**Feature interrelation profiling**

I. Čmelo, Prague/CZ, M. Voršilák, Prague/CZ, W. Dehaen, Prague/CZ, D. Svozil, Prague/CZ

Dr. Ivan Čmelo, UCT Prague, Technická 5, 166 28 Prague 6 – Dejvice

Chemical structures are routinely represented by vectors of their binary features, such as structural fingerprints. The feature vectors often serve as a basis for direct structure comparisons in similarity searches and visualizations, or as an input to more advanced methods and models. Feature interrelation profiling (FIP) is a generic information-theory based extension of these feature vector core concepts. FIP uses various forms of Kullback–Leibler divergence matrices to store information about feature co-occurrence probabilities within various sets of chemical structures, thus forming their interrelation profiles.

The resulting interrelation profiles can be used to directly compare chemical structures and their sets based not only on the amount of individual shared features, but also on how consistent is a given feature combination with a reference interrelation profile. To quantify said consistency of given feature combination against a reference interrelation profile, we propose a measurement of “relative feature tightness” (RFT). For a given chemical structure or their set, RFT weighs all observed feature co-occurrences against their co-occurrence scores within a given reference interrelation profile. RFT yields a scalar value that describes the overall match in present feature combinations, much like Tanimoto similarity describes the overall match in present features.

This method of quantifying the relative “favorableness” of feature combinations, as well as the interrelation profiles themselves, present additional layers of information that can be easily used in similarity searches, visualizations and other methods as a complement to the standard feature vectors. This poster shows the theoretical background of FIP, its uses for visualization and its pilot application on assessing synthetic accessibility of chemical structures, with performance close to that of dedicated methods like SAScore and SYBA. [1]

Literature:

[1] I. Čmelo, M. Voršilák, D. Svozil, J Cheminform 13, 3 (2021).