Improved multimodal protein language model-driven universal biomolecules-binding protein design with EiRA

Wenwu Zeng[1], Haitao Zou[1], Xiaoyu Li[1], Xiaoqi Wang[2,3,*], Shaoliang Peng[1,*]

[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China

[2] School of Computer Science, Northwestern Polytechnical University, Xi'an, 710129, China

[3] Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University, Ministry of Industry and Information Technology, Xi'an, 710129, China

*To whom correspondence should be addressed. xqw@nwpu.edu.cn and slpeng@hnu.edu.cn

Abstract

The interactions between proteins and other biomolecules, such as nucleic acids, form a complex system that supports life activities. Designing proteins capable of targeted biomolecular binding is therefore critical for protein engineering and gene therapy. In this study, we propose a new generative model, EiRA, specifically designed for universal biomolecular binding protein design, which undergo two-stage post-training, i.e., domain-adaptive masking training and binding site-informed preference optimization, based on a general multimodal protein language model. A multidimensional evaluation reveals the SOTA performance of EiRA, including structural confidence, diversity, novelty, and designability on eight test sets across six biomolecule types. Meanwhile, EiRA provides a better characterization of biomolecular binding proteins than generic models, thereby improving the predictive performance of various downstream tasks. We also mitigate severe repetition generation in the original language model by optimizing training strategies and loss. Additionally, we introduced DNA information into EiRA to support DNA-conditioned binder design, further expanding the boundaries of the design paradigm.

Introduction

Over hundreds of millions of evolutionary years, proteins have emerged as the primary molecular mediators of biological processes. As of 28 March 2025, the UniParc database [1] archives 916,871,247 non-redundant protein sequences, yet this constitutes a vanishingly small fraction of the theoretically accessible sequence space. The engineering of novel proteins with tailored functionalities holds transformative potential for biomedical innovation, particularly in therapeutic development like immunotherapy and gene editing. Conventional approaches, including energy minimization-based rational design and physics-driven modeling, however, struggle to efficiently navigate this vast combinatorial landscape. Artificial Intelligence (AI)-enabled high-throughput design paradigms now promise to overcome this bottleneck. Post-2022, AI-generated protein candidates have surged at an unprecedented rate, outpacing the cumulative progress of the preceding decade, a testament to the catalytic role of AI in reshaping protein design paradigms [2-4].

Contemporary AI-driven de novo protein design frameworks typically follow a three-stage pipeline: (1) Functionally oriented backbone generation prioritizes structural stability and designability through geometric constraints derived from target motifs (e.g., binding interfaces or catalytic sites), employing probabilistic sampling via diffusion models (RFdiffusion [5], Chroma [6], and SCUBA-D [7]) or flow-matching (Proteína [8]) to explore conformationally diverse scaffolds; (2) Spatial-aware sequence design tools such as ProteinMPNN [9], CarbonDesign [10], MapDiff [11], and ESM-IF1 [12] then populate amino acid residues onto these backbones using graph neural networks to encode atomic-level spatial constraints, followed by autoregressive decoding to optimize sequence-structure compatibility; (3) Computational validation via protein structure predictors like AlphaFold2/3 [13, 14] or RoseTTAFold [15] assesses structural fidelity and stability for early screening. Beyond these efforts, multimodal methods like PROTSEED [16] attempt to jointly design protein sequence and structure [17, 18], while cross-modal architectures like ProteinDT [19] and protein language models (PLM) ProGen [20] bridge functional semantics (e.g., "ATP-binding hydrolase") with protein sequence for natural language-guided sequence generation. Collectively, these efforts underscore the transformative capacity of AI to navigate the combinatorial vastness of protein space, accelerating the discovery of functional proteins with unprecedented precision.

To the best of our knowledge, ESM3 [21] proposes the first large-scale multimodal protein language model (MPLM) that unifies sequence, structural, and functional information within a unified neural architecture (The largest version has 98 billion parameters). The model processes diverse biological features, including primary sequences, tertiary structures, secondary structure (SS), solvent accessibility (SA), and InterPro annotations [22] through specialized tokenization strategies, subsequently constructing a multi-track Transformer framework capable of joint input processing and multimodal output generation. Through self-supervised learning on extensive protein datasets, ESM3 effectively captures cross-modal dependencies and coupling relationships, enabling coherent multimodal protein editing and generation. Consequently, the model successfully explored the latent evolutionary space of protein sequences and generated a functional green fluorescent protein (GFP) variant with a sequence similarity of 58% to natural counterparts by prompting the core catalytic residues, effectively traversing millions of evolutionary years.

However, most biological processes involve intricate molecular interactions rather than isolated protein functions. Critical biological systems such as the Cas9-sgRNA complex mediating DNA binding and cleavage through protospacer adjacent motif (PAM) recognition, hemoglobin's cooperative oxygen transport via heme-globin coordination, and MHC-TCR-peptide interactions driving immune activation [23]. All of these crucial activities require precise biomolecular interplay. Designing proteins to interact with various biomolecules is critical for many life science fields, such as drug development and gene editing. The universal ESM3, built on indiscriminate training data, remains underexplored regarding the complex binding modes of different ligands. Moreover, biomolecular interactions are typically guided by conserved motif sequences. Generating novel proteins while understanding these key sites can preserve the native core functions of the template protein and efficiently explore the potential sequence

space.

In this study, we propose EiRA, an MPLM specialized for biomolecule-binding protein generation. EiRA undergoes two-stage post-training on ESM3: biomolecular-binding domain adaptive masking training followed by binding site-guided preference optimization (named EiRAD). Unconditional generation evaluations reveal that designed proteins exhibit high diversity, novelty, foldability, and functionality. Systematic and multidimensional assessments across 8 test sets demonstrate that our method designs molecular-binding proteins with higher confidence than that of ESM3 (including small/medium/large versions) and RFdiffusion+ProteinMPNN in both single-sequence and complex evaluations. Simultaneously, EiRA's embeddings effectively characterize natural proteins to support diverse biomolecule-binding downstream tasks, achieving dual benefits in representation and generation. Notably, we identified severe duplicate generation under binding-motif conditions in ESM3 medium (ESM3m, 7B) and large (ESM3l, 96B) models, substantially undermining structural confidence and sequence foldability. This issue emerged during both training stages and was mitigated through repetition penalties and loss adjustment based on direct preference optimization (DPO). Furthermore, whereas ESM3 lacks non-protein biomolecular knowledge, EiRA integrates DNA sequence information via cross-attention to enable multimodal protein editing conditioned on target DNA sequence, addressing critical real-world target-based protein generation scenarios.

Methods

Data curation

We systematically constructed all datasets for this study through a multi-stage curation pipeline, comprising two training sets for self-supervised adaptation and DPO training and 8 independent biomolecular-binding test sets (see Tables 1 and 2).

To align the native knowledge space of the ESM3 with universal biomolecular-binding characteristics, we developed a specialized post-training dataset named UniBind40 via the following protocol (top of Fig. 1): Initial ~54 million biomolecular-interacting protein sequences were retrieved from UniProtKB database [1] up to August 17, 2024, followed by MMseqs2 [24] clustering at 0.4 sequence identity threshold to derive ~6.4 million non-redundant cluster representatives. Structural validation proceeded through two pathways: for cluster representatives with existing AlphaFold2 [25] predictions (~4.6 million), we retained only high-confidence structures (pLDDT > 0.7); for sequences lacking precomputed AlphaFold2 structures, we implemented a cascaded prediction strategy using ESM3 (1.5B) for initial high-throughput modeling, followed by ESMFold refinement for cases with suboptimal confidence (ESM3 pLDDT < 0.7), ultimately requiring ESMFold pLDDT > 0.7 for structural acceptance. In addition, samples in which the number of unknown amino acids exceeded 10% of the total sequence length were discarded. This rigorous filtration yielded 3,735,303 qualified biomolecule-binding proteins with 1,264,943,457 residues, designated as UniBind40. The dataset construction emphasizes structural reliability through confidence validation while maintaining sequence diversity via aggressive deduplication.

Protein-ligand binding data were obtained from the BioLip2 database [26] (accessed November 2, 2024), including five categories: DNA-binding proteins (DBPs, 45,686 records), RNA-binding proteins (RBPs, 160,800), Metal-binding proteins (MBPs, 197,122), Peptide-binding proteins (PBPs, 39,065), and Regular-binding proteins (ReBPs, 460,725; e.g., ATP/HEM-binding). After merging redundant chains, we derived 24,459 (DBP), 128,464 (RBP), 125,873 (MBP), 33,755 (PBP), and 252,976 (ReBP) unique chains. Sequence redundancy was reduced using MMseqs2 clustering (40% identity threshold) followed by CD-HIT-2D [27] filtering (40% identity) against UniBind40, yielding 1,490 DBP, 2,585 RBP, 1,656 MBP, 2,667 PBP, and 2,677 ReBP chains. Random sampling selected 1,214 DBP, 2,332 RBP, 1,419 MBP, 2,420 PBP, and 2,319 ReBP chains for the DPO training set (named BioDPO), with remaining chains assigned to independent test sets. For protein-protein interaction (PPI) data, sequences with annotated interaction domains were sourced from Swiss-Prot [28]. After redundancy removal with UniBind40 and filtering for AlphaFold2 pLDDT>0.7, 1,047 sequences were added to BioDPO, and 351 were reserved for the PPI test set.

Independent test samples for DBPs, RBPs, MBPs, PBPs, and ReBPs employed binding motif prompts containing sequences and coordinates of 4 flanking residues around the real binding site (8 for MBPs/ReBPs). PPI test samples directly utilized full domain sequences/coordinates as prompts. To evaluate multi-domain design capability, multi-domain DBP/RBP test sets were constructed by: (1) selecting UniProtKB entries with ⩾2 annotated domains; (2) removing redundant sequences and those with max identity>40% against UniBind40 via MMseqs2 and CD-HIT-2D; (3) retaining sequences with total domain length <40% of the full sequence. This produced 97 multi-domain RBPs and 64 multi-domain DBPs, processed with domain-level prompts following PPI samples in BioDPO. All of this data is available at https://github.com/pengsl-lab/EiRA.

Feature generation

For each protein in the UniBind40, the ESM3 feature representation is applied. The DSSP [29] is utilized to calculate eight protein secondary structure (SS) states. The solvent accessibility (SA) is calculated by the Shrake-Rupley algorithm in the Biotite package [30]. We downloaded the InterPro annotations of all UniProtKB sequences at https://ftp.ebi.ac.uk/pub/databases/interpro/current_release/protein2ipr.dat.gz and then extracted the corresponding UniBind40 records from them. For BioDPO and all test sets, only sequences and atomic coordinates are utilized as prompts. All these feature sources are fed into the ESM3 tokenizer to form numerical tokens for training.

Evaluation metrics

We use the evaluation metrics commonly used in previous studies about protein design to assess our method. For a designed protein sequence, we used three models, i.e., **ESM3 structure decoder**, **ESMFold** [31], and **AlphaFold3** (AF3, re-implementation version of Boltz [32]) to predict its tertiary structure and obtained a series of metrics. The ESM3 structure decoder is used to convert the structural tokens output by MPLM into backbone atomic coordinates; ESMFold is an MSA-free structure predictor based on a protein sequence language model; AF3 is a diffusion model-based MSA-depend complex structure predictor. The evaluation of the designed sequences by different structural predictors can prevent biased assessments of a single structure prediction model. Specifically, predicted Template Modeling Score (**pTM** ↑) assessed the overall stability and rationality of the predicted structure; predicted Local-Distance Difference Test (**pLDDT** ↑) evaluated the local positional validity of residues; and predicted Aligned Error (**pAE** ↓) depicts the confidence of the relative positions of pairs of residues in the predicted structure. In addition, we calculated the self-consistent Root Mean Square Deviation (**scRMSD** ↓) of the designed backbone against the ESMFold and AF3 structures to assess the designability of the backbone.

For the evaluation of the design sequence and its ligand binding ability, the interface pTM (**ipTM** ↑), interface pLDDT (**ipLDDT** ↑), and interface pAE (**ipAE** ↓) of AF3 were employed to assess the accuracy of the interface interactions between subunits in the predicted complexes. Additionally, we conducted sequence validation using three complementary approaches: CD-HIT clustering to assess **diversity**, MMseqs2 similarity searches to evaluate **novelty**, and InterProScan scans to verify **functionality**.

