

***Supplemental Information:***

**A comprehensive thermodynamic model for RNA binding by the  
*Saccharomyces cerevisiae* Pumilio protein PUF4**

Christoph Sadée, Lauren D. Hagler, Winston R. Becker, Inga Jarmoskaite, Pavanapuresan P. Vaidyanathan, Sarah K. Denny, William J. Greenleaf, and Daniel Herschlag

## Supplemental Figures and Tables:

```

PUF4 TNSNSAEKQRK-IEESSRFADAVLDQYIGSIHSLCKDQHGCRFLQKQLDILGSKAADAIFEETKDYTVEL 69
PUF3 HRSPLLEQLRNSSSDKNSNSNMSLKDIFGHSLEFCCKDQHGSRFIQRELATSPASEKEVIFNEIRDDAIEL 70
PUM2 GRSRLLED FRN-----NRFPNLQLRDLIGHIVEFSQDQHGSRFIQKLERATPAERQMFVNEILQAAYQL 65

PUF4 MTDSFGNYLIQKLLLEEVTTEQRIVLTKISSPHFVEISLNPHGTRALQKLIIECIKTDEEA--QIVVDSL RP 137
PUF3 SNDVFGNYVIQKFFFEFGSKIQKNTLVDQFKGNMQLSLQMYACRVIQKALEYIDSNQRI--ELVLE-LSD 137
PUM2 MTDVFGNYVIQKFFFEFGSLDQKLALATRIRGHVLPALQMYGCRVIQKALESISSDQQVISEMVKE-LDG 134

PUF4 YTVQLSKDLNGNHVIQKCLQRLKPENFQFIFDAISDSCIDIATHRHGCCVLQRCLDHGTTEQCDNLCDKL 207
PUF3 SVLQMIKDQNGNHVIQKAIETIPIEKLPFILSSLTGHIYHLSTHSYGC RVIQRLLLEFGSSEDQESILNEL 207
PUM2 HVLKCVKDQNGNHVVQKCIIECVQPQSLQFIIDAFKGVFVLSSTHPYGC RVIQRILEHCTAEQTLPILEEL 204

PUF4 LALVDKLTLDPFNGNYVVQYIITKEAEKN--KYDYTHKIVHLLKPRAIELSIHKFGSNVIEKILKTAIVSE 275
PUF3 KDFIPYLIQDQYGNVVIQYVVLQQDQFTNKEMVDIKQEIIETVANNVVEYSKHKFASNVVEKSILYGSKNQ 277
PUM2 HQHTEQLVQDQYGNVVIQHVLEHGRPED-----KSKIVSEIRGKVLALSQHKFASNVVEKCVTHASRAE 268

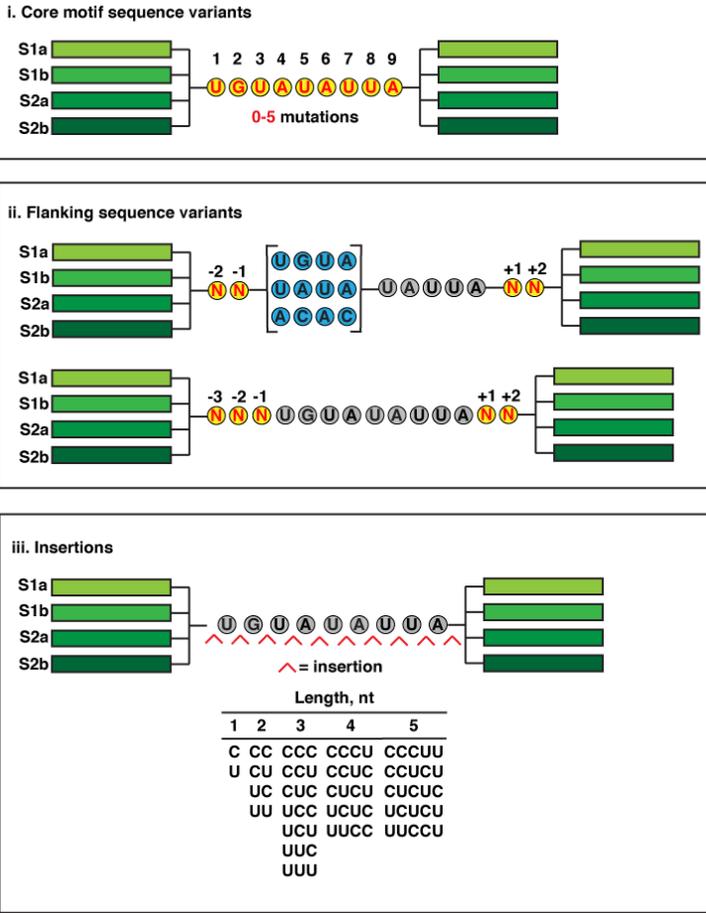
PUF4 PMIL--EIL--NNG-----GETGIQSLNDSYGNVVLQTALDISHKQNDYLYKRLSEIVAPLLVGP I R N 335
PUF3 KDLIISKILPRDKNHALNLEDDSPMILMIKDQFANYVIQKLVNVSEGE GK---KLIVIAIRAYLDKLNKS 344
PUM2 RALLIDEVCCQNDG-----PHSALYTMKDKQYANYVVQK MIDMAEPAQR---KIIMHKIRPHITTL-RK 328

PUF4 TPHGKRIIGMLHLD S 350
PUF3 NSLGNRHLASVEKLA 359
PUM2 YTYGKHILAKLEKYY 343

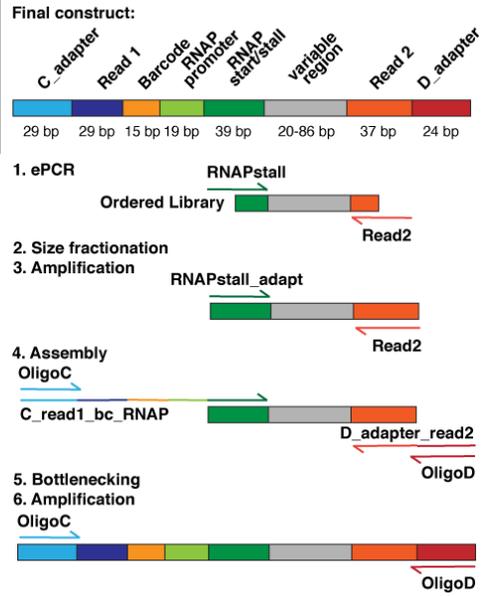
```

