Supplementary Figures

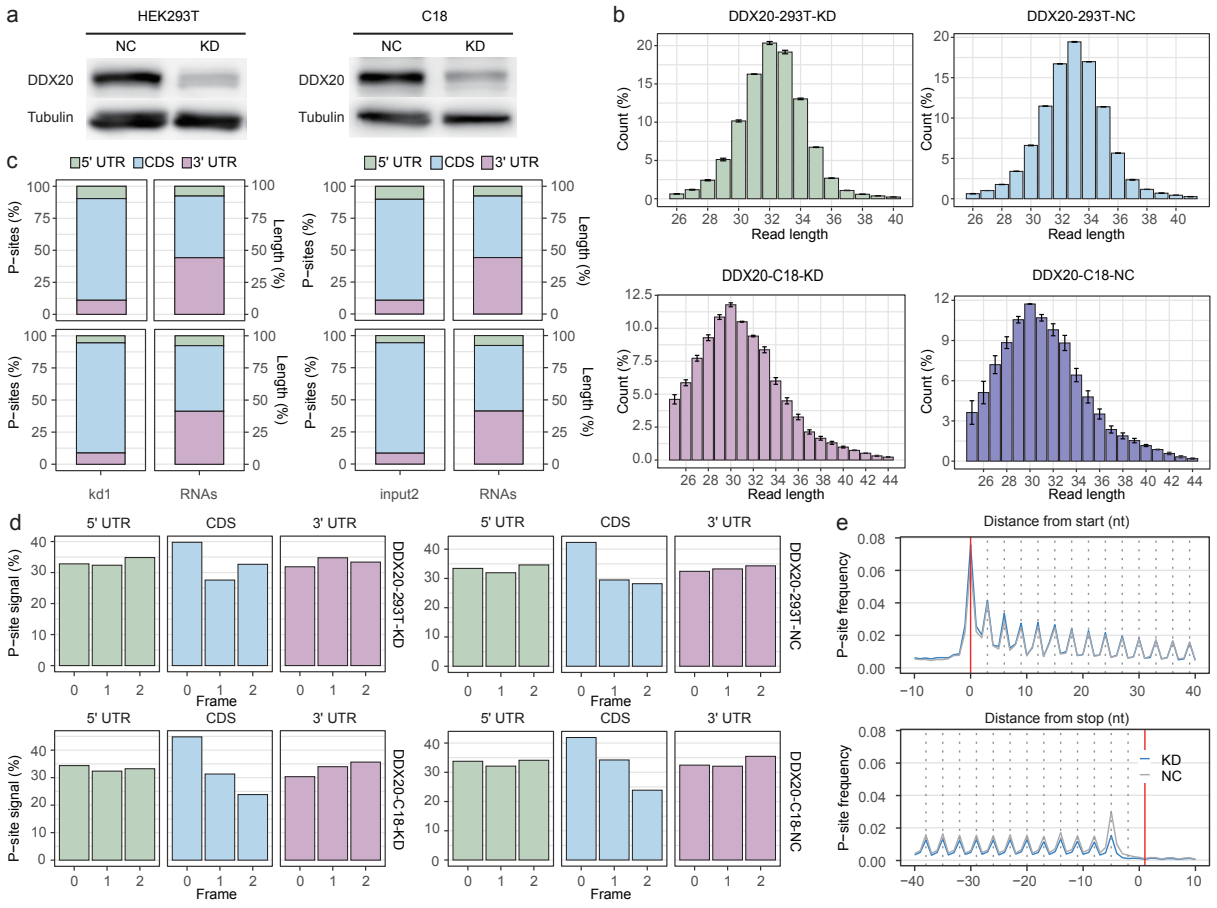

Fig. S1 Quality control and validation of DDX20 knockdown and ribosome profiling data. a, Western blot analysis confirming the loss of DDX20 protein expression in DDX20 knockdown (KD) compared with negative control (NC) HEK293T and C18 cell lines. **b,** Bar plot showing the length distribution of ribosome-protected fragments (RPFs) in our ribosome profiling data. **c,** Stacked bar plots showing the distribution of mapped ribosome profiling reads across 5′ untranslated regions (5′ UTR), coding sequences (CDS), and 3′ untranslated regions (3′ UTR). **d,** Bar plots showing the percentages of ribosome footprints aligned to each of the three reading frames, indicating the expected 3-nt periodicity. **e,** Overlay meta-profiles of ribosome footprint densities along mRNAs enriched in the DDX20 knockdown (KD) group (blue) and the control group (grey).

