bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

A high-resolution protein-RNA binding atlas for structure-aware interaction prediction and post-transcriptional regulatory modeling

Mehran Karimzadeh[*,2,3,4], Lisa Fish[*,1,2,3,4], Heather Karner[*,2,3,4], Johannes Linder[*,5,6], Ali Emadi[*,1], Aldo Hernandez-Corchado[8,9], Fatima Zohra Smaili[8,9], Khashayar Ghavami[2], Edna S. Stewart[10], David Kelley[†,7], Bo Wang[†,11], Hamed Shateri Najafabadi[†,8,9], Manuel D. Leonetti[†,10], Hani Goodarzi[†,1,2,3,4]

[1]Arc Institute, Palo Alto, CA, USA;
[2]Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, CA, USA;
[3]Department of Urology, University of California San Francisco, San Francisco, CA, USA;
[4]Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, CA, USA;
[5]Institute of Computational Biology, Helmholtz Munich, Neuherberg, Germany;
[6]School of Medicine and Health, Technical University of Munich, Munich, Germany;
[7]Calico Life Sciences LLC, South San Francisco, CA, USA;
[8]Department of Human Genetics, McGill University, Montreal, Quebec, Canada;
[9]Victor P. Dahdaleh Institute of Genomic Medicine, Montreal, Quebec, Canada;
[10]Biohub, San Francisco, CA, USA;
[11]Vector Institute, Toronto, Ontario, Canada

Abstract

RNA-binding proteins (RBPs) control post-transcriptional gene expression by regulating RNA processing, localization, stability, and translation. However, our understanding of RNA regulation remains limited by the resolution, scale, and uniformity of available RBP binding maps. Here we introduce eclair CLIP (ecCLIP), a high-throughput CLIP-seq platform that combines endogenous protein tagging for uniform affinity capture with ligation-free library preparation to generate high-resolution RNA-protein binding profiles. We applied ecCLIP to 67 endogenously tagged RBPs in HEK293T cells, generating a high-quality transcriptome-scale atlas that recovered known binding preferences and captured post-transcriptional regulons linked to RNA half-life. To test the utility of the atlas, we incorporated ecCLIP into Borzuki, a Borzoi-derived sequence-to-function model trained on complementary RNA degradation, expression, CLIP, and Perturb-seq data. Inclusion of ecCLIP data improved RNA half-life prediction and assessment of 3′ UTR variant effects on RNA stability, demonstrating that our atlas helps capture RNA dynamics. To directly model the cis-regulatory grammar underlying these RBP-RNA interactions, we developed Pythia, a structure-aware neural network architecture that incorporates context-free grammar-inspired fixed convolutions to explicitly encode RNA base-pairing features. Pythia predicted RBP binding from sequence and structural context and decomposed predictions into sequence- and structure-dependent components. We tested this decomposition for five RBPs using massively parallel reporter CLIP (MPRC) binding assays in cells, using synthetic RNA libraries containing mutations designed to preserve or disrupt predicted RNA structures across nominated binding sites. We used the same MPRC strategy to confirm allele-specific binding effects for variants prioritized by Pythia. Together, ecCLIP and Pythia provide a resource for structure-aware decoding of RNA-protein interaction grammar and interpretation of post-transcriptional regulatory variation.

[*]These authors contributed equally to this work.
[†]Corresponding authors: David Kelley (drk@calicolabs.com), Bo Wang (bowang@vectorinstitute.ai), Hamed Shateri Najafabadi (hamed.najafabadi@mcgill.ca), Manuel Leonetti (manuel.leonetti@biohub.org), Hani Goodarzi (hani.goodarzi@arcinstitute.org).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

1. Introduction

RNA-binding proteins (RBPs) regulate gene expression after transcription by controlling RNA processing, localization, stability, and translation. These interactions define post-transcriptional regulatory programs that determine when, where, and how transcripts are expressed. Disruption of RBP abundance, binding specificity, or cis-regulatory RNA elements that they recognize contributes to diverse human diseases (Gerstberger et al., 2014; Hentze et al., 2018; Gebauer et al., 2021). Understanding and predicting these mechanisms requires transcriptome-scale maps of RBP binding and models that connect binding events to RNA fate.

Crosslinking immunoprecipitation followed by sequencing (CLIP-seq) is a central approach for mapping RBP-RNA interactions in cells (Ule et al., 2003; Licatalosi et al., 2008; König et al., 2010; Hafner et al., 2010; Van Nostrand et al., 2016, 2020). Existing CLIP resources have provided valuable maps of RBP occupancy, but high-quality profiling across many RBPs remains technically demanding. Antibody-specific immunoprecipitation conditions, radioactive visualization, ligation-dependent library preparation, and protocol heterogeneity limit throughput, consistency, and recovery. These limitations are especially consequential for sequence-to-function modeling, where each experimental track must be sufficiently resolved, uniform, and biologically informative to improve generalization.

To address this need, we developed eclair CLIP (ecCLIP), a high-throughput CLIP-seq platform for systematic RBP profiling that combines endogenous split-fluorescent-protein tagging for uniform affinity capture with infrared visualization and ligation-free library preparation. We applied ecCLIP to 67 endogenously tagged RBPs in HEK293T cells, generating a high-resolution transcriptome-scale atlas that captures RBP binding specificity and regulatory signal linked to RNA half-life. We reasoned that a high quality RBP binding atlas should measurably improve our ability to decode post-transcriptional regulatory grammars beyond what existing public binding and RNA processing datasets allow. We built Borzuki to test this directly. Borzuki is a Borzoi-derived (Linder et al., 2025) sequence-to-function model for RNA expression and dynamics that extends the sequence-to-stability framing introduced by Saluki (Agarwal and Kelley, 2022). It takes ecCLIP profiles for stability-modulating RBPs as its central new input, alongside complementary and publicly available ENCODE RBP eCLIP (Van Nostrand et al., 2020), RBP Perturb-seq (Kowalski et al., 2024), and RNA decay data. In matched ablations, adding ecCLIP data improved prediction of RNA half-life and of MPRA-measured 3′UTR variant effects on RNA stability, showing that the ecCLIP atlas carries regulatory information about RNA fate that existing data does not.

Beyond their utility for predicting RNA regulation, nucleotide-resolution RNA interaction profiles also provide an opportunity to learn the underlying *cis*-regulatory grammars. RBP-RNA binding is shaped not only by primary sequence but also by RNA base-pairing, structural accessibility, and higher-order structural context (Eddy and Durbin, 1994; Dowell and Eddy, 2004; Lange et al., 2012; Goodarzi et al., 2014; Kazan et al., 2010). Earlier work has shown that structural RNA regulatory elements can encode post-transcriptional programs (Goodarzi et al., 2014, 2012). However, most deep learning models of RBP binding represent RNA sequence explicitly while treating structure implicitly, if at all.

To address this we developed Pythia, a structure-aware neural network architecture that explicitly models the cis-regulatory grammar of RBP-RNA recognition. Pythia encodes RNA base-pairing features through context-free grammar-inspired fixed convolutions while preserving a dedicated sequence arm for primary sequence recognition, enabling both prediction of RBP binding and decomposition of individual binding sites into sequence- and structure-dependent components. To experimentally validate these predictions, we developed massively parallel reporter CLIP (MPRC), the first approach enabling pooled, quantitative measurement of sequence- and structure-dependent binding effects alongside allele-specific effects for prioritized regulatory variants.

This study is organized in three parts. First, we describe the generation of the ecCLIP atlas and characterize its quality and resolution. Second, we evaluate its functional utility through Borzuki, demonstrating that ecCLIP profiles improve prediction of RNA half-life and 3′ UTR variant effects beyond what existing data afford. Third, we deploy Pythia to dissect the sequence and structural determinants of RBP recognition and interpret the binding consequences of regulatory variants. Together, these results establish an integrated framework in which high-resolution RBP binding maps sharpen models of RNA fate and enable mechanistic interpretation of RNA-protein interaction grammar at nucleotide resolution.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

2. Results

2.1. Generating a high-resolution RBP binding atlas with eclair CLIP

CLIP-seq remains a central technique for transcriptome-wide mapping of RBP-RNA interactions, but applying CLIP systematically across large RBP panels is limited by several practical and technical bottlenecks. Conventional workflows typically require protein-specific antibodies and immunoprecipitation conditions, and differences in antibody availability, specificity, affinity, and lot-to-lot quality make atlas-level comparisons across RBPs difficult. Other steps, especially ligation-dependent library preparation, can reduce recovery and introduce bias. CLIP signal is also influenced by crosslinking efficiency, which can complicate interpretation of broad peaks and lower-resolution binding profiles (Park et al., 2018). Subsequent methods have addressed subsets of these limitations, including size-matched input controls to account for background abundance (Van Nostrand et al., 2016) and photoactivatable ribonucleoside labeling to identify binding sites through diagnostic mutational signatures (Ascano et al., 2012). However, a uniform and efficient workflow is still needed to generate high-quality, high-resolution RBP binding maps at scale.

To enable systematic profiling of RBP binding, we developed eclair CLIP (ecCLIP), an efficient and streamlined CLIP-seq workflow optimized for high-throughput, high-resolution mapping of ribonucleoprotein complexes (**Fig. 1a**). ecCLIP combines endogenous split mNeonGreen (mNG) tagging for uniform affinity capture, infrared labeling in place of radioisotope-based visualization, and ligation-free, template-switch-based library preparation to improve recovery and reduce ligation-associated bias (Feng et al., 2017; Zarnegar et al., 2016). We evaluated recovery of mNG-tagged RBPs, optimized RNA labeling, and compared RNA isolation methods (**Supp. Fig. S1a–e**).

We selected RBPs primarily from the OpenCell endogenous tagging resource (Cho et al., 2022) and supplemented this set with additional proteins to represent major RBP families. The final panel included RBPs containing a wide variety of RNA-binding domains, including RNA recognition motifs, zinc finger motifs, KH-domains, DEAD-box helicases, YTH domains and double-stranded RNA-binding domains (**Supp. Fig. S1f**). We also designed the atlas to complement the ENCODE eCLIP resource in two ways: by profiling RBP binding in HEK293T cells, rather than the K562 and HepG2 cell contexts used for most ENCODE eCLIP datasets, and by including 40 RBPs not represented in ENCODE. In total, we profiled 67 endogenously tagged RBPs and defined their transcriptome-wide binding landscapes in HEK293T cells.

Across all 67 RBPs, we identified 9,097,672 ecCLIP peaks (**Supp. Fig. S1g**). Peak counts varied substantially across proteins, which can be attributed to a variety of factors, including RBP-specific differences in expression, crosslinking efficiency, RNA occupancy, and biochemical recovery (**Fig. 1b**, left). The resulting binding profiles also showed distinct regional preferences, with examples of RBPs enriched in UTRs (ELAVL1), coding sequences (YBX1), or intronic regions (HNRNPC) (**Fig. 1b**, middle). Most peaks were unique to individual RBPs, whereas a relative minority was shared between two or more RBPs (**Fig. 1b**, right), consistent with the atlas capturing largely protein-specific binding programs.

We compared pairwise distances between ESM2 protein embeddings and ecCLIP-derived binding preferences to assess correspondence between these representations (Ray et al., 2013). Here, binding preferences were represented as dense 5-mer enrichment profiles of ecCLIP peaks relative to flanking background, a compact description of each RBP's RNA sequence specificity, and proteins were represented by ESM2 embeddings of their annotated RNA-binding domains. Distances in the two representations showed limited but non-random correspondence (Mantel $\rho = 0.083$, modified RV coefficient $= 0.110$, distance correlation $= 0.605$; each exceeding its permutation null, $P \leq 0.002$; **Fig. 1c**), and clustering the 67 proteins by domain embeddings and by ecCLIP-derived sequence preferences produced largely different groupings (**Fig. 1d**). Whole-protein embeddings gave a similar, slightly stronger correspondence (Mantel $\rho = 0.117$, modified RV $= 0.121$, distance correlation $= 0.641$; $P = 0.0001$; **Supp. Fig. S1i,j**). This comparison does not establish the limits of motif prediction from protein sequence: RNA-binding-domain similarity and models trained specifically to predict RNA preferences address that question more directly. Rather, it shows that distances between general-purpose ESM2 embeddings, which are not trained on RNA-binding data, only weakly reflect the in vivo sequence preferences measured by ecCLIP, and therefore should not be used on their own as a proxy for shared binding specificity (Ray et al., 2013; Sasse et al., 2026). These results establish ecCLIP as a high-resolution atlas for empirically measuring RBP binding programs across a broad panel of proteins.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

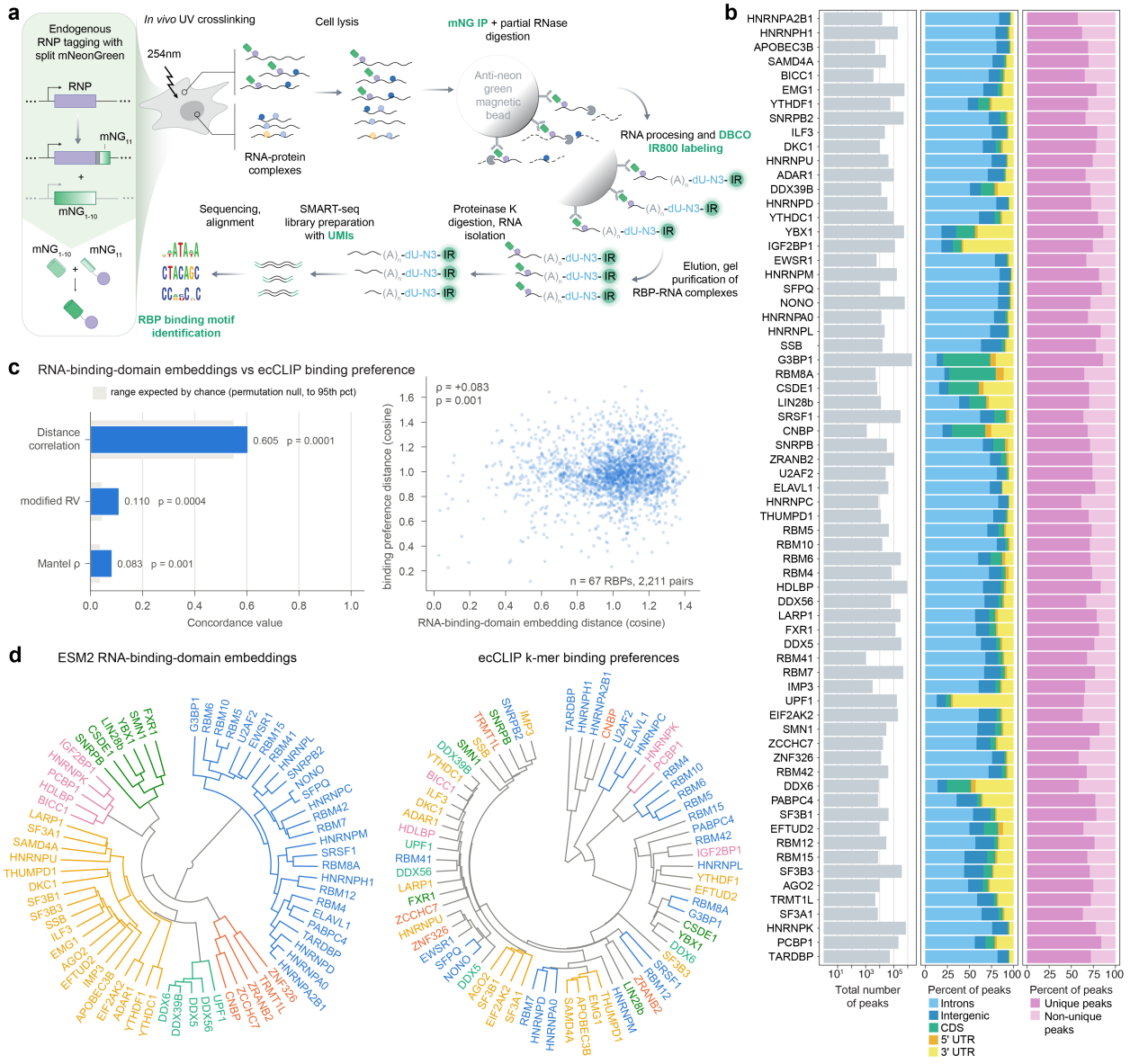


Figure 1 | **Efficient, high-resolution profiling of RBP-RNA interactions with eclair CLIP. (a)** Schematic of the ecCLIP workflow. Endogenous split mNeonGreen tagging provides a uniform affinity-capture strategy across RBPs. After UV crosslinking, cell lysis, partial RNase digestion, anti-mNG immunoprecipitation, infrared RNP labeling, gel separation, proteinase K digestion, RNA isolation, and ligation-free template-switch library preparation with UMIs, sequencing reads are aligned to identify transcriptome-wide RBP binding profiles. **(b)** Barplots show the number of peaks detected per RBP (left), the fraction of peaks overlapping UTRs, coding sequence, intergenic regions, and introns (middle), and the fraction of peaks that are unique or shared across RBPs (right). **(c)** Correspondence between ESM2 RNA-binding-domain embeddings and ecCLIP-derived binding preferences for 67 RBPs. Left, concordance statistics (distance correlation, modified RV coefficient, and Mantel ρ); grey bands show the range expected by chance (95th percentile of 9,999 permutations of RBP labels). Right, pairwise cosine distances between RBP domain embeddings versus between binding-preference profiles (2,211 RBP pairs). **(d)** Circular dendrograms of the 67 RBPs clustered (Ward, k = 6) by ESM2 RNA-binding-domain embeddings (left) and by ecCLIP-derived 5-mer binding preferences (right). Both trees are colored by the same grouping, defined from the protein-embedding clusters, so scattered colors in the right tree show where the two representations disagree.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

2.2. Establishing ecCLIP data quality and utility for modeling RNA dynamics

We next evaluated the resolution and signal quality of the ecCLIP atlas using two complementary metrics. First, we quantified the fraction of reads in peaks (FRiP) across window sizes centered on peak summits, an analysis that measures how tightly the sequencing signal is concentrated around inferred binding sites. For nine RBPs (TARDBP, HNRNPK, HNRNPC, PCBP1, EWSR1, HNRNPL, IGF2BP1, LIN28B, and ZRANB2), peak-centered FRiP curves and aggregate read profiles showed concentrated ecCLIP signal around inferred binding sites (**Fig. 2a**; **Supp. Fig. S2a**). Published eCLIP profiles are included as a reference rather than as a controlled comparison of assay performance, because they differ in cellular context, library preparation, and peak definition (Van Nostrand et al., 2020). Second, to test whether ecCLIP peaks capture biologically meaningful RBP binding preferences, we compared sequence preferences inferred from cell-derived ecCLIP peaks to orthogonal in vitro measurements from RNA Bind-n-Seq (RBNS; Lambert et al., 2014). For TARDBP (TDP-43), ecCLIP recovered GAAUG-containing 5-mers consistent with its RNAcompete consensus motif, providing an example of recognizable sequence specificity (**Fig. 2b**; Ray et al., 2013). We then assessed concordance with RBNS across up to nine RBPs with available eCLIP, ecCLIP, and RBNS measurements. For each RBP, we computed the absolute Pearson correlation between CLIP-derived k-mer enrichment and RBNS affinity at the highest reported RBNS protein concentration (1300 nM), using $k = 4$, 5, and 6 and peak-centered windows of 10 or 20 nucleotides. ecCLIP-derived k-mer enrichments agreed with RBNS preferences across these parameter settings; the corresponding eCLIP measurements provide a contextual reference (**Fig. 2c**). This result was robust when RBNS affinity was evaluated at the optimal protein concentration for each RBP rather than at a fixed concentration (**Supp. Fig. S2b**). Together, these results show that ecCLIP produces binding maps that are both high resolution and informative of RBP sequence specificity.

