WHAT CAN A SMALL RNA MOLECULE TEACH US ABOUT MOLECULAR SWITCHES? DISTINGUISHING CONTACT AND ENSEMBLE EPISTASIS IN THE ADENINE RIBOSWITCH

by

DARIA WONDERLICK

A THESIS

Presented to the Department of Chemistry & Biochemistry
and the Robert D. Clark Honors College
in partial fulfillment of the requirements for the degree of
Bachelor of Science

June 2021

An Abstract of the Thesis of

Daria Wonderlick for the degree of Bachelor of Science
in the Department of Chemistry & Biochemistry to be taken June 2021

Title: What Can a Small RNA Molecule Teach Us About Molecular Switches? Distinguishing Contact and Ensemble Epistasis in the Adenine Riboswitch

Approved: *Dr. Michael J. Harms*
Primary Thesis Advisor

The ability to engineer biomolecules with new and improved functions could revolutionize medicine, industry, and biotechnology. Efficient bioengineering requires a predictive understanding of the sequence changes needed to produce a desired functional effect. Our work develops a theoretical framework for handling the complications that arise when multiple sequence changes are introduced into a molecular switch. The functional effects of paired mutations can be coupled to each other, and communication between the sites can be mediated by both direct (contact-based) and indirect (ensemble-based) pathways.

We designed an experiment to distinguish between these two coupling mechanisms in the adenine riboswitch, a small RNA molecule that switches between different conformations depending on adenine availability. We measured the coupled effects of paired mutations throughout the riboswitch with respect to binding the fluorescent base analog 2-aminopurine. By perturbing the riboswitch's chemical environment with magnesium ions, we could manipulate the riboswitch's conformations to isolate direct (within-conformation contacts) and indirect (across-conformation ensemble redistribution) coupling components.

While we observed experimental signatures corresponding to both direct and indirect coupling mechanisms, we were unable to successfully tease them apart. This may point to their fundamental interdependence in molecular switches. Efforts to engineer these biomolecules may therefore benefit from predictive models that take both communication pathways into account.

Acknowledgements

First and foremost, I wish to thank my primary thesis advisor and principal investigator Dr. Michael Harms for his mentorship and guidance throughout my four years in the Harms lab. Thank you for challenging and inspiring me to grow both as a person and as a scientist. It has been a joy to work on this project with you.

I want to thank my second reader and graduate student mentor Anneliese Morrison for taking me under her wing as an undergraduate researcher. Thanks to you and the other members of the Harms lab, I looked forward to coming into lab each day. I am grateful to the Department of Chemistry & Biochemistry for fostering a culture that welcomes and celebrates undergraduate research. I am also deeply indebted to Dr. Julia Widom for generously opening her lab to me and training me in proper RNA techniques.

Thank you to Dr. Samantha Hopkins for being my CHC advisor throughout my four years at the University of Oregon and for generously offering to be the CHC representative on my thesis committee. I am blessed to be a part of a Clark Honors College family that stretches me to think and articulate ideas in new ways.

I am grateful to both the Presidential Undergraduate Research Scholars (PURS) program and the Anita and Friedhelm Baitis Summer Fellowship for generously funding this research project.

Finally, I am beyond thankful for my wonderful family and friends who always have words of encouragement at the ready. Thank you for believing in me before I believe in myself.

Table of Contents

Introduction 1

A biomolecule's sequence determines its function 1

Complications from conformational ensembles and epistatic coupling 2

Broad research goals 3

Background I: Thermodynamic modelling of a conformational ensemble 4

Background II: Thermodynamic modelling of coupled sites 7

Background III: Contact and ensemble mechanisms for epistatic coupling 8

Specific research goals 12

Project Design 13

Experimental Hypotheses 14

Results 16

Modelling the effects of mutations on conformations within the ensemble 22

Analysis of epistatic coupling between sites that form a base pair (G-C) 23

Analysis of epistatic coupling between sites in opposite loops (G-U) 29

Analysis of epistatic coupling between sites in distant, linked regions (G-A) 31

Single-conformation contacts describe total epistatic coupling 35

Discussion 38

Contact- and ensemble-based coupling pathways may not be fully separable 38

Model improvement and validation 39

Magnitudes and patterns of contact- and ensemble-based coupling 40

Conclusions 41

Methods 42

Glossary 45

Bibliography 46

List of Figures

Figure 1. Conformational switching in the adenine riboswitch. 5
Figure 2. Three-state adenine riboswitch ensemble. 10
Figure 3. Positions of selected mutations to the adenine riboswitch. 14
Figure 4. Comparison of the four-, three-, two-state, and apparent binding models. 19
Figure 5: Three-state model fits of 2AP binding data for all riboswitch genotypes. 20
Figure 6: Apparent binding fits of 2AP binding data for all riboswitch genotypes. 21
Figure 7. Fractional populations of riboswitch conformations. 23
Figure 8. Epistatic coupling upon pairing mutations G and C. 26
Figure 9. Epistatic coupling upon pairing mutations G and U. 30
Figure 10. Epistatic coupling upon pairing mutations G and A. 33
Figure 11. Contact-based coupling matches total epistasis. 36

Introduction

The molecules that keep us alive are elegantly simple yet extraordinarily complex. Through the actions of DNA, RNA, and proteins—henceforth referred to as biomolecules—all living things maintain cellular order and organization amidst an onslaught of environmental stressors. Life's molecular catalysts, switches, and sensors are often more selective and efficient than their man-made counterparts. If we understand the physical rules that dictate their function, we can build better predictive models for engineering biomolecules that could revolutionize medicine, industry, and biotechnology.

A biomolecule's sequence determines its function

Engineering a biomolecule first requires understanding its blueprint. All biomolecules are constructed by linking chemical groups together like stringing letters together to form a word. Each class of biomolecules has its own alphabet of chemical groups: four deoxyribose nucleotides for DNA, four ribose nucleotides for RNA, and twenty amino acids for proteins. This work uses an RNA molecule called the adenine riboswitch, which can be thought of as a 71-letter chemical word spelled with the chemical letters adenine (A), cytosine (C), guanine (G), and uracil (U). Just as English letters have their own sound, each RNA nucleotide letter has its own chemical properties. Most importantly, adenine forms a base pair with uracil (A-U) while cytosine forms a base pair with guanine (C-G). These base pairs bring non-neighboring letters within an RNA word together, forming two-dimensional loops from the one-dimensional sequence of nucleotide letters. Further base pairing and other interactions between these loops can twist the RNA molecule into a three-dimensional shape.

In the simplest case, a biomolecule's sequence of chemical letters gives the blueprint for a single shape, or structure. This structure allows the molecule to carry out a specific function, like forming a channel in a membrane or interacting with another molecule. One way to alter the function of a biomolecule is to change its sequence with a mutation. If we thoroughly understand the physical rules linking sequence to function, we should be able to predict how a mutation will affect a molecule's function. Such predictive models would allow us to design sequences with specific functions and select the best mutations for improving desirable functions or reducing undesirable ones.

Complications from conformational ensembles and epistatic coupling

There are two main challenges associated with predicting the function of a biomolecule from its sequence or predicting the functional effect of a sequence change. First, a biomolecule's sequence of chemical letters might form a blueprint for multiple structures, or conformations, with different functional roles. In fact, the functions of many important biomolecules involve switching amongst a group, or ensemble, of conformations (Nussinov et al. 2019). For example, signal-processing biomolecules switch between "on" and "off" conformations depending on whether the signal is present. In the case of the adenine riboswitch, the riboswitch primarily adopts an "off" conformation until an adenine molecule nestles into the riboswitch binding pocket and promotes a transition to the "on" conformation (Figure 1, Lemay et al. 2006). This is similar to how the four-letter English word "bear" specifies an ensemble of meanings (a large animal, to carry, to tolerate, to produce, etc…), one of which is selected based on the word's context. The function of a molecular catalyst, switch, or sensor depends both on the structures of each conformation and its ability to transition between them.

Predicting its function from its sequence, therefore, requires the added challenge of understanding its conformational ensemble.

The second challenge to engineering a biomolecule's sequence for a desired function is that sites within a biomolecule can be interconnected. This means that a chemical group at a given position in a biomolecule contributes differently to the molecule's function depending on the identities of chemical groups at the connected positions. These connected positions can be located both adjacent to and distant from the site of interest. This is similar to how the pronunciation of an English letter can depend on the letter that follows it (consider the c's in "cheese" versus "scooter") or on a letter farther away (consider the a's in "scraps" versus "scrape"). In a phenomenon known as epistasis, the functional effect of mutating a position to a new chemical letter depends on the letters at all its connected positions (Starr and Thornton 2016). Such epistastic coupling between interconnected sites makes it difficult to predict how a biomolecule's function will be impacted when multiple mutations are introduced.

Broad research goals

Interconnected sites within a biomolecule complicate efforts to predict how paired sequence changes will impact its function. The multiple structures within conformational ensembles add further complexity to these predictions because each conformation might have its own network of connected sites. As a first step towards overcoming these challenges, this project investigates the mechanisms by which sites are epistatically coupled in a molecular switch. What are the pathways of communication within individual conformations and across the ensemble? Can we tease apart these coupling mechanisms in a real system?

Background I: Thermodynamic modelling of a conformational ensemble

Before we can consider coupling between sites in a molecular switch, we first need to be able to describe the conformations that make up its ensemble. Thermodynamics offers a mathematical language to describe the transitions between these conformations. In thermodynamic terms, transitions depend on energy. A biomolecule randomly fluctuates between all the conformations in its ensemble, but it has a higher probability of being in a conformation with a lower energy. Scaling back to consider a large group of biomolecules that are interconverting between possible conformations, the largest fraction of the group will be in the lowest energy conformation. Low energy conformations are considered energetically stable because the biomolecule spends more time populating them.

We can describe the adenine riboswitch as transitioning between adenine-unbound *U* and adenine-bound *B* conformations (Figure 1). The Boltzmann probability distribution describes the relative probability *(p)* of each conformation as an exponential function of its free energy *G*, the gas constant *R*, and the temperature *T*. The higher the probability, the more favorable is the conformation.

$$p_U = e^{-G_U/RT},\ p_B = e^{-G_B/RT}$$

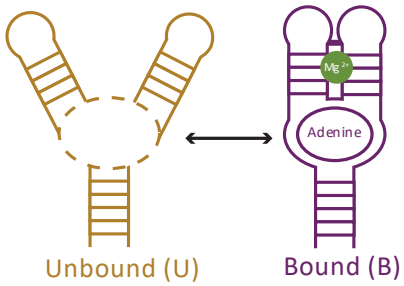

Figure 1. Conformational switching in the adenine riboswitch.

