StrucTrace: Fourier Watermarking for Traceable Bio-molecular Assets

Xu Wang[1] **Tin-Yeh Huang**[1,2,3] **Yiquan Wang**[1,4] **Yafei Yuan**[1,*]

[1]Beijing Frontier Research Center for Biological Structure, Tsinghua University
[2]Xinya College, Tsinghua University
[3]Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University
[4]College of Mathematics and System Science, Xinjiang University

[*]Correspondence: yuanyf@tsinghua.edu.cn

Abstract

The rise of generative artificial intelligence (GenAI) in protein and nucleic acid design has created unprecedented opportunities for synthetic biology, but also heightened the need for reliable provenance and intellectual-property protection. To meet this challenge, we present a Fourier domain watermarking framework that encodes digital identifiers directly into three-dimensional biomolecular structures. By perturbing only flexible backbone atoms and embedding information through frequency domain modulation, the method achieves imperceptible alterations while ensuring deterministic and reversible decoding. Large-scale validation on over 40,000 protein structures demonstrates its robustness: structural deviations remain orders of magnitude below biological thresholds, watermarks are recovered with perfect accuracy, and functional analyses confirm stability at both thermodynamic and dynamic levels. Beyond technical performance, the approach provides a foundation for a broader ecosystem of secure biomolecular asset management, integrating provenance verification, access control, and digital rights management. Together, these advances establish biomolecules as traceable and auditable digital assets, aligning the future of bio-design with emerging standards for trustworthy AI.

Keywords: Bio-molecular Watermarking, Protein Design, Molecular Provenance

1 Introduction

Generative artificial intelligence (GenAI) is accelerating discovery in synthetic biology by enabling the rapid creation of novel proteins and nucleic acids [1–4]. However, it also presents a critical challenge: ensuring traceability. As these molecules become structurally indistinguishable from their natural counterparts, verifying their origin is essential for intellectual-property attribution [5, 6], regulatory compliance [7], and biosecurity [8–10].

While combining GenAI with biomacromolecule watermarking is a promising research direction, current methods face several fundamental limitations. First, their applicability is restricted, as they are often tightly coupled to specific generative models [11–13] and cannot be used on experimentally derived structures. Second, their ”black-box” nature hinders interpretability, preventing a transparent assessment of any structural perturbations. These pose a practical challenge to collaboration, as the verification process often requires full molecular disclosure, which conflicts with data privacy requirements [14, 15].

To overcome these challenges, we introduce StrucTrance (**Struc**ture **Trace**), a watermarking framework that directly embeds digital identifiers into 3D biomolecular structures. By applying a Fourier transform to atomic coordinates, the method creates a deterministic geometric encoding, confining perturbations to targeted regions to preserve functional integrity. This approach ensures unambiguous watermark retrieval, achieving 100% bit accuracy while reducing structural deviation 300-fold (scRMSD) compared to state-of-the-art methods. StrucTrance thus establishes a robust foundation for auditable and privacy-preserving biomolecular design, in line with emerging trustworthy AI standards.

