MARCO: THE RADIOACTIVE WATERMARK FOR PROTEIN GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein Generative Models (PGMs) have revolutionized structural biology by enabling the design of complex 3D protein structures from sequence data. However, this breakthrough introduces a dual-use challenge, exposing high-value PGMs to economic risks like unauthorized model extraction and biosecurity threats such as biohazard synthesis. To mitigate these threats, we propose **MARCO** (CONFORMATION WATERMARK), the first radioactive watermarking framework specifically tailored for PGMs. MARCO establishes a Dual-Layer defense that simultaneously protects intellectual property and ensures the forensic traceability of potential biosecurity misuses. (i) To preserve efficiency, MARCO iteratively embeds watermarks during diffusion reverse denoising via an auxiliary encoder-decoder, allowing the original PGM parameters to remain frozen for broad compatibility. (ii) To preserve biophysical fidelity and maximize robustness, we employ specialized loss functions targeting C_α-atom pairwise distances and torsion angles (ψ, ϕ) within an adversarial training framework integrated with stochastic attack simulations. (iii) Crucially, MARCO exhibits "radioactivity" where the watermark automatically transfers to the outputs of any pirate models trained on the watermarked data, effectively countering model extraction attacks. Comprehensive experiments demonstrate that MARCO achieves superior fidelity and robustness while successfully validating watermark transferability.

1 INTRODUCTION

Protein generative models (PGMs), including AlphaFold Abramson et al. (2024) and RoseTTAFold Baek et al. (2021), have transformed structural biology by predicting 3D structures from sequences Baker & Sali (2001). Recognized by the 2024 Nobel Prize in Chemistry, these advances also enable precise de novo design and accelerate the development of therapeutics and biomaterials beyond the reach of traditional screening Luo et al. (2021).

Their rapid deployment, however, creates a dual-use challenge. Open access fosters innovation but may enable malicious users to engineer enhanced pathogens or toxins Wang et al. (2025a), prompting organizations including OpenAI, Anthropic, and Google DeepMind to highlight biosecurity risks Wentzel et al. (2026). Meanwhile, extracting proprietary models Chen et al. (2025) threatens commercial advantages and investment in drug discovery Zhang et al. (2025a). PGMs therefore require tailored protection.

Although watermarking is established for large language models Kirchenbauer et al. (2023) and diffusion-based image generators Fernandez et al. (2023a), proteins impose distinct biophysical and geometric constraints. Embedded signals in coordinates or amino acid sequences must preserve conformation—and thus folding stability and binding affinity Jumper et al. (2021). Because protein structures occupy continuous space and are invariant to rigid transformations, robust watermarks Li et al. (2021) must also withstand $SE(3)$ operations and conformational flexibility. This requires geometry-aware encoding in invariant features such as torsion angles or residue-residue distance maps, rather than simple coordinate noise Miñán et al. (2025).

Our Work. We introduce **MARCO** (CONFORMATION WATERMARK), the first generalized radioactive watermarking framework compatible with contemporary PGM architectures. MARCO adds robust signatures to high-fidelity structures from pretrained models without fine-tuning or re-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

training. Its "radioactive" watermark transfers to pirate models trained on protected structures, preserving data provenance.

We study three questions: **RQ1)** Can a general watermarking framework adapt to existing PGMs without weight fine-tuning or retraining? **RQ2)** Can it preserve biological fidelity while resisting common structural perturbations? **RQ3)** Can the watermark propagate to derivative models trained on protected data, establishing a "radioactive" chain of custody?

A1) Contemporary diffusion-based PGMs generate protein backbones as rigid-body sequences defined by translations and rotations Baek et al. (2021); Lin et al. (2022); Yim et al. (2023c;a). MARCO perturbs both components during reverse diffusion, steering generation toward a watermarked conformation at inference time without retraining. **A2)** MARCO combines a joint encoder-decoder and discriminator Goodfellow et al. (2020) with a frozen PGM. The encoder watermarks intermediate diffusion predictions under fidelity constraints; the decoder recovers signals from stochastically distorted structures; and the discriminator distinguishes clean from watermarked structures. Their competition keeps perturbations biophysically indistinguishable from natural proteins. **A3)** By intervening throughout reverse diffusion rather than watermarking a static structure, MARCO shifts the output distribution Sander et al. (2024). A pirate model trained on these samples learns the biased distribution, so its outputs inherit the watermark and remain traceable to the source.

Our contributions are listed as follows:

- We propose the first "radioactive" watermarking framework tailored to the biophysical and geometric constraints of PGMs with specifically designed biological loss terms.
- MARCO embeds high-fidelity structures with signatures robust to $SE(3)$ transformations and malicious attacks; surrogate models trained on protected data inherit the signal, enabling persistent intellectual-property protection.
- MARCO is an efficient, architecture-agnostic, plug-and-play module that requires no PGM weight fine-tuning.

2 Preliminaries and Related Work

2.1 Protein Generative Models

Overview. Models such as RFdiffusion Watson et al. (2023), ESMFold Lin et al. (2022), and AlphaFold 3 Abramson et al. (2024) learn a joint distribution $p_\theta(\mathbf{X}, \mathbf{S})$ over sequence $\mathbf{S}$ and structure $\mathbf{X}$. A structure comprises N rigid frames $\mathbf{X} = \{(\mathbf{R}_i, \mathbf{x}_i)\}_{i=1}^N$, where $\mathbf{R}_i \in SO(3)$ is the backbone orientation and $\mathbf{x}_i \in \mathbb{R}^3$ the C_α position. They enforce $SE(3)$-equivariance, $\Theta(g \cdot \mathbf{X}) = g \cdot \Theta(\mathbf{X})$, for any global transformation g Fuchs et al. (2020), making predictions coordinate-frame independent and supporting physical validity and conformational generalization. $SO(3)$ is the special orthogonal group of 3D rotations, $SO(3) = \mathbf{R} \in \mathbb{R}^{3\times3} \mid \mathbf{R}^\top\mathbf{R} = I,\ \det(R) = 1$, whose rotations preserve distances, angles, and orientation. $SE(3)$ is the group of 3D rigid transformations, $SE(3) = (\mathbf{R}, \mathbf{x}) \mid \mathbf{R} \in SO(3),\ \mathbf{x} \in \mathbb{R}^3$, which maps $\mathbf{a} \in \mathbb{R}^3$ to $\mathbf{R}\mathbf{a} + \mathbf{x}$ and has the homogeneous form $\mathbf{X} = \begin{bmatrix} \mathbf{R} & \mathbf{x} \\ 0 & 1 \end{bmatrix}$, where $SO(3)$ represents rotation and $\mathbf{x}$ translation.

Learning Objective. Most models sample through denoising diffusion Ho et al. (2020) or flow matching Lipman et al. (2023), whose forward process corrupts $\mathbf{X}_0$ into noise $\mathbf{X}_T$ via an SDE or interpolant. RFdiffusion Watson et al. (2023), for example, recovers $\mathbf{X}_0 = (\mathbf{R}_0, \mathbf{x}_0)$ from $\mathbf{X}_t$ by minimizing weighted translation and rotation errors:

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{X}_0,\eta}\left[w_{\text{trans}}(t)\|\hat{\mathbf{x}}_0 - \mathbf{x}_0\|^2 + w_{\text{rot}}(t)\|\hat{\mathbf{R}}_0 - \mathbf{R}_0\|_{\mathcal{F}}^2\right], \tag{1}$$

where the terms measure C_α coordinate error and the Frobenius distance between orientation matrices, respectively.

Sampling. At step t, Θ predicts the clean structure from $\mathbf{X}_t$: $(\hat{\mathbf{R}}_0, \hat{\mathbf{x}}_0) = \Theta(\mathbf{X}_t, t, \mathbf{S})$. This iterative refinement improves topological coherence. The sampler computes the posterior mean for $\mathbf{X}_{t-1}$,

with linear-interpolation translational mean

$$\boldsymbol{\mu}_{\mathbf{x}} = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \tag{2}$$

and rotational mean obtained by geodesic interpolation between $\mathbf{R}_t$ and $\hat{\mathbf{R}}_0$ on $SO(3)$. Gaussian noise yields $\mathbf{x}_{t-1} = \boldsymbol{\mu}_{\mathbf{x}} + \sigma_t\eta$, where $\eta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, while rotation updates as

$$\mathbf{R}_{t-1} = \mathbf{R}_t \cdot \mathrm{Exp}\left(c_t \cdot \log\left(\mathbf{R}_t^\top \hat{\mathbf{R}}_0\right) + \sigma_t \cdot \xi\right), \tag{3}$$

where c_t is the posterior coefficient, σ_t the noise scale, and $\xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The prediction becomes $\hat{\mathbf{X}}_0^{\text{prev}}$ at the next step. At $t = 0$, a sequence design module such as ProteinMPNN typically optimizes $\mathbf{S}$ for the final backbone $\mathbf{X}_0$.

2.2 Watermarking Diffusion Models

Diffusion watermarking follows two paradigms: *inference-time injection* and *model-time imprinting*. Both bias the generated distribution $p_\theta(\mathbf{x}_0)$ to make outputs verifiable while preserving fidelity.

Inference-Time Injection. Tree-Ring Watermarking Wen et al. (2023) embeds a pattern m_{TR} in a frequency ring M of the initial noise $\mathbf{x}_T$, exploiting low-frequency control of global image layout:

$$\mathcal{F}(\mathbf{x}'_T) = \mathcal{F}(\mathbf{x}_T) \odot (1 - M) + m_{\text{TR}} \odot M \tag{4}$$

Detection recovers $\hat{\mathbf{x}}_T$ through DDIM inversion and tests for m_{TR}. Gaussian Shading Yang et al. (2024) instead uses optimal transport to move $\mathbf{x}_T$ toward a watermarked latent region, avoiding inversion during detection.

Model-Time Imprinting. Stable Signature Fernandez et al. (2023b) fine-tunes the VAE decoder D_ϕ to embed a robust, invisible signature w in every image, balancing fidelity and detection:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{perceptual}}(\mathbf{x}, \mathbf{x}_w) + \lambda\mathcal{L}_{\text{detect}}(E(\mathbf{x}_w), w), \tag{5}$$

where E extracts the watermark. ProMark Asnani et al. (2024) uses learnable prompt tokens to trigger watermarked outputs, whereas rejection-sampling watermarking Zhao et al. (2024) retains outputs satisfying $h(\mathbf{x}) \in \mathcal{K}$.

Limitations in Watermarking Protein Structures. Image methods Zhao et al. (2023); Huang et al. (2024); Liu et al. (2023); Ci et al. (2025) assume fixed 2D grids, image-specific latents or text conditioning, and perceptual invisibility, so they do not transfer directly to variable-length 3D proteins. Protein watermarks must respect $SE(3)$ symmetry and preserve inter-residue distances, torsion angles, stereochemistry, folding stability, and function. FoldMark Zhang et al. (2025b) applies watermark-conditioned LoRA fine-tuning to PGMs, but modifies model weights and does not sufficiently constrain the predicted structures biologically.

3 Threat Model

We consider two parties: the victim and the adversary. Based on this setting, we define MARCO's design properties in Appendix A.

