RNA conformational propensities determine cellular activity

Megan L. Ken[1], **Rohit Roy**[2], **Ainan Geng**[1], **Laura R. Ganser**[3], **Akanksha Manghrani**[1], **Bryan R. Cullen**[4], **Ursula Schulze-Gahmen**[5,*], **Daniel Herschlag**[6,*], **Hashim M. Al-Hashimi**[7,*]

[1]Department of Biochemistry, Duke University School of Medicine, Durham, NC 27710, USA

[2]Center for Genomic and Computational Biology, Duke University School of Medicine, Durham, NC 27710, USA

[3]Department of Biophysics, Johns Hopkins University, Baltimore, MD 21218, USA

[4]Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, NC 27710, USA

[5]Gladstone Institute of Virology, San Francisco, CA 94158, USA

[6]Department of Biochemistry, Stanford University, Stanford, CA 94305, USA

[7]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York 10032, USA

Abstract

Cellular processes are the product of interactions between biomolecules, which associate to form biologically active complexes [1]. These interactions are mediated by intermolecular contacts, which if disrupted, lead to alterations in cell physiology. Nevertheless, the formation of intermolecular contacts nearly universally requires changes in the conformations of the interacting biomolecules. As a result, binding affinity and cellular activity crucially depend not only on the strength of the contacts, but also on the inherent propensities to form binding-competent conformational states[2,3]. Thus, conformational penalties are ubiquitous in biology and must be known in order to quantitatively model binding energetics for protein and nucleic acid interactions[4,5]. However, conceptual and technological limitations have hindered our ability to dissect and quantitatively measure how conformational propensities impact cellular activity. Here, we systematically altered

*Corresponding author. Correspondence and requests for material should be addressed to : Hashim M. Al-Hashimi, ha2639@cumc.columbia.edu; Daniel Herschlag, herschla@stanford.edu; and Ursula Schulze-Gahmen, ursula.schulzegahmen@gladstone.ucsf.edu.

and determined the propensities for forming the protein-bound conformation of HIV-1 TAR RNA. These propensities quantitatively predicted the binding affinities of TAR to the RNA-binding region of the Tat protein and predicted the extent of HIV-1 Tat-dependent transactivation in cells. Our results establish the role of ensemble-based conformational propensities in cellular activity and reveal an example of a cellular process driven by an exceptionally rare and short-lived RNA conformational state.

There is a growing database of nucleic acid and protein structures[6]. However, conformational propensities can only be deduced from conformational ensembles specifying the probability of forming bound states in the absence of binding partners[1]. These bound states are lowly-populated and short-lived, falling outside detection of standard biophysical methods[7,8]. Moreover, we currently lack approaches for determining conformational ensembles within cells, and conformational propensities may differ within the physiologically relevant cellular environment relative to *in vitro* conditions used to measure them[9,10]. Even under ideal *in vitro* conditions, determining ensembles has been time-consuming and requires sophisticated approaches for resolving minor conformational states[11], making it difficult to systematically examine how changes in conformational propensities impact cellular activity. Compounding these limitations are unique challenges in obtaining quantitative and accurate measurements of interactions and their consequences inside cells.

Modeling conformational propensities

We have taken on this set of challenges, spanning biophysics to cellular function, by quantitatively and systematically examining the role of the transactivation response element (TAR) RNA conformational propensities in Tat-dependent transactivation of the human immunodeficiency virus 1 (HIV–1) genome (Fig. 1a). TAR is a highly conserved and structured[12] RNA element located at the 5' terminal end of the retroviral genome. Transactivation is a multi-step cellular process initiated by binding of the viral protein Tat and the human super elongation complex (SEC) to the active conformation of TAR (Fig. 1a)[13–17]. Productive Tat binding to TAR and cellular transactivation depend on coaxial stacking of the two TAR helices[18] and the formation of a U23•A27-U38 base triple (Fig. 1b,d), in which bulge residue U23 forms a reverse Hoogsteen base pair with A27 in the upper stem[19–21]. The Tat protein forms several critical contacts with this base-triple[22,23], which require TAR to be in the stacked conformation. Two arginine residues (R49 and R52) form cation-pi interactions involving their guanidinium groups and the aromatic rings of TAR bases A22 and U23 (Fig. 1c)[20,24]. R52 is flanked on both sides by A22 and U23, forming an arginine sandwich motif, while A49 has only one cation-pi interaction with U23[22,25] (Fig. 1c). R52 also forms an arginine fork motif involving hydrogen bonds between their guanidinium groups and the G26 base, between bridging and non-bridging phosphate groups[24,26].

Prior studies[7] showed that in the absence of Tat, the free wild-type TAR (*wt*) ensemble very rarely forms the base-triple bound conformation (Fig. 1b), which was estimated to be energetically disfavored by a difference in Gibbs free energy (ΔG) of greater than 7 kcal/

mol. To quantify how the propensity to form the base-triple impacts Tat-dependent cellular transactivation, we decomposed the energetics of protein-TAR binding (ΔG_{prot}) into two independent contributions (Fig. 1a): the conformational propensity to form the base-triple TAR bound state (ΔG_{conf}) and favorable binding of the protein to this conformational state (ΔG_{bind}) such that

$$\Delta G_{prot} = \Delta G_{conf} + \Delta G_{bind}.$$

We further decomposed G_{conf} into two contributions (Fig. 1b): the energetic cost of redistributing the ensemble to form the stacked state ($\Delta G_{penalty,\,stack}$) and formation of the base triple (ΔG_{triple}) in the stacked state, again assuming independence,

$$\Delta G_{conf} = \Delta G_{penalty,\,stack} + \Delta G_{triple}.$$

We then examined how mutations that are predicted to alter $\Delta G_{penalty,\,stack}$ and/or ΔG_{triple} without affecting contacts between TAR and Tat:SEC, and therefore ΔG_{bind}, impact cellular transactivation. This approach directly links RNA conformational preferences to cellular function (Fig 1d). Assuming the TAR variants predominantly bind Tat:SEC in a base-triple conformation, the difference in the measured protein binding energetics, ΔG_{prot}, between the reference *wt* TAR and a variant *j* should only depend on the difference in the conformational propensities ($\Delta\Delta G_{conf} = \Delta\Delta G_{penalty,\,stack} + \Delta\Delta G_{triple}$) as follows:

$$\Delta G_{prot}(wt) = \Delta G_{penalty,\,stack}(wt) + \Delta G_{triple}(wt) + \Delta G_{bind}(wt)$$

$$\Delta G_{prot}(j) = \Delta G_{penalty,\,stack}(j) + \Delta G_{triple}(j) + \Delta G_{bind}(wt)$$

$$\Delta\Delta G_{prot}(j - wt) = \Delta\Delta G_{penalty,\,stack}(j - wt) + \Delta\Delta G_{triple}(j - wt).$$

Our model assumes that Tat binds the TAR variants in a similar conformation, and the differences in the penalty to form the binding-competent Tat-ARM conformation is similar across the TAR variants. This predictive model allowed us to dissect and learn more about the contribution of conformational propensities to protein binding and cellular function.

$\Delta G_{penalty,\,stack}$ was incrementally increased by increasing the TAR bulge length[18] from zero to seven nucleotides ($U_0 - U_7$) (Fig 1d). Increasing the bulge length leads to increased sampling of kinked conformational states and lowered stacking propensities, most likely due to unfavorable conformational entropy[27], and therefore disfavors productive binding to the base-triple bound state. The G_{triple} penalty was increased for each bulge variant ($U_0 - U_7$) by destabilizing the U23•A27 base-pair in two ways, by replacing A27-U38 with U27-A38 and by replacing A27 with a deaza-N7-modified adenosine (Fig 1d). Due to topological constraints, the U_1 variant is also expected to increase ΔG_{triple} relative to *wt* (i.e., render

it less favorable) since base triple formation typically requires an additional spacer bulge residue[28,29].

Assuming the simplest model of ΔG_{conf} (Fig. 1a,b), once a TAR variant becomes stacked, the energy required to form the base triple does not vary with bulge length and is a constant given by $\Delta G_{triple}(wt)$. Our model assumes that all TAR variants are predominantly bound by Tat:SEC in a base-triple conformation (binding to kinked conformations has been detected to form as a minor state with smaller Tat fragments, a point we return to)[7]. Thus, for the base-triple forming variants, the difference between the protein-binding energetics ΔG_{prot} comparing the reference wt and a variant j is predicted to be equal to the corresponding difference in stacking propensities, $\Delta\Delta G_{penalty, stack}(j - wt)$, which can be measured experimentally[18],

$$\Delta\Delta G_{prot}(j - wt) = \Delta\Delta G_{penalty, stack}(j - wt) + \Delta\Delta G_{triple}(wt - wt)$$

$$\Delta\Delta G_{prot}(j - wt) = 0 + \Delta\Delta G_{penalty, stack}(j - wt).$$

As the model above predicts independent energetic effects from changes in stacking and base triple formation, the base-triple destabilizing variants (denoted *) are predicted to increase ΔG_{conf} relative to their unmodified counterparts by a constant amount (c_{triple}) for each stacking variant (j) where c_{triple} is the amount that the base triple is destabilized. Thus, the following relationships are predicted:

$$\Delta\Delta G_{prot}(j* - j) = \Delta\Delta G_{penalty, stack}(j* - j) + \Delta\Delta G_{triple}(j* - j)$$

$$\Delta\Delta G_{prot}(j* - j) = 0 + \Delta\Delta G_{triple}(j* - j),$$

and

$$\Delta\Delta G_{prot}(j* - j) = \Delta\Delta G_{triple}(j* - j) = c_{triple}.$$

Our model therefore makes strong quantitative predictions: (1) for the bulge variants, G_{prot} will vary linearly with $\Delta\Delta G_{penalty, stack}$ with a slope of 1 and intercept of 0; (2) for the base-triple destabilized variants ($j*$), the binding energy of each variant will be weakened relative to its unmodified counterpart (j) by a constant amount, c_{triple}.

$G_{penalty,stack}$ predicts TAR-Tat binding

Before examining how changing ΔG_{conf} impacted Tat-dependent cellular transactivation, we first tested a more direct prediction of the model on *in vitro* binding energetics (ΔG_{pep}) of TAR to a 12-amino acid arginine-rich motif (ARM) peptide containing the RNA-binding region of Tat (Tat-ARM peptide) (Fig. 1c). This peptide captures the full TAR-bulge-binding region of Tat, which covers all sites of TAR being mutated, and we assume that any other

energetic contributions to binding from the full-length protein would be similar across the different TAR variants. We shall see that our findings support this assumption (vide infra). The most informative approach to measure RNA stacking would be through determination of the full atomic-resolution experimental ensemble. The measured fractional population of the stacked conformation (p_{stack}) is then:

$$p_{\text{stack}} = [\text{stack}]/\{[\text{kink}] + [\text{stack}] + [\text{triple}]\}\,.$$

The free energy cost to redistribute the unbound TAR ensemble to the stacked state is given by (see Methods)[1,30]:

$$\Delta G_{\text{penalty, stack}} = -\text{RTln}\, p_{\text{stack}}\,.$$

However, determining an ensemble typically takes several years per variant[11]. We therefore sought a simpler, but still quantitative approach to determine $\Delta\Delta G_{\text{penalty, stack}}$ and applied it to our TAR library. We developed such an approach using the NMR chemical shifts of U23-C6 and A22-C8[18] to measure the fractional population of the stacked conformational state $p_{\text{stack}} \sim [\text{stack}]/\{[\text{kink}] + [\text{stack}]\}$ (see methods) (Fig 2a). This chemical shift perturbation (CSP) approach allowed us to rapidly measure p_{stack} and $\Delta G_{\text{penalty, stack}}$ without having to determine complete ensembles.

