

AI Driven Structure-Aware RAG for Cytochrome P450: Hybrid Embeddings and Cross Enzyme Attention

Taqdees Khan*, Kim Hee Cheol*

* *Department of Digital Anti-Aging Healthcare, Inje University, Republic of Korea*

nevakhan57@gmail.com, heeki@inje.ac.kr

Abstract—The enzymes known as cytochrome P450 (CYP) process 75% of clinical drugs, and therefore, the precision of drug enzyme interactions should be key in developing pharmaceuticals. Nonetheless, state-of-the-art AI technologies are unable to integrate 3D protein structure with textual information, which limits their application in drug metabolism forecasting. Our structure-aware retrieval augmented generation (RAG) model integrates AlphaFold structures, experimental PDB data, and scientific literature data via two text geometric projections. Our weight optimized hybrid retrieval formula with $\alpha=0.7$ attains $NDCG@10=0.847$, an improvement of 35.9 per cent over text only methods. It uses biology conscious chunking based on functional protein domains (31.3% of the performance), geometric embeddings (26.5%), and cross enzyme attention mechanisms to process CYP3A4, CYP2D6, and CYP2C19. It is much better than current techniques, such as D-CyPre (F1: 0.669) and DEEPCYPs, and performs quantitative structural measurements that are crucial for drug design. Provenance tracking is implemented using blockchain, which ensures scientific reproducibility and addresses the gap between general purpose language models and specialized molecular modelling tools.

Keywords—Drug Discovery, Cytochrome P450, Retrieval-Augmented Generation, Protein Structure, Machine Learning, Drug Metabolism.



Taqdees Khan received the Bachelor of Science degree in Computer Science from the University of Lahore, Pakistan, in 2022, and completed the Korean Language and Culture Program at Inje University, South Korea, in 2024. She is currently pursuing the M.S. degree at the Institute of Digital Anti-Aging Healthcare and working as a Research Assistant at the Smart Computing Laboratory (SCL), Inje University, South Korea. Her research interests include computer vision, artificial intelligence, machine learning, deep learning, visual recognition, generative modeling, bioinformatics, and AI-driven drug discovery and design.



HEE-CHEOL KIM was born in Busan, Korea. He has a BSc from the Department of Mathematics, an MSc from the Department of Computer Science at SoGang University in Korea, and a Ph.D. in Numerical Analysis and Computing Science at Stockholm University in Sweden. He is a professor and Head of the Department of the Institute of Digital Anti-aging Healthcare, Inje University, S: Korea. His research interests include Artificial intelligence, Text mining, and Bioinformatics.