No.	K25-2171	
Project Title	Deep learning-based approaches for non-coding variant effect prediction	
Principal Investigator	Jiangning Song (Monash University・Professor)	
Project Member(s)	IMSUT Host Researcher	Seiya Imoto (The Institute OF Medical Science, The University Of Tokyo・Professor)
	Project Member (s)	Professor Seiya Imoto (Human Genome Center・Professor, Director of Human Genome Center)
	Project Member (s)	Associate Professor Yao-zhong Zhang (Human Genome Center・Associate Professor)
	Project Member (s)	Associate Professor Kotoe Katayama (Human Genome Center・Associate Professor)
	Project Member (s)	Professor Rui Yamaguchi (Aichi Cancer Center / Division of Cancer Systems Biology・Professor, Chief, Division of Cancer Systems Biology, Aichi Cancer Center)
	Project Member (s)	Xiaoyu Wang (Monash University / Monash Biomedicine Discovery Institute・PhD student)
	Project Member (s)	Tong Pan (Monash University / Monash Biomedicine Discovery Institute・PhD student)
	Project Member (s)	Yue Bi (Monash University / Monash Biomedicine Discovery Institute・PhD student)
	Project Member (s)	Yumeng Zhang (Monash University / Monash Biomedicine Discovery Institute・PhD student)
	Project Member (s)	Zehua (Cameron) Sun (Monash University / Monash Biomedicine Discovery Institute・PhD student)

(Form 2)

Principal Investigator
Jiangning Song

IMSUT International Joint Usage/Research Center Project <International>

Joint Research Report (Annual/Project Completion)

Project Completion Report

Report

1. Research Background and Objectives In recent years, the biomedical field has seen a rapid accumulation of large-scale data, such as whole slide images (WSIs) and protein sequences. However, establishing methodologies that can analyze these data with high accuracy and explainability remains a significant challenge. This research aimed to develop novel deep learning frameworks to improve predictive performance and interpretability in two distinct domains: phenotype identification in pathology images and protein function prediction.

2. Research Methods and Achievements First, in the field of pathology image analysis, we proposed a "weakly supervised neuron selection" approach to extract disentangled representations from CLIP-derived pathology foundation models. By utilizing Sparse Autoencoders (SAEs), we decomposed image features into individual neurons that are interpretable by humans. We then identified key neurons using slide-level labels within a Multiple Instance Learning (MIL) framework. Experiments on datasets such as Camelyon16 demonstrated that our method could effectively identify tumor patches without detailed manual annotations, providing superior explainability compared to conventional attention-based models.

Second, for protein function prediction, we developed "PFresGO," a deep learning model that integrates the hierarchical structure of Gene Ontology (GO) with residue-level protein information. The model employs a multi-head cross-attention mechanism, using GO terms as queries to pinpoint functional sites within protein sequences. Our evaluation showed that PFresGO outperformed state-of-the-art methods (in terms of

Fmax and AUPRC) and exhibited high generalization capabilities, even for uncharacterized proteins with low sequence homology.

3. Significance and Future Outlook The significance of this research lies in its ability to "untangle" the internal representations of complex AI models, providing biological evidence directly applicable to clinical diagnosis and drug discovery. Moving forward, we plan to integrate these methodologies and expand their application to multi-omics data integration.