The Victim. The victim, typically a pharmaceutical company or research laboratory, develops or hosts a proprietary PGM Θ (e.g., RFdiffusion) trained on valuable structural data. It embeds a recoverable watermark $\mathbf{w}$ in generated structures $\hat{\mathbf{X}} \sim \Theta(\cdot)$, producing $\tilde{\mathbf{X}}$ such that $D(\tilde{\mathbf{X}}) \approx \mathbf{w}$. The victim has white-box access to Θ, including its parameters θ, activations, and gradients. It also has sufficient computing resources for biophysical validation through forward-folding simulations and energy minimization.

The Adversary. The adversary is a malicious user or competitor who exploits model outputs while hiding their provenance to infringe intellectual property or evade biosecurity screening for restricted agents such as toxins. It transforms a watermarked structure as $\mathbf{X}' = \mathcal{A}(\tilde{\mathbf{X}})$. An attack succeeds

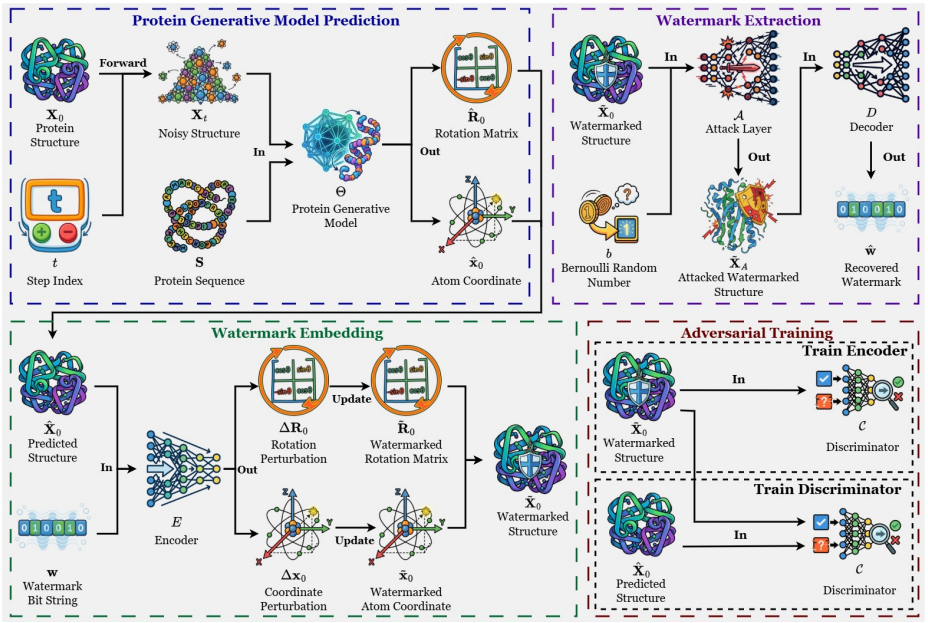


Figure 1: **MARCO training.** *i) PGM prediction*: The frozen PGM Θ predicts a clean structure $\hat{\mathbf{X}}_0$ from noisy state $\mathbf{X}_t$ and sequence $\mathbf{S}$. *ii) Watermark embedding*: Encoder E embeds $\mathbf{w}$ through geometric perturbations, producing $\tilde{\mathbf{X}}_0$ and steering the generation trajectory. *iii) Watermark extraction*: A Bernoulli-controlled attack layer produces $\tilde{\mathbf{X}}_A$, from which decoder D recovers $\hat{\mathbf{w}}$ for robustness. *iv) Adversarial training*: Discriminator $\mathcal{C}$ distinguishes $\hat{\mathbf{X}}_0$ from $\tilde{\mathbf{X}}_0$, while E learns biophysically imperceptible perturbations that deceive $\mathcal{C}$.

only if *i)* detection fails, $D(\mathbf{X}') \not\approx \mathbf{w}$, and *ii)* $\mathbf{X}'$ retains the structural integrity and biological function of $\tilde{\mathbf{X}}$. The adversary knows the watermarking algorithm and model architecture, but not the victim's training data, training algorithm, or watermark parameters. Its access to the victim model is black-box and query-only, but it has unlimited access to watermarked structures. It has standard computing resources: an A100 GPU with 80GB VRAM and 64GB RAM suffices to fine-tune AlphaFold2, and molecular-modeling software such as Rosetta and PyMOL[1].

Adaptive Attacker. An adaptive adversary knows MARCO fully and uses optimization to erase its watermark while preserving fidelity. It may use OpenMM molecular dynamics Eastman et al. (2013) or Rosetta Relax Tyka et al. (2011) to drive structures toward local thermodynamic minima. With white-box access to its pirate model, the adversary may also use Differentially Private Stochastic Gradient Descent (DPSGD) Abadi et al. (2016) to scrub the watermark during training. Its calibrated gradient noise prevents the PGM from retaining the watermarked distributional features. It may also forge a watermark on the structures Chen et al. (2024); Zhang et al. (2021) to obscure ownership verification.

4 MARCO METHOD

4.1 OVERVIEW

We introduce MARCO, a watermarking framework for Protein Generative Models (PGMs) that formulates watermarking as multi-objective optimization on the rigid-body $SE(3)$ manifold using an adversarially trained encoder-decoder network. At each denoising step, the encoder embeds the watermark into the predicted protein backbone, steering diffusion toward a verifiable "watermark manifold". A discriminator penalizes distributional divergence so that watermarked structures remain biophysically indistinguishable from natural ones. During training, a stochastic attack layer

[1]https://rosettacommons.org/, https://pymol.org/

Algorithm 1: Training Process of MARCO

Data: Frozen PGM Θ; encoder E; decoder D; discriminator $\mathcal{C}$; attack simulator $\mathcal{A}$; training structures and sequences; watermark $\mathbf{w}$
Result: Trained watermark encoder E and decoder D
Freeze Θ;
while *training has not converged* **do**
  Sample a noisy structure and use Θ to predict its clean backbone;
  Use E to embed $\mathbf{w}$ into the prediction with a timestep-aware geometric perturbation;
  Apply a stochastic simulated attack and use D to recover the watermark;
  Update E and D to balance watermark recovery, structural fidelity, and covertness;
  Update $\mathcal{C}$ to distinguish clean predictions from watermarked ones;
end
return E, D

Algorithm 2: Evaluation Process of MARCO

Data: PGM Θ; trained encoder E and decoder D; protein sequence $\mathbf{S}$; owner and model identities; watermark length; authoritative ledger; suspect structure $\mathbf{X}$
Result: Watermarked structure $\tilde{\mathbf{X}}$; recovered watermark $\hat{\mathbf{w}}$; detection decision d_{wm}
/* Provenance-Backed Watermark Generation */
Derive a target watermark $\mathbf{w}$ of the specified length from the model identity;
Timestamp and register the ownership record with the authoritative ledger;
/* Watermarked Structure Generation */
Initialize a random noisy structure;
foreach *denoising step* **do**
  Use Θ to predict the clean protein structure;
  Use E to embed $\mathbf{w}$ through a timestep-aware geometric perturbation;
  Advance the denoising trajectory using the watermarked prediction;
end
Set $\tilde{\mathbf{X}}$ to the final generated structure;
/* Watermark Validation */
Use D to recover $\hat{\mathbf{w}}$ from the suspect structure $\mathbf{X}$;
Compare $\hat{\mathbf{w}}$ with $\mathbf{w}$ using a statistical confidence test;
Set d_{wm} according to whether the detection threshold is met;
return $\tilde{\mathbf{X}}$, $\hat{\mathbf{w}}$, d_{wm}

encourages redundant, persistent features that survive geometric perturbations and malicious attacks, while the decoder recovers the signature from the distorted structures to reinforce robustness. For flow-matching and deterministic models (e.g. ESMFold), MARCO can also provide robust watermark with fewer steps of perturbations. In Appendix B, we prove that MARCO provides certified robustness.

4.2 Training MARCO

As shown in Algorithm 1 and Figure 1, MARCO jointly optimizes a watermark encoder E, decoder D, and discriminator $\mathcal{C}$ with a frozen PGM Θ in four phases: *i)* PGM prediction, *ii)* watermark embedding, *iii)* watermark extraction, and *iv)* adversarial training.

Protein Generative Model Prediction. The forward process converts a ground-truth structure $\mathbf{X}_0$ into $\mathbf{X}_t$ at timestep t, from which Θ estimates the clean baseline $\hat{\mathbf{X}}_0 = (\hat{\mathbf{R}}_0, \hat{\mathbf{x}}_0)$ using $\mathbf{X}_t$, t, and sequence $\mathbf{S}$ (Subsection 2.1).

Watermark Embedding. The encoder E embeds $\mathbf{w}$ into this estimate through residual learning, predicting a perturbation field $(\Delta\mathbf{R}_0, \Delta\mathbf{x}_0)$ rather than absolute watermarked coordinates:

$$(\Delta\mathbf{R}_0, \Delta\mathbf{x}_0) = E(\hat{\mathbf{R}}_0, \hat{\mathbf{x}}_0, \mathbf{w}) \tag{6}$$

The watermark is embedded in $\hat{\mathbf{x}}_0$ by adding $\Delta\mathbf{x}_0$ under a temporal scaling schedule:

$$\tilde{\mathbf{x}}_0 = \hat{\mathbf{x}}_0 + \cos\left(\frac{\pi(T-t)}{2T}\right)\Delta\mathbf{x}_0 \tag{7}$$

The cosine schedule anneals the perturbation magnitude with timestep t.

Because $SO(3)$ is a non-linear manifold, element-wise addition violates rotational orthogonality. We instead define $\Delta\mathbf{R}_0 \in \mathbb{R}^3$ in the tangent space $\mathfrak{so}(3)$, with $\|\Delta\mathbf{R}_0\|$ giving the angular magnitude in radians, and map it onto the manifold via the matrix exponential with the same temporal

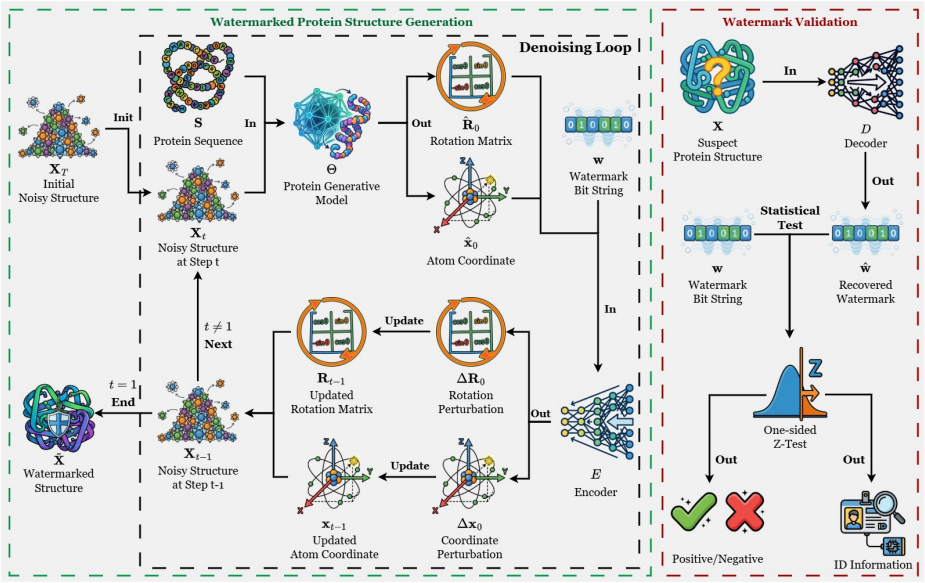


Figure 2: **MARCO evaluation.** *i) Watermarked structure generation*: At each denoising step, watermark perturbations are applied to the PGM's clean predictions, yielding a fully watermarked structure $\tilde{\mathbf{X}}$. *ii) Watermark validation*: The decoder recovers $\hat{\mathbf{w}}$ from a suspect structure and statistically compares it with the original watermark. Successful verification confirms provenance and identifies the embedded user.

modulation:

$$\Delta\mathbf{R}_0 = \exp(\cos(\frac{\pi(T-t)}{2T}) \cdot \Delta\mathbf{R}_0) \tag{8}$$

The perturbed rotation is then obtained by matrix multiplication:

$$\tilde{\mathbf{R}}_0 = \hat{\mathbf{R}}_0 \Delta\mathbf{R}_0 \tag{9}$$

The watermarked estimates $(\tilde{\mathbf{R}}_0, \tilde{\mathbf{x}}_0)$ replace $(\hat{\mathbf{R}}_0, \hat{\mathbf{x}}_0)$ in the standard diffusion updates (Subsection 2.1 and Equation 3), propagating the watermark to $\mathbf{X}_{t-1}$.