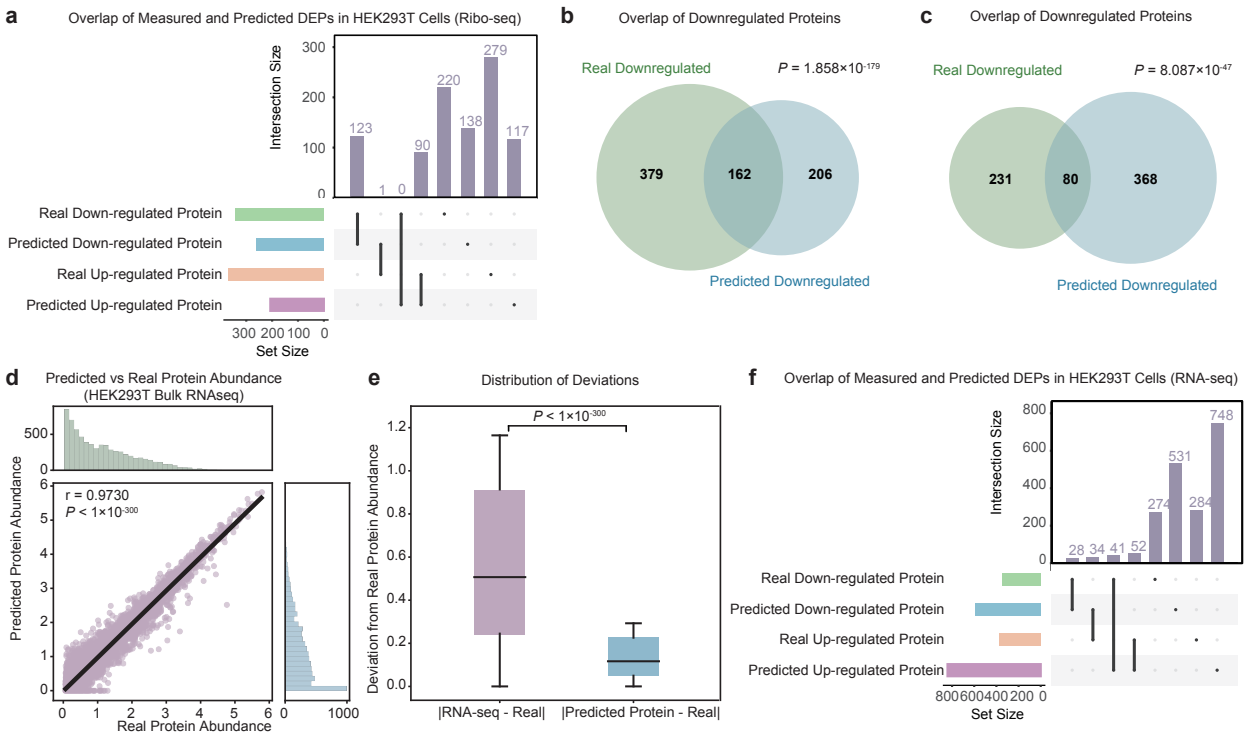

Fig. S2 Extended analyses of bulk protein prediction from Ribo-seq and RNA-seq. a, UpSet plot illustrating the overlap between measured and predicted DEPs in HEK293T cells based on Ribo-seq data. **b,** Venn diagram of downregulated DEPs in HEK293T cells reveals 162 overlapping proteins (P, hypergeometric test). **c,** Venn diagram shows 80 overlapping downregulated DEPs (P values from hypergeometric test). **d,** Predicted and measured protein abundance in DDX20 knockdown (KD) HEK293T cells using RNA-seq show a strong correlation, with marginal histograms illustrating the expression distributions (P value was computed using a two-sided t-test for Pearson correlation). **e,** Boxplots demonstrate that RECIPE substantially reduces the difference between predicted and measured protein abundance compared with raw RNA-seq (P value was computed using a one-sided Welch's t-test). **f,** UpSet plot illustrating the overlap between measured and predicted DEPs in HEK293T cells based on RNA-seq data.

