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

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.

Published online August 2, 2019.

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

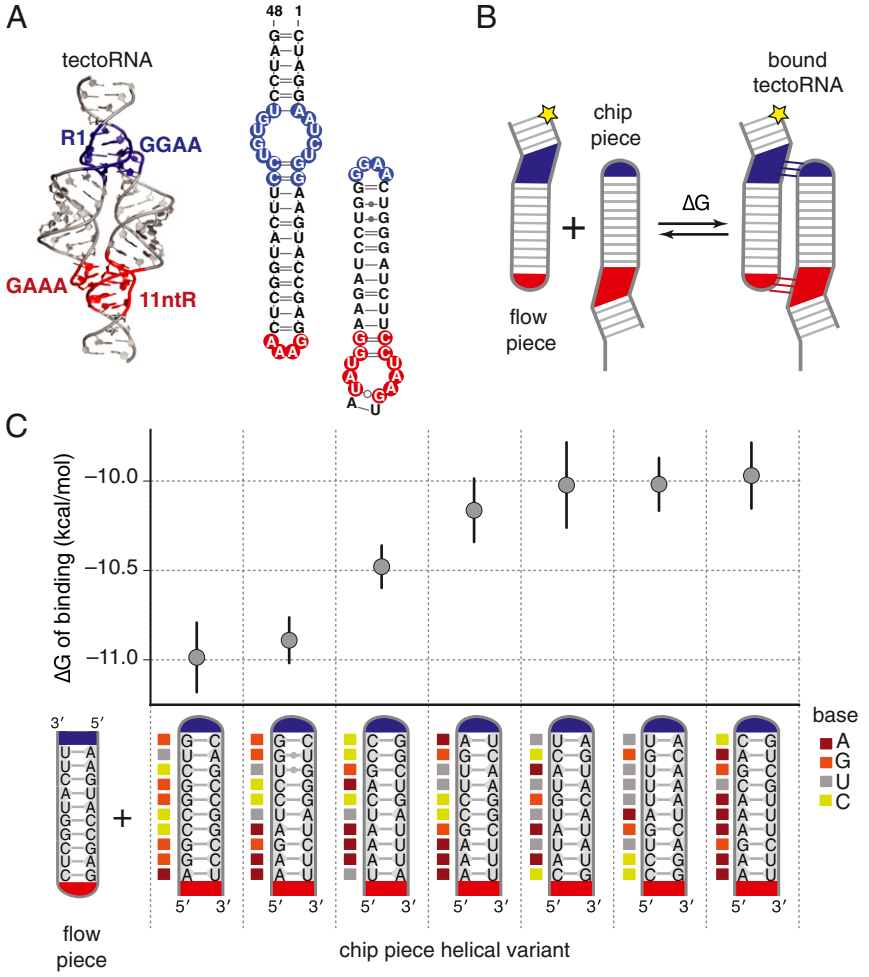

Fig. 1. Free energy of tectoRNA binding depends on helix sequence. (*A*) Structure of tectoRNA homodimer [Protein Data Bank (PDB): 2ADT] with 2 tertiary contacts (GAAA-11nt). One of these tertiary contacts is replaced (GGAA-R1; blue) to convert the complex to the heterodimer used in this study (32). 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 $\Delta G = 0.16$ 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

