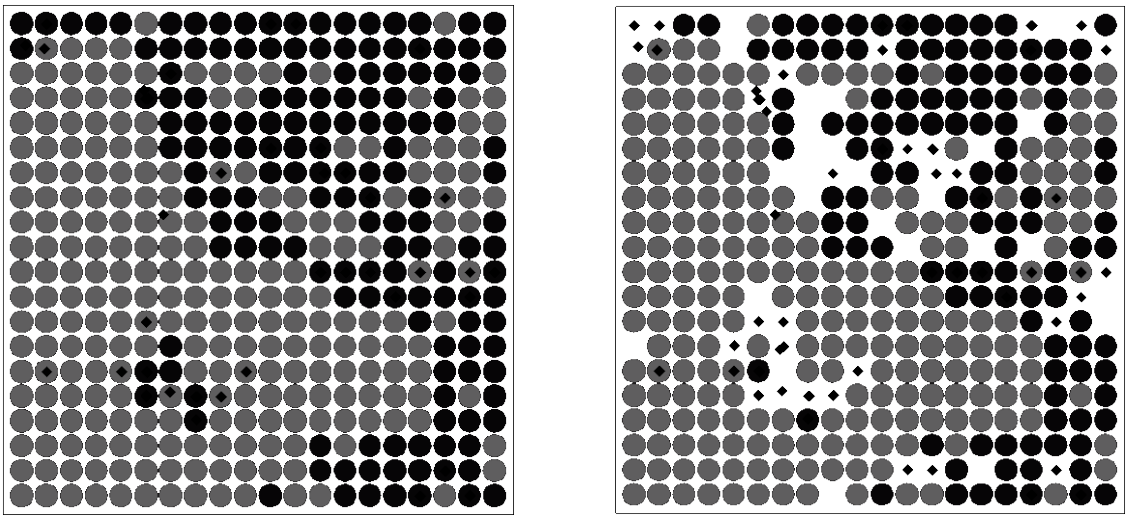
[P51] GTM Applicability Domain for Classification and Regression models

Héléna Gaspar[1], Timur Gimadiev[1,2], Gilles Marcou[1], Dragos Horvath[1], Alexandre Varnek[1]

[1]*Laboratoire de Chémoinformatique, UMR 7140 CNRS, Université de Strasbourg, 1 rue B. Pascal, 67000 Strasbourg, France.*
[2]*Butlerov Institute of Chemistry, Kazan Federal University, Kazan, Russia*

Generative Topographic Mapping (GTM) is a dimensionality reduction method, and the probabilistic counterpart of Kohonen maps. Each *i*-th molecule in *N*-dimensional initial space is projected into the *k*-th node of 2D latent space with a probability R_{ik}. so that each compound is represented both by a mean position (a point) on a 2D map, and a probability distribution R_i, which may be used for predictions of activity (property) of new compounds.
Here, we suggest several different GTM-based definitions of applicability domain (AD) of both regression and classification models. This concerns the approaches involving: (i) a likelihood threshold, (ii) relative population of nodes, (iii) class entropy, and, (iv) ratio of classes' probabilities. These approaches are demonstrated for regression models for stability constants of ligand-metal complexes, and GTM-based classification models [1] for Biopharmaceutics Drug Disposition Classification System (BDDCS) [2] and inhibitors of P-glycoprotein 1 (Pgp), an ATP-dependent efflux pump.



CPF = 1, BA= 0.84 , coverage = 100 % ***CPF = 4, BA= 0.90, coverage = 78 %***

Figure 1. Graphical interpretation of the applicability domain for GTM classification models. On the map prepared for the entire set of 1568 molecules of inhibitors (dark grey) and non inhibitors (light grey) of P-glycoprotein 1, the color stands for the class having the highest probability compared to the other in a given node. Black points correspond to incorrectly classified molecules. The increase the class prevalence factor (**CPF** = *Probability of major class / Probability of another class*) from CPF=1 (left) to 4 (right) results in shrinking the AD area. This leads to the decrease of the number of molecules inside AD (coverage), on one hand, and to the increase of the model's performance (Balance Accuracy, BA), on the other hand.

[1] Kireeva, N., Baskin, I. I., Gaspar, H. A., Horvath, D., Marcou, G. and Varnek, A., Mol. Inf., 31 (2012): 301–312. doi: 10.1002/minf.201100163
[2] Gaspar, H. A., Marcou, G., Horvath, D., Arault, A., Lozano, S., Vayer, P., and Varnek, A. J. Chem. Inf. Model. 53 (2013), 3318–3325.