

Worked Example:

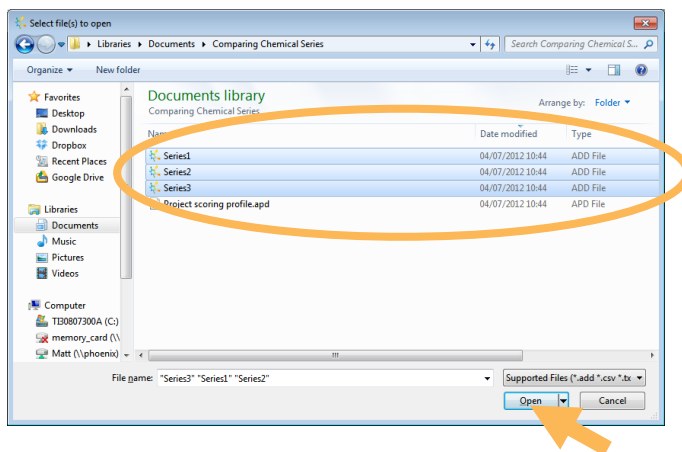
Comparing Chemical Series with Probabilistic Scoring

This example is taken from a project in which screening of a diverse library resulted in hits from multiple chemistries. Without the resources to follow-up all of the hit chemistries, the project team wished to focus on a small number of series which were most likely to yield high quality leads with appropriate physicochemical and ADME properties. In the following steps, we will compare three virtual series, resulting from expansion around these hits, using predictions from *in silico* ADMET models and probabilistic scoring to prioritise them for future exploration.

Exercise

- Start StarDrop
- From the **File->Open** menu item, open the three files, **Series1.add**, **Series2.add** and **Series3.add**

Hint: You can open all three data sets simultaneously by selecting the files while holding the Ctrl key and then clicking **Open**.



Optibrium™, StarDrop™, Nova™ Glowing Molecule™ and Auto-Modeller™ are trademarks of Optibrium Ltd.

© 2014 Optibrium Ltd.


You will note that none of the data sets contain compound structures, for confidentiality reasons. However, a range of physicochemical and ADME properties have been predicted for each compound.

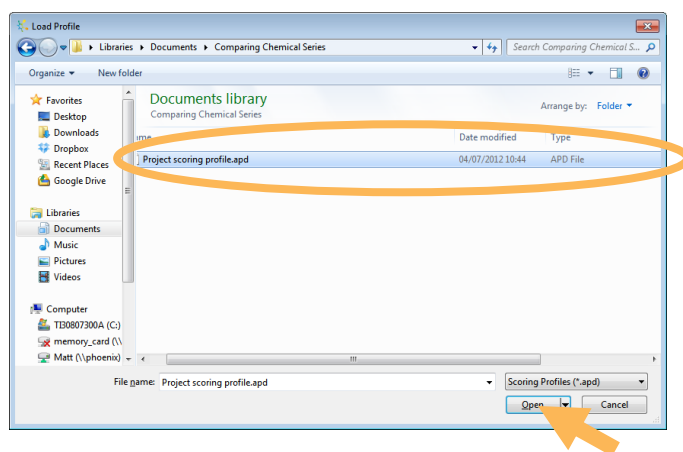
The screenshot shows the StarDrop software interface. On the left is a sidebar with 'Available Models' including StarDrop, logS, logS @ pH7.4, logP, logD, 2C9 pKi, hERG pIC50, BBB log([brain]:[blood]), BBB category, HIA category, P-gp category, 2D6 affinity category, PPB90 category, MW, HBD, HBA, TPSA, Flexibility, Rotatable Bonds, Legacy models, and Derek Nexus. The main window displays a table with the following data:

Identifier	logS	logP	logD	hERG pIC50	PPB category
1 Array3-1	-1.473	6.72	5.286	4.82	high
2 Array3-2	-1.29	6.938	5.737	4.86	high
3 Array3-3	-1.057	6.315	5.364	4.745	high
4 Array3-4	-1.548	7.208	6.109	4.91	high
5 Array3-5	-0.438	5.521	4.58	4.599	high
6 Array3-6	-0.6729	6.055	4.984	4.697	high
7 Array3-7	-0.7745	6.296	5.242	4.742	high
8 Array3-8	-0.8919	6.064	5.653	4.699	high
9 Array3-9	-0.09283	5.044	4.675	4.511	high

At the bottom, it shows 'Series1', 'Series2', and 'Series3' tabs, and a status bar indicating 'Rows 994 (0) Columns 12 (0) Selected 0'.

