

Function-preserving watermarking of AI-generated proteins

<https://doi.org/10.1038/s41586-026-10965-y>

Received: 29 September 2025

Accepted: 29 July 2026

Published online: 30 September 2026

Open access

 Check for updates

David Stutz^{1,2}, Alexander I. Cowen-Rivers^{1,2}, Guillermo Ortiz-Jimenez¹, Jeremy Ratcliff¹, Vinicius Zambaldi¹, Lindsay Willmore¹, Josh Abramson¹, Harshnira Patani¹, Christina Kouridi¹, Florian Stimberg¹, Mel Vecerik¹, Alex Chu¹, Sukhdeep Singh¹, Sumanth Dathathri¹, Eliseo Papa¹, Valentin De Bortoli¹, Arnaud Doucet¹, Demis Hassabis¹, Jue Wang¹, Sven Goyal¹ & Pushmeet Kohli¹

Generative artificial intelligence (AI) models are revolutionizing biology, with tools such as AlphaFold 3 and protein design models accelerating breakthroughs in protein structure prediction and the creation of new functional proteins¹. Tracking and establishing the provenance of AI-generated protein sequences and structures is becoming increasingly important to tackle a range of emerging challenges, including biosecurity and concerns about information veracity^{2–4}. Here we introduce SynthIDBio, a family of methods for watermarking protein sequences and structures to establish the provenance of those generated with AI. SynthIDBio-sequence actively embeds a watermark into protein sequences while preserving function. We demonstrate this by creating watermarked, functional designed protein binders with binding affinity comparable with non-watermarked counterparts and near-perfect watermark detection accuracy. Furthermore, SynthIDBio-structure, a fine-tuned AlphaFold 3 model, embeds an imperceptible watermark into biomolecular structures. Our work is a proof-of-concept that function-preserving biological watermarking is feasible, introducing a potential tool for provenance in the rapidly expanding era of AI-driven biological engineering.

AI is accelerating biological discovery and design, with wide-ranging applications in therapeutics, diagnostics and biomedical research⁵. Specialist models for protein design and structural predictions are becoming an integral part of biological workflows. Ensuring their safe and responsible use involves the ability to determine and track the original model of AI-generated sequences and structures. This holds promise to address potential challenges that stem from the widespread adoption of ever more powerful AI models in biology, including biosecurity and information veracity. Here, information on provenance may become a useful signal for DNA synthesis providers^{2,4,6–8} and institutions managing biological databases such as the Protein Data Bank (PDB)⁹, UniProt¹⁰ or GenBank¹¹.

Several means of establishing provenance for biological objects have been proposed, including through centralized databases or by associating metadata about the design process with the biological sequence^{2,12}. Databases require central coordination, are prone to false positives at scale and may cause privacy challenges. Cryptographically signed metadata, by contrast, can preserve privacy, but is easy to remove. An alternative approach, watermarking, embeds signatures into the AI-generated objects themselves. This approach has successfully been applied to AI-generated text and multimedia and deployed in various products, including Google's generative AI^{13,14}. Watermarks promise high detectability and quality preservation by being imperceptible. They offer a privacy-preserving alternative where information on provenance is more challenging to remove and no central coordination

is required. Moreover, watermarking can establish provenance for open model predictions. However, it is unclear whether existing watermarking methods transfer to biological modalities where preserving quality relates to the protein's intended use: for protein design, this requires maintaining experimentally verifiable biological function; for structural prediction, it involves maintaining the accuracy of predicted structural features, ensuring a researcher's interpretation and subsequent biological function hypotheses remain unchanged. Concurrent work for watermarking protein sequences is limited to *in silico* experiments with wild-type monomer structures from PDB³. Current methods result in a noticeable decrease in accuracy¹⁵ when watermarking protein structures. Combined, these show that function-preserving watermarking is not yet available.

We introduce SynthIDBio, a family of methods for embedding highly detectable yet function-preserving watermarks directly into biological sequences and structures. SynthIDBio-sequence is a method for designing watermarked protein binders. Following Fig. 1a, SynthIDBio-sequence integrates watermarking into ProteinMPNN¹⁶, which is a commonly used sequence design model applied on top of a backbone structure before structural metrics are used to filter designs to predicted binders^{17,18}. We show that detectability of the watermark can be ensured with minimal *in silico* performance reductions and explore the watermark's robustness to intentional removal. Through *in vitro* validation, we obtained biologically watermarked, functional protein binders: low nanomolar binders for the SARS-CoV-2 receptor

¹Google DeepMind, London, UK. ²These authors contributed equally: David Stutz, Alexander I. Cowen-Rivers. ✉e-mail: dstutz@google.com; pushmeet@google.com

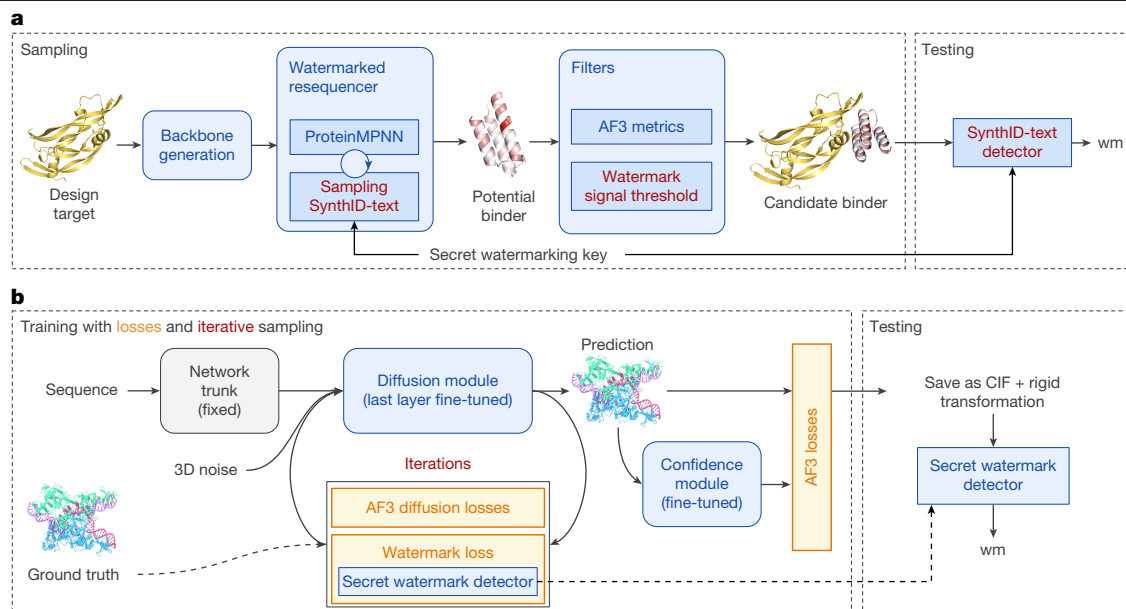


Fig. 1 | Graphical overviews of SynthIDBio-sequence and SynthIDBio-structure. a, End-to-end protein design with SynthIDBio-sequence. During sampling, based on a generated backbone, we use ProteinMPNN with SynthID-text's tournament sampling to sample watermarked sequences that are then filtered using metrics derived from AF3 and a further watermark (wm) detectability threshold that is calibrated to a target FPR. At test time, we use the secret watermarking key used during sampling for watermark detection. **b**, We fine-tune the diffusion and confidence modules of AF3 (blue) to obtain

SynthIDBio-structure, which includes an imperceptible watermark within each structure prediction. We also train a watermark detector (blue) by adding an additional watermarking loss (orange) to AF3's diffusion losses. The sampling process remains unchanged using the fine-tuned diffusion module. At test time, the watermark detector can be used in a standalone fashion to identify watermarked structures. The detector remains secret and is not shared with users of the fine-tuned AF3 model.

binding domain (SC2RBD) and subnanomolar binders for vascular endothelial growth factor A (VEGF-A) and programmed death ligand 1 (PD-L1). Our experimental design leveraged known backbones from AlphaProteo¹⁷, which were resequenced using ProteinMPNN with or without SynthIDBio-sequence. Across targets and backbones, we find that watermarking does not affect the binding affinity distribution or binding hit rates.

We also present SynthIDBio-structure—a model fine-tuned from AlphaFold 3 (AF3)¹⁹ that predicts watermarked structures while preserving structural accuracy. Following Fig. 1b, we fine-tune AF3's diffusion module, co-training it with a watermark detector that is guided by a watermarking objective integrated directly into AF3's diffusion loss. SynthIDBio-structure offers near-perfect detectability with a negligible effect on global accuracy and a minimal effect on bond geometry, as validated on AF3's evaluation set. The watermarks are robust to basic manipulations including rigid transformations and noise. Qualitative validation shows that SynthIDBio-structure successfully hides watermarking information in atom–atom distances, indistinguishable from the natural variance of AF3's own predictions relative to the ground truth.

SynthIDBio presents a technical proof-of-concept that function- and quality-preserving biological watermarking is possible. We also discuss potential use-cases, including for biosecurity and scientific integrity, where we expect the ability to reliably track provenance to be important in the future; however, operationalizing SynthIDBio for these applications will require further innovation in addition to industry-wide coordination and standardization.

Tournament sampling for ProteinMPNN

Recent approaches for protein binder design rely on target-conditional structure generation followed by sequence design using ProteinMPNN^{16–18,20} (Fig. 1a). ProteinMPNN, with left-to-right decoding, generates discrete sequences of amino acids autoregressively by sampling tokens (residues) x_t at each step t , from a distribution

$p(\cdot|x_{<t})$ conditioned on the preceding context. To watermark sampled sequences, we apply the sampling strategy from SynthID-text, called tournament sampling, which biases the sampling distribution $p(\cdot|x_{<t})$ using a random seed r_t derived from the preceding context and a secret watermarking key k . Specifically, the preceding context can be H -grams, that is, $x_{t-H}, \dots, x_{t-1}$, and the random seed^{21,22} is derived from a hashing function $f_r(x_{t-H:t-1}, k)$ applied to the H -gram and k . The distribution is biased by using so-called scoring functions (also called g -functions, $g(x_t, r_t)$) that, conditioned on the random seed, assign a binary score to every possible token x_t . Tournament sampling explicitly maximizes these g -values. With knowledge of the secret key k , the watermark can be detected by recomputing H -grams and evaluating the scoring functions at the test time. Watermarked sequences will be expected to score higher, averaged across the sequence, with non-watermarked sequences scoring near the null expectation of 0.5. In text watermarking, detectability improves with longer sequences or higher entropy in the distributions $p(\cdot|x_{<t})$ ¹³. The latter depends on the task and can be influenced by modulating ProteinMPNN's sampling temperature.

