MMAtt-DTA: A MULTIMODAL ATTENTION-BASED APPROACH TO PREDICT DRUG-TARGET AFFINITIES ACROSS SEVEN TARGET SUPERFAMILIES

Aron Schulman, Juho Rousu, Tero Aittokallio, Ziaurrehman Tanoli*

*ziaurrehman.tanoli@helsinki.fi

FIMM
Institute for Molecular Medicine Finland
Nordic EMBL Partnership for Molecular Medicine

Drug-target affinity (DTA) prediction is key for drug discovery and repurposing. This study introduces a multimodal, attention-based method to predict DTAs for human proteins across seven superfamilies (Figure 1). We explored nine descriptor sets to identify optimal representations for drug-target pairs (Figure 2). Using independent testing, our method showed promising performance in three prediction scenarios and outperformed several state-of-the-art solutions (Figure 3). We applied our models to predict the complete interaction matrix between 3492 FDA-approved drugs and 1251 human proteins (Figure 4). The data and models are freely accessible via the provided QR-code.

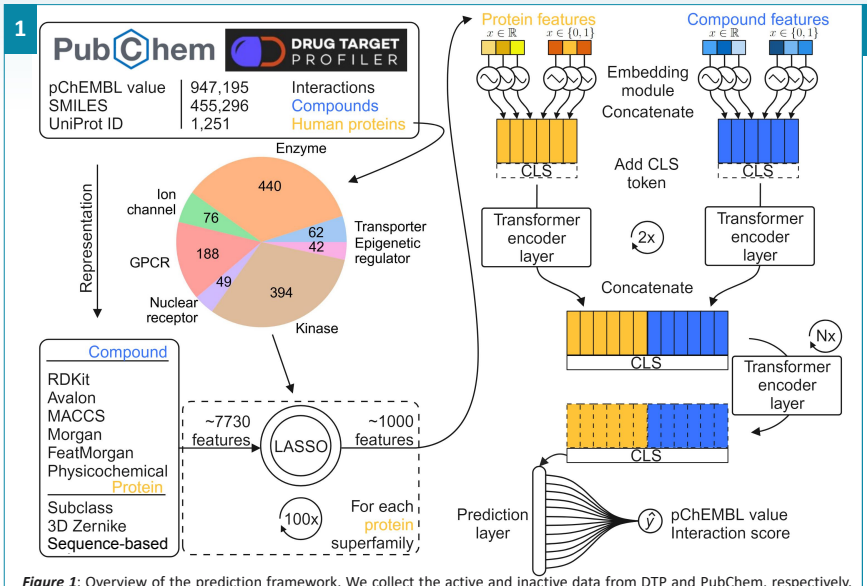

Figure 1: Overview of the prediction framework. We collect the active and inactive data from DTP and PubChem, respectively. We generate a high-dimensional feature vector and select the best descriptors with LASSO separately for each protein superfamily. We separate each descriptor set into protein and compound features, and further divide them into continuous and binary variables. The grouped features are passed into the predictive model to yield pairwise binding affinity point predictions.