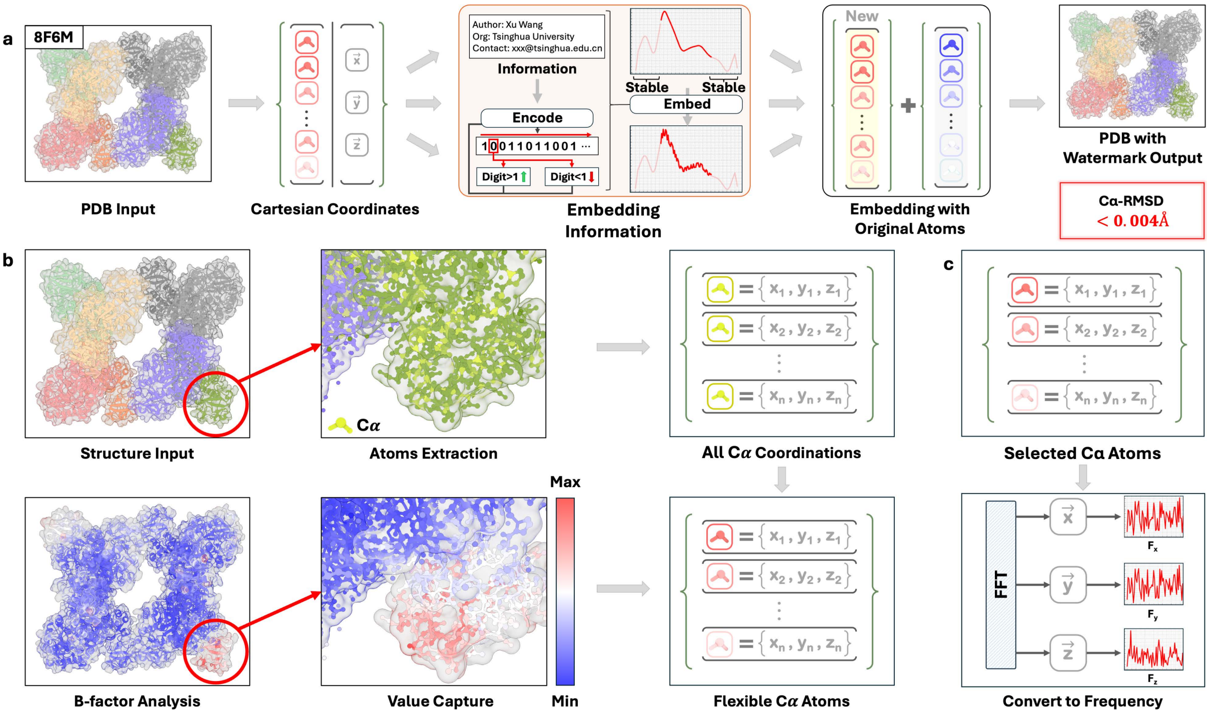

Figure 1. Workflow for frequency-domain digital watermarking of protein structures. (a) The overall watermarking pipeline. Starting with an input PDB file, the 3D coordinates of target C_α atoms are extracted. These coordinates are then transformed into the frequency domain using a discrete Fourier transform (DFT) for watermark embedding. This frequency range is optimal as it preserves the low-frequency signals that define the global fold while offering greater robustness to noise than high-frequency signals. Phase information is preserved and conjugate symmetry is enforced to ensure the inverse transform yields real-valued coordinates. Then, an inverse DFT (IDFT) converts the frequency-domain signal back into real-valued spatial coordinates, generating a new PDB file that contains the watermark while preserving all original metadata. **(b) Target atom selection based on B-factors.** To preserve the protein's biological function, perturbations are confined to backbone C_α atoms in structurally flexible regions. These atoms are identified by their high B-factor values in the source PDB file. Their inherent mobility allows for minor coordinate shifts to be tolerated without compromising the protein's global fold or the conformation of its active sites. **(c) Watermark encoding via Fast Fourier Transform (FFT).** The coordinate vectors (x, y, z) of the selected C_α atoms are transformed into the frequency domain. A binary watermark is embedded by modulating the amplitudes of specific mid-range frequency components.

2 Methods

We introduce a frequency-domain method to embed digital watermarks into 3D protein structures by applying imperceptible coordinate perturbations to structurally flexible C_α atoms. This process, designed to preserve biological function, is illustrated in Figure 1.

The watermarking algorithm proceeds as follows:

2.1 Target Atom Selection

We identify C_α atoms in structurally flexible regions based on their B-factor values from the input PDB file. High B-factors indicate thermal mobility, allowing coordinate perturbations to be tolerated without disrupting the global fold or active site geometry.

2.2 Frequency Domain Transformation

The coordinate vectors (x, y, z) of selected atoms are independently transformed using the Discrete Fourier Transform (DFT):

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-i2\pi kn/N}, \tag{2.1}$$

where N is the number of selected atoms, $x[n]$ represents the spatial coordinates, and $X[k]$ denotes the frequency-domain representation.

