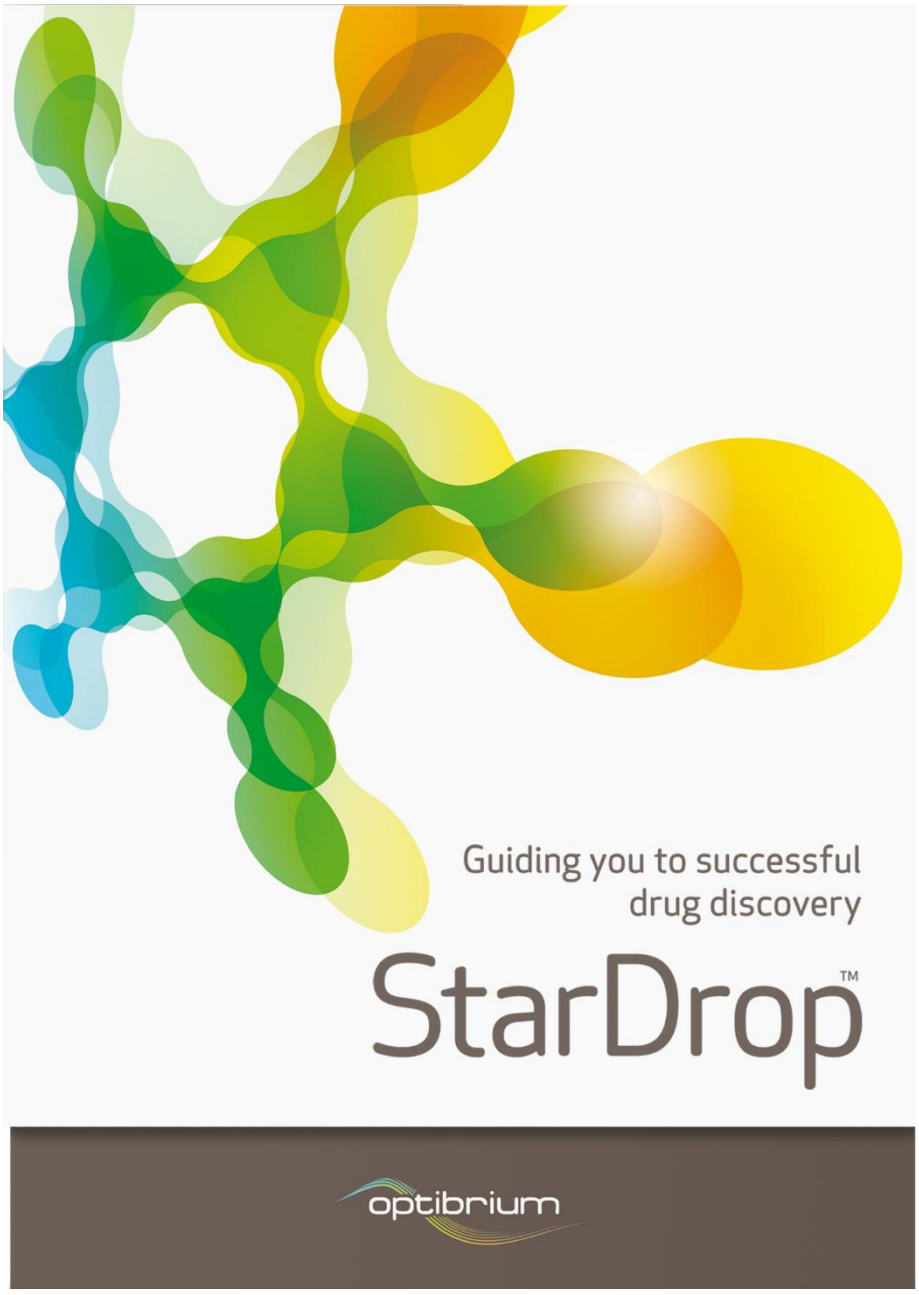
Guided Multi-Parameter Optimisation of 2D and 3D SAR

RSC/SCI Workshop on Computational Tools for Drug Discovery



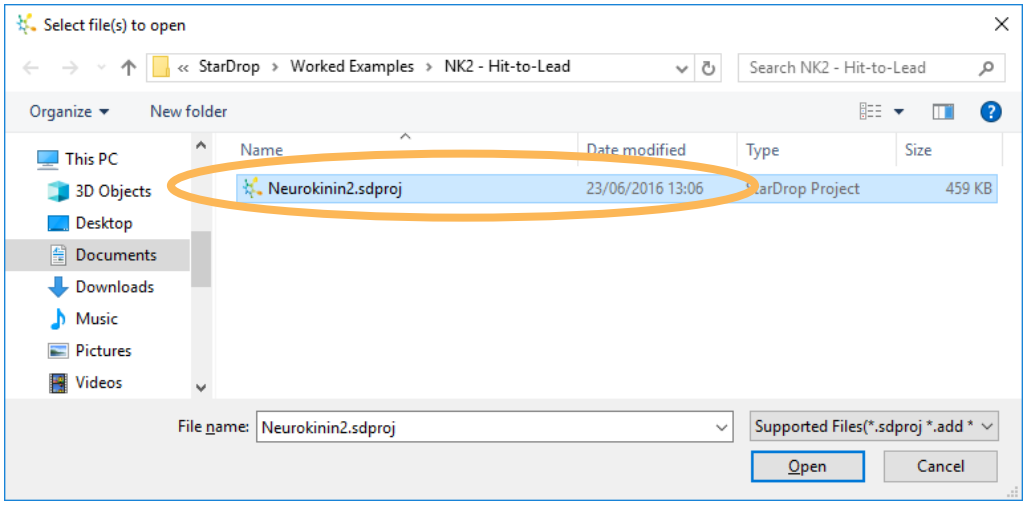
Guiding Compound Selection in Hit-to-Lead

This example explores some of the challenges typically encountered in a hit-to-lead project. The objective in this case is to identify one or more high quality chemistries for progression to detailed *in vitro* and *in vivo* studies, based on initial screening data for potency; ideally the compounds chosen for progression should not only be potent, but also have appropriate ADME properties to result in a high-quality lead series.

During this exercise we will use a variety of StarDrop's capabilities to explore the data in order to select compounds with a good balance of properties. Step-by-step instructions for all the features you will need to use in StarDrop are provided, along with screenshots and examples of the output you are likely to generate. If you have any questions, please feel free to contact stardrop-support@optibrium.com.

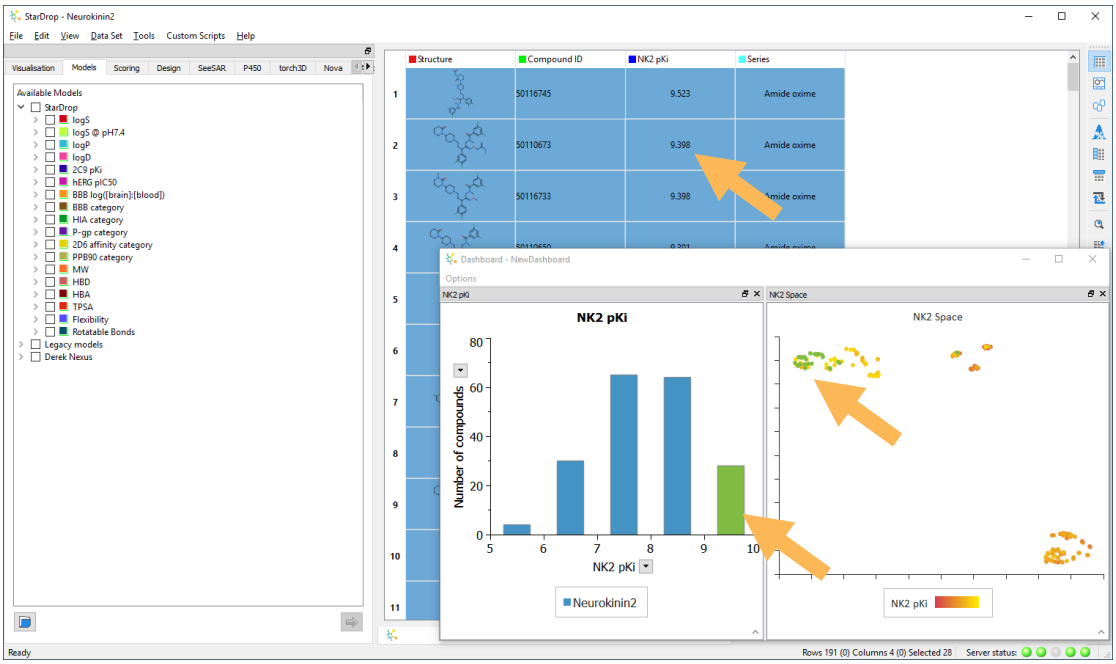
Exercise

- In StarDrop, open the file **Neurokinin2.sdproj** by selecting **Open** from the **File menu**.



You will see a spreadsheet containing 191 structures and their measured affinities for Neurokinin 2 (in the column **NK2 pKi**).

A dashboard is also displayed, containing a histogram showing the distribution of pK_i values for the compounds in the data set and a 'chemical space' showing the distribution of these activities across the diversity of compounds that have been screened. In this chemical space, each point corresponds to a compound and the distance between points represents their structural similarity.

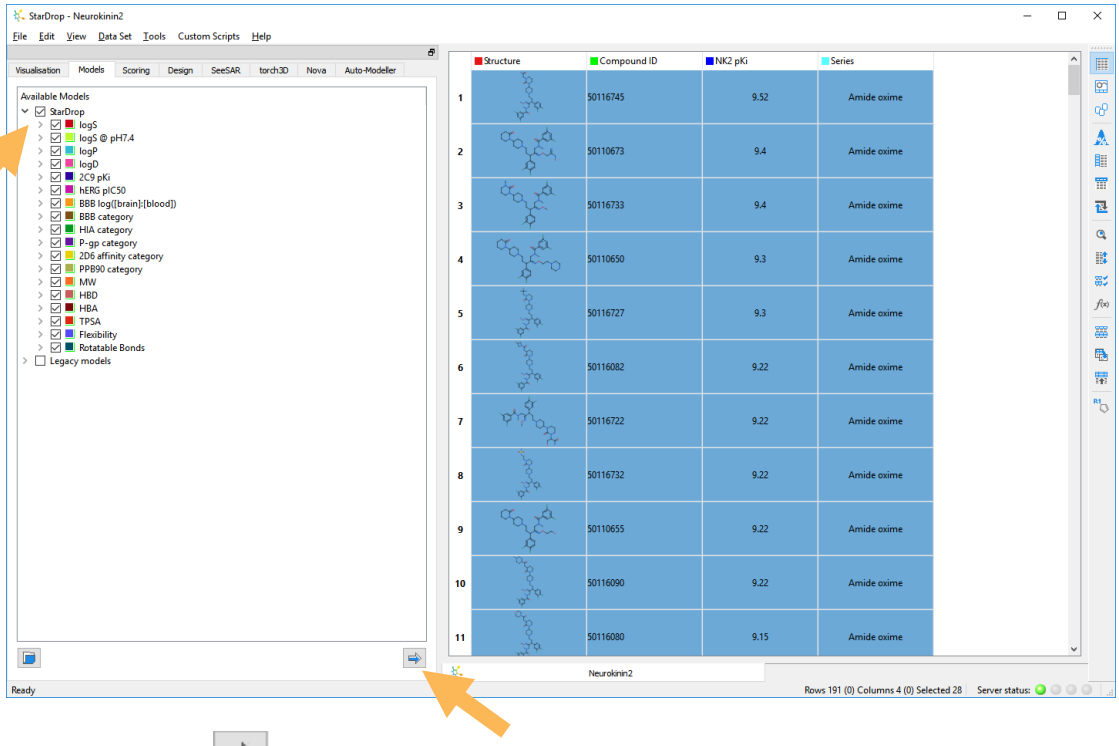


You will notice that a large proportion of the compounds with the highest pK_i values (yellow) come from one of the clusters in the chemical space.

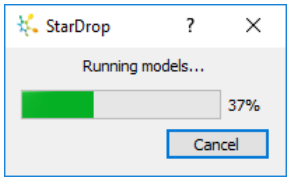
- To confirm this, click on the right-hand histogram bar which represents the highest pK_i values.

This will highlight the most potent compounds in the chemical space, as well as the data set. This region of the chemical space may be the best area in which to focus for selecting compounds, but first we should consider the other properties that are important in a high-quality lead (you can minimise the dashboard window to keep it out of the way for now). We have no further measured data but can generate predictions of ADME and physicochemical properties.

- In the **Models** area on the left, select all the StarDrop models by ticking the box next to the word StarDrop.