Table 1. The composition of the pre-training and DPO training sets

Data set	UniBind40 (for post-training)	BioDPO (for DPO)					
Type	Mixed biomolecular binding	DBP	RBP	PBP	ReBP	MBP	PPI
N_{seq}	3,735,303	1,214	2,332	2,420	2,319	1,419	1,047
N_{res}	1,264,943,457	395,728	695,230	739,251	818,405	530,400	513,642
Mean length	338	326	298	305	353	373	490
Source	UniProt/AF DB/InterPro	BioLip2/PDB					SwissProt/AF DB

Table 2. The composition of the 8 test sets across 6 biomolecular types

Type	DBP	RBP	PBP	ReBP	MBP	PPI	MDBP	MRBP
N_{seq}	275	252	247	358	237	346	63	97
N_{res}	122,578	114,985	79,513	124,423	83,761	179,072	37,772	67,459
Mean length	445	456	321	347	353	517	599	695
Mask rate	0.893	0.805	0.870	0.878	0.942	0.784	0.771	0.739
Source	BioLip2/PDB					SwissProt/AF DB		

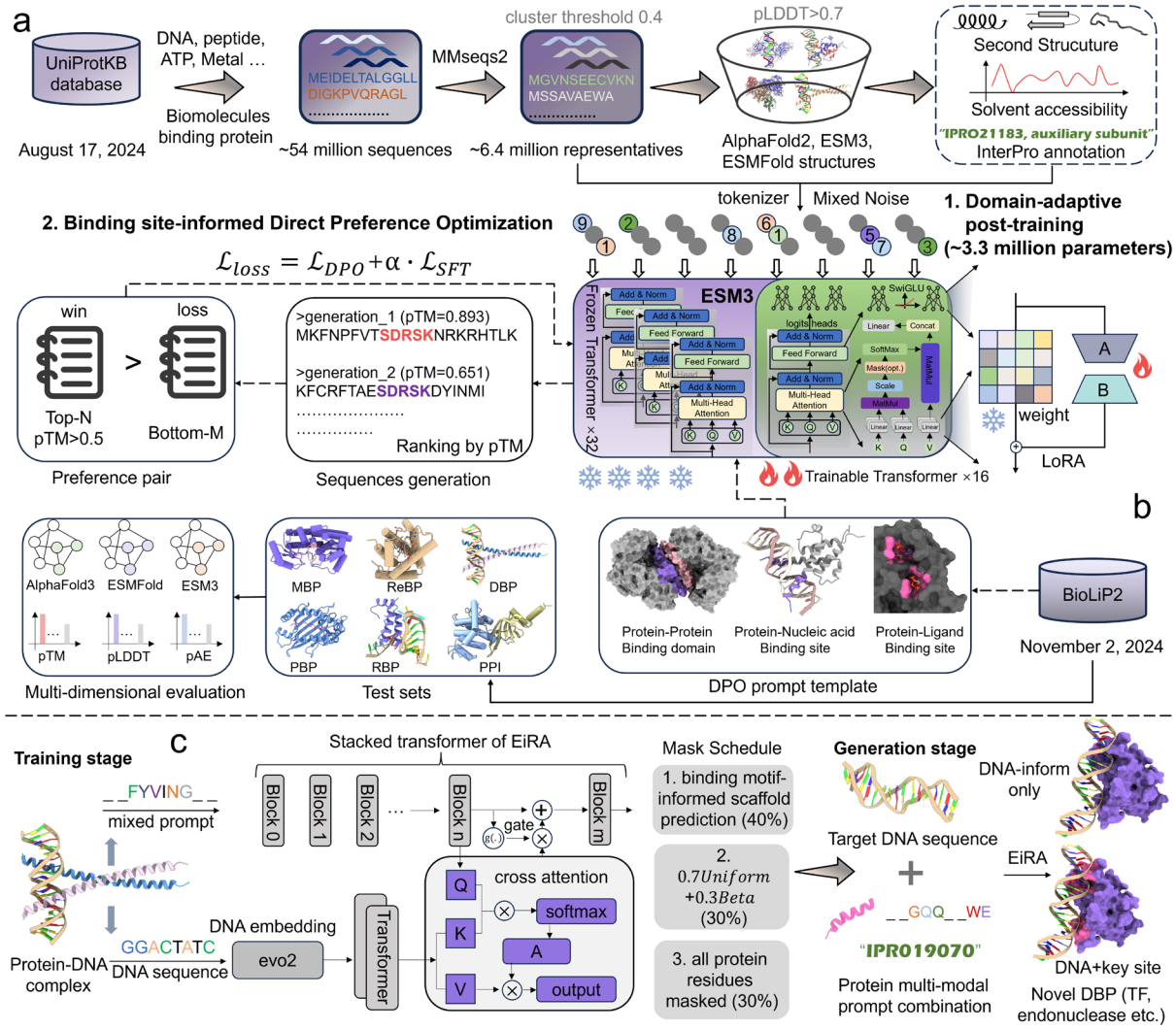

Fig.1 | EiRA is a multimodal protein language model specifically designed for generating biomolecule-binding proteins. a. We collected ~54 million universal biomolecule binding sequences from UniProtKB and selected ~6.4 million representatives using MMseqs2; then AF2, ESM3, and ESMFold were used to filter proteins with low structural confidence (pLDDT<0.7, remaining 3,735,303, named UniBind40); and finally, corresponding SS, SA, and functional annotations were calculate using the DSSP, Shrake-Rupley, and InterProScan tools. **b.** We first trained the post-32 transformer blocks (LoRA strategy) of ESM3s on UniBind40 by mixing noise; then, protein-biomolecule complex data collected from the BioLip database were used to perform binding site-informed preference optimization. **c.** The large genome language model Evo2 is used to extract the embedding of DNA sequences, which is subsequently introduced into EiRA's post-4-layer transformer for mask training using a gated cross-attention mechanism; the resulting EiRA model enable DNA-conditioned DBP generation.

Domain-adaptive self-supervised post-training

The general MPLM ESM3 has learned a vast amount of multimodal world knowledge about proteins but lacks focus on specific domains such as biomolecular binding. To transfer the general ESM3 knowledge space and focus it on the field of biomolecular binding, we continue to perform a mask-based self-supervised post-training strategy for ESM3-1.4B MPLM on our well-curated UniBind40 training dataset. Although ESM3 includes three versions: ESM3 small (ESM3s), ESM3 medium(ESM3m), and ESM3 large (ESM3l) with 1.4 billion, 7 billion, and 96 billion paramters respectively and Hayes *et al.* [21] reports that larger model have stronger generative capabilities, only the 1.4B version of ESM3 has open weights and source code (https://huggingface.co/EvolutionaryScale/esm3-sm-open-v1).

For different modalities, we employed different noise scales as ESM3 to simulate conditional generation tasks, i.e., cosine for structure track, cubic for structure coordinates, and square root for SS, SA, and function track. Unlike ESM3, for the sequence track, we used a betalinear70 noise composed 30% $\beta eta(3,9)$ and 70% $Uniform(0,1)$, which offers a higher mask rate than ESM3 training strategy to optimize generation tasks (Fig. S1). We experimented with various fine-tuning strategies, including directly updating the original weights, Low-Rank Adaptation (LoRA) [33], and various variants. Ultimately, it was found that the vanilla LoRA achieved the best performance. Previous study has demonstrated that partial parameter fine-tuning can outperform full parameter fine-tuning on specific tasks and datasets [34]. Here, we apply LoRA to the last 16 transformer blocks and the token classification head of each track, totaling 3,368,046 trainable parameters (only 0.24% of all ESM3 1.4B parameters). The rank of LoRA is set to 8, the scaling factor a is 32, and the dropout rate is 0.2. Cross-entropy is used to calculate the loss between the true tokens and the predicted tokens for each track. Since we observed a severe phenomenon of repeated generation of tokens conditioned at binding sites in ESM3, if the predicted tokens at 7 consecutive adjacent positions are consistent, the loss at these positions is multiplied by a penalty factor of 2. The 8-bit quantized AdamW optimizer is used to update the parameters, with a learning rate of 4e-04, β_1 of 0.9, β_2 of 0.95, and a weight decay of 5e-04. The context length is 768. Long sequences are split, and short sequences are filled with pad tokens. The batch size is 160. We implemented training using the standard Distributed Data Parallel strategy of Pytorch and employed a mixed precision training with BF16 and FP32. All training was performed on 8 NVIDIA H20 GPUs.

Binding site-informed Preference optimization

The Reinforcement Learning from Human Feedback (RLHF) strategy introduced by ChatGPT aims to align the output of large language models with human expectations [35]. In simple terms, this paradigm forces the model to favor those responses among multiple generated answers that align with human expectations. The recent DeepDeek-R1 [36] achieved SOTA performance for LLM using only a RL strategy called Group Relative Policy Optimization without any supervised fine-tuning (SFT), which demonstrated the advantages of RL. This concept has been borrowed in the field of PLM [21, 37]. ESM3 uses an Iterative Reasoning Preference Optimization [38] strategy composed of a Direct Preference Optimization (DPO) [39] loss and an NLL loss conjunctions to bias the model towards design proteins with high pTM and low backbone constrained site RMSD (cRMSD).

Algorithm description: To enhance the comprehensive generation quality of EiRA under the condition of binding motif, we initially used only the DPO strategy to train the EiRA model, but crashed duplicate generation occurred similar to ESM3. Therefore, we introduce an additional SFT loss to impose a larger penalty on mispredictions of repeatedly generated regions. Consequently, we propose a binding site-informed DPO strategy combined with an SFT loss as follows (Equations 1, 2, and 3):

$$Loss = \mathcal{L}_{DPO} + a \cdot \mathcal{L}_{SFT} \tag{1}$$

$$\mathcal{L}_{DPO} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\left(\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)\right)\right] \tag{2}$$

$$\mathcal{L}_{SFT} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}, \log(p_{i,c}) \tag{3}$$

where π_θ and π_{ref} are respectively the learnable policy model and the frozen reference model; x, y_w, and y_l are the input prompt sequence, the chosen and rejected designed sequences separate; a is the weight of SFT loss and set to 0.2; b is used to control the degree of deviation between the policy model and the reference model and set to 0.1. $\mathcal{L}_{SFT}$ directly employs a multi-class cross entropy loss. N is the total number of residues in the current batch; C is the number of categories; $y_{i,c}$ indicates whether the residue i belongs to category c, and $p_{i,c}$ indicates the probability that the model predicts that the residue i belongs to category c. To suppress duplication, when the maximum predicted probabilities of 5 consecutive positions (or more) all point to the same amino acid type, the loss of this region is multiplied by a penalty coefficient of 2.

Preference pair generation: For BioDPO, we extracted the amino acid types and coordinates of the binding site and *n* (*n*=4 for DBP, RBP, and PBP; *n*=8 for MBP and ReBP) positions upstream and downstream as hints (binding domain for PPI), and other positions were masked and then predicted. Each prompt generated 30 designed proteins. We construct preference pairs according to the following rules: **1.** If the maximum pTM of the current generated set is greater than threshold *p*, the Top-M sequences with the largest pTM are chosen as "good generation", and the Bottom-6 with the smallest pTM are defined as "bad generation". The minimum pTM in the chosen set needs to be greater than *q*. Note that (*p*, M, *q*)=(0.7, 5, 0.6) or (0.6, 4, 0.5) or (0.5, 3, 0.5). **2.** The sequence whose minimum (or maximum) pTM of the generated set of a sequence is greater than 0.8 (or lower than 0.5) is discarded. **3.** The sequence whose number of unknown amino acids is greater than 8% of the sequence length is considered a "bad generation". **4.** If the length of the largest repeated substring is greater than 7 and the total number of the Top-2 amino acids in length is greater than 40% of the sequence length, the current generation is considered a bad sample. As a result, we generate a training dataset of 224,891 preference pairs for preference optimization.

Results

EiRA generates biomolecular binder with higher confidence

We assessed the generative capabilities of EiRA against ESM3 using three distinct evaluation strategies: (1) unconditional generation, (2) single-sequence scoring, and (3) evaluation of binding ability to specific biomolecules.

Unconditional generation: Unconditional generation is a key benchmark for protein design models, as it evaluates their capacity to navigate the vast potential sequence space. In this task, the model generates sequences based solely on a specified length, with all other contextual information masked. We generated a dataset of 100,000 sequences by first sampling 1,000 distinct lengths from three intervals: [50, 500], [501, 1000], and [1001, 1500], with sampling probabilities of 0.6, 0.3, and 0.1, respectively. For each sampled length, 100 unique sequences were generated. Subsequently, the tertiary structure for each of these 100,000 designed sequences was predicted using the ESM3 structure decoder to calculate key quality metrics (see Fig. 2A). EiRA demonstrated superior performance, achieving average pTM and

