[BIOINFO 2023]

The Dr.Emb Appyter: A Web Platform for Drug Discovery using Embedding Vectors

Songhyeon Kim[1], Wootaek Lim[1], and Minji Jeon[1,2,*]

[1] *Department of Medicine, Korea University College of Medicine*
[2] *Biomedical Research Center, Korea University Anam Hospital*
**Corresponding author: mjjeon@korea.ac.kr*

An embedding vector is a form of dense representation of data in a low-dimensional latent space. It captures the semantic meanings or relationships between data points, making that similar data are close to each other in the embedding space. In drug discovery, a lot of studies proposed embedding methods to capture chemical compound information including structures, chemical properties, and drug-induced gene expression. These embedding vectors can be used as input for downstream tasks such as predicting compound properties and drug-target interaction. Moreover, they can be also utilized for novel drug candidate discovery by using the characteristics of embedding: similar data points are located closely. For example, drug property-based embedding vectors can be used to identify novel compounds that are closely located to well-known drugs and they may have similar properties. In this way, the embedding method has the advantage of discovering novel compounds by computing distances between hundreds of millions of compounds without actually conducting experiments. However, huge computational resources and programming skills are essential to use them and these are major obstacles for scientists who are not in this field. Therefore, it is necessary to develop a user-friendly web application that enables chemical compound discovery using embedding methods without the need for significant computing resources or programming skills.

Here, we developed a web platform for drug discovery using embedding vectors (Dr.Emb) Appyter that leverages various embedding methods to discover novel drug candidates. Built on Appyter, which transforms Jupyter notebooks into web applications, the Dr.Emb Appyter offers a user-friendly interface. It generates embedding vectors for a large number of compounds in a library, as well as for a query compound, using SMILES, molecular graph, and drug-induced gene expression-based embedding methods. These vectors facilitate the identification of compounds that are close to the query compound in the embedding space through a FAISS-based similarity search. Additionally, the Dr.Emb Appyter provides detailed information on the top compounds, offers various visualizations including scatter plots, heat maps, and UpSet plots, and returns drug set enrichment analysis results.