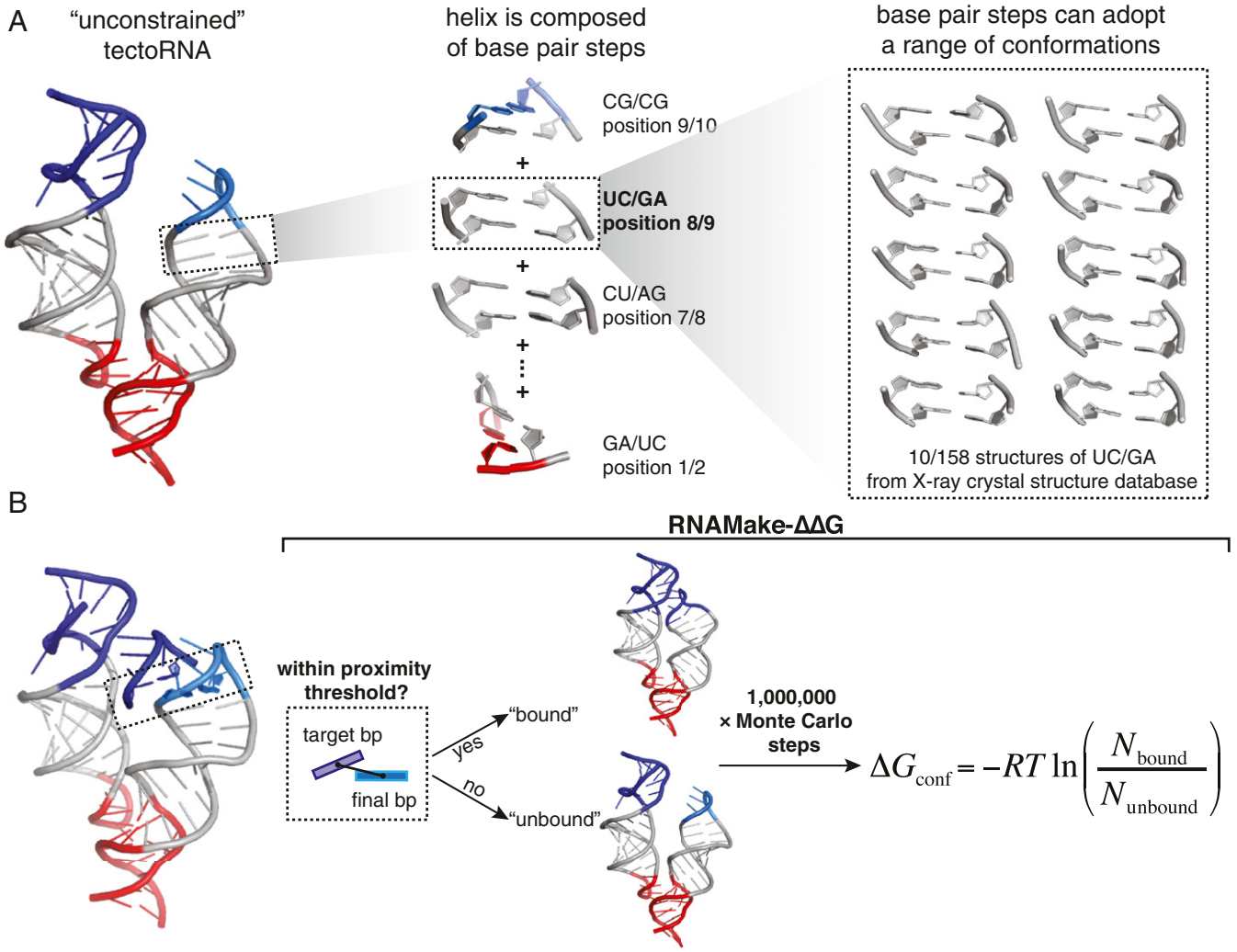

Fig. 2. Ensemble model for RNA helices allows prediction of tectoRNA assembly energetics. (*A*, *Left*) The modeled structure of the unconstrained tectoRNA (i.e., with one contact formed) is shown. The global structure was assembled from the structures of its constituent elements, including the base-pair 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,

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

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 ($\Delta\Delta G_{binding}$) to differences in this conformational alignment term,

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

where $\Delta G_{conf,1}$ and $\Delta G_{conf,2}$ 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),

