

FAQ

StarDrop: Compound Selection

Essential decision points in drug discovery involve the choice of a set of compounds ('selection') for progression from a larger collection ('library'). This may be based on predicted or measured properties for the compounds in the library and other factors such as chemical diversity. StarDrop comprises an automated compound selection algorithm which allows the user to explore different strategies based on chemical diversity, quality or a combination of the two.

While a scoring scheme, such as probabilistic scoring, allows the 'quality' of the compounds in the library to be assessed against a project's target profile across a broad range of properties, it may not be appropriate to choose only the 'best' compounds from the library for progression. In particular, early in a project, consideration should be given to exploring a diversity of chemistry. The purpose of this is to provide additional structure-activity data, validation of the accuracy of property predictions and, if possible, multiple chemical series to provide alternatives should an unexpected difficulty arise with the lead chemical series.

In combination with the chemical space plots and probabilistic scoring, the compound selection algorithm provides the ability to explore the impact of different compound selection strategies on the likelihood of success. There may be a good argument for manual selection of specific compounds, to test a specific hypothesis or even out of curiosity, and this should not be ruled out. However, the objective balancing of diversity and quality encourages the majority of research effort to be expended in those areas of chemistry most likely to yield progress toward a project's objectives.

Frequently Asked Questions

What are the requirements to run a 'biased' selection?

The data set should contain structures and have had a scoring profile applied. This enables molecules to be selected that have the best possible balance between performance against the scoring profile and structural diversity. In this case the user can determine the bias between 'Rank' and 'Diversity'.

How is chemical diversity assessed?

Chemical diversity is defined in terms of the patterns of atoms present in their chemical structures. The patterns of atoms along 'paths' through the 2D chemical structure of a compound are encoded in a binary 'fingerprint' and the similarity of two compounds, A and B, can then be defined in terms of the Tanimoto coefficient. The advantage of a path-based fingerprint approach to similarity and diversity is that it provides a 'generic' method of comparing compounds.

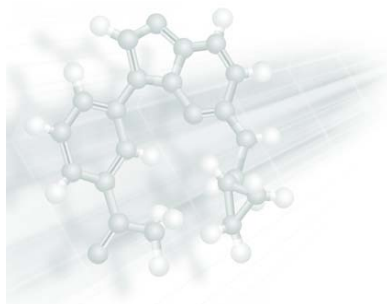
Why is the compound selection based on a genetic algorithm?

The number of possible selections increases exponentially with the size of a virtual library, e.g. there are 2.6×10^{23} ways of choosing 10 compounds from a library of 1,000. Therefore, when considering diversity, it rapidly becomes impossible to perform an exhaustive search for the optimal selection for a given set of criteria. Instead, a 'stochastic' approach must be taken, which cannot guarantee to identify the optimal solution but will find the optimal or a near-optimal selection with high probability. Genetic algorithms are a well known and robust approach commonly used in this context.

Should we wait for the selection algorithm to reach 1?

The maximum value that can be achieved for the optimal selection will depend on the balance of rank and diversity requested and the characteristics of the compound set from which the selection is being made. Commonly, this optimal value will be less than 1, unless your data set is small. Even in the cases where it is possible it could take a long time to achieve so we recommend you wait until the plot reaches a plateau.





FAQ

StarDrop: Compound Selection

What is the appropriate balance of quality and diversity?

It is usually beneficial to explore the sensitivity of a selection to the degree of bias chosen before making a final decision. Often, a significant degree of added diversity can be explored for a small decrease in the overall quality of the compounds selected. In this case, it is advisable to spread the risk across diverse compounds, provided synthetic resources permit. Conversely, in some cases, the selection of compounds will remain the same until a large bias toward diversity is selected. In this case, the selection of a diverse set may require an unacceptable decrease in the overall quality of the compounds.

As a general rule, at the earlier stages of a project where little is known about the SAR of the target, it is advisable to bias a selection in favour of diversity. Typically a diversity:rank ratio of 80:20 will sample across the extremes of chemical diversity, whilst still ensuring that top scoring compounds are represented within the selection. As the project moves towards the candidate stage, it will become more important to bias the selection towards 'good' compounds. In this case a diversity:rank ratio of 20:80 may be more appropriate.

Note that a diversity:rank ratio of 100:0 will select molecules on the basis of their structural diversity and the newly selected set will mirror the diversity of the original set (assuming a reasonable sample size of molecules has been selected). A diversity:rank ratio of 0:100 will select molecules entirely on their 'Rank' and the top molecules will be selected.

Can I do a random selection?

Yes, there is an option for random selection. Unlike the biased options, random selection option does not need either chemical structure information or scoring profile results to be run.

Can I save the selection?

Yes, any selection automatically creates a new data set than can in turn be saved as a StarDrop file or exported as .sd or .csv files.

What is the maximum number of compounds that can be selected?

Selection of compounds with a consideration of diversity can be computationally demanding. The computational cost scales as the square of the number of compounds being selected. In practice, a reasonable limit is approximately 1,000 compounds.