Version 8

StarDrop™ Worked Example: Generating new optimisation ideas using Matched Series Analysis

Contents

Introduction

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 Ki data and searches the ChEMBL pIC50 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 identical compounds except for different substituents at a single point (see Section 10.3 of the StarDrop Reference Guide for more details). The suggestions are derived by 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 support@optibrium.com.

Running a Matched Series Analysis

• Open the file **HistamineH1.add** by selecting **Open Data Set** 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**).
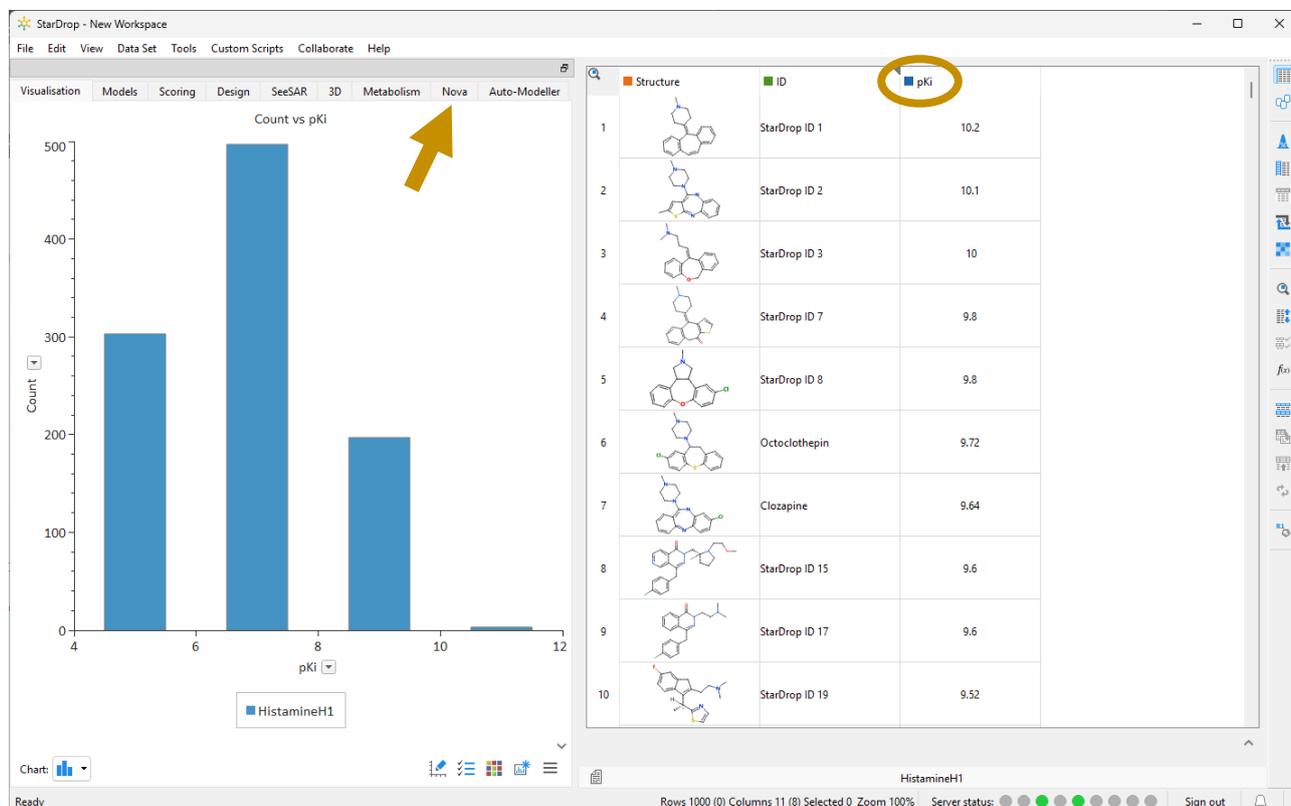
- To start the matched series analysis, click on the **Nova** tab and then, at the bottom of the Nova area, click the ➡ button.

