# ESFOLD: ADAPTING BIOMOLECULAR STRUCTURE MODELS WITH EXPERIMENTAL SIGNALS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Protein–RNA interactions are fundamental to gene regulation and RNA function, yet their structural modeling is limited by scarce experimentally resolved complexes. In contrast, enhanced crosslinking and immunoprecipitation (eCLIP) assays provide abundant interaction evidence but lack atomic resolution. Here we introduce ESFold, a two-stage framework that uses eCLIP profiles to post-train Protenix for protein–RNA structure prediction. The model does not consume experimental profiles at inference. In the first stage, supervised post-training combines structural and eCLIP data to learn binding-aware representations. In the second stage, we adapt Diffusion Negative-aware Fine-Tuning (DiffusionNFT) to molecular structure generation, using eCLIP agreement and structural plausibility as feedback to optimize the distribution of generated conformations. We construct 265,070 candidate protein–RNA pairs spanning 53,014 RNA windows and 84 proteins. On the protein–RNA subset of FoldBench, ESFold improves the prediction success rate from 61.43% with Protenix to 74.29% and reduces interface RMSD from 7.61 Å to 5.02 Å. On 190 held-out protein–RNA inputs in experimentally localized RNA windows, peak F1 averaged across RNA-binding proteins (RBPs) improves from 0.332 to 0.459. These gains are achieved while broadly maintaining Protenix's performance on structural prediction tasks beyond protein–RNA complexes. Together, the findings highlight experimental-signal post-training as an approach to integrating functional genomics into three-dimensional biomolecular modeling.

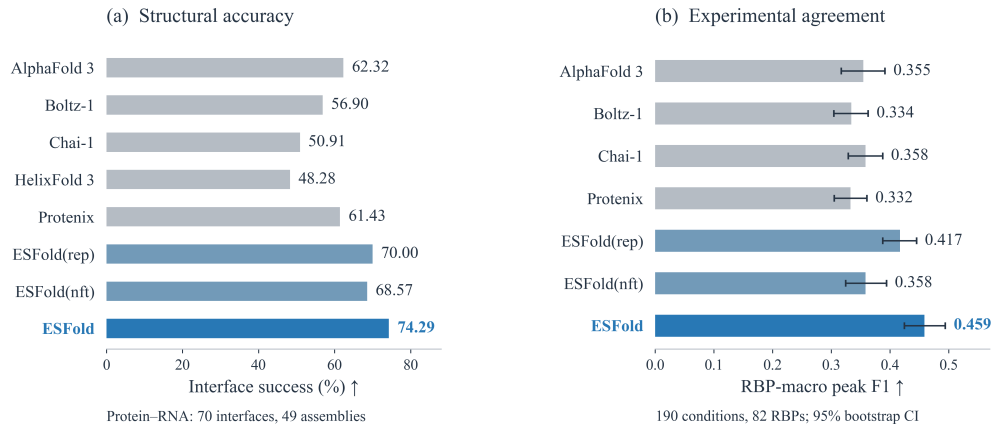


Figure 1: **Structural accuracy and experimental agreement.** (a) Protein–RNA interface success (DockQ $\geq 0.23$). (b) RBP-macro peak F1, with 95% intervals from 2,000 RBP-bootstrap resamples. Gray: external baselines; muted blue: single-stage variants; blue: ESFold. Full results are in Tables 4 and 3.

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

# 1 INTRODUCTION

Protein–RNA interactions regulate RNA processing, translation, and decay, connecting molecular recognition to cellular function (Van Nostrand et al., 2020). RoseTTAFoldNA, AlphaFold 3, and Protenix have advanced joint modeling of proteins and nucleic acids from molecular inputs (Baek et al., 2024; Abramson et al., 2024; Zhang et al., 2026). However, the available coordinate data cover only a limited portion of RNA-mediated interactions. RNA flexibility further complicates the connection between a predicted structure and a functional binding event: measurements of individual RNA molecules reveal distinct conformers that are difficult to capture with a single structural description (Degenhardt et al., 2025). Improving interface prediction therefore calls for supervision that complements resolved structures with evidence of where interactions occur.

Recent work shows how experimental observations can refine learned structural priors. GRASP integrates heterogeneous experimental information to assist protein complex prediction (Xie et al., 2025). Experiment-guided AlphaFold3 generates protein ensembles consistent with experimental measurements (Maddipatla et al., 2026). These advances establish that experimental consistency is a useful modeling objective even when the evidence does not specify every coordinate. They motivate a complementary question: can experimental interaction profiles improve predictions for held-out inputs without providing experimental profiles to the model at inference?

Transcriptome-wide binding assays offer a route to such supervision. eCLIP provides position-aligned evidence of RNA association with specific proteins across many transcripts (Van Nostrand et al., 2016; 2020). RBPNet shows that these profiles contain binding information that can be learned from sequence (Horlacher et al., 2023). Their structural interpretation nevertheless differs from that of assigned distance restraints: an enriched RNA region does not identify its protein-side partner or specify an orientation of the complex. The learning problem is therefore to retain the regional information supplied by the assay while using a structural prior to resolve the remaining geometric degrees of freedom.

A generative model provides candidate protein–RNA conformations that can be scored against regional binding profiles through their implied RNA contacts, even when reference complex coordinates are unavailable. BioEmu shows that experimental protein stabilities can help train a reusable model of protein ensembles (Lewis et al., 2025). Turning this feedback into learning requires both a representation sensitive to binding regions and an update that favors compatible, geometrically plausible conformations.

We introduce ESFold, an experimental-signal post-training framework that first learns binding-aware representations and then optimizes complete conformations. Stage 1 mixes standard structural supervision on PDB examples with alignment of distogram-derived contact probabilities to eCLIP signal profiles, emphasizing experimental peak regions. Stage 2 freezes the adapted conditioning representation and updates the diffusion transformer (DiT) using experimental agreement and geometric plausibility. We adapt the existing DiffusionNFT objective (Zheng et al., 2026) to molecular coordinate space, allowing discrete contact-region rewards to guide denoising without reward gradients. At inference, ESFold requires molecular inputs alone.

Our contribution is this integration of regional supervision and conformation-level feedback, together with an aligned corpus of 265,070 candidate pairs spanning 53,014 RNA windows and 84 proteins. Figure 1 summarizes complementary evaluations of structure and experimental agreement: ESFold improves protein–RNA interface success over Protenix by 12.86 percentage points and RBP-macro peak F1 from 0.332 to 0.459. These gains accompany broadly maintained performance across other complex and monomer tasks. Single-stage variants compare the evaluated training strategies.

# 2 BACKGROUND

## 2.1 CONDITIONAL DIFFUSION FOR MOLECULAR COORDINATES

Diffusion models learn to reconstruct clean data from progressively perturbed inputs (Ho et al., 2020). In biomolecular prediction, the generated variable is an atomic coordinate array $X \in \mathbb{R}^{N\times 3}$, and molecular sequences and associated features define the condition $c$. AlphaFold 3 uses diffusion

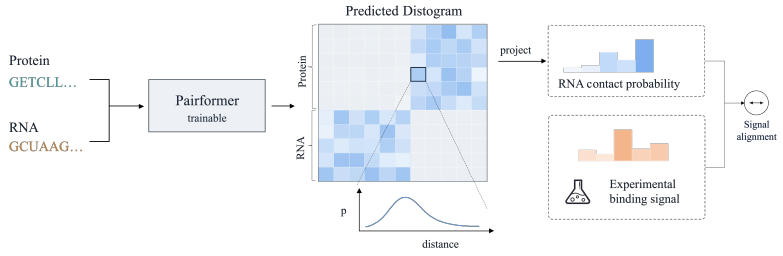


Figure 2: **Stage 1: binding-aware supervised adaptation.** The eCLIP branch projects predicted distograms into RNA contact probabilities and aligns them with experimental binding profiles through a signal-weighted cross-entropy objective. PDB examples are interleaved with eCLIP examples and retain standard structural supervision to stabilize training (PDB branch not shown).

to generate biomolecular coordinates, and Protenix provides the architecture adapted in this work (Abramson et al., 2024; Zhang et al., 2026). The Protenix Pairformer produces single representations for individual tokens and pair representations for relationships between tokens. A DiT uses these representations to condition iterative coordinate denoising.

Using an additive Gaussian noise convention, a clean structure is perturbed as

$$X_\sigma = X + \sigma\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I). \tag{1}$$

A schematic coordinate-denoising objective is

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{(X,c)\sim\mathcal{D}_{\text{coord}},\sigma,\varepsilon}\left[\lambda(\sigma)\,\|D_\theta(X_\sigma,\sigma,c) - X\|_M^2\right], \tag{2}$$

where $\mathcal{D}_{\text{coord}}$ provides coordinate supervision, $M$ masks valid coordinate entries, and $\lambda(\sigma)$ weights the noise levels. In the EDM formulation, noise schedules, output preconditioning, and numerical solvers are separate design choices (Karras et al., 2022). Sampling starts from noise and repeatedly applies the conditional denoiser. Equation 2 describes the underlying regression problem; Section 3 specifies the adaptive error and online targets used for fine-tuning.

Flow matching learns a velocity field along a path between noise and data (Lipman et al., 2023). AlphaFlow repurposes pretrained structure predictors within this framework and learns conformational variation from structural and simulation data (Jing et al., 2024). DiffusionNFT also uses velocity prediction along a Gaussian path; here we retain Protenix's denoised-coordinate output and EDM noise convention while applying the same forward-noising principle.

### 2.2 Reward-guided diffusion optimization

For molecular inputs $c$, an experimental profile $y$, and the model-induced distribution $p_\theta(X \mid c)$, reward-guided adaptation targets

$$J(\theta) = \mathbb{E}_{(c,y)\sim\mathcal{D}_{\text{exp}},\, X\sim p_\theta(\cdot|c)}\left[R(X; y, c)\right]. \tag{3}$$

The experimental profile is used to score generated structures. DiffusionNFT (Zheng et al., 2026) turns these rewards into a denoising objective using forward-noised samples from an old policy. This accommodates nondifferentiable rewards without storing reverse-sampling trajectories. Section 3 defines our coordinate-space adaptation.

# 3 ESFOLD: EXPERIMENTAL-SIGNAL POST-TRAINING

## 3.1 TWO-STAGE PROBLEM SETUP

Let $c$ contain protein and RNA sequences and their associated features, and let $y$ be the aligned eCLIP profile. Starting from Protenix, Stage 1 learns binding-aware representations from structural and experimental supervision. Stage 2 fixes these representations and optimizes the generated conformations $X$ using agreement with $y$ and geometric plausibility.

## 3.2 STAGE 1: LEARNING BINDING-AWARE REPRESENTATIONS

Stage 1 mixes PDB structures with eCLIP sequence–signal examples. PDB examples retain standard structural supervision to stabilize training and preserve the model's structural prior. For eCLIP examples, predicted distograms are converted into contact probabilities and projected onto RNA positions. The resulting profile is aligned with the experimental signal using a signal-weighted cross-entropy objective, so signal-rich positions contribute more strongly to the supervision (Figure 2). This signal-alignment objective trains the model to identify experimentally supported RNA regions while structural supervision anchors the adaptation.

