Supplementary Information

RNA conformational propensities determine cellular activity

Megan L. Ken, Rohit Roy, Ainan Geng, Laura R. Ganser, Akanksha Manghrani, Bryan R. Cullen, Ursula Schulze-Gahmen, Daniel Herschlag, and Hashim M. Al-Hashimi

Correspondence to: Ursula Schulze-Gahmen, ursula.schulzegahmen@gladstone.ucsf.edu; Daniel Herschlag, herschla@stanford.edu; Hashim M. Al-Hashimi, ha2639@cumc.columbia.edu

**Methods**

**Preparation and purification of RNA**

Unlabeled wtTAR and TAR mutants (full sequences listed in Supplementary Table 7) for NMR experiments and *in vitro* displacement assays were synthesized with the MerMade 6 DNA/RNA synthesizer (Bioautomation) using standard phosphoramidite chemistry and 2'-hydroxyl deprotection protocols. Samples were purified using 20% (w/v) denaturing polyacrylamide gel electrophoresis (PAGE) with 8M urea and 1X TBE. RNA was excised and then electroeluted (Whatman, GE Healthcare) in 1X TAE buffer. Eluted RNA was concentrated and ethanol precipitated. RNA was then dissolved in water to a concentration of ~50 μM and annealed by heating at 95°C for five minutes and cooling on ice for 1 hour. For NMR experiments, RNA constructs were buffer exchanged using centrifugal concentration (3 kDa molecular weight cutoff, EMD Millipore) into NMR buffer consisting of 15 mM NaH2PO4/Na2HPO4, 25 mM NaCl, 0.1 mM EDTA, 10% (v/v) D2O at pH 6.4. For *in vitro* assays, RNA constructs were again buffer exchanged using centrifugal concentration into Tris-HCl assay buffer consisting of 50 mM Tris-HCl, 100 mM NaCl, 0.01% (v/v) Triton X-100 at pH 7.4.

**NMR Experiments**

All NMR experiments were performed at 5°C or 25°C, on a Bruker Avance III 700 MHz spectrometer equipped with triple resonance HCN cryogenic probes. Natural abundance RNA samples were exchanged into NMR buffer consisting of 15 mM NaH2PO4/Na2HPO4, 25 mM NaCl, 0.1 mM EDTA, 10% (v/v) D2O at pH 6.4. Final concentrations ranged from 0.4 to 1.0 mM NMR spectra were processed with NMRPipe35 and visualized with SPARKY (version 3.115)36.

**Chemical shift perturbation (CSP) measurement of stacking energetics**

Stacking populations were calculated from the observed chemical shifts (δobs) of U23-C6 and A22-C8 for all mutants, measured from 2D HMQC spectra (Extended Data Fig. 2), as described previously18. Briefly, δobs for a given variant represents a population-weighted average of the fractional populations of the stacked (S) and kinked (K) states, such that:

δobs = *p*stackδstack + *p*kinkδkink

In which the contribution of the base-triple state [T] is assumed to be negligible because [T]<<[K] or [S]. Thus,

$p\_{stack}=\frac{\left[S\right]}{\left[K\right]+\left[S\right]+[T]}$~ $\frac{\left[S\right]}{\left[K\right]+\left[S\right]}$

The values of δstack (U23-C6) = 143.81 ppm and δstack (A22-C8 ) = 140.19 ppm were obtained from a prior study18 based on Mg2+ titrations that alter the extent of stacking. The value of δkink (A22-C8) = 139.00 ppm was obtained from a prior study based on low-salt conditions37. The value of δkink (U23-C6)= 141.50 ppm was obtained based on the most upfield shifted resonance in our set of mutants. The *p*stack(A22) and *p*stack(U23) values, which agreed within 2-fold for most variants, were averaged to obtain *p*stack for each mutant (Extended Data Fig. 1) and used to determine the ΔGpenalty,stack = -RTln*p*stack

**TAR-Tat-ARM peptide binding assay**

The fluorescence-based binding assay employed chemically synthesized unlabeled RNA constructs (*wt* and mutants) prepared in-house and a peptide mimic (Genscript) of the Tat RNA binding domain N-AAARKKRRQRRR-C containing the arginine rich motif (ARM), an N-terminal fluorescein label, and a C-terminal Carboxytetramethylrhodamine (TAMRA) label. The peptide is highly flexible when free in solution, allowing the two terminal fluorophores to interact and quench the fluorescent signal32. Upon binding to TAR the peptide becomes structured, the two fluorophores are held apart, and fluorescence resonance energy transfers from fluorescein to TAMRA. In this assay, RNA was plated in 1:3 serial dilutions in a 384-well plate and a constant concentration of 20 nM Tat-ARM peptide was used for initial experiments. Both RNA and Tat-ARM peptide were diluted in assay buffer consisting of 50 mM Tris-HCl, 100 mM NaCl, 0.01% (v/v) Triton X-100 at pH 7.4. The plate was left to incubate in the dark for 15 minutes before reading. Fluorescence was measured in triplicate with a CLARIOstar plate reader (BMG Labtech) using a 485 nm excitation wavelength and 590 nm emission wavelength. This assay was repeated five times with distinct samples for each RNA mutant (n = 5).

To ensure the assay accurately reports dissociation constants, the following controls were performed according to Jarmoskaite *et al*, 202038. We first repeated the assay for the tightest binders, *wt* and U2, with varying concentrations of the constant component (Tat-ARM peptide) from 2 nM to 100 nM (Extended Data Fig. 3c). Over this 50-fold change in the constant component, we observed at most 2-fold changes in the dissociation constants. These assays were done in triplicate, 1-3 times for each condition. This result provides strong assurance that our assay accurately relates binding constants (is in the so-called “binding regime”38). We then tested that the binding reaction reaches equilibrium during the 15-minute incubation. With *wt* TAR, we varied the incubation time from 5 minutes to 2 hours and observed no change in the *K*d (Extended Data Fig. 3d). This was done one time in triplicate, with the same assay plate being read at each timepoint. This constancy provides strong evidence that the reaction has reached equilibrium even at the shortest timepoint for the strongest binder. Baseline fluorescence shifts with increasing incubation times likely result from photobleaching that occurs with each reading of the assay.