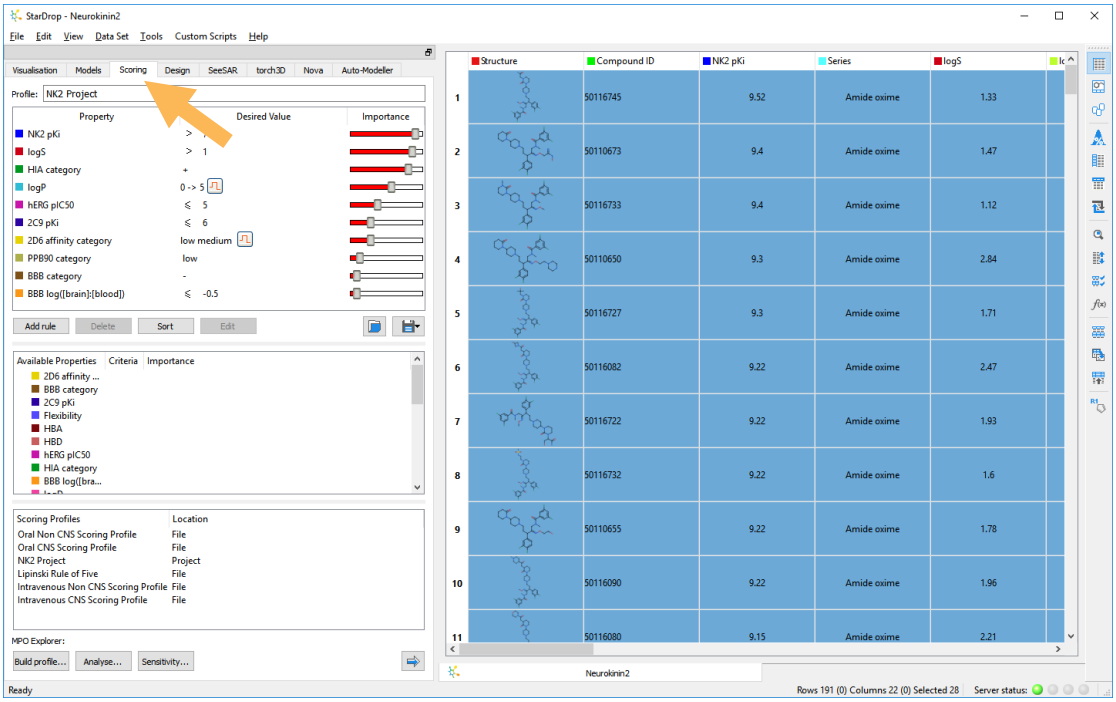


- Click the  button at the bottom of the **Models** area. A progress bar will be displayed while the predictions are calculated.



When this process is complete you will see that a new column has been added to the data set for each property calculated. Due to the volume and complexity of the data, it is challenging to find the compounds which have the best balance of properties and so we're going to use StarDrop's approach to multi-parameter optimisation (called Probabilistic Scoring) which makes it easy to assess all this information.

- Click on the **Scoring tab**.



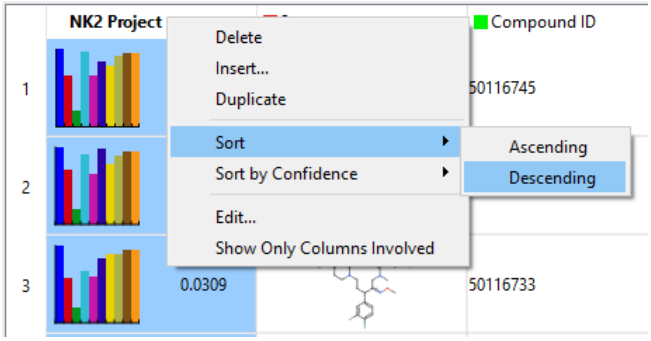
A scoring profile enables you to define a set of criteria that are important for your project. StarDrop provides some example profiles and, as part of the project file, we have already loaded a profile designed for this Neurokinin 2 project.

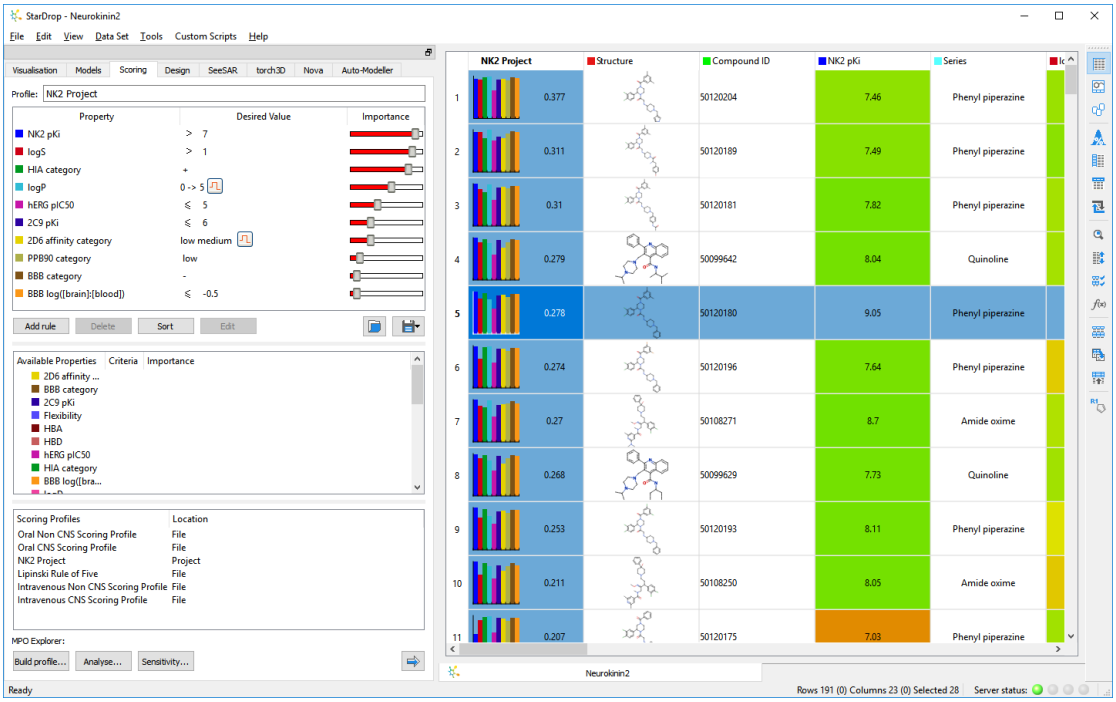
A scoring profile contains a series of properties along with criteria describing desired values and their relative importance. In this profile we are looking for compounds with good affinity for Neurokinin 2 and which are suitable for a peripheral target.

- Run scoring by clicking the  button at the bottom of the **Scoring area**.

A new column is added to the data set containing a score for each compound, taking into account each property criteria, its relative importance in the profile, and the uncertainty in the underlying experimental and predicted data. The score is a value between 0 and 1, representing the likelihood of the compound meeting all the criteria in the profile.

- To find the highest scoring compounds, right-click on the scoring column header and choose **Descending** from the **Sort menu** to sort the data set from high to low score.

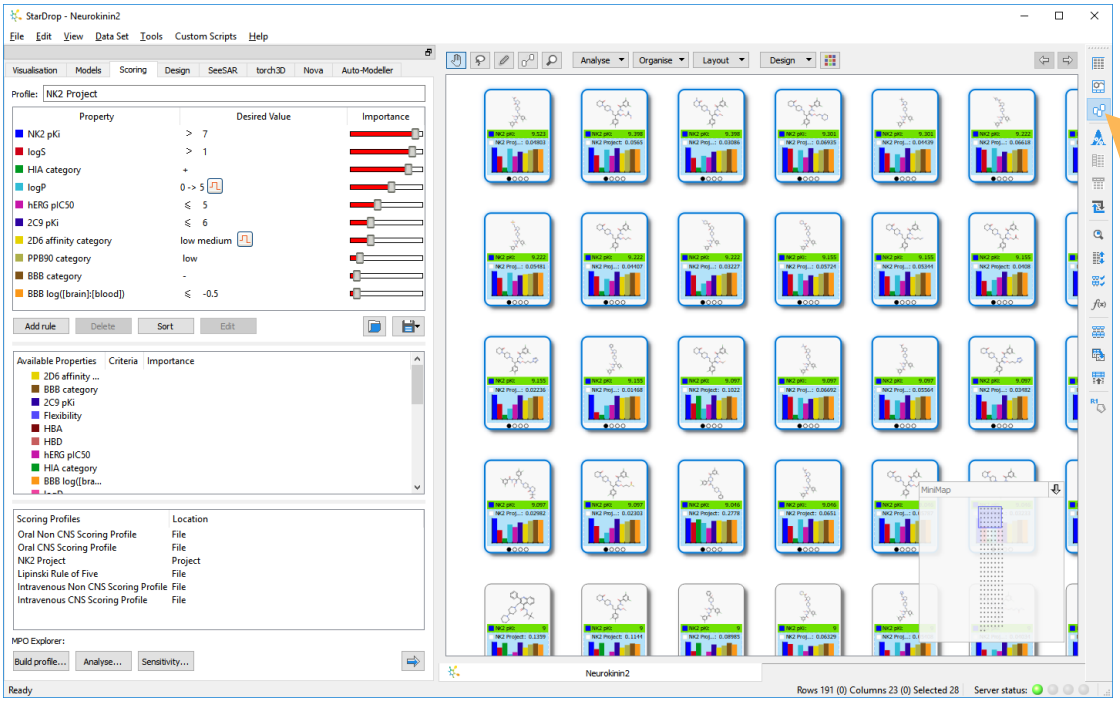




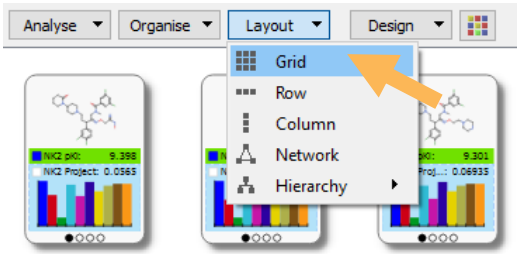
The histogram in each cell gives a quick overview of the impact of each property on the score for a compound. As you scroll down the data set, and the scores become lower, the low bars in the histograms indicate properties that have not met the requirements defined in the project profile, taking into account the confidence in the data and importance of each property.

Having scored the compounds, we're now going to compare some compounds using StarDrop's Card View. This provides a convenient way to work with our compounds and data in the context of this project, by representing compounds on cards that can be moved, stacked and linked however we wish.

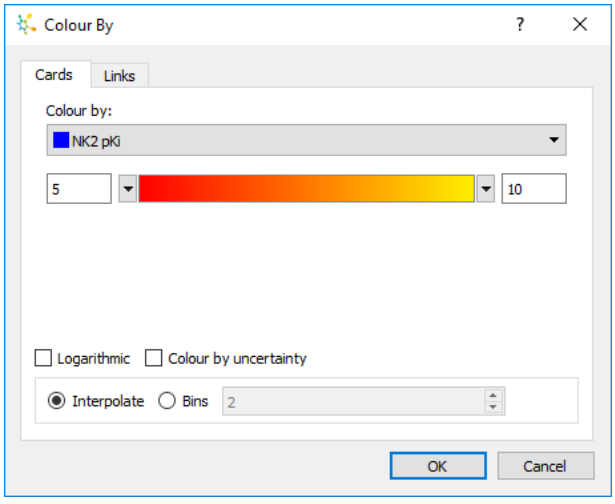
- Switch to **Card View** by clicking the **Card View** button in the toolbar to the right. This will display the data set as a grid of cards. You can choose which properties you would like to see on a card, but in this case, we have already created a card design that shows the scoring profile and some of the properties in which we are interested.



- Sort the grid in order of the score from top-left to bottom-right by choosing the **Grid** option from the **Layout menu** at the top of Card View.

