Quantifying the hardness of bioactivity prediction tasks for transfer learning

Hosein Fooladi[1,2,3], Steffen Hirte[2,3], and Johannes Kirchmair[1,2]

[1]*Christian Doppler Laboratory for Molecular Informatics in the Biosciences, Department for Pharmaceutical Sciences, University of Vienna, 1090 Vienna, Austria*
[2]*Department of Pharmaceutical Sciences, Division of Pharmaceutical Chemistry, University of Vienna, Josef-Holaubek-Platz 2, 1090 Vienna, Austria*
[3]*Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences (PhaNuSpo), University of Vienna, 1090 Vienna, Austria*

Today, machine learning methods are widely employed in drug discovery. However, the chronic lack of data continues to hamper their further development, validation, and application. Several modern strategies aim to mitigate the challenges associated with data scarcity by learning from data on related tasks.[1,2] These knowledge-sharing approaches encompass transfer learning, multi-task learning, and meta-learning. A key question remaining to be answered for these approaches is about the extent to which their performance can benefit from the relatedness of available source (training) tasks, in other words, how difficult ("hard") a test task is to a model, given the available source tasks.

We present a new method for quantifying and predicting the hardness of a bioactivity prediction task based on its relation to the available training tasks.[3] The approach involves the generation of protein and chemical representations and the calculation of distances between the bioactivity prediction task and the available training tasks. In the example of meta-learning on the FS-Mol data set,[4] we demonstrate that the proposed task hardness metric is inversely correlated with performance (Pearson's correlation coefficient r=-0.72). In practice, the new metric can be used to (i) predict the advantage of knowledge-sharing vs. single-task approaches for specific prediction tasks and (ii) select relevant source tasks for optimum performance gain of knowledge-sharing methods.

Bibliography:

(1) Finn, C.; Abbeel, P.; Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *Int. Conf. Mach. Learn.* **2017**, 1126–1135. https://doi.org/10.48550/ARXIV.1703.03400.
(2) Cai, C.; Wang, S.; Xu, Y.; Zhang, W.; Tang, K.; Ouyang, Q.; Lai, L.; Pei, J. Transfer Learning for Drug Discovery. *J. Med. Chem.* **2020**, *63* (16), 8683–8694. https://doi.org/10.1021/acs.jmedchem.9b02147.
(3) Fooladi, H.; Hirte, S.; Kirchmair, J. Quantifying the Hardness of Bioactivity Prediction Tasks for Transfer Learning. **2024**. https://doi.org/10.26434/chemrxiv-2024-871mt.
(4) Stanley, M.; Bronskill, J.; Maziarz, K.; Misztela, H.; Lanini, J.; Segler, M.; Schneider, N.; Brockschmidt, M. FS-Mol: A Few-Shot Learning Dataset of Molecules. *Thirty-Fifth Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track* **2021**.