For this evaluation of data utility, we developed Borzuki as an expansion of Borzoi focused on post-transcriptional regulation ("Borzuki" being a combination of Borzoi and Saluki; Agarwal and Kelley, 2022). Borzoi established that long-range RNA-seq coverage can be predicted directly from genomic sequence, providing a general framework for learning regulatory programs from functional genomics tracks (Linder et al., 2025). However, RNA abundance also reflects RNA stability, RBP binding, and processing, suggesting that Borzoi-style models could be tuned for RNA dynamics by incorporating direct measurements of post-transcriptional control.

We collected RNA-seq degradation time-course training tracks from human and mouse systems (Sasse et al., 2026) and paired these data with ecCLIP profiles, in addition to broader ENCODE eCLIP resources for half-life-associated RBPs (Van Nostrand et al., 2020) and Perturb-seq measurements for 3′-end processing regulators (Linder et al., 2025). We combined these with ENCODE DNase-seq/RNA-seq and FANTOM5 CAGE data for matching cell types (The ENCODE Project Consortium, 2012; Noguchi et al., 2017). We included the transcriptional readouts because steady-state RNA abundance is set jointly by synthesis and decay, and the two are confounded in expression data alone. A model that learns initiation rate from sequence can account for the transcriptional component of expression and attribute the remaining variation to post-transcriptional control, which sharpens the estimate of how much of RNA fate is set by stability. The resulting 40-million-parameter transformer U-net predicts RBP binding, RNA stability, and steady-state expression across 452 human and 160 mouse tracks from 393-kb input sequences and was trained as an ensemble of six replicate models (**Fig. 2d**). This multi-task design allows Borzuki to score sequence changes across linked post-transcriptional readouts, including expression, degradation, and RBP binding, rather than requiring separate specialized models for each output.

Borzuki accurately predicted held-out RNA-seq and CLIP-seq (i.e. both eCLIP and ecCLIP) coverage profiles. In ablations evaluated against HEK293T 4sU-based half-life measurements, adding ecCLIP profiles to the training data improved the prediction of RNA-seq degradation time-course tracks, and adding ENCODE eCLIP and 3′-regulator Perturb-seq measurements further improved performance (**Fig. 2e**). For the held-out gene ZNF611 in a K562 actinomycin D transcriptional shutoff experiment, in which inhibition of transcription allows RNA loss to be followed over time, predicted coverage recapitulated both gene-level log-coverage and position-specific signal across time points (**Fig. 2f**; Melé et al., 2017). For both ecCLIP and ENCODE eCLIP tracks, predicted and measured coverage agreed at gene-level and within-gene bin-level resolution (**Fig. 2g**; **Supp. Fig. S2c**), with ELAVL1 examples shown in **Supplementary Fig. S2e,f**. These analyses establish that the model can learn post-transcriptional regulatory tracks from sequence across held-out loci.

In a complementary comparison across model sizes and training-data compositions, models trained with

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

CLIP-seq and Perturb-seq data consistently outperformed matched models trained without these modalities when predicting RNA-seq coverage, with the strongest performance observed for the 40-million-parameter model trained on the full dataset (**Fig. 2h**). Using K562 actinomycin D half-life tracks, Borzuki-predicted mRNA half-lives correlated with experimental measurements (Spearman R = 0.745) and with Saluki's cell type-agnostic half-life predictions (Spearman R = 0.728; **Supp. Fig. S2d**; Agarwal and Kelley, 2022).

To extend the data-utility assessment to regulatory variation, we asked whether the learned features generalize to sequence perturbations not used for model training. Across fine-mapped 3′ UTR eQTLs, Borzuki distinguished high-confidence causal variants from matched negatives more effectively than Saluki (AUPRC = 0.694 for Borzuki vs 0.626 for Saluki), with the strongest performance from the 40-million-parameter model trained with the full CLIP- and Perturb-seq-expanded data set (**Fig. 2i**; Kerimov et al., 2021; Lonsdale et al., 2013). Borzuki also performed well on causal Mendelian 3′ UTR variants from TraitGym, particularly when variants were scored using the gene context with the largest predicted effect rather than only the annotated target gene (AUPRC = 0.545 for Borzuki vs 0.422 for Saluki) (**Supp. Fig. S2g**; Benegas et al., 2025). We further assessed variant scoring capabilities using massively parallel reporter assay (MPRA) data measuring 3′ UTR variant effects on RNA stability (**Supp. Fig. S2h**; Siegel et al., 2022). To score variants, we inserted reference and variant oligos in native 3′ UTRs and predicted differences in exon-aggregated log-coverage across RNA-seq and CLIP-seq tracks using the Borzuki ensemble (Methods). Borzuki was less accurate than Saluki on this task, but averaging the two models' scores outperformed either model alone, suggesting that they capture complementary aspects of variant effects (**Supp. Fig. S2i**). Importantly, model ablations trained without ecCLIP data performed worse at half-life variant effect prediction (**Supp. Fig. S2j**). The ablation directly supports the contribution of ecCLIP to RNA-stability variant prediction; the eQTL and Mendelian comparisons describe the performance of the combined model on those specific benchmarks.

Together, these analyses establish ecCLIP as a high-quality, high-resolution resource for RBP-RNA interaction mapping. Our dataset shows concentrated signal at inferred binding sites, recovers independently measured RBP sequence preferences, and provides information that improves sequence-to-function modeling of RNA dynamics. The ablations demonstrate that the knowledge of RBP binding sites provided by ecCLIP adds predictive information about RNA half-life and RNA-stability variant effects.

2.3. Modeling RBP recognition and regulatory variation with Pythia

Having established the quality and utility of ecCLIP, we next used its nucleotide-resolution maps to learn the sequence and structural grammar that underlie RBP-RNA interactions. This problem is more complex than DNA motif discovery because RBP binding is shaped not only by primary RNA sequence, but also by base-pairing, accessibility, and local folding context. Linear motif models, including k-mers, regular expressions, and position weight matrices, provide a natural representation for sequence preferences. In contrast, RNA structural elements couple distant nucleotide positions, making context-free grammars (CFGs) a natural formalism for representing paired sequence constraints. This idea has a long history in RNA modeling, including stochastic CFG approaches and RNApromo for discovering structural RNA motifs (Searls, 2002; Rabani et al., 2008), and was central to our earlier work where structural cis-regulatory elements were represented as families of CFGs and scored for association with post-transcriptional regulatory programs (Goodarzi et al., 2014, 2012; Fish et al., 2021).

We developed Pythia to bring this CFG-inspired representation into a neural network architecture for RNA-protein interaction modeling. Pythia uses two parallel arms: a sequence arm that processes one-hot encoded RNA sequence with trainable convolutional layers, and a structure arm that first passes the same sequence through a fixed dilated convolutional layer (**Fig. 3a**). The fixed layer enumerates candidate Watson-Crick and wobble base-pairing rules across a range of distances and bulge configurations. Conceptually, each kernel corresponds to a CFG production rule: a paired nucleotide constraint at a defined separation, with intervening positions left unconstrained. Because these kernels are frozen rather than learned, they provide an explicit structural inductive bias without requiring an external folding algorithm, experimentally measured structure, or task-specific structure labels. Downstream trainable layers then learn which combinations of these candidate structural rules are informative for a given supervised task.

We first evaluated whether this representation accurately captured RNA structural information independently of RBP binding. When sequences from canonical RFAM structural families were passed through the fixed dilated convolutional layer, principal component analysis separated families with distinct structural organiza-

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

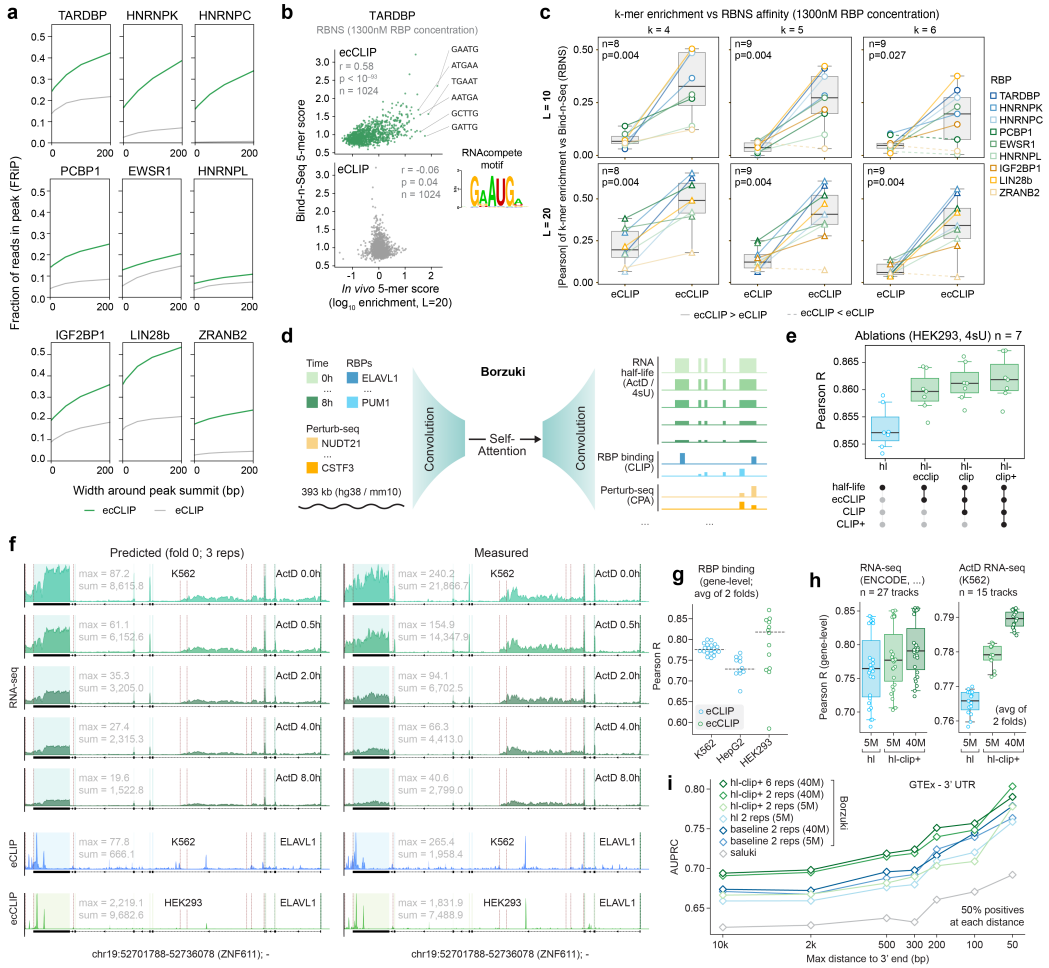


Figure 2 | **ecCLIP quality metrics and genome-scale modeling of RNA dynamics. (a)** Comparison of peak resolution in ecCLIP and eCLIP. Fraction of reads in peaks (FRiP) is shown as a function of window width around each peak summit for nine RBPs with comparable datasets (TARDBP, HNRNPK, HNRNPC, PCBP1, EWSR1, HNRNPL, IGF2BP1, LIN28B, and ZRANB2). **(b)** Example comparison between CLIP-derived and RBNS-derived sequence preferences for TARDBP. Scatterplots compare 5-mer enrichment scores (log ratio of peak vs flanking background, 20-nt peak-centered windows) from eCLIP or ecCLIP peaks to RBNS 5-mer affinity scores at 1300 nM. The top enriched 5-mers are labeled, and the TDP-43 RNAcompete motif is shown for reference. **(c)** Systematic comparison of k-mer concordance between eCLIP/ecCLIP and RBNS across up to nine RBPs. The absolute Pearson correlation between CLIP-derived k-mer enrichment and RBNS affinity at 1300 nM is shown for k = 4, 5, and 6, using peak-centered windows of 10 or 20 nucleotides. Lines connect matched RBPs between eCLIP and ecCLIP; solid lines indicate stronger concordance for ecCLIP and dashed lines indicate stronger concordance for eCLIP. P-values correspond to paired comparisons between eCLIP and ecCLIP. **(d)** Borzuki training data and architecture overview. Borzuki is trained to jointly predict sequencing coverage from RNA-seq degradation time courses, RBP CLIP-seq, and 3′-processing regulator Perturb-seq, together with ENCODE RNA-seq/DNase-seq and FANTOM5 CAGE data. The 40-million-parameter transformer U-net predicts coverage profiles for 452 human and 160 mouse tracks from 393-kb input sequences and is trained as an ensemble of six replicate models. **(e)** Borzoi-style model ablation for HEK293T RNA half-life prediction. Boxplots summarize Pearson correlations (n = 7 RNA-seq tracks) for models trained with half-life tracks ('hl'), half-life tracks plus ecCLIP ('hl_ecclip'), half-life tracks plus ecCLIP and eCLIP ('hl_clip'), or half-life tracks plus both CLIP resources and 3′-end processing Perturb-seq ('hl_clip+'). Reported metrics were averaged across 2 model replicates. **(f)** Predicted and measured coverage for the held-out gene ZNF611 using a three-model ensemble. Top: RNA-seq coverage across K562 actinomycin D time points (0-8 h). Bottom: ELAVL1 binding from two independent experiments, ENCODE eCLIP in K562 and ecCLIP in HEK293T, shown as representative CLIP-seq tracks predicted by the same model. **(g)** Test-set Pearson correlation for CLIP-seq gene-level prediction across held-out tracks. **(h)** Model size and training-data composition comparison. Pearson correlation metrics are shown for gene-level log-coverage prediction across two model sizes (5 million and 40 million parameters) and two training-data compositions, with or without CLIP-seq and Perturb-seq modalities. **(i)** Comparison between Saluki and Borzuki at distinguishing fine-mapped 3′ UTR eQTLs (PIP > 0.7) from matched negative SNPs (Kerimov et al., 2021; Lonsdale et al., 2013). Borzuki model size, ensemble size, and training-data composition are varied.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

tion (**Fig. 3b**; Kalvari et al., 2021). We also evaluated synthetic sequences containing an injected CFG-defined structural element and matched controls lacking the element, without fixed primary sequence identity. Their embedding distributions overlapped in the UMAP projection (**Supp. Fig. S3a**), which alone does not establish classification performance. As a more challenging structural classification task, we tested whether Pythia could distinguish preserved from disrupted pseudoknots generated by in silico mutation. Pythia achieved high performance on this task (AUROC = 0.95, AUPRC = 0.87), outperforming supervised sequence models including LSTM, DeepBind, DeepSEA, and ResNet architectures (**Supp. Fig. S3b,c**; Alipanahi et al., 2015; Zhou and Troyanskaya, 2015; He et al., 2016).

To benchmark the architecture more broadly, we evaluated Pythia on four RNA structural tasks of the BEACON suite: contact map prediction (CMP), distance map prediction (DMP), structural score imputation (SSI), and secondary structure prediction (SSP; Ren et al., 2024). The 17 BEACON baselines comprise conventional CNN, ResNet, and LSTM architectures; pretrained RNA language models, including RNA-FM, RNABERT, RNA-MSM, SpliceBERT, 3UTRBERT, and UTR-LM variants; and BEACON-B/BEACON-B512 (Methods). Against this panel, Pythia achieved the top overall performance on the two contact and secondary-structure prediction tasks: CMP (Top-L precision 66.6%) and SSP (F1 71.5%), surpassing even the purpose-built BEACON-B baseline. For SSI (R^2 38.6%) and DMP (R^2 51.3%), Pythia exceeded the supervised CNN, ResNet, and LSTM baselines, while trailing the top pretrained RNA language models on DMP (**Fig. 3c**). Together, these benchmarks show that Pythia's fixed base-pairing prior captures RNA secondary-structure constraints, and motivated our application of Pythia to RBP binding sites from ecCLIP data to simultaneously capture the sequence and structural determinants of RBP-RNA interactions.

2.4. Pythia predicts ecCLIP binding sites from sequence and structural context

We next trained Pythia on the ecCLIP atlas to predict RBP binding sites. For each of the 67 RBPs, we trained a separate binary classifier to discriminate ecCLIP peaks from dinucleotide-shuffled versions of the same sequences, which serve as composition-matched negatives, with held-out chromosomes used for evaluation. Thus, RBP identity was not provided as a model input; instead, each model learned an RBP-specific binding function using the same sequence-plus-structure architecture. This design allowed us to test whether the fixed structural arm improved binding-site prediction while preserving a common framework for comparing sequence and structure contributions across RBPs. Per-RBP auROC and auPR values summarize variation in prediction performance across the atlas (**Supp. Fig. S3e**).

We compared Pythia with four supervised baselines (DeepBind, DeepSEA, GraphProt, and PRIESSTESS) and with supervised linear probes on frozen embeddings from six RNA language models (RiNALMo, RNA-FM, UTR-LM-TE_EL, UTR-LM-MRL, RNA-MSM, and RNAErnie). Across matched RBP binding tasks, Pythia outperformed the four baselines trained on sequence features and/or externally supplied structure features (**Fig. 3d**; Maticzka et al., 2014; Laverty et al., 2022). For HNRNPA0, a representative RBP with clear separation among models, Pythia achieved auROC = 0.93 and auPR = 0.72, compared with auROC = 0.86 and auPR = 0.54 for the next-best supervised comparator, DeepSEA (**Supp. Fig. S3f**). For the embedding comparisons, Pythia and the linear probes used matched labeled data splits, but the foundation models may have encountered test sequences during pretraining. Thus, labeled-data supervision is matched while pretraining exposure is not controlled. With that caveat, Pythia performed comparably to the strongest foundation-model embeddings while retaining a compact supervised architecture with an explicit structural representation (**Fig. 3e**; **Supp. Fig. S3g**). These results show that Pythia learns accurate RBP binding-site predictors from the ecCLIP atlas and that an explicit structural inductive bias can match or exceed learned sequence representations on this task. Pythia's downstream weights are learned with supervision and the fixed component is the initial base-pairing feature layer.

