04-2

RetroSynthWAVE: An Open-Source Software Platform for Efficient Chemical Synthesis Research

Haris Hasic [1, 2]
hasic@cb.cs.titech.ac.jp

Takashi Ishida [1]
ishida@c.titech.ac.jp

[1] Department of Computer Science, School of Computing, Tokyo Institute of Technology, W8-85, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
[2] Elix Inc., Daini Togo Park Building 3F, 8-34 Yonbancho, Chiyoda-ku, Tokyo 102-0081, Japan

Keywords: Machine Learning – AI Method Development, Data Curation, Data Visualization

The RetroSynthWAVE project aims to establish a systematic, open-source software platform focused on the field of chemical synthesis. It enables quick and efficient research for beginners as well as advanced users by providing software packages that cover the following synthesis-related functionalities: **W**idgets and helpers, **A**ggregated chemical compound and chemical reaction data, **V**ariety of popular existing model implementations, and **E**valuation metrics. Each of the software packages can be used independently, as well as within the project software stack.

RetroSynthWAVE: HANA is the first and fundamental software package of the project. The name is derived as an acronym for **H**elpers **AN**d **A**ccessories, and it represents a utility wrapper module that encapsulates all of the essential libraries (e.g., RDKit [1], RDChiral [2]) and offers additional fundamental functionalities while being easy to use.

RetroSynthWAVE: COCORO is the second software package of the project which is developed using the functionalities from the previous one. The name is derived as an acronym for **CO**llection of Chemical **CO**mpound and **R**eacti**O**n Data, and it represents an easy-to-use data platform that automates the retrieval, cleaning and featurization of available chemical information datasets. (e.g., ChEMBL [3], USPTO [4])

RetroSynthWAVE: CO-OP is the third software package of the project which is developed using the functionalities from the previous two. The name is derived as an acronym for **CO**llection **O**f **P**opular Models, and it represents an easy-to-use model platform for the reimplementation of popular existing (retro)synthesis models using the PyTorch library. It enables other users to submit the implementation of new models in a standardized fashion.

RetroSynthWAVE: REFEREE is the final software package of the project. The name is derived as an acronym for **R**etroactive **EF**ficiency M**E**t**R**ics **E**valuation Fram**E**work, and it represents a utility module that enables the definition of user-defined, pre-existing, and novel evaluation metrics for (retro)synthesis-focused models. Furthermore, by using the functionalities from all of the previous software packages, it enables the retroactive application of new metrics on pre-existing models thus enabling more advanced benchmarking of all relevant models.

[1] RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org/. (2021.08.13)
[2] Coley, C.W.; Green, W.H.; Jensen, K.F.; RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application, *Journal of Chemical Information and Modeling*, 2019, 59, 6, 2529-2537.
[3] ChEMBL Database - EMBL-EBI: https://www.ebi.ac.uk/chembl/. (2021.08.13)
[4] Lowe, D. M. Extraction of Chemical Structures and Reactions from the Literature. *Ph.D. Thesis*, University of Cambridge, Department of Chemistry, Pembroke College, 2012.