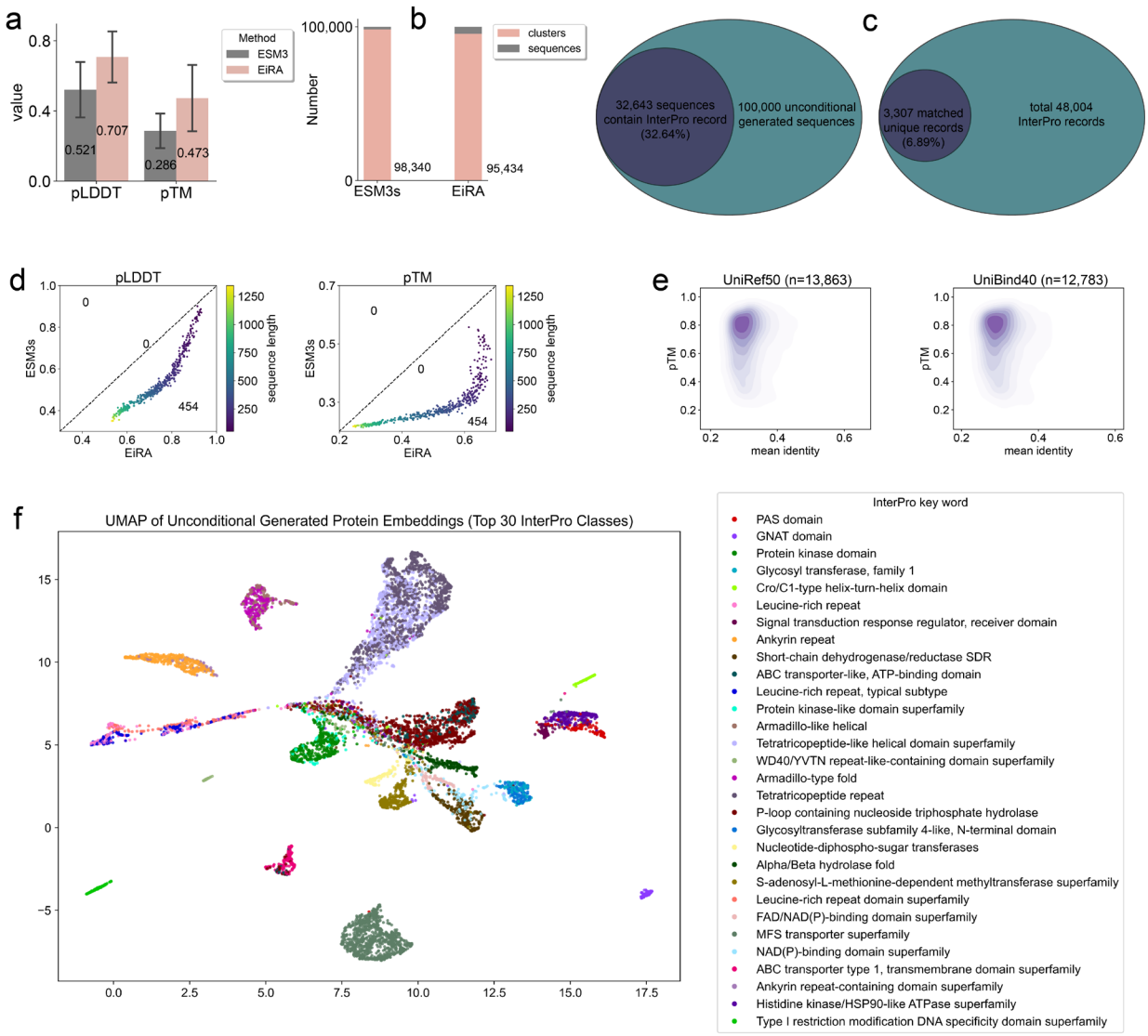

Fig.2 | Unconditional generation evaluation. a. Direct comparison of pLDD and pTM of 10,000 generated sequences. **b.** Number of clusters of 100,000 generated sequences (CD-HIT tool with a sequence identity threshold of 0.4 for clustering). **c.** Venn diagram of the InterPro annotation of the designed sequence. **d.** Head-to-Head comparison of mean pTM (or pLDDT) for sequences of equal length. **e.** Kernel density for the average sequence identity (against UniRef50 and UniBind40, using MMseqs2 easy-search) and pTM. **f.** Embedding visualization of sequences with Top-30 InterPro annotation numbers, including several typical biomolecular-binding domains.

pLDDT scores of 0.473 and 0.707, which are 35.70 and 65.38% higher than those of ESM3s, respectively. A direct comparison of metrics for sequences of identical lengths reveals that EiRA consistently outperforms ESM3 across all length ranges (see Fig. 2d). An observed trend for both models is the degradation of pTM and pLDDT scores with increasing sequence length, suggesting that the modeling of long-range interactions remains a challenge for both generative and predictive models.

The **diversity** of generated sequences is a critical indicator of a model's generative scope. Following the methodology of Proteína [8], we quantified diversity by clustering the 100,000 designed sequences using the CD-HIT program with a sequence identity threshold of 40%. This analysis yielded 95,434 clusters for EiRA and 98,340 for ESM3s (see Fig. 2B). A high cluster count signifies low sequence similarity among the generated samples. In addition, the MMseqs2 tool was used to search the similar sequences of 100,000 designed sequences against the UniRef50 and UniBind40 databases. Consequently, for UniRef50 and UniBind40, there are 13,863 and 12,783 designed sequences that were matched to a total of 293,3813 and 149,5174 records, respectively. We averaged the identities of the different records matched to the identical sequence and plotted the kernel densities against pTM (see Fig. 2E). The densities in the upper left corner indicate generally **low similarity** and **high structural confidence**. These results indicate that EiRA not only designs structurally superior proteins but also maintains a high degree of sequence diversity and novelty, comparable to that of ESM3s.

To validate **functionality** across 100,000 unconditionally generated sequences, we performed functional annotation prediction using InterProScan. Of these, 32,643 sequences matched known InterPro entries (see left of Fig. 2C), yielding 149,571 redundant InterPro records (Fig. S8) comprising 3,307 unique IDs (see right of Fig. 2C). This demonstrates EiRA's capacity to generate functional protein sequences unconditionally. We ranked the Top-30 non-redundant InterPro annotations (assigning each protein its most frequent InterPro ID) and visualized corresponding sequences via UMAP (see Fig. 2F). Notably, these include multiple biomolecule-binding domains (PAS, ATP-binding, and Cro/C1-type HTH domains),

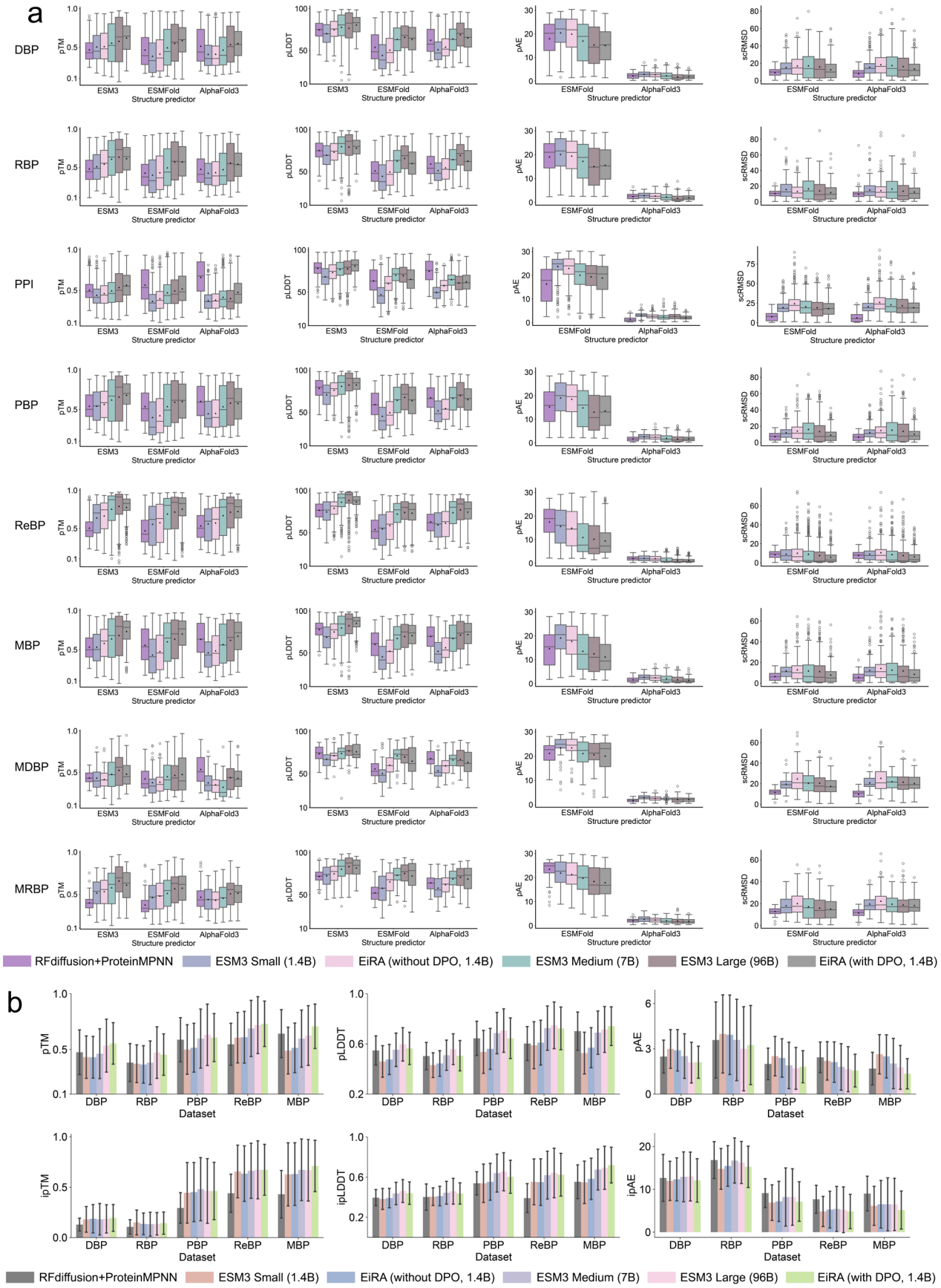

Fig.3 | EiRA outperforms ESM3 on biomolecule binder generation over single sequence (a) and complex evaluation (b). Since ESM3 medium and ESM3 large are not open-source, the generations come from the Forge service at https://forge.evolutionaryscale.ai/. Boltz was employed as a replacement for AF3. MSA for complex structure prediction was generated using MMseqs2 against UniRef50 and Small BFD (taking only the first non-consensus sequence from each cluster in BFD) database.

exhibiting distinct clustering patterns. The PAS domain (ubiquitous in eukaryotes and prokaryotes) mediates protein homo-/heterodimerization, while the Cro/C1-type HTH domain represents one of nature's most efficient compact DNA recognition modules. Visualization of 16 representative samples (Fig. S7) further confirms EiRA's generation of biomolecule-binding domains (targeting DNA, RNA, ATP, GTP, UDP, NAD, and AMP), including classical DNA-binding architectures (HTH, BTB, and zinc finger C2H2 domains) with rich secondary structure. Significantly, gen_545045 (Fig. S7, lower right) integrates three distinct domains (NAD-binding, UDP-binding, and a central domain), highlighting EiRA's capacity for complex multi-domain protein generation.

Collectively, large-scale validation demonstrates that EiRA unconditionally generates proteins with enhanced structural confidence and biomolecule-binding functionality over ESM3s, while maintaining high diversity.

Monomer evaluation: This experimental protocol involved systematic evaluation of five MPLMs across 8 benchmark datasets: ESM3 small (1.4B, ESM3s), ESM3 medium (7B, ESM3m), ESM3 large (96B, ESM3l), EiRA without DPO (1.4B), and DPO-enhanced EiRA (1.4B, EiRAD). In addition, a mainstream scheme, i.e., RFdiffusion (backbone generation)+ProteinMPNN (sequence design), was used as the SOTA control. Generated sequences were subsequently processed through three structural predictors to obtain quantitative metrics (Fig. 3A).

Notably, EiRA without DPO demonstrated superior performance over ESM3s, confirming that domain-adaptive post-training effectively transfers MPLM knowledge to biomolecular binding contexts, yielding more physiologically plausible biomolecule-binding protein designs. The observed performance hierarchy (ESM3l >ESM3m > EiRA >ESM3s) aligns with established scaling laws [40], where increased model parameters correlate with enhanced learning capacity. However, EiRAD produced sequences with metric improvements surpassing ESM3m and achieving parity with ESM3l. Intriguingly, comparative analysis revealed EiRAD's superior pTM scores versus ESM3l (indicative of enhanced global stability) alongside lower pLDDT values (suggesting reduced local structural precision). Both ESMFold and AF3 evaluations confirmed EiRAD's design superiority through significantly lower scRMSD values, validating methodological reliability.

Dataset stratification revealed pronounced metric disparities: MBP and ReBP ligands exhibited elevated pTM/pLDDT values, likely attributable to evolutionary conservation facilitating pattern recognition in training data, and the average length is relatively short. Conversely, PPI, MDBP, and MRBP designs showed reduced performance, potentially stemming from template sourcing limitations (AlphaFold DB vs. experimental PDB structures) that propagate structural inaccuracies. Crucially, EiRAD achieved ESM3l-comparable pTM scores in multi-domain protein tests, demonstrating effective handling of complex structural architectures.

An analysis of performance reveals a notable exception to EiRAD's general superiority: it is inferior to RFdiffusion on the PPI and MDBP test sets. We attribute this to the nature of these two benchmarks, which contain a relatively large number of prompt residues (prompt rates: 21.59% and 22.89%) that heavily constrain the design task. RFdiffusion's PDB-centric training predisposes it to generate sequences that conservatively "infill" the structural scaffold in a manner similar to natural proteins. This yields designs that are readily validated with high confidence by structure predictors. Conversely, EiRA, which leverages a more diverse set of predicted structures, maintains its tendency to generate novel sequences. For these specific, highly-constrained tasks, this novelty is a double-edged sword. The resulting unfamiliar sequences challenge structure predictors, leading to lower confidence scores and masking EiRA's underlying generative capability.

In general, following domain-adapted post-training, the EiRA framework exhibited enhanced capabilities in biomolecular-binding protein design compared to the native ESM3s architecture. Building upon this foundation, the binding site-informed DPO-enhanced EiRAD demonstrates remarkable parameter efficiency, achieving performance surpassing (or parity with) the 96B-parameter ESM3l model despite operating with merely 1.4B parameters. Additionally, EiRAD's superior performance over the SOTA method, RFdiffusion, across most test sets underscores the effectiveness of our method.

Complex Evaluation: We input the designed sequences from the DBP, RBP, PBP, ReBP, and MBP test sets, along with their respective ligand chains in the PDB template, into AF3 to obtain predicted complex structures and evaluation metrics (see Fig. 3b). EiRA consistently outperformed the original ESM3s models across all test sets in terms of pLDDT, even without DPO enhancement. Furthermore, the DPO-enhanced EiRAD surpassed ESM3m and ESM3l in multiple comparisons. For instance, EiRA achieved higher pTM than both ESM3l and ESM3m on three out of five test sets (DBP, ReBP, MBP). Additionally, EiRA with DPO exhibited lower ipAE than ESM3m and ESM3l across all test sets. Notably, EiRA demonstrated superior performance on the MBP test set for all metrics, indicating unique advantages. Regarding ipTM, while EiRA was significantly higher than other methods on MBP, performance differences were negligible on other datasets. This pattern may arise since the generative prompts directly inherit interface residue properties from the PDB templates. We also observed performance that was comprehensively superior to RFdiffusion, both in terms of overall pTM and interface ipTM scores.

