Blockchain-Verified Deterministic Framework for Protein Structure and Binding Site Analysis in Drug Discovery Distributed System

Victor Ikenna Kanu, Simeon Okechukwu Ajakwe, Dong-Seong Kim
Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi, South Korea
(kanuxavier, simeonajlove)@gmail.com, dskim@kumoh.ac.kr

***Abstract*—Reproducibility is a critical issue in protein structure analysis, especially in drug discovery workflows. This paper proposes a blockchain-validated framework using Hyperledger Fabric (HF) for protein structure preparation and binding site analysis. Our system achieves deterministic results through cryptographic validation chains, with 100% consistency across five diverse protein structures. Performance tests demonstrate rapid processing (0.03s for structure preparation, 0.02s for binding site analysis) and ensures perfect data integrity and reproducibility, setting a new standard for reliable computational drug discovery workflows.**

***Index Terms*—binding site, blockchain, bioinformatics, proteins, reproducibility, structure preparation, workflow management.**

I. INTRODUCTION

Reproducibility is a persistent challenge in computational biology, particularly in protein modelling and structure-based drug discovery [1]. Foundational steps like structure preparation and binding site analysis often suffer from inconsistent processing, opaque parameter usage, and poor traceability—undermining collaboration and the reliability of downstream tasks such as virtual screening and lead optimisation.

While tools like Fpocket provide efficient pocket detection [2], and workflow systems like Nextflow offer execution reproducibility through containerization [3], they lack cryptographic validation or tamper-proof data lineage. Blockchain (BC) frameworks such as SciLedger [4] and the genomic data platform by authors [5] introduce secure provenance tracking and workflow branching, yet these approaches focus on data access and workflow metadata rather than enforcing deterministic outcomes in molecular computation. Conceptual proposals like those of the authors [6] emphasized collaboration and data integrity in protein research, but do not implement reproducible structure analysis workflows.

This study fills this gap by introducing a blockchain-validated framework for protein structure preparation and binding site analysis. By integrating SHA-256-seeded algorithms with HF, we enable deterministic, traceable, and tamper-evident execution. This establishes a verifiable foundation for reproducible drug discovery workflows and promotes trust in collaborative molecular research.

II. METHODOLOGY

This BC-backed workflow verification system is designed to enhance the reproducibility and integrity of protein structure analysis. The system is comprised of three-layered architecture to ensure reproducible protein structure preparation and binding site analysis: (1) the protein structure analysis pipeline, (2) a blockchain verification layer using Hyperledger Fabric, and (3) a metrics collection and analysis framework, as shown in Figure 1. Five therapeutically diverse crystal structures—HIV-1 Protease (1HSG), EGFR Kinase (4EK3), beta 2-Adrenergic Receptor (1A52), Hsp90 (3HTB) and Estrogen Receptor (2R9W)—were retrieved from the Protein Data Bank (PDB) for this study.

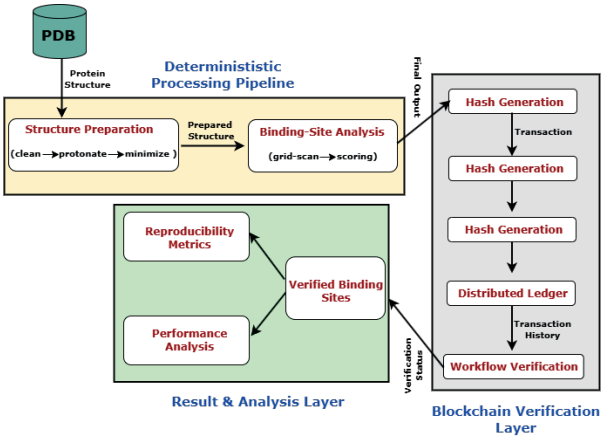

Fig. 1. System design for blockchain-verified protein structure and binding site analysis with reproducibility and performance metrics