Stage 1 updates the Pairformer and distogram head while freezing the DiT and confidence-head parameters. The resulting representations condition coordinate generation in Stage 2. The eCLIP objective aligns RNA-position profiles without assigning an exact protein-side contact partner. Section 4.1 describes the training data and structural references; Appendix A gives the mixed-training setup.

## 3.3 EXPERIMENTAL AND GEOMETRIC REWARDS

We map each conformation to an RNA contact profile $q(X)$ using representative-atom distances, then align it with the experimental profile $y$. The feedback combines signal agreement with geometric plausibility:

$$\mathbf{r}_k = \big(R^{(k)}_{\text{signal}}, R^{(k)}_{\text{marginal}}, R^{(k)}_{\text{local}}, P^{(k)}_{\text{phys}}\big), \quad \boldsymbol{\omega} = (1.0, 0.30, 0.50, -0.30). \tag{4}$$

Here $R_{\text{signal}}$ measures regional overlap and positional agreement, $R_{\text{marginal}}$ measures each sample's contribution to group-level agreement, $R_{\text{local}}$ preserves chain-local geometry from a predicted distillation reference, and $P_{\text{phys}}$ penalizes clashes and chain discontinuities. The reference is not experimental ground truth. Components are normalized separately before combination with $\boldsymbol{\omega}$ in Equation 5; Appendix A gives their definitions and validity checks.

## 3.4 STAGE 2: OPTIMIZING COMPLETE CONFORMATIONS

Stage 2 adapts DiffusionNFT to molecular coordinates (Figure 3). An old policy samples $K$ conformations per input. Experimental and geometric rewards determine their relative weights, and fresh forward noise creates the denoising inputs for each update.

We maintain a trainable denoiser $D_\theta$, an old sampling policy $D_{\bar{\theta}}$, and a fixed reference $D_{\theta_0}$. All three are initialized from the Stage 1 checkpoint and use the same detached representations from the adapted Pairformer. Sampling and reward computation are performed without gradient tracking. The Pairformer, confidence head, and distogram head remain frozen throughout Stage 2, so only the diffusion module is updated.

**Relative reward weights.** We normalize each reward component independently before combining the components. For component $m$, let $r^{(m)}_{gk}$ denote the reward of sample $k$ in condition group $g$, $\bar{r}^{(m)}_g$ its group mean, and $s_m$ its standard deviation across the current collection of valid samples. The advantage is

$$A_{gk} = \sum_m \omega_m \frac{r^{(m)}_{gk} - \bar{r}^{(m)}_g}{\max(s_m, \epsilon_A)}. \tag{5}$$

Centering within each group compares conformations generated for the same molecular input. Scaling each component reduces sensitivity to the numerical units of individual rewards. Signed weights

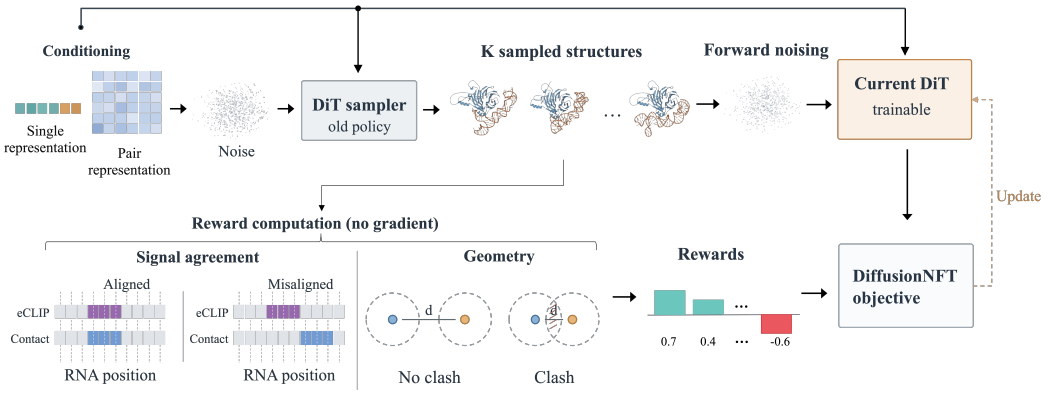


Figure 3: **Stage 2: conformation optimization.** Frozen representations condition sampled structures. Experimental and geometric rewards weight the positive and negative branches of the coordinate-space DiffusionNFT update.

$\omega_m$ favor experimental agreement and penalize geometric violations. The combined advantage is detached from the computation graph and mapped to a bounded branch weight,

$$r_{gk} = \frac{1}{2} + \frac{\mathrm{clip}(A_{gk}, -a, a)}{2a} \in [0, 1]. \tag{6}$$

Positive relative advantages give $r_{gk} > 1/2$ and increase the positive-branch contribution; negative advantages increase the negative-branch contribution.

**Coordinate-space update.** A generated conformation $X$ undergoes a random rigid transformation and is perturbed with fresh noise as $X_\sigma = X + \sigma\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, I)$ and $\sigma$ follows the Protenix training noise distribution. All three policies receive identical noisy coordinates and molecular conditioning. Omitting these common inputs from the notation, the implicit positive and negative denoising branches are

$$D^+ = \beta D_\theta + (1-\beta)\,\mathrm{sg}(D_{\bar\theta}), \qquad D^- = (1+\beta)\,\mathrm{sg}(D_{\bar\theta}) - \beta D_\theta. \tag{7}$$

Here sg stops gradient propagation; both branches retain the coordinate parameterization of Protenix.

Let $\ell$ be the masked adaptive coordinate error defined in Appendix A. With $r_{gk}$ from Equation 6, the NFT objective is

$$\mathcal{L}_{\mathrm{NFT}} = \mathbb{E}_{g,k,\sigma,\varepsilon}\left[\frac{a}{\beta}\left\{r_{gk}\ell(D^+, X^{(gk)}) + (1-r_{gk})\ell(D^-, X^{(gk)})\right\}\right]. \tag{8}$$

The old policy generates the coordinate targets online. Experimental feedback enters through the branch weights, so the update does not require reward gradients.

**Reference constraint and policy refresh.** The total loss includes a fixed-policy output constraint, $\mathcal{L} = \mathcal{L}_{\mathrm{NFT}} + \lambda_{\mathrm{ref}}\mathcal{L}_{\mathrm{ref}}$. The reference term penalizes deviations from $D_{\theta_0}$, with scaling that accounts for the EDM output preconditioning in Protenix. After optimization, the old policy is updated as $\bar\theta \leftarrow \gamma_t\bar\theta + (1-\gamma_t)\theta$. This update allows the sampling distribution to evolve while the fixed reference continues to constrain deviations from the initialization. Algorithm 1 summarizes the training cycle. Appendix A provides the numerical settings and reference-loss definition.

## 4 Experiments

We assess ESFold using a structural benchmark and an experimental-binding test set. The former measures agreement with resolved complexes; the latter measures agreement between predicted RNA contacts and eCLIP profiles. Figure 1 summarizes the primary results. Appendices D and C provide full results, including eight structural tasks beyond protein–RNA prediction.

**Algorithm 1** ESFold Stage 2: coordinate-space adaptation of DiffusionNFT with experimental signals

**Require:** Stage 1 checkpoint with adapted frozen Pairformer, denoiser $D_{\theta_0}$, aligned examples $(c, y)$
Initialize $\theta \leftarrow \theta_0$ and $\bar{\theta} \leftarrow \theta_0$
**for** each optimization cycle **do**
  Sample $K$ conformations per condition using $D_{\bar{\theta}}$ without gradients
  Compute contact profiles, reward components, and sample-validity masks
  Retain groups with at least two valid samples
  Compute detached advantages $A_{gk}$ and weights $r_{gk}$ using Eqs. 5–6
  **for** each retained coordinate minibatch **do**
    Apply a random rigid transform and draw fresh forward noise
    Evaluate current, old, and reference denoisers on identical noisy inputs
    Form $D^+$ and $D^-$ using Eq. 7
    Update $\theta$ to minimize $\mathcal{L}_{\mathrm{NFT}} + \lambda_{\mathrm{ref}}\mathcal{L}_{\mathrm{ref}}$
  **end for**
  Refresh $\bar{\theta} \leftarrow \gamma_t\bar{\theta} + (1 - \gamma_t)\theta$; keep $\theta_0$ fixed
**end for**
**return** Fine-tuned denoiser $D_\theta$

### 4.1 Training data and geometric references

Stage 1 uses PDB structures with a cutoff of 30 September 2021 and eCLIP sequence–signal examples from the training partition of a processed candidate corpus. Across its training, validation, and test partitions, this corpus contains 265,070 candidate protein–RNA pairs spanning 53,014 source RNA windows and 84 proteins. The sequence–signal data come from version v1 of the PanRBP-Net ENCODE Dataset on Zenodo (Horlacher, 2024). We preserve the release's training, validation, and test assignments, with RNA-level redundancy control and aligned sequence–signal cropping. The experimental-binding evaluation uses 190 held-out test conditions excluded from both training stages. Appendix B describes the source dimensions, filtering, and partition counts.

For a subset of training pairs, Protenix generates 50 conformations using 10 seeds and five samples per seed. We retain the conformation with the highest experimental signal agreement as a structural pseudo-label, yielding 25,529 unique distillation structures. These structures provide the local geometric references used by the Stage 2 rewards (Appendix A). Stage 1 eCLIP supervision instead aligns distogram-derived contact probabilities with experimental signal profiles. Signal-guided structure selection is confined to training data; benchmark predictions are selected by model confidence.

### 4.2 Baselines and inference protocol

We compare ESFold with AlphaFold 3 (Abramson et al., 2024), Boltz-1 (Wohlwend et al., 2024), Chai-1 (Chai Discovery Team, 2024), HelixFold 3 (Liu et al., 2024), and Protenix (Zhang et al., 2026) on the structural benchmark (Xu et al., 2026). Results for the first four baselines are taken from the published structural benchmark; results for Protenix and the three ESFold variants are computed from the evaluated model predictions. Appendix E summarizes data alignment and evaluation consistency.

The Protenix baseline and ESFold initialization use Protenix v1.0.0 (Appendix A). ESFold(rep) denotes supervised adaptation from Protenix, ESFold(nft) denotes direct NFT adaptation from Protenix, and ESFold denotes the complete two-stage model.

The Protenix baseline and the final two-stage ESFold model use the same input construction, sequence features, and inference budget of 25 candidates per input. Candidates are generated with five seeds and five samples per seed. All models and single-stage variants evaluated in this work use the same conformation-selection criterion: retain the candidate with the highest ranking score (Top-1). Neither native structural scores nor agreement with test eCLIP profiles is used to select the reported conformation. Appendix A provides additional inference settings. The structural benchmark defines low homology relative to structural data preceding 13 January 2023. We retain this benchmark protocol when evaluating the two-stage pipeline.

### 4.3 Structural accuracy on protein–RNA interfaces

The structural benchmark contains 70 annotated protein–RNA interfaces from 49 assemblies (Xu et al., 2026). We evaluate one confidence-ranked Top-1 prediction per assembly, retaining the full assembly context. The primary metric is interface success, defined by DockQ $\geq 0.23$; complete structural metrics and aggregation details appear in Appendix D.