We will use a scoring profile, defined by the project team, to prioritise these chemical series.


- Change to the **Scoring** tab in StarDrop and click the  button to load a new scoring profile. Select the file **Project scoring profile.apd** and click **Open**.

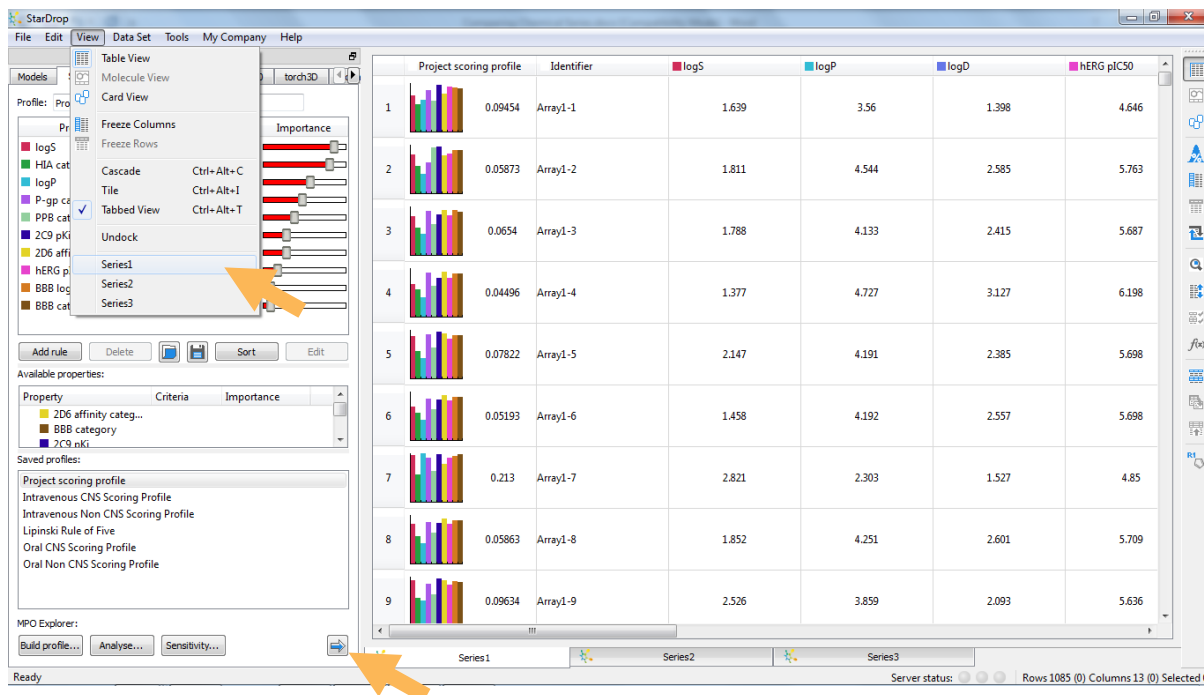


The scoring profile (shown to the right) will be displayed, showing the criterion for each property and the importance of the property to the overall objective of the project, as defined by the project team.


