Fraunhofer SCAI

MultiGML – multimodal graph machine learning for drug target prioritization

Finding the right drug target

Identifying a good target is key for the effectiveness and safety of a drug candidate in pharmaceutical research. Traditionally, the identification of targets is based on human knowledge and understanding of fundamental disease mechanisms. However, given the growing flood of scientific literature, this manual approach is prone to overlooking relevant data and information, which may lead to sub-optimal choices.

Adverse drug events as a risk for clinical trials

Adverse drug events (ADEs) are defined as an injury resulting from the use of a drug, including harm caused by the drug (adverse drug reactions and overdoses) and harm from the use of the drug (including dose reductions and discontinuations of drug therapy). The appearance of an ADE can at least partially be associated with the choice of the primary target protein or properties of the chemical structure of a drug. Experimental approaches to address potential ADEs (e.g., liver toxicity) based on animal and tissue models are well established. Yet, results obtained in such model systems may not always reflect the situation in humans, and there are ethical concerns regarding the use of animal models. Furthermore, reliable model systems do not exist for all ADEs and indication areas.

Limitations of human genetics

One possible way to address the abovementioned concerns is to check whether genetic variants in a candidate drug target have been associated with an unfavorable phenotype. While this approach has been reported to be highly effective in cases where such an association could be identified, it is crucial to see that there is a high risk of missing relevant associations due to a lack of statistical power.

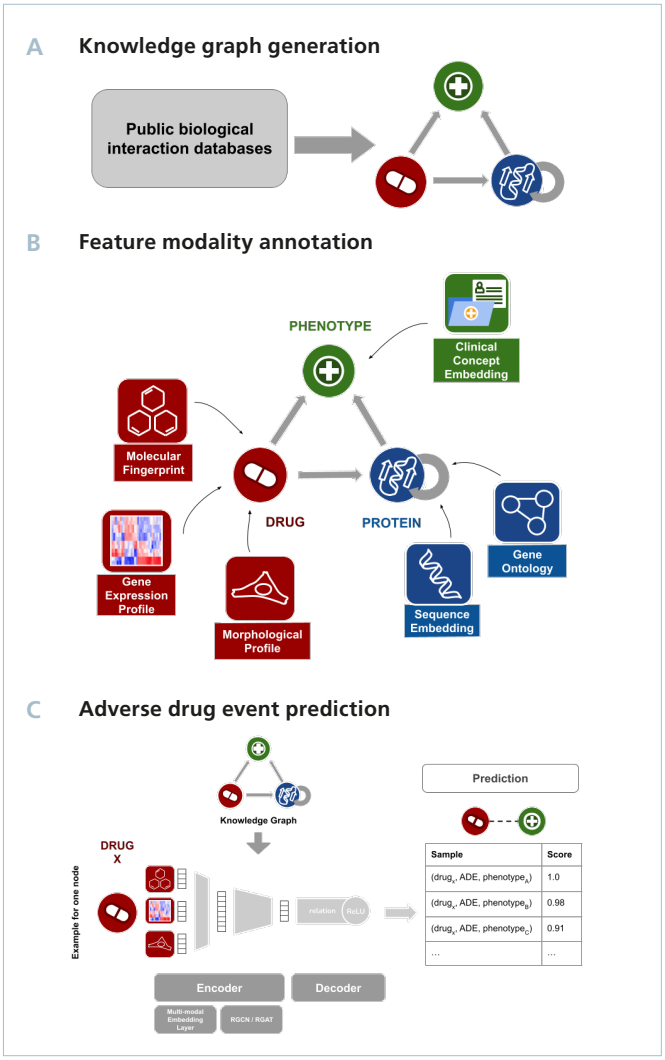

Figure 1. Workflow. We compile a comprehensive biomedical knowledge graph from public databases (A), and enrich it with quantitative data and information (B). A multimodal Graph Neural Network architecture is trained to predict new links between compounds and phenotypes or targets and phenotypes (C).