ESFold achieves 74.29% success, compared with 61.43% for Protenix and 62.32% for AlphaFold 3 (Figure 1(a)). The full model also improves over ESFold(rep) at 70.00% and ESFold(nft) at 68.57%, and achieves the best reported LDDT and RMSD metrics among the evaluated models. These results favor the full adaptation sequence on protein–RNA interfaces.

To assess whether protein–RNA adaptation compromises broader structural prediction, we compare complex success rates and monomer LDDT with Protenix (Tables 6 and 7). Relative to Protenix, the full ESFold model matches or improves the primary metric in seven of eight additional categories; antibody–antigen success decreases slightly from 50.61% to 50.00%. Thus, the protein–RNA gains are not accompanied by broad degradation in these primary metrics. Effects remain task- and metric-dependent: for example, DNA monomer LDDT improves while RMSD increases (Table 9).

### 4.4 Agreement with held-out eCLIP profiles

We evaluate 190 held-out protein–RNA conditions spanning 82 RBPs within experimentally localized RNA windows. Experimental profiles define the cropped benchmark inputs and evaluation targets; models receive only the molecular inputs, and candidate selection uses model confidence. All seven models receive identical inputs and complete all evaluations. The primary metric is peak F1 averaged first within each RBP and then equally across RBPs. Figure 1(b) shows the means and RBP-bootstrap confidence intervals; Appendix C provides metric definitions, protocol details, and full results.

ESFold achieves RBP-macro peak F1 of 0.459, compared with 0.332 for Protenix and 0.358 for the strongest external baseline, Chai-1. ESFold(rep) reaches 0.417, while direct NFT adaptation reaches 0.358. The full model also leads on the remaining signal metrics, supporting improved agreement with experimentally supported RNA regions.

**Case studies: localizing experimentally supported RNA regions.** Figure 4 compares AlphaFold 3 and ESFold on NIP7 and IGF2BP1. For NIP7, AlphaFold 3 places its main contacts away from the experimental peak, whereas ESFold concentrates contacts near G46. For IGF2BP1, ESFold places contacts around the dominant region at positions 46–50 and reduces several off-peak contacts visible in the AlphaFold 3 prediction. Positional offsets and off-peak contacts remain, so the improvement concerns regional agreement rather than exact recovery of every signal position. These selected examples illustrate the aggregate results; Appendix F provides five-model comparisons for NIP7 and IGF2BP1.

## 5 Related Work

**Structural priors and experimental restraints.** Structural archives underpin learned molecular prediction (Berman et al., 2000), and AlphaFold established the value of structural priors and self-distillation (Jumper et al., 2021). Integrative approaches such as HADDOCK and the Integrative Modeling Platform combine prior geometry with experimental restraints (Dominguez et al., 2003; Russel et al., 2012). AlphaLink and AlphaLink2 incorporate crosslinking mass-spectrometry evidence into learned protein and complex prediction (Stahl et al., 2023; 2024). GRASP extends restraint-assisted prediction to crosslinking, covalent labeling, chemical shift perturbation, and deep mutational scanning (Xie et al., 2025). This broader treatment is relevant to eCLIP because experimental evidence need not identify a unique residue pair to inform an interface. The distinction in ESFold is how that evidence is used: binding profiles supervise adaptation across training examples, rather than serving as required restraints for each test complex.

**Measurement-guided structures and ensembles.** Experiment-guided AlphaFold3 uses NMR, crystallographic, and cryo-EM measurements to guide reverse diffusion through gradients of mea-

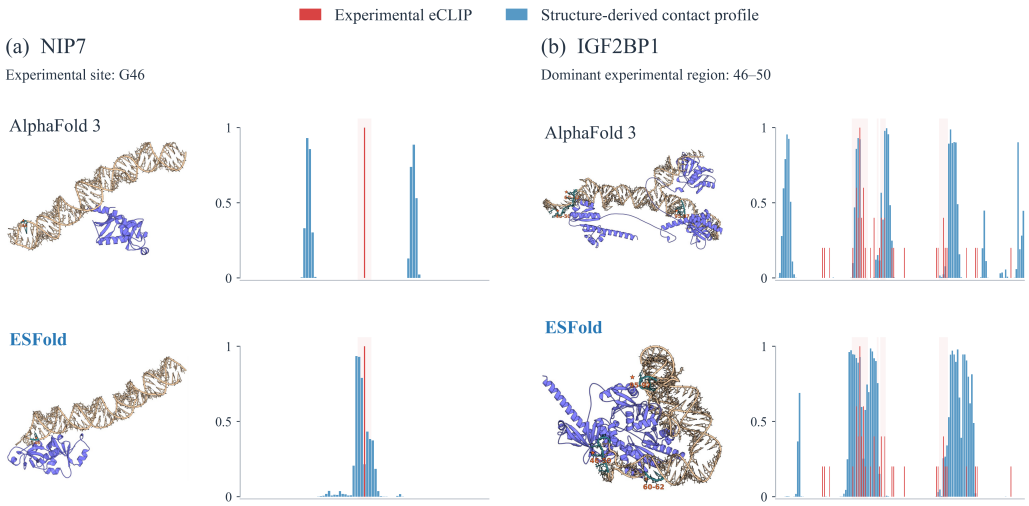


Figure 4: **AlphaFold 3 and ESFold on (a) NIP7 and (b) IGF2BP1.** Structures show protein in purple, RNA in tan, and experimentally supported RNA positions in teal. Red bars: maximum-normalized eCLIP; blue bars: contact scores; shading: experimental peaks. Profiles use unsmoothed positional values and a shared 0–1 scale.

surement likelihoods (Maddipatla et al., 2026). For RNA, HORNET combines atomic force microscopy with modeling and neural-network assessment to recover distinct conformers (Degenhardt et al., 2025), while DRRAFTER builds RNA coordinates into cryo-EM maps (Kappel et al., 2018). NPDock and RNP-denovo address protein–RNA assembly through docking and coupled folding–docking, respectively (Tuszynska et al., 2015; Kappel & Das, 2019). These methods address structure reconstruction under different information sources and resolutions.

**Functional maps and RNA representations.** Large-scale eCLIP studies complement broader functional-genomics resources by measuring protein associations along RNA (ENCODE Project Consortium, 2012; Van Nostrand et al., 2016; 2020). DeepBind and GraphProt learn binding preferences from sequence or sequence–structure features (Alipanahi et al., 2015; Maticzka et al., 2014). Shared representations support learning across related tasks (Caruana, 1997), as illustrated by DeepRiPe across RBPs (Ghanbari & Ohler, 2020). RBPNet predicts position-resolved profiles with separate protein-specific and background components (Horlacher et al., 2023), whereas ERNIE-RNA and NucleicBERT emphasize reusable sequence representations for structural and functional tasks (Yin et al., 2025; Upadhyay et al., 2026). iCLIP and PureCLIP address localization of crosslink events and assay bias (König et al., 2010; Krakau et al., 2017).

**Diffusion adaptation and inference-time steering.** Reward-weighted training, diffusion preference optimization, differentiable reward optimization, and online policy gradients adapt diffusion models through different feedback mechanisms (Lee et al., 2023; Wallace et al., 2024; Clark et al., 2024; Black et al., 2024; Liu et al., 2025). Feynman–Kac steering instead resamples interacting trajectories according to reward-derived potentials at inference (Singhal et al., 2025). This distinction between steering a sampler and updating model parameters also applies to molecular prediction: AlphaFlow learns conformational distributions from structural and simulation data (Jing et al., 2024). ESFold uses the existing forward-process, negative-aware objective of DiffusionNFT (Zheng et al., 2026), with regional signal rewards and molecular geometry constraints.

## 6 Discussion and Limitations

ESFold has two main limitations. First, eCLIP provides regional binding evidence rather than atomic restraints and is influenced by assay and cellular context. We therefore combine signal supervision with geometric regularization and evaluate structural accuracy independently of signal agreement.

Second, our experimental evaluation focuses on selected protein–RNA pairs and signal-rich RNA windows. The results establish performance within this setting; transfer to unseen RBP families and transcriptome-wide binding-site discovery remains to be evaluated. Extending the framework to additional experimental modalities and broader molecular systems is a promising direction.

## 7 Conclusion

ESFold combines supervised representation adaptation with a molecular coordinate-space adaptation of DiffusionNFT. The resulting model improves both protein–RNA interface accuracy and agreement with held-out experimental binding signals, outperforming either single-stage variant on the primary metrics. These results support using regional functional evidence to post-train biomolecular structure models without requiring experimental signals at inference.

## Reproducibility Statement

Appendices A–C describe the two training stages, candidate-data processing, and signal metrics. Evaluation records comprise benchmark inputs, model-selected predictions, chain mappings, and per-sample scores. Appendix E documents data alignment and evaluation consistency.

### AI use statement

Generative AI tools assisted with manuscript drafting and revision, language and organization, literature identification and summarization, reference formatting, and the preparation of plotting code and figures from existing experimental results. AI assistance also supported the presentation and interpretation of results and provided feedback on methodological descriptions. The authors retain responsibility for all scientific judgments and the final content of this work, including the text, references, results, and AI-assisted artifacts.

## References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Augustin Žídek, Alex Bridgland, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630 (8016):493–500, 2024.

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33:831–838, 2015. doi: 10.1038/nbt.3300. URL https://doi.org/10.1038/nbt.3300.

Minkyung Baek, Ryan McHugh, Ivan Anishchenko, Hanlun Jiang, David Baker, and Frank DiMaio. Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nature Methods*, 21:117–121, 2024. doi: 10.1038/s41592-023-02086-5. URL https://www.nature.com/articles/s41592-023-02086-5.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.13301.

Rich Caruana. Multitask Learning. *Machine Learning*, 28:41–75, 1997. doi: 10.1023/a:1007379606734. URL https://doi.org/10.1023/a:1007379606734.

Chai Discovery Team. Chai-1 technical report. Technical report, Chai Discovery, September 2024. URL https://chaiassets.com/chai-1/paper/technical_report_v1.pdf.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.17400.

Maximilia F. S. Degenhardt, Hermann F. Degenhardt, Yuba R. Bhandari, Yun-Tzai Lee, Jienyu Ding, Ping Yu, William F. Heinz, Jason R. Stagno, Charles D. Schwieters, Norman R. Watts, Paul T. Wingfield, Alan Rein, Jinwei Zhang, and Yun-Xing Wang. Determining structures of RNA conformers using AFM and deep neural networks. *Nature*, 637:1234–1243, 2025. doi: 10.1038/s41586-024-07559-x. URL https://doi.org/10.1038/s41586-024-07559-x.

Cyril Dominguez, Rolf Boelens, and Alexandre M. J. J. Bonvin. HADDOCK: A protein–protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125 (7):1731–1737, 2003. doi: 10.1021/ja026939x. URL https://doi.org/10.1021/ja026939x.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

Mahsa Ghanbari and Uwe Ohler. Deep neural networks for interpreting RNA-binding protein target preferences. *Genome Research*, 30:214–226, 2020. doi: 10.1101/gr.247494.118. URL https://doi.org/10.1101/gr.247494.118.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2006.11239.

