# Fingerprint-Informed Doc2Vec for Activity Prediction and Interpretation

Yuuto Matsumoto
YOKOHAMA National University

Nowadays, activity prediction methods assisted by natural language processing (NLP) have performed better than descriptor-based methods. Chemical language models (CLMs) and multimodal learning like CLIP are promising methods, but they have several issues. CLMs, which are transformer models that process molecular structures as sequences, such as SMILES and SELFIES, suffer from limited interpretability due to complex and opaque architectures. Multimodal models have two problems: limited chemical interpretability and dependence on textual input, which limits their applicability to unseen compounds.

In this study, we propose an NLP-based compound representation method using embeddings generated by Doc2Vec, which is trained on molecular descriptions with fingerprints as document tags.

We show that these embeddings perform better than both the original fingerprints and embeddings generated solely from textual descriptions when predicting the activity of unseen compounds.

Notably, the proposed method enables prediction of unseen compounds without the need for additional textual input. The input embeddings for such compounds can be directly reconstructed from their fingerprints by summing the partial embeddings corresponding to each fingerprint bit.

Furthermore, we evaluated the interpretability of the learned embeddings using SHAP. The important features identified by SHAP were consistent with previously reported structure–activity relationships.