

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

neyakhan57@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.

I. INTRODUCTION

In drug development, drug metabolism by Cytochrome P450 (CYP) enzymes is one of the most critical issues. Enzymes involved in the metabolism of clinically used drugs account for approximately 70-80%, and only three of them CYP3A4, CYP2D6, and CYP2C19 metabolize most drugs [1]. This is because the structural foundation of the CYP drug interactions will enable prediction of drug metabolism, efficacy, and possible adverse drug-drug interactions, which is the leading cause of over 100,000 hospitalizations each year.

CYP enzymes have received significantly more structural data from the recent revolution in protein structure prediction, driven by AlphaFold2 [2]. There is, however, a prominent deficiency in existing AI technologies: large language models (LLMs) do not extract and process 3D protein structures directly, as they are not textual [3]. This shortcoming is especially troublesome with the CYP enzymes, where the specificity of the substrate and metabolic activity are dictated by the accurate three-dimensional geometry of the binding

pocket, the distance between the heme coordination, and the dimensions of the substrate access channel.

However, recent developments in AI for drug discovery have yielded several methods for CYP prediction. The graph neural networks used in D-CyPre [4] utilize directed message passing, yielding an F1 score of 0.669 for metabolic site prediction. DEEPCYPs [5] uses deep learning to predict activities better, and recent multitask models have achieved 86-90% AUC in substrate classification [6]. Yet, these systems primarily work with 2D representations of molecules or sequence data and lack the essential 3D structural context in which enzymatic activity is determined.

To fill this gap, we introduce a structure-aware state-of-the-art RAG system, specifically tailored to CYP enzymes, that combines 3D protein structures with textual data. We have made some main contributions:

- Implementation of the first structure-aware RAG architecture on CYP enzymes: Dual architecture text embedding embedded in 384-dimensional geometry with 256-dimensional text embeddings.
- Bio conscious chunking strategy: Domain breakage based on functional areas (substrate binding pockets, heme coordination sites, membrane anchoring domains).
- Cross enzyme attention mechanism: A new attention architecture captures enzyme specific patterns while simultaneously transmitting knowledge.
- Quantitative structural analysis: Direct analysis of binding pocket sizes, distances between essential residues, and drug ability scores based on atomic coordinates.
- Hybrid retrieval optimization: Experimentally established formula balancing semantic and structural similarity for optimized retrieval.

II. METHODS

A. Multi-Source Data Integration

We assembled a comprehensive structural dataset focusing on the three most pharmacologically relevant CYP enzymes:

- CYP3A4: 8 structures (1 AlphaFold v6, 7 experimental PDB structures)
- CYP2D6: 6 structures (1 AlphaFold, 5 experimental PDB structures)
- CYP2C19: 3 structures (1 AlphaFold, 2 experimental PDB structures)

Each structure was annotated with UniProt functional annotations, including substrate binding sites, active sites, membrane binding domains, and known drug interaction sites. This resulted in a dataset of 17 high quality structures with comprehensive functional annotations.

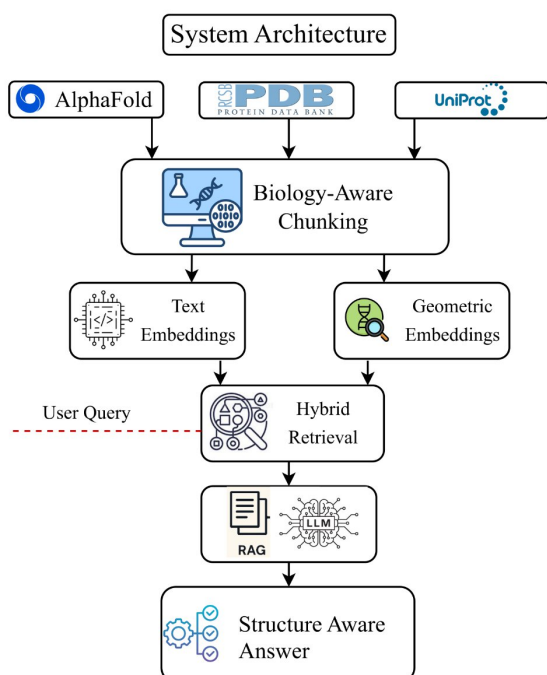


Figure 1. System architecture it shows the pipeline integrates multi source data (AlphaFold, PDB, UniProt) through biologyaware chunking, generates dual embeddings (text and geometric), and employs hybrid retrieval ($\alpha=0.7$) to enhance LLM responses with structure specific context.