Marc Horlacher. PanRBPNet ENCODE Dataset (filtered by min max-positional CL count of 5, across eCLIP and SMInput). Zenodo, November 2024. URL https://doi.org/10.5281/zenodo.14176118. Dataset, version v1.

Marc Horlacher, Nils Wagner, Lambert Moyon, Klara Kuret, Nicolas Goedert, Marco Salvatore, Jernej Ule, Julien Gagneur, Ole Winther, and Annalisa Marsico. Towards in silico CLIP-seq: predicting protein-RNA interaction via sequence-to-signal learning. *Genome Biology*, 24:180, 2023. doi: 10.1186/s13059-023-03015-7. URL https://doi.org/10.1186/s13059-023-03015-7.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. AlphaFold meets flow matching for generating protein ensembles. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 22277–22303, 2024. URL https://proceedings.mlr.press/v235/jing24a.html.

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

Kalli Kappel and Rhiju Das. Sampling Native-like Structures of RNA-Protein Complexes through Rosetta Folding and Docking. *Structure*, 27:140–151.e5, 2019. doi: 10.1016/j.str.2018.10.001. URL https://doi.org/10.1016/j.str.2018.10.001.

Kalli Kappel, Shiheng Liu, Kevin P. Larsen, Georgios Skiniotis, Elisabetta Viani Puglisi, Joseph D. Puglisi, Z. Hong Zhou, Rui Zhao, and Rhiju Das. De novo computational RNA modeling into cryo-EM maps of large ribonucleoprotein complexes. *Nature Methods*, 15:947–954, 2018. doi: 10.1038/s41592-018-0172-2. URL https://doi.org/10.1038/s41592-018-0172-2.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2206.00364.

Julian König, Kathi Zarnack, Gregor Rot, Tomaž Curk, Melis Kayikci, Blaž Zupan, Daniel J Turner, Nicholas M Luscombe, and Jernej Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Structural & Molecular Biology*, 17:909–915, 2010. doi: 10.1038/nsmb.1838. URL https://doi.org/10.1038/nsmb.1838.

Sabrina Krakau, Hugues Richard, and Annalisa Marsico. PureCLIP: capturing target-specific protein–RNA interaction footprints from single-nucleotide CLIP-seq data. *Genome Biology*, 18:240, 2017. doi: 10.1186/s13059-017-1364-2. URL https://doi.org/10.1186/s13059-017-1364-2.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. doi: 10.48550/arXiv.2302.12192. URL https://arxiv.org/abs/2302.12192.

Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew Y. K. Foong, Victor García Satorras, Osama Abdin, Bastiaan S. Veeling, Iryna Zaporozhets, Yaoyi Chen, Soojung Yang, Adam E. Foster, Arne Schneuing, Jigyasa Nigam, Federico Barbero, Vincent Stimper, Andrew Campbell, Jason Yim, Marten Lienen, Yu Shi, Shuxin Zheng, Hannes Schulz, Usman Munir, Roberto Sordillo, Ryota Tomioka, Cecilia Clementi, and Frank Noé. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389:eadv9817, 2025. doi: 10.1126/science.adv9817. URL https://doi.org/10.1126/science.adv9817.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1362. URL https://arxiv.org/abs/2505.05470.

Lihang Liu, Shanzhuo Zhang, Yang Xue, Xianbin Ye, Kunrui Zhu, Yuxin Li, Yang Liu, Jie Gao, Wenlai Zhao, Hongkun Yu, Zhihua Wu, Xiaonan Zhang, and Xiaomin Fang. Technical report of HelixFold3 for biomolecular structure prediction. *arXiv preprint arXiv:2408.16975*, 2024. doi: 10.48550/arXiv.2408.16975. URL https://arxiv.org/abs/2408.16975.

Advaith Maddipatla, Nadav Sellam Bojan, Meital Bojan, Volodymyr Masalitin, Sanketh Vedula, Paul Schanda, Ailie Marx, and Alex M. Bronstein. Experiment-guided AlphaFold3 resolves measurement-consistent protein ensembles. *Nature Biotechnology*, 2026. doi: 10.1038/s41587-026-03166-5. URL https://doi.org/10.1038/s41587-026-03166-5.

Daniel Maticzka, Sita J Lange, Fabrizio Costa, and Rolf Backofen. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biology*, 15:R17, 2014. doi: 10.1186/gb-2014-15-1-r17. URL https://doi.org/10.1186/gb-2014-15-1-r17.

Daniel Russel, Keren Lasker, Ben Webb, Javier Velázquez-Muriel, Elina Tjioe, Dina Schneidman-Duhovny, Bret Peterson, and Andrej Sali. Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. *PLoS Biology*, 10:e1001244, 2012. doi: 10.1371/journal.pbio.1001244. URL https://doi.org/10.1371/journal.pbio.1001244.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 55810–55827, 2025. URL https://proceedings.mlr.press/v267/singhal25b.html.

Kolja Stahl, Andrea Graziadei, Therese Dau, Oliver Brock, and Juri Rappsilber. Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. *Nature Biotechnology*, 41:1810–1819, 2023. doi: 10.1038/s41587-023-01704-z. URL https://doi.org/10.1038/s41587-023-01704-z.

Kolja Stahl, Robert Warneke, Lorenz Demann, Rica Bremenkamp, Björn Hormes, Oliver Brock, Jörg Stülke, and Juri Rappsilber. Modelling protein complexes with crosslinking mass spectrometry and deep learning. *Nature Communications*, 15:7866, 2024. doi: 10.1038/s41467-024-51771-2. URL https://doi.org/10.1038/s41467-024-51771-2.

Irina Tuszynska, Marcin Magnus, Katarzyna Jonak, Wayne Dawson, and Janusz M. Bujnicki. NPDock: A web server for protein–nucleic acid docking. *Nucleic Acids Research*, 43(W1):W425–W430, 2015. doi: 10.1093/nar/gkv493. URL https://academic.oup.com/nar/article/43/W1/W425/2467938.

Utkarsh Upadhyay, Julian Herold, Markus Götz, and Alexander Schug. NucleicBERT interprets RNA sequence space through self-supervised language modelling. *Nature Machine Intelligence*, 8:1426–1440, 2026. doi: 10.1038/s42256-026-01295-9. URL https://doi.org/10.1038/s42256-026-01295-9.

Eric L. Van Nostrand, Gabriel A. Pratt, Alexander A. Shishkin, Chelsea Gelboin-Burkhart, Mark Y. Fang, Balaji Sundararaman, Steven M. Blue, Thai B. Nguyen, Christine Surka, Keri Elkins, Rebecca Stanton, Frank Rigo, Mitchell Guttman, and Gene W. Yeo. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, 13(6):508–514, 2016. doi: 10.1038/nmeth.3810. URL https://doi.org/10.1038/nmeth.3810.

Eric L. Van Nostrand, Peter Freese, Gabriel A. Pratt, Xiaofeng Wang, Xintao Wei, Rui Xiao, Steven M. Blue, Jia-Yu Chen, Neal A. L. Cody, Daniel Dominguez, Sara Olson, Balaji Sundararaman, Lijun Zhan, Cassandra Bazile, Louis Philip Benoit Bouvrette, Julie Bergalet, Michael O. Duff, Keri E. Garcia, Chelsea Gelboin-Burkhart, Myles Hochman, Nicole J. Lambert, Hairi Li, Michael P. McGurk, Thai B. Nguyen, Tsultrim Palden, Ines Rabano, Shashank Sathe, Rebecca Stanton, Amanda Su, Ruth Wang, Brian A. Yee, Bing Zhou, Ashley L. Louie, Stefan Aigner, Xiang-Dong Fu, Eric Lécuyer, Christopher B. Burge, Brenton R. Graveley, and Gene W. Yeo. A large-scale binding and functional map of human RNA-binding proteins. *Nature*, 583(7818):711–719, 2020. doi: 10.1038/s41586-020-2077-3. URL https://doi.org/10.1038/s41586-020-2077-3.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024. URL https://arxiv.org/abs/2311.12908.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, 2024. doi: 10.1101/2024.11.19.624167.

Yuhao Xie, Chengwei Zhang, Shimian Li, Xinyu Du, Yanjiao Lu, Min Wang, Yingtong Hu, Zhenyu Chen, Sirui Liu, and Yi Qin Gao. Integrating diverse experimental information to assist protein complex structure prediction by GRASP. *Nature Methods*, 22:2362–2374, 2025. doi: 10.1038/s41592-025-02820-1. URL https://doi.org/10.1038/s41592-025-02820-1.

Sheng Xu, Qiantai Feng, Lifeng Qiao, Hao Wu, Tao Shen, Yu Cheng, Shuangjia Zheng, and Siqi Sun. Benchmarking all-atom biomolecular structure prediction with FoldBench. *Nature Communications*, 17:442, 2026. doi: 10.1038/s41467-025-67127-3. URL https://doi.org/10.1038/s41467-025-67127-3.

Weijie Yin, Zhaoyu Zhang, Shuo Zhang, Liang He, Ruiyang Zhang, Rui Jiang, Gan Liu, Jingyi Wang, Xuegong Zhang, Tao Qin, and Zhen Xie. ERNIE-RNA: an RNA language model with structure-enhanced representations. *Nature Communications*, 16:10076, 2025. doi: 10.1038/s41467-025-64972-0. URL https://doi.org/10.1038/s41467-025-64972-0.

Yuxuan Zhang, Chengyue Gong, Hanyu Zhang, Wenzhi Ma, Zhenyu Liu, Xinshi Chen, Jiaqi Guan, Lan Wang, Yanping Yang, Yu Xia, and Wenzhi Xiao. Protenix-v1: Toward high-accuracy open-source biomolecular structure prediction. *bioRxiv*, 2026. doi: 10.64898/2026.02.05.703733.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, 2026. doi: 10.48550/arXiv.2509.16117. URL https://arxiv.org/abs/2509.16117.

# A Implementation Settings

## A.1 Stage 1: supervised adaptation of binding-aware representations

Initialization uses the checkpoint `protenix_base_default_v1.0.0.pt`.

Stage 1 samples experimentally determined RNA-containing PDB structures and eCLIP sequence–signal examples with probabilities of 0.7 and 0.3, respectively. PDB examples use the standard structural supervision objectives and their original loss coefficients. Interleaving these examples with eCLIP data anchors training to resolved structures and supports stable adaptation.

The eCLIP branch uses the training partition of the filtered, pair-specifically cropped corpus described in Appendix B. During Stage 1, the confidence head and diffusion module (DiT) are frozen, while the Pairformer and distogram head remain trainable. Diffusion-module parameters are not updated on either PDB or eCLIP examples; diffusion-module fine-tuning is performed only in Stage 2. The eCLIP quality objective propagates gradients through the frozen confidence head to its conditioning representations.

**Distogram-derived RNA contact profile.** Let $z_{ijb}$ be the predicted distogram logit for RNA token $i$, protein token $j$, and distance bin $b$, with bin center $\mu_b$. The pairwise contact probability and RNA-position score are

$$p_{ijb} = \mathrm{softmax}_b(z_{ijb}), \qquad C_{ij} = \sum_{b:\mu_b<d_{\mathrm{contact}}} p_{ijb}, \qquad s_i = \max_{j\in\mathcal{P}} C_{ij}, \tag{9}$$

where $\mathcal{P}$ is the set of protein tokens and $d_{\mathrm{contact}} = 8$ Å is the contact threshold. Taking the maximum over protein tokens produces one contact score per RNA position. These scores are computed directly from the distogram; Stage 2 instead scores contact profiles derived from sampled coordinates.

