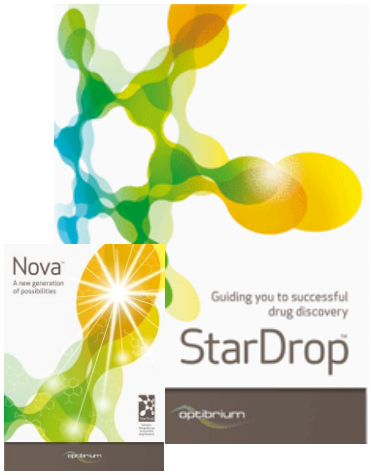


Worked Example:

Generating New Optimisation Ideas Using Matched Series Analysis

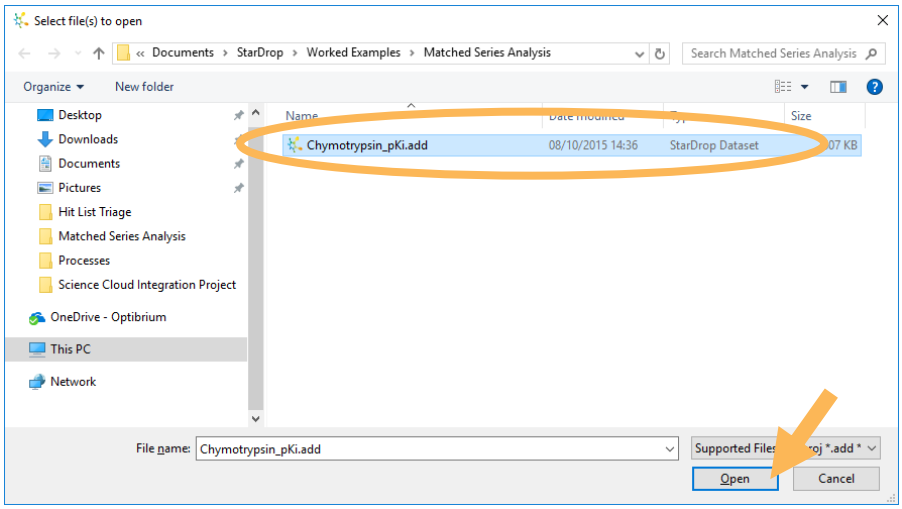
The objective in this worked example is to identify new derivatives that are likely to improve activity at their target, given the SAR already generated on a project. This example uses a publically available set of Human Chymotrypsin K_i data and searches the ChEMBL pIC_{50} knowledge base (generated by NextMove Software) to find matched series that indicate new substitutions with a high likelihood of improving the binding at Chymotrypsin.

A matched series is a series of compounds that are identical except for different substituents at a single point (see Section 10.3 of the StarDrop Reference Guide for more details). The suggestions are derived from comparing matched series in the input data with those in a knowledge base, which are measured across diverse target proteins. The suggestions are based on the premise that a matched series with similar activity order in the input data set and the knowledge base implies that those groups occupy a similar binding environment created by their target proteins. Given a similar binding environment, groups that have been shown to be better binders within the knowledge base, have a strong likelihood of being better binders to the target behind the input data set, in this case Chymotrypsin.