To intuitively illustrate the advantages of the proposed method, we visualized four designed cases: an RBP (PDB ID: 6dt8, Bacteriophage N4 RNA polymerase II elongation complex 1), a DBP (PDB ID: 1gm5, RecG protein, DNA-specific helicase), an MBP (PDB ID: 8on7, FMRFa-bound Malacoceros FaNaC1), and a PBP (PDB ID: 6wtb, Sort-Tagged Drosophila Cryptochrome). As shown in Fig. S1-S4, EiRA generated proteins with higher confidence (pTM: 0.79, 0.73, 0.828, 0.901; ipTM: 0.84, 0.68, 0.912, 0.835) than all ESM3 versions in all cases. Crucially, AF3 predicted successful target ligand binding for all four EiRA designs, whereas designs for the RBP (by ESM3l) and DBP (by ESM3m and ESM3s) were predicted not to bind their target nucleic acids. For DBP (1gm5) and MBP (8on7), the structural confidence of the EiRA generations is even higher than that of the AF3 structure of the natural sequence. Furthermore, EiRA designs maintained contact between the key prompt motif (highlighted in pink) and target ligand in the RBP, PBP, and MBP cases, indicating successful inheritance of the template's core interactions. Low sequence identities (18.37, 17.73, and 19.60%) for these three cases confirm the ability to explore broader sequence and structural space while preserving original interaction.

Relief from repeated generation

Repetitive generation in large language models (LLMs) has long been recognized within NLP. In our experiments, significant repetitive generation was observed in biomolecule-binding motif-conditioned ESM3, particularly in the medium (ESM3m) and large (ESM3l) versions. This resulted in a severe skew in amino acid type distribution (Fig. 4a), with ESM3l and ESM3m exhibiting a pronounced bias towards Alanine and Leucine, respectively. In contrast, ESM3s and EiRAD produced more balanced distributions. Such bias towards specific residues compromises sequence diversity and foldability.

Formally, a protein sequence was defined as repetitive if its longest single-amino acid repeat substring exceeded length 5 and the combined frequency of the top two amino acids exceeded 40% of the

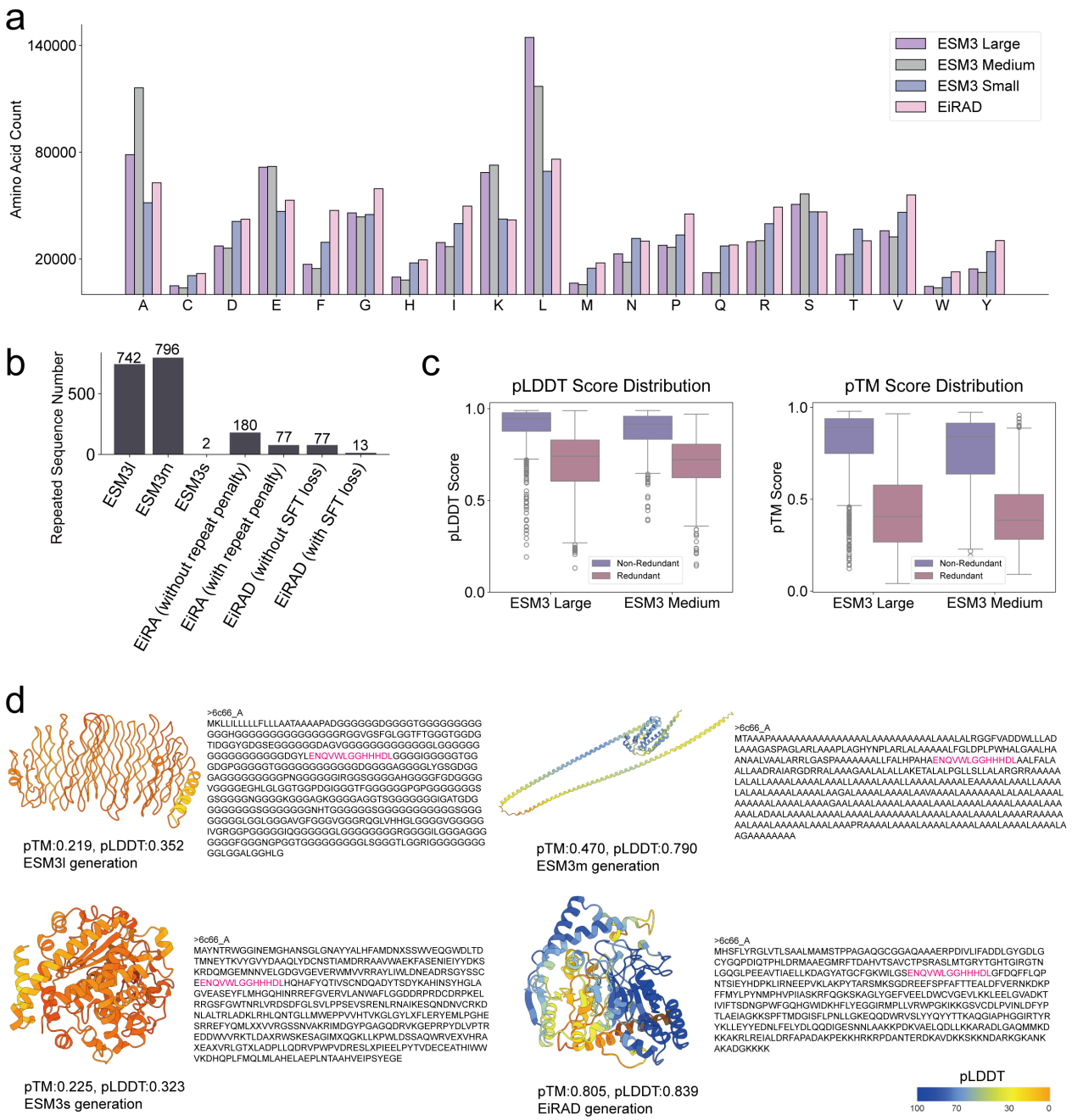

Fig.4 | EiRAD rescues the duplicate generation. a. Amino acid distribution of the generated sequences for the test sets. **b.** Statistics on the number of duplicate sequences on the test sets. **c.** Repeated designed sequences lead to extremely deteriorated structure confidence. **d.** The case (RNA-binding protein, PDB ID: 6c66_A) of repeated sequences generated by ESM3l and ESM3m. The prompt residues are highlighted in pink. The structures are predicted by Boltz.

protein length. Counting repetitive sequences generated by different models across 8 test sets (Fig. 4b) revealed three key phenomena: **1.** ESM3l and ESM3m generated a high number of repetitive sequences (742 and 796, respectively); **2.** post-training directly on ESM3s exacerbated repetitive generation; and **3.** our proposed loss-penalization and DPO+SFT strategies effectively mitigated repetitive generation during post-training and DPO phases, respectively. Sequences designed by ESM3l and ESM3m were categorized into repetitive and non-repetitive groups. Visualization of their pLDDT and pTM scores (Fig. 4c) shows that the repetitive group scored significantly lower on both metrics than the non-repetitive group, indicating reduced structural confidence in repetitively sequences.

Illustratively, the structures generated for target 6c66_A (chain A of CRISPR RNA-guided surveillance complex) by the four models clearly demonstrate ESM3's failure modes (Fig. 4d). Under identical RNA-binding residue prompts, ESM3l and ESM3m exhibited biases towards Glycine and Alanine, respectively, failing to form a stable tertiary structure. Although ESM3s avoided repetitive generation, their structural confidence remained low (pLDDT: 0.323, pTM: 0.225). Conversely, EiRAD generated proteins with diverse amino acid residues that formed a stable tertiary structure (pLDDT: 0.839, pTM: 0.805). In conclusion, the higher-parameter ESM3l and ESM3m models, typically considered more generative than ESM3s, suffer from severe repetitive generation under biomolecule-binding motif conditioning. Our improved loss strategy mitigates the repetitive generation arising during ESM3s-based training, achieving simultaneous improvements in structural confidence and amino acid diversity.

Table 1. The composition of the datasets for representation learning

Type	DataSet	$N_{protein}$	$N_{positive}$	$N_{negative}$	Ratio
DNA-binding interface	DNA573 (Training)	573	14,479	145,404	0.100
	DNA129 (Test)	129	2,240	35,275	0.064
	DNA181 (Test)	181	3,208	72,050	0.044
RNA-binding interface	RNA495 (Training)	495	14,609	122,290	0.119
	RNA117 (Test)	117	2,031	35,314	0.058
	RNA285 (Test)	285	4,245	41,399	0.102
ATP-binding interface	ATP388 (Training)	388	5,657	142,086	0.039
	ATP41 (Test)	41	674	14,159	0.047
	ATP549 (Training)	549	8,502	205,035	0.041
	ATP202 (Test)	202	3,063	78,667	0.039
DNA-binding protein	UniSwiss-Tr (Training)	9,000	4,500	4,500	0.500
	UniSwiss-Tst (Test)	762	381	381	0.500
	Tr22063 (Training)	22,063	7,927	14,136	0.560
	TEST474 (Test)	474	183	291	0.628

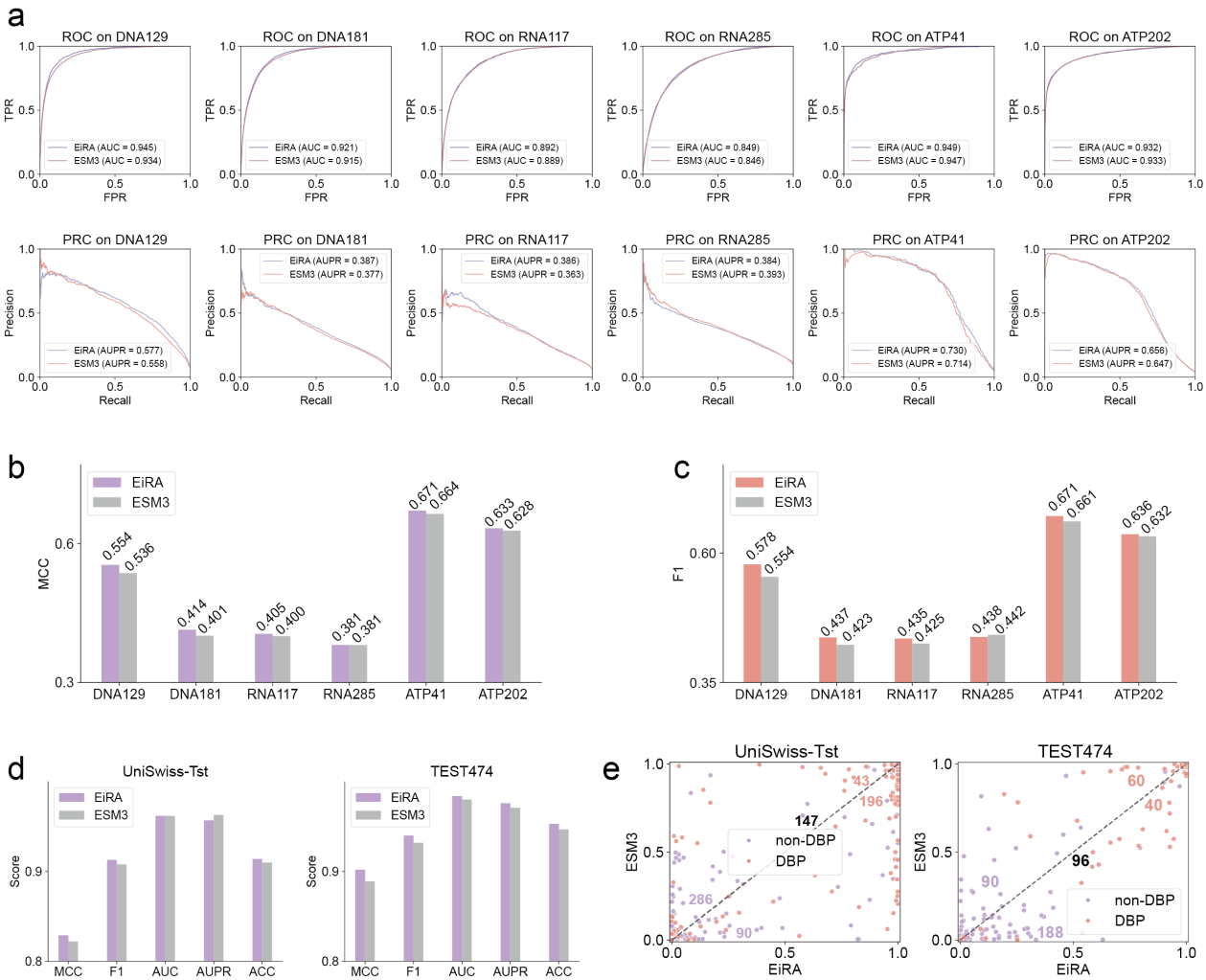

Fig.5 | Representation learning evaluation on four biomolecules-binding-related downstream tasks. a-c. Evaluation on 6 test sets for DNA-binding, RNA-binding, and ATP-binding interface predictions. **d.** Evaluation on 2 DNA-binding protein prediction test sets, i.e., UniSwiss-Tst and TEST474. **e.** Head-to-head comparison of the prediction probabilities between EiRA and ESM3 embedding on 2 DBP prediction test sets. Each point represents a protein, and the number indicates the number of DBP or non-DBP located in the diagonal, upper, or lower triangle.