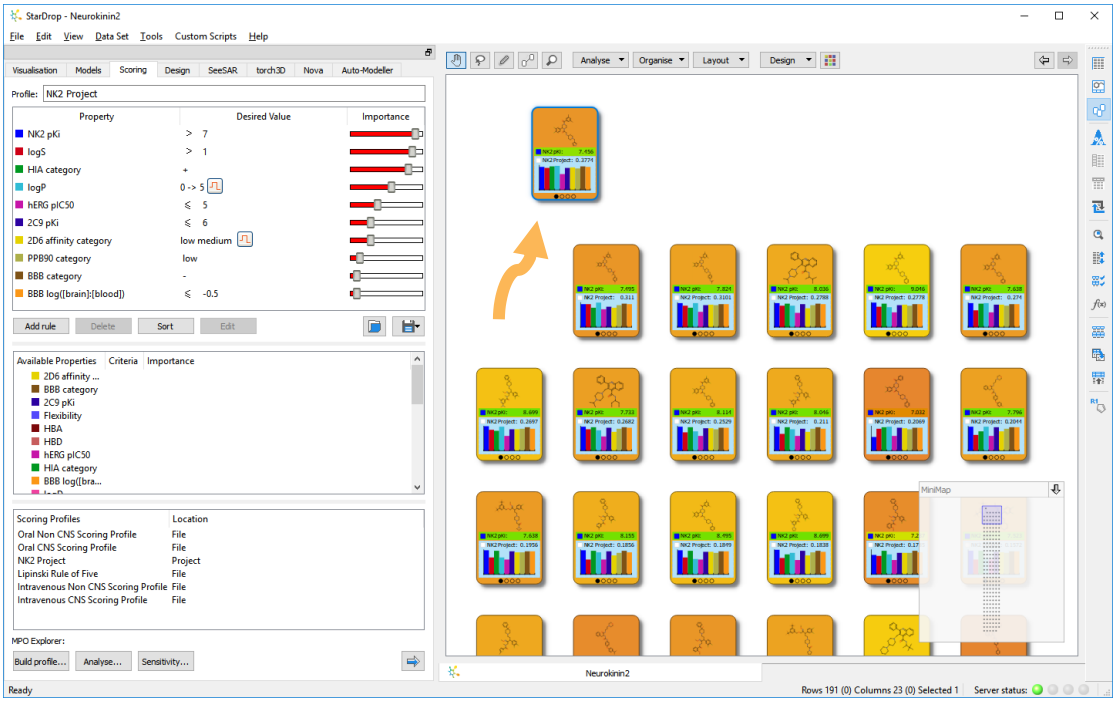


- We can now colour the cards to highlight the most potent, by clicking on the **Format button**  and in the **Colour By** dialogue choose to colour the cards by **NK2 pKi**.

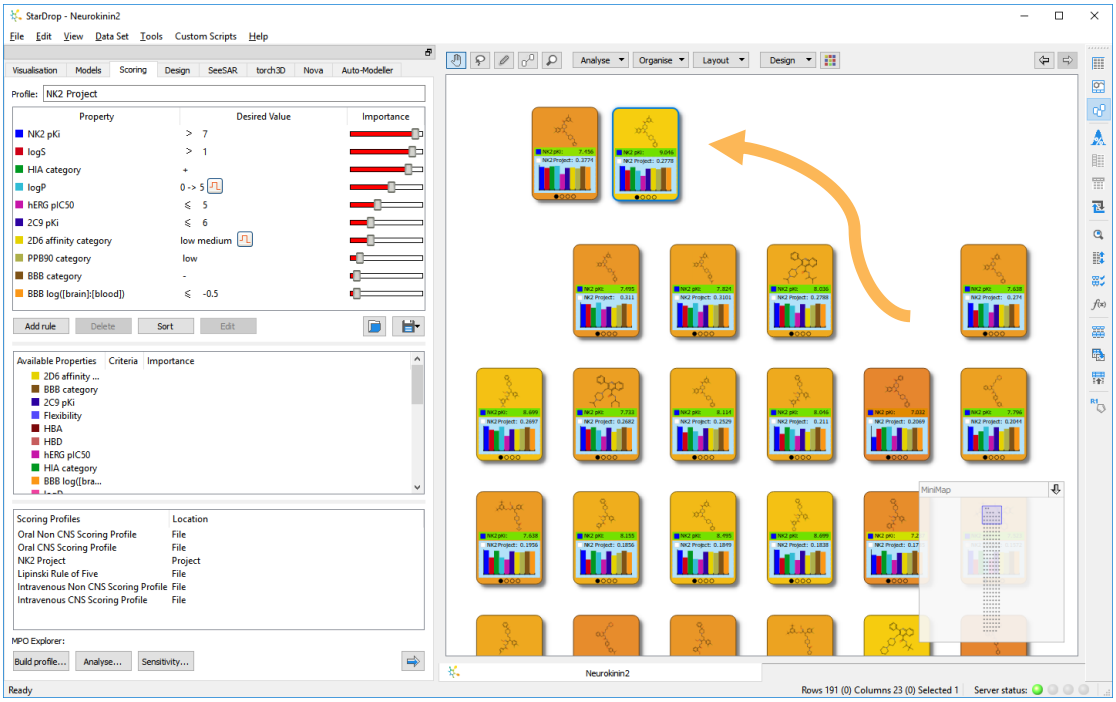


We can easily compare two compounds by placing their corresponding cards side-by-side.

- Click on the highest scoring compound and drag it to the top. You can make space by dragging the background if necessary.



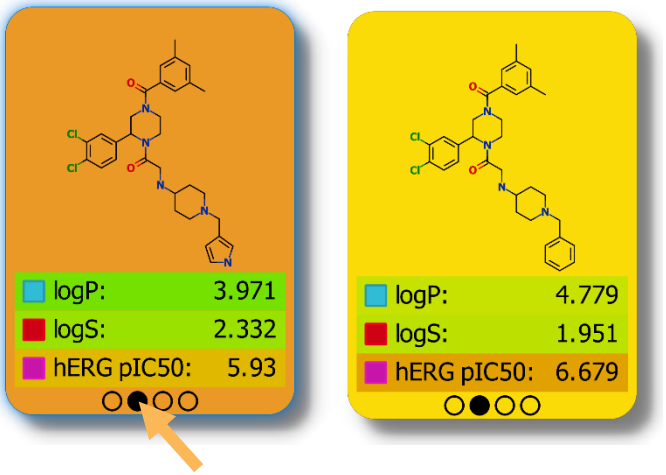
- Now select the fifth compound, which is one of the most potent, and drag it to lie beside the first.



- You can now zoom in to these cards to look at them in more detail by pointing at the cards and using the mouse wheel (or if you don't have a mouse, by holding down the **CTRL** key and pressing the **=** key). Dragging the background will also help you to position the cards conveniently.

Here we can see that the compound on left has a higher score, even though the compound on the right is one of the most potent. We can investigate this further:

- Click on the second circle on the bottom of a card to change to the second 'page'.

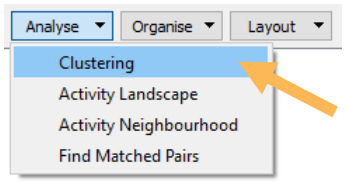


Here we can clearly see that the compound on the right has higher values for logP and hERG pIC_{50}, which increase the risk for this compound, and hence result in a lower score.

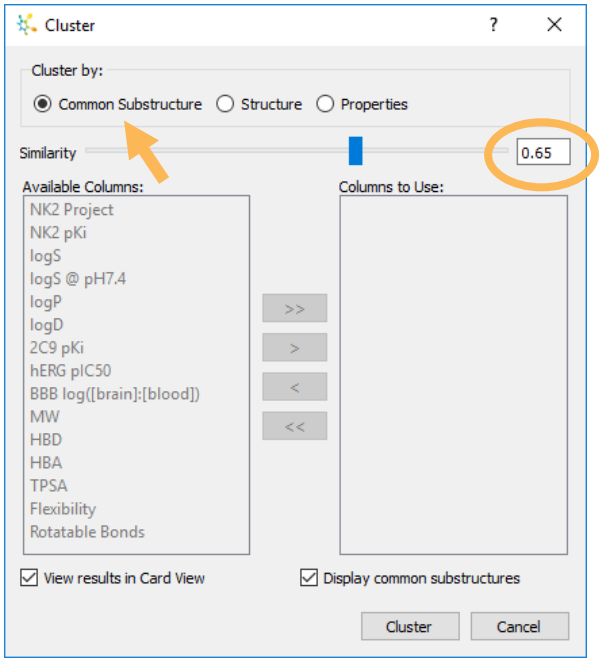
- Click on the fourth circle now to go to the notes page and then double-click on ***Add notes*** to add your own text which will be saved in a **Notes** column in the data set.

Having compared individual compounds, we would also like to explore how the properties vary across the different chemical series in this library. Clustering provides a convenient way to group compounds by chemical similarity.

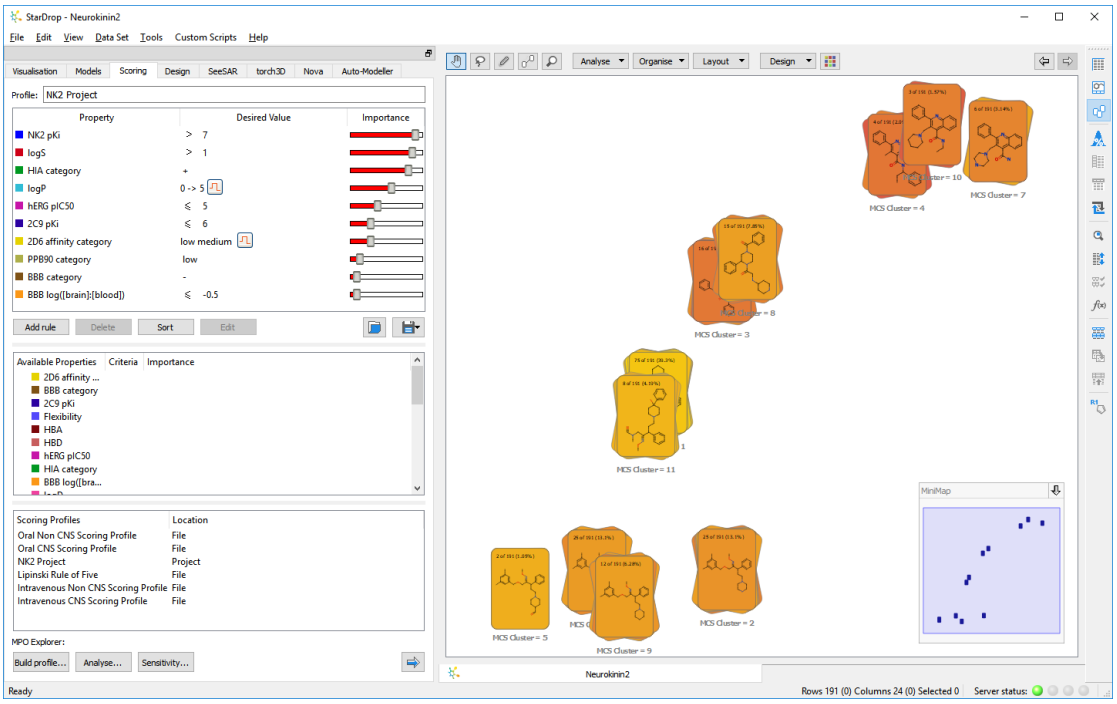
- From the **Analyse menu** at the top of the Card View, select **Clustering**.



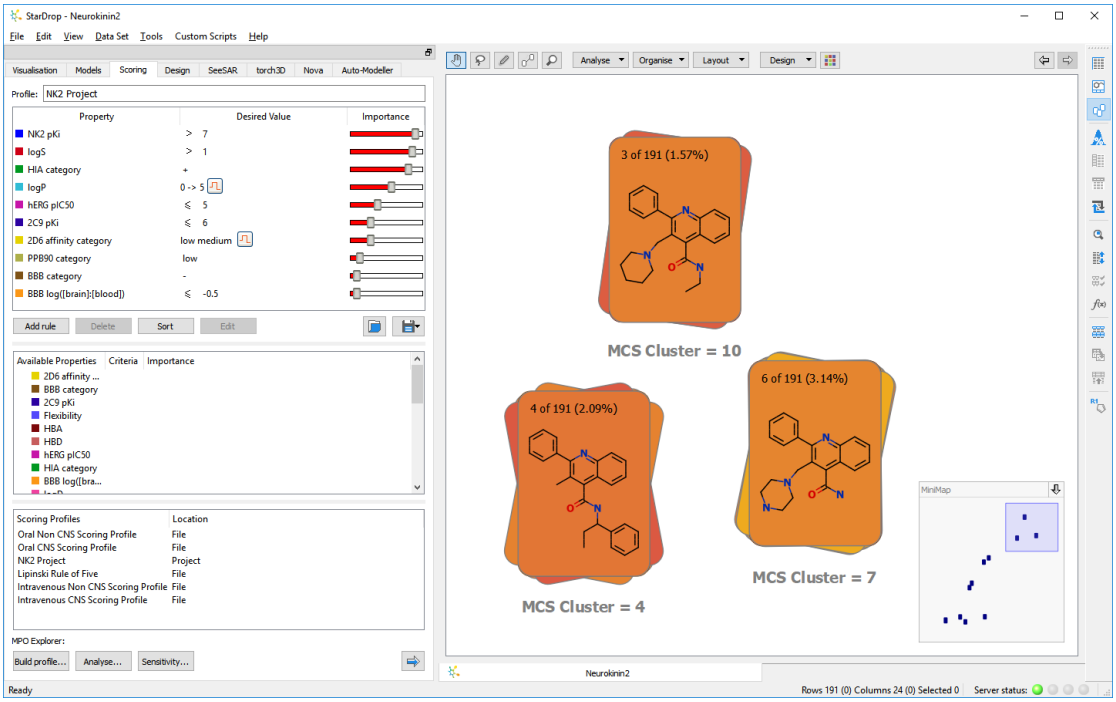
- In the **Cluster** dialogue, we will use the default **Common Substructure** method. Set the **Similarity** to **0.65**, as shown right, and click the **Cluster button**.



The compounds will be grouped by common substructure to form 'stacks' containing multiple compounds. The stacks will be positioned such that stacks with similar common substructures will be close together.