We benchmarked the CSP approach by comparing measured $\Delta\Delta G_{\text{penalty, stack}}$ values with counterparts obtained from full atomic-resolution ensemble models previously developed using NMR-aided Fragment Assembly of RNA with Full-Atom Refinement (FARFAR-NMR)[11,31]. FARFAR-NMR determines the RNA conformational ensemble by constraining empirical structure models[31] with experimental data from NMR spectroscopy[11]. The measured p_{stack} values were converted into $\Delta G_{\text{penalty, stack}}$ for *wt* and U_7, the two ensembles with stacked populations in the range of detection of FARFAR-NMR. While the CSP approach systemically (and slightly) underestimated the stacked population determined using FARFAR-NMR, the differences in stacking penalty $\Delta\Delta G_{\text{penalty, stack}}$ for U_7 relative to the reference *wt* were < 0.2 kcal/mol (Fig. 2b, Supplementary Table 1). Thus, by comparing the differences in $\Delta G_{\text{penalty, stack}}$ we reduce systematic errors between the two techniques.

Using the CSP approach, we measured p_{stack} for all twenty-seven TAR variants used in this study (Extended Data Fig. 1, Supplementary Table 2). The measured p_{stack} values across base-triple competent variants varied by 5-fold and corresponds to differences in stacking propensities of $\Delta\Delta G_{\text{penalty, stack}} = 0.9$ kcal/mol (Fig. 2c). As expected from entropic considerations and prior results[18], increasing the bulge length decreased the stacking propensity (Fig. 2c). Unexpectedly, U_7 did not follow the trend, showing more stacking than U_6. Additional NMR data provided evidence that U_7 forms a short uridine-rich helix, which could extrude the bulge and increase the stacking propensity (Extended Data Fig. 2, Supplementary Discussion 1). This observation also highlights other contributions to stacking propensities including the effects of solvation and hydrogen bonding. In agreement with our additivity model, for a given bulge variant, introducing base-triple

destabilizing mutations A27U-U38A or A27-deaza-N7 minimally impacted the stacking penalty (differences in $\Delta\Delta G_{penalty,stack}$ <0.3 kcal/mol) for all bulge variants (r = 0.99 and p < 0.0001 for comparisons of each with the *wt* base triple constructs; Fig. 2c and Extended Data Fig. 1b).

We determined ΔG_{pep} by measuring the binding affinities of the TAR variants to the Tat-ARM peptide using a FRET-based *in vitro* binding assay (Extended Data Fig. 3, Supplementary Table 3)[32]. The range of peptide binding propensities ($\Delta\Delta G_{pep}$) for base-triple competent variants was ~1.2 kcal/mol (Fig. 2d), similar to the range measured for $\Delta\Delta G_{penalty,stack}$ of 0.9 kcal/mol.

As predicted by our model, the variation in the binding energetics $\Delta\Delta G_{pep}$ across these variants was in excellent agreement with the differences in stacking propensities, $\Delta\Delta G_{penalty,stack}$ (Fig. 2c–d). As expected, the binding affinities decreased with increasing bulge length up to U_6, and U_7 showed a higher binding affinity relative to U_6. We quantitatively compared two aspects of the propensities $\Delta\Delta G_{pep}$ and $\Delta\Delta G_{penalty,stack}$. We determined how strongly correlated $\Delta\Delta G_{pep}$ and $\Delta\Delta G_{penalty,stack}$ are using the Pearson correlation (r) and how well each agrees with our model, quantified using both R^2 and root mean squared error (RMSE).

A strong linear correlation was observed between $\Delta\Delta G_{pep}$ and $\Delta\Delta G_{penalty,stack}$ for base-triple forming mutants (Pearson correlation r = 0.96; p = 0.0006) (Fig. 2e). U_1 is excluded as its ΔG_{triple} is significantly increased relative to *wt* and U_{2-7} due to topological constraints[28,29]. The predictions from our model, without any adjustable parameters, agree with the experimental data with RMSE = 0.24 kcal/mol and $R^2 = 0.62$ (Fig. 2e, Supplementary Table 4). Additionally, the model predicted the experimental data within the 95% confidence intervals (CI) for the best-fit slope and y-intercept.

In contrast to the above results that were consistent with our model predictions, the binding energetics of the base-triple destabilized $U_2 - U_7$ variants were not weakened relative to their unmodified counterparts by a constant amount, c_{triple}, as predicted by the model (Fig. 2d). Rather, they bound Tat-ARM with similar affinities (~100 – 500 nM). In addition, the correlation between $\Delta\Delta G_{pep}$ and $\Delta\Delta G_{penalty,stack}$ vanished for these variants (Extended Data Fig. 4). As binding is nearly uniform across the base-triple variants destabilized by up to ~2.0 kcal/mol relative to *wt*, the results suggest that base triple formation is not required for Tat-ARM binding. Presumably, this different binding mode is favored because the energetic penalty to form the base triple exceeds that needed to bind in the kinked state that lacks the base triple[7] (Supplementary Discussion 2). Indeed, the distinct fluorescence intensities of the Tat-peptide complex with the base-triple destabilized TAR mutants is consistent with a different Tat-peptide binding mode (Extended Data Fig. 5a–b). However, additional studies would be required to rule out alternative models in which the Tat peptide binds to these TAR variants in a base-triple-like conformation possibly using a different set of interactions. Because our model was quantitative and predictive, we could determine when it failed for the base triple mutants and verify that a kinked state can also bind the Tat peptide[7] (Supplementary Discussion 2).

$G_{penalty,stack}$ predicts G_{cell}

Next, we examined whether changes in TAR's propensity to form the stacked, base-triple conformation (ΔG_{conf}), due to our conformational propensity-altering mutations, result in corresponding quantitative changes in Tat-dependent cellular transactivation ($\Delta Gcell$) as predicted by our simple model (Fig 1a). $\Delta Gcell$ is defined as the Tat contribution to the overall TAR-Tat:SEC binding energetics inside cells (see Methods, "Calculation of $\Delta\Delta Gcell$"). According to our model, the probability of transcriptional activation should be proportional to the probability of forming TAR in its active stacked conformation that contains the *wt* base triple (Fig 1a). Assuming that the multi-step cellular process of Tat-dependent transactivation is under thermodynamic control with respect to Tat:SEC binding to TAR (Fig 3a; Extended Fig. 6) and that the empirically adjusted[33] cellular concentrations of TAR and Tat:SEC are low enough that most molecules are unbound ("subsaturating") our model predicts that the observed differences in Tat-dependent cellular transactivation ($\Delta\Delta G_{cell}$) for two TAR variants will be equal to the energetic difference with which they bind the Tat-peptide (see Methods),

$$\Delta\Delta G_{cell} = \Delta\Delta G_{pep}.$$

We used a gene-reporter assay to quantitatively measure ΔG_{cell} for the library of TAR variants[33], by transiently transfecting (i) the TAR variants driving *Firefly* luciferase (FLuc) expression (ii) Tat under control of a constitutive CMV promoter, and (iii) *Renilla* luciferase (RLuc) also being driven by the CMV promotor to control for transfection efficiency (Fig. 3a). We quantified Tat-dependent cellular transactivation across the different TAR variants from luminescence measurements in the presence and absence of Tat (Extended Data Fig. 6). The cellular concentration of Tat was empirically adjusted by varying the amount of Tat plasmid transfected to avoid saturating TAR (see Methods, Extended Fig. 6). We then converted the measured fold-differences in cellular transactivation into free energy differences ($\Delta\Delta G_{cell}$) (Fig. 3b, Supplementary Table 5).

Remarkably, as predicted by our simple thermodynamic model, we observed quantitative agreement between $\Delta\Delta G_{pep}$ and $\Delta\Delta G_{cell}$ for *wt* and the $U_{n \geq 0}$ variants (Pearson correlation r = 0.89, p = 0.001) (Fig. 3c). Our experimental data fit the model with an RMSE of 0.41 kcal/mol and an R^2 value of 0.71 (Fig. 3c, Supplementary Table 4). The model predicts the experimental data to within the 95% CIs of the best-fit slope and y-intercept. These results indicate that the differences in conformational propensities to form the stacked base-triple conformation ($\Delta\Delta G_{conf}$) is similar *in vitro* and in cells and that this propensity determines differences in cellular transactivation ($\Delta\Delta G_{cell}$) (see Fig. 4c).

To further test that the observed differences in ΔG_{cell} for the TAR variants originate from differences in TAR-Tat:SEC binding as predicted by our model, we semi-quantitatively measured the TAR-Tat:SEC binding *in vitro* for a subset of TAR variants (*wt* and $U_{1,2,4,6,7}$) (Fig. 3d, Extended Data Fig. 7, Supplementary Table 6). We observed excellent agreement (Pearson correlation r = 0.86, p = 0.012) between $\Delta\Delta G_{prot, app}$ and $\Delta\Delta G_{cell}$ (Fig. 3e).

Our experiments testing the functional effects of base triple mutations on cellular Tat-dependent transactivation also gave results consistent with our simple model. Specifically, G_{cell} for base-triple disrupted mutants was uniformly reduced by a constant (c_{triple}) relative to counterparts lacking the base-triple destabilizing mutation ($\Delta\Delta G_{j^*-j} = c_{triple} \sim 1.2\ kcal/mol$; Fig. 4a), in quantitative agreement with predictions from our model (Supplementary Discussion 2). These results suggest that in contrast to Tat-ARM interactions, the Tat:SEC complex preferentially binds these TAR variants in a stacked base-triple like conformation, possibly to ensure contacts also form between the TAR apical loop and the cyclin T1 protein component of Tat:SEC (Supplementary Discussion 2, Extended Data Fig. 5).

Taken together, these results indicate that the difference between the conformational propensities to form the stacked, bound state across the TAR variants are quantitatively the same, within error, *in vitro* and in cells (Fig. 4b). Thus, this fundamental property of the isolated RNA is not altered by differences in *in vitro* versus cellular conditions. We also observed deviations from the model with the U_7 variant, which indicate directions for further biophysical investigations, and the resulting explanations of these behaviors led to a more complete understanding of this system (Supplementary Discussion 1).

Conclusions

The impact of mutations on RNA cellular activity is typically interpreted in terms of their effects on intermolecular contacts with proteins or on overall RNA folding[34]. Our study revealed a generally hidden mechanism for tuning RNA cellular activity by altering the propensities to form rare and low-abundance biologically active conformational states (Fig. 4c). Our results demonstrate that conformational propensities depend on both sequence and secondary structure and are needed to quantitatively model RNA cellular activity and understand the evolutionary pressures on RNA sequence and structure. In this regard, it is noteworthy that some of the most common natural variants of HIV-1 TAR, U_2 and U_3, have the highest stacking propensities and, in some cases, have activities exceeding that of wild-type TAR (Fig. 2c–d, Fig. 3b). Future studies should dissect the microscopic thermodynamic contributions to overall conformational properties measured here, including enthalpy and entropy contributions and those arising from the RNA, protein, ions, and solvent, and also assess any contribution from cellular crowding.