**Supplementary Fig. 1 BLAST sequence alignment of *S. cerevisiae* PUF4, *S. cerevisiae* PUF3, and human PUM2.** Full sequence alignment of the Pumilio homology domain of *S. cerevisiae* PUF4 (top row), *S. cerevisiae* PUF3 (middle row), and human PUM2 (bottom row) using blastp (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>). The amino acid reference positions are shown on the right. Grey boxes represent the amino acids that contact the nucleic acids in the eight Pumilio repeats.

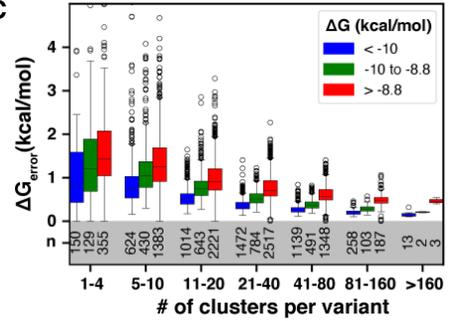
**A**



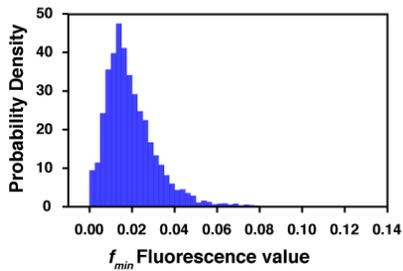
**B**



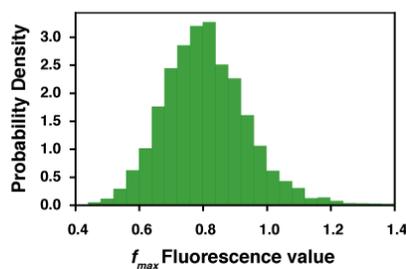
**C**



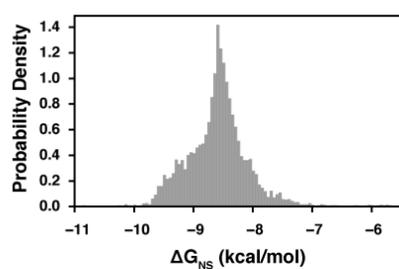
**D**



**E**

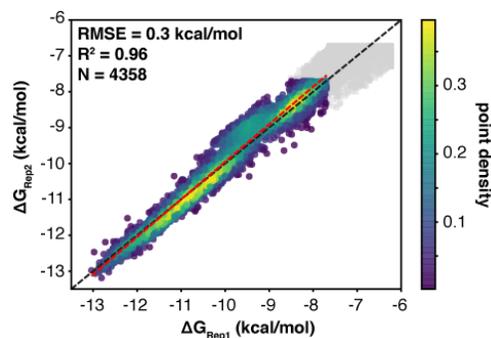


**F**

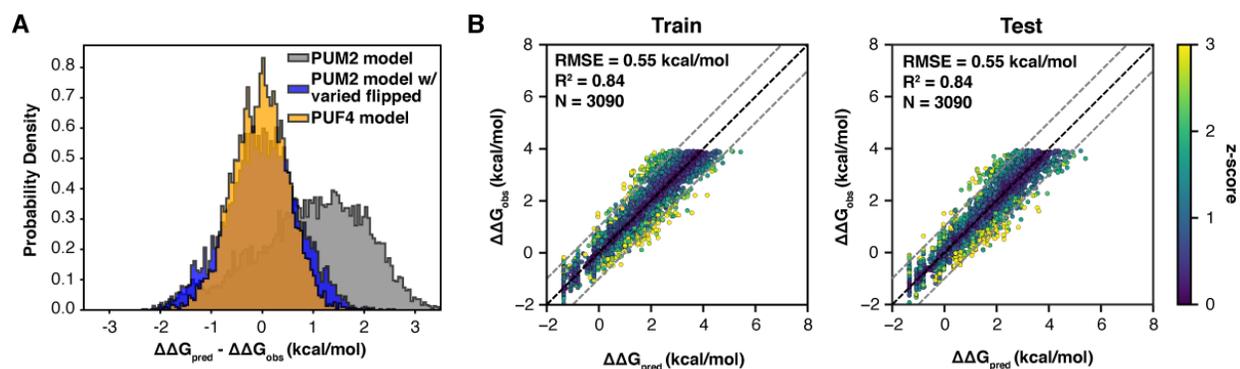


**Supplementary Fig. 2 Designed RNA library preparation and experiments (Related to Fig. 2a).**

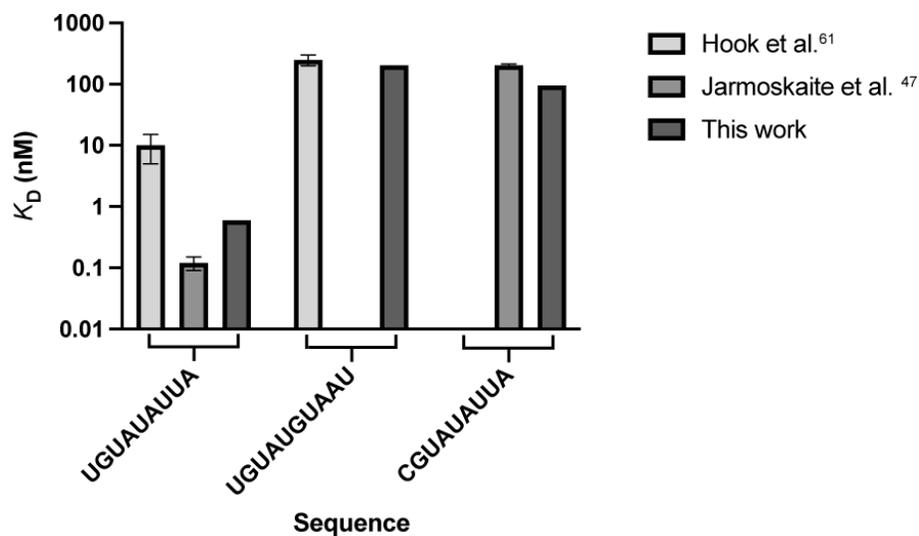
**a** Designed RNA libraries used in this study. The RNA library was designed based on three Pumilio consensus sequences (PUF4, PUM2, and PUF5). For simplicity, only the variants relative to the PUF4 consensus motif are shown. A full list of all sequences and mutants in all RNA libraries studied herein can be found in Supplementary Table 1 and in <sup>20</sup>. Each variant in the designed library was inserted in 2-4 scaffolds (S1a-S2b, Fig. 2a). Variations relative to the consensus motif can be grouped into three categories. i. Motif mutations, where 1-5 bases were changed. A maximum of two bases were mutated in the conserved 5' core UGUA to avoid a predominance of unmeasurable low affinity binders. ii. Flanking sequence variants, where bases upstream and downstream of the consensus are varied to assess their binding effects. iii. Insertions, where 0-5 C or U bases were inserted between the residues of the consensus motif. **b** The DNA array library was prepared as described by <sup>20</sup>. See "Methods" for detailed sequence information and description of each step. **c** The number of clusters per variant from the chip used in low-salt measurements and affinity correlate with measurement certainty. Box plots are grouped into high, medium, and low affinity binders, with the lowest affinity binder group representing RNAs with <15% bound RNA at highest protein concentration. The orange line represents the median, the box the interquartile range, whiskers are max and min values and circles show outliers. **d-f** Distribution of initial fit parameters for variant-independent parameters. Histogram of high affinity variants with >85% RNA bound at highest concentration based on initial per variant  $K_D$  and other criteria as described in Methods. (D)  $f_{min}$  (E)  $f_{max}$  (F)  $\Delta G_{NS}$ .



