Chemical Language Models for Molecular Design

Jürgen Bajorath[1,2,3]

[1]Department of Life Science Informatics and Data Science, B-IT, University of Bonn, Friedrich-Hirzebruch-Allee 5/6, D-53115 Bonn.
[2]Lamarr Institute for Machine Learning and Artificial Intelligence, Friedrich-Hirzebruch-Allee 5/6, D-53115 Bonn.
[3]Department of Biological Structure, School of Medicine, University of Washington, Seattle, WA 98195.
E-mail: bajorath@bit.uni-bonn.de

In chemoinformatics, computational models based on textual representations of chemical structure are becoming increasingly popular. These models were adapted from the field of natural language processing and are therefore referred to as language models (LMs). Generally, LMs translate one sequence of characters into another. Therefore, chemical LMs must learn the vocabulary and syntax used to represent molecules as well as conditional probabilities for the occurrence of characters in sequences depending on the preceding characters. LMs are built using neural network variants to enable alternative learning strategies. Preferred architectures include recurrent neural networks (RNNs) and transformers, which consist of multiple encoder and decoder modules with attention (importance) functions. The versatility of LMs in addressing various machine translation tasks and in conditioning them on diverse properties provides new opportunities for generative molecular design. Different chemical and biochemical LMs are introduced that are derived for specific applications in medicinal chemistry including fragment-based generative modeling following molecular hierarchies, the design of three-dimensional scaffolds, the prediction of new active compounds from protein sequence motifs, prediction of activity cliffs and highly potent compounds, or the extension of analogue series with increasingly potent compounds. Analogue series extension is complemented with computational methods for the systematic identification of structure-activity relationship (SAR) transfer events.