Table 1. Cross-Method Comparison of Bit Accuracy and Structural Deviation

Embedding Method		4bit		8bit		16bit		32bit	
		BitAcc ↑	scRMSD ↓	BitAcc ↑	scRMSD ↓	BitAcc ↑	scRMSD ↓	BitAcc ↑	scRMSD ↓
FoldFlow	No watermark	-	1.926	-	1.926	-	1.926	-	1.926
	WaDiff[12]	88.2%	2.107	85.0%	2.115	80.1%	2.397	64.3%	2.630
	AquaLoRA[13]	73.5%	2.056	72.6%	2.210	71.7%	2.446	62.8%	2.718
	FoldMark[11]	99.9%	1.937	99.7%	1.980	98.9%	2.114	94.5%	2.307
FrameDiff	No watermark	-	2.850	-	2.850	-	2.850	-	2.850
	WaDiff	76.8%	2.919	73.3%	3.235	62.2%	3.810	50.4%	4.058
	AquaLoRA	64.3%	3.150	59.1%	3.431	56.2%	3.890	51.6%	4.179
	FoldMark	98.7%	2.795	98.3%	2.914	88.4%	3.045	82.0%	3.428
FrameFlow	No watermark	-	1.855	-	1.855	-	1.855	-	1.855
	WaDiff	77.1%	1.883	76.4%	2.270	63.5%	2.456	54.6%	2.823
	AquaLoRA	63.6%	1.920	61.4%	2.317	54.5%	2.680	52.1%	2.953
	FoldMark	99.6%	1.860	99.5%	1.939	96.7%	2.019	95.4%	2.192
RCSB PDB (N>40000)	Our Method	**100%**	**0.0003**	**100%**	**0.0004**	**100%**	**0.001**	**100%**	**0.0015**

2.3 Watermark Embedding

A binary watermark $W = \{w_1, w_2, \ldots, w_M\}$ is embedded by modulating the amplitudes of mid-range frequency components. Specifically, for each bit w_j, we modify the amplitude at frequency index k_j:

$$|X'[k_j]| = |X[k_j]| + \alpha \cdot w_j, \tag{2.2}$$

where α is a scaling factor controlling embedding strength. Phase information is preserved to ensure the inverse transform yields real-valued coordinates. Conjugate symmetry is enforced to maintain the real-valued property of spatial coordinates.

2.4 Inverse Transformation

The modified frequency-domain signal is converted back to spatial coordinates via the Inverse DFT (IDFT):

$$x'[n] = \frac{1}{N} \sum_{k=0}^{N-1} X'[k] \cdot e^{i2\pi kn/N}. \tag{2.3}$$

The resulting coordinates $x'[n]$ form the watermarked structure, which is exported as a new PDB file with all original metadata preserved.

2.5 Watermark Decoding

The decoding process is deterministic and does not require access to the original structure. We re-select the same C_α atoms using B-factor criteria, apply DFT, and extract the watermark by analyzing amplitude patterns at the targeted frequency indices. Bit recovery is achieved by thresholding amplitude differences against expected embedding patterns.

3 Results

We validated our watermarking method across over 40,000 RCSB PDB structures, achieving three key breakthroughs (Tab. 1).

First, ultra-low structural perturbation was demonstrated with scRMSD $\leq$ 0.0015 Å, significantly lower than the structural biology threshold of 2.0 Å for indistinguishable folds. This ensures watermarked proteins remain visually and computationally identical to their originals, as confirmed by alignment tools and manual inspection.