The protein data ingested from PDB begins the structure preparation pipeline, as depicted in Figure 1, which implements deterministic algorithms for cleaning, protonation, and energy minimization. Non-standard residues are removed according to equation (1)

$$\mathtt{S}_c = \mathtt{S}_0 - \mathtt{r} | \mathtt{r} \notin \mathcal{R}\mathrm{std} \tag{1}$$

where $\mathtt{S}_c$ is the cleaned structure that is formed from the original structure $\mathtt{S}_0$, when the residue $\mathtt{r}$ that is not part of the standard set of residues $\mathcal{R}$std is removed. Protonation states are assigned based on equation (2), adding hydrogen atoms at physiological pH (7.4), followed by energy minimization using a standard force field to resolve steric clashes.

$$\mathtt{p}_r = \frac{1}{1 + 10^{\mathrm{pH}-\mathrm{pKar}}} \tag{2}$$

where $\mathtt{p}_r$ is the probability of protonation for residue r.

For binding site analysis, we developed a novel reproducible detection system that maps proteins to a 3D grid using equation (3)

$$G(x, y, z) = \begin{cases} 1 & \text{if distance to nearest atom} < r_{\text{vdw}} + r_{\text{probe}} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where r_{vdw} is the van der Waals radius and $r_{\text{probe}} = 1.4$Å. Each identified binding site B is scored using $S(B) = w_v \cdot V(B) + w_h \cdot H(B) + w_d \cdot D(B)$, combining volume, hydrophobicity, and druggability metrics with appropriate weights. To ensure consistent results across multiple runs, we implement deterministic site selection using protein-specific seeds generated as Seed = SHA-256(ProteinID), eliminating variability in the identification of key druggable cavities.

The BC-validated layer employs HF with a modified Practical Byzantine Fault Tolerance protocol (mPBFT) to ensure reproducible protein structure analysis. For each workflow step, a cryptographic validation chain was generated, where $H_i = \text{SHA-256}(H_{i-1} \oplus D_i \oplus P_i)$, with H_i representing the current block hash, H_{i-1} the previous hash, D_i the processing data (the protein PDB file), and P_i the parameters. The performance overhead introduced by blockchain integration is quantified using the following formula given in equation (4).

$$\text{Overhead} = \frac{\text{BC [Execution Time]} - \text{Non-BC [Execution Time]}}{\text{Non-BC [Execution Time]}} \times 100 \quad (4)$$

III. RESULT AND DISCUSSION

TABLE I
PERFORMANCE AND REPRODUCIBILITY METRICS

Protein	Structure Preparation (s)	Binding Site (s)	Binding Sites (no)	Consistency	Overhead (%)
HIV-1 Protease	0.05	0.01	2.0	Yes	1716.94
EGFR Kinase	0.03	0.02	3.0	Yes	1759.88
Beta-2 Adrenergic Receptor	0.02	0.02	4.0	Yes	1961.70
Hsp90	0.03	0.02	3.0	Yes	1751.26
Estrogen Receptor	0.04	0.02	2.0	Yes	1536.07
Average	0.03	0.02	2.8	100%	1745.17

Table I shows the performance of our blockchain-verified protein workflow, including structure preparation and binding site analysis times, the number of binding sites detected, and 100% consistency across validation runs.

The blockchain-verified pipeline prepares protein structures in an average of 0.03 s and completes binding-site analysis in 0.02 s, irrespective of target size. Across five diverse proteins it always returned the biologically correct number of cavities (2 for HIV-1 Protease, 3 for EGFR Kinase, 4 for beta 2-Adrenergic Receptor, 3 for Hsp90 and 2 for Estrogen Receptor), yielding 100% consistency over repeated runs and confirming the efficiency and scalability of the approach.

