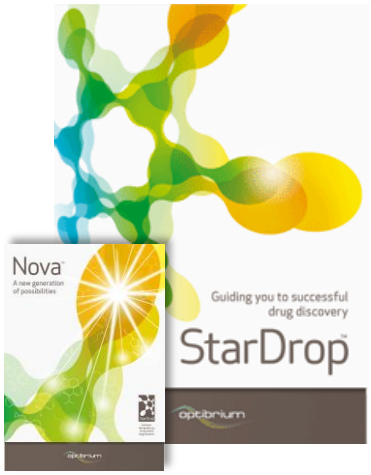


Worked Example:

Generating New Optimisation Ideas Using Matched Series Analysis

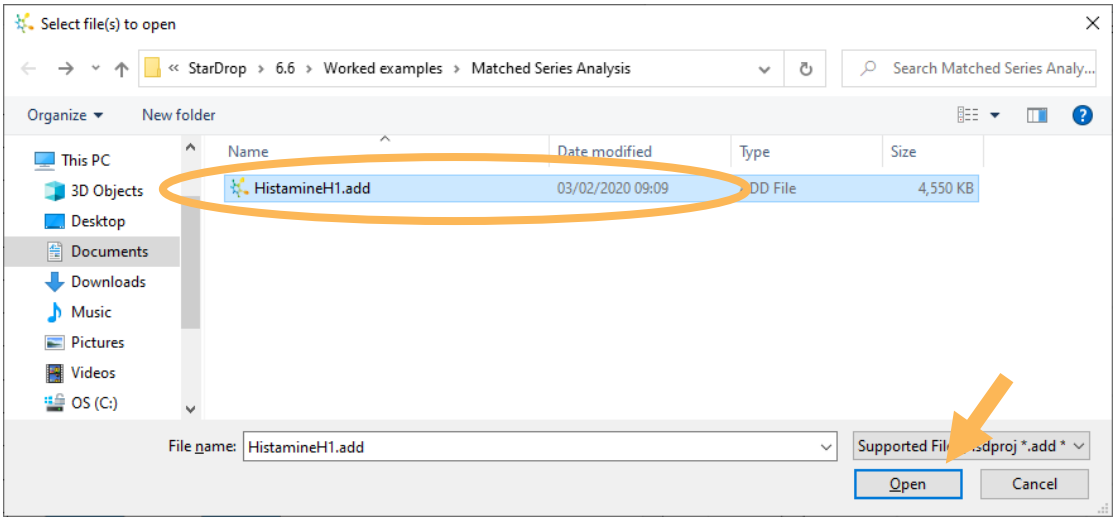
The objective in this worked example is to identify new derivatives that are likely to improve activity at their target, given the SAR already generated on a project. This example uses a publicly available set of Human Histamine H1 K_i data and searches the ChEMBL pIC_{50} knowledge base (generated by NextMove Software) to find matched series that identify new substitutions with a high likelihood of having improved binding.

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 Histamine H1.

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.