The adenine riboswitch transitions between adenine-unbound (U) and bound (B) conformations. The bound conformation is stabilized by magnesium ions (green).

If we watch a group of riboswitch molecules for a long time (such that they reach thermodynamic equilibrium), we see each riboswitch molecule randomly fluctuate between its bound and unbound conformations. The ratio of *B* relative to *U* in the whole population, however, remains constant and can be described by the equilibrium constant K_B. The free energy difference between the bound conformation relative to the unbound conformation, ΔG_B, is also a constant.

$$K_B = \frac{[B]}{[U]} = \frac{e^{-G_B/RT}}{e^{-G_U/RT}} = e^{-\Delta G_B/RT}$$

When *U* and *B* have equal concentrations, their equilibrium ratio $K_B = 1$ and energy difference $\Delta G_B = 0$. This means the riboswitch has no preference either way for *U* or *B*. A positive value of ΔG_B means that *B* has a higher energy and is therefore populated less frequently than *U*. A negative value of ΔG_B means the opposite: *B* is more populated than *U* because it is the more energetically stable conformation.

Environment and sequence perturbations can promote ensemble redistribution

The concentrations of the adenine bound and unbound riboswitch conformations depend on the free energy difference between them. We now consider two ways to

redistribute the ensemble by altering this free energy difference: changing the riboswitch's chemical environment by adding ions and changing its sequence by introducing mutations.

Perturbations to the chemical environment of a molecular switch can impact its multiple conformations differently, leading to changes in their equilibrium concentrations. In the adenine riboswitch, the binding pocket of the bound conformation is organized around the adenine molecule. Binding pocket organization is further promoted by collapsing the entire molecule into a more compact structure by docking its two distant loops together (Leipply and Draper 2011). Such compaction is energetically unfavorable because it brings negative charges on the RNA phosphate backbone into close proximity. The interaction of positively charged magnesium ions with the RNA backbone helps to neutralize its negative charge. This electrostatic neutralization is more stabilizing to the bound conformation than the unbound conformation because it is more structurally compact. Adding magnesium ions will decrease ΔG_B and redistribute the riboswitch ensemble: there will be a larger bound concentration and smaller unbound concentration at equilibrium than before the ions were added.

In a similar fashion, perturbing the sequence of a molecular switch with mutations can also redistribute its conformational ensemble. As previously mentioned, docking of the riboswitch's two loops helps to organize the binding pocket and stabilize the bound conformation. Introducing a mutation at a position in the loop-loop interface that interferes with loop docking will destabilize the bound conformation but will have little effect on the unbound conformation that does not have docked loops. Introducing

this loop docking mutation will increase ΔG_B, reshaping the ensemble in the opposite direction of magnesium ions: there will be a smaller bound concentration and larger unbound concentration as compared to the wildtype, or unmutated, riboswitch.

Background II: Thermodynamic modelling of coupled sites

Now that we can calculate the concentrations of conformations in an ensemble, we can further use thermodynamics to quantify the functional effect of epistatic coupling between sites in a biomolecule.

As described previously, a biomolecule is a sequence of linked chemical building blocks that chemically and physically interact with each other, causing the molecule to fold into a structure (or ensemble of structures) that determines its function. Each site in the molecule contributes to the function by nature of its location. For a molecular switch with an ensemble of conformations, a given site might be in a different location in different conformations, such as how a site at the loop-loop interface of the adenine riboswitch bound conformation is not part of an interface in the unbound conformation.

Introducing a mutation replaces a chemical group with a different one, altering the chemical and physical environment surrounding the site. In each conformation, new interactions can form with the mutated chemical group that change the conformation's energy. But these new interactions could be slightly different in conformations where the site is in a slightly different location, resulting in a different energy change.

We can measure the effect of a mutation on the function of the biomolecule. For a molecular switch, this function often depends on the relative concentration of "functional" relative to "nonfunctional" conformations. If we consider the adenine-

binding function of the adenine riboswitch to be determined by the bound conformation, then the functional effect of mutation *X* depends on how it differentially affects the energies of *U* and *B*, readjusting their equilibrium concentrations as quantified by K_{bind}^{X} and ΔG_{bind}^{X}.

$$K_{bind}^{X} = \frac{[B]}{[U]} = \frac{e^{-(G_B + \Delta G_B^X)/RT}}{e^{-(G_U + \Delta G_U^X)/RT}}$$

Taking the natural logarithm of this equation gives

$$\Delta G_{bind}^{X} = (G_B + \Delta G_B^X) - (G_U + \Delta G_U^X)$$

We can measure the effect of a different mutation *Y* in the same way. If the sites of mutations *X* and *Y* are independent of each other, we expect their individual energetic effects on each conformation to sum together when the mutations are paired.

$$\Delta G_{bind}^{XY} = (G_B + \Delta G_B^X + \Delta G_B^Y) - (G_U + \Delta G_U^X + \Delta G_U^Y)$$

However, if the sites are coupled within a conformation, we might expect to see an extra effect on the function that only occurs when both mutations are present. This coupling-related effect is quantified as epistasis, ϵ. Epistasis measures how the functional effect of a mutation appears to change when it is introduced alone *(X)* versus with its coupled partner *(XY)*.

$$\epsilon^{XY} = (\Delta G_{bind}^{XY} - \Delta G_{bind}^{Y}) - (\Delta G_{bind}^{X} - \Delta G_{bind})$$

Background III: Contact and ensemble mechanisms for epistatic coupling

Contact-based epistatic coupling

There are two potential ways to get epistatic coupling between mutations at different sites in a biomolecule. One mechanism is that the coupled positions are

connected by a network of sites that directly contact each other within a conformation. In what we call "contact epistasis," a mutation at one site introduces energetic and structural perturbations that are propagated through the contact network to the mutation at the second site. In the simplest case, the coupled sites make direct contact with each other, such as by forming a base pair in an RNA molecule. The energetic effects of within-conformation contacts can be described by adding a coupling term $\Delta\Delta G^{XY}$ the energy of each conformation.

$$\Delta G^{XY}_{contact} = (G_B + \Delta G^X_B + \Delta G^Y_B + \Delta\Delta G^{XY}_B) - (G_U + \Delta G^X_U + \Delta G^Y_U + \Delta\Delta G^{XY}_U)$$

All $G, \Delta G^X$, and ΔG^Y terms are given by independent measurements of mutations *X* and *Y,* so they cancel out when plugged into the equation for epistasis. The $\Delta\Delta G^{XY}$ contact terms cannot be predicted from these independent measurements, so they compose extra the epistatic effect that occurs upon pairing the coupled *X* and *Y* mutations.

$$\epsilon^{XY}_{contact} = \Delta\Delta G^{XY}_B - \Delta\Delta G^{XY}_U$$

Ensemble-based epistatic coupling

Thus far, we have been considering the ensemble of the adenine riboswitch to consist solely of a bound and an unbound conformation. In this two-state system, all epistatic coupling comes from within-conformation contacts. If there are no contact pathways connecting the sites of the paired mutations, then each contact-based coupling term is zero, resulting in no epistasis.

We can expand the riboswitch ensemble, however, by considering a third conformation. Experimental evidence suggests that "unbound" conformation actually consists of two interconverting conformations: an extended *(E)* conformation with a low

affinity for adenine and a docked *(D)* conformation poised to bind adenine (Figure 2, Leipply and Draper 2011).

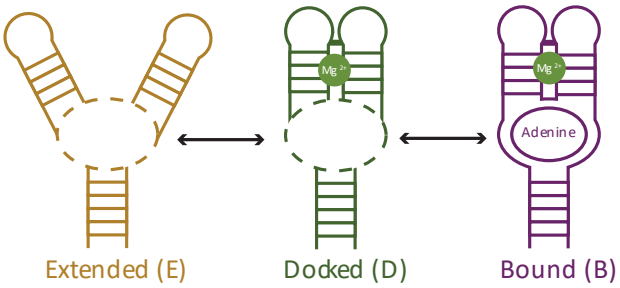

Figure 2. Three-state adenine riboswitch ensemble.

Unbound conformations of the adenine riboswitch can be separated into one with docked loops (D) and another with a more extended structure (E)

With the function of the riboswitch still proportional to the concentration of the adenine-bound conformation *(B)*, the functional effect of mutation *X* now depends on its energetic impacts to *B, D,* and *E.*

$$K_{bind}^{X} = \frac{[B]}{[D]+[E]} = \frac{e^{-(G_B+\Delta G_B^X)/RT}}{e^{-(G_D+\Delta G_D^X)/RT} + e^{-(G_E+\Delta G_E^X)/RT}}$$

Taking the natural logarithm of this equation to give ΔG_{bind}^{X} does not result in the neat linear form of the two-state system. The Boltzmann energy terms for the *D* and *E* conformations cannot be separated from each other under the logarithmic operation. Therefore, the effect of mutation *X* on adenine binding, ΔG_{bind}^{X}, is given by

$$\Delta G_{bind}^{X} = (G_B + \Delta G_B^X) + RTln\left(e^{-(G_D+\Delta G_D^X)/RT} + e^{-(G_E+\Delta G_E^X)/RT}\right)$$

In the case that mutations *X* and *Y* are independent from each other within each conformation (that is, there are no $\Delta\Delta G^{XY}$ contact-based coupling terms), the functional effect of pairing *X* and *Y* together is given by ΔG_{ens}^{XY}.

$$\Delta G_{ens}^{XY} = (G_B + \Delta G_B^X + \Delta G_B^Y) + \mathrm{RTln}\left(e^{-(G_D+\Delta G_D^X+\Delta G_D^Y)/RT} + e^{-(G_E+\Delta G_E^X+\Delta G_E^Y)/RT}\right)$$

While all $G, \Delta G^X$, and ΔG^Y terms are given by independent measurements of mutations *X* and *Y*, the energetic terms corresponding to the *D* and *E* conformations are mathematically "trapped" within the logarithmic term, which we can denote $G_{D,E}$. Therefore, the only terms that cancel upon calculating ensemble-based epistasis are the terms corresponding to the mathematically accessible *B* conformation.

$$\epsilon_{\mathrm{ens}}^{\mathrm{XY}} = (\Delta G_{ens}^{XY} - \Delta G_{bind}^{Y}) - (\Delta G_{bind}^{X} - \Delta G_{bind})$$
$$= \left(G_{D,E}^{XY} - G_{D,E}^{Y}\right) - \left(G_{D,E}^{X} - G_{D,E}\right)$$

In the case that one of the conformations in logarithmic term, such as *E*, has such a high energy relative to *D* that its concentration is practically zero, the ensemble collapses back to a two-state system:

$$\Delta G_{bind}^{\mathrm{XY}} = (G_B + \Delta G_B^X + \Delta G_B^Y) - (G_D + \Delta G_D^X + \Delta G_D^Y)$$

Since all $G, \Delta G^X$, and ΔG^Y terms are now mathematically accessible and there are no $\Delta\Delta G^{XY}$ contact-based coupling terms, the epistasis between mutations *X* and *Y* is zero.