To preserve structural fidelity, we minimize the translational and rotational deviations $\|\hat{\mathbf{x}}_0^i - \tilde{\mathbf{x}}_0^i\|^2$ and $\|\hat{\mathbf{R}}_0^i - \tilde{\mathbf{R}}_0^i\|^2$. Under Equation 6, this reduces to constraining the predicted perturbations through:

$$\mathcal{L}_{\text{wm}} = \frac{1}{N}\sum_{i=1}^{N} \|\Delta\mathbf{R}_0^i\| + \|\Delta\mathbf{x}_0^i\| \tag{10}$$

Minimizing $\mathcal{L}_{\text{wm}}$ constrains the geometric distance between $\hat{\mathbf{X}}_0$ and $\tilde{\mathbf{X}}_0$, preserving biophysical consistency.

We further promote embedding fidelity with two novel losses: the distance matrix loss $\mathcal{L}_{\text{dm}}$ and torsion angle loss $\mathcal{L}_{\text{tor}}$.

The distance matrix loss enforces the preservation of global structural topology by penalizing deviations in pairwise C_α distances:

$$\mathcal{L}_{\text{dm}} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\|\hat{\mathbf{x}}_0^i - \hat{\mathbf{x}}_0^j\| - \|\tilde{\mathbf{x}}_0^i - \tilde{\mathbf{x}}_0^j\|)^2 \tag{11}$$

Minimizing this term preserves the clean backbone's internal organization and relative atomic positions, preventing global fold distortion.

The torsion angle loss $\mathcal{L}_{\text{tor}}$ constrains local geometry:

$$\mathcal{L}_{\text{tor}} = \frac{1}{N-2}\sum_{i=2}^{N-1}[1-\cos(\hat{\phi}_i - \tilde{\phi}_i) + 1 - \cos(\hat{\psi}_i - \tilde{\psi}_i)], \tag{12}$$

where $(\hat{\phi}_i, \hat{\psi}_i)$ and $(\tilde{\phi}_i, \tilde{\psi}_i)$ are the clean and watermarked backbone dihedral angles, computed using standard protocols Watson et al. (2023). The loss penalizes sterically implausible conformations to preserve local structural integrity.

Watermark Extraction. Before extraction, an attack layer $\mathcal{A}$ simulates adversarial conditions through manipulations such as rotation, cropping, and noise addition, producing $\tilde{\mathbf{X}}_A = \mathcal{A}(\tilde{\mathbf{X}}_0, b)$. For $b \sim \text{Bern}(0.5)$, the attack is applied when $b = 1$; otherwise, the structure is unchanged.

The decoder D recovers $\hat{\mathbf{w}}$ from the potentially distorted $\tilde{\mathbf{X}}_A$ using the extraction loss:

$$\mathcal{L}_{\text{ext}} = -\frac{1}{l}\sum_{i=1}^{l} \mathbf{w}_i \log(\hat{\mathbf{w}}_i) + (1 - \mathbf{w}_i)\log(1 - \hat{\mathbf{w}}_i) \tag{13}$$

Minimizing $\mathcal{L}_{\text{ext}}$ enables robust recovery from both pristine and structurally compromised proteins.

Adversarial Training. A discriminator $\mathcal{C}$, trained jointly with E and D, minimizes the classification error between clean and watermarked structures in an adversarial minimax game:

$$\mathcal{L}_{\mathcal{C}} = -\big(\log \mathcal{C}(\hat{\mathbf{X}}_0) + \log\big(1 - \mathcal{C}(\tilde{\mathbf{X}}_0)\big)\big) \tag{14}$$

The encoder simultaneously minimizes an adversarial loss to deceive $\mathcal{C}$:

$$\mathcal{L}_{\text{adv}} = -\log\big(\mathcal{C}(\tilde{\mathbf{X}}_0)\big) \tag{15}$$

Thus, $\mathcal{L}_{\mathcal{C}}$ trains the discriminator to separate the distributions, while $\mathcal{L}_{\text{adv}}$ trains the encoder to make $\tilde{\mathbf{X}}_0$ statistically indistinguishable from $\hat{\mathbf{X}}_0$.

The overall MARCO training loss is

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{wm}} + \lambda_2\mathcal{L}_{\text{dm}} + \lambda_3\mathcal{L}_{\text{tor}} + \lambda_4\mathcal{L}_{\text{ext}} + \lambda_5\mathcal{L}_{\text{adv}}. \tag{16}$$

4.3 Evaluating MARCO

As shown in Algorithm 2 and Figure 2, MARCO establishes a provenance-backed watermark, generates a watermarked protein structure, and validates the watermark. To prevent ownership claims based on forged watermarks, MARCO cryptographically timestamps and registers the model identity and timestamp with a trustworthy ledger; unregistered watermarks cannot establish ownership.

Watermarked Protein Structure Generation. Starting from standard Gaussian noise $\mathbf{X}_T$ and protein sequence $\mathbf{S}$, the reverse diffusion process ($t = T \to 1$) uses Θ to estimate a clean baseline $\hat{\mathbf{X}}_0$ at each step. Encoder E injects perturbations $(\Delta\mathbf{R}_0, \Delta\mathbf{x}_0)$ to produce $(\tilde{\mathbf{R}}_0, \tilde{\mathbf{x}}_0)$, which guides the transition to $\mathbf{X}_{t-1}$. The process yields the releasable watermarked structure $\tilde{\mathbf{X}}$.

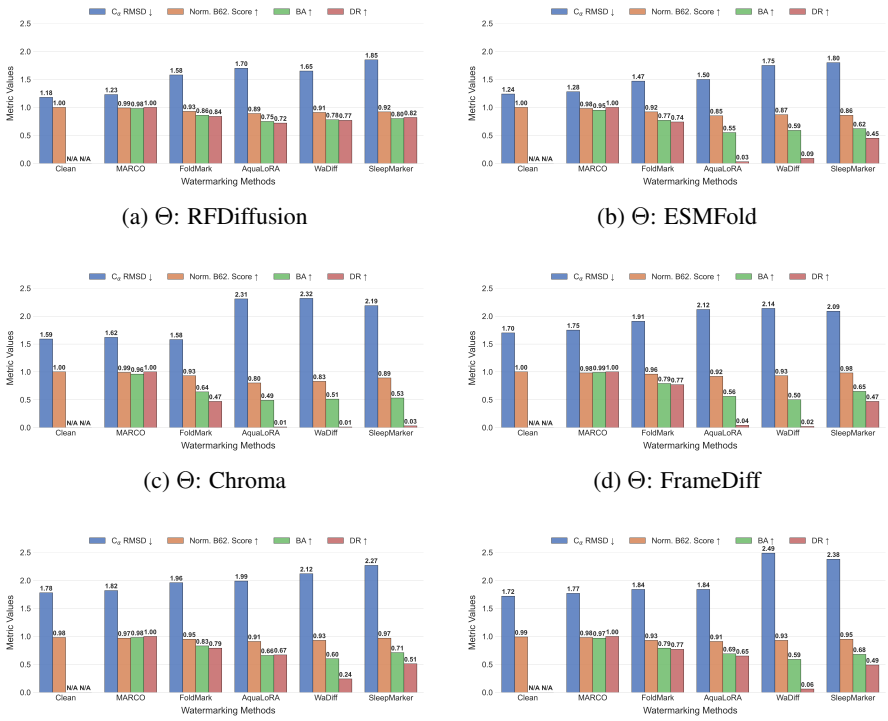
Watermark Validation. The victim uses decoder D to recover $\hat{\mathbf{w}}$ from a suspect structure $\mathbf{X}$ and counts the k matching bits between $\hat{\mathbf{w}}$ and the original watermark $\mathbf{w}$. A one-sided Z-test assesses significance against random coincidence using $z = (k - \mu)/\sigma$, where $\mu = 0.5l$ and $\sigma = 0.5\sqrt{l}$ are the mean and standard deviation for a random binary sequence of length l with $p = 0.5$. The threshold $\hat{z} = 3.0$ gives a false positive rate of $\leq 10^{-3}$. MARCO sets $d_{\text{wm}} = 1$ if $z \geq \hat{z}$ and 0 otherwise, returning both d_{wm} and $\hat{\mathbf{w}}$.

5 Evaluation

The experimental settings and the other experiments about robustness, radioactivity, robustness and ablation study are presented in Appendix C and Appendix D.

5.1 Fidelity

We compare MARCO with four baselines using C_α-RMSD, Normalized BLOSUM62 score, Bit Accuracy (BA), and Detection Rate (DR) across six pre-trained PGMs with official weights and



(a) Θ: RFDiffusion (b) Θ: ESMFold

(c) Θ: Chroma (d) Θ: FrameDiff

(e) Θ: FrameFlow (f) Θ: FoldFlow2

Figure 3: **Evaluation of Fidelity of PGM Watermarking Methods**.

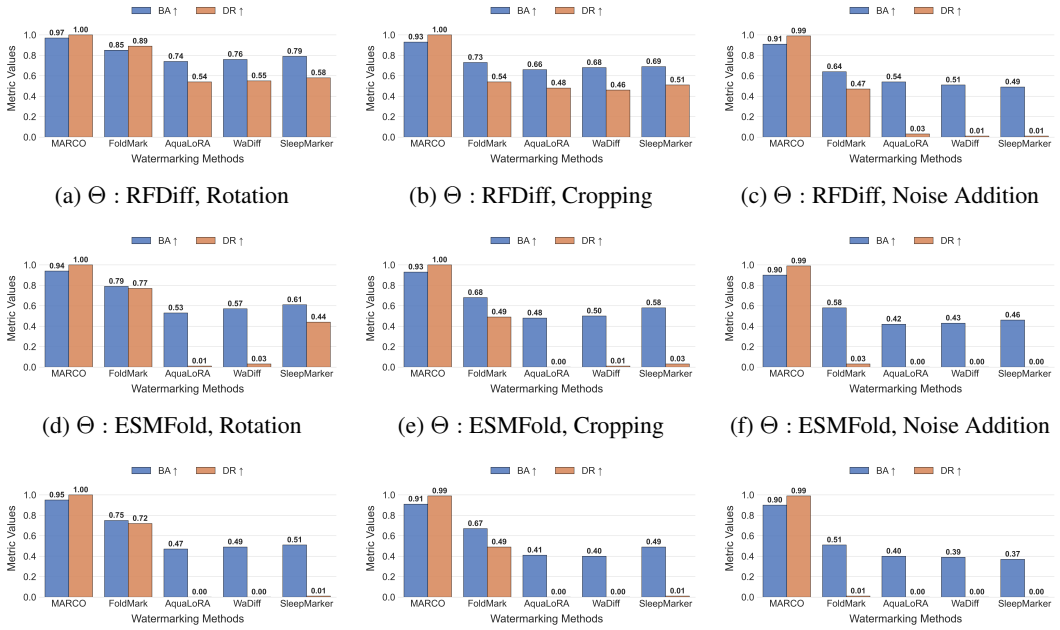
published baseline hyperparameters. We set the watermark length to $l = 64$, providing a 2^{64} identification space for robust validation, and generate 1000 clean–watermarked structure pairs $(\hat{\mathbf{X}}, \tilde{\mathbf{X}})$. Results compare these structures and the ground-truth and recovered watermarks $(\mathbf{w}, \hat{\mathbf{w}})$.