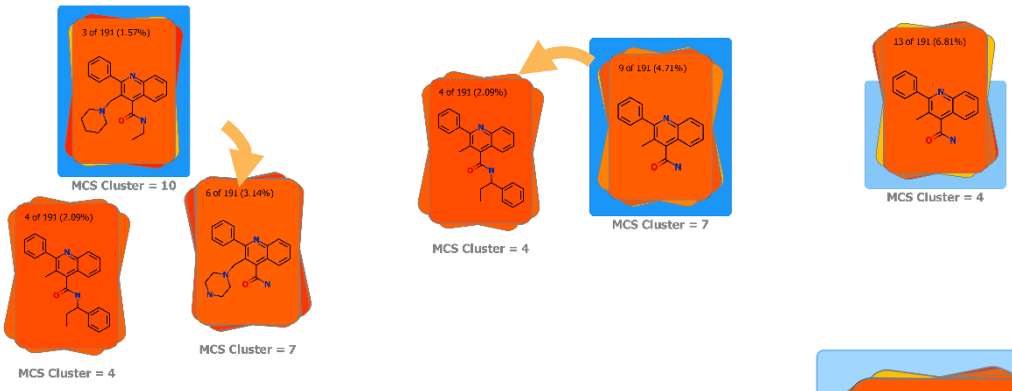


- Zoom in on the stacks corresponding to clusters 4, 7 and 10, which will be close together and drag the background to centre these in the display. You can move the stacks to separate them slightly by clicking and dragging.

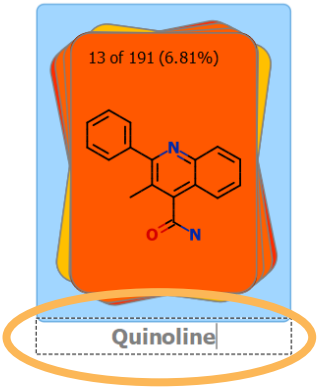


Here you can see that, although the compounds are in different stacks, this is an artefact of the cut-off assigned to the clusters and these groups all share the same quinolone scaffold and could be considered a single series.

- We can easily combine these by dragging one stack on top of another.

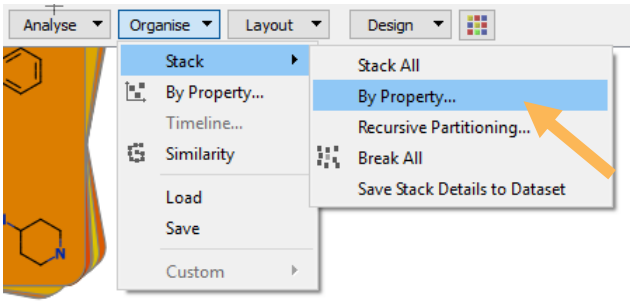


- We can then assign a name to this series by clicking and editing the label under the stack.

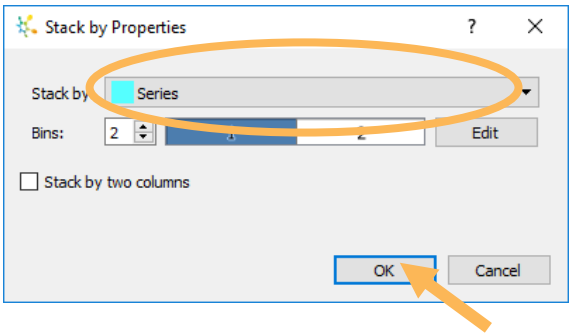


We could examine the other stacks and repeat this process, where necessary, to refine the clustering results. However, to save time, we have saved the series definitions in the data set and can use this to stack all of the compounds.

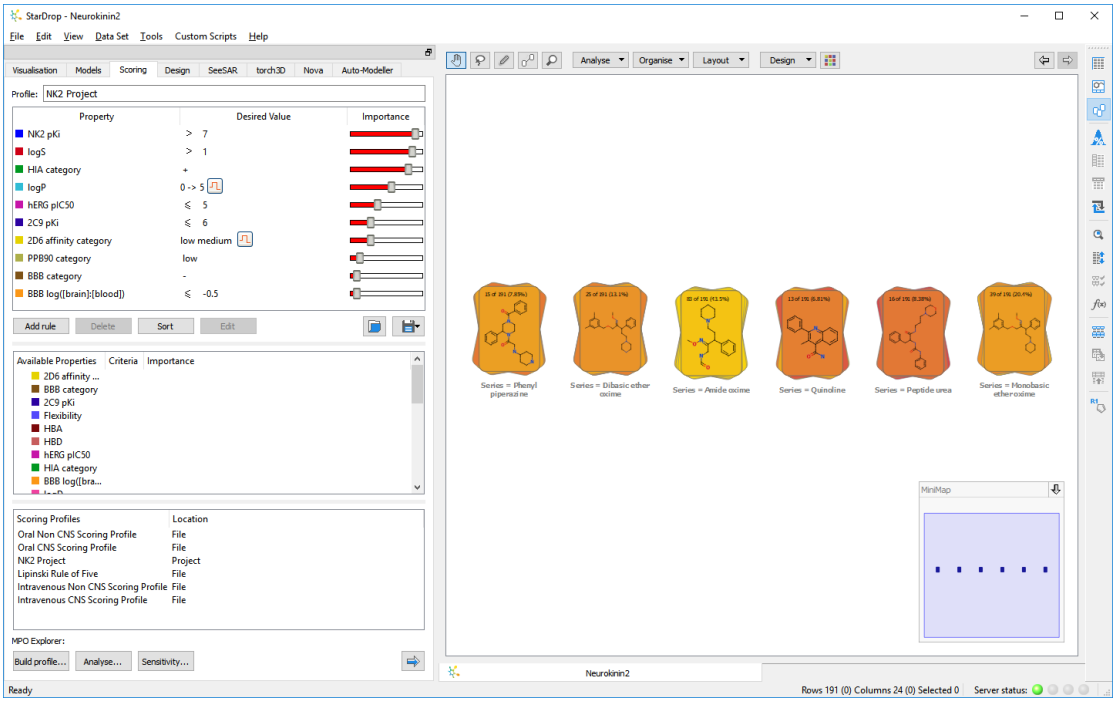
- From the **Organise menu** at the top of Card View, select **By property** from the **Stack menu**.



- In the **Stack by Properties dialogue** choose **Series** from the **Stack by menu** and click the **OK button**.

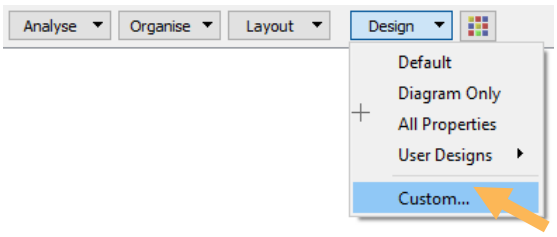


Here we can see that there are 6 series, of which the Amide oxime series has the highest average potency (it is the most yellow).



We can display more information on a stack. To illustrate this, we can choose a new stack design.

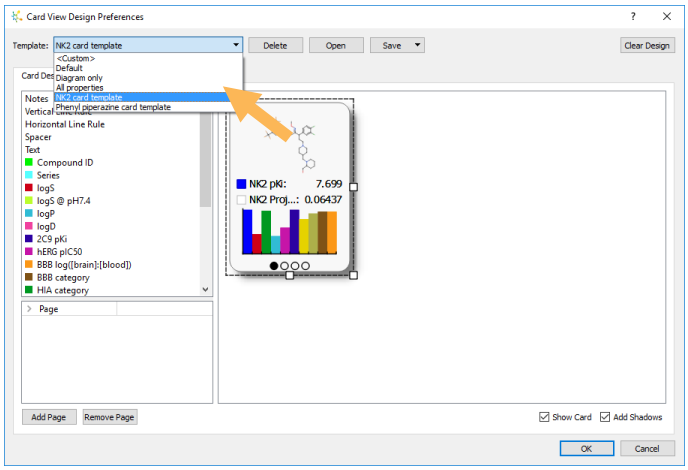
- Select **Custom** from the **Design menu** at the top of Card View.



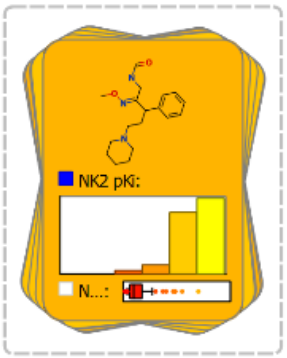
In the **Card View Design Preferences dialogue**, you can change card sizes, choose which properties should be displayed and how the properties should be arranged, but in this case, we are going to choose a pre-defined template saved as part of this project.

- Select **NK2 card template** and click the **OK button**.

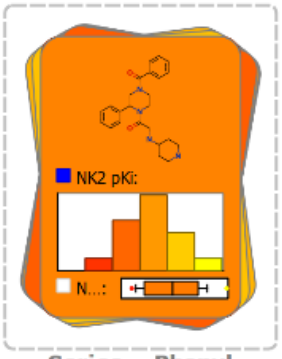
Note: You can also select previously saved designs directly from Card View's **Design menu** under **User Designs**.



The resulting stacks show a histogram of the distribution of pK_i values of the compounds within each stack and, below this, a box plot showing the distribution of scores. This makes it straightforward to compare different series, for example the Amide oxime series has an excellent distribution of potency; however, most scores are very low due to poor results for other properties. In contrast, the Phenyl piperazine series has fewer potent compounds but a significant proportion of high scoring compounds, suggesting that this would be a good series to investigate further.