Profile: Project scoring profile

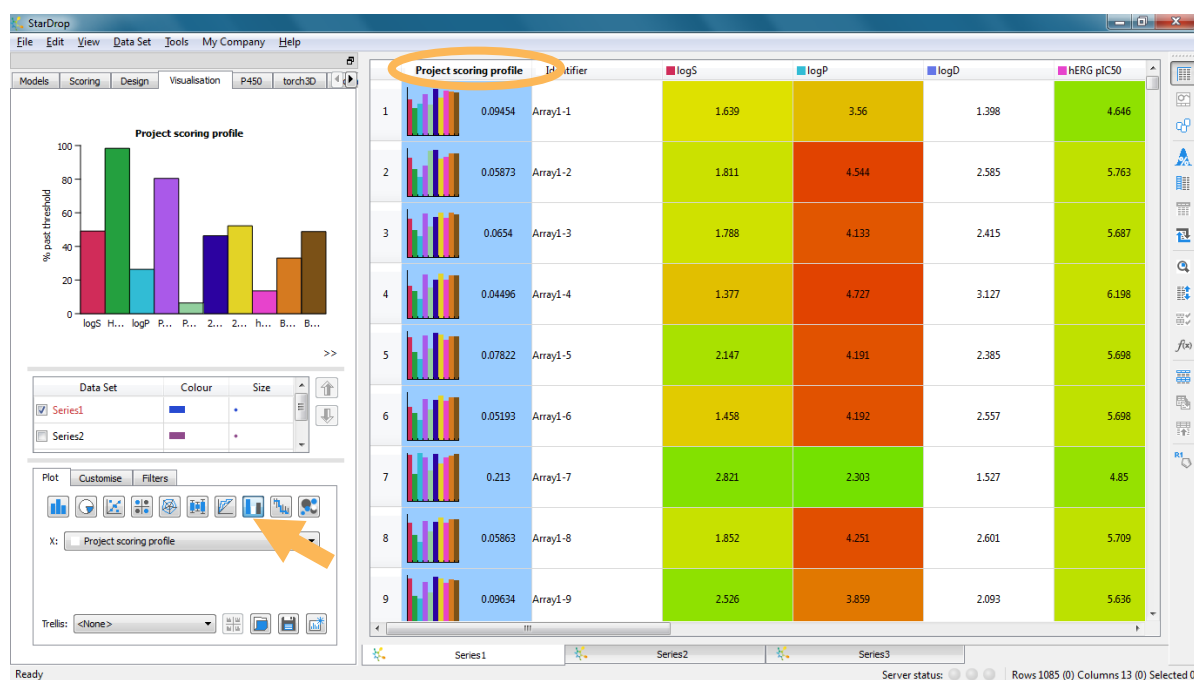
Profile	Desired Value	Importance
logS	> 1	
HIA category	+	
logP	≤ 3.5	
P-gp category	no	
PPB category	low	
2C9 pKi	≤ 6	
2D6 affinity category	low medium	
hERG pIC50	≤ 5	
BBB log([brain]:[blood])	≤ -0.5	
BBB category	-	

- Select each of the three data sets in turn from the **View** menu and run the scoring profile using the  button on the **Scoring** tab.



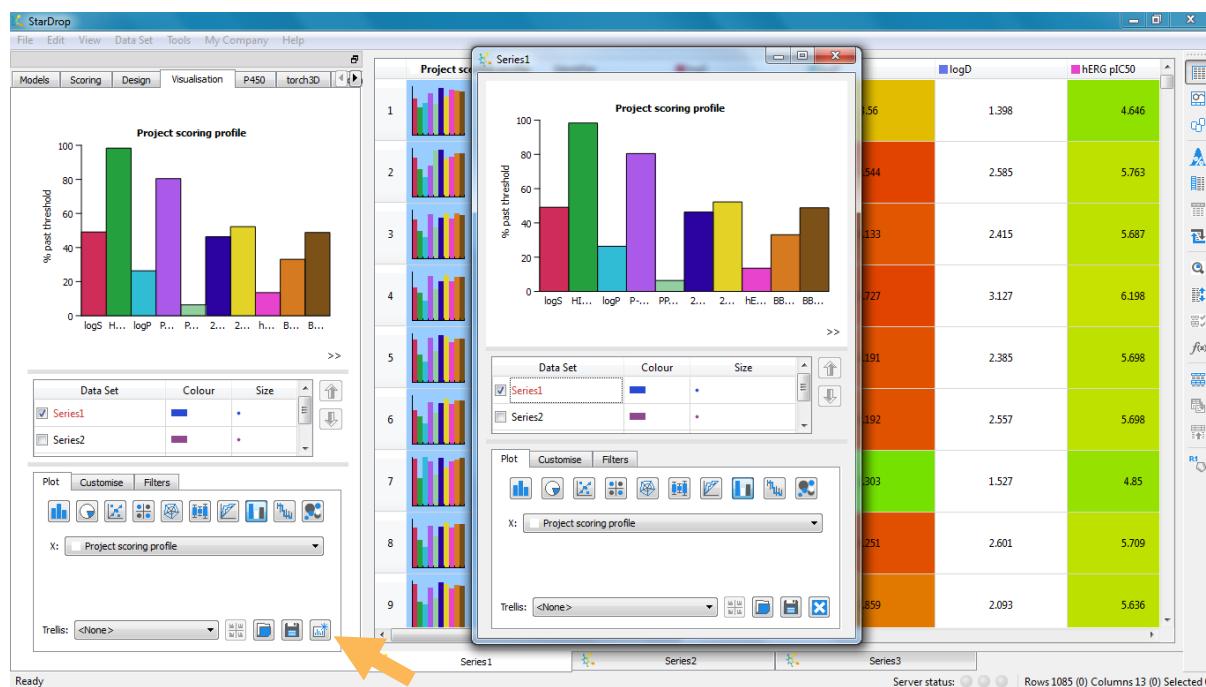
Some of these data sets contain approximately 1000 compounds, so we will now use some of the visualisation tools in StarDrop to help us to explore the distributions of properties and scores for these different series.

