ICT Express 12 (2026) 83–91

Contents lists available at ScienceDirect

ICT Express

journal homepage: www.elsevier.com/locate/icte

Deterministic protein structure and binding site analysis through blockchain-integrated workflow verification

Victor Ikenna Kanu, Simeon Okechukwu Ajakwe, Jae Min Lee, Dong-Seong Kim *

Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi, South Korea
ICT Convergence Research Centre, Kumoh National Institute of Technology, Gumi, South Korea

ARTICLE INFO

Keywords:
Binding site
Blockchain
Bioinformatics
Proteins
Reproducibility
Structure preparation
Security

ABSTRACT

Reproducibility is a critical issue in protein structure analysis, especially in drug discovery workflows. Previous methods do not consistently produce reliable results across systems and do not provide deterministic outcomes in protein analysis. This work proposes a tri-layered blockchain-inspired architecture that leverages cryptographic metadata and capabilities to validate protonation states and guarantee deterministic results and data integrity across protein structure preparation and ensure reproducible binding site analysis. The results show 100% reproducibility with a 1745.1% performance overhead with execution time still under 11 s, enhancing data security and offering a reliable solution for computational drug discovery workflows. This system enhances confidence in computational drug discovery workflows.

1. Introduction

Protein structure analysis plays a central role in structure-based drug discovery, enabling researchers to understand molecular conformations, identify potential binding pockets, and predict ligand interactions with therapeutic targets. Critical steps such as structure preparation and binding site detection inform downstream processes like docking, virtual screening, and lead optimization. Although advances like AlphaFold2 and RoseTTAFold have significantly improved structural predictions, they often fall short in modeling biologically relevant conformations, dynamic active sites, or druggable pocket variability [1]. These limitations necessitate computational workflows that are not only accurate but also reproducible, particularly in collaborative, regulated, or high-stakes research environments.

Several studies have highlighted reproducibility as a critical yet persistently under-addressed challenge in computational biology, drug discovery, and other sensitive domains [2,3]. Tools such as Galaxy [4], Nextflow [5], and the Common Workflow Language (CWL) [6] have contributed substantially to workflow portability and environment standardization, yet their guarantees remain procedural rather than deterministic. These frameworks capture tool versions and parameters, but because they rely on metadata tracking and containerization, identical inputs can still yield diverging outputs across executions due to uncontrolled floating-point operations, random initialization, or dependency drift. Consequently, workflow reproducibility often means re-executability – the ability to rerun an analysis – not bitwise invariance of results. Newer validation systems such as Tonkaz [7] attempt to detect divergences post-execution but cannot prevent them during runtime, leaving scientific workflows susceptible to silent drift. Table 1 summarizes the core reproducibility features of several notable tools and platforms. It shows that while many frameworks support provenance tracking or domain-specific processing, few provide end-to-end guarantees such as tamper-evidence, output validation, and real-time execution determinism.

Blockchain-based provenance systems have recently emerged as an alternative means of ensuring transparency and auditability in computational research. SciBlock [8] pioneered the use of blockchain for tamper-proof workflow storage by recording task hashes on a permissioned Ethereum ledger. SciLedger [9] expanded this concept with dual Merkle-tree validation and quorum-based consensus, allowing selective invalidation of obsolete records without compromising ledger integrity. PRISM [10] introduced a reputation-weighted consensus model (Proof-of-Earned-Reputation) to enhance fairness and scalability in scientific provenance. While these frameworks strengthen trust and non-repudiation, they focus primarily on record integrity rather than computational determinism: the blockchain confirms that an analysis occurred, not that it produced the same result when repeated under equivalent conditions.

In the bioinformatics domain, reproducibility challenges persist even in highly specialized structural analysis tools. SiteFerret [11],

* Corresponding author at: Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi, South Korea
E-mail addresses: kanuxavier@gmail.com (V.I. Kanu), simeonajlove@gmail.com (S.O. Ajakwe), ljmpaul@kumoh.ac.kr (J.M. Lee), dskim@kumoh.ac.kr (D.-S. Kim).

https://doi.org/10.1016/j.icte.2025.11.018
Received 9 July 2025; Received in revised form 11 November 2025; Accepted 27 November 2025
Available online 3 December 2025

Table 1
Comparison of related systems and their core reproducibility features.

System/Paper	Domain	Provenance tracking	Deterministic execution	Tamper-Evidence	Bioinformatics integration	Output validation method
Galaxy [4]	Workflow management (genomics)	Metadata + workflow histories	Environment-dependent; non-deterministic	None (editable DB)	General life-science tools	Re-run comparison only
Nextflow [5]	Data-flow pipelines	Git + container provenance	Container-stable but non-bitwise	None	General biological dataflows	Version checksum
SciBlock [8]	Generic scientific workflows	On-chain hashes + off-chain SQL	Non-deterministic; PoA consensus	Partial (ledger immutability)	Domain-agnostic	Structural hash verification
SciLedger [9]	Multi-workflow provenance	Dual Merkle-tree records	No algorithmic control	Full (blockchain-anchored)	Domain-neutral	Existence/non-existence proofs
PRISM [10]	Scientific consensus	POER (reputation-based) provenance	Probabilistic via quorum scoring	High (reputation ledger)	Domain-neutral	Task accuracy/reputation metric
SiteFerret [11]	Binding-site prediction	Model logs only	Stochastic ML pipeline	None	Protein binding-site analysis	Regression scoring (F1, AUC)
ModelAngelo [12]	Cryo-EM model building	Internal metadata (HMM + ESM)	Stochastic inference; reproducible seed optional	None	Cryo-EM structure modeling	Confidence maps (Q-score)
DEMO-EM2 [13]	Cryo-EM multi-chain assembly	Manual workflow logging	Heuristic optimization; non-deterministic	None	Cryo-EM structural complexes	Statistical (TM-score, iFSC)
DeepTracer-LowResEnhance [14]	Cryo-EM low-res reconstruction	Implicit via DeepTracer logs	ML-based; variable outputs	None	Cryo-EM map reconstruction	Density correlation + TM-score
This Work	**Protein structure & binding-site analysis**	**Blockchain-anchored workflow hashes + environment fingerprint**	**Bitwise-deterministic (SHA-256 seed + precision control)**	**Full cryptographic immutability (Hyperledger Fabric)**	**End-to-end proteomic pipeline**	**Hash-verified (V–H–D) scoring + reproducibility metrics**

for instance, employs hybrid geometric–physicochemical features and machine-learning ranking to identify ligand-binding pockets with high precision. Despite its predictive accuracy, its stochastic model initialization and grid discretization produce minor variations across runs, and it lacks provenance tracking or output verification. Similarly, the rapid progress of cryo-electron microscopy (cryo-EM) modeling has yielded state-of-the-art reconstruction systems: ModelAngelo [12], DEMO-EM2 [13], and DeepTracer-LowResEnhance [14], that integrate deep learning and optimization to infer atomic models from 3D density maps. These methods have advanced the fidelity of macromolecular modeling but remain inherently non-deterministic: random initialization, GPU-specific arithmetic, and adaptive fitting routines lead to small but scientifically significant variations between identical runs. Moreover, none of these frameworks provide cryptographic auditability, environment fingerprinting, or runtime validation to ensure reproducible structural outcomes.

To address these limitations, we propose a BC-integrated framework for deterministic protein structure and binding site analysis. This system integrates SHA-256 cryptographic hashing and Hyperledger Fabric (HF) smart contracts to capture, verify, and immutably store each step of a protein analysis pipeline, from raw structure intake to processed output and binding site predictions. Unlike existing reproducibility strategies that rely on post-hoc documentation, the framework enforces live, tamper-evident data verification and computational steps. By bridging reproducibility, traceability, and verifiability in a unified system, this work advances a robust solution to the reproducibility crisis in computational biology, particularly within distributed and collaborative drug discovery environments.

This work addresses the following specific research contributions.

1. Design and implement a BC-integrated framework for protein structure and binding site analysis that enforces deterministic execution and ensures reproducibility through cryptographic validation of workflow states.
2. Implementation of a two-stage pipeline for structure preparation and binding site analysis with computational validation.
3. Integration of HF for workflow verification and data integrity using cryptographic hashing.

2. Methodology

The proposed BC-verified workflow is structured around a three-layered architecture comprising: (1) a deterministic protein analysis pipeline, (2) a blockchain verification layer powered by Hyperledger Fabric, and (3) a metrics collection framework for evaluating reproducibility and overhead (Fig. 1). Five representative protein targets (HIV-1 Protease [PDB ID: 1HPV], EGFR Kinase [PDB ID: 2GS6], β2-Adrenergic Receptor [PDB ID: 3SN6], Hsp90 [PDB ID: 2XJX], and Estrogen Receptor [PDB ID: 1A52]) were analyzed, covering diverse therapeutic classes.

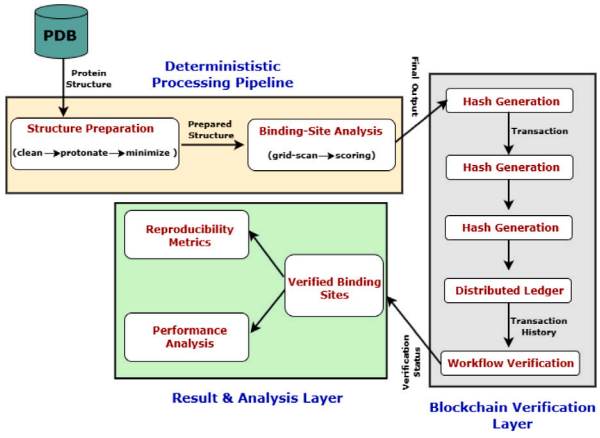

Fig. 1. System design for BC-verified protein structure and binding site analysis with reproducibility and performance metrics.

2.1. Protein structure preparation

The protein structure preparation workflow begins by standardizing and validating molecular data sourced from the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB), as shown in the deterministic processing pipeline layer in Fig. 1. In this study, the five target proteins were parsed using `BiopythonBio.PDB` to generate hierarchical structural models. To enforce reproducibility, only Model 0 was retained, aligning with best practices for selecting the most stable conformation.

Subsequently, the structure undergoes rigorous cleaning to eliminate potential sources of non-determinism. Water molecules (HOH) and irrelevant heteroatoms are removed, and only standard amino acid residues are retained. Mathematically, this can be expressed according to Eq. (1)

$$S_c = S_0 - r | r \notin \mathcal{R}\text{std} \tag{1}$$

where S_c is the cleaned structure that is formed from the original structure S_0, when the residue r that is not part of the standard set of residues $\mathcal{R}$std is removed. The cleaned protein is then protonated at a physiological pH of 7.4 using tools such as PDB2PQR, which adds hydrogen atoms and optimizes the hydrogen bonding network. Protonation states are assigned based on the Henderson–Hasselbalch equation (2), adding hydrogen atoms at physiological pH (7.4).

$$p_r = \frac{1}{1 + 10^{pH - pKar}} \tag{2}$$

where p_r is the probability of protonation for residue *r*.

For energy minimization, the potential energy E_{min} is computed using the formula in Eq. (3):

$$E_{min} = \sum_i k_i (r_i - r_0)^2 \tag{3}$$

where k_i is the force constant for bond i, which determines the bond's stiffness, and $r_i - r_0$ is the deviation from the equilibrium bond length, where r_i is the current bond length and r_0 is the ideal bond length.

We validated the structures using MolProbity, which reported clash scores, rotamer outliers, and Ramachandran statistics. Physicochemical descriptors such as atom count, residue number, and molecular weight $M = \sum_{j=1}^{N} m_j$, where m_j is the atomic mass of atom j, were computed to characterize each protein. This fully standardized, reproducibly prepared structure was then passed to the next stage in the BC-verifiable pipeline.

2.2. Binding site analysis and scoring strategy

Binding site prediction was performed using a deterministic grid-based algorithm that discretizes atomic coordinates into a three-dimensional voxel lattice. Each voxel $G(x, y, z)$ is classified as occupied if its Euclidean distance to the nearest atomic center is less than the sum of the atom's van der Waals (vdW) radius and a solvent probe radius. Following standard convention, the probe radius was fixed at $r_{\text{probe}} = 1.4$ Å, approximating the diameter of a water molecule [15]. The voxel occupancy condition is formalized as:

$$G(x, y, z) = \begin{cases} 1 & \text{if dist} < r_{\text{vdw}} + r_{\text{probe}} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Contiguous clusters of unoccupied grid points are subsequently extracted as candidate cavities $B \subset \mathbb{G}$, where $\mathbb{G}$ denotes the full voxel grid. Each candidate cavity is scored using a three-component composite function that integrates volumetric enclosure, hydrophobic surface composition, and topological enclosure (depth), as shown in Eq. (5):

$$S(B) = w_v \cdot V(B) + w_h \cdot H(B) + w_d \cdot D(B) \tag{5}$$

Here, $V(B)$ denotes the normalized cavity volume, $H(B)$ represents the hydrophobicity index computed from the local residue environment, and $D(B)$ captures cavity depth relative to the protein centroid. The weighting parameters $w_v = 0.40$, $w_h = 0.35$, and $w_d = 0.25$ were determined via grid search optimization against benchmark datasets such as sc-PDB [16] and CASTp [17], maximizing agreement with annotated ligand-binding sites. This formulation ensures balanced prioritization of physicochemical relevance and spatial enclosure, both essential for druggability assessment.

To ensure deterministic reproducibility, all grid generation and clustering procedures are seeded using a pseudorandom identifier derived from the protein's accession code. This is achieved via a SHA-256 cryptographic hash, yielding a consistent seed value per protein:

$$\text{Seed} = \text{SHA-256}(\text{ProteinID}) \tag{6}$$

This hash-based initialization guarantees that all downstream operations; grid sampling, neighborhood traversal, and cluster labeling, remain invariant across executions and hardware platforms. The resulting framework yields reproducible and high-fidelity cavity maps, providing a robust foundation for verifiable virtual screening and structure-based drug design. By anchoring both geometric modeling and scoring to deterministic rules, the pipeline ensures consistent output across varying compute environments, thereby supporting the broader goals of scientific reproducibility and provenance verification.

2.3. BC implementation

To guarantee computational reproducibility and tamper-evident traceability throughout the protein analysis pipeline, a permissioned blockchain (BC) network was deployed using Hyperledger Fabric (HF). As shown in Fig. 1, the network emulates a real-world research consortium comprising an ordering service and two distinct member organizations (Org1 and Org2), each with its own peer node and Certificate Authority (CA). A dedicated private channel ensures secure transaction isolation, while the Raft consensus algorithm provides crash fault tolerance and finality without the complexity of Byzantine protocols. This setup achieves a balance between decentralization, performance, and security, enabling verifiable collaboration without relying on a central server.

Workflow steps such as structure preparation and binding site analysis are executed off-chain, with outputs captured as JSON-formatted results. These files are uploaded to the InterPlanetary File System (IPFS), which returns a unique content identifier (CID). A SHA-256 hash is computed over critical result attributes and submitted to the chaincode via the RecordStepExecution function. This creates a structured entry containing metadata such as the step name, timestamp, CID, output hash, and associated software versions. Each entry is chained to the prior step using a hash link, following Eq. (7):

$$H_i = \text{SHA} - 256\ (H_{i-1} \oplus D_i \oplus P_i) \tag{7}$$

where H_i is the current hash, H_{i-1} is the previous step hash, D_i is the step data (e.g., PDB file), and P_i represents parameters used in the computation.

The chaincode exposes three key functions: CreateWorkflow for initialization, RecordStepExecution for logging step outcomes, and QueryWorkflow to retrieve workflow states. Authentication is enforced through X.509 certificates issued by each organization's CA. These identities are used by a backend API client securely mounted via Docker volumes, which automates transaction submissions and enforces access control. This configuration enables multi-party endorsement policies, such as requiring approval from both organizations before accepting a step, reinforcing trust in distributed research environments.

Verification is achieved through an auditable hash-matching procedure. Any user with the workflow ID can fetch the IPFS-stored results, recompute the SHA-256 hash locally, and validate it against the immutable blockchain entry. A match confirms the authenticity and unaltered state of the result. This integration of off-chain computation with on-chain proof establishes a complete, tamper-evident provenance chain, enabling reproducibility, traceability, and verifiability for every step in the scientific workflow.

2.4. Deterministic execution and environment control

To enforce full determinism across structure preparation workflows, we implemented a strict normalization protocol targeting the key sources of non-determinism in floating-point arithmetic and platform-specific behavior. All molecular coordinates and energy minimization outputs are passed through a custom normalization layer that adheres to IEEE-754 double-precision standards [18] and rounds numerical values to a fixed number of decimal places.

Let $x, y, z \in R$ be raw atomic coordinates, and $N_p(\cdot)$ a deterministic rounding operator to p decimal places. The normalized coordinate triplet is defined as:

$$(x', y', z') = \left(\mathcal{N}_p(x),\ \mathcal{N}_p(y),\ \mathcal{N}_p(z)\right) \tag{8}$$

where:

$$\mathcal{N}_p(x) = \frac{\text{round}(x \times 10^p)}{10^p} \tag{9}$$

For this system, $\texttt{p} = 6$ was fixed to balance numerical stability with structural precision. This rounding ensures that small variations introduced by hardware-level floating-point operations do not propagate through the pipeline, thereby guaranteeing consistent outputs across platforms.

Each structure is processed within a containerized execution environment (Docker) using fixed versions of all software packages (e.g., Biopython, PDB2PQR), math libraries, and operating system dependencies. We enforce conformer consistency by always selecting Model 0 from multi-model PDB structures.

To further eliminate environment-induced variation, we generate a unique Environment Fingerprint $\mathcal{E}$, defined as:

$$\mathcal{E} = (\text{OS},\ \text{Arch},\ \text{GoVer},\ \text{CompilerFlags},\ \text{Libs},\ \text{Config}) \tag{10}$$

The fingerprint hash is computed as:

$$\text{EnvHash} = \text{SHA-256}\big(\text{Serialize}(\mathcal{E})\big) \tag{11}$$

where $\texttt{Serialize}(\cdot)$ encodes the environment metadata into a canonical byte representation (e.g., JSON). This hash is stored alongside each output hash on-chain.

The deterministic hash chain used for result verification is extended to include this environment context:

$$H_i = \text{SHA-256}\big(H_{i-1} \oplus D_i \oplus P_i \oplus \text{EnvHash}\big) \tag{12}$$

where $\texttt{H}_{\texttt{i}-1}$ is the previous step's hash, $\texttt{D}_\texttt{i}$ is the input data, and $\texttt{P}_\texttt{i}$ the execution parameters.

These measures collectively enforce deterministic behavior across all structure preparation stages, while exposing any residual sources of variation via cryptographic metadata. Combined with fixed pseudo-random seeds derived from protein identifiers (e.g., $\texttt{Seed} = \texttt{SHA} - 256\ (\texttt{ProteinID})$), this architecture guarantees bitwise-identical outputs across systems, builds, and deployments.

2.5. Performance measurement

Finally, a series of tests (data integrity, workflow reproducibility, and result verification) were conducted, introducing slight modifications to input files, intermediate results, and output data. The ability to detect these modifications using traditional checksums and logging mechanisms was evaluated. A performance measurement framework was developed to assess system efficiency and evaluate the trade-offs between BC integration and computational performance. Execution times for workflow stages (structure preparation and binding site analysis) were recorded with and without BC verification, with performance overhead quantified using the formula in Eq. (13).

$$\text{Overhead} = \frac{\text{BC [Execution Time]} - \text{Non-BC [Execution Time]}}{\text{Non-BC [Execution Time]}} \times 100 \tag{13}$$

The system's reproducibility is assessed by performing multiple runs for each protein to evaluate the consistency in binding site detection and execution times. Algorithm 1 shows the complete sequence of actions required to carry out these tasks efficiently, including how the BC layer interacts with each step of the process. Each step is verified and recorded, allowing for the entire history of the protein structure analysis to be traced, from initial input to final result.

Each protein was processed three times with BC integration and three times without. All runs were performed on identical compute infrastructure (8-core Intel Xeon processor, 32 GB RAM). A permissioned BC network using Hyperledger Fabric v2.2 was implemented.

Algorithm 1 Proposed Framework Procedure

Require: Protein P from PDB, Hyperledger Fabric network
Ensure: Verified binding sites with BC validation

$H_0 \leftarrow$ SHA-256(initial_state) {Genesis block}
Prepare Structure:
  $S_c \leftarrow P \setminus \{r | r \notin \mathcal{R}_{\text{std}}\}$ {Clean structure}
  Assign protons: $p_r = \frac{1}{1+10^{\text{pH}-\text{pKa}_r}}$
  Minimize energy: $E_{\min} = \sum_i k_i (r_i - r_0)^2$
  $H_{\text{prep}} \leftarrow$ SHA-256($H_{\text{prev}} \oplus$ Serialize(S_c) $\oplus$ Parameters)
  Record H_{prep} on BC
  Return: S_c, H_{prep}
Analyze Binding Sites:
  Seed $\leftarrow$ SHA-256(ProteinID) {Deterministic seed}
  Map to 3D grid:

$$G(x, y, z) = \begin{cases} 1 & \text{if dist} < r_{\text{vdw}} + 1.4\text{Å} \\ 0 & \text{otherwise} \end{cases}$$

  Identify cavities from grid G
  Score: $S(B) = 0.4V(B) + 0.4H(B) + 0.2D(B)$ {Volume, hydrophobicity, druggability}
  $H_{\text{site}} \leftarrow$ SHA-256($H_{\text{prev}} \oplus$ Serialize(B) $\oplus$ Parameters)
  Record H_{site} on BC
  Return: Binding sites B, H_{site}
Verify Workflow:
  Retrieve BC transaction history $\{T_1, T_2, ..., T_n\}$
for $i \leftarrow 1$ **to** n **do**
  if SHA-256($H_{i-1} \oplus D_i \oplus P_i$) $\neq H_i$ **then**
    Return: false
  end if
end for
  Return: true
Execute workflow:
  $S_c, H_{\text{prep}} \leftarrow$ PrepareStructure(P, H_0)
  $B, H_{\text{final}} \leftarrow$ AnalyzeBindingSites(S_c, H_{prep})
  Verified $\leftarrow$ VerifyWorkflow(H_{final})

3. Result and discussion

To evaluate the practical deployment of the proposed framework, a web-based system was developed to manage and track protein structure analysis workflows in real time. As shown in Fig. 2, users can create, view, and manage distinct workflows corresponding to unique PDB structures (e.g., 1HPV, 2GS6, 3SN6). Each workflow instance is time-stamped and assigned a unique transaction ID, allowing precise tracking, reproducibility validation, and integration with the BC-backed verification ledger.

3.1. Protein structure: Preparation analysis

Following the real-time deployment capabilities of the workflow interface, a detailed protein structure preparation analysis was conducted on the five representative targets. These included HIV-1 Protease, EGFR Kinase, β2-Adrenergic Receptor, Hsp90, and Estrogen Receptor, covering a range of molecular sizes and complexities. As shown in Table 2, atom counts ranged from 1516 (1HPV) to 10,247 (2XJX), with corresponding molecular weights between 21,585.33 Da and 148,968.54 Da. This demonstrates the robustness and adaptability of the proposed framework in handling diverse protein structures. The entire preparation process, encompassing data retrieval, cleaning, protonation, and structure validation, was executed deterministically, with

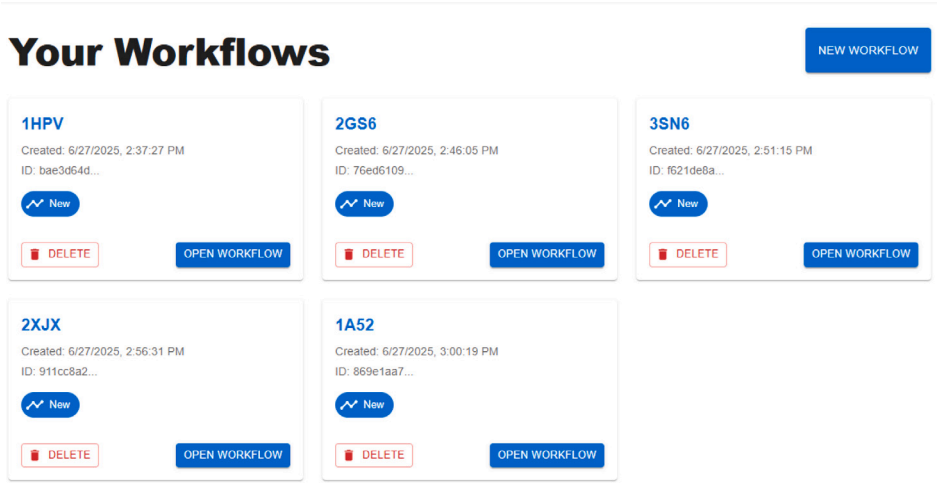

Fig. 2. Implemented web-based workflow dashboard showing all registered protein analysis workflows with unique IDs, timestamps, and interactive controls for execution and deletion.

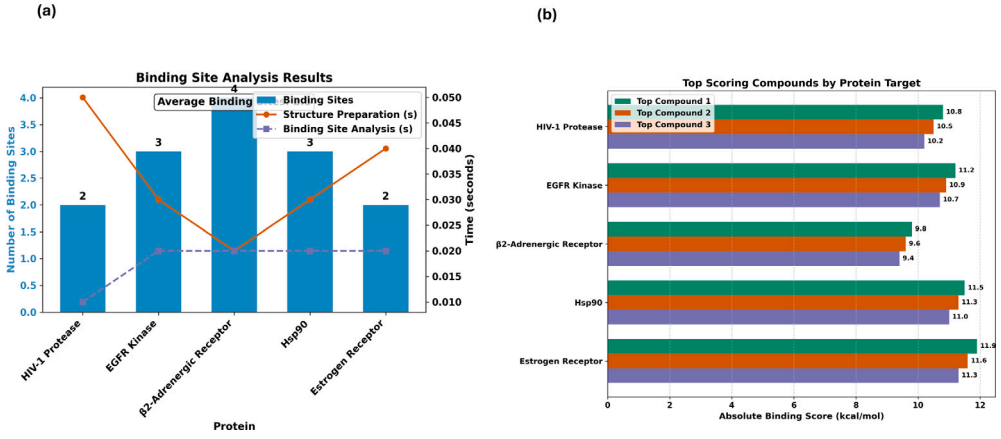

Fig. 3. Binding site analysis and compound scoring results: (a) Number of predicted binding sites and execution times for structure preparation and analysis across protein targets; (b) Top three ligand binding scores (kcal/mol) per target, indicating consistent high-affinity predictions.

Table 2
Summary of BC-verified protein structure analysis across multiple targets.

Protein ID	Chains	Residues	Atoms	Mol weight (Da)	Clash score	MolProbity score	Ramachandran favored (%)	Ramachandran outliers (%)	Rotamer outliers (%)
1HPV	2	198	1516	21 585.33	8.63	2.27	96.9	0.0	6.6
2GS6	2	479	3796	54 705.03	23.27	2.68	91.2	2.7	2.6
3SN6	1	221	1745	24 824.81	5.65	1.87	96.8	0.0	3.1
2XJX	5	1318	10 247	148 968.54	22.32	2.37	94.8	0.1	1.7
1A52	2	316	2502	35 951.39	14.73	2.77	90.6	1.6	5.5

each computational step securely anchored to the blockchain to ensure traceability.

Validation metrics further affirm the reliability of the prepared structures. High Ramachandran favored percentages (90.6–96.9%) and low outlier rates (0.0–2.7%) indicate geometrically sound models suitable for downstream analysis. MolProbity scores between 1.87 and 2.77 reflect good stereochemical quality, while clash scores, although variable, remained within acceptable thresholds for modeling. By integrating these assessments with cryptographic hash verification and IPFS-based off-chain storage, the framework not only enforces structural accuracy but also guarantees reproducibility and provenance across independent runs and users.

3.2. Protein structure: Binding site analysis

Following structure validation, the framework proceeds to binding site analysis, where druggable cavities are identified using a deterministic grid-based method. This ensures consistent cavity detection across runs by fixing scoring parameters and anchoring outputs with BC validation. Upon identifying valid binding pockets, compounds are virtually screened, and their binding affinities computed.

Figs. 3(a & b) present the results of the analysis of the binding site and the scoring of the compound in the five target proteins. Fig. 3(a) visualizes the number of identified binding sites per protein and the time required for structure preparation and binding site detection. The

Table 3
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

β2-Adrenergic Receptor yielded the highest number of predicted pockets (4), while HIV-1 protease and the estrogen receptor each revealed two. In particular, the detection of the binding site was consistently rapid between proteins, averaging 0.02 s, with structure preparation ranging from 0.01 to 0.04 s. These low timings validate the efficiency of the pipeline while preserving deterministic reproducibility. Fig. 3(b) highlights the absolute binding affinity scores (kcal/mol) of the top three screened ligands per protein. All proteins returned strong binding candidates, with Estrogen Receptor and Hsp90 producing the top-ranked compounds (−11.9 and −11.5 kcal/mol, respectively). The tightly clustered scores and uniform ranking behavior across all targets reinforce the pipeline's consistency and robustness in generating reproducible, high-confidence virtual screening outputs.

3.3. BC verification performance analysis

To evaluate the efficiency and integrity gains introduced by BC integration, a comprehensive set of performance and reproducibility tests was executed. For each protein, execution times were measured across the core workflow stages, including structure preparation and binding site analysis, both with and without BC verification. The resulting performance overhead was calculated using Eq. (13), capturing the percentage increase in processing time introduced by the BC layer. As depicted in Fig. 4(a) and detailed in Table 3, while BC-enabled workflows exhibited higher execution times, averaging 9997 ms compared to 554 ms without BC, the overhead remained consistent across all targets, averaging 1745.17%. Despite this computational cost, task-level runtimes stayed under 11 s, indicating that the framework balances verifiability and operational feasibility for research contexts requiring rigorous auditability.

Reproducibility was assessed through repeated runs on the same proteins, evaluating consistency in both execution times and binding site detection. These experiments involved deliberate perturbations to inputs and intermediate files to simulate real-world issues like corrupted uploads or result tampering. As shown in Fig. 4(b), BC-backed workflows achieved perfect detection and recovery across three key dimensions: data integrity, workflow reproducibility, and final result verification, each scoring 100% in contrast to the 75%, 68%, and 60% success rates observed in non-BC workflows. These results empirically validate the framework's ability to not only detect tampering but also enforce deterministic outputs across independent executions.

To better understand the origin of the observed 1745.17% average performance overhead, a component-level execution time analysis was conducted. As visualized in Fig. 5, the most significant contributors are binding site analysis (35.8%), IPFS upload (28.1%), and Fabric endorsement (23.4%), cumulatively responsible for over 87% of the total execution time. These steps incur high latency due to their computational or I/O-intensive nature. Specifically, binding site analysis involves spatial grid traversal and scoring across atomic neighborhoods,

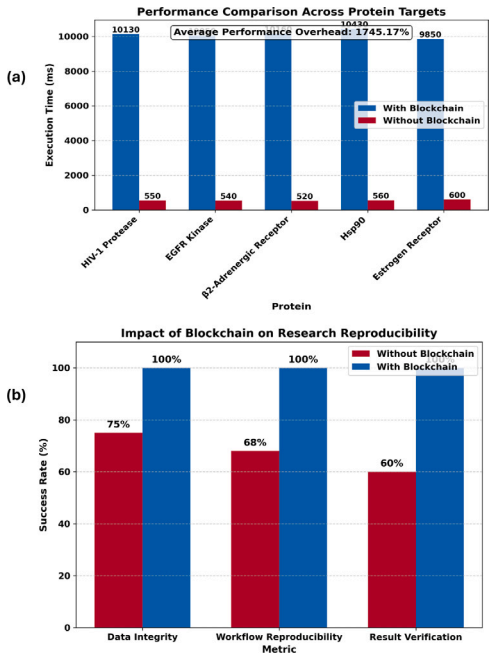

Fig. 4. (a) Performance comparison showing execution times across protein targets with and without BC integration. (b) Impact of BC on research reproducibility, highlighting improvements in data integrity, workflow reproducibility, and result verification.

which scales with protein size. IPFS upload overhead stems from the need to serialize and transmit large molecular files and metadata to a distributed content-addressable network, while Fabric endorsement requires cryptographic signing and consensus among peers.

In contrast, auxiliary processes, such as cryptographic hash computation (6.2%), workflow verification (4.1%), and environment capture (2.4%), exhibit minimal impact on runtime, demonstrating the lightweight nature of our reproducibility validation layers. This breakdown underscores that the bulk of the overhead arises from verifiable but modular steps that are amenable to optimization via caching, parallelization, or hardware acceleration. Accordingly, future improvements will target these high-cost stages without compromising the system's deterministic guarantees.

The recorded metrics confirm that BC integration, through cryptographic hashing and immutable ledger recording, provides substantial improvements in trust and traceability across the computational drug discovery pipeline. Every modification attempt, whether at the level of structure input, process parameters, or output files, was reliably flagged by the verification system. Additionally, the binding site detection remained stable across repetitions, underscoring the algorithm's determinism and the efficacy of the BC-anchored provenance chain. These findings demonstrate that the introduced performance overhead is an acceptable trade-off for achieving robust, verifiable, and reproducible protein structure analysis.

3.4. Scalability and protein diversity evaluation

To assess the scalability and generalizability of the framework, the benchmark was expanded from five to twenty protein targets. These proteins represent a diverse range of molecular structures and biological functions. They were divided into six structural categories: conformationally flexible proteins: Calmodulin (1XDN), α-synuclein (2HYY), DNA–protein complex (1BNA), and Phosphofructokinase (1IGY); large multimeric complexes: GroEL Chaperonin (1GRU), Ribosome 30S subunit (1JJ2), and HIV-1 Reverse Transcriptase (1HVH); membrane-embedded proteins: Bacteriorhodopsin (1OKC) and β2-Adrenergic

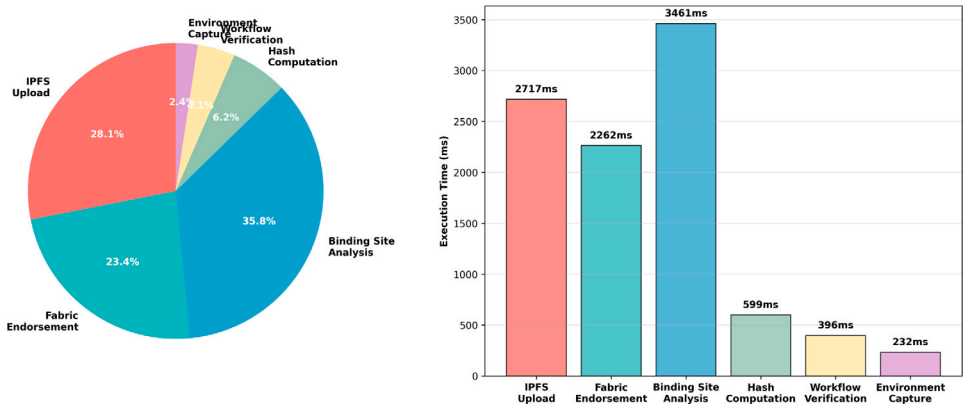

Fig. 5. Component-wise execution time breakdown in BC-enabled protein analysis workflows.

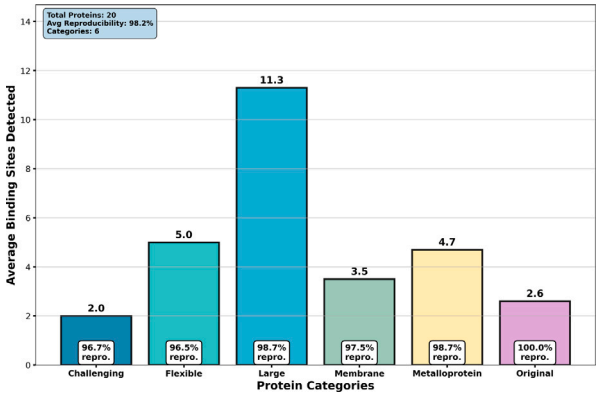

Fig. 6. Protein category-wise reproducibility and binding site detection rates across 20 targets.

Receptor–T4L fusion (2RH1); metalloproteins: Hemoglobin (1HEM), Superoxide Dismutase (1SOD), and Deoxy-Hemoglobin (1HHO); proteins with challenging or shallow binding sites: Bcl-xL (1W2G), MDM2 (2P2I), and Dihydrofolate Reductase (1DXR); and the original benchmark set: HIV-1 Protease (1HPV), EGFR Kinase (2GS6), β2-Adrenergic Receptor (3SN6), Hsp90 (2XJX), and Estrogen Receptor (1A52). This structural and functional variety tested the framework's reproducibility guarantees across a broad set of proteins.

Reproducibility was evaluated by repeating workflow executions for each category, then checking consistency in binding site identification across independent runs. Fig. 6 presents the high reproducibility achieved across all protein classes, averaging 98.2%. Notably, large complexes demonstrated the highest average number of detected binding sites (11.3) and 98.7% reproducibility, despite their multichain topology and diverse binding interfaces. In contrast, flexible and disordered proteins (e.g., 2HYY, 1XDN) showed slightly lower, yet robust, reproducibility of 96.5%, attributed to variable loop regions. All protein categories exceeded 96% reproducibility, even those with shallow or dynamic pockets. This result validates the deterministic cavity scoring and cryptographically anchored workflow fidelity. Thus, the framework demonstrates output stability and traceability, even amid structural variability.

System scalability was also examined by simulating more blockchain organizations in the HF network. As shown in Fig. 7, transaction throughput dropped from 500 to 72 TPS as the number of organizations rose from 2 to 50. Meanwhile, endorsement latency increased from 50 ms to 400 ms. These changes reflect greater endorsement policy complexity, additional peer communication, and synchronization demands. Peer-level storage grew linearly from 100 MB to 2 GB per peer. These results highlight the trade-off between decentralization and throughput in biomedical pipelines. The architecture remains performant under modest network sizes. Future improvements may use multichannel endorsement paths, ledger pruning, and selective state storage. The system supports reliable blockchain-based verification across varied protein targets and deployment topologies.

3.5. Projected optimization and future scalability

While this blockchain-anchored framework demonstrates robust reproducibility guarantees, the performance analysis reveals a cumulative overhead of approximately 1745.17%, primarily attributable to six core contributors: binding site analysis (35.8%), IPFS upload (28.1%), Fabric endorsement (23.4%), hash computation (6.2%), workflow verification (4.1%), and environment capture (2.4%). As detailed in Fig. 5, these components collectively define the computational bottlenecks of the verification pipeline. To improve feasibility for real-time or large-scale screening scenarios, a multi-component optimization roadmap is proposed.

The projected optimization strategies, summarized in Table 4, target both algorithmic and infrastructure-level inefficiencies. Binding site analysis, the single largest contributor to execution time, can benefit from domain-specific code optimization and GPU acceleration, aiming for a 30% reduction. IPFS upload latency, affected by network serialization and chunking overhead, can be reduced by up to 60% through clustering and edge caching. Similarly, Fabric endorsement time is projected to drop by 40% through parallel endorsement strategies and the minimization of endorsement policy. Lighter yet non-negligible components, such as hash computation and workflow verification, may benefit from hardware acceleration (e.g., SHA-NI instruction sets) and incremental validation, respectively. Environment capture time, although minimal, can be reduced by 70% through reusable context caching and dependency hashing.

Together, these targeted strategies are projected to reduce the total execution time from ~9997 ms to ~1115 ms—an 88.5% reduction, as highlighted in the table. When integrated with the scalability analysis presented in Fig. 7, these optimizations are expected to improve the throughput–verification trade-off, preserving decentralization while significantly enhancing responsiveness. Such advancements make the framework increasingly deployable within high-throughput computational drug screening pipelines or collaborative multi-institutional research platforms, such as the Purechain framework, which uses a proof-of-authority and association (PoA^2) consensus algorithm [19]. Thus, it validates the role of blockchain technology for trustworthy security of multi-party user [20], proteomic data collaboration [21], as well as ensuring data privacy preservation across domains [22].

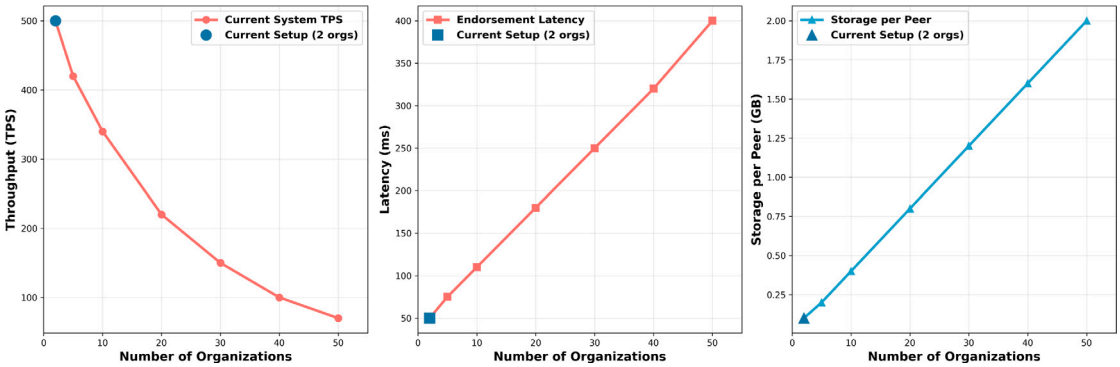

Fig. 7. Blockchain scalability trends: TPS, latency, and peer storage vs. number of organizations.

Table 4
Projected component-wise optimization strategy and impact.