Ablation study on fine-tuning strategy and parameter

ESM3 is trained on multiple databases, including UniRef [41], Mgnify [42], JGI [43], OAS [44], and PDB [45], with approximately 3,143 million unique samples. Since our training set UniBind40 is only a specific subset of the natural protein space, directly fine-tuning all parameters may compromise the fundamental biochemical and protein knowledge held by the ESM3. Here, we observe the performance of the generation task with different levels of fine-tuning parameter counts (see Table S1).

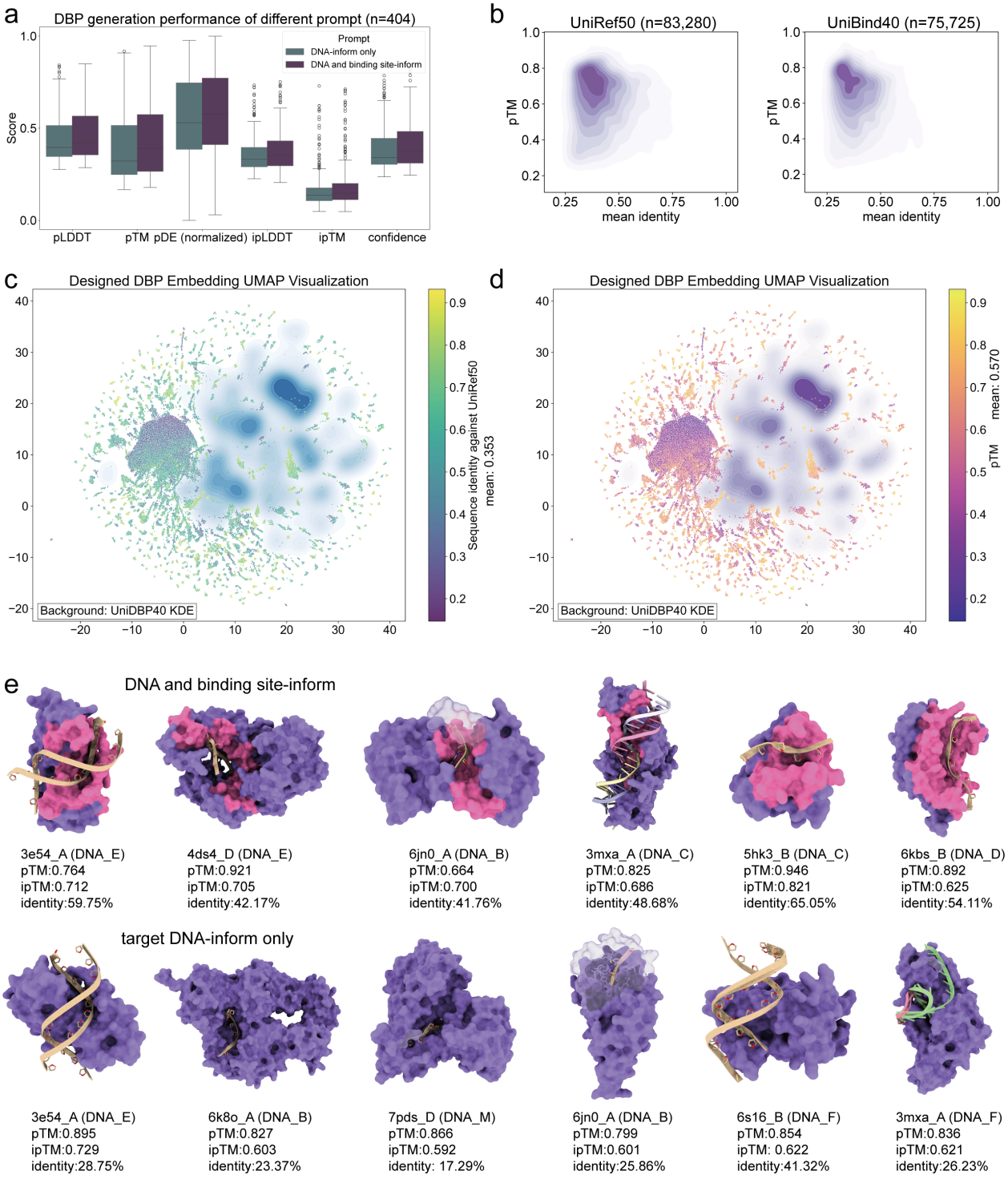

Fig.6 | DNA-conditioned binder generation. a. Comparison of the generation performance on the DBP test set of different prompts, i.e., DNA and binding site-inform and DNA-inform only. **b.** The kernel density maps between the sequence identity and pTM. For the 101,950 generated sequences, MMseqs2 was used to search the UniRef50 and UniBind40 databases, and similar sequences were found for 83,280 and 75,725 samples, respectively. For each sample, the average sequence identity was taken. **c-d.** UMAP visualization of the embedding of 101,950 generated DBP sequences and UniDBP40 background (170,264 native non-redundant DBPs). **e.** Examples of DBPs generated by different prompts. Prompted residues are highlighted in pink. Complex structures were predicted using boltz. Sequence identity was calculated against the natural sequence of PDB templates.

Most intuitively, the improved structural confidence compared to ESM3 after fine-tuning demonstrates the advantages of our domain-adaptive training strategy. Furthermore, fine-tuning 16 (or 32) Transformer blocks achieves leading structural confidence. Considering that the latter requires more training and inference resources, the former is easier to scale, optimize, and apply in practice.

In addition, we performed different fine-tuning schemes and observed the generated performance (see Table S2). It is worth noting that directly updating the original ESM3 weights resulted in mediocre performance. Among the many LoRA techniques (vanilla LoRA [33],

LoRA+ [46], LoKr [47], hydraLoRA [48], and AdaLoRA [49]), vanilla LoRA demonstrates higher structural confidence than other variants. A potential reason may be that these variants are task-specific and lack generalizability.

Representation learning

PLMs encode evolutionary information and biochemical features within protein families, and their embedding representations support diverse downstream tasks, such as secondary structure prediction, tertiary structure prediction, and mutation effect prediction. ESM3 demonstrated that high mask rates improve generation performance but impair representation learning. In the EiRA training strategy, we therefore increased the mask rate to optimize the generative task. Here, we evaluate EiRA's representation learning capability specifically for biomolecular binding prediction tasks (see Fig. 5). We extracted a feature matrix of size $L \times 1536$ from the final transformer block of both ESM3 and EiRA. The normalized matrix was then applied to four downstream tasks: DNA-binding interface (DBI), RNA-binding interface (RBI), ATP-binding interface (ABI), and DNA-binding protein (DBP) prediction. Details of the training and test sets are provided in Table 3. To accommodate different scenarios: for DBI and RBI prediction, we employed an EGNN model based on tertiary structure; for ABI prediction, a lightweight sequence-based BiLSTM model was used; and for the DBP task, we first averaged the feature matrix along the first dimension (sequence length) before feeding it into a simple multilayer perceptron (MLP) model. EiRA consistently outperformed ESM3 across most metrics and datasets. For instance, EiRA achieved AUPR values of 0.577 (DNA129), 0.387 (DNA181), 0.386 (RNA117), 0.730 (ATP41), and 0.656 (ATP202), representing improvements of 3.40, 2.65, 6.33, 2.24, and 1.39%, respectively, over ESM3 (Fig. 5a). Furthermore, for DBP prediction on UniSwiss-Tst, EiRA provided higher prediction probabilities for 196 DBPs and lower probabilities for 286 non-DBPs (Fig. 5e).

Furthermore, we present three representative cases (Fig. S9): the HSDR subunit of the ECOR124I restriction enzyme (ATP-binding, PDB ID: 4XJX_B), the DNA repair protein RAD4 (DNA-binding, PDB ID: 7m2u_A), and Mitochondrial ribosomal protein L45 (12s rRNA-binding, PDB ID: 6ydw_Bi). EiRA achieved MCC values of 0.529, 0.592, and 0.471 for these samples, respectively, exceeding those of ESM3 (0.280, 0.202, and 0.000). Specifically, ESM3's high number of false negatives (FNs) on 7m2u_A indicates frequent misclassification of non-DNA-binding sites (non-DBS) as DNA-binding sites (DBS), a failure largely avoided by EiRA. For 6ydw_Bi, a component of the large mitochondrial ribosomal complex, ESM3 failed to identify any RNA-binding residues. Conversely, EiRA successfully predicted multiple native binding sites; moreover, its predicted FN residues clustered near the target rRNA, indicating its distinct advantages indicating the distinctive advantage.

Collectively, the above experiments demonstrate that EiRA, enhanced by training on numerous molecular binding proteins, captures meaningful biological patterns that safeguard representation learning capability while simultaneously enhancing generative task performance.

DNA sequence-guide binder generation

Target-based binder generation represents a core challenge in protein design [2-4]. While ESM3 leverages extensive protein data, it lacks information on other biomolecules. Here, using DNA as an example, we integrate DNA embedding knowledge from Evo2 [50] (a state-of-the-art DNA language model) into ESM3 as a generative condition via cross-attention (Fig. 1c). Specifically, Evo2 embeddings are first processed through a 2-layer transformer block. The outputs of EiRA's final four transformer blocks are then residually fused with the DNA transformer via gated cross-attention, and the combined representation is used for masked pre-training. Three distinct masking strategies were applied: (1) Masking residues at untied localization sites (40% of cases); (2) EiRA's original pre-training masking strategy (30% of cases); (3) Full masking of protein residues with DNA input retained for DNA-only generation (30% of cases). We fine-tuned the final 32 transformer blocks of EiRA using all 1,214 DNA-binding proteins (DBPs) from BioDPO. DNA sequences exceeding 50 residues (the average training length is 26) were truncated; shorter sequences were zero-padded. Protein processing mirrored EiRA's original protocol. Training spanned 25 epochs on 8 H100 GPUs with a batch size of 10.

Using binding sites and DNA chains from BioDPO complexes as prompts, we generated 101,950 sequences (50 per prompt). MMseqs2 searches against UniRef50 and UniBind40 identified similar sequences for 83,280 and 75,725 samples, respectively. Kernel density plots of mean identity versus pTM reveal a pronounced cluster in the upper-left quadrant, indicating designed sequences achieve both novelty and high predicted structural confidence (Fig. 6b). UMAP visualization further compares generated sequences against EiRA embeddings of UniDBP40 (from ESM-DBP [34]. Clustering all UniProt DBPs at 40% similarity (CD-HIT) yielded 170,264 representatives native DBPs.), assessed by sequence identity (Fig. 6c) and pTM (Fig. 6d). The designed DBPs show minimal overlap with nature DBPs, confirming novelty. Most designs exhibit low sequence similarity while retaining good pTM scores. Notably, a cluster with low pTM and low sequence similarity arises due to structure prediction's reliance on homology, assessing the foldability of such sequences remains challenging.

EiRA fully retains ESM3's capacity for flexible multi-modal input, combining sequence and structural information at any residue. This enables protein generation conditioned solely on DNA by masking protein sequence/structure. Using 404 sample pairs (DBP275 protein sequences and interacting DNA chains from PDB complexes), we performed generation tasks conditioned on: 1) DNA + binding site, and 2) DNA alone. Structural confidence (pTM) decreased only marginally when only DNA was provided, despite the absence of protein information (Fig. 6a). Representative structures for different prompts are shown (Fig. 6e: DNA+binding site above; DNA-only below). Given the binding sites, the target DNA is embedded within its groove (pink highlight). Given only the target DNA, the model generates proteins with pockets accommodating the DNA. Moreover, DNA-only designs exhibit significantly greater sequence divergence from natural binders compared to DNA+binding site designs, providing a powerful method for expanding novel DBP libraries.

Discussion

In this study, we introduce EiRA, a multimodal protein language model specifically designed for biomolecular-binding protein generation. Built upon the general foundation model ESM3, EiRA undergoes two rounds of post-training: domain-adaptive masking training and binding site-informed preference optimization. The main contributions of this work are summarized as follows:

1. We curated UniBind40, a large-scale dataset of biomolecular-binding proteins, to support customized masking training;
2. Comprehensive multi-dimensional evaluations demonstrate that EiRA outperforms the native ESM3, including its larger-scale variants ESM3m and ESM3l, across multiple ligand-binding protein design tasks;
3. The optimized loss function effectively alleviates ESM3's propensity for repetitive generation, leading to improved design quality;
4. EiRA's representations facilitate direct predictions for diverse biomolecular binding-related downstream tasks and offer superior protein characterization compared to the native ESM3;
5. By integrating DNA information, EiRA supports the design of

DNA-binding proteins under DNA conditions, thereby broadening the design paradigm.

We summarized 4 challenges for AI protein design, hoping to provide potential insights for future studies.

1. Developing screening strategies integrated with biological context and microenvironment to bridge the gap between design and practical application.
2. Incorporating dynamic and state information of proteins and complexes to achieve more precise and controllable design outcomes.
3. Establishing alternative screening methods for long and orphan proteins, where conventional structure prediction remains unreliable.
4. Balancing the inherent trade-off between diversity and reliability in generative protein design.

Although there is still room for improvement, the proposed EiRA has the potential to become a powerful protein design tool, thereby accelerating the development of biomedicine, such as gene editing, immunotherapy and drug discovery.

Data Availability

The data, source code, and standalone program of EiRA are freely accessible at https://github.com/wwzll123/EiRA . The pretraining data set and the weight of all EiRA models are available for academic use at https://huggingface.co/zengwenwu/EiRA/tree/main.