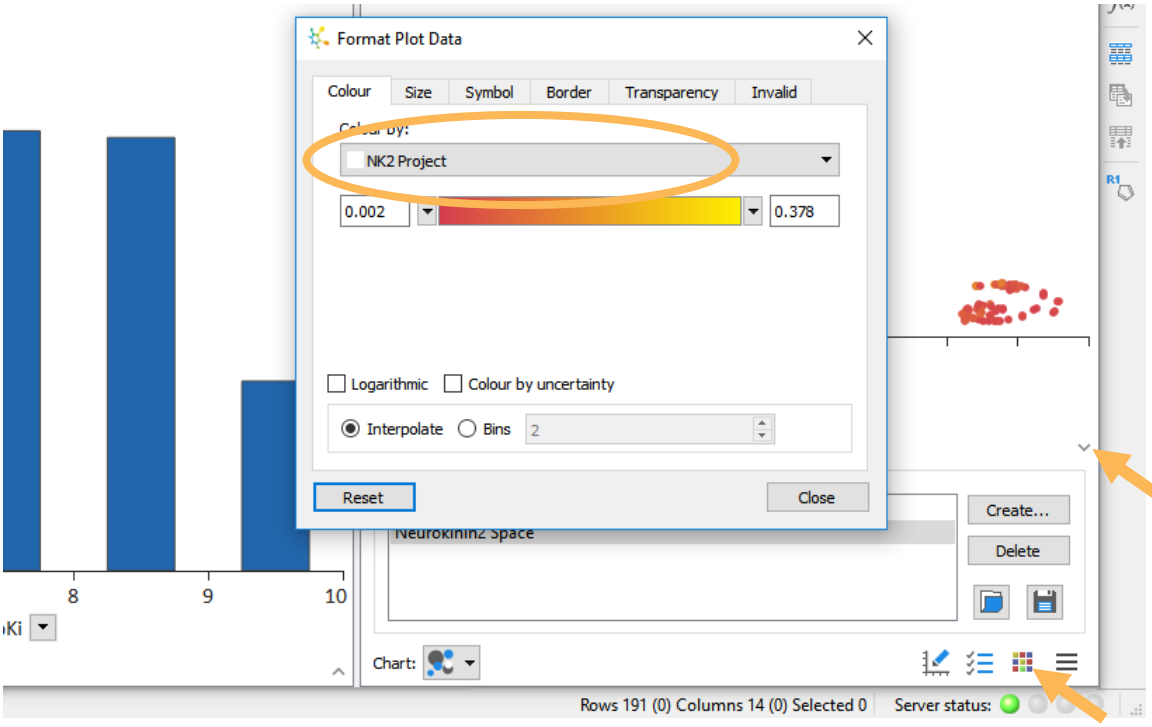
Series = Amide oxime



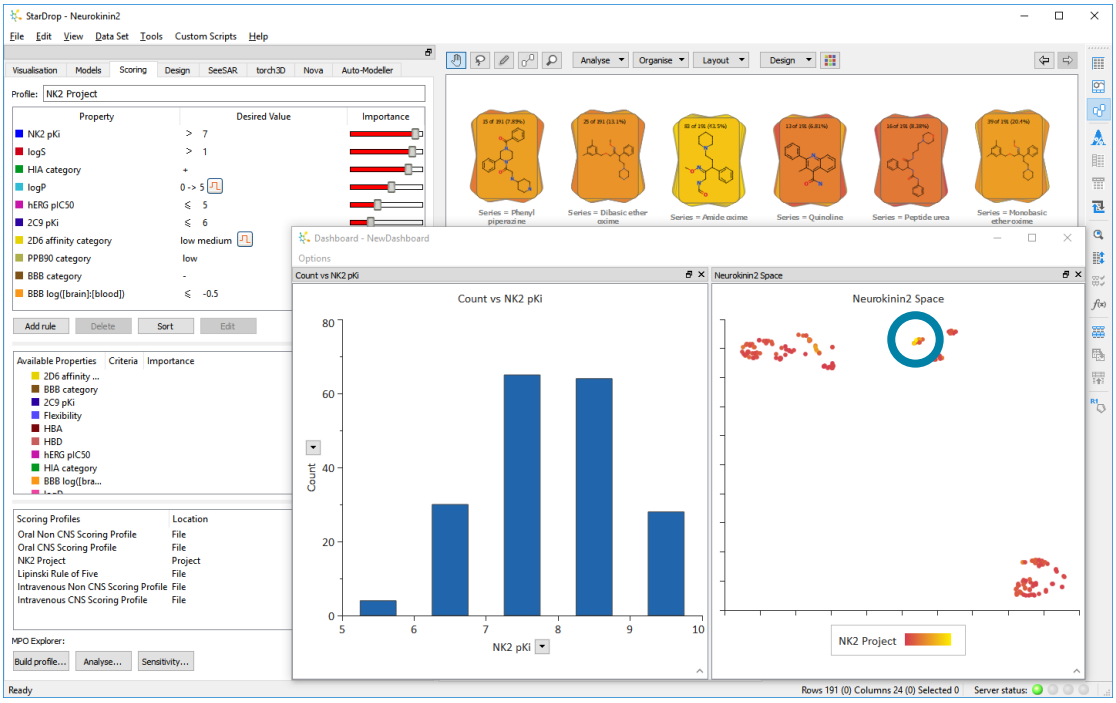
Series = Phenyl piperazine

We can explore the distribution of scores across the whole data set by returning to the dashboard we created earlier and formatting the chemical space.

- Click the arrow in the bottom corner to display the controls, click on the **Format button**  at the bottom to bring up the **Format Plot Data dialogue** again.

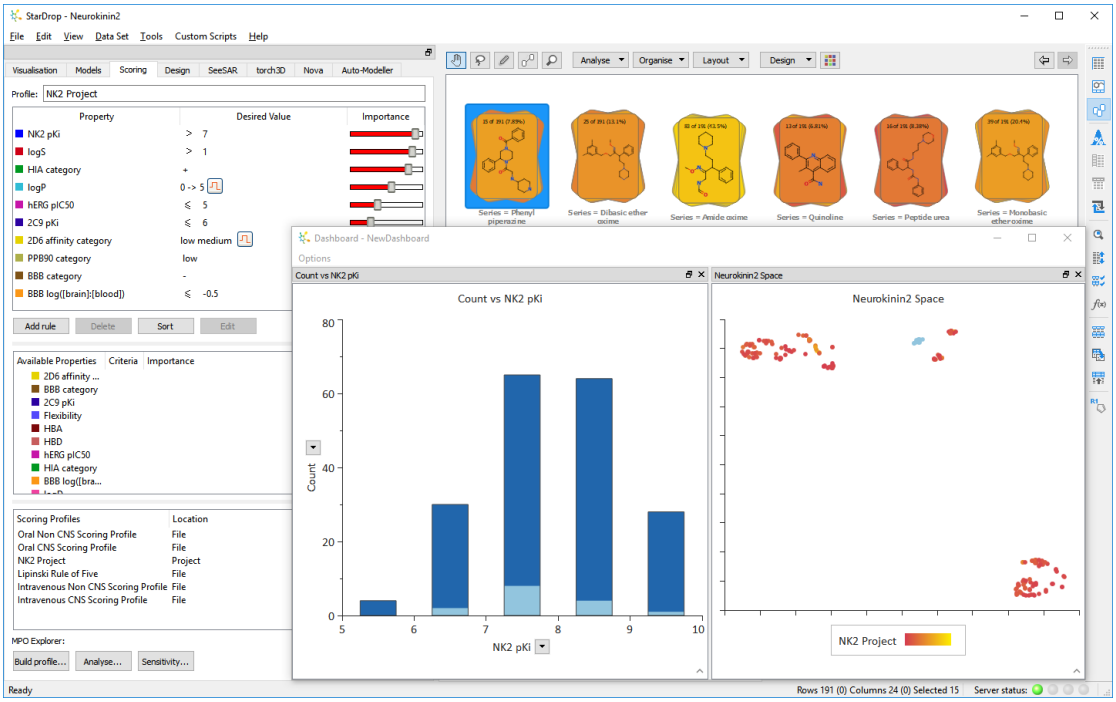


- Change the **Source** property to be the **NK2 Project** score values, before clicking the **Close button**.



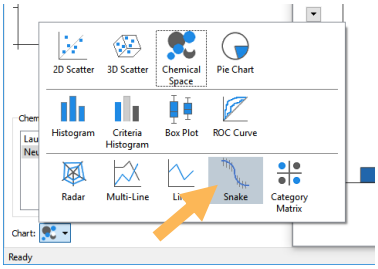
The points are now coloured based upon the overall scores of the compounds, not just the pK_i values. We can see that the pattern has changed dramatically from that which we could see when looking only at potency. There is just one small cluster of yellow points (highlighted by the blue circle) as a 'hot spot' of high scoring compounds.

- Click with the left mouse button and draw around this yellow cluster in the chemical space. You will see that the stack corresponding to the Phenyl piperazine series is highlighted, confirming that the hot spot corresponds to this series.

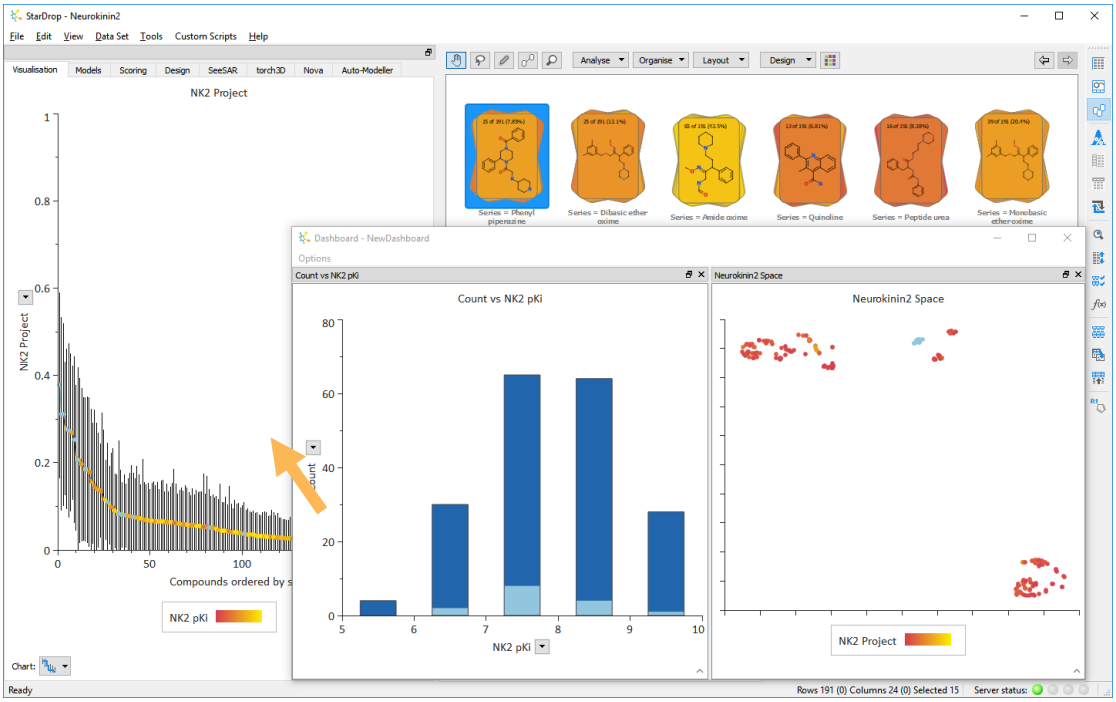


Although this appears to be the best series, we should consider the impact of the uncertainty in the underlying data on our ability to confidently distinguish between compounds. Therefore, we will use another type of chart to explore this.

- Go to the **Visualisation area** again.
- Now open the **Chart menu** at the bottom and choose a **Snake** plot (as shown right).

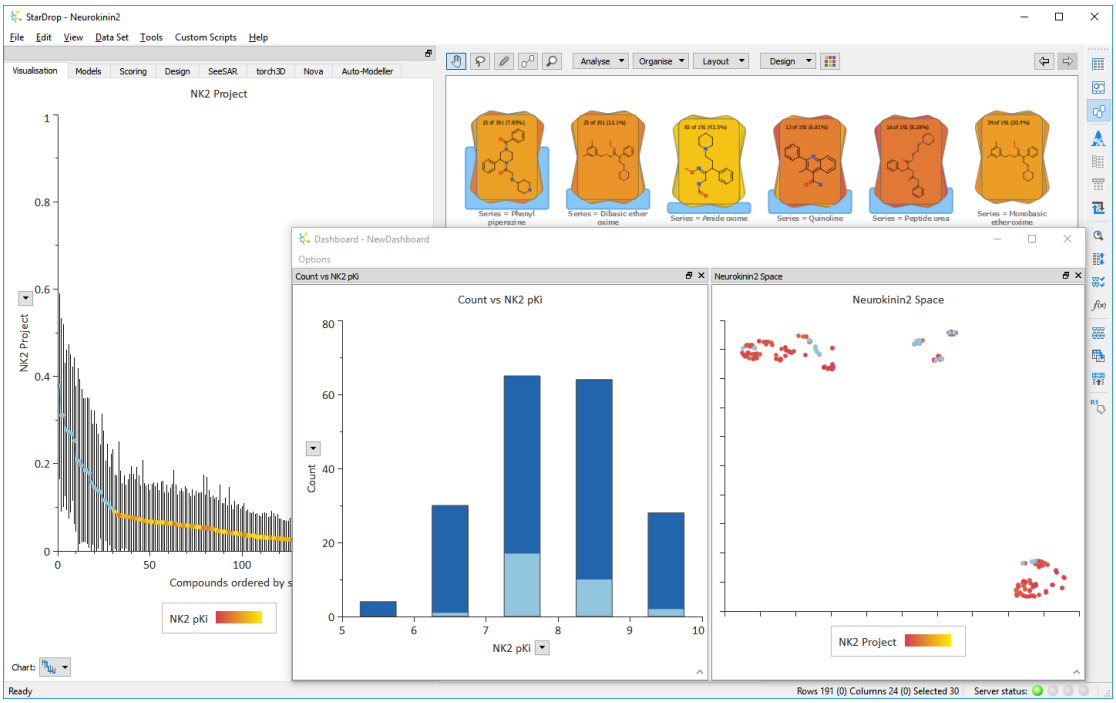


A 'snake plot' shows the scores (on the y-axis) for all compounds in order from highest score to lowest score (along the x-axis). The overall uncertainty in each score, due to the uncertainty in the underlying data, is also displayed as an error bar around each point.



The top 30 compounds in the snake plot cannot be confidently distinguished from the top scoring compound (notice that the error bar for the top-scoring compound overlaps with the error bars of approximately the top 30 compounds). Therefore, we should consider exploring the properties of the compounds in this range further so that we can make a confident selection of a potential lead series.

- Select the top 30 compounds in the snake plot by drawing a ring round them with the mouse.



Notice that this selection includes compounds from different regions of the chemical space and a small number of compounds from every stack except the Monobasic ether oxime series. This suggests that some of these chemistries cannot be rejected with confidence. Therefore, it may be more appropriate to sample a small number of compounds from some of these alternative chemistries to generate some experimental ADME data. These data will have lower uncertainty than predicted values enabling us to identify with greater confidence the chemistries that will yield a high-quality lead series.

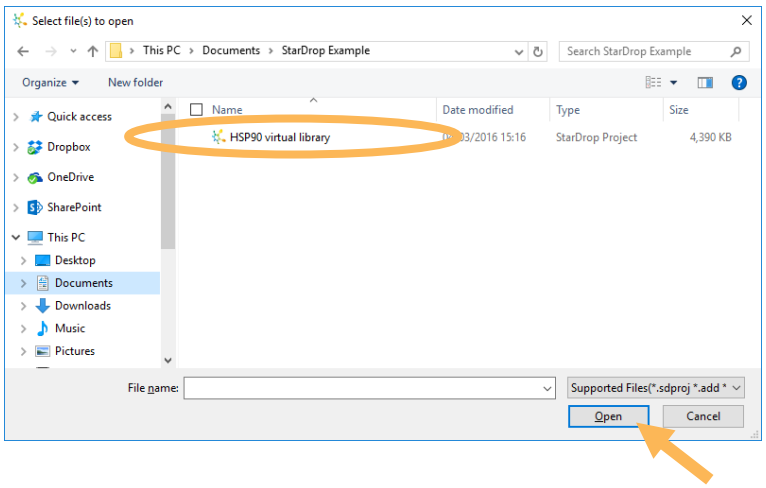
Linking 2D and 3D SAR to Guide Design

This example uses a combination of 2D and 3D methods to understand and optimise a virtual library of Heat Shock Protein 90 (HSP90) inhibitors. The library, created by a *de novo* design process, is based around an amide substitution on a beta resorcylic acid core. The objective in this example is to use the SeeSAR™ module to develop an understanding of the 3D structure-activity relationships (SAR) and then use multi-parameter optimisation to further optimise the absorption, distribution, metabolism and excretion (ADME) and physicochemical properties of a potent inhibitor, without losing efficacy.