Our findings underscore the importance of going beyond the current paradigm of probing dominant structures *in vitro* and *in vivo* to quantitatively measuring ensembles that describe the propensities to form rare and short-lived biologically active conformational states. By bridging ensembles determined *in vitro* with measurements of cellular activity, the approach established in this study can broadly be applied to quantitatively study the role of conformational propensities in the cellular activities of RNAs and the effect of the cellular environment on these propensities.
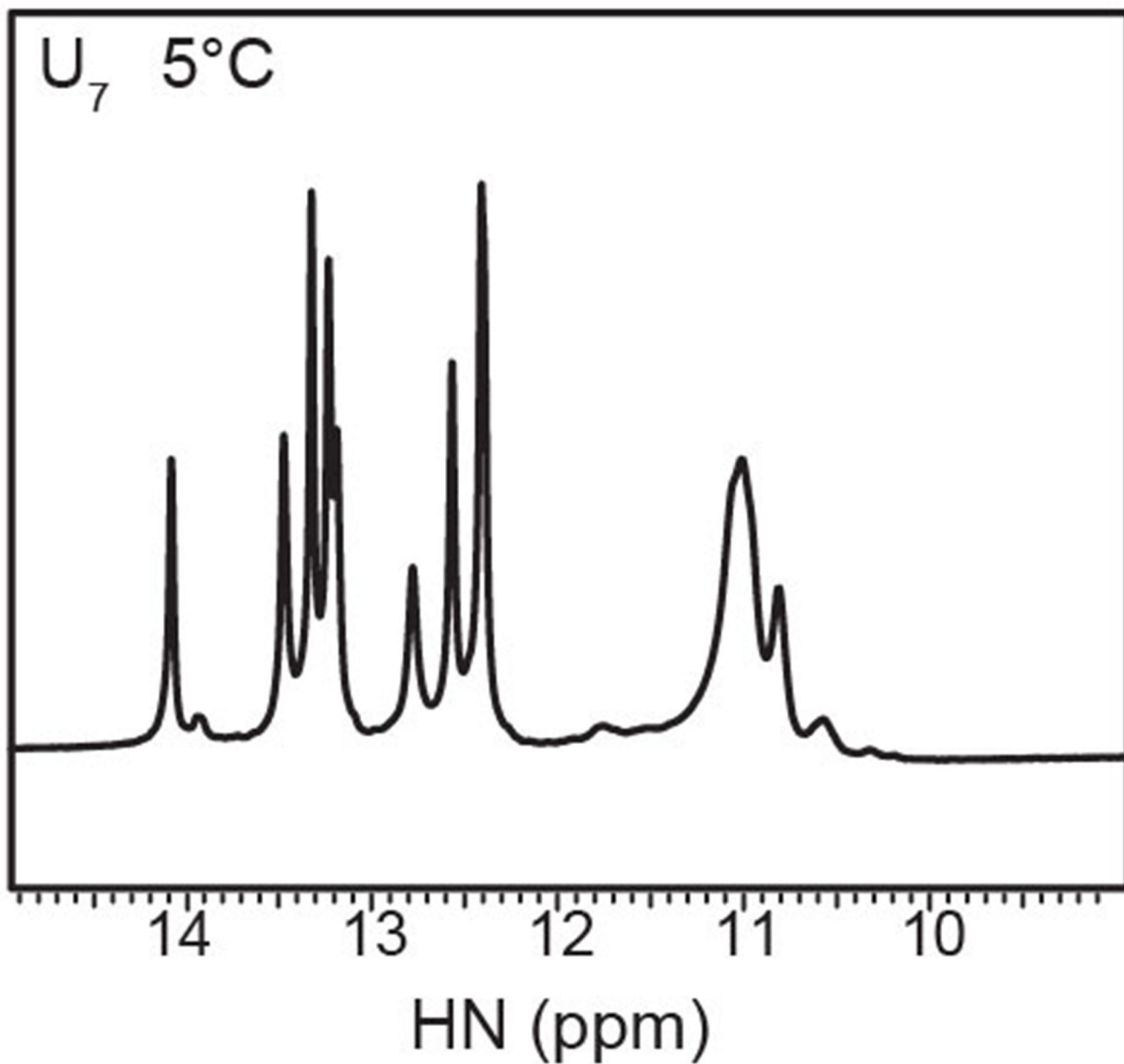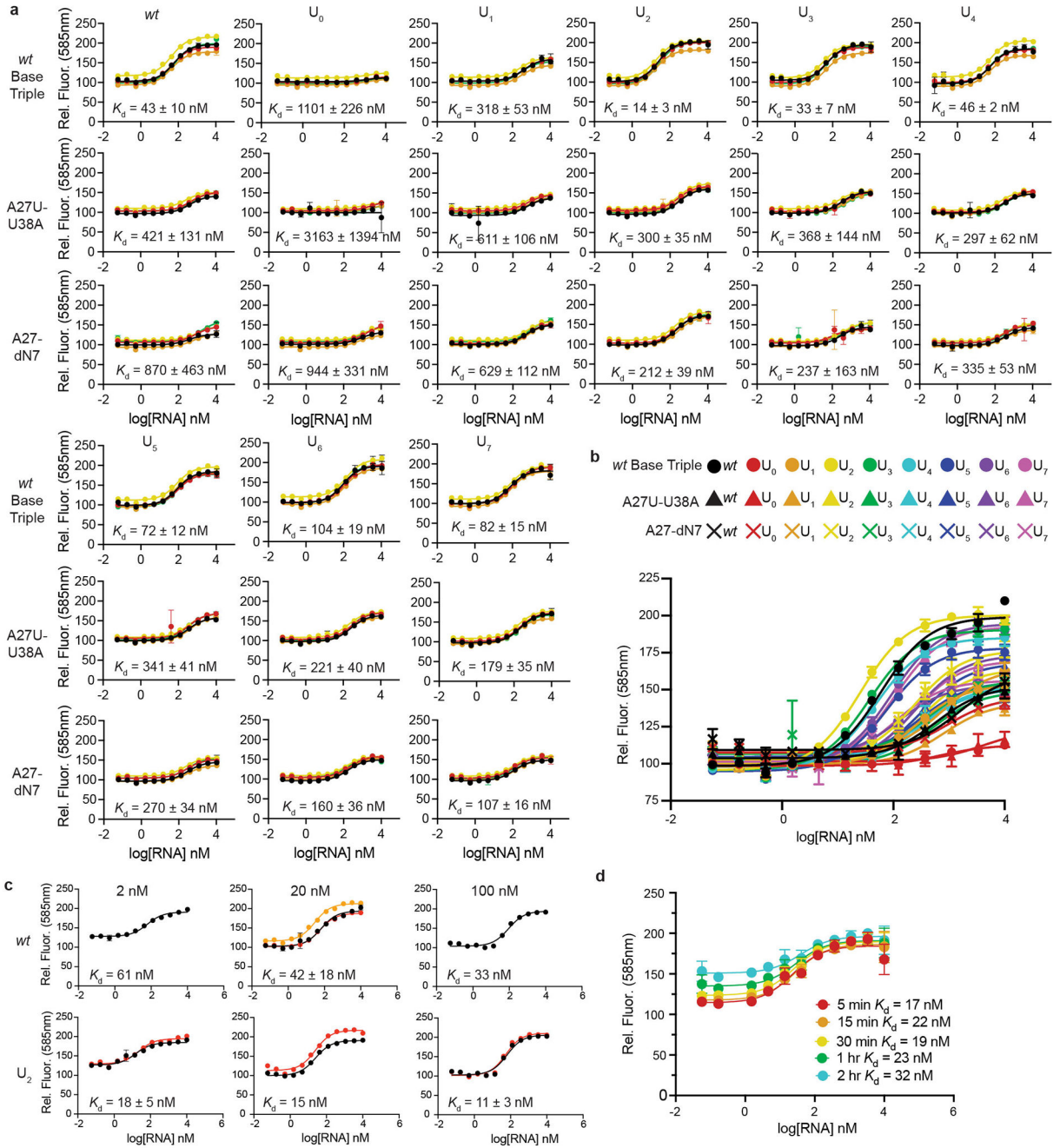
Extended Data
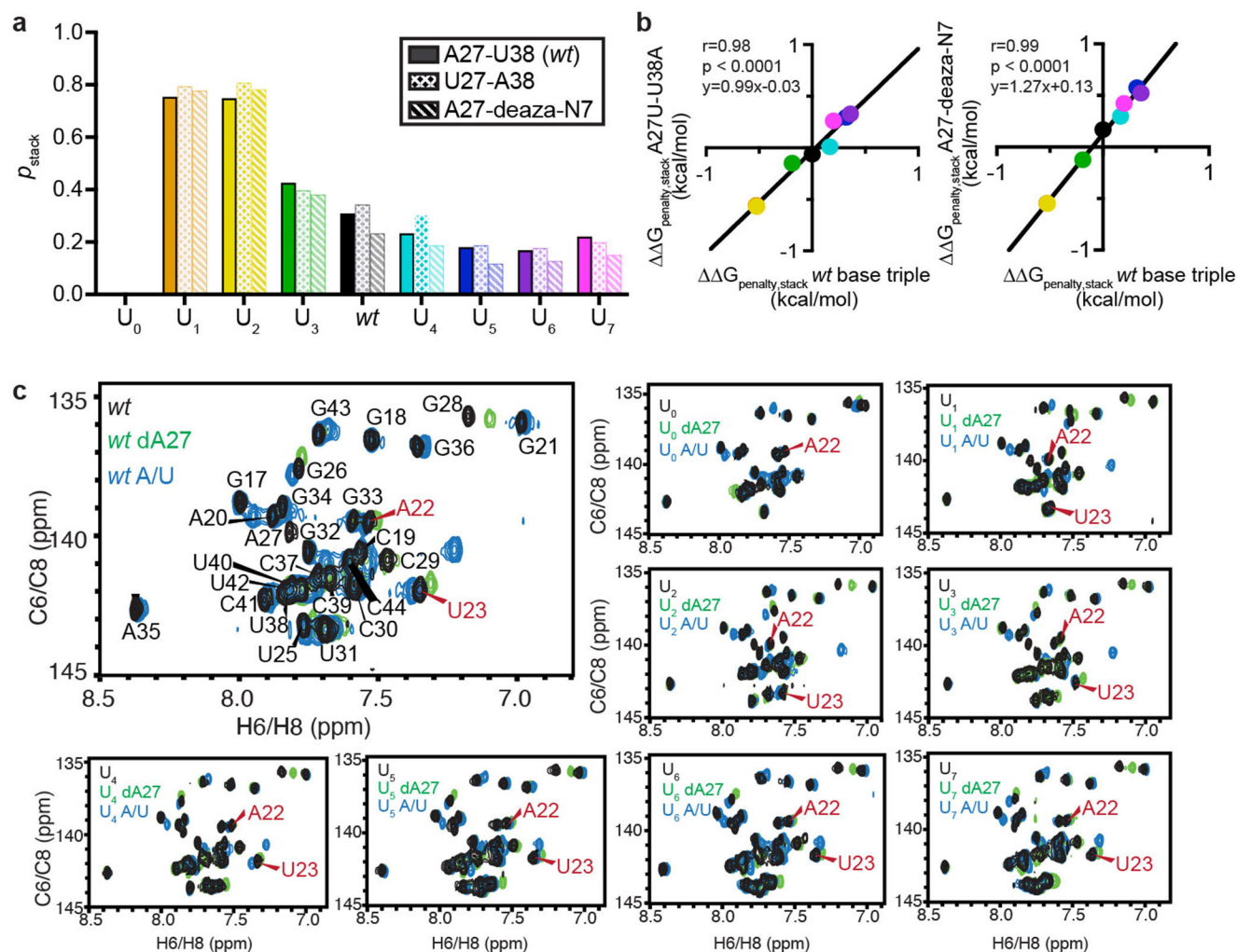
Extended Data Figure 1 |. Measurement of p_{stack} by 2D aromatic [^{13}C, ^{1}H] SOFAST-HMQC[44].
a, p_{stack} (see Methods) for all TAR mutants U_{0-7} and *wt* with and without the A27U-U38A and A27-deaza-N7 base-triple disrupting mutations. **b,** Differences in $\Delta\Delta G_{penalty, stack}$ between the *wt* A27-U38 version of each bulge variant to its two corresponding base-triple disrupted variants, U27-A38 and N7-deaza-A27, are small (within +/−0.3 kcal/mol). This correspondence in stacking is indicated by the strong linear correlations observed between stacked populations for the *wt* base triple mutants versus their base triple disrupted counterparts, Pearson correlation (r) and line of best fit shown, where the colors correspond to the bulge length as shown in panel a. **c,** Sets of overlayed spectra for *wt* and all bulge mutants U_{0-7}. For each, the *wt* base triple construct is black and the base-triple disrupting mutants are overlayed, A27U-U38A in blue and A27-deaza-N7 in green. The *wt* spectrum is fully assigned, for the bulge constructs the stacking reporter residues A22 and U23 are indicated.