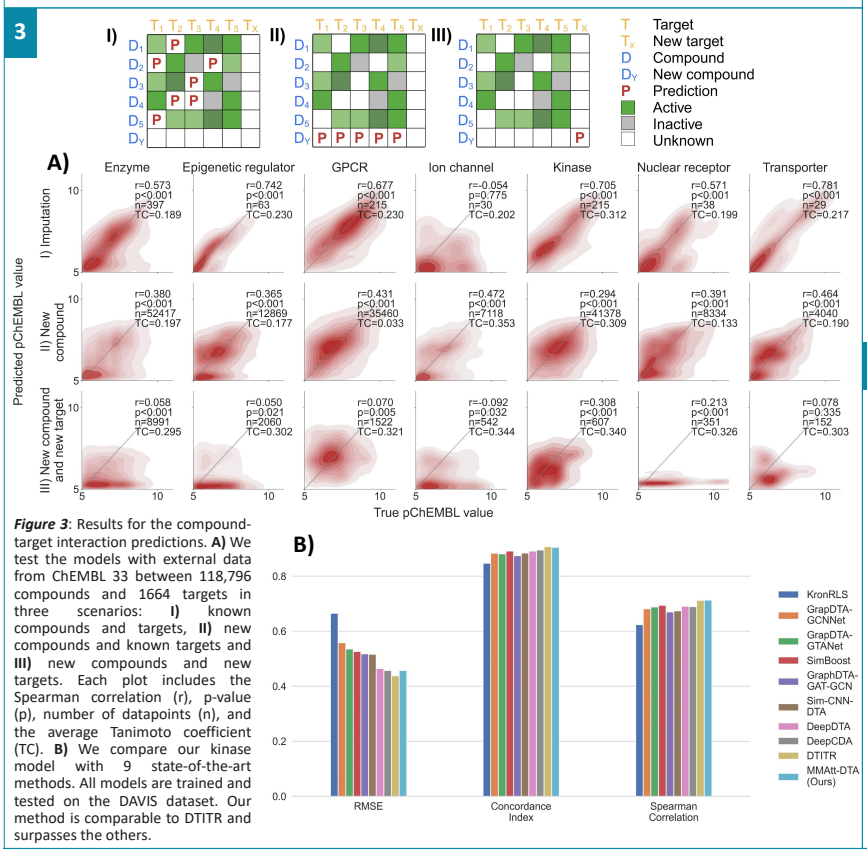

Figure 3: Results for the compound-target interaction predictions. **A)** We test the models with external data from ChEMBL 33 between 118,796 compounds and 1664 targets in three scenarios: **I)** known compounds and targets, **II)** new compounds and known targets and **III)** new compounds and new targets. Each plot includes the Spearman correlation (r), p-value (p), number of datapoints (n), and the average Tanimoto coefficient (TC). **B)** We compare our kinase model with 9 state-of-the-art methods. All models are trained and tested on the DAVIS dataset. Our method is comparable to DTITR and surpasses the others.

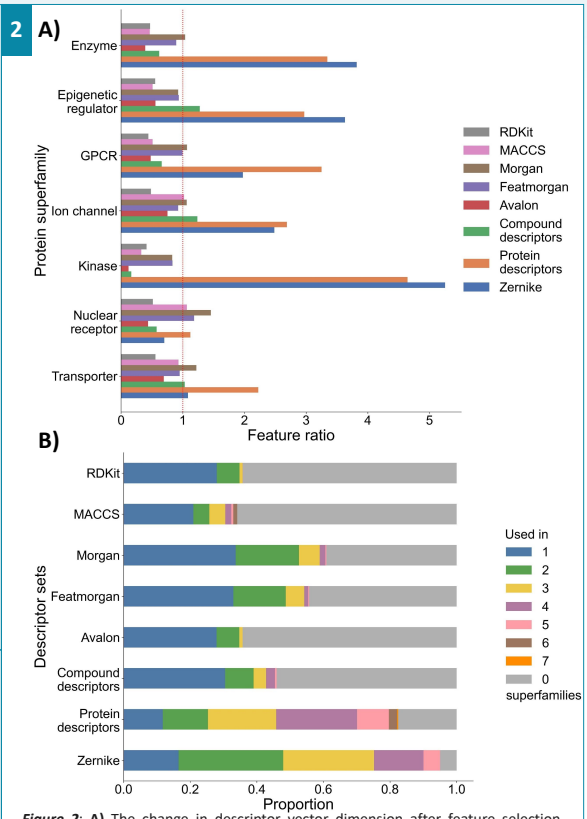

Figure 2: **A)** The change in descriptor vector dimension after feature selection. Originally, each compound-protein pair was represented by a 7719-dimensional vector. We reduced the dimension to 1000 with LASSO. The selected descriptors are different across superfamilies. Bars that cross the dotted reference line indicate an increased contribution of a descriptor set to the overall composition of a feature-selected vector than to the original vector. **B)** Proportions of how often individual descriptors belonging to a set are selected by LASSO into a model for each protein superfamily. The descriptors are grouped by their generality, i.e., how often they are selected for different superfamilies. Protein subclass labels are always included in the descriptor vectors and are thus omitted from the figure.

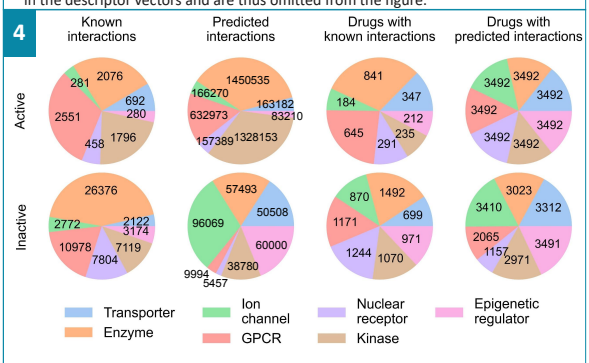

Figure 4: The proportions of known and predicted interactions by protein superfamily for the completed interaction matrix between 3492 FDA-approved drugs and 1251 human proteins. An interaction with pChEMBL value of 5.0 or smaller is considered inactive. For the predicted values, we set the inactivity threshold at 5.1 due to the regression model predicting many values close to, but not exactly, 5.0.

FIMM - Institute for Molecular Medicine Finland
P.O. Box 20, FI-00014 University of Helsinki, Finland
Biomedicum Helsinki 2U, Tukholmankatu 8, 00290 Helsinki, Finland

www.fimm.fi

UNIVERSITY OF HELSINKI
Aalto University
HiLIFE HELSINKI INSTITUTE OF LIFE SCIENCE
CSC
ACADEMY OF FINLAND
RE MED 4ALL