AI/ML approaches for generative chemical design and virtual screening acceleration.

Alexander Tropsha

UNC Eshelman School of Pharmacy,
University of North Carolina at Chapel Hill, Chapel Hill, NC 27516, USA.

Recent advances in high throughput technologies have lead to the “Bing Bang” expansion of the purchasable chemical universe, which created substantial computational challenges for traditional approaches to virtual screening (VS) such as similarity searching and molecular docking. I will describe novel resource- and cost-effective approaches to both ligand-based (LB) and structure-based (SB) VS. To modernize LBVS, we have developed a novel approach to molecular embedding dubbed SALSA (Semantically-Aware Latent Space Autoencoder)[1], a transformer-autoencoder modified with a contrastive task, tailored specifically to learn and preserve graph-to-graph similarity between molecules in the latent space. Building upon SALSA, we also developed SmallSA (Small Structurally-Aware embeddings)[2] employing a combination of low-dimensional chemical embeddings by SALSA and a k-d tree data structure to achieve ultra-fast nearest neighbor searches. Searches on over one billion chemicals execute in less than a second on a single CPU core, five orders of magnitude faster than the brute-force approach. In SBVS space, we have developed a novel computational methodology termed HIDDEN GEM (HIt Discovery using Docking ENriched by GEnerative Modeling)[3]. Our workflow uniquely integrates machine learning, generative chemistry, massive chemical similarity searching (which takes advantage of SmallSA) and molecular docking of small, selected libraries in the beginning and the end of the workflow. For each target, HIDDEN GEM nominates a small number of top-scoring virtual hits prioritized from ultra-large purchasable libraries. I will discuss the above methods and their application to identifying novel compounds active experimentally antiviral targets. Methods and tools discussed in this presentation contribute to the overarching movement in cheminformatics in support of democratizing computational drug discovery.

References

1. Kirchoff, K. E., Maxfield, T., Tropsha, A. & Gomez, S. M. SALSA: Semantically-Aware Latent Space Autoencoder. *Proc. AAAI Conf. Artif. Intell.* **38**, 13211–13219 (2024).
2. Kirchoff, K. E. *et al.* Utilizing Low-Dimensional Molecular Embeddings for Rapid Chemical Similarity Search. in *Advances in Information Retrieval* (eds. Goharian, N. et al.) 34–49 (Springer Nature Switzerland, Cham, 2024). doi:10.1007/978-3-031-56060-6_3.
3. Popov, K. I., Wellnitz, J., Maxfield, T. & Tropsha, A. HIt Discovery using docking ENriched by GEnerative Modeling (HIDDEN GEM): A novel computational workflow for accelerated virtual screening of ultra-large chemical libraries. *Mol. Inform.* **43**, e202300207 (2024).