

Poster presentation

Open Access

## Exploring benchmark dataset bias in ligand based virtual screening

Knut Baumann\* and SG Rohrer

Address: Institute of Pharmaceutical Chemistry, Braunschweig University of Technology, Beethovenstr. 55, 38106 Braunschweig, Germany

\* Corresponding author

from 3rd German Conference on Chemoinformatics  
Goslar, Germany. 11-13 November 2007

Published: 26 March 2008

*Chemistry Central Journal* 2008, **2**(Suppl 1):PI doi:10.1186/1752-153X-2-S1-PI

This abstract is available from: <http://www.journal.chemistrycentral.com/content/2/S1/PI>

© 2008 Baumann and Rohrer

A common finding of many reports evaluating VS methods is that validation results vary considerably with changing datasets, i.e. chemical space of the active ligands. It is assumed that these dataset specific effects are caused by the self-similarity and cluster structure inherent to these datasets.

As a first step, an experimental setup was developed that isolated dataset composition as the sole factor of variance influencing VS performance. The Hert-Willet benchmark datasets have been widely used for the validation of ligand based VS protocols. Various sampling strategies (D-optimum design, Onion-design, minimum distance design) were employed to generate archetypal subsamples from these datasets: (1) maximum diversity subsets, (2) space filling samples and (3) subsets with the minimum intra-set diversity. The analysis of the varying VS performance on these prototype datasets showed that dataset composition does indeed exert a critical influence on VS validation and identified local clustering and global spread of the datasets with respect to the set of decoys as the factors with the highest impact on VS performance.

Keeping the concept of chemical space in mind, it is reasonable to make use of the field of spatial statistics, which offers a wealth of methods for the analysis of clustering, patchiness and dispersion of datasets. By employing these, we were able to analyse the spatial composition of the benchmark datasets in more detail and derive several rules of thumb for choosing unbiased datasets for evaluating ligand based VS methods.