Therefore, for a biomolecule with an ensemble with three or more conformations, we might expect to see epistatic coupling between two mutations do not make within-conformation contacts as long as multiple conformations within the logarithmic term have similar energies. In a recent manuscript, we termed this across-conformation coupling mechanism "ensemble epistasis" (Morrison et al. 2020).

Ensemble-based coupling reflects the fact that the exponential Boltzmann energy function is nonlinear and can only be broken into additive components when conformations are distinguishable from one another. Since measuring adenine binding does not distinguish between the two unbound conformations, mutations that change the

energies of these conformations redistribute their concentrations in a nonlinear way. This nonlinear redistribution cannot be predicted from adding up the effects of paired mutations within each conformation, resulting in epistatic coupling at the level of function since it depends on the relative concentrations of all conformations in the ensemble.

In Morrison et al. (2020), we proposed a simple experimental test for ensemble-based coupling. As described earlier, one way to redistribute an ensemble is by changing the biomolecule's environment such as by adding magnesium ions to the adenine riboswitch. We expect significant ensemble-based epistatic coupling between paired mutations in environmental conditions where conformations in the logarithmic term have equal energies. Conversely, we expect negligible ensemble-based coupling in conditions where the logarithmic term collapses to a single conformation. Our work suggests that environment-dependent patterns of epistasis could be a key signature of this ensemble-based mechanism of epistatic coupling between mutations.

Specific research goals

The two mechanisms for epistastic coupling, within-conformation contacts and across-conformation ensemble redistribution, need not be mutually exclusive. What if paired mutations in a three-conformation molecular switch make within-conformation contacts? The functional effect of the paired mutations in the adenine riboswitch would be given by

$$\Delta G^{XY}_{ens+contact} = (G_B + \Delta G^X_B + \Delta G^Y_B + \Delta\Delta G^{XY}_B) - \mathrm{RT}\ln\left(e^{-(G_D+\Delta G^X_D+\Delta G^Y_D+\Delta\Delta G^{XY}_D)/RT} + e^{-(G_E+\Delta G^X_E+\Delta G^Y_E+\Delta\Delta G^{XY}_E)/RT}\right)$$

While this equation was presented in Morrison et al. (2020), we had yet to investigate epistatic coupling resulting from both the within-conformation contact terms ($\Delta\Delta G^{XY}$) and the across-conformation ensemble term ($-RTln(e^{G_D} + e^{G_E})$). Since the contact terms are embedded within the ensemble term, we asked whether it is possible to distinguish between contact- and ensemble-based epistatic coupling in a real molecule. If so, what are the relative magnitudes of each? Are there different patterns of contact- and ensemble-based coupling depending on the relative locations of the paired sites within the molecule?

Project Design

We set out to answer these questions using the adenine riboswitch as a simple model system. As contact- and ensemble-based epistatic coupling is likely to get more convolved in large molecules with large ensembles, we selected the smallest, simplest biomolecule with a well-studied conformational ensemble (Porter et al. 2014). We took advantage of the riboswitch's high binding affinity for a fluorescent base analog of adenine, 2-aminopurine (2AP), to get a fluorescence-based readout of mutational effects on adenine binding.

We also capitalized on the ability of positively charged magnesium ions to promote the collapse of RNA molecules into more compact structures by offsetting the negative charge of the RNA phosphate backbone. As described previously, the addition of magnesium ions selectively stabilizes the docked and bound riboswitch conformations because of their compact, organized regions. We can therefore manipulate the magnesium ion concentration to shift the conformation energies into regimes where only the extended conformation is appreciatively populated (low

ensemble coupling), both extended and docked are populated (high ensemble coupling), and only the docked conformation is appreciatively populated (low ensemble coupling). Such an approach also allows us to measure the magnitude of ensemble epistasis in this system. Any remaining epistatic coupling can then be attributed to within-conformation contacts.

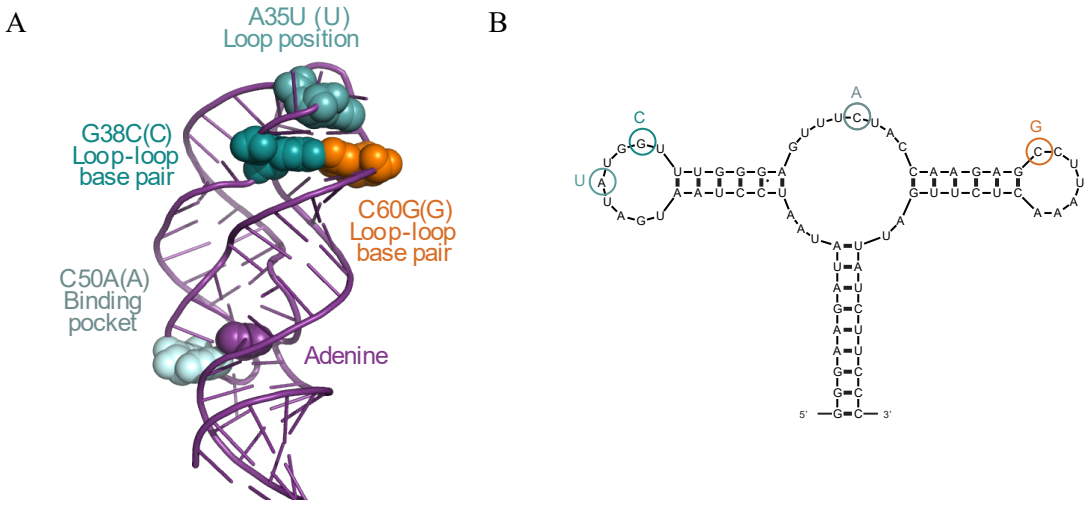

Figure 3. Positions of selected mutations to the adenine riboswitch.

A) Crystal structure of riboswitch bound conformation with mutated bases shown as colored spheres (PDB: 4XNR). B) Riboswitch sequence with mutations identified.

Experimental Hypotheses

We selected mutation pairings at different locations throughout the riboswitch to sample different combinations of the two types of epistatic coupling (Figure 3). We chose loop-docking disrupting mutation C60G (G) as our reference mutation and paired it with G38C (C), its loop-docking base pair partner, A35U (U), located nearby in the opposite loop, and C50A (A), located at a distant position in the adenine binding pocket.

We expected that the G-C pairing would be dominated by contact-based coupling due to the ability of the paired mutations to form a base pair. As we did not foresee an obvious contact network connecting the loop positions of G and U, we predicted that the G-U pairing would be dominated by ensemble-based coupling. The positions of loop mutation G and binding pocket mutation A are structurally distant, but because loop docking promotes binding pocket organization, we predicted that a contact network might connect this mutation pair, leading to a mix of contact- and ensemble-based epistatic coupling.

Results

We set out to measure and analyze the mechanisms responsible for epistatic coupling between mutations to the adenine riboswitch. As environment-dependent epistasis has been a proposed signature of ensemble-based coupling (Morrison et al. 2020), we wanted to measure the effects of mutations on adenine binding across a Mg^{2+} concentration range in which the riboswitch ensemble transitions between populating different conformations. The cellular concentration of magnesium ions is approximately 1 mM (Tyrell et al. 2013), so we chose to measure adenine binding at 0.1, 1, 10, and 100 mM Mg^{2+}.

We performed an adenine binding assay for the wildtype riboswitch using the fluorescent base analog 2AP as previously described (see Methods, Lemay et al. 2006, Stoddard et al. 2013) at these four magnesium concentrations. As the fluorescence of 2AP is quenched upon binding to the riboswitch, full fluorescence quenching means that all 2AP molecules have been bound up while no quenching means all 2AP molecules are free in solution. We kept the 2AP concentration constant while titrating in excess RNA to capture the full dynamic range of 2AP fractional saturation levels for the wildtype riboswitch (Figure 4 B,D,F,H).

Modelling a magnesium-dependent riboswitch ensemble

We first sought to extract information about the underlying riboswitch ensemble from these 2AP binding curves in order to later decompose mutational effects and epistatic couplings into components for each conformation. We turned to the literature to find thermodynamic models of the adenine riboswitch conformational ensemble. Leipply and Draper (2011) identified four conformations: the extended and docked

conformations described previously in both adenine bound and unbound forms. They described the system with equilibrium constants for the transition between extended and docked conformations (K_{dock}), the transition between adenine bound and unbound forms (K_{bind}), and a linkage term between loop docking and adenine binding (K_{link}).

As we wanted to measure the magnesium-dependence of the riboswitch ensemble, we needed to add parameters to describe RNA-ion interactions. In another study, Leipply and Draper (2010) reported that the transition from the extended to docked conformation involves the uptake of magnesium ions. These ions, however, tend to form ion clouds around riboswitch molecules with many long-range electrostatic interactions rather than a few short-range binding interactions at well-defined positions in the riboswitch. Leipply and Draper modelled the RNA-ion interaction with an ion-uptake coefficient that could be equated with a hill coefficient within narrow magnesium concentration ranges.

