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

Joseph D. Yesselman[a,1], Sarah K. Denny[b,1,2], Namita Bisaria[a,3], Daniel Herschlag[a,c,d,4], William J. Greenleaf[b,c,e,f,g,4], and Rhiju Das[a,h,4]

[a]Department of Biochemistry, Stanford University, Stanford, CA 94305; [b]Program in Biophysics, Stanford University, Stanford, CA 94305; [c]Department of Chemistry, Stanford University, Stanford, CA 94305; [d]Stanford ChEM-H (Chemistry, Engineering, and Medicine for Human Health), Stanford University, Stanford, CA 94305; [e]Department of Genetics, Stanford University, Stanford, CA 94305; [f]Department of Applied Physics, Stanford University, Stanford, CA 94305; [g]Chan Zuckerberg Biohub, San Francisco, CA 94158; and [h]Department of Physics, Stanford University, Stanford, CA 94305

Edited by Hashim M. Al-Hashimi, Duke University Medical Center, Durham, NC, and accepted by Editorial Board Member Michael F. Summers June 25, 2019 (received for review February 1, 2019)

Structured RNAs and RNA complexes underlie biological processes ranging from control of gene expression to protein translation. Approximately 50% of nucleotides within known structured RNAs are folded into Watson–Crick (WC) base pairs, and sequence changes that preserve these pairs are typically assumed to preserve higher-order RNA structure and binding of macromolecule partners. Here, we report that indirect effects of the helix sequence on RNA tertiary stability are, in fact, significant but are nevertheless predictable from a simple computational model called RNAMake-ΔΔG. When tested through the RNA on a massively parallel array (RNA-MaP) experimental platform, blind predictions for >1500 variants of the tectoRNA heterodimer model system achieve high accuracy (rmsd 0.34 and 0.77 kcal/mol for sequence and length changes, respectively). Detailed comparison of predictions to experiments support a microscopic picture of how helix sequence changes subtly modulate conformational fluctuations at each base-pair step, which accumulate to impact RNA tertiary structure stability. Our study reveals a previously overlooked phenomenon in RNA structure formation and provides a framework of computation and experiment for understanding helix conformational preferences and their impact across biological RNA and RNA-protein assemblies.

blind prediction | RNA energetics | high-throughput data | indirect readout

BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Structured RNAs perform a wealth of essential biological functions, including the catalysis of peptide bond formation, gene expression regulation, and genome maintenance. In each case, the RNA folds into a complex 3D structure whose thermodynamics governs its function (1–5). Interrogation of the folding process has yielded a general picture in which the RNA structure generally forms hierarchically, first through the formation of Watson–Crick (WC) double helices—the RNA secondary structure—and then through assembly of these helices through non-WC interactions into tertiary structures (6–8). Extensive in vitro measurements have enabled a thermodynamic model that can generally predict the RNA secondary structure from the RNA sequence (9, 10). However, no thermodynamic model exists to predict tertiary structure formation from a secondary structure, even though this final step is fundamental to RNA function.

Understanding RNA tertiary structure requires methods to predict possible 3D structures and to estimate their relative energetics; both steps require careful accounting of the geometric preferences and flexibility of the individual elements that compose the RNA (6–8, 11–14). In recent years, the major focus of RNA modeling groups has been outside canonical base-paired helices and instead on noncanonical motifs, such as structured junctions and tertiary contacts, which are the hallmarks of complex tertiary structure (11–13, 15–20). Nevertheless, within structured RNAs, over 50% of residues are still contained within WC base-paired helices (21), implying that even subtle conformational variation in WC base pairs (as observed in refs. 22–24) might accumulate to substantially influence tertiary structure folding. Several lines of evidence suggest that such sequence-dependent conformational variations in RNA helices could exist. Depending on their sequences, RNA helices have different mechanical properties (22) and distinct chemical shift profiles as determined by NMR (25). In addition, there is extensive work on sequence-dependent conformational preferences of nucleic acid helices in the related field of DNA-protein assembly. Such preferences underlie "indirect readout" effects in which sequence changes in double helix segments in between, but not directly at, protein-DNA contacts can change DNA-protein binding affinities by up to 200-fold (3 kcal/mol at 37 °C), and modeling studies that explicitly consider conformational ensembles can partially reproduce these data (26–28). For RNA tertiary structure, analogous

Significance

Structured RNAs fold into complex tertiary structures to perform critical roles in a multitude of biological functions. Over half the nucleotides in structured RNAs form simple Watson–Crick (WC) double helices, which can then assemble through non-WC interactions into elaborate tertiary structures. Here, we report the serendipitous discovery that sequence changes in WC base pairs impact the energetics of RNA tertiary folding. These observations led to a computational model for helix conformational fluctuations that then blindly predicted the results of thousands of high-throughput experiments with surprisingly high accuracy. Our study reveals that sequence-specific helix structure preferences are needed for understanding RNA folding quantitatively and outlines a route for dissecting the impact of helix conformational fluctuations across general RNA biophysical events.

Author contributions: J.D.Y., S.K.D., D.H., W.J.G., and R.D. designed research; J.D.Y., S.K.D., and N.B. performed research; J.D.Y. and S.K.D. analyzed data, and J.D.Y., S.K.D., and R.D. wrote the paper, with participation by all authors.

Conflict of interest statement: D.H. and R.D. have coauthored manuscripts with H.M.A.-H. in the past 2 years.

This article is a PNAS Direct Submission. H.M.A.-H. is a guest editor invited by the Editorial Board.

[1]J.D.Y. and S.K.D. contributed equally to this work.

[2]Present address: Scribe Therapeutics, Berkeley, CA 94704.

[3]Present address: Whitehead Institute for Biomedical Research, Cambridge, MA 02142.

[4]To whom correspondence may be addressed. Email: herschla@stanford.edu, wjg@stanford.edu, or rhiju@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1901530116/-/DCSupplemental.

changes in RNA double helix conformational ensembles in between, but not directly at, tertiary contacts could impact the stability of RNA tertiary structure assemblies (27, 29). However, such effects have not yet been tested, partially due to the difficulty of separating out such effects from other complicating factors in RNA structure formation, including possible changes in secondary structure, the typical presence of multiple tertiary contacts, and the involvement of single-stranded RNA regions.

Overcoming these difficulties, the tectoRNA model system involves binding 2 RNA pieces with well-defined secondary structures through 2 well-understood tetraloop/receptor tertiary contacts that are connected by 10 base-pair helices (Fig. 1*A*) (30–32). We recently reported that the tectoRNA is amenable to quantitative experiments involving thousands of distinct variants through the RNA-MaP technology (14, 33). Here, we describe how serendipitous early observations of helix-dependent effects in tectoRNA RNA-MaP measurements led us to develop a computational method that models the sequence-dependent conformations of WC base-pair steps and uses these conformations to quantitatively predict the energetics of the tertiary assembly. Computational simulations generated blind predictions of the relative affinity of all possible helix sequence variants of one piece of the tectoRNA heterodimer ($>10^5$ predictions). We then measured >1500 of these previously uncharacterized tectoRNA variants, including comprehensive changes in base-pair sequence and length of 1 helix. Our results establish that sequence- and length-dependent conformational effects of helical elements influence the thermodynamic stability of tertiary structures over unexpectedly wide ranges of 40-fold and 2,000-fold, respectively, and that these effects can be predicted with high accuracy.

Results

High-Throughput Platform to Measure Thermodynamic Stability of TectoRNAs. Our model system is shown in Fig. 1*A*. Each piece of the tectoRNA heterodimer is composed of a 10-bp RNA helix flanked by a tetraloop (TL) and by a tetraloop receptor (TLR) (30). The TL of 1 monomer binds selectively to the TLR of the other monomer, forming 2 tertiary contacts that stabilize the heterodimer (Fig. 1*A*). Suboptimal positioning of the 2 tertiary contact interfaces by the intervening helices destabilizes the heterodimer (32, 34). The tectoRNA system is thus sensitive to

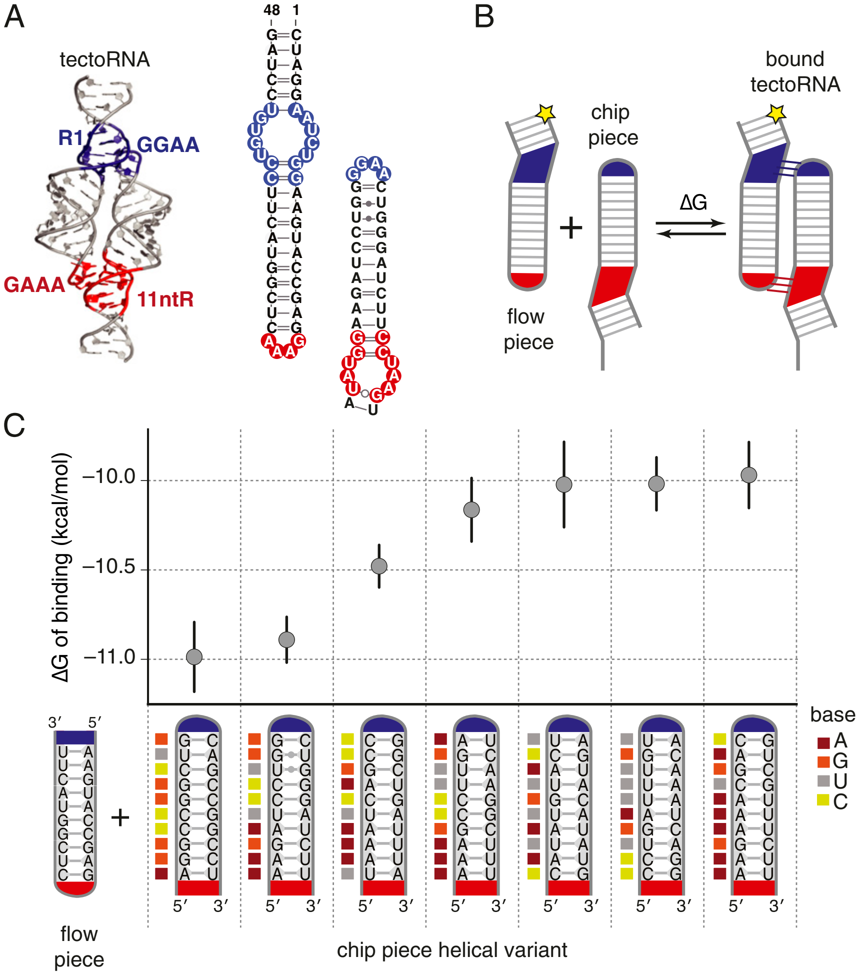


Fig. 1. Free energy of tectoRNA binding depends on helix sequence. (*A*) Structure of tectoRNA homodimer [Protein Data Bank (PDB): 2ADT] with 2 tertiary contacts (GAAA-11nt). One of these tertiary contacts is replaced (GGAA-R1; blue) to convert the complex to the heterodimer used in this study (32). On the right is the sequence and secondary structure of the wild-type tectoRNA interaction. Numbers indicate the "position" within the chip-piece helix. (*B*) In our experimental setup, one piece of the heterodimer was fluorescently labeled and free in solution (the "flow piece"), while the other was immobilized on the surface of a sequencing chip (chip piece). Quantification of the bound flow piece to the chip surface allowed determination of the free energy of binding (ΔG) to form the bound tectoRNA. (*C*) Free energy of binding of the flow piece to 7 distinct chip-piece variants. Error bars are 95% CI on the measured ΔG. The sequence of the flow- and chip-piece helices is indicated (*Bottom*).

the conformational preferences of RNA helices and provides a quantitative thermodynamic readout in the form of heterodimer binding affinity.

A library of sequence variants of one piece of the tectoRNA heterodimer was designed, synthesized, and sequenced (Fig. 1*B* and *SI Appendix*, Fig. S1*A*). We leveraged a modified sequencing platform to in situ transcribe the library into RNA directly on the surface of the sequencing chip (*Methods*), enabling the display of sequence-identified clusters of RNA (*SI Appendix*, Fig. S1*B*) (33). This piece of the tectoRNA heterodimer was thus called the chip piece. The binding partner of the chip piece (the flow piece) was fluorescently labeled and introduced to the sequencing chip flow cell at a series of increasing concentrations, and the amount of bound fluorescence to each cluster of RNA was quantified after equilibration (*Methods*). These fluorescence values were used to derive the affinity of the flow piece to each chip piece variant in terms of the equilibrium dissociation constant (K_d) and binding free energy ($\Delta G = RT \log K_d$). Values for ΔG obtained in 2 independent experiments were highly reproducible ($R^2 = 0.92$; rmsd = 0.15 kcal/mol; *SI Appendix*, Fig. S2*A*). Each chip piece variant was present in multiple locations per chip ($n \geq 5$), allowing estimation of confidence intervals for each affinity measurement [median uncertainty on ΔG = 0.16 kcal/mol (95% CI); *SI Appendix*, Fig. S2*B*]. In previously tested systems, RNA-MaP measurements correspond directly to gel-shift assays (33, 35), and the binding affinities for the tectoRNA are similar to those measured for the original constructs (4 nM for the 10-bp heterodimer measured in ref. 32 compared with 6–30 nM measured for 10-bp heterodimers in our experiment) (32).