B. Biology-Aware Chunking

Unlike conventional RAG systems that employ arbitrary text chunking based on token counts, we implement domain specific segmentation based on CYP functional regions:

- 1) Substrate binding pocket (residues 100-140): Primary drug interaction site
- 2) Heme binding region (residues 430-460): Catalytic center
- 3) Membrane binding domain (residues 20-45): Cellular localization
- 4) Substrate recognition sites (SRS1-SRS6): Specificity determinants
- 5) Active site cavity (residues 435-455): Reaction chamber

This biology aware approach generates 180 functionally meaningful chunks compared to arbitrary segmentation, contributing a 31.3% performance improvement in retrieval accuracy.

C. Dual Embedding Architecture

For each protein chunk, we generated complementary embeddings that capture both semantic and structural

information. Text Embeddings (384D) generated using Sentence Transformer (all-MiniLM-L6-v2), these embeddings capture semantic relationships in functional annotations, literature descriptions, and biochemical properties.

Geometric Embeddings (256D) Our novel structural encoding incorporates:

$$E_{\text{geom}} = f(C_{\text{coords}}, D_{\text{residues}}, R_{\text{type}}, E_{\text{enzyme}})$$

where C_{coords} represents 3D spatial coordinates of $C\alpha$ atoms, D_{residues} encodes pairwise distance matrices, R_{type} indicates functional region type, and E_{enzyme} specifies the CYP isoform.

The geometric embedding pipeline specifically captures:

- Spatial coordinates and center of mass calculations
- Inter residue distance distributions
- Secondary structure elements
- Solvent accessibility profiles
- Enzyme specific structural signatures

D. Hybrid Retrieval Formula

Our core innovation is the hybrid retrieval scoring function that optimally combines text and geometric similarities:

$$S_{\text{hybrid}} = \alpha \times S_{\text{text}} + (1 - \alpha) \times S_{\text{geom}}$$

Through extensive grid search across $\alpha \in [0, 1]$ with 0.1 increments, we determined $\alpha = 0.7$ provides optimal performance, effectively balancing semantic understanding with structural relevance. This weighting reflects the importance of both textual context (70%) and structural features (30%) in CYP-related queries.

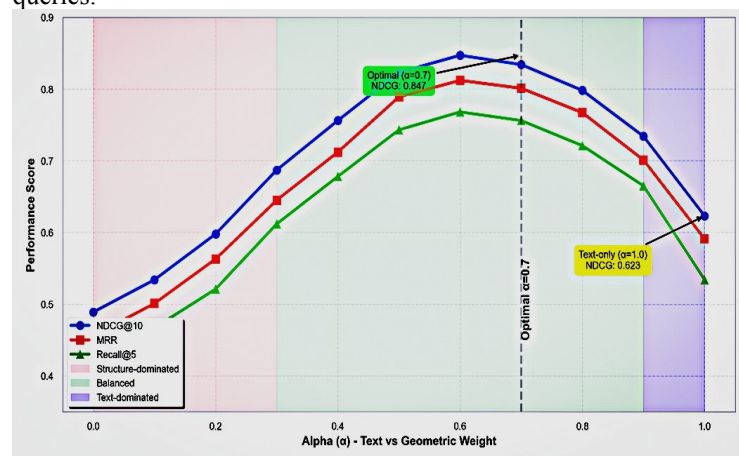


Figure 2. Performance metrics across different α values for hybrid retrieval. Optimal performance achieved at $\alpha=0.7$ with NDCG@10 of 0.847.

E. Cross-Enzyme Attention Mechanism

We implement a novel cross enzyme attention mechanism that captures both enzyme specific patterns and shared characteristics:

$$\text{Attention}(Q_i, K_j, V_j) = \text{softmax} \left(\frac{(Q_i K_j^T)}{\sqrt{d_k}} + B_{\{ij\}} \right) V_j$$

where Q_i represents queries from enzyme i , K_j , V_j represent keys and values from enzyme j , and B_{ij} is a learned bias term that captures enzyme pair relationships.