The averages and standard deviations of the *K*d values for all variants are reported in Extended Data Fig. 3a and Supplementary Table 3. The. *K*d values were obtained by fitting the binding curves to equation 1 using GraphPad Prism (version 9.3.1),

$A=A\_{free}+(A\_{bound}-A\_{free})(\frac{\left[TAR\right]+\left[TatARM\right]+K\_{d}-\sqrt{\left(\left[TAR\right]+\left[TatARM\right]+K\_{d}\right)^{2}-4\left[TAR\right]\left[TatARM\right]}}{2\left[TatARM\right]}) $(1)

where A is the measured fluorescence; Afree is the fluorescence in the absence of TAR-Tat-ARM binding; Abound is the fluorescence at saturated TAR-Tat-ARM binding; *K*d is the measured binding affinity and [TAR] and [TatARM] are the concentrations of TAR and the peptide, respectively.

**Plasmid construction**

pcTat39, pFLuc-TAR33, pcRLuc33, and pBC12-CMV39 expression plasmids were constructed as described previously. The pFLuc-TAR plasmid was modified to increase expression of FLuc and introduce a BsrgI restriction site for generating TAR mutants. To construct many mutants by inserting pre-annealed oligos with the mutated sequences, we required two restriction sites within 100 bps of each other surrounding the TAR sequence. A HindIII site existed 3’ of TAR, and we then modified the plasmid by making a single base change immediately 5’ of TAR to create a BsrgI site. This was done using two overlapping primers with the desired mutation (Supplementary Table 7). We then eliminated the native BsrgI site in the FLuc gene with a single silent mutation using the same strategy (Supplementary Table 7). This modified plasmid, termed pFLuc-TAR-140, is the wtTAR plasmid in this study. The pFLuc-TAR-140 plasmid was then modified to prepare all the mutants in this study. TAR mutants were introduced at BsrgI and HindIII restriction sites using annealed DNA oligos (Supplementary Table 7) that were designed to have overhangs complimentary to BsrgI and HindIII.

**TAR-Tat dependent trans-activation assay**

HeLa cells were maintained in Dulbecco’s modified Eagle medium (DMEM) supplemented with 10% fetal bovine serum and 0.1% gentamicin at 37°C and 5% CO2. Cells were plated to 1.5 x 105 cells per well in 12-well plates 24 hours prior to transfection with polyethylenimine PEI (Polysciences). The primary transfection mixtures contained 125 ng pFLuc-TAR-140 reporter plasmid, 10 ng RLuc control plasmid, +/– 20 ng pcTat expression plasmid, and pBC12-CMV filler DNA plasmid up to a total of 1385 ng total DNA per well. Media was replaced at 24 hours post-transfection. Cells were lysed at 48 hours post-transfection with 250 μL passive lysis buffer (Promega) and incubated 20 minutes at room temperature. FLuc and RLuc activity was measured using a Dual-Luciferase Reporter Assay System (Promega). Assay was repeated in duplicate five times for all mutants on independent days (n = 5).

**Protein Expression**

We co-expressed P-TEFb, composed of human CDK9 1–330 and human cyclin T1 1–264, with  HIV-1 Tat 1–57 in High5 insect cells using recombinant baculovirus infections. Baculovirus generation and High5 cell infections were described in detail previously40. AFF4 fragments 2–73  with an N-terminal TEV-protease-cleavable His-tag were expressed in E. coli40. Note that Tat 1-57 contains all conserved domains required for TAR binding and transcriptional activation. We avoid only working with full-length Tat because *in vivo* Tat binds TAR as part of the Tat:SEC complex and because on its own full-length Tat is natively unfolded and extremely difficult to prepare and purify for biophysical studies41.

**TAR-Tat:SEC electrophoretic mobility shift assay**

Synthetic TAR RNA was resuspended at 0.1 mg/ml in 20 mM Na HEPES pH 7.3, 100 mM KCl, 3 mM MgCl2 and annealed by heating the RNA at 75 °C for 2 min, followed by rapid cooling on ice. Refolded synthetic TAR (nucleotides 17-45) was radioactively labeled with 32P--ATP using T4-polynucleotide kinase. A 10 µl reaction was prepared with 200 nM TAR, 0.3 mCi 32P--ATP (7000 Ci/mmol, MP Biomedicals, Sohon, OH), and 10 units of T4-polynucleotide kinase (New England BioLabs, Ipswich, MA) in 70 mM Tris/HCl pH 7.6, 10 mM MgCl2, 2 mM DTT. After incubating at 37 °C for 1 hour, 25 µl of annealing buffer (20 mM Na HEPES pH 7.3, 100 mM KCl, 3 mM MgCl2) were added to the reaction. The mixture was purified twice over Illustra G25 spin columns (GE Healthcare, Piscataway, NJ) to remove free nucleotides. The purified labeled TAR was diluted to 10 nM (3000-5000 cpm/ µl) with annealing buffer for storage and use in EMSAs.