A preliminary experiment measured 7 chip-piece RNA variants with different arbitrarily chosen WC base-pair compositions. We observed a 5-fold range of binding affinities (1 kcal/mol; Fig. 1*C*), contrary to our initial expectation that these assemblies would have the same affinity and thereby act as controls. The

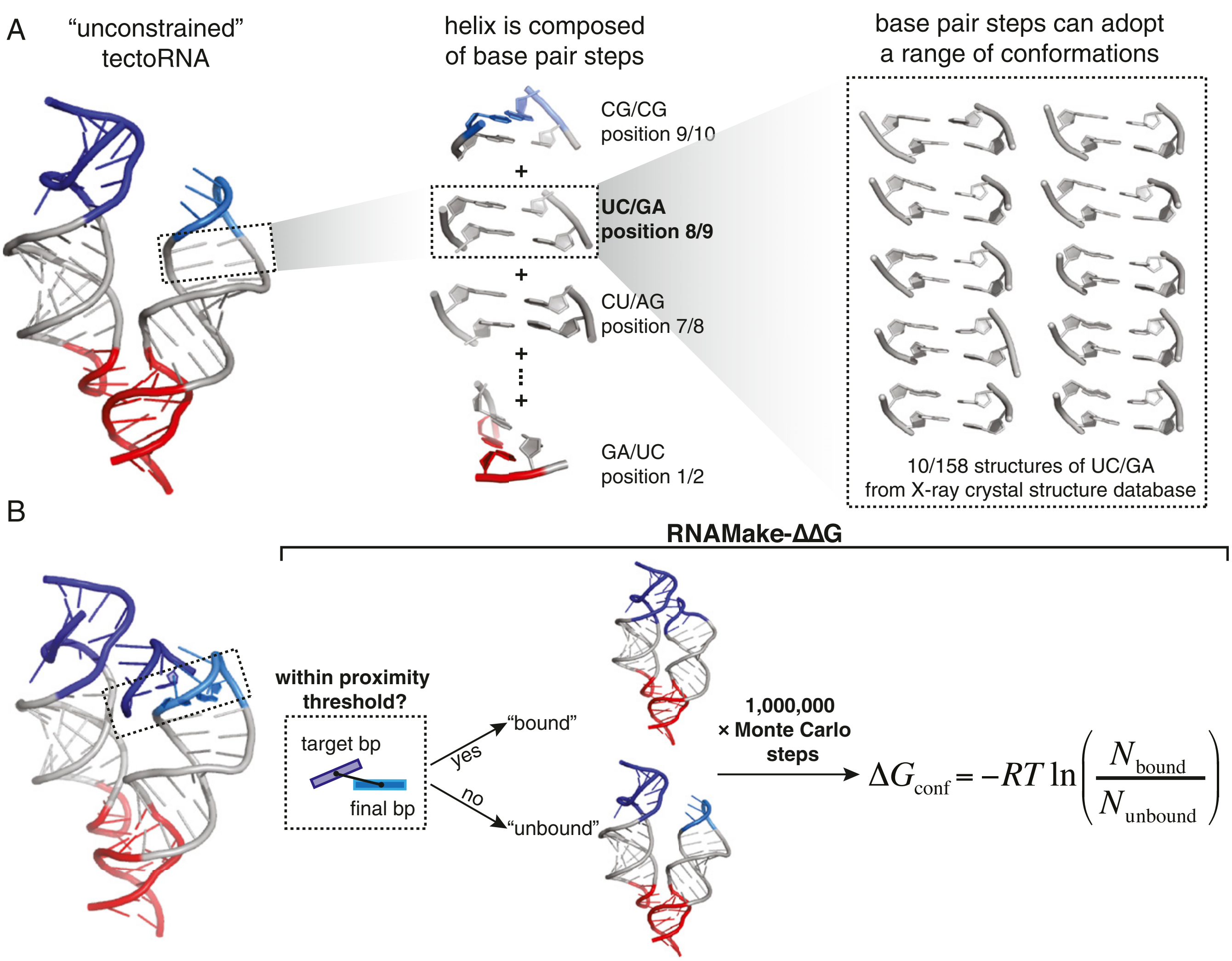


Fig. 2. Ensemble model for RNA helices allows prediction of tectoRNA assembly energetics. (*A*, *Left*) The modeled structure of the unconstrained tectoRNA (i.e., with one contact formed) is shown. The global structure was assembled from the structures of its constituent elements, including the base-pair steps that compose the helical regions. (*A*, *Center*) Example base-pair steps are shown for the chip-piece helix. Each base-pair step can adopt an ensemble of many possible conformations, which were derived from examples of that base-pair step in the crystallographic database. (*A*, *Right*) Example conformations within the UC/GA conformational ensemble are shown. (*B*) Starting with the unconstrained tectoRNA as shown in *A*, a Monte Carlo simulation was performed. At each step of the simulation, the structure of one base-pair step in the tectoRNA was replaced with a new state from its conformational ensemble. The new structure of the unconstrained tectoRNA assembly was evaluated for whether it was "bound" or "unbound," according to the translational and rotational distances from the target base pair to the final base pair. One million steps were performed, and the total number of computed bound and unbound tectoRNA conformations were used to calculate the free energy change between the bound and the unbound tectoRNAs (ΔG_{conf}).

serendipitous observation of these affinity differences inspired the development of a computational model (described below) to relate helix structure to tectoRNA stability, based on structural differences between WC base pairs.

Conformational Ensembles of RNA Helices Predict TectoRNA Stability in RNAMake-ΔΔG. We developed a computational model for tectoRNA stability that explicitly models the conformational ensemble for each RNA helix sequence, i.e., the distribution of conformations that the unconstrained helix explores in solution. Inspired by previous modeling procedures pioneered by Olson and colleagues (see refs. 22 and 36), we divided each helix into a set of base-pair steps (i.e., 2 sequential base pairs) (Fig. 2*A*). Decomposition of helices in this manner allows for modeling of arbitrary helix sequences using a minimal set of structural states. Base-pair step conformational ensembles were determined by compiling all instances of that base-pair step in structured RNAs from the RNA crystal structure database (Fig. 2*A*, *Right* and *Methods*) (22, 37–39). These base-pair step structures were then clustered based on structural similarity to form a set of 50–250 discrete conformational states, each weighted according to its frequency (*Methods* and *SI Appendix*, Table S1).

Modeling the tectoRNA additionally required structures for each of the TL/TLR tertiary contacts, which we modeled as single structural conformations, as this type of tertiary contact appears nearly structurally identical across all extant crystallographic structures (40). These conformations were derived from a crystal structure and Rosetta modeling (41) for the GAAA-11nt and GGAA-R1 TL/TLR interactions, respectively (see *Methods*).

With this model we generated the "unconstrained" tectoRNA—i.e., the intermediate state of tectoRNA binding where only a single tertiary contact is formed (Fig. 2*A*). In this unconstrained state, the helices explore their full sterically allowed conformational ensembles and occasionally bring the loop and receptor of the second tertiary contact in close enough proximity to form the closed tectoRNA assembly (Fig. 2*B*). We sampled conformations explored by the unconstrained tectoRNA with a Monte Carlo simulation by swapping the conformation of one randomly chosen base-pair step per simulation iteration. Each sampled conformation of the tectoRNA was assessed for whether the closing base pair of the unbound TL was in close proximity to its position in the bound TL/TLR (Fig. 2*B*), based on a proximity threshold of 5 Å and a rotational alignment term (see *Methods* and refinement below), to define whether the structure was closed with both contacts formed (bound) or not (unbound) (Fig. 2*B*). This assessment was used to calculate the free energy of conformational alignment of the tertiary contacts,

$$\Delta G_{\mathrm{conf}} = -RT\ \log(N_{\mathrm{bound}}/N_{\mathrm{unbound}}),$$

where T is the temperature, R is the universal gas constant, and N_{bound} and N_{unbound} are the number of simulated structures annotated as bound or unbound, respectively. We attributed differences in binding affinity between any 2 tectoRNA variants ($\Delta\Delta G_{\mathrm{binding}}$) to differences in this conformational alignment term,

$$\Delta\Delta G_{\mathrm{binding}} = \Delta G_{\mathrm{conf,2}} - \Delta G_{\mathrm{conf,1}},$$

where $\Delta G_{\mathrm{conf,1}}$ and $\Delta G_{\mathrm{conf,2}}$ are the conformational alignment terms for 2 variants (indicated by 1 and 2, respectively). For a more detailed justification of how other physical effects cancel out in this difference, see ref. 40. This model was built as an extension of RNAMake, a toolkit for the design of the RNA 3D structure (42), to predict thermodynamics of tertiary structure formation; thus we call the method RNAMake-ΔΔG.

We generated ΔG_{conf} for all possible sequences of the 4 canonical base pairs within the chip-piece helix using RNAMake-ΔΔG (*Methods*), and these calculations predicted a substantial effect of helix sequence on tectoRNA assembly of 2.5 kcal/mol, corresponding to a 70-fold effect on affinity (*SI Appendix*, Fig. S3).

Blind Tests of Sequence-Dependent TectoRNA Stability. We next tested the predictions of RNAMake-ΔΔG in a blind prediction challenge. We selected 2000 tectoRNA sequences that were predicted (by author J.D.Y.) to uniformly span the predicted range of affinity. Two authors (S.K.D. and N.B.) then carried out high-precision measurements for 1,596 of these sequences (the remaining sequences were not sufficiently represented in our library). The tested sequences gave experimental tertiary stabilities spanning a range of affinity of 2.1 kcal/mol (corresponding to a 40-fold effect on K_d) between the lowest and the highest affinity binders, similar to the predicted range of 2.5 kcal/mol (a 70-fold range in K_d). These data confirmed that sequence-dependent conformations of RNA helices can have a substantial effect on tertiary structure formation.

Strikingly, we observed a high correlation between the observed and the predicted affinities ($R^2 = 0.71$) with rmsd of 0.34 kcal/mol to the predicted line of fixed slope = 1 (Fig. 3*A*). Allowing the slope to vary gave a slightly better prediction (rmsd = 0.21 kcal/mol; best-fit slope = 0.54) (Fig. 3*A*). The accuracy of these blind predictions of tertiary energetics was better than the scale of thermal fluctuations ($RT = 0.6$ kcal/mol). The good agreement between our observed and the predicted values suggests that this computational model captures structural differences among helices that, in turn, influence the thermodynamics of tertiary structure formation.

After our blind predictions, we investigated whether the magnitude of the proximity threshold used to evaluate base-pair overlap, the choice of base pair at which to evaluate overlap, and the choice of starting conformation affected the accuracy of the model. There is a large range of proximity thresholds that give similar R^2 values, although the slope between our predictions and the observed values changes slightly (*SI Appendix*, Fig. S4). In addition, our predictions are largely independent of the base pair at which we evaluated overlap as well as the starting conformation for simulations (*SI Appendix*, Fig. S5).

To help visualize the formation of the tectoRNA assembly, we present in Fig. 3 *B* and *C* the modeled conformational ensembles of 2 tectoRNA variants from the extremes of the range of tectoRNA affinity measurements (magenta = −10.2 kcal/mol, cyan = −12.0 kcal/mol). Fig. 3*B* shows a subset of the chip-piece helix trajectories, while Fig. 3*C* shows the modeled distribution of the final base pair of the flow and chip-piece helix, projected on the *x-y* plane. Both the low- and the high-affinity chip-piece helices sample a wide range of RNA backbone trajectories in the unconstrained tectoRNA ensembles with variation in the position of the final base pair of more than 7 Å (full width at half maximum in the *x* and *y* directions; Fig. 3*C*). The median position of the final base pair differed by 5.3 Å between the 2 chip-piece helices with the end of the helix being substantially farther from the flow piece for the low-affinity variant (Fig. 3*C*). For both cases and especially the destabilized case (magenta), our modeling suggested that the chip piece was bound to the flow piece only in the subset of conformational states making more extreme conformational excursions (i.e., compare black and gray trajectories in Fig. 3*B*). Further supporting this picture, attempting to model binding affinity using only a single most populated structure for each base-pair step produced worse predictions ($R^2 = 0.42$; *SI Appendix*, Fig. S6 *A* and *B*). Finally, our modeling suggested that certain structural differences between helix sequences had large effects on thermodynamic stability, while others had minimal effects (*SI Appendix*, Fig. S6 *C* and *D* and the next section). By taking the difference between the centroid of the bound states and the unconstrained states, we determined a spatial projection of the structural differences most coupled to thermodynamic effects. Differences between helices along this projection were highly correlated to the observed ΔΔG values ($R^2 = 0.71$),