Extended Data Figure 2 |. NMR evidence for U-U wobbles in the U$_7$ TAR variant.
The ^{1}H 1D imino NMR spectrum of the U$_7$ variant shows resonances in the 10-12 ppm region, suggesting the U-rich bulge might transiently form a short helix comprised of U-U wobble mismatches which could in turn promote stacking of the TAR helices.

Extended Data Figure 3 |. TAR-Tat-ARM peptide binding assay.
a, Binding curves for individual TAR variants, with all five independent experiments
overlayed (black: experiment 1, red: experiment 2, orange: experiment 3, yellow: experiment
4, green: experiment 5). The data points for each individual curve represent the mean
fluorescence values, and the error bars represent the standard deviation, of 3 technical
replicates. Each individual curve was fit to equation 1, and average K_d values +/−
the standard deviation over the five independent experiments are displayed for each
mutant. **b,** One experiment (experiment 5) of representative fluorescence binding curves

for all TAR mutants overlayed. The data points for each individual curve represent the mean fluorescence values, and the error bars represent the standard deviations, of 3 technical replicates. **c,** Observed dissociation constants do not change as the concentration of the constant component (Tat-ARM peptide) is varied, as expected for accurate K_d measurements[38]. Dissociation constants were measured for *wt* and U_2 at multiple concentrations of Tat-ARM peptide, varying 50-fold. The dissociation constants for *wt* and U_2 remain vary < 2-fold over this range. The data points for each individual curve represent the mean fluorescence values, and the error bars represent the standard deviation, of 3 technical replicates. Each individual curve was fit to equation 1, and average K_d values +/− the standard deviation over the 1 (*wt*-2 nM, *wt*-100 nM), 2 (U_2-2 nM, U_2-20 nM, U_2-100 nM), or 3 (*wt*-20 nM) independent experiments are displayed for each mutant. **d,** Observed dissociation constants do not change as the equilibration time is varied, as expected for accurate K_d measurements[38]. Shown are K_d measurements for *wt* at varying timepoints to demonstrate the reaction has reached equilibrium. The K_d value does not decrease with increasing incubation times, indicating the reaction has reached equilibrium at the lowest timepoint. The same assay plate was read at each time point, creating a photobleaching effect at each subsequent timepoint, which is evident in the increasing baseline values. The data points for each individual curve represent the mean fluorescence values, and the error bars represent the standard deviation, of 3 technical replicates. Each individual curve was fit to equation 1, with the resulting K_d values displayed.
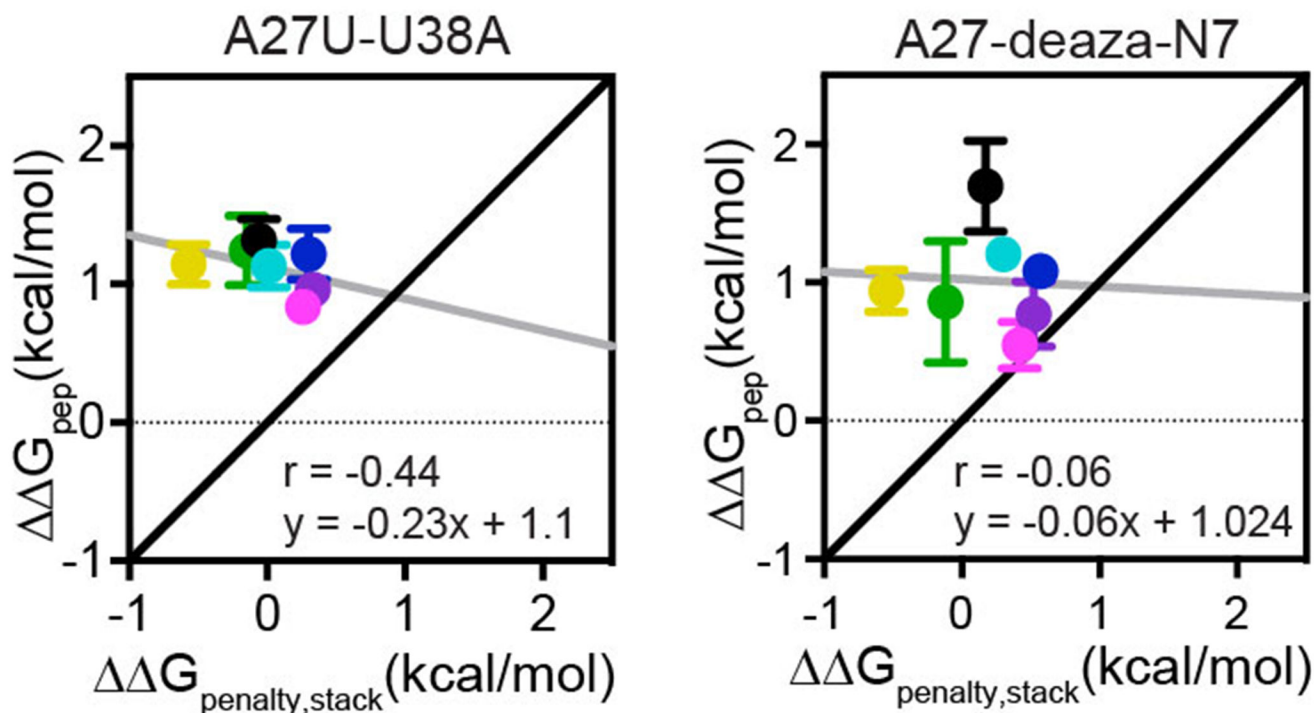
Extended Data Figure 4 |. Stacking and peptide binding energetics for *wt* and U_{2-7}.
$\Delta\Delta G_{pep}$ *versus* $\Delta\Delta G_{penalty, stack}$ for base-triple destabilized mutants, A27U-U38A (left) and A27-deaza-N7 (right), correlates poorly (Pearson correlation shown). Grey lines indicate the best

fit (equation shown), and black lines indicate slope of 1, which is the prediction of the model in the absence of the base triple disrupting mutations. Error bars represent the standard deviation of 5 independent experiments measuring $\Delta\Delta G_{pep}$.
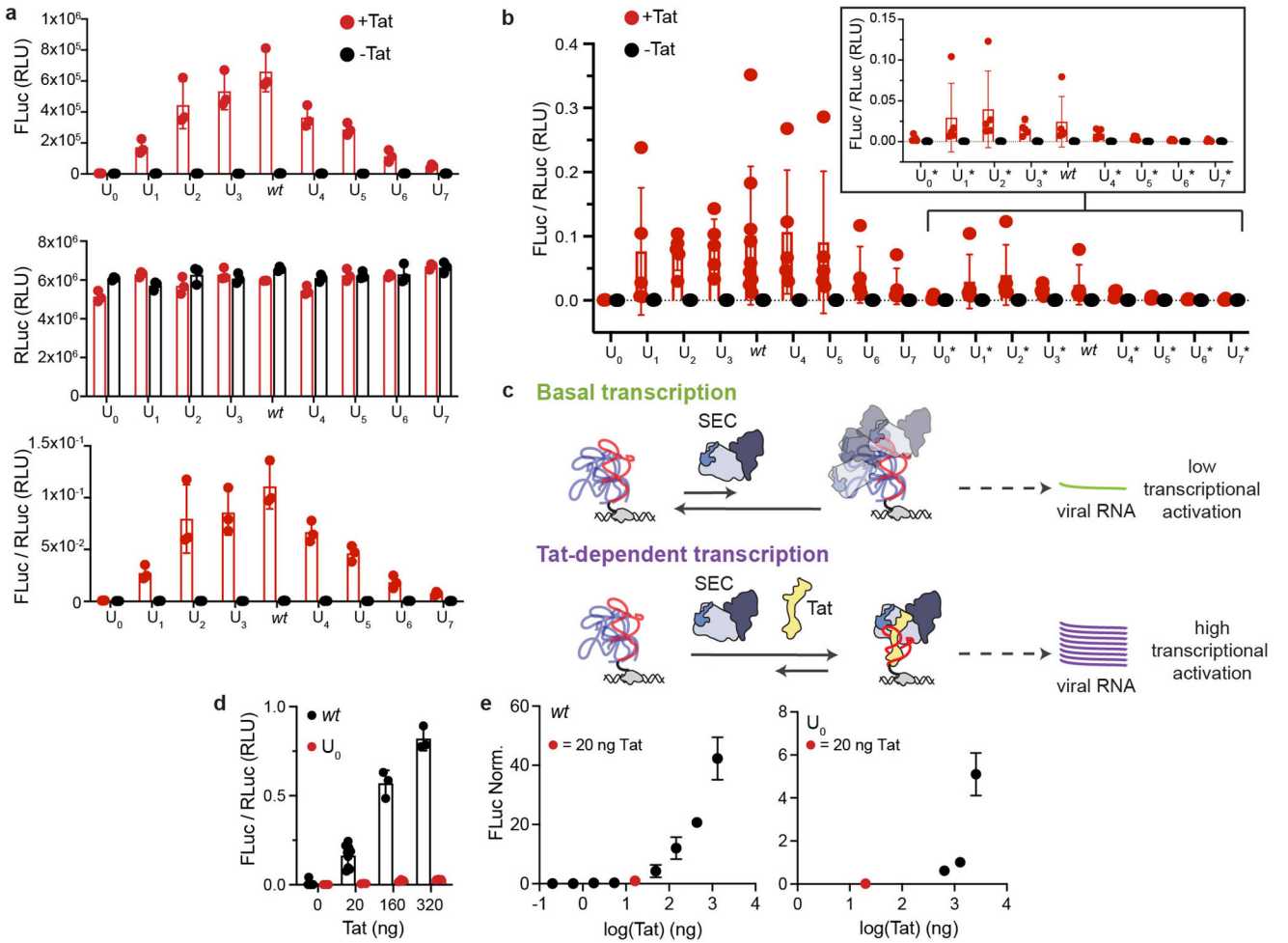
Extended Data Figure 5 |. Energetics of base-triple disruption in Tat-ARM binding and cellular transactivation