Using binary scoring functions g , where each possible token x_t is assigned a score of 0 or 1, tournament sampling allows candidate tokens to compete against each other. In its simplest, one-layer ($L = 1$) instantiation, we sample two candidate tokens and select the candidate with the higher score (or uniformly sample in the case of a tie). This can be extended to $L > 1$ layers, introducing a stronger watermark, as follows. Initially, for layer $l = 1$, we sample 2^l candidates and let pairs compete against each other using a scoring function g with key k_1 . For layer $l = 2$, using the remaining 2^{l-1} candidates, we repeat this process with key k_2 and so on. After L layers, only one candidate remains. In practice, this can be implemented as a transformation on top of ProteinMPNN's logits, avoiding the need to explicitly sample 2^l candidate tokens. As a result, we obtain a set of L binary scores per token; averaging across layers l and tokens t gives an average g -value used for detection. Using unique keys k_l for each layer is argued to be non-distortionary in the sense that the token and sequence distributions remain unchanged in expectation over the keys k_l .

(refs. 13,23). Repeating keys, in contrast, explicitly introduces distortion and can boost detectability. However, it is important to emphasize that this and similar notions of distortion consider distributions over tokens, not semantics. It is entirely possible for a distortionary watermark to have no effect on protein function; this can only be determined through experimental testing. As ProteinMPNN's temperature is typically low for protein design tasks^{17,18} to reduce entropy during design, we make use of distortionary watermarking for low temperatures.

Integrating into protein design

Typically in protein design, a model generates a design backbone structure, which is then resequenced using ProteinMPNN to produce the final design. These designs are then filtered using thresholds on various structural metrics from a protein folding model (for example, AF3), as illustrated in Fig. 1a. Application of the filters can result in pass rates in single or low double digits depending on the target¹⁷. As filters are applied on top of the designs, watermarking can affect the distribution of metrics and thus reduce pass rates while filters can drop strongly watermarked designs, thereby decreasing watermark detectability. We tackle the former by generating sufficient candidates and the latter by boosting detectability as follows: we increase the sampling temperature from the default of 0.1, which also counters the reduction in diversity from watermarking, or by using distortionary watermarking²⁴. For the former, we consider temperatures 0.3 and 0.5 and for the latter we use the same key k_i across layers (both with $H = 4$ and $L = 25$). For our in vitro validation, we chose both non-distortionary and distortionary watermarking with temperature 0.5 and 0.1, respectively. These settings were chosen as they enabled the evaluation of both high temperatures (increasing entropy) and stronger watermarking (increasing distortion) on binding affinity. We compare against designs without watermarking at temperature 0.1. Furthermore, designs are filtered on the basis of a calibrated g -value threshold for both in silico and in vitro analyses. As detailed in the Methods, these thresholds are calibrated to a target false positive rate (FPR) on a set of held-out, non-watermarked designs across 23 different targets as well as natural protein sequences from PDB and UniRef50 (refs. 25,26). This threshold provides an expected 100% true positive rate (TPR) at the expense of stricter filtering.

Fine-tuning AF3 for structure watermarks

In contrast to protein design, which predicts sequences, folding models such as AF3 (ref. 19) predict structures of continuous 3D coordinates of atoms. These models are commonly made publicly available, meaning that sampling-based or post-hoc watermarking can easily be bypassed. To prevent this, our approach embeds the means of watermarking directly into the model's weights, by fine-tuning the model using an additional watermarking objective. AF3 consists of a trunk that creates feature representations of protein sequences followed by a diffusion module that iteratively generates the folded structures and a confidence head that estimates the prediction accuracy (Fig. 1b).

We focus on AF3's diffusion module which, conditioned on intermediate features, generates 3D coordinates in T iterations of denoising 3D Gaussian noise. The diffusion module is trained with a standard denoising objective, $\mathcal{L}_{\text{diff}}$, which is based on a previous work²⁷. To embed the watermark, we introduce a detector network and fine-tune the diffusion module using a composite loss function. The detector is trained to distinguish between watermarked, non-watermarked and ground-truth structures via a binary cross-entropy loss, $\mathcal{L}_{\text{wm}}$. We combine this with the diffusion loss to form the total training objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + w_{\text{wm}}\mathcal{L}_{\text{wm}}, \quad (1)$$

where w_{wm} is a hyperparameter weighting the contribution of the watermarking objective, which we set to $w_{\text{wm}} = 0.1$. The loss of the confidence head remains unchanged.

The detector's architecture is inspired by PointNet²⁸. It operates by first computing intra-residue geometric features such as atom–atom distances and torsion angles. These features are then processed through several layers of 1D convolutions on a per-residue basis. The resulting feature maps are aggregated by average pooling before being passed to a final linear head for classification.

To enhance watermark robustness during training, we inject noise into the predicted structure's coordinates before they are passed to the detector. This is achieved by adding uniform noise sampled from $U(-s, s)$, with $s \in \{0.1, 0.01, 0.001\}$ Å. We generally found that a minimal perturbation ($s \geq 0.001$) is required to ensure the watermark persists when structures are saved in common file formats such as crystallographic information file (CIF)²⁹, which typically rounds atom coordinates to three decimal points. As the atom–atom distances and torsion angles that the detector uses are invariant to rigid transformations, our watermarks are robust to such transformations by construction.

SynthIDBio results

SynthIDBio-sequence in vitro validation

For protein sequence watermarking to be a feasible intervention, it must not disrupt the intended use of the protein. To verify that watermarking did not affect the function of protein binders, we performed in vitro binding affinity measurements of non-watermarked and watermarked binders against three targets: SC2RBD, VEGF-A and PD-L1, as a representative subset of a past study¹⁷. For each target, we used 15 backbones from existing designs with known binding affinity from historical AlphaProteo campaigns. These backbones were resequenced through further application of ProteinMPNN. We started from backbones of known binders, rather than performing a full de novo design to increase the likelihood of binding success per sequence and thus the amount of useful affinity data gathered. Furthermore, the use of multiple backbones per target enabled evaluation of the impact of sequence and structural diversity on watermarking performance. For each backbone, we tested 18 designs total: one parent sequence (previously validated design corresponding to a selected backbone), five non-watermarked sequences; six from non-distortionary watermarking with temperature 0.5; and six from distortionary watermarking with temperature 0.1. Three sequences per target were randomly replaced with a negative control (one per model setting) consisting of a resequenced binder for an alternative target. In total, without controls, our in vitro evaluation included 222 non-watermarked sequences and 267 watermarked sequences for each watermarking setting. Refer to the Methods for full details on the characterization set-up and experimental protocols.

In Fig. 2 we find minimal impact of either watermarking setting on mean binding affinity as measured by surface plasmon resonance (SPR) spectroscopy and expressed using the dissociation constant K_D (Fig. 2a), or hit rates stratified by binding affinity (Fig. 2b). There were no significant population-level differences in SPR-derived binding affinities between non-watermarked and watermarked binders (with either setting) as measured by a two-tailed Wilcoxon rank-sum test. However, binders targeting SC2RBD that were watermarked at distortionary 0.1 had a significantly lower median K_D than those watermarked at non-distortionary 0.5. With a g -value threshold calibrated for a 0.1% FPR, we automatically obtain a 100% TPR for detecting these designs. There were no significant differences in hit rates for binding affinity thresholds $K_D \leq 10^{-7}$. However, there was a significant difference between non-distortionary 0.5 and non-watermarked hit rates at the $K_D \leq 10^{-6}$ threshold, with non-watermarked binders having a higher hit rate. All negative controls ($n = 9$) had no detectable binding. We also found that resequenced designs often had better binding affinity than the parent design, as shown in the Supplementary Information. This improvement on the parent sequence in affinity is not attributable to the watermarking procedure but is rather a consequence of the iterative, multi-round sampling from ProteinMPNN. We further show