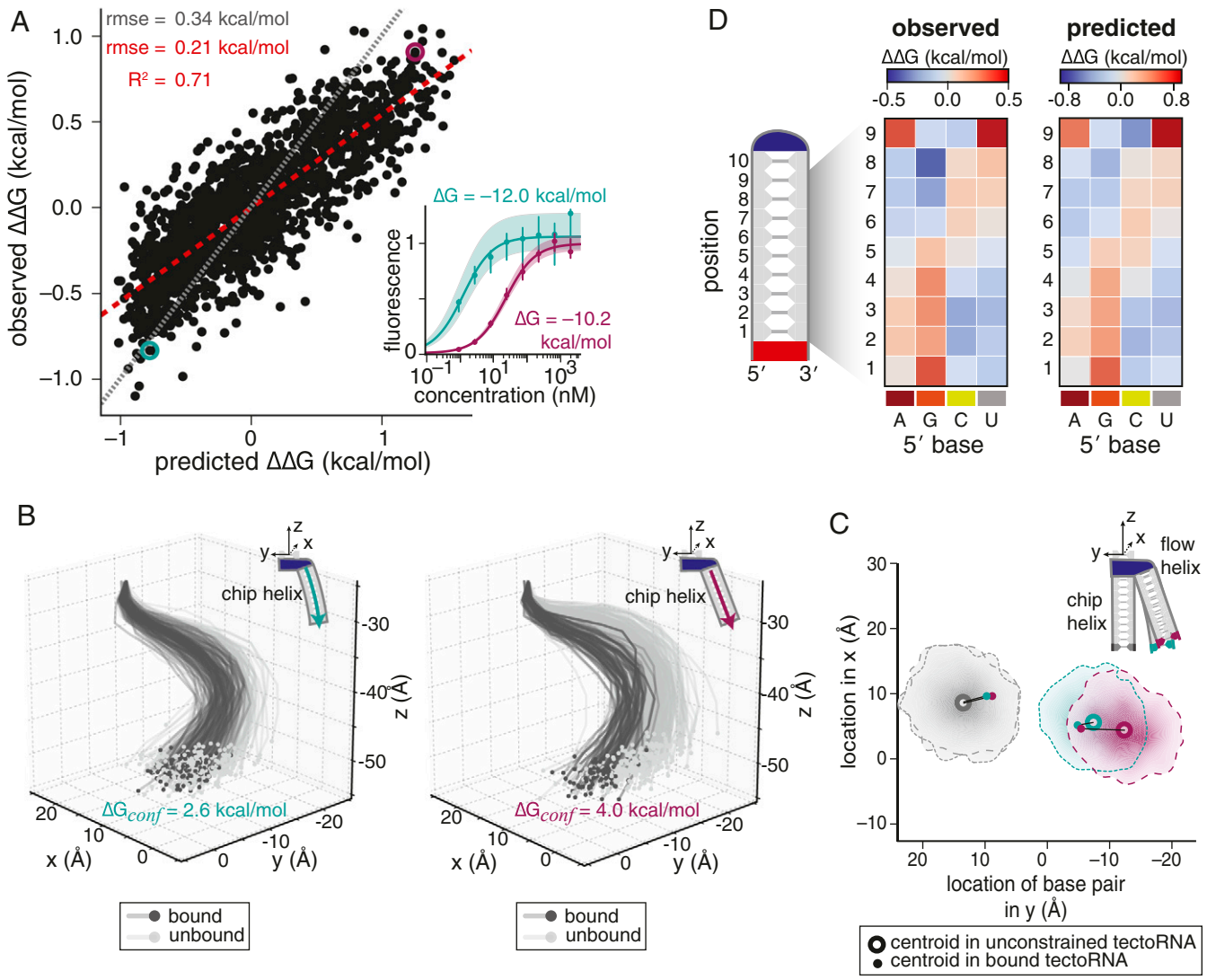

Fig. 3. RNAMake-ΔΔG accounts for changes in tectoRNA affinity in a blind prediction challenge. (*A*) Blind predictions generated with the RNAMake-ΔΔG model agree well with observed values of tectoRNA binding ΔG for 1,536 chip-piece variants (R^2 = 0.71). Each set of ΔG values is compared with their respective 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

