ChORISO: a highly curated organic reaction SMILES dataset

V. Sabanza Gil[1], A. M. Bran[1], M. Franke[1], P. Schwaller[1,2]*, J. Luterbacher[1,2]*

[1]LIAC, EPFL EPFL Lausanne, [2]NCCR Catalysis

Artificial Intelligence (AI) and machine learning (ML) have been successfully applied to core problems in organic chemistry, like reaction outcome prediction and synthesis planning (1). Many of the underlying ML models are commonly trained using data in the form of Simplified Molecular Input Line Entry System notation2 (SMILES). Despite the popularity and use of some public available datasets like the US Patent and Trademark Office3 (USPTO), high quality data are scarce and difficult to extract manually. Therefore, new publicly available curated data could improve and leverage the existing ML models for chemical reaction tasks.

This work presents ChORISO (Chemical Organic ReactIon SMILES Omnibus), a new benchmark containing highly curated organic chemistry reaction SMILES. We have completed, processed and cleaned 3.2 M reaction SMILES extracted from chemical literature4, obtaining a final dataset with around 700k examples. We have analysed the most relevant features and compared them to the standard USPTO dataset. In addition, we have used our new dataset to train and benchmark different reaction prediction models (molecular transformer and graph-2-SMILES). This work offers a new high-quality organic reaction SMILES dataset which can contribute to the development and assessment of new chemical reaction prediction models.

References:
1. Schwaller et al.; WIREs Comput Mol Sci. 12 (2022) e1604.
2. Weininger D. J Chem Inf Comput Sci. 28 (1988) 31-6
3. Lowe, D.; Extraction of Chemical Structures and Reactions from the Literature. Ph.D. Thesis, University of Cambridge, 2012
4. S. Jiang et al. IEEE Access, 9 (2021) 85071-85083.