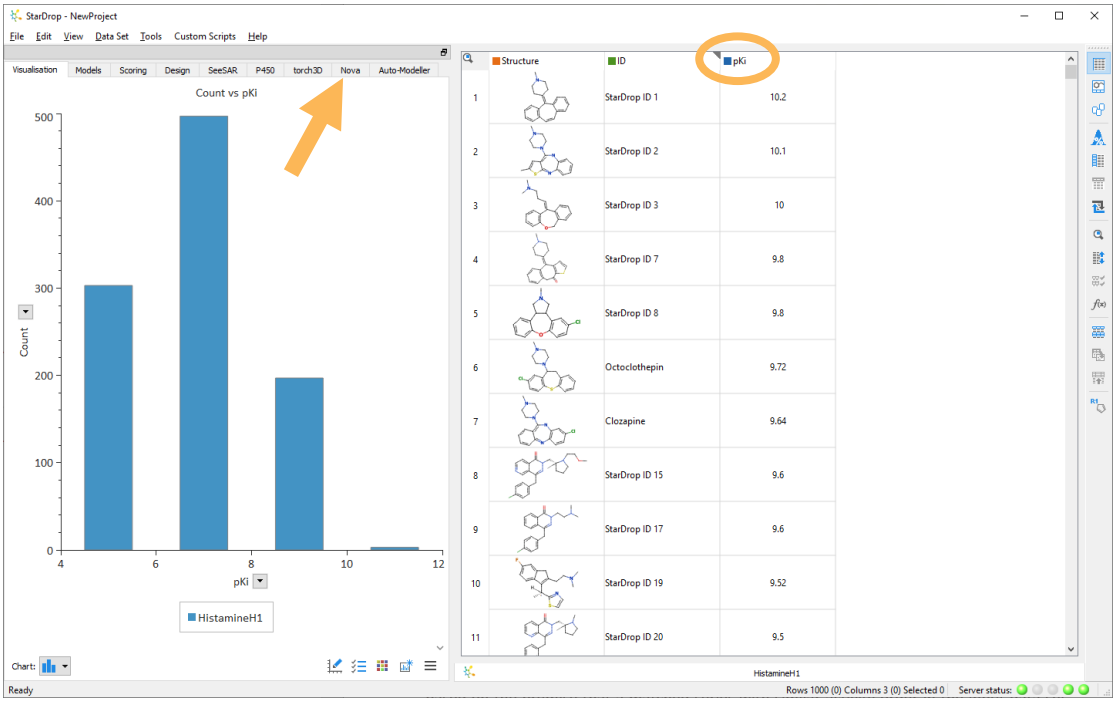
optibrium

Exercise



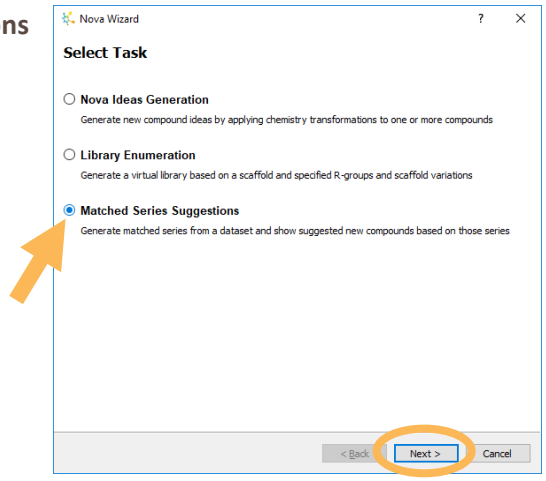
Open the file **HistamineH1.add** by selecting **Open** from the **File menu**.

- You will see a spreadsheet containing 1000 structures and their measured affinities for Human Histamine H1 (in the column labelled **pKi**).

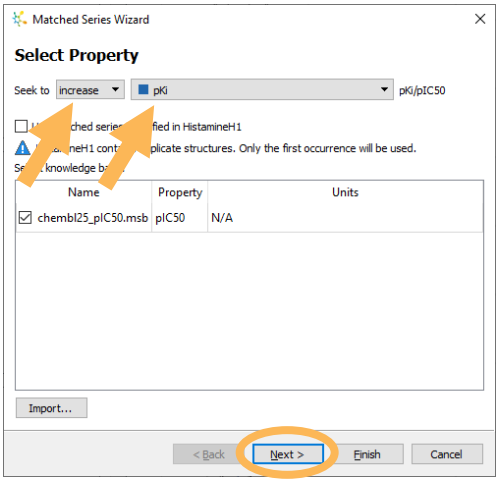


- To start the matched series analysis, click on the **Nova tab** and then at the bottom of the **Nova area** click the  button.

- 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.

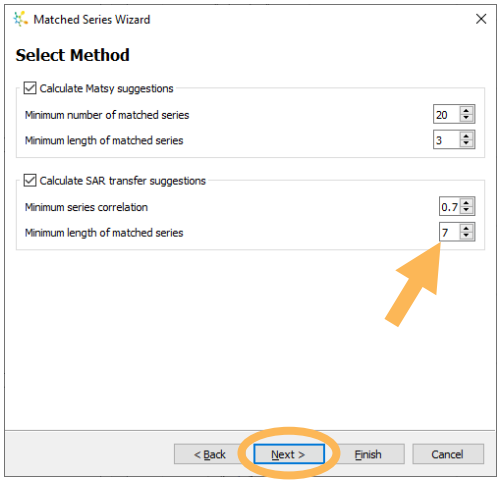


At this point you can change the limitations placed on the suggestions returned. In Matsy™ the support for a suggestion comes from the number of times it has been seen; the more frequent the occurrence of the order of the input series, the more likely it is that the suggestion will be an improvement. 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.

