

Three way Comparison of Chemical Spaces avoiding Structure Exchange

Jens Loesel¹

¹ Pfizer, Sandwich, United Kingdom

The comparison of chemical space between two compound collections tends to involve knowledge about the individual structures forming the collections. Unfortunately the exchange of structure data between pharmaceutical companies can pose such big administrative hurdles that the scientific effort of a comparison sometimes seems not worthwhile. In addition a generic comparison between two compound collections tends to outline the difference and novelty between them. To investigate the quality of an unknown set more labor intense work is often necessary.

We tried to address both these issues by developing a generic three way comparison between chemical spaces using a fixed six dimensional BCUT space as a frame of reference. Into this space we map a background set, a reference set as well as an unknown set of compounds.

The background and reference set is used to generate scores for individual bins between zero and one – based on enrichment or lack of it in the reference set compared to the background set. The unknown set is then mapped into the same bins and an overall score between zero and one is generated which is based on the occupancy figures of these set in the individual bins.

This work flow allows a comparison of compound collections by exchanging the frame of reference and the binning operation in one way and by getting bin occupancy values back. Multiple reference sets can be mapped into the same space. These reference sets represent the space of known drugs, active compounds in specific target classes or property spaces like high MW or ClogP areas. Each set will generate an individual score for the unknown set – giving an indication for the suitability of the unknown set to be fit for a set of properties or targets.

Once multiple reference sets are generated the process can be automated – allowing the scoring and evaluation of a new collection in mere minutes without manual intervention.

One often cited disadvantage of BCUT descriptor is the inability to use them for a targeted design. For the purpose of this work the obfuscation of chemical structure into Eigenvalues is seen as an advantage. The exchange of raw descriptors like fingerprints always bears the danger of re-engineering of the descriptor via a GA approach into a (similar) structure. This problem doesn't exist in the outlined work flow.

Some examples of scores for Gene classes and properties are shown in the presentation. A discussion will focus on limitations – as for some analysis you still need structure information. The sensibility of the score will be demonstrated using two sets with (known) different Molecular Weight distributions.

1. BCUT reference needed here