Author Contributions

Wenwu Zeng: Conceptualization, Methodology, Software, Visualization, Writing original draft. Haitao Zou: Methodology, Software. Xiaoyu Li: Visualization. Xiaoqi Wang: Conceptualization, Writing-review & editing. Shaoliang Peng: Funding acquisition, Resources, Writing-review & editing, Supervision.

Acknowledgment

This work was supported by National Key R&D Program of China 2022YFC3400400; NSFC Grants U19A2067; Key R&D Program of Hunan Province 2023GK2004, 2023SK2059, 2023SK2060; Top 10 Technical Key Project in Hunan Province 2023GK1010; Key Technologies R&D Program of Guangdong Province (2023B1111030004 to FFH). We would like to thank the Fund of the National Supercomputing Center in Changsha (http://nscc.hnu.edu.cn/). Thanks to EvolutionaryScale for developing the ESM3 foundational language model and ESMFold. Thanks to DeepMind for developing the AlphaFold3 and the MIT Jameel Clinic for reproducing the Boltz version. X.Q.W and S.L.P are the corresponding authors of this paper.

Competing interests

The authors declare no competing interests.

References

1. Consortium, U. UniProt: a hub for protein information. *Nucleic acids research* **43**, D204-D212 (2015).
2. Chronowska, M., Stam, M.J., Woolfson, D.N., Di Costanzo, L.F. & Wood, C.W. The Protein Design Archive (PDA): insights from 40 years of protein design. *Nature Biotechnology* **43**, 669-671 (2025).
3. Notin, P., Rollins, N., Gal, Y., Sander, C. & Marks, D. Machine learning for functional protein design. *Nature biotechnology* **42**, 216-228 (2024).
4. Kortemme, T. De novo protein design—From new structures to programmable functions. *Cell* **187**, 526-544 (2024).
5. Watson, J.L. et al. De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089-1100 (2023).
6. Ingraham, J.B. et al. Illuminating protein space with a programmable generative model. *Nature* **623**, 1070-1078 (2023).
7. Liu, Y. et al. De novo protein design with a denoising diffusion network independent of pretrained structure prediction models. *Nature Methods* **21**, 2107-2116 (2024).
8. Geffner, T. et al. Proteina: Scaling Flow-based Protein Structure Generative Models. *arXiv preprint arXiv:2503.00710* (2025).
9. Dauparas, J. et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49-56 (2022).
10. Ren, M., Yu, C., Bu, D. & Zhang, H. Accurate and robust protein sequence design with CarbonDesign. *Nature Machine Intelligence* **6**, 536-547 (2024).
11. Bai, P. et al. Mask-prior-guided denoising diffusion improves inverse protein folding. *Nature Machine Intelligence* **7**, 876-888 (2025).
12. Hsu, C. et al. Learning inverse folding from millions of predicted structures. *In International conference on machine learning*, 8946-8970 (PMLR, 2022).
13. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024).
14. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *nature* **596**, 583-589 (2021).
15. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876 (2021).
16. Shi, C., Wang, C., Lu, J., Zhong, B. & Tang, J. Protein sequence and structure co-design with equivariant translation. *arXiv preprint arXiv:2210.08761* (2022).
17. Wang, C., Alamdari, S., Domingo-Enrich, C., Amini, A.P. & Yang, K.K. Toward deep learning sequence–structure co-generation for protein design. *Current Opinion in Structural Biology* **91**, 103018 (2025).
18. Lisanza, S.L. et al. Joint Generation of Protein Sequence and Structure with RoseTTAFold Sequence Space Diffusion. *bioRxiv*, 2023.2005.2008.539766 (2023).
19. Liu, S. et al. A text-guided protein design framework. *Nature Machine Intelligence* **7**, 580-591 (2025).
20. Madani, A. et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology* **41**, 1099-1106 (2023).
21. Hayes, T. et al. Simulating 500 million years of evolution with a language model. *Science*, eads0018 (2025).
22. Blum, M. et al. InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research* **53**, D444-D456 (2024).
23. Liu, B. et al. Design of high-specificity binders for peptide–MHC-I complexes. *Science* **389**, 386-391 (2025).
24. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence

searching for the analysis of massive data sets. *Nature Biotechnology* **35**, 1026-1028 (2017).

25. Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439-D444 (2021).
26. Zhang, C., Zhang, X., Freddolino, P.L. & Zhang, Y. BioLiP2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic acids research* **52**, D404-D412 (2024).
27. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).
28. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research* **31**, 365-370 (2003).
29. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules* **22**, 2577-2637 (1983).
30. Kunzmann, P. & Hamacher, K. Biotite: a unifying open source computational biology framework in Python. *BMC bioinformatics* **19**, 1-8 (2018).
31. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
32. Wohlwend, J. et al. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, 2024.2011.2019.624167 (2024).
33. Hu, E.J. et al. Lora: Low-rank adaptation of large language models. *ICLR* **1**, 3 (2022).
34. Zeng, W., Dou, Y., Pan, L., Xu, L. & Peng, S. Improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein. *Nature Communications* **15**, 7838 (2024).
35. Bai, Y. et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).
36. Guo, D. et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).
37. Liu, X., Liu, Y., Chen, S. & Hu, W. Controllable Protein Sequence Generation with LLM Preference Optimization. *Proceedings of the AAAI Conference on Artificial Intelligence* **39**, 505-513 (2025).
38. Pang, R.Y. et al. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems* **37**, 116617-116637 (2024).
39. Rafailov, R. et al. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* **36**, 53728-53741 (2023).
40. Bahri, Y., Dyer, E., Kaplan, J., Lee, J. & Sharma, U. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences* **121**, e2311878121 (2024).
41. Suzek, B.E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926-932 (2015).
42. Richardson, L. et al. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic acids research* **51**, D753-D759 (2023).
43. Nordberg, H. et al. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic acids research* **42**, D26-D31 (2014).
44. Olsen, T.H., Boyles, F. & Deane, C.M. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science* **31**, 141-146 (2022).
45. Sussman, J.L. et al. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Biological Crystallography* **54**, 1078-1084 (1998).
46. Hayou, S., Ghosh, N. & Yu, B. Lora+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354* (2024).
47. Yeh, S.-Y. et al. Navigating text-to-image customization: From lycoris fine-tuning to model evaluation. *In The Twelfth International Conference on Learning Representations* (2023).
48. Tian, C., Shi, Z., Guo, Z., Li, L. & Xu, C.-Z. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *Advances in Neural Information Processing Systems* **37**, 9565-9584 (2024).
49. Zhang, Q. et al. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512* (2023).
50. Brixi, G. et al. Genome modeling and design across all domains of life with Evo 2. *BioRxiv*, 2025.2002. 2018.638918 (2025).

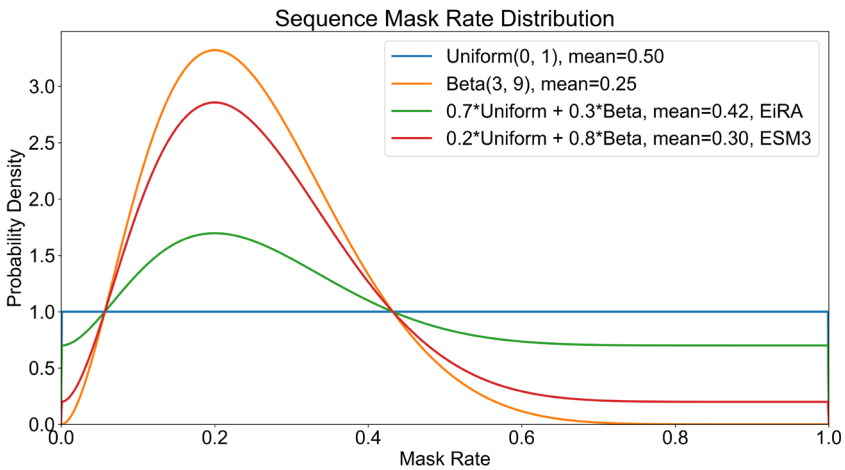

Fig. S1 | EiRA adopts a higher sequence mask rate than that of ESM3.

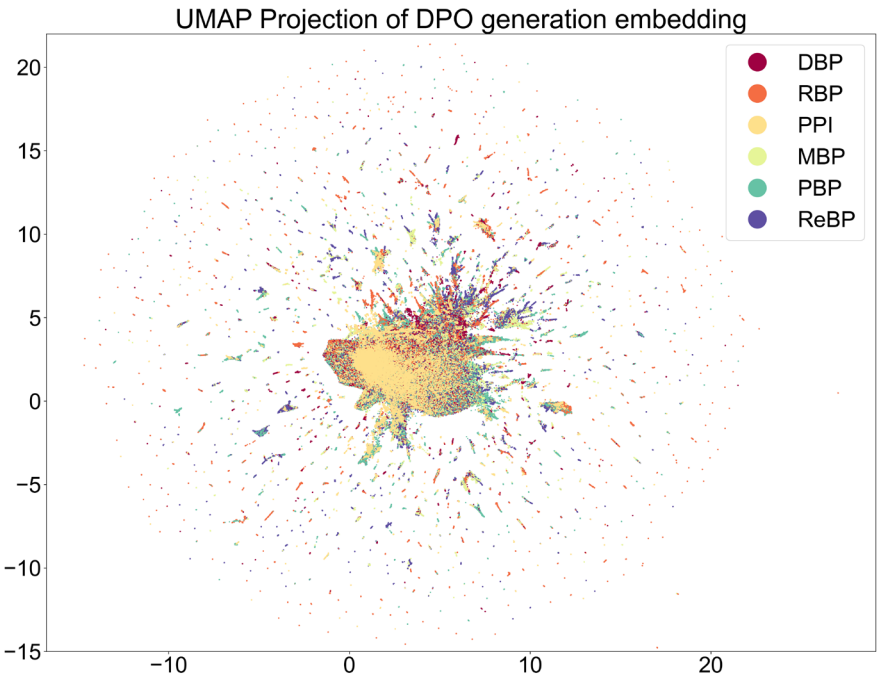

Fig. S2 | UMAP visualization of the DPO training set colored by ligand type.

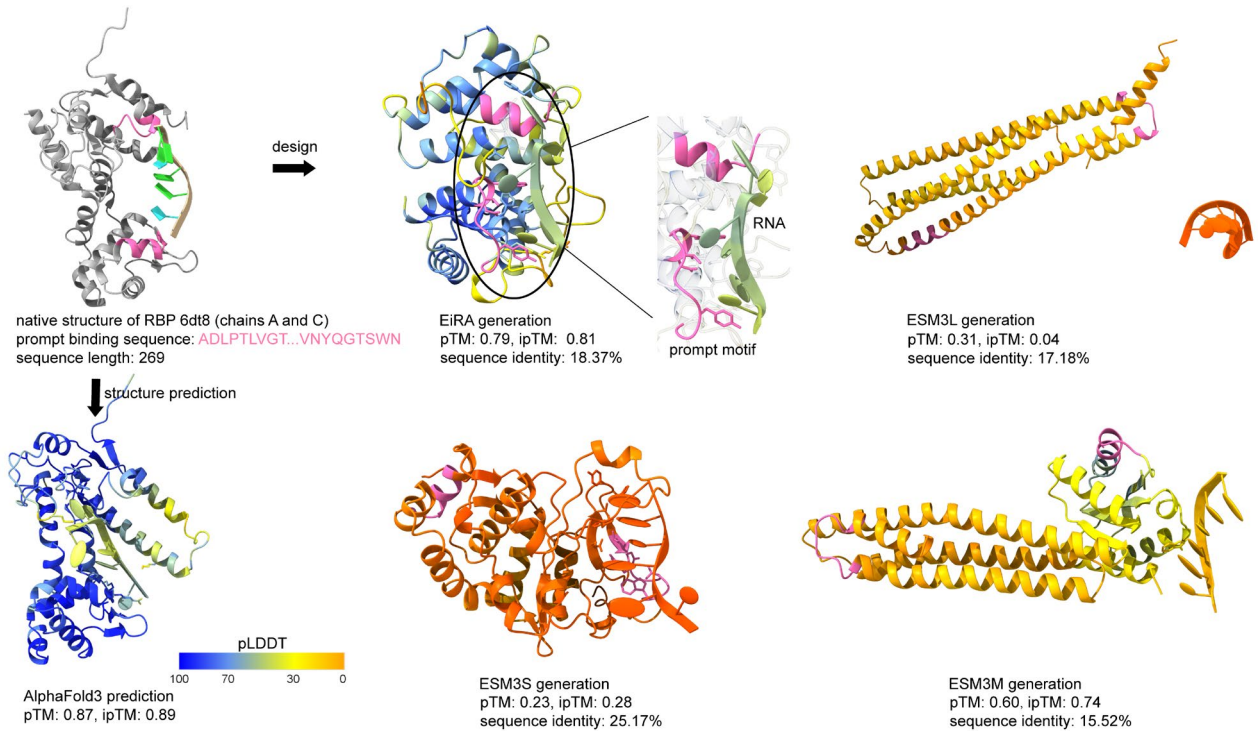

Fig. S3 | Case study on a RBP (PDB ID: 6dt8, Bacteriophage N4 RNA polymerase II elongation complex 1). The new protein (a sequence identity of 18.37% against PDB template) designed by EiRA maintains high stability while retaining the ability of the template motif to contact RNA ligands. In contrast, all three versions of ESM3 produced low-confidence proteins. The protein sequences were designed by four MPLMs from the template motif (highlight with pink) and then the complex structures were predicted by Boltz.