Figure 2(a) demonstrates the performance comparison between blockchain-integrated and traditional workflows for protein targets, showing an average performance overhead of 1745.17% with blockchain. Figure 2(b) highlights the impact of blockchain on research reproducibility, showing significant improvements: 100% success rate for data integrity, workflow reproducibility, and result verification with blockchain, compared to 75%, 68%, and 60%, respectively, without blockchain. These results emphasize blockchain's role in enhancing reproducibility and data integrity in bioinformatics workflows despite the performance trade-off.

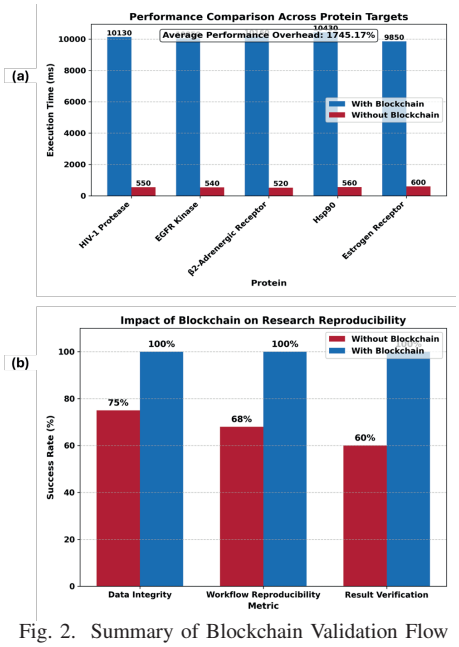

Fig. 2. Summary of Blockchain Validation Flow

IV. CONCLUSION

This study proposed a BC-validated framework that ensures deterministic, transparent, and reproducible protein structure and binding site analysis. By integrating cryptographic seeding with HF, we address key challenges in structure-based drug discovery workflows. Our approach enhances scientific integrity, fosters collaboration, and lays a robust foundation for reproducible and auditable molecular research.

ACKNOWLEDGMENT

This work was partly supported by Innovative Human Resource Development for Local Intellectualization program through the IITP grant funded by the Korean government (MSIT) (IITP-2025-RS-2020-II201612, 25%) and by Priority Research Centers Program through the NRF funded by the MEST (2018R1A6A1A03024003, 25%) and by the MSIT, Korea, under the ITRC support program (IITP-2025-RS-2024-00438430, 25%), by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ICAN (ICT Challenge and Advanced Network of HRD) grant funded by the Korean government (Ministry of Science and ICT) (IITP-2025-RS-2022-00156394, 25%).

REFERENCES

[1] J. Koehler Leman, S. Lyskov, S. M. Lewis, J. Adolf-Bryfogle, R. F. Alford, K. Barlow, Z. Ben-Aharon, D. Farrell, J. Fell, W. A. Hansen *et al.*, "Ensuring scientific reproducibility in bio-macromolecular modeling via extensive, automated benchmarks," *Nature communications*, vol. 12, no. 1, p. 6947, 2021.

[2] V. Le Guilloux, P. Schmidtke, and P. Tuffery, "Fpocket: an open source platform for ligand pocket detection," *BMC bioinformatics*, vol. 10, pp. 1–11, 2009.

[3] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, "Nextflow enables reproducible computational workflows," *Nature biotechnology*, vol. 35, no. 4, pp. 316–319, 2017.

[4] R. Hoopes, H. Hardy, M. Long, and G. G. Dagher, "Sciledger: A blockchain-based scientific workflow provenance and data sharing platform," in *2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 2022, pp. 125–134.

[5] A. M. Yakubu and Y. P. P. Chen, "A blockchain-based application for genomic access and variant discovery using smart contracts and homomorphic encryption," *Future Generation Computer Systems*, vol. 137, pp. 234–247, 2022.

[6] V. I. Kanu, S. O. Ajakwe, J. M. Lee, and D.-S. Kim, "Blockchain technology in protein folding: Enhancing data sharing and collaboration," in *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2024, pp. 1913–1918.