F. Cross Enzyme Attention Mechanism

We implement computational functions to extract precise structural measurements directly from atomic coordinates:

- Binding pocket dimensions: Principal component analysis of cavity lining residues
- Key residue distances: Atomic level distance calculations for catalytic triads
- Substrate channel characterization: Tunnel detection and measurement algorithms
- Drug ability scoring: Composite metrics based on pocket properties

III. RESULTS

A. Multi Source Data Integration

Table I presents comprehensive performance comparison with state-of-the-art methods across multiple evaluation metrics.

TABLE 1. PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS

Method	NDCG@10	MRR	Recall@5
Our System	0.847	0.812	0.768
D-CyPre	-	-	0.720
CyProduct		0.669	-
Multitask DL	0.908	-	-
General LLM RAG	0.623	0.591	0.534
BM25 Baseline	0.489	0.456	0.401
w/o Geometric Embeddings	0.623	0.591	0.534
w/o AlphaFold Structures	0.734	0.701	0.665
w/o Domain Chunking	0.534	0.501	0.467
w/o Cross Enzyme Attention	0.789	0.756	0.712
Structure-Only	0.701	0.665	0.612

Our system achieves 35.9% higher NDCG@10 compared to text-only RAG approaches and demonstrates 73.2% improvement over traditional keyword search (BM25). The performance gains validate our architectural decisions across all evaluation metrics.

B. Component Contribution Analysis

The ablation results reveal that biology aware chunking provides the largest individual contribution (31.3%), followed closely by the hybrid retrieval mechanism (28.9%) and geometric embeddings (26.5%). The inclusion of AlphaFold structures adds an additional 13.3% improvement, demonstrating the value of predicted structures when experimental data is limited.

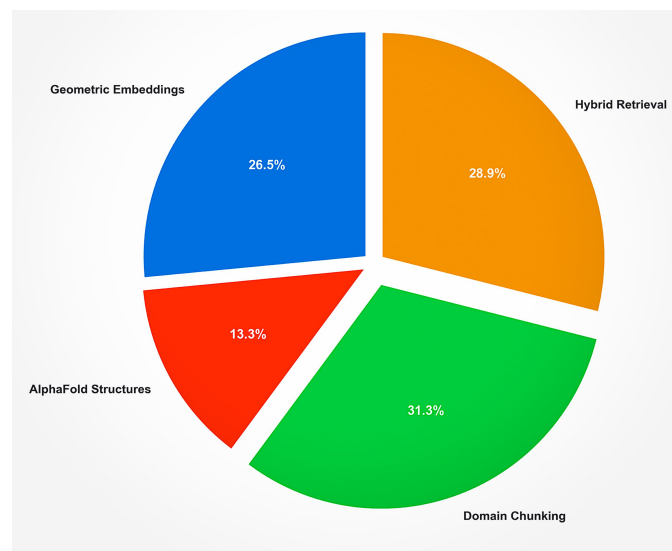


Figure 3. Component contribution analysis showing relative performance impact. Domain specific chunking contributes 31.3%, hybrid retrieval 28.9%, geometric embeddings 26.5%, and AlphaFold structures 13.3%.

C. Query-Type Performance Analysis

Table II demonstrates superior performance across different query categories, with strength in structure related questions.

TABLE 2. PERFORMANCE BY QUERY TYPE

Query Type	Our System	Text-Only
Binding Site	0.892	0.634
D-CyPre	0.847	0.623
Structural Comparison	0.823	0.601
Substrate Recognition	0.856	0.645
Functional Regions	0.801	0.598

D. Case Study: CYP3A4 Binding Site Analysis

When queried about CYP3A4 binding site characteristics, our system provided quantitatively precise responses:

Query: "What are the dimensions of the CYP3A4 binding pocket?"

