Language Models for Drug Design

Jürgen Bajorath[1,2,3]

[1]Department of Life Science Informatics and Data Science, B-IT,
[2]LIMES Institute – Program Chemical Biology and Medicinal Chemistry,
[3]Lamarr Institute for Machine Learning and Artificial Intelligence,
University of Bonn, 53115 Bonn, Germany

In drug discovery, language models (LMs) are used for a variety of predictions that can be framed as machine translation tasks. These LMs must learn the vocabulary and syntax of tokenized molecular representations and conditional probabilities for the occurrence of characters in sequences depending on the preceding characters. Preferred LM architectures include recurrent neural networks and transformers comprising multiple encoder and decoder modules with attention functions. The versatility of LM learning using compound and target information combined with context-dependent rules provides new opportunities for attempting predictions that were difficult to address or infeasible thus far. Exemplary applications using LMs of different architecture are discussed.