2.5. Pythia attribution and MPRC test structural determinants of RBP binding

Having trained Pythia to predict ecCLIP binding sites, we next asked whether the model could identify which binding events depend primarily on RNA sequence and which depend on structural context. Because Pythia contains separate sequence and structure arms, we can quantify the contribution of each arm to individual RBP binding predictions. For each RBP-specific model, we applied DeepLIFT separately to the sequence arm and the fixed-dilated-convolution structure arm, comparing the effect of perturbing each input representation on the predicted binding score (Shrikumar et al., 2019). The resulting sequence-to-structure attribution ratio

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

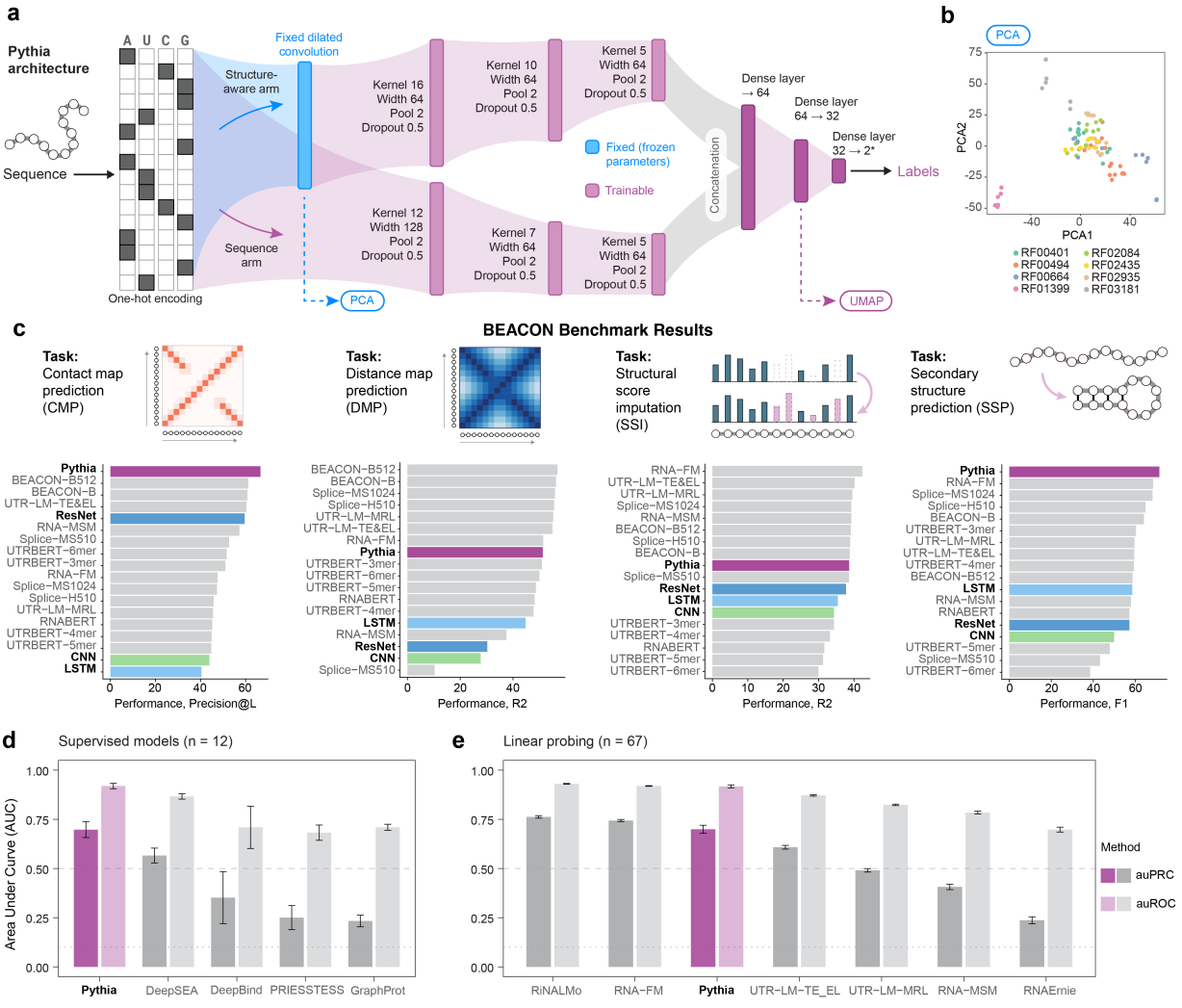


Figure 3 | **Pythia leverages the context-free grammar of RNA secondary structure.** **(a)** Model architecture. Pythia incorporates a fixed dilated convolutional layer with frozen parameters (blue) that captures hierarchical structural patterns in RNA. The resulting representation is processed alongside raw sequence input by additional convolutional layers and three downstream dense layers for task-specific prediction. **(b)** RNA structural family organization. The first principal components of Pythia's fixed convolutional layer embeddings cluster members of canonical RFAM RNA structural motif families, reflecting inherent structural grammar capture. **(c)** Benchmarking results across four BEACON RNA structure tasks; Contact Map Prediction (CMP), Distance Map Prediction (DMP), Structural Score imputation (SSI), and Secondary Structure prediction (SSP). Pythia achieves the top performance in CMP and SSP, while remaining highly competitive in DMP and SSI. **(d)** Binding-site prediction compared with four supervised baselines (DeepSEA, DeepBind, PRIESSTESS, and GraphProt) across the 12 RBPs shown. **(e)** Binding-site prediction compared with supervised linear probes on six frozen RNA foundation models (RiNALMo, RNA-FM, UTR-LM-TE_EL, UTR-LM-MRL, RNA-MSM, and RNAErnie; n=67 RBPs). For the binding tasks, supervised baselines and linear probes use the same labeled data splits as Pythia; foundation-model pretraining exposure is not controlled. For each model, opaque and transparent bars show auPRC and auROC, respectively.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

summarized the model's relative reliance on each arm at each binding site, with high ratios corresponding to sequence-dominant predictions and low ratios corresponding to structure-dominant predictions (**Fig. 4a**).

Across the ecCLIP atlas, DeepLIFT revealed substantial heterogeneity in the sequence and structure contributions to RBP binding. Positive-control sequences with fixed primary-sequence motifs showed strong sequence dominance, whereas pseudoknot-derived structural controls showed strong structure dominance, providing controls for the direction of the attribution scale (**Fig. 4a**). We assigned RBP-level sequence-dominant or structure-dominant labels from Pythia attributions, requiring an absolute median $\log_2$ sequence-to-structure ratio > 0.05 and a one-sample Wilcoxon test against zero with Benjamini–Hochberg FDR $< 10\%$ (Methods). Models that did not meet these criteria were labeled mixed; the site-level scores nevertheless form a continuum: some RBPs showed predominantly sequence-driven binding, others showed stronger structural dependence, and many exhibited site-specific variation. This heterogeneity indicates that structural dependence is not only an RBP-level property: individual sites bound by the same RBP can differ in the degree to which Pythia relies on primary sequence versus structural context.

To test whether these model-derived attributions correspond to experimentally measurable binding determinants, we developed massively parallel reporter CLIP (MPRC). For five RBPs spanning the sequence-structure continuum (EMG1, HNRNPC, PCBP1, HNRNPU, and SFPQ), we selected representative ecCLIP binding sites and designed synthetic reporter libraries containing the reference sequence together with targeted mutant alleles (**Fig. 4b,c**). For each site, the library included mutations designed to preserve the predicted RNA structure, mutations designed to disrupt the predicted structure, and multi-mutation negative controls. We introduced these libraries into HEK293T cells expressing the corresponding endogenously tagged RBP and measured relative RBP binding by reporter-targeted CLIP. This design allowed us to directly test whether sites predicted to depend on structural context were specifically sensitive to structure-disrupting mutations while retaining binding under structure-preserving mutations.

We quantified the MPRC output using UMI-deduplicated reporter counts, normalized IP and input libraries, and summarized binding for each construct as $\log_2$(IP/input). Because these reporters move endogenous binding sites outside their native transcript context, we first asked whether the reference reporter retained measurable binding in the CLIP assay. Of 30 structure-validation sites tested, 19 showed clear reference enrichment (Ref $\log_2$(IP/input) > 0.25), including 11 of 18 sites predicted to be structure-dependent and 8 of 12 sites not predicted to be structure-dependent. We restricted the rescue analysis to this reference-enriched subset and grouped reporters as reference, structure-preserving, structure-disrupting, or negative-control constructs. Read-depth distributions for the reference-enriched sites were comparable between predicted structure-dependent and not-structure-dependent classes, and rescue effects were not restricted to low-count sites (**Supp. Fig. S4a**). Because a nonsignificant difference from the reference can reflect limited power rather than true preservation of binding, we used effect-size rescue as the primary validation statistic. Specifically, we asked whether structure-preserving mutations remained closer to the reference $\log_2$(IP/input) than structure-disrupting mutations.

MPRC confirmed that Pythia-nominated structural features correspond to functional determinants of RBP binding. Among sites predicted by Pythia to be structure-dependent, structure-preserving reporters were closer to the reference than structure-disrupting reporters at all 11 reference-enriched sites, using the same number of mutations in both conditions (one-sided paired Wilcoxon $P = 4.9 \times 10^{-4}$; **Fig. 4c**). In contrast, sites not predicted to be structure-dependent showed no significant paired difference ($n = 8$, $P = 0.527$; **Fig. 4c**). We additionally tested whether structure-preserving and structure-disrupting constructs differed in IP/input enrichment using binomial logistic regression adjusted for replicate (Methods). Requiring a significant difference between these groups yielded the expected pattern at 10 of 10 predicted structure-dependent sites with valid model fits, compared with 4 of 8 sites not predicted to be structure-dependent (one-sided Fisher exact $P = 0.023$). At the continuous-score level, more structure-dependent Pythia scores were associated with stronger rescue by structure-preserving mutations (**Supp. Fig. S4b**), and additional representative validation sites across HNRNPC, PCBP1, and EMG1 illustrated the range of responses across sites (**Supp. Fig. S4c–f**). One nominated example was an intronic GALNT13 sequence bound by HNRNPC, in which the reference sequence and structure-preserving variants showed strong enrichment (1.15 and 0.90 mean $\log_2$(IP/input), respectively), whereas structure-disrupting and negative-control variants showed reduced binding (-0.20 and -0.62 mean $\log_2$(IP/input); **Fig. 4d**). This example is consistent with previous work showing that local RNA structural changes can expose HNRNPC binding motifs and modulate HNRNPC-RNA interactions (Liu et al., 2015). Predicted RNA structures illustrate possible changes in poly-U accessibility in this reporter, providing

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

structural context for the binding measurements (**Supp. Fig. S4g**).

These results support the use of Pythia's structure-arm attributions to nominate binding sites that are sensitive to structure-disrupting mutations in the reporter context. More broadly, they establish that structure-aware interaction models can nominate specific RNA structural features whose disruption alters binding while compensatory or structure-preserving sequence changes retain activity.

2.6. Pythia models predict allele-specific effects of trait-associated variants on RBP binding

Having shown that RBP-specific Pythia models can nominate structural determinants of RBP binding, we next asked whether the same model collection could predict the effects of naturally occurring genetic variants on RBP binding. This is a central use case for structure-aware RBP binding models: many disease- and trait-associated variants fall in transcribed noncoding regions, but their effects on post-transcriptional regulation are difficult to interpret from sequence context alone. We therefore focused on phenotype-annotated variants located within ecCLIP binding sites for five RBPs assayed in the MPRC validation framework (EMG1, HNRNPC, HNRNPU, PCBP1, and SFPQ). For each RBP-variant pair, we used the corresponding RBP-specific Pythia model to score the reference and alternate alleles and prioritized variants spanning predicted losses, gains, and minimal effects on binding (**Fig. 5a,b**).

We experimentally tested 88 RBP-variant pairs using the same pooled reporter CLIP strategy (i.e. MPRC). Reference and alternate 50-nt sequences were synthesized into a pooled library, introduced into HEK293T cells expressing the corresponding endogenously tagged RBP, and quantified by MPRC. We then compared allele-specific RBP enrichment between each matched reference and alternate sequence testing whether the alternate allele changes RBP enrichment relative to its matched reference sequence while controlling for input abundance and replicate effects.

Through MPRC, we identified allele-specific RBP binding effects across all five RBPs. Using a global FDR threshold of 10% and an absolute effect-size threshold of 0.25 $\log_2$ units, 24 of 88 tested RBP-variant pairs showed significant allele-specific binding (**Fig. 5c**). Variant effects predicted by the RBP-specific Pythia models were positively correlated with experimentally measured effects across the full validation set (Pearson r = 0.33, P = 0.0019; Spearman ρ = 0.37, P = 4.6×10^{-4}; **Fig. 5d**). Directional concordance was also above chance with 53 of 86 non-zero predicted variants showing concordant direction (one-sided binomial P = 0.020). Performance varied across RBPs, with the strongest concordance for PCBP1 (Pearson r = 0.51, P = 0.021; Spearman ρ = 0.80, P = 1.9×10^{-5}) and EMG1 (Pearson r = 0.77, P = 0.025; Spearman ρ = 0.90, P = 0.002), and weaker correlations for HNRNPC, HNRNPU, and SFPQ (**Fig. 5e**). This RBP-level heterogeneity reinforces the need to pair model-based variant prioritization with experimental calibration in the relevant assay system. These measurements evaluate Pythia's variant predictions in the selected reporter set; the binding-site classification benchmarks in **Fig. 3** do not establish comparative variant-effect performance for the other models. Observed effects stratified by predicted direction and examples with large predicted effects are shown in **Supplementary Fig. S5a,b**.

To highlight a few examples, we applied a strict concordance filter requiring |predicted effect| > 0.1 $\log_2$ units, significant allele-specific binding, and matched predicted and observed direction (**Supp. Fig. S5b**). This highlighted a small set of functional candidates with trait-relevant annotations. Two intronic variants were predicted and validated to reduce HNRNPC binding: rs9323624, whose locus is annotated to non-steroidal anti-inflammatory drug-induced acute urticaria and angioedema, and rs56302494, annotated to heel bone mineral density. Because HNRNPC binding can influence RNA processing, these variants nominate local RNA-processing hypotheses for follow-up in native transcript and disease-relevant cellular contexts. The same strict filter also identified a gain of HNRNPU binding at the BCKDHB-locus variant rs9361614, annotated to waist circumference adjusted for body mass index, and loss of PCBP1 binding at rs11668313, annotated to low-density lipoprotein cholesterol levels. These links are speculative and should not be interpreted as causal disease mechanisms; rather, they illustrate how allele-specific RBP binding measurements can generate testable hypotheses for mechanistic dissection of trait-associated loci.

These results show that an ecCLIP-trained collection of RBP-specific, structure-aware models can prioritize naturally occurring variants that alter RBP binding in a pooled experimental assay. The validation should be interpreted specifically as evidence for allele-specific binding effects in the HEK293T cellular context, rather than as direct evidence of disease mechanism in the relevant tissue or cell type. Nonetheless, by linking trait-

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

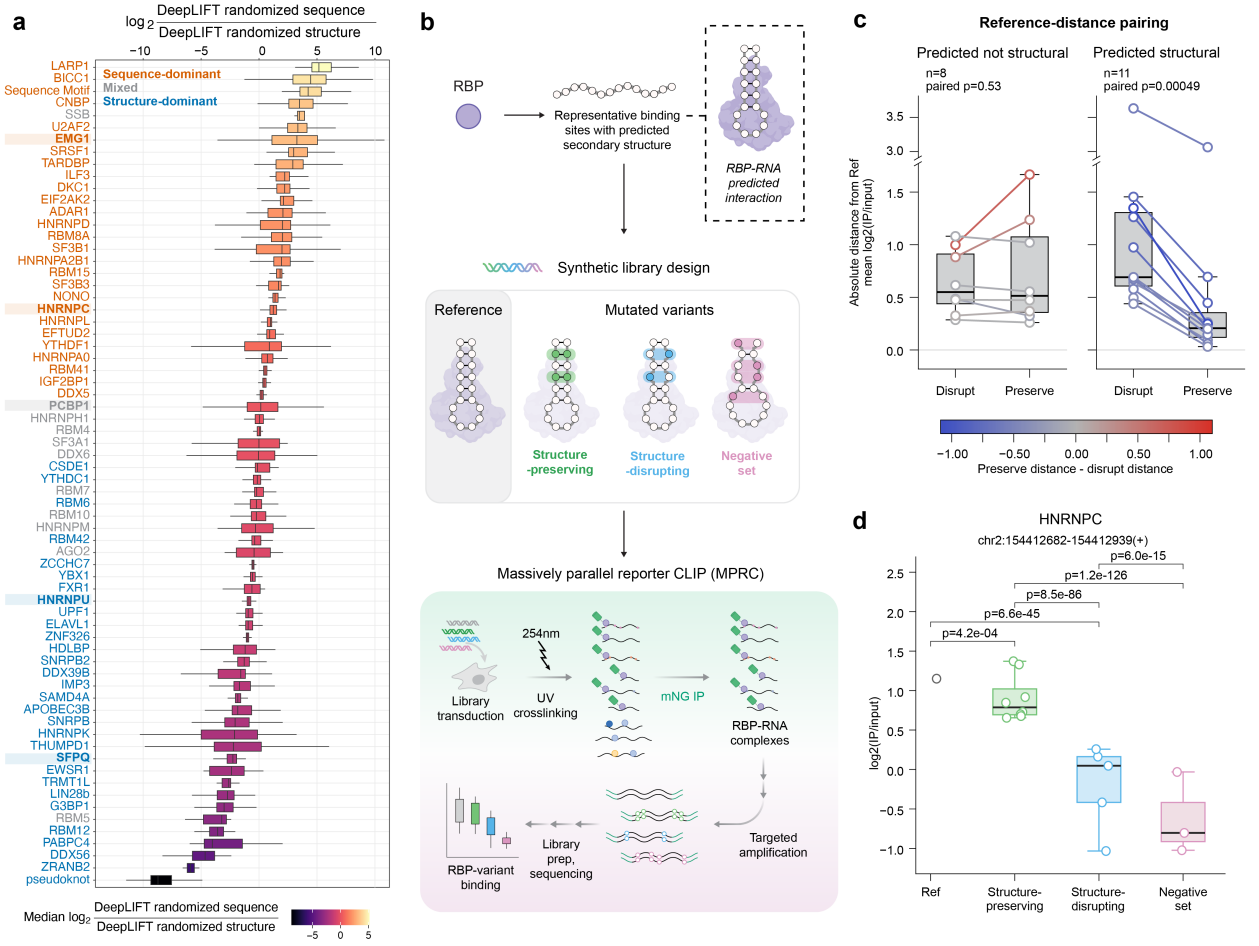


Figure 4 | **DeepLIFT and massively parallel reporter CLIP validate structure-dependent RBP binding determinants.** **(a)** DeepLIFT-based sequence and structure attribution across RBP binding models. For each RBP-specific Pythia model, DeepLIFT scores were computed after perturbing the sequence arm or the fixed-dilated-convolution structure arm. The $\log_2$ sequence-to-structure attribution ratio summarizes relative model reliance on each arm. Each box shows, for one RBP, the distribution of per-peak $\log_2$ ratios across its ecCLIP peaks, restricted to the 10% of peaks with the highest variance across the sequence, structure, and full-model DeepLIFT scores (Methods). Boxes show the median and interquartile range, whiskers extend to 1.5x the interquartile range, outliers are not shown, and fill color indicates the RBP's median $\log_2$ ratio. RBP-level labels are derived from these Pythia scores: an absolute median ratio >0.05 and Benjamini-Hochberg FDR $<10\%$ for a one-sample Wilcoxon test against zero define sequence-dominant (positive) or structure-dominant (negative) models; the remaining models are labeled mixed (Methods). Fixed sequence motifs and pseudoknot-derived controls illustrate the positive and negative directions of the attribution scale, respectively. **(b)** Schematic of massively parallel reporter CLIP (MPRC). Pythia-nominated RBP binding sites are synthesized as pooled reporter libraries containing reference sequences, structure-preserving mutations, structure-disrupting mutations, and multi-mutation negative controls. Reporter libraries are introduced into HEK293T cells expressing the corresponding endogenously tagged RBP, and reporter-targeted CLIP is used to quantify relative RBP binding. **(c)** Paired reference-distance analysis of MPRC validation sites whose reference reporters retained binding. Of 30 structure-validation sites tested, 19 showed reference enrichment (Ref $\log_2$(IP/input) > 0.25) and were included in this analysis. For each reference-enriched site, the absolute distance from the reference $\log_2$(IP/input) was compared for structure-disrupting and structure-preserving reporters. Sites predicted by Pythia to be structure-dependent showed significantly smaller reference distance for structure-preserving reporters than structure-disrupting reporters ($n = 11$, one-sided paired Wilcoxon $P = 4.9 \times 10^{-4}$), whereas sites not predicted to be structure-dependent showed no significant paired difference ($n = 8$, $P = 0.527$). **(d)** Representative HNRNPC validation site. For an intronic GALNT13 site at chr2:154,412,682-154,412,939(+), structure-preserving mutations maintain binding near the reference sequence, whereas structure-disrupting mutations and multi-mutation negative controls show reduced binding.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

associated sequence variation to measured changes in RBP binding, this analysis establishes a practical route from ecCLIP-trained, structure-aware prediction to experimentally testable post-transcriptional regulatory hypotheses.