Our solution: Learning from comprehensive knowledge and experimental data in an unbiased manner

Our solution combines relevant knowledge and experimental data (e.g., gene expression data, microscopy images, protein sequences) in a comprehensive, structured, and unbiased manner. Technically, this is realized via a semantically harmonized knowledge graph, which we compiled from 14 curated biological databases, resulting in around 400.000 relations between proteins, drugs, and phenotypes, including ADEs (Figure 1). Based on this wealth of information, we trained a multimodal Graph Neural Network architecture capable of accurately predicting new associations between compounds and phenotypes or between protein targets and phenotypes. A distinction to alternative solutions is the possibility of MultiGML to deliver explanations for model predictions (Figure 2).

Our offering

We offer our customers MultiGML in two variants:

1. **MultiGML_Model:** Our fully trained MultiGML model in both RGCN and RGAT variants (PyTorch version 1.9.1) is ready for use with instructions for installing the environment and applying the pre-trained model.
2. **MultiGML_Code:** The code to enable running MultiGML with a custom knowledge graph, including our scripts to generate node features of the knowledge graph, employed in our previous publication, as well as scripts to explain predictions post-hoc using the Integrated Gradients method. The node feature generation scripts include scripts for the following node-type features:
   - Drugs: molecular fingerprint, gene expression signature, morphological fingerprint
   - Proteins: sequence embedding, gene ontology fingerprint
   - Phenotypes: medical concept embedding

Instructions for the installation of environments and the usage of the command-line interface are included. This will give you full flexibility, including the possibility to enrich your knowledge graph according to your wishes, train your customized MultiGML instance, and subsequently explain your predictions.

Reference

Sophia Krix, Lauren Nicole DeLong, Sumit Madan, Daniel Domingo-Fernández, Ashar Ahmad, Sheraz Gul, Andrea Zaliani, Holger Fröhlich, MultiGML: Multimodal graph machine learning for prediction of adverse drug events, Heliyon, Volume 9, Issue 9, 2023, e19441, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2023.e19441

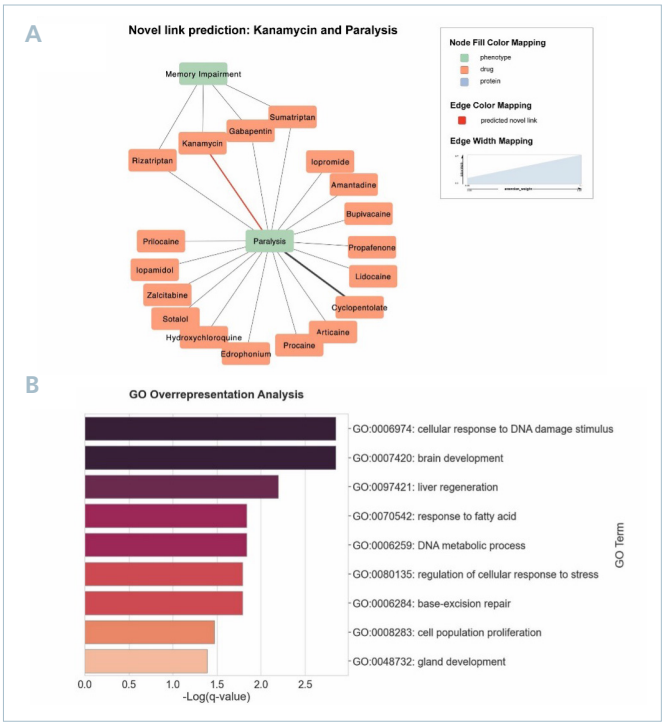

Figure 2: Example of an explanation of a model prediction provided by MultiGML. A) MultiGML predicted an association between the compound karamycin and the phenotype paralysis. The graph depicts the first-order neighborhood of the compound and the phenotype of interest and highlights the links that MultiGML considered most relevant to make this prediction. The edges' thickness indicates the attention MultiGML gave to individual relations to make the prediction.

Contact

Fraunhofer Institute for Algorithms and Scientific Computing SCAI
Schloss Birlinghoven 1
53757 Sankt Augustin
Germany

Prof. Dr. Holger Fröhlich
Phone +49 2241 14-4219
holger.froehlich@scai.fraunhofer.de

Sophia Krix
Phone +49 2241 14-4234
sophia.krix@scai.fraunhofer.de
www.scai.fraunhofer.de/bio

Distributed by

scapos AG
Schloss Birlinghoven 1
53757 Sankt Augustin
Germany

Phone +49 2241 14-4400
info@scapos.com
www.scapos.com