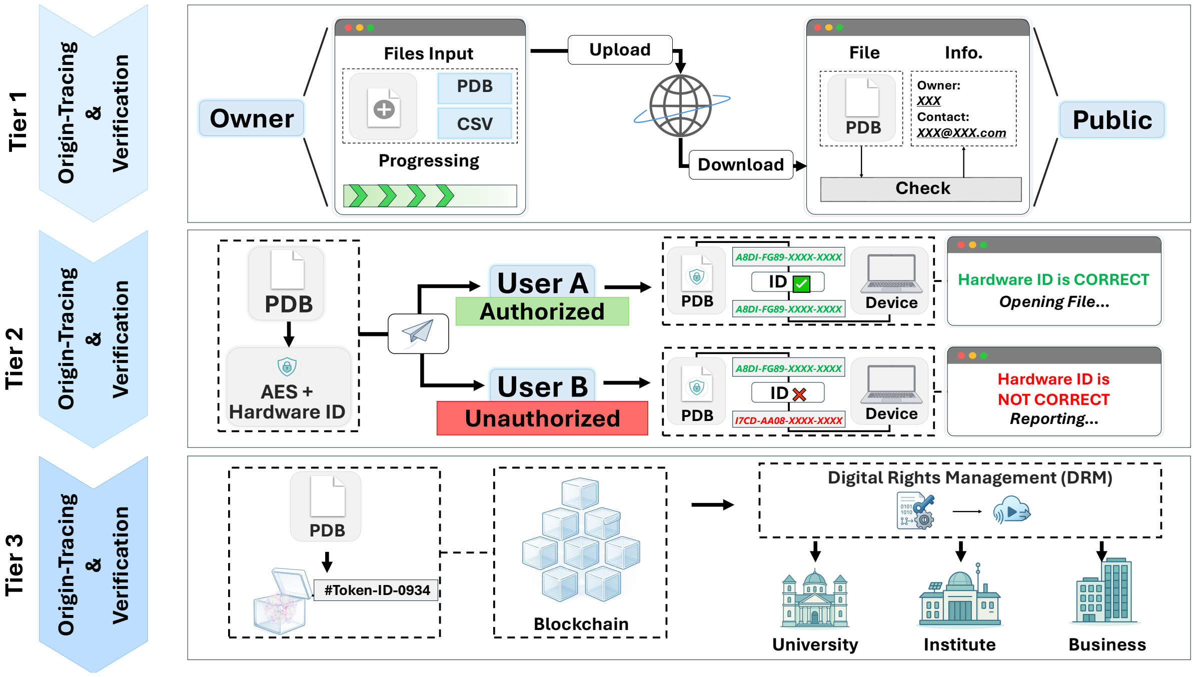

Figure 2. A three-tiered framework for biomolecular provenance, access control, and digital rights management. Our watermarking technology underpins a tiered system for managing biomolecular assets across different use cases. **(Tier 1) Provenance Verification for Public Release.** The watermark links a structure to verifiable metadata, establishing a public record of intellectual priority for academic and open-source applications. **(Tier 2) Hardware-Bound Secure Access.** For proprietary industrial designs, the watermark functions with an encryption layer to restrict access to authorized hardware. Unsuccessful access attempts are logged, providing a security audit trail. **(Tier 3) Blockchain-based Digital Rights Management (DRM).** To facilitate a commercial market, each watermarked structure is assigned a unique blockchain token. This token represents ownership and enables transparent, licensed distribution and rights management on a decentralized ledger.

Second, high-capacity information embedding was achieved with 100% bit accuracy (BitAcc) for 4–32 bit payloads. Structural deviation (scRMSD) increased by only 0.0003 to 0.0015 Å across payload scales, a 300-fold improvement in stability compared to existing AI methods [11–13]. This decouples information density from structural distortion, overcoming a critical limitation of prior approaches.

Third, thermodynamic and dynamic analyses confirmed functional preservation. As Rosetta's `ddg_monomer` protocol showed $\Delta\Delta G \approx 0$ across all cases, there was no indication of destabilization. Molecular dynamics simulations (50–100 ns) revealed stable backbone RMSD trajectories without unfolding or large-scale drift, with RMSF analysis localizing perturbations to non-functional loop regions.

The method's universal applicability was tested on diverse systems, from monomeric actin to multi-subunit RNA polymerase II complexes, with consistent performance. It outperformed state-of-the-art models while requiring no database-specific training.

This combination of minimal structural noise, scalable information density, and preserved biological function establishes a new standard for imperceptible watermarking in bio-molecular assets, addressing IP protection needs in structural and synthetic biology.

4 Discussion

This work introduces a significant advancement in biomolecular intellectual property protection by embedding information directly into the molecular framework. Our approach achieves functional imperceptibility, creating an intrinsic link between the watermark and the molecule without compromising biological function. To translate our watermarking technology into practice, we propose a three-tiered ecosystem for comprehensive biomolecular asset management. This framework, detailed in Figure 2, is designed to address the distinct needs of academic provenance, industrial security, and commercial licensing for bio-designs.

