Medoid-based “satellite” sampling methods in chemical space visualization

Kenneth Lopez-Perez

Department of Chemistry, University of Floirda, Gainesville FL

Visualization of the chemical space is useful in drug discovery for library screening, and as a guide in rational design. However, the sheer volume of even comparatively small sub-sections of chemical space implies that we need to use approximations at the time of navigating through chemical space. “ChemMaps” is a popular visualization methodology that approximates the distribution of compounds in large libraries based on the selection of “satellite” compounds that yield a similar mapping of the whole dataset when PCA on similarity matrix is performed. While in previous works the “satellite” sampling is done by choosing molecules randomly, in this work we propose three new sampling methods: uniform, medoid-centered, and mediod-diversity. As well, extended similarity indices are used to speed up similarity matrices and medoid calculations. These approaches render a more reproducible and faster way of building the ChemMaps.