As shown in Figure 3, MARCO leads most metrics, achieving the lowest RMSD, near-perfect BLOSUM62 scores, BA > 0.95, and DR $= 1.0$ without attacks. Image-domain baselines lack protein-specific thermodynamic constraints and therefore compromise fidelity and robustness by distorting molecular conformations. FoldMark is the strongest baseline, but MARCO's design performs better. Moreover, the baselines require PGM fine-tuning, whereas MARCO is a plug-and-play module that leaves the backbone model unchanged. Their partial watermark recovery also remains substantially less accurate than MARCO's, especially under adversarial attacks. The RFDiffusion overlays in Figure 5 (Appendix D) show strong alignment between clean and watermarked structures, confirming that MARCO preserves native conformations with minimal geometric alteration.

5.2 Robustness

MARCO vs. Conformational Manipulations. We evaluate MARCO against random rotation, cropping, and Gaussian noise injection. Random Rotation: For $\mathbf{X} = \{(\mathbf{R}_i, \mathbf{x}_i)\}_{i=1}^{N}$, we sample a global rotation $\mathbf{Q} \sim \text{Haar}(SO(3))$, the uniform distribution over $SO(3)$, and apply $\mathbf{x}'_i = \mathbf{Q}(\mathbf{x}_i - \mathbf{c}) + \mathbf{c}$ and $\mathbf{R}'_i = \mathbf{Q}\mathbf{R}_i$, where $\mathbf{c}$ is the geometric center. Cropping: We remove a contiguous 30% of the backbone from either end. Gaussian Noise Injection: For $\mathbf{x} \in \mathbb{R}^{N\times 3}$, we sample $\chi \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\chi \in \mathbb{R}^{N\times 3}$, and add it to $\mathbf{x}$.

As shown in Figure 4 and Figure 6 (Appendix D), MARCO significantly outperforms all four baselines in Bit Accuracy (BA) and Detection Rate (DR), maintaining BA > 0.90 and DR > 0.99 under every perturbation. Rotation has relatively little effect, whereas cropping and noise sharply degrade the baselines by removing structural information and globally distorting molecular conformations, respectively. MARCO's robustness suggests that the attack layer is critical for learning persistent features; adversarial training likely also diffuses the watermark throughout the structure to evade the discriminator and prevent localized failures.



(a) Θ : RFDiff, Rotation (b) Θ : RFDiff, Cropping (c) Θ : RFDiff, Noise Addition

(d) Θ : ESMFold, Rotation (e) Θ : ESMFold, Cropping (f) Θ : ESMFold, Noise Addition

(g) Θ : Chroma, Rotation (h) Θ : Chroma, Cropping (i) Θ : Chroma, Noise Addition

Figure 4: Part A – **Evaluation of Robustness of PGM Watermarking Methods against Conformation Manipulations**.

6 Discussion

PGMs face economic and biosecurity threats: extracting proprietary models causes economic loss, while bypassing safeguards to synthesize toxins or bioweapons risks severe societal harm. MARCO addresses both with a dual-layer defense that embeds user-specific intellectual property as a watermark, serving as both a Digital Rights Management (DRM) tool and a Biological Weapons Convention (BWC)-compliant digital chain of custody for PGM accountability. MARCO balances fidelity, preserved by geometric constraints, with robustness, enhanced by the attack layer and extraction loss; adversarial training reinforces both. Adjusting the loss weights and probability of conformational manipulations tailors this trade-off to specific security requirements. Capacity remains a performance constraint: the ablation study shows that increasing watermark length l reduces fidelity and robustness. Choosing an optimal l therefore balances scalable user identification with structural integrity. Model extraction attacks aim to learn a PGM's output distribution. MARCO makes this distribution radioactive through stepwise reverse-diffusion perturbations, and cross-architecture training experiments confirm that the watermark propagates to pirate models. This radioactivity is essential for identifying stolen models and validating ownership. In Appendix E, we discuss potential extensions of MARCO to protein language models and wet-lab verification.

7 Conclusion

We propose MARCO, a radioactive watermark designed to efficiently adapt to mainstream Protein Generative Models without modifying their pre-trained weights. MARCO achieves a critical balance of fidelity and robustness, generating attack-resilient structures with negligible geometric deviation from clean predictions. Sequence validation via ProteinMPNN confirms that these structural watermarks exert minimal influence on the recovered amino acid sequence. Furthermore, the watermark proves robust against conformational manipulations and even adaptive attacks while supporting high-capacity payloads with minimal performance trade-offs. Future work will focus on further enhancing conformational fidelity and extending the framework's applicability to a broader spectrum of generative architectures and other biological modalities.

AI USE STATEMENT

In the preparation of this manuscript, we employed generative AI tools strictly for the purpose of linguistic refinement. Specifically, these tools were used to check for grammatical errors and to improve the flow and readability of the text. We affirm that:

- **No Content Generation:** The core concepts, experimental design, data analysis, and scientific conclusions are entirely the work of the authors.
- **No Synthetic Text:** No substantial portions of the text were generated by AI. The tools acted solely as a copy-editing assistant.
- **Human Oversight:** All suggestions made by the AI were manually reviewed and verified by the authors, who take full responsibility for the accuracy and integrity of the final publication.

ETHICS STATEMENT

In conducting this research, we strictly adhered to a "no-harm" principle. Our methodology was designed to protect the copyrights of the model owner and trace biosecurity misuses of protein generative models in a purely theoretical and isolated context, ensuring that no individuals, organizations, or operational systems were negatively impacted.

Isolation from Operational Systems. All experiments described in this paper were performed in a fully offline, local environment. We utilized open-source implementations of dataset distillation protocols and deployed them on our own hardware. At no point did our method interact with, query, or stress-test live commercial APIs or external web services. Consequently, this research caused no service disruption, latency, or financial cost to any model providers or platforms.

Use of Public, Non-Sensitive Data. The evaluation was conducted exclusively using standard, publicly available academic datasets. We did not harvest data from private users, nor did we target specific content creators. By relying solely on established benchmarks, we ensured that no personal data was processed, and no individual's privacy or intellectual property rights were infringed upon during the course of this study.

Safe Security Assessment. Our work treats the sensitive information of the protein datasets and the protein generative models as a mathematical artifact to be analyzed, rather than attacking the infrastructure that hosts it. By confining our "red-teaming" efforts to a sandbox environment, we demonstrate that security vulnerabilities can be identified and documented without risk to the broader digital ecosystem or its stakeholders.

REPRODUCIBILITY STATEMENT

To promote transparency and reproducibility, we will release artifacts including all the code of all empirical studies employed in this paper. We currently provide the artifact in the following anonymized link: https://anonymous.4open.science/r/MARCO-0116. All datasets we use are included in the artifact as they are public accessible.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael

Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493—-500, 2024. doi: 10.1038/s41586-024-07487-w.

Vishal Asnani, John Collomosse, Tu Bui, Xiaoming Liu, and Shruti Agarwal. Promark: Proactive diffusion watermarking for causal attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10802–10811, 2024.

Ziga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, Raina Thomas, Vincent Dutordoir, Matteo Perino, Soham De, Alexander Karollus, Adam Gayoso, Toby Sargeant, Anne Mottram, Lai Hong Wong, Pavol Drotár, Adam Kosiorek, Andrew Senior, Richard Tanburn, Taylor Applebaum, Souradeep Basu, Demis Hassabis, and Pushmeet Kohli. Advancing regulatory variant effect prediction with alphagenome. *Nature*, 649(8099):1206–1218, January 2026. ISSN 1476-4687. doi: 10.1038/s41586-025-10014-0. URL http://dx.doi.org/10.1038/s41586-025-10014-0.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

David Baker and Andrej Sali. Protein structure prediction and structural genomics. *Science*, 294 (5540):93–96, 2001.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Huajie Chen, Tianqing Zhu, Chi Liu, Shui Yu, and Wanlei Zhou. High-frequency matters: attack and defense for image-processing model watermarking. *IEEE Transactions on Services Computing*, 17(4):1565–1579, 2024.

Huajie Chen, Tianqing Zhu, Lefeng Zhang, Bo Liu, Derui Wang, Wanlei Zhou, and Minhui Xue. Queen: Query unlearning against model extraction. *IEEE Transactions on Information Forensics and Security*, 2025.

Hai Ci, Yiren Song, Pei Yang, Jinheng Xie, and Mike Zheng Shou. WMAdapter: Adding WaterMark control to latent diffusion models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10901–10919. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/ci25a.html.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

David R Davies, Eduardo A Padlan, and Steven Sheriff. Antibody-antigen complexes. *Annual review of biochemistry*, 59(1):439–473, 1990.

SD Durbin and G Feher. Protein crystallization. *Annual review of physical chemistry*, 47(1):171–204, 1996.

Peter Eastman, Mark S Friedrichs, John D Chodera, Randall J Radmer, Christopher M Bruns, Joy P Ku, Kyle A Beauchamp, Thomas J Lane, Lee-Ping Wang, Diwakar Shukla, et al. Openmm 4: a reusable, extensible, hardware independent library for high performance molecular simulation. *Journal of chemical theory and computation*, 9(1):461–469, 2013.

Weitao Feng, Wenbo Zhou, Jiyan He, Jie Zhang, Tianyi Wei, Guanlin Li, Tianwei Zhang, Weiming Zhang, and Nenghai Yu. Aqualora: Toward white-box protection for customized stable diffusion models via watermark lora. *arXiv preprint arXiv:2405.11135*, 2024.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22466–22477, 2023a.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22466–22477, 2023b.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, 33:1970–1981, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Huayang Huang, Yu Wu, and Qian Wang. Robin: Robust and invisible watermarks for diffusion models with adversarial optimization. *Advances in Neural Information Processing Systems*, 37: 3937–3963, 2024.

Xiufeng Huang, Ziyuan Luo, Qi Song, Ruofei Wang, and Renjie Wan. Marksplatter: Generalizable watermarking for 3d gaussian splatting model via splatter image structure. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12189–12198, 2025.

Guillaume Huguet, James Vuckovic, Kilian Fatras, Eric Thibodeau-Laufer, Pablo Lemos, Riashat Islam, Cheng-Hao Liu, Jarrid Rector-Brooks, Tara Akhound-Sadegh, Michael Bronstein, et al. Sequence-augmented se (3)-flow matching for conditional protein backbone generation. *Advances in neural information processing systems*, 2024.