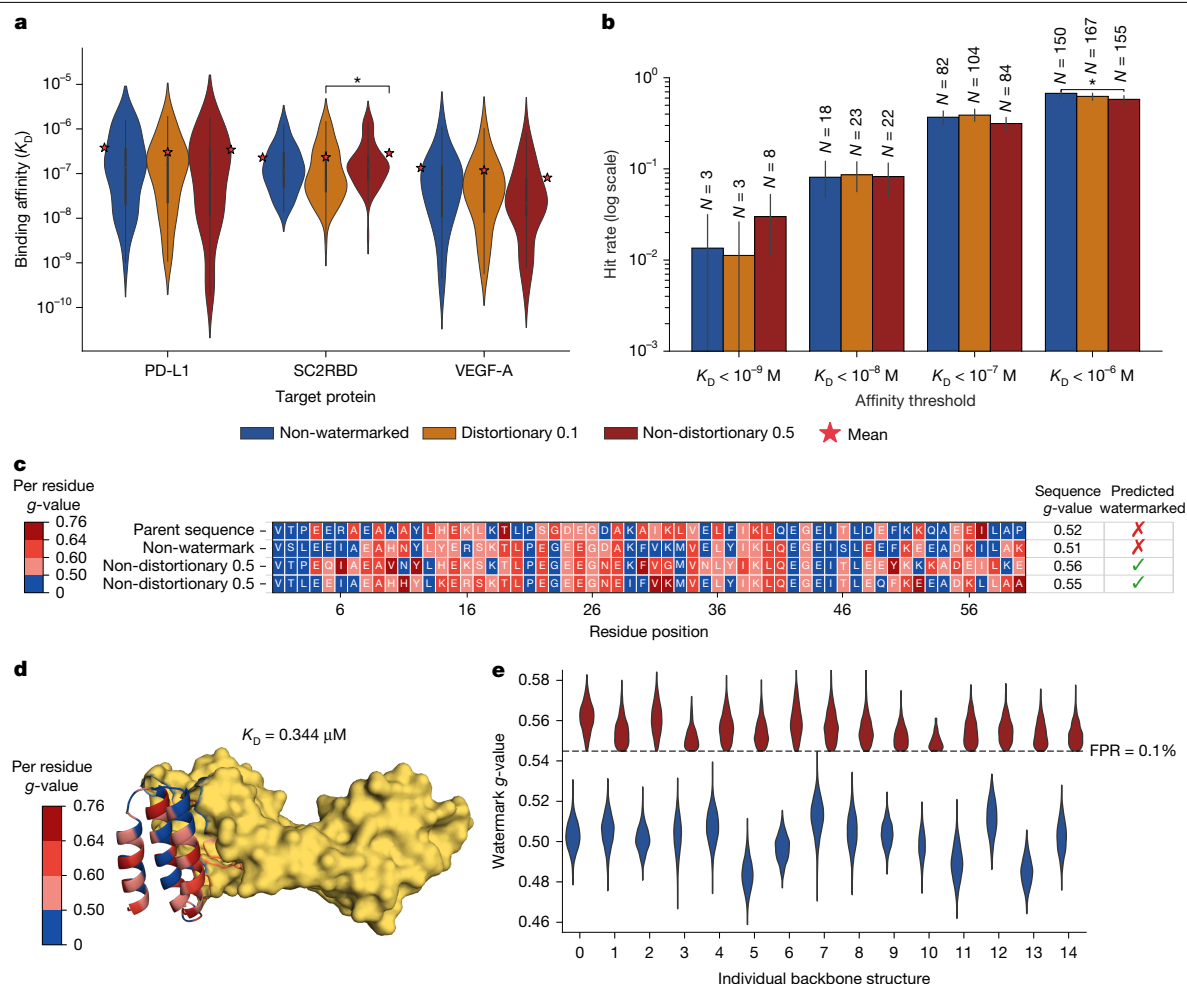


Fig. 2 | In vitro results for SynthIDBio-sequence. **a**, Violin plots of binding affinity calculated as the mean of at least two technical replicates for all three targets. We compare non-watermarked settings with our watermarked settings for all designs that had quantifiable binding. Violins include a box-and-whisker plot depicting the interquartile range. The significance—if found via a two-tailed Wilcoxon rank-sum test—is indicated by * ($P \leq 0.05$). The total sample numbers per bin ranged from $n = 43$ to $n = 69$. **b**, Hit rates stratified by binding affinity thresholds. **c**, Representative multiple sequence alignment of non-watermarked parent sequence (a previously validated design corresponding to a selected backbone), non-watermarked and watermarked binders targeting

VEGF-A. The colours correspond to g -values. Watermarked sequences are detected on the basis of a sequence-level, average g -value threshold of 0.545. **d**, Visualization of a distortionary 0.5 watermarked VEGF-A binder, with per-residue g -values indicated by colours. **e**, Per-backbone g -value distributions for non-watermarked (blue) and non-distortionary 0.5 watermarked binders (red) targeting PD-L1. Non-distortionary 0.5 watermarked binders are limited to those that pass the g -value threshold for an FPR of 0.1%. Total sample numbers per bin ranged from $n = 124$ to $n = 4,037$. g -value distributions can vary widely by backbone; our additional g -value-based filter ensures that synthesized designs are detectable.

exemplar watermarked sequences in a multiple sequence alignment and the predicted structure of a watermarked binder interacting with VEGF-A (Fig. 2c,d). As shown in both panels, the watermarking signal is distributed throughout the protein, including in all types of secondary structure. As we started with fixed backbones, Fig. 2e shows non-distortionary 0.5 and non-watermarked g -value distributions across all backbones for PD-L1. This highlights that the ability to watermark sequences through a resequencer is constrained by the backbone structure we input into ProteinMPNN. For example, we show a significantly different distribution of g -values between non-watermarked backbones 5 and 12 (two-tailed Wilcoxon rank-sum test).

Detectability and compute trade-offs

For in silico analysis, we measure watermark detectability and the impact of g -value filtering on in silico pass rates across 23 different targets with 10,000 samples each. Figure 3a shows TPR without g -value thresholding across non-distortionary watermarking at temperatures 0.1, 0.3 and 0.5, as well as distortionary watermarking for temperature 0.1. These were obtained at the same, calibrated FPR thresholds

used in our in vitro validation. Without a g -value threshold, TPR is not guaranteed to be 100% but depends on how well the designs can be watermarked. Specifically, TPR generally improves with increasing temperature or when using distortionary watermarking. As these TPRs may not be high enough to ensure good detectability, we introduce a further threshold on average g -value. In Fig. 3b, TPR is guaranteed to be 100% and we show the impact of this threshold on in silico pass rates. The reduction is higher for lower temperatures or distortionary watermarking. A reduced pass rate ultimately translates to higher computational cost of the protein design pipeline. Intuitively, a 41.7% decrease in pass rate, as for non-distortionary 0.5 at 0.1% FPR induces a $100\% / (100\% - 41.7\%) = 171.5\%$ increase in candidates needed. However, as structural validation via AF3 is the most significant element in computational cost and we do not need to run AF3 on candidates not passing the g -value threshold, we found that SynthIDBio-sequence adds negligible cost in practice. Refer to the Supplementary Information for a detailed analysis. Figure 3c shows that we typically do not see a significant impact on key metric distributions such as AF3 confidence between watermarked and non-watermarked designs.

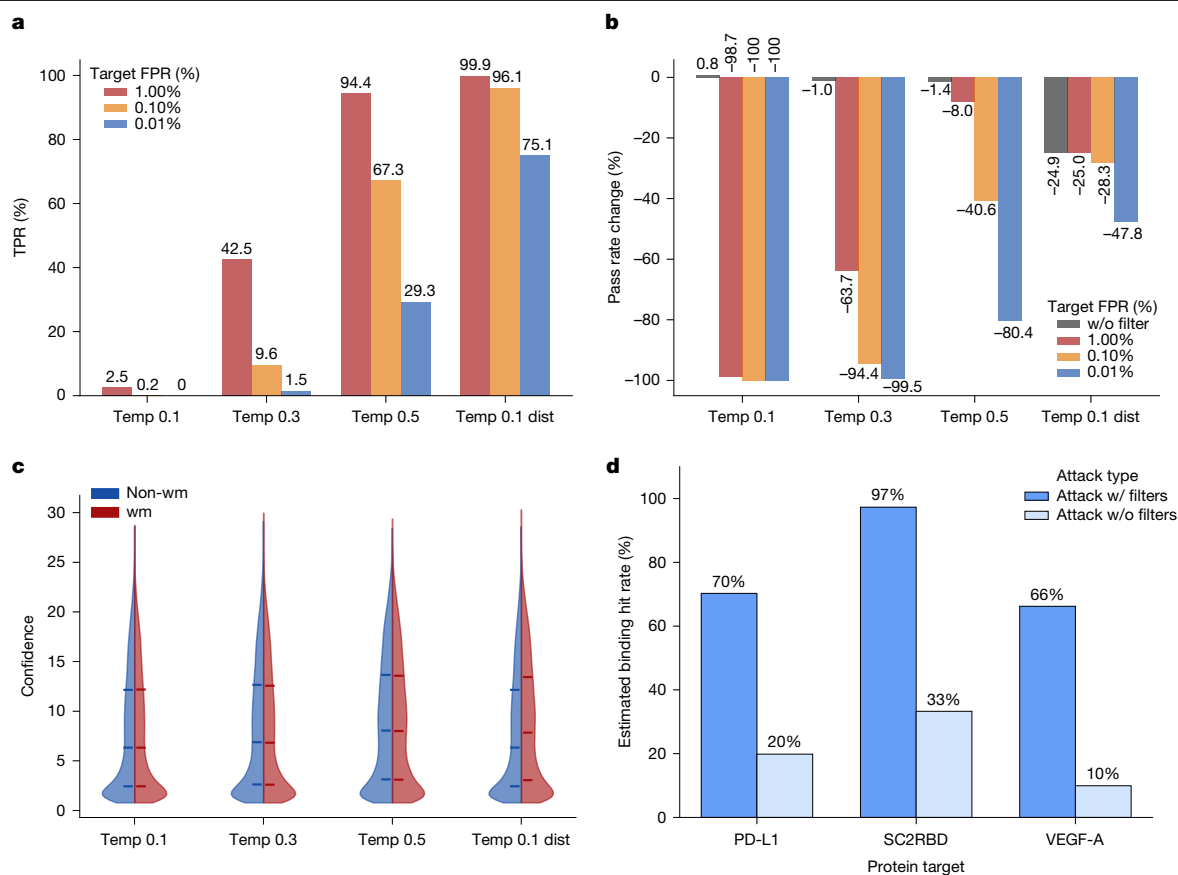


Fig. 3 | In silico results for SynthIDBio-sequence. **a**, The TPR for different watermarking settings across FPRs without g -value filtering. Results were obtained for 23 different binding targets using 10,000 designs each. Either higher sampling temperatures (Temp.) or distortionary watermarking (dist.) is required for a high rate of designs to clear the detectability threshold. **b**, Pass rates for different watermarking settings—with and without g -value based filtering—for different FPRs. Without g -value filtering, SynthIDBio-sequence

has limited effect on pass rates (except for the distortionary setting), but results in variable TPR. With filtering, the TPR is guaranteed to be 100%, but pass rates are reduced. **c**, SynthIDBio-sequence has a negligible impact on AF3 confidence distributions except when using distortionary watermarking (see quantiles). **d**, We show estimated binder hit rates, starting from a known binder for resequencing attacks with or without knowledge of our filters. These attacks largely remove the watermark (Supplementary Information).

SynthIDBio-sequence robustness

We consider several threat models for removing the SynthIDBio-sequence watermark, classifying attackers on the basis of their objective, expertise in biology and AI, and access to computational and wet-lab resources. Here we present results for a resequencing attack via ProteinMPNN. Based on a resequencing attack performed on 38,396 binders, this approach effectively removes the watermark. Assuming the attacker starts with a known, watermarked binder, Fig. 3d shows an estimation of hit rates on the basis of our in vitro experiments for two attacks: resequencing with and without structure-based filters when starting with a known binder (the parent sequence known to bind). Resequencing reduces the estimated hit rate to 97% (SC2RBD), 70% (PD-L1) and 66% (VEGF-A) with filters, but only 33% (SC2RBD), 20% (PD-L1) and 3% (VEGF-A) without filters. Hit rates further reduce when the attacker does not know whether the watermarked sequence binds. Refer to the Supplementary Information for results for this and substitution-based attacks.