a, Changes in fluorescence upon peptide binding is greater for base-triple competent variants than for base-triple disrupted variants. Shown are the fitted minimum and maximum fluorescence values (from equation 1, see Methods) from the TAR-Tat-ARM peptide binding assay for 5 independent experiments. Red dotted lines indicate average maximum values for the base-triple competent variants (190), and base-triple disrupted variants (155). U_{0-1} are shown in grey as they are unable to form the base-triple. **b,** Energy diagram of Tat-ARM peptide binding to base triple competent and base-triple disrupted variants. The peptide can bind a bulge-independent kinked TAR conformation lower in energy than the base-triple disrupted stacked conformation. **c**, Energy diagram of Tat:SEC binding to TAR in the cellular context. The favorable interactions between Cyclin T1 and the TAR apical loop are unable to form in the kinked state of TAR, and so each base-triple disrupted

variant is destabilized by the same amount (c_{triple}) and binds its non-base triple stacked state (demarcated with an asterisk*). **d,** Proposed model for an alternative sheared base-triple conformation in the A27U-U38A base-triple disrupting mutants with hydrogen bonds shown as black dashed line (left). Two views of the 3D structural model for the alternative sheared base-triple conformation obtained by replacing A27 with U and U38 with A in the PDBID:6MCE[22] U$_2$ TAR structure (right).
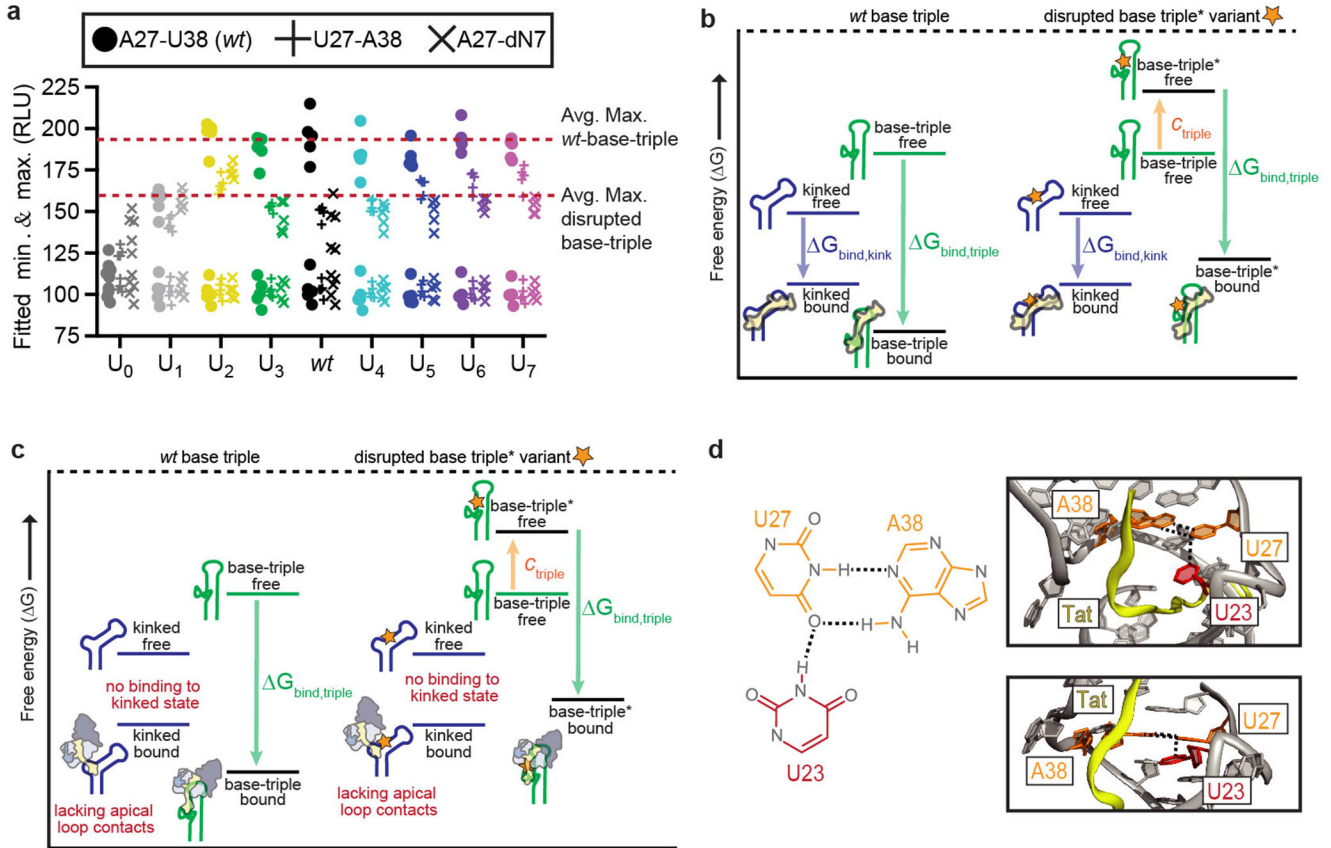

Extended Data Figure 6 |. Cellular transactivation assay.

a, Representative example of luminescence data for one biological replicate of U$_{0-7}$ and *wt* (3 technical replicates). Shown are luminescence values for Firefly luciferase, reporting on transactivation (top), luminescence values for *Renilla* luciferase under control of a CMV promoter, used as a control for transfection (middle), and the ratio FLuc/RLuc to normalize for differences arising from transient transfection (bottom), with the error bars representing the standard deviations of those values over 3 technical replicates. **b,** Aggregate FLuc/RLuc data for all TAR mutants over 5 independent experiments (biological replicates). Mutants labelled with (*) indicate the A27U/U38A base-triple disrupting mutation. In all graphs, red data are values when Tat is co-transfected and black data are values without Tat, representing Tat-independent baseline activity. Error bars represent the standard deviation in FLuc/RLuc values over 5 biological replicates. **c,** Model of Tat-dependent versus Tat-independent transactivation energetics in cells. (Top) The observed level of basal transcription is likely due to many nonspecific binding interactions of the preformed SEC complex to TAR,

which does not alter the conformational propensities of the TAR ensemble and has a low probability of achieving an active bound conformation leading to transactivation and transcription. (Bottom) In Tat-dependent transactivation, the presence of Tat increases the binding affinity to form the active bound state, leading to higher levels of transactivation and transcription. **d,** Tat plasmid titration. In this experiment, the concentration of Tat was varied for *wt* (black), one of the most transactivating constructs, and U_0 (red), one of the least transactivating constructs. We see that for both *wt* and U_0, the level of transactivation (FLuc/RLuc) increases with an increase in Tat, indicating that the reaction is not saturated at the level of Tat we are using (20 ng). Dots are the individual FLuc/RLuc values and error bars represent the standard deviation in these values over 3 independent experiments. **e,** Larger scale Tat plasmid titrations for *wt* and U_0 covering four orders of magnitude, with the y-axis being FLuc signal normalized to the average FLuc value measured for *wt* at 20 ng Tat. Again, for both mutants, the level of transactivation continually increases with an increase of Tat plasmid; the value we use in our assays (20 ng, red dot) is at the low end of this spectrum. Dots represent the average, and error bars the standard deviation, of normalized FLuc luminescence values over 3 independent experiments.
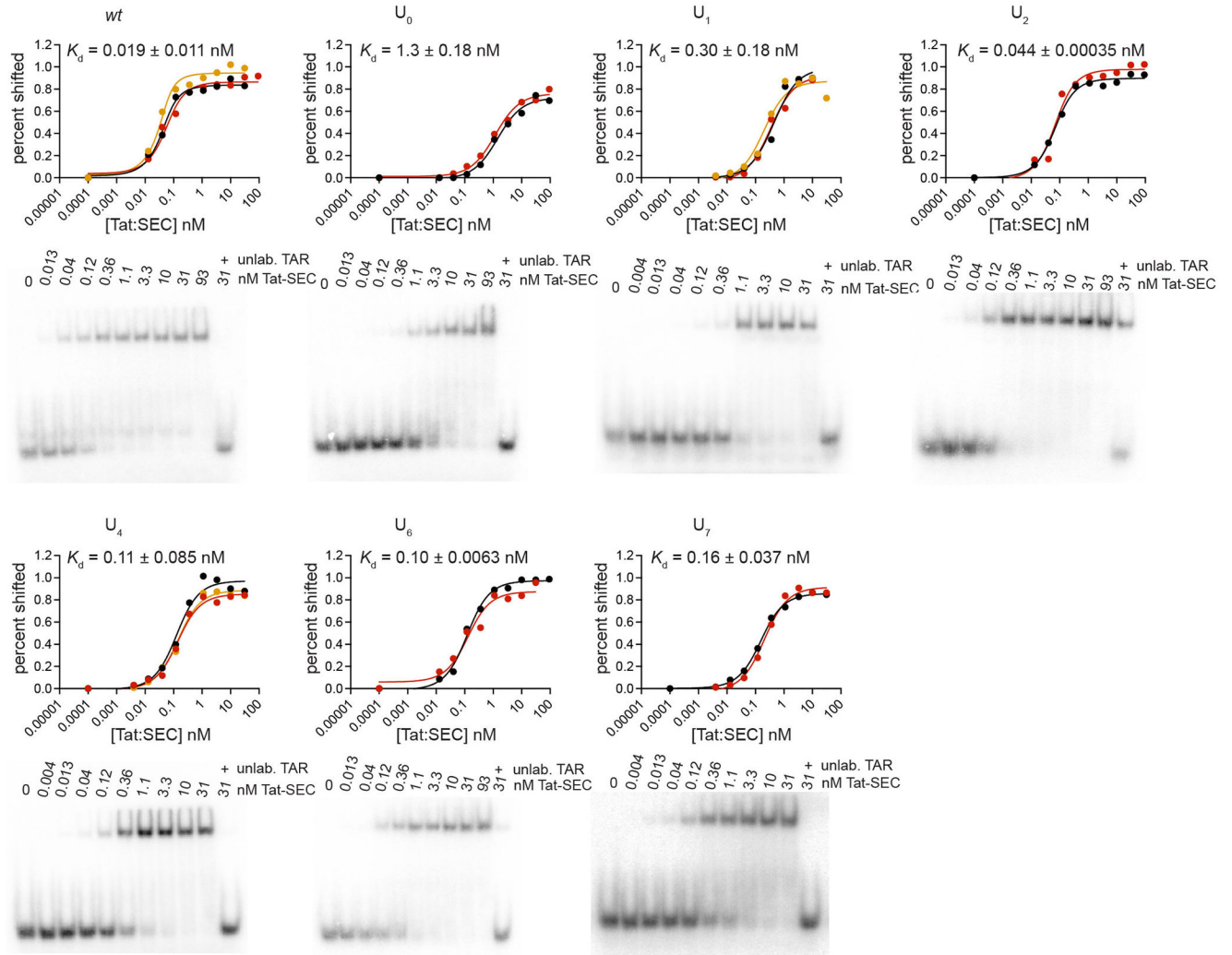
Extended Data Figure 7 |. Measurements of TAR-Tat:SEC binding using electrophoretic mobility shift assay (EMSA).