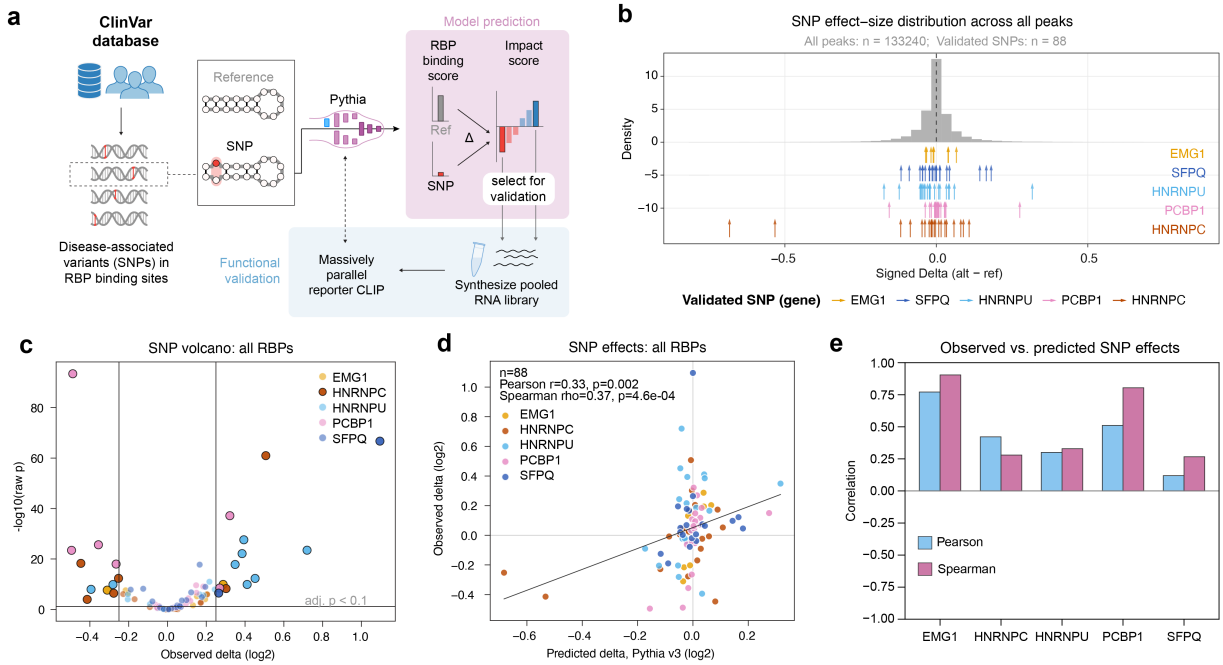


Figure 5 | **RBP-specific Pythia models predict allele-specific effects of trait-associated variants on RBP binding. (a)** Schematic of the targeted variant-validation workflow with MPRC. Phenotype- or trait-annotated variants overlapping ecCLIP binding sites are scored with RBP-specific Pythia models, reference and alternate alleles are cloned into pooled reporter libraries, and MPRC is used to quantify allele-specific RBP binding. **(b)** Variant selection strategy. For each RBP, the distribution of allele effects predicted by the corresponding Pythia model is shown for candidate variants overlapping ecCLIP binding sites, with selected predicted-loss, predicted-gain, and low-effect variants highlighted. **(c)** MPRC measurements of allele-specific binding. Volcano plot shows the observed allelic effect, defined as $\log_2(\text{IP/input})_{\text{Alt}}$ - $\log_2(\text{IP/input})_{\text{Ref}}$, versus -$\log_{10}$(raw P) from replicate-adjusted binomial logistic regression. Variants passing global FDR < 0.1 and $|\text{observed effect}| > 0.25$ are highlighted. **(d)** Concordance between predicted and experimentally measured variant effects across all tested RBP-variant pairs. Each point represents one RBP-variant pair, colored by RBP; predictions were generated by the RBP-specific Pythia model corresponding to the assayed RBP. Across 88 tested pairs, predicted and observed effects were positively correlated (Pearson r = 0.33, P = 0.0019; Spearman rho = 0.37, P = 4.6×10^{-4}). **(e)** RBP-specific concordance between predicted and observed variant effects. Bars show Pearson and Spearman correlations for EMG1 (n = 8), HNRNPC (n = 20), HNRNPU (n = 20), PCBP1 (n = 20), and SFPQ (n = 20).

3. Discussion

RNA-binding proteins regulate gene expression through sequence recognition, RNA structural context, transcript position, and cellular state. Dissecting these mechanisms requires experimental maps that are uniform enough for atlas-scale comparison and resolved enough to support quantitative modeling. Here we developed eclair CLIP (ecCLIP), a scalable CLIP-seq framework that combines endogenous split-fluorescent-protein tagging with ligation-free library preparation. Applied to 67 RBPs in HEK293T cells, ecCLIP generated a high-quality RBP binding atlas that recovers known sequence preferences, captures regulatory signal linked to RNA half-life, and provides a substrate for structure-aware modeling of RBP-RNA interaction grammar.

A central technical feature of ecCLIP is that endogenous tagging creates a common and stringent affinity-capture strategy across RBPs while preserving endogenous expression control. This addresses two persistent limitations of atlas-scale CLIP studies: antibodies differ substantially in availability, specificity, and capture efficiency, whereas approaches based on exogenous tagged constructs can perturb RBP abundance, stoichiometry, localization, or binding behavior (Ule et al., 2003; König et al., 2010; Van Nostrand et al., 2020). The shared endogenous capture handle does not remove all sources of variation, but it reduces major RBP-specific

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

technical variables without requiring ectopic over-expression from a transgene. Because many profiled RBPs were selected from the OpenCell collection, ecCLIP also adds a transcriptome-wide RNA-occupancy layer to a broader endogenous-tagging resource that includes protein localization and interaction measurements (Cho et al., 2022). This creates opportunities to interpret RBP binding alongside other OpenCell modalities while contributing a complementary data type to this resource.

The atlas shows that high-resolution RBP occupancy profiles are useful beyond binding-site annotation. Incorporating ecCLIP tracks into Borzoi-style models improved prediction of HEK293T RNA half-life, and expanding the training set with broader CLIP and Perturb-seq data improved it further. Borzuki serves here as a test of data utility: it extends the Borzoi framework to incorporate these measurements into prediction of RNA stability, RBP binding, and expression dynamics from genomic sequence (Linder et al., 2025). The ecCLIP ablations on RNA half-life and MPRA variant effects provide direct evidence for the added information in these binding maps. Performance on fine-mapped eQTLs and causal Mendelian 3′ UTR variants further assesses the combined model, while the MPRA results show complementarity with Saluki. These results should still be viewed as a step toward, rather than a complete solution for, post-transcriptional variant interpretation, because 3′ UTR variant effects remain dependent on cell type, assay, molecular readout, and transcript context.

Pythia provides the modeling framework for asking which sequence and structural features determine RBP recognition and how their perturbation changes binding. Deep learning models of transcription factor-DNA binding are motivated not only by prediction in new sequences, but also by the possibility of learning regulatory syntax from functional genomics data (Alipanahi et al., 2015; Kelley et al., 2018; Avsec et al., 2021). RBP-RNA interactions add a structural layer because binding can depend on both primary sequence and RNA accessibility or pairing. Building on prior work showing that structural context shapes post-transcriptional regulatory elements, including TEISER and pyTEISER (Goodarzi et al., 2012), Pythia explicitly separates sequence and structure arms so that model predictions can be decomposed into these components. The DeepLIFT and MPRC analyses show that this decomposition is experimentally meaningful: structure-preserving mutations were more likely to retain binding than structure-disrupting mutations at Pythia-nominated sites. The HNRNPC/GALNT13 example illustrates this logic, with structure-preserving mutations retaining HNRNPC binding and structure-disrupting mutations reducing binding, consistent with prior evidence that local RNA structure modulates HNRNPC accessibility (Liu et al., 2015).

RBP-specific Pythia models also prioritized naturally occurring variants with allele-specific binding effects in the reporter assay. Across 88 tested RBP-variant pairs, predicted and observed effects were positively correlated, and directional concordance was above chance after excluding variants with zero predicted effect. The strongest examples included HNRNPC binding loss at rs9323624 and rs56302494, HNRNPU binding gain at the BCKDHB-locus variant rs9361614, and PCBP1 binding loss at rs11668313. These findings should be interpreted as molecular effects on RBP binding in the HEK293T reporter system, not as causal disease mechanisms. Their trait annotations nominate hypotheses for follow-up, but establishing organism-level consequences will require testing native transcripts, relevant cellular states, and downstream RNA-processing or expression effects.

Several limitations define the next steps. The atlas was generated in HEK293T cells, so extending ecCLIP across cell types, cell perturbations, and disease contexts will be important for modeling context-specific regulation. Endogenous tagging provides a uniform capture strategy but still modifies each RBP locus, making orthogonal functional checks important for proteins sensitive to terminal tags. Pythia's structure arm encodes structural potential rather than directly measured RNA structure, and reporter assays simplify native transcript context. Pairing ecCLIP with structure probing, long-read isoform measurements, perturbations of trans-acting RBPs, and endogenous genome editing will help connect local interaction grammar to native regulatory outcomes.

Together, these results establish ecCLIP as both an experimental atlas and a modeling substrate for post-transcriptional regulation. The data improve genome-scale modeling of RNA output through Borzuki and enable structure-aware interaction modeling through Pythia. As RNA-focused sequence-to-function models continue to expand, their utility will depend on the quality, resolution, and consistency of the biological measurements used to train and evaluate them. ecCLIP addresses this need for RBP-RNA interactions and provides a foundation for decoding how RNA sequence and structure jointly shape post-transcriptional gene regulation.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

4. Methods

4.1. Generation of stable mNG-tagged RBP HEK293T cell lines

The 67 HEK293T cell lines expressing split mNeonGreen-tagged proteins were generated as previously described (Cho et al., 2022). The Methods Table lists the 67 proteins in the ecCLIP atlas; UniProt identifiers refer to the protein sequences used for analysis.

4.2. Cell Culture

HEK293T cell lines were cultured in Dulbecco's Modified Eagle Medium supplemented with 10% FBS, 1μg/mL Amphotericin (Corning), and 100 U/mL penicillin, 100 μg/mL streptomycin (Corning) in a 5% CO_2 humidified incubator.

4.3. eclair CLIP-seq

4.3.1. Cell Crosslinking

Cells were harvested at 80-90% confluency, two 15cm plates were used per replicate. Cells were washed 1X with ice-cold PBS and then irradiated with 400mJ/cm^2 254nm UV. On ice, cells were collected by scraping, and cell pellets frozen at −80°C.

4.3.2. Immunoprecipitation

Cell pellets were lysed on ice in 600μL cold low salt wash buffer (1X PBS, no Mg++ or Ca++, 0.1% SDS, 0.5% sodium deoxycholate, 0.5% IGEPAL CA-630) with 1X halt EDTA-free protease inhibitor (Thermo) and 6μL of SuperaseIn RNase inhibitor (ThermoFisher). 60μL DNase I (1 U/μL, Promega) was added to the lysate and incubated at 37°C for 10 minutes with shaking at 1000rpm. Lysate was then divided equally into 2 tubes. To tube 1, 10μL of a high concentration RNase mix (3.3ng/μL RNase A, 1 U/μL RNase I) was added; to tube 2 10μL of a low RNase mix (0.67ng/μL RNase A, 0.2 U/μL RNase I) was added. RNase mixes were diluted in low salt wash buffer before adding to lysate. Lysate was then incubated at 37°C for 5 minutes. The two RNase-treated tubes were then combined and spun at 4°C at 20,000xg for 20 minutes. The resulting clarified lysate was added to 25μL anti-neon green magnetic beads (Proteintech) that had been pre-washed 2X with PBST (PBS + 0.02% tween-20) and 1X with low salt wash buffer. The bead-lysate mixture was rotated end-over-end for 1.5-2 hours at 4°C. Beads were washed 2X with cold low salt wash buffer, 2X with cold high salt wash buffer (5X PBS no Mg++ or Ca++, 0.1% SDS, 0.5% sodium deoxycholate, 0.5% IGEPAL CA-630), and 2X with cold PNK wash buffer (50mM Tris-HCl pH 7.5, 10mM $MgCl_2$, 0.5% IGEPAL CA-630).

4.3.3. Dephosphorylation

After removing the final wash, RNA was dephosphorylated by resuspending the beads in 2.5μL 10X PNK buffer (500mM Tris pH 6.8, 5mM $MgCl_2$, 50mM DTT), 2μL T4 PNK (10u/μL, NEB), 0.5μL SuperaseIn (ThermoFisher), 20μL H_2O. Samples were incubated at 37°C for 20 minutes with interval shaking at 1350rpm 15 seconds / 5 minutes off. Beads were washed 1X with cold PNK wash buffer, 1X with cold high salt wash buffer and 2X with PNK wash buffer.

4.3.4. PolyA tailing RNP complexes

Following dephosphorylation, RNA was poly(A) tailed. On ice, beads were resuspended in 0.8μL of yeast poly(A) polymerase (Jena 600U/μL), 4μL 5X yeast PAP buffer, 1μL 10mM ATP (unlabeled), 0.5μL SuperaseIn, and 13.7μL H_2O. Samples were then incubated at 22°C for 5 minutes with a 15 second shake at 1350rpm after 2.5 minutes. The beads were then washed 2X with cold high salt wash buffer and 2X with cold PNK wash buffer.

4.3.5. Azide end labeling of RNP complexes

RNA was end labeled with azide by adding a mix containing 0.4μL yeast poly(A) polymerase (Jena 600U/μL), 2μL 5X yeast PAP buffer, 0.25μL SuperaseIn, 2μL 10mM 2′-Azido-2′-Deoxyuridine-5′-Triphosphate (TriLink

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

BioChemicals), and 5.35μL H_2O to the beads. Samples were incubated at 37°C for 20 minutes and interval shaking at 1350rpm 15 seconds / 5 minutes off. Beads were washed 2X with cold high salt wash buffer and 2X with cold 1X PBS

4.3.6. IR labeling of RNP complexes

Samples were then infrared labeled by adding 20μL of 0.2mM 800CW DBCO (LiCor Biosciences) in PBS. Beads were incubated at 22°C for 30 minutes with interval shaking at 1350rpm for 15 seconds / 5 minutes off, protected from light. Beads were then washed 1X with cold high salt wash buffer and 1X with cold PNK wash buffer. After removal of final wash, beads were resuspended in loading buffer (1X NuPAGE loading buffer (ThermoFisher), 50mM DTT). Beads were incubated at 75°C for 10 minutes shaking at 1000rpm protected from light. Beads were magnetized and the eluate collected.

4.3.7. PAGE Gel and Transfer

The eluates were run on a NuPAGE 4-12% Bis-Tris gel in MOPS running buffer (ThermoFisher) at 4°C, protected from light. Samples were transferred to Protran BA-85 nitrocellulose membrane (Whatman) in 1X NuPAGE transfer buffer (ThermoFisher) with 10% ethanol. Membrane was imaged on a LI-COR Odyssey instrument. To isolate the RNA bound to the protein of interest, the region ~20-45kDa above the molecular weight of the protein of interest was excised from the membrane.

4.3.8. Proteinase K digest and RNA capture

RNA was released by submerging the membrane pieces in 200μL proteinase K digestion buffer (100mM Tris-HCL pH 7.5, 100mM NaCl, 1mM EDTA, 0.2% SDS), and 12.5μL proteinase K (Ambion, 20U/μL) and incubating at 55°C for 45 minutes with shaking at 1100rpm. The eluate was transferred to clean tubes and the NaCl concentration of the eluate was adjusted to 0.5M NaCl. This eluate was then transferred to oligo d(T) magnetic beads (NEB) that had been pre-washed 2X with proteinase K buffer. Samples were then incubated at 25°C for 20 minutes with interval shaking at 300rpm 10 min / 1350rpm 15 seconds. Beads were then washed 2X with cold high salt wash buffer and 2X with cold 1X PBS. Samples were eluted off of the beads by adding 8μL of bead elution buffer (20mM Tris-HCl pH 7.5, 1mM EDTA) and incubating at 50°C for 5 minutes. The eluate was then transferred to PCR tubes.

4.3.9. Library preparation

Library preparation was performed using the Takara SMARTer smRNA-Seq Kit with modifications. To the RNA from the previous step, 2.5μL smRNA Mix 1 and 1μL of 10μM ecCLIP UMI RT primer was added and incubated at 72°C for 3 minutes followed by 2-5 minutes on ice. Then 6.5μL kit smRNA Mix 2, 0.5μL RNase inhibitor, and 2μL PrimeScript RT (200U/μL) was added and samples incubated at 42°C for 60 minutes, 70°C for 10 minutes, and then placed back on ice. To this cDNA, 50μL of 2X SeqAmp CB PCR Buffer, 2μL SeqAmp DNA Polymerase, 2μL 10μM ecCLIP LP reverse primer and 24μL H_2O was added. Samples were indexed by adding 2μL of a unique 10μM Takara forward index primer (provided in kit) to each sample. Samples were amplified using the following program: 98°C 1 minute followed by 98°C 10 seconds - 60°C 5 seconds - 68°C 10 seconds, repeated for the number of cycles determined by qPCR to ensure no overamplification. Libraries were cleaned up using SPRI beads at a 1.1X bead : PCR reaction ratio. Library size distribution was assessed with a Tapestation instrument (Agilent), and pooled libraries were sequenced on a HiSeq 4000 (Illumina) at the UCSF Center for Advanced Technologies.

4.4. Massively Parallel Reporter CLIP-seq (MPRC)

4.4.1. CLIP reporter cell line construction

Sequences containing the binding site for the RBP of interest, along with mutant variants (Methods Table), were cloned immediately 3′ of an EGFP CDS in a lentiviral backbone. Using lentiviral transduction, each variant sequence library was stably integrated into a HEK293T cell line expressing the split mNeonGreen-tagged target RBP.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

4.4.2. Cell Crosslinking

Cells were harvested at 80-90% confluency, two 15cm plates were used per replicate. Cells were washed 1X with ice-cold PBS and then irradiated with 400mJ/cm^2 254nm UV. On ice, cells were collected by scraping, and cell pellets frozen at −80°C.