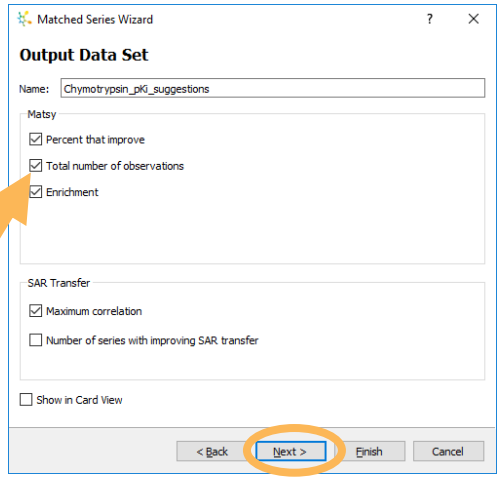
With SAR transfer, the support for a suggestion comes from a long series of derivatives that shows a consistent trend with that seen in the input data set. This example data set is too small to have matched series with the default minimum number of derivatives (8), so for this example we will decrease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

- Click on the **Minimum length of matched series** box in the **SAR transfer** section and change the value to **7**.
- Click the **Next** button to continue.

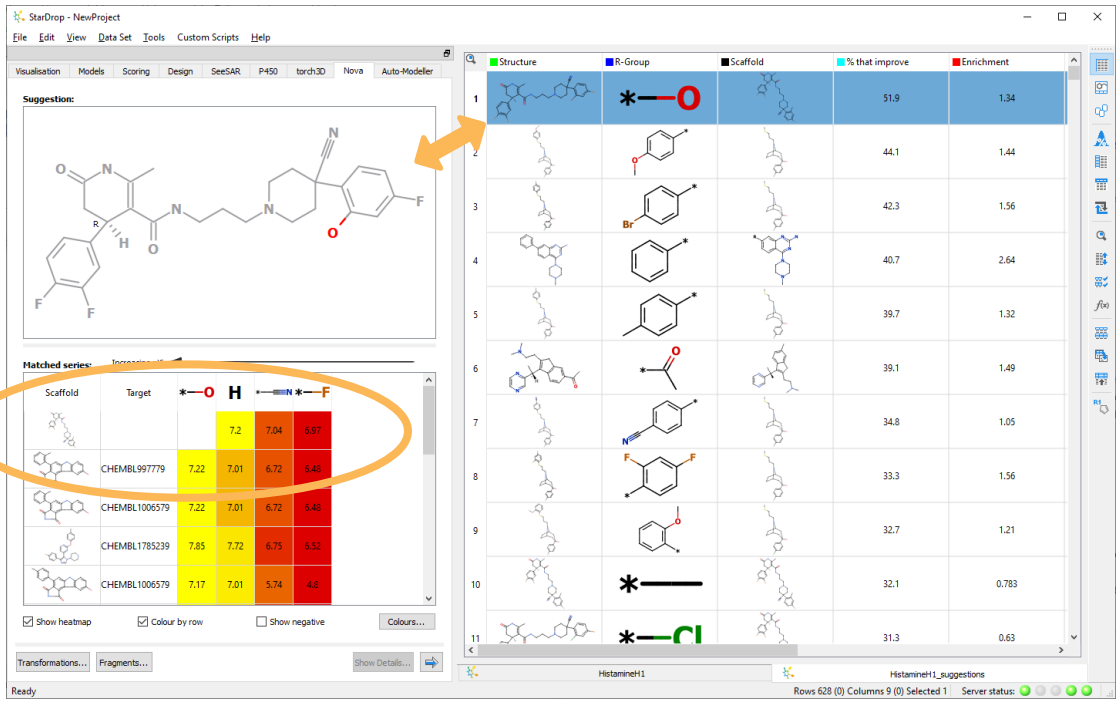


Here you can give the output data set of suggestions a different name and control what else is reported in the output.

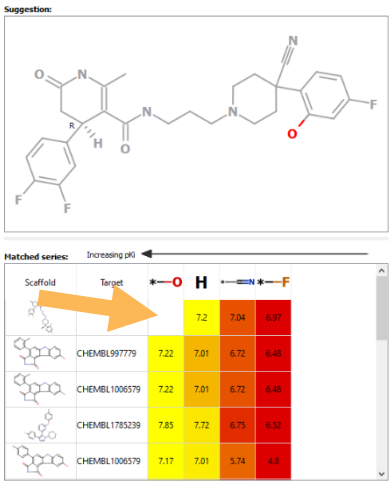
- Check all the boxes for the **Matsy** output as shown above and select **Next**
- Here you can choose structural filters to exclude certain chemical groups. In this case we will use the defaults, so click **Finish** to begin the matched series analysis.