Binding reactions (10 μl) were carried out in 20 mM Na HEPES pH 7.3, 100 mM KCl, 3 mM MgCl2, 1 mM DTT, 4% glycerol with 12 units RNasin (Promega, Madison, WI), 10 µg/ml BSA, and 5 µg/ml poly(I:C) (Invitrogen, San Diego, CA). Each reaction contained 50-100 pM labeled TAR RNA and a concentration of Tat:SEC protein complex ranging from 0.004 nM to 31 nM. Reactions were incubated at 20 °C for 30 mins and RNA-binding complexes were separated on a pre-run 6% polyacrylamide gel in 0.5x TBE (100 V, 1 hr at 4 °C). Gels were dried, exposed to storage phosphor screens for 12-24 hrs, and imaged on a Typhoon FLA 9000 phosphorimager (GE Healthcare, Piscataway, NJ). The intensity of bands for unshifted and shifted TAR were measured using the program ImageJ (version 1.52)42 and the fraction of shifted TAR calculated by dividing the intensity of the shifted TAR by the intensity of the unshifted TAR in the negative control without SEC.

Each EMSA was repeated two to three times on independent days (n = 2-3) and analyzed with GraphPad Prism (version 9.3.1), fitting the EMSA data to equation 2 to calculate apparent *K*d values (Extended Data Fig. 7, Supplementary Table 6):

$A=A\_{free}+(A\_{bound}-A\_{free})(\frac{\left[Tat:SEC\right]+\left[TAR\right]+K\_{d}-\sqrt{\left(\left[Tat:SEC\right]+\left[TAR\right]+K\_{d}\right)^{2}-4\left[Tat:SEC\right]\left[TAR\right]}}{2\left[TAR\right]}) $(2)

where A is the measured intensity; Afree is the intensity in the absence of Tat:SEC binding; Abound is the background corrected phosphor intensity at saturated TAR-Tat:SEC binding; *K*d is the measured, semi-quantitative binding affinity, and [TAR] and [Tat:SEC] are the concentrations of TAR and the Tat:SEC (Tat:AFF4:P-TEFb) complex, respectively.

The values obtained are referred to as apparent affinities and give apparent free energy values because we were not able to rigorously test whether the tightest binders were subject to titration distortions38.

**Calculating *p*stack and interhelical Euler angles for FARFAR-NMR TAR ensembles**

The ensemble models for *wt* and variants U2 and U7, consisting of N = 2000 conformations, used in this study were described previously11. In this study, we used these previously derived FARFAR-NMR ensemble models to determine the *p*stack value for *wt* and U7. Coaxial stacking of the helices was determined using X3DNA-DSSR (Dissecting the Spatial Structure of RNA) (version 1.9.9)43. DSSR identifies helices as contiguous segments of base-pairs with stacking interactions, regardless of base-pair geometry and backbone connectivity. Conformations in which both upper and lower stems of TAR were identified as part of the same helical stack were assigned to be stacked, all other conformations were assigned to be kinked. The percentage of the conformations in the FARFAR-NMR generated ensembles that were assigned as stacked determined the stacked population (*p*stack) value for *wt* and U7.

To investigate the deviation of the U2 variant from the model predictions (Supplementary Discussion 1) we analyzed the previously determined *wt* and U2 FARFAR-NMR ensembles11 with DSSR (version 1.9)43. We used DSSR to assign individual conformations in each ensemble model as stacked or kinked (as above), and as base-triple like or not based on hydrogen bonding between U23 and A27.

**Equations relating ΔGconf and ΔGpep**

The binding reaction between TAR and Tat (or Tat:SEC) is given by:

TAR + Tat ⇌ TAR-Tat

The binding association constant is given by:

$$K\_{A}=\frac{[TAR-Tat]}{[TAR][Tat]}$$

The TAR conformational ensemble includes kinked (K), stacked (S); and base-triple (T) conformational states (Fig 1a). The equilibrium concentration of unbound TAR is given by the sum over the three conformational states:

[TAR] = [K]+ [S] + [T]

Assuming only the base-triple conformational state T binds Tat, the apparent $K\_{A} $is given by:

$$K\_{A}=\frac{[T-Tat]}{(\left[K\right]+\left[S\right]+[T]) [Tat]}$$

Multiplying by [T]/[T]:

$K\_{A}=\frac{[T-Tat][T]}{(\left[K\right]+\left[S\right]+[T]) [Tat][T]}=\frac{[T]}{\left[K\right]+\left[S\right]+[T]}×\frac{[T-Tat]}{\left[T\right][Tat]} =$ *p*triple x *K*bind

In which *K*bind is the association constant for binding to the pre-formed base-triple (T) state (Fig 1A). *p*triple is the equilibrium population of the base-triple conformational state:

*p*triple $= \frac{[T]}{\left[K\right]+\left[S\right]+[T]}$

Therefore, relative to binding to a preformed conformational state, the association constant is reduced by an amount given by the equilibrium population of the bound species in the unbound conformational ensemble.

K, S, and T are in dynamic equilibrium (Fig 1a):

K⇌S⇌T

*p*triple can be expressed in terms of the population of the stacked state *p*stack and the equilibrium constant ($K\_{triple}$ = $\frac{[T]}{\left[S\right]}$ ) for forming the base-triple state (T) starting from the stacked state (S):

*p*triple = $ \frac{[T]}{\left[K\right]+\left[S\right]+[T]}=\frac{[S]}{\left[K\right]+\left[S\right]+[T]}×\frac{[T]}{\left[S\right]} $= *p*stack x *K*triple

$K\_{A}$ is given by:

*K*A = *p*stack x *K*triple x *K*bind

The overall free energy of binding is given by:

ΔG = -RTln*K*A

= -RTln(*p*stack x *K*triple x *K*bind)

= -RTln*p*stack -RTln*K*triple -RTln*K*bind

= ΔGpenalty,stack +ΔGtriple +ΔGbind

And

ΔGconf = ΔGpenalty,stack + ΔGtriple

ΔGpenalty,stack = -RTln*p*stack is the free energy cost needed to redistribute the unbound TAR conformational ensemble from a starting 3-state conformational ensemble (K, S, and T) to an ensemble in which the stacked S state is 100% populated1. Because it is an energetic penalty the value is always positive (ΔGpenalty,stack ≥ 0) unless *p*stack = 1 in which case ΔGpenalty,stack = 0. Note that even if the base-triple conformational state T were to become the dominant state, making $∆G\_{triple}$ more favorable does not result in a correspondingly smaller ΔGconf penalty and overall higher binding affinity ($K\_{A}$) as ΔGpenalty,stack becomes less favorable due to a concomitant reduction in *p*stack as given by the relationship:

*K*A = *p*stack x *K*triple x *K*bind

**Calculation of ΔΔGcell**

The gene-reporter assay with a TAR variant (i) quantifies the amount of FLuc mRNA produced during a time interval (t) based on the luminescence signal (*F*(i)) due to the translated FLuc protein:

*F* (i) = *c* x [FLuc mRNA](i)

in which *c* is a proportionality constant relating the luminescence signal *F*(i) and the concentration of FLuc mRNA. The FLuc mRNA is produced from Tat-dependent transactivation during which Tat:SEC binds to TAR with association constant *K*bind,cell, triggering a cascade of phosphorylation events mediated by the Tat-activated P-TEFb, which modifies positive and negative cellular elongation factors, and results in efficient transcription elongation17. This multi-step reaction can be reduced to the following two-step reaction mechanism:

TAR + Tat:SEC $⇔$ TAR – Tat:SEC $ →$ FLuc mRNA

in which *k*net is a kinetic rate constant subsuming all the steps following Tat:SEC binding to TAR leading to FLuc transcription, which is assumed to be equal for all TAR variants. The rate of FLuc mRNA production as a function of time is given by

$\frac{d\left[FLuc mRNA\right]}{dt}$ = *k*net[TAR – Tat:SEC](t),

in which [TAR-Tat:SEC](t) is the time-dependent concentration of TAR-Tat:SEC. Assuming that binding is not rate limiting, and *k*net is much slower than the rate of TAR-Tat:SEC dissociation (*k*net << *k*-1), we can apply the pre-equilibrium approximation, and obtain an expression for [TAR-Tat:SEC] in terms of *K*bind,cell:

[TAR – Tat:SEC](t) = *K*bind,cell [TAR]free (t) [Tat:SEC]free (t)

$\frac{d\left[FLuc mRNA\right]}{dt}$ *= k*net*K*bind,cell [TAR]free (t) [Tat:SEC]free (t)

When the total concentrations of [TAR]total and [Tat:SEC]total are much lower than the binding dissociation constant i.e. [TAR]total and [Tat:SEC]total $\ll \frac{1}{K\_{bind,cell}}$, the free concentrations of [TAR]free and [Tat:SEC]free is approximately equal to their total concentrations:

[TAR]free (t) ≈ [TAR]total

[Tat:SEC]free (t) ≈ [Tat:SEC]total

$\frac{d\left[FLuc mRNA\right]}{dt}$ *= k*net*K*bind,cell [TAR]total [Tat:SEC]total

Integrating the above differential equation with respect to time, we obtain the desired expression relating the amount of FLuc mRNA produced over a time interval τ and *K*bind,cell:

*F* (i) = *c* x [FLuc mRNA](i) = *c* x *k*net(i)*K*bind,cell (i) [TAR]total [Tat:SEC]total τ

Because only the TAR sequence is altered when comparing results across TAR variants, and since we can control for variations in *k*net, [TAR]total, and [Tat:SEC]total from one TAR variant to another (see below), the ratio of the FLuc signal (*F*(i)) measured for a mutant (i) relative to *wt* *F*(*wt*) will be equal to the fold difference in *K*bind,cell:

*F* (i) = *c* x [FLuc mRNA] (τ) (i) = *c* x *k*net (i) *K*bind,cell (i) [TAR]total [Tat:SEC]total τ

*F* (*wt*) = *c* x [FLuc mRNA] (τ) (*wt*) = *c* x *k*net *K*bind,cell (*wt*) [TAR]total [Tat:SEC]total τ

$\frac{F(i)}{F(wt)}=\frac{\left[FLuc mRNA\right](τ)(i)}{\left[FLuc mRNA\right](τ)(wt)}=\frac{K\_{bind}(i)}{K\_{bind}(wt)}\_{ }$

To control for variations in [TAR]total and [Tat:SEC]total due to differences in transfection efficiencies of the DNA plasmids from experiment to experiment, we normalized the FLuc signal by the signal (*R*(i)) obtained for a CMV-driven *Renilla* luciferase (RLuc) reporter plasmid, which is not under the control of the TAR promoter,providing a measure of transfection efficiency and therefore [TAR]total and [Tat:SEC]total:

*R*(i) = *c*’*k*’net τ,

in which c’ is the proportionality constant relating RLuc luminescence to RLuc mRNA, and *k’*net is the net rate of RLuc mRNA production. Provided that the ratio *k*net / *k’*net is constant across TAR variants, the ratio of the RLuc normalized FLuc signal for a mutant (*i*) relative to *wt* will be equal to the fold difference in *K*bind,cell:

$\frac{c k\_{net}(i)}{c^{'}k\_{net}^{'}(i)}=\frac{c k\_{net}(wt)}{c^{'}k\_{net}^{'}(wt)}$

$\frac{F(i)/R(i)}{F(wt)/R(wt)}=\frac{K\_{bind,cell }(i)}{K\_{bind,cell}(wt)}\_{ }$

To measure the contribution to *K*bind,cell due to the interaction between TAR and Tat component of the Tat:SEC complex, as well as control for variations in *k*net and *k*net’ arising due to differences in the metabolic state of the cells at the time of harvesting, we normalized the FLuc/RLuc signal by the corresponding signal obtained in the absence of Tat:

$\frac{F(+Tat)}{R(+Tat)}=\frac{c x \left[FLuc mRNA\right]\left(τ\right)\left(+Tat\right)}{c^{'}x \left[RLuc mRNA\right]\left(τ\right)\left(+Tat\right)}=\frac{c x k\_{net}K\_{bind,cell}(+Tat)}{ c^{'}x k'\_{net}}$

$\frac{F(-Tat)}{R(-Tat)}=\frac{c x \left[FLuc mRNA\right]\left(τ\right)\left(-Tat\right)}{c^{'}x \left[RLuc mRNA\right]\left(τ\right)\left(-Tat\right)}=\frac{c x k\_{net}K\_{bind,cell}(-Tat)}{ c^{'}x k^{'}\_{net}}$

$\frac{F(+Tat)/R(+Tat)}{F\left(-Tat\right)/R(-Tat)}=\frac{K\_{bind,cell}(+Tat)}{K\_{bind,cell}(-Tat)}$

in which *K*bind,cell(+Tat) and *K*bind,cell(-Tat) are the equilibrium association constants for binding of Tat:SEC and SEC only to TAR in cells, respectively:

*K*bind,cell (+Tat) = $\frac{[TAR-Tat:SEC]}{[TAR]\_{total} [Tat:SEC]\_{total}}$

*K*bind,cell (-Tat) = $\frac{[TAR-SEC]}{[TAR]\_{total} [SEC]\_{total}}$

ΔGcell+Tat = -RTln(*K*bind(+Tat))

ΔGcell-Tat = -RTln(*K*bind(-Tat))

Based on our additivity model, we can decompose the energetics of Tat:SEC binding to TAR into energetic contributions arising from interactions between TAR and Tat (ΔGTat) and between TAR other components of SEC (ΔGSEC). The binding energetics with and without Tat are given by:

ΔGcell(+Tat) = ΔGSEC + ΔGTat,cell

ΔGcell(-Tat) = ΔGSEC

The Tat contribution to the overall TAR-Tat:SEC binding energetics, which we define as the Tat-dependent cellular transactivation (ΔGcell) is then given by:

ΔGcell = ΔGcell+Tat - ΔGcell-Tat = ΔGTat,cell

ΔGcell = ΔGcell+Tat - ΔGcell-Tat = -RTln$ $($\frac{K\_{bind,cell}(Tat:SEC)(+Tat)}{K\_{bind,cell}(SEC)(-Tat)})=$ -RTln$\left\{\frac{F(+Tat)/R(+Tat)}{F\left(-Tat\right)/R(-Tat)}\right\}$

To ensure [TAR]total and [Tat:SEC]total $\ll \frac{1}{K\_{bind,cell}}$, the concentration of the Tat plasmid was empirically adjusted to ensure that we (i) observe differences between *wt* and the mutants33 and (ii) fall in the linear proportionality regime in which increasing the Tat plasmid over a 10,000-fold range leads to a corresponding linear increase in the FLuc signal (Extended Fig. 6c-d).

**Statistical comparisons of observed *vs.* predicted ΔΔG values**

Statistical analysis was done using GraphPad Prism (version 9.3.1). For all ΔΔG dataset comparisons, three statistical analyses were performed: (1) a Pearson correlation (r) and associated *p* value was computed using a two-tailed t-test; (2) a linear regression to the line of best fit, accounting for the number of replicates and standard deviations, was performed, and the 95% confidence intervals of the best fit slope and y-intercept were calculated; and (3) a least squares regression was used to fit the comparisons to our model with a fixed slope of 1 and fixed y-intercept of 0, which output R2 and RMSE values.

**Figure design**

All figures were designed in Adobe Illustrator 2020 (version 24.1). Images originally created from other sources were exported to Adobe Illustrator to be incorporated into final figures. Bar graphs were plotted in GraphPad Prism (version 9.3.1). Correlation plots were plotted using Python3 (version 3.7.3). Chemical structures were created using MarvinSuite version 19.21.0 (2019) from ChemAxon (https://www.chemaxon.com). Structural models of RNA and proteins were formatted in MacPyMOL (version 1.5.0.4). NMR spectra were processed using NMRPipe (version 10.8)35, visualized using Sparky (version 3.115)36, and then exported to Adobe Illustrator.

**Methods References**

35. Delaglio, F. *et al.* NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).

36. Goddard, T. D. & Kneller, D. G. Sparky—NMR assignment and integration software. *Univ. California, San Fr.* (2006).

37. Casiano-Negroni, A., Sun, X. & Al-Hashimi, H. M. Probing Na+-induced changes in the HIV-1 TAR conformational dynamics using NMR residual dipolar couplings: New insights into the role of counterions and electrostatic interactions in adaptive recognition. *Biochemistry* **46**, 6525–6535 (2007).

38. Jarmoskaite, I., AlSadhan, I., Vaidyanathan, P. P. & Herschlag, D. How to measure and evaluate binding affinities. *Elife* **9**, e57254 (2020).

39. Tiley, L. S., Madore, S. J., Malim, M. H. & Cullen, B. R. The VP16 transcription activation domain is functional when targeted to a promoter-proximal RNA sequence. *Genes Dev.* **6**, 2077–2087 (1992).

40. Schulze-Gahmen, U. *et al.* The AFF4 scaffold binds human P-TEFb adjacent to HIV Tat. *Elife* **2,** e00327 (2013).

41. Shojania, S. & O’Neil, J. D. HIV-1 Tat is a natively unfolded protein: The solution conformation and dynamics of reduced HIV-1 Tat-(1-72) by NMR spectroscopy. *J. Biol. Chem.* **281**, 8347–8356 (2006).

42. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of Image Analysis. *Nat. Methods* **9**, 671 (2012).

43. Lu, X. J., Bussemaker, H. J. & Olson, W. K. DSSR: An integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.* **43**, e142 (2015).

