
Vincent Le Guilloux¹; Guillaume Guenegou¹; Stéphane Bourg²; Julie Dubois-Chevalier¹,³; Luc Morin-Allory¹; Lionel Colliandre¹

1) Institut de Chimie Organique et Analytique, ICOA, Laboratoire de Chémoinformatique, UMR CNRS 6005, 2) Fédération de recherche “Physique et Chimie du Vivant”, FR CNRS 2708, 3) Laboratoire d’Informatique Fondamentale d’Orléans, LIFO, Université d’Orléans, BP 6759, F-45067 Orléans Cedex2, France

The concept of chemical space provides a way to represent a set of molecules in a given reference space having a specific metric. Each molecule is generally represented by its coordinates in the reference space, and most often mapped in a visual and interpretable way. Besides pure visual description, this concept is useful to tackle various aspects related to chemical library characterization: How diverse is the library? Which regions of the chemical space are filled by the library? Which compounds should be selected to get a better chemical space coverage / chemical diversity? Where are located biologically active compounds?

Several methods have been proposed to define chemical spaces and describe / compare chemical libraries. Each method has its own specificity, and allows describing molecules at various levels and using different metrics. Interestingly, only few of these methods are available for direct use, and we are still lacking common reference chemical spaces which could be used by everyone to describe chemical libraries.

We present herein a study in which reference spaces designed to be representative of commercially available HTS compounds are defined. We have extracted several random subsets from a database of 4.2 millions unique commercial compounds coming from 51 providers, and applied a strict protocol to obtain stable and representative chemical spaces using Principal Component Analysis on various 2D and 3D sets of descriptors. We illustrate the usefulness of these reference chemical spaces by characterizing several well known compounds libraries.

These reference chemical spaces will be bundled with the Screening Assistant 2 (SA2) open source software, and thus freely available to anyone.