Generative Topographic Mapping approach as a Ligand-based Virtual Screening tool

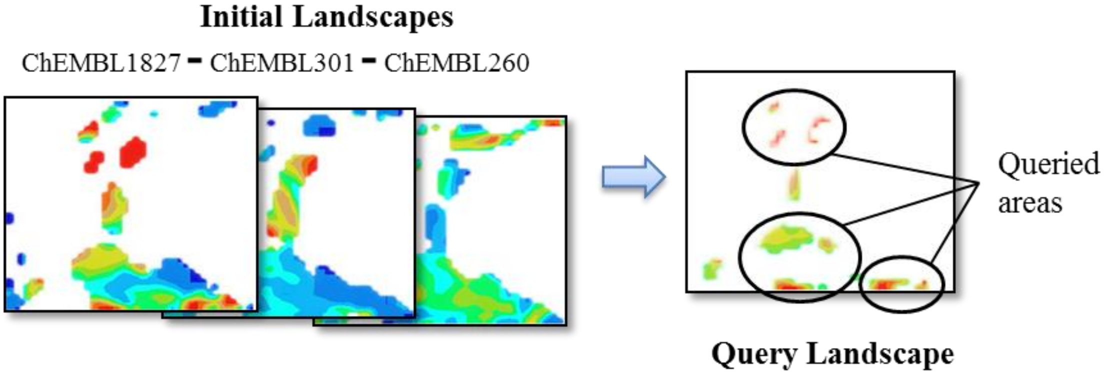
Arkadii Lin[a,b], Dragos Horvath[a], Gilles Marcou[a], Alexandre Varnek[a], and Bernd Beck[b]

[a] *Laboratory of Chemoinformatics, Faculty of Chemistry, University of Strasbourg, 4, Blaise Pascal str., 67081 Strasbourg, France*

[b] *Department of Medicinal Chemistry, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorferstrasse 65, 88397 Biberach an der Riss, Germany*

Generative Topographic Mapping (GTM) is a dimensionality reduction method that can be used for large chemical data visualization and analysis. [1] Recently it was tested as a tool for large chemical databases comparison (PubChem-17, ChEMBL-17, and FDB-17). [2] It was also tested as a machine learning method for Quantitative Structure-Activity Relation (QSAR) tasks. [3, 4] However, it was not fully tested as a tool for Ligand-based Virtual Screening (LBVS) procedure, where large chemical databases are used.

In this project, GTM is compared with the most popular methods for LBVS such as Random Forest, Neural Networks, and Similarity search with data fusion. Within the usual GTM approach, where each model is built for the particular target, a “universal” map approach is also tested as a method where only one map is used to represent activity landscapes of any number of targets or properties. This enables the querying by activity profile (focusing on zones with jointly favorable predictions for all targeted properties, see Figure below).



Benchmarking results show that GTM is competitive in terms of performance. For example, “universal” maps built and having activity landscapes calibrated on > 1.5M ChEMBL compounds are excellent discriminators for the Directory of Useful Decoys (DUD) compounds (excluding the ones present in ChEMBL, to ensure strict “external” validation). For 9 biological targets ROC AUC values ranged within 0.7÷0.8.

Furthermore, GTM has some important advantages in terms of usage, notably the ability to intuitively visualize the chemical space, and its support of multiple predictive landscapes on a single map. Calculation times are independent of reference set sizes (unlike in pairwise similarity searching).

[1] Bishop CM, Svensén M, Williams CK. GTM: *Neural computation*. **1998**, 10(1), 215-34.
[2] Lin, A., Horvath, D., Afonina, V., Marcou, G., Reymond J.L. and Varnek, A. *ChemMedChem*. doi:10.1002/cmdc.201700561.
[3] Gaspar HA, Baskin II, Marcou G, Horvath D, Varnek A. *Journal of Chemical Information and Modeling*. **2015**, 55(1), 84-94.
[4] Sidorov P, Gaspar H, Marcou G, Varnek A, Horvath D. *Journal of Computer-Aided Molecular Design*. **2015**, 29(12), 1087-108.