4.4.3. Immunoprecipitation

Cell pellets were lysed on ice in 1.2mL cold low salt wash buffer (1X PBS, no Mg++ or Ca++, 0.1% SDS, 0.5% sodium deoxycholate, 0.5% IGEPAL CA-630) with 1X halt EDTA-free protease inhibitor (Thermo) and 12μL of SuperaseIn RNase inhibitor (ThermoFisher). 120μL DNase I (1 U/μL, Promega) was added to the lysate and incubated at 37°C for 10 minutes with shaking at 1000rpm. The lysate was spun at 4°C 20,000xg for 20 minutes. 5% of the resulting clarified lysate was retained to use as the input sample. The remaining clarified lysate was added to 25μL anti-neon green magnetic beads (Proteintech) that had been pre-washed 2X with low salt wash buffer. The bead-lysate mixture was rotated end-over-end for 1.5 hours at 4°C. Beads were then washed 3X with cold low salt wash buffer and then 3X with cold high salt wash buffer (5X PBS no Mg++ or Ca++, 0.1% SDS, 0.5% sodium deoxycholate, 0.5% IGEPAL CA-630).

4.4.4. Proteinase K digest and RNA isolation

RNA was released from the beads by adding 200μL proteinase K digestion buffer (100mM Tris-HCL pH 7.5, 100mM NaCl, 1mM EDTA, 0.2% SDS), and 12.5μL proteinase K at 37°C for 1 hour with interval shaking at 1100rpm 30 seconds / 5 minutes off. The lysate input samples were processed in parallel. Acid phenol chloroform (Sigma) extraction was used to isolate the RNA. The RNA was ethanol precipitated, washed 2X with 75% ethanol, air dried and resuspended in 18μL H_2O. The input RNA was resuspended in 54μl H_2O.

4.4.5. Library preparation

cDNA was synthesized by mixing 1μL of 2μM reporter MPRC UMI RT primer, 1μL 10mM dNTPs, and 11.5μL of the IP RNA isolated above. For the input samples, 1μg of RNA was used as input. Samples were incubated at 65°C 5min then placed immediately on ice. To this, 4μL 5X RT buffer, 1μL 100mM DTT, 0.5μL SuperaseIN and 1μL SuperScript IV RT (ThermoFisher) were added. Samples were incubated at 55°C 10min then 80°C for 10 min. In parallel, reactions without reverse transcriptase were performed to assess for DNA contamination. To digest free oligo, 1μL thermolabile exonuclease I (20 U/μL, NEB) was added to each sample and incubated at 37°C 4 min then 80°C 1 min. Libraries were amplified by combining 50μL 2X KAPA HiFI HotStart Ready mix (Roche), 2μL 10μM reporter MPRC forward PCR1 primer, 2μL 10μM reverse index primer, 34μL H_2O and 10μL of the ExoI-treated cDNA. Samples were cycled using the following program: 1. 95°C 3min 2. 98°C 20sec 3. 67°C 15sec 4. 72°C 15sec. 5. 72°C 1min. Steps 2-4 were repeated for the number of cycles determined by qPCR to prevent overamplification. Libraries were cleaned-up using SPRI beads and eluted in 22μL H_2O. A second index was added by combining 50μL 2X KAPA HiFi HotStart Ready mix, 2μL 10μM forward index primer, 2μL 10μM reverse index primer, 20μL PCR product (from previous step), 24μL H_2O. Samples were cycled using the following program: 1. 95°C 3min 2. 98°C 20sec 3. 67°C 15sec 4. 72°C 15sec repeat 2-4 X 8 5. 72°C 1min. Libraries were cleaned-up using SPRI beads, eluted in 22μL H_2O, and subsequently pooled and size-selected using PAGE. Library size distribution and quantity was assessed using a tapestation instrument (Agilent). Sequencing was performed on a NovaSeqX (Illumina) at the UCSF Center for Advanced Technologies.

4.5. Computational analysis of eclair CLIP libraries

Raw paired-end FASTQ files were first converted to single-read ecCLIP input files while preserving molecular barcode information. For each R1 FASTQ file, the matching R2 file was used to append a unique molecular identifier (UMI). Datasets were then analyzed using the CLIP Tool Kit (CTK) based scripts. Unless otherwise noted, reads were trimmed with cutadapt (v2.3) by removing the first 3 nt (-u 3), applying a quality cutoff of 15, retaining reads of at least 20 nt, and trimming a poly(A) adapter sequence (AAAAAAAAAAAAAAAAAA). Trimmed reads were aligned to the human hg38 genome with bwa (v0.7.17) using -n 0.06 and -q 20. Alignments were sorted and indexed with samtools (v1.7), and PCR duplicates were collapsed with umi_tools (v1.1.6) dedup using the UMI encoded in the read name.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Deduplicated alignments were converted to CTK (v1.1.2) tag BED files using parseAlignment.pl with minimum mapping quality 1 and minimum read length 18. Mutation files were generated during parsing, and mutations corresponding to unique tags were retained with selectRow.pl. Biological replicates were then concatenated into pooled tag BED files. CTK tag2profile.pl was used to generate strand-specific bedGraph profiles, and CTK tag2peak.pl was used for peak calling with strand specificity, valley seeking, $p = 0.05$, valley depth 0.9, hg38 annotation, and multiple-test correction. Both full peak regions, peak boundaries, and half peak-height regions were exported.

Mutations in unique tags were split into substitutions, strand-aware C-to-T conversions, deletions, and insertions. Deletion-based crosslinking-induced mutation sites (CIMS) were called from pooled unique tags with CTK CIMS.pl using -n 10, permutation mode, and cached deletion backgrounds. Sites with FDR ≤ 0.1 were retained for the main RBP analyses, and ±10 nt windows around CIMS sites were generated for sequence and motif analyses.

Background-filtered peak and CIMS sets were generated by removing strand-matched intervals overlapping the tag-only negative-control BED set using intersectBed -s -v (bedtools v2.26.0). These filtered BED files were the primary inputs for the atlas-level summaries.

4.6. RBP atlas summaries

Filtered CIMS and peak BED files were collected across RBPs. Genomic-region annotations were assigned with strand-specific intersectBed against hg38 3′ UTR, 5′ UTR, exon, and intron BED files. Categories were assigned hierarchically: 3′ UTR first, then 5′ UTR after excluding 3′ UTR intervals, then coding/exonic intervals after excluding UTRs, then introns after excluding UTRs and exons. Remaining intervals were classified as intergenic.

For atlas-level overlap analysis, filtered intervals were concatenated, sorted, and merged strand-specifically with mergeBed -s. Merged intervals were named by genomic coordinates and the set of contributing RBPs. Per-RBP binary/count matrices were generated by intersecting each RBP's filtered BED file with the merged interval set using intersectBed -wo -s, followed by counting overlaps per merged interval. A separate tag-count matrix was produced by intersecting each pooled unique-tag BED file with the merged interval set. Heatmaps and clustered correlation plots were generated with pandas, seaborn, matplotlib, and scipy.

Biological replicate concordance was estimated by converting filtered BED intervals to SAF annotation files and counting deduplicated BAM reads over those intervals with featureCounts (v1.6.2) -F SAF -R BAM -s 1 -T 20. For each RBP, replicate count columns were read into pandas and Pearson correlation coefficients were calculated with scipy.

4.7. Sequencing Read Processing

Paired-end sequencing reads were processed separately for each RBP, sample type, and replicate. Read pairs were first merged with PEAR using multithreading. PEAR intermediate FASTQ outputs were compressed after merging. The terminal 8 nt UMI was extracted from the 3′ end of each assembled read using umi_tools extract –3prime with the pattern NNNNNNNN. The extracted UMI was appended to the read name and the UMI bases were removed from the read sequence.

Library linkers were removed from UMI-trimmed assembled reads using cutadapt with a linked-adapter model matching the expected 5′ and 3′ library linkers. Reads were retained only if the expected linker structure was detected. After linker removal, inserts were filtered to retain only exact designed insert lengths: 50 nt for SNP constructs and 75 nt for structure-variant constructs.

4.8. Reference Construction and Alignment

Per-RBP reference FASTA files were generated from the construct metadata. Metadata sequences were provided as RNA sequences and were converted to DNA by replacing U with T. SNP references included both reference and alternate 50 nt alleles. Structure references included reference controls, structure-preserving variants, structure-disrupting variants, and negative-control variants. FASTA headers were sanitized to Bowtie-compatible construct identifiers, with a separate metadata table preserving the mapping to the original con-

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

struct IDs and annotations.

Filtered inserts were aligned to the corresponding per-RBP reference using Bowtie2 in end-to-end mode. Alignments were restricted to exact insert matches by using –score-min L,0,0; in end-to-end alignment this requires a Bowtie2 score of 0. Reads were aligned with –norc, –very-sensitive, -L 18, -k 2, and –no-unal. Only reads with a unique best exact alignment were counted. PCR duplicates were collapsed by counting unique UMIs per construct per sample.

4.9. Count Normalization

For each RBP, a construct-by-sample count matrix was generated from unique UMI counts across IP and input samples for both biological replicates. SNP and structure constructs were normalized together because they were present in the same library and sequencing samples.

Counts were normalized with a conservative DESeq2-style median-of-ratios procedure. Size factors were estimated from construct count ratios and then rescaled so that the minimum size factor was 1.0. This ensures that normalization only scales counts downward and does not inflate counts or statistical power. Normalized counts were floored to integers before logistic-regression analyses.

4.10. SNP Effect Analysis

For each SNP, reference and alternate allele constructs were compared using a binomial logistic-regression model adapted from the Ribolog framework. For each replicate and allele, IP counts were modeled as successes and input counts as failures. The model included allele and replicate terms:

```
cbind(IP, input) ~ allele + replicate
```

The reference allele was used as the treatment reference level. The allele coefficient therefore estimates the log-odds-ratio of IP/input for the alternate allele relative to the reference allele. This coefficient was converted to $\log_2$ units:

$$\texttt{observed_delta_log2} = \log_2(\text{IP/input})_{\text{Alt}} - \log_2(\text{IP/input})_{\text{Ref}}$$

SNP model p-values were obtained from the Wald test for the alternate-allele coefficient. Multiple-testing correction was performed using Benjamini-Hochberg FDR both globally across all tested SNPs and separately within each RBP. Positive predicted values indicate higher predicted RBP binding for the alternate allele relative to the reference allele, while negative values indicate lower predicted binding for the alternate allele.

4.11. Structure-Variant Analysis

Structure-variant constructs were grouped by reference peak. Each construct was assigned to one of four analysis groups: Ref (reference), Structure-preserv. (structure preserving mutations), Structure-disrupt. (structure disrupting mutations), Neg. set (negative set; multiple mutations to disrupt both sequence and structure). For each construct, $\log_2$ enrichment was calculated from normalized IP and input counts using a small pseudocount. Replicates were combined at the construct level for plotting, so each construct contributed a single plotted value. For pairwise group comparisons, binomial logistic regression modeled normalized IP counts as successes and normalized input counts as failures, with construct-group and replicate terms. The construct-group coefficient tests a difference in IP/input enrichment between the two groups after accounting for replicate effects. We required $\log_2$(IP/input) for Ref to be >0.25, so that we are confident the sequence is being bound by the cognate RBP within the context of the reporter.

Predicted structural and predicted not-structural peaks were compared using effect-size summaries, one-sided Mann-Whitney tests, Welch t-tests, and Cliff's delta. A categorical expected-pattern analysis was also performed, where a peak was counted as matching the expected structure-dependent pattern when it was reference-enriched, Structure-preserv. was closer to Ref than Structure-disrupt. was, and Structure-preserv. exceeded Structure-disrupt.. Enrichment of this pattern among predicted structural peaks was evaluated with

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

a one-sided Fisher exact test. For the primary paired rescue analysis, we compared absolute distances from reference enrichment for structure-preserving and structure-disrupting constructs within each reference-enriched site using a one-sided paired Wilcoxon test, separately for the two predicted structural classes (**Fig. 4c**).

4.12. Pythia model

4.12.1. Overview

Pythia is a dual-arm neural network architecture that jointly processes RNA sequence and potential base-pairing information to predict either RBP binding (binary classification) or RNA structural properties (four BEACON tasks: contact map prediction, distance map prediction, secondary structure prediction, and structural score imputation). All models share a common structural inductive bias: a frozen fixed dilated convolutional layer that explicitly encodes candidate Watson–Crick and wobble base pairs across a range of stem distances and bulge configurations.

4.12.2. Input representation

RNA sequences are encoded as one-hot tensors of shape (B, 4, L), where B is batch size, L is sequence length, and the four channels correspond to A, C, G, and U (T is treated as U). All sequences are zero-padded or truncated to the task-specific maximum length prior to batching.

4.12.3. Fixed dilated convolutional layer

The core structural feature extractor is a single frozen 1-D convolutional layer with pre-designed kernels that score potential base pairs at multiple distances and bulge configurations. For each dilation radius d ∈ [d_start, d_end] and each Watson–Crick or wobble pair type (A-U, U-A, C-G, G-C, G-U, U-G), the layer places one kernel that activates when a complementary nucleotide pair is present at positions separated by d. To accommodate imperfect stems, each dilation radius is further expanded into bulge variants: for bulge size 1, offsets of {−1, 0, +1} are applied to each position in the pair; for bulge size 2, additional offsets of ±2 are included. Kernel weights are fixed and not updated during training. The total number of output channels depends on the dilation range and bulge size; for the default configuration (d_start=5, d_end=24, bulge_size=2) this yields 600 channels.

4.12.4. Dual-arm feature extraction

All Pythia models process each input sequence through two parallel convolutional arms:

Structure arm (arm1): receives the output of the fixed dilated convolutional layer and applies three successive 1-D convolutional blocks (kernel sizes 16, 10, 5) with BatchNorm and ReLU activations between each layer. For the RBP model, an adaptive max-pool to 60 positions is inserted before arm1 to produce a fixed-length representation from the variable-length FD output; for BEACON tasks, the FD output is symmetrically padded and center-cropped to preserve sequence length L.

Sequence arm (arm2): receives the raw one-hot input and applies three successive 1-D convolutional blocks (kernel sizes 12, 7, 5) with BatchNorm and ReLU activations.

After both arms, the feature maps are concatenated along the channel dimension. A BatchNorm1d transition layer is applied to the concatenated representation before downstream processing in all BEACON models (the RBP model omits this).

4.12.5. Pythia RBP model

Task: Binary classification of ecCLIP peak sequences (Bound) versus dinucleotide-shuffled control sequences (Unbound). The negative class is a computational control, not an experimentally verified set of unbound transcripts.

Architecture: Input sequences are 256 nt. The structure arm (arm1) processes fixed dilated convolutional features (d_start=5, d_end=24, bulge_size=2) after adaptive max-pooling to 60 positions (BatchNorm + ReLU), then applies three Conv1d layers with widths (64, 64, 64), kernel sizes (16, 10, 5), and max-pool

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

stride 2 after each layer. The sequence arm (arm2) applies three Conv1d layers directly to one-hot input with widths (128, 64, 64), kernel sizes (12, 7, 5), and max-pool stride 2 after each. Both arms use dropout (p=0.25). The outputs of both arms are concatenated and flattened, then passed through three fully connected layers: Dense(64) → ReLU → Dense(32) → ReLU → Dense(2). Weights are initialized with Kaiming normal (Conv1d), Xavier uniform (Linear), and constant 0.01 bias (Linear).

Training: Models are trained with Adam (lr=1×10^{-3}, weight_decay=1×10^{-4}), batch size 64, for up to 50 epochs with early stopping (patience=5). A linear learning-rate warmup over 2 epochs is followed by cosine decay. Gradient norms are clipped to 1.0. Training uses 16-bit mixed precision. The loss is CrossEntropyLoss with class weights [0.2, 1.0] (Unbound, Bound), giving the positive (Bound) class 5× more influence to compensate for the ~10:1 negative-to-positive class imbalance within the training datasets. For each of the 67 RBPs, a separate model is trained. Negative (Unbound) examples are dinucleotide-shuffled sequences derived from bound peaks.

4.13. Pythia BEACON models

All four BEACON task models share the same dual-arm backbone (FixedDilatedConv d_start=5, d_end=24 unless noted; bulge_size=2; dropout=0.15) and diverge at the fusion and prediction head. For the three 2-D tasks (CMP, DMP, SSP), per-position 1-D features from both arms are lifted into pairwise (L × L) representations via outer product, outer sum, and absolute outer difference operations, then projected into a shared 2-D representation and refined by a 2-D ResNet. All pairwise 2-D outputs are symmetrized. For the 1-D task (SSI), per-position features are refined by a stack of 1-D residual blocks and decoded by a scalar regression head.

Learned distance encoding: For all four tasks, a learned positional encoding of pair distance $|i-j|$ is added to the pairwise feature map. This is parameterized as a lookup table over binned distances (dist_enc_dim=64 for all models), allowing the model to learn distance-dependent priors for structural contact formation.

4.13.1. Contact map prediction (PythiaCMP)

Task: Binary prediction of long-range RNA base contacts ($|i-j| \geq 23$).

Architecture: Arm1 widths: (256, 192, 128); Arm2 widths: (256, 128, 128). Total 1-D feature dimension at arm outputs: 128 + 128 = 256. Pairwise feature channels: outer_prod + outer_sum + outer_diff (3×256) + dist_enc (64) = 832. A 1×1 Conv2d (BatchNorm + ReLU) projects pairwise features to hidden_dim=256. The 2-D ResNet contains 7 residual blocks (BatchNorm + ReLU) with a multi-scale dilation pattern [1, 2, 4, 8, 1, 2, 4] applied in sequence. The output head applies Conv2d(256→128, kernel 3) + BatchNorm + ReLU + Dropout + Conv2d(128→1, kernel 1), followed by symmetrization. Sequences are padded to max_len=1024.

Loss and metric: BCEWithLogitsLoss. Evaluation uses the BEACON official Top-L precision: for each RNA of length L, the top-L predicted positions (by sigmoid probability) with score > 0.5 among positions with $|i-j| \geq 23$ are taken, and precision is micro-averaged across the test set.

Training: Adam (lr=1×10^{-4}, weight_decay=0.01), batch size 1, gradient accumulation over 8 steps (effective batch size 8), up to 50 epochs, early stopping patience=5, linear warmup for 1 epoch.

4.13.2. Distance map prediction (PythiaDMP)

Task: Regression of pairwise C3′–C3′ distances (normalized to [0,1]).

Architecture: Arm1 widths: (192, 128, 128); Arm2 widths: (256, 192, 128). Total 1-D feature dimension: 128 + 128 = 256. Pairwise feature channels: outer_prod + outer_sum + outer_diff + distance-weighted outer_diff + inverse-distance-weighted outer_sum (5×256) + raw sequence distance (1) + multi-scale exponential decay features at scales 10, 50, 100 nt (3) + dist_enc (64) = 1348. A 1×1 Conv2d (BatchNorm + ReLU) projects to hidden_dim=228. The 2-D ResNet contains 7 blocks with dilation pattern [1, 2, 4, 8, 1, 2, 4]. The output head applies Conv2d(228→114) + BatchNorm + ReLU + Dropout + Conv2d(114→1), followed by sigmoid (to enforce [0,1] output) and symmetrization. Sequences are padded to max_len=1024.

Loss and metric: Huber loss (δ=0.1). Evaluation uses global Pearson R^2 computed over all non-padded pairwise positions across the test set (micro-averaged).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Training: Adam (lr=5×10^{-5}, weight_decay=0.01), batch size 1, gradient accumulation over 8 steps (effective batch size 8), up to 50 epochs, early stopping patience=5, linear warmup for 5 epochs.

4.13.3. Secondary structure prediction (PythiaSSP)

Task: Binary prediction of base-paired positions (full L × L contact matrix, including short-range pairs).