We began by making the simplifying assumption that a magnesium-binding hill coefficient would hold over our 1000-fold Mg^{2+} concentration range. We therefore added an equilibrium constant for magnesium ion binding to the docked conformation (K_{MG}) with a corresponding hill coefficient (n) representing the number of magnesium ions taken up in the extended to docked transition.

Four-state ensemble model

We used this four-state model to describe the 2AP binding data at all four magnesium concentrations simultaneously (Figure 4 A,B). We took a maximum-likelihood fitting approach to find estimates of the model parameters K_{link}, K_{dock}, K_{bind}, K_{MG}, and n that best reproduced the 2AP binding data. Since we only have access to the

total concentrations of RNA, 2AP, and magnesium ions ($[R]_T$, $[A]_T$, and $[M]_T$), this fitting process also involved estimating the free 2AP concentration [A] and free magnesium ion concentration [M]. Given four-state model estimates for all of these values, we can calculate the concentrations of each of the four conformations in the riboswitch ensemble.

Three- and two-state ensemble models

We then repeated the 2AP binding assay to measure the independent effects of each selected mutation (G, C, U, A) and mutation pair (GC, GU, GA). After fitting the four-state model to all riboswitch genotypes, or sequence variants, we found that the concentration of the adenine-bound extended conformation *(EA)* was never more than 0.02% of the total RNA concentration. We therefore asked if we could simplify our model by ignoring the *EA* conformation and eliminating the fitting parameter K_{link}.

The resulting three-state model is shown in Figure 4 (C,D). As this three-state model appeared to fit the 2AP binding data just as well, if not better, than the four-state model, we asked if we could simplify the model further by ignoring the unbound docked conformation *(D)* and eliminating the fitting parameter K_{dock}. As seen in Figure 4 (E,F), the two-state model was unable to successfully fit the 2AP binding data.

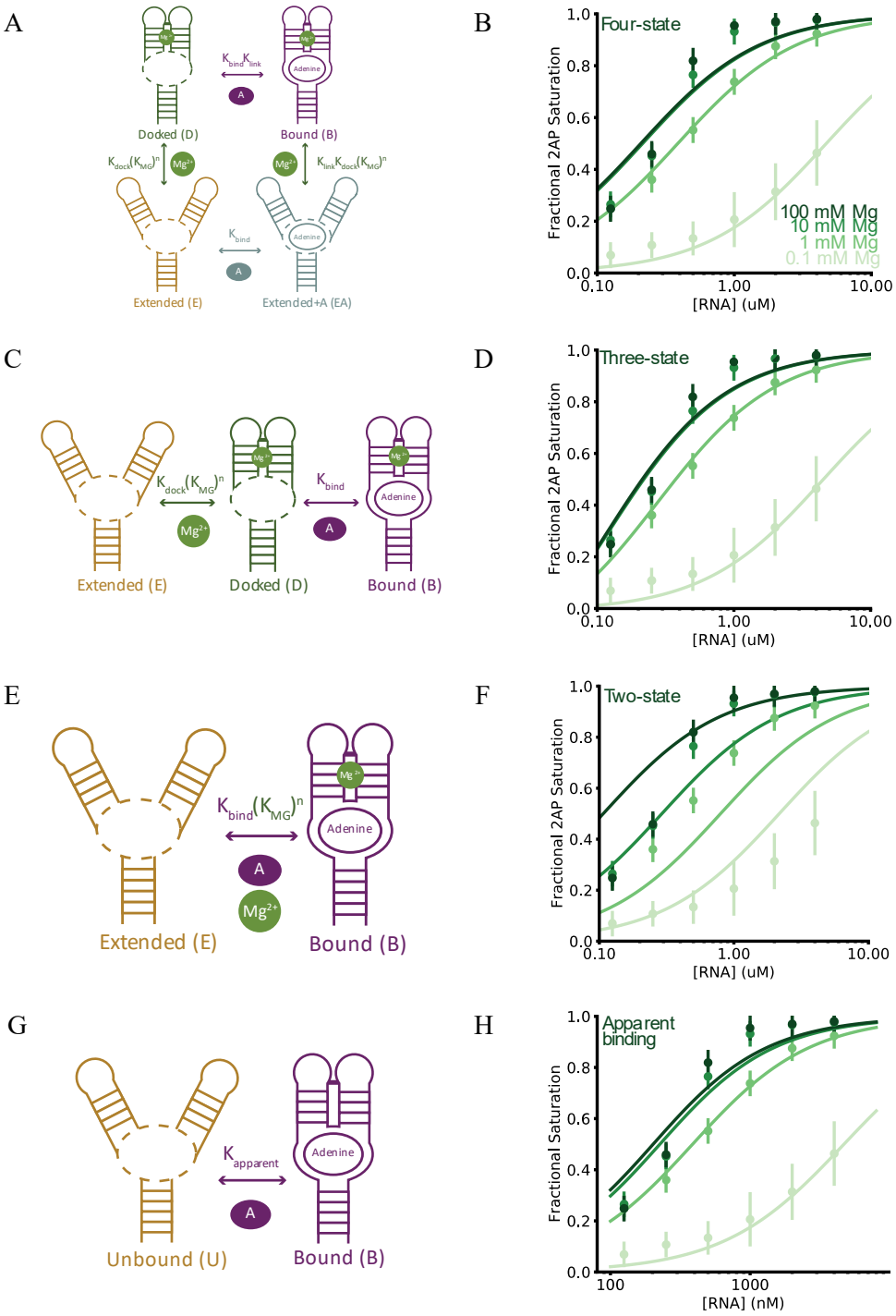

Figure 4. Comparison of the four-, three-, two-state, and apparent binding models.

A-B) Four-state thermodynamic model and fit of 2AP binding data for the wildtype riboswitch at 0.1, 1, 10, and 100 mM Mg^{2+}. C-D) Three-state model and fit. E-F) Two-state model and fit. G-H) Apparent binding model and fit.

Model selection

To select the best model of the adenine riboswitch ensemble, we performed an AIC statistical test (see Methods) across all genotypes for each model. The AIC test favors models with high likelihoods while penalizing models with more parameters that are prone to overfitting. Relative to a three-state AIC probability of 1, the AIC test gave a four-state probability of $\frac{1}{10^{49}}$ and a two-state probability of 0. Therefore, the remaining analyses were conducted using a three-state model of the adenine riboswitch composed of magnesium- and adenine-dependent transitions between extended *(E)*, docked *(D)*, and bound *(B)* conformations. Figure 5 shows three-state model fits of the 2AP binding data for all genotypes, with the "clouds" representing Bayesian sampling from parameter uncertainty.

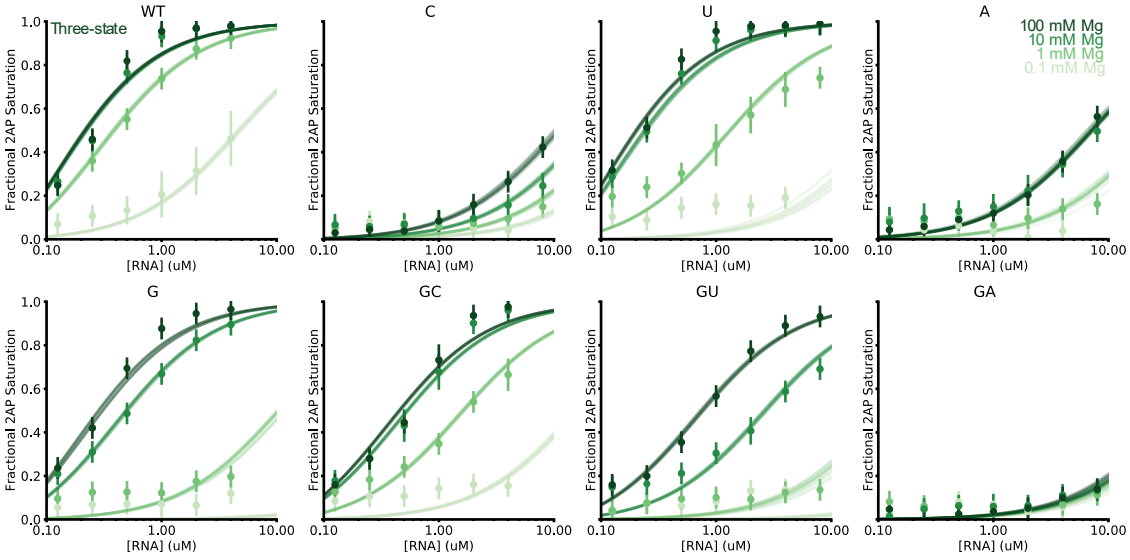

Figure 5: Three-state model fits of 2AP binding data for all riboswitch genotypes.

Magnesium-independent model check

We wanted to check whether we were justified in our simplifying assumption that a single hill-coefficient adequately describes RNA-ion interactions over our 1000-fold Mg^{2+} concentration range. We therefore defined a two-state model to describe the apparent binding equilibrium (K_{app}) between all adenine bound *(B)* and unbound conformations *(U)* at isolated magnesium ion concentrations (Figure 4 G,H).

We observed that the three-state ensemble model and the apparent binding model similarly fit the 2AP binding data at 1, 10, and 100 mM Mg^{2+} concentrations (Figures 5 and 6). The three-state model, however, systematically underestimates 2AP fractional saturation values at 0.1 mM Mg^{2+} while the apparent binding model does not, which suggests that the three-state model does not accurately model RNA-ion interactions at this low magnesium ion concentration. We therefore restricted our following analyses to a magnesium ion concentration range between 1 and 100 mM Mg^{2+}.

