MolSpace: Visualization and Analysis of De Novo Generation

Maksim Vendin[1], Arkadii Lin[1], Ien An Chan[1], Yury Kashkur[1], Vladimir Aladinskiy[1], Alex Aliper[1], Alex Zhavoronkov[1,2,3]

[1] Insilico Medicine AI Limited, Level 6, Unit 08, Block A, IRENA HQ Building, Masdar City, Abu Dhabi, UAE

[2] Insilico Medicine Canada Inc., 3710-1250 Ren´ e-L´ evesque Blvd W, Montreal, Quebec, H3B 4W8, Canada

[3] Insilico Medicine Hong Kong Ltd., Unit 310, 3/F, Building 8W, Phase 2, Hong Kong Science Park, Pak Shek Kok, New Territories, Hong Kong, HongKong SAR, China.

Traditionally, de novo design novelty is assessed by dissimilarity to the reference data [1]. This approach ensures the generated structures are new, even though only a tiny change in the scaffold led to that. This single-number approach (i.e., similarity metric) loses the context of chemotype neighboring in the chemical space, which may make the results interpretation confusing. To keep the full picture, one may need to compare the generated structures with vast public chemical spaces to see how far the generated structures are from the public domain. This requires specialized tools capable of handling big data.

We propose MolSpace, which is a part of the Chemistry42 platform [2]. It delivers a highly GPU-optimized PyTorch implementation of Generative Topographic Mapping (GTM) [3, 4], which makes it possible to train a map of 100x100 nodes using thousands of training structures in minutes. A Comparison Landscape integrated into MolSpace allows projecting generated and reference structures onto the same map, providing both local and global perspectives on the chemical space organization. The local map indicates whether any reference compounds occupy the generated data space, highlighting regions of local novelty. In contrast, the global map assesses whether these generated structures occupy previously unpopulated regions of the entire reference space or overlap with known chemistry. Beyond spatial occupancy, MolSpace enables evaluation of target property distributions and extracting promising structures from the target property(ies) profile perspective.

MolSpace was applied to compare 3,309 generated putative JAK3 inhibitors against the 2.4-million-compound ChEMBL database. By correlating local Comparison and multi-parameter Regression Landscapes, we successfully isolated 16 target-exclusive clusters containing 306 molecules. We categorized these as "hidden treasures" [4]—structurally unique chemotypes that lack public analogs yet still exhibit highly optimized drug-like properties. The most similar to ChEMBL generated structure had a Tanimoto value of 0.61. The selected structures were further sent to retrosynthesis, and 185 were solved using the 308K in-stock building blocks.

Bibliography:

[1] Özçelik, R.; Brinkmann, H.; Criscuolo, E.; Grisoni, F. J. Chem. Inf. Model. 65 (2025) 7352–7372.
[2] Ivanenkov, Y.A.; Polykovskiy, D.; Bezrukov, D.; Zagribelnyy, B.; Aladinskiy, V.; Kamya, P.; Aliper, A.; Ren, F.; Zhavoronkov, A. J. Chem. Inf. Model. 63 (2023) 695–701.
[3] Lin, A.; Beck, B.; Horvath, D.; Marcou, G.; Varnek, A. J. Comput. Aided Mol. Des. 34 (2020) 805–815.
[4] Zabolotna, Y.; Lin, A.; Horvath, D.; Marcou, G.; Volochnyuk, D.M.; Varnek, A. J. Chem. Inf. Model. 61 (2021) 179–188.