AZIMUTH: A Hierarchical GTM-Based Chemical-Space Atlas for Adaptive Zooming, Fast Similarity Search, and Library Comparison

Orkhan Abdullayev, Alexey Orlov, Dragos Horvath, Alexandre Varnek

Laboratory of Chemoinformatics, UMR 7140 CNRS, University of Strasbourg, Strasbourg, France

As chemical datasets expand to as many as 10^{26} accessible or synthetically feasible structures, conventional chemical-space analysis pipelines become increasingly difficult to apply at industrial scale. For chemical space visualization, atlas-based chemography provides a natural strategy for handling ultra large datasets: instead of relying on a single projection, chemical space is represented as a hierarchically organized collection of maps, enabling both a global overview and high-resolution local navigation. Recent studies have shown that such hierarchical atlases can be applied to ultra large chemical collections, including ones containing billions of compounds[1,2]. However, scaling these methods to encompass ever-larger combinatorial collections faces severe bottlenecks, as the required memory and computational time quickly become prohibitive. Therefore, new methods capable of navigating these massive, unenumerated spaces without explicit construction are urgently required.

Here we introduce AZIMUTH, a scalable hierarchical framework for Generative Topographic Mapping (GTM)-based chemical space visualization and analysis that extends chemical space atlases to ultra large spaces, including explicitly enumerated libraries and non-enumerated combinatorial spaces. AZIMUTH introduces four main innovations: an enhanced implementation of CoLiNN method[3] for enumeration-free projection of combinatorial libraries from synthon feature vectors and reaction representations, a new non-maximum suppression (NMS)-based zooming strategy for constructing non-redundant hierarchical GTM atlases, a fast atlas-guided molecular similarity search method, and a hierarchy-aware library comparison approach. The resulting methodology was successfully applied to the visualization of a dataset comprising 388 diverse DNA-encoded library (DEL) subsets and 2.1 million ChEMBL33 compounds. Relative to the global root map, AZIMUTH increased mean neighborhood preservation in the leaf GTMs by about 27%, with the best leaf improving by up to 67%. In hierarchical similarity search, AZIMUTH pruned 99% of the search space while retaining about 70% threshold-conditioned recall at K=5 for exact neighbors at $T_c \geq 0.85$. For library similarity, it showed high efficiency in selecting DELs best covering the reference space of CDK2 inhibitors with Spearman $\rho = 0.86$ between coverage fractions in the GTM-based and initial descriptor spaces.[4]

Bibliography:

[1] Zabolotna, Y. ; et al. J. Chem. Inf. Model. 62 (2022) 4537–4548.
[2] Flores Sepúlveda, A. ; Reymond, J.-L. J. Chem. Inf. Model. (2026) 6c00420.
[3] Pikalyova, R. ; Akhmetshin, T. ; Horvath, D. ; Varnek, A. Mol. Inform. 44 (2025) e202400263.
[4] Plyer, L. ; Orlov, A. A. ; Akhmetshin, T. N. ; Yeghyan, E. ; Bonachera, F. ; Horvath, D. ; Varnek, A. Mol. Inform. 45 (2026) e70026.