Probing the Generalisability of Single-Step Retrosynthesis Models

Sara Tanovic[1], Fernanda Duarte[1]

[1]*Chemistry Research Laboratory, 12 Mansfield Road, Oxford, OX1 3TA, UK*

Single-step retrosynthesis models are integral to computer-aided synthesis planning (CASP) research, yet assessing their performance and limits remains challenging. Experimental validation of predictions is impractical due to laboratory constraints and the large size of test sets. Moreover conventional metrics, such as top-*k* accuracy, have been criticized for their simplicity and oversight of key aspects of model performance such as prediction diversity and reaction feasibility.[1-3]

This contribution evaluates a variety of metrics to assess the generalisability of different machine learning algorithms across diverse reactions and product structures. Multiple databases are used to expand the region of chemical space tested and understand how similarity between the training and test sets influences performance. The effect of training set sizes on generalisability and performance are also discussed. Our preliminary results identify key features impacting accuracy and show that template-free models show little ability to extrapolate beyond reaction templates seen in the training set. Future directions are proposed for the development of new algorithms and evaluation standards.

Bibliography

[1] V. S. Gil, A. M. Bran, M. Franke, R. Schlama, J. S. Luterbacher and P. Schwaller, *Proceedings of NeurIPS 2023 AI for Science Workshop*, 2023.

[2] P. Torren-Peraire, A. K. Hassen, S. Genheden, J. Verhoeven, D. Clevert, M. Preuss and I. V. Tetko, *Digital Discovery*, 2024, **3**, 558-572.

[3] P. Schwaller, R. Petraglia, V. Zullo, V. Nair, R. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano and T. Laino, *Chem. Sci.*, 2020, **11**, 3316-3325.