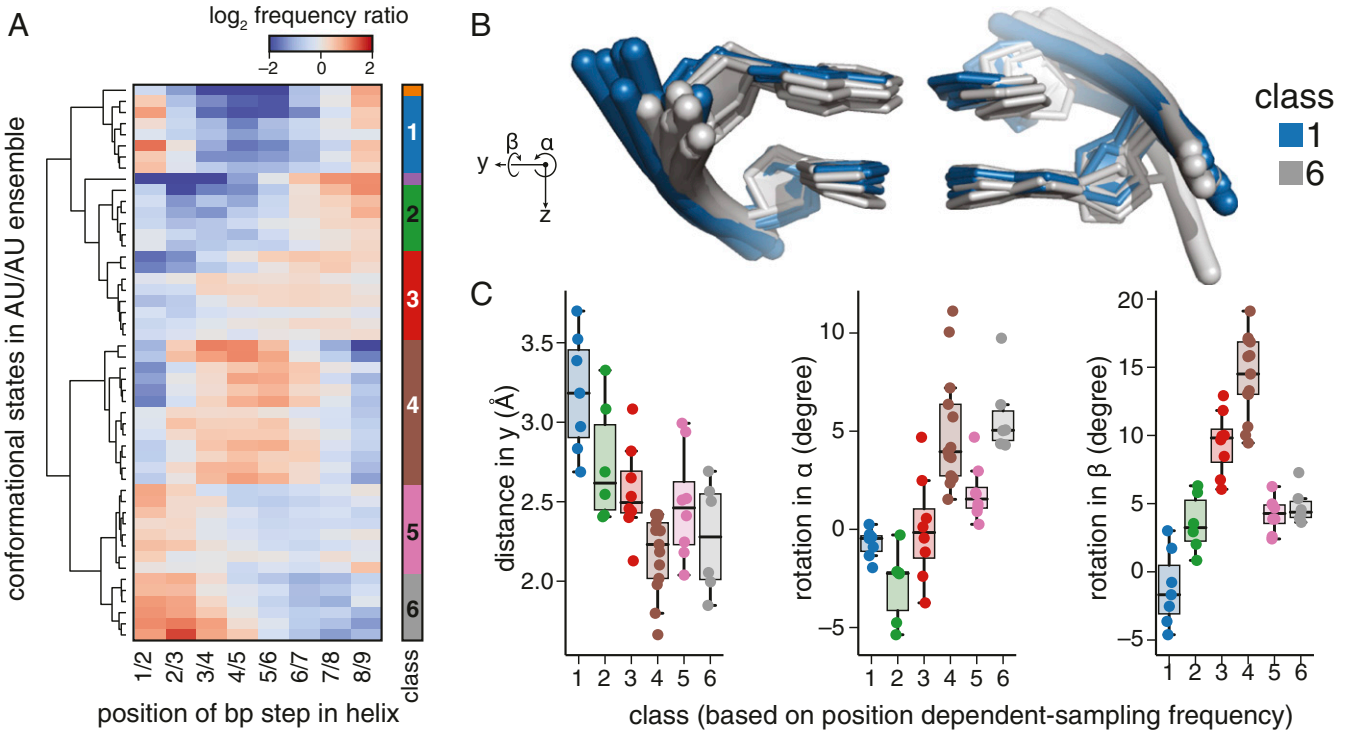

Fig. 4. Base-pair conformations differ by position within the helix. (*A*) Change in sampling frequency of conformational states in the AU/AU ensemble in the bound versus the partially bound. (*B*) Example structures of base-pair step conformations that are enriched and depleted at 2 positions. (*C*) 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.