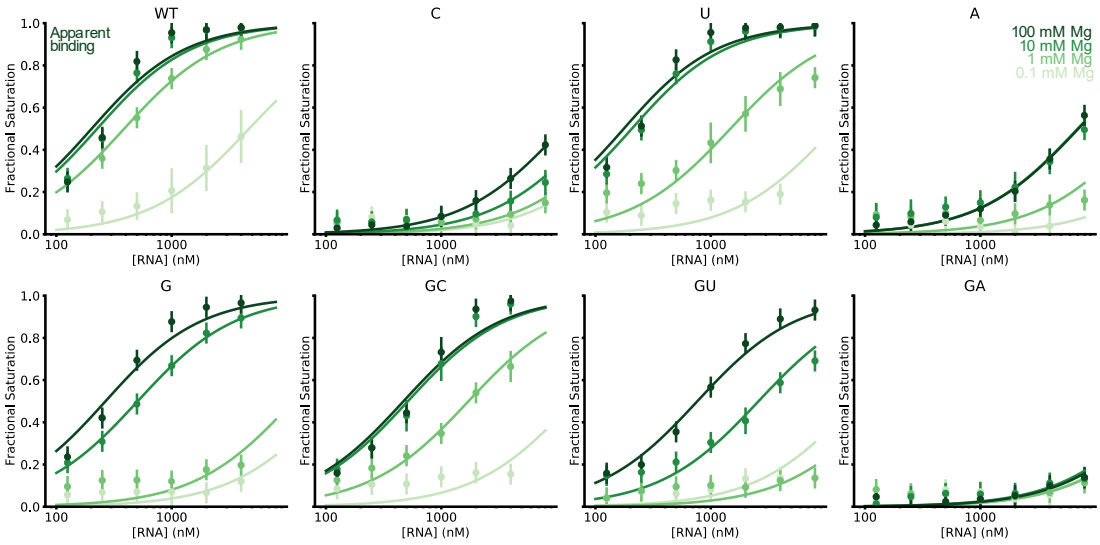

Figure 6: Apparent binding fits of 2AP binding data for all riboswitch genotypes.

Modelling the effects of mutations on conformations within the ensemble

We then attempted to dissect the mechanisms of epistatic coupling between a loop-docking mutation G and its partner base-pair mutation C, opposite loop mutation U, and binding pocket mutation A. The magnitude of epistatic coupling represents the change in functional effect when mutation G is introduced independently into the wildtype background versus when it is introduced into a mutant riboswitch already carrying the C, U, or A mutations.

The function of the adenine riboswitch, as discussed in the introduction, depends on the concentration of the adenine bound *(B)* conformation relative to the unbound docked *(D)* and extended *(E)* conformations. The energy of each of these conformations can be altered by mutations, which may change its relative concentration in solution. We must take all conformations into account to decompose epistatic coupling into within-conformation contacts and across-conformation ensemble redistribution.

Figure 7 uses equilibrium constants derived from the three-state model to calculate how mutation G redistributes the relative populations of all three conformations as a function of magnesium ion concentration when introduced into the WT, C, U, and A genetic backgrounds. For this and all subsequent analysis, the total RNA concentration is held at 1 uM to ensure at least a 100-fold magnesium ion excess. This allows us to simplify our analysis by using the total magnesium ion concentration to approximate the free concentration, $[M] = [M]_T$. Further, at this RNA concentration, the three-state model suggests that the concentration of free adenine is never less than 95% of the total adenine concentration, such that we can make the simplifying approximation $[A] = [A]_T$. Therefore, the energies and concentrations of all riboswitch

conformations can be calculated relative to the wildtype extended conformation as follows.

$$G_E = -\mathrm{RTln}\left(\frac{[E]}{[E_{WT}]}\right), [\mathrm{E}] = \frac{[R]_T}{1 + K_{dock}(K_{MG}[M]_T)^n}$$

$$G_D = -\mathrm{RTln}\left(\frac{[D]}{[E_{WT}]}\right), [D] = K_{dock}(K_{MG}[M]_T)^n[E]$$

$$G_B = -\mathrm{RTln}\left(\frac{[B]}{[E_{WT}]}\right), [B] = K_{bind}[A]K_{dock}(K_{MG}[M]_T)^n[E]$$

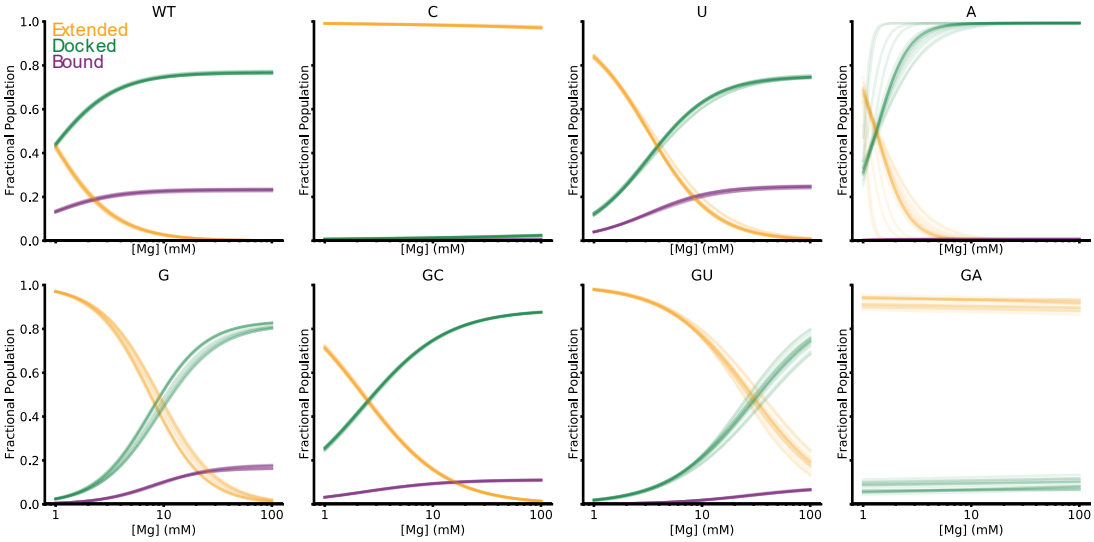

Figure 7. Fractional populations of riboswitch conformations.

Populations were calculated from a three-state model of the riboswitch ensemble as a function of the total magnesium concentration.

Analysis of epistatic coupling between sites that form a base pair (G-C)

Ensemble redistribution from mutations G and C

We first used our three-state model to analyze epistatic coupling between sites that make direct contact with each other. From the population plots in Figure 7, we can

see that mutation G requires a higher magnesium concentration that the wildtype riboswitch to transition from populating the extended conformation (yellow curves) to populating the docked and bound conformations (green and purple curves, respectively). This observation makes sense considering that this mutation changes a strong G-C base pair spanning the docked loops to an unfavorable G-G mismatch. Without this base pair to stabilize the loop-loop interface, high magnesium concentrations are required to collapse them together.

Similarly, mutation C disrupts the base pair, but the resulting C-C mismatch is much more destabilizing to the docked and bound states than the G-G mismatch. The extended state remains the only state appreciatively populated in the magnesium range used for these experiments.

Pairing mutations G and C results in the formation of a reversed C-G base pair in the docked and bound conformations. While mutation G moved the extended to docked transition to a higher magnesium concentration when introduced into the wildtype (WT) background, it moved the transition to a lower magnesium concentration when added into the C background. This is a clear example of epistatic coupling, as mutation G's effect depends on whether it is alone or paired with mutation C.

Total G-C epistatic coupling

Before attempting to decompose this epistatic coupling into contact- and ensemble-based components, we looked at the total coupling between the paired mutations across our magnesium range. We used fitted equilibrium constants from the three-state model to calculate the free energy of adenine binding, ΔG_{bind}, as a function of magnesium ion concentration for the WT, G, C, and GC genotypes.

$$\Delta G_{bind} = G_B + \mathrm{RTln}\left(e^{-G_D/RT} + e^{-G_E/RT}\right)$$

The total epistasis resulting from pairing mutations G and C, given by the following equation, is plotted in dark blue in Figure 8B.

$$\epsilon_{\mathrm{total}}^{\mathrm{GC}} = \left(\Delta G_{bind}^{GC} - \Delta G_{bind}^{C}\right) - \left(\Delta G_{bind}^{G} - \Delta G_{bind}^{WT}\right)$$

We see that the total epistasis from the G-C pairing is magnesium-dependent, which suggests that ensemble redistribution may be a source of the epistatic coupling. We also see that the average epistasis across the magnesium range is approximately -3 kcal/mol, which means that the functional effect of mutation G on adenine binding is 3 kcal/mol more favorable when the mutation is introduced into the C background rather than the wildtype background.

To confirm that this magnesium-dependent pattern of epistatic coupling is not an artifact of poor RNA-ion interaction modelling with the three-state model, we also calculated epistasis in adenine binding at 1, 10, and 100 mM Mg^{2+} with the magnesium-independent apparent binding model (cyan dots, Figure 8B). We see that the magnitude and pattern of epistatic coupling are identical for both models, supporting the validity of the three-state model's simplified representation of RNA-ion interaction within the magnesium concentration range.

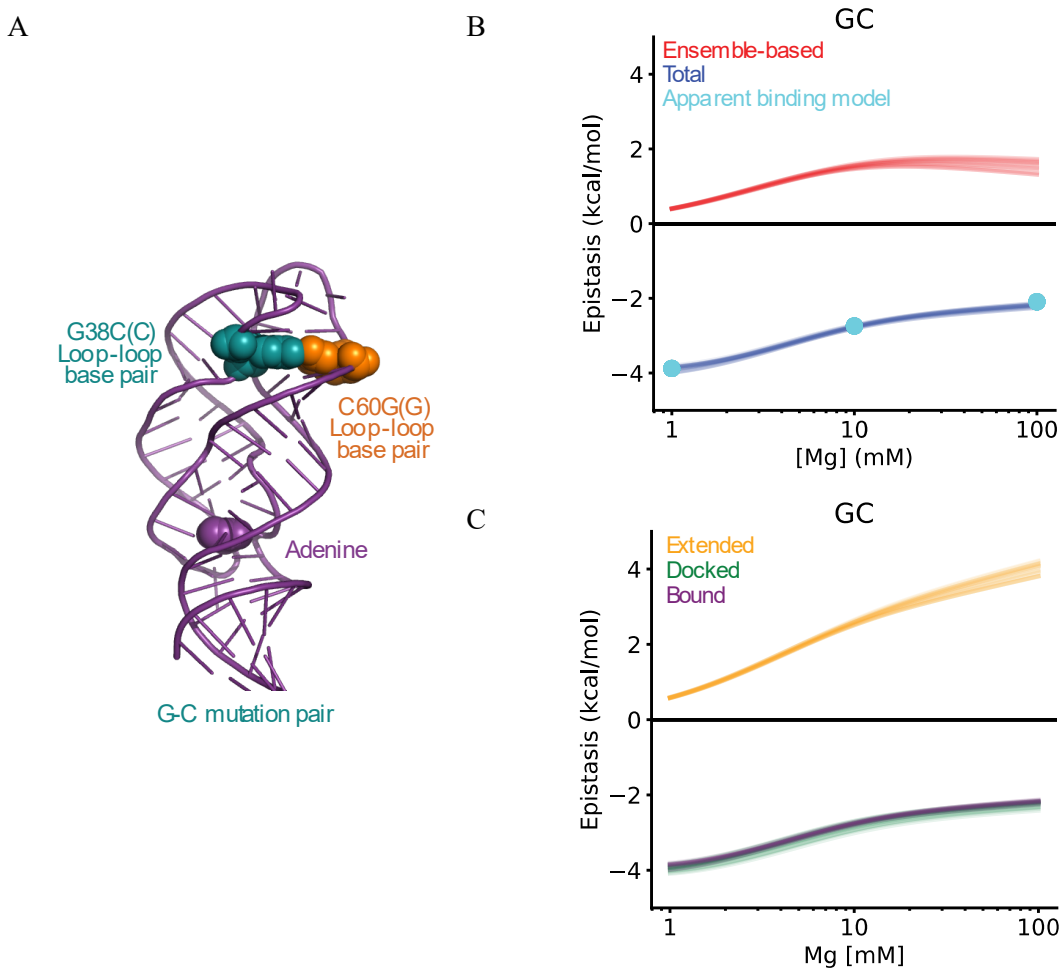

Figure 8. Epistatic coupling upon pairing mutations G and C.

