[P36] Mapping of the Chemical Universe vs Available Chemical Space of Lead-Like Compounds

Afonina V.A.,[1,2] Varnek A.,[1,2] Horvath D.[1]

[1] *Laboratoire de Chemoinformatique, UMR 7140, CNRS-Univ. Strasbourg, 1 rue Blaise Pascal, 67000 Strasbourg, France*
[2] *Laboratory of Chemoinformatics, Butlerov Institute of Chemistry, Kazan Federal University, Kazan, Russia*

The chemical space or "Universe" [1] is the ensemble of all possible molecules, which is believed to contain at least 10^{60} organic molecules of possible interest for drug discovery. The ''available drug-like chemical space'' is formed by all the ~10^8 synthetized and tested molecules, and is a dwindling small subset of the entire chemical space – yet, a huge collection by the standards of the effort needed to process the chemical information it contains. It is also extremely biased, due to public health and economic factors, in favor of products of cheap chemical synthesis targeted against major diseases, with a robust business potential (kinase and GPCR inhibitor candidates forming together the vast majority of compounds reported in public bioactivity databases and corporate databases alike).

By contrast, the recent exhaustive enumeration of all compounds with up to 17 atoms[2] (GDB17) provides a much needed, unbiased view of the chemical Universe. Albeit restricted to lower-size drug-like compounds, this collection is a much-needed baseline against which one may highlight the various biases and diversity "holes" affecting the so-far available compound collections. This is the undertaking of the present work, aimed at comparing the "baseline" GDB17 to size-matching subsets (≤ 17 heavy atoms) of publicly available compound libraries ChEMBL and PubChem, respectively.

A sample of ~10^7 compounds from GDB17 was compared to the complete ChEMBL17 (10^5 entries of ≤ 17 heavy atoms) and PubChem17 (10^7 entries) subsets. This comparison is driven by Generative Topographic Mapping (GTM), a probabilistic topology-preserving dimensionality reduction method, which projects the D-dimensional chemical space (here, ISIDA atom-pairs descriptors) onto a two-dimensional space.[3,4] Mapping of compound sets of above-mentioned sizes required an in-depth – and still ongoing – rethinking of map generation strategies. Both the use of representative, small frame sets and the use of the entire GDB17-subset were shown to be technically envisageable strategies for map building, and results highlighted various interesting differences between the three compound collections.

References:

1. Reymond, J.-L. van Deursen, R., Blum, L. C., Ruddigkeit, L. Medicinal Chemistry Communications 2010, 1, 30-38.
2. Ruddigkeit, L., van Deursen, R., Blum, L. C., Reymond, J.-L. J. Chem. Inf. Mod. 2012, 52, 2864-2875.
3. Bishop, C.M., Svensén, M., Williams, C.K.I. Neurocomputing 1998, 21, 203–224.
4. Gaspar, H. A., Baskin, I.I., Marcou, G., Horvath, D., Varnek, A. J. Chem. Inf. Mod. 2015, 55(1), 84–94.