John B. Ingraham, Max Baranov, Zak Costello, Karl W. Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M. Lord, Christopher Ng-Thow-Hing, Erik R. Van Vlack, Shan Tie, Vincent Xue, Sarah C. Cowles, Alan Leung, João V. Rodrigues, Claudio L. Morales-Perez, Alex M. Ayoub, Robin Green, Katherine Puentes, Frank Oplinger, Nishant V. Panwar, Fritz Obermeyer, Adam R. Root, Andrew L. Beam, Frank J. Poelwijk, and Gevorg Grigoryan. Illuminating protein space with a programmable generative model. *Nature*, 2023. doi: 10.1038/s41586-023-06728-8.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pp. 17061–17084. PMLR, 2023.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-antoine Gourraud, Mickaël Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. In *62th Annual Meeting of the Association for Computational Linguistics (ACL'24)*, 2024.

Xiaolong Li, Yijia Weng, Li Yi, Leonidas J Guibas, A Abbott, Shuran Song, and He Wang. Leveraging se (3) equivariance for self-supervised category-level object pose estimation from point clouds. *Advances in neural information processing systems*, 34:15370–15381, 2021.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Yugeng Liu, Zheng Li, Michael Backes, Yun Shen, and Yang Zhang. Watermarking diffusion model. *arXiv preprint arXiv:2305.12502*, 2023.

Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3d generative model for structure-based drug design. *Advances in Neural Information Processing Systems*, 34:6229–6239, 2021.

Rui Min, Sen Li, Hongyang Chen, and Minhao Cheng. A watermark-conditioned diffusion model for ip protection. In *European Conference on Computer Vision*, pp. 104–120. Springer, 2024.

Raúl Miñán, Javier Gallardo, Álvaro Ciudad, and Alexis Molina. Informed protein–ligand docking via geodesic guidance in translational, rotational and torsional spaces. *Nature Machine Intelligence*, 7(9):1555–1560, 2025.

Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. Watermarking makes language models radioactive. *Advances in Neural Information Processing Systems*, 37: 21079–21113, 2024.

Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, et al. Boltzgen: Toward universal binder design. *bioRxiv*, pp. 2025–11, 2025.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *The Twelfth International Conference on Learning Representations*.

Michael D Tyka, Daniel A Keedy, Ingemar André, Frank DiMaio, Yifan Song, David C Richardson, Jane S Richardson, and David Baker. Alternate states of proteins revealed by detailed energy landscape mapping. *Journal of molecular biology*, 405(2):607–618, 2011.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Mengdi Wang, Zaixi Zhang, Amrit Singh Bedi, Alvaro Velasquez, Stephanie Guerra, Sheng Lin-Gibson, Le Cong, Yuanhao Qu, Souradip Chakraborty, Megan Blewett, et al. A call for built-in biosecurity safeguards for generative ai tools. *Nature Biotechnology*, 43(6):845–847, 2025a.

Zilan Wang, Junfeng Guo, Jiacheng Zhu, Yiming Li, Heng Huang, Muhao Chen, and Zhengzhong Tu. Sleepermark: Towards robust watermark against fine-tuning text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8213–8224, 2025b.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023.

Josh Wentzel, Zach Graves, and Dean Ball. In Support of Mandatory Nucleic Acid Synthesis Screening and Recordkeeping. https://www.thefai.org/posts/in-support-of-mandatory-nucleic-acid-synthesis-screening-and-recordkeeping, 2026. [Online; accessed 05-June-2026].

Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12162–12171, 2024.

Jason Yim, Andrew Campbell, Emile Mathieu, Andrew YK Foong, Michael Gastegger, Jose Jimenez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S Veeling, Frank Noe, et al. Improved motif-scaffolding with se (3) flow matching. *Transactions on Machine Learning Research*.

Jason Yim, Andrew Campbell, Andrew YK Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S Veeling, Regina Barzilay, Tommi Jaakkola, et al. Fast protein backbone generation with se (3) flow matching. *arXiv preprint arXiv:2310.05297*, 2023a.

Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se (3) diffusion model with application to protein backbone generation. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 40001–40039, 2023b.

Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023c.

Jie Zhang, Dongdong Chen, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, and Nenghai Yu. Deep model intellectual property protection via deep watermarking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4005–4020, 2021.

Zaixi Zhang, Souradip Chakraborty, Amrit Singh Bedi, Emilin Mathew, Varsha Saravanan, Le Cong, Alvaro Velasquez, Sheng Lin-Gibson, Megan Blewett, Dan Hendrycs, et al. Generative ai for biosciences: Emerging threats and roadmap to biosecurity. *arXiv preprint arXiv:2510.15975*, 2025a.

Zaixi Zhang, Ruofan Jin, Guangxue Xu, Xiaotong Wang, Marinka Zitnik, Le Cong, and Mengdi Wang. Foldmark: Safeguarding protein structure generative models with distributional and evolutionary watermarking. *bioRxiv*, pp. 2024–10, 2025b.

Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative ai. *Advances in neural information processing systems*, 37:8643–8672, 2024.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023.

APPENDIX

A DESIGN PROPERTIES

We define the following core properties as essential prerequisites for any effective PGM watermarking framework. MARCO is explicitly designed to satisfy these constraints:

Fidelity quantifies the preservation of structural integrity and biological function. A viable watermark must ensure that the modified protein structures remain indistinguishable from the original distribution in terms of biophysical plausibility. Mathematically, we require that the watermarked samples $\tilde{\mathbf{X}}$ reside within comparable low-energy states to the clean predictions $\hat{\mathbf{X}}$:

$$\mathbb{E}_{\mathbf{X}_w}[\mathcal{E}_{\text{phys}}(\tilde{\mathbf{X}})] \approx \mathbb{E}_{\mathbf{X}}[\mathcal{E}_{\text{phys}}(\hat{\mathbf{X}})], \tag{17}$$

where $\mathcal{E}_{\text{phys}}$ denotes a validated physical energy potential. Specifically, Equation 1 is an example of defining fidelity in detail.

Robustness evaluates the resilience of the embedded signal against adversarial structural perturbations. A watermarking scheme is deemed robust if the watermark $\mathbf{w}$ remains reliably recoverable even after the protein undergoes any utility-preserving transformation or attack $\mathcal{A} \in \mathcal{T}$ (e.g., rigid body rotation, noise injection, or energy minimization). Mathematically, we require that the Bit Error Rate (BER) of the recovered signal remains strictly bounded by a tolerance threshold ρ:

$$\sup_{\mathcal{A}\in\mathcal{T}} \text{BER}\Big(\mathbf{w},\, D\big(\mathcal{A}(\tilde{\mathbf{X}})\big)\Big) \leq \rho \tag{18}$$

where D represents the decoding function, and $\text{BER}(\mathbf{w}, \hat{\mathbf{w}}) = \frac{1}{l}\sum_{i=1}^{l} \mathbb{1}_{(\mathbf{w}_i \neq \hat{\mathbf{w}}_i)}$ measures the fraction of mismatched bits.

Radioactivity guarantees the persistent propagation of the watermark through model extraction attacks. It ensures that if an adversary trains a pirate model Υ using a dataset $\mathcal{C}_m = \{\tilde{\mathbf{X}}^{(i)}\}_{i=1}^{N}$ derived from watermarked structures, the signature is effectively encoded into the surrogate's parameters. Consequently, any structure $\mathbf{X}_{\text{out}}$ generated by this derivative model $\Upsilon(\cdot)$ must inherently carry the watermark signal, satisfying the condition:

$$\sup \text{BER}\Big(\mathbf{w}, D(\mathbf{X}_{\text{out}})\Big) \leq \rho \tag{19}$$

Covertness mandates that the embedded watermark remains statistically imperceptible to adversarial inspection. The watermarked distribution p_m must be effectively indistinguishable from the clean model distribution p_θ. Formally, this requires minimizing the divergence between the two distributions as measured by an optimal discriminator $\mathcal{C}$:

$$\max_{\mathcal{C}} \left| \mathbb{E}_{\hat{\mathbf{X}}\sim p_\theta}[\mathcal{C}(\hat{\mathbf{X}})] - \mathbb{E}_{\tilde{\mathbf{X}}\sim p_m}[\mathcal{C}(\tilde{\mathbf{X}})] \right| \approx 0 \tag{20}$$

In the context of protein structures, this implies that key geometric statistics, namely bond lengths, torsion angles, and residue packing densities, must strictly adhere to natural biological distributions, thereby preventing detection via statistical anomaly analysis.

Capacity requires the length of the watermark must be sufficient to carry the identification information.

Efficiency requires that the watermark embedding and extraction processes must not consume too much time.

Generalizability ensures that the watermarking algorithm is agnostic to specific model architectures and protein topologies.

B CERTIFIED ROBUSTNESS VIA RANDOMIZED SMOOTHING

While we empirically demonstrate robustness against various attacks, we further provide a theoretical guarantee for watermark persistence using Randomized Smoothing. This framework certifies

that the watermark decoder D will reliably recover the correct bit string m under any adversarial perturbation δ bounded by a radius r (in terms of L_2 norm or Root Mean Square Deviation (RMSD)).

Let $D_i : \mathbb{R}^{3N} \to \{0, 1\}$ be the base decoder for the i-th watermark bit. We define a smoothed classifier g_i as the majority vote of D_i under isotropic Gaussian noise $\mathcal{N}(0, \sigma^2 I)$:

$$g_i(X) = \underset{c \in \{0,1\}}{\operatorname{argmax}} \mathbb{P}_{\epsilon \sim \mathcal{N}(0,\sigma^2 I)}(D_i(X + \epsilon) = c) \tag{21}$$

Theorem 1 (Certified Radius). *Let p_A be the probability of the most likely class (the correct bit) predicted by D_i under noise distribution $\mathcal{N}(0, \sigma^2 I)$, and let $p_B = 1 - p_A$. If $p_A > 0.5$, then $g_i(X)$ is provably robust against any perturbation δ satisfying $||\delta||_2 < r$, where:*

$$r = \frac{\sigma}{2}(\Theta^{-1}(p_A) - \Theta^{-1}(p_B)), \tag{22}$$

where Θ^{-1} is the inverse standard normal CDF.

Connection to RMSD. In the context of protein structures with N atoms, an L_2-bounded perturbation relates to the RMSD as RMSD $= ||\delta||_2/\sqrt{N}$. Therefore, for a certified L_2 radius r, MARCO guarantees that the watermark bit m_i will not flip for any attack that modifies the structure within an RMSD threshold of:

$$R_{\text{RMSD}} = \frac{\sigma}{2\sqrt{N}}(\Theta^{-1}(p_A) - \Theta^{-1}(1 - p_A)) \tag{23}$$

This theoretical bound confirms that the robustness of MARCO is not merely an artifact of specific attack simulations, but an intrinsic geometric property enforced by the noise-aware training of the decoder.

C Experiment Settings

Pre-trained Protein Generative Models. We evaluate our proposed method across a diverse suite of state-of-the-art protein generative models, ranging from structure prediction networks to de novo backbone generators.