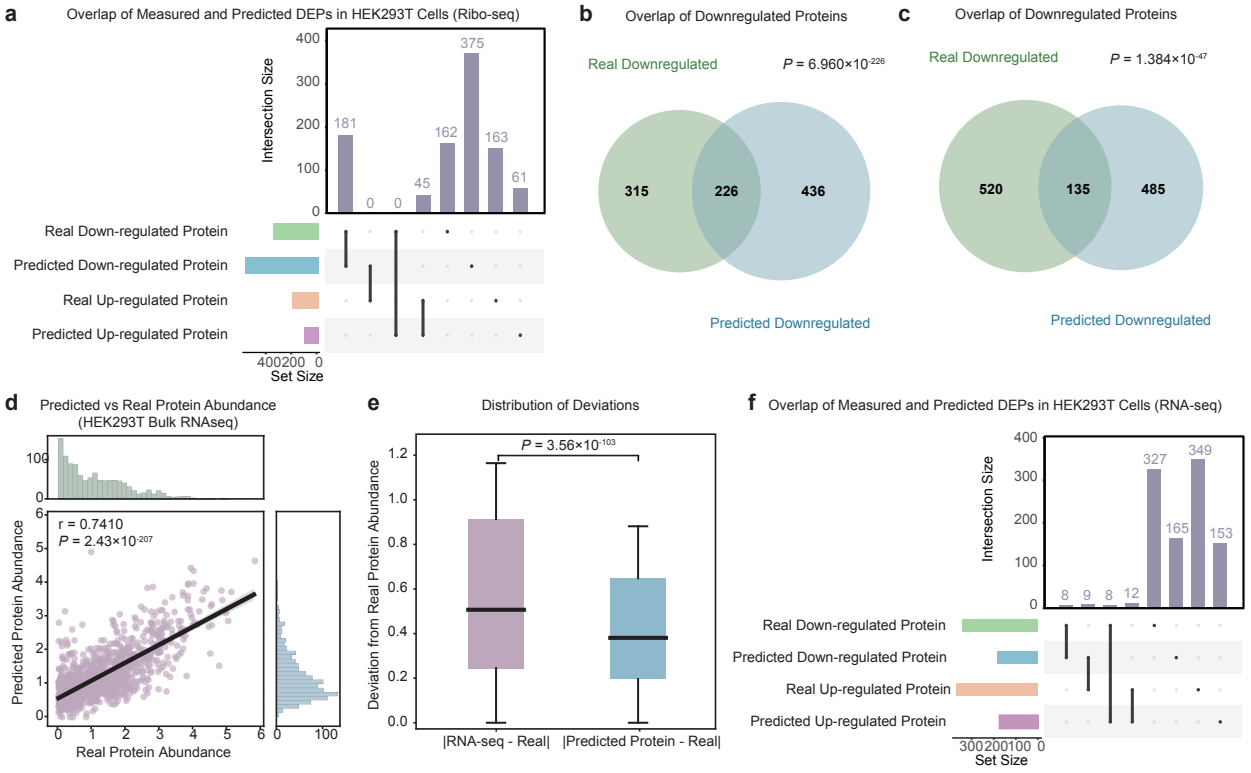

Fig. S3 RECIPE extrapolates unmeasured proteins and extends to protein prediction from RNA-seq. a, UpSet plot illustrating the overlap between measured and predicted DEPs in HEK293T cells based on Ribo-seq data. **b,** Venn diagram reveals 226 overlapping downregulated DEPs (P values from hypergeometric test). **c,** Venn diagram identifies 135 overlapping downregulated DEPs (P values from hypergeometric test). **d,** Predicted and real abundance of unmeasured proteins in DDX20 knockdown (KD) HEK293T cells using RNA-seq show a strong correlation, with marginal histograms illustrating the expression distributions (P value was computed using a two-sided t-test for Pearson correlation). **e,** Boxplots demonstrate that RECIPE substantially reduces the difference between predicted and real protein abundance for unmeasured proteins compared with raw RNA-seq (P value was computed using a one-sided Welch's t-test). **f,** UpSet plot illustrating the overlap between measured and predicted DEPs in HEK293T cells based on RNA-seq data.