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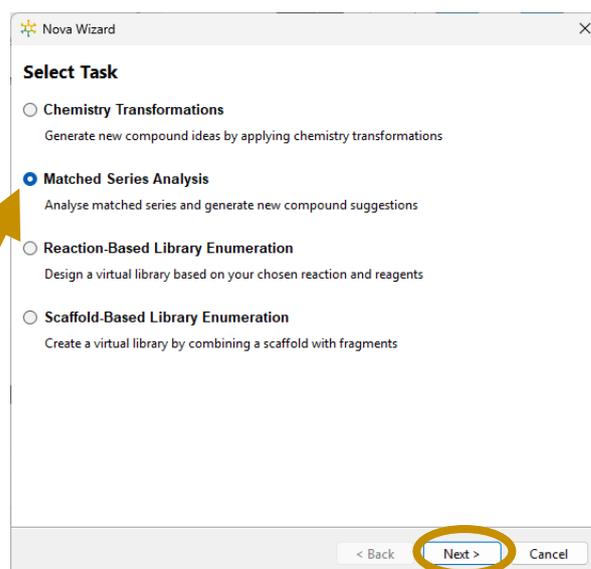
- In the dialog box that appears, you can specify the column containing the property you wish to improve. In this case, the column we are interested in, **pKi**, is already chosen, and we want to find suggestions that **increase** this value, so this default option is also correct.

- Click the **Next** button to continue.
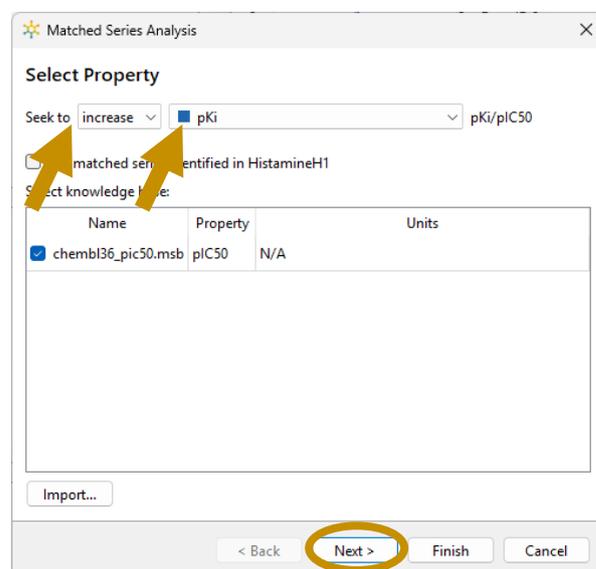
You can now change the limitations placed on the suggestions returned. In Matsy™, the support for a suggestion comes from the number of times it has been seen; the more frequent the occurrence of the order of the input series, the more likely it is that the suggestion will be an improvement. Hence, the compared series are generally short in order to find sufficient examples in the ChEMBL knowledge base.

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.

You can provide a name for the output data set of suggestions (we will keep the default name for this example) and control what else is reported in the output.

- Check all the boxes for the **Matsy** output and click **Next**.

- If not selected by default, tick the box to filter out **Undesirable Fragments** and click **Finish**.

The suggestions are returned in a table with the Matsy suggestions first, followed by the SAR transfer suggestions.

Matsy results

The Matsy suggestions consider short series that occur frequently in the literature. The results in the data set are ordered by the **% that improve** column, which tells what percentage of the total observations of this matched series in the knowledge base show improvement for the suggested functional group. When a row is selected in the data set, the suggestion is displayed in the Nova area, and the supporting evidence is shown in the table below.

• Select **row 2** of the data set.

Row 2 shows one of the suggestions most likely to improve pKi: the creation of the hydroxy derivative. In the main data table, we can see that this matched series has been found 41 times in the ChEMBL knowledge base, and more than 48% of those 41 observations have shown an increase in activity for the hydroxy derivative.


The table in the bottom panel of the Nova area shows the specific evidence from ChEMBL for the suggestion. 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).