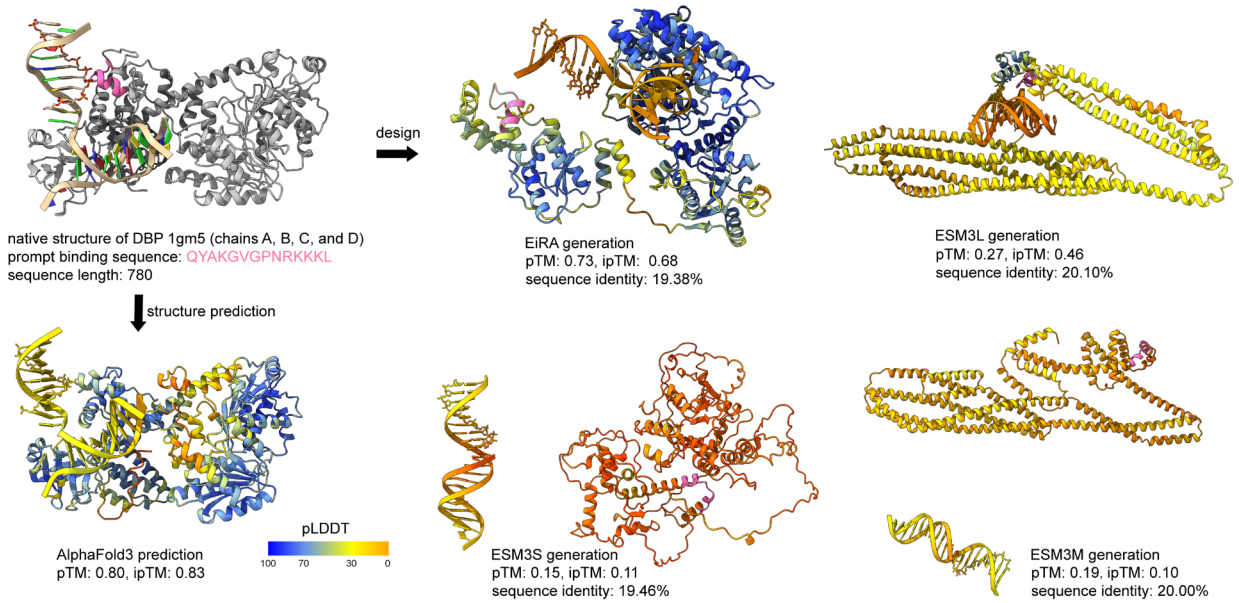

Fig. S4 | Case study on a DBP (PDB ID: 1gm5, RecG protein, DNA-specific helicase). The novel proteins generated by EiRA have high confidence and are predicted to interact with DNA, while the confidence of the other three is low, and the ones generated by ESM3s and ESM3m are predicted not to bind DNA.

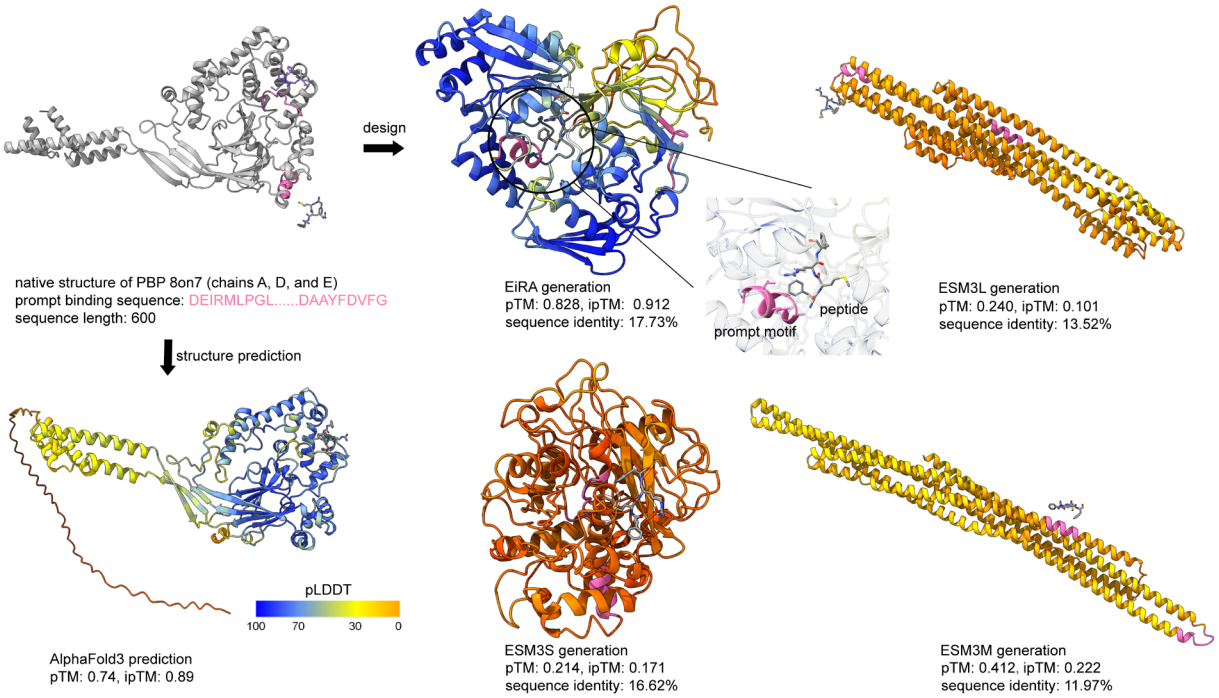

Fig. S5 | Case study on a PBP (PDB ID: 8on7, Sort-Tagged Drosophila Cryptochrome). The pTM and ipTM scores of EiRA designed protein are higher than that of the AF3 structure of native sequence.

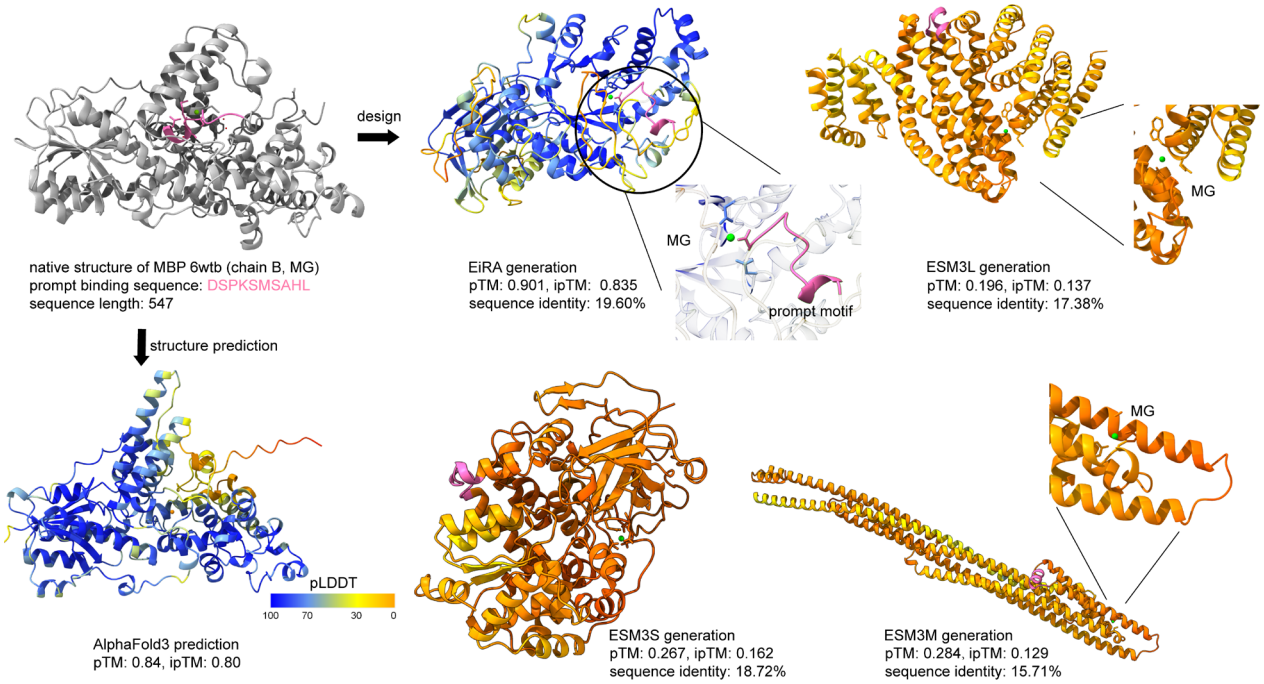

Fig. S6 | Case study on a MBP (PDB ID: 6wtb, FMRFa-bound Malacoceros FaNaC1).

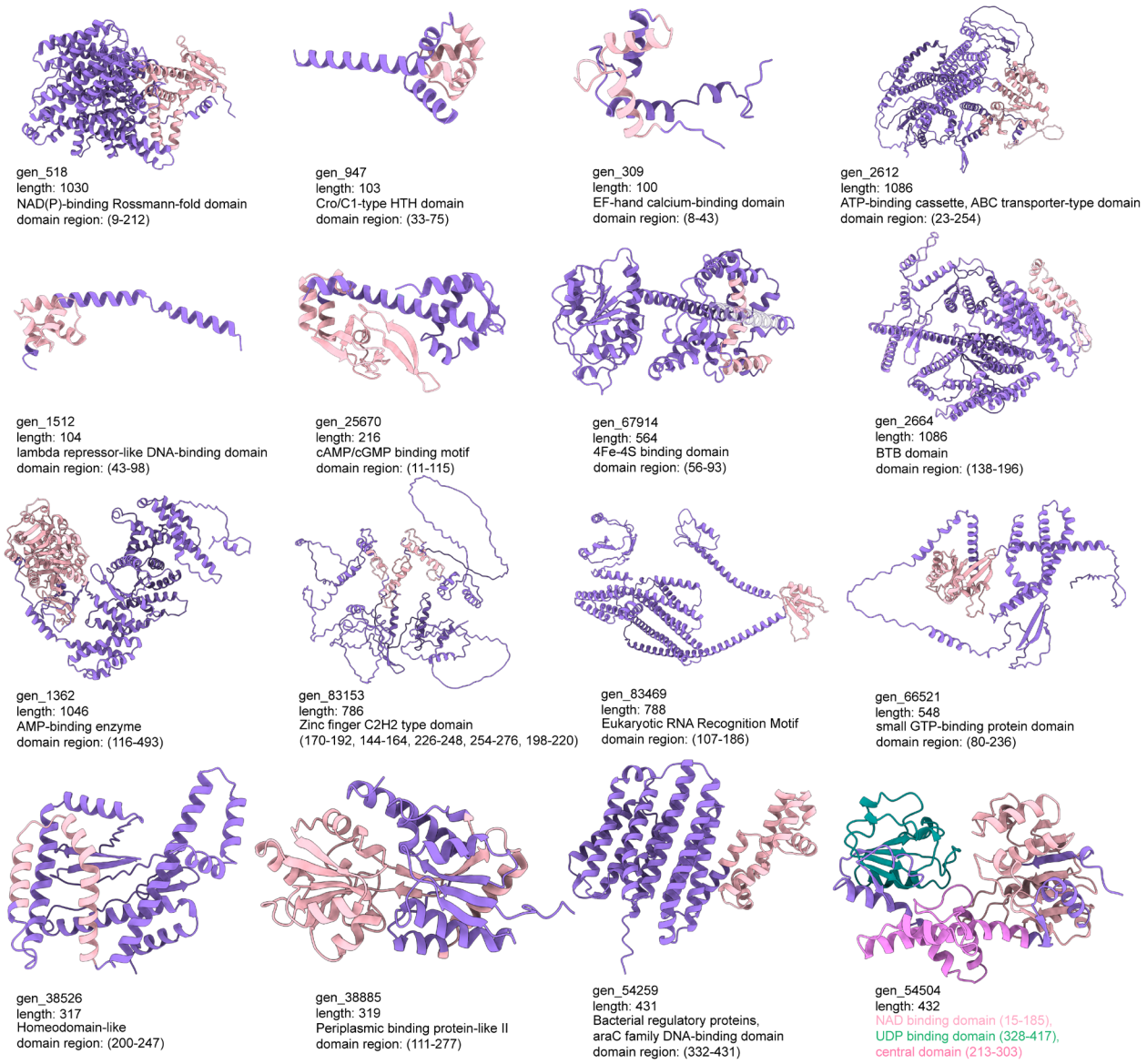

Fig. S7 | EiRA generates various biomolecule binding domains unconditionally. The samples are selected from 100,000 unconditionally generated sequences, and AF3 is used to predict the tertiary structure. The domain annotations are predicted using InterProScan tool.