Ultimately, this integrated system provides the cornerstone of trust for the burgeoning bioeconomy. As AI-driven protein design proliferates, it addresses the urgent need for provenance in both academic research and industrial collaborations. By establishing the technical foundation for structure-embedded intellectual property, our work redefines biomolecules as secure, verifiable, and tradable digital assets. This framework fosters innovation by ensuring proper attribution for creators, paving the way for a more transparent and equitable future in biotechnology. The system has been deployed at our center, securing our entire database and our flagship peptide design competition.

Acknowledgments

We thank the Beijing Frontier Research Center for Biological Structure, Tsinghua University, for their support and assistance with this research.

References

[1] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596 (2021), pp. 583–589.

[2] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, et al. “Accurate structure prediction of biomolecular interactions with AlphaFold 3”. In: *Nature* 630.8016 (2024), pp. 493–500.

[3] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, et al. “De novo design of protein structure and function with RFdiffusion”. In: *Nature* 620 (2023), pp. 1089–1100.

[4] J. B. Ingraham, M. Baranov, Z. Costello, K. W. Barber, W. Wang, A. Ismail, V. Frappier, D. M. Lord, C. Ng-Thow-Hing, E. R. Van Vlack, et al. “Illuminating protein space with a programmable generative model”. In: *Nature* 623 (2023), pp. 1070–1078.

[5] A. Jo. “The promise and peril of generative AI”. In: *Nature* 614 (2023), pp. 214–216.

[6] R. Moulange, M. Langenkamp, T. Alexanian, S. Curtis, and M. Livingston. “Towards responsible governance of biological design tools”. In: *arXiv*, 2311.15936 (2023).

[7] S. P. Ikonomova, B. J. Wittmann, F. Piorino, D. J. Ross, S. W. Schaffter, O. Vasilyeva, E. Horvitz, J. Diggans, E. A. Strychalski, S. Lin-Gibson, et al. “Experimental Evaluation of AI-Driven Protein Design Risks Using Safe Biological Proxies”. In: *bioRxiv*, 2025.05.15.654077 (2025).

[8] D. Bloomfield, J. Pannu, A. W. Zhu, M. Y. Ng, A. Lewis, E. Bendavid, S. M. Asch, T. Hernandez-Boussard, A. Cicero, and T. Inglesby. “AI and biosecurity: The need for governance”. In: *Science* 385 (2024), pp. 831–833.

[9] D. Baker and G. Church. “Protein design meets biosecurity”. In: *Science* 383 (2024), pp. 349–349.

[10] M. Wang, Z. Zhang, A. S. Bedi, A. Velasquez, S. Guerra, S. Lin-Gibson, L. Cong, Y. Qu, S. Chakraborty, M. Blewett, et al. “A call for built-in biosecurity safeguards for generative AI tools”. In: *Nature Biotechnology* 43.6 (2025), pp. 845–847.

[11] Z. Zhang, R. Jin, G. Xu, X. Wang, M. Zitnik, L. Cong, and M. Wang. “FoldMark: Safeguarding Protein Structure Generative Models with Distributional and Evolutionary Watermarking”. In: *bioRxiv*, 2024.10.23.619960 (2025).

[12] R. Min, S. Li, H. Chen, and M. Cheng. “A watermark-conditioned diffusion model for ip protection”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 104–120.

[13] W. Feng, W. Zhou, J. He, J. Zhang, T. Wei, G. Li, T. Zhang, W. Zhang, and N. Yu. “Aqualora: Toward white-box protection for customized stable diffusion models via watermark lora”. In: *arXiv*, 2405.11135 (2024).

[14] Y. Chen, Z. Hu, Y. Wu, R. Chen, Y. Jin, M. Zhan, C. Xie, W. Chen, and H. Huang. “Enhancing privacy in biosecurity with watermarked protein design”. In: *Bioinformatics*, btaf141 (2025).

[15] J. E. Gallegos, D. M. Kar, I. Ray, I. Ray, and J. Peccoud. “Securing the exchange of synthetic genetic constructs using digital signatures”. In: *ACS Synthetic Biology* 9 (2020), pp. 2656–2664.