**Experimental profile alignment.** The eCLIP objective is a signal-weighted cross-entropy over annotated RNA positions. Let $\mathcal{V}$ contain the valid positions, and let $y_i$ denote the aligned experimental signal. Averaging and rescaling yield nonnegative real-valued signal weights rather than necessarily integer counts. The implementation clamps negative signal values to zero, caps each position at $c_{\max}$, and rescales the resulting profile when its total exceeds $T_{\max}$:

$$v_i = \min(\max(y_i, 0), c_{\max}), \qquad \widetilde{y}_i = v_i \min\left(1, \frac{T_{\max}}{\max(\sum_{j\in\mathcal{V}} v_j, \epsilon)}\right). \tag{10}$$

The predicted positional distribution is

$$\pi_i = \frac{\max(s_i, \epsilon)}{\sum_{j\in\mathcal{V}} \max(s_j, \epsilon)}. \tag{11}$$

For profiles whose maximum nonnegative signal reaches the eligibility threshold $h_{\min}$, the parameter-dependent signal-alignment term is

$$\mathcal{L}_{\mathrm{eCLIP}} = -\sum_{i\in\mathcal{V}} \widetilde{y}_i \log \pi_i. \tag{12}$$

Target-only additive terms are omitted because they do not affect model gradients. For $T = \sum_{i\in\mathcal{V}} \widetilde{y}_i > 0$, this objective equals $T$ times the cross-entropy between the normalized target profile $\widetilde{\mathbf{y}}/T$ and $\boldsymbol{\pi}$; it therefore retains the total-signal weighting across examples. Profiles below the eligibility threshold, or without valid positions, contribute zero signal loss. We use $c_{\max} = 30$, $T_{\max} = 100$, $h_{\min} = 3$, and $\epsilon = 10^{-8}$, with a profile weight of 1.0. Larger experimental signals receive greater weight in the positional objective, emphasizing signal-rich regions. No additional signal smoothing is applied within the loss.

**PDB auxiliary contact-profile objective.** In addition to standard structural supervision, PDB examples use a KL-divergence objective between structure-derived RNA contact profiles and the predicted positional distribution. Let $t_i \geq 0$ be the structure-derived target at a valid RNA position and

$\widehat{t}_i = t_i / \sum_{j \in \mathcal{V}} t_j$. For targets with a maximum value of at least 0.5 and a total signal greater than $\epsilon$, the PDB-branch loss is

$$\mathcal{L}_{\mathrm{PDB,total}} = \mathcal{L}_{\mathrm{structure}} + 0.05\, D_{\mathrm{KL}}(\widehat{\mathbf{t}} \,\|\, \boldsymbol{\pi}). \tag{13}$$

The auxiliary term uses a profile weight of 1.0 and a target threshold of zero; it is zero for targets that do not meet the eligibility conditions or contain no valid positions. The standard structural loss remains active.

**Confidence quality term.** The eCLIP branch also includes a confidence-based quality term. A sampled conformation is evaluated by the confidence head, and the predicted atom-level pLDDT scores are averaged on their normalized scale to obtain $\bar{c}$. The total eCLIP-branch loss is

$$\mathcal{L}_{\mathrm{eCLIP,total}} = \lambda_{\mathrm{signal}} \mathcal{L}_{\mathrm{eCLIP}} + \lambda_{\mathrm{quality}} \max(0, c_{\mathrm{target}} - \bar{c}), \tag{14}$$

where $\lambda_{\mathrm{signal}} = 0.05$, $\lambda_{\mathrm{quality}} = 0.2$, and $c_{\mathrm{target}} = 0.8$. The quality penalty is active when the mean predicted confidence falls below this target. Each eCLIP example uses one conformation sampled with 20 reverse-diffusion steps and 10 recycling iterations. Sampling does not retain gradients through the reverse diffusion trajectory.

Stage 1 uses AdamW with a learning rate of $10^{-4}$, 2,000 warmup steps, bfloat16 arithmetic, an EMA decay of 0.999, seven GPUs, seed 42, and a maximum training crop of 650 tokens. The diffusion batch size is 32. The resulting checkpoint initializes Stage 2, in which the adapted conditioning representation is frozen.

### A.2 Contact mapping and experimental rewards

**Reward composition.** The signal reward combines peak F1, position-level Jaccard overlap $J$, contact enrichment $E$, rank agreement $S$, and a coverage penalty $C$:

$$R_{\mathrm{signal}}(q, y) = w_F F_{\mathrm{peak}} + w_J J + w_E E + w_S S - w_C C. \tag{15}$$

Peak F1 measures overlap between contact regions and experimental peaks. The remaining terms address misplaced, weakly enriched, or excessively broad contacts. Their definitions follow below.

**Auxiliary group feedback.** The marginal reward measures how each sample changes agreement between the group-averaged contact profile and the experiment:

$$R^{(k)}_{\mathrm{marginal}} = R_{\mathrm{signal}}\left(\frac{1}{K}\sum_{j} q^{(j)}, y\right) - R_{\mathrm{signal}}\left(\frac{1}{K-1}\sum_{j \neq k} q^{(j)}, y\right). \tag{16}$$

This auxiliary term is used during training; evaluation uses confidence-ranked Top-1 predictions. Conditions require usable protein–RNA content and experimental signal, and groups with fewer than two valid samples are excluded from within-condition comparisons.

**Coordinate-to-contact mapping.** For RNA position $i$, let $d_i$ be the minimum distance from its representative atom to a protein representative atom. The contact score is

$$q_i = \sum_{b:\, \mu_b < d_c} \frac{\exp[-(d_i - \mu_b)^2/(2\tau^2)]}{\sum_{b'} \exp[-(d_i - \mu_{b'})^2/(2\tau^2)]}, \tag{17}$$

where $\mu_b = d_{\min} + (b + \frac{1}{2})\Delta d$ are the distance-bin centers, $b = 0, \ldots, B-1$, and $\Delta d = (d_{\max} - d_{\min})/B$. Here $d_{\min}$ and $d_{\max}$ denote the outer bin edges rather than the centers. Table 1 specifies $B$, the edge bounds, $d_c$, and $\tau$. Signal evaluation uses the same mapping for all models.

Experimental peaks are defined separately within each contiguous annotated RNA segment after moving-average smoothing. Positions at or above the 90th percentile of the smoothed segment values are retained only if their signal is positive. Adjacent retained positions form peak regions. Narrow regions are expanded symmetrically toward a target width, subject to bounded expansion on each side. Predicted regions comprise consecutive positions with $q_i \geq 0.6$. Segment boundaries prevent regions from merging across missing positions or chains.

The reward in Equation 15 uses position-level Jaccard overlap $J$ between predicted contacts and expanded peaks. Enrichment is defined as $E = \{1 + \tanh[\log((\bar{q}_{\text{peak}} + \epsilon)/(\bar{q}_{\text{bg}} + \epsilon))]\}/2$ and compares mean contact scores inside and outside the expanded peak mask. Rank agreement is $S = (1 + \rho_{\text{Spearman}})/2$, computed relative to the smoothed signal. If $f$ denotes the fraction of positions predicted as positive, the coverage penalty is $C = \max(0, f_{\min} - f) + \max(0, f - f_{\max})$.

**Local geometry preservation.** The pseudo-local lDDT component uses pairs of representative atoms from the same chain in the signal-selected distillation structure described in Section 4.1. Stored coordinates and residue confidence scores provide the reference geometry and confidence weights; no additional reference structure is generated for this reward. Both residues must have confidence scores of at least 70, and their reference distance must lie between $10^{-3}$ and 30 Å. For each eligible pair, the score averages indicators of whether the absolute distance error is below 0.5, 1, 2, and 4 Å. Pair scores are weighted by the product of the two residue confidence scores after normalization to $[0, 1]$. This component is set to zero if confidence annotations are unavailable or fewer than eight eligible pairs remain. It preserves local geometry from the pseudo-structure without specifying an experimentally determined protein–RNA interface. The physical penalty separately evaluates clashes and chain continuity.

**Physical penalty.** Let $\mathcal{A}$ denote the set of valid heavy atoms and $\mathcal{N}$ the set of eligible unordered nonlocal atom pairs. We exclude from $\mathcal{N}$ atom pairs within the same token, atom pairs on the same chain separated by at most one residue index, and atom pairs belonging to explicitly bonded token pairs. The clash count per valid heavy atom is

$$c(X) = \frac{\sum_{(a,b)\in\mathcal{N}} \mathbf{1}[\|x_a - x_b\| < 1.5\,\text{Å}]}{\max(|\mathcal{A}|, 1)}. \tag{18}$$

Let $\mathcal{C}$ denote pairs of consecutive valid representative atoms on the same chain with residue indices differing by one, and let $e_{ij} = |\|x_i - x_j\| - \|x_i^{\text{dist}} - x_j^{\text{dist}}\||$ measure the distance deviation from the distillation reference. The chain-continuity penalty and total physical penalty are

$$b(X) = \frac{1}{|\mathcal{C}|} \sum_{(i,j)\in\mathcal{C}} \frac{\max(0, e_{ij} - \delta)}{\delta}, \qquad \delta = 3\,\text{Å}, \tag{19}$$

$$P_{\text{phys}}(X) = 10 \min(c(X), 1) + \min(b(X), 1) + \mathbf{1}[X \text{ contains nonfinite coordinates}]. \tag{20}$$

We set $b(X) = 0$ when $\mathcal{C}$ is empty. The nonnegative penalty $P_{\text{phys}}$ enters component-wise reward normalization with a weight of $-0.30$. Samples with nonfinite coordinates or $c(X) > 0.02$ are rejected independently of this graded penalty. Chain-continuity violations do not trigger hard rejection; they contribute through $b(X)$. Table 1 summarizes the numerical settings.

### A.3 Coordinate objective and optimization

The training noise distribution is $\sigma = \sigma_{\text{data}} \exp(p_{\text{mean}} + p_{\text{std}} z)$, where $z \sim \mathcal{N}(0, 1)$. Averaging valid coordinate entries under mask $M$ gives the adaptive error

$$\ell(Y, X) = \frac{\langle (Y - X)^2 \rangle_M}{\max(\text{sg}(\langle |Y - X| \rangle_M), \epsilon_\ell)}. \tag{21}$$

The fixed-policy penalty is

$$\mathcal{L}_{\text{ref}} = \mathbb{E}\left[ \frac{\langle (D_\theta - \text{sg}(D_{\theta_0}))^2 \rangle_M}{\max(c_{\text{out}}(\sigma)^2, \epsilon_\ell^2)} \right], \qquad c_{\text{out}}(\sigma) = \frac{\sigma \sigma_{\text{data}}}{\sqrt{\sigma^2 + \sigma_{\text{data}}^2}}. \tag{22}$$

Dividing by $c_{\text{out}}^2$ expresses deviations on the scale of the underlying network outputs. The old-policy update schedule uses $\gamma_t = \min(0.001t, 0.5)$, where $t$ denotes the optimizer step. Only the diffusion module is optimized.