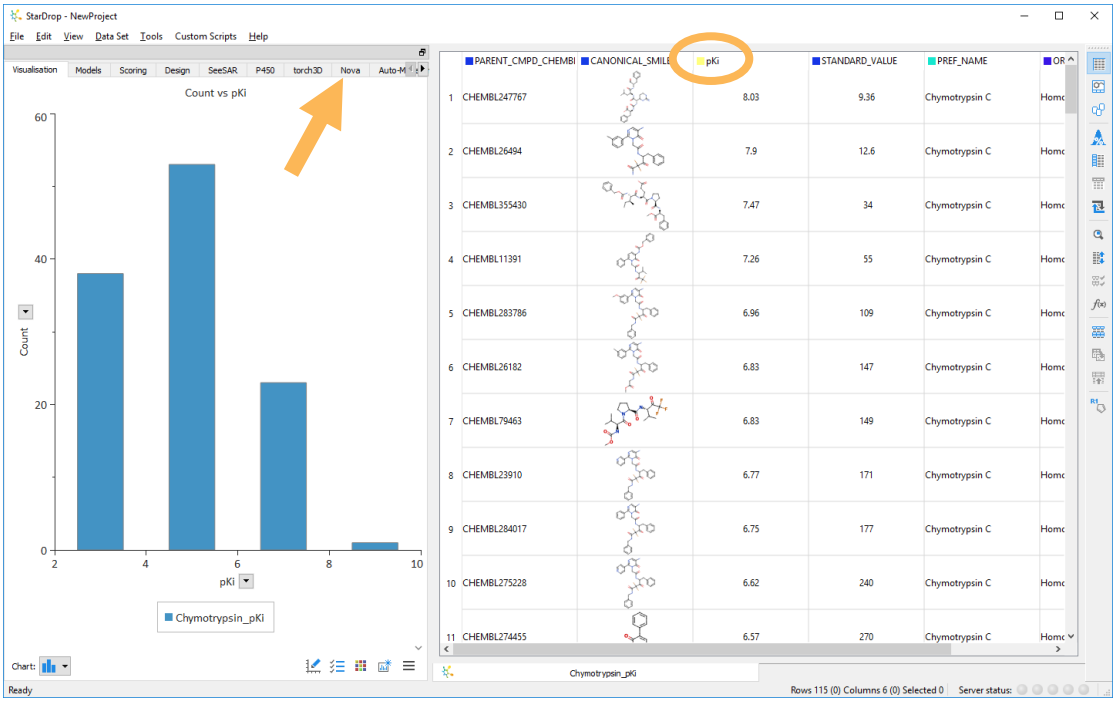
Step-by-step instructions for all the features you will need to use in StarDrop are provided, along with screenshots and examples of the output you are likely to generate. If you have any questions, please feel free to contact stardrop-support@optibrium.com.

Exercise

- Open the file **Chymotrypsin_pKi.add** by selecting **Open** from the **File menu**.

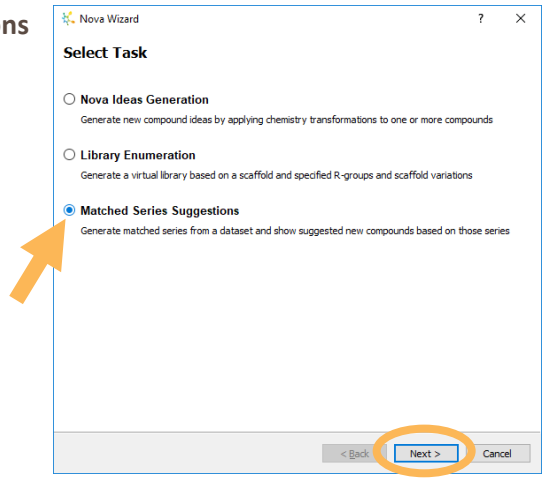


- You will see a spreadsheet containing 115 structures and their measured affinities for Human Chymotrypsin C (in the column labelled pKi).

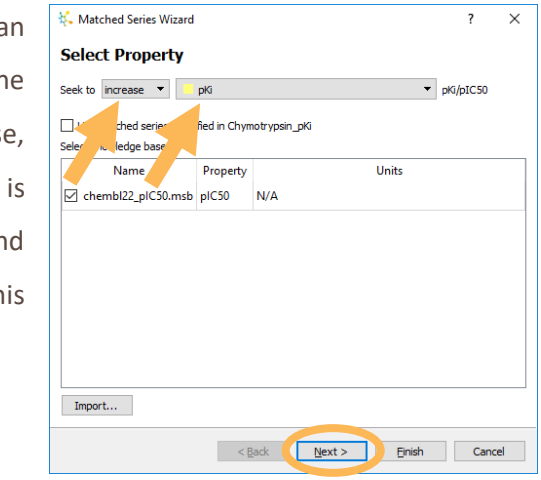


- To start the matched series analysis, click on the **Nova tab** and then at the bottom of the **Nova area** click the  button.