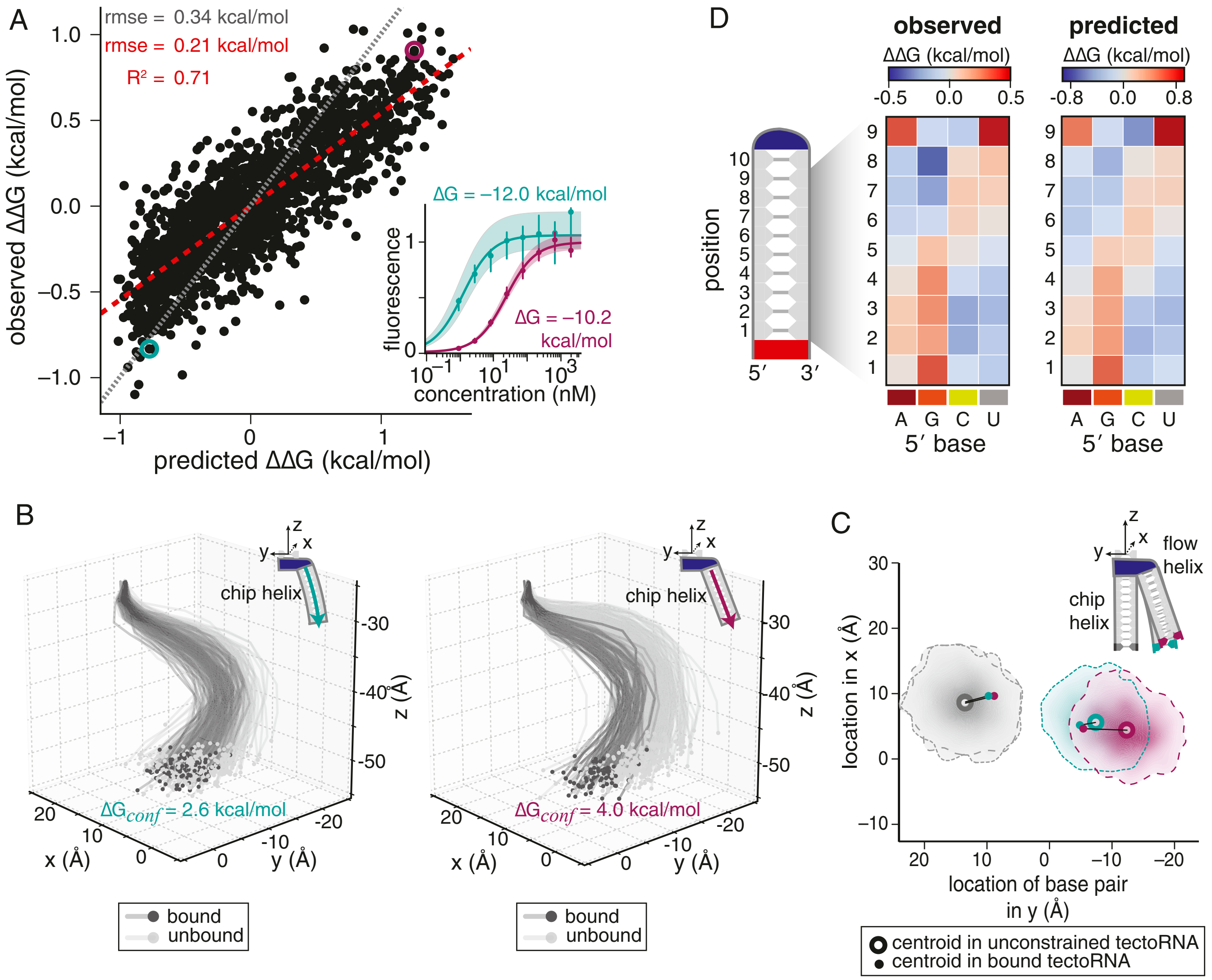


Fig. 3. RNAMake-ΔΔG accounts for changes in tectoRNA affinity in a blind prediction challenge. (*A*) Blind predictions generated with the RNAMake-ΔΔG model agree well with observed values of tectoRNA binding ΔG for 1,536 chip-piece variants ($R^2 = 0.71$). Each set of ΔG values is compared with their respective medians to obtain ΔΔGs. The red dashed line indicates the best-fit line (slope = 0.54); the gray dotted line indicates the line of slope 1. *Inset* shows the measured binding affinity curves of 2 chip-piece variants. (*B*) Example 3D trajectories of the chip-piece helix produced during the Monte Carlo sampling for 2 variants whose binding curves are shown in A). For each variant, 250 unbound trajectories (light gray) and 100 bound trajectories are shown (dark gray). All trajectories are aligned by the top base pair. The traces are through the center of each base pair in the helix. (*C*) Distribution of the terminating base pair of the chip and flow pieces in the partially bound tectoRNA projected on the *x*-*y* plane. Distributions were determined using bivariate kernel density estimate smoothing of ∼1,000 bound or partially bound structures sampled from the simulation. The centroids of the distributions are shown as open circles; the black lines connect the centroid of the partially bound structures to the centroid of the bound structures (black dot). (*D*) Observed (*Left*) and predicted (*Right*) affinities for chip-piece helices with the indicated base pair at each position within the helix. Affinities are given as the deviation from the median observed or predicted affinity across all 1,536 variants.

while differences along a perpendicular axis were uncorrelated (*SI Appendix*, Fig. S6*D*). Thus, specific differences between static structures may be used to predict and understand thermodynamic effects, albeit less directly than with the full computational model.

Base-Pair Elements Adopt Distinct Structures at Different Positions. To gain insight into the how primary sequence affects binding probability in this system, we determined the average effect on tectoRNA affinity (ΔΔG) of having any given base pair at each position within the helix, compared with the average affinity of all 1,594 tested variants (Fig. 3*D* and *SI Appendix*, Fig. S7). These effects were highly correlated between the observed and the predicted values ($R^2 = 0.93$; *SI Appendix*, Fig. S7*A* and Fig. 3*D*). Each base pair has either stabilizing or destabilizing effects depending on its position within the helix (Fig. 3*D*). Base pairs with a purine residue on the 5′ side of the helix (i.e., A-U and G-C base pairs) were destabilizing when placed closer to the receptor (positions 1–3) but stabilizing when placed closer to the loop (positions 6–8), while the reverse was true for base pairs with a purine on the 3′ side of the helix (i.e., U-A and C-G base pairs; Fig. 3*D*). This observed position dependence of sequence preference strongly contrasts with the "nearest-neighbor rules" governing secondary structure energetics in which each base-pair step contributes an additive free energy term toward the overall free energy of folding, regardless of its position within a helix (43). This observation also suggests that partial unfolding of the secondary structure is not responsible for the differences in tectoRNA assembly formation (see also *SI Appendix*, Fig. S8).

The overall trend in position dependence suggests a simplifying rule that conformational preferences of purine pyrimidine base pairs are similar but are distinct from pyrimidine purine base pairs. However, an exception to this rule is evident at position 9 where A-U and U-A were both destabilizing. This base pair is adjacent to the closing base pair of the loop, leading us to consider whether this base pair adopted substantially different conformations in the bound tectoRNA due to the proximity of the tertiary contact. However, the observed effect was highly correlated with the effect predicted by the RNAMake-ΔΔG model (Fig. 3 *D*, position 9 row). Therefore, even at this loop-proximal base-pair step, our data can be understood without invoking any physical effects beyond the intrinsic base-pair step conformational preferences used in RNAMake-ΔΔG.

To achieve a more granular understanding of the position-dependent structural preferences of base-pair steps, we quantified the contribution of each of the base-pair step's conformational states in the bound tectoRNA. States with an increased representation (over and above the expected sampling frequency from the Monte Carlo simulation) in the bound tectoRNA should correspond to the states that promote binding and vice versa for those with a decrease in representation (*SI Appendix*, Fig. S9). We observed disproportionate representation of certain states within each base-pair step's ensemble in the bound tectoRNA (illustrated for the AU/AU ensemble in Fig. 4*A* and for all base pairs in *SI Appendix*, Fig. S10). Notably, these changes were highly position dependent such that the majority of states could be over-represented or under-represented, depending on their position within the chip-piece helix (Fig. 4*A*). To illustrate further, conformational states of the AU/AU ensemble were clustered based on their position-dependent representation (shown in a dendrogram and colors in Fig. 4*A*). Conformational states in different clusters were each associated with distinct structural behaviors with small but consistent structural differences between structures in different clusters (>1-Å differences; Fig. 4 *B* and *C*). For example, conformers in class 6, which promote binding in positions 1–3 in the helix, are more twisted and thus span less translational distance than conformers in class 1, which promote binding only in the very first or last base pair in the helix (Fig. 4 *B* and *C*). These results would predict that the same base-pair element adopts different conformations in the bound state depending on its location within the helix, thereby accounting for the differential base-pair preferences along the helix (Fig. 3*D*). These different conformational preferences further underscore the necessity of an ensemble to account for thermodynamic effects in RNA tertiary structure formation.

Testing RNAMake-ΔΔG at More Extreme Helical Distortions. We next explored RNAMake-ΔΔG's capacity to predict the thermodynamic effects of helix length changes by adding or deleting base pairs on both the flow and the chip RNAs. We generated chip RNAs with helix lengths of 8–12 bp (n = 32–96 sequence variants per length) and tested each chip RNA against flow RNAs with helix lengths of 9–11 bp, yielding 15 length-pair combinations (Fig. 5*A*). For each of these complexes, we calculated ΔΔG values relative to the original assemblies with 10-bp flow and 10-bp chip helices, which we abbreviate as "10/10 bp." Certain highly mismatched length combinations were so destabilizing that no binding was detectable (ΔΔG > 4.4 kcal/mol relative to 10/10 bp; 8 length-pair combinations; *SI Appendix*, Fig. S11). The remaining length-pair combinations had effects spanning this 4.4 kcal/mol range. The thermodynamic stability of each length-pair complex with observable binding was calculated with RNAMake-ΔΔG. Comparisons to measurements demonstrated a correlation of R^2 = 0.66 and rmsd = 0.72 kcal/mol for these predictions, with the best-fit line having a slope indistinguishable

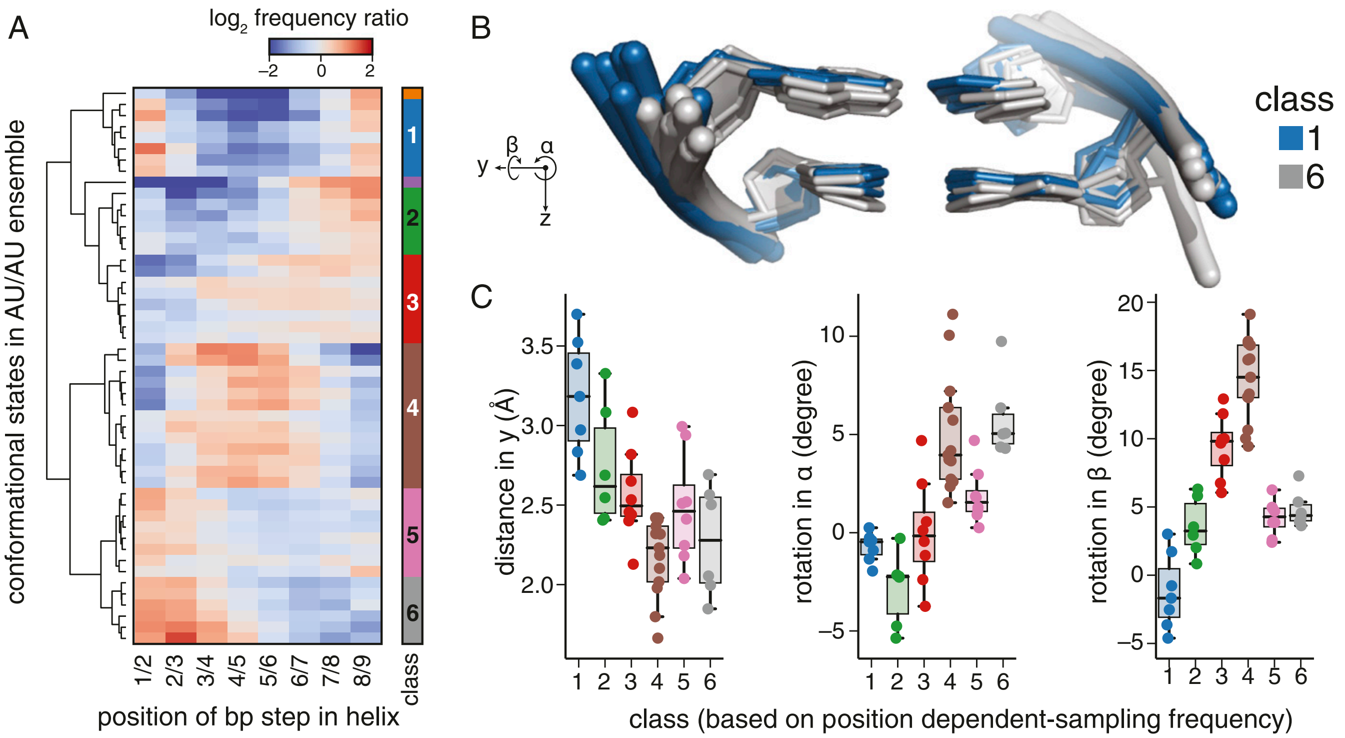


Fig. 4. Base-pair conformations differ by position within the helix. (*A*) Change in sampling frequency of conformational states in the AU/AU ensemble in the bound versus the partially bound. (*B*) Example structures of base-pair step conformations that are enriched and depleted at 2 positions. (*C*) Change in positioning between enriched and depleted conformational states at each position of any base-pair type. (see *SI Appendix*, Fig. S8 for other coordinates). Enriched = sampling frequency more than 2-fold greater than expected, and depleted = sampling frequency less than 2-fold less than expected.