Protein MSA and template features are reused from shared resources, with cached protein-MSA search results used as a fallback. RNA MSA features are included when available in the shared cache. The Protenix baseline and ESFold use identical input construction and sequence-feature resources during evaluation.

Table 1: Numerical settings for the Stage 2 NFT objective, experimental rewards, and inference. Geometric thresholds specify the validity checks and physical penalty.

| Setting | Value |
|---|---|
| *Contact and signal rewards* | |
| Distance bins $B$; outer edges $[d_{\min}, d_{\max}]$ | 64; 2.3125–21.6875 Å |
| Contact cutoff $d_c$; smoothing scale $\tau$ | 8 Å; 2 Å |
| Experimental smoothing; peak quantile | 5 nt; 0.90 |
| Dynamic target width; slack bounds | 5 nt; 0–2 nt per side |
| Predicted positive threshold | 0.6 |
| $(w_F, w_J, w_E, w_S, w_C)$ | $(0.45, 0.20, 0.20, 0.15, 0.20)$ |
| Contact-fraction bounds $(f_{\min}, f_{\max})$ | $(0.02, 0.35)$ |
| Enrichment stabilizer $\epsilon$ | $10^{-6}$ |
| *Geometric checks* | |
| Nonlocal heavy-atom clash distance | 1.5 Å |
| Maximum clash count per valid heavy atom | 0.02 |
| Clash-ratio penalty scale; upper clipping | 10; 1 |
| Chain-continuity tolerance; penalty cap | 3 Å; 1 |
| Hard rejection on chain-continuity violations | Disabled |
| Nonfinite-coordinate handling | Reject sample |
| *NFT optimization* | |
| Component weights (signal, marginal, pseudo-local lDDT, physical penalty) | $(1.0, 0.30, 0.50, -0.30)$ |
| Structures per condition $K$ | 8 |
| Reverse-sampling steps per rollout | 50 |
| GPUs; condition groups per GPU per outer cycle | 2; 1 |
| Forward-noise draws per structure; inner passes | 2; 1 |
| $\sigma_{\text{data}}$; $(p_{\text{mean}}, p_{\text{std}})$ | 16 Å; $(-1.2, 1.5)$ |
| NFT interpolation $\beta$; advantage clip $a$ | 1; 5 |
| Reference-loss weight $\lambda_{\text{ref}}$ | 0.1 |
| Advantage and adaptive-error floors $(\epsilon_A, \epsilon_\ell)$ | $(10^{-4}, 10^{-5})$ |
| Optimizer; learning rate | AdamW; $10^{-5}$ |
| Adam moments; weight decay | $(0.9, 0.999)$; $10^{-4}$ |
| Gradient-norm clip; arithmetic | 1.0; bfloat16 |
| Coordinate augmentation | Centering and random rigid transforms |
| *Protenix benchmark inference* | |
| Inference seeds | 42, 66, 101, 2024, 8888 |
| Samples per seed | 5 |
| Recycling iterations; diffusion steps | 10; 200 |
| Conformation selection | Model-ranked Top-1 |

## B CANDIDATE-DATA PROCESSING DETAILS

**Released sequence–signal examples.** We use version v1 of the PanRBPNet ENCODE Dataset (Horlacher, 2024), distributed as `encode.filtered.5.hfds.tar.gz` in Zenodo record 14176118 (DOI: 10.5281/zenodo.14176118). The released data contain RNA genomic windows with sparse one-hot sequences and aligned eCLIP and control matrices. Each 600-nt window has 223 experimental tracks spanning 150 protein symbols in K562 and HepG2 cells. The source partitions contain 512,946 training entries, 116,542 validation entries, and 70,626 test entries, totaling 700,114. A task is a protein–cell-line combination. Export decodes the sparse arrays and creates task-level rows containing the RNA sequence, protein symbol, cell line, and aligned eCLIP signal vector. Protein amino acid sequences are supplied separately through a protein-symbol mapping. The export used here retains the experimental eCLIP tracks.

**Source partitions and retained representatives.** The training, validation, and test assignments supplied with the released data are preserved throughout filtering, protein selection, and pair-specific cropping. The retained source-RNA counts are 41,287 for training, 7,010 for validation, and 4,717 for testing.

**Additional filtering for structure modeling.** RNA scoring groups source records by RNA sequence, yielding 670,340 unique sequences. We retain the top 10% of source RNA windows by the composite signal score below, then apply global and hotspot-level redundancy filters. Proteins are restricted to at most 600 amino acids. Five protein-associated examples are selected per retained RNA by the signal-pattern procedure below; all selected pairs survive pair-specific cropping.

The composite RNA score is $S(x) = 0.18C_{\text{total}} + 0.18C_{\text{peak}} + 0.12C_{\text{count}} + 0.22C_{\text{concentration}} + 0.18C_{\text{length}} + 0.12C_{\text{cross}}$, with each component expressed as a percentile rank. The components measure total signal, peak intensity, preference for an intermediate number of peaks, hotspot concentration, signal normalized by effective length, and hotspot recurrence across proteins, respectively.

Global near-duplicate filtering uses sequence identity and coverage thresholds of 0.9. Local filtering compares binding hotspots using an identity threshold of 0.85 and an overlap threshold of 0.7. These additional filters are applied across the supplied partitions before pair-specific cropping, and retained representatives preserve their original partition labels. These filters are applied in this work after the source-data processing.

**Protein-associated example selection.** For each retained RNA, tracks for the same protein are averaged across available cell lines into one positional signal vector. Active protein profiles are transformed with $\log(1 + y)$, normalized by their positional sums, and grouped by $k$-means. Clusters are ranked by their size multiplied by the Euclidean distance of their center from the global mean profile. The protein closest to each selected cluster center supplies a representative example; remaining slots are filled by unselected proteins with the highest total signal. The retained examples use the mean protein signal vectors.

**Pair-specific cropping.** For pair-specific cropping, the eCLIP signal is smoothed with an 11-nt window. The selection rule combines a relative threshold of 0.2 times the peak value with the 0.6 quantile of positive signal values. Peak segments separated by at most 30 nt are merged, and 40 nt of flanking context is added on each side, subject to a final length of 80–300 nt. The sequence and signal are sliced at identical start and end coordinates. Source RNA identifiers and partition assignments are preserved. Table 2 summarizes the candidate corpus, and Figure 5 reports the numbers of unique RNAs and candidate protein–RNA pairs at each filtering stage.

**Distillation accounting.** The 25,529 selected structures each correspond to a distinct protein–RNA pair; chain and interface sampling entries for the same structure are counted once. The 50-candidate training-label selection budget is separate from the 25-candidate benchmark inference budget.

Table 2: Candidate corpus construction, including the number of candidate protein–RNA pairs retained after cropping and the number of unique structures used for distillation.

| Quantity | Value |
|---|---|
| Unique RNAs entering signal scoring | 670,340 |
| Top 10% signal-rich RNAs | 67,034 |
| RNA representatives after global near-duplicate filtering | 66,825 |
| Non-redundant RNAs after global and hotspot filtering | 53,014 |
| Proteins retained after length filtering | 84 |
| Protein-associated samples per retained RNA | 5 |
| Candidate protein–RNA pairs after cropping | 265,070 |
| Unique distillation training structures | 25,529 |
| Crop length range | 80–300 nt |

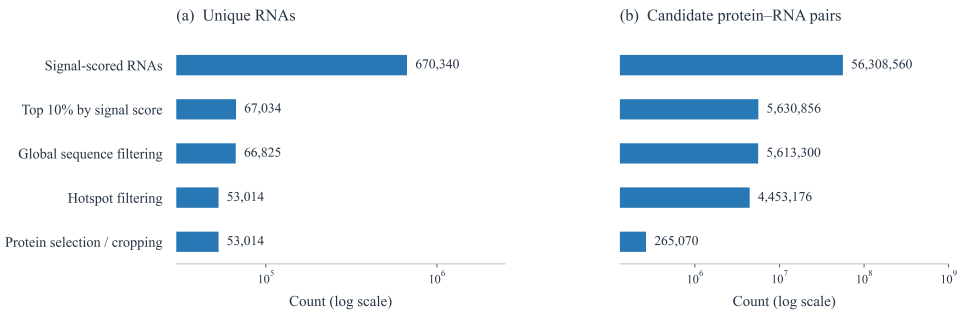


Figure 5: **Candidate-data filtering.** Counts of (a) unique source RNAs and (b) eligible protein–RNA pairs, on logarithmic axes. Pair counts sum eligible proteins per RNA; these are candidate combinations, not verified interactions.

## C Signal Evaluation Details

**Evaluation protocol.** A fixed random sample of 190 conditions is drawn from the supplied test partition and excluded from training. Experimental profiles are used to construct the cropped benchmark inputs and to evaluate predictions. All seven models receive only the same resulting molecular inputs, and candidates are selected by model confidence without access to experimental profiles. This benchmark assesses prediction within experimentally localized RNA windows rather than transcriptome-wide binding-site discovery. Selected-prediction records contain 25 candidates per condition for AlphaFold 3, Chai-1, Protenix, and ESFold; Boltz-1 also selects one confidence-ranked prediction per condition. Experimental peaks require positive signal at or above the 90th percentile after five-position smoothing. The Protenix, ESFold(rep), and ESFold peak-F1 95% RBP-bootstrap intervals are [0.305, 0.361], [0.388, 0.445], and [0.424, 0.494], respectively.

Table 3: Agreement with experimental binding signals on the experimental-binding test set (190 conditions, 82 RBPs). Values are RBP-macro means; higher values indicate better agreement for all metrics. $n$ is the number of completed evaluations out of 190. $J$ denotes peak-contact Jaccard, $\rho$ denotes Spearman correlation with the smoothed eCLIP signal, and $E$ denotes top-signal contact enrichment.

| Method | $n$ | Precision | Recall | Peak F1 | $J$ | $\rho$ | $E$ |
|---|---|---|---|---|---|---|---|
| AlphaFold 3 | 190 | 0.287 | 0.606 | 0.355 | 0.125 | 0.044 | 2.048 |
| Boltz-1 | 190 | 0.230 | 0.764 | 0.334 | 0.122 | 0.044 | 1.331 |
| Chai-1 | 190 | 0.263 | 0.735 | 0.358 | 0.135 | 0.081 | 1.681 |
| Protenix | 190 | 0.237 | 0.719 | 0.332 | 0.126 | 0.051 | 1.732 |
| ESFold(rep) | 190 | 0.313 | 0.793 | 0.417 | 0.155 | 0.144 | 3.044 |
| ESFold(nft) | 190 | 0.265 | 0.715 | 0.358 | 0.129 | 0.071 | 1.881 |
| ESFold | 190 | **0.359** | **0.817** | **0.459** | **0.162** | **0.172** | **5.651** |

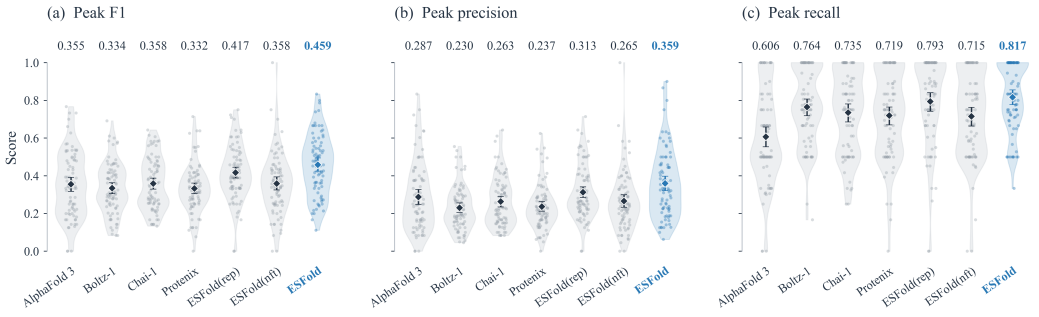


Figure 6: **RBP-level signal agreement.** Points are per-RBP means; violins show their distributions. Diamonds and values denote macro means, with 95% RBP-bootstrap intervals. All seven models use the same 190 conditions and top-40% contact rule.