Shown are EMSA binding curves for TAR bulge mutants $U_{0,1,2,4,6,7}$ and UCU along with average apparent K_d values (see Methods) for each variant, obtained by fitting data to equation 2 using GraphPad Prism (version 9.3.1). Binding curves from 2 (U_0, U_2, U_6, U_7) or 3 (*wt*, U_1, U_4) independent experiments are overlayed (black: experiment 1, red: experiment 2, orange: experiment 3). Below the binding curves for each variant is one representative EMSA gel (experiment 1) of 2 total gels (U_0, U_2, U_6, U_7) or 3 total gels (*wt*, U_1, U_4) for each variant.
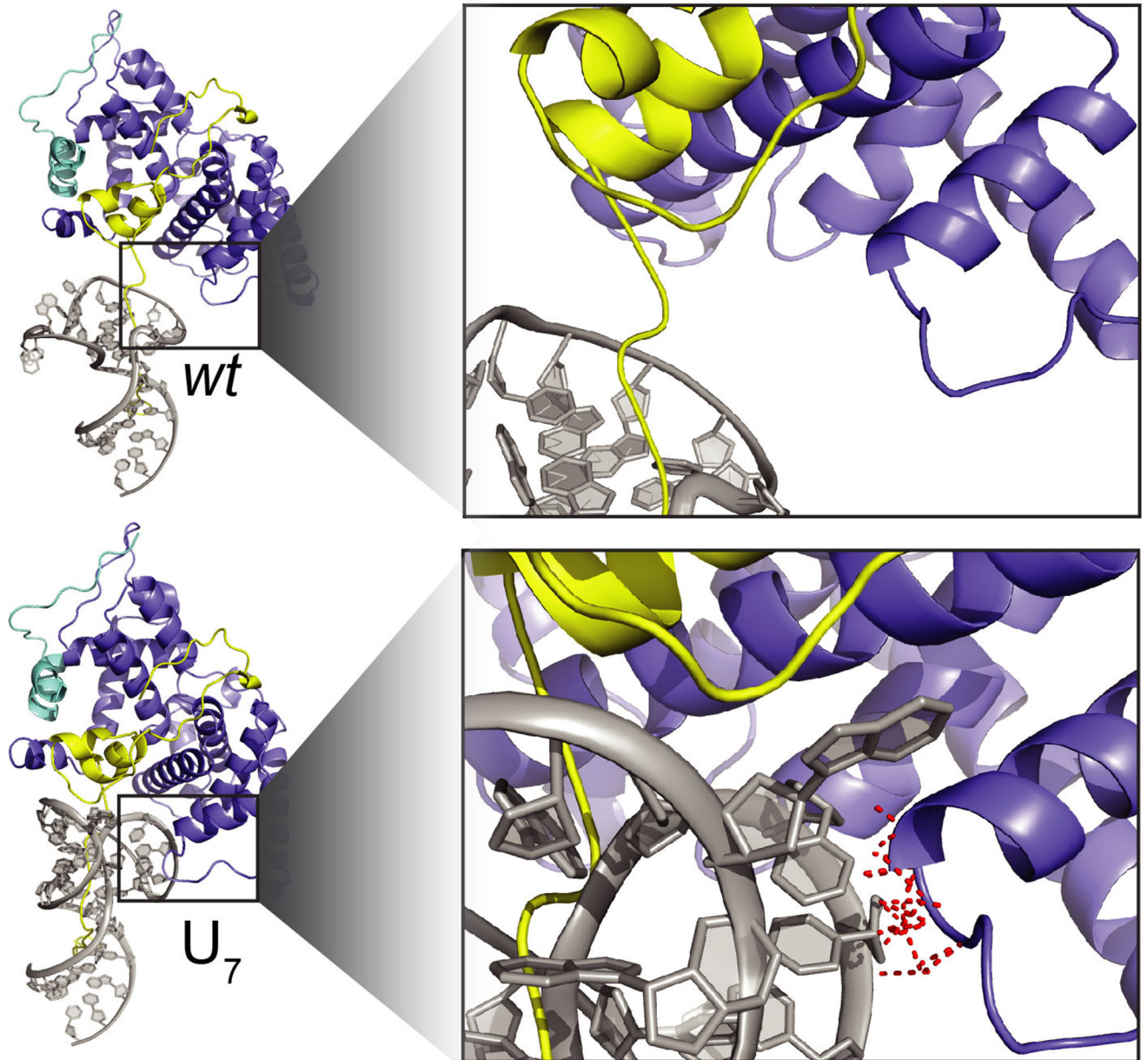
Extended Data Figure 8 |. Model of steric interaction between the U_7 bulge and P-TEFb.
(Left) FARFAR models of representative base-triple conformations of *wt* and U_7 bound to the Tat:SEC complex. (Right) Zoomed in view of the bulge interaction with P-TEFb. In dashed red lines are atom distances between bulge residues and P-TEFb that are within 2.5 Å, representing steric overlap. U_7 (bottom) has multiple steric overlaps, whereas *wt* (top) does not.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Main References

1. Ganser LR, Kelly ML, Herschlag D & Al-Hashimi HM The roles of structural dynamics in the cellular functions of RNAs. Nat. Rev. Mol. Cell Biol 20, 474–489 (2019). [PubMed: 31182864]

2. Fischer M, Coleman RG, Fraser JS & Shoichet BK Incorporation of protein flexibility and conformational energy penalties in docking screens to improve ligand discovery. Nat. Chem 6, 575–583 (2014). [PubMed: 24950326]

3. Parvin JD, Mc Cormick RJ, Sharp PA & Fisher DE Pre-bending of a promoter sequence enhances affinity for the TATA-binding factor. Nature 373, 727 (1995).

4. Denny SK et al. High-Throughput Investigation of Diverse Junction Elements in RNA Tertiary Folding. Cell 174, 377–390.e20 (2018). [PubMed: 29961580]

5. Afek A et al. DNA mismatches reveal conformational penalties in protein–DNA recognition. Nat. 2020 5877833 587, 291–296 (2020).

6. Berman HM et al. The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000). [PubMed: 10592235]

7. Orlovsky NI, Al-Hashimi HM & Oas TG Exposing Hidden High-Affinity RNA Conformational States. J. Am. Chem. Soc 142, 907–921 (2020). [PubMed: 31815464]

8. Alderson TR & Kay LE NMR spectroscopy captures the essential role of dynamics in regulating biomolecular function. Cell 184, 577–595 (2021). [PubMed: 33545034]

9. Guo JU & Bartel DP RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. Science 353, aaf5371 (2016). [PubMed: 27708011]

10. Rouskin S, Zubradt M, Washietl S, Kellis M & Weissman JS Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature 505, 701–705 (2014). [PubMed: 24336214]

11. Shi H et al. Rapid and accurate determination of atomistic RNA dynamic ensemble models using NMR and structure prediction. Nat. Commun 11, 5531 (2020). [PubMed: 33139729]

12. Tomezsko P, Swaminathan H & Rouskin S Viral RNA structure analysis using DMS-MaPseq. Methods 183, 68–75 (2020). [PubMed: 32251733]

13. Schulze-Gahmen U & Hurley JH Structural mechanism for HIV-1 TAR loop recognition by Tat and the super elongation complex. Proc. Natl. Acad. Sci. USA 115, 12973–12978 (2018). [PubMed: 30514815]

14. D'Orso I et al. Transition Step during Assembly of HIV Tat:P-TEFb Transcription Complexes and Transfer to TAR RNA. Mol. Cell. Biol 32, 4780 (2012). [PubMed: 23007159]

15. Liu Y, Suñé C & Garcia-Blanco MA Human Immunodeficiency Virus Type 1 Tat-Dependent Activation of an Arrested RNA Polymerase II Elongation Complex. Virology 255, 337–346 (1999). [PubMed: 10069959]

16. Marciniak RA, Calnan BJ, Frankel AD & Sharp PA HIV-1 Tat protein trans-activates transcription in vitro. Cell 63, 791–802 (1990). [PubMed: 2225077]

17. Karn J, Stoltzfus CM, Bushman FD, Nabel GJ & Swanstrom R Transcriptional and posttranscriptional regulation of HIV-1 gene expression. Cold Spring Harbor perspectives in medicine 2, a006916 (2012). [PubMed: 22355797]

18. Merriman DK et al. Increasing the length of poly-pyrimidine bulges broadens RNA conformational ensembles with minimal impact on stacking energetics. RNA 24, 1363–1376 (2018). [PubMed: 30012568]

19. Puglisi JD, Chen L, Frankel AD & Williamson JR Role of RNA structure in arginine recognition of TAR RNA. Proc. Natl. Acad. Sci. USA 90, 3680–3684 (1993). [PubMed: 7682716]

20. Puglisi JD, Tan R, Calnan BJ, Frankel AD & Williamson JR Conformation of the TAR RNA-arginine complex by NMR spectroscopy. Science 257, 76–80 (1992). [PubMed: 1621097]

21. Tao J, Chen L & Frankel AD Dissection of the proposed base triple in human immunodeficiency virus TAR RNA indicates the importance of the Hoogsteen interaction. Biochemistry 36, 3491–3495 (1997). [PubMed: 9131998]

22. Pham VV et al. HIV-1 Tat interactions with cellular 7SK and viral TAR RNAs identifies dual structural mimicry. Nat. Commun 9, 4266 (2018). [PubMed: 30323330]

23. Brodsky AS & Williamson JR Solution structure of the HIV-2 TAR-argininamide complex. J. Mol. Biol 267, 624–639 (1997). [PubMed: 9126842]

24. Calnan BJ, Tidor B, Biancalana S, Hudson D & Frankel AD Arginine-mediated RNA recognition: The arginine fork. Science 252, 1167–1171 (1991). [PubMed: 1709522]

25. Chavali SS, Bonn-Breach R & Wedekind JE Face-time with TAR: Portraits of an HIV-1 RNA with diverse modes of effector recognition relevant for drug discovery. J. Biol. Chem 294, 9326–9341 (2019). [PubMed: 31080171]

26. Chavali SS, Cavender CE, Mathews DH & Wedekind JE Arginine Forks Are a Widespread Motif to Recognize Phosphate Backbones and Guanine Nucleobases in the RNA Major Groove. J. Am. Chem. Soc 142, 19835–19839 (2020). [PubMed: 33170672]

27. Mustoe AM, Al-Hashimi HM & Brooks CL Coarse Grained Models Reveal Essential Contributions of Topological Constraints to the Conformational Free Energy of RNA Bulges. J. Phys. Chem. B 118, 40 (2014).

28. Karn J, Dingwall C, Finch JT, Heaphy S & Gait MJ RNA binding by the tat and rev proteins of HIV-1. Biochimie 73, 9–16 (1991). [PubMed: 1903308]

29. Barthel A & Zacharias M Conformational Transitions in RNA Single Uridine and Adenosine Bulge Structures: A Molecular Dynamics Free Energy Simulation Study. Biophys. J 90, 2450–2462 (2006). [PubMed: 16399833]

30. Qi Y et al. Continuous Interdomain Orientation Distributions Reveal Components of Binding Thermodynamics. J. Mol. Biol 430, 3412–3426 (2018). [PubMed: 29924964]

31. Watkins AM, Rangan R & Das R FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds. Structure 28, 963–976.e6 (2020). [PubMed: 32531203]

32. Matsumoto C, Hamasaki K, Mihara H & Ueno A A high-throughput screening utilization intramolecular fluorescence resonance energy transfer for the discovery of the molecules that bind HIV-1 TAR RNA specifically. Bioorganic Med. Chem. Lett 10, 1857–1861 (2000).