44. Sathyamoorthy, B., Lee, J., Kimsey, I., Ganser, L. R. & Al-Hashimi, H. Development and application of aromatic [13C, 1H] SOFAST-HMQC NMR experiment for nucleic acids. *J. Biomol. NMR* **60**, 77–83 (2014).

**Supplementary Discussion 1: TAR variant U7 deviates from the model**

The U7 variant deviated from the model by showing a difference in relative free energy change between the Tat-ARM peptide binding assay (Fig. 2d) and the cell-based assay (Fig. 3b). In the peptide assay, U7 binds with higher affinity than U6, breaking the trend of decreasing affinity with bulge length. However, in the cell-based assay and the SEC-binding assay, U7 had a lower level of transactivation relative to U6, maintaining the trend of decreased activity with increasing bulge length. As noted in the main text, there is NMR evidence to suggest that the U7 uridine-rich bulge may be forming an additional helix consisting of U-U mismatches that improves stacking and does not sterically hinder peptide binding. Our observation of several imino resonances in the region expected for U-U mismatches supports this hypothesis (Extended Data Fig. 2). However, in the Tat-SEC complex this stacking benefit may be counteracted by steric interactions, which could account for the decrease in affinity of U7 relative to U6 in the Tat:SEC binding assay and cell-based assay. Indeed, structural modelling revealed steric overlap between the bulge and Cyclin T1 for U7 variant that are not present in models for *wt* or U2 TAR (Extended Data Figure 8).

**Supplementary Discussion 2: Effect of base-triple destabilizing mutations on Tat-ARM peptide binding and cellular transactivation.**

A model that can account for the similar binding affinities measured for Tat-ARM to the base-triple destabilizing TAR mutants is that Tat-ARM binds these TAR mutants in a kinked rather than the stacked conformation (Extended Data Fig. 5a-b). A prior study provided evidence that the Tat-ARM ligand-mimic argininamide can bind TAR in a kinked conformation7. If the conformational penalty due to the base-destabilizing mutation (ΔGtriple) exceeds the differences between how well Tat-ARM binds the stacked versus kinked conformation (ΔΔG = ΔGbind,stack – ΔGbind,kinked), then the Tat-ARM peptide would predominantly bind to TAR in the kinked conformation (Extended Data Fig. 5b). Moreover, since the kinked state predominates for all bulge lengths, binding to the kinked conformation would be independent of bulge-length, as observed in Fig. 2d. This model is supported by the lower maximum fluorescence value with saturating Tat-ARM peptide for all the bulge mutants (Extended Data Fig. 5a). Because the fluorescence value is dependent on the orientation of the C and N terminal ends of the peptide relative to one another, this difference suggests that those variants with a lower maximum fluorescence are binding the peptide in a different conformation in which the peptide termini are closer together than when in the base-triple formation.

In contrast, the behavior observed in the cellular context is as predicted from our thermodynamic model which assumes that Tat:SEC binds to the TAR variants predominantly in a stacked base-triple like conformation so that the differences between the binding energetics for the *wt* and disrupted base triple variants is a constant (*c*triple) across all bulge lengths. In this case, it appears that the conformational penalty due to the base-destabilizing mutation (ΔGtriple) does not exceed the difference between how well Tat:SEC binds the stacked versus kinked conformation (ΔΔG = ΔGbind,stack – ΔGbind,kinked). Tat:SEC could have a stronger preference for binding the stacked versus kinked conformation because a stacked conformation may be required to properly orient the TAR apical loop so that it can form key contacts with the cyclin-T1 component of Tat:SEC15, which is absent from Tat-ARM (Extended Data Fig. 5b-c). One possibility is that Tat:SEC binds to the base-triple disrupting mutants in a stacked TAR conformation in which U23 forms a U23•U27-A38 base-triple-like conformation with U23 forming a single hydrogen bond with U27 but otherwise retaining the overall *wt* TAR conformation in the quinary complex (Extended Data Fig. 5d).

**Supplementary Table 1. Comparison of the interhelical stacked state population (*p*stack) determined using CSP and FARFAR-NMR.** *p*stack was determined using these two methods for the *wt* and U7 FARFAR-NMR ensembles11.

|  |  |  |
| --- | --- | --- |
| **RNA**  | ***p*stack CSP** | ***p*stack** **FARFAR-NMR** |
| *wt* | 0.31 | 0.43 |
| U7 | 0.22 | 0.26 |

**Supplementary Table 2. Measuring stacking energetics using CSPs.** Shown for each TAR variant are the measured chemical shifts (δ) for U23-C6 and A22-C8, the computed population of the stacked state (*p*stack), and the difference in stacking energetics ΔΔGpenalty,stack referenced to *wt*.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RNA**  | **δU23-C6 (ppm)** | **δA22-C8 (ppm)** | ***p*stack (%)** | **ΔΔGpenalty,stack (*j-wt*) (kcal/mol)** |
| *wt* | 141.94 | 139.51 | 0.31 | (0) |
| U0 | n/a | 139.07 | n/a | n/a |
| U1 | 143.24 | 139.90 | 0.75 | -0.53 |
| U2 | 143.27 | 139.87 | 0.75 | -0.52 |
| U3 | 142.58 | 139.46 | 0.43 | -0.19 |
| U4 | 141.85 | 139.38 | 0.23 | 0.17 |
| U5 | 141.71 | 139.32 | 0.18 | 0.32 |
| U6 | 141.64 | 139.33 | 0.17 | 0.36 |
| U7 | 141.74 | 139.40 | 0.22 | 0.20 |
| *wt*-A27U/U38A | 141.99 | 139.56 | 0.34 | -0.06 |
| U0-A27U/U38A | n/a | 139.20 | n/a | n/a |
| U1-A27U/U38A | 143.35 | 139.94 | 0.79 | -0.56 |
| U2-A27U/U38A | 143.27 | 140.01 | 0.81 | -0.57 |
| U3-A27U/U38A | 142.50 | 139.43 | 0.40 | -0.15 |
| U4-A27U/U38A | 141.98 | 139.48 | 0.30 | 0.01 |
| U5-A27U/U38A | 141.69 | 139.35 | 0.19 | 0.30 |
| U6-A27U/U38A | 141.68 | 139.33 | 0.18 | 0.33 |
| U7-A27U/U38A | 141.70 | 139.37 | 0.20 | 0.26 |
| *wt*-A27deazaN7 | 141.72 | 139.44 | 0.23 | 0.17 |
| U0-A27deazaN7 | n/a | 139.04 | n/a | n/a |
| U1-A27deazaN7 | 143.33 | 139.91 | 0.78 | -0.55 |
| U2-A27deazaN7 | 143.15 | 140.01 | 0.78 | -0.55 |
| U3-A27deazaN7 | 142.28 | 139.51 | 0.38 | -0.12 |
| U4-A27deazaN7 | 141.71 | 139.34 | 0.19 | 0.30 |
| U5-A27deazaN7 | 141.54 | 139.26 | 0.12 | 0.57 |
| U6-A27deazaN7 | 141.55 | 139.28 | 0.13 | 0.52 |
| U7-A27deazaN7 | 141.59 | 139.32 | 0.15 | 0.42 |