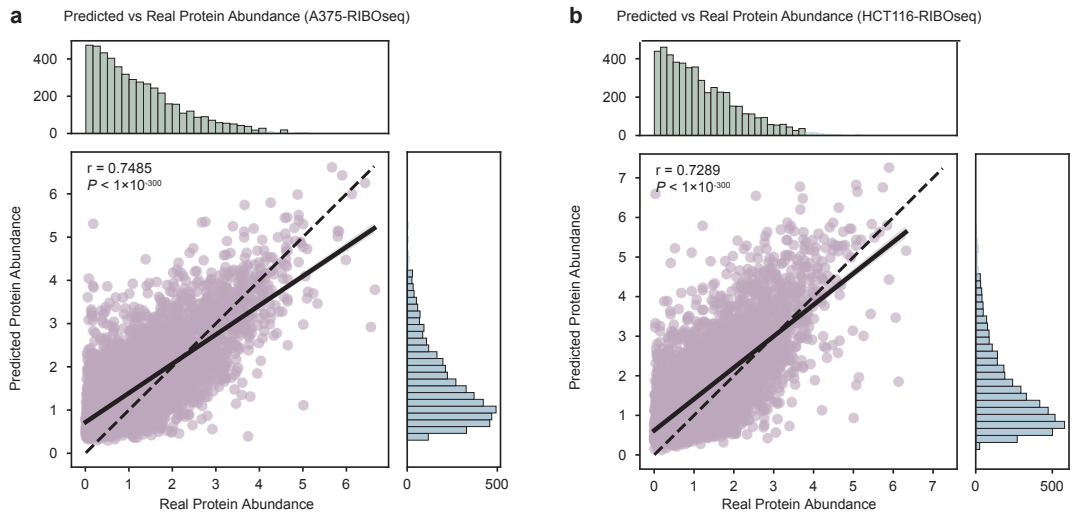

Fig. S4 RECIPE extrapolates to unmeasured proteins in additional Ribo-seq datasets. a, Predicted and measured protein abundance in A375 cells show strong correlation, with marginal histograms of expression distributions (P value was computed using a two-sided t-test for Pearson correlation). **b,** Predicted and measured protein abundance in HCT116 cells show strong correlation, with marginal histograms of expression distributions (P value was computed using a two-sided t-test for Pearson correlation).

