Enhancing Compound Similarity Prediction: A Novel Approach

E. Borzone, L. Di Persia, M. Gerard
Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH/UNL-CONICET, (3000) SF, ARG

Background:
Predicting similarity between compounds is a challenge when the structure of the compounds involved is unknown. To address this issue, we have proposed a model based on graph neural networks (GNNs) to generate embeddings that capture compound relationships. Despite the good results obtained up to now, our analyses have shown that embeddings are not clustered according to the similarity of the compounds they represent.

Results:
In order to improve the embeddings, we introduced a modification in the cost function We expect that by encouraging the formation of more well-defined and cohesive clusters within the latent space, we could establish a robust negative correlation between spatial distance and compound similarity.

Preliminary results provide compelling evidence of the effectiveness of our proposed modification. The embeddings produced by our modified cost function now exhibit significantly improved clustering, resulting in spatial similarity that closely aligns with compound similarity, expecting an improvement for compounds with unknown structure due to the better correlation between the measured distance in embedding space and the similarity index used (Tanimoto index). This crucial enhancement has translated into a boost in the accuracy of similarity predictions, with an increase of 0.54%. The tighter clustering and enhanced spatial organization suggest that our approach effectively addresses the limitations encountered in previous models.

Furthermore, a notable challenge tackled in this study pertains to predicting similarity among compounds lacking known structural information. To assess the practical implications of our model, we performed an analysis by plotting the distances in the embedding space against the Tanimoto index. In this analysis, we aimed to observe the presence of a negative correlation, as enhanced embeddings should ideally exhibit this relationship. The results clearly demonstrate substantial improvements compared to previous methodologies. This breakthrough represents a major step forward in our ability to predict the similarity, even when dealing with compounds lacking known structures. These findings underscore the potential utility of our model in diverse applications across various domains, including drug discovery and metabolic pathway analysis.

Conclusion:
Our study marks a significant leap forward in the prediction of compound similarity within metabolic pathways. Leveraging Graph Neural Networks (GNNs) and introducing an innovative embedding modification, we have achieved substantial progress. Our modified cost function has substantially boosted the accuracy of compound similarity predictions, resulting in more coherent and interpretable compound representations. This breakthrough not only promises a deeper understanding of metabolic pathways but also holds great potential for practical applications.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
E. Borzone, L. Di Persia & M. Gerard; "Enhancing Compound Similarity Prediction: A Novel Approach"
A2B2C 13th Meeting, 2023.

Looking ahead, our future work will focus on expanding the scope of our research. Incorporating additional metabolic pathways into our dataset will enrich our model's ability to handle diverse biochemical contexts. Furthermore, we plan to explore the incorporation of edge features to capture finer nuances in compound interactions. These steps will not only enhance the comprehensiveness of our model but also open doors to a broader range of applications, reinforcing our commitment to advancing the field of compound similarity prediction within metabolic pathways.