Structural accuracy and detectability

Figure 4a (left) shows results for SynthIDBio-structure on the AF3 evaluation set described in supplementary section 6.1 of ref. 19, highlighting high detection accuracy with negligible impact on accuracy. Specifically, we show TPR at 0.1% FPR for models trained with $s = 0.001$, 0.01 and 0.1. The TPR exceeds 99.8% for all models. Moreover, measured in terms of local distance difference test (LDDT) and template

modelling scores, we show that SynthIDBio-structure predicted structures are close to both the ground truth and AF3 predictions. In fact, for $s = 0.001$, SynthIDBio-structure does not lead to a reduction in LDDT or template modelling score, while watermarking with larger s leads to a small decrease compared with the AF3 baseline. Figure 4a (right) demonstrates that watermark detection is effective across various biomolecules (proteins, RNA, DNA). In Fig. 4b we investigated how the watermark works: we found that the watermarked models commonly shift the distributions of C- α atom distances; the watermarking also affects some torsion or bond angles. However, the examples shown in Fig. 4b are meant to illustrate the worst case, meaning we picked features where we qualitatively saw significant changes. We emphasize that the vast majority of atom-atom distances, torsion and bond angles are preserved, especially for the $s = 0.001$ model, which we recommend. Overall, we argue that SynthIDBio-structure is indistinguishable from AF3 as key metrics such as LDDT or template modelling score are unaffected and shifts in feature distributions are within AF3's error margin.

SynthIDBio-structure robustness

Perturbing the generated structures mildly with Gaussian noise of s.d. 0.01 Å (Fig. 4a, left) keeps the watermark mostly intact, with TPR reducing insignificantly. Using a higher s.d. of 0.1 destroys the watermark for smaller s , but our model trained with $s = 0.1$ still achieves a TPR of 88.5% (Supplementary Information). Although we recommend $s = 0.001$, this is a common trade-off between robustness and accuracy in watermarking. Moreover, Fig. 4a (right) shows that the watermark can be detected

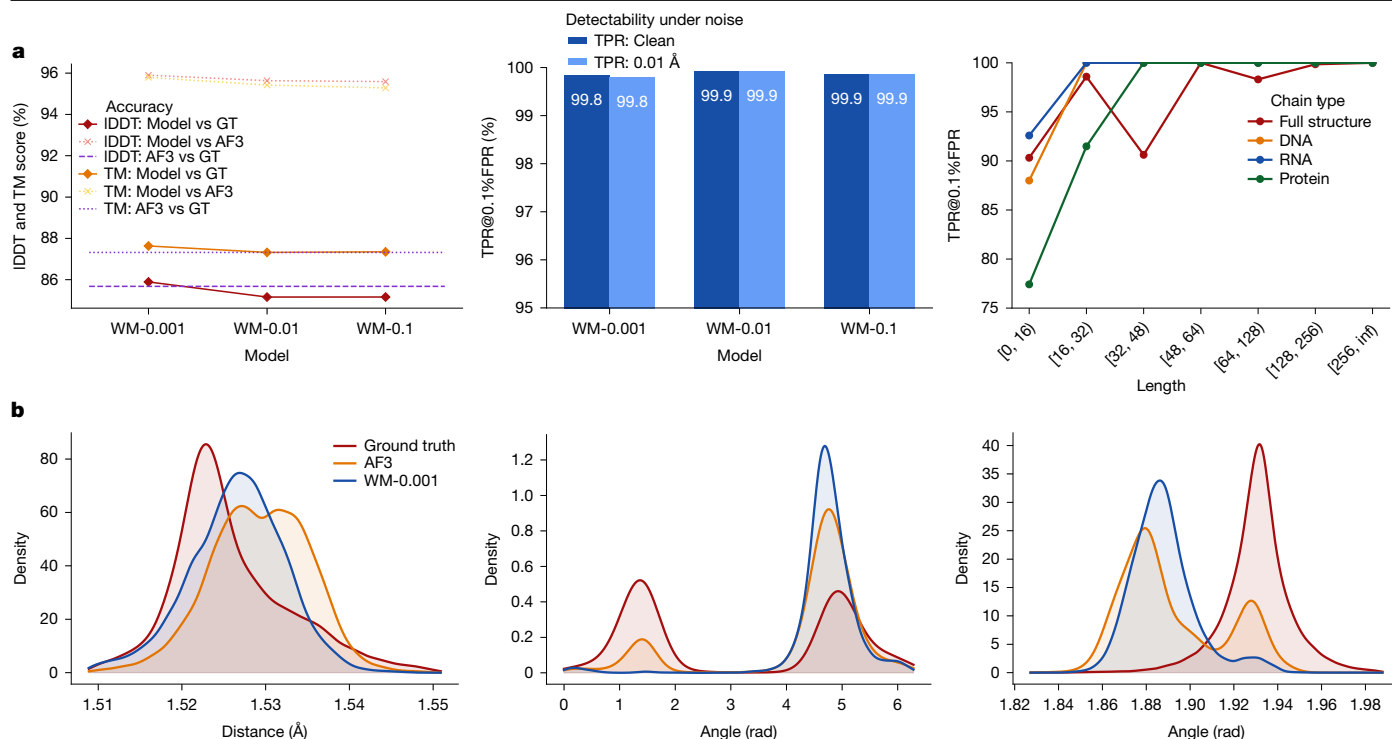


Fig. 4 | SynthIDBio-structure results compared to AF3 baseline. Display includes metrics from three SynthIDBio-structure models fine-tuned with $s \in \{0.001, 0.01, 0.1\}$, called WM-0.001, WM-0.01 and WM-0.1, respectively. **a**, SynthIDBio-structure accuracy (left), detectability (middle) and robustness (right). The LDDT and template modelling scores between the watermarked and ground-truth structures are not lower than the AF3 baseline for $s = 0.001$. The TPR calibrated at 0.1% FPR is very high, exceeding 99.8% for all three models; TPRs remain high for lower FPRs, for example, 98.99% for $s = 0.001$ at 0.01% FPR. The TPR drops insignificantly when applying Gaussian noise of

s.d. 0.01 Å to test the watermark's robustness. We also show the TPR by number of residues and for different biomolecules; detection is generally lower for very short structures with less than 16 residues but quickly reaches over 99% for larger structures. The watermark performs well across biomolecules, including proteins, RNA and DNA. **b**, We show examples of how feature distributions change across 1,000 structures, including the C–C $_{\alpha}$ atom distances in methionine (left), the χ^2 torsion angle of phenylalanine (middle) and the CG1–CB–CG2 angle of arginine (right).

even in small structures starting between 32 and 48 residues. This is also because the detector can handle structures of varying lengths, ignoring the order of residues, by construction. Similarly, our detector is inherently invariant to rigid transformations due to the features used. We also conducted experiments considering constrained relaxation of watermarked structures, using OpenMM with the Amber99sb force-field³⁰, which successfully destroys the watermark.

Discussion

We introduced SynthIDBio-sequence and SynthIDBio-structure, new methods for watermarking AI-generated protein sequences and structures, respectively. SynthIDBio-sequence integrates SynthID-text's tournament sampling and further watermark score-based filtering into a protein design pipeline, whereas SynthIDBio-structure fine-tunes an AF3 compatible model's diffusion module alongside a PointNet-inspired watermark detector for structural watermarking. Both methods achieve high detectability while critically preserving biological utility of the outputs: binding affinity for designed binder sequences and key structural metrics for predicted structures.

Compared with related proposals to establish and track provenance of biological objects^{2,12}, watermarking with SynthIDBio has several advantages worth highlighting. Tool-generated metadata, as discussed in refs. 12,31,32, enable a privacy-preserving and data-rich way to track provenance. However, metadata are trivial to remove and, without cryptographic signatures, vulnerable to forgery. Provenance information tracked in databases² can also be data-rich and cannot be removed. Retrieval in databases is, however, prone to false positives at scale³³. Moreover, databases pose challenges for privacy and

intellectual property protection and, depending on implementation, may require a central, trustworthy authority. Although this may be overcome using cryptographic hashing^{34,35}, watermarking can readily avoid this problem. SynthIDBio only requires the exchange of detection keys with trusted parties. The detectability of watermarks is also known to scale better and watermarks are harder to remove than metadata.

We believe that SynthIDBio is a first step towards reliable provenance for biological objects in the age of AI-enhanced biological research. Although this work has to be understood as a technical proof-of-concept, we also discuss potential use-cases across biosecurity and scientific integrity as well as adoption of watermarking for these purposes, especially in light of powerful open weight models. Although these are hypothetical at the time of writing and will require further research and careful implementation to operationalize, they demonstrate the potential impact SynthIDBio may have if adopted.

Biosecurity perspective

In a suite of layered interventions, SynthIDBio-sequence could support nucleic acid synthesis screening. These technical and policy interventions could help secure the digital-to-physical interface by limiting access to synthetic nucleic acids that could be used for harm³⁶. Screening algorithms compare orders with databases of sequences of known risk. A technical challenge arises when orders have low sequence homology to sequences in reference databases. In these instances, synthesis providers must decide whether to fill orders containing apparently new sequences³¹, driving increased operational costs and longer synthesis timelines. The emergence of advanced AI tools enabling creative and extensive biological design has altered the threat landscape; a recent study demonstrated that resequencing could generate

malicious sequences that bypassed deployed screening algorithms⁴. Although there are ongoing efforts to migrate from homology-based to function-based³⁷ detection algorithms, which should be resilient to resequencing attacks, these algorithms remain in development. In the interim, establishing AI-generated provenance may support these assessments.

A punitive implementation of SynthIDBio-sequence would involve further work by the user if the watermark is detected. For example, this could involve requesting further design details. However, the widespread availability of open-source tools severely limits the feasibility of this approach. First, there is little incentive for good actors to use SynthIDBio-sequence. Second, and more importantly, a malicious actor would simply use a non-watermarking tool in the first place or conduct the resequencing attack when using a watermarked design service.

Similar to proposals for a metadata exchange framework^{12,32}, watermarking could establish which tools were used during design. As design capabilities continue to expand, there are growing calls for model developers to incorporate guardrails into their tools to reduce the risk of misuse³¹. Here, watermarking could serve as a verification signal that a ‘trusted’ tool, which a third-party verifies contains sufficient mitigations and guardrails to prevent nefarious use, was used to generate an order (for example, a resequencer deployed via an application programming interface that blocks input structure parent sequences aligning with a database of harmful proteins). For screening providers, an intact watermark provides low-overhead, positive evidence of compliance; it simultaneously signals that a sequence region is engineered and that the submitter has not actively attempted to obscure its provenance. Establishing processes for this verification is outside the scope of this paper. Were evasion techniques to grow in complexity with the computational cost of screening rising, watermark detection could enable bypassing more computationally expensive portions of future analysis pipelines (for example, structure prediction). Synthesis companies could then incentivize the use of these trusted tools by passing on these cost savings to consumers. However, careful tuning of FPR thresholds would be required if watermark detection were to be used to bypass screening algorithms. From a tool developer perspective, this could increase scientific uptake and thus warrant the engineering effort required to embed SynthID-text within their tool. It is important to note that this practice could also be enabled by tool-specific metadata paired with cryptographic signatures.