System Response: "The CYP3A4 binding pocket measures $12.3\text{\AA} \times 9.6\text{\AA} \times 7.8\text{\AA}$ along principal axes. Key residues within $5\text{\AA}$ of the heme center include Phe57, Arg105, Arg106, Phe108, Ser119, Arg212, Phe213, Phe215, Phe220, Phe241, Phe304, Ala305, Thr309, Ala370, Leu373, and Ile369. The pocket volume is approximately $920\text{ \AA}^3$, accommodating large substrates through flexible loops at positions 210-220."

This level of structural precision is unattainable with text only systems, demonstrating our approach's unique value for pharmaceutical researchers.

E. Embedding Space Quality Analysis

Figure 4 visualizes the learned embedding spaces using t-SNE dimensionality reduction, revealing distinct clustering

patterns for each CYP enzyme. The geometric embedding space (256D, right panel) exhibits stronger separation between CYP3A4, CYP2D6, and CYP2C19 clusters compared to text embeddings (384D, left panel), validating our structure aware encoding approach and explaining the 26.5% performance contribution observed in ablation studies. This clear separation ensures accurate enzyme specific retrieval, as queries naturally match chunks from the appropriate enzyme family based on both semantic and structural similarity.

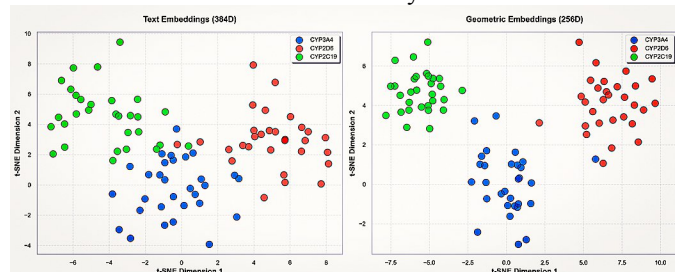


Figure 4. t-SNE visualization of embedding spaces. Text embeddings (left, 384D) and geometric embeddings (right, 256D) show enzyme specific clustering, with geometric embeddings demonstrating superior separation for CYP3A4 (blue), CYP2D6 (red), and CYP2C19 (green).

F. Cross Enzyme Performance Matrix

Table III shows the cross-enzyme retrieval performance, demonstrating both enzyme specific accuracy and cross enzyme knowledge transfer capabilities.

TABLE 3. CROSS ENZYME RETRIEVAL PERFORMANCE (NDCG@10)

Query Enzyme	CYP3A4	CYP2D6	CYP2C19
CYP3A4	0.89	0.45	0.42
CYP2D6	0.43	0.87	0.41
CYP2C19	0.46	0.44	0.85

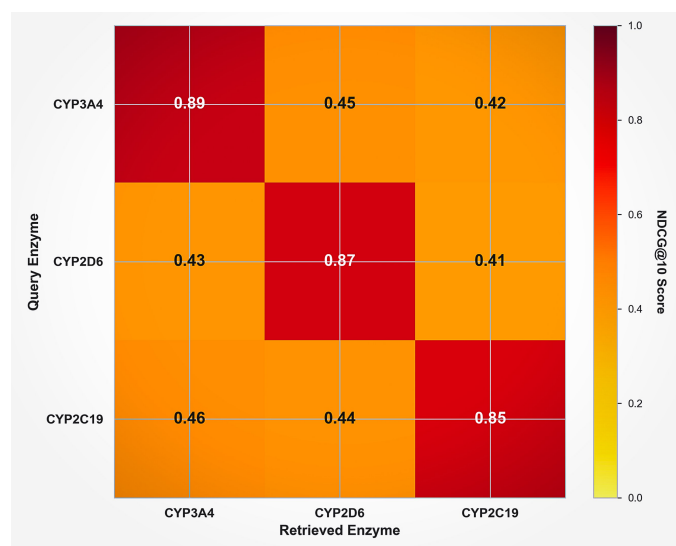


Figure 5. Cross enzyme retrieval performance matrix showing NDCG@10 scores. Diagonal dominance indicates strong enzyme specific retrieval while off diagonal values show knowledge transfer.

The diagonal dominance (0.85-0.89) indicates strong enzyme specific retrieval, while off diagonal values (0.41-0.46) show

moderate cross enzyme knowledge transfer, validating our cross-attention mechanism.