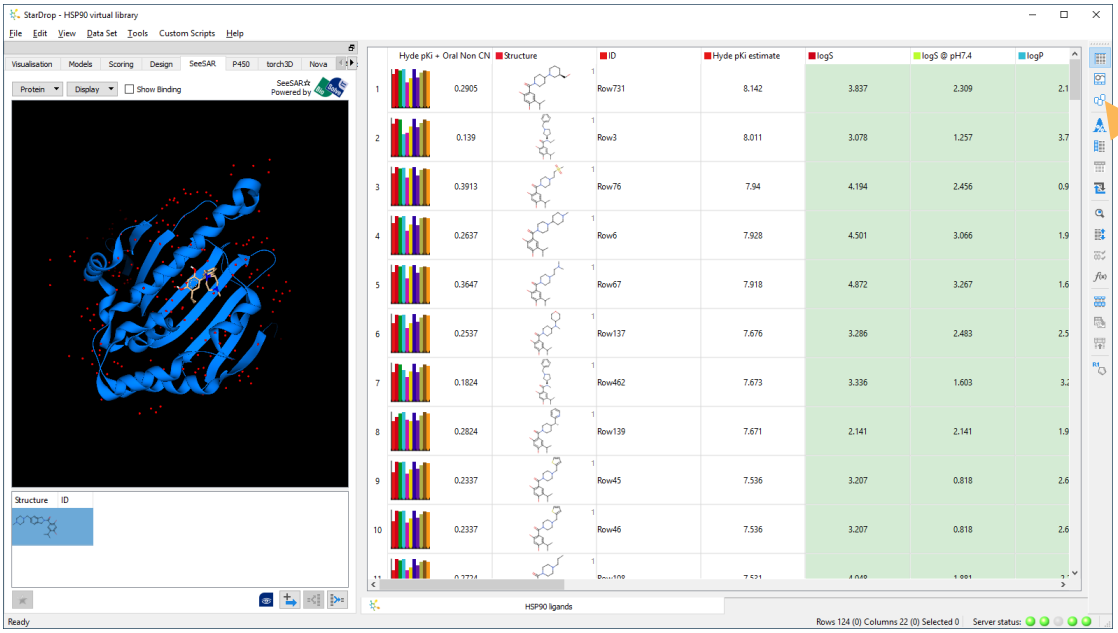
During this exercise, we will use a variety of StarDrop's capabilities to explore the data in order to understand the SAR and design compounds with a good balance of properties. Step-by-step instructions for all the features you will need to use in StarDrop are provided, along with screenshots and examples of the output you are likely to generate. If you have any questions, please feel free to contact stardrop-support@optibrium.com.

Exercise

- In StarDrop, open the project file **HSP90 virtual library.sdproj** by selecting **Open** from the **File menu**.



You will see a spreadsheet containing 124 structures with their estimated affinities for HSP90 (in the column **Hyde pKi estimate**).

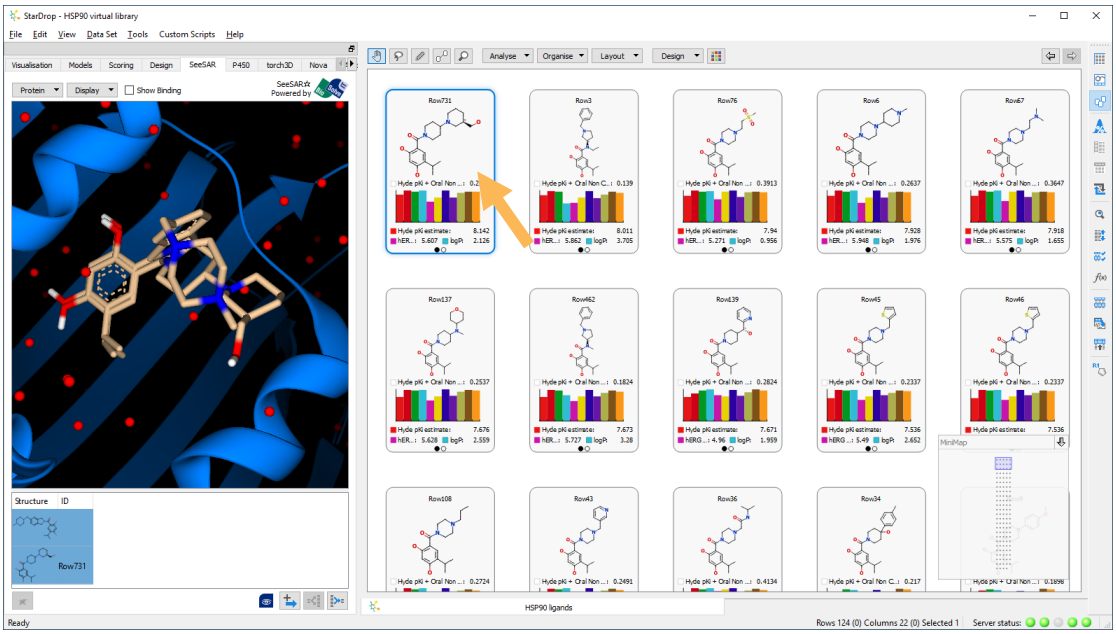


We'll use Card View again to explore the SAR in the results from the virtual screen.

- Click the **Card View button** on the right-hand toolbar to switch into Card View.

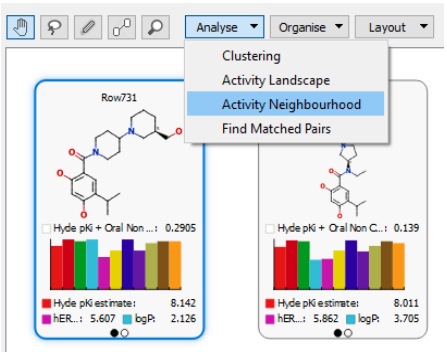
The cards are displayed in a grid in the order of estimated pK_i from top-left to bottom-right.

- Click on the top-left card, which is estimated to be the most active.

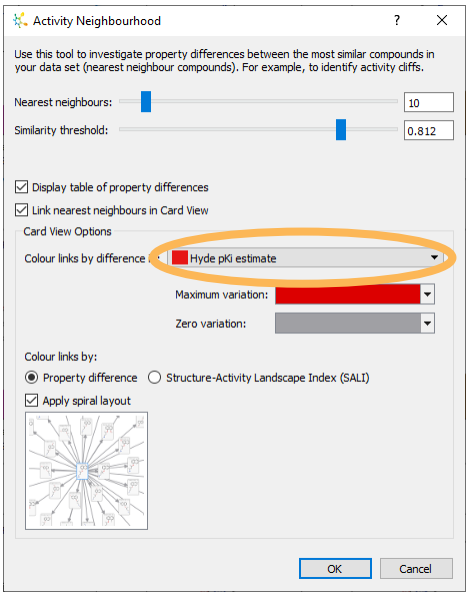


We're going to explore the SAR of the nearest neighbours to this compound. An Activity Neighbourhood analysis compares each compound with the selected 'reference' compound and arranges these in order of structural similarity. The nearest neighbours are linked to the reference and that link can be coloured by the change in a selected property.

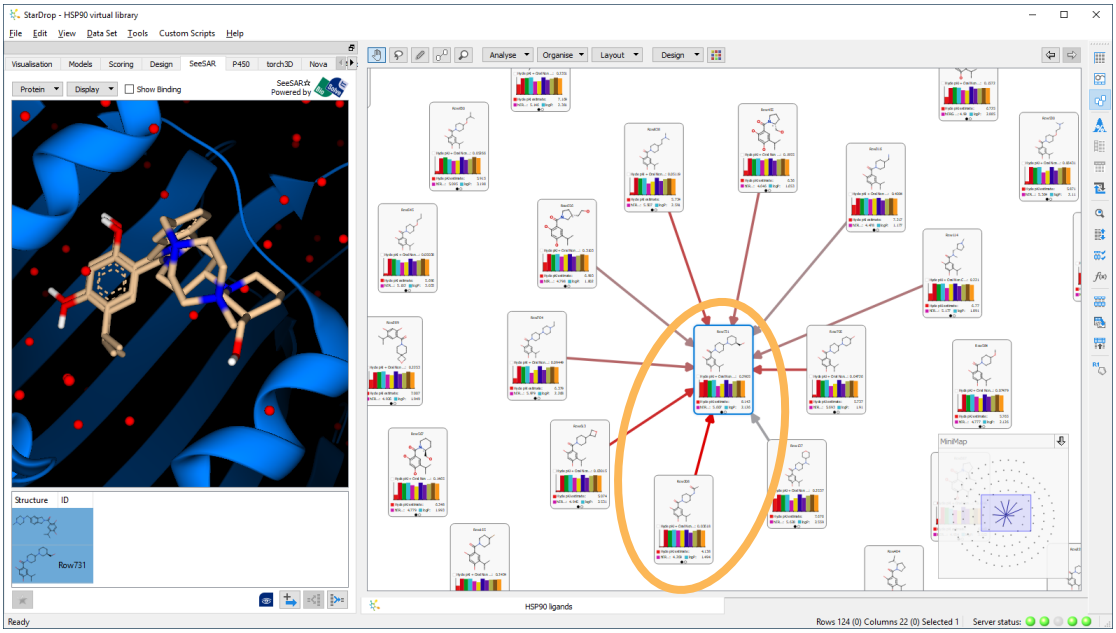
- In the **Analyse** menu at the top of Card View, select the **Activity Neighbourhood** option.



- In the **Activity Neighbourhood** dialogue, choose **Hyde pKi estimate** in the **Colour links by difference in** menu.



The resulting display shows the reference compound in the centre and the remaining compounds arranged in a spiral in order of decreasing similarity (increasing distance) from the reference. The 10 closest compounds are joined to the reference by a link; the arrow indicates the direction of *increase* of the selected property, in this case Hyde pK_i estimate. The colour indicates the magnitude of the change, in this case from red being the largest to grey being the smallest. Therefore, a short link with a bright, red colour indicates an activity cliff, i.e. a small change in structure that gives rise to a large change in activity. One example of this can be seen between the reference compound and its third nearest neighbour, to the bottom left, **Row 208**.



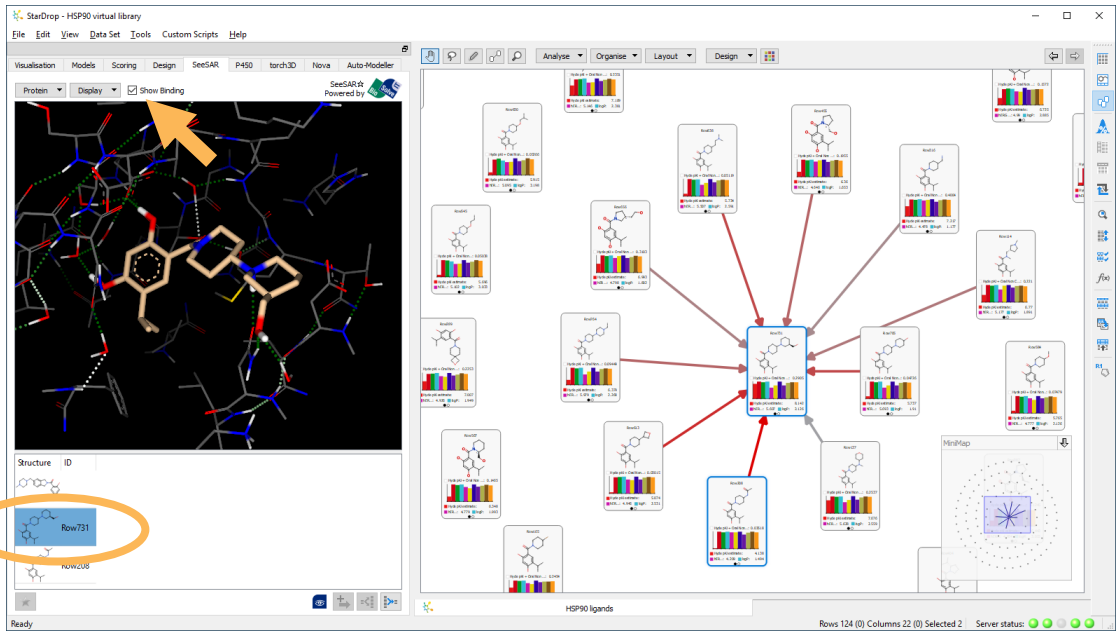
Note that a table summarising the property changes for the nearest neighbours will also be produced, but we won't use it in this example, so you can close it if you wish.