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. More information for each entry can be obtained by clicking on the target name in the target column (which will bring up the ChEMBL web page for that target). Alternatively, additional information can be seen by hovering the mouse over the scaffold image in the table to give an enlarged view.
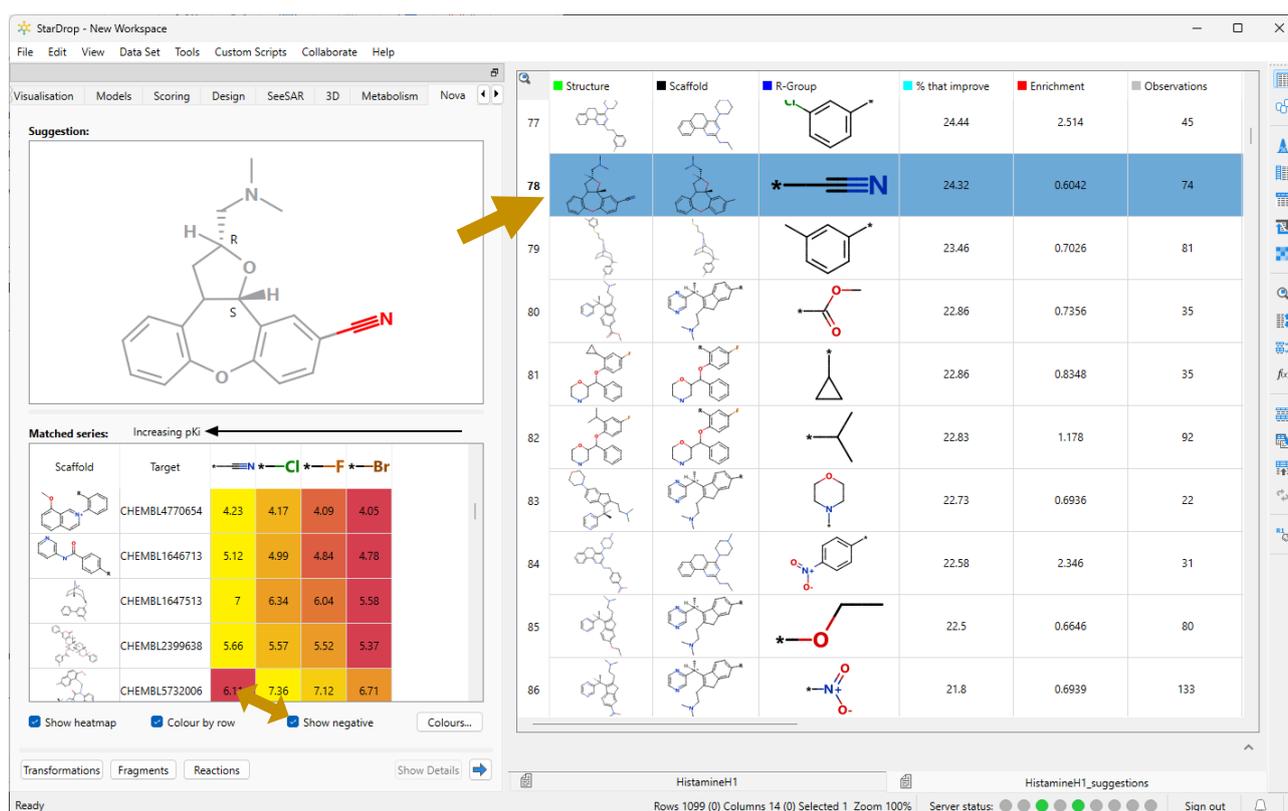
Two of the matching series from the ChEMBL knowledge base (targets CHEMBL864855 and CHEMBL997779, both variants of WEE1 kinase) occur three times each in the results. These matched series show the same behaviour for substitutions at the ortho-, meta-, and para-positions on the ring, indicating that this change might relate more to influencing the aryl ring electrostatics rather than any direct effects on target binding.
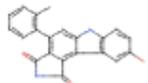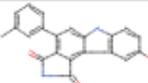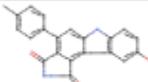
- Select **row 78** in the data set.

If we examine further down the list at suggestions with a lower % that improve, we find that some might be substitutions that are not immediately obvious. In row 78, we have an interesting matched series in our input data set because it does not follow the traditional order of halogens by molecular weight. The recommended cyano group is not a group one might typically consider next in the series, yet there are 74 observations of this series in the knowledge base, which results in an increase in activity 24% of the time.