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



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

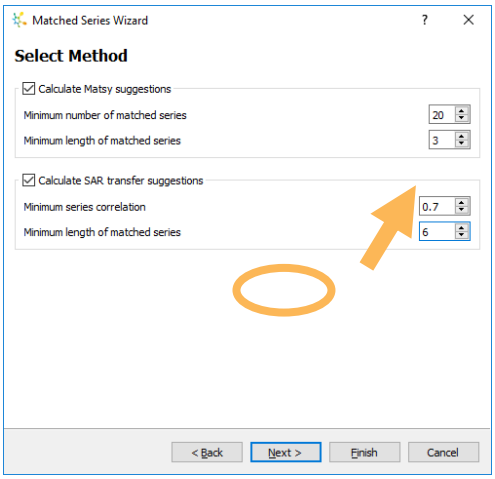


At this point you can change the limitations placed on the suggestions returned. In Matsy™ the support for a suggestion comes from the number of times it has been seen; the more frequent the occurrence of the order of the input series, the more likely it is that the suggestion will be an improvement.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

The default options are to match a series of 3 derivatives and that series should have been seen at least 20 times in the knowledge base and these are acceptable for this data set.

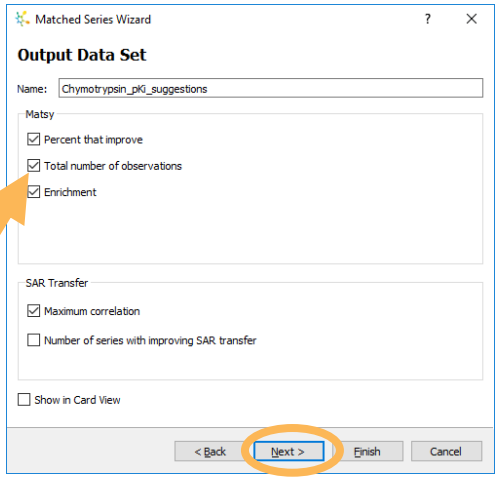
With SAR transfer, the support for a suggestion comes from a long series of derivatives that shows a consistent trend with that seen in the input data set. This example data set is too small to have matched series with the default minimum number of derivatives (8), so for this example we will decrease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

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

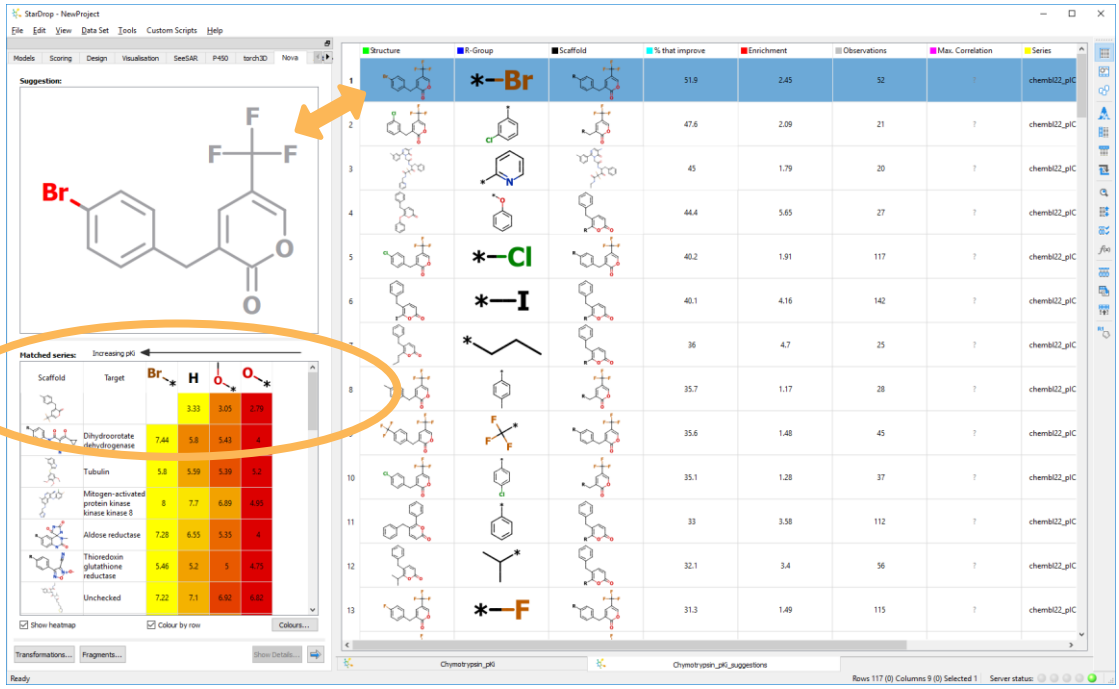


Here you can give the output data set of suggestions a different name and control what else is reported in the output.

