G. Marcou[1]
D. Horvath[1]
A. Varnek[1]

A MEASURE FOR QSAR MODELABILITY AND FOR DATA ORGANIZATION

[1] Laboratory of Chemoinformatics, Faculty of Chemistry, University of Strasbourg, 1, rue Blaise Pascal, 67000 Strasbourg, France;

The lack of adequacy between a target property and molecular descriptors encoding chemicals structures may significantly deteriorate structure-activity models. Thus, an *a priori* assessment of data modelibility within a particular descriptor space may significantly reduce computational costs of QSAR modeling [1]. In this work, we propose to assess chemical data modelibility using the Hilbert Schmidt Independence Criterion (HSIC) [2] still rarely used in chemoinformatics. HSIC is an empirical estimate of the Hilbert- Schmidt norm of the cross-covariance operator. Here we demonstrate that HSIC can efficiently be used to estimate *a priori* the adequacy of ensemble of molecular descriptors to a given chemical library both in simple QSAR and in multi-task learning studies.

The HSIC could also be used for data visualization purpuses. Thus, this parameter helps to reorder a dataset in such a way that chemical structures located near a given position are chemicaly related (see Figure 1).

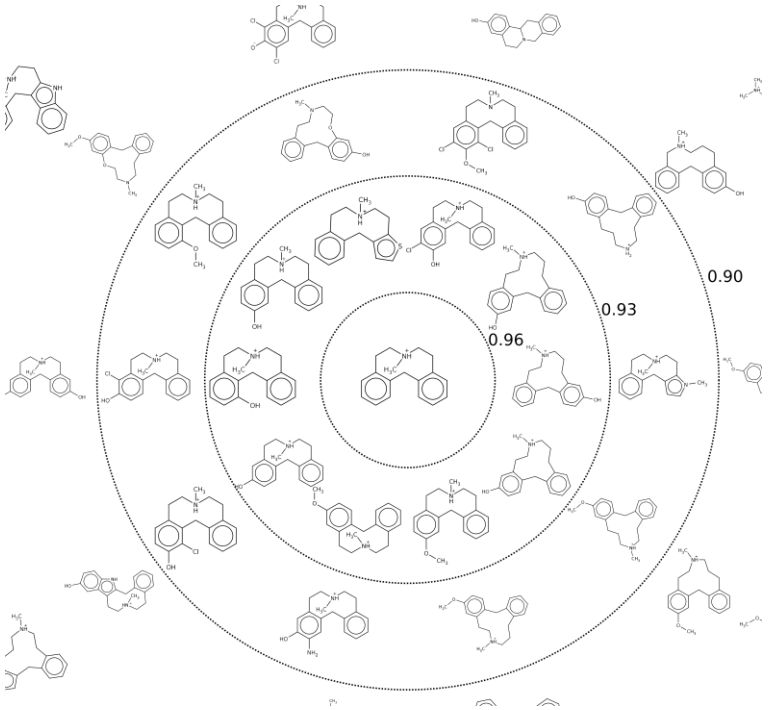

Figure 1 : Ordering of ligands of the human dopamine D5 receptor around a central scaffold. The dot lines indicate the minimum level of similarity of the enclosed compounds with the central scaffold.

1. Golbraikh, A.; Muratov, E.; Fourches, D.; Tropsha, A., *Journal of chemical information and modeling* **2014,** *54* (1), 1-4.
2. Gretton, A.; Bousquet, O.; Smola, A.; Schölkopf, B. In *Measuring statistical dependence with Hilbert-Schmidt norms*, Algorithmic learning theory, Springer: 2005; pp 63-77.

We thank Dr I. Baskin from Lomonsov Moscow State University (Russia) for stimulating discussion