Supporting Information

Sequence-dependent RNA helix conformational preferences predictably impact tertiary structure formation

Joseph D. Yesselman
Sarah K. Denny
Namita Bisaria
Daniel Herschlag
William J. Greenleaf
Rhiju Das*

*Correspondence should be addressed to R.D. (rhiju@stanford.edu).

Table of Contents

www.pnas.org/cgi/doi/10.1073/pnas.1901530116

SI Methods

All software and source code used in this work are freely available for non-commercial use. RNAMake software and documentation are at https://github.com/jyesselm/RNAMake.

Flow piece labeling.

Three distinct flow pieces were used to probe the chip piece library, with helices of length 9, 10 (wildtype), and 11 base pairs (see SI Appendix, Table S3). Flow pieces were ordered as RNA oligos from Integrated DNA Technologies (Coralville, Iowa) with a 5′-Amino Modifier C6 modification, with HPLC purification. Each flow piece was ethanol precipitated at –20 °C overnight, followed by resuspension to a final concentration of 2 mM with 2 mM of NHS-conjugated Cy3b dye in 50 mM phosphate buffer (pH 8.7). This reaction was incubated at 37 °C for 1 hour, followed by PAGE purification (8% PAGE, 8 M Urea, 1x TBE: 89 mM Tris-HCl, 89 mM Boric Acid, pH 7.4, 2 mM sodium EDTA). RNA was eluted from the gel in water using three freeze-thaw cycles. To reduce aggregation on the chip surface, flow piece solutions were spun in a 50K Amicon filter two times and collected on a 3K Amicon filter. Flow pieces were quantified after purification using Qubit RNA high sensitivity kit (Thermofisher).

Chip piece library design, amplification, and sequencing.

The tectoRNA library was designed by replacing the chip piece helix with a set of defined WC base pair sequences. This library of chip piece variants (~2000 sequences) was ordered together with other tectoRNA variants not discussed here, to form a final library of ~45,000 variants. The library was ordered with common priming sequences across chip piece variants from CustomArray (Bothell, WA). This pool of DNA oligonucleotides was PCR amplified with primers oligopool_left and oligopool_right (see SI Appendix, Table S4 and Figure S1A), with 1:400 dilution of the synthesized oligo pool, 200 nM of each primer, 200 µM dNTPs, 3% DMSO, 1x Phusion HF buffer, 0.01U/µl of HS Phusion (NEB). Primers were purchased from Integrated DNA Technologies (Coralville, Iowa). The reaction proceeded for 9 cycles of 98 °C for 10 seconds, 62 °C for 30 seconds, and 72 °C for 30 seconds, followed by cleanup of the reaction mixture using Qiagen PCR Cleanup Kit (elution into 20 µl). To append sequencing adapters to this PCR product as well as include unique molecular identifier (UMI, in the form of a 16 nt random N-mer), a five-piece PCR assembly was performed, with 1 µl of the previous reaction, 137 nM of primers (short_C and short_D; SI Appendix, Table S4), 3.84 nM of the adapter sequences (C1_R1_BC_RNAP and D_Read2; SI Appendix, Table S4), 200 µM dNTPs, 3% DMSO, 1x Phusion HF buffer, and 0.02U/µl of Phusion Hot Start Flex enzyme (NEB). The reaction proceeded for 14 cycles of 98 °C for 10 seconds, 63 °C for 30 seconds, and 72 °C for 30 seconds, followed by cleanup with Qiagen PCR Cleanup Kit, as above.

After amplification and assembly, the library was bottlenecked to reduce the representation of UMIs to ~700K distinct 16 nt N-mers. First, the library was diluted 1:5000 in 0.1% Tween20, and this dilution was quantified against a standard library of PhiX (Illumina, Hayward, CA), which was diluted two-fold seven times to form a dilution series from 25 pM to 0.2 pM. The standard series and the library dilution were amplified in a qPCR assay to determine their relative cycle threshold (CT) values; these values were used to determine the concentration of the diluted library by linear regression analysis of the CT values against the known concentrations of the standards. The volume associated with 700K molecules was PCR amplified, with 1.25 µM of primers (short_C and short_D), and 1x NEBNext Master Mix (NEB, M0541S), for 21 cycles of 98 °C for 10 seconds, 63 °C for 30 seconds, and 72 °C for 60 seconds, followed by cleanup with Qiagen PCR Cleanup Kit. The final library was sequenced on an Illumina Miseq instrument at 10-30% of the total sequencing chip, with the rest of the chip consisting of high-complexity genomic sequences. Sequencing cycles were performed as follows: 75 bases in read 1, 75 bases in read 2, and an 8 bp i7 index read, resulting in demultiplexed, paired-end sequences.

The output of the Illumina sequencing included the read1 and read2 sequence associated with each cluster ID. This information was processed to extract the UMI sequence from read1 for each cluster (by extracting the sequence preceding the RNAP initiation site; see SI Appendix, Figure S1A). Clusters with common UMI sequences were processed to obtain a consensus read2 sequence, by taking the most common base at each position (i.e. per-base voting consensus). UMIs of poor quality, with poor representation or poor agreement across sequences, were removed, by assessing the number of clusters with read2 sequences matching the consensus sequence. Poor quality was defined if the number of matches (or successes) could be explained by a null model with p value > 0.01, where the null model was a binomial distribution with probability of success of 0.25. This filter removed UMIs associated with diverse unrelated sequences, or with relatively few reads per UMI.

Finally, the consensus sequence of each UMI was associated with each designed library variant by searching for an exact match of the reverse complement of the designed sequence within the read2 consensus (starting at the first base).

Experimental platform for parallel measurements on a sequencing chip.

The sequencing chip used for Illumina Miseq sequencing was directly used on a custom-built imaging station, made from a combination of parts from an Illumina Genome Analyzer IIx and parts that were custom-designed, as described originally in (1), and modified as in (2). The flow cell surface was imaged with a total internal reflection fluorescence (TIRF) setup, allowing measurement of the bound fluorescence on the chip surface with minimal background from fluorescent molecules in solution. Custom scripts were used to control the laser power, stage, temperature, fluidics, and camera. Images could be taken in one of two channels, the “red” channel (660 nm laser, with 664 nm long pass filter from Semrock), and the “green” channel (530 nm laser and a 590 (104) nm band pass filter from Semrock). To image the flow cell surface, 16 images were taken to overlap tiles 1 through 16 taken of the Miseq sequencing output. Each image was taken for 400 ms exposure time with 200 mW input laser power.

RNA was generated in situ on the surface of the Illumina Miseq chip by a series of enzymatic reactions carried out through fluidic application and temperature control, as described in (1–3). In brief, covalently attached ssDNA was converted to dsDNA through extension of a biotinylated primer, followed by incubation with streptavidin to create a streptavidin roadblock (see SI Appendix, Figure S1B). *E. coli* RNA Polymerase (NEB M0551S) was applied to the flow cell with limiting concentrations of NTPs (2.5 μM each of ATP, GTP, and UTP), allowing only very limited extension and preventing initiation by more than one polymerase per molecule. Excess polymerase was washed out of the flow cell, followed by incubation with the full suite of NTPs at high concentration (1 mM each NTP) to allow extension. Encountering the streptavidin roadblock causes polymerases to stall, resulting in stable display of the nascent transcript (SI Appendix, Figure S1B). Detailed descriptions of each of these steps may be found in (3).

After RNA extension, blocking oligos were annealed to common regions on the nascent transcript (see SI Appendix, Figure S1A) to limit the formation of alternate secondary structure, as well as to fluorescently label clusters of transcribed RNA (fluorescent_stall and dark_read2; SI Appendix, Table S4). Oligos were purchased from Integrated DNA Technologies (Coralville, Iowa) with RNase-Free HPLC Purification.

On chip experiments to determine tectoRNA affinity.

For each experiment, a fluorescently-labeled tectoRNA flow piece was serially diluted three-fold to form a concentration series from 2000 nM to 0.91 nM in binding buffer (89 mM Tris-Borate, pH 8.0, 30 mM $MgCl_2$, 0.01 mg/ml yeast tRNAs (ThermoFisher Scientific AM7119), 0.01% Tween20. To fold the flow piece, it was initially diluted to 10 uM in water, and denatured by incubating for 1 minute at 95 °C, followed by refolding for 2 minutes on ice (preceding the dilution to 2 uM and serial dilution). Each flow piece solution was applied to the flow cell, and after waiting for sufficient time for equilibration, the flow cell was imaged in the red and green channels, with the red channel capturing the annealed oligo corresponding to any transcribed RNA, and the green channel capturing the bound flow piece. Experiments were carried out at at 22 °C. Equilibration times were as follows: 3 hours, 2 hours, 1 hour, 45 min, 30 min, 20 min, 20 min, and 20 min, for 0.91 nM, 2.7 nM, 8.2 nM, 25 nM, 74 nM, 222 nM, 667 nM, and 2000 nM, respectively. These times were calculated to allow equilibration for the most stable variants (i.e. ΔG of –12 kcal/mol or K_d of 1 nM), assuming a common association rate constant (k_{on}) of ~$6x10^4$ $M^{-1}s^{-1}$ (3).

Quantification of ΔG from image series.

Each image taken during the course of an experiment was processed to extract the fluorescence values of the Illumina Miseq clusters. First, the Miseq tile and x-y-positions of each sequenced cluster was determined (from the Miseq output). Because of differences in the optics of the Miseq and the imaging station, these coordinates did not correspond 1:1 to the pixel values of our images. To account for this, sequence data coordinates were scaled by an overall scale factor (of 10.96 imaging-station pixels to Miseq x-y position units). A global registration offset was determined by cross-correlation of the images and subsequent fitting of the cross-correlation matrix to a 2D Gaussian to obtain the x-y- position that maximized the cross correlation coefficient. Finally, to correct for nonlinear aberrations, this cross-correlation procedure was repeated for 256 subdivisions of the overall image to obtain corrections on the global x-y- position as a function of the location within the image. These corrections were fit to 2D surfaces for the x- and y- corrections, as a function of x- and y- position.