Architecture: Arm1 widths: (128, 128, 128); Arm2 widths: (256, 128, 128). Total 1-D feature dimension: 128 + 128 = 256. Pairwise feature channels: outer_prod + outer_sum + outer_diff (3 × 256) + dist_enc (64) = 832. A 1×1 Conv2d (GroupNorm(8) + GELU) projects to hidden_dim=256. The 2-D ResNet contains 8 residual blocks (GroupNorm(8) + GELU) with dilation pattern [1, 2, 4, 1, 2, 4, 1, 1]. The output head applies Conv2d(256→128, kernel 3) + GroupNorm(8) + GELU + Conv2d(128→1, kernel 1), followed by symmetrization. Sequences are padded to max_len=512.

Loss and metric: BCEWithLogitsLoss. Evaluation uses global micro-averaged F1 (with precision and recall) at a 0.5 sigmoid threshold across all non-padded positions.

Training: Adam (lr=3×10^{-4}, no weight decay), batch size 4, gradient accumulation over 2 steps (effective batch size 8), up to 100 epochs, early stopping patience=3, linear warmup for 3 epochs.

4.13.4. Structural score imputation (PythiaSSI)

Task: Per-position regression of structural accessibility scores (e.g., SHAPE reactivity or predicted unpairedness), imputing masked positions from context.

Architecture: Extended dilation range: d_end=48, giving the fixed dilated conv layer sensitivity to potential base pairs up to 48 nt apart. Arm1 widths: (128, 128, 128); Arm2 widths: (256, 128, 128). Combined 1-D feature dimension after concatenation and transition BatchNorm: 256. A 1×1 Conv1d projects to hidden_dim=256. Six 1-D residual blocks (BatchNorm + ReLU, dilation=1) refine per-position features. A final LayerNorm is applied before the regression head: Linear(256→128) + GELU + Dropout(0.15) + Linear(128→1), producing a per-position scalar. Sequences are padded to max_len=440.

Loss and metric: MSELoss computed only at masked (imputed) positions. Evaluation uses Pearson R^2 between predicted and true structural scores at masked positions.

Training: Adam with two learning-rate groups — backbone (lr_feat=3×10^{-4}) and regression head (lr_head=1.5×10^{-3}) — weight_decay=0.01; batch size 64, gradient accumulation over 2 steps; up to 40 epochs, early stopping patience=10, linear warmup for 5 epochs. Gradient norms are clipped to 1.0.

4.14. DeepLIFT analysis of sequence versus structure contributions

To quantify the relative contributions of sequence and structural features to individual RBP binding predictions, we applied DeepLIFT to models trained on each RBP included in the attribution analysis. For each peak sequence, DeepLIFT attribution scores were computed separately by replacing the sequence arm input or the structure arm input with an all-zero (no-nucleotide) baseline, holding the other arm fixed. The resulting attribution scores reflect the contribution of each arm to the binding prediction. For each sequence, the $\log_2$ ratio of the sequence-arm attribution to the structure-arm attribution was computed as $\log_2$(DeepLIFT_seq_randomized / DeepLIFT_str_randomized). Positive values indicate sequence dominance; negative values indicate structural dominance. To focus on the most informative peaks, only the top 10% of peaks ranked by variance across the three scores (structure, sequence, DeepLIFT) were retained per RBP. The resulting distribution of $\log_2$ ratios per RBP was tested against a median of zero using a one-sample Wilcoxon signed-rank test. P-values were corrected for multiple testing using the Benjamini–Hochberg method. RBPs with $|\text{median } \log_2 \text{ ratio}| > 0.05$ and FDR < 10% were classified as sequence-dominant (positive median) or structure-dominant (negative median); the remaining models were labeled mixed. For the reporter validation in **Supplementary Fig. S4b**, this per-site $\log_2$ score ratio is termed the Structural Reliance Index (SRI) and we defined predicted structure-dependent sites by SRI < −0.1.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

4.15. Protein-embedding comparison with ecCLIP binding preferences

Canonical sequences for the 67 RBPs were retrieved from UniProt and embedded with ESM2 model (esm2_t33_650M_UR50D; Lin et al., 2023). Each protein was represented by the layer-30 embedding averaged over its annotated RNA-binding-domain residues (UniProt domain annotations; full-length protein for the four RBPs without an annotated domain) (**Fig. 1c,d**), or over all residues for the whole-protein comparison (**Supp. Fig. S1i,j**). Binding preferences were demonstrated as the log enrichment of each of the 1,024 5-mers in 20-nt windows centered on ecCLIP peak summits relative to flanking background windows, using up to 100,000 peaks per RBP. After standardizing features within each space, concordance was measured with the modified RV coefficient, a Mantel test on pairwise cosine distances (Spearman ρ), and distance correlation (reported as R). Significance was assessed against a null from 9,999 permutations of RBP labels. Dendrograms were built with Ward's method and cut at k = 6, and both trees were colored by the protein-embedding clusters.

4.16. Linear probing for RBP prediction

To validate Pythia on RBP-binding prediction, we benchmarked it against large pre-trained RNA foundation models, where we evaluated six RNA language models by linear probing: RNA-FM (Chen et al., 2022), RiNALMo (Penić et al., 2025), RNA-MSM (Zhang et al., 2024), RNAErnie (Wang et al., 2024), and the two UTR-LM variants UTR-LM-MRL and UTR-LM-TE_EL (Chu et al., 2024). All models were accessed as pre-trained weights through the MultiMolecule framework (Chen, 2026), an open-source library that distributes RNA foundation models via the Hugging Face Hub and were kept frozen throughout. For each of the 67 RBPs, every RNA sequence was tokenized and passed through the frozen model, and a fixed-length representation was obtained by mean-pooling the final hidden-layer token embeddings over all non-special, non-padding positions ("mean-last" pooling). On top of these embeddings, we trained an L2-regularized logistic-regression probe (scikit-learn (Pedregosa et al., 2011); lbfgs solver, class-balanced weighting, up to 1,000-2,000 iterations) to classify bound versus unbound sequences. The inverse regularization strength C was selected per RBP and per model over a grid spanning 10^{-3} to 10^{2}, choosing the value that maximized area under the receiver-operating-characteristic curve (auROC) on the held-out tuning set; no foundation-model weights were updated at any stage. Probes were then evaluated on the validation and held-out sets using auROC and area under the precision-recall curve (auPR).

To ensure a controlled comparison, all foundation-model probes and Pythia were trained, tuned, and evaluated on identical per-RBP data splits (training, tuning, validation, and held-out sets) drawn from the eclairCLIP dataset, with unbound (negative) sequences generated by dinucleotide-preserving shuffling of bound peaks, matching the Pythia RBP training procedure. Using the same splits, we additionally trained four supervised baselines from scratch on each RBP: DeepBind (Alipanahi et al., 2015), DeepSEA (Zhou and Troyanskaya, 2015), PRIESSTESS (Laverty et al., 2022), and GraphProt (Maticzka et al., 2014), evaluated with the same auROC and auPR metrics.

4.17. BEACON Benchmark Suite

To place Pythia's structure-awareness on a standardized external footing, we also evaluated it on the BEACON RNA-benchmark suite (Ren et al., 2024), focusing on its four structure-oriented tasks: contact map prediction (CMP), distance map prediction (DMP), structural score imputation (SSI), and secondary structure prediction (SSP). For each task we used BEACON's provided datasets, train/validation/test splits, and evaluation protocol without modification, so that Pythia's results are directly comparable to the 17 baseline models reported in the benchmark; three conventional architectures (CNN, ResNet, LSTM), twelve pre-trained RNA language models (RNA-FM, RNABERT, RNA-MSM, three SpliceBERT and four 3UTRBERT variants, and two UTR-LM variants), and the benchmark's own BEACON-B and BEACON-B512 models. The Pythia task-specific architectures and training configurations are described above (Pythia BEACON models); in every case only the task-specific arms and prediction head were trained while the fixed dilated-convolution layer remained frozen. For each task the best-performing epoch on the validation split was selected and evaluated once on the held-out test set. Task-specific metrics followed the BEACON definitions: CMP was scored by Top-L precision over long-range contacts (contacts defined at an 8 Å threshold), DMP by the coefficient of determination (R^2) over valid pairwise positions, SSI by R^2 over masked positions, and SSP by F1 over base-paired positions.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

4.18. Secondary-structure accessibility of the HNRNPC poly-U binding site

HNRNPC recognizes poly-U only when the tract is single-stranded and presented across the RRM β-sheet (PDB 2MXY; Cieniková et al., 2014; Cienikova et al., 2015), so we used RNA secondary structures to test whether the structure-preserving and structure-disrupting mutations remain compatible with binding. Minimum-free-energy secondary structures for each 100-nt construct were predicted from sequence with ViennaRNA (RNAfold v2.7.2, default Turner 2004 parameters; Lorenz et al., 2011), and per-nucleotide base-pairing probabilities were derived from the equilibrium partition function. Structures were drawn with VARNA (Darty et al., 2009) and accessibility was quantified as the number of U52-U58 bases left unpaired in the minimum-free-energy structure and their mean probability of being unpaired.

4.19. Deep Learning-based joint modeling of RBP binding, mRNA degradation and expression

4.19.1. Model architecture

The Borzuki model follows the same overall architecture as the original Borzoi model (Linder et al., 2025). However, since we trained Borzuki on a considerably smaller set of coverage profiles, we reduced the number of parameters. Specifically, we trained two different model sizes, each with a fixed input window size of 393,216 bp:

40 million parameters:

- Convolution tower: Starts with 320 channels, ends with 768 channels.
- Self-attention: 8 blocks, 64 key size, 2x dense expansion.

5 million parameters:

- Convolution tower: Starts with 192 channels, ends with 384 channels.
- Self-attention: 4 blocks, 32 key size, 1.5x dense expansion.

4.19.2. Training data

We curated RNA-seq degradation time-course experiments from recount3, and paired these data with auxiliary post-transcriptional and epigenomic assays from Borzoi's original training set. These data include:

RNA-seq degradation time-course experiments:

- 14 experiments in human HEK293T, HeLa, MCF7, HepG2, LCL, Jurkat, K562, and HUES9 cells.
- 11 experiments in mouse 3T3, mESC, cortical neuron, and dendritic cells.

CLIP-seq experiments:

- 31 ENCODE eCLIP experiments in human K562 and HepG2 cells.
- 11 ecCLIP-seq experiments from the Goodarzi Lab in HEK293T cells.

Perturb-seq experiments:

- 10 perturbations of 3′-cleavage regulators in HEK293T cells (Kowalski et al., 2024).

Baseline training data from the original Borzoi model:

- FANTOM5 CAGE, ENCODE DNase-seq, and RNA-seq matching any of the above 12 cell types.

See the 'Data and code availability' section for access to a detailed list of data identifiers. We only included CLIP-seq targets that overlapped the set of stability-correlated RBPs identified by Agarwal and Kelley (2022).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

4.19.3. Marginalized variant effect predictions

To make predictions for genetic variants measured in an MPRA context, we inserted the reference and alternate oligo sequences into multiple native genomic loci and averaged the variant effect score across these contexts. The model makes variant effect predictions for multiple RNA-seq tracks, and these values must be distilled into a scalar variant effect score. We tested two methods for making such marginalized predictions, which are detailed below.

As preliminaries, let $\boldsymbol{s}^{\text{ref}}$ and $\boldsymbol{s}^{\text{alt}}$ be the reference and alternate oligos. Insert $\boldsymbol{s}^{\text{ref}}$ and $\boldsymbol{s}^{\text{alt}}$ into K distinct genomic sequences (each representing a unique gene's 3′ UTR), producing a set of reference and alternate one-hot-coded sequences $\{\boldsymbol{x}^{\text{ref}}\}_{k=1}^{K}$ and $\{\boldsymbol{x}^{\text{alt}}\}_{k=1}^{K}$, each of shape $L \times 4$ (L = 393,216 for Borzuki). Obtain coverage predictions $\boldsymbol{y}^{\text{ref}}, \boldsymbol{y}^{\text{alt}} \in \mathbb{R}^{M \times K \times T}$ for all T RNA-seq tracks (M = output length) and compute exon-aggregated log-coverage statistics $\boldsymbol{u}^{\text{ref}}, \boldsymbol{u}^{\text{alt}} \in \mathbb{R}^{K \times T}$:

$$\boldsymbol{u}^{\text{ref}}_{k,t} = \log_2\left(\sum_{h=1}^{H} \boldsymbol{y}^{\text{ref}}_{\mathcal{B}_k(h),k,t}\right), \quad \boldsymbol{u}^{\text{alt}}_{k,t} = \log_2\left(\sum_{h=1}^{H} \boldsymbol{y}^{\text{alt}}_{\mathcal{B}_k(h),k,t}\right) \tag{4.1}$$

The set $\mathcal{B}_k = \{b_1, \ldots, b_H\}$ denotes the indices of exon-overlapping output bins for the target gene context k.

4.19.4. Approach 1 - Fit to Saluki PC1:

In this method, we fit a linear regression model to predict Saluki's cell type-agnostic half-life measurements, which were originally derived from a principal component analysis of multiple uniformly processed RNA degradation datasets (Agarwal and Kelley, 2022). The idea is to use this regressor as a heuristic to distill scalar half-life variant effect predictions from Borzuki's many output tracks. First, we assume access to Borzuki's predicted exon-aggregated log-coverage statistics for all N human genes of the original training dataset $\boldsymbol{u} \in \mathbb{R}^{N \times T}$. We fit a linear regression model f to predict Saluki's PC1 half-life measurements $\boldsymbol{v} \in \mathbb{R}^{N}$ given $\boldsymbol{u}$ as input features. Using this model, we can predict a scalar effect $z \in \mathbb{R}$ for a given MPRA variant as:

$$\hat{z} = (1/K) \times \sum_{k=1}^{K} \left(f\left(\boldsymbol{u}^{\text{alt}}_{k}\right) - f\left(\boldsymbol{u}^{\text{ref}}_{k}\right)\right) \tag{4.2}$$

In other words, the variant scores derived from each RNA-seq track t are scaled by the regression weight $\boldsymbol{w}_t$ corresponding to that track's importance for predicting the Saluki measurement, averaged across K contexts.

4.19.5. Approach 2 - Fit to MPRA reference values:

This method is simpler than Approach 1, but requires access to individual MPRA measurements of the reference oligo sequences. To this end, let $\boldsymbol{v}^{\text{ref}} \in \mathbb{R}^{N}$ be the log-transformed MPRA reference values (e.g. reference oligo log-decay rates), and let $\boldsymbol{U}^{\text{ref}} \in \mathbb{R}^{K \times N \times T}$ be the log-coverage predictions computed for all N oligos. For a given genomic context k, the oligos are all inserted into the same 3′ UTR. As a result, the columns in $\boldsymbol{U}^{\text{ref}}_{k}$ should correlate with $\boldsymbol{v}^{\text{ref}}$, although they will be offset by a constant specific to the particular gene k into which we inserted the oligos. Hence, for each context k, we fit a linear regression model f_k to predict $\boldsymbol{v}^{\text{ref}}$ using $\boldsymbol{U}^{\text{ref}}_{k}$ as input features:

$$f_k\left(\boldsymbol{U}^{\text{ref}}_{k,n}\right) = \left(\sum_{t=1}^{T} \boldsymbol{w}_{k,t} \times \boldsymbol{U}^{\text{ref}}_{k,n,t}\right) + \boldsymbol{w}_{k,0}, \quad \boldsymbol{v}^{\text{ref}}_{n} \sim f_k\left(\boldsymbol{U}^{\text{ref}}_{k,n}\right) + \epsilon, \quad 1 \le n \le N \tag{4.3}$$

This model is more flexible than Approach 1, as it learns a specific weight $\boldsymbol{w}_{k,t}$ for a given context k and RNA-seq output track t. The model might learn to rely on a particularly salient cell type (e.g. K562) for gene context k, while for another context j it might rely on different cell types (for example, if context j is not expressed in K562 cells).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Finally, to make variant effect prediction $z \in \mathbb{R}$ for a given variant, we calculate the difference in weighted log-coverage, averaged across the K linear models:

$$\hat{z} = (1/K) \times \sum_{k=1}^{K} \left(f_k(\boldsymbol{u}_k^{\text{alt}}) - f_k(\boldsymbol{u}_k^{\text{ref}}) \right) \tag{4.4}$$

For ensemble predictions, we re-compute and average the predicted effect z across multiple Borzuki replicates.

References

V. Agarwal and D. R. Kelley. The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biology*, 23(1):245, 2022. doi: 10.1186/s13059-022-02811-x.

B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, 2015. doi: 10.1038/nbt.3300.

M. Ascano, M. Hafner, P. Cekan, S. Gerstberger, and T. Tuschl. Identification of RNA–protein interaction networks using PAR-CLIP. *WIREs RNA*, 3(2):159–177, 2012. doi: 10.1002/wrna.1103.

Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D. R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, 2021. doi: 10.1038/s41592-021-01252-x.

G. Benegas, G. Eraslan, and Y. S. Song. Benchmarking DNA Sequence Models for Causal Regulatory Variant Prediction in Human Genetics. *bioRxiv*, 2025. doi: 10.1101/2025.02.11.637758.

J. Chen, Z. Hu, S. Sun, Q. Tan, Y. Wang, Q. Yu, L. Zong, L. Hong, J. Xiao, T. Shen, et al. Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function Predictions. *arXiv preprint arXiv:2204.00300*, 2022. doi: 10.48550/arXiv.2204.00300.

Z. Chen. DLS5-Omics/multimolecule: v0.2.1. Zenodo, 2026. Software, version 0.2.1.

N. H. Cho, K. C. Cheveralls, A.-D. Brunner, K. Kim, A. C. Michaelis, P. Raghavan, H. Kobayashi, L. Savy, J. Y. Li, H. Canaj, et al. OpenCell: Endogenous tagging for the cartography of human cellular organization. *Science*, 375(6585):eabi6983, 2022. doi: 10.1126/science.abi6983.

Y. Chu, D. Yu, Y. Li, K. Huang, Y. Shen, L. Cong, J. Zhang, and M. Wang. A 5′ UTR language model for decoding untranslated regions of mRNA and function predictions. *Nature Machine Intelligence*, 6(4):449–460, 2024. doi: 10.1038/s42256-024-00823-9.

Z. Cieniková, F. F. Damberger, J. Hall, F. H.-T. Allain, and C. Maris. Structural and Mechanistic Insights into Poly(uridine) Tract Recognition by the hnRNP C RNA Recognition Motif. *Journal of the American Chemical Society*, 136(41):14536–14544, 2014. doi: 10.1021/ja507690d.

Z. Cienikova, F. Damberger, J. Hall, F.-T. Allain, and C. Maris. Solution structure of hnRNP C RRM in complex with 5'-AUUUUUC-3' RNA. Protein Data Bank (PDB ID: 2MXY), 2015.

K. Darty, A. Denise, and Y. Ponty. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15):1974–1975, 2009. doi: 10.1093/bioinformatics/btp250.

R. D. Dowell and S. R. Eddy. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, 5(1):71, 2004. doi: 10.1186/1471-2105-5-71.

S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11): 2079–2088, 1994. doi: 10.1093/nar/22.11.2079.