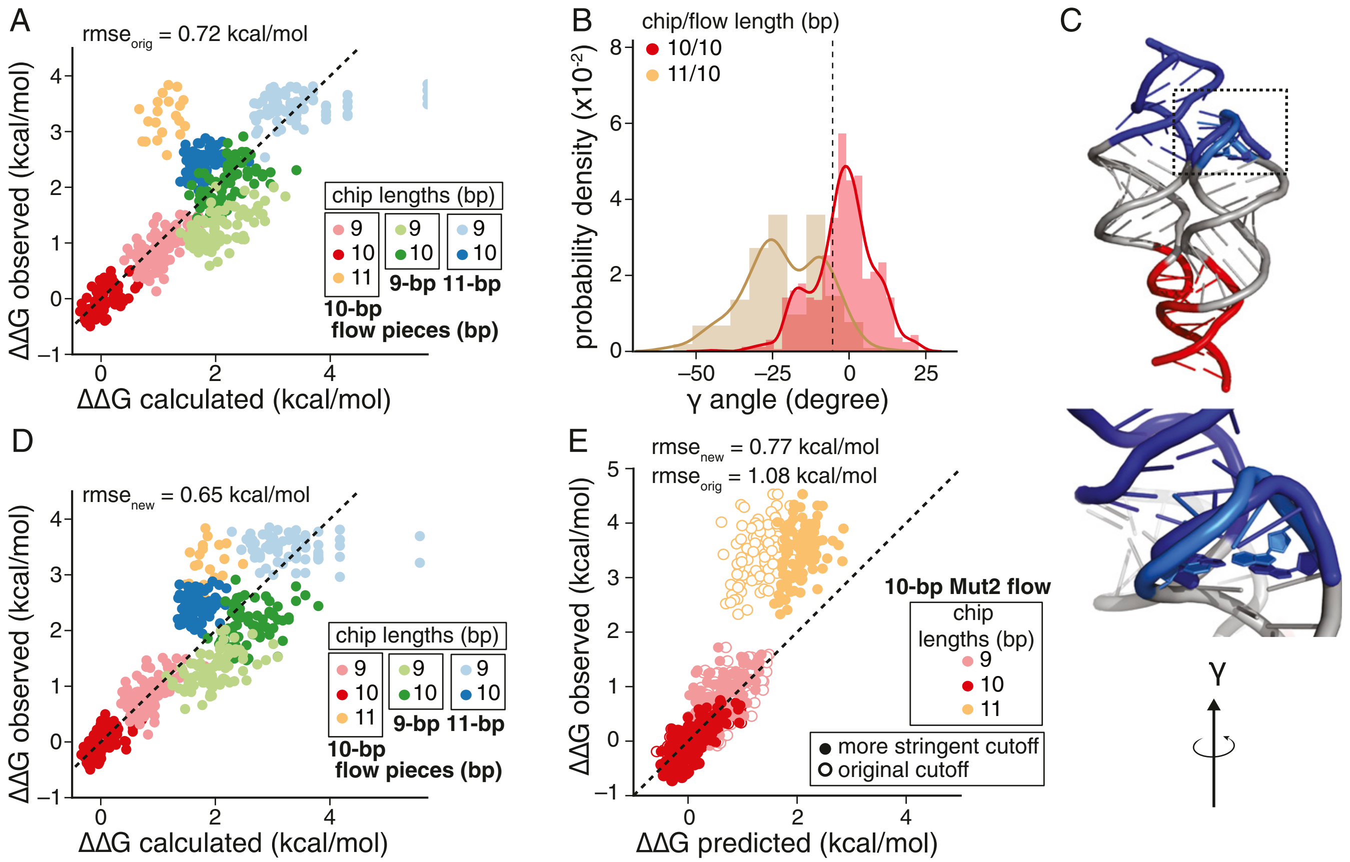


Fig. 5. Increased prediction accuracy of different length pairs with refinement of the bound state cutoff. (*A*) Observed versus calculated affinities for chip- and flow-piece variants with altered lengths. The colors indicate the length of the flow- and chip-piece helices. (*B*) Distribution of the value for Euler angle γ within 2 bound tectoRNA complexes, where γ represents the rotation between the final bp and the target bp around the *z* axis. The 11-bp chip-piece variant has distinct values for γ compared with the 10-bp chip-piece variant. The vertical dashed line indicates γ = −10°. (*C*) Structure of the bound complex with the original cutoff (light blue) or a more stringent cutoff (blue) where γ has to be > −10°. (*D*) Observed and calculated affinities for length-pair combinations with the more stringent cutoff which excluded overtwisted conformations (i.e., γ > −10° in the bound complex). The colors indicate the length of the flow- and chip-piece helices as in *A*; observed values are the same as in *A*. (*D*) Observed versus predicted (blind prediction values) of a new set of chip-piece sequences against a distinct 10-bp flow piece using either the original model (open circles) or the updated model with the more stringent cutoff (closed circles).

from 1 (Fig. 5*A*). The larger rmsd compared with the 10/10-bp sequence predictions appears due to systematic deviations between the observed and the predicted effects for specific length pairs. For example, the 10/11-bp flow/chip complexes uniformly bound more weakly than predicted, while the 9/9-bp flow/chip complexes were bound slightly tighter than predicted.

One possible explanation for predicting stronger binding than is observed is an overly accommodating proximity threshold for determining bound tectoRNA structures during prediction. Such a loose threshold would allow unrealistic structures to be considered bound during the RNAMake-ΔΔG simulation. To assess this possibility, we analyzed the distribution of bound tectoRNA conformations in the 6 values describing the overlap in our proximity threshold [the difference in position (x,y,z) and alignment Euler angles (α,β,γ)]. There was a striking difference in the distributions in bound tectoRNA of a 10/11-bp flow/chip complex compared with other topologies with respect to the twist Euler angle γ: conformations with $\gamma < -10$ were significantly enriched (Fig. 5*B* and *SI Appendix*, Fig. S12). We hypothesized that these states were binding incompetent, leading to the discrepancy between observed and predicted values for these length-pair complexes. To avoid classifying these states as bound, we tested a more stringent cutoff by implementing an additional criteria that the helix within the bound complex cannot be substantially undertwisted (Euler angle $\gamma > -10°$; see Fig. 5 *B* and *C*). With this additional constraint, the agreement between our calculated and the observed ΔΔG for all length pairs improved significantly ($R^2 = 0.71$; rmsd = 0.65 kcal/mol; Fig. 5*D*). Additionally, we applied this cutoff to the sequence-dependent set and observed no significant difference in predictions (*SI Appendix*, Fig. S13).

To test this refined proximity threshold, we carried out a second blind prediction challenge with calculations and experiments carried out independently by authors J.D.Y. and S.K.D., respectively. The affinity of additional 300 chip variants of 3 different lengths (9, 10, and 11 bp) were measured against a distinct 10-bp flow piece. These tectoRNA variants represented a wider diversity of sequences than those used to refine the proximity criterion. The blind predictions using the additional constraint demonstrated a significantly improved relationship between the observed and the predicted binding affinities, although it did not completely account for the destabilizing effect of this length-pair complex (rmse, original model = 1.08 kcal/mol; rmse updated model = 0.77 kcal/mol; Fig. 5*E*). The development of this additional constraint on the bound conformation suggests the utility of an iterative protocol for refining the anisotropic binding landscape of a tertiary contact.

Discussion

A major goal in understanding the many fundamental biological complexes containing RNA has been to develop a model for predicting RNA structure and energetics from a primary sequence.

We have presented here extensive experimental and computational evidence for a factor that has largely been neglected in these studies: RNA double helix conformational preferences that depend on helix sequence can impact RNA tertiary structure energetics. RNAMake-ΔΔG gives quantitative estimates for how helix sequence and length can change the favorability of bringing together segments that make RNA–RNA tertiary contacts and makes thousands of testable predictions for the tectoRNA heterodimer model system for tertiary assembly. High-throughput measurements with RNA-MaP allowed rigorous blind tests of this model and confirmed its predictions with accuracies of 0.34 and 0.77 kcal/mol for effects of sequence and length changes, respectively. These RNAMake-ΔΔG accuracies are somewhat better than those achieved in post hoc modeling efforts for protein-DNA indirect readout (0.9 kcal/mol, ref. 26) and are similar to those achieved in recent blind prediction of nearest-neighbor parameters for the RNA secondary structure (44, 45).

The conformational ensembles arising from RNAMake-ΔΔG modeling gives a detailed physical description of how RNA helices "look" inside tertiary assemblies. For example, the same base-pair sequence is predicted to have different physical structures when embedded at different positions in the tertiary assembly, and this phenomenon explains the qualitatively different sequence preferences at each position, observed in both computation and experiment (Fig. 3). The model gives a view of such structural effects as spread throughout the helix and not focused at 1 particular "kink" within the helix, providing support that small deviations can accumulate to cause larger energetic effects. Importantly, this view implies that most current schemes to model the RNA tertiary structure through optimization of local pairwise interactions will be unable to model such long-range cumulative effects without including a new term analogous to the RNAMake-ΔΔG calculations herein. It will be important to expand the RNAMake model to include conformational ensembles for RNA structural elements beyond helices; preliminary work on G•U wobble pairs and other "mismatches" suggests that such modeling will be feasible (*SI Appendix*, Table S2).

We anticipate that our computational framework will be useful for understanding the energetic costs and sequence preferences associated with RNA double helix distortions that occur