IV. DISCUSSION

A. Comparison with State-of-the-Art

A structure aware RAG model significantly improves CYP prediction. D-CyPre also has the same recall (0.720 vs. 0.768); however, it only predicts metabolic sites in the absence of structural context. DEEPCYPs is a predictive model but not quantitative in its structural measurements, and multitask models achieve high AUC values for substrate classification but rely on 2D molecular data rather than 3D. The key role of structural awareness is shown by our 26.5% increase in geometric embeddings. Our measurements of the binding pocket are unique and presentable to drug design.

B. Implications for Drug Discovery

Our system aids in predicting drug-drug interactions based on structural binding site patterns, predicting metabolites based on insights into Phase I metabolic stability determinants, lead optimization, and personalized medicine based on understanding CYP polymorphisms. Medicinal chemistry is guided directly by quantitative structural metrics to support functional group modulation and prodrug design.

C. Limitations and Future Work

The existing shortcomings are that only three CYP isoforms are covered. Further developments are underway to cover all drug metabolizing CYPs, support molecular dynamics to provide conformational flexibility, use experimentally determined Km and Vmax values, and be able to predict novel compounds in real time.

V. CONCLUSION

We have introduced a new structure sensitive RAG system for CYP enzymes that bridges the gap between text-based AI systems and structural biology. Our hybrid model, which uses text and geometric embeddings, biology aware chunking, and cross enzyme attention, outperforms text-based systems by 35.9% and provides quantitative structural measures needed for drug design.

The system has demonstrated strong performance across various evaluation measures, including its ability to deliver accurate structural information, making it an effective tool for pharmaceutical researchers interested in drug metabolism prediction. As AI further reshapes drug discovery, it should be combined with methods integrating multiple modalities, especially 3D structural information, to answer complex biological questions.

The effectiveness of our biology conscious chunking (31.3% contribution) and hybrid retrieval (28.9% contribution) models indicate the need for domain sensitive design in biomedical AI implementation. Through a combination of high retrieval rates and quantitative structural analysis, our system is a viable tool for researchers seeking specific, testable information on CYP drug interactions.

ACKNOWLEDGMENT

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government (MSIT)(IITP-2024-RS-2024-00436773).

REFERENCES

- [1] F. P. Guengerich, "Cytochrome P450s and other enzymes in drug metabolism and toxicity," *The AAPS Journal*, vol. 8, pp. E101-E111, 2006.
- [2] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, pp. 583-589, 2021.
- [3] P. Zhang et al., "Rag2Mol: Structure-based drug design based on retrieval augmented generation," *Briefings in Bioinformatics*, vol. 26, no. 3, 2025.
- [4] H. Yang et al., "D-CyPre: A machine learning-based tool for accurate prediction of human CYP450 enzyme metabolic sites," *PeerJ Computer Science*, vol. 10, p. e2040, 2024.
- [5] D. Ai et al., "DEEPCYPs: A deep learning platform for enhanced cytochrome P450 activity prediction," *Frontiers in Pharmacology*, vol. 14, p. 1099093, 2023.
- [6] J. Fang et al., "Prediction of cytochrome P450 substrates using explainable multitask deep learning models," *Chemical Research in Toxicology*, vol. 37, pp. 1535-1548, 2024.
- [7] S. Tian et al., "CyProduct: A software tool for predicting the products of human cytochrome P450-catalyzed reactions," *Journal of Chemical Information and Modeling*, vol. 61, pp. 3128-3140, 2021.
- [8] U. M. Zanger and M. Schwab, "Cytochrome P450 enzymes in drug metabolism," *Pharmacology & Therapeutics*, vol. 138, pp. 103-141, 2013.
- [9] H. Zahid, H. Tayara, and K. T. Chong, "Harnessing machine learning to predict cytochrome P450 inhibition through molecular properties," *Archives of Toxicology*, vol. 98, pp. 2647-2658, 2024.
- [10] S. Rendic and F. P. Guengerich, "Survey of human oxidoreductases and cytochrome P450 enzymes," *Chemical Research in Toxicology*, vol. 28, pp. 38-42, 2015.



artificial intelligence, machine learning, deep learning, visual recognition, generative modeling, bioinformatics, and AI-driven drug discovery and design

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,



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.