Reminiscence about the Future of Cheminformatics.

Alexander Tropsha

UNC Eshelman School of Pharmacy, UNC-Chapel Hill, Chapel Hill, USA.

The field of Cheminformatics has always been one of the earliest adopters of innovations in computational methods. Multiple algorithms leveraging fundamental advances in ML such as (deep) neural networks, multi-dimensional scaling, generative topographic mapping, support vector machines, natural language processing, generative AI and other approaches have contributed to the evolution of the field. However, one may argue that its major challenges remained unchanged including tasks such as chemical similarity searching, QSAR modeling, molecular docking, data visualization, and rational design of new chemical entities predicted to have the desired property or activity. Thus, the history of the field can provide hints about its future. I will review how computational tools that address fundamental cheminformatics challenges have evolved with the major transformative component of the field: the continuing growth in the size of molecular datasets including recent Big Bang expansion of synthetically feasible and purchasable Chemical Universe. I will discuss recent tools developed in our laboratory and elsewhere that address challenges posed by Big Data. Examples include a novel approach to molecular embedding dubbed SALSA (Semantically-Aware Latent Space Autoencoder), a transformer-autoencoder modified with a contrastive task, tailored specifically to learn and preserve graph-to-graph similarity between molecular representations in the latent space.[1] I will also describe SmallSA[2] (Small Structurally-Aware embeddings) that leverages SALSA by employing a combination of low-dimensional chemical embeddings and a k-d tree data structure to achieve ultra-fast nearest neighbor searches in ultra-large libraries. I will describe a novel computational methodology termed HIDDEN GEM (HIt Discovery using Docking ENriched by GEnerative Modeling)[3] that greatly accelerates structure based virtual screening. This workflow uniquely integrates machine learning, generative chemistry, massive chemical similarity searching and molecular docking of small, selected libraries in the beginning and the end of the workflow. In place of using LLM models, in the end I will attempt to hallucinate about the future of cheminformatics including greater integration with experiment via self-driving labs, proliferation of cheminformatics tools and concepts across multiple disciplines, and democratization of drug discovery.

Bibliography :

1. Kirchoff, K. E. et al. SALSA: Semantically-Aware Latent Space Autoencoder https://arxiv.org/abs/2310.02744v1 (accessed Feb 15, 2024).

2. Kirchoff, K. E. et al. Utilizing Low-Dimensional Molecular Embeddings for Rapid Chemical Similarity Search. https://arxiv.org/abs/2402.07970

3. Popov, K. I. et al. *Mol. Inform.* **2024**, *43*, e202300207.