33. Ganser LR et al. Probing RNA Conformational Equilibria within the Functional Cellular Context. Cell Rep. 30, 2472–2480.e4 (2020). [PubMed: 32101729]

34. Taliaferro JM et al. RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation. Mol. Cell 64, 294–306 (2016). [PubMed: 27720642]
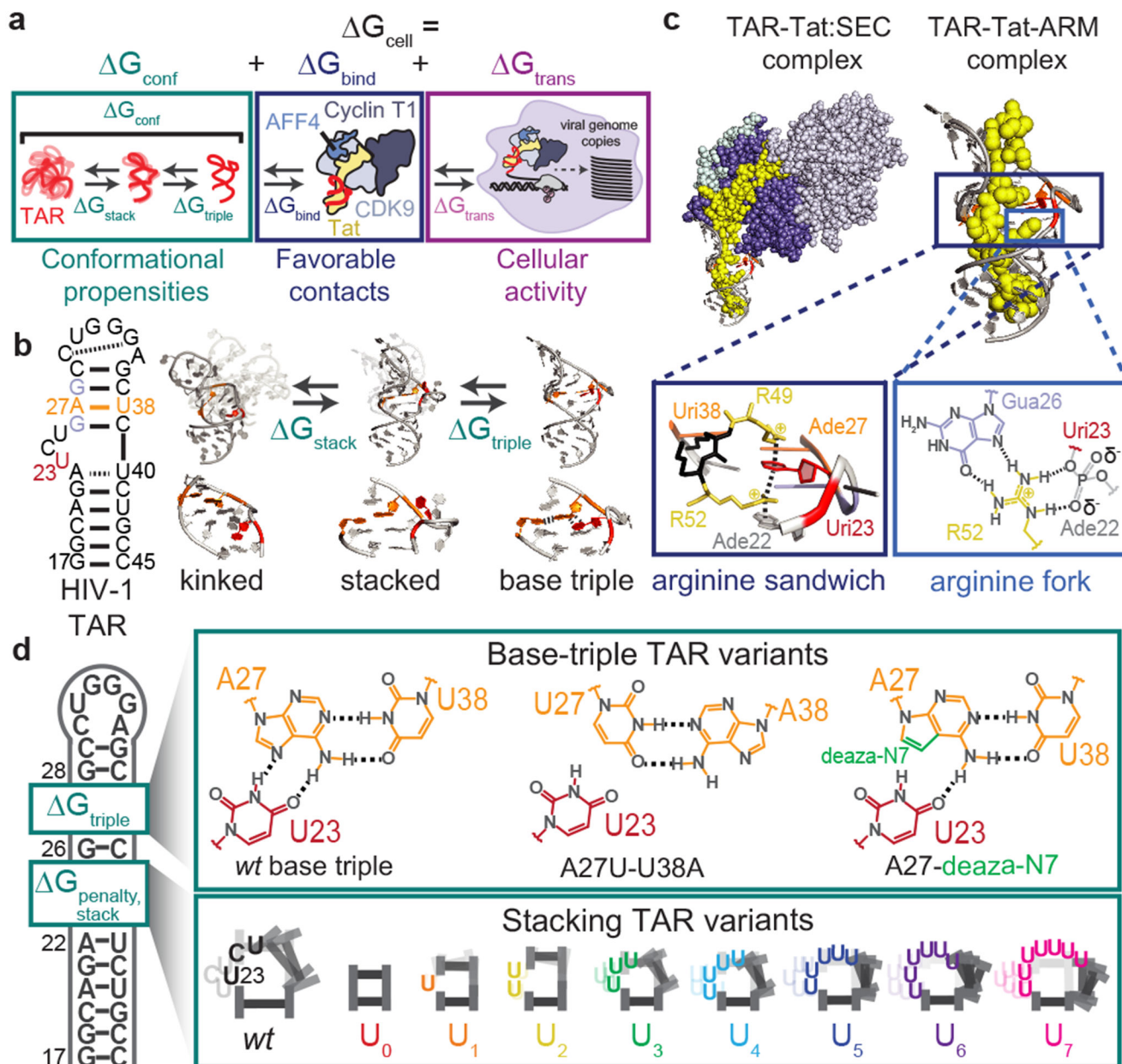
Fig. 1: Revealing the role of conformational propensities in HIV-1 Tat-dependent cellular transactivation.

a, Thermodynamic model of HIV-1 Tat-dependent transactivation. The energetics of cellular transactivation (ΔG_{cell}) is decomposed into contributions from conformational penalty to redistribute the ensemble into the base-triple bound TAR conformational state ($\Delta G_{conf} = -\text{RT}\ln p_{stack} - \text{RT}\ln K_{triple}$), binding of Tat:SEC to TAR ($\Delta G_{bind} = -\text{RT}\ln K_{bind}$), and the several steps leading to transactivation (ΔG_{trans}). ΔG approximately holds for unfavorable and sub-saturating conditions (see methods for derivation of equations). **b,** Secondary structure of HIV-1 TAR, FARFAR[31] models of the bent and stacked ensembles (see Methods), and base triple conformation (PDB entry 6MCE[22]) with close-ups of the base-triple-forming