**Supplementary Table 3. Measured *K*d and calculated ΔGpep and ΔΔGpep(*j*-*wt*) values.** Shown for each TAR variant are the measured *K*d values and standard deviations for binding to the Tat-ARM peptide, the corresponding ΔGpep (T= 298.15°K) and standard deviations, and the differences in binding energetics ΔΔGpep referenced to *wt*.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **RNA**  | **Mean *K*d (nM)** | **Std. Dev. (nM)** | **Mean ΔGpep (kcal/mol)** | **Std. Dev. (kcal/mol)** | **Mean ΔΔGpep(*j*-*wt*) (kcal/mol)** | **Std. Dev. (kcal/mol)** |
| *wt* | 42 | 10 | -9.9 | 0.13 | (0) | (0) |
| U0 | 1100 | 230 | -8.0 | 0.13 | 1.9 | 0.23 |
| U1 | 320 | 50 | -8.7 | 0.10 | 1.7 | 0.17 |
| U2 | 14 | 3 | -10.5 | 0.14 | -0.66 | 0.051 |
| U3 | 33 | 7 | -10.0 | 0.12 | -0.15 | 0.14 |
| U4 | 46 | 2 | -9.8 | 0.026 | 0.056 | 0.14 |
| U5 | 72 | 12 | -9.6 | 0.11 | 0.31 | 0.12 |
| U6 | 100 | 19 | -9.3 | 0.11 | 0.52 | 0.23 |
| U7 | 82 | 15 | -9.5 | 0.11 | 0.39 | 0.15 |
| *wt*-A27U/U38A | 420 | 130 | -8.6 | 0.20 | 1.3 | 0.16 |
| U0-A27U/U38A | 3200 | 1400 | -7.4 | 0.35 | 2.5 | 0.31 |
| U1-A27U/U38A | 610 | 110 | -8.3 | 0.10 | 1.6 | 0.12 |
| U2-A27U/U38A | 300 | 30 | -8.7 | 0.065 | 1.1 | 0.15 |
| U3-A27U/U38A | 380 | 140 | -8.6 | 0.23 | 1.2 | 0.25 |
| U4-A27U/U38A | 300 | 62 | -8.7 | 0.12 | 1.1 | 0.15 |
| U5-A27U/U38A | 340 | 41 | -8.7 | 0.066 | 1.2 | 0.18 |
| U6-A27U/U38A | 220 | 40 | -8.9 | 0.11 | 0.96 | 0.083 |
| U7-A27U/U38A | 180 | 35 | -9.0 | 0.13 | 0.84 | 0.083 |
| *wt*-A27deazaN7 | 870 | 460 | -8.2 | 0.32 | 1.7 | 0.33 |
| U0-A27deazaN7 | 940 | 330 | -8.1 | 0.18 | 1.8 | 0.22 |
| U1-A27deazaN7 | 630 | 110 | -8.3 | 0.10 | 1.6 | 0.14 |
| U2-A27deazaN7 | 210 | 39 | -8.9 | 0.11 | 0.94 | 0.15 |
| U3-A27deazaN7 | 240 | 160 | -9.0 | 0.50 | 0.86 | 0.43 |
| U4-A27deazaN7 | 340 | 53 | -8.7 | 0.094 | 1.2 | 0.13 |
| U5-A27deazaN7 | 270 | 34 | -8.8 | 0.076 | 1.1 | 0.10 |
| U6-A27deazaN7 | 160 | 36 | -9.1 | 0.14 | 0.77 | 0.24 |
| U7-A27deazaN7 | 110 | 16 | -9.3 | 0.08 | 0.54 | 0.17 |

**Supplementary Table 4. Statistics assessing the quality of the fit for various ΔΔG comparisons.** (Top) Shown are RMSE and R2 values for the fit of each ΔΔG comparison to our thermodynamic model. (Bottom) Shown are the Pearson correlation and p-value (two-tailed), the RMSE and R2 for the best-fit linear regression, along with the parameters of that regression line including, slope, y-intercept, and their 95% confidence intervals.