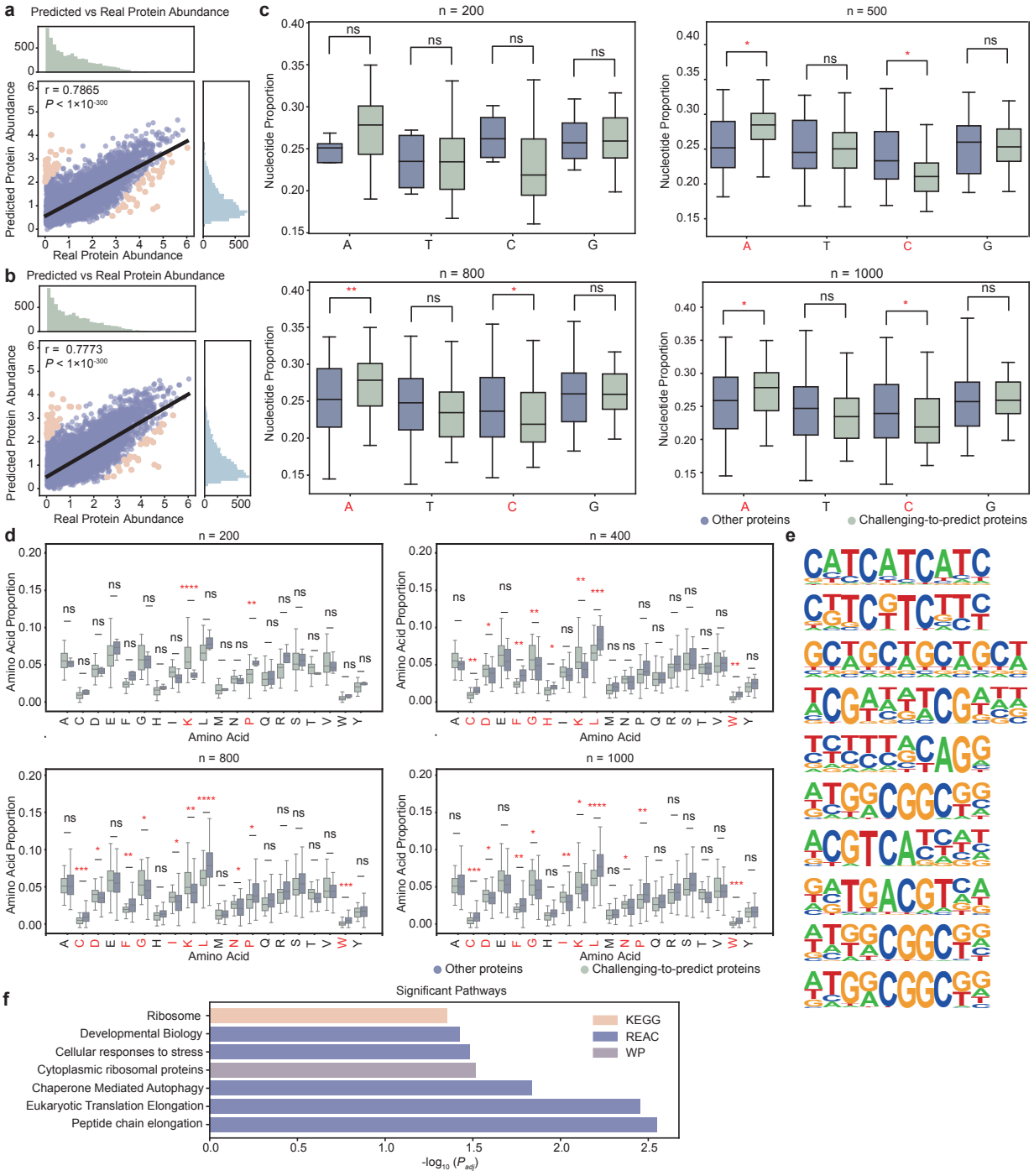

Fig. S5 Characterization of challenging-to-predict proteins reveals mechanisms beyond ribosome occupancy. a, Correlation between predicted and real protein abundance in DDX20 knockdown samples from model training with random seed 0, with proteins of lower prediction accuracy highlighted in yellow (P value was computed using a two-sided t-test for Pearson correlation). **b,** Correlation between predicted and real protein abundance in DDX20 knockdown samples from model training with random seed 12, with proteins of lower prediction accuracy highlighted in yellow (P value was computed using a two-sided t-test for Pearson correlation). **c,** Nucleotide composition differs between challenging-to-predict proteins and other proteins across different easy-protein cutoffs (200, 500, 800, and 1,000) (two-sided Welch's t-test; significance codes: **** ($P < 0.0001$), *** ($P < 0.001$), ** ($P < 0.01$), * ($P < 0.05$), and ns ($P \geq 0.05$)). **d,** Amino acid composition also distinguishes hard-to-predict proteins from easy-to-predict proteins across the same cutoff settings (two-sided Welch's t-test; significance as in c). **e,** Sequence logo representation of transcript sequences corresponding to challenging-to-predict proteins. **f,** Pathway enrichment analysis of challenging proteins (values shown as $-\log_{10}$ of adjusted P values from a one-sided hypergeometric test).