component conformations below. **c,** TAR-Tat:SEC complex (modelled using PDB entries 6CYT and 6MCE)[25], TAR-Tat-ARM peptide, and critical contacts between TAR and the Tat arginine rich motif (Tat-ARM). Tat arginine residues R49 and R52 (yellow) form cation-pi interactions (dashed lines) with TAR bases U23 (red) and A22 (grey), with R52 forming a A22/R52/U23 arginine sandwich motif. R52 also forms an arginine fork involving hydrogens bonds (dashed lines) between the guanidinium group and the base of G26 (purple) as well as bridging and non-bridging phosphate groups[24,26]. **d,** Library of TAR variants with two types of mutations which incrementally increase $\Delta G_{penalty,\,stack}$ through replacement of the *wt* UCU bulge with increasingly longer uridine bulges ($U_0 - U_7$) or increase ΔG_{triple} through replacement of A27-U38 with either U27-A38 or deaza-N7 modified A27. Dotted black lines indicate hydrogen bonds in **c** and **d**.
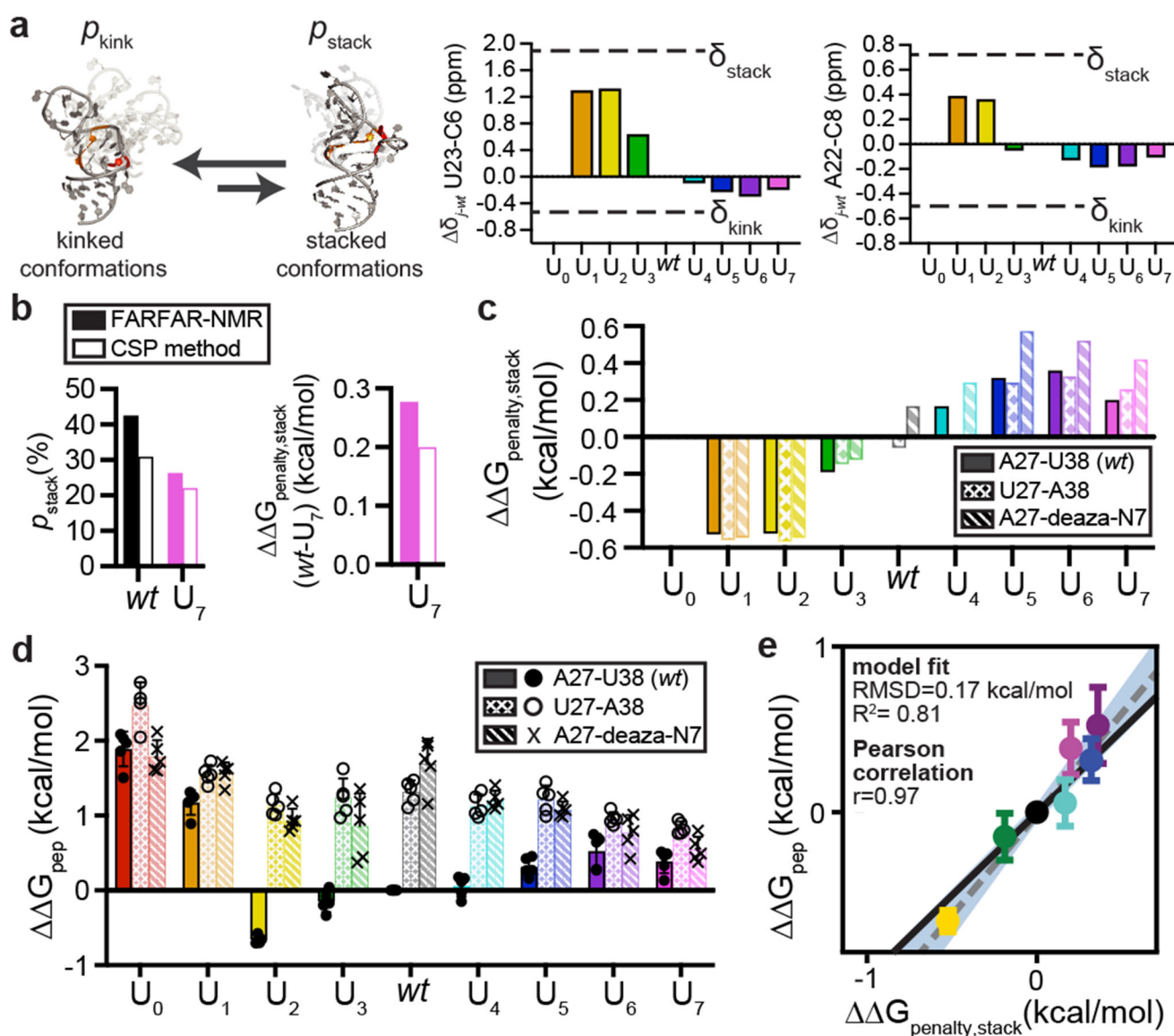
Fig. 2: Differences in stacking propensities predict differences in TAR-Tat-ARM binding.
a, TAR exists in dynamic equilibrium between populations of kinked (p_{kink}) and stacked (p_{stack}) inter-helical conformations[18]. $\Delta G_{penalty, stack}$ is the free energy cost to redistribute the unbound TAR ensemble to the stacked state (see Methods). Chemical shift perturbations at reporter resonances U23-C6 and A22-C8[18] are used to measure p_{stack} (Uncertainty in the ^{13}C chemical shifts are <1% and the chemical shift derived stacked populations <0.02%). **b,** Comparison of p_{stack} and $\Delta\Delta G_{penalty, stack}(wt - U_7)$ deduced using FARFAR-NMR and NMR CSPs (see Methods). **c,** Differences in stacking propensities ($\Delta\Delta G_{penalty, stack}$, referenced to wt) for the bulge variants with (solid bars) and without the base triple destabilizing A27-deaza-N7 and A27U-U38A mutations (stippled bars) obtained from NMR CSPs. Absolute values of $\Delta\Delta G_{penalty, stack}$ are given in Supplementary Table 2. **d,** Differences between the Tat-ARM binding energetics ($\Delta\Delta G_{pep}$, referenced to wt) for TAR bulge variants with and without the
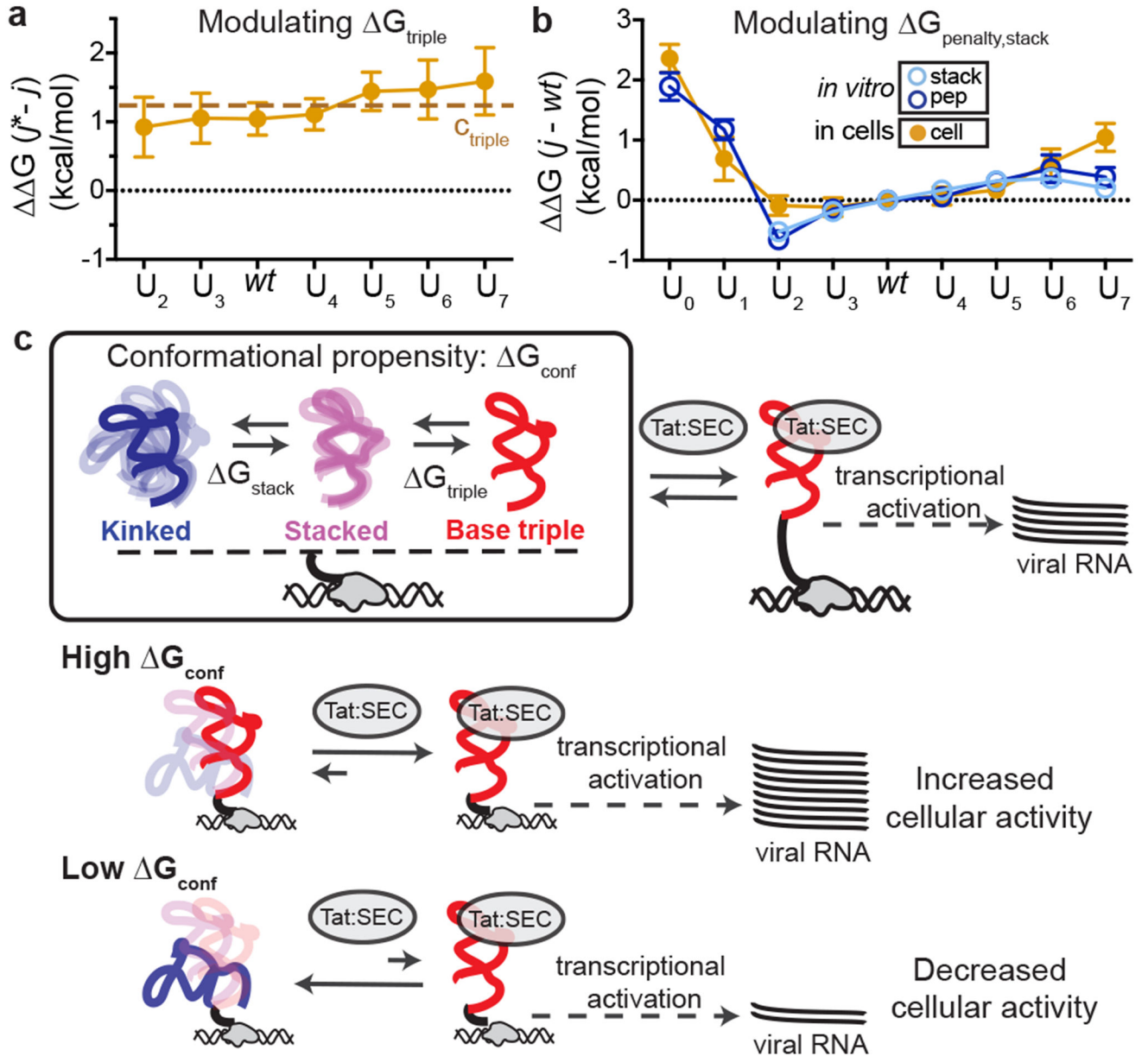
A27-deaza-N7 and A27U-U38A base-triple destabilizing mutations. Bar height represents the mean and error bars represent standard deviations for 5 independent experiments. **e,** Comparison between $\Delta\Delta G_{pep}$ and $\Delta\Delta G_{penalty, stack}$ relative to *wt*, for base-triple forming constructs, with Pearson correlation shown. The black line of slope one indicates predictions from our model. Shown is the fit to this model (RMSE and R^2), as well as the best fit line (dotted, grey) with the region encompassing the 95% confidence intervals for slope and y-intercept shaded in blue. Each data point represents the measured $\Delta\Delta G_{penalty, stack}$ value (x-axis), and average $\Delta\Delta G_{pep}$ (y-axis) values. Vertical error bars represent the standard deviation in $\Delta\Delta G_{pep}$ measurements over 5 independent experiments.
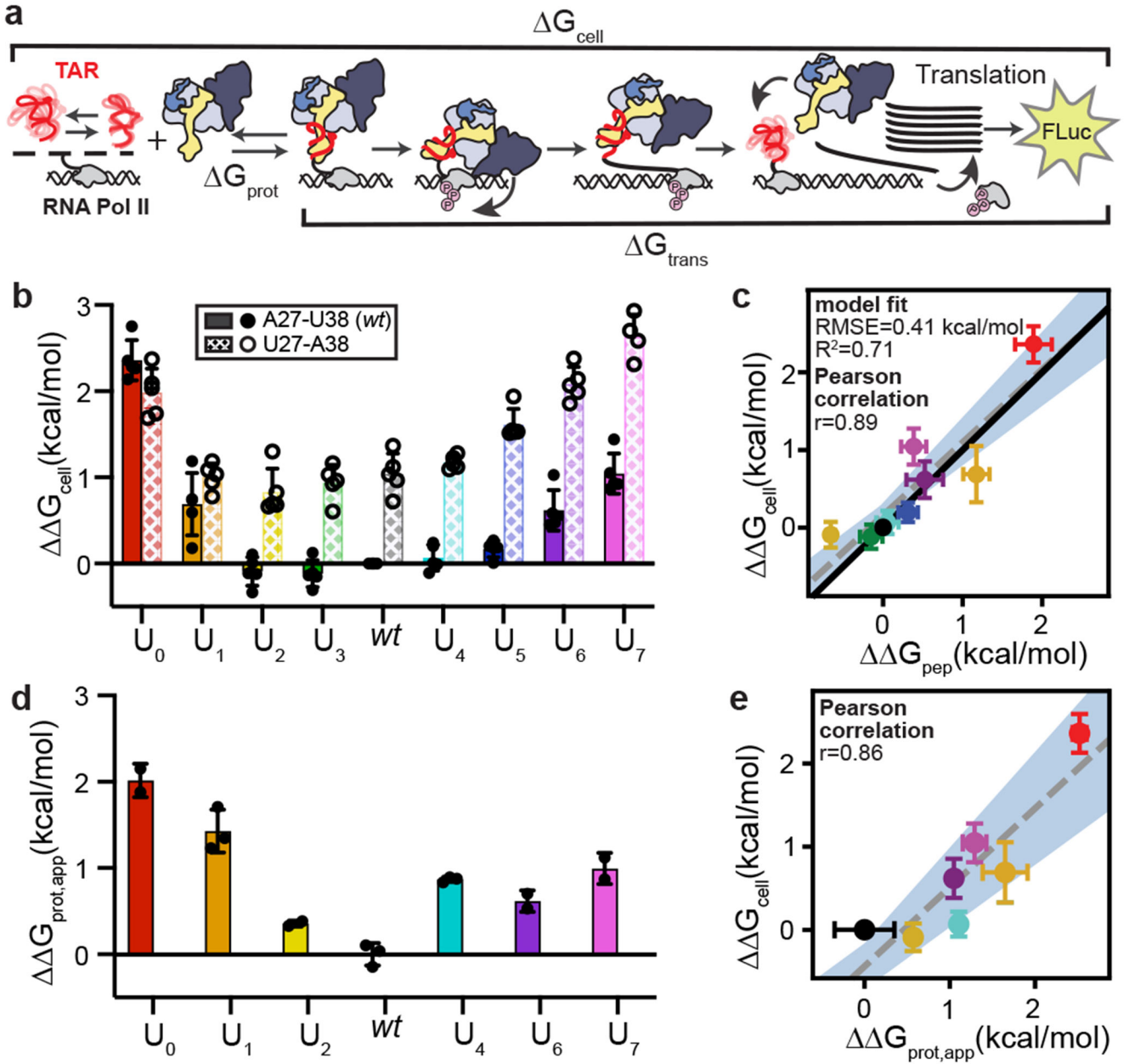
Fig. 3: TAR-Tat-ARM binding predicts differences in Tat-dependent cellular transactivation.
a, Transcriptional activation is a multi-step cellular process which is initiated by binding of the Tat:SEC complex to TAR. The cyclin-dependent kinase 9 (Cdk9) in this complex is then activated, which in turn phosphorylates negative (NELF) and positive (C-terminal domain of RNAP II and Spt5) elongation factors to increase the processivity of RNAPII and activate transcription of the retroviral genome. The energetics of Tat-dependent cellular transactivation (ΔG_{cell}) can be decomposed into the conformational penalty of assuming the bound state, mutation sensitive TAR binding to Tat:SEC ($\Delta G_{prot, app}$), and contributions from other transactivation steps (ΔG_{trans}) assumed to be unaffected by the mutations in our TAR library. **b,** Differences between cellular transactivation ($\Delta\Delta G_{cell}$, referenced to *wt*) for the

bulge variants with (stippled) and without (solid) the base triple destabilizing A27U-U38A mutation. Bar height represents the mean and error bars represent standard deviation for 5 biologically independent experiments. **c,** Comparison between $\Delta\Delta G_{cell}$ and $\Delta\Delta G_{pep}$ for bulge variants U_{0-7} *without* the base-triple destabilizing mutation, Pearson correlation shown. The black line indicates the prediction from our model. Shown is the fit to this model (RMSE and R^2), as well as the best-fit line (dotted, grey) with the region encompassing the 95% confidence intervals for slope and y-intercept shaded in blue. Each data point represents the average $\Delta\Delta G_{pep}$ (x-axis) and $\Delta\Delta G_{cell}$ (y-axis) values, and error bars represent the standard deviation in $\Delta\Delta G_{pep}$ and $\Delta\Delta G_{cell}$, each over 5 independent experiments. **d,** Differences in the apparent Tat:SEC binding energetics ($\Delta\Delta G_{prot,app}$, referenced to *wt*) for the TAR variants. Bar height represents the mean and error bars represent standard deviation over 2 (U_0, U_2, U_6, U_7) or 3 (*wt*, U_1, U_4) independent experiments. **e,** Comparison between $\Delta\Delta G_{cell}$ and $\Delta\Delta G_{prot,app}$ across the TAR variants, Pearson correlation shown. The line of best fit is grey and dotted with the region encompassing the 95% confidence intervals for slope and y-intercept shaded in blue. Each data point represents the average $\Delta\Delta G_{pep}$ (x-axis) and $\Delta\Delta G_{cell}$ (y-axis) values. Error bars represent the standard deviation in $\Delta\Delta G_{prot,app}$ over 2 (U_0, U_2, U_6, U_7) or 3 (*wt*, U_1, U_4) independent experiments, and the standard deviation in $\Delta\Delta G_{cell}$ over 5 independent experiments.

Fig. 4: The role of conformational propensities in Tat-dependent cellular transactivation.
a, Differences in transactivation for *wt* and U_{2-7} variants with the wt-base triple intact (j) and with the A27U-U38A base-triple destabilizing mutation (j^*), with dots representing the average value and the errors bars representing the standard deviation. Orange dashed line is the value of $\Delta\Delta G_{j^*-j}$ predicted by the model (~1.2 kcal/mol) *), with dots representing the average value and the errors bars representing the standard deviation for 5 independent experiments. **b,** Comparison of $\Delta\Delta G_{\text{cell}}$ measured in cells with $\Delta\Delta G_{\text{pep}}$, and $\Delta\Delta G_{\text{penalty, stack}}$ measured *in vitro* for the base-triple forming variants. Dots represent the average value and errors bars represent the standard deviation for 5 independent experiments in the case of $\Delta\Delta G_{\text{cell}}$ and $\Delta\Delta G_{\text{pep}}$, and one NMR chemical shift experiment in the case of $\Delta\Delta G_{\text{penalty, stack}}$

c, Schematic illustrating how conformational propensities shape cellular activity using Tat-dependent transactivation as an example. Increasing or decreasing the conformational propensities to form the RNA conformations bound in the active complex results in corresponding increases or decreases in cellular activity.