In each of the 256 subtiles, all clusters within the subtile were fit to a sum of 2D Gaussians, with x-y- positions given by the sequencing data coordinates, nonlinearly corrected as described above, as in (1). The integrated fluorescence associated with each cluster is then: $2\pi A\sigma^2$, where A is the amplitude and σ the standard deviation of the 2D Gaussian. The fluorescence associated with the bound flow piece was normalized by dividing by the fluorescence in the red channel, to account for variability of cluster size.

The series of concentration values for each cluster were fit to a binding isotherm, according to the equation:$f(x) = f_{min} + f_{max}\left(\frac{x}{x+exp(\Delta G/RT)}\right)$, where f is the normalized fluorescence, f_{min}, f_{max}, and ΔG are free parameters, x is the concentration, R is the gas constant, and T is the temperature in Kelvin. Following single cluster fits, the values for f_{min}, f_{max}, and ΔG per variant were obtained by finding the median of these values across single clusters associated with each variant. An additional fitting step refined these values by applying a distribution for f_{max} for those variants that did not achieve saturation, based on the values for f_{max} of variants that did, as described in (3), ultimately allowing consistent attribution of the change in fluorescence values to changes in ΔG rather than f_{max}. In brief, this fit refinement took the median fluorescence values across a set of clusters (resampled from all clusters associated with the variant). This set of median fluorescence values was fit to the binding isotherm equation, with f_{min} set to the median fluorescence value across clusters that did not achieve saturation, and f_{max} either allowed to float or set to a random value generated from the distribution of f_{max}, depending on if the maximum fluorescence in the binding series did or did not exceed the lower bound of the 95% confidence interval of the f_{max} distribution, respectively. This resampling and refitting was repeated 100 times for each variant, allowing determination of confidence intervals on the fit values of ΔG per variant.

Combining experimental replicates.

Data for the wildtype, 10-bp flow piece comes from two replicate experiments (shown in SI Appendix, Figure S2A). Values reported for this flow piece represent the average of the two replicate values, weighted by the inverse of the variance on each measurement. If the 95% confidence interval on ΔG is $\delta\Delta G$, then the variance on the measurement is: $\sigma^2 =(\delta\Delta G/1.96)^2$. Thus, the weighted average on ΔG is then: $\Delta G_{avg} = \left(\frac{\Delta G_1}{\sigma_1^2} + \frac{\Delta G_2}{\sigma_2^2}\right)\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1}$. The combined error is then: $\sigma = \left(\sqrt{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}\right)^{-1}$.

Building base pair step ensembles.

To build a curated library of base-pair step components, we obtained the set of non-redundant RNA crystal structures managed by the Leontis and Zirbel groups (4) (version 1.45: http://rna.bgsu.edu/rna3dhub/nrlist/release/1.45). This set specifically removes redundant RNA structures that are identical to previously solved structures, such as ribosomes crystallized with different antibiotics. We processed each RNA structure to extract every motif using Dissecting the Spatial Structure of RNA (DSSR) (5) with the following command:

```
x3dna-dssr –i file.pdb –o file_dssr.out
```

We manually checked each extracted motif to confirm that it was the correct type, as DSSR sometimes classifies tertiary contacts as higher order junctions and vice versa. For each motif collected from DSSR, we ran the X3DNA find_pair and analyze programs to determine the reference frame for the first and last base pair of each motif to allow for alignment between motifs:

```
find_pair file.pdb 2> /dev/null stdout | analyze stdin >& /dev/null
```

We defined a base pair step as two consecutive residues on one chain base-paired to two consecutive residues on another chain, where both base pairs are in Watson-Crick orientation. Each instance of this pairing was collected from every structure. See SI Appendix, Table S1 for a summary of all total instances of each base-pair step.

Clustering procedure for base pair step ensembles.

To cluster the base-pair steps, all structures were first translated and rotated so that the first base pair was situated with its origin at (0,0,0) and its axes aligned with x, y and z orientation of the identity matrix, definition of base pair center and coordinate systems are as in (6). Fixed radius clustering was performed using a radius of a $distance\_score$ of 1.50, which was ideal according to optimization, although other radii did not greatly affect the final results. The $distance\_score$ between a cluster center and a new base-pair step is calculated below, where d_1 and R_1 are the translation and orientation of the cluster center's second base pair, respectively. d_2 and R_2 are the translation and orientation of the second base pair in the base-pair step to be clustered.

$$distance\_score = |\overrightarrow{d_1} - \overrightarrow{d_2}| + 2\sum_i^3 \sum_j^3 abs\,(R_{1ij} - R_{2ij}) \quad (1)$$

The number of clusters generated for each base-pair step sequence is shown in SI Appendix, Table S1. Each cluster was assigned a relative energy (Eq. 2) based on its population. $N_{members}$ is the number of base-pair steps in a given cluster, and N_{total} is the number of base-pair steps of the current identity, i.e. AU/AU. This energy is used during our Monte Carlo simulations to allow swapping based on population.

$$E = -RTln\,(\frac{N_{members}}{N_{total}}) \quad (2)$$

TectoRNA simulation protocol.

The simulation is set up by supplying a sequence and secondary structure for both tecto heterodimers. With this information, a 3D system is built up by representing each base-pair step with a corresponding structural ensemble and representing both tertiary contacts as single structures. The structure of the GAAA tetraloop/tetraloop-receptor (TTR) was isolated from the P4-P6 domain of the *Tetrahymena* ribozyme (PDB: 1GID). There is no known solved structure of the GGAA TTR; therefore, a structure was generated by modeling using stepwise Monte Carlo (7). The simulation proceeds by attempting to swap a randomly selected base-pair step from one conformation to another. If the new conformation has a lower energy, it is accepted; if not, it is selected by the Metropolis criterion. All motifs are connected to each other by shared base pairs, so if a base-pair step is swapped from one conformation to another, the orientation change will propagate throughout the structure accordingly. In total, one million swaps are attempted during our standard simulation. To determine whether a conformation is bound, we calculate the distance_score (Eq. 1) between the final base pair of the chip helix and its original position (Figure 2B). If this score is lower than 5, we consider the conformation to be bound.

Calculating the relative binding free energy of the tecto system.

`rnamake_ddg_tecto` is part of a larger toolkit known as RNAMake. For instructions on installing RNAMake as well as extensive documentation available at http://jyesselm.github.io/RNAMake/. An example of running `rnamake_ddg_tecto` is shown below.

```
rnamake_ddg_tecto \
     -fseq “CTAGGAATCTGGAAGTACCGAGGAAACTCGGTACTTCCTGTGTCCTAG” \
     -fss  “((((((....((((((((((((....))))))))))))....))))))” \
     -cseq “CTAGGATATGGAAGATCCTCGGGAACGAGGATCTTCCTAAGTCCTAG” \
     -css  “(((((((..((((((((((((....))))))))))))...)))))))” \
     -s 1000000
```

The tecto system is composed of two distinct RNA molecules that dimerize. First is the “chip” piece, which is transcribed from the DNA on a MiSeq sequencing chip. There are up to one hundred thousand distinct sequences on each chip in a given experiment. The second sequence is the “flow” piece, which is titrated in during the experiment and can bind to all chip sequences. We maintain this nomenclature while running `rnamake_ddg_tecto`. “`-fseq`” specifies the sequence of the flow RNA, and “`-fss`” specifies the corresponding secondary structure in dot-bracket notation. If a new sequence has the default secondary structure, “`-fss`” does not need to be used again. The flow sequence must include the GGAA tetraloop-receptor sequence and secondary structure or it will return an error. “`-cseq`” and “`-css`” are analogues to “`-fseq`” and “`-fss`”, but for the chip RNA. This RNA must include the GAAA tetraloop-receptor sequence or the output secondary structure will return an error. “`-s`” specifies the number of Monte Carlo steps to perform. The default is one million. The output of the program is the number of times that the Monte Carlo simulation sampled a “bound” conformation.

Using the output of the `rnamake_ddg_tecto` program, it is possible to calculate the relative binding free energy of each sequence compared to the wild-type (WT) sequence where N_bound

values are evaluated as the number of simulated conformations given distance score (eq. 1) compared to the target conformation of 5. Alternative forms of the distance score in (1), including more standard rotationally invariant metrics to define rotation matrix differences (8) or base-pair-to-base-pair RMSDs based on quaternions (9), but these were not tested in the current study.

Generation of 2000 helix sequences for blind predictions.

To computationally assess the effect of the primary sequences of helices on relative binding, we generated all possible Watson-Crick helices. We put an A-U, U-A, G-C or C-G base pair at 9 positions in the chip sequence for a total of 4^9 (262,144) sequences. For each generated sequence, we utilized RNAFold from ViennaFold (10) to confirm that the sequence folds into the target secondary structure. Then, we ran `rnamake_ddg_tecto` on each new sequence with the following command.

```
rnamake_ddg_tecto -cseq new_sequence
```

Estimating free energy of secondary structure formation.

Secondary structure of each tecto RNA sequence was calculated using RNAfold (v. 2.1.8) from ViennaFold (10) to obtain the free energy of the ensemble at 20 °C, using the command:

```
RNAfold --noPS -p0 -T20
```

Computing ΔΔGs with mismatch base pairs.

We utilized a set of 305 unique chip sequences with a single mismatched base pair (see SI Appendix, Table S2; ref: (3)) with measured binding affinities to bound to the 9 bp, 10 bp or 11 bp flow piece leading to 628 unique measurements (SI Appendix, Dataset S2). For each chip peice / flow piece combination we ran `rnamake_ddg_tecto` with the following arguments shown below.

```
rnamake_ddg_tecto \
     –fseq “CTAGGAATCTGGAAGTACCGAGGAAACTCGGTACTTCCTGTGTCCTAG” \
     -fss  “((((((....((((((((((((....))))))))))))....))))))” \
     -cseq “CTAGGATATGGAAGATCCTCGGGAACGAGGATCTTCCTAAGTCCTAG” \
     -css  “(((((((..((((((((((((....))))))))))))...)))))))” \
     -s 1000000
```

These are the same ones described in Method Section: Calculating the relative binding free energy of the tecto system. `rnamake_ddg_tecto` automated identifies if there is a non-canonical motif and uses an ensemble representation with existing examples found from the PDB. We directly compared each mismatch-containing sequence to a corresponding chip sequence with the same base pairs except for a Watson-Crick base pair instead of the mismatch. This comparison allows us to compute a ΔΔG of introducing a mismatch base pair, canceling out all other effects (SI Appendix, Table S2).