A) Mutations G and C form a base pair spanning the docked loops in the bound and docked riboswitch conformations. B) Total and ensemble-based coupling between the G-C mutation pair as a function of magnesium concentration. C) Within-conformation coupling for the G-C mutation pair.

Within-conformation G-C contacts

We next used our three-state model to calculate the amount of epistatic coupling that occurs within each of the three conformations when mutations G and C are paired together.

By looking at the energies of each conformation individually, we can calculate the magnitude of within-conformation contacts generated when mutations G and C are paired together. For example, contact-based coupling within the docked conformation is given by a $\Delta\Delta G_D^{GC}$ contact energy term.

$$\epsilon_{\mathrm{D,contact}}^{\mathrm{GC}} = \Delta\Delta \mathrm{G}_{\mathrm{D}}^{\mathrm{GC}} = (\Delta G_D^{GC} - \Delta G_D^{C}) - (\Delta G_D^{G} - G_D)$$

The contact-based coupling for each conformation is shown in Figure 8C. We see that the extended state loses favorable contacts (or gains unfavorable contacts) as the magnesium concentration increases (yellow curve). Perhaps this destabilization comes from the unfavorability of maintaining an extended form while magnesium ions promote structural collapse.

We see that the docked and bound states experience identical within-conformation coupling, likely because the reformed C-G base pair is the same in both conformations (green and purple curves, Figure 8C). The contact-based coupling is strongest at low magnesium concentrations (-4 kcal/mol) but weakens with increasing magnesium to approximately 2 kcal/mol. This trend might be explained by high ion concentrations changing the dielectric constant surrounding the reformed base pair such that it is less stabilizing. Interestingly, we noted that the contact-based epistasis within the docked and bound conformations (green and purple curves, Figure 8C) exactly matches the total epistatic coupling for the G-C mutation pair (dark blue curve, Figure 8B).

Ensemble-only G-C epistatic coupling

We then attempted to uncover the contribution of ensemble-based coupling to the total epistasis measured for the G-C pair. In the absence of any within-conformation

contacts, the independent energetic effects of mutations G and C on each conformation sum together when they are combined. If ensemble epistasis is the only source of coupling, $\Delta G_{ens}^{GC}\ gives$ the functional effect upon pairing them together.

$$\Delta G_{ens}^{GC} = (G_B + \Delta G_B^G + \Delta G_B^C) + \mathrm{RTln}\left(e^{-(G_D+\Delta G_D^G+\Delta G_D^C)/RT} + e^{-(G_E+\Delta G_E^G+\Delta G_E^C)/RT}\right)$$

We replaced our experimentally measured ΔG_{bind}^{GC} with this ensemble-only ΔG_{ens}^{GC} term in the equation for epistasis, which produced the red curve in Figure 8B.

$$\epsilon_{ens}^{GC} = \left(\Delta G_{ens}^{GC} - \Delta G_{bind}^{C}\right) - \left(\Delta G_{bind}^{G} - \Delta G_{bind}^{WT}\right)$$

The ensemble-based contribution to the epistatic coupling between mutations G and C appears to match the general pattern of the total epistasis but is offset by approximately +4 kcal/mol. Ensemble-based coupling approaches zero at low magnesium concentrations (red curve, Figure 8B), as only the wildtype riboswitch has multiple conformations populated in this regime (see Figure 7). As the magnesium concentration increases, the docked and bound conformations become appreciatively populated in the wildtype, G, and GC riboswitches while the extended conformation remains dominant in the C riboswitch. At these higher magnesium concentrations with more populated conformations, ensemble-based coupling reaches a maximum value near +2 kcal/mol. This means ensemble redistribution makes adenine binding 2 kcal/mol less favorable when mutation G is introduced into the C background relative to the wildtype background.

Analysis of epistatic coupling between sites in opposite loops (G-U)

We followed the same analysis pipeline as used for the G-C pairing to decompose the epistatic coupling between mutation G and mutation U located in opposite loops.

We see from the population plots in Figure 7 that mutation U, which also forms interactions to bridge the docked loops, requires higher magnesium concentrations than the wildtype riboswitch to promote the extended to docked transition (yellow and green curves, respectively). This observation suggests that it also destabilizes the loop-loop interface, but to a lesser degree than mutation G that breaks a strong spanning base pair. Pairing mutation G with mutation U results in a greatly shifted extended to docked transition, with a much lower population of the bound state at high magnesium concentrations than seen with either mutation individually (purple curves, Figure 7).

Total G-U epistatic coupling

We used the energies of each conformation in the WT, G, U, and GU genotypes to calculate the total epistatic coupling for the G-U mutation pair.

$$\epsilon_{\text{total}}^{\text{GU}} = \left(\Delta G_{bind}^{GU} - \Delta G_{bind}^{U}\right) - \left(\Delta G_{bind}^{G} - \Delta G_{bind}^{WT}\right)$$

The resulting blue curve in Figure 9B is magnesium dependent, again suggesting that ensemble redistribution may contribute to epistatic coupling. The coupling pattern goes from no epistasis at low magnesium to a maximum near 1 kcal/mol before starting to drop again. This pattern matches that of the apparent binding model (cyan dots, Figure 9B), again supporting the validity of the three-state model.

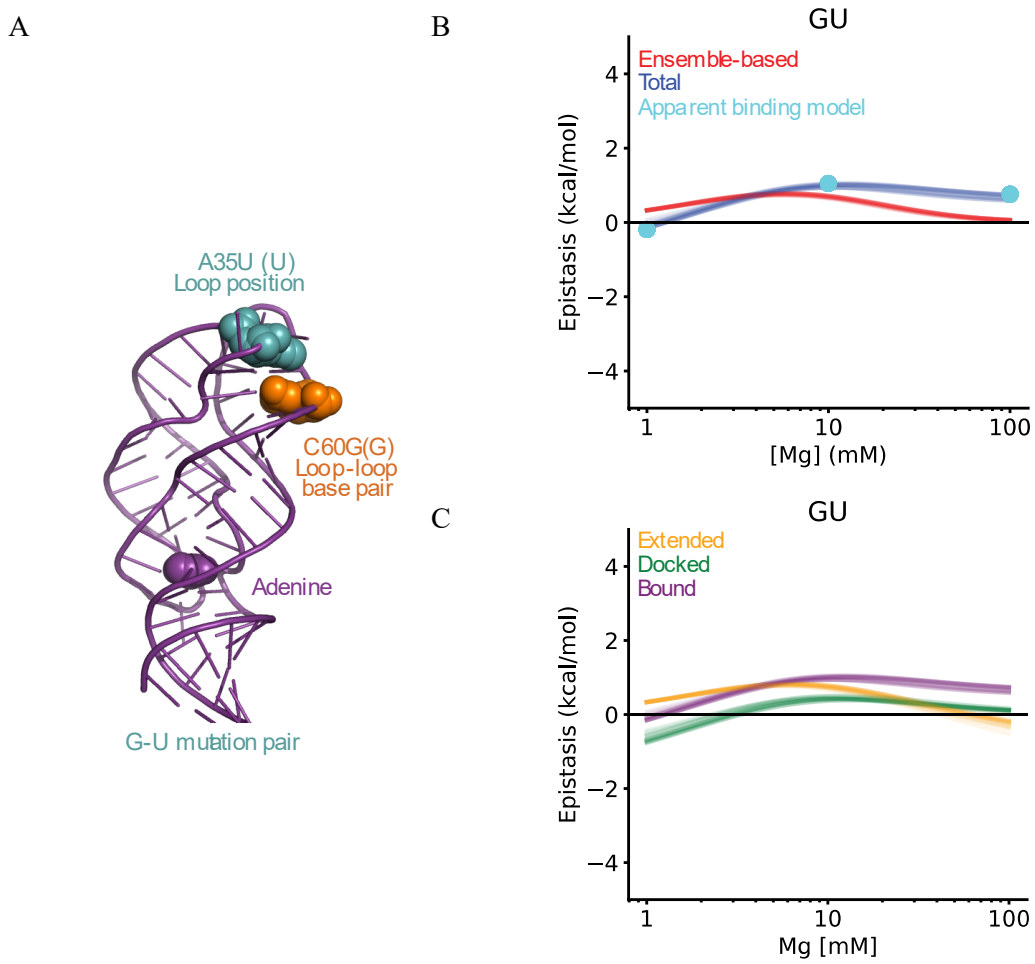

Figure 9. Epistatic coupling upon pairing mutations G and U.

A) Mutations G and U are located nearby, but in opposite loops. Both participate in loop-loop interface interactions in the bound and docked riboswitch conformations. B) Total and ensemble-based coupling between the G-U mutation pair as a function of magnesium concentration. C) Within-conformation coupling for the G-U mutation pair.