- Change to the **Visualisation** tab in StarDrop and choose the **Series1** data set from the **View** menu.
- Select the column containing the score by clicking on the header labelled **Project scoring profile** and click the Criteria Histogram button () in the **Visualisation** tab.

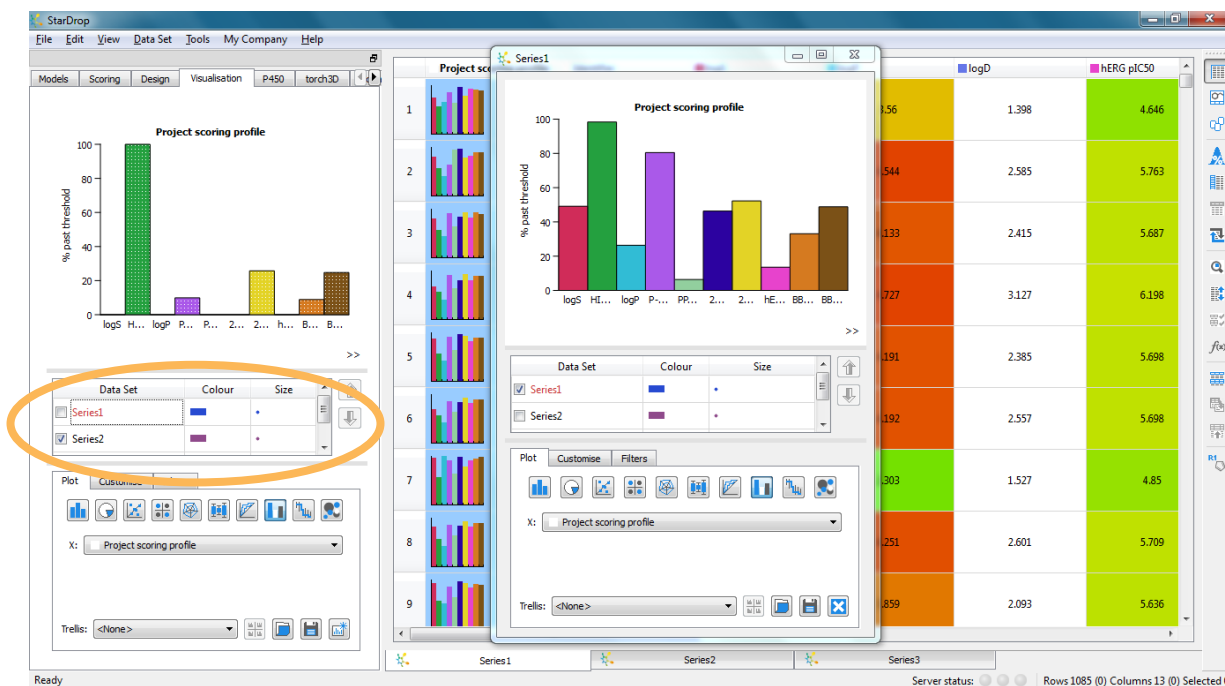



The Criteria Histogram shows the percentage of compounds in the data set that meet the ideal property criteria, as defined in the scoring profile. A high bar for a property indicates that the majority of the compounds in the data set meet a criterion, while a low bar indicates a consistent issue across the data set. The colours correspond to the key in the scoring profile and the bars are ordered from most important on the left to least important on the right, as defined in the scoring profile.

- Click on the detach button () in the **Visualisation** tab to create a free-floating copy of this plot.