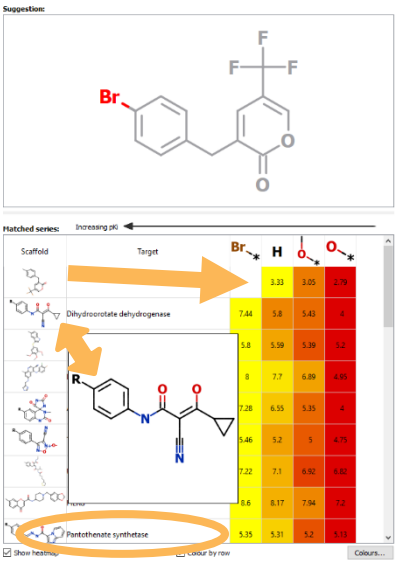
- Check all the boxes for the **Matsy** output as shown above and select **Next**
- Here you can choose structural filters to exclude certain chemical groups. In this case we will use the defaults, so click **Finish** to begin the matched series analysis.



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



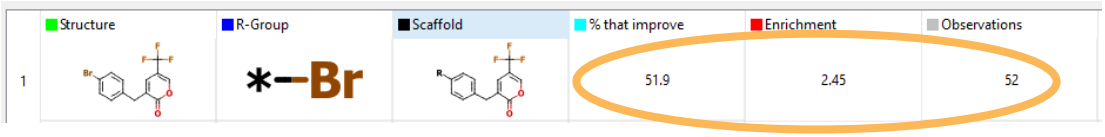
The SAR data from the input data set is in the first row (which is why the target and first substituent columns are empty) and the SAR data are ordered with the least active/desirable on the right to the most active/desirable on the left (as indicated by the colour coding in the table cells).



The first row of the data set shows one of the suggestions that is most likely to improve the pK_i which is the creation of the Bromine derivative. This suggestion is based on the order of activity seen for the hydroxyl, methoxy, and unsubstituted derivatives, at that position on the displayed scaffold, seen in the input data set.

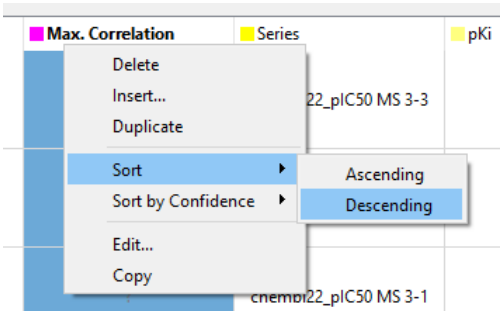
The supporting evidence for the suggestion comes from a variety of target sources and scaffolds and more information for each entry can be obtained by either clicking on the target name in the target column (which will bring up the ChEMBL web page for that target) or by hovering the mouse over the scaffold image in the table to give an enlarged view.

Whilst one might be surprised to find that Bromine is a suggestion, given the fact that the SAR shows that the smallest group is the most active in the series, this effect has been seen many times (nearly 52% of the 52 observations of this series) and may well be due to this group binding to a flexible hydrophobic area of a protein that can move to accommodate the larger Bromine.