Within-conformation G-U contacts

We see that the $\Delta\Delta G^{GU}$ contact energy terms within the three conformations have similarly low magnitudes ranging between -0.5 and 1 kcal/mol across the magnesium concentration range (Figure 9C). Unlike the G-C pairing in which the docked and bound conformations exhibited identical contact-based coupling, pairing

mutation G with mutation U results in contacts that are slightly more stabilizing to the docked state than to the bound state (green and purple curves, Figure 9C). A possible explanation for this difference is that alterations to the allosteric pathway connecting loop docking to binding pocket organization is more pronounced when adenine is bound in the pocket. As with the G-C pair, we noticed that the pattern of contact-based coupling in the docked and bound conformations is similar to the pattern of the total epistatic coupling (blue curve, Figure 9B).

Ensemble-only G-U epistatic coupling

To isolate ensemble-based epistatic coupling, we described the functional effect of the paired G and C mutations as the summed independent energetic effects of mutations G and C on each conformation. We used this ΔG_{ens}^{GU} term to calculate coupling due solely to ensemble redistribution (red curve, Figure 9B).

$$\Delta G_{ens}^{GU} = (G_B + \Delta G_B^G + \Delta G_B^U) + \mathrm{RTln}\left(e^{-(G_D+\Delta G_D^G+\Delta G_D^U)/RT} + e^{-(G_E+\Delta G_E^G+\Delta G_E^U)/RT}\right)$$

$$\epsilon_{ens}^{GU} = \left(\Delta G_{ens}^{GU} - \Delta G_{bind}^{U}\right) - \left(\Delta G_{bind}^{G} - \Delta G_{bind}^{WT}\right)$$

We noted that the ensemble-based coupling was of similar magnitude as the total epistasis, remaining between 0 and 1 kcal/mol throughout the magnesium concentration range, but peaking at an intermediate magnesium concentration where the WT, G, U, and GU riboswitches all populate multiple conformations (see Figure 7).

Analysis of epistatic coupling between sites in distant, linked regions (G-A)

Before analyzing the sources of epistatic coupling for the G-A mutation pair, we first studied how mutations readjust the populations within the riboswitch ensemble as shown in Figure 7. Since mutation A is not located in the docked loop region, it has

little effect of the transition from the extended to docked conformation relative to the wildtype riboswitch (yellow and green curves). It does, however, greatly destabilize the bound conformation such that is never appreciatively populated across the magnesium range (purple curves, Figure 7). This makes sense considering that mutation A is located within the binding pocket, where it can directly disrupt favorable interactions with the adenine molecule in the bound conformation.

Pairing mutation G, which destabilizes the docked loops by breaking a spanning base pair, with mutation A produces a riboswitch ensemble dominated by the extended conformation across the magnesium concentration range. The docked conformation is less populated in GA than in G, suggesting that the binding pocket disorganization caused by mutation A is propagated to the loop region where it produces further destabilization of the docked loops.

Total G-A epistatic coupling

We calculated the total epistatic coupling for the G-A mutation pair using the energies of each conformation in WT, G, A, and GA genotypes (blue curve, Figure 10B).

$$\epsilon_{\text{total}}^{\text{GA}} = \left(\Delta G_{bind}^{GA} - \Delta G_{bind}^{A}\right) - \left(\Delta G_{bind}^{G} - \Delta G_{bind}^{WT}\right)$$

We observed that the total coupling between mutations G and A transitions from a value of -2 kcal/mol at low magnesium concentrations to approximately +1 kcal/mol at the highest concentrations. This sign flip means that the effect of introducing G into the A background changes from making adenine binding more favorable to less favorable relative the wildtype background as the magnesium concentration increases. Again, the

total epistastic coupling generated by the three-state model aligns perfectly with the magnesium-independent apparent binding model (cyan dots, Figure 10B).

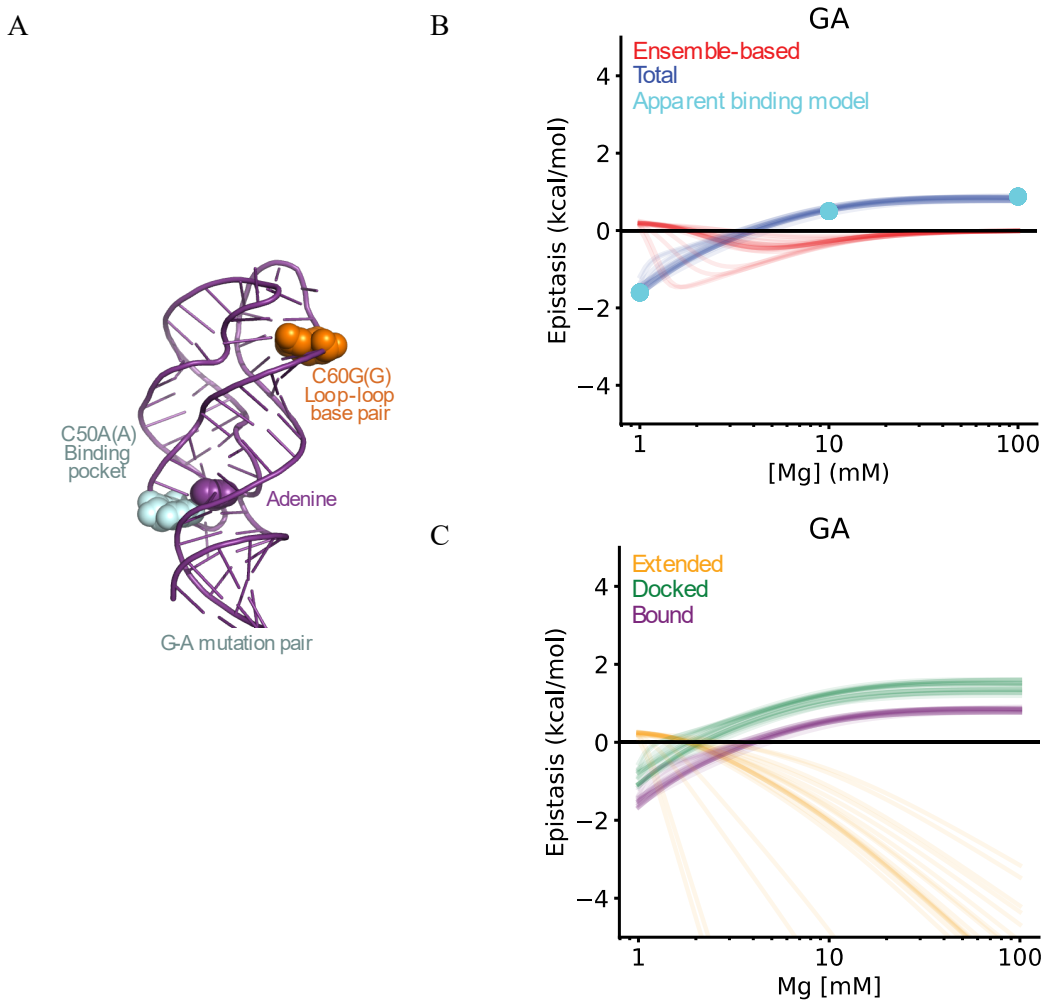

Figure 10. Epistatic coupling upon pairing mutations G and A.

A) Mutations G and A are in different regions of the riboswitch, but they may be connected by an allosteric network linking loop docking and adenine binding. B) Total and ensemble-based coupling between the G-A mutation pair as a function of magnesium concentration. C) Within-conformation coupling for the G-A mutation pair.

Within-conformation G-A contacts

We next isolated $\Delta\Delta G^{GA}$ contact energy terms for each conformation (Figure 10C). The extended conformation possesses stabilizing contacts that increase in strength with the magnesium concentration (yellow curve). Perhaps increased flexibility in both the loop and binding pocket regions as a result of the paired G and A mutations increases flexibility throughout the entire pathway linking the loop and binding pocket regions. This entropic boost to the extended conformation might be stronger at high magnesium concentrations where it helps to prevent magnesium-induced structural collapse.

The within-conformation contacts between mutations G and A in the docked and bound conformations follow a similar pattern as the total epistasis. (green and purple curves, Figure 10C). These contacts are stabilizing at low magnesium concentrations but destabilizing at higher concentrations. Interestingly, the contacts in the bound conformation are similar but slightly more stabilizing than those in the docked conformation. Perhaps the presence of a bound adenine molecule adds stabilizing interactions to the binding pocket region that make the pathway linking the binding pocket to the loop region more robust in the bound conformation relative to docked.

Ensemble-only G-A epistatic coupling

As ensemble-based coupling between the G-A mutation pair is generated solely from the additive energetic effects of mutations within each conformation, we used ΔG_{ens}^{GA} as the functional effect of the paired mutations to calculate the ensemble-based coupling contribution.

$$\Delta G_{ens}^{GA} = (G_B + \Delta G_B^G + \Delta G_B^A) + \mathrm{RTln}\left(e^{-(G_D+\Delta G_D^G+\Delta G_D^A)/RT} + e^{-(G_E+\Delta G_E^G+\Delta G_E^A)/RT}\right)$$

$$\epsilon_{ens}^{GA} = \left(\Delta G_{ens}^{GA} - \Delta G_{bind}^{A}\right) - \left(\Delta G_{bind}^{G} - \Delta G_{bind}^{WT}\right)$$

The resulting red curve in Figure 10B shows that ensemble redistribution contributes minimally to epistatic coupling between the G-A mutation pair, with a maximum value near -0.5 kcal/mol at an intermediate magnesium concentration where the WT, G, A, and GA riboswitches all have a small amount of ensemble diversity (see Figure 7). This ensemble contribution does not appear to have any relation to within-conformation contacts or the total epistasis, which is different from the other two mutation pairs. The ensemble contribution for the G-C pair appears to be a constant offset from the total epistasis (red and blue curves, Figure 8B) and the ensemble contribution for the G-U pair matches the contact-based coupling within the extended conformation (red and yellow curves, Figure 9 B,C).

Single-conformation contacts describe total epistatic coupling

As mentioned briefly in the above analyses for each mutation pair, we noticed that contact-based coupling within the bound conformation tended to reflect the magnesium-dependent pattern of the total epistasis. We confirmed this intuition by plotting both the total epistasis and the contact-based coupling within the bound conformation in Figure 11. The resulting curve is bluish-purple due to the complete overlap of dark blue (total epistasis) and purple (bound conformation contact) curves.