- **RFDiffusion**Watson et al. (2023) : A leading method for de novo protein design that combines the RoseTTAFold structure prediction network with a Denoising Diffusion Probabilistic Model (DDPM) framework. It is widely used for designing motif scaffolds and ligand binders.
- **ESMFold** Lin et al. (2022): A high-speed structure prediction model based on the ESM protein language model. Unlike AlphaFold, it generates atomic-level structures end-to-end directly from sequence data without requiring explicit Multiple Sequence Alignment (MSA) retrieval, making it approximately ten times faster than MSA-based methods.
- **Chroma** Ingraham et al. (2023): A diffusion-based generative model designed to enable the creation of novel programmable protein structures and functions, pushing the boundaries of de novo design beyond natural protein space.
- **FrameDiff** Yim et al. (2023b): An $SE(3)$-invariant diffusion model specifically designed for protein backbone generation. It operates on rigid bodies (frames) in 3D space rather than just point clouds, allowing it to generate designable protein monomers up to 500 amino acids long without relying on pre-trained folding networks.
- **FrameFlow** Yim et al.: A flow-matching model that adapts the FrameDiff architecture to a deterministic generative paradigm. By utilizing $SE(3)$ flow matching, it significantly reduces the number of sampling steps required for inference while maintaining the geometric rigor of frame-based updates9.
- **FoldFlow2** Huguet et al. (2024): A state-of-the-art generative model that integrates sequence encoding (from ESM-2) with structure encoding (via flow matching) in a multi-modal trunk. Trained on a dataset eight times larger than its predecessor, it effectively addresses the interdependence of sequence and structure to generate diverse and high-quality proteins.

Datasets. Following the data curation protocols of RFdiffusion Watson et al. (2023), we utilize a subset of monomeric structures derived from the Protein Data Bank (PDB) Berman et al. (2000). The inclusion criteria require sequence lengths ranging from 60 to 512 residues and a structural resolution finer than 5Å. Additionally, structures exhibiting greater than 50% loop content are excluded to ensure structural integrity. The final curated dataset comprises 20,312 protein entries.

Implementation. For the encoder, decoder, and discriminator modules within MARCO, we adopt a 6-layer architecture adapted from FrameDiff Yim et al. (2023b). This architecture integrates Invariant Point Attention (IPA) Jumper et al. (2021), Transformer blocks Vaswani et al. (2017), and Multi-Layer Perceptrons (MLP). The IPA module, configured with 16 hidden channels, processes spatial relationships while strictly preserving rotational invariance. Following this, Transformer blocks process the node embeddings to capture global contextual dependencies. Finally, the MLP blocks, composed of three linear layers with ReLU activation and LayerNorm, perform the final node and edge feature updates.

Model training is conducted using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 64. The training protocol spans 50 epochs, during which the weights of the pre-trained PGM remain frozen; optimization is exclusively applied to the encoder, decoder, and discriminator parameters. All experiments were executed on a high-performance server equipped with 8 NVIDIA A100 GPUs (80GB VRAM each). The weight parameters λs are determined by grid search in pre-experiments.

Watermark Baselines. To provide a rigorous comparative analysis, we adapt state-of-the-art watermarking methods originally developed for image diffusion models (e.g., Stable Diffusion) to the protein domain.

- **AquaLoRA** Feng et al. (2024) embeds watermarks directly into the diffusion U-Net via a Low-Rank Adaptation (LoRA) module using a two-stage process: Latent Watermark Pre-training and Prior Preserving Fine-Tuning.
- **WaDiff** Min et al. (2024) integrates user-specific watermarks directly into the diffusion sampling process. The method functions as a "watermark-conditioned" diffusion model, where the watermark is treated as a conditioned input by expanding the model's input channels.
- **SleeperMark** Wang et al. (2025b) explicitly disentangles watermark information from semantic concepts, guiding the model to embed a multi-bit message only when triggered, while regular prompts yield watermark-free, high-fidelity outputs.
- **FoldMark** Zhang et al. (2025b) trains an encoder-decoder and fine-tunes each PGM using watermark-conditioned LoRA, evolutionary information, and structural consistency losses to make the output carry the watermark.

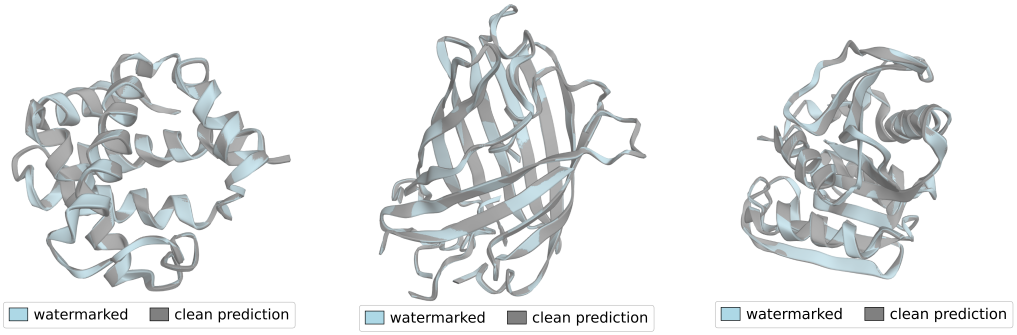
We follow the hyperparameter settings in the original papers.

Metrics. For the two most important properties: fidelity and robustness, we use the following metrics to measure them:

- **Structure Similarity**: Root Mean Square Deviation (RMSD) is the standard measure of the average distance between the atoms of superimposed proteins.

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2}, \tag{24}$$

  where N is the number of atoms and d_i is the distance between the i-th pair of corresponding atoms. A lower RMSD indicates a better structural similarity, and a value less than 2.0Å is usually considered a very close match.
- **Sequence Similarity**: Instead of directly computing match over no match, we employ BLOSUM62 Henikoff & Henikoff (1992) to compute a substitution matrix to score aligned pairs between the original sequence and the translated sequence derived by feeding the watermarked structure into the ProteinMPNN Dauparas et al. (2022). In short, for identical pairs or mutations that are not going to significantly affect the protein structure, BLOSUM62 gives positive scores, whereas negative scores are given to the mutations significantly altering the protein structure.



(a) 1A3N – Deoxy Human Hemoglobin (b) 1EMA – Green Fluorescent Protein (c) 4AO6 – Cold-Adapted Esterase

Figure 5: **Visualization of the Clean Prediction vs. Watermarked Structure**. The conformation of the watermarked structure highly aligns to the clean prediction, demonstrating the high fidelity of MARCO.

However, raw BLOSUM62 scores scale with the length of the protein, where a longer protein yields a higher score. Thus, they cannot be directly compared across different pairs. To measure similarity between different pairs, we normalize the score into a range $[0, 1]$. We employ the Geometric Mean Normalization to derive the normalized BLOSUM62 score

$$S_{\text{norm}}(\mathbf{S}_a, \mathbf{S}_b) = \frac{S(\mathbf{S}_a, \mathbf{S}_b)}{\sqrt{S(\mathbf{S}_a, \mathbf{S}_a) \times S(\mathbf{S}_b, \mathbf{S}_b)}}, \tag{25}$$

where $S(\cdot, \cdot)$ denotes the raw BLOSUM62 score between two protein sequences. A high $S_{\text{norm}}(\mathbf{S}_a, \mathbf{S}_b)$ indicates that the two sequences $\mathbf{S}_a$ and $\mathbf{S}_b$ are biologically similar, whereas a low value means the opposite.

- **Watermark Bit Accuracy**: We simply count the number of matching bits between the original and the recovered watermarks, and divide it by the watermark length to get the Bit Accuracy (BA):

$$\text{BA} = \frac{\text{CountMatch}(\mathbf{w}, \hat{\mathbf{w}})}{l} \tag{26}$$

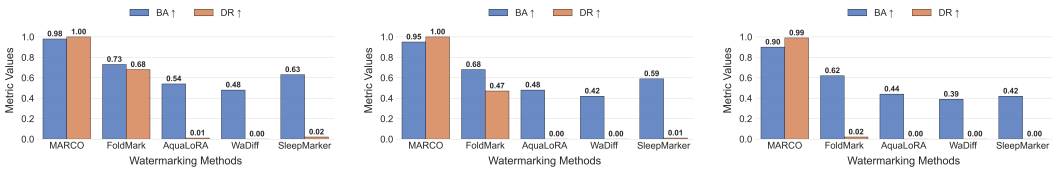
- **Watermark Detection Rate**: We compute the Z-score as described in Subsection 4.3, and if $z \geq 3$ (BA ≥ 0.6875), it is considered a successful detection. The detection rate (DR) is derived by dividing the number of successful detections by the total number of detections.

D Additional Experiments

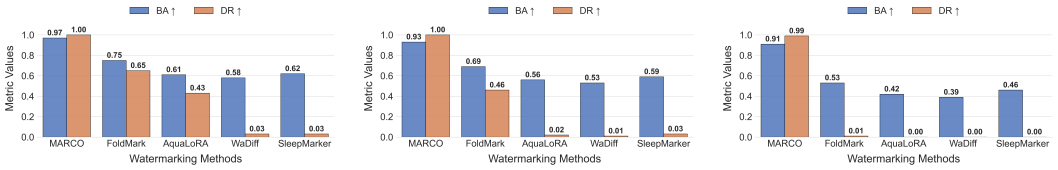
D.1 Robustness

MARCO vs. Energy-Based Scrubbing. To simulate an adaptive attacker, we employ an energy-based scrubbing attack implemented via OpenMM Eastman et al. (2013). We apply a harmonic positional restraint $E_{\text{rest}} = \frac{k}{2} \sum_i \|\mathbf{x}_i - \mathbf{x}_i^0\|_2^2$, where $\mathbf{x}_i^0$ denotes the initial watermarked coordinate and k is the restraint stiffness in $\text{kJ mol}^{-1}\,\text{nm}^{-2}$. We evaluate MARCO's robustness by generating a watermarked dataset containing 1000 watermarked structures using RFdiffusion. These structures are subjected to scrubbing across a stiffness range of $k \in \{0, 50, 100, 500, 1000, 5000\}$ and subsequently used to fine-tune a FrameDiff pirate model. As illustrated in Figure 7, MARCO demonstrates impressive robustness to this adaptive threat. Crucially, although extremely low k values can compromise watermark recovery, they simultaneously destroy structural fidelity, rendering the resulting pirate model functionally unusable.

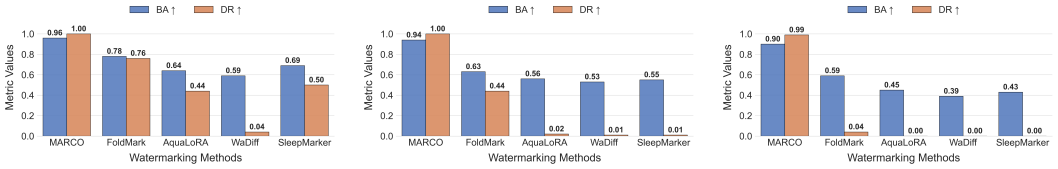
Additionally, we evaluate MARCO under finite-temperature molecular dynamics using 100-ns OpenMM simulations of the same watermarked structures. Each protein is parameterized with the AMBER ff14SB force field, solvated in a TIP3P water box with a 1.0-nm buffer and 0.15-M NaCl, energy-minimized, equilibrated for 100 ps under NVT and 1 ns under NPT, and simulated without positional restraints at 300 K and 1 bar using a Langevin thermostat, Monte Carlo barostat,



(a) Θ : FrameDiff, Rotation (b) Θ : FrameDiff, Cropping (c) Θ : FrameDiff, Noise Addition



(d) Θ : FrameFlow, Rotation (e) Θ : FrameFlow, Cropping (f) Θ : FrameFlow, Noise Addition



(g) Θ : FoldFlow2, Rotation (h) Θ : FoldFlow2, Cropping (i) Θ : FoldFlow2, Noise Addition

Figure 6: Part B – **Evaluation of Robustness of PGM Watermarking Methods against Conformation Manipulations**. After conformation manipulations, MARCO survives and outperforms the other four baselines with impressive robustness.