Select the card corresponding to **Row 208** in addition to the reference compound, **Row 731**, by clicking on it while holding down the **CTRL** key.

The PDB file 2XJX has already been loaded into this StarDrop project and can be seen in the **SeeSAR area** to the left. Below the protein you will see a table which contains three ligands. The first is Onalespib, an HSP90 inhibitor which was included in the PDB file. The other two ligands are the pair we selected in Card View. Any ligands we select in our main data set will be available in the table below SeeSAR for viewing in the protein.

Note that you can make the tabbed area wider by dragging the grey bar between the SeeSAR and Card View areas.

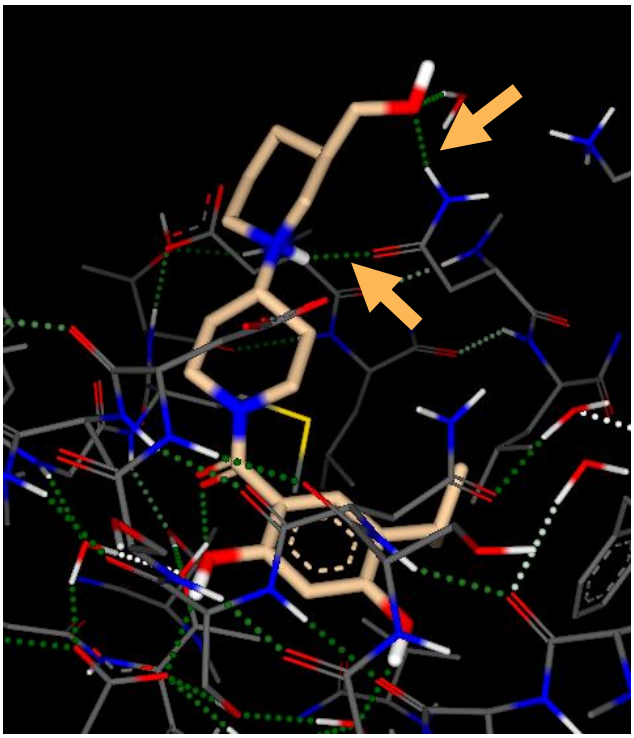
- Select the second row in the table below SeeSAR – this is the ligand we identified with the highest estimated activity (Row 731).
- Now click the **Show Binding checkbox** above the protein to change the view of the protein to focus on the binding pocket.



Using the mouse, you can interact with the view of the protein:

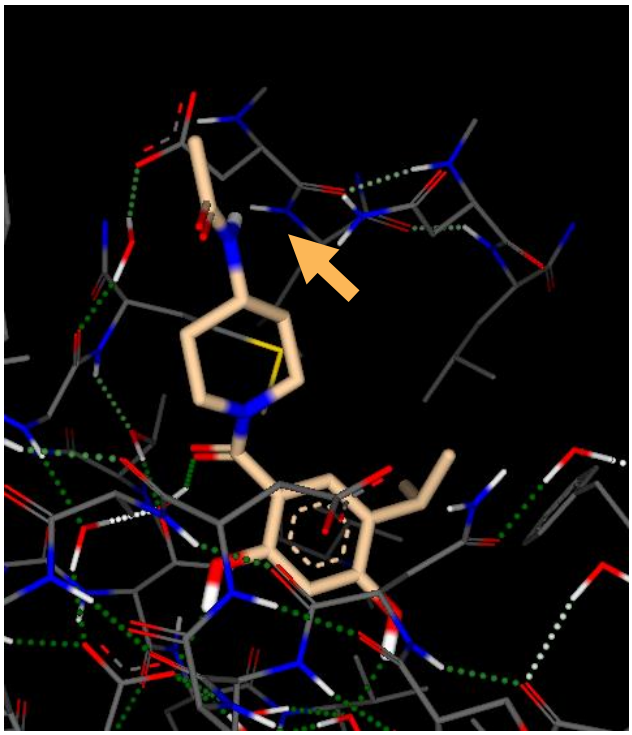
- Use the mouse-wheel to zoom in and out
- Use the left-mouse button and drag to rotate the view
- Use the right-mouse button and drag to pan the view

When you choose the **Show Binding** option, hydrogen bond interactions between the ligand and protein are displayed by green dotted lines. Focussing in particular on the hydroxy-methyl piperidine portion of the ligand we can see its bi-dentate interaction with HSP90.



- Click on the bottom row of the table below the protein to display the N-acetyl derivative instead (Row 208).

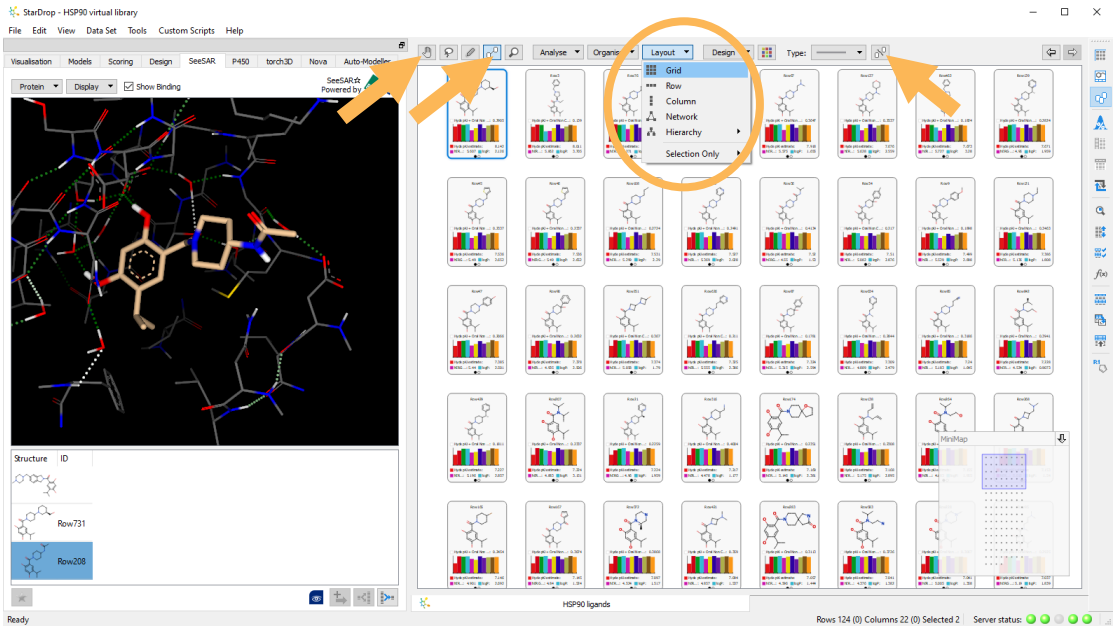
Here we can see that there are no equivalent interactions between this part of the ligand and the protein which suggests that this may be responsible for the significant reduction in the estimated activity.



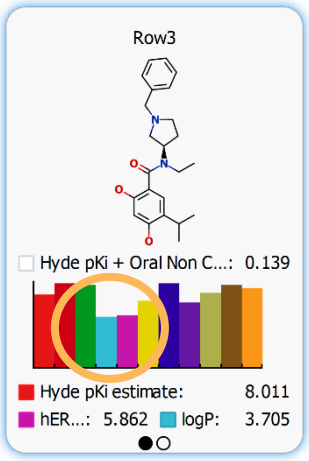
Having considered some of the reasons underlying the estimated activity of the ligands in this virtual library, let's take a look at how we can combine our view of the 3D SAR with QSAR models to find strategies to design potent compounds with a good balance of ADME and physicochemical properties.

Let's begin by cleaning up the display in Card View.

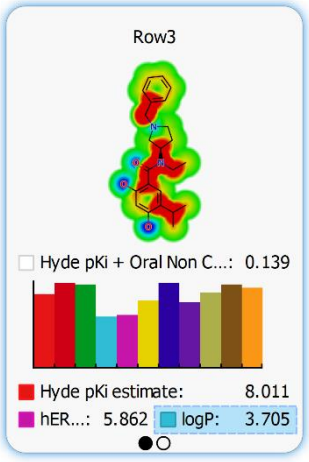
- Click on a space somewhere between cards to deselect everything.
- Zoom in to make the cards a little larger using the mousewheel or the **Ctrl** and **=** keys.
- Click on the **Link button**  to show the link tool and then click the **Remove Links button**  to remove the network.
- Now select **Grid** from the **Layout menu** at the top of Card View to tidy the cards into a grid.
- Finally, click the  button at the top of Card View to change back to the general mode in which we can move cards around.



In Card View, zoom in to look at the properties of the second compound, called **Row3** (use the middle mouse wheel or **Ctrl** with **-** and **=** keys to zoom in and out. The zoom will be centred on the mouse position.)



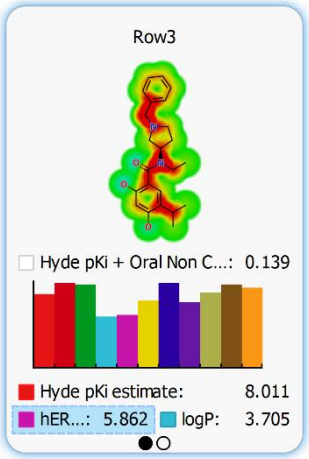
In the scoring histogram we can see that the two lowest bars, which indicate potential problems, are cyan and pink. These correspond to logP and hERG pIC_{50} respectively and the values of these two properties are shown below. When scoring the compounds we indicated that we would ideally like a logP value between 0 and 3.5. We also indicated that an ideal hERG pIC_{50} value would be below 5.



StarDrop's Glowing Molecule™ highlights regions of the molecule that are having a significant effect upon a predicted property. This can help to guide the design of new compounds by indicating regions where a small change in structure may have a significant impact on the predicted property.

- Click on the logP prediction on the card to display the Glowing Molecule representation for logP.

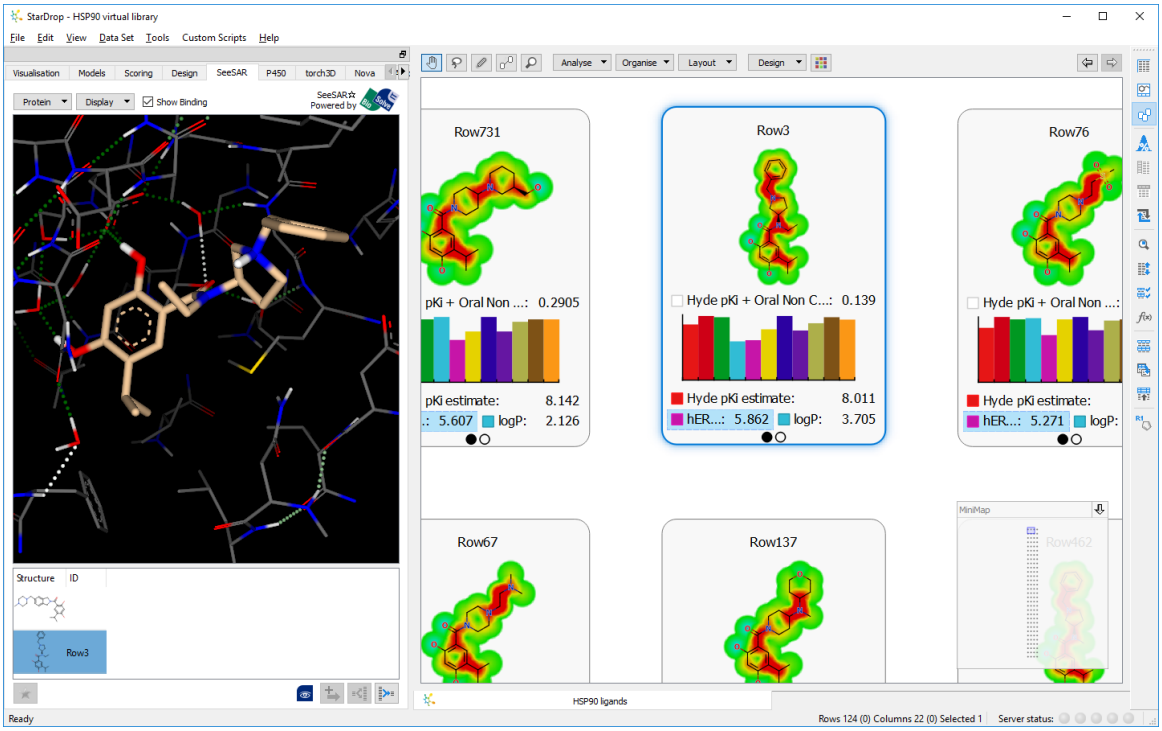
Regions glowing red are tending to increase the predicted value whereas regions glowing blue are tending to decrease the predicted value.



