

Maximum Unbiased Validation (MUV) of Ligand Based Virtual Screening

K. Baumann¹

S.G. Rohrer¹

¹ *Institute of Pharmaceutical Chemistry, Beethovenstr. 55, Braunschweig University of Technology, 38106 Braunschweig, Germany. Phone: +49-531-3912751, Fax: +49-531-3912799.*

E-mail: k.baumann@tu-braunschweig.de

A common finding of many reports evaluating ligand-based virtual screening methods is that validation results vary considerably with changing benchmark datasets. Such effects are caused by the redundancy and self-similarity inherent to those datasets. These phenomena manifest themselves in the datasets' representation in descriptor space, which is termed the dataset *topology*. Three key findings that allow the design of MUV datasets are presented.

1.) A methodology for the characterization of dataset topology based on spatial statistics is introduced. The method is non-parametric and can deal with arbitrary distributions of descriptor values. It utilizes two cumulative distribution functions of distances in chemical space, called the “nearest-neighbor function” $G(t)$ and the “empty space function” $F(t)$, which reflect the distributions of active-active and decoy-active distances, respectively. With this methodology it is possible to associate differences in virtual screening performance on different datasets with differences in dataset topology (correlation coefficient: 0.92, $n = 234$). Moreover, the better virtual screening performance of certain descriptors can be explained by their ability of representing the benchmark datasets by a more favorable topology (correlation coefficient: 0.91, $n = 234$).

2.) It is shown, that the topologies of certain benchmark datasets cause over-optimistic validation results. Spatial statistics analysis as proposed here allows the detection of such biased datasets.

3.) $G(t)$ and $F(t)$ can effectively be used as objective functions in the design of unbiased benchmark datasets. In order to design benchmark datasets with minimum bias, datasets should exhibit the lowest possible level of self-similarity, which can be monitored by $G(t)$. Conversely, the set of decoys should be selected as similar to the benchmark set as possible. This process can efficiently be guided by $F(t)$. Here, we apply this design strategy to a collection of datasets carefully selected from the bio-activity data available in PubChem.

The resulting maximum unbiased benchmark datasets are validated by retrospective virtual screening simulations and spatial statistics analysis.

The presented benchmark datasets will be available for download on our web-page. (<http://www.pharmchem.tu-bs.de/forschung/baumann/>)