Let $\mathcal{E}_s$ denote the set of expanded experimental peak regions and $\mathcal{P}_s$ the set of predicted positive regions for sample $s$. During evaluation, positions are classified as positive when $q_i \geq Q_{0.6}(\{q_j : j \text{ valid}\})$, where the threshold is the 60th percentile of the unsmoothed contact scores. Ties at the threshold are retained, so the fraction of positive positions may exceed 40%. This evaluation rule differs from the absolute $q_i \geq 0.6$ threshold used in the training reward. Peak precision and peak recall are

$$P_s = \frac{|\{p \in \mathcal{P}_s : \exists e \in \mathcal{E}_s,\ p \cap e \neq \varnothing\}|}{|\mathcal{P}_s|},$$
$$R_s = \frac{|\{e \in \mathcal{E}_s : \exists p \in \mathcal{P}_s,\ p \cap e \neq \varnothing\}|}{|\mathcal{E}_s|}, \qquad F_{1,s} = \frac{2P_sR_s}{P_s + R_s}. \tag{23}$$

Each region contributes at most one hit to the corresponding numerator; overlap matching does not require one-to-one assignments between regions. If no positive region is predicted, all three scores are set to zero. Peak F1 is also set to zero whenever $P_s + R_s = 0$.

The peak-contact Jaccard score measures intersection over union between the top-40% predicted-positive positions and the expanded experimental peak mask. Spearman correlation compares the continuous $q_i$ profile with the eCLIP signal smoothed over five positions, using annotated RNA positions. Top-signal contact enrichment compares mean $q_i$ inside the expanded experimental peak mask with mean $q_i$ outside it. These profile-based metrics complement region-level F1; enrichment is a ratio and may exceed one. Top-40% post-processing affects the region metrics, Jaccard, and the fraction of signal on predicted contacts, but leaves continuous-profile correlation and enrichment unchanged.

All structures use the same representative-atom convention and are aligned to the input RNA sequence. Positions without experimental annotations are masked consistently across models. Incomplete coordinate mapping is recorded as a prediction failure. For each metric, valid condition scores are averaged within each RBP, and the resulting RBP means receive equal weight in the overall average. The reported 95% percentile intervals are obtained by resampling the 82 RBP means 2,000 times using the fixed benchmark seed. Failed conditions are reflected in the coverage count and are not replaced. Per-sample scores, per-RBP scores, selected prediction paths, and contact profiles accompany the fixed panel identifiers.

**RBP-level metric distributions.** Figure 6 reports peak F1, precision, and recall for all seven models under the same evaluation protocol. Each metric is averaged within each RBP before macro aggregation; consequently, macro F1 is not the harmonic mean of macro precision and recall. Recall depends on the contact-selection rule, and high recall alone does not establish precise localization. The violin distributions show variation between RBPs, while the bootstrap intervals quantify uncertainty in each macro mean. These intervals are not a paired test of differences between models.

# D Detailed Structural Benchmark Results

Protein–RNA predictions retain full assembly context; the benchmark pipeline determines chain correspondence and evaluates all annotated interfaces using the assembly's selected prediction. LDDT measures local structural agreement, while iRMSD and lRMSD measure interface and ligand-chain coordinate error under benchmark conventions. Each metric uses its corresponding valid-evaluation set.

Table 4 reports the complete protein–RNA results corresponding to Figure 1(a).

Table 4: Protein–RNA structure prediction on the structural benchmark (70 interfaces, 49 assemblies). Success is defined as DockQ $\geq 0.23$ for predictions selected by confidence-ranked Top-1 selection. RMSD values are in Å.

| Method | Success (%) ↑ | LDDT ↑ | iRMSD ↓ | lRMSD ↓ |
|---|---|---|---|---|
| AlphaFold 3 | 62.32 | 0.78 | 6.96 | 16.99 |
| Boltz-1 | 56.90 | 0.70 | 9.29 | 23.43 |
| Chai-1 | 50.91 | 0.73 | 9.19 | 25.84 |
| HelixFold 3 | 48.28 | 0.73 | 9.87 | 23.73 |
| Protenix | 61.43 | 0.74 | 7.61 | 18.66 |
| ESFold(rep) | 70.00 | 0.78 | 5.37 | 13.24 |
| ESFold(nft) | 68.57 | 0.76 | 6.56 | 17.19 |
| ESFold | **74.29** | **0.80** | **5.02** | **12.29** |

The additional structural benchmark tasks characterize model behavior beyond protein–RNA interfaces. Together with the primary protein–RNA subset, they cover 1,522 unique biological assemblies. Table 5 reports counts from the evaluation target manifests. Task counts refer to annotated interfaces or monomer targets; a single assembly may contribute multiple interfaces and appear in more than one task.

Table 5: Additional structural benchmark evaluation subsets. Counts refer to the annotated evaluation targets.

| Task | Targets | Primary metric |
|---|---|---|
| Protein–protein | 279 | DockQ success rate |
| Antibody–antigen | 172 | DockQ success rate |
| Protein–ligand | 558 | Joint lRMSD/LDDT-PLI success rate |
| Protein–peptide | 51 | DockQ success rate |
| Protein–DNA | 330 | DockQ success rate |
| Protein monomer | 334 | LDDT |
| RNA monomer | 15 | LDDT |
| DNA monomer | 14 | LDDT |

Input construction, assembly-level Top-1 selection, and metric aggregation follow Section 4. Under the published benchmark definitions, success for protein–protein, antibody–antigen, protein–peptide, protein–DNA, and protein–RNA interfaces requires DockQ $\geq 0.23$. For protein–ligand interfaces, success requires both ligand RMSD (lRMSD) $< 2$ Å and LDDT-PLI $> 0.8$. Monomer tasks use mean LDDT as the primary metric; Table 9 also reports RMSD, TM-score, and GDT-TS. Table 8 reports additional interface metrics. All task scores are computed from valid metric evaluations.

Beyond protein–RNA interfaces, success rates increase relative to Protenix for protein–protein (71.22% to 71.96%), protein–ligand (59.68% to 60.66%), and protein–DNA (70.22% to 77.74%) interactions. Antibody–antigen success decreases from 50.61% to 50.00%, while protein–peptide success remains at 88.89%. For RNA monomers, LDDT increases from 0.60 to 0.63 and RMSD decreases from 8.55 to 7.75 Å. For DNA monomers, LDDT increases slightly, from 0.46 to 0.47, while RMSD increases from 7.79 to 8.47 Å. Protein monomer LDDT remains at 0.89. These mixed results indicate task-dependent effects rather than uniform improvement.

Table 6: Success rates (%, higher is better) on additional structural benchmark interface tasks. P–P: protein–protein; Ab–Ag: antibody–antigen; P–L: protein–ligand; P–Pep: protein–peptide; P–DNA: protein–DNA.

| Method | P–P | Ab–Ag | P–L | P–Pep | P–DNA |
|---|---|---|---|---|---|
| AlphaFold 3 | **72.93** | 47.90 | **64.90** | 82.98 | **79.18** |
| Boltz-1 | 68.25 | 33.54 | 55.04 | 79.17 | 70.97 |
| Chai-1 | 68.53 | 23.64 | 51.23 | 66.67 | 69.97 |
| HelixFold 3 | 66.27 | 28.40 | 51.82 | **89.47** | 50.00 |
| Protenix | 71.22 | 50.61 | 59.68 | 88.89 | 70.22 |
| ESFold(rep) | 72.32 | **51.22** | 60.17 | 88.24 | 76.97 |
| ESFold(nft) | 69.00 | 49.39 | 56.28 | 84.44 | 70.53 |
| ESFold | 71.96 | 50.00 | 60.66 | 88.89 | 77.74 |

Table 7: Monomer structure prediction on the structural benchmark, measured by mean LDDT (higher is better). $N$ denotes the subset size, not the number of valid metric evaluations.

| Method | Protein ($N = 334$) | RNA ($N = 15$) | DNA ($N = 14$) |
|---|---|---|---|
| AlphaFold 3 | 0.88 | 0.61 | **0.53** |
| Boltz-1 | 0.87 | 0.44 | 0.34 |
| Chai-1 | 0.87 | 0.49 | 0.46 |
| HelixFold 3 | 0.86 | 0.55 | 0.29 |
| Protenix | **0.89** | 0.60 | 0.46 |
| ESFold(rep) | **0.89** | **0.64** | 0.46 |
| ESFold(nft) | **0.89** | 0.58 | 0.48 |
| ESFold | **0.89** | 0.63 | 0.47 |

## E Data Alignment and Evaluation Consistency

Each processed example retains its source RNA and protein identifiers, crop coordinates, partition assignment, and aligned signal. Sequence and signal use the same crop boundaries. Predicted coordinates are mapped to these input RNA positions before contact scoring, with missing annotations masked consistently across models. Appendix B describes source-data processing, Appendix C defines signal scoring and aggregation, and Appendix D describes structural evaluation.