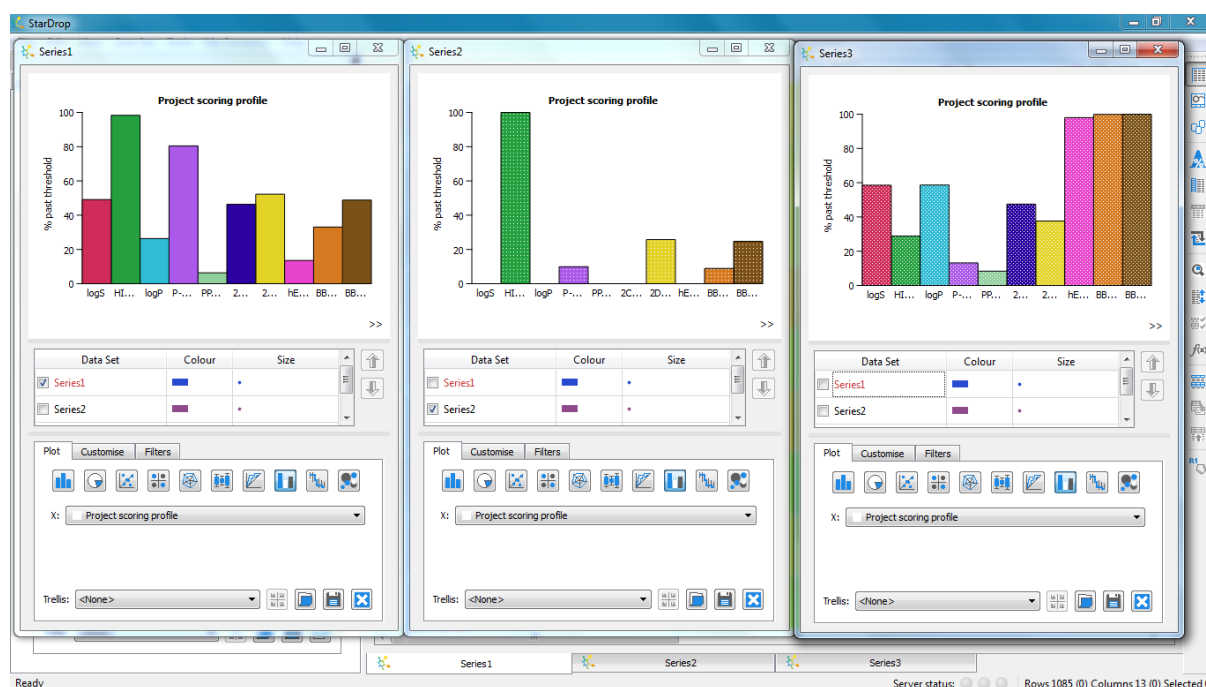


- Tick the box next to **Series2** in the key below the plot in the **Visualisation** tab and untick the box next to **Series1**, to plot a criteria histogram for Series 2.



- Click on the detach button () in the **Visualisation** tab again to create a free-floating copy of the plot for Series 2 and then repeat the process to plot a Criteria Histogram for Series 3.

We can then easily compare the property profiles for the three chemical series side-by-side, as shown below:

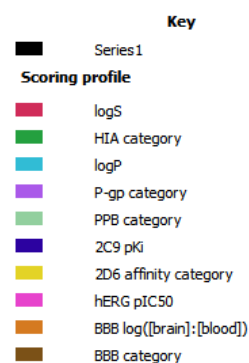


From these simple profiles, we can see that Series 1 and Series 3 both contain a reasonable proportion of compounds that are likely to meet each of the property criteria. The most consistent issues in Series 1 are likely to be high plasma protein binding and inhibition of the hERG ion channel. For Series 3, the most consistent issue is predicted to be high plasma protein binding.