Computing ΔΔGs for RNA acceptor helix while bound to aspartyl-tRNA synthetase.

To compute the sequence dependence of tRNA-AspRS binding free energy on acceptor stem sequence we used RNAMake's `rnamake_ddg_helix_sampler` which can compute the likelihood of a helix sequence adopting a supplied conformation. For each PDB we extracted the acceptor stem of the tRNA (PDB 1IL2: C901-C907, C966-C972. PDB 1C0A: B601-B607, B666-B672. PDB 1ASZ: S601-S607, S666-S672) (11–13). Using this extracted helix we supplied to `rnamake_ddg_helix_sampler` with the following commands.

For PDB 1IL2:

```
rnamake_ddg_helix_sampler \
     -seq "AAAAAAA&UUUUUUU" \
     -start_bp "C901-C972"  \
     -end_bp "C907-C966   \
     -pdb "1il2_aceptor_helix.pdb" \
     -all
```

For PDB 1C0A:

```
rnamake_ddg_helix_sampler \
     -seq "AAAAAAA&UUUUUUU" \
     -start_bp "B601-B672"  \
     -end_bp "B607-B666   \
     -pdb "1c0a_aceptor_helix.pdb" \
     -all
```

For PDB 1ASZ:

```
rnamake_ddg_helix_sampler \
     -seq "AAAAAAA&UUUUUUU" \
     -start_bp "S601-S672"  \
     -end_bp "S607-S666   \
     -pdb "1asz_aceptor_helix.pdb" \
     -all
```

In each case "`-start_bp`" denotes where the first base pair will be aligned to and correspondingly "`-end_bp`" is the base pair that is that target of the last base pair of the generated helix. In both "`-start_bp`" and "`-end_bp`" accept their base pair by "name" which is the name of the two residues contained in it in the format chain id appended to its residue number. In the case of A141-A162 that declares that there is a base pair between residue 141 on chain A to residue 161 also on chain A. Argument "`-pdb`" supplies the path of the PDB that contains the coordinates at least the start and end base pair. Argument “-seq” supplies the sequence of the helix to build, if “-all” is supplied, all sequences will be checked but “-seq” is still required. rnamake_ddg_helix_sampler outputs the raw number (“count”) of conformations that were within a cutoff of the target base pair specified with “-end_bp”. To calculate a ΔΔG we compared the outputted count to the wild-type sequence of both the yeast (UCCGUGA&UCGCGGA) and *E. coli* sequences (GGAGCGG&CCGUUCC) which were determined to be 54243 and 64219 respectively. All computed ΔΔGs can be found in SI Appendix, Dataset S4.

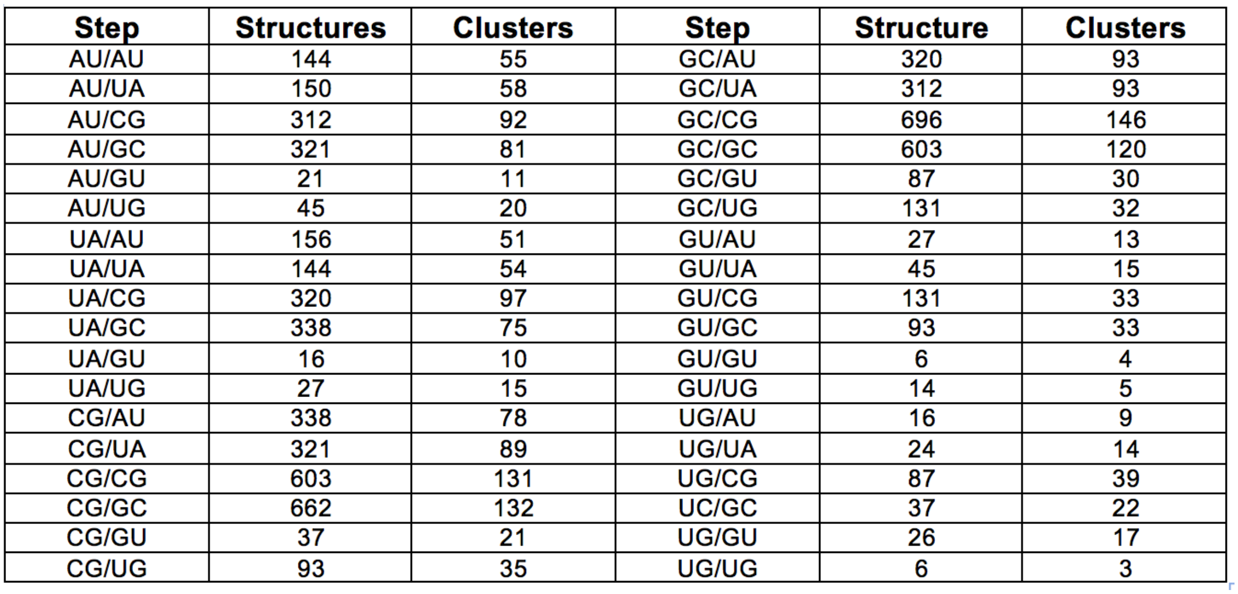
Computing ΔΔGs for anticodon helix for aminoacyl-tRNA•EF-Tu accommodation during ribosome codon recognition.

Similarly to compute the sequence dependence of anticodon helix for aminoacyl-tRNA•EF-Tu accommodation during ribosome codon recognition we also used rnamake_ddg_helix_sampler. The structures came from PDBs 4V5G, 4V5P, 4V5Q, 4V5R and 4V5S (14,15). For each PDB we extracted the acceptor stem of the tRNA (residues AY27-AY31 and AY39-AY43 in all PDBs). Using these extracted helices we supplied to rnamake_ddg_helix_sampler with the following command.

```
rnamake_ddg_helix_sampler \
     -seq "AAAAA&UUUUU" \
     -start_bp "A31-A39"  \
     -end_bp "A27-A43"   \
     -pdb "4v5g_anticodon_helix.pdb" \
     -all
```

To calculate a ΔΔG we compared the outputted count to the wild-type sequence of the $tRNA^{Thr}$ anticodon helix (GGGUG&CACCC) with a count of 10291. See SI Appendix, Dataset S4, for all computed ΔΔGs.

SI Figures and Tables



Step	Structures	Clusters	Step	Structure	Clusters
AU/AU	144	55	GC/AU	320	93
AU/UA	150	58	GC/UA	312	93
AU/CG	312	92	GC/CG	696	146
AU/GC	321	81	GC/GC	603	120
AU/GU	21	11	GC/GU	87	30
AU/UG	45	20	GC/UG	131	32
UA/AU	156	51	GU/AU	27	13
UA/UA	144	54	GU/UA	45	15
UA/CG	320	97	GU/CG	131	33
UA/GC	338	75	GU/GC	93	33
UA/GU	16	10	GU/GU	6	4
UA/UG	27	15	GU/UG	14	5
CG/AU	338	78	UG/AU	16	9
CG/UA	321	89	UG/UA	24	14
CG/CG	603	131	UG/CG	87	39
CG/GC	662	132	UC/GC	37	22
CG/GU	37	21	UG/GU	26	17
CG/UG	93	35	UG/UG	6	3

Table S1. Base pair steps collected from RNA crystallographic structures.

Summary of the total number of structures found and number of structural clusters determined for each base-pair step.

Mismatch	Strand 1	Strand 2	Conformations	Number of measurements	RMSE, kcal/mol
A-A	CAG	CAG	3	31	0.85
A-G	CAG	CGG	1	25	1.65
C-C	CCG	CCG	1	37	0.29
C-U	CCG	CUG	2	36	0.99
G-A	CGG	CAG	1	27	1.68
G-G	CGG	CGG	5	33	0.36
G-U	CGG	CUG	1	35	0.40
U-C	CUG	CCG	2	35	1.10
U-G	CUG	CGG	3	37	0.32
U-U	CUG	CUG	14	37	1.28
A-G	GAC	GGC	1	24	0.84
C-C	GCC	GCC	2	35	0.68
C-U	GCC	GUC	1	31	0.67
G-A	GGC	GAC	1	34	0.75
G-G	GGC	GGC	3	34	0.60
G-U	GGC	GUC	5	35	0.38
U-C	GUC	GCC	1	30	1.02
U-G	GUC	GGC	3	35	0.43
U-U	GUC	GUC	21	36	0.49

Table S2. Preliminary predictions of tectoRNA binding affinities with mismatched base pairs.

Three consecutive base pairs were replaced with the sequences described, which harbor a non-Watson-Crick mismatch between two flanking G-C pairs. The substitution was made at all positions of the tectoRNA chip piece and observed ΔΔG's were compared to RNAMake predictions assuming an ensemble for the three-base-pair segment derived from observations of the sequence in the crystallographic database. Note excellent RMSE accuracies for constructs with G-U pairs; worse predictions for other mismatches may be due to poor representation of the segments in the crystallographic database (as few as 1 observation). See also analysis in ref. (3).

Name	Sequence
9-bp	CUAGGAAUCUGGAAGACCGAGGAAACUCGGUCUUCCUGUGUCCUAG
10-bp	CUAGGAAUCUGGAAGUACCGAGGAAACUCGGUACUUCCUGUGUCCUAG
11-bp	CUAGGAAUCUGGAAGUACACGAGGAAACUCGUGUACUUCCUGUGUCCUAG

Table S3. Flow piece sequences

The name and sequence of each of the flow pieces used in this study.

Name	Sequence
oligopool_left	TTGTATGGAAGACGTTCCTGGAT
oligopool_right	GCTGAACCGCTCTTCCGATCT
short_C	AATGATACGGCGACCACCGA
short_D	CAAGCAGAAGACGGCATACGA
C_R1_BC_RNAP	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNNNNNNNTTTATGCTATAATTATTTCATGTAGTAAGGAGGTTGTATGGAAGACGTTCCTGGAT
D_Read2	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
Fluorescent_stall	GGATCCAGGAACGTCTTCCATACAACCTCCTTACTACAT-3’Alexa647 (NHS ester)
Dark_read2	CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

Table S4. Primers used to amplify library for sequencing

The name and sequence of the primers used to amplify the library for sequencing.