Implementing the above framework, however, requires distributed detection via the secure sharing of detection algorithms and keys between tool developers and synthesis companies. To prevent spoofing, where a malicious user fraudulently claims the use of a verified tool, these keys must be hidden from the user. Consequently, the trusted tool framework is most immediately applicable to models deployed via hosted interfaces (for example, programming interfaces, web portals) where the watermarking process is obscured from the user. Considering open-weight models, further watermark innovations would be required, as discussed in the ‘Enabling and incentivizing adoption’ section. Furthermore, the proposed approach is probably to be most applicable to de novo biological design, wherein the watermarking signal can be embedded across the entire modality. Instances where only portions of the order are watermarked have an increased risk of false negatives. Relatedly, manual modification after a sequence is designed, such as addition of a C-terminal expression tag, will reduce the watermarking signal proportionally on the basis of the modification’s relative size. To reduce this risk, submissions to synthesis providers could include metadata identifying the coordinates for designed fragments. Finally, the robustness of SynthIDBio-sequence itself could be a source of risk. Sophisticated actors could append short malicious sequences to watermarked designs, betting that the proportional dilution of watermarking signal would still allow the order to be fast-tracked. This vulnerability can be mitigated through a layered detection approach, wherein orders are still screened using standard

similarity-based approaches regardless of the watermark. Detection of low-homology segments outside of the design coordinates can motivate additional scrutiny. Further discussion on this topic is available in the Supplementary Information.

The objective of SynthIDBio-sequence aligns with the historical precedent of the gene synthesis industry: to raise the barrier to entry rather than guarantee absolute security. Defeating a robust watermark introduces an asymmetric operational cost on malicious actors, requiring further technical sophistication, increasing computational overhead and introducing the risk of functionally degrading the protein. Furthermore, widespread adoption of watermarking among model developers may systematically restrict malicious users to less performant, non-watermarked legacy models.

Scientific integrity perspective

SynthIDBio and related technologies may also be useful for ensuring information veracity in public scientific databases. Practically all major AI-enabled tools for biological research—including AF3, RFdiffusion and AlphaProteo—depend on widely accessible and well-maintained databases of protein structures and sequences for model training, such as the PDB⁹ and UniProt¹⁰. Many of these same tools, as well as common workhorses for bioinformatics (for example, Kraken2; ref. 38), further depend on databases such as GenBank¹¹ or MGnify³⁹ for model inference or tool application. The submission of entries to these databases is open to members of the public and in some instances is closely tied to professional scientific advancement, including academic publications. This reality presents a hypothetical opportunity for misuse of generative AI. It is conceivable that malicious actors could submit machine-generated sequences with falsified metadata to intentionally degrade the performance of tools reliant on these databases—a potential risk posited elsewhere⁴⁰. The submission of improperly labelled synthetic DNA to GenBank has already been shown to degrade the utility of the Basic Local Alignment Search Tool (BLAST) for identifying pathogens⁴¹. Automated detection of AI-generated sequences through identification of a watermark could enable additional scrutiny before submitted information is exposed to the public.

Enabling and incentivizing adoption

Adoption of watermarking for these applications is driven by both incentives (as alluded to in the ‘Biosecurity perspective’ section) and technical feasibility.

Incentives go beyond the implementation model, that is, whether watermarking results in negative (that is, punitive) or positive action. We expect that incentives will be very sensitive to concrete operating points chosen by synthesis providers or database maintainers, which are incentivized to optimize for very low FPR and/or very high TPR. In the scientific integrity use case, false negatives would lead to misinformation, whereas false positives would primarily lead to extra work performing verification and checks and a risk of spurious accusations of AI assistance. For biosecurity, false positives would avoid more intensive screening (following our proposal of the ‘Biosecurity perspective’ section), which itself poses biosecurity risks. On the other hand, they also have an interest in encouraging adoption by users. For these users, guaranteeing low FPR and high TPR usually comes at a cost. Concretely, for SynthIDBio-sequence, guaranteeing 100% TPR via *g*-value thresholding for extremely low FPR will result in significant reductions of pass rates, meaning higher compute cost. If this cost exceeds the perceived value of faster or cheaper processing by the synthesis provider, they will not choose to use a watermarking tool. We hope that this tension will result in selection of thresholds that benefit both parties.

Technical practicability includes protocols for sharing detection keys and models, which are considered secret and thus limited to trusted parties. Further innovations for public-key watermarking would be required^{42,43} to share detectors or keys publicly, and to ensure differentiability of watermarks from multiple providers is prioritized.

Furthermore, SynthIDBio-sequence and SynthIDBio-structure have different trade-offs that are worth emphasizing. SynthIDBio-sequence integrates directly into ProteinMPNN's sampling algorithm, meaning no changes to the underlying model are needed and adoption is technically easy. Moreover, we expect SynthIDBio-sequence to generalize to other models that sample protein sequences auto-regressively, irrespective of model training or architecture. However, it also implies that the watermarking mechanism can easily be removed or turned off by users. This is in contrast to SynthIDBio-structure, which embeds the information of how to watermark in the model weights via fine-tuning. Although this requires fine-tuning, which presents a high upfront cost, it means users cannot remove the watermarking mechanism, even when the model weights are shared openly. In the case of SynthIDBio-structure, our approach is largely adapted to AF3's diffusion module. As diffusion models are common for these types of prediction tasks, we expect this approach to generalize well. Ultimately, however, we consider this a proof-of-concept that function-preserving watermarking of protein sequences and structures is possible. We believe that other choices of these trade-offs are possible. For example, approaches based on fine-tuning ProteinMPNN to watermark sequences in the spirit of ref. 44 or a sampling-based approach for AF3 similar to ref. 45 are both realistic.

Related work

Concurrent work also considers watermarking in the context of protein sequences³ or structures¹⁵. However, a past work³ does not go beyond an *in silico* study, showing that resequencing wild-type monomer structures from PDB does not affect ESMfold-predicted⁴⁶ LDDT distributions. Moreover, the Gumbel sampling used for watermarking²² has been shown to be inferior to SynthID-text's tournament sampling in terms of detectability and diversity¹³. By contrast, SynthIDBio-sequence integrates watermarking into a protein design pipeline, which also involves resequencing AI-generated, synthetic structures. Crucially, through *in vitro* validation, we show that this preserves the function and binding strength of designed binders. Another past work¹⁵ relies on training a post-hoc watermarking model and detector first, before distilling it into ESMfold using low-rank adaptation fine-tuning⁴⁷ resulting in a noticeable increase in the root mean square deviation (r.m.s.d.). SynthIDBio-structure is considerably simpler in that we directly fine-tune the diffusion module of AF3 jointly on structure and watermark losses. We believe this is key in enabling watermarking at state-of-the-art structural accuracy.

Limitations and future work

SynthIDBio still has several limitations that need to be addressed to operationalize any of the above use-cases. SynthIDBio-sequence does not support encoding more information than the presence of the watermark (that is, a zero-bit watermarking scheme). Moreover, it can incur computational overhead in protein design and is susceptible to further resequencing through ProteinMPNN. Although we show that such removal attempts also affect the estimated binder hit rate, future work is needed to investigate more robust watermarking schemes for protein design and to evaluate the impact on detectability and pass rates of alternative attacks such as partial resequencing. In particular, the availability of browser-based instances of ProteinMPNN lowers the technical threshold for conducting the resequencing attack. It is worth highlighting that similar attacks on text or image watermarking—known as regeneration attacks—remain effective despite significant recent advancements. Refining the quantitative impact of resequencing on binder hit rates would also benefit from further *in vitro* studies. Moreover, we do not address a random or right-to-left decoding order that is supported by ProteinMPNN. Similarly, for SynthIDBio-structure, we show significant robustness to noise, rigid transformations and cropping. However, robustness to relaxation is lacking, but we expect that this could be addressed by explicitly considering relaxation while training SynthIDBio-structure. Moreover,

SynthIDBio-structure currently operates as a zero-bit watermarking scheme, which does not allow differentiation of multiple users. Finally, differentiability from other structure-based watermarks such as FoldMark¹⁵ has not been studied.

Conclusion

The ability to reliably watermark AI-designed protein sequences and structures without affecting their intended use, as demonstrated with SynthIDBio, represents a notable step towards establishing and tracking provenance of biological objects. These tools offer a practical, technically sound, proof-of-concept method to enhance traceability and accountability in the rapidly advancing field of AI-driven protein design, with potential implications for biosecurity and scientific integrity.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-026-10965-y>.

- Winnifrid, A., Outeiral, C. & Hie, B. L. Generative artificial intelligence for de novo protein design. *Curr. Opin. Struct. Biol.* **86**, 102794 (2024).
- Baker, D. & Church, G. Protein design meets biosecurity. *Science* **383**, 349–349 (2024).
- Chen, Y. et al. Enhancing privacy in biosecurity with watermarked protein design. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btaf141> (2025).
- Wittmann, B. J. et al. Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science* **390**, 82–87 (2025).
- Marchand, A., Van Hall-Beauvais, A. K. & Correia, B. E. Computational design of novel protein–protein interactions—an overview on methodological approaches and applications. *Curr. Opin. Struct. Biol.* **74**, 102370 (2022).
- National Academies of Sciences, Engineering, and Medicine. *The Age of AI in the Life Sciences: Benefits and Biosecurity Considerations* (National Academies Press, 2025).
- Bloomfield, D. et al. AI and biosecurity: the need for governance. *Science* **385**, 831–833 (2024).
- Hunter, P. Security challenges by AI-assisted protein design. *EMBO Rep.* **25**, 2168–2171 (2024).
- Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide protein data bank. *Nat. Struct. Mol. Biol.* **10**, 980–980 (2003).
- The UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2024).
- Benson, D. A. et al. Genbank. *Nucleic Acids Res.* **41**, D36–D42 (2012).
- Wheeler, N. E. Responsible AI in biotechnology: balancing discovery, innovation and biosecurity risks. *Front. Bioeng. Biotechnol.* <https://doi.org/10.3389/fbioe.2025.1537471> (2025).
- Dathathri, S. et al. Scalable watermarking for identifying large language model outputs. *Nature* **634**, 818–823 (2024).
- SynthID: a tool to watermark and identify content generated through AI <https://deepmind.google/science/synthid/> (Google, accessed 26 August 2025).
- Zhang, Z. et al. FoldMark: Protecting protein generative models with watermarking. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.10.23.619960> (2025).
- Dauparas, J. et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
- Zambaldi, V. et al. De novo design of high-affinity protein binders with AlphaProteo. Preprint at <https://arxiv.org/abs/2409.08022> (2024).
- Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
- Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
- Pacesa, M. et al. One-shot design of functional protein binders with BindCraft. *Nature* <https://doi.org/10.1038/s41586-025-09429-6> (2025).
- Kirchenbauer, J. et al. A watermark for large language models. Preprint at <https://arxiv.org/abs/2301.10226> (2024).
- Aaronson, S. & Kirchner, H. Watermarking of large language models (2022) <https://www.scottaaronson.com/talks/watermark.ppt>. Accessed 9 May 2025.
- Hu, Z. et al. Unbiased watermark for large language models. Preprint at <https://arxiv.org/abs/2310.10669> (2023).
- Venugopal, A., Uszkoreit, J., Talbot, D., Och, F. & Ganitkevitch, J. Barzilay, R. & Johnson, M. Watermarking the outputs of structured prediction with an application in statistical machine translation. In *Proc. 2011 Conference on Empirical Methods in Natural Language Processing* (eds Barzilay, R. & Johnson, M.) 1363–1372 (Association for Computational Linguistics, 2011).
- Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2014).
- Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).