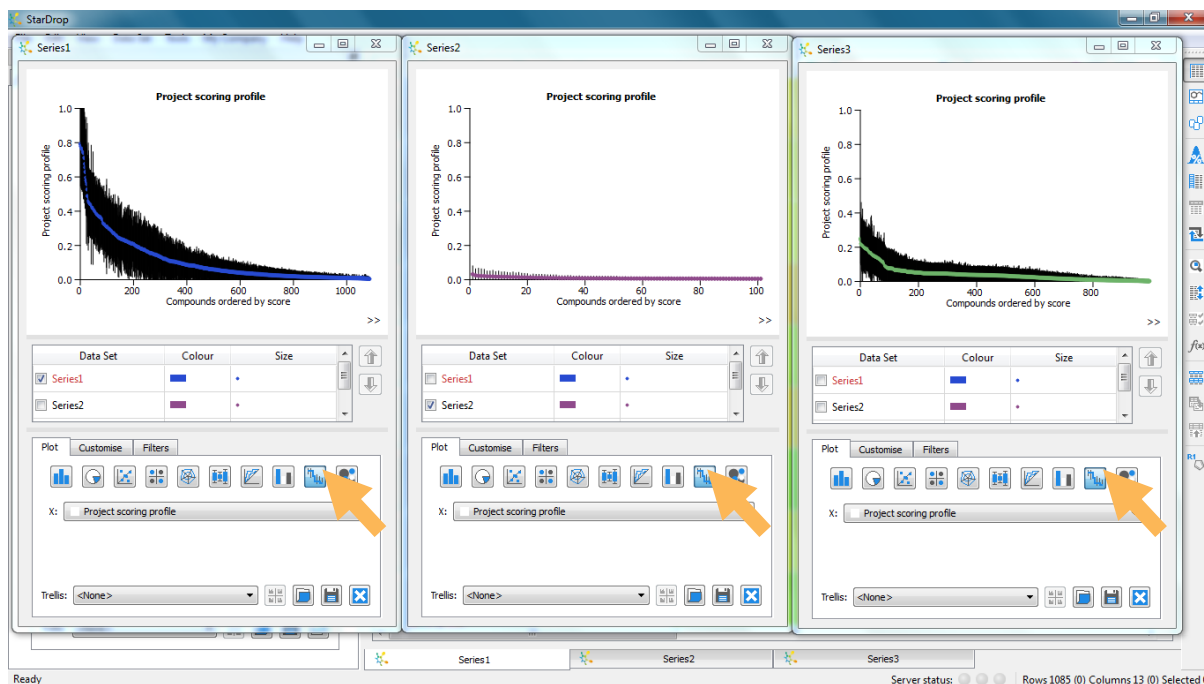
However, all of the compounds in Series 2 fail to meet several criteria including solubility and logP, which are two of the most important properties. This indicates that Series 2 is very likely to be a high risk chemistry. Furthermore, in a chemical series where all of the compounds have poor values for a property, it is likely that the issue lies with the common scaffold of the series, because varying the substituent groups does not affect the outcome.

From this simple analysis, Series 1 and Series 3 look to be the most promising, however it is difficult to confidently choose between these chemical series.

To help us to explore these chemistries further, we will use the *Snake Plots* generated by Probabilistic Scoring.



- On each of the detached plots, click on the Snake Plot button () to change the plot type:



A Snake Plot shows the compounds in a data set ordered along the x-axis, from the highest scoring on the left to the lowest scoring on the right. The score for each compound is plotted on the y-axis and an error bar around each point shows the uncertainty in the overall score, given the uncertainty in the underlying data used to calculate the score.

From these Snake Plots we can clearly see that there are compounds in Series 1 that have a high chance of meeting all of the property criteria, indicating that this series is most likely to yield a high quality compound. Resources should be focussed on following up this series and generating experimental data to validate this predicted hypothesis.

Although the criteria histograms suggest that Series 3 appears to meet more criteria overall, most of the compounds in this series have poor predicted absorption, which is one of the most important properties. Furthermore, those that are predicted to be well absorbed are unlikely to meet other criteria. This results in a lower likelihood of success than for Series 1. However, given the uncertainty in the data, as illustrated by the error bars for the highest scoring compounds in this series, it may be worthwhile sampling a small number of compounds and generating experimental data as a backup strategy.

Finally, the Snake Plot for Series 2 confirms the picture we saw from the criteria histogram for this series. The chances of success of the compounds in this series, against the profile of required properties, are very low. Furthermore, the confidence in these assessments is high (as indicated by the small error bars). This is because the chance of all of the models being incorrect simultaneously, resulting in an unexpected success, is very low.

Conclusion

This example illustrated how we can rigorously compare three chemical series by assessing their properties against a profile required for a successful compound in the project for which they are intended. By taking into account the importance of each property to the objective of the project and the uncertainty in each prediction, probabilistic scoring can quickly and confidently identify the chemistry with the highest chance of success.