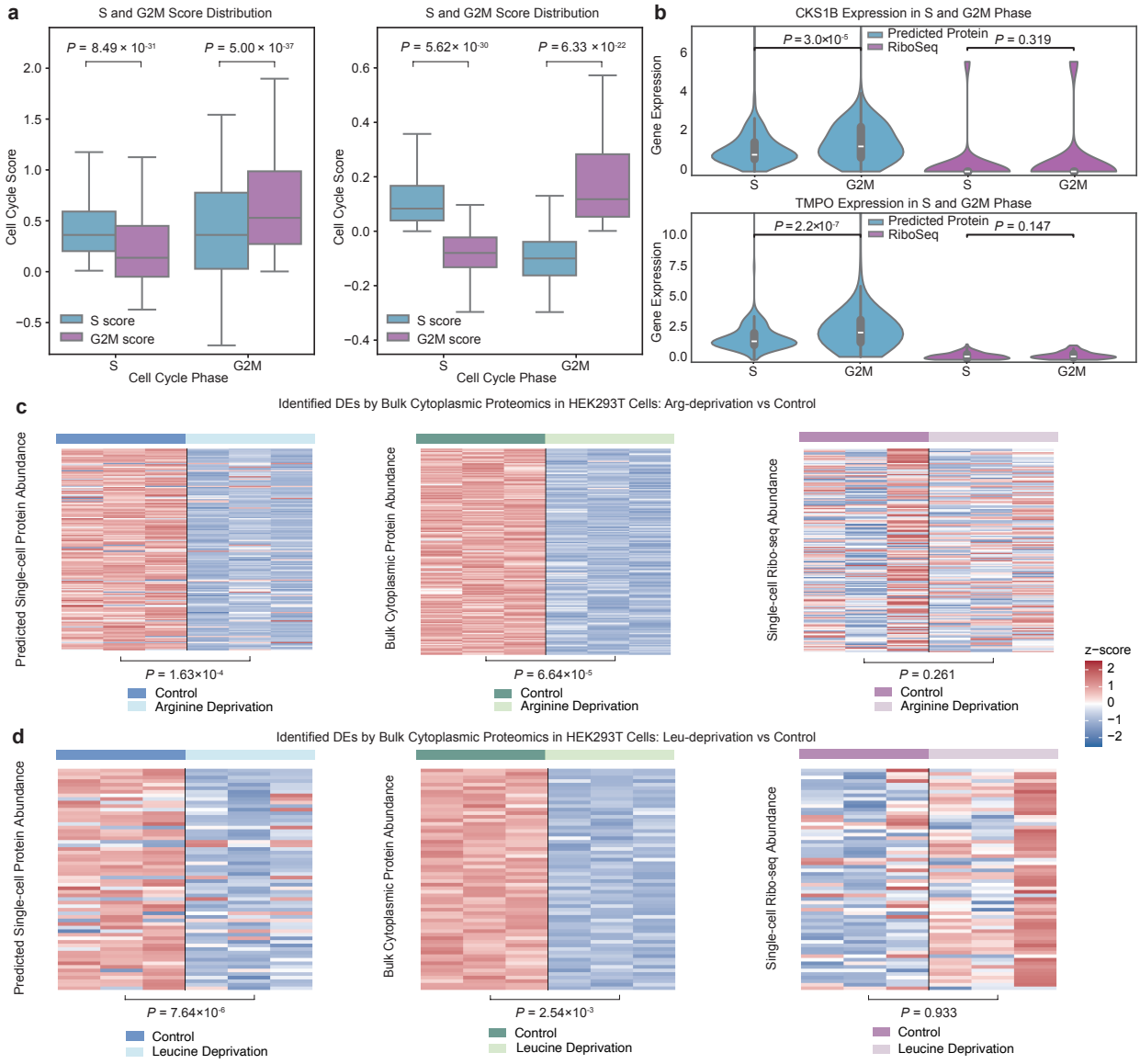

Fig. S6 cell-cycle analyses. a, Boxplots of S-phase and G2/M-phase scores from predicted protein abundance (left) versus Ribo-seq (right). Significance was assessed using a one-sided Wilcoxon rank-sum test. **b,** Violin plots showing distributions of G2/M-phase marker expression in predicted protein expression versus Ribo-seq data. Significance was assessed using a one-sided Wilcoxon rank-sum test. **c-d** Heatmaps of bulk-selected downregulated proteins across bulk proteomics, RECIPE-predicted single-cell protein abundance, and single-cell Ribo-seq under arginine (**c**) and leucine (**d**) deprivation. Proteins were selected from the corresponding bulk differential analysis (adjusted $P < 0.05$ and logFC < -1) and ordered by bulk logFC. The bulk panels show normalized expression across three amino acid deprivation replicates and three control replicates. The predicted and Ribo-seq panels show bootstrap replicates generated by sampling 80% of cells from the deprivation group and the rich group three times and calculating the mean expression for each subsample. Heatmaps display row-wise z-scored expression within each modality. The panel-level P value shown in each title was calculated using a one-sided Welch's t-test on the mean expression of the selected protein set between deprivation and control replicates.