27. Karras, T., Aittala, M., Aila, T. & Laine, S. Elucidating the design space of diffusion-based generative models. Preprint at <https://arxiv.org/abs/2206.00364> (2022).
28. Qi, C., Su, H., Mo, K. & Guibas, L. J. PointNet: deep learning on point sets for 3D classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* 77–85 (2016).
29. R. K. Green. Beginner's Guide to PDBx/mmCIF. *PDB-101, RCSB Protein Data Bank* <https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/beginner%E2%80%99s-guide-to-pdbx-mmCIF> (2023).
30. Hornak, V. et al. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure* <https://doi.org/10.1002/prot.21123> (2006).
31. Carter, S., Wheeler, N., Isaac, C. & Yassif, J. Developing guardrails for AI biodesign tools. *Nuclear Threat Initiative* <https://www.nti.org/analysis/articles/developing-guardrails-for-ai-biodesign-tools/> (2024).
32. Densmore, D., Isaac, C., Wheeler, N. & Yassif, J. A proposal for biodesign metadata exchange for use in biosecurity. *Nuclear Threat Initiative* <https://www.nti.org/analysis/articles/white-paper-a-proposal-for-biodesign-metadata-exchange-for-use-in-biosecurity/> (2025).
33. Saberi, M., Sadasivan, V. S., Zarei, A., Mahdavi, H. & Feizi, S. DREW: towards robust data provenance by leveraging error-controlled watermarking. Preprint at <https://arxiv.org/abs/2406.02836> (2024).
34. Baum, C. et al. A system capable of verifiably and privately screening global DNA synthesis. Preprint at <https://arxiv.org/abs/2403.14023> (2024).
35. Gretton, D. W. et al. Random adversarial threshold search enables automated DNA screening. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.03.20.585782> (2024).
36. Biosecurity in the age of AI. *Helena* <https://helena.org/projects/helena-biosecurity> (accessed 5 September 2025).
37. Godbold, G. D., Proescher, J. & Gaudet, P. New and revised gene ontology biological process terms describe multiorganism interactions critical for understanding microbial pathogenesis and sequences of concern. *J. Biomed. Sem.* <https://doi.org/10.1186/s13326-025-00323-8> (2025).
38. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* <https://doi.org/10.1186/s13059-019-1891-0> (2019).
39. Richardson, L. J. et al. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2022).
40. Ratcliff, J. Transformer model generated bacteriophage genomes are compositionally distinct from natural sequences. *NAR Genomics and Bioinformatics* <https://doi.org/10.1093/nargab/lqae129> (2024).
41. Beal, J., Clore, A. J. & Manthey, J. A. Studying pathogens degrades BLAST-based pathogen identification. *Sci. Rep.* <https://doi.org/10.1038/s41598-023-32481-z> (2022).
42. Fairuze, J. et al. Publicly detectable watermarking for language models. *IACR Commun. Cryptol.* **1**, 31 (2023).
43. Fairuze, J., Ortiz-Jiménez, G., Vecerik, M., Jha, S. & Goyal, S. On the difficulty of constructing a robust and publicly-detectable watermark. Preprint at <https://arxiv.org/abs/2502.04901> (2025).
44. Gu, C., Li, X. L., Liang, P. & Hashimoto, T. On the learnability of watermarks for language models. Preprint at <https://arxiv.org/abs/2312.04469> (2023).
45. Yang, Z. et al. Gaussian shading: provable performance-lossless image watermarking for diffusion models. Preprint at <https://arxiv.org/abs/2404.04956> (2024).
46. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
47. Hu, J. E. et al. LoRA: low-rank adaptation of large language models. Preprint at <https://arxiv.org/abs/2106.09685> (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026

Methods

SynthIDBio-sequence

Following recent de novo protein design approaches^{16,18,48,49} that combine conditional structure generation, sequence generation and filtering, SynthIDBio-sequence integrates SynthID-text¹³ into a commonly used sequence design model, ProteinMPNN¹⁶. The ability to detect the watermark within the generated designs relies on the underlying entropy during sampling from ProteinMPNN. If there is sufficient entropy, SynthID-text can reach good detectability while being non-distortionary^{13,23}, meaning that watermarking does not change the distribution over sequences in expectation over a secret watermarking key. As protein design typically uses low temperature, limiting entropy, we also consider SynthID-text's distortionary variant and introduce another filter on the basis of the watermark signal.

Non-distortionary SynthID-text. Like large language models, ProteinMPNN is an autoregressive model, providing a probability distribution $p(x_t|x_{<t})$ over the next token (amino acid, in this case) to sample. We use $p(x_t|x_{<t})$ to refer to the distribution after any modifications through top- p sampling^{50,51}. SynthID-text then further modifies this distribution to add a detectable signature through a sampling procedure called tournament sampling. This strategy consists of three components: a random seed generator, a scoring function and a sampling algorithm. At each step t of sampling, the random seed generator provides a random seed r_t that is conditioned on a secret key k and possibly the context, that is, the history of $x_{<t}$. The scoring function assigns a so-called g -value to each possible token based on the random seed and the sampling algorithm then tries to bias sampling towards high g -values.

The random seed generator is a deterministic function $f_r(x_{<t}, k)$ applied to the random secret key k . A common choice is using a hash function $f_r = h(x_{t-H:t-1}, k)$ that hashes the secret key together with the H -gram $x_{t-H:t-1} = (x_{t-H}, \dots, x_{t-1})$ observed before the token x_t to be sampled^{21,22}. The scoring functions are set as binary functions $g(x, r) \in \{0, 1\}$ that assign either 0 or 1 to every possible token x on the basis of the current seed r . Although there are various ways to define these g -values, we follow¹³ and use a binary function. Again, we make use of a hash function h that takes as input the token to be generated, x_t , the seed r_t and the H -gram context:

$$g(x_t, r_t) = \begin{cases} 0 & \text{if } \frac{h(x_t, x_{t-H:t-1}, r_t)}{2^n} \leq 0.5 \\ 1 & \text{if } \frac{h(x_t, x_{t-H:t-1}, r_t)}{2^n} > 0.5 \end{cases} \quad (2)$$

where n is the number of bits of the hashing function h used.

Tournament sampling then considers multiple scoring functions g_l indexed by the layer $l \in [L]$ (generated by using a hash function h that also depends on l). We use different keys (k_l) per layer. The sampling algorithm takes as input a probability distribution over tokens $p(x_t|x_{<t})$ and the seed r_t , and outputs the next token, that is, the next amino acid for the designed sequence. For $l=1$, that is, a single tournament layer, sampling works as follows: given the distribution $p(x_t|x_{<t})$, we sample two candidate tokens x_{t_1} and x_{t_2} and let them compete in terms of their scores $g(x_{t_1}, r_t)$, $g(x_{t_2}, r_t)$ and pick the winner, or a random token in case of a tie. This is generalized to multiple layers by starting with $l=1$, sampling 2^l candidates and 2^{l-1} pairs and letting them compete this way using g_l . For $l=2$, $2^{l-1} = \frac{2^l}{2}$ candidates are left over; pairs now compete using g_2 . This process concludes at $l=L$ where only one candidate remains. To ensure non-distortionary sampling (see algorithm 3 in ref. 13), we refrain from biasing sampling for repeating contexts, that is, repeating seeds r_t that typically stem from seeing the same H -gram multiple times throughout sampling²³. Biasing is also not applied to the first $H-1$ tokens. In the protein design context this is beneficial to ensure higher diversity. Note that this approach generalizes to any

residue that may be fixed beforehand as part of the protein design pipeline.

Watermark detection is based on the fact that tournament sampling maximizes the scoring functions g_l across all layers $l \in [L]$ throughout the sequence. As we can evaluate these scoring functions given the secret key k and the full sequence, we can compute an overall average g -value:

$$\frac{1}{LT} \sum_{t=1}^T \sum_{l=1}^L m_l g_l(x_t, r_t) \quad (3)$$

where m is a mask indicating which tokens to evaluate ($m_t = 1$) or skip ($m_t = 0$) due to repeating contexts.

Distortionary watermarking. The detectability of SynthID-text, especially with low entropy and shorter sequences, can be improved through distortionary watermarking. In tournament sampling, this can be accomplished using two equivalent strategies: we can sample more than two candidates per layer, resulting in s^L candidates when using s leaves at each level, or using the same keys $k_l = k_{l'}$ across multiple layers, $l \neq l'$. The latter is equivalent to the former in that using the same key twice in a row will effectively use $s = 4$ leaves. However, by using the former strategy, we can also use an s that is not a power of two. We follow the latter strategy for simplicity, taking an extreme approach of using $L = 25$ layers, all using the same key. This typically results in significant biasing of the generated sequences (conditioned on the H -gram context).