Among the list of targets, we find the dopamine transporter (CHEMBL1647514), noradrenaline transporter (CHEMBL1647512), 5-hydroxytryptamine receptor (CHEMBL615782), acetylcholinesterase (CHEMBL1109363) and butyrylcholinesterase (CHEMBL4770654), indicating that this suggestion may merit further investigation as a histamine antagonist.


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.

SAR Transfer

In contrast to Matsy, SAR transfer looks for correlations across long series of substitutions. Naturally, there will be fewer observations of these in the literature, so the important metric is the correlation of the order of improvement for each R-group in the series.

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

- Select **row 3** in the newly sorted data set.
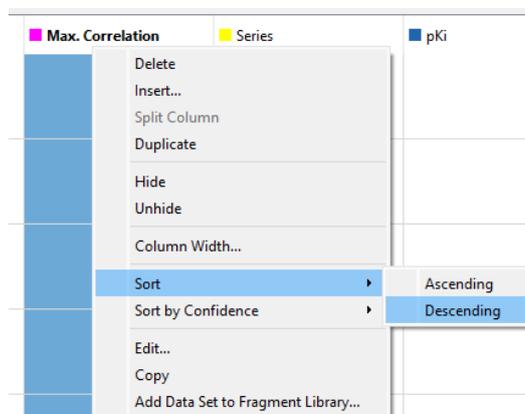
For this suggestion, there is only one example from the knowledge base, showing that the series of aryl derivatives from the input data correlates very well with the activities of derivatives at MAP3K8.

Note: You may need to make your Nova area wider to see the complete series by dragging the divider between the data set and the Nova area with the mouse.

- Select **row 6** in the data set.

This series is very long and has multiple matches that correlate, though not all correlate perfectly. This is an example where it would have been difficult to determine such evidence based on database searching, particularly given the various differences and gaps in the example series.

Scaffold	Target	Correlation	5.824	5.638	5.638	5.602	5.444	5.357	5.347	5.26	5.229	5.208	5.18	5.051	5.041	5.036	5.027	4.796	4.699
	CHEMBL2345676	0.991	5.16			5.1			4.87	4.71					4.71	4.63	4.61	4.56	
	CHEMBL5730662	0.7369	7.26	7.26	7.26	7.26	7.26		6.26		7.26				6.26	6.26			
	CHEMBL5166901	0.8571	5.25				5.2		4.49	4.89	4.6			4.08	3.96		4.19		3.57
	CHEMBL4257289	0.866	6			5	5	5	5	5				4	4	4			4
	CHEMBL4233573	0.7832	6.08	5.27		5.9	4.98	4.31	4.11	4.24	4.57		4.07		3.88	4.72	3.92	3.81	
	CHEMBL890119	0.8394	6.32		6.13	5.68	5.39	5	5.39	5.24			5	5	5				
	CHEMBL1947960	0.919	4.66	4.56		4.57	4.07	4.42		4.05					4	4			
	CHEMBL685883	0.7857	6.55			6.04	5.49	5.46	5.96		5.36			4.8	5.43				
	CHEMBL5730483	0.8833	8.32			7.35	7.51		7.49	6.76	7.03	7.13			6.56	6.55	6.24		
	CHEMBL3108471	0.7167	7.6	6.85		6.96	5.28	6.46	5.64	6.44					5.21	6.3		4.92	
	CHEMBL5730966	0.8833	8.32			7.35	7.52		7.49	6.76	7.03	7.13			6.56	6.55	6.24		

- Select **row 8** in the data set.

Here we can see an example where the suggested compound is a benzoic acid, which will make the suggested scaffold zwitterionic. 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).

Scaffold	Target	Correlation	5.824	5.638	5.602	5.444	5.357	5.26	5.18	5.041
	CHEMBL2046255	0.9762	7.4	6.6	6.55	6.22	5.94	6.21	5.78	5.71	5.6

Conclusion

This worked example demonstrates 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 and correlations with the ChEMBL knowledge base. 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 rely 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 other StarDrop worked examples, which can be found in the customer hub section of our website. If you have any questions or would like to see a demonstration, please contact us by email at support@optibrium.com.