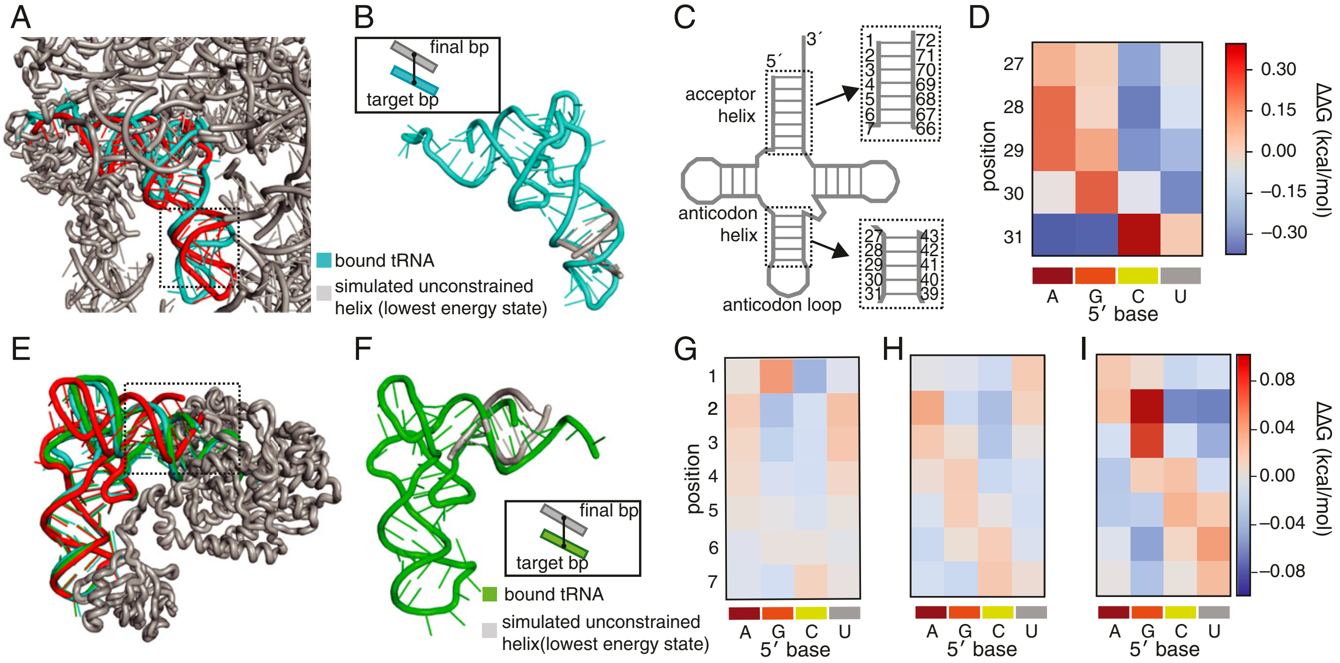


Fig. 6. Prediction of RNA double helix distortions that occur during ribosomal A-site accommodation and amino acid charging. (*A*) When complexed with EF-Tu and being loaded into the A-site of the ribosome (the A/T state), *Thermus thermophilus* tRNA^{Thr} appears bent (cyan, PDB: 4V5G) compared with *Escheria coli* tRNA^{Phe} only complexed with EF-Tu (red, PDB: 1OB2); (*B*) Overlay of the target fully A/T-bound configuration of the anticodon helix (cyan) and example RNAMake-modeled configuration (gray); *Inset* shows how scoring occurs between the target base pair from the bound tRNA and the last base pair in the RNAMake built model. (*C*) The secondary structure of tRNA and the location of the anticodon helix and acceptor helix (boxed). (*D*) Predicted dependence of A/T-tRNA^{Thr} binding free energy on the sequence of the anticodon helix with the indicated base pair at each position within the helix. Additional heat maps from independently solved structures give indistinguishable sequence dependences (*SI Appendix*, Fig. S14). RNAMake-calculations were performed over all 4^5 anticodon helix sequences (Dataset S4). Rigorous tests of the RNAMake predictions will require high-precision presteady-state or single molecule measurements that isolate the binding equilibrium of EF-Tu-bound tRNA into the A/T state. (*E*) tRNA^{asp} from either *E. coli* (cyan, 1C0A) or yeast (green, 1IL2) form similar conformations when bound to *E. coli* aspartyl-tRNA synthetase (AspRS). This conformation is bent at the acceptor helix compared with a structure of a partially bound yeast tRNA^{asp} that does not make contact to the synthetase at its acceptor end and was cocrystallized with the bound conformation (red, 1IL2). (*F*) Overlay of the target fully bound configuration (green) and example RNAMake-modeled configuration (gray); the inset shows how scoring occurs between the target base pair from the bound tRNA and the last base pair in the RNAMake built model. (*G*–*I*) Predicted dependence of tRNA-AspRS binding free energy on the acceptor stem sequence with the indicated base pair at each position within the helix. RNAMake calculations were performed over all 4^7 acceptor helix sequences (Dataset S3). While the predicted effects are small in magnitude, calculations with target-bound conformations drawn from (*G*) *E. coli* tRNA/*E. coli* AspRS (1C0A) and (*H*) yeast tRNA/*E. coli* AspRS (1IL2) give similar predicted preferences with slight differences arising from the slightly different sequences and AspRS-bound structures taken by the 2 tRNAs in nucleotides outside the acceptor stem. The sequence preference map for (*F*) binding of yeast tRNA^{asp} to the yeast aspartyl-tRNA synthetase (1ASZ) is quite distinct. Reference binding free energies for ΔΔG are based on RNAMake calculations with the *E. coli* tRNA^{asp} sequence (*G* and *H*) and the yeast tRNA^{asp} sequence (*I*). Note that the scale of effects (0.2 kcal/mol or less) is smaller than the differences in enzymatic rates (1 to 2 kcal/mol) for the few tRNA combinations reported in refs. 29 and 47, suggesting that effects beyond conformational bending account for those results, such as the differences in chemical modification or processing in tRNAs prepared in vivo. Rigorous tests of the RNAMake predictions will require high-precision thermodynamic measurements using in vitro prepared tRNA substrates.

throughout RNA biological processes, such as the amino acid charging and multistage ribosomal readout of tRNAs (27, 29, 46). However, the effects of changing any single helix base pair on the energetics of RNA structure or complex formation may be <1 kcal/mol, and so qualitative, low-throughput measurements will not be sufficient for understanding the energetics of such distortions. Indeed, in our own paper, it has been critical to make predictions and measurements across thousands of sequences to convincingly demonstrate our model of helix conformational preferences as well as its quantitative limits.

To aid future studies, we have made extensive predictions for 2 RNA systems in which "indirect readout" effects have been previously hypothesized (29, 46): anticodon helix sequence effects on aminoacyl-tRNA•EF-Tu accommodation during ribosome codon recognition (Fig. 6 *A–D* and *SI Appendix*, Fig. S14) and acceptor helix sequence effects on tRNA^{Asp} aminoacylation (Fig. 6 *E–I*). We look forward to upcoming advances in RNA-MaP and other high-throughput biophysical methods that will enable stringent tests of these quantitative predictions for fundamental events in RNA molecular biology.

Methods

Detailed methods for the design, preparation, and experimental measurements of binding affinities for the tectoRNA library as well as the simulation protocol of RNAMake-ΔΔG (including basic equations, simulation parameters, and scoring function) are presented in the *SI Appendix*.

ACKNOWLEDGMENTS. We thank Curtis Layton and Johan Andreasson for developing, building, and maintaining the imaging station and W.J.G., D.H., and R.D. laboratory members for reagents and critical feedback. This work was supported by the National Institute of Health (Grant P01 GM066275 to D.H.; R01 GM111990 to W.J.G.; R01 GM100953 and R35 GM122579 to R.D.; and R01 GM121487 to P. Bradley [lead principal investigator], supporting W.J.G.). W.J.G. acknowledges support as a Chan-Zuckerberg Investigator. S.K.D. was supported, in part, by the Stanford Biophysics training Grant (T32 GM008294) and by the NSF Graduate Research Fellowship Program (GRFP). N.B. was supported, in part, by the NSF GRFP. J.D.Y. was supported by the Ruth L. Kirschstein National Research Service Award Postdoctoral Fellowships GM112294.

1. M. J. Moore, From birth to death: The complex lives of eukaryotic mRNAs. *Science* **309**, 1514–1518 (2005).
2. J. L. Rinn, H. Y. Chang, Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145–166 (2012).
3. H. F. Noller, RNA structure: Reading the ribosome. *Science* **309**, 1508–1514 (2005).
4. P. Nissen, J. Hansen, N. Ban, P. B. Moore, T. A. Steitz, The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**, 920–930 (2000).
5. T. H. D. Nguyen *et al*., The architecture of the spliceosomal U4/U6.U5 tri-snRNP. *Nature* **523**, 47–52 (2015).
6. I. Tinoco, Jr, C. Bustamante, How RNA folds. *J. Mol. Biol.* **293**, 271–281 (1999).
7. P. Brion, E. Westhof, Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.* **26**, 113–137 (1997).
8. D. Herschlag, S. Bonilla, N. Bisaria, The story of RNA folding, as told in epochs. *Cold Spring Harb. Perspect. Biol.* **10**, a032433 (2018).
9. M. G. Seetin, D. H. Mathews, RNA structure prediction: An overview of methods. *Methods Mol. Biol.* **905**, 99–122 (2012).
10. P. P. Gardner, R. Giegerich, A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* **5**, 140 (2004).
11. R. Das, J. Karanicolas, D. Baker, Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods* **7**, 291–294 (2010).
12. J. Bernauer, X. Huang, A. Y. L. Sim, M. Levitt, Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA* **17**, 1066–1075 (2011).
13. M. J. Boniecki *et al*., SimRNA: A coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.* **44**, e63 (2016).
14. S. K. Denny *et al*., High-throughput investigation of diverse junction elements in RNA tertiary folding. *Cell* **174**, 377–390.e20 (2018).
15. A. T. Frank, A. C. Stelzer, H. M. Al-Hashimi, I. Andricioaei, Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: New insights into RNA dynamics and adaptive ligand recognition. *Nucleic Acids Res.* **37**, 3670–3679 (2009).
16. X. Shi, P. Walker, P. B. Harbury, D. Herschlag, Determination of the conformational ensemble of the TAR RNA by X-ray scattering interferometry. *Nucleic Acids Res.* **45**, e64 (2017).
17. X. Shi, L. Huang, D. M. J. Lilley, P. B. Harbury, D. Herschlag, The solution structural ensembles of RNA kink-turn motifs and their protein complexes. *Nat. Chem. Biol.* **12**, 146–152 (2016).
18. L. Salmon, G. Bascom, I. Andricioaei, H. M. Al-Hashimi, A general method for constructing atomic-resolution RNA ensembles using NMR residual dipolar couplings: The basis for interhelical motions revealed. *J. Am. Chem. Soc.* **135**, 5457–5466 (2013).
19. L. Salmon *et al*., Modulating RNA alignment using directional dynamic kinks: Application in determining an atomic-resolution ensemble for a hairpin using NMR residual dipolar couplings. *J. Am. Chem. Soc.* **137**, 12954–12965 (2015).
20. C. D. Eichhorn, H. M. Al-Hashimi, Structural dynamics of a single-stranded RNA-helix junction using NMR. *RNA* **20**, 782–791 (2014).
21. J. Stombaugh, C. L. Zirbel, E. Westhof, N. B. Leontis, Frequency and isostericity of RNA base pairs. *Nucleic Acids Res.* **37**, 2294–2312 (2009).
22. F.-C. Chou, J. Lipfert, R. Das, Blind predictions of DNA and RNA tweezers experiments with force and torque. *PLoS Comput. Biol.* **10**, e1003756 (2014).
23. J. A. Abels, F. Moreno-Herrero, T. van der Heijden, C. Dekker, N. H. Dekker, Single-molecule measurements of the persistence length of double-stranded RNA. *Biophys. J.* **88**, 2737–2744 (2005).
24. J. Lipfert *et al*., Double-stranded RNA under force and torque: Similarities to and striking differences from double-stranded DNA. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 15408–15413 (2014).
25. S. Barton, X. Heng, B. A. Johnson, M. F. Summers, Database proton NMR chemical shifts for RNA signal assignment and validation. *J. Biomol. NMR* **55**, 33–46 (2013).
26. N. B. Becker, L. Wolff, R. Everaers, Indirect readout: Detection of optimized subsequences and calculation of relative binding affinities using different DNA elastic potentials. *Nucleic Acids Res.* **34**, 5638–5649 (2006).
27. D. E. Draper, Protein-RNA recognition. *Annu. Rev. Biochem.* **64**, 593–620 (1995).
28. R. Rohs *et al*., Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* **79**, 233–269 (2010).
29. J. J. Perona, Y.-M. Hou, Indirect readout of tRNA for aminoacylation. *Biochemistry* **46**, 10419–10432 (2007).
30. L. Jaeger, N. B. Leontis, Tecto-RNA: One-Dimensional Self-Assembly through Tertiary Interactions. *Angew. Chem. Int. Ed. Engl.* **39**, 2521–2524 (2000).
31. L. Nasalean, S. Baudrey, N. B. Leontis, L. Jaeger, Controlling RNA self-assembly to form filaments. *Nucleic Acids Res.* **34**, 1381–1392 (2006).
32. C. Geary, S. Baudrey, L. Jaeger, Comprehensive features of natural and in vitro selected GNRA tetraloop-binding receptors. *Nucleic Acids Res.* **36**, 1138–1152 (2008).
33. J. D. Buenrostro *et al*., Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nat. Biotechnol.* **32**, 562–568 (2014).
34. L. Jaeger, E. Westhof, N. B. Leontis. TectoRNA: Modular assembly units for the construction of RNA nano-objects. *Nucleic Acids Res.* **29**, 455–463 (2001).
35. I. Jarmoskaite *et al*., A quantitative and predictive model for RNA binding by human pumilio proteins. *Mol. Cell* **74**, 966–981.e18 (2019).
36. W. Olson, A. Colasanti, L. Czapla, G. Zheng, "Insights into the sequence-dependent macromolecular properties of DNA from base-pair level modeling" in *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Gregory A. Voth, Ed. (CRC Press, 2009).
37. W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock, V. B. Zhurkin, DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 11163–11168 (1998).
38. H. M. Berman *et al*., The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
39. A. I. Petrov, C. L. Zirbel, N. B. Leontis, Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *RNA* **19**, 1327–1340 (2013).
40. N. Bisaria, M. Greenfeld, C. Limouse, H. Mabuchi, D. Herschlag, Quantitative tests of a reconstitution model for RNA folding thermodynamics and kinetics. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E7688–E7696 (2017).
41. A. M. Watkins *et al*., Blind prediction of noncanonical RNA structure at atomic accuracy. *Sci. Adv.* **4**, eaar5316 (2018).
42. J. D. Yesselman *et al*., Computational design of asymmetric three-dimensional RNA structures and machines. bioRxiv:10.1101/223479 (21 November 2017).
43. D. H. Turner, N. Sugimoto, S. M. Freier. RNA structure prediction. *Ann Rev Biophys Biophys Chem*. **17**, 167–192 (1988).
44. F.-C. Chou, W. Kladwang, K. Kappel, R. Das, Blind tests of RNA nearest-neighbor energy prediction. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 8430–8435 (2016).
45. D. J. Wright, C. R. Force, B. M. Znosko, Stability of RNA duplexes containing inosine·cytosine pairs. *Nucleic Acids Res.* **46**, 12099–12108 (2018).
46. T. M. Schmeing *et al*., The crystal structure of the ribosome bound to EF-Tu and aminoacyl-tRNA. *Science* **326**, 688–694 (2009).

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.

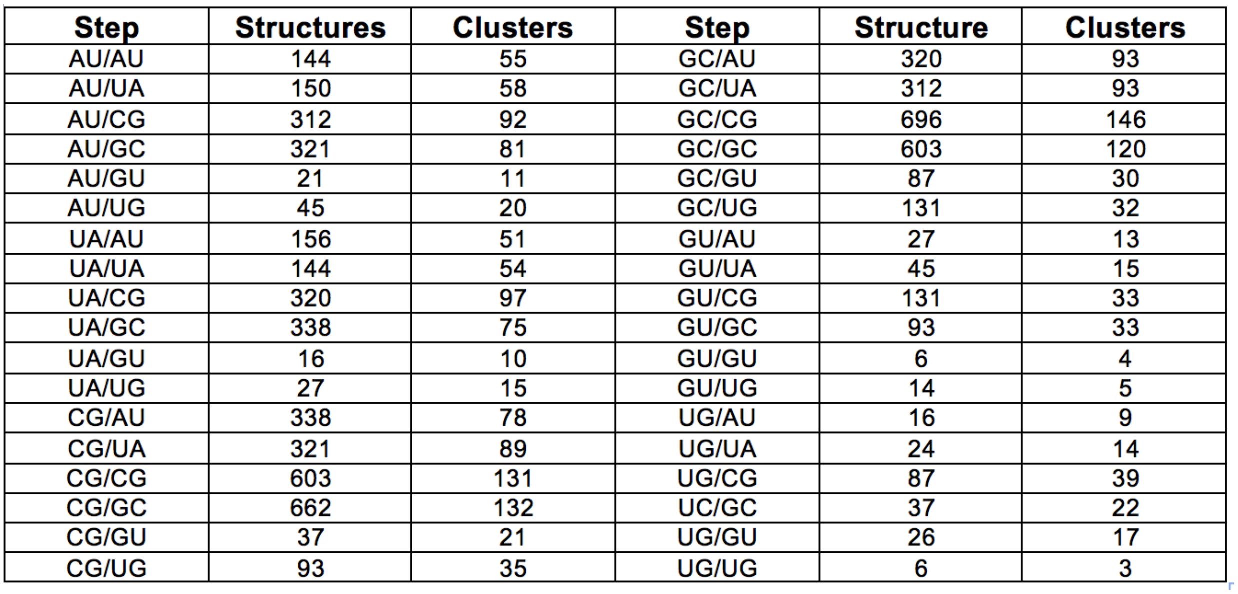
Computing ΔΔGs for anticodon helix for aminoacyl-tRNA•EF-Tu accommodation during ribosome codon recognition.