Integration into protein design pipeline. We make two key adjustments to the common protein design and filtering pipeline. First, we use distortionary or non-distortionary tournament sampling on top of ProteinMPNN, which may change the distributions of key metrics relevant for protein design. Second, to guarantee sufficient detectability, we introduce another g -value filter to the filtering stage. This filter is meant to ensure a specific target FPR and essentially trades computational resources for better detectability. We collected a representative population of non-watermarked protein binder and natural protein sequences to determine our FPR thresholds. As average g -values are more variable for shorter sequences, we enriched the natural sequence dataset for shorter sequences by clipping select UniRef50 sequences up to the first m residues. In total, this dataset consisted of $n = 700,000$ non-watermarked binder sequences (of length 40 to 140 residues due to length parameters); $n = 661,581$ full-length sequences that are over 40 residues in length from the full PDB dataset downloaded on 18 April 2025 (ref. 9); $n = 700,000$ full-length sequences from the March 2022 build of UniRef50 (refs. 25,26); $n = 700,000$ UniRef50 sequences clipped to $m = 40$; and $n = 700,000$ UniRef50 sequences clipped to $m = 140$ (98.5% of sequences were precisely of length 140) for a total of nearly 3.5 million sequences. We consider FPR values of 0.01%, 0.1% and 1% for in silico experiments, whereas we use 0.1% for in vitro validation. This essentially trades computational resources for better detectability by having to generate significantly more design candidates to ensure a similar number of passed designs after filtering. Exact thresholds used can be found in the Supplementary Information.

Design selection and filtering. We chose the three targets that had the largest number of experimentally validated binders in past AlphaProteo campaigns for our in vitro study. These targets and their cropped sequences were VEGF-A (UniProt no. P15692-4 (V40-K133)), PD-L1 (UniProt no. Q9NZQ7-1 (N17-A132)) and SC2RBD (UniProt no. A0A8B6RKS7 (T221-G414)). For each target, we first filter all designs to those that pass our strictest quality checks and then selected the top 15 parent sequences based on binding affinity as previously measured using bio-layer interferometry¹⁷. These parent sequences were used as backbone structures for the resequencer experiments.

From each backbone, we then produced 6,000 designs through resequencing per setting: non-watermarked sequences; distortionary watermarked sequences with temperature 0.1; and non-distortionary watermarked sequences with temperature 0.5. These sequences were filtered on the basis of AF3 metrics (Supplementary Information) and a FPR-based g -value threshold of 0.1%. Finally, we computed confidence value and agreement value per design. The confidence value was defined as the average of the following metrics: complex interface predicted template modelling, binder's predicted template modelling in complex with the target, and the predicted template modelling of the binder in isolation. The agreement value was defined as the average of the following metrics: complex LDDT, complex interface LDDT, and binder's LDDT with the generated backbone. Designs were sorted by the product of the agreement and confidence values and the top six were selected for each setting.

In vitro validation set-up. Synthetic gene fragments encoding the designed protein binders were codon-optimized and purchased from Twist Bioscience. For binders targeting SC2RBD and VEGF-A, gene fragments were ordered as complete constructs containing the binder sequence and C-terminal GFP11 (RDHMLHEYNAAAGIT) and Twin-Strep (WSHPQFEKGGGSGGGSGGSAWSHPQFEK) tags. These constructs were used directly after concentration normalization, and sequence validation consisted of batch Sanger sequencing (Microsynth) of 10–15% of designs selected randomly. For binders targeting PD-L1, the C-terminal tags were assembled into constructs using Gibson-type assembly (NEBuilder HiFi DNA Assembly Kit, New England Biolabs) with 24-bp overlaps. Assembly products were amplified by polymerase chain reaction (Q5 High-Fidelity DNA Polymerase, NEB), quantified using a dye-based DNA quantification assay (Qubit DNA Quantification Kit, Invitrogen), and analysed by capillary electrophoresis on a ZAG DNA Analyzer (Agilent) to confirm fragment size and integrity. For sequence validation, 10–15% of constructs were randomly selected for batch Sanger sequencing (Microsynth). Furthermore, Illumina short-read amplicon sequencing (Amplicon-EZ, Genewiz) was performed on a subset of the library to confirm construct identity before expression.

Protein binders were expressed using an optimized prokaryotic in vitro transcription–translation system (8 μ l reactions, 1 nM DNA, 37 °C, 8 h). Lysates were centrifuged at 10,000 r.p.m. for 2 min to remove aggregates before downstream assays. Soluble expression levels were quantified directly in clarified lysates using the incorporated splitGFP reporter assay⁵². Briefly, GFP1-10 was added to a final concentration of 0.13 mg ml⁻¹ and mixed with clarified lysate diluted in PBS pH 7.4 eightfold, in a total reaction volume of 6 μ l in a black 384-well plate. A protein standard of known concentration—measured across eight concentration points—was used to generate a standard curve. The plate was incubated for 90 min at 30 °C to allow GFP11–GFP1-10 complementation. Fluorescence was recorded at 485 nm excitation and 520 nm emission in a plate reader (Pherastar FS, BMG Labtech). Plates were specific to single targets and included one positive control, which was the parent sequence from the original AlphaProteo campaign, and one negative control, which was a positive control for a different target. These controls randomly replaced generated binders.

Binding kinetics of the recombinantly expressed designed protein binders to their cognate protein targets were characterized using SPR spectroscopy on a Carterra LSA XT instrument, using a carboxymethylated sensor chip functionalized with Strep-Tactin XT (IBA Lifesciences). Proteins were captured directly onto biosensor surfaces via the C-terminal Twin-Strep affinity tag and wash steps removed non-specific background material, providing effective in situ purification before affinity characterization. The analyte, containing a dilution series of protein target (either PD-L1, VEGF-A or SC2RBD), was prepared in running buffer (10 mM HEPES, 150 mM NaCl, 3 mM EDTA, 0.05% Tween-20, pH 7.4) at seven concentrations in a half-log dilution series between 1,000 nM and 1 nM. Single cycle kinetic assays were

performed (with a flow rate of 50 μ l per minute, increasing analyte concentrations, no intermediate regeneration). The chip surface was regenerated (10 mM glycine-HCl, pH 1.5, 5 min) and washed (running buffer, 20 min) following each full single kinetic cycle. The data were processed and fitted globally to a 1:1 Langmuir binding model using custom fitting software. Final kinetic parameters (k_{on} , k_{off} , K_D) were selected on the basis of fit quality. Designs were classified as binders if their association signal was significantly above the negative control across all replicates. Although all such designs are labelled as binders (binding = TRUE), we only report K_D values for samples that produced quantifiable binding curves and stable kinetic fits. Samples labelled as binders with a null K_D typically represent weak interactors (often ≥ 1 μ M) or cases where low expression or rapid kinetics precluded accurate affinity estimation. These should be interpreted as binders with insufficient signal for quantification rather than designs lacking affinity.

Resequencer attack robustness. We estimated the design hit rate after using resequencing to remove the watermark based on a combination of data generated as part of this study and historical AlphaProteo data. The resequencer attack was performed by running non-watermarked ProteinMPNN (with default 0.1 temperature) on a watermarked binder in complex with its target. In favour of the attacker, we assume a threat model where they obtain a watermarked design that is known to bind. We use the product of the in silico pass rate and the resequencer hit rate to estimate their hit rate. In the more realistic scenario of the design having unknown binding behaviour, percentages shown in Fig. 3d will be multiplied by AlphaProteo's historical hit rate on PD-L1, VEGF-A and SC2RBD, which are 15%, 33% and 12%, respectively. Further details are available in the Supplementary Information.

Statistical testing. We use the two-tailed Wilcoxon rank-sum test for significance tests evaluating the difference in distributions (affinity or g -value). We use Fisher's exact test for significance tests evaluating the differences in discrete data, such as the differences in hit rates.

SynthIDBio-structure

For watermarking predictions from AF3, SynthIDBio-structure fine-tunes AF3's diffusion module to predict watermarked 3D atom coordinates jointly with a new watermark detector. As the watermark is not added post-hoc or during sampling, this has the advantage that the watermarking procedure cannot be removed once the model weights are publicly released.

Protein structure watermark detectors. Our watermark detector is inspired by early PointNet architectures²⁸ that operate on individual 3D coordinates, or groups thereof, using a subnetwork, before aggregating these features in a way that is agnostic to the number of 3D coordinates of the input. For protein structures, this means we apply a per-residue network across the whole protein and then average pool the resulting features. Intra-residue distances, that is, pairwise distances between all atoms, are the inputs into the per-residue network, together with torsion angles computed in a sliding window fashion across all quadruples of consecutive atoms. This choice of features makes the whole detector invariant to rigid transformations such as translations, rotations and reflections (or combinations thereof) because features only rely on intra-residue characteristics. Features are stacked into an array and we apply five layers of one-dimensional convolutional layers with 288 channels, each followed by rectified linear unit activations. We appropriately pad atoms and residues throughout the process of feature computation and 1D convolutions. The convolutions compute a fixed size feature vector per residue. We aggregate these features across residues using average pooling, followed by a dense layer predicting a single binary output. That is, we train a zero-bit watermark. Although we expect that the same approach will generalize to multi-bit settings, a multi-bit watermark is not intended in our open-model context.

Fine-tuning AF3 for watermarking. In an abstracted view, AF3 consists of three key modules: the backbone (input embedder, template module, MSA module and pairformer; see fig. 1 in ref. 19), the diffusion module and the confidence module. Our approach embeds an imperceptible watermark into the protein structures generated by AF3 by fine-tuning its diffusion module. This is achieved by introducing a watermarking objective into the model's training process, which modifies the network weights to directly incorporate the watermarking signal. Moreover, as the confidence module is based on the output of the predicted structure, it also needs to be fine-tuned alongside the diffusion module.

The diffusion module generates the 3D atomic coordinates of a protein structure. It is conditioned on features computed by the model's backbone and operates by iteratively denoising a structure that begins as random 3D Gaussian noise. The denoising network, D , is trained to predict a clean structure x from a noised version $x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. The training objective minimizes the reconstruction error over a continuous schedule of noise scales σ , defined by the diffusion loss $\mathcal{L}_{\text{diff}}$:

$$\mathcal{L}_{\text{diff}} = \int \lambda(\sigma) \mathbb{E}_{x, \epsilon} [\|D(x + \epsilon, \sigma) - x\|_2^2] d\sigma. \quad (4)$$

where $\lambda(\sigma)$ is a noise scale dependent weight. In practice, the noise scales are discretized into T timesteps according to the schedule $\sigma_T = 0$ and

$$\sigma_t = \left(\sigma_{\max}^{\frac{1}{\rho}} + \frac{t}{T-1} \left(\sigma_{\min}^{\frac{1}{\rho}} - \sigma_{\max}^{\frac{1}{\rho}} \right) \right)^{\rho} \quad (5)$$

for $t \in [0, T-1]$. AF3 implements this schedule with parameters $\rho = 7$, $\sigma_{\max} = 160$, $\sigma_{\min} = 0.0004$ and $T = 200$. To introduce the watermark, we co-train a detector network, d , that learns to distinguish between watermarked and non-watermarked structures. As SynthIDBio-structure is supposed to allow sharing of the model's weights, embedding messages into the watermark is not required, and so a zero-bit watermark is sufficient.