S. Feng, S. Sekine, V. Pessino, H. Li, M. D. Leonetti, and B. Huang. Improved split fluorescent proteins for endogenous protein labeling. *Nature Communications*, 8(1):370, 2017. doi: 10.1038/s41467-017-00494-8.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

L. Fish, M. Khoroshkin, A. Navickas, K. Garcia, B. Culbertson, B. Hänisch, S. Zhang, H. C. B. Nguyen, L. M. Soto, M. Dermit, et al. A prometastatic splicing program regulated by SNRPA1 interactions with structured RNA elements. *Science*, 372(6543):eabc7531, 2021. doi: 10.1126/science.abc7531.

F. Gebauer, T. Schwarzl, J. Valcárcel, and M. W. Hentze. RNA-binding proteins in human genetic disease. *Nature Reviews Genetics*, 22(3):185–198, 2021. doi: 10.1038/s41576-020-00302-y.

S. Gerstberger, M. Hafner, and T. Tuschl. A census of human RNA-binding proteins. *Nature Reviews Genetics*, 15(12):829–845, 2014. doi: 10.1038/nrg3813.

H. Goodarzi, H. S. Najafabadi, P. Oikonomou, T. M. Greco, L. Fish, R. Salavati, I. M. Cristea, and S. Tavazoie. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature*, 485 (7397):264–268, 2012. doi: 10.1038/nature11013.

H. Goodarzi, S. Zhang, C. G. Buss, L. Fish, S. Tavazoie, and S. F. Tavazoie. Metastasis-suppressor transcript destabilization through TARBP2 binding of mRNA hairpins. *Nature*, 513(7517):256–260, 2014. doi: 10.1038/nature13466.

M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano, A.-C. Jungkamp, M. Munschauer, et al. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, 141(1):129–141, 2010. doi: 10.1016/j.cell.2010.03.009.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, 2016. IEEE. doi: 10.1109/CVPR.2016.90.

M. W. Hentze, A. Castello, T. Schwarzl, and T. Preiss. A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology*, 19(5):327–341, 2018. doi: 10.1038/nrm.2017.130.

I. Kalvari, E. P. Nawrocki, N. Ontiveros-Palacios, J. Argasinska, K. Lamkiewicz, M. Marz, S. Griffiths-Jones, C. Toffano-Nioche, D. Gautheret, Z. Weinberg, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, 49(D1):D192–D200, 2021. doi: 10.1093/nar/gkaa1047.

H. Kazan, D. Ray, E. T. Chan, T. R. Hughes, and Q. Morris. RNAcontext: A New Method for Learning the Sequence and Structure Binding Preferences of RNA-Binding Proteins. *PLoS Computational Biology*, 6(7): e1000832, 2010. doi: 10.1371/journal.pcbi.1000832.

D. R. Kelley, Y. A. Reshef, M. Bileschi, D. Belanger, C. Y. McLean, and J. Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–750, 2018. doi: 10.1101/gr.227819.117.

N. Kerimov, J. D. Hayhurst, K. Peikova, J. R. Manning, P. Walter, L. Kolberg, M. Samoviča, M. P. Sakthivel, I. Kuzmin, S. J. Trevanion, et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nature Genetics*, 53(9):1290–1299, 2021. doi: 10.1038/s41588-021-00924-w.

J. König, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D. J. Turner, N. M. Luscombe, and J. Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Structural & Molecular Biology*, 17(7):909–915, 2010. doi: 10.1038/nsmb.1838.

M. H. Kowalski, H.-H. Wessels, J. Linder, C. Dalgarno, I. Mascio, S. Choudhary, A. Hartman, Y. Hao, A. Kundaje, and R. Satija. Multiplexed single-cell characterization of alternative polyadenylation regulators. *Cell*, 187 (16):4408–4425.e23, 2024. doi: 10.1016/j.cell.2024.06.005.

N. Lambert, A. Robertson, M. Jangi, S. McGeary, P. A. Sharp, and C. B. Burge. RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Molecular Cell*, 54 (5):887–900, 2014. doi: 10.1016/j.molcel.2014.04.016.

S. J. Lange, D. Maticzka, M. Möhl, J. N. Gagnon, C. M. Brown, and R. Backofen. Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Research*, 40(12):5215–5226, 2012. doi: 10.1093/nar/gks181.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

K. U. Laverty, A. Jolma, S. E. Pour, H. Zheng, D. Ray, Q. Morris, and T. R. Hughes. PRIESSTESS: interpretable, high-performing models of the sequence and structure preferences of RNA-binding proteins. *Nucleic Acids Research*, 50(19):e111–e111, 2022. doi: 10.1093/nar/gkac694.

D. D. Licatalosi, A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. W. Chi, T. A. Clark, A. C. Schweitzer, J. E. Blume, X. Wang, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456 (7221):464–469, 2008. doi: 10.1038/nature07488.

Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

J. Linder, D. Srivastava, H. Yuan, V. Agarwal, and D. R. Kelley. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nature Genetics*, 57(4):949–961, 2025. doi: 10.1038/s41588-024-02053-6.

N. Liu, Q. Dai, G. Zheng, C. He, M. Parisien, and T. Pan. N6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature*, 518(7540):560–564, 2015. doi: 10.1038/nature14234.

J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, et al. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013. doi: 10.1038/ng.2653.

R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011. doi: 10.1186/1748-7188-6-26.

D. Maticzka, S. J. Lange, F. Costa, and R. Backofen. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biology*, 15(1):R17, 2014. doi: 10.1186/gb-2014-15-1-r17.

M. Melé, K. Mattioli, W. Mallard, D. M. Shechner, C. Gerhardinger, and J. L. Rinn. Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Research*, 27(1):27–37, 2017. doi: 10.1101/gr.214205.116.

S. Noguchi, T. Arakawa, S. Fukuda, M. Furuno, A. Hasegawa, F. Hori, S. Ishikawa-Kato, K. Kaida, A. Kaiho, M. Kanamori-Katayama, et al. FANTOM5 CAGE profiles of human and mouse samples. *Scientific Data*, 4 (1):170112, 2017. doi: 10.1038/sdata.2017.112.

S. Park, S. H. Ahn, E. S. Cho, Y. K. Cho, E.-S. Jang, and S. W. Chi. CLIPick: a sensitive peak caller for expression-based deconvolution of HITS-CLIP signals. *Nucleic Acids Research*, 46(21):11153–11168, 2018. doi: 10.1093/nar/gky917.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

R. J. Penić, T. Vlašić, R. G. Huber, Y. Wan, and M. Šikić. RiNALMo: general-purpose RNA language models can generalize well on structure prediction tasks. *Nature Communications*, 16(1):5671, 2025. doi: 10.1038/s41467-025-60872-5.

M. Rabani, M. Kertesz, and E. Segal. Computational prediction of RNA structural motifs involved in post-transcriptional regulatory processes. *Proceedings of the National Academy of Sciences*, 105(39):14885–14890, 2008. doi: 10.1073/pnas.0803169105.

D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457): 172–177, 2013. doi: 10.1038/nature12311.

Y. Ren, Z. Chen, L. Qiao, H. Jing, Y. Cai, S. Xu, P. Ye, X. Ma, S. Sun, H. Yan, et al. BEACON: Benchmark for Comprehensive RNA Tasks and Language Models. *arXiv preprint arXiv:2406.10391*, 2024. doi: 10.48550/ARXIV.2406.10391.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

A. Sasse, D. Ray, K. U. Laverty, C. L. Tam, M. Albu, H. Zheng, Y. Levdansky, O. Lyudovyk, T. Dalal, K. Nie, et al. A resource of RNA-binding protein motifs across eukaryotes reveals evolutionary dynamics and gene-regulatory function. *Nature Biotechnology*, 44(7):1226–1236, 2026. doi: 10.1038/s41587-025-02733-6.

D. B. Searls. The language of genes. *Nature*, 420(6912):211–217, 2002. doi: 10.1038/nature01255.

A. Shrikumar, P. Greenside, and A. Kundaje. Learning Important Features Through Propagating Activation Differences. *arXiv preprint arXiv:1704.02685*, 2019. doi: 10.48550/arXiv.1704.02685.

D. A. Siegel, O. Le Tonqueze, A. Biton, N. Zaitlen, and D. J. Erle. Massively parallel analysis of human 3′ UTRs reveals that AU-rich element length and registration predict mRNA destabilization. *G3: Genes, Genomes, Genetics*, 12(1):jkab404, 2022. doi: 10.1093/g3journal/jkab404.

The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012. doi: 10.1038/nature11247.

J. Ule, K. B. Jensen, M. Ruggiu, A. Mele, A. Ule, and R. B. Darnell. CLIP Identifies Nova-Regulated RNA Networks in the Brain. *Science*, 302(5648):1212–1215, 2003. doi: 10.1126/science.1090095.

E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, et al. 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.

E. L. Van Nostrand, P. Freese, G. A. Pratt, X. Wang, X. Wei, R. Xiao, S. M. Blue, J.-Y. Chen, N. A. L. Cody, D. Dominguez, et al. 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.

N. Wang, J. Bian, Y. Li, X. Li, S. Mumtaz, L. Kong, and H. Xiong. Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence*, 6(5):548–557, 2024. doi: 10.1038/s42256-024-00836-4.

B. J. Zarnegar, R. A. Flynn, Y. Shen, B. T. Do, H. Y. Chang, and P. A. Khavari. irCLIP platform for efficient characterization of protein–RNA interactions. *Nature Methods*, 13(6):489–492, 2016. doi: 10.1038/nmeth.3840.

Y. Zhang, M. Lang, J. Jiang, Z. Gao, F. Xu, T. Litfin, K. Chen, J. Singh, X. Huang, G. Song, et al. Multiple sequence alignment-based RNA language model and its application to structural inference. *Nucleic Acids Research*, 52(1):e3–e3, 2024. doi: 10.1093/nar/gkad1031.

J. Zhou and O. G. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, 2015. doi: 10.1038/nmeth.3547.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

A high-resolution protein-RNA binding atlas for structure-aware interaction prediction and post-transcriptional regulatory modeling

Supplementary Material

A Acknowledgements 2

B Funding 2

C Data, code, and materials availability 2

D Supplementary figures 3

Figure S1. Optimization and validation of the eclair CLIP workflow. 3

Figure S2. Extended ecCLIP quality metrics and Borzuki validation. 5

Figure S3. Performance of AI models in predicting RBP binding. 6

Figure S4. Extended MPRC support for Pythia-nominated structure-dependent RBP binding. 7

Figure S5. Predicted-effect stratification and high-leverage SNPs for RBP-specific Pythia variant-effect models. 8

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

A. Acknowledgements

We thank A. Pawluk and C. Ricci-Tam for reading earlier versions of this manuscript and helping us with the content. We thank Albertas Navickas for input on eclair CLIP method development.

B. Funding

L.F. was a Zena Werb Scholar supported by T32CA108462. H.K. was a Zena Werb Scholar supported by T32CA108462. Sequencing was performed at the UCSF Center for Advanced Technology (CAT), supported by UCSF PBBR, RRP IMIA, and NIH 1S10OD028511-01 grants. H.S.N. is supported by the Canadian Institutes of Health Research (CIHR) grant PJT-173317 and resource allocations from Digital Research Alliance of Canada, and holds a CIHR Canada Research Chair. E.M.S. and M.D.L. are supported by Biohub. H.G. is a Core Investigator at Arc Institute, and this research was supported by Arc.

C. Data, code, and materials availability

All data and code needed to evaluate and reproduce the results in the paper are present in the paper and/or the Supplementary Materials.

- Pythia codebase is available at https://github.com/goodarzilab/pythia.
- Pre-trained Borzuki model weights and code are available at https://github.com/calico/borzoi-paper/tree/main/extensions/borzuki.
- All raw and processed sequencing data are available from the Gene Expression Omnibus (GEO) under accession number GSE349934.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

D. Supplementary figures

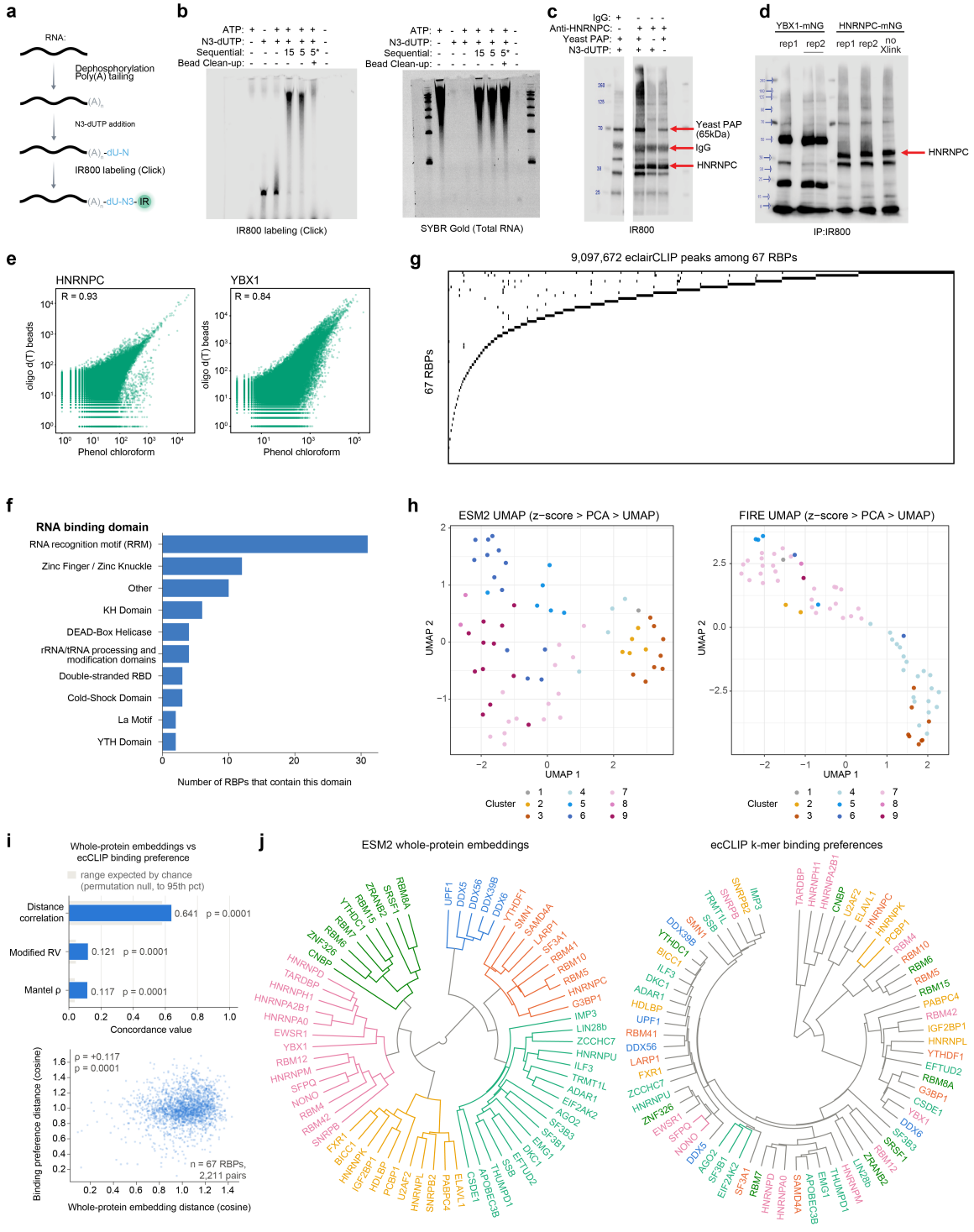


Figure S1 | **Optimization and validation of the eclair CLIP workflow.** Legend continued on the next page.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Figure S1 | **Optimization and validation of the eclair CLIP workflow. (a)** Schematic of eclair CLIP RNA labeling strategy. **(b)** An RNA oligo was subjected to poly(A) tailing with Yeast PAP, N_3-dUTP azide incorporation, and IR800 click labeling. Denaturing PAGE comparing labeling efficiency across conditions varying ATP, N_3-dUTP concentration, sequence and timing of steps, and bead cleanup, visualized by IR800 (left) and SYBR Gold stain (right). Sequential (-) indicates co-addition of ATP and N_3-dUTP. Sequential (+) indicates ATP addition followed by N_3-dUTP addition with the number indicating time in minutes of N_3-dUTP incubation. Sequential* indicates ATP addition followed by SPRI clean-up followed by N_3-dUTP incubation for 5 minutes. **(c)** Blot showing immunoprecipitation of HNRNPC from UV-crosslinked HEK293T cells using an anti-HNRNPC antibody or anti-mouse IgG. Shown are IgG control compared to anti-HNRNPC pulldown with and without yeast PAP-mediated poly(A) tailing and N_3-dUTP labeling. All labeled with DBCO-IR800. Bands corresponding to yeast PAP (65 kDa), IgG, and HNRNPC are indicated. **(d)** Blot showing immunoprecipitation of split mNeonGreen-tagged RBPs (YBX1 and HNRNPC) from HEK293T cells. A neon green nanobody was used to immunoprecipitate and the ecCLIP workflow used to label the complexes. A no-crosslink control is included. **(e)** Comparison of RNA isolation methods for eclair CLIP. Scatter plots show per-gene read counts for RNA isolated by oligo d(T) magnetic beads versus phenol chloroform extraction for HNRNPC (Pearson R = 0.93) and YBX1 (Pearson R = 0.84), demonstrating high concordance between methods. **(f)** Distribution of RNA binding domains among proteins in the ecCLIP dataset. Some proteins contain multiple RNA binding domain types. **(g)** Distribution of ecCLIP peaks across 67 RBPs. Dot plot shows the number of peaks called per RBP, ranked by peak count. A total of 9,097,672 peaks were identified across all 67 RBPs. **(h)** UMAP projections of ESM2 protein embeddings (left) and ecCLIP-derived sequence preferences labeled FIRE (right), following z-score normalization and principal component analysis. Points represent RBPs and are colored by clusters within each representation. **(i)** Whole-protein version of the analysis in Fig. 1c: concordance statistics between ESM2 whole-protein embeddings and ecCLIP-derived binding preferences, shown against the permutation-null range (grey bands), with the corresponding pairwise cosine-distance scatter (2,211 RBP pairs). **(j)** Circular dendrograms as in Fig. 1d, built from whole-protein embeddings (left) and ecCLIP-derived binding preferences (right), both colored by the protein-side grouping.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