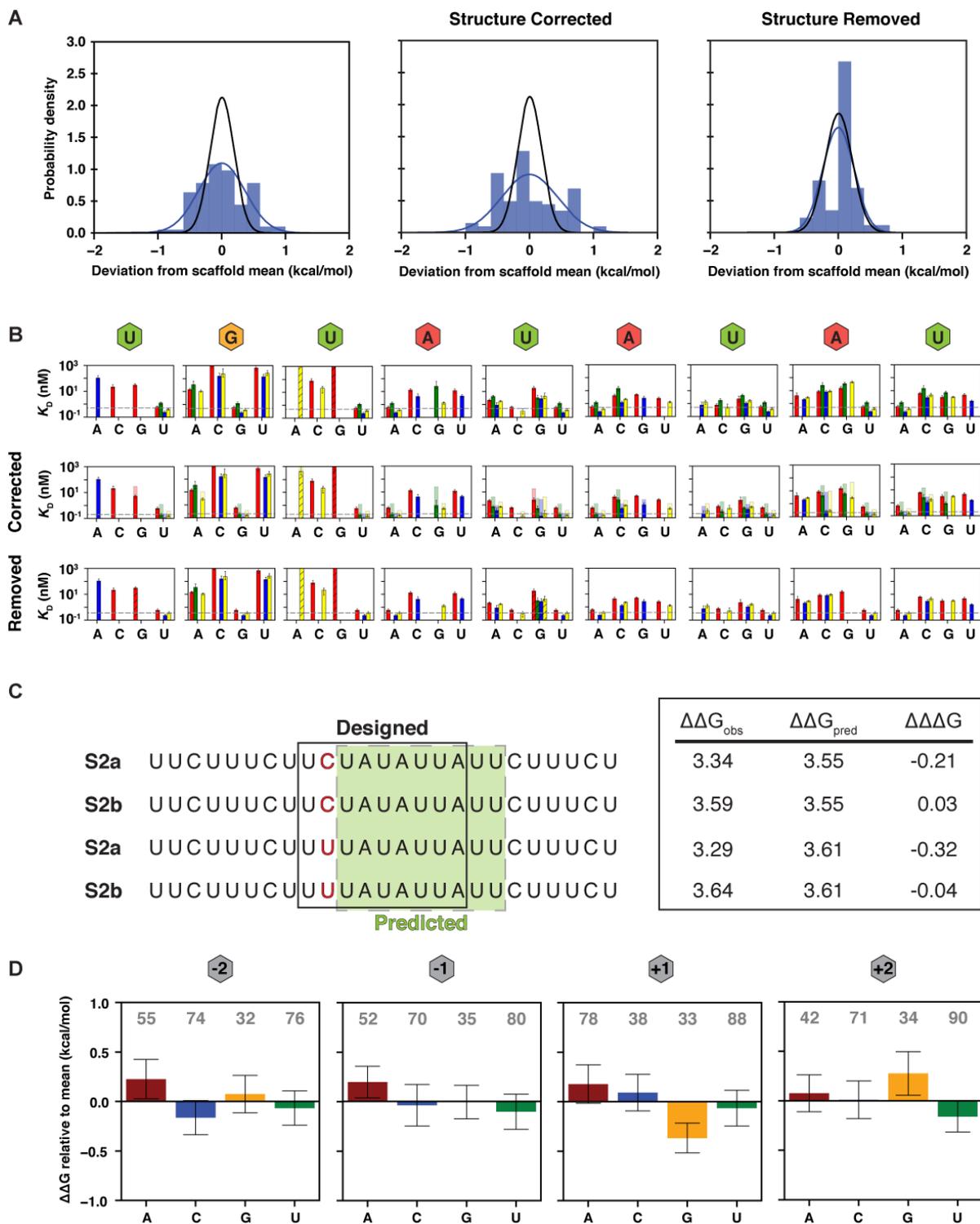
**Supplementary Fig. 3 Comparison of replicate data.** Two technical replicates of Puf4 binding measurements were fit as described in the Methods (binding conditions: 2 mM  $\text{MgCl}_2$ , 100 mM KOAc ). Note these are higher salt conditions than used in the main text; lower salt was used to enhance binding and thus the number of RNAs that gave measurable  $K_D$  values]. The transparent region (above -7.7 kcal/mol) corresponds to  $\Delta G$  values that cannot be distinguished from background with high confidence. N is the number of values in the high confidence range. The black dashed line represents a correlation line ( $x = y$ ). The red dashed line is the best fit line with a slope of 1.05 and intercept of 0.51 kcal/mol.