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.



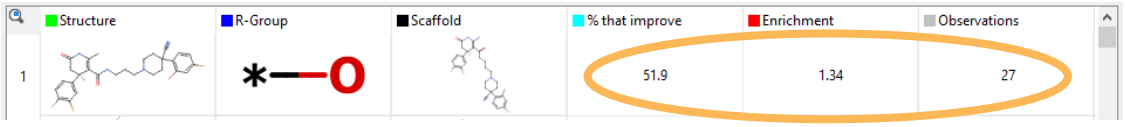
The SAR data from the input data set is in the first row (which is why the target and first substituent columns are empty) and the SAR data are ordered with the least active/desirable on the right to the most active/desirable on the left (as indicated by the colour coding in the table cells).



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 hydroxy derivative. This suggestion is based on the order of activity seen for the fluoro, cyano, and unsubstituted derivatives, at that position on the displayed scaffold, seen in the input data set.

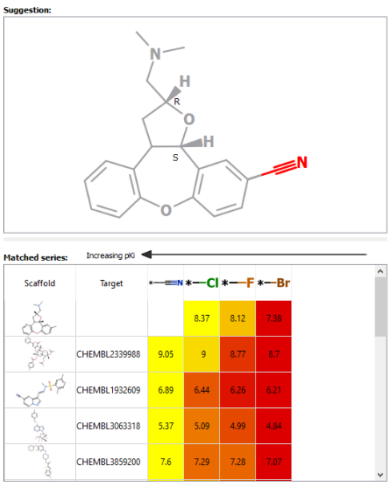
The supporting evidence for the suggestion comes from a variety of target sources and scaffolds and more information for each entry can be obtained by either clicking on the target name in the target column (which will bring up the ChEMBL web page for that target) or by hovering the mouse over the scaffold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

The first of the matching series from the ChEMBL knowledgebase is for target CHEMBL997779 (which is human WEE1 kinase) where the series also occurs at the ortho position, matching the scaffold in our series. Note that further examples are also seen in the same order for the meta and para positions, which may relate to influencing the aryl ring electronics rather than any direct effects. This pattern has been seen may times and nearly 52% of the 27 observations have shown an increase in activity.

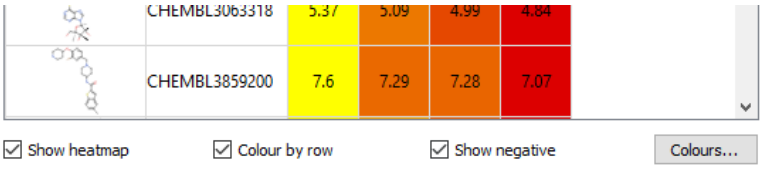


- Select row 13 in the data set.

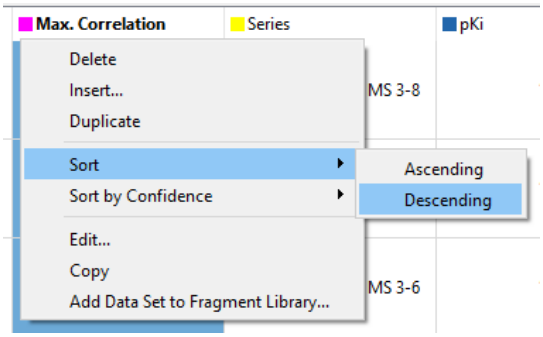
Here we have an interesting example because it does not follow the traditional order of halogens by molecular weight. In addition, the cyano group is not a group one might typically consider next in the series, yet there are 49 observations of this series in the knowledgebase where this results in an increase in activity nearly 31% of the time.



