Multi-instance learning as a response to the complexity of molecular entities

Pavel Polsihchuk[1], Dmitry Zankov[2], Timur Madzhidov[3], Alexandre Varnek[2,4]

[1] Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University Olomouc, Olomouc, Czech Republic.

[2]ICReDD, Hokkaido University, Sapporo, Japan

[3]Chemistry Solutions, Elsevier, Oxford, United Kingdom

[4]Laboratory of Chemoinformatics, University of Strasbourg, Strasbourg, France

Molecules are complex dynamic objects that can exist in different molecular forms (conformations, tautomers, stereoisomers, protonation states, etc.) and often it is not known which molecular form is responsible for observed physicochemical and biological properties of a given compound. This raises the problem of the selection of the correct molecular form for machine learning modeling of target properties. Most commonly the molecular forms selected for modeling are determined by specific standardization protocols by choosing most probable tautomeric or protonation states or the most stable conformer. However, this selection is prone to errors that may reduce model predictive performance.

Multi-instance learning (MIL) is an efficient approach for solving problems where objects under study cannot be uniquely represented by a single instance, but rather by a set of multiple alternative instances. Developed in 1997 to predict biological activities MIL was not widely adopted by the chemoinformatic community, but found a lot of applications in other domains, such as information retrieval, computer vision, signal processing, bankruptcy prediction, etc. Here, we revisited the MIL concept and its applications in chemo- and bioinformatics [1]. The advantages of MIL over commonly used single-instance learning will be demonstrated on several recent examples of prediction of biological activity [2] and enantioselectivity of chemical reactions [3].

With introduction of modern neural networks MIL provides a lot of flexibility to construct architectures the most suitable to solve particular tasks. Despite of higher computational requirements, MIL is very competitive to existing modeling approaches and provide unique features. One of such features is the ability to identify key instances (molecular forms) which are responsible for the modeling response. In particular, MIL models could identify biologically relevant conformations of molecules comparably well or better than molecular docking.

Bibliography :

[1] Zankov, D.; Madzhidov, T.; Varnek, A.; Polishchuk, P. WIREs Comput. Mol. Sci. 14 (2024) e1698.
[2] Zankov, D. V.; Matveieva, M.; Nikonenko, A. V.; Nugmanov, R. I.; Baskin, I. I.; Varnek, A.; Polishchuk, P.; Madzhidov, T. I. J. Chem. Inf. Model. 61 (2021) 4913-4923.
[3] Zankov, D.; Madzhidov, T.; Polishchuk, P.; Sidorov, P.; Varnek, A. J. Chem. Inf. Model. 63 (2023) 6629-6641.