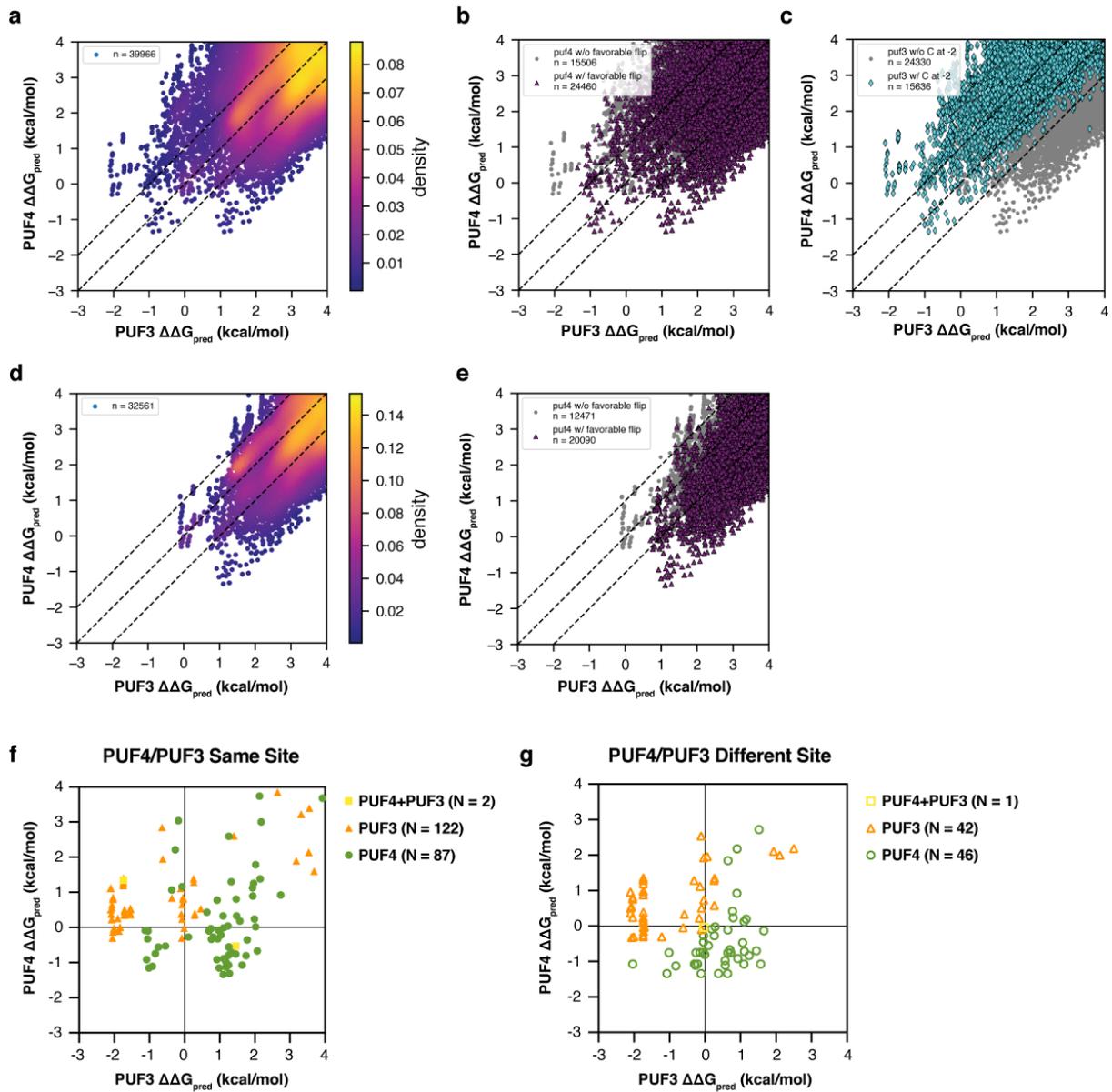
**Supplementary Fig. 4 Comparison of predictive models for PUM2/PUF4 specificity (Related to Fig. 3).** **a** Distribution of differences between model predictions and observed affinities. The grey plot corresponds to the data fit to the PUM2 model as described in ref <sup>20</sup> (Fig. 3b). The blue plot corresponds to the data fit to the PUM2 model with all flipping parameters optimized as described in the Methods (Fig. 3C). The orange plot corresponds to the PUF4 model derived herein (Fig. 3d). **b** To test the model for overfitting, data were initially randomly and split into equal-sized training and test datasets. The model was optimized on the training dataset and the resulting parameters were used to predict the test data. RMSE and  $R^2$  were used to assess goodness of fit, and these parameters were equal for the two datasets.