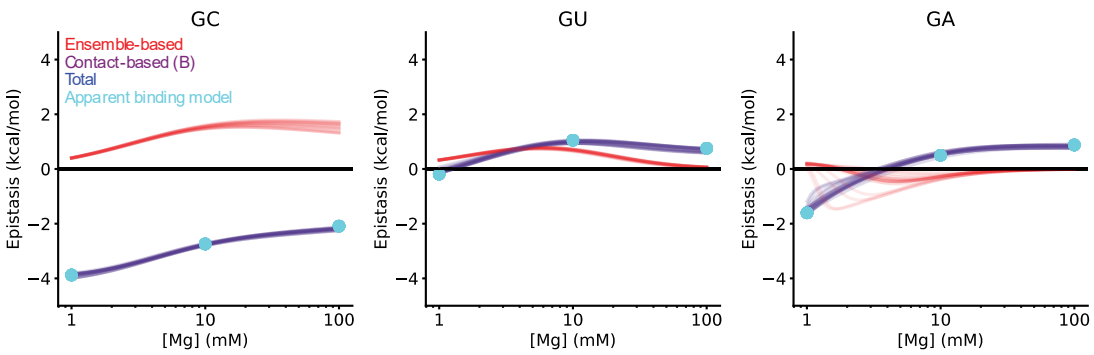

Figure 11. Contact-based coupling matches total epistasis.

Contact-based coupling in the riboswitch bound conformation explains total epistatic coupling for all mutation pairs, even when ensemble-based coupling is nonzero.

As described previously, we defined the function of the riboswitch, ΔG_{bind}, as the energy of the bound conformation relative to a nonlinear function of the energies of the extended and docked conformations.

$$\Delta G_{bind} = G_B + \mathrm{RTln}\left(e^{-G_D/RT} + e^{-G_E/RT}\right)$$

If the total epistatic coupling is generated entirely by the bound conformation, then the logarithmic term has no epistatic contribution.

The nonzero ensemble-based coupling shown as the red curve in Figure 11 for all mutation pairs, however, conflicts with the conclusion that the docked and extended conformations do not contribute to epistasis. They are the only conformations driving ensemble epistasis since the mathematically accessible energetic terms for the bound conformation cancel out.

$$\epsilon_{\mathrm{ens}}^{\mathrm{XY}} = \left(G_{D,E}^{XY} - G_{D,E}^{Y}\right) - \left(G_{D,E}^{X} - G_{D,E}\right)$$

A possible way to collapse the total epistatic coupling for each mutation pair into a bound-conformation contact term is by reducing the nonzero ensemble-based coupling term to zero through docked and extended $\Delta\Delta G^{XY}$ within-conformation

contacts. While this mechanism is plausible, it requires such careful balancing between contact- and ensemble-based coupling that it seems like something that might occur in special circumstances rather than across our entire magnesium range for all three mutation pairs.

Discussion

The primary goal of this project was to determine whether it is possible to distinguish between contact- and ensemble-based epistatic coupling mechanisms in a real molecule. After selecting a magnesium-dependent three-state model to describe the conformational ensemble of the adenine riboswitch, we decomposed the epistastic coupling between three mutation pairs into contact-based components within each conformation and an ensemble-based component across multiple conformations.

Our results reman in tension. Analyzed from one perspective, we see clear evidence of ensemble-based coupling for all mutation pairs. Analyzed from a different perspective, the total epistatic coupling can be entirely explained by contact-based coupling within one conformation. There are possible technical explanations for this inconsistency, discussed below, but these contradictory observations could also reflect the fundamental interconnectedness of coupling pathways in the riboswitch.

Contact- and ensemble-based coupling pathways may not be fully separable

When we measure the functional effect of a mutation to the riboswitch, our readout is integrated over all the contact networks within each conformation as well as across the relative probabilities all conformations in the ensemble. Perhaps we have yet to fully separate the contact- and ensemble-based components of the system, meaning that they remained intertwined throughout the analyses we present. We may think we are looking at two nickels that should add together to give a dime when, in reality, we are looking at two sides of the same coin.

By using contacts within a single conformation to describe the total epistatic coupling for a mutation pair, we may well be overlooking the ways in which those

within-conformation contacts induce ensemble redistribution. Similarly, when we identify ensemble-based coupling by describing how additive energetic effects of mutations to each conformation produce nonlinear ensemble redistribution, we may glaze over important roles that within-conformation contacts play in modulating this effect. A more detailed mathematical treatment of the link between contact- and ensemble-based epistasis may reveal whether decomposition into separate terms is meaningful in an absolute sense, or whether contact- and ensemble-based perspectives are equally valid for constructing a complete description of the same system.

Model improvement and validation

There may be technical reasons for our failure to separate contact- and ensemble-based epistasis. We have identified a possible source of error in the fitting process, as the free adenine concentration might not be properly constrained. Better equilibrium constant estimates from an improved regression model might adjust our modelling of the riboswitch ensemble such that it becomes clear how within-conformation contacts and across-conformation ensemble redistribution combine to produce the total epistatic coupling.

Moving forward, we hope to validate our model-based calculations of within-conformation coupling in the adenine riboswitch with structure-based insight from molecular dynamics (MD) simulations. These molecular movies of a single conformation may allow us to quantify the contact energies for paired mutations without ensemble contamination from the other conformations. Such an external check on the contact-based coupling described our model could help us decide whether the model is fully separating contact- and ensemble-based coupling components.

Magnitudes and patterns of contact- and ensemble-based coupling

Assuming our model of the riboswitch ensemble is correct, we can begin to answer the remaining questions posed at the beginning of the project. We asked about the relative magnitudes of ensemble and contact epistasis in a real molecule, and we found that the magnitude of ensemble-based coupling for our riboswitch mutation pairs ranges between 0 and 2 kcal/mol. The magnitude of total coupling, which can be described entirely by contact contributions, reaches 4 kcal/mol for the mutation pair in direct contact but does not exceed 1 kcal/mol for distantly paired mutations.

We also asked whether the relative location of paired mutations influences the partitioning of epistatic coupling into contact- and ensemble-based components. We did not find a neat pattern linking mutation distance to coupling mechanism, as the base pair-forming mutation pair generated the largest magnitudes of both types of epistasis. We could, however, rationalize most of the observed contact- and ensemble-based coupling patterns from structural and biochemical knowledge of each mutation.

Conclusions

In broad strokes, we set out to understand the pathways of communication between sites in a molecular switch. Paired sites at different locations in our simple adenine riboswitch model system produced experimental signatures of coupling through both direct (contact-based) and indirect (ensemble-based) mechanisms.

While we have yet to successfully tease apart these mechanisms, our observation that both mechanisms can contribute at least 2 kcal/mol to epistatic coupling in a real molecule suggests that efforts to engineer molecular switches, sensors, and catalysts need to take both communication pathways into account. To better predict which sequence changes to combine to generate a desired functional effect, our work points to the need to accurately model how the sequence changes impact each conformation individually and the ensemble of conformations as a whole.

Methods

RNA synthesis

Sequence variants of the *V. vulnificus* adenine riboswitch aptamer domain were synthesized by T7 RNA polymerase *in vitro* transcription of the corresponding DNA oligonucleotides ordered from Eurofins. The wildtype RNA sequence is shown with mutated positions in boldface followed by the identity of the mutation in parentheses.

3'GGGAAGAUAUAAUCCUAAUGAU**A(U)**UG**G(C)**UUUGGGAGUUU**C(A)**UACC AAGAG**C(G)**CUUAAACUCUUGAUUAUCUUCCC

RNA purification

The riboswitch product was purified from the *in vitro* transcription reaction mixture by separating the reaction mixture on a 12% denaturing polyacrylamide gel, identifying the 71-nt product by UV shadowing, and extracting the product from the gel by electroelution. The purified RNA product was concentrated by ethanol precipitation, desalted, quantified by Nanodrop absorbance spectroscopy, and resuspended in 50 mM Tris-HCl before being stored at -80^{oC}.

2-aminopurine binding assays

2AP binding assays were based on ones previously described by Lemay et al. (2006) and Stoddard et al (2013). Riboswitch aliquots were thermally denatured at 90^{oC} for 1 min before refolding on ice for approximately 10 min. Increasing concentrations of RNA were combined with 100 mM KCl, 50 nM 2AP, and 0.1-100 mM $MgCl_2$ in 96-well plate. The assay plate was shaken in a fluorescence plate reader at 37^{oC} for 5 minutes allow the reaction mixtures to reach equilibrium. 2AP was excited at 310 nm,

and emission spectra were recorded from 335-450 nm. Integrated emission spectra for RNA titrations at each experimental condition were normalized relative to corresponding negative (no RNA, no 2AP) and positive (no RNA, 50 nM 2AP) controls to give a fluorescence readout between 0 and 1. Two technical replicates for each experimental condition were averaged together during a given experiment, and the binding assays were repeated to give at least three biological replicates for each experimental condition. The standard deviation across biological replicates was capped at a minimum value of 0.05.

Maximum Likelihood and Bayesian fitting of 2AP binding data

Software written in the Python programming language was used to calculate 2AP fractional saturation values at each experimental condition. The likelihood nonlinear regression package (https://github.com/harmslab/likelihood) was used to produce maximum likelihood and Bayesian estimates of equilibrium constant fitting parameters to describe the 2AP binding data.

Likelihood values from maximum likelihood regression for the two-, three-, and four-state models across all measured genotypes were compared via an Akaike information criterion (AIC) statistical test. For a given model with *k* fitting parameters and a vector *L* of likelihoods over all measured genotypes, $\mathrm{AIC} = 2\mathrm{k} - 2\ln(L)$. The AIC probability of a given *k*-parameter model relative to the model with the minimum AIC value is $e^{(AIC_{min} - AIC_k)/2}$.

Epistasis analysis of riboswitch conformational ensembles

Python scripts for the two-, three-, four-state and apparent binding models as well as energy, concentration, and epistasis calculations for riboswitch conformational ensembles are available upon request from daria.wonderlick@gmail.com.

Glossary

Adenine riboswitch: Small RNA molecule that can bind adