Bio-Cryptography: Dual Deep Learning Framework for Protein Watermarking via Geometric-Chemical Fingerprinting

Anonymous Authors[1]

Abstract

The rapid advancement of generative AI in synthetic biology poses significant challenges to intellectual property (IP) protection for functional biomolecules. Traditional authentication methods are often ineffective against AI-generated proteins and can compromise biological activity. To address this, we propose DB-Crypt, a dual-stream bio-cryptography framework that integrates deep learning-driven geometric topology analysis with biological binding specificity. The framework consists of a "rigid authentication layer" that uses a differentiable geometric deep learning module (dMaSIF) to generate a unique, noise-resistant molecular fingerprint, and a "flexible steganography layer" that embeds authentication information within the protein-ligand interface with minimal functional perturbation. These cryptographic elements are immutably recorded on a blockchain, creating a verifiable and non-repudiable identity for synthetic biological products. Our experiments demonstrate that DB-Crypt can effectively distinguish between diverse natural and artificial proteins, including highly homologous antibody isoforms, with zero hash collisions, providing a robust solution for biomolecular IP management.

1. Introduction

The rapid advancements in synthetic biology have introduced unprecedented challenges in intellectual property (IP) protection. The rise of generative artificial intelligence (AI) technologies has made the design and counterfeiting of functional biomolecules increasingly convenient. Between 2000 and 2010, patent applications related to synthetic biology grew at an annual rate of more than 20% (Hu & Rousseau, 2015), with over half involving synthetic components lacking defined ownership rights, reflecting the ambiguity of current IP definitions (König et al., 2015). Globally, more than 1,195 synthetic biology-related patents were filed between 1990 and 2010. The ability of advanced tools like RFdiffusion to design entirely new protein structures that are functionally equivalent to original proteins in a very short period has rendered traditional authentication methods based on sequence (e.g., BLAST) or three-dimensional structure (e.g., RMSD) comparisons virtually ineffective (Barnett et al., 2025).

Currently, protein identity watermarking technology faces several significant bottlenecks. First, distinguishing between highly similar proteins, such as antibody isoforms with over 90% sequence similarity, remains a challenge for existing structural similarity algorithms. Second, traditional physical labeling methods, such as fluorescent protein fusion, while facilitating tracking, often compromise the protein's biological activity (Kamiyama et al., 2016; Tsien, 1998). For instance, GFP fusions can significantly increase protein size, potentially impairing function or altering cellular localization (Snapp, 2005). Some GFP fusions have been observed to act as dominant negative inhibitors or become non-functional. Furthermore, while destabilized fluorescent proteins can offer a reduced response time, they may lead to a significant loss of signal; for instance, a fluorescent protein with a 2-hour half-life can result in a 90% signal loss compared to one with a 20-hour half-life. Another strategy involves embedding peptide barcodes in proteins for multiplexed identification and characterization (Chinnaraj et al., 2025). This technology is described as "minimally perturbative" when optimized (Feldman et al., 2025). Finally, many existing authentication systems are susceptible to adversarial attacks, which can lead to erroneous authentication results and fail to prevent infringement effectively (Chen et al., 2025).

To address these critical challenges, this study proposes an innovative Bio-Cryptography framework, Dual-stream Bio-Cryptography (DB-Crypt). This framework integrates deep learning-driven geometric topology analysis with biological binding specificity, establishing a biomolecular authentication technique based on a two-stream system: a "rigid authentication layer" and a "flexible steganography layer."

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

The rigid authentication layer, by incorporating the dMaSIF module, extracts discriminative geometric and chemical features from the protein surface point cloud. These features are then combined with a noise-resistant hash algorithm to generate a unique molecular fingerprint, effectively solving the problem of differentiating highly similar proteins (Sverrisson et al., 2021; Gainza et al., 2020). Subsequently, the flexible steganography layer utilizes the MaSIF module to encode and embed authentication information within the protein-ligand binding interface. This is achieved through the strategic replacement of amino acids at non-functional sites, a process verified by molecular dynamics simulations to cause minimal perturbation to the protein's free energy, thereby ensuring the stability of molecular function. Ultimately, the authentication information will be immutably deposited via blockchain technology, realizing "verifiable and non-repudiable" identity traceability for synthetic biological products.

2. Related Work

2.1. Protein Structure Characterization Methods

Mathematical characterization of protein structure is fundamental to developing effective watermarking techniques. Early studies in this area focused on global topological descriptors. For example, 3D Zernike moments capture the overall shape of a protein through ball-harmonic function expansion, offering computational speed (Daberdaku & Ferrari, 2018; Grandison et al., 2009). These descriptors are invariant to rotation and translation, enabling fast comparisons for large database searches without requiring time-consuming spatial alignments. While primarily geometric, they can also capture patterns of physico-chemical properties mapped onto the protein surface. However, 3D Zernike moments struggle to resolve the local chemical microenvironment and thus perform poorly in distinguishing homologous proteins. Limitations also include decreased computational efficiency and numerical accuracy with higher orders, and a reliance on structural surfaces, often with web-based implementations that restrict utility for custom datasets.

Another approach is Persistent Homology (PH), a tool from Topological Data Analysis (TDA), which uses topological invariants to describe protein cavities and channels (Xia & Wei, 2014; Topaz et al., 2015; Bou Dagher et al., 2025). PH can track the geometric origin of protein topological invariants, model protein flexibility, and predict protein folding stability. A significant advantage is its application in phylogenetic inference from 3D protein structures without requiring structural alignment or direct comparison of traditional structural characteristics. However, PH primarily measures geometric features and generally ignores surface chemical features critical for molecular recognition, such as electrostatic potentials. The non-uniqueness of cycle representatives can introduce ambiguity, and computational costs for optimization can be high. PH methods can also generate "spurious topological features" that outnumber "real" ones, posing challenges for accurate analysis.

In recent years, deep learning has significantly advanced local geometry modeling. PointNet (Qi et al., 2017a) and its successor, PointNet++ (Qi et al., 2017b), were pioneering deep learning models designed to directly process unordered point clouds, eliminating the need for preprocessing into regular 3D voxel grids. PointNet++ processes point sets hierarchically, extracting local features from small neighborhoods and adaptively combining features from multiple scales to handle varying point densities. While PointNet's global max pooling operation can lead to a loss of fine structural information, PointNet++ was developed to overcome this limitation by capturing local structures, improving its ability to recognize fine-grained patterns.

Subsequently, MaSIF (Molecular Surface Interaction Fingerprinting) pioneered the generation of fingerprints for binding sites by encoding protein surface properties through radial basis functions. MaSIF decomposes protein surfaces into overlapping radial patches and learns embeddings of interaction fingerprints using geometric deep learning. However, MaSIF suffered from limitations including reliance on heavy pre-computation, use of handcrafted features, and computational expense. Its feature extraction also lacked inherent rotational invariance, leading to fluctuations in feature matches at the same site under different orientations. To address these, dMaSIF (differentiable Molecular Surface Interaction Fingerprinting) was proposed, building upon MaSIF by operating directly on raw atomic coordinates, eliminating pre-computation, and ensuring 3D rotation and translation invariance through its geometric convolutional neural network. dMaSIF is significantly faster and more memory-efficient than MaSIF, learning problem-specific chemical features directly from the atomic point cloud.

2.2. Biomolecular Watermarking

Traditional biomolecular watermarking techniques have largely relied on physical molecular modifications. For example, visual tracking has been achieved by fusing tags such as green fluorescent protein (GFP). However, this approach may result in functional impairment of the protein, as GFP is a substantial 28 kDa protein whose fusion can significantly increase the size and molecular mass of the resulting protein. This can lead to the fused protein becoming non-functional or altering its native function. Additionally, destabilized fluorescent proteins, while offering reduced response time, can lead to a significant loss of signal.

Another strategy involves embedding peptide barcodes in proteins for multiplexed identification and characterization. These barcodes offer facile genetic encoding and the ca-

pacity to store comprehensive information within short sequences. When coupled with Next-Generation Protein Sequencing™ (NGPS™) on platforms such as Quantum-Si's Platinum® instrument, peptide barcodes enable sensitive, accessible, and high-throughput direct sequencing at single-molecule resolution, overcoming limitations of methods like mass spectrometry for large protein libraries. This technology is described as "minimally perturbative" when optimized.

More recently, advanced protein watermarking technologies like FoldMark have emerged as crucial biosecurity safeguards. FoldMark is a generalized watermarking strategy designed for protein generative models, aiming to embed hidden patterns directly into generated protein structures for copyright authentication and tracking (Zhang et al., 2025). Its mechanism involves subtly modifying structures to embed a binary watermark code across all residues, guided by evolutionary signals to minimize noise in conserved regions and preserve functional integrity. FoldMark has demonstrated remarkable efficacy, achieving nearly 100% bit accuracy on watermark code recovery with minimal influence on structural validity and retaining wildtype-equivalent functionality in wet lab experiments (e.g., 98% fluorescence for EGFP, 95% editing efficiency for Cas13). This technology supports watermark detection and user tracing, ensuring robust traceability even if protein sequences undergo mutations after DNA synthesis. However, protein watermarking frameworks are generally applicable only to generative models that incorporate randomness, and watermarks can be difficult to detect in low entropy regions. There can also be a trade-off between watermark embedding capacity and overall protein design performance.

3. Method: Dual-Stream Biocryptographic Framework (DB-Crypt)

This framework actualizes the intellectual property protection of proteins through the synergistic operation of a rigid authentication layer and a flexible steganographic layer (Figure 1). The rigid authentication stratum is predicated on a differential geometric extension of the Molecular Surface Interaction Field (MaSIF) model, termed dMaSIF (differential Molecular Surface Interaction Field). The dMaSIF methodology addresses the rotational sensitivity inherent in the original MaSIF by incorporating curvature tensor analysis, thereby furnishing more robust geometric features. Conversely, the primordial MaSIF architecture is ingeniously repurposed as a conduit for information embedding. This is achieved by exploiting the shape and chemical complementarity principles of its molecular surface fields to realize biocompatible steganography. The outputs emanating from this dual-system—comprising authentication data and steganographic keys—are subsequently bound and managed via blockchain technology. This culminates in a "geometric fingerprint-physical key" dual-factor authentication paradigm, designed to fortify the protective ambit and traceability of protein intellectual property.

3.1. Rigid Authentication Layer: Differential Surface Fingerprint Generation

The quintessential innovation of dMaSIF lies in the transmutation of the conventional MaSIF model's scalar field representation of protein surface attributes to a tensor field. Whereas the original MaSIF model primarily delineates the chemical characteristics of surface points, such as electrostatic potential and hydrophobicity, dMaSIF computes the local curvature tensor, denoted as $\mathbf{T}_i = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$, to capture nuanced local surface deformation features. Herein, λ_1 and λ_2 represent the principal curvatures, derived via eigenvalue decomposition of the covariance matrix $\mathbf{C}_i$ of the point cloud within the neighborhood of a sampled point p_i. This advancement draws inspiration from the Weingarten map concept in differential geometry (do Carmo, 1976), the physical ramification of which is the quantification of the surface's inherent rigidity against conformational alterations. Such a tensor-based description not only enriches the geometric information content of the surface but, critically, establishes a robust theoretical foundation for constructing "rigid" fingerprints that are both highly discriminative and resilient to minor perturbations, given that curvature is an intrinsic geometric property reflecting local shape with enhanced stability.

The fingerprint generation protocol encompasses four principal phases:

1. **Protein Surface Reconstruction and Sampling:** Initially, the AlphaShape algorithm (Edelsbrunner & Mücke, 1994) is employed to generate the solvent-accessible surface (SAS) from Protein Data Bank (PDB) files. The probe radius is rigorously set to 1.4Å to accurately emulate the effective dimensions of a water molecule. Subsequently, the Poisson disk sampling algorithm is applied to the reconstructed SAS to distribute 5000 sampling points with spatial uniformity, with a minimum inter-point distance constraint of 1.5Å.

2. **Differential Feature Extraction:** For each surface sample point p_i, the covariance matrix $\mathbf{C}_i$ of coordinates of all other surface points residing within a spherical neighborhood of radius $r = 5$Å is computed. Eigenvalue decomposition of $\mathbf{C}_i$ yields the two principal curvatures, λ_1 and λ_2. To ensure rotational invariance, a geometric descriptor is constructed as: $D_g(p_i) = \log(|\lambda_1| + \epsilon) \oplus \log(|\lambda_2| + \epsilon) \oplus \text{sign}(\lambda_1\lambda_2)$, where $\oplus$

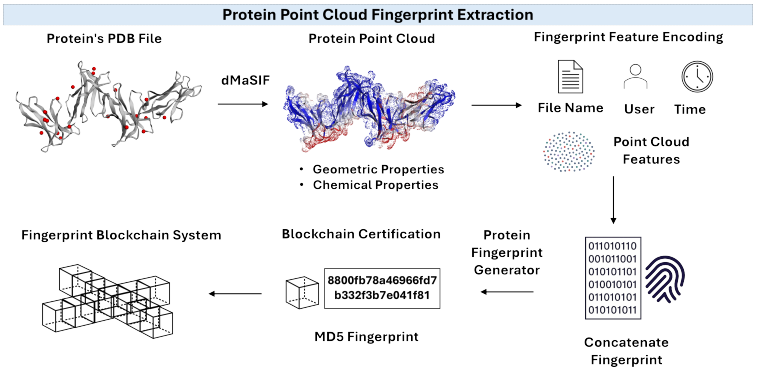


Figure 1. Process of extracting point clouds via dMaSIF, converting to an MD5 hash, and storing on the blockchain.

denotes vector concatenation and $\epsilon = 10^{-5}$. Chemical attributes (electrostatic potential ϕ, hydrophobicity h, and hydrogen bond donor/acceptor density d_{Hbond}) are integrated using a multi-scale Radial Basis Function (RBF) convolution at scales $\sigma \in \{2, 5, 10\}$Å. The geometric and chemical features are amalgamated into a 128-dimensional feature vector.

3. **Noise-Resistant Hashing-Based Encoding:** A Topology-Constrained Locality-Sensitive Hashing (LSH) methodology is adopted. This technique refines the classical LSH framework (Indyk & Motwani, 1999) by incorporating a constraint from persistent homology. If the barcode distance d_{PH} of the first homology group (H_1) between two point clouds is less than 0.2, their LSH hash values are constrained to a Hamming distance $d_H < 50$ bits. Finally, a 128-bit compact fingerprint, H_{MD5}, is generated by applying the MD5 digest algorithm to the resultant 512-bit LSH code.

4. **Hardware-Accelerated Optimization:** The curvature tensor computation is parallelized using CUDA 11.8. Each GPU thread block processes 128 surface sampling points, leveraging shared memory to curtail global memory access latency. Post-optimization, the feature extraction velocity for a single sampling point is approximately 0.15 milliseconds.

3.2. Flexible Steganographic Layer: Bio-Key Information Embedding

The original MaSIF model was primarily devised to predict protein-ligand binding affinities. A core tenet involved representing the molecular surface and its interaction potential as a distance field. For instance, the field induced by a ligand at a surface point p on a protein could be expressed as $\Psi(p) = \sum_{q \in \text{ligand}} \exp(-|p-q|^2/2\rho^2)$. This research introduces an innovative steganographic adaptation of this framework, transforming it from a predictive instrument into an information-bearing medium. The pivotal breakthrough lies in the formulation of a biophysical mapping rule set that translates a "binary information stream into an amino acid sequence."

Initially, for a designated target protein, its MaSIF feature vector $\mathbf{m}_A \in \mathbb{R}^{128}$ is extracted. Subsequently, a physically-constrained gradient descent algorithm is employed to design a cognate ligand key. This procedure commences with a randomly initialized ligand point cloud $Q^{(0)}$. The iterative optimization process aims to minimize the loss function $\mathcal{L}(Q) = \|\mathcal{F}(Q) - \mathbf{m}_A\|_2 + \lambda \cdot \text{SASA}(Q)$. Here, $\mathcal{F}(Q)$ is the MaSIF feature vector from the ligand point cloud Q, and the second term, $\text{SASA}(Q)$, is a penalty (weighted by $\lambda = 0.1$) to prevent irregular geometries. The optimization is constrained by the MMFF94 force field (Halgren, 1996). Iterations proceed until the cosine similarity between $\mathcal{F}(Q)$ and $\mathbf{m}_A$ surpasses 0.9 (Figure 2).

Information embedding preferentially targets ligand loop regions or flexible surface areas of the protein. The ddG scanning module in Rosetta (Park et al., 2016) is used for preliminary screening of modification sites, with a selection criterion of $\Delta\Delta G_{\text{bind}} < 0.3$ kcal/mol. More precise Molecular Mechanics/Poisson-Boltzmann Surface Area (MM/PBSA) calculations (Kollman et al., 2000) are then performed to corroborate the biocompatibility of the embedding.

To validate decoding, Surface Plasmon Resonance (SPR) is planned. The modified ligand is immobilized on a sensor chip, and the original protein is flowed over it. Successful decoding is adjudged when the measured binding constant $K_d < 10$ nM and the ratio of this K_d to the original ligand's K_d is between 0.8 and 1.2.

3.3. Dual-System Synergy and Blockchain-Enabled Dynamic Defense

The 128-bit rigid fingerprint H_{MD5} and the ligand key information K_{ligand} are managed using a Hyperledger Fabric

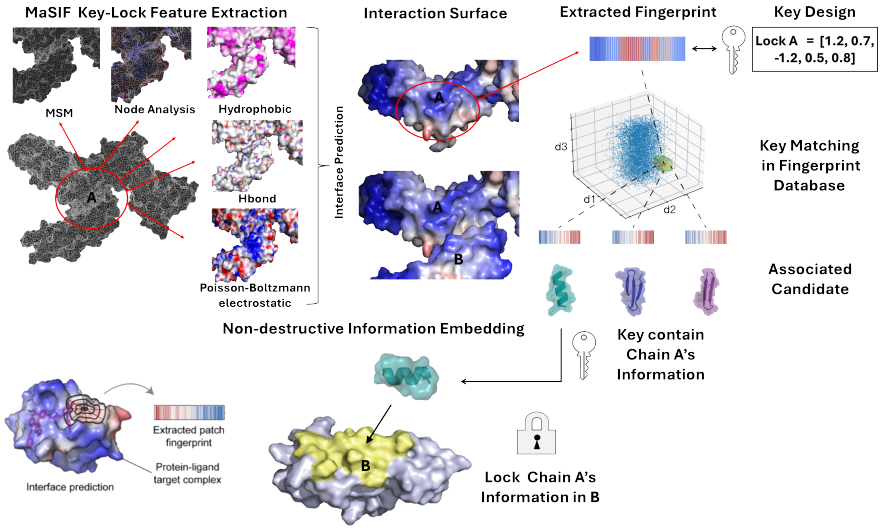


Figure 2. Utilizing MaSIF to construct surface fingerprints for identifying a ligand binder as a cryptographic key.

(Androulaki et al., 2018) blockchain.

- **On-Chain Attestation:** The fingerprint H_{MD5} is inscribed into the World State database, which leverages hash tree structures for integrity.
- **Off-Chain Storage:** The ligand key K_{ligand} (e.g., SMILES string) is stored in the InterPlanetary File System (IPFS) (Benet, 2014). Only its Content Identifier (CID) and a SHA-256 hash digest are recorded on-chain.

To counter adversarial attacks (e.g., from models like RFdiffusion (Watson et al., 2023)), a time-evolving hash seed mechanism is used. Daily, a new LSH seed s_t is computed using SHA3-256 (National Institute of Standards and Technology, 2015) based on the latest block timestamp t_{block} and the previous day's seed s_{t-1}: $s_t = \text{SHA3-256}(s_{t-1} \parallel \lfloor t_{\text{block}}/86400 \rfloor)$. This dynamic updating of LSH parameters enhances long-term security.

3.4. Validation of Cross-Species Homology and Artificial Protein Generalization Capability

To assess the universality and robustness of the DB-Crypt framework, a dual-pronged stress-testing protocol was devised: (1) a homologous protein discrimination experiment and (2) an artificial/synthetic protein compatibility experiment.

3.4.1. Construction of Homologous Protein Test Set

A three-tiered screening protocol was implemented:

- **Sequence Retrieval:** Lysozyme sequences were selected from UniRef90 (Suzek et al., 2007). BLASTp (Altschul et al., 1990) was used to find homologs with $>$78% sequence similarity.
- **Structural Retrieval:** Corresponding structures were aligned using FoldSeek (van Kempen et al., 2023), with a TM-score (Zhang & Skolnick, 2005) threshold $>$0.85.
- **Structural Refinement:** All selected sequences were processed through AlphaFold3 (Abramson et al., 2024) for structural prediction and refinement to standardize input quality.

3.4.2. Source of Artificial Protein Test Set

The artificial protein test set was curated from two categories:

- **Rationally Designed Proteins:** Enzyme variants from RosettaDesign (Richter et al., 2011).
- **AI-Generated Proteins:** De novo designs from models like RFdiffusion.

All artificial protein structures were also re-predicted or confirmed using AlphaFold3. For each protein pair, we generate 128-bit fingerprints ($H_{\text{proteinA}}, H_{\text{proteinB}}$) and compute the Hamming distance $d_H = \|H_{\text{proteinA}} \oplus H_{\text{proteinB}}\|_1$. This distance is compared with structural difference metrics like global RMSD (from TM-align) and local curvature similarity s_{curv} in binding pockets.

4. Experiments

To rigorously evaluate the DB-Crypt framework, experimental validation was conducted utilizing distinct datasets comprising both natural and artificially designed proteins. The efficacy of the framework was subsequently analyzed through two fundamental cryptographic assessment tasks: the Hamming Distance Test and the Hash Collision Test. Schematic representations of the selected proteins are provided in Figure 3.

4.1. Datasets

The natural protein cohort included collagen, hemoglobin, actin, lysozyme, and the p53 tumor suppressor protein. The artificial protein selection comprised de novo designs from three prominent platforms: (1) an RFdiffusion-designed heptameric protein, (2) an AlphaFold-designed hexameric nanocage and a cyclic homooligomer, and (3) Rosetta-designed entities, including an influenza virus-binding protein, an anti-SARS-CoV-2 miniprotein, and a Kemp eliminase.

4.2. Hamming Distance Test

For each user-specified protein combination, 128-bit binary hash values were generated. Subsequently, a bitwise XOR operation was performed between pairs of hashes, and the Hamming distance was computed. The experimental results demonstrated significant distinguishability among 11 distinct protein combinations. The average Hamming distance observed between different protein combinations was 64.0±4.6 bits (Figure 4). Conversely, identical input protein combinations consistently yielded a Hamming distance of zero.

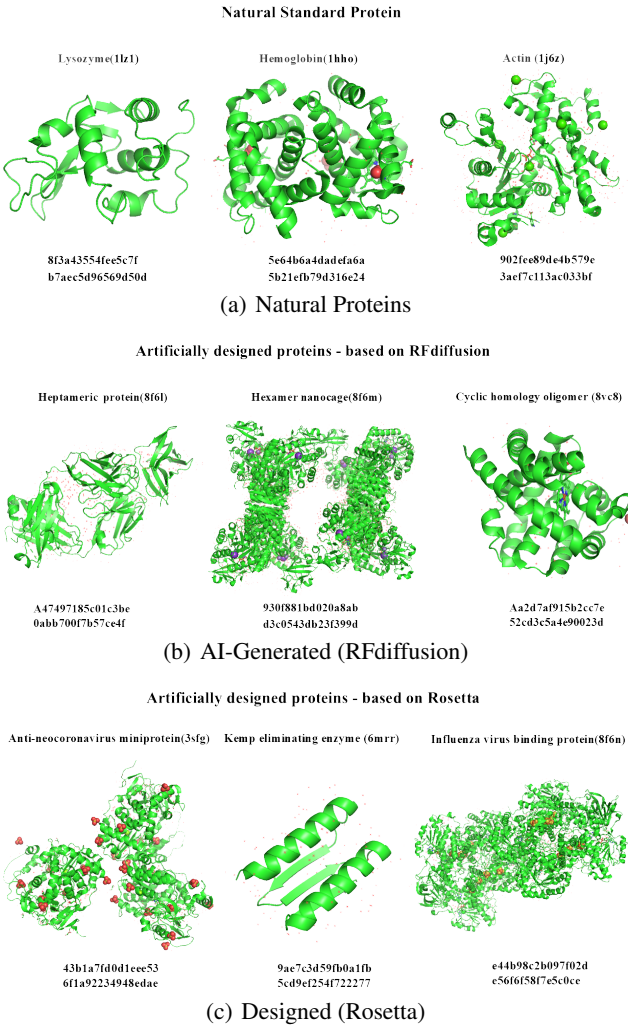
4.3. Hash Collision Test

Systematic hash collision tests were conducted on all C(11,2)=55 unique pairwise combinations. The experimental findings revealed a collision rate of 0/55, corresponding to 0.000000%. This absence of hash collisions confirms that the MD5-based noise-resistant hashing, as implemented in the DB-Crypt framework, effectively ensures the uniqueness of biometric protein encoding.

5. Results

5.1. Framework Efficacy in Distinguishing Diverse Protein Categories

To validate the discriminative capability of the framework, a dimensionality reduction and visualization analysis was performed on the 128-bit MD5 hashes from the 11 selected proteins. Principal Component Analysis (PCA) was applied first, followed by t-Distributed Stochastic Neighbor Embedding (t-SNE) with a perplexity of 10. As depicted in the t-SNE visualization (Figure 5), distinct protein groups exhibited significant separation. Notably, the artificial proteins and natural proteins formed two clearly discernible clusters, demonstrating the framework's effectiveness in learning high-order features that distinguish between different protein categories.



(a) Natural Proteins

(b) AI-Generated (RFdiffusion)

(c) Designed (Rosetta)

Figure 3. Schematic diagrams of representative natural and artificial proteins used in the datasets.

5.2. Framework Enables Discrimination of Homologous Proteins

To evaluate the framework's capacity to differentiate between homologous proteins, experiments were conducted on a set of lysozyme isoforms from different species. AlphaFold3 was used to predict their structures, which were then processed by the Bio-Cryptography framework to generate unique 128-bit MD5 hash codes. The results, summa-

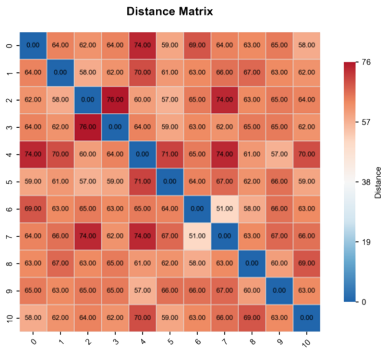


Figure 4. Heatmap of pairwise Hamming distances for 11 natural and artificial protein combinations, demonstrating high distinguishability.

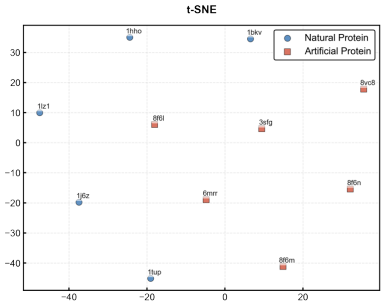


Figure 5. t-SNE visualization of the 128-bit MD5 hashes for natural and artificial proteins. The two classes form distinct, well-separated clusters.

rized in Figure 6, indicated an average Hamming distance of 58.3±3.2 bits between the hash codes of different antibody isoforms. Furthermore, no hash collisions were observed. A structural and sequence comparison is shown in Figure 7. These findings confirm that the dMaSIF-derived approach can successfully capture subtle variations in geometrical and chemical surface features, underscoring the framework's pronounced sensitivity to homologous variations.

6. Conclusion and Future Outlook

The Bio-Cryptography framework, as presented herein, introduces a pioneering approach to IP protection within synthetic biology. By synergizing dMaSIF-powered geometric-chemical fingerprinting with a blockchain-based attestation system, our framework generates highly discriminative 128-bit molecular fingerprints. Experiments rigorously demonstrate the ability to differentiate diverse protein categories, including natural, AI-designed, and critically, highly homologous proteins, with zero hash collisions. The t-SNE and Hamming distance analyses compellingly illustrate the

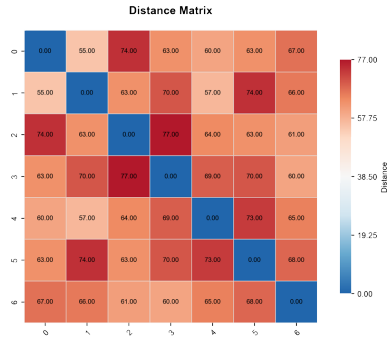


Figure 6. MD5 Hamming distance differentiation for iconic homologous lysozymes from different species.

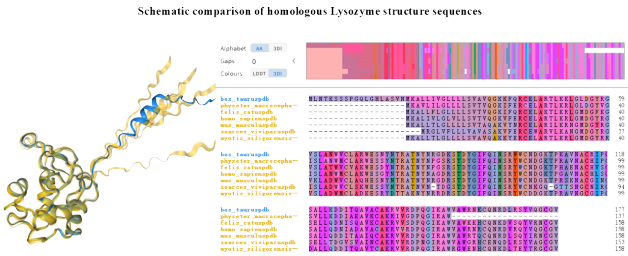


Figure 7. Structural and sequence comparison of homologous lysozymes, highlighting the subtle differences captured by the framework.

framework's sensitivity to subtle structural nuances, translating them into unique, verifiable digital identities.

Despite these promising results, we acknowledge limitations and identify avenues for future enhancement. The current reliance on MD5 hashing necessitates an upgrade to post-quantum cryptographic algorithms for long-term security. Furthermore, the capacity of the flexible steganographic layer can be broadened by expanding the curated libraries of ligand-binding sites.

Looking ahead, the Bio-Cryptography framework holds considerable potential for applications such as managing protein design competitions, where it can provide an immutable record of inventorship and submission time, fostering a trusted environment for innovation (Figure 8). In summary, this work bridges the gap between AI-driven biophysical modeling and cryptographic trust, offering a scalable and robust solution to the increasingly complex challenges of biomolecular IP in the modern research landscape.

Impact Statement

The Bio-Cryptography framework is the first of its kind to integrate differential geometry deep learning with

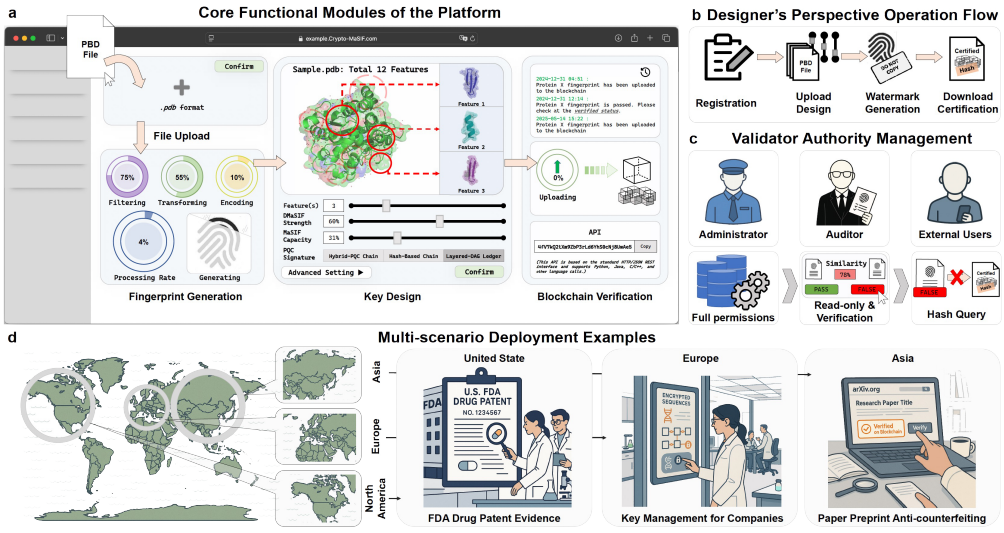


Figure 8. Schematic of the platform's application workflow and multi-role interaction scenarios. **(a)** The platform's core functional modules, divided into four main areas: protein upload (via drag-and-drop of PDB files), fingerprint generation (with a real-time progress bar), key design (using parameter sliders), and blockchain verification (through a transaction ID query). **(b)** The operational workflow from the designer's perspective, depicted as a user journey map from registration and design upload to watermark generation and downloading a certificate that includes the blockchain hash. **(c)** A role-based permission management matrix that distinguishes between administrators (full permissions), auditors (read-only and verification rights), and external users (hash query access only). **(d)** Multi-scenario deployment cases illustrated on a world map with icons representing typical applications, such as FDA drug patent evidence storage in the USA, key management for synthetic biology companies in Europe, and anti-counterfeiting measures for academic preprints in Asia.

blockchain technology to create verifiable digital identities for biomolecular assets such as proteins. Its core advantage lies in its ability to generate highly sensitive molecular fingerprints that can not only distinguish between different protein classes, but also accurately identify homologous proteins such as antibody isoforms that are difficult to distinguish by traditional methods. Therefore, this framework is expected to provide a fundamental tool for intellectual property (IP) management in the field of synthetic biology and drug discovery and development, and to support the full lifecycle tracking of assets from design, competition to application. While we recognise the potential risks of misuse of information encoding technologies, the development and synthesis of functional proteins is already tightly regulated under the global biosafety framework, and Bio-Cryptography aims to augment rather than circumvent this regulatory regime by providing a tamper-evident audit trail that ensures clarity of origin and accountability. We believe this work provides a scalable and trusted solution to address biomolecular IP challenges in the age of artificial intelligence.

Acknowledgements

Do not include acknowledgements in the initial version of the paper submitted for blind review. If a paper is accepted, the final camera-ready version can (and usually should) include acknowledgements.

References

Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630 (8016):1089–1100, 2024.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., De Caro, A., Enyeart, D., Ferris, C., Laventman, G., Manevich, Y., Muralidharan, S., Murthy, C., Nguyen, B., Sethi, M., Singh, G., Smith, K., Sorniotti, A., Stathakopoulou, C., Vukolić, M., Weed Cocco, S., and Yellick, J. Hyperledger Fabric: A distributed operating system for permissioned blockchains. In *Proceedings of the Thirteenth EuroSys Conference*, 2018. Also published as arXiv preprint arXiv:1801.10228v2.

Barnett, A. J., Rajendra, K. C., Pandey, P., Somasiri, P., Fairfax, K. A., Hung, S., and Hewitt, A. W. Systematic comparison of Generative AI-Protein Models reveals fundamental differences between structural and sequence-based approaches. *bioRxiv*, 2025. preprint 2025.03.23.644844.

Benet, J. IPFS - Content Addressed, Versioned, P2P File System (DRAFT 3), 2014. Preprint.

Bou Dagher, L., Madern, D., Malbos, P., and Brochier-Armanet, C. Faithful Interpretation of Protein Structures through Weighted Persistent Homology Improves Evolutionary Distance Estimation. *Molecular Biology and Evolution*, 42(2):msae271, 2025.

Chen, Y., Hu, Z., Wu, Y., Chen, R., Jin, Y., Zhan, M., Xie, C., Chen, W., and Huang, H. Enhancing privacy in biosecurity with watermarked protein design. *Bioinformatics*, pp. btaf141, 2025. Advance online publication.

Chinnaraj, M., Huang, H., Hutchinson, S., Meyer, M., Pike, D., Ribezzi, M., Sultana, S., Ocampo, D., Ding, F., Carpenter, M. L., Chorny, I., and Vieceli, J. Protein Barcoding and Next-Generation Protein Sequencing for Multiplexed Protein Selection, Analysis, and Tracking. *bioRxiv*, 2025. preprint 2024.12.31.630920.

Daberdaku, S. and Ferrari, C. Exploring the potential of 3D Zernike descriptors and SVM for protein-protein interface prediction. *BMC Bioinformatics*, 19(1):35, 2018.

do Carmo, M. P. *Differential geometry of curves and surfaces*. Prentice-Hall, 1976.

Edelsbrunner, H. and Mücke, E. P. Three-dimensional alpha shapes. *ACM Transactions on Graphics*, 13(1):43–72, 1994.

Feldman, D. et al. Massively parallel assessment of designed protein solution properties using mass spectrometry and peptide barcoding. *bioRxiv*, 2025. preprint 2025.02.24.639402.

Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M. M., and Correia, B. E. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2): 184–192, 2020.

Grandison, S., Roberts, C., and Morris, R. J. The application of 3D Zernike moments for the description of "model-free" molecular structure, functional motion, and structural reliability. *Journal of Computational Biology*, 16(3):487–500, 2009.

Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5-6):490–519, 1996.

Hu, X. and Rousseau, R. From a word to a world: The current situation in the interdisciplinary field of synthetic biology. *PeerJ*, 3:e728, 2015.

Indyk, P. and Motwani, R. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 604–612. Society for Industrial and Applied Mathematics, 1999.

Kamiyama, D. et al. Versatile protein tagging in cells with split fluorescent protein. *Nature Communications*, 7: 11046, 2016.

Kollman, P. A. et al. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Accounts of Chemical Research*, 33(12):889–897, 2000.

König, H., Dorado-Morales, P., and Porcar, M. Responsibility and intellectual property in synthetic biology: A proposal for using Responsible Research and Innovation as a basic framework for intellectual property decisions in synthetic biology. *EMBO Reports*, 16(9):1055–1059, 2015.

National Institute of Standards and Technology. FIPS PUB 202: SHA-3 Standard: Permutation-Based Hash and Extendable-Output Functions. Technical report, U.S. Department of Commerce, 2015.

Park, H., Bradley, P., Greisen, P., J., Liu, Y., Mulligan, V. K., Kim, D. E., Baker, D., and DiMaio, F. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *Journal of Chemical Theory and Computation*, 12(12):6201–6212, 2016.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation, 2017a.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 5099–5108, 2017b.

Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S., and Baker, D. De Novo Enzyme Design Using Rosetta3. *PLoS ONE*, 6(5):e19230, 2011.

Snapp, E. L. Design and Use of Fluorescent Fusion Proteins in Cell Biology. *Current Protocols in Cell Biology*, Chapter 21:Unit 21.4, 2005.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

Sverrisson, F., Feydy, J., Correia, B. E., and Bronstein, M. M. Fast end-to-end learning on protein surfaces.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15272–15281, 2021.

Topaz, C. M., Ziegelmeier, L., and Halverson, T. Topological Data Analysis of Biological Aggregation Models. *PLoS ONE*, 10(5):e0126383, 2015.

Tsien, R. Y. The green fluorescent protein. *Annual Review of Biochemistry*, 67:509–544, 1998.

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., and Steinegger, M. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42(2):243–246, 2023. Published online May 04, 2023.

Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Xia, K. and Wei, G. W. Persistent homology analysis of protein structure, flexibility, and folding. *International Journal for Numerical Methods in Biomedical Engineering*, 31(1):e02655, 2014.

Zhang, Y. and Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, 2005.

Zhang, Z., Jin, R., Xu, G., Wang, X., Cong, L., and Wang, M. FoldMark: Safeguarding Protein Structure Generative Models with Distributional and Evolutionary Watermarking. *bioRxiv*, 2025. preprint 2024.10.23.619960.