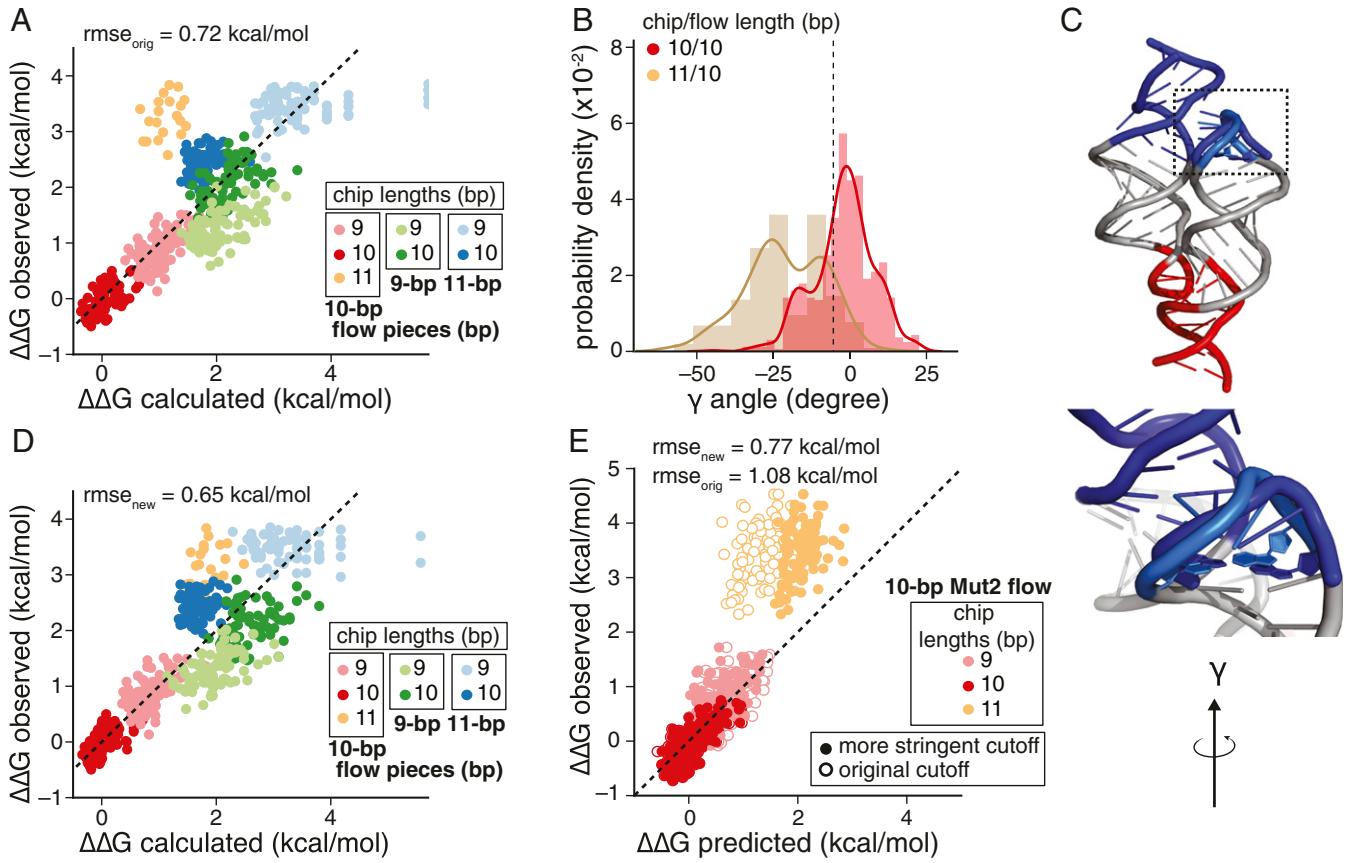

Fig. 5. Increased prediction accuracy of different length pairs with refinement of the bound state cutoff. (*A*) Observed versus calculated affinities for chip- and flow-piece variants with altered lengths. The colors indicate the length of the flow- and chip-piece helices. (*B*) Distribution of the 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 γ < −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 γ > −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

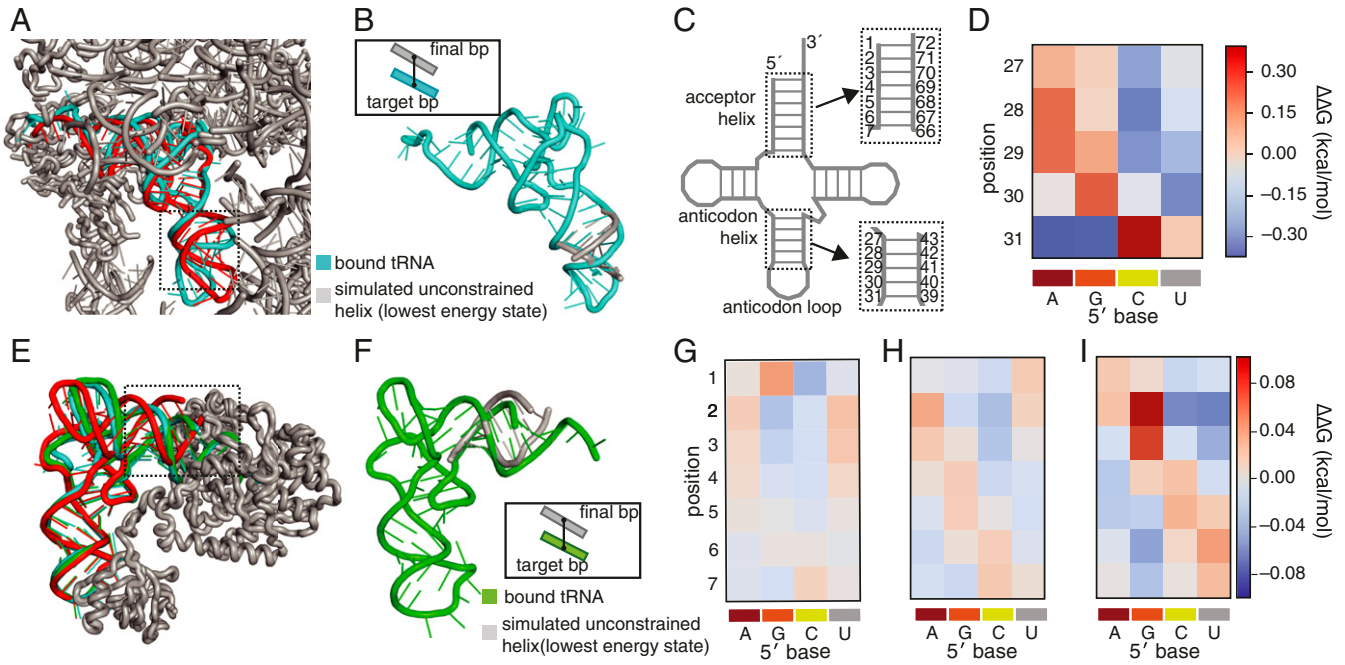

Fig. 6. Prediction of RNA double helix distortions that occur during ribosomal A-site accommodation and amino acid charging. (*A*) When complexed with EF-Tu and being loaded into the A-site of the ribosome (the A/T state), *Thermus thermophilus* $tRNA^{Thr}$ appears bent (cyan, PDB: 4V5G) compared with *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).