**Supplementary Fig. 5. Comparison to literature values.** Published literature affinities for PUF4 binding to RNAs for measurements at 25 °C and cases where equilibration was >30 min<sup>47,61</sup> compared to the binding of similar RNA variants from this work. Other literature values for PUF4<sup>37</sup> are not directly comparable because they were measured at different temperatures or their equilibration times were not long enough to ensure equilibration (see criteria outlined in reference<sup>47</sup>).



**Supplementary Fig. 6. Analysis of single-mutant variants binding to PUF4.** **a** The effects of scaffold differences were assessed by determining the deviation of the individual scaffold affinities from scaffold mean for that variant. The deviation was calculated for each binding site variant (single mutant and the PUF4 consensus UGUAUAUUA). The distribution of these deviations (blue line) was compared to the distribution expected from experimental error (black line). Distributions were compared without structure correction (left). Inter-scaffold variation is larger than that predicted by experimental error (blue vs. black lines). Correcting for structure effects using ViennaRNA fold did not resolve the deviation (center). Removing structured variants with cutoff  $\Delta G_{\text{fold}} < -0.5$  (as described in “Methods”) lowered inter-scaffold variation within experimental error (right). **b** PUF4 affinities for UGUAUAUUA single mutants before accounting for structure (top), with structure correction (middle), and after structured variants were removed (bottom). Scaffolds are depicted with colored bars (red = S1a, green = S1b, blue = S2a, yellow = S2b). Hashed lines indicate variants not satisfying all of the quality control cutoffs, but were included because they were the only single mutant variant at that position. **c** Assessment of alternative binding registers. Alternative registers were predicted based on the register with the highest binding affinity. PUF4 predominantly bound in designed registers with minor exception for single mutants at position two, deviating no more than 0.25 kcal/mol and within experimental error. **d** Flanking sequence effects on PUF4 binding. Four bases, two downstream and two upstream of the PUF4 consensus were randomized (see also Supplementary Fig. 2) and embedded into different scaffolds ( $n = 256$ ). Effects were assessed by taking the average affinity of all variants with a given base at each position. The number of variants for each is shown above the bar. The values were subtracted from the average of all combined variants to assess deviation from mean. Deviations were within 0.5 kcal/mol with the most significant increase at position +1 for guanine. Hence, an additional base was added to the PUF4 consensus to account for the additional energy term in global fitting, yielding this 10-mer UGUAUAUUAU. U at position 10 was chosen over G due to its large representation in the library. Structured variants with  $\Delta G_{\text{fold}} < -0.5$  were removed for this analysis.



