

Use of data mining to help identify compounds that are unstable in DMSO

J Hussain, G Harper, Z Blaxill, I Areri, F Saremi-Yarahmadi, S Pickett, P Sidebottom
GlaxoSmithKline Research & Development Limited, Stevenage, UK

Dimethyl sulfoxide (DMSO) has the ability to dissolve a wide range of organic compounds hence it is not surprising that it is the solvent of choice for the solution storage of large compound sets for use in High Throughput Screening (HTS). One of the issues that need to be addressed when new compounds are added to a HTS compound collection is the likely stability of the compounds in DMSO. GSK and other companies [1,2], employ a set of *in silico* substructural filters to remove known unstable compounds from their HTS compound collections. However, the diversity of these compound sets is large hence it is likely that many compounds that degrade in DMSO are not captured by these filters. The presentation will explain some of the work done within GSK to identify further compounds and substructures that are unstable within DMSO.

GSK has done a significant amount of compound quality assurance (QA) work to make sure its legacy HTS compound collection does not contain any impure compounds [1]. Additionally any new compounds that are purchased by GSK go through the QA process before entering the HTS collection. Because of this, GSK now has a large volume of data on compounds that do not pass the QA process. This data is inherently noisy as compound degradation is just one of the many reasons why a compound can fail QA. The talk will describe the use of a data-driven algorithm [3] to mine this noisy data and retrieve substructures that may be unstable in DMSO.

The data driven algorithm will be explained, along with the range of molecular descriptors used in the algorithm, from simple substructures to more complex pharmacophoric representations. The talk will present the results of the data driven analysis and describe subsequent experimental work which shows that a selection of the substructures selected by the data driven algorithm do indeed degrade in DMSO. The future direction of the work will also be discussed; describing how the data-driven technique has potential to highlight those compounds that need to be re-checked for purity to help maintain a high quality compound collection.

1. Lane, S. J.; Eggleston, D. S.; Brinded, K. A.; Hollerton, J. C.; Taylor, N. L.; Readshaw, S. R. Defining and maintaining a high quality screening collection: the GSK experience. *Drug Discovery Today*. **2006**, 11, 267-272.
2. Schopfer, U.; Engeloch, J.; Stanek, J.; Girod, M.; Schuffenhauer, A.; Jacoby, E.; Acklin, P. The Novartis Compound Archive – From Concept to Reality. *Combinatorial Chemistry & High Throughput Screening*. **2005**, 8, 513-519.
3. Harper, G.; Bravi, G. S.; Pickett, S. D.; Hussain, J.; Green, D. V. S.; The Reduced Graph Descriptor in Virtual Screening and Data-Driven Clustering of High-Throughput Screening Data. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 2145-2156.