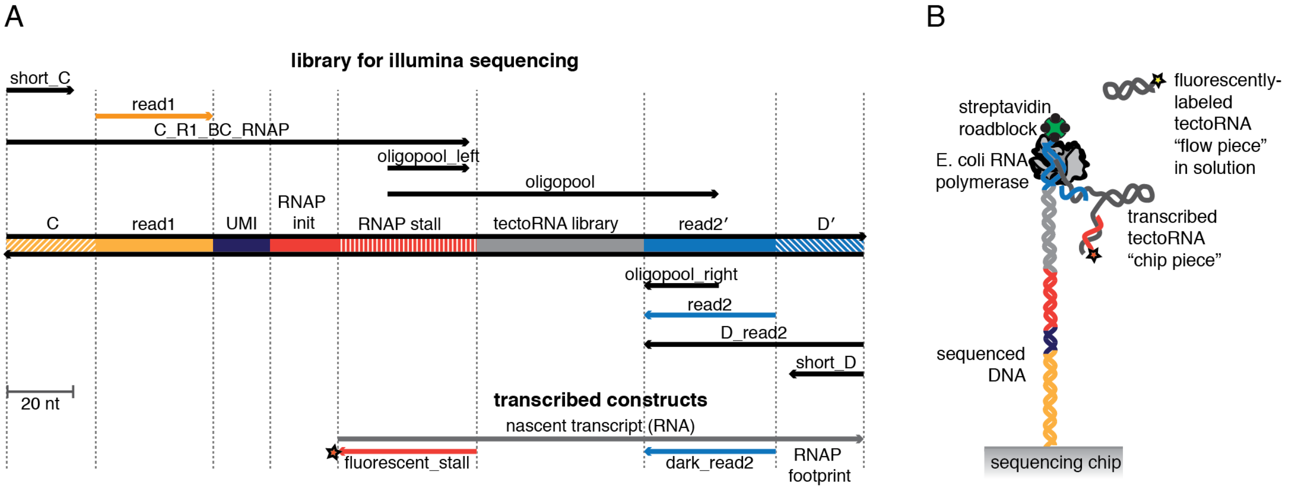


Figure S1. Library construction and experimental setup.

A) Schematic of the sequencing library containing the tectoRNA chip piece variants. Regions encoding an RNAP initiation site and stall sequence are included, as well as sequencing adapters, and a unique molecular identifier (UMI). B) The configuration of the *in situ* transcribed tectoRNA on the surface of the sequencing chip. After initiation at the RNAP initiation site, the *E. coli* RNAP transcribes the tectoRNA chip piece variant, eventually stalling due to a streptavidin-biotin linkage at the 3’ end of the DNA. A fluorescently-labeled DNA oligo annealed to the 5’ end of the transcript labels transcribed RNA (Alexa-647). Fluorescently-labeled tectoRNA “flow” piece introduced to the sequencing chip flow cell binds to the tectoRNA “chip” piece.

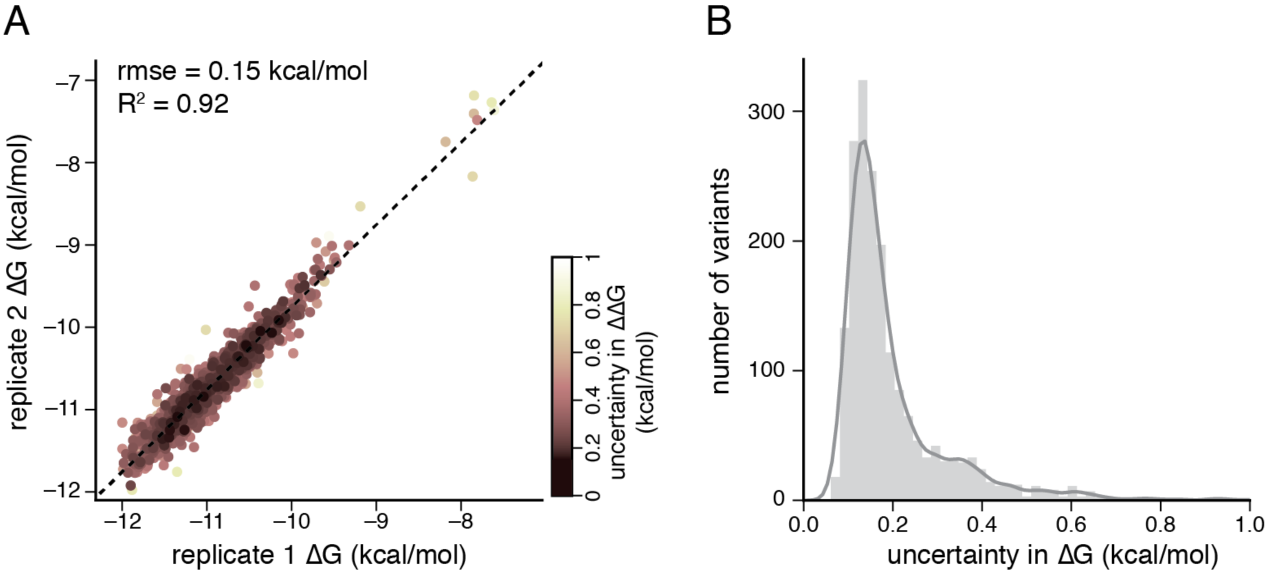


Figure S2. Measured ΔG values are reproducible and precise.

A) Experiments measuring the free energy of binding between the tectoRNA flow piece and 1,455 chip piece variants that were measured in at least 5 clusters in both experiments. The chip piece variants had a different composition of WC base pairs and different lengths. Each measurement is colored by the combined uncertainty in the ΔΔG (i.e. $\sqrt{\delta\Delta G_1{}^2 + \delta\Delta G_2{}^2}$ where $\delta\Delta G$ is the uncertainty in ΔG (95% confidence interval; CI) in each experiment. B) Distribution of the uncertainty of the measured ΔG (95% CI) per variant, after combining the replicate experiments (Methods).

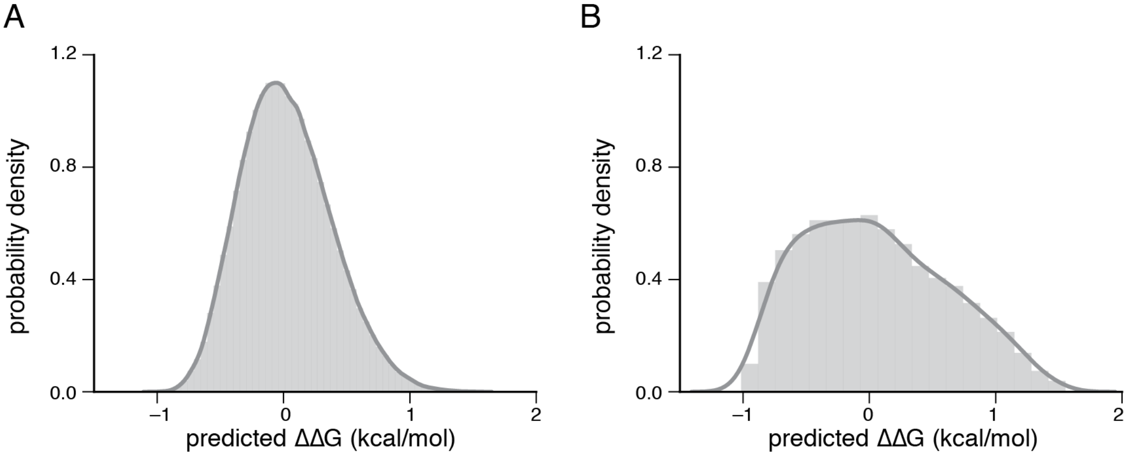


Figure S3. Predicted effect of helix sequence on tectoRNA binding free energy.

Distribution of the predicted ΔΔG across all possible chip tectoRNA sequences of length 10 bp (A) or the subset of ~2000 tectoRNA sequences of length 10 bp tested in the tectoRNA library. The effect is relative to the median.

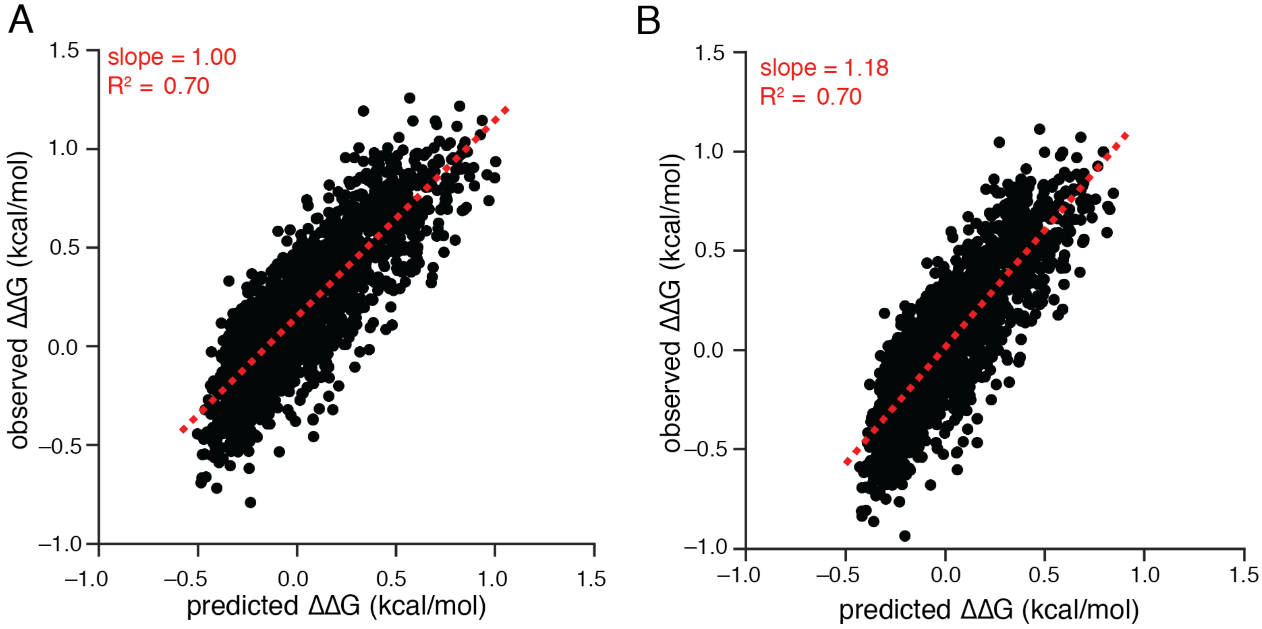


Figure S4. Simulation parameter sweeps

(A-B) Examples of predicted ΔΔGs with different slopes as a function of changing the proximity cutoff (A) 8.75 Å (B) 10.0 Å.