**Supplementary Fig. 7. Transcriptome-wide analysis of PUF4 and PUF3 binding sites in *S. cerevisiae* 3'UTRs (Related to Fig. 6).** **a-c** PUF4/PUF3 affinity–specificity landscape for 13 nt windows of *S. cerevisiae* 3'UTRs in the Saccharomyces genome database. See also Methods.  $\Delta\Delta G_{\text{pred}}$  is the ensemble binding energy for each window relative to the consensus sequence for each RBP (PUF4 = UGUAUAUUA, PUF3 = CNUGUAUAUA), obtained using the appropriate mathematical model (see Table 1 and Supplementary Table 3); all sites with  $\Delta\Delta G_{\text{ensemble}} < 4$  kcal/mol are included. **a** is colored by point density; **b** compares binding sites where the PUF4 model utilizes the 6/7 flipping term (purple triangle); **c** compares binding sites where the PUF3 model utilizes the C at position -2 term (cyan diamond). **d,e** PUF4/PUF3 affinity–specificity landscape for 11 nt windows of *S. cerevisiae* 3'UTRs in the Saccharomyces genome database as in **a,b**, with the consensus sequence for PUF3 being UGUAUAUA. See also Methods. **f,g** PUF4/PUF3 affinity–specificity landscapes for the highest affinity binding site for previously identified PUF3 and PUF4 3'UTR targets<sup>9</sup>. In **f**, the highest affinity site is the same for PUF3 and PUF4, and in **g**, the highest affinity sites are different for PUF3 and PUF4. PUF3 targets are depicted by an orange triangle; targets that bind PUF3 and PUF4 with a yellow square, and PUF4 targets with a green circle. N is the number of known targets for each.

**Supplementary Table 1:** All sequences and associated affinities

**Supplementary Table 2:** “Flipped” Parameters for PUM2 model with optimized flips (Related to Fig. 3c)

Term II	$\Delta\Delta G_f^Y$ (kcal/mol)				
	Y =				
Flipped Residue Position f =	A	C	U	G	NN <sup>a</sup>
<b>3/4</b>	2.00	1.12	2.24	1.02	2.50
<b>4/5</b>	2.00	0.30	1.23	1.71	2.50
<b>5/6</b>	2.00	0.47	1.32	1.57	2.38
<b>6/7</b>	0.03	-0.82	-0.82	-0.07	1.88
<sup>a</sup> 2- nt flip of any sequence					

**Supplementary Table 3.** Parameters used for PUF3 Model (Related to Fig. 6e,f and Supplementary Fig. 6)

<b>PUF3 Thermodynamic values</b>					
Term I	$\Delta\Delta G_b^X$ (kcal/mol)				
	X =				
Bound Residue Position b =	A	C	U	G	
<b>-2</b>	0.00	-2.00	0.00	0.00	
<b>-1</b>	0.0	0.00	0.00	0.00	
<b>1</b>	3.08	2.91	0.00	3.04	
<b>2</b>	1.93	3.14	3.14	0.00	
<b>3</b>	2.39	2.49	0.00	2.92	
<b>4</b>	0.00	1.92	1.46	1.71	
<b>5</b>	-0.03	0.17	0.00	0.79	
<b>6</b>	0.00	1.83	1.49	1.82	
<b>7</b>	1.55	1.78	0.00	1.59	
<b>8</b>	0.00	1.57	1.01	1.52	
<b>9</b>	0.30	0.29	0.00	-0.07	
Term II	$\Delta\Delta G_f^Y$ (kcal/mol)				
	Y =				
Flipped Residue Position f =	A	C	U	G	NN <sup>a</sup>
<b>3/4</b>	>2.00	1.79	1.41	>1.5	>2.50 <sup>b</sup>
<b>4/5</b>	>2.00	>3.00	>2.50	>2.50	>2.50
<b>5/6</b>	1.22	1.05	0.81	1.57	2.18
<b>6/7</b>	>2.00	1.77	>2.02	>2.0	2.04
<sup>a</sup> 2- nt flip of any sequence					
<sup>b</sup> “>” indicates a lower limit					