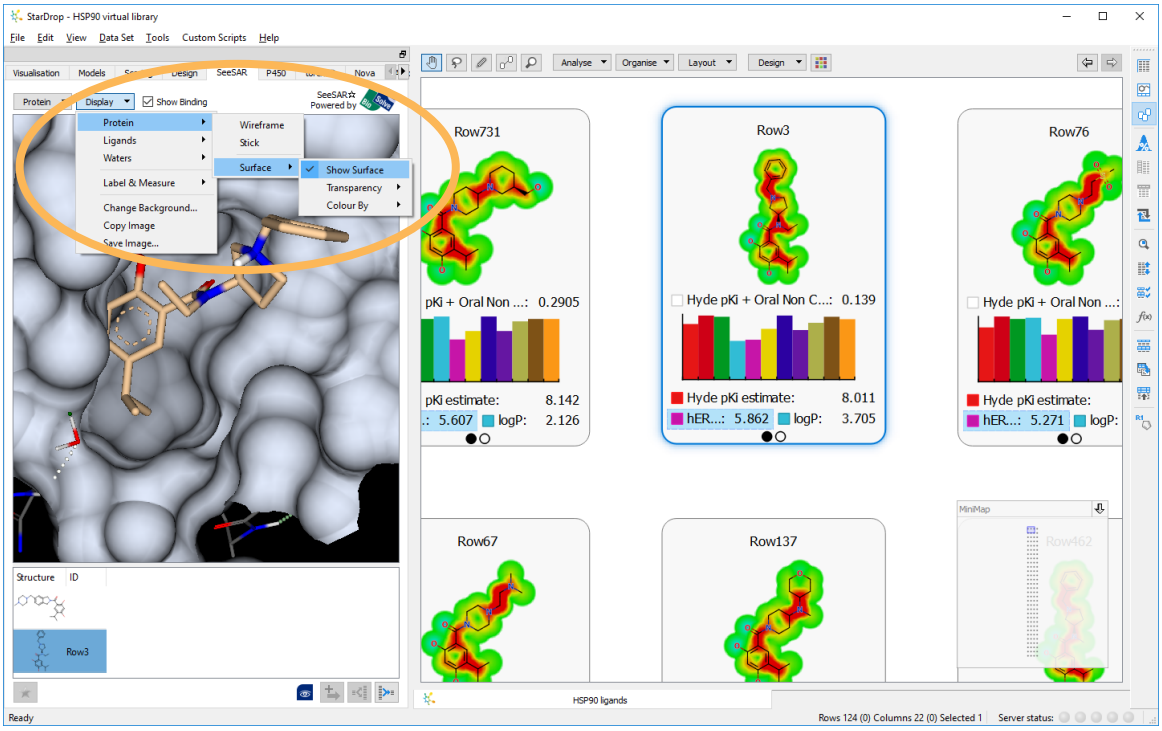
Now select the hERG pIC50 property to view the Glowing Molecule representation for this prediction.

For both the logP and hERG pIC_{50} we can see that regions around the beta resorcylic, phenyl and pyrrolidine rings are red, indicating that these parts of the molecule are contributing to an increase in these properties.

- We can now easily compare the 2D SAR indicated by the predictive models with the 3D SAR indicated by the docking pose. Change back to the **SeeSAR** area and you will see this compound shown in SeeSAR.

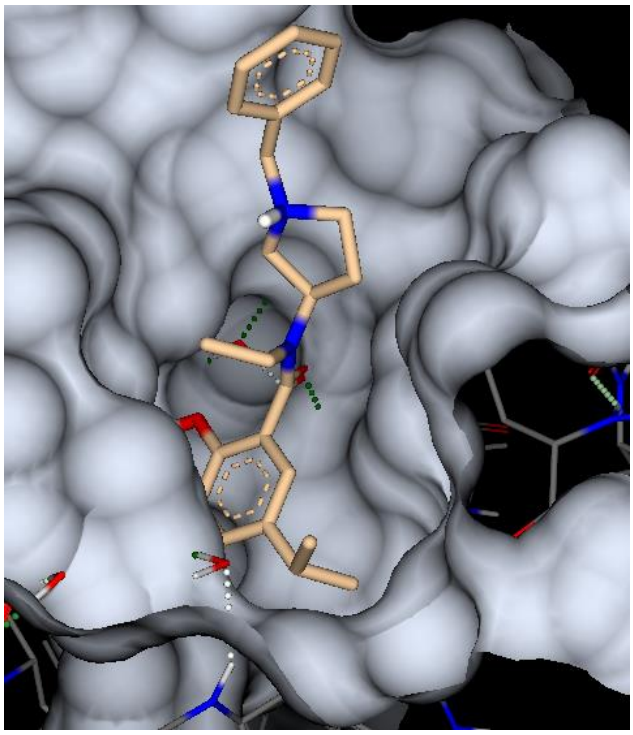


- From the **Display menu** at the top of the SeeSAR area choose **Protein**, **Surface** and then **Show Surface**.



- Rotate the structure within the SeeSAR area to see how the ligand fits into the binding site.

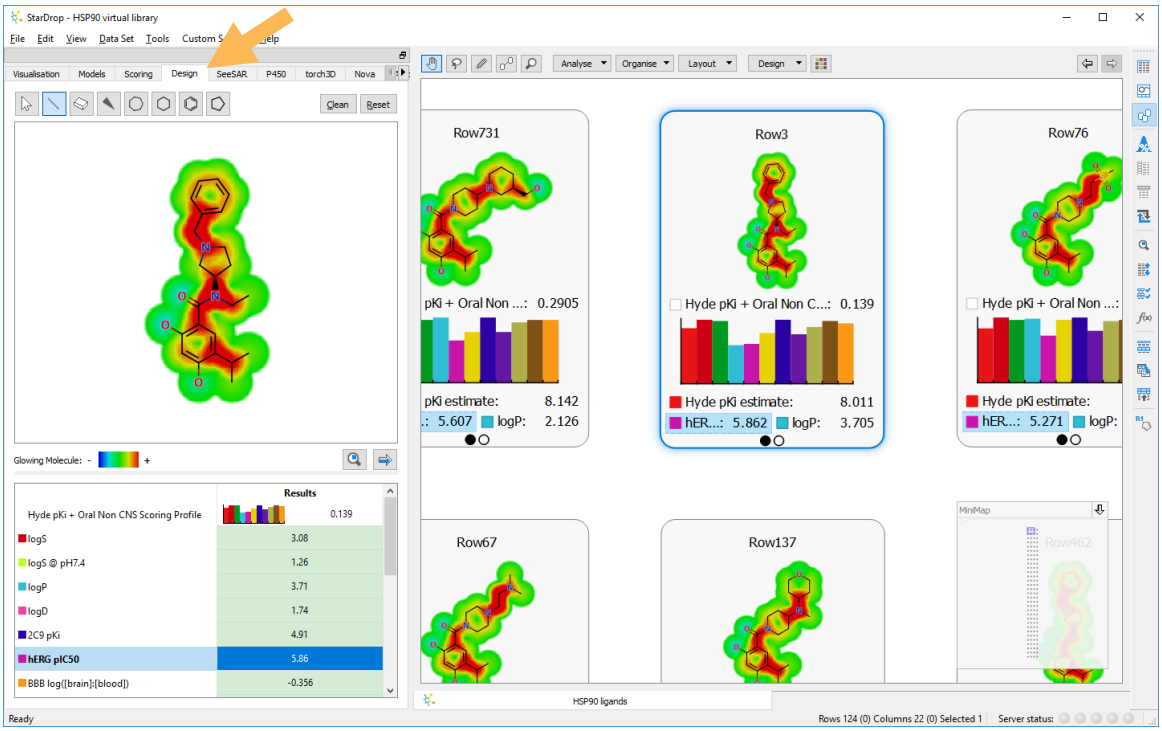
Whereas the resorcylic acid core sits deep within the pocket and forms a strong hydrogen-bonding network with the protein, the phenyl and pyrrolidine rings sit outside the pocket, showing no obvious strong interactions.



This suggests that these regions may not contribute strongly to the compound's binding affinity, so these rings may well be a good place to consider making modifications in order to mitigate the risks of high logP and hERG inhibition without having a large affect on activity.

We can use StarDrop's interactive design tool to explore the impact of possible changes.

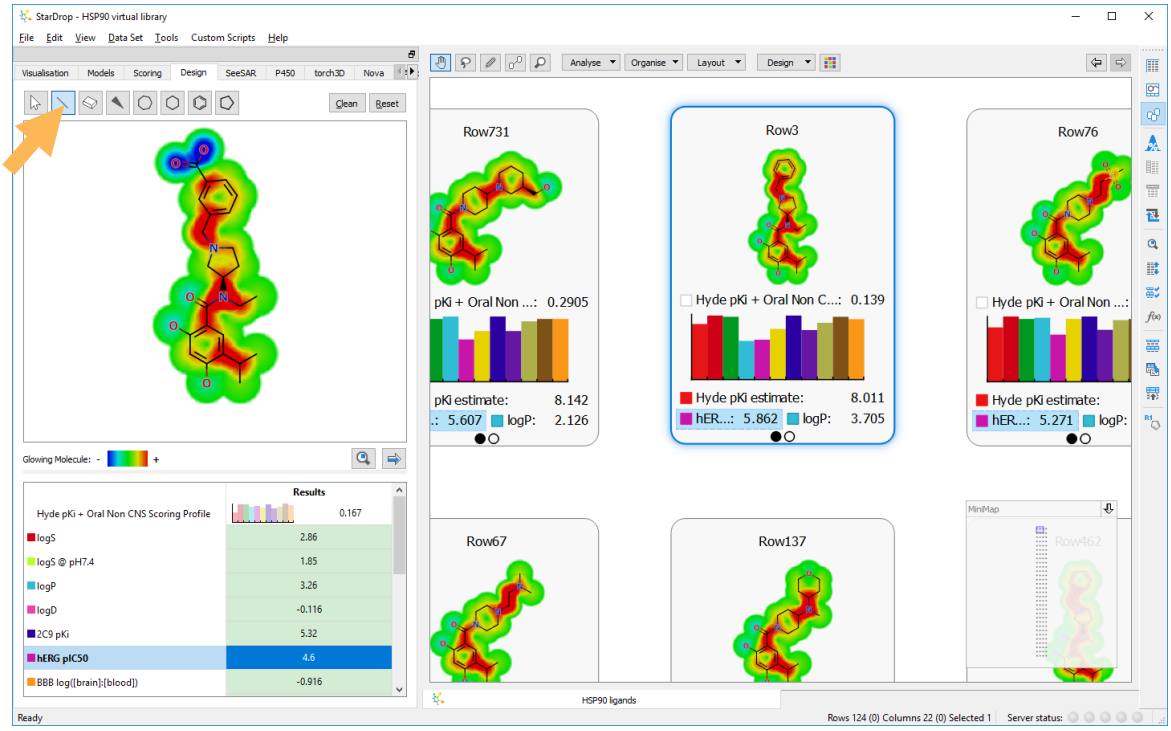
- Change to the **Design area**.



Here we can edit and design new compounds and the selected molecule, Row3, is already shown. Below the molecule editor you can also see predicted properties which will update as we modify the compound structure.

- Click on the **Bond tool**  and add a carboxylic acid at the meta position on the phenyl ring.

Note that to draw a bond you simply hover the bond tool over the attachment point and then click and drag. To draw a heteroatom, hover the bond tool over the atom you wish to modify and type the element symbol. To change to a double bond, click the bond tool on the bond you wish to turn into a double bond. To undo any changes you make type **Ctrl-Z**.



While editing the molecule you will see that the properties below the design area will update to show new predicitons for the displayed molecule. If a property in the list is selected then you can also see the Glowing Molecule for that propety update while editing. The carboxylic acid derivative has a logP of 3.26 and a hERG pIC_{50} of 4.6 and so this compound would appear to have a better balance of ADME properties (you can see all the other properties as well to confirm that none of these have become significantly worse).

- Add this new molecule to the data set by clicking the  button.

We can continue to investigate other alternatives and to find out the effect any changes have had on the predicted activity we could use the SeeSAR Pose module in StarDrop or StarDrop's Pose Generation Interface to link StarDrop with your preferred docking program to generate poses and score the resulting fit to the binding pocket. For an example of the SeeSAR Pose moodule, go to

http://www.optibrium.com/community/videos/introduction-to-stardrop-modules-and-features/469-seesarposemodule

and for an example of the Pose Generation Interface, go to:

http://www.optibrium.com/community/videos/introduction-to-stardrop-modules-and-features/375-pgi

This example illustrated how we are able to combine our understanding of the 2D and 3D SAR to suggest improvements to compounds from a virtual library designed as inhibitors for HSP90. StarDrop contains many more features for selecting and designing compounds and further worked examples are available from the Optibrium community at http://www.optibrium.com/community/tutorials. If you have any questions or feedback, please contact stardrop-support@optibrium.com and we will be happy to help.