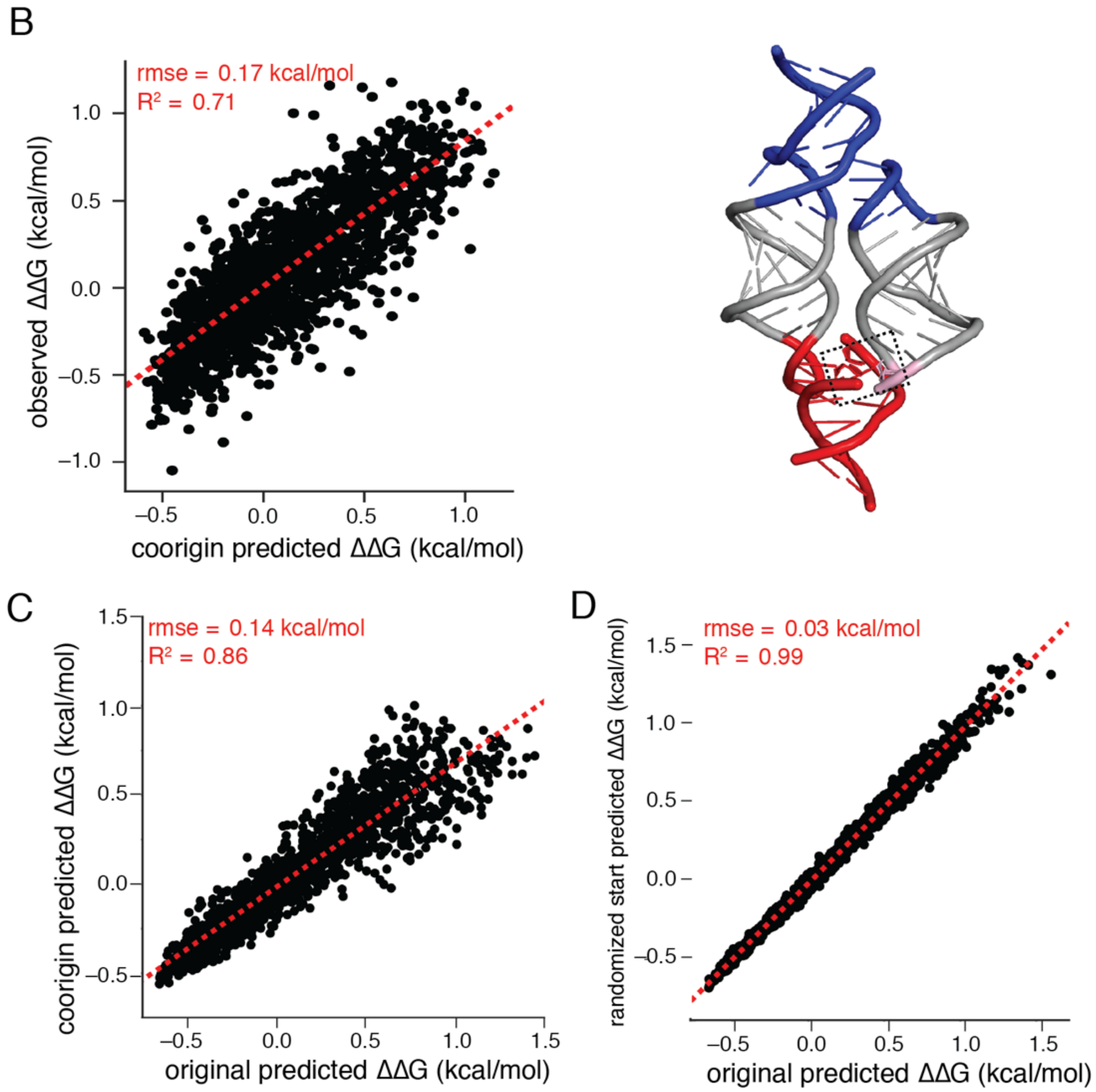


Figure S5. Comparison of simulation topology and starting conformation

(A) Co-origin model for predicted ΔΔG and unconstrainted tectoRNA, the proximity threshold is now measured in the receptor (seen in B) instead of the tetraloop see in the main text. (C) Correlation between the original and co-origin predicted ΔΔGs. (D) Instead of starting the simulation by using the lowest energy conformation for each base pair step (as done in all simulations reported throughput the study), randomly select one.

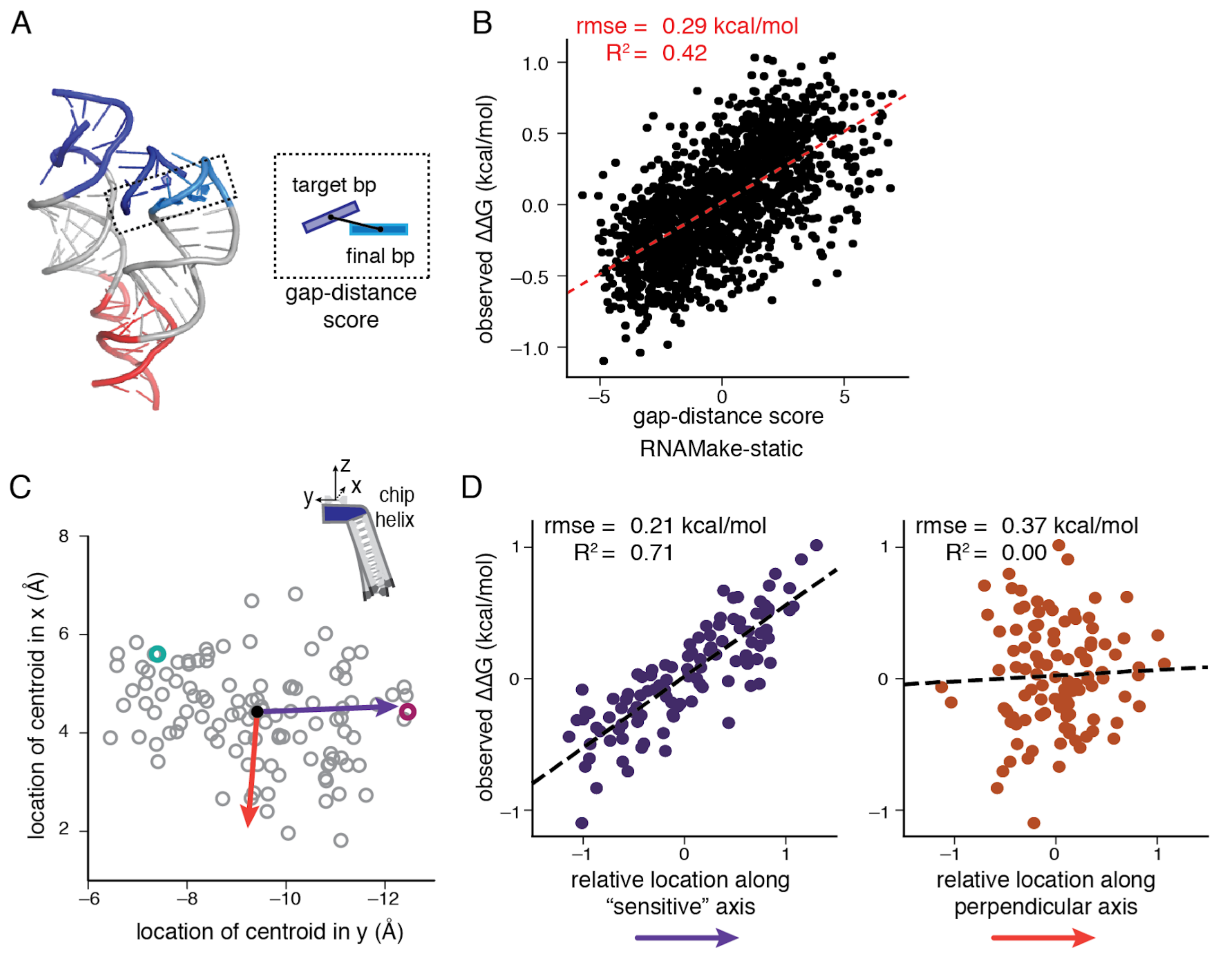


Figure S6. Non-ensemble models for tectoRNA affinity do not consistently predict observed effects.

A) Schematic, as in Figure 2B, of the unconstrained tectoRNA, that shows the final bp of the chip piece helix (turquoise) compared to where it should be to allow binding of the GGAA tetraloop to the R1 receptor (blue). The distance between the final base pair position and the target base pair position is quantified as the 'gap-distance' score, as in Eq. 1. B) Scatterplot comparing the observed free energy of binding to the gap-distance score of a single structure of the unconstrained state, i.e. using only the single lowest energy structure for each base pair step. Both the observed ΔG and the gap distance score are shown as relative to their respective medians. C) Positions of the centroid of the final bp of the chip piece helix in the partially bound tectoRNA of 100 different chip pieces that vary in affinity. Arrows indicate two axes that differentiate the centroids. The purple axis is defined by finding the average difference between the unconstrained and bound structure centroids. The orange is a perpendicular vector. Each vector is defined in all six translational or rotational coordinates, but only the projection into the x-y plane is shown. D) Calculation of relative affinity depends on the location of each chip piece partially bound centroid when projected into one axis (left, "sensitive" axis), but not a perpendicular axis (right). These results illustrate one reason that the ensemble was required for accurate energetic prediction: the binding landscape is anisotropic and without simulating the ensemble of the global assembly, we could not have demarcated these sensitive and insensitive axes of positional variation.