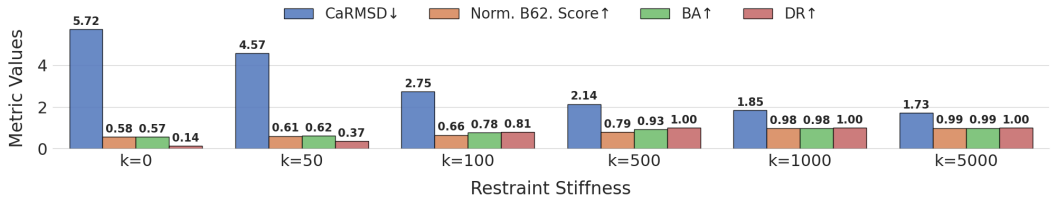


Figure 7: **MARCO vs. Energy-Based Scrubbing**. A sufficiently low k results in a compromised watermark but at a non-negligible cost of fidelity.

PME electrostatics, a 1.0-nm nonbonded cutoff, constrained hydrogen bonds, and a 2-fs timestep. Snapshots are collected every 10 ns and evaluated using C_α-RMSD, normalized BLOSUM62, BA and DR. As depicted in Figure 8, as time flows, both the fidelity and robustness of the watermarked structures decrease, because the shape of the structures slightly changes to reach a lower-energy state. However, the performance reaches a balance from 80ns, when the structures become stable. This suggests that MARCO is robust against molecular dynamic simulations.

MARCO vs. DPSGD. We further simulate an adaptive attacker utilizing local regularization techniques to remove the watermark. Specifically, we employ DPSGD Abadi et al. (2016) to fine-tune FrameDiff on the watermarked dataset. We fixed $\delta = 10^{-5}$ and varied the privacy budget across $\epsilon \in \{0.5, 1.0, 2.0, 4.0, 8.0, 16.0\}$. As illustrated in Figure 9, MARCO maintains its robustness with BA ≥ 0.91 and DR ≥ 0.99 when the privacy budget $\epsilon \geq 4.0$. Similar to the energy-based scrubbing results, low ϵ may degrade the watermark signal. However, this comes at the cost of severe fidelity loss, rendering the resulting pirate model functionally unusable.

MARCO vs. Refolding. In this scenario, an adaptive attacker queries the PGM protected by MARCO to unconditionally generate de novo protein structures. The structures are then mapped back to sequences using some mapping functions. The sequences are fed into a clean PGM to remove the watermark inside. To simulate this attack, we employ RFDiffusion as the original PGM

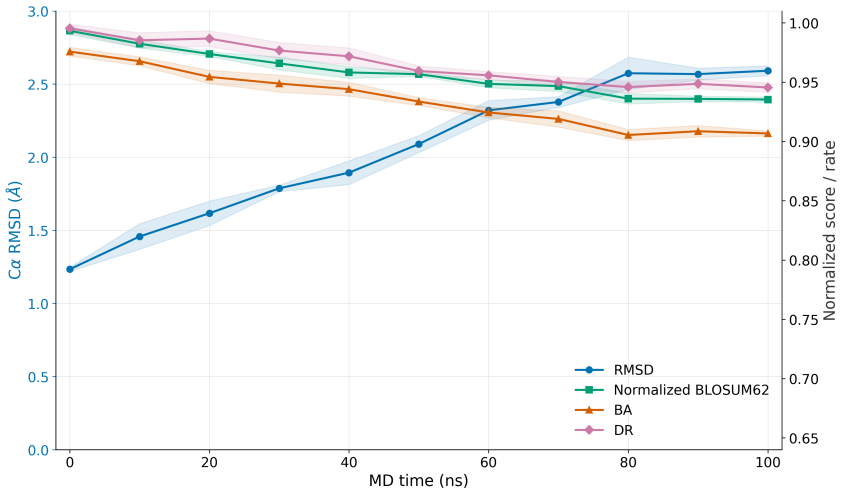


Figure 8: **MARCO vs. Molecular Dynamic Simulation**

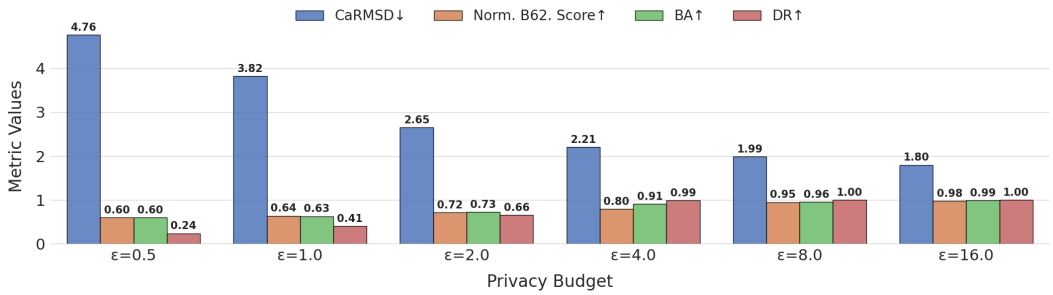


Figure 9: **MARCO vs. DPSGD**. Even if the watermark is removed with a low privacy budget, the pirate model greatly sacrifices its fidelity.

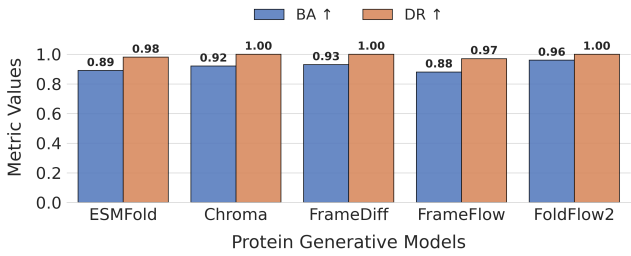


Figure 10: **MARCO vs. Refolding**. By selectively altering the translated sequences, MARCO gains sufficient robustness against refolding attack.

protected by MARCO to generate 1000 de novo watermarked structures. The structures are then mapped back to sequences using ProteinMPNN Dauparas et al. (2022) and refolded into protein structures using the other five clean pre-trained PGMs. We derive the metrics values by extracting the recovered watermark from the refolded structures.

The experimental results are depicted in Figure 10, where MARCO shows sufficient robustness against refolding attack. As demonstrated in the previous experiments, MARCO does not present perfect sequential fidelity, because MARCO trades off this attributes to gain resistance against refolding attack. The watermark signal affects the sequences translated from the watermarked structure such that structures generated by other clean PGMs given these sequences inherit the watermark.

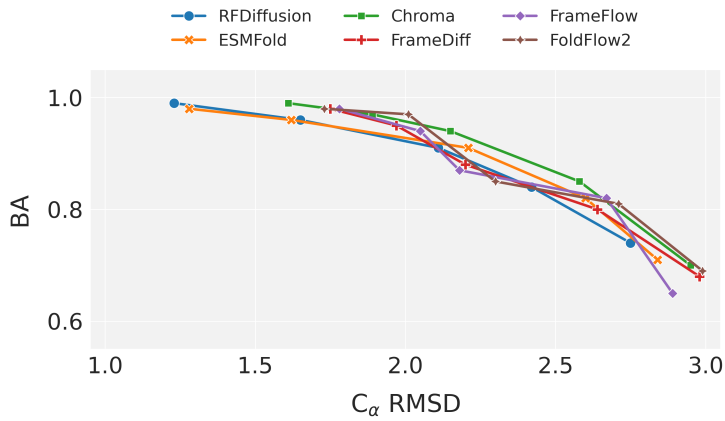


Figure 11: **Certifiable Robustness**. The robustness of MARCO is an intrinsic noise-aware geometric property.

Table 1: **MARCO vs. Dilution.**

Watermarked Data Ratio	BA↑	DR↑
1.00	0.94	1.00
0.75	0.88	0.99
0.50	0.81	0.92
0.25	0.78	0.85
0.10	0.73	0.68

Certified Robustness. To prove the certified robustness of MARCO, we conduct an experiment where we use different MARCO-guarded PGMs to generate 1000 watermarked structures. The watermarked structures are then perturbed with random noises multiplied by different weights. Next, we extract the watermark and compute the RMSD between the ground truth and the perturbed structures and the BA to derive a curve of the correlation between RMSD and BA.

In Figure 11, we observe that the robustness of MARCO is not just an artifact of attack simulation, but an intrinsic noise-aware property. The average BA of all PGMs remains above 0.8 when RMSD < 2.5, which suffices for watermark verification. Moreover, this relation between RMSD and BA shows that the adversary cannot guarantee watermark removal without significantly downgrade the fidelity. This indicates that the manifold of the watermarked structures reflects the watermark itself.

MARCO vs. Dilution. We evaluate MARCO when faced with fine-tuning attacks that attempt to remove the watermark by fine-tuning the pirate model with watermarked structures mixed with clean structures. We fine-tuned a FrameDiff model using varying ratios of MARCO-watermarked structures to clean structures. As shown in Table 1, the ratio does affect the effectiveness of MARCO. However, with only 25% watermarked data, MARCO achieves 0.78 BA and 0.85 DR, demonstrating its strong robustness. Additionally, we conduct another dilution experiment where we replace the clean structures with watermarked structures generated via FoldMark. The results in Table 2 demonstrate that the watermarked structures do not differ from the clean structures in terms of removing MARCO's watermark.

MARCO vs. Geometry-Aware Baselines. We elevate the level of robustness evaluation of MARCO by comparing it with a geometry-aware baseline. We compare MARCO with MarkSplatter Huang et al. (2025), a SOTA 3D point-cloud watermark. We adapted MarkSplatter by treating the 3D atomic coordinates of 1,000 RFdiffusion-generated proteins as unstructured point clouds to fit into its workflow. The results are listed in Table 3, from which we conclude that MARCO out-

Table 2: **MARCO vs. FoldMark Dilution.**

MARCO Data Ratio	BA↑	DR↑
1.00	0.92	0.99
0.75	0.85	0.97
0.50	0.82	0.92
0.25	0.75	0.73
0.10	0.71	0.65

Table 3: **MARCO vs. Geometry-Aware Baselines.**

Methods \ Metrics	C_αRMSD↓	Norm. B62↑	BA↑	DR↑	BA(Rotation)↑	DR(Rotation)↑	BA(Crop)↑	DR(Crop)↑
MARCO	1.23	0.99	0.98	1.00	0.97	1.00	0.93	0.99
MarkSplatter	2.71	0.74	0.95	1.00	0.73	0.68	0.63	0.35

performs MarkSplatter in watermarking PGMs due to its original design aiming at protein structure prediction.

D.2 Radioactivity

MARCO vs. Various Architectures. To evaluate radioactivity, we randomly create a subset of 1,000 sequences from the training data. We then generate watermarked datasets using RFdiffusion, ESMFold, and Chroma as the source PGMs Θ with a fixed watermark. Subsequently, we fine-tune three computationally efficient pirate models Υ including FrameDiff, FrameFlow, and FoldFlow2 using these paired sequences and watermarked structures. Finally, we quantify radioactivity by extracting the recovered watermark from the outputs of Υs.