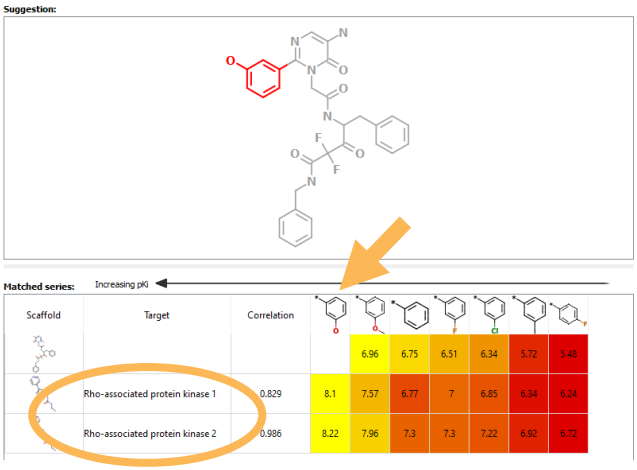


The increased lipophilicity of the bromo-derivative may also help drive the increase in binding by hydrophobic collapse of the ligand with the protein in water.

- To see the SAR transfer suggestions, one **right-click** on the **Max correlation** column in the data set and choose **Descending** from the **Sort menu**.

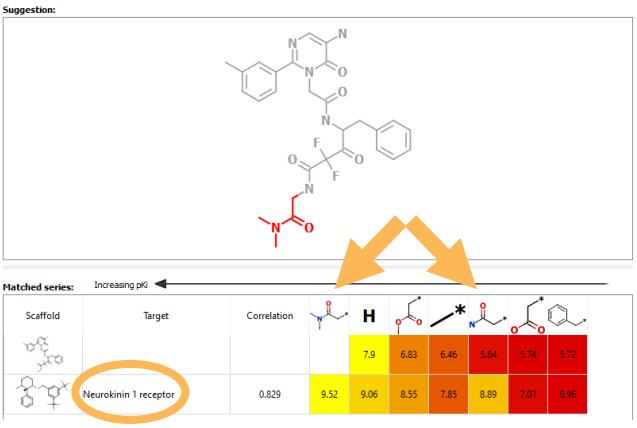


In the top suggestion you can see that the series of aryl derivatives from the input data correlates very well with the activities of derivatives at Rho Kinase.



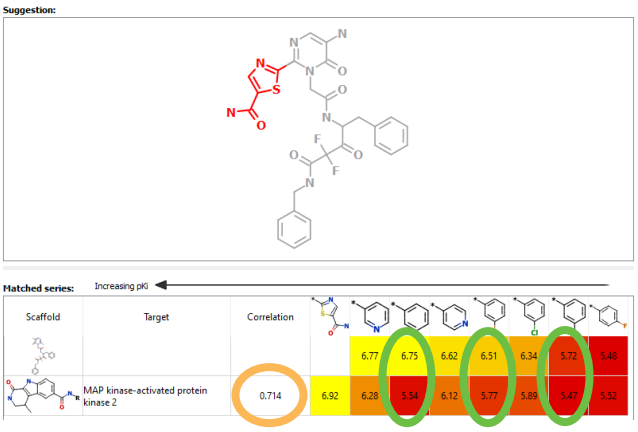
Given the improvement in activity seen in the input data set, in moving from the more hydrophobic aryls to the anisole, the suggested phenol follows that trend for more polarity.

The second suggestion, shown right, on the other end of the pseudo-peptide series, is derived from a GPCR whose endogenous ligand is the peptide, Substance P.



Given the input SAR on its own, the suggestion of the N-dimethyl glycine derivative would not be obvious. However the evidence from the NK1 receptor shows that it might be worth considering. The correlation is not perfect because the unsubstituted glycine derivative is more active in NK1 than in the input data set, but this might be due to a poor assay value in the input set.

The suggestion in row 19, shown right, is more unusual and would be likely to lead the project team into a different area of chemical space for this substituent.



The correlation is lower in this example, due to a few differences in the ordering of the derivatives in the MAPK2 SAR. However, the variations are (with the exception of the phenyl derivative) well within assay variability.

This worked example has shown how matched series analysis can generate suggestions for novel derivatives to improve the binding at your target, based on the data already generated within your own project. The applicability and suitability of suggestions also relies on many other compound properties so the data set of suggestions can be further prioritised in StarDrop, for example using predictive models of physicochemical and ADME properties or target activity and Probabilistic Scoring for multi-parameter optimisation.

Examples of the use of these methods can be found in further worked examples, but if you would like to see a demonstration please contact stardrop-support@optibrium.com.