|  |
| --- |
| **Model Fit** |
| Parameter | ΔΔGpenalty,stack *vs.* ΔΔGpep | ΔΔGpep *vs.* ΔΔGcell |
| RMSE (kcal/mol) | 0.17 | 0.41 |
| R2 | 0.81 | 0.71 |
| **Best Fit** |
| Parameter | ΔΔGpenalty,stack *vs.* ΔΔGpep | ΔΔGpep *vs.* ΔΔGcell |
| RMSE (kcal/mol) | 0.16 | 0.40 |
| R2 | 0.84 | 0.74 |
| Slope | 1.22 | 0.94 |
| Slope 95% CI | 1.03 to 1.41 | 0.77 to 1.11 |
| y-intercept (kcal/mol) | 0.009 | 0.16 |
| y-int 95% CI | -0.05 to 0.07 | 0.02 to 0.29 |
| Pearson correlation | 0.97 | 0.89 |
| p-value | 0.0004 | 0.0013 |

**Supplementary Table 5. Measured luminescence values and calculated Tat-dependent ΔGcell and ΔΔGcell (*j*-*wt*) values:** Shown are the measured FLuc/RLuc luminescence ratios in the presence and absence of the co-transfected Tat plasmid, the calculated Tat-dependent ΔGcell value (using the incubator T = 37°C / 310.15°K) and their standard deviations, as well as the differences in Tat-dependent transactivation energies ΔΔGcell referenced to *wt* and their standard deviations.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **RNA**  | **Mean FLuc/RLuc + Tat (RLU)** | **Mean FLuc/RLuc - Tat (RLU)** | **Mean Tat-dependent ΔGcell (kcal/mol)** | **Std. Dev. (kcal/mol)** | **Mean ΔΔGcell (kcal/mol)** | **Std. Dev. (kcal/mol)** |
| *wt* | 1.0 x 10-1 | 5.9 x 10-4 | -2.9 | 0.68 | (0) | (0) |
| U0 | 8.7 x 10-4 | 3.4 x 10-4 | -0.6 | 0.20 | 2.4 | 0.23 |
| U1 | 5.6 x 10-2 | 6.8 x 10-4 | -2.4 | 0.87 | 0.69 | 0.36 |
| U2 | 7.5 x 10-2 | 5.1 x 10-4 | -3.0 | 0.31 | -0.091 | 0.17 |
| U3 | 8.4 x 10-2 | 6.0 x 10-4 | -3.1 | 0.25 | -0.12 | 0.16 |
| U4 | 1.1 x 10-1 | 6.4 x 10-4 | -3.0 | 0.31 | 0.067 | 0.15 |
| U5 | 9.1 x 10-2 | 5.4 x 10-4 | -3.0 | 0.42 | 0.17 | 0.097 |
| U6 | 4.0 x 10-2 | 5.1 x 10-4 | -2.5 | 0.40 | 0.62 | 0.24 |
| U7 | 2.2 x 10-2 | 5.2 x 10-4 | -2.1 | 0.45 | 1.0 | 0.23 |
| *wt*-A27U/U38A | 2.4 x 10-2 | 6.5 x 10-4 | -2.1 | 0.42 | 1.0 | 0.24 |
| U0-A27U/U38A | 3.7 x 10-3 | 5.8 x 10-4 | -1.1 | 0.22 | 2.0 | 0.28 |
| U1-A27U/U38A | 3.0 x 10-2 | 5.7 x 10-4 | -2.1 | 0.50 | 1.0 | 0.15 |
| U2-A27U/U38A | 4.0 x 10-2 | 7.2 x 10-4 | -2.3 | 0.47 | 0.83 | 0.27 |
| U3-A27U/U38A | 1.6 x 10-2 | 6.1 x 10-4 | -2.0 | 0.24 | 0.94 | 0.21 |
| U4-A27U/U38A | 9.5 x 10-3 | 5.1 x 10-4 | -1.8 | 0.25 | 1.2 | 0.077 |
| U5-A27U/U38A | 4.1 x 10-3 | 4.4 x 10-4 | -1.3 | 0.16 | 1.6 | 0.18 |
| U6-A27U/U38A | 1.8 x 10-3 | 4.5 x 10-4 | -0.85 | 0.094 | 2.1 | 0.19 |
| U7-A27U/U38A | 1.4 x 10-3 | 6.7 x 10-4 | -0.28 | 0.58 | 2.6 | 0.26 |

**Supplementary Table 6. Measured apparent *K*d values and corresponding ΔGprot and ΔΔGprot(*j*-*wt*) values.** Shown for each TAR variant are the apparent measured *K*d values and standard deviations for binding to the Tat:SEC protein complex, the corresponding ΔGprot (T= 298.15°K) and standard deviation, and the differences in binding energetics ΔΔGprot referenced to *wt*.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **RNA**  | **Mean *K*d (nM)** | **Std. Dev. (nM)** | **Mean ΔGprot (kcal/mol)** | **Std. Dev. (kcal/mol)** | **Mean ΔΔGprot (kcal/mol)** | **Std. Dev. (kcal/mol)** |
| *wt* | 0.036 | 0.0076 | -14.0 | 0.13 | (0) | 0.13 |
| U0 | 1.2 | 0.38 | -12.0 | 0.19 | 2.02 | 0.19 |
| U1 | 0.45 | 0.20 | -12.5 | 0.25 | 1.43 | 0.25 |
| U2 | 0.066 | 0.0042 | -13.6 | 0.037 | 0.36 | 0.037 |
| U4 | 0.16 | 0.0082 | -13.1 | 0.030 | 0.87 | 0.030 |
| U6 | 0.10 | 0.022 | -13.3 | 0.12 | 0.62 | 0.12 |
| U7 | 0.20 | 0.062 | -13.0 | 0.18 | 1.00 | 0.18 |

**Supplementary Table 7. RNA sequences and oligonucleotides.** This Excel spreadsheet contains in tab 1 the sequences of all synthesized RNAs in this study; in tab 2 the sequences of the DNA oligo primers used to construct the TAR-FLuc-140 plasmid; in tab 3 the sequences of the inserted oligos used to create the TAR mutant plasmids used in the cellular transactivation assay.