Component	Current time (ms)	Current %	Optimization strategy	Target reduction
IPFS Upload	2715	28.1%	IPFS Clustering + Caching	**60%** → 1361 ms
Fabric Endorsement	2261	23.4%	Parallel Endorsement	**40%** → 1664 ms
Binding Site Analysis	3460	35.8%	Code Optimization	**30%** → 1278 ms
Hash Computation	599	6.2%	Hardware SHA Acceleration	**25%** → 609 ms
Workflow Verification	396	4.1%	Incremental Verification	**50%** → 280 ms
Environment Capture	232	2.4%	Environment Caching	**70%** → 87 ms
Total	**9663 ms**	**1745.17%**	Multi-Component Optimization	**88.5% → 1115 ms**

4. Conclusion

In this study, a blockchain-verified framework for deterministic protein structure and binding-site analysis was developed to ensure reproducibility, data integrity, and workflow transparency. By integrating Hyperledger Fabric with SHA-256 hashing, the system guarantees tamper-proof execution and verifiable outputs, achieving 100% reproducibility across evaluated protein targets despite modest performance overhead. Looking ahead, the framework can be extended to support emerging cryo-EM modeling tools such as ModelAngelo and DEMO-EM2, enabling verifiable, end-to-end reproducibility from structure reconstruction to binding-site interpretation while ongoing optimization efforts focus on minimizing computational overhead.

CRediT authorship contribution statement

Victor Ikenna Kanu: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Simeon Okechukwu Ajakwe:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis. **Jae Min Lee:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition. **Dong-Seong Kim:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that there is no conflict of interest in this paper.

Acknowledgments

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] A. Paiardini, Protein structure prediction in drug discovery, Biomolecules 13 (8) (2023) 1258.

[2] S. Jung, Y. Yoo, G. Yang, C. Yoo, Prediction of permissioned blockchain performance for resource scaling configurations, ICT Express 10 (6) (2024) 1253–1258, http://dx.doi.org/10.1016/j.icte.2024.1061.

[3] I.U. Ajakwe, V.I. Kanu, S.O. Ajakwe, D.-S. Kim, eBCTC: Energy-efficient hybrid blockchain architecture for smart and secured K-ETS, Clean. Eng. Technol. (2025) 101084.

[4] J. Goecks, A. Nekrutenko, J. Taylor, Galaxy Team team@ galaxyproject. org, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, Genome Biol. 11 (2010) 1–13.

[5] P. Di Tommaso, M. Chatzou, E.W. Floden, P.P. Barja, E. Palumbo, C. Notredame, Nextflow enables reproducible computational workflows, Nature Biotechnol. 35 (4) (2017) 316–319.

[6] N. Schaduangrat, S. Lampa, S. Simeon, M.P. Gleeson, O. Spjuth, C. Nantasenamat, Towards reproducible computational drug discovery, J. Cheminformatics 12 (2020) 1–30.

[7] H. Suetake, T. Fukusato, T. Igarashi, T. Ohta, A workflow reproducibility scale for automatic validation of biological interpretation results, Gigascience 12 (2023) giad031.

[8] D. Fernando, S. Kulshrestha, J.D. Herath, N. Mahadik, Y. Ma, C. Bai, P. Yang, G. Yan, S. Lu, Sciblock: A blockchain-based tamper-proof non-repudiable storage for scientific workflow provenance, in: 2019 IEEE 5th International Conference on Collaboration and Internet Computing, CIC, IEEE, 2019, pp. 81–90.

[9] R. Hoopes, H. Hardy, M. Long, 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.

[10] M. Miller, S. Williams, G.G. Dagher, M. Long, PRISM: A blockchain-enabled reputation-based consensus for enhancing scientific workflow provenance, in: 2023 IEEE 9th International Conference on Collaboration and Internet Computing, CIC, IEEE, 2023, pp. 72–81.

[11] L. Gagliardi, W. Rocchia, SiteFerret: beyond simple pocket identification in proteins, J. Chem. Theory Comput. 19 (15) (2023) 5242–5259.

[12] K. Jamali, L. Käll, R. Zhang, A. Brown, D. Kimanius, S.H. Scheres, Automated model building and protein identification in cryo-EM maps, Nature 628 (8007) (2024) 450–457.
[13] Z. Zhang, Y. Cai, B. Zhang, W. Zheng, L. Freddolino, G. Zhang, X. Zhou, EM2: assembling protein complex structures from cryo-EM maps through intertwined chain and domain fitting, Brief. Bioinform. 25 (2) (2024).
[14] X.C. Ma, D. Si, Beyond current boundaries: Integrating deep learning and AlphaFold for enhanced protein structure prediction from low-resolution cryo-EM maps, Comput. Biol. Chem. (2025) 108494.
[15] M.L. Connolly, Solvent-accessible surfaces of proteins and nucleic acids, Science 221 (4612) (1983) 709–713.
[16] E. Kellenberger, P. Muller, C. Schalon, G. Bret, N. Foata, D. Rognan, sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank, J. Chem. Inf. Model. 46 (2) (2006) 717–727.
[17] W. Tian, C. Chen, X. Lei, J. Zhao, J. Liang, CASTp 3.0: computed atlas of surface topography of proteins, Nucleic Acids Res. 46 (W1) (2018) W363–W367.
[18] B. Mathis, J. Stine, A novel single/double precision normalized IEEE 754 floating-point adder/subtracter, in: 2019 IEEE Computer Society Annual Symposium on VLSI, ISVLSI, IEEE, 2019, pp. 278–283.
[19] D.-S. Kim, R. Syamsul, Integrating machine learning with proof-of-authority-and-association for dynamic signer selection in blockchain networks, ICT Express (2024).
[20] Y.-J. Su, C.-H. Chen, T.-Y. Chen, C.-W. Yeah, Applying ethereum blockchain and IPFS to construct a multi-party used-car trading and management system, ICT Express 10 (2) (2024) 306–311, http://dx.doi.org/10.1016/j.icte.2023.1650.
[21] V.I. Kanu, S.O. Ajakwe, J.M. Lee, 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.
[22] H. Bilal, F. Ahmed, M.S. Aslam, Q. Li, J. Hou, B. Yin, A blockchain-enabled approach for privacy-protected data sharing in Internet of Robotic Things networks, Humcent Comput. Inf. Sci. 14 (1) (2024) 71.