No.	K25-2188	
Project Title	Investigation of drug molecular representation using advanced AI models and its application in drug-target identification	
Principal Investigator	Chen Li (Zhejiang University School of Medicine・Associate professor)	
Project Member(s)	IMSUT Host Researcher	Seiya Imoto (The Institute OF Medical Science, The University Of Tokyo・Professor)
	Project Member (s)	Yaozhong Zhang (Division of Health Medical Intelligence・Associate Professor)
	Project Member (s)	Rui Yamaguchi (Division of Cancer Systems Biology, Aichi Cancer Center・Division Chief)
	Project Member (s)	Chen Zhiang (Zhejiang University School of Medicine・Student)

(Form 2)

Principal Investigator
Seiya Imoto

IMSUT International Joint Usage/Research Center Project <International>

Joint Research Report (Annual/Project Completion)

Annual Report

Report

Title: Investigation of drug molecular representation using advanced AI models and its application in drug-target identification

AI methods have become increasingly important in predicting drug–target interactions (DTIs). Over the past year, we evaluated several molecular representation approaches to understand how different input features influence model performance. Building on these findings, we further examined the redundancy present in large pretrained representations and developed a supervised vector quantization (SVQ) approach to identify features that are most relevant for DTI prediction.

(1) Systematic analysis of pretrained representations for DTI

Using benchmark DTI datasets (e.g., BIOSNAP, BindingDB, and DAVIS), we combined pretrained molecular embeddings (MMELON) and protein embeddings (ESM-C) and investigated which dimensions in these high-dimensional representations are truly informative for DTI prediction. We applied random forest–based feature importance analysis together with the Boruta algorithm to identify discriminative features in the concatenated drug–protein embedding space. The results showed that only a small subset of features contributed to DTI prediction, while most dimensions were redundant or noisy.

(2) Development of a supervised vector quantization framework (SVQ-MLP)

Motivated by the above findings, we designed a SVQ module that can be inserted between pretrained representations and a simple multilayer perceptron (MLP) classifier. In this framework, drug and protein embeddings are split into low-dimensional segments and each segment is mapped to a discrete codeword from a learnable codebook. The quantized sub-vectors are then concatenated and passed to an MLP classifier trained with a hybrid loss combining quantization loss and classification loss. The overall architecture of this SVQ-MLP framework for DTI prediction is illustrated in Figure 1.

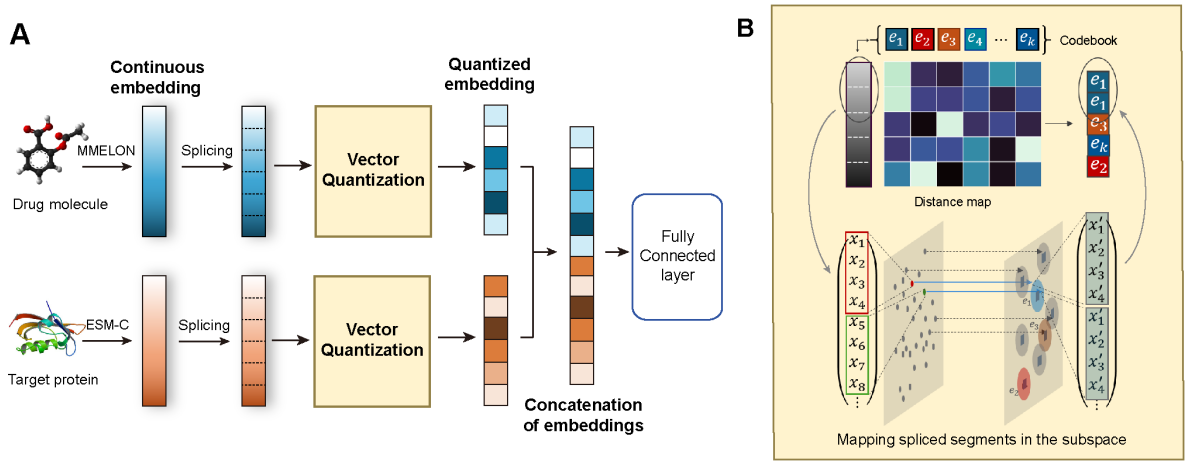

Fig. 1. Schematic overview of SVQ-MLP framework for DTI prediction. A. The framework integrates a SVQ module into the baseline MLP layer, enabling structured compression and selection of informative features for improved performance and interpretability. B. In the subspace of spliced embedding, each continuous segment is projected onto a nearest codeword from a discrete codebook, effectively quantizing the representation.

3) Performance evaluation and comparison with state-of-the-art methods

We systematically evaluated three settings: (i) using all pretrained features, (ii) using Boruta-selected features, and (iii) using features processed by the SVQ module. Linear probing experiments showed that removing irrelevant features significantly improved performance, confirming that selective use of pretrained features is beneficial. Incorporating the SVQ module allowed the MLP classifier to recover and even surpass the performance achieved with Boruta-based feature selection, while maintaining an end-to-end training pipeline. Compared with recent deep learning–based DTI methods, the SVQ-MLP framework achieved competitive or superior AUPR on the benchmark datasets, and also showed good generalization in zero-shot scenarios.

(4) Interpretability analysis of quantized representations

We further analyzed the learned codebook and codeword usage patterns. By examining codeword frequency profiles across drugs interacting with proteins containing specific Pfam domains, we observed clear domain-specific activation patterns, suggesting that the codebook captures biologically meaningful interaction signals at the protein-domain level. Moreover, a bag-of-words–like representation of codeword usage preserved much of the structural information of the original drug embeddings, indicating that co-occurrence patterns of codewords, rather than the exact semantic content of individual dimensions, encode most of the useful information. These findings provide a more interpretable view of how pretrained representations can be compressed without losing DTI-relevant structure.