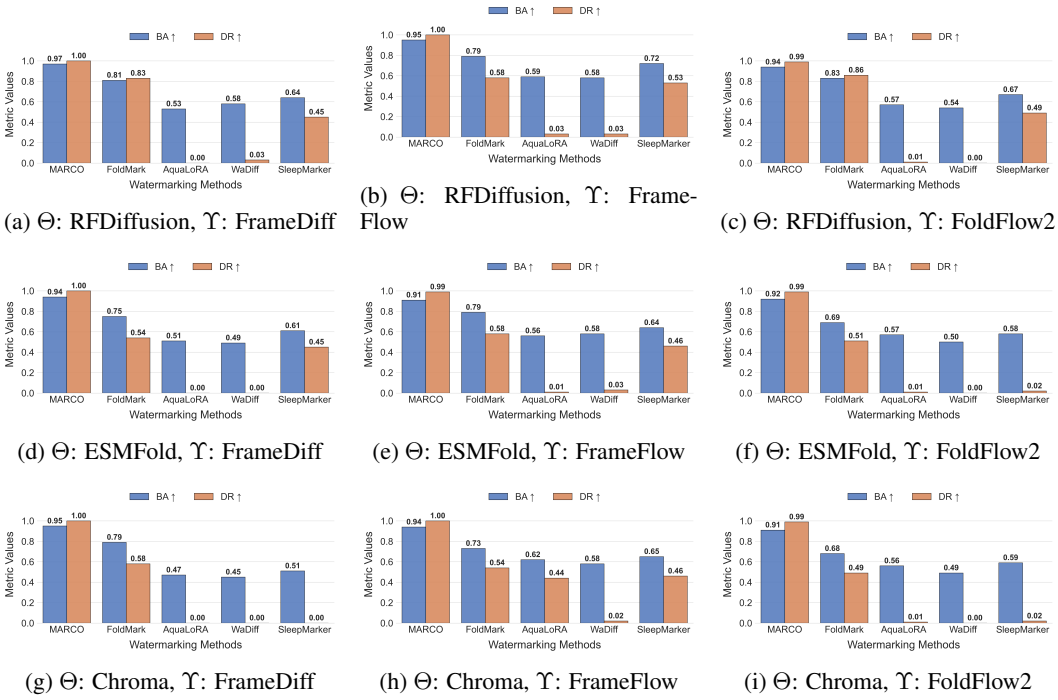
As illustrated in Figure 12, MARCO demonstrates superior radioactivity, consistently achieving BA ≥ 0.90 and DR ≥ 0.99. This confirms that MARCO effectively propagates the watermark signal to the pirate model, regardless of the architectural disparities between the source and pirate models. In contrast, the baselines fail to exhibit comparable radioactivity. Their sporadic success likely stems from the pirate models learning simple distributional shifts introduced during the baselines' fine-tuning process. However, MARCO's robust performance indicates that the bias injected iteratively during the reverse diffusion process establishes a more persistent and transferable signature.

D.3 Capacity

To assess the capacity of MARCO, we varied the watermark length l across the set $\{32, 48, \ldots, 128\}$ and evaluated the impact on fidelity and robustness. As shown in Figure 13, MARCO effectively encodes identity information without compromising performance. Both the normalized BLOSUM62 score and DR remain stable as l increases, indicating that the induced perturbations are sufficiently redundant to preserve sequential similarity and detection reliability. However, RMSD and Bit Accuracy (BA) exhibit greater sensitivity; specifically, for $l \geq 96$, we observe significant fluctuations, suggesting a saturation threshold where capacity begins to degrade fidelity. While the reduced BA at high l values still supports basic detection, it may compromise precise user identification. Nonetheless, the default length of $l = 64$ offers an optimal trade-off, providing a vast coding space with robust fidelity.

D.4 Ablation Study

Impact of the Loss Terms and the Attack Layer. To quantify the contribution of individual components, we conduct an ablation study by sequentially setting the weights $\lambda = 0$ and disabling the Attack Layer $\mathcal{A}$. We utilize RFdiffusion as the fixed PGM and evaluate robustness specifically under noise addition, the most disruptive attack observed in prior experiments.



(a) Θ: RFDiffusion, ϒ: FrameDiff (b) Θ: RFDiffusion, ϒ: FrameFlow (c) Θ: RFDiffusion, ϒ: FoldFlow2

(d) Θ: ESMFold, ϒ: FrameDiff (e) Θ: ESMFold, ϒ: FrameFlow (f) Θ: ESMFold, ϒ: FoldFlow2

(g) Θ: Chroma, ϒ: FrameDiff (h) Θ: Chroma, ϒ: FrameFlow (i) Θ: Chroma, ϒ: FoldFlow2

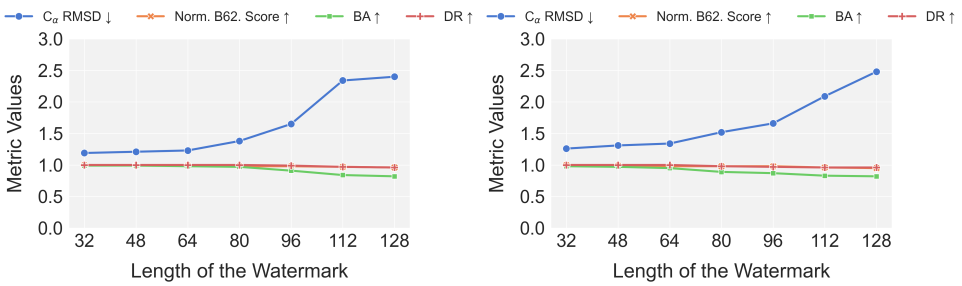
Figure 12: **Evaluation of Radioactivity of PGM Watermarking Methods with Various Model Architectures**. MARCO shows superior radioactivity by transferring its watermark signal to models trained on its watermarked data regardless of model architectures.

As presented in Table 5, every component is critical to MARCO's performance. The watermark loss $\mathcal{L}_{\text{wm}}$, which constrains perturbation magnitudes, is essential for fidelity. Its removal causes significant structural degradation, even though robustness remains high. Similarly, omitting the distance matrix Loss $\mathcal{L}_{\text{dm}}$ and torsion angle Loss $\mathcal{L}_{\text{tor}}$ results in substantial fidelity drops. The extraction loss $\mathcal{L}_{\text{ext}}$ is foundational; without it, robustness vanishes entirely. The adversarial loss $\mathcal{L}_{\text{adv}}$ serves a dual role, enhancing both fidelity and robustness. Finally, the attack layer proves vital for generalization. Its exclusion leaves the watermark highly vulnerable to conformational manipulations.

Distributional Indistinguishability. We further measure the distributional indistinguishability between clean predictions and watermarked structures using the Fréchet Distance (FD). We trained two MARCOs with and without $\mathcal{L}_{\text{adv}}$. We generated 1,000 triplets of unwatermarked and watermarked proteins via RFdiffusion using the two MARCO variants to compute the FD. With $\mathcal{L}_{\text{adv}}$, the FD is 2.27, whereas the FD increases to 16.31 without it. This suggests that $\mathcal{L}_{\text{adv}}$ significantly improves covertness.

Impact of the Training Epochs. We further investigate the impact of training duration on MARCO's performance by varying the epoch count under identical experimental conditions. As detailed in Table 4, performance gains diminish significantly beyond 25 epochs. Notably, the model exhibits negligible improvement when scaling from 50 to 100 epochs, indicating convergence. Consequently, we determine that 50 epochs are sufficient to achieve optimal performance stability.

Inference Overhead. Regarding computational overhead, training MARCO for 50 epochs requires approximately 16.71 hours on a single NVIDIA A100 GPU. This translates to a cloud computing cost of $57.32 (based on standard AWS pricing), underscoring the framework's training efficiency and economic feasibility for deployment. Moreover, we measure the inference overhead of MARCO and present the results in Table 6. We discover that the ratio of the increased time to the clean reference time is low and is therefore trivial to the protein structure generation.



(a) Θ: RFDiffusion (b) Θ: ESMFold

(c) Θ: Chroma (d) Θ: FrameDiff

(e) Θ: FrameFlow (f) Θ: FoldFlow2

Figure 13: **Evaluation of Capacity of MARCO**. As the length of the watermark increases, both the fidelity and robustness drops.

Table 4: **Impact of Training Epochs**.

Epochs \ Metrics	C_αRMSD↓	Norm. B62↑	BA↑	DR↑
5	2.44	0.71	0.62	0.17
10	1.62	0.92	0.84	0.96
25	1.41	0.97	0.96	1.00
50	1.23	0.98	0.98	1.00
100	1.23	0.98	0.99	1.00

Negative Control. We evaluate MARCO's specificity against negative structures. We make the trained watermark decoder extract a pre-defined 64-bit watermark from 100 natural structures and clean predicted structures to compute the BA and DR. The results in Table 7 show that MARCO

Table 5: **Impact of Loss Terms and the Attack Layer**.

Terms \ Metrics	C_αRMSD↓	Norm. B62↑	BA↑	DR↑
With All	1.23	0.99	0.98	1.00
W/O $\mathcal{L}_{\text{wm}}$	2.13	0.79	0.97	1.00
W/O $\mathcal{L}_{\text{dm}}$	1.63	0.92	0.98	1.00
W/O $\mathcal{L}_{\text{tor}}$	1.56	0.87	0.97	1.00
W/O $\mathcal{L}_{\text{ext}}$	1.25	0.99	0.02	0.00
W/O $\mathcal{L}_{\text{adv}}$	1.34	0.97	0.92	0.99
W/O $\mathcal{A}$	1.22	0.99	0.05	0.00

Table 6: **Inference Overhead (s/structure)**.

Model	Clean Inference Time	MARCO Inference Time
RFDiffusion	28.8	32.2
ESMFold	8.7	8.9

Table 7: **Negative Control Evaluation.**

Structure Type	BA ↓	DR ↓
Natural Structures	0.48	0.00
Clean Predicted Structures	0.49	0.00

cannot falsely retrieve the pre-defined watermark from the natural or the clean predicted structures, demonstrating strong specificity.

E Limitations & Future Work

Currently, MARCO is specialized for diffusion-based architectures. Auto-regressive Protein Language Models (PLMs) such as ESM3 Hayes et al. (2025), SaProt Su et al., and BioMistral Labrak et al. (2024) remain outside our current scope. However, inspired by Kirchenbauer et al. Kirchenbauer et al. (2023), we can design a new paradigm for PLM watermarking by dynamically dividing the amino acid dictionary into a red and a green set with some biological constraints. Additionally, we have not yet conducted wet-lab experiments to validate the recovery of watermarks from physically synthesized and re-digitized proteins. Future research will focus on generalizing MARCO to discrete modalities such as PLMs and the latest genome language model AlphaGenome Avsec et al. (2026), and adapting to emerging architectures like BoltzGen Stark et al. (2025).

For wet-lab verification of a synthesized protein, we envision two complementary approaches. First, after confirming the sequence by DNA sequencing and LC–MS/MS, the protein structure could be resolved using X-ray crystallography, cryo-EM, or NMR and processed by MARCO's decoder to recover the registered watermark Durbin & Feher (1996). Second, a lower-cost antibody-based assay could be enabled by concentrating part of the watermark into a biologically tolerant, surface-exposed conformational epitope. A specific monoclonal antibody could then detect the watermarked protein through native ELISA, immunoprecipitation, or surface plasmon resonance Davies et al. (1990).