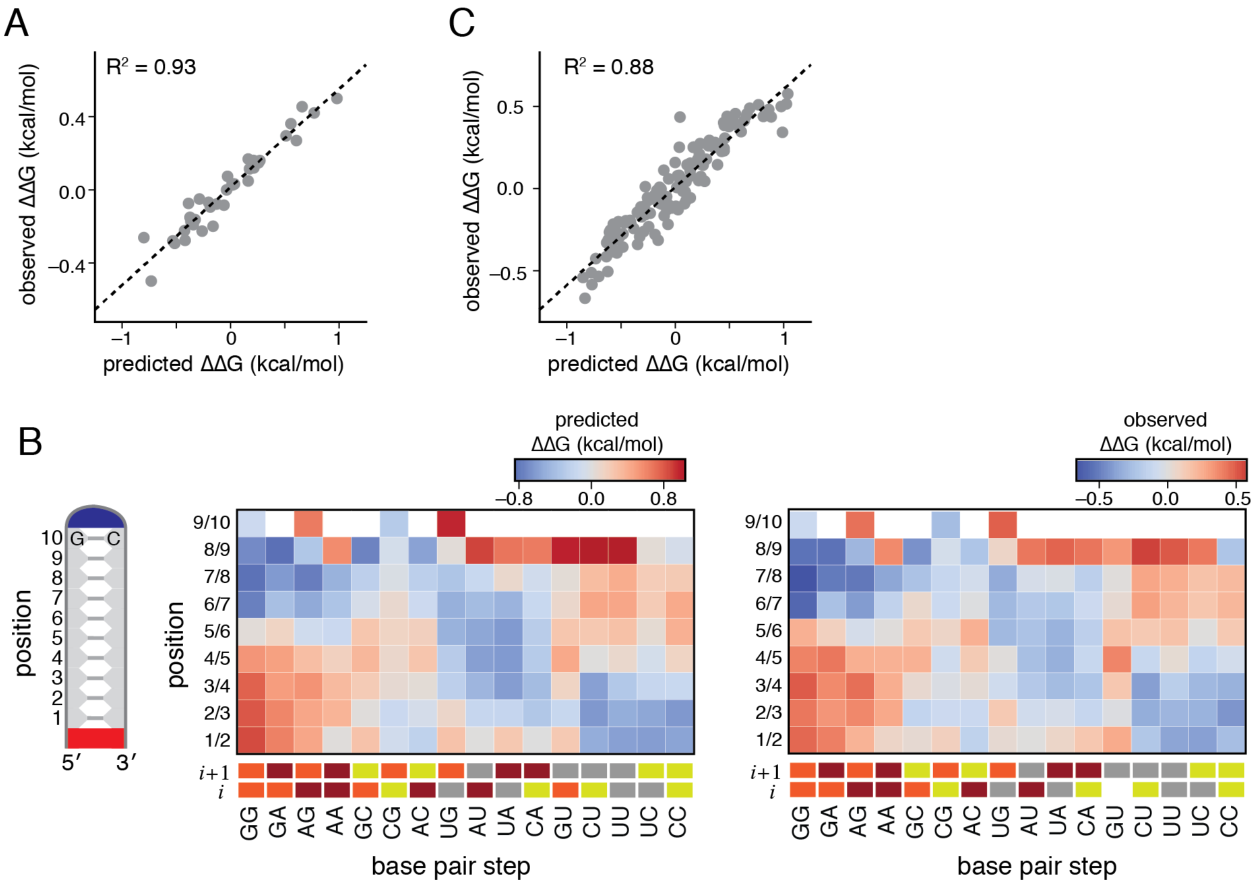


Figure S7. Predicted effect of each base pair step at each position within the tectoRNA chip piece helix.

A) Scatterplot compares the observed to the predicted effect of having each base at each position (effects are shown in Figure 3D). B) (Left) Schematic shows the position of each base pair within the chip piece helix. (Heatmaps) Either the predicted (left) or observed (right) free energy of binding for chip piece sequences with the indicated base pair step at each pair of positions, for the effect for the subset of sequences tested in the tectoRNA library. ΔG is given as a deviation from the median ΔG of all possible chip piece variants. Color below each heatmap indicates the two bases from 5’ to 3’ of the base pair step on the 5’ side of the tetraloop. White indicates missing values. All chip sequences had a G at position 10, thereby limiting the base pair step types that were evaluated at this position. C) Scatterplot comparing the observed and predicted effect of having each base pair step at each pair of positions (effects are as in (B)).

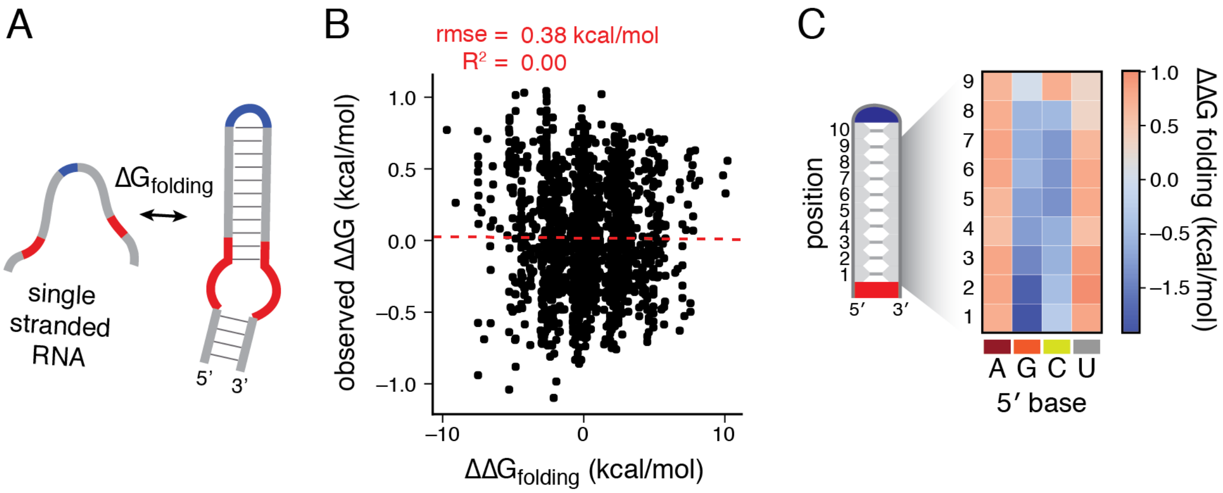


Figure S8. Observed changes in tectoRNA affinity are not dependent on predicted changes in free energy of secondary structure formation.

A) Schematic shows the secondary structure formation of the tectoRNA chip piece. B) Scatterplot comparing the dependence of observed free energy of binding to the tectoRNA flow piece ($\Delta\Delta G_{bind}$) on the predicted free energy of secondary structure folding for each tectoRNA chip piece ($\Delta\Delta G_{fold}$). C) Predicted free energy of secondary structure folding for chip pieces with the indicated base pair at each position in the helix. ΔG_{fold} is given as a deviation from the median ΔG_{fold} of all 1536 chip piece variants. Position is as indicated in Figure 1A and Figure 3D. The secondary structure folding calculations show no correlation with observed tertiary assembly measurements, supporting the assumption that the molecules of the tectoRNA dimer have pre-formed secondary structure.

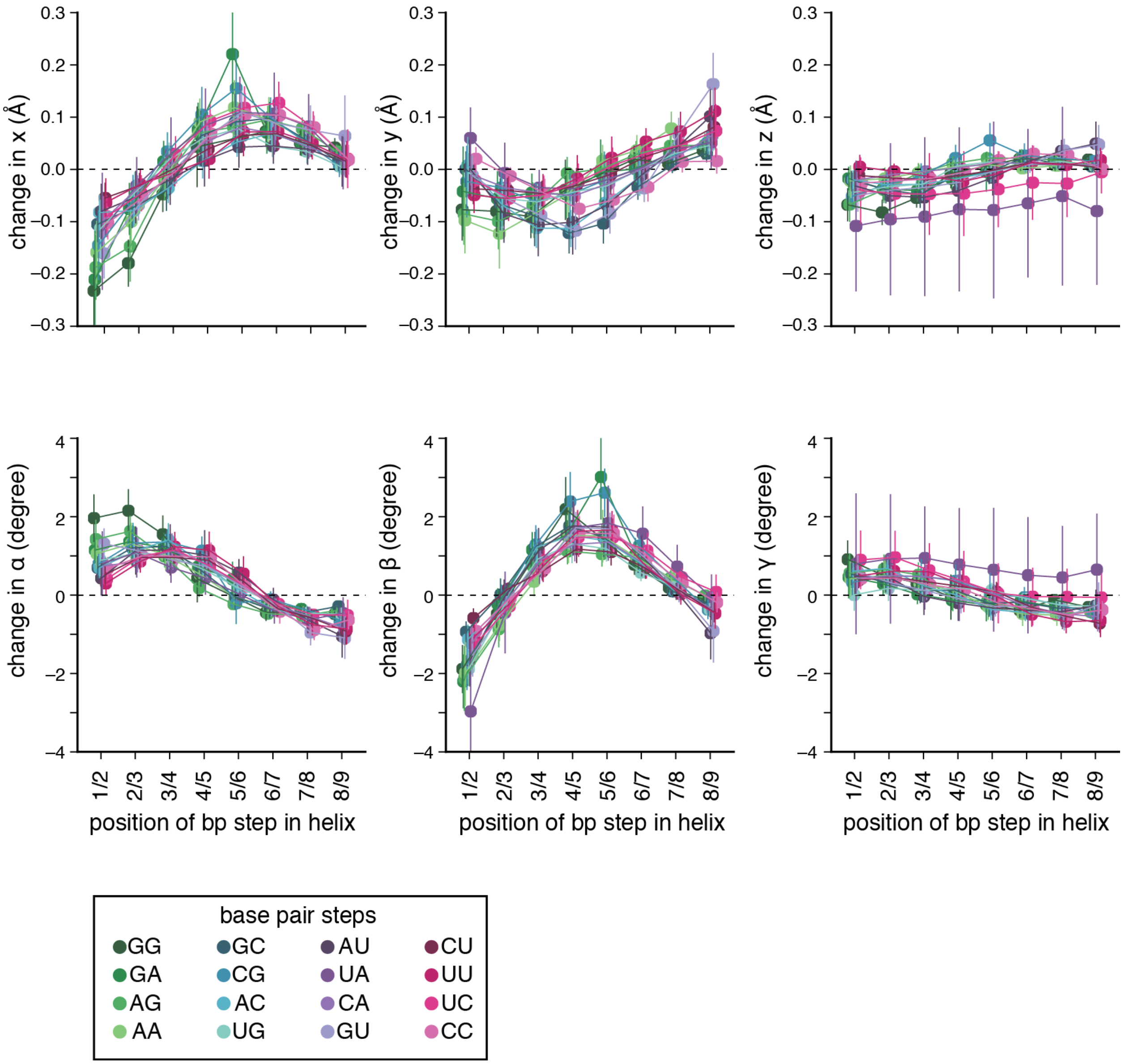


Figure S9. Structural preferences of conformational states across positions.

The difference in structural coordinates of base pair steps in the bound tectoRNA versus the unconstrained tectoRNA (i.e. free helix). The average structure of each base pair step was determined by taking the weighted average of each structural coordinate across the base pair step's conformational states. Weights were the number of times that conformational state was sampled at that position (across 100 different chip piece variants that spanned the range of affinity) in the bound tectoRNA and unconstrained tectoRNA. Error bars are 95% confidence intervals determined through bootstrapping. In legend, base pair step refers to the 5´ strand of the helix sub-sequence, e.g. 'GG' corresponds to 5´-GG-3´/5´-CC-3´.

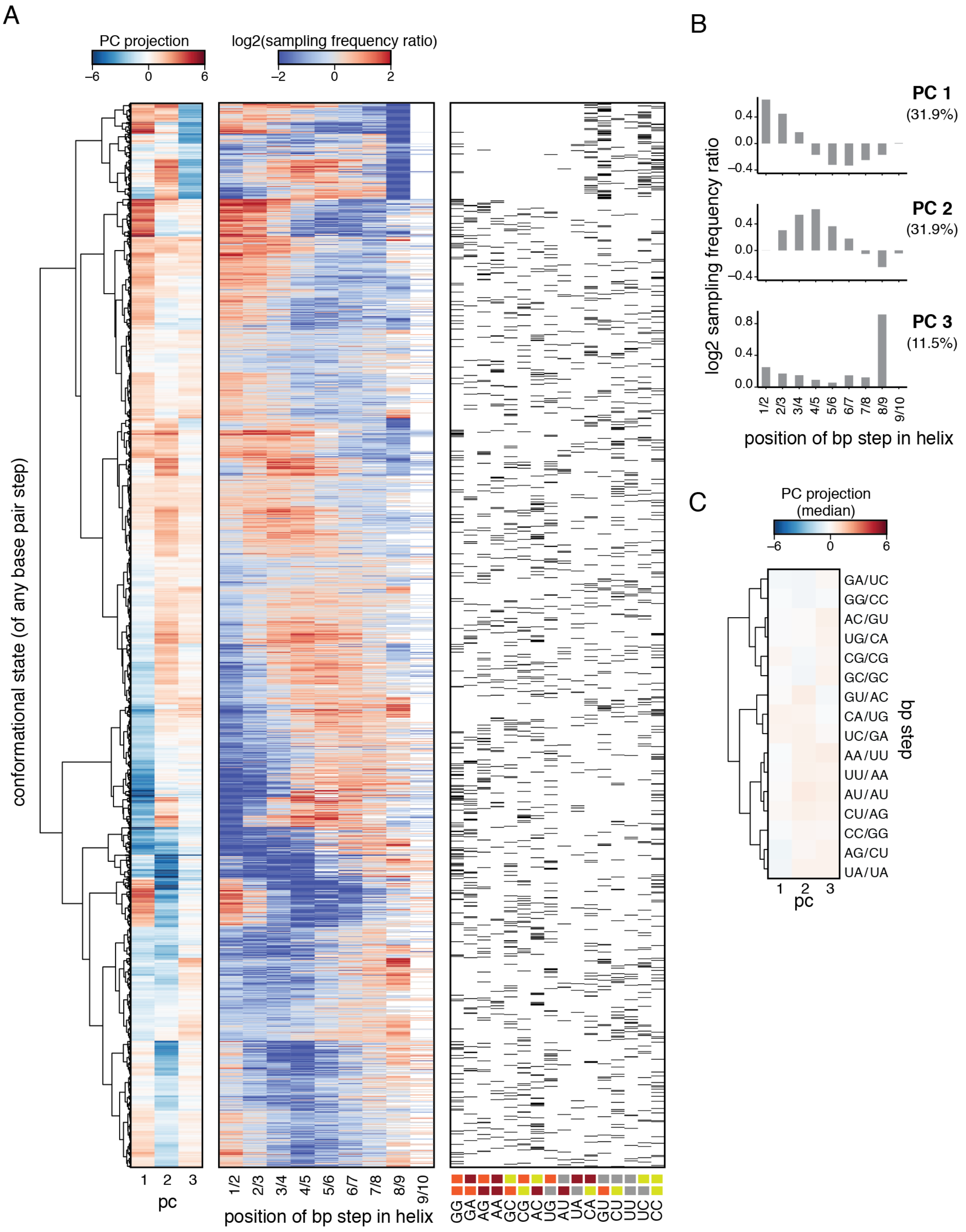

A
PC projection
−6
0
6
log2(sampling frequency ratio)
−2
0
2
conformational state (of any base pair step)
1
2
3
pc
1/2
2/3
3/4
4/5
5/6
6/7
7/8
8/9
9/10
position of bp step in helix
GG
GA
AG
AA
GC
CG
AC
UG
AU
UA
CA
GU
CU
UU
UC
CC
B
log2 sampling frequency ratio
PC 1
(31.9%)
0.4
0.0
−0.4
PC 2
(31.9%)
0.4
0.0
−0.4
PC 3
(11.5%)
0.8
0.4
0.0
1/2
2/3
3/4
4/5
5/6
6/7
7/8
8/9
9/10
position of bp step in helix
C
PC projection
(median)
−6
0
6
GA/UC
GG/CC
AC/GU
UG/CA
CG/CG
GC/GC
GU/AC
CA/UG
UC/GA
AA/UU
UU/AA
AU/AU
CU/AG
CC/GG
AG/CU
UA/UA
bp step
1
2
3
pc

Figure S10. Base pair step conformers have position-dependent sampling frequencies.

A) The sampling frequency of each base pair step conformational state was compared to the expected (i.e. within the unconstrained tectoRNA) to obtain the conformational state's sampling frequency ratio for each position within the tectoRNA. (Left heatmap) Sampling frequency ratios across positions were projected into the top three principal components (PCs; shown in (B)) determined with PC analysis. These values were hierarchically clustered to obtain dendrogram at left. (Middle heatmap) Shown are the sampling frequency ratios for each conformational state across positions. (Right heatmap) The base pair step identity is shown for each of the conformational states (black corresponds to a match). While some structure was evident (i.e. certain conformational states of the GU, CU, UC, and CC ensembles are not sampled at position 8/9), in general, the conformation states associated with particular position-dependent sampling behaviors could belong to any base pair step type. In legend, base pair step refers to the 5´ strand of the helix sub-sequence, e.g. 'GG' corresponds to 5´-GG-3´/5´-CC-3´. B) Values for the log2 of the sampling frequency ratio associated with each PC. C) The median value of the PC projections (shown in A; left) for each of the base pair steps.

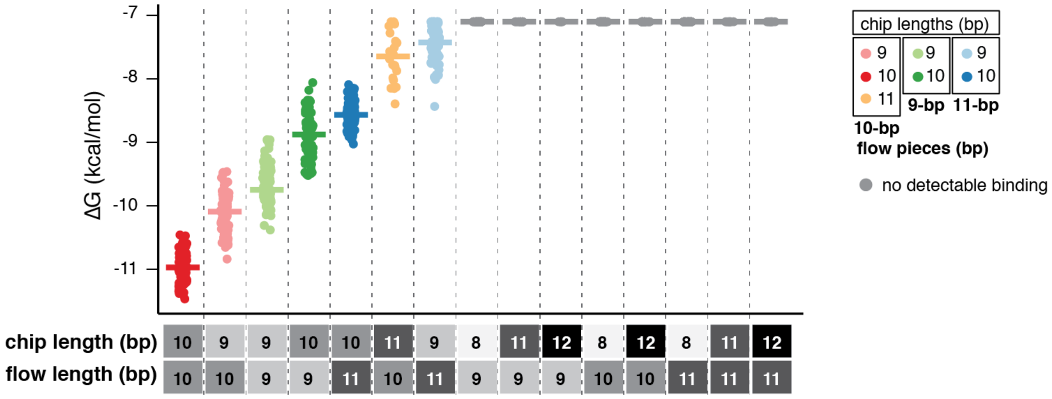


Figure S11. Measured binding affinity (ΔG) of different length-paired complexes.

Between 32 and 96 different WC sequences were measured for each chip length. The length of the chip- and flow- piece helices is indicated. Chip pieces of length 8 bp or length 12 bp have dissociation constants were too destabilized to observe.

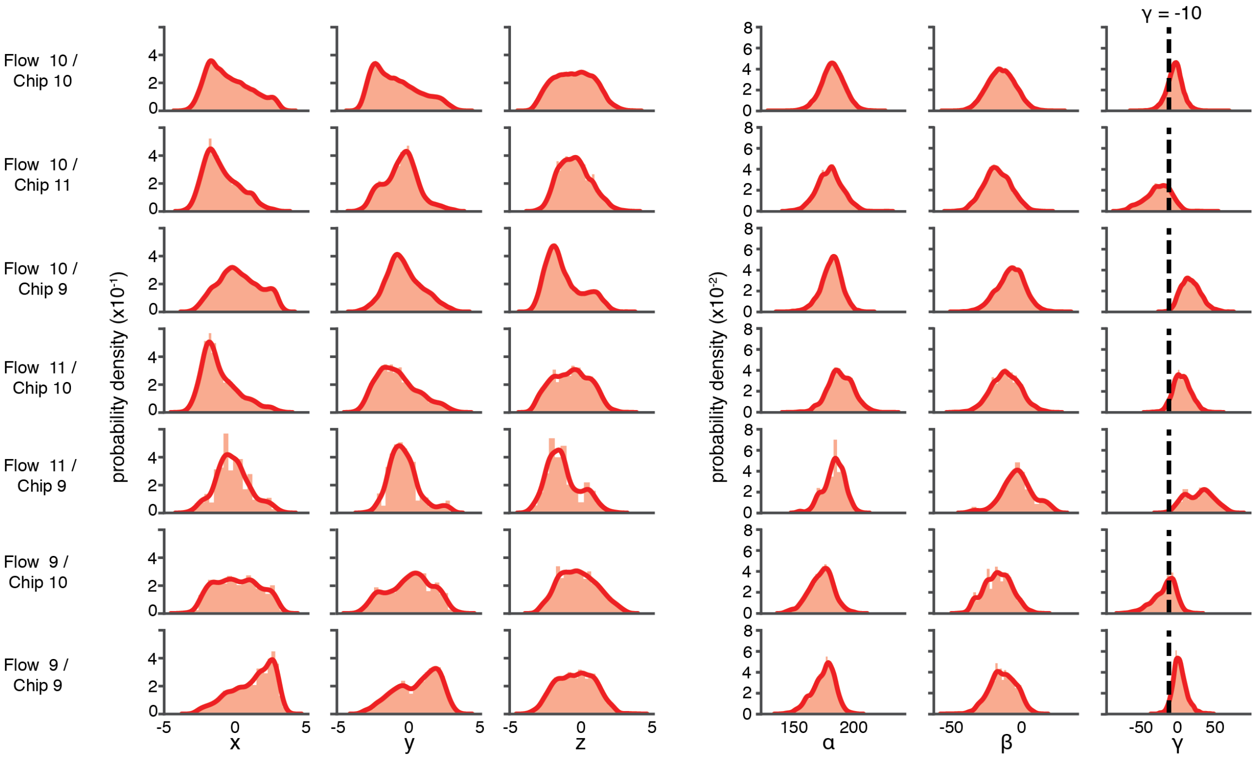


Figure S12. Distribution of six-dimensional values of bound conformation of for each tectoRNA length topology.

Distribution of values describing the position (i.e. x, y, z) and alignment (α, β, γ) of the final base pair of the partially bound tectoRNA, compared to where it would be in the closed tectoRNA structure, for the set of conformations determined to be bound (i.e. distance score < 5, see Eq. 1) for each tectoRNA length topology. Vertical dashed line indicates the more stringent cutoff applied to identify "bound" conformations in Figure 5B (right), where in addition to the distance score < 5, the gamma parameter had to be greater than this value (–10 degrees). Only Flow 10 / Chip 11 and Flow 9 / Chip 10 were significantly affected.

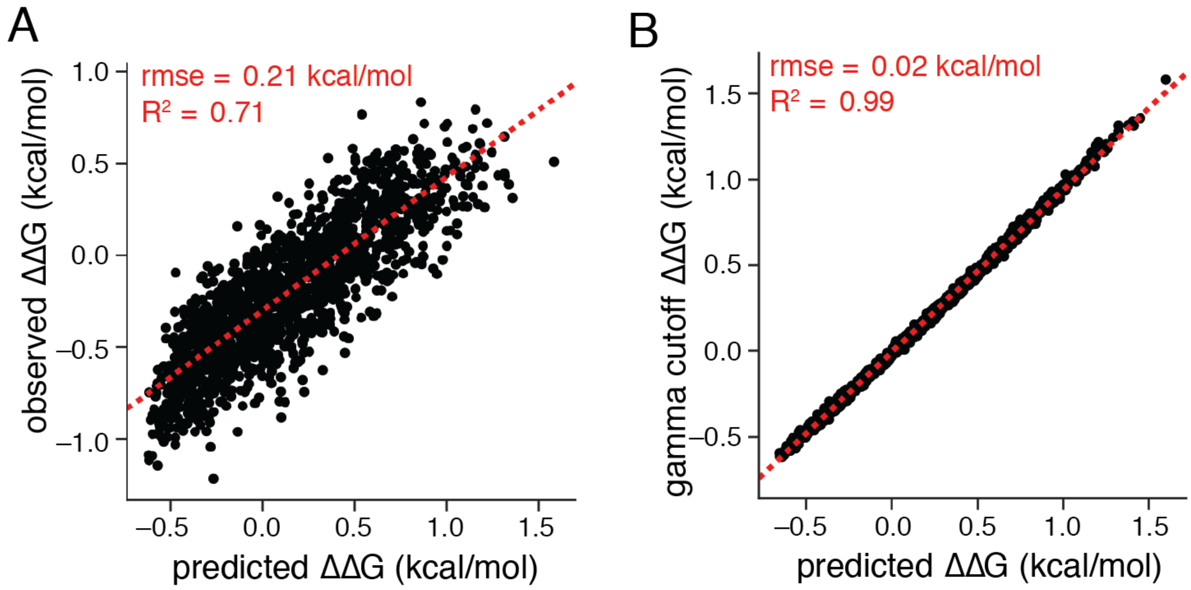


Figure S13. Gamma corrected predictions of sequence-dependent set

(A) Predicted ΔΔGs with gamma cutoff compared to observed values. (B) Comparison between predictions before and after gamma cutoff.

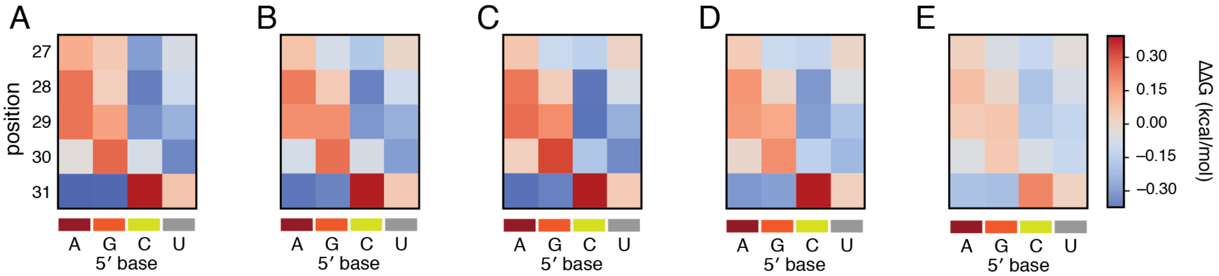


Figure S14. Prediction of helical sequence preference of anticodon helix for aminoacyl-tRNA•EF-Tu accommodation during ribosome codon recognition.

A-E) Predicted dependence of A/T-tRNA^{Thr} binding free energy on sequence of the anticodon helix with the indicated base pair at each position within the helix. Each heatmap is from an independently solved structure, yet the sequence dependence is consistent across all models. RNAMake calculations were performed over all 4^5 anticodon helix sequences (see SI Appendix, Dataset S4) A) 4V5G. B) 4V5P. C) 4V5Q. D) 4V5R. E) 4V5S. Rigorous tests of the RNAMake predictions will require high-precision pre-steady-state or single molecule measurements that isolate the binding equilibrium of EF-Tu-bound tRNA into the A/T state.

Supplemental References

1. Buenrostro JD, Araya CL, Chircus LM, Layton CJ, Chang HY, Snyder MP, et al. Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. Nat Biotechnol. 2014 Jun;32(6):562–8.

2. She R, Chakravarty AK, Layton CJ, Chircus LM, Andreasson JOL, Damaraju N, et al. Comprehensive and quantitative mapping of RNA-protein interactions across a transcribed eukaryotic genome. Proc Natl Acad Sci USA. 2017 Apr 4;114(14):3619–24.

3. Denny SK, Bisaria N, Yesselman JD, Das R, Herschlag D, Greenleaf WJ. High-Throughput Investigation of Diverse Junction Elements in RNA Tertiary Folding. Cell. 2018 Jul 12;174(2):377–390.e20.

4. Petrov AI, Zirbel CL, Leontis NB. Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. RNA. 2013 Oct;19(10):1327–40.

5. Lu X-J, Bussemaker HJ, Olson WK. DSSR: an integrated software tool for dissecting the spatial structure of RNA. Nucleic Acids Res. 2015 Dec 2;43(21):e142.

6. Lu X-J, Olson WK. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res. 2003 Sep 1;31(17):5108–21.

7. Watkins AM, Geniesse C, Kladwang W, Zakrevsky P, Jaeger L, Das R. Blind prediction of noncanonical RNA structure at atomic accuracy. BioRxiv. 2017 Nov 22;

8. Huynh DQ. Metrics for 3D rotations: comparison and analysis. J Math Imaging Vis. 2009 Oct;35(2):155–64.

9. Karney CFF. Quaternions in molecular modeling. J Mol Graph Model. 2007 Jan;25(5):595–604.

10. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The Vienna RNA websuite. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W70-4.

11. Moulinier L, Eiler S, Eriani G, Gangloff J, Thierry JC, Gabriel K, et al. The structure of an AspRS-tRNA(Asp) complex reveals a tRNA-dependent control mechanism. EMBO J. 2001 Sep 17;20(18):5290–301.

12. Eiler S, Dock-Bregeon A, Moulinier L, Thierry JC, Moras D. Synthesis of aspartyl-tRNA(Asp) in Escherichia coli--a snapshot of the second step. EMBO J. 1999 Nov 15;18(22):6532–41.

13. Cavarelli J, Eriani G, Rees B, Ruff M, Boeglin M, Mitschler A, et al. The active site of yeast aspartyl-tRNA synthetase: structural and functional aspects of the aminoacylation reaction. EMBO J. 1994 Jan 15;13(2):327–37.

14. Schmeing TM, Voorhees RM, Kelley AC, Gao Y-G, Murphy FV, Weir JR, et al. The crystal structure of the ribosome bound to EF-Tu and aminoacyl-tRNA. Science. 2009 Oct 30;326(5953):688–94.

15. Schmeing TM, Voorhees RM, Kelley AC, Ramakrishnan V. How mutations in tRNA distant from the anticodon affect the fidelity of decoding. Nat Struct Mol Biol. 2011 Apr;18(4):432–6.

16. Perona JJ, Hou Y-M. Indirect readout of tRNA for aminoacylation. Biochemistry. 2007 Sep 18;46(37):10419–32.