universität innsbruck
Digital Science Center (DiSC)

universität innsbruck
Department of
Molecular Biology

The Department of Molecular Biology and the Digital Science Center would like to invite you to the following talk:

Prof. Markus List
Technical University of Munich (TUM)

Data leakage is a widespread issue in machine learning in the biomedical domain: How to spot it, avoid it, and fix it

Machine learning has become indispensable in modern biomedical research, and various user-friendly tools have been developed to make model training accessible to a broad user base. However, uninformed or uneducated use of machine learning often leads to reporting wrong or, at best, overoptimistic results. Basic strategies to avoid widely known biases, such as overfitting, are frequently not used. Studies in which no data splitting was performed (into training and test sets) still slip through the review process at a worrying rate. However, even machine learning experts can create models that are not fit for practical application due to issues such as data leakage. Data leakage occurs when information from the test set (inadvertently) influences the training process. Recognizing this problem often requires expert domain knowledge. In this seminar talk, I will show two examples where not-so-obvious machine learning problems have masked the true extent of the machine learning challenge and where overoptimistic reporting is pervasive in the literature. The first example (PMID: 38446741) is the task of predicting protein-protein interactions based solely on the amino acid sequences of the proteins. Here, I will show that methods achieve only random performance when the data splitting strategy does not allow similar sequences to be found in the training and test set. The second example is the drug response prediction for cell lines based on molecular profiles or drug information, where methods rarely do better than a naïve predictor of mean drug response. In both examples, an immense amount of literature has been published where data leakage issues were not recognized, and appropriate baseline methods have not been included. I hope to convince you in my talk that cross-domain expertise is essential for successful applications of machine learning in biomedicine, and, finally, I will present guidelines (PMID: 39122953) that we have developed that help recognize and avoid data leakage in future research projects.

About the speaker

Prof. Dr. Markus List is Professor of Data Science of Systems Biology at TUM School of Life Sciences, Technical University of Munich, Germany. He gives his talk upon invitation by Francesca Finotello.

Date, Time & Place:
Tuesday, 27 May 2025, 17:00 (CEST)
Campus Technik, Hörsaal D, Viktor-Franz-Hess-Haus, Technikerstr. 25a, 1. Stock