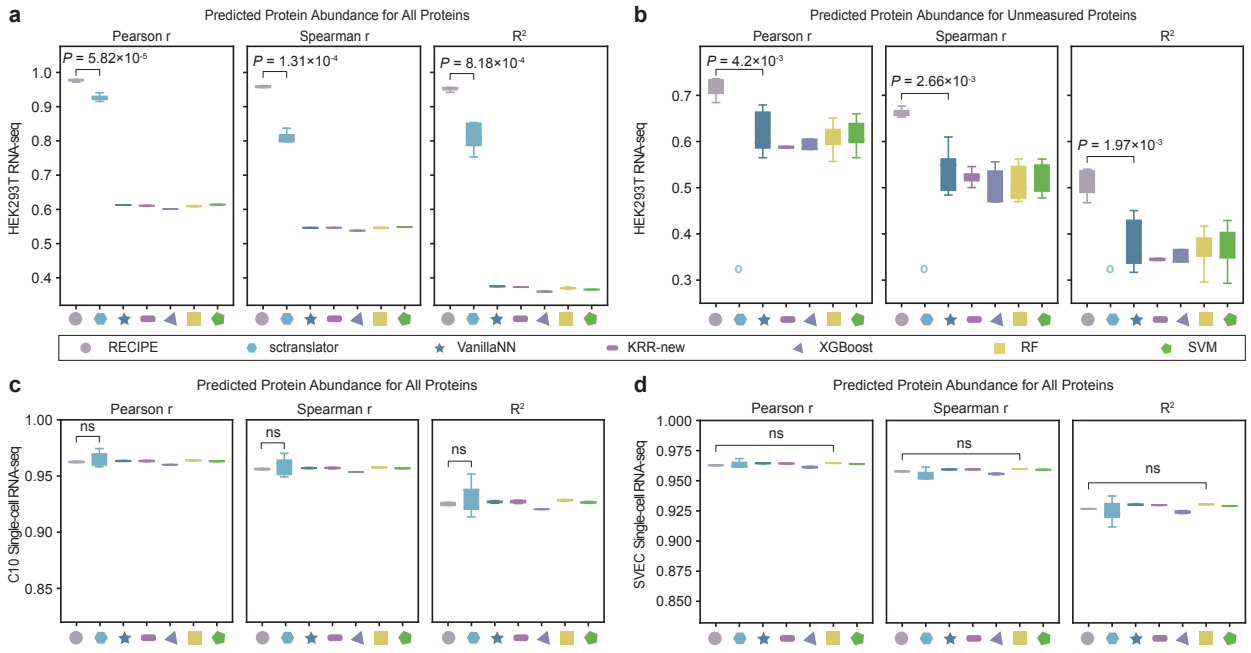

Fig. S7 Comparison of RECIPE with existing methods for prediction on RNA-seq. RECIPE was benchmarked against six alternative methods (XGBoost, RF, SVM, scTranslator, VanillaNN, and KRR-new). Predictive performance was evaluated using Pearson correlation, Spearman correlation, and R^2. To simulate proteins not quantified by proteomics, a subset of proteins was masked during training; these proteins are referred to here as unmeasured proteins. **a–d,** Bulk protein prediction. **a,b,** HEK293T cells. **a,** Models were trained on a subset of genes in DDX20-NC and evaluated on all proteins in DDX20-KD. **b,** Same setting as in **a**, with evaluation restricted to unmeasured proteins. **c-d,** Single-cell protein prediction. **c,** Prediction of all proteins using C10-trained models evaluated on SVEC. **d,** Prediction of all proteins using SVEC-trained models evaluated on C10. Statistical significance was assessed using a one-sided Welch's t-test. Blue circle denotes values < 0. ns, not significant ($P \geq 0.05$).

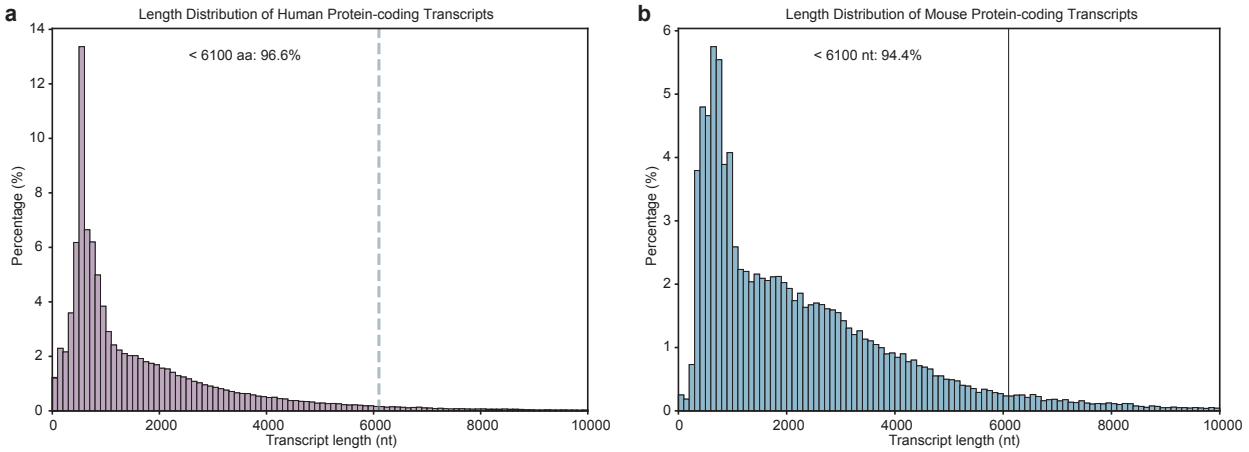

Fig. S8 Length distributions of human protein-coding transcripts and mouse protein-coding transcripts. a, Distribution of lengths for human protein-coding transcripts. The x-axis indicates protein length in nucleotides, and the y-axis indicates the percentage of proteins. The dashed vertical line marks 6,100 nucleotides. **b,** Distribution of lengths for mouse protein-coding transcripts. The x-axis indicates transcript length in nucleotides, and the y-axis indicates the percentage of transcripts. The dashed vertical line marks 6,100 nucleotides.