Table 8: Additional structural benchmark interface metrics for the evaluated models. RMSD values are in Å.

| Task | Metric | AlphaFold 3 | Boltz-1 | Chai-1 | HelixFold 3 | Protenix | ESFold(rep) | ESFold(nft) | ESFold |
|---|---|---|---|---|---|---|---|---|---|
| Protein–ligand | LDDT-LP ↑ | **0.91** | 0.89 | 0.89 | 0.89 | 0.90 | 0.90 | 0.90 | **0.91** |
| | LDDT-PLI ↑ | **0.77** | 0.71 | 0.70 | 0.69 | 0.73 | 0.74 | 0.71 | 0.74 |
| Protein–protein | iRMSD ↓ | **5.49** | 6.63 | 6.68 | 7.14 | 6.33 | 6.19 | 6.68 | 6.21 |
| | lRMSD ↓ | **12.78** | 16.16 | 15.23 | 16.46 | 14.99 | 15.37 | 16.13 | 15.22 |
| | LDDT ↑ | 0.84 | 0.76 | 0.80 | 0.78 | 0.82 | 0.84 | 0.82 | **0.85** |
| Antibody–antigen | iRMSD ↓ | 7.40 | 8.80 | 10.48 | 9.97 | 7.45 | **6.72** | 7.72 | 7.24 |
| | lRMSD ↓ | 24.54 | 30.70 | 33.44 | 33.02 | 24.56 | **22.01** | 25.36 | 23.39 |
| | LDDT ↑ | **0.88** | 0.84 | 0.85 | 0.84 | 0.87 | 0.87 | 0.87 | 0.87 |
| Protein–DNA | iRMSD ↓ | 4.05 | 4.90 | 5.20 | 7.67 | 4.82 | **3.81** | 4.87 | 3.91 |
| | lRMSD ↓ | 9.37 | 11.30 | 11.74 | 17.06 | 10.89 | **8.83** | 11.00 | 9.13 |
| | LDDT ↑ | 0.85 | 0.81 | 0.81 | 0.71 | 0.84 | 0.85 | 0.84 | **0.86** |
| Protein–peptide | iRMSD ↓ | 2.81 | 2.88 | 3.36 | 3.72 | 2.61 | 2.67 | **2.30** | 2.53 |
| | lRMSD ↓ | 6.56 | 8.82 | 8.77 | 9.79 | 7.18 | 7.33 | **5.94** | 7.25 |
| | LDDT ↑ | **0.84** | 0.83 | 0.82 | 0.83 | 0.82 | 0.83 | 0.82 | 0.83 |

Table 9: Full structural benchmark monomer metrics for the evaluated models. RMSD values are in Å; the other scores are dimensionless.

| Task | Metric | AlphaFold 3 | Boltz-1 | Chai-1 | HelixFold 3 | Protenix | ESFold(rep) | ESFold(nft) | ESFold |
|---|---|---|---|---|---|---|---|---|---|
| Protein | GDT-TS ↑ | 0.92 | 0.91 | 0.90 | 0.91 | **0.93** | 0.92 | 0.92 | 0.92 |
| | TM-score ↑ | 0.94 | 0.94 | 0.93 | 0.94 | 0.94 | **0.95** | 0.94 | 0.94 |
| | RMSD ↓ | 1.95 | 1.97 | 2.10 | 2.03 | 1.88 | **1.84** | 1.86 | 1.87 |
| | LDDT ↑ | 0.88 | 0.87 | 0.87 | 0.86 | **0.89** | **0.89** | **0.89** | **0.89** |
| RNA | GDT-TS ↑ | 0.53 | 0.51 | 0.42 | 0.50 | 0.54 | **0.58** | 0.51 | 0.56 |
| | TM-score ↑ | 0.41 | 0.37 | 0.29 | 0.36 | 0.39 | **0.44** | 0.36 | 0.43 |
| | RMSD ↓ | 8.66 | 11.16 | 13.73 | 11.51 | 8.55 | **7.62** | 9.61 | 7.75 |
| | LDDT ↑ | 0.61 | 0.44 | 0.49 | 0.55 | 0.60 | **0.64** | 0.58 | 0.63 |
| DNA | GDT-TS ↑ | **0.58** | 0.55 | 0.48 | 0.38 | 0.46 | 0.48 | 0.45 | 0.48 |
| | TM-score ↑ | **0.31** | 0.28 | 0.22 | 0.19 | 0.23 | 0.25 | 0.25 | 0.26 |
| | RMSD ↓ | 6.48 | **5.88** | 8.57 | 10.73 | 7.79 | 8.54 | 8.13 | 8.47 |
| | LDDT ↑ | **0.53** | 0.34 | 0.46 | 0.29 | 0.46 | 0.46 | 0.48 | 0.47 |

## F Qualitative Protein–RNA Case Studies

The two examples below are selected from the matched experimental-binding test panel. Each model uses the highest-confidence prediction selected under the protocol in Section 4. RNA positions are indexed from one within the cropped input sequence.

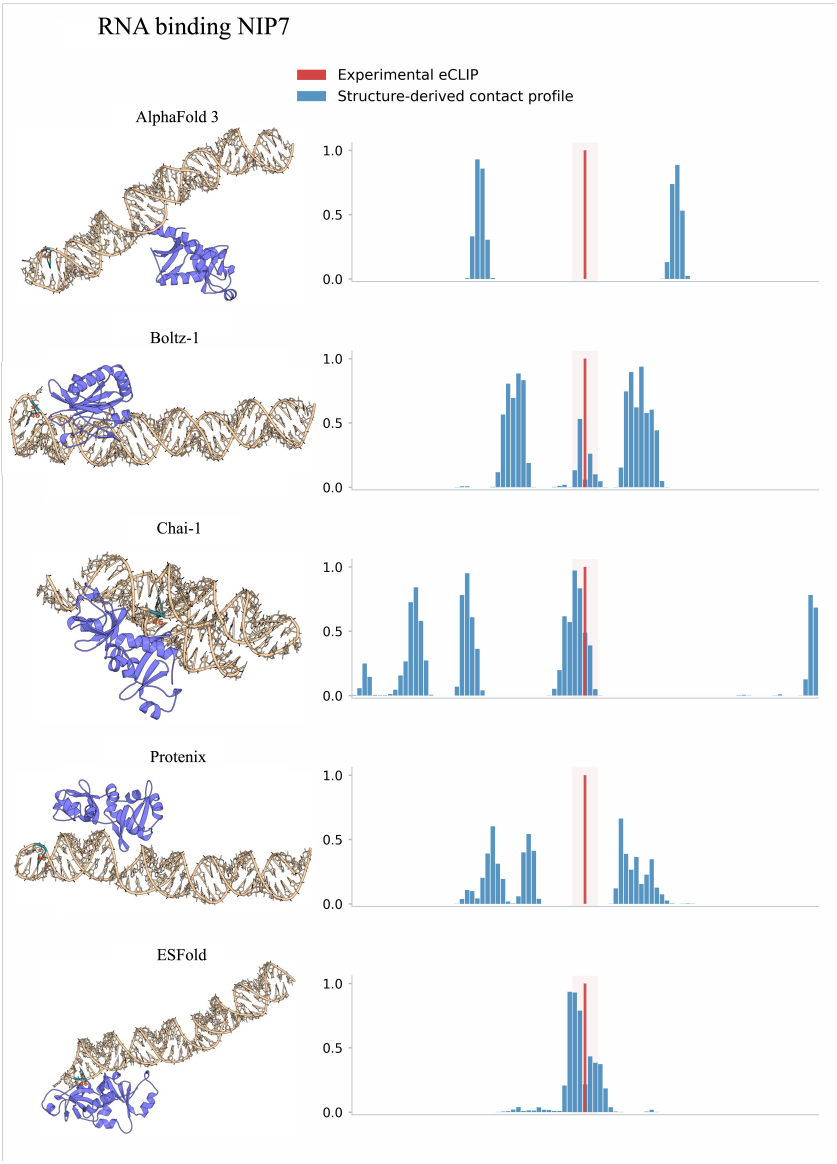


Figure 7: **NIP7: five-model comparison.** Each row shows a predicted protein–RNA complex (left) and the experimental eCLIP signal (red) alongside the structure-derived contact profile (blue; right). Rows compare AlphaFold 3, Boltz-1, Chai-1, Protenix, and ESFold. Shading marks the experimental peak region.

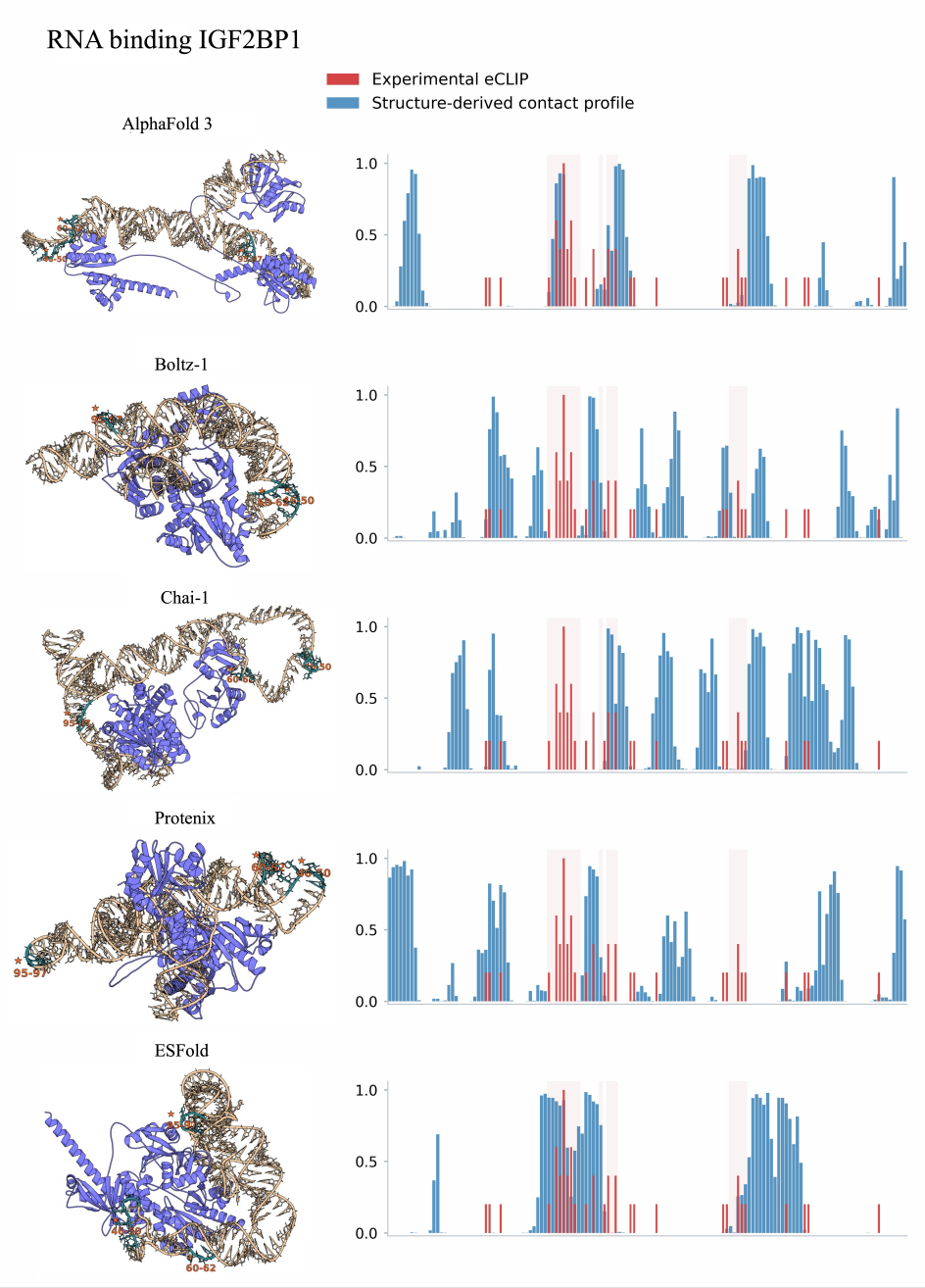


Figure 8: **IGF2BP1: five-model comparison.** The 140-nt RNA has its strongest signal at U48 (region 46–50), with secondary regions at 60–62 and 95–97. Layout and colors follow Figure 7; bars show unsmoothed positional values.