We use two versions of the denoiser during fine-tuning: the original, pre-trained denoiser with frozen weights, denoted D_θ , and a copy with trainable weights, $D_{\tilde{\theta}}$. The watermarking objective trains the detector d to assign a high probability to structures generated by $D_{\tilde{\theta}}$ (watermarked) and a low probability to ground truth structures (x) and those from the original D_θ (non-watermarked). The watermarking loss, $\mathcal{L}_{\text{wm}}$, is formulated as a sum of binary cross-entropy losses:

$$\mathcal{L}_{\text{wm}}(x, \sigma, \epsilon) = \text{CE}(D_\theta(x + \epsilon, \sigma), D_{\tilde{\theta}}(x + \epsilon, \sigma)) + \text{CE}(x, D_{\tilde{\theta}}(x + \epsilon, \sigma)) \quad (6)$$

with

$$\text{CE}(y, \tilde{y}) = \log(1 - d(T(y))) - \log d(T(\tilde{y})). \quad (7)$$

Here, T is a transformation applied to enhance watermark robustness. This involves adding uniform noise, $T(y) = y + U(-s, s)$ with $s \in \{0.1, 0.01, 0.001\}$ Å, to the coordinates before they enter the detector. This step ensures the watermark is resilient to small perturbations, which is generally desirable in watermarks and also makes sure the folded structures can be saved in common file formats that often round to three decimal points.

We explored two strategies for integrating the watermarking objective. The first involved applying $\mathcal{L}_{\text{wm}}$ only to fully denoised structures (that is, after unrolling multiple denoising iterations during training). The second, which we found trained significantly faster and more smoothly, was to apply the watermarking loss at every step of the denoising process. The final training objective, $\mathcal{L}_{\text{total}}$, combines the diffusion and watermarking losses:

$$\min_{\tilde{\theta}} \int \mathbb{E}_{x, \epsilon} [\lambda(\sigma) \|D_{\tilde{\theta}}(x + \epsilon, \sigma) - x\|_2^2 + w_{\text{wm}} \cdot \mathcal{L}_{\text{wm}}(x, \sigma, \epsilon)] d\sigma, \quad (8)$$

where $\mathcal{L}_{\text{wm}}(x, \sigma, \epsilon)$ represents the watermarking loss evaluated at a specific noise level, comparing the outputs of $D_\theta(x + \epsilon, \sigma)$ and $D_{\tilde{\theta}}(x + \epsilon, \sigma)$. Note that we moved the weight $\lambda(\sigma)$ into the expectation,

ensuring a constant weighting of the watermark detection loss. By integrating $\mathcal{L}_{\text{wm}}$ directly into the denoising schedule, the model learns to embed the watermark throughout the entire structure generation process.

Fine-tuning was performed on 256 A100 GPUs for approximately one day. Our fine-tuning set-up mirrors the 'Fine tuning 3' stage used in AF3 (see supplementary table 6 of ref. 19) with a reduced sampling weight for disorder PDB distillation data (from 0.02 to 0.01). Compared with the overall cost of training AF3 (about 20 days), SynthIDBio-structure is fairly cheap to train; however, we also found that training for less than one day results in a noticeable reduction in detectability (data not shown). Note that SynthIDBio-structure does not introduce any overhead during inference.

Confirming watermark imperceptibility. As watermarks trained following the above approach are usually not visible when looking purely at aggregate metrics such as r.m.s.d., LDDT or template modelling score, we also look at common per-residue feature distributions. These include: pairwise intra-residue atom distances, inter-residue atom distances between backbone atoms, bond angles and torsion angles⁵³. We compute these separately for watermarked, non-watermarked and natural structures from PDB in order to qualitatively and quantitatively measure shifts in distributions of these features in the form of histograms as shown in Fig. 4b.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The in vitro validation data for SynthIDBio-sequence are available at <https://github.com/google-deepmind/synthidbio>. For SynthIDBio-structure, all scientific datasets used to create training and evaluation inputs are freely available from public sources. Data from the worldwide PDB were used for training and as templates. Structures are available at <https://files.wwpdb.org/pub/pdb/data/assemblies/mmCIF/>; 40% sequence identity clustering data at <https://cdn.rcsb.org/resources/sequence/clusters/clusters-by-entity-40.txt> and the derived data dictionary at https://files.wwpdb.org/pub/pdb/derived_data/. Training used a version of the PDB that was downloaded on 12 January 2023, whereas template search used a version downloaded on 28 September 2022. We also used the Chemical Components Dictionary, which was downloaded on 19 October 2023. Source data are provided with this paper.

Code availability

SynthIDBio-sequence code is available at <https://github.com/google-deepmind/synthidbio>. Instructions for access to the weights for SynthIDBio-structure with $s = 0.001$, our recommended model, are also available at <https://github.com/google-deepmind/synthidbio>.

48. Kortemme, T. De novo protein design—From new structures to programmable functions. *Cell* **187**, 526–544 (2024).
49. Chu, A. E., Lu, T. & Huang, P.-S. Sparks of function by de novo protein design. *Nat. Biotechnol.* **42**, 203–215 (2024).
50. Ackley, D. H., Hinton, G. E. & Sejnowski, T. J. A learning algorithm for Boltzmann machines. *Cogn. Sci.* **9**, 147–169 (1985).
51. Holtzman, A., Buys, J., Du, L., Forbes, M. & Choi, Y. The curious case of neural text degeneration. Preprint at <https://arxiv.org/abs/1904.09751> (2019).
52. Cabantous, S. et al. A new protein-protein interaction sensor based on tripartite split-gfp association. *Sci. Rep.* <https://doi.org/10.1038/srep02854> (2013).
53. Harder, T. et al. Beyond rotamers: a generative, probabilistic model of side chains in proteins. *BMC Bioinform.* **11**, 306 (2010).

Acknowledgements We thank R. Bunel, A. Cupani, R. Fergus, T. Frerix, S. Ghalebikesabi, J. Jumper, J. Kelly, D. La, V. Martin, S. Nowozin, S. Petersen, A. Petrov, U. Okereke, S.-A. Rebuffi, R. Schneider, A. Senoner, R. Shuai, A. Thillaisundaram, E. Wong, Z. Wu and A. Židek for their

contributions. We thank the team at Adaptv Bio for their contributions to the in vitro validation. We acknowledge J. Diggans for productive conversation regarding the biosecurity perspective. Finally, we thank five reviewers for their helpful comments that substantially strengthened the manuscript.

Author contributions P.K., S.G., J.W., D.H. and E.P. provided strategic guidance. S.S. provided organizational support. The SynthIDBio-sequence project was conceptualized by A.I.C.-R., D.S. and J.R. who also performed the associated research, experiments and coding. A.I.C.-R., H.P., J.R., V.Z., J.W. and A.C. conducted the SynthIDBio-sequence wet lab experiments. A.I.C.-R., D.S., J.R., H.P., V.Z. and A.D. wrote the SynthIDBio-sequence section of the manuscript, with writing feedback provided by V.Z., H.P., J.W., P.K., S.D. and A.C. For SynthIDBio-structure, D.S., G.O.-J., L.W., J.A., C.K., F.S. and M.V. performed the research, experiments and coding. D.S., V.D.B. and A.D. conceptualized SynthIDBio-structure methodology. D.S., G.O.-J., A.D. and A.I.C.-R. wrote SynthIDBio-structure section of the manuscript, and L.W., J.A., S.G., P.K., A.D. and J.R. provided feedback on the writing.

Funding This study was funded by Alphabet Inc and/or a subsidiary thereof ('Alphabet').

Competing interests This study was funded by Alphabet Inc and/or a subsidiary thereof ('Alphabet'). All authors are employees of Alphabet and may own stock as part of the standard compensation package.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-026-10965-y>.

Correspondence and requests for materials should be addressed to David Stutz or Pushmeet Kohli.

Peer review information *Nature* thanks James Diggans, Eric Horvitz, Florian Praetorius, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☐	☒ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	No software used for data collection.
Data analysis	Data analysis was conducted using Python v3.13.12 (https://www.python.org/), NumPy v2.4.6 (https://github.com/numpy/numpy), SciPy v1.17.1 (https://www.scipy.org/), seaborn v0.13.2 (https://github.com/mwaskom/seaborn), Matplotlib v3.10.9 (https://github.com/matplotlib/matplotlib), Pandas v2.3.3 (https://github.com/pandas-dev/pandas), Scikit-learn v1.9.0 (https://github.com/scikit-learn/scikit-learn), JAX v0.11.0 (https://github.com/google/jax), PyArrow v14.0.2 (https://github.com/apache/arrow), statsmodels v0.12.2 (https://github.com/statsmodels/statsmodels), Biopython v1.87 (https://github.com/biopython/biopython), PyYAML v5.4.1 (https://github.com/yaml/pyyaml), tqdm v4.64.1 (https://github.com/tqdm/tqdm), etils v1.14.0 (https://github.com/google/etils), Kalign (version 2.0), TensorFlow v2.22.0 (https://github.com/tensorflow/tensorflow), zstandard v0.25.0 (https://github.com/indygreg/python-zstandard), and Colab (https://research.google.com/colaboratory ; multiple releases from 2025-06-16 to 2026-06-10).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Provide your data availability statement here.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

N/A

Reporting on race, ethnicity, or other socially relevant groupings

N/A

Population characteristics

N/A

Recruitment

N/A

Ethics oversight

N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

For SynthIDBio-sequence, our sample size was determined by a power analysis on the difference in mean binding affinity between watermarked and non-watermarked binders using a two-sample t-test. Assuming a standard deviation in log binding affinity, $\log(K_D)$ of 2.5, we determined the number of designs needed to detect changes in $\log(K_D) \geq 1$ with 80% power. Taking into account practical constraints, we thus selected 90 designs on a single plate per target, per setting.

Data exclusions

No data was excluded from this study.

Replication

For SynthIDBio-sequence, each binder evaluated in vitro had at least two technical replicates.

Randomization

N/A

Blinding

N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A