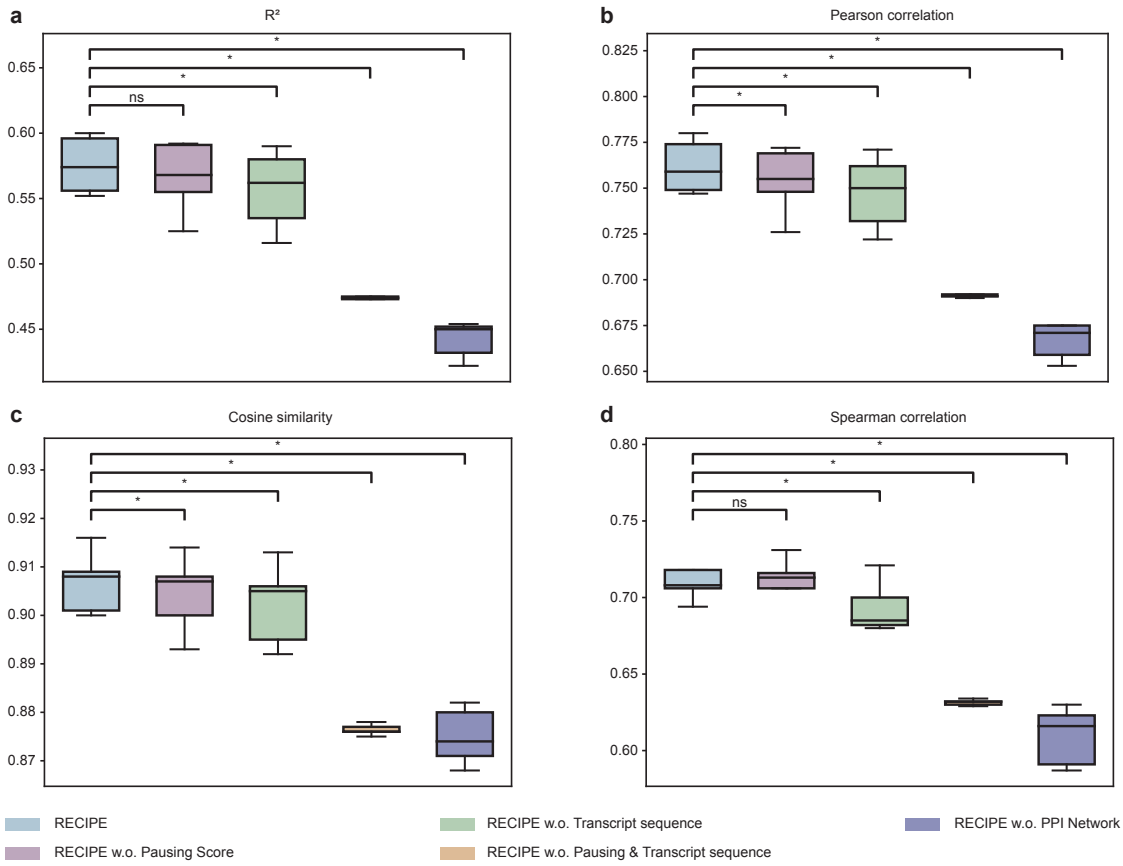

Fig. S9 Ablation analysis of RECIPE. a, Test-set coefficient of determination (R^2), **b,** Pearson correlation, **c,** cosine similarity, and **d,** Spearman correlation for the RECIPE model and its ablated variants, including RECIPE w.o. Pausing Score, RECIPE w.o. Transcript sequence embeddings, RECIPE w.o. Pausing & Transcript sequence embeddings, and RECIPE w.o. PPI Network. Statistical significance was assessed using a one-sided Welch's t-test. P values were annotated as follows: P values were annotated as follows: $P < 0.0001$, ****; $P < 0.001$, ***; $P < 0.01$, **; $P < 0.05$, *; and $P \geq 0.05$, ns (not significant).