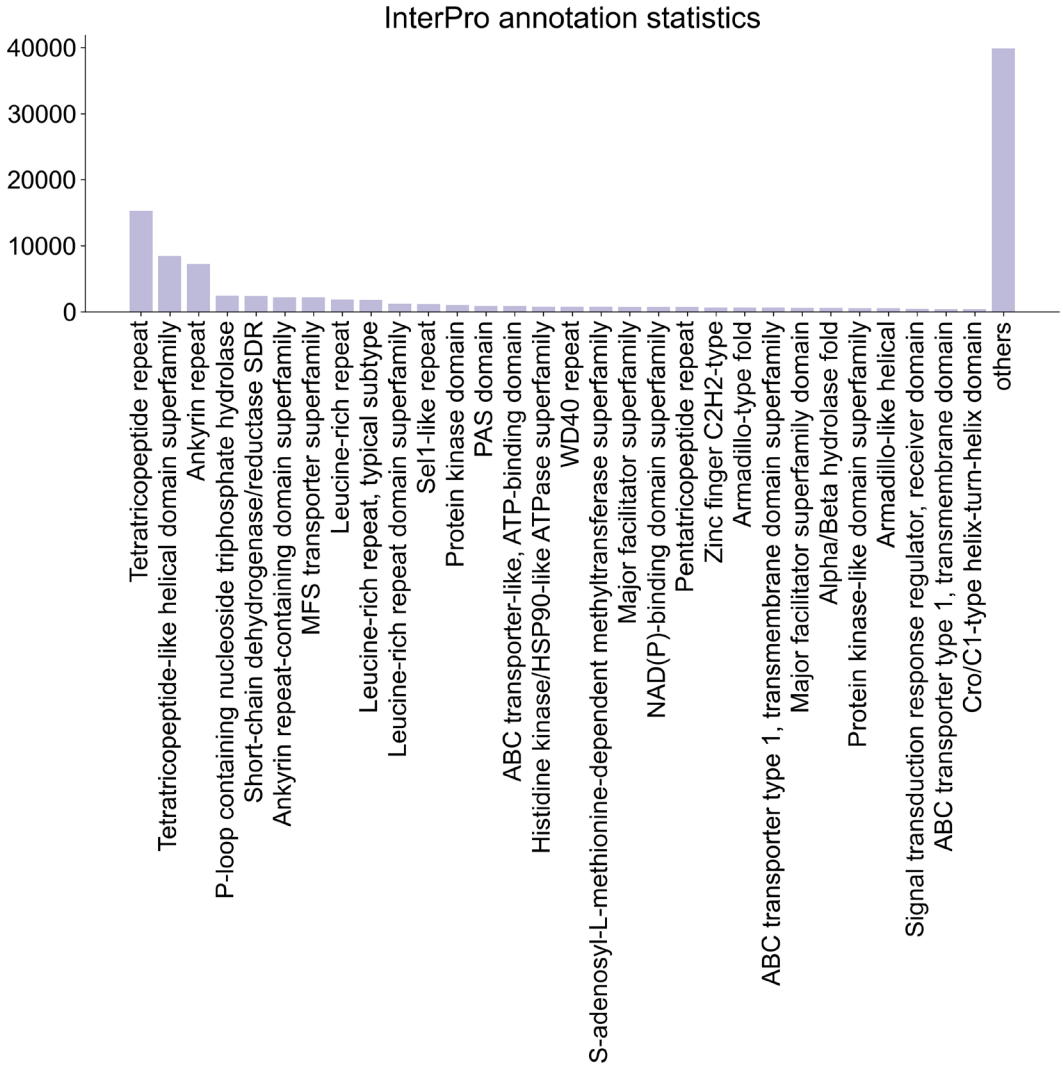

Fig. S8 | InterProScan records distribution of 100,000 unconditional generated sequences (no redundancy removed).

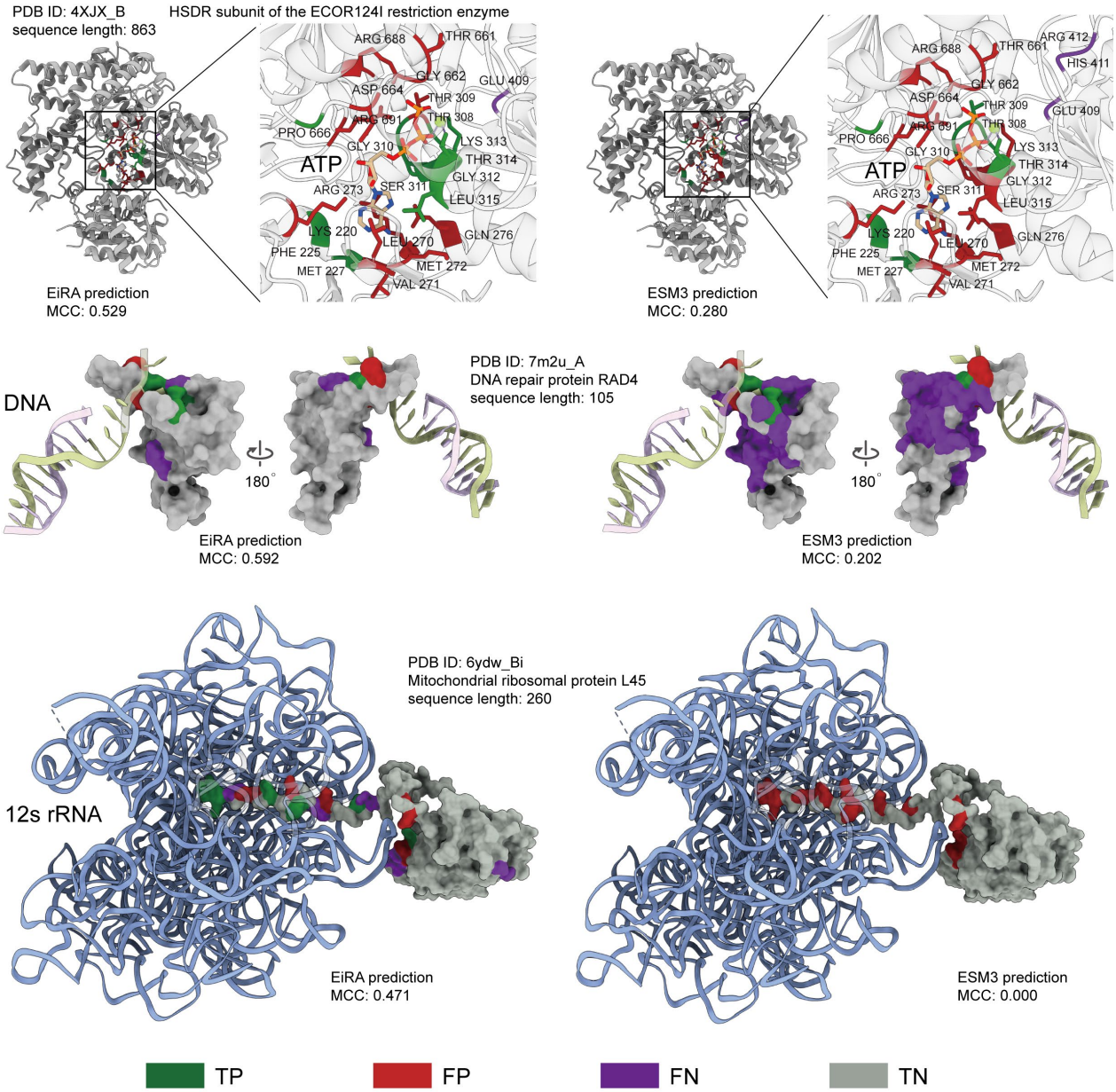

Fig. S9 | Case studies of DNA-, RNA-, and ATP-binding interface predictions.

Table S1 | The impact of updatable Transformer blocks on performance

Structure predictor	Evaluation index	Fine-tuning layers	DBP	RBP	PBP	ReBP	MBP	PPI	MDBP	MRBP
ESM3	pTM	0 (ESM3)	0.503	0.502	0.540	0.630	0.530	0.439	**0.442**	0.516
		16	0.502	0.496	0.559	0.656	0.578	0.445	0.406	0.498
		24	0.453	0.487	0.519	0.600	0.511	0.415	0.383	0.472
		32	**0.506**	**0.534**	0.563	**0.657**	**0.580**	**0.462**	0.431	**0.532**
		48	0.485	0.529	**0.566**	0.654	0.572	0.461	0.430	0.517
	pLDDT	0 (ESM3)	0.700	0.691	0.708	0.742	0.680	0.681	0.679	0.727
		16	0.742	0.722	**0.780**	**0.802**	**0.775**	0.732	0.722	0.746
		24	0.680	0.683	0.719	0.732	0.694	0.719	0.709	0.726
		32	**0.745**	0.727	0.764	0.787	0.751	0.733	0.729	0.752
		48	0.744	**0.738**	0.772	0.794	**0.775**	**0.769**	**0.759**	**0.775**
ESMFold	pTM	0 (ESM3)	0.418	0.429	0.396	0.548	0.444	0.363	0.374	0.317
		16	**0.443**	**0.434**	**0.449**	**0.603**	**0.528**	0.397	**0.405**	0.473
		24	0.387	0.405	0.377	0.509	0.410	0.345	0.343	0.430
		32	0.416	0.426	0.419	0.570	0.477	**0.398**	0.396	**0.480**
		48	0.380	0.392	0.425	0.564	0.479	0.389	0.388	0.461
	pLDDT	0 (ESM3)	0.441	0.437	0.452	0.535	0.453	0.469	0.524	0.448
		16	**0.504**	0.492	**0.521**	**0.601**	**0.557**	0.563	0.616	0.626
		24	0.440	0.449	0.443	0.512	0.443	0.462	0.485	0.552
		32	0.502	**0.494**	0.496	0.584	0.524	**0.599**	**0.630**	**0.655**
		48	0.485	0.484	0.512	0.583	0.531	0.584	0.605	0.617
	pAE	0 (ESM3)	19.83	19.54	19.121	15.92	18.94	23.80	23.47	26.47
		16	**19.36**	19.57	**17.98**	**14.14**	**16.45**	**22.87**	22.83	21.71
		24	20.80	20.31	19.88	17.03	19.79	24.38	24.91	23.13
		32	19.97	**19.50**	18.61	15.01	17.72	22.94	**22.81**	**21.36**
		48	20.60	20.17	18.46	15.13	17.77	22.94	22.97	21.75

Table S2 | The impact of fine-tuning strategy on performance

Structure predictor	Evaluation index	Fine-tuning strategy	DBP	RBP	PBP	ReBP	MBP	PPI	MDBP	MRBP
ESM3	pTM	base [a]	0.463	0.484	0.536	0.602	0.529	0.432	0.370	0.415
		vanilla LoRA [b]	**0.502**	0.496	**0.559**	**0.656**	**0.578**	0.445	0.406	0.498
		LoRA+ [c]	0.471	**0.501**	0.557	0.641	0.569	0.446	0.411	**0.512**
		LoKr [d]	0.421	0.459	0.504	0.596	0.524	0.394	0.365	0.427
		hydraLoRA [e]	0.475	0.493	0.551	0.623	0.549	**0.453**	**0.427**	0.504
		AdaLoRA [f]	0.449	0.477	0.518	0.612	0.523	0.406	0.373	0.430
	pLDDT	base	0.667	0.663	0.684	0.707	0.667	0.707	0.678	0.691
		vanilla LoRA	**0.742**	**0.722**	**0.780**	**0.802**	**0.775**	0.732	0.722	0.746
		LoRA+	0.723	0.715	0.755	0.779	0.752	0.741	**0.733**	**0.754**
		LoKr	0.681	0.681	0.730	0.755	0.725	0.668	0.652	0.683
		hydraLoRA	0.688	0.677	0.754	0.767	0.748	**0.752**	0.673	0.715
		AdaLoRA	0.688	0.682	0.725	0.755	0.700	0.671	0.636	0.656
ESMFold	pTM	base	0.359	0.354	0.349	0.468	0.381	0.336	0.305	0.344
		vanilla LoRA	**0.443**	**0.434**	**0.449**	**0.603**	**0.528**	**0.397**	**0.405**	0.473
		LoRA+	0.418	0.417	0.419	0.565	0.491	0.385	0.400	**0.475**
		LoKr	0.394	0.408	0.388	0.540	0.467	0.357	0.363	0.428
		hydraLoRA	0.381	0.389	0.400	0.552	0.451	0.369	0.359	0.459
		AdaLoRA	0.416	0.428	0.393	0.543	0.437	0.356	0.339	0.399
	pLDDT	base	0.426	0.413	0.432	0.480	0.423	0.487	0.494	0.495
		vanilla LoRA	**0.504**	**0.492**	**0.521**	**0.601**	**0.557**	**0.563**	**0.616**	**0.626**
		LoRA+	0.484	0.475	0.496	0.567	0.524	0.537	0.599	0.619
		LoKr	0.444	0.445	0.459	0.548	0.499	0.516	0.558	0.591
		hydraLoRA	0.424	0.420	0.474	0.550	0.488	0.519	0.510	0.586
		AdaLoRA	0.456	0.460	0.471	0.550	0.470	0.500	0.507	0.540
	pAE	base	21.27	21.42	20.55	18.21	20.52	24.52	25.50	25.51
		vanilla LoRA	**19.36**	**19.57**	**17.98**	**14.14**	**16.45**	**22.87**	**22.83**	21.71
		LoRA+	19.96	20.03	18.79	15.29	17.56	23.22	22.89	**21.63**
		LoKr	20.66	20.34	19.58	16.13	18.22	24.44	24.35	23.28
		hydraLoRA	20.64	20.48	19.14	15.74	18.56	23.52	24.21	22.03
		AdaLoRA	19.98	19.67	19.40	16.02	19.04	24.36	25.02	24.21

a. Update the original weights directly, rather than applying adapters and introducing additional parameters.
b. Low-rank matrix adaptation constructed by multiplying two miniature matrices.
c. LoRA+ sets different learning rates for the matrices A and B of LoRA.
d. LoKr employs Kronecker product decomposition.
e. Asymmetric hybrid expert LoRA, containing a shared matrix A, multiple matrices B, and a router.
f. AdaLoRA dynamically adjusts the rank of low-rank matrices based on parameter importance scores, optimizing parameter allocation.