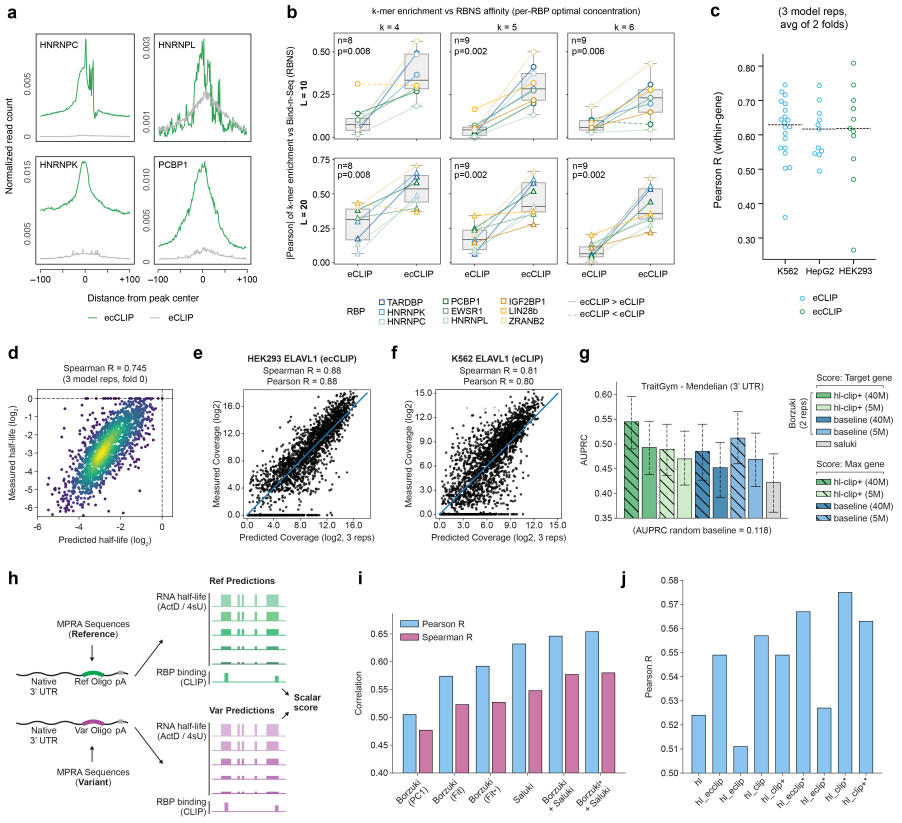


Figure S2 | **Extended ecCLIP quality metrics and Borzuki validation.** **(a)** Additional aggregate profiles comparing ecCLIP and eCLIP read signal centered on peak summits for matched RBPs. **(b)** Robustness of k-mer concordance to RBNS concentration choice. The absolute Pearson correlation between CLIP-derived k-mer enrichment and RBNS affinity is shown using the optimal RBNS protein concentration for each RBP, rather than the fixed 1300 nM concentration used in Fig. 2c. Results are shown for k = 4, 5, and 6 and peak-centered windows of 10 or 20 nucleotides. Lines connect matched RBPs between eCLIP and ecCLIP; solid lines indicate stronger concordance for ecCLIP and dashed lines indicate stronger concordance for eCLIP. P-values correspond to paired comparisons between eCLIP and ecCLIP. **(c)** Test set Pearson R when predicting CLIP-seq profiles at bin-level resolution within genes for held-out test data, shown for both ecCLIP and ENCODE eCLIP tracks. **(d)** Predicted versus measured mRNA degradation rates derived from a K562 actinomycin D transcriptional shutoff experiment. Predicted half-lives were derived by fitting exponential decay curves to Borzuki-predicted exon-aggregated coverage across time points. Spearman R = 0.745 with experimental measurements; Spearman R = 0.728 with Saluki's cell type-agnostic PC1 half-life summary. **(e)** Predicted versus measured gene-level exon-aggregated log-coverage for the ELAVL1 CLIP-seq experiment in the ecCLIP dataset. **(f)** Predicted versus measured gene-level exon-aggregated log-coverage for an ENCODE eCLIP ELAVL1 experiment. **(g)** Comparison between Saluki and Borzuki when classifying putative causal 3′ UTR variants associated with Mendelian traits from TraitGym (Benegas et al., 2025). Variant scoring is compared using the annotated target gene or the gene context with the largest predicted effect. **(h)** Illustration of variant scoring procedure when predicting mRNA half-life effects measured in an MPRA experiment in Beas2B cells (Siegel et al., 2022). Reference and variant oligos are inserted in endogenous 3′ UTRs and their predicted effect on RNA half-life and CLIP binding is calculated using Borzuki. For each RNA-seq or CLIP-seq track predicted by the model, the total log fold-change in coverage across the gene's exons due to the mutation is calculated. These functional multi-track scores are combined into a scalar variant impact using a secondary regressor model, which is fitted either on reference measurements from the MPRA assay or PCA values from the Saluki model's training data (Methods). **(i)** Comparison between Saluki and Borzuki at the task of predicting half-life MPRA variant effects ('PC1' / 'Fit' = Borzuki's multi-track scores are either fit to Saluki's PC1 measurements or to the MPRA's reference values, '*' = variant scores are derived from CLIP-seq tracks in addition to half-life RNA-seq tracks). **(j)** Correlation metrics when predicting MPRA-derived half-life variant effects using versions of Borzuki models trained on subsets of training data ('hl' = trained on base ENCODE/FANTOM5 DNase/RNA-seq/CAGE tracks and RNA half-life time-course tracks, 'hl_ecclip' = additionally trained on ecCLIP experiments, 'hl_eclip' = additionally trained on ENCODE eCLIP, 'hl_clip' = trained on ecCLIP and eCLIP tracks, 'hl_clip+' = trained on all CLIP-seq experiments and polyadenylation Perturb-seq. '*' = variant scores are derived from CLIP-seq tracks in addition to half-life RNA-seq tracks).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

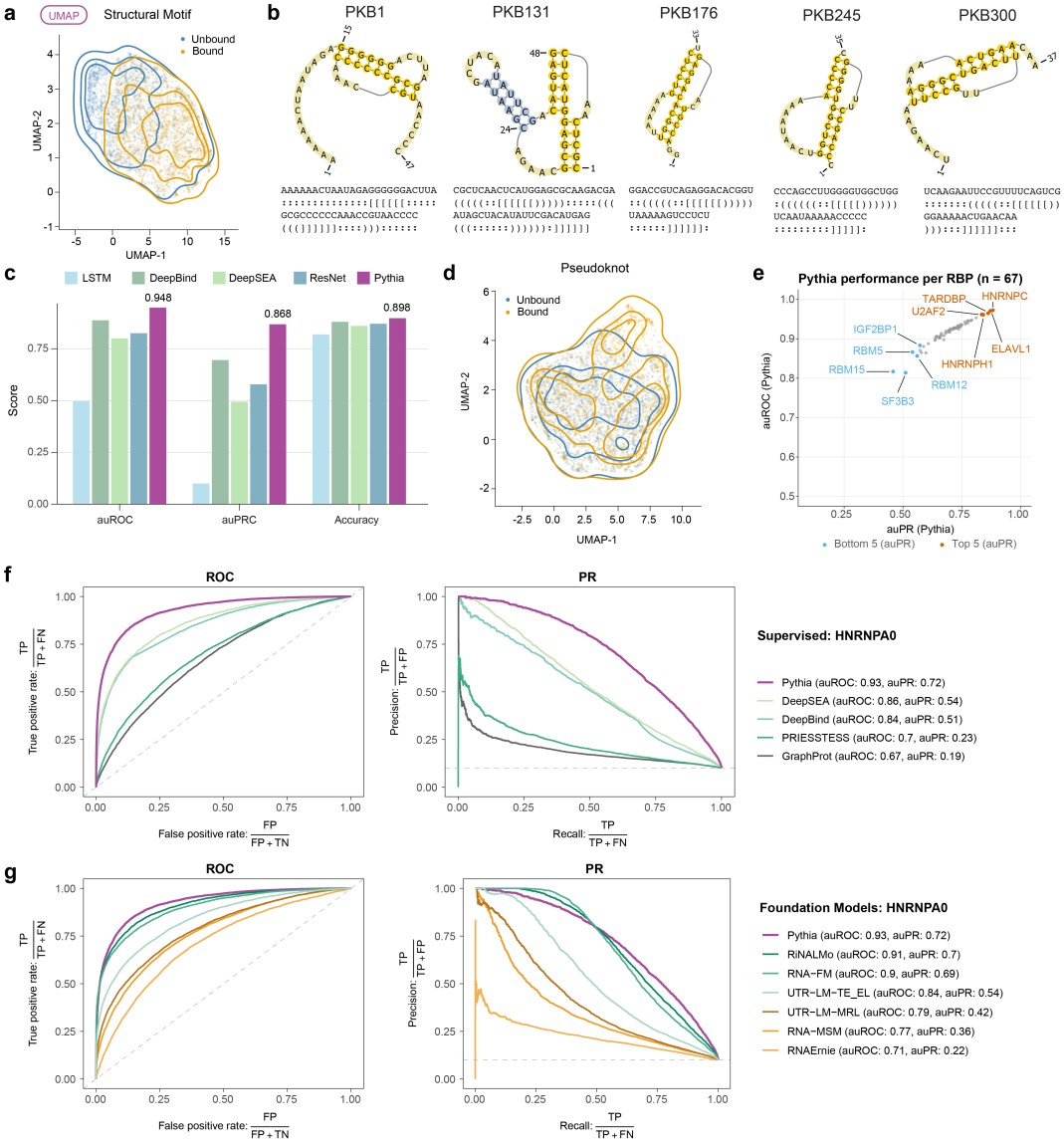


Figure S3 | **Performance of AI models in predicting RBP binding. (a)** Generalization to sequence-independent motifs. UMAP projection of Pythia embeddings for held-out synthetic sequences with or without the CFG-defined structural motif. The two groups overlap; this visualization alone does not establish motif classification accuracy. **(b)** Pseudoknot structure prediction. We performed in silico mutations on five distinct pseudoknot structures, generating positive examples that preserved base-pairings and negative examples that disrupted them. **(c)** Pythia distinguished preserved from disrupted pseudoknot structures with an AUROC = 0.95 and AUPRC = 0.87, outperforming other models (LSTM, DeepBind, DeepSEA, ResNet) and demonstrating structural sensitivity. **(d)** Latent space organization. UMAP projection of Pythia embeddings for preserved and disrupted pseudoknot examples. The overlapping distributions are shown descriptively; classification performance is quantified in (c). **(e)** Scatterplot of Pythia auPR and auROC performance per RBP. **(f)** ROC (left) and precision-recall (PR) curves for supervised models predicting genome-wide binding of HNRNPA0. The y-axis of the ROC plot shows true positive rate [TP/(TP+FN)] and the x-axis shows false positive rate [FP/(FP+TN)]. The y-axis of the PR plot shows precision [TP/(TP+FP)] and the x-axis shows recall [TP/(TP+FN)]. Pythia (purple; auROC = 0.93, auPR = 0.72) outperforms DeepSEA (auROC = 0.86, auPR = 0.54), DeepBind (auROC = 0.84, auPR = 0.51), PRIESSTESS (auROC = 0.70, auPR = 0.23), and GraphProt (auROC = 0.67, auPR = 0.19). **(g)** Same as (f), showing performance of Pythia and the linear probing results of the RNA foundation models for HNRNPA0. Pythia (auROC = 0.93, auPR = 0.72) performs comparably to top foundation models RiNALMo (auROC = 0.91, auPR = 0.70) and RNA-FM (auROC = 0.90, auPR = 0.69), and outperforms UTR-LM-TE_EL (auROC = 0.84, auPR = 0.54), UTR-LM-MRL (auROC = 0.79, auPR = 0.42), RNA-MSM (auROC = 0.77, auPR = 0.36), and RNAErnie (auROC = 0.71, auPR = 0.22).

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

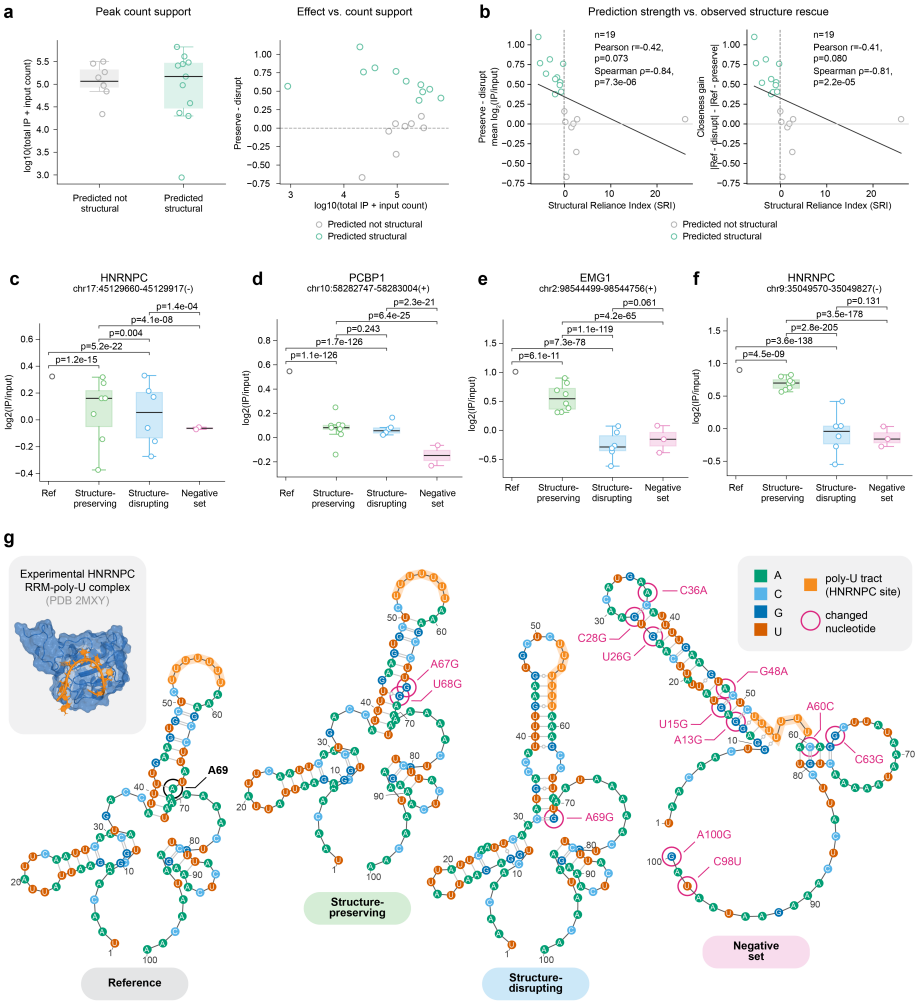


Figure S4 | **Extended MPRC support for Pythia-nominated structure-dependent RBP binding. (a)** Count support for structure-rescue summaries. Each point is one reference-enriched peak with Ref $\log_2$(IP/input) > 0.25. The left panel compares total peak counts, computed as summed IP plus input counts across constructs, between predicted not-structure-dependent and predicted structure-dependent groups on a $\log_{10}$ scale. The right panel plots the preserve-minus-disrupt rescue metric against total peak count to test whether observed rescue effects are concentrated in low- or high-count peaks. Points are colored by predicted structural class. **(b)** Pythia prediction strength versus observed structure-rescue metrics. Each point is one reference-enriched peak with Ref $\log_2$(IP/input) > 0.25. The x-axis shows a Structural Reliance Index (SRI), defined per peak as log ratio of sequence-arm to structure-arm DeepLIFT attribution under a zero-baseline reference, and the dashed vertical line marks the prediction threshold of −0.1 used to define predicted structure-dependent peaks. The left panel shows preserve-minus-disrupt, defined as mean structure-preserving enrichment minus mean structure-disrupting enrichment. The right panel shows closeness gain, defined as abs(Ref - Structure-disrupting) - abs(Ref - Structure-preserving). Positive y-values indicate a stronger structure-preserving rescue pattern. Points are colored by threshold-defined prediction group; fitted lines and Pearson/Spearman statistics summarize continuous-score associations. **(c–f)** Additional representative structure-variant validation examples for HNRNPC at chr17:45,129,660-45,129,917(-), PCBP1 at chr10:58,282,747-58,283,004(+), EMG1 at chr2:98,544,499-98,544,756(+), and HNRNPC at chr9:35,049,570-35,049,827(-). Each point is one construct summarized across replicates as $\log_2$(IP/input). Constructs are grouped as reference, structure-preserving, structure-disrupting, or negative-control reporters. Boxes are shown for groups with multiple constructs, overlaid open points show individual construct values, and brackets report pairwise binomial logistic-regression P-values where valid. **(g)** Structural context for the HNRNPC validation example. The experimentally determined HNRNPC RRM-poly-U complex (PDB 2MXY; Cieniková et al., 2014; Cienikova et al., 2015) is shown alongside sequence-based secondary-structure predictions for reference, structure-preserving, structure-disrupting, and negative-control constructs. The poly-U tract is highlighted in orange and altered nucleotides are circled. RNA structures were predicted with ViennaRNA and displayed with VARNA (Methods; Lorenz et al., 2011; Darty et al., 2009). These predictions provide structural context and are not direct measurements of reporter folding in cells.

bioRxiv preprint doi: https://doi.org/10.64898/2026.10.12.758580; this version posted October 13, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

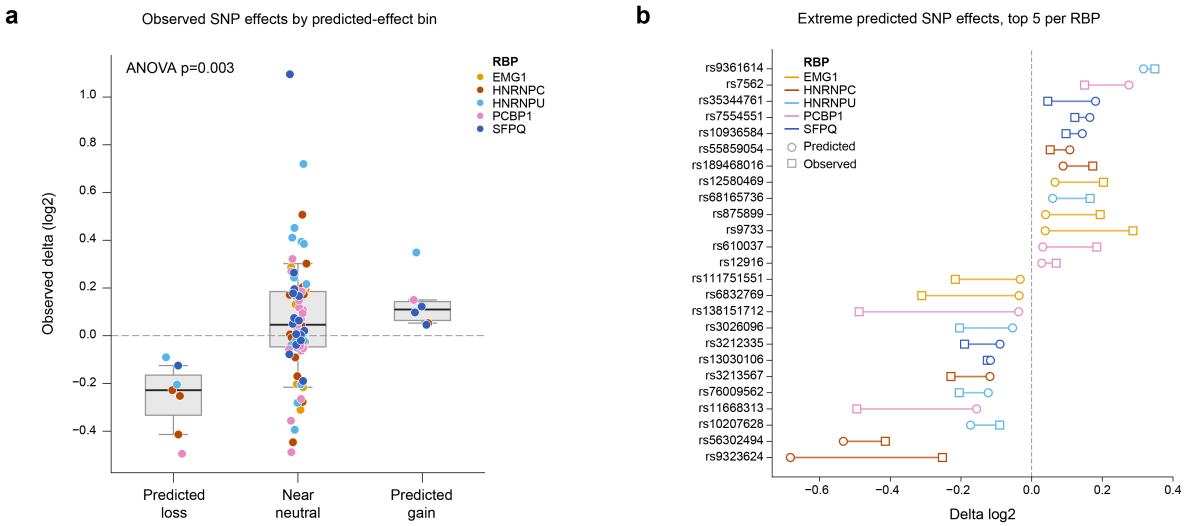


Figure S5 | **Predicted-effect stratification and high-leverage SNPs for RBP-specific Pythia variant-effect models. (a)** Observed SNP effects grouped by predicted-effect bin. SNPs with valid, converged model fits were binned by Pythia predicted change into predicted loss (delta < -0.1), near neutral ($-0.1 \leq$ delta ≤ 0.1), and predicted gain (delta > 0.1). The y-axis shows the observed logistic-regression SNP effect in $\log_2$ units. Boxes summarize each bin, with the box spanning the 25th to 75th percentiles, the median line marking the 50th percentile, and whiskers set to the 10th and 90th percentiles. Overlaid points show individual SNPs colored by RBP. The dashed horizontal line marks zero observed effect. The text box reports the one-way ANOVA P-value across the three observed-delta distributions. **(b)** Extreme predicted SNP effects. For each RBP, the top five SNPs by absolute Pythia v3 predicted effect are shown. Each horizontal segment connects the predicted SNP effect (circle) to the observed logistic-regression effect (square), both in $\log_2$ units. The y-axis shows SNP names, and segments and markers are colored by RBP. The vertical dashed line marks zero effect. This plot highlights high-leverage predicted SNPs and shows whether the observed direction and magnitude agree with the prediction.