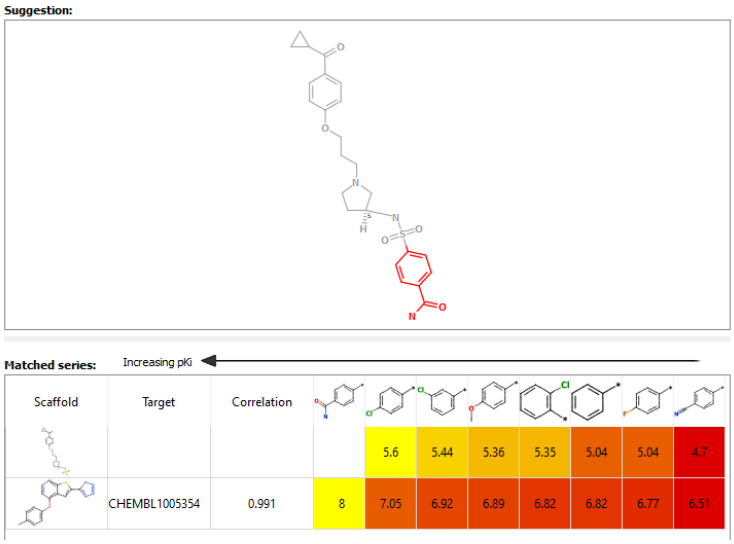
Note: You can view the examples where the order does not match the input series for any of the suggestions by ticking the **Show negative** option below the table of examples.



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



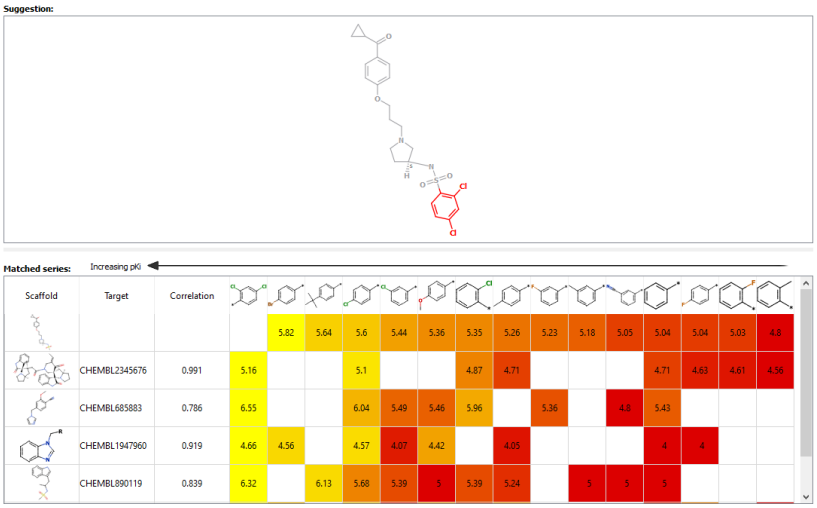
In the top suggestion you can see that the series of aryl derivatives from the input data correlates very well with the activities of derivatives at MAP3K8. **Note:** You may need to make your Nova area wider to see the complete series by dragging with the mouse.



For this suggestion there is only one example, whereas some suggestions are based on multiple long series.

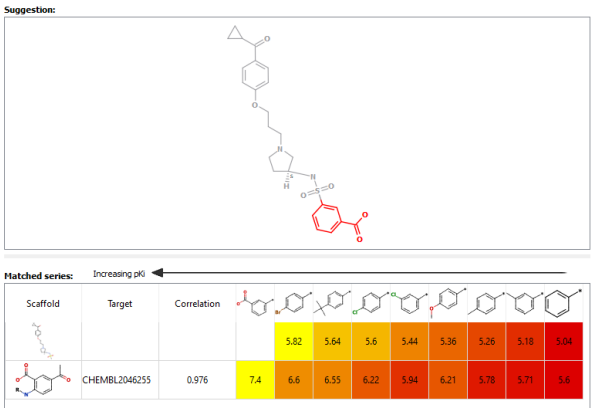
- Select row 4 in the data set.

This series is very long and has multiple matches that correlate well. This is a good example where it would probably be difficult to determine



such evidence for the suggestion based upon database searching, particulalry given the various differences, and gaps, in the example series.

- Select row 6 in the data set.



Here we can see an example where the suggested compound is a benzoic acid, which will make the suggested scaffold zwitterioninc. We may consider this suggestion surprising based on the fact that the reference sequence is all hydrophobic. While interesting and not the obvious next derivative to make, this may come from the exemplar structure already being a benzoic acid and the target CHEMBL2046255 being aldo-keto-reductase (i.e. this may be a special case specific to that protein).

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. For any suggestions that appear interesting, it is important to consider the evidence provided by the example series to determine how this may translate to 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.