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.

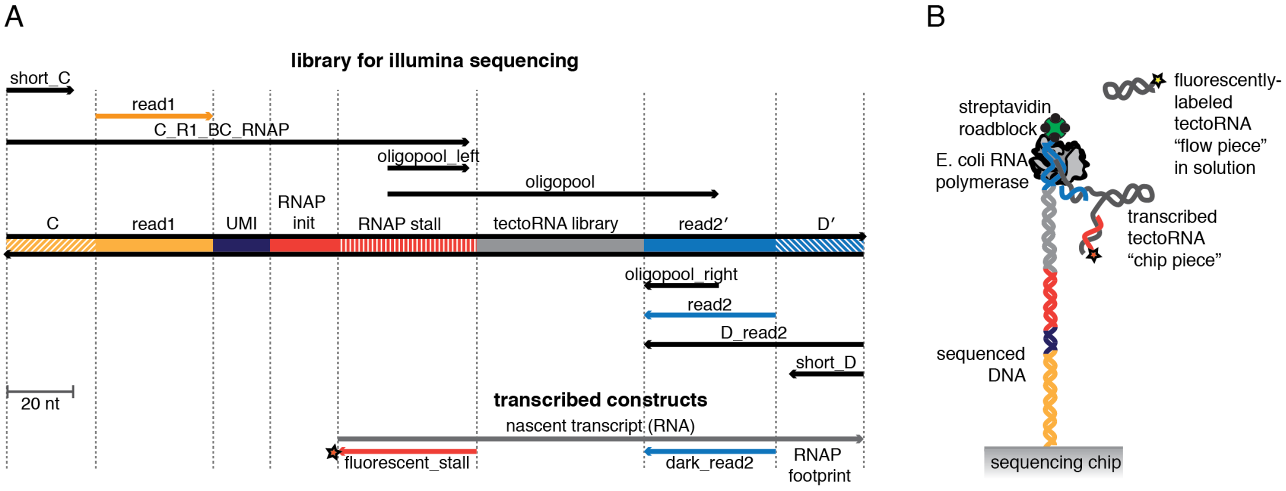


Figure S1. Library construction and experimental setup.

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.

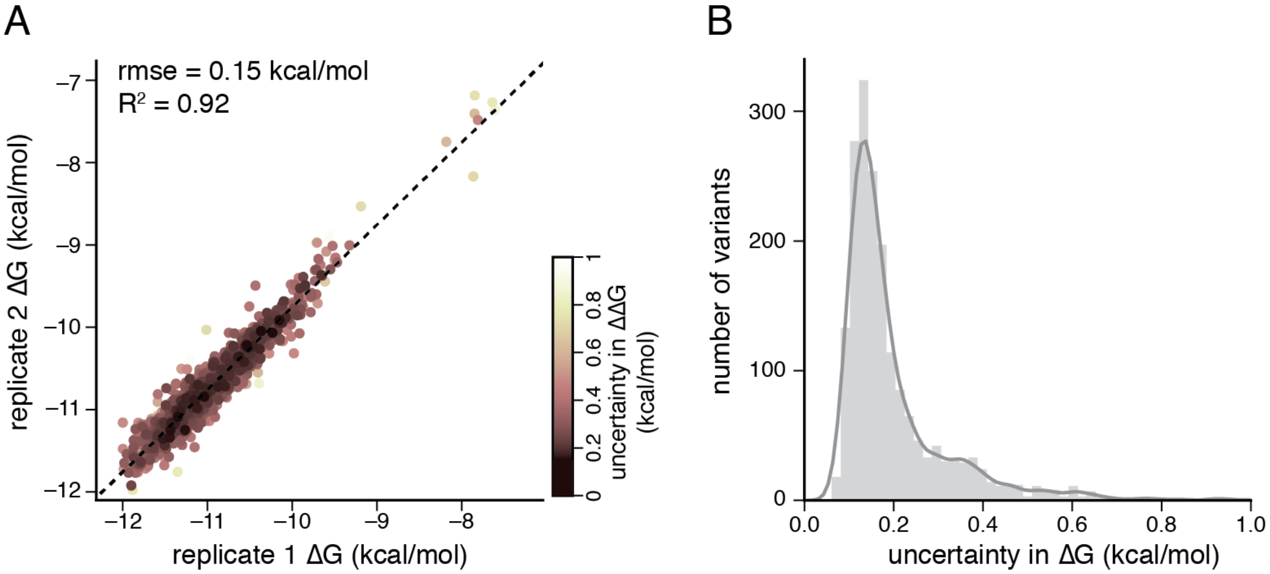


Figure S2. Measured ΔG values are reproducible and precise.

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).

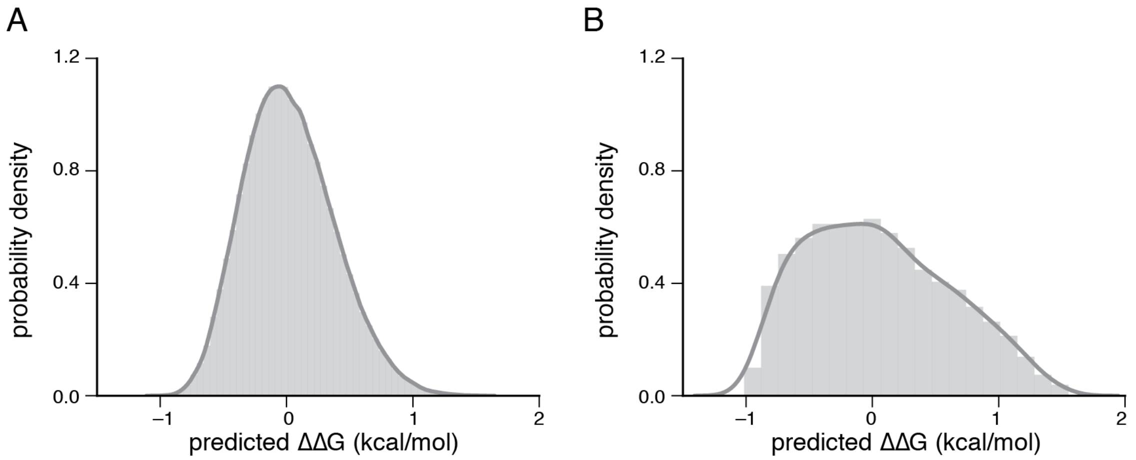


Figure S3. Predicted effect of helix sequence on tectoRNA binding free energy.

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.

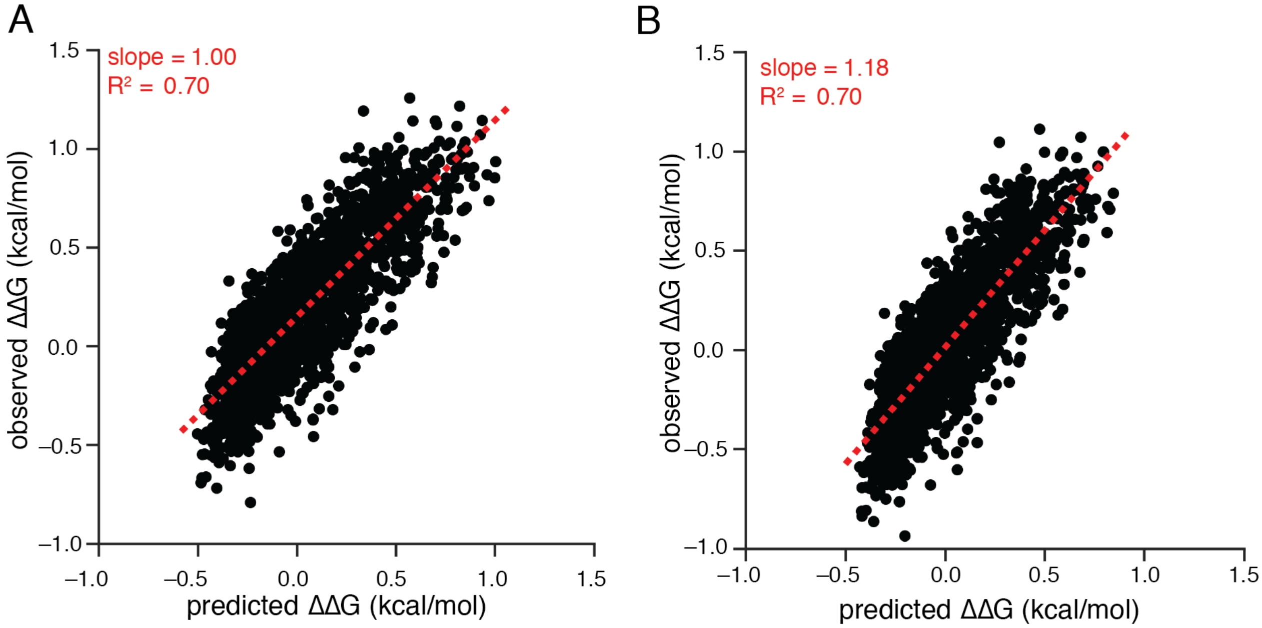


Figure S4. Simulation parameter sweeps

(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.

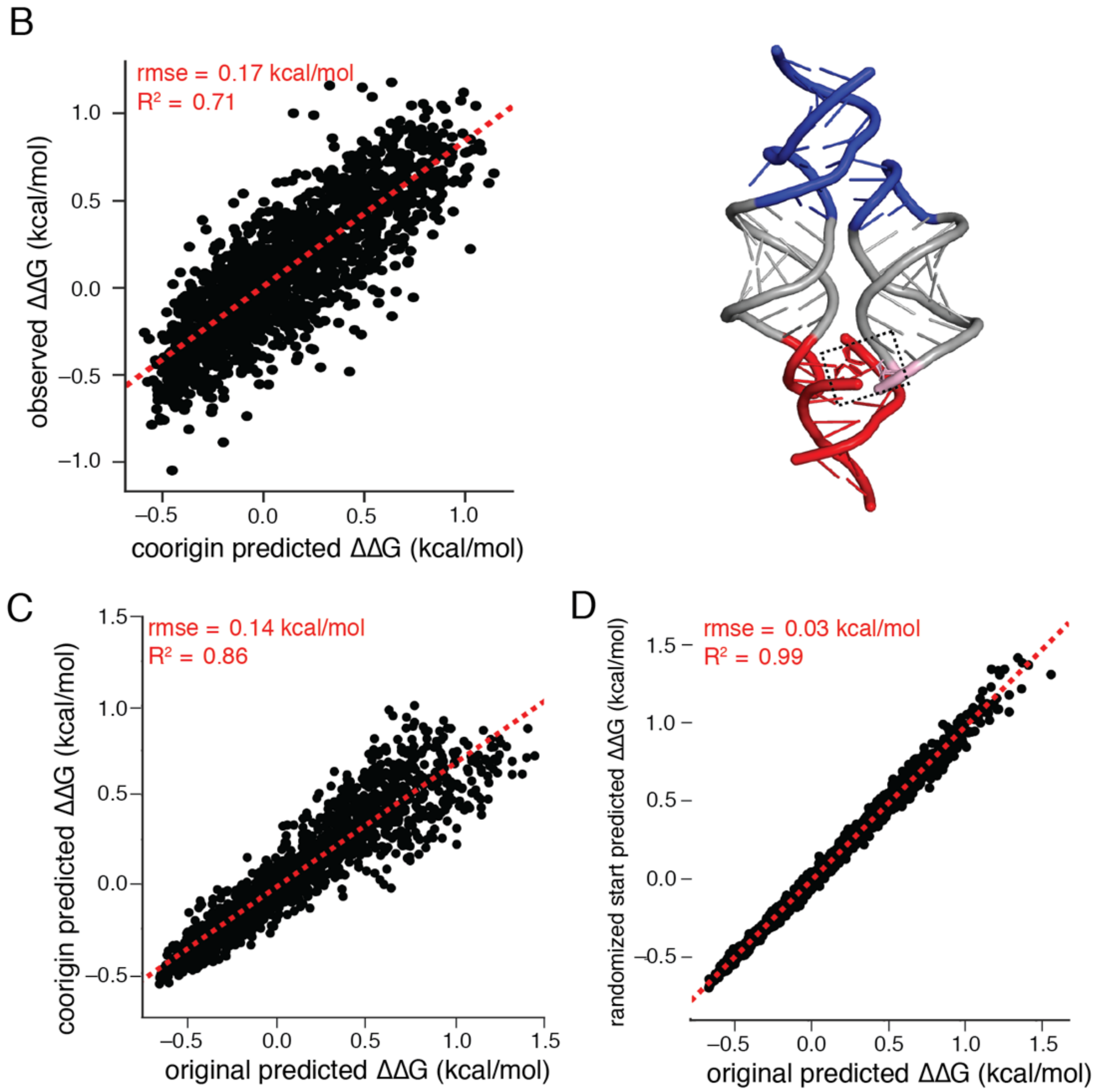


Figure S5. Comparison of simulation topology and starting conformation

(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.

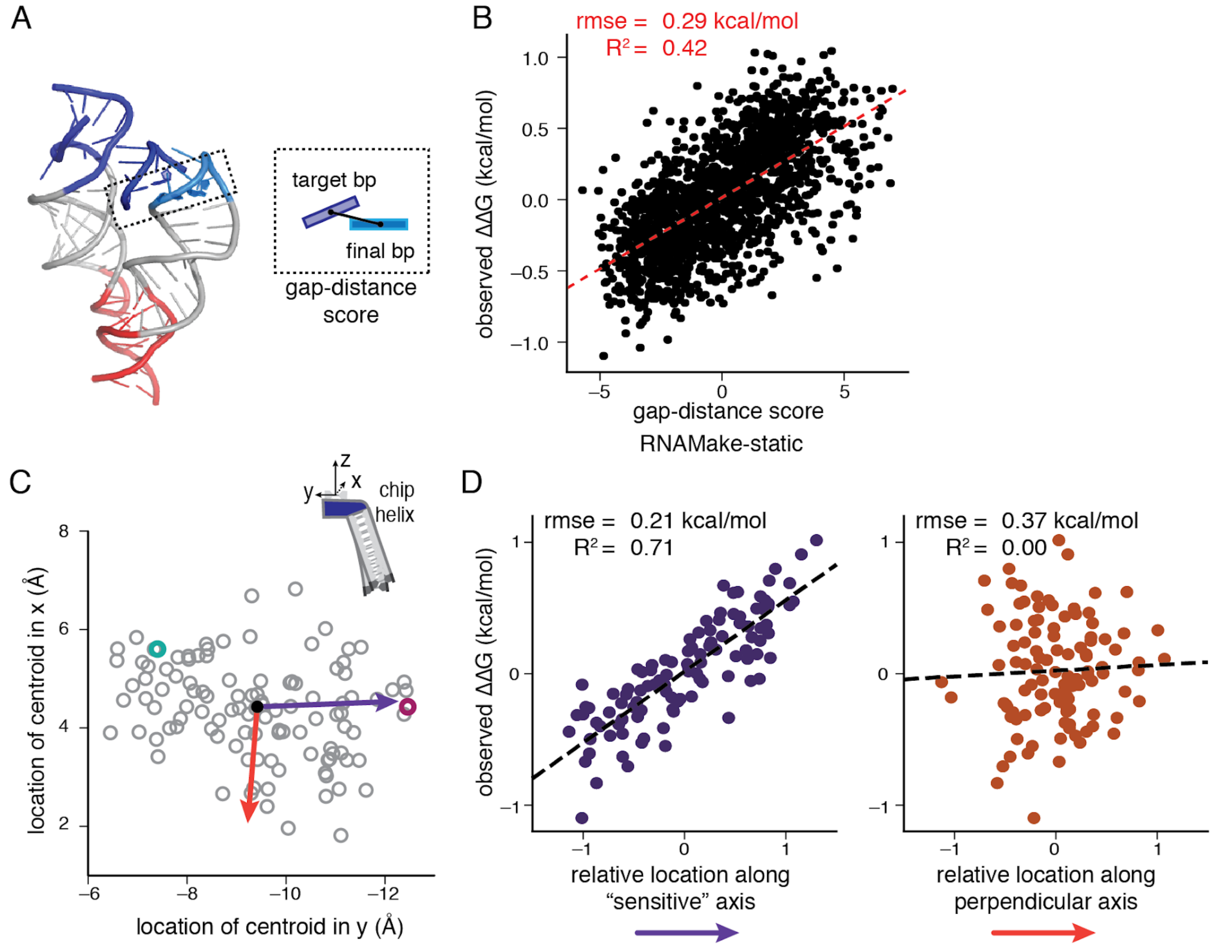


Figure S6. Non-ensemble models for tectoRNA affinity do not consistently predict observed effects.

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.

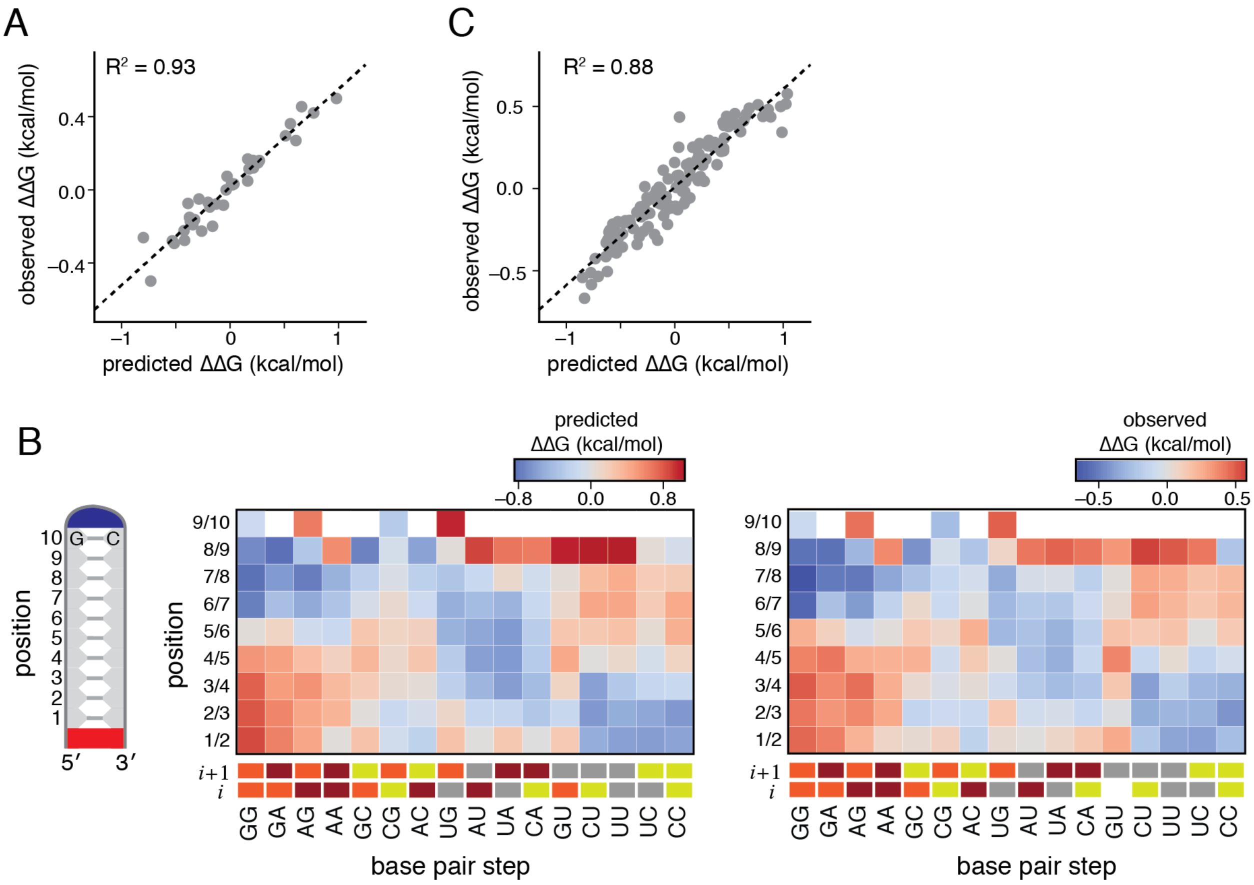


Figure S7. Predicted effect of each base pair step at each position within the tectoRNA chip piece helix.

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)).

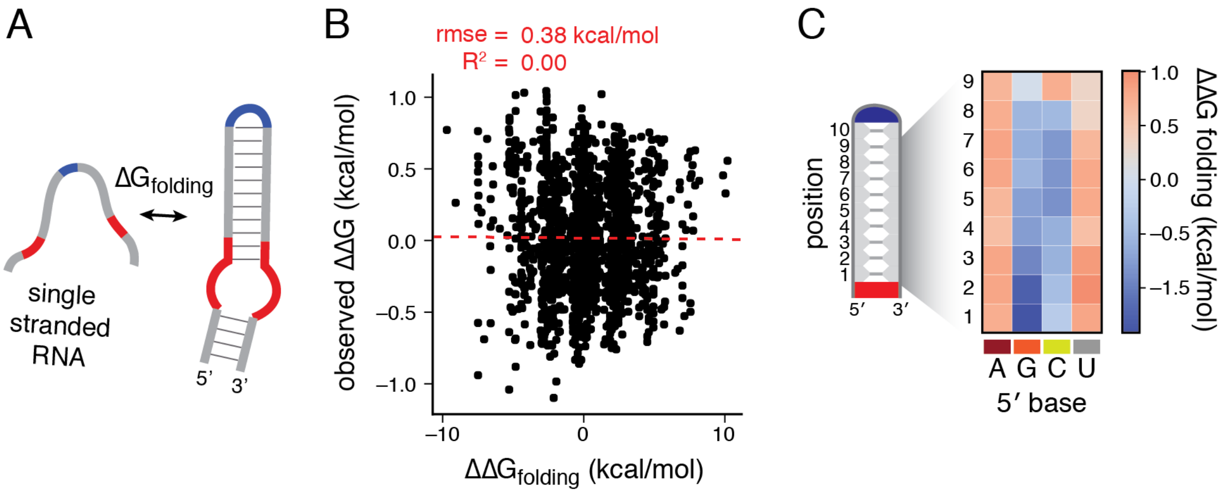


Figure S8. Observed changes in tectoRNA affinity are not dependent on predicted changes in free energy of secondary structure formation.

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.

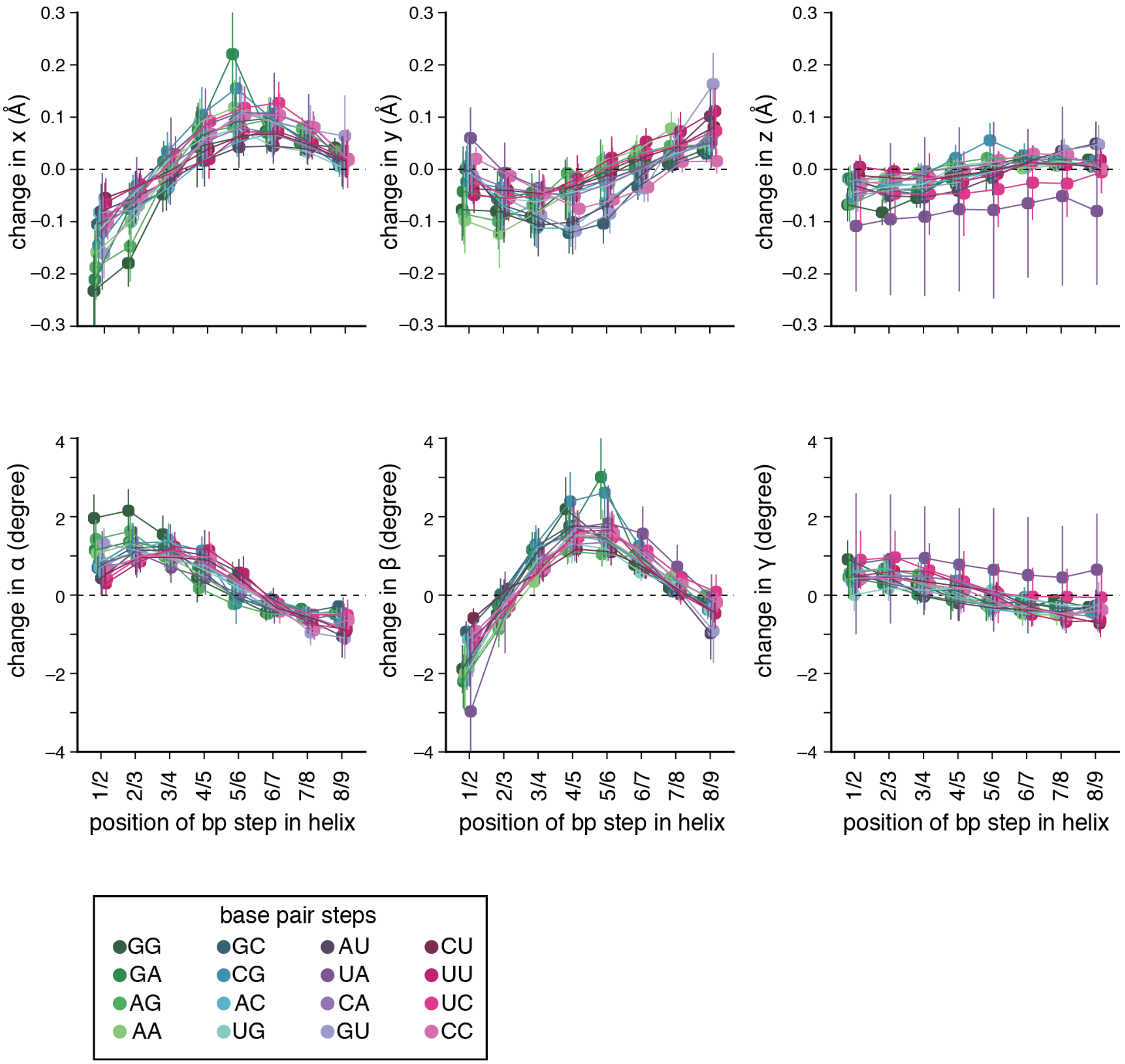


Figure S9. Structural preferences of conformational states across positions.

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´.

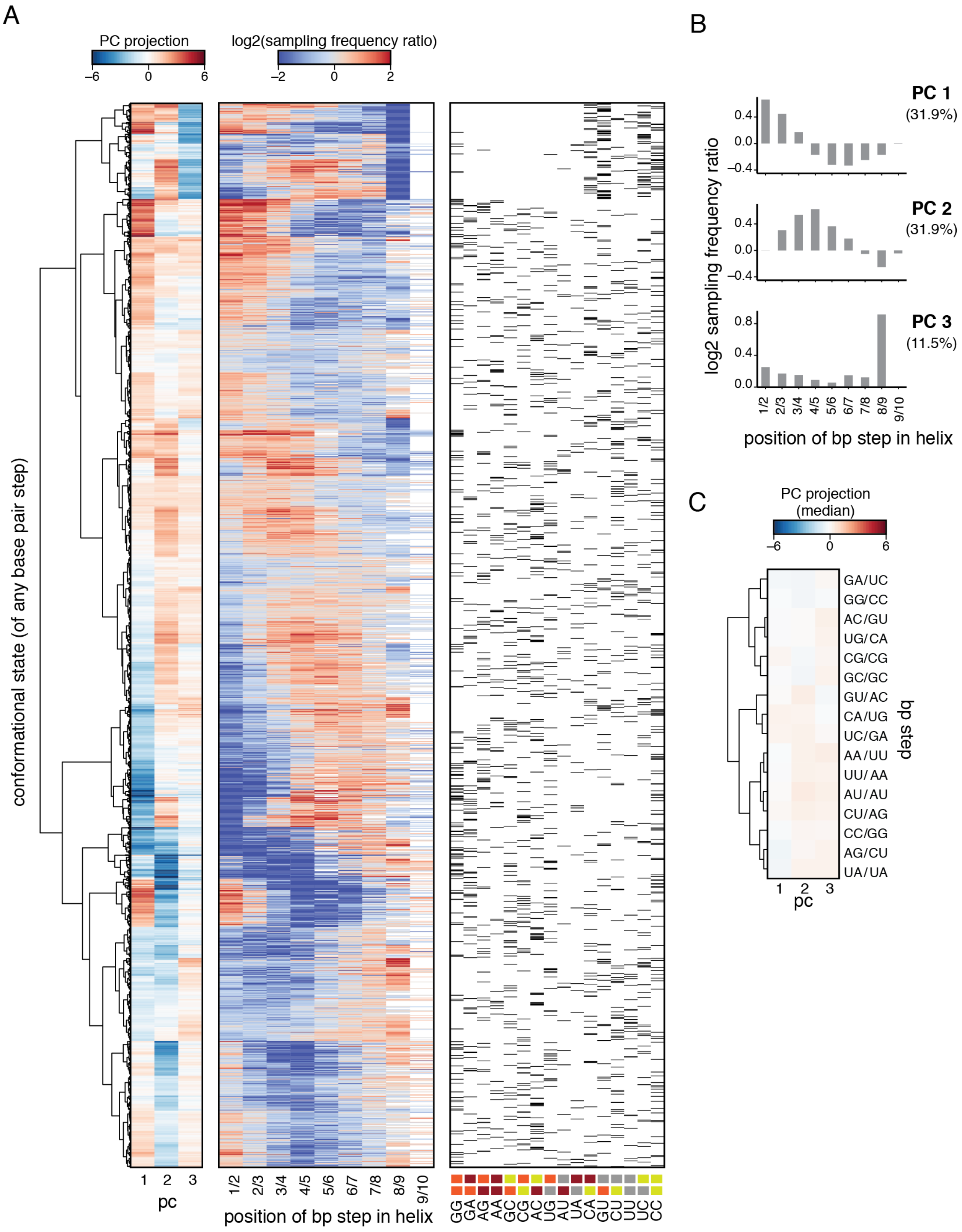

A
PC projection
−6
0
6
log2(sampling frequency ratio)
−2
0
2
conformational state (of any base pair step)
1
2
3
pc
1/2
2/3
3/4
4/5
5/6
6/7
7/8
8/9
9/10
position of bp step in helix
GG
GA
AG
AA
GC
CG
AC
UG
AU
UA
CA
GU
CU
UU
UC
CC
B
log2 sampling frequency ratio
PC 1
(31.9%)
0.4
0.0
−0.4
PC 2
(31.9%)
0.4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

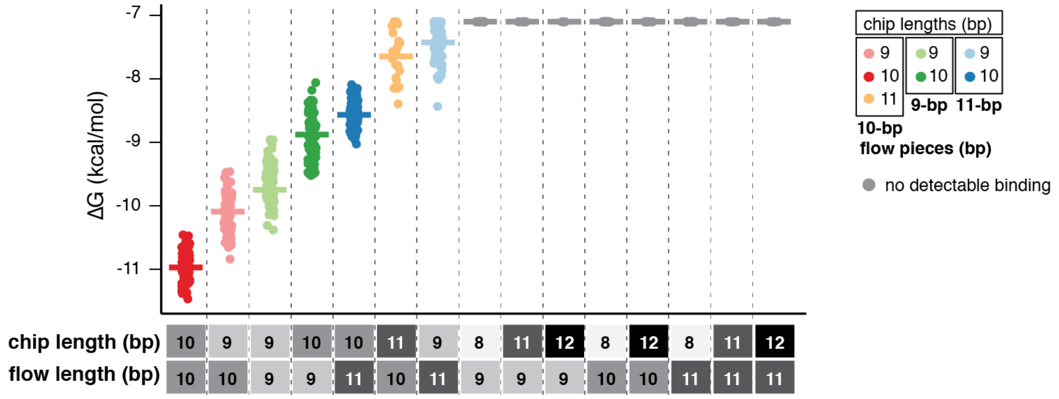


Figure S11. Measured binding affinity (ΔG) of different length-paired complexes.

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.

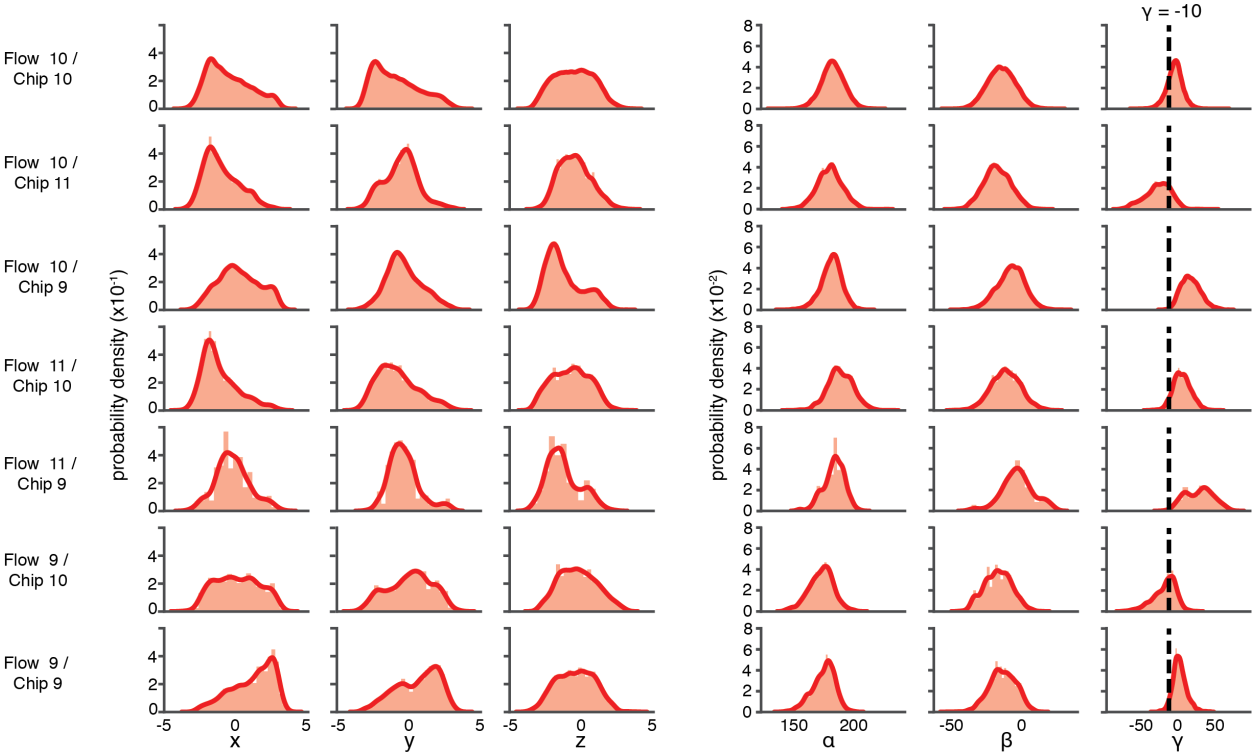


Figure S12. Distribution of six-dimensional values of bound conformation of for each tectoRNA length topology.

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.

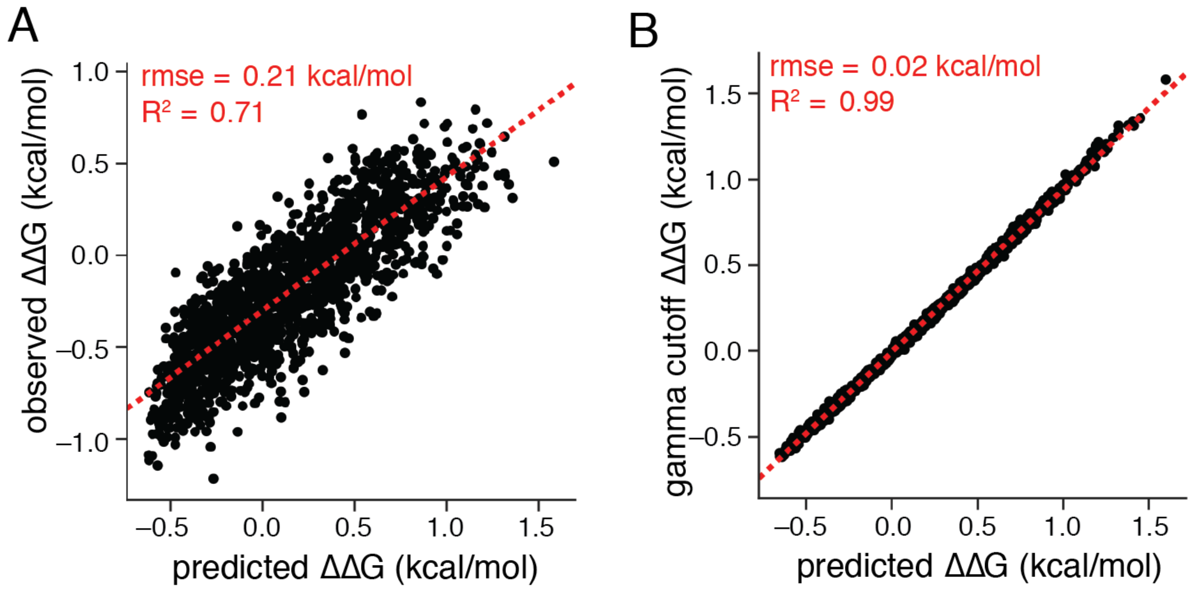


Figure S13. Gamma corrected predictions of sequence-dependent set

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

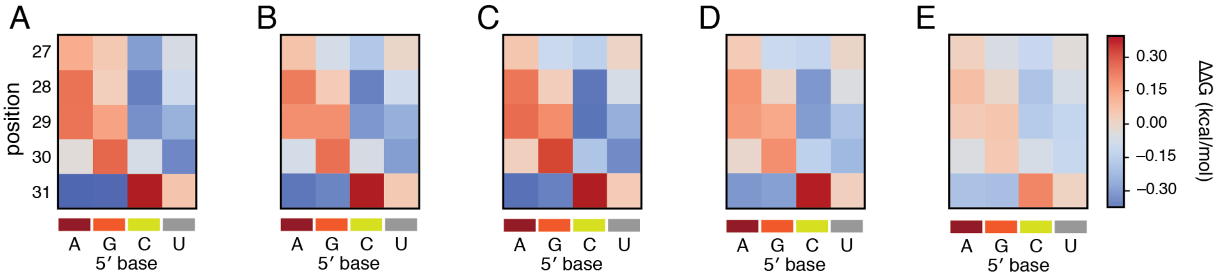


Figure S14. Prediction of helical sequence preference of anticodon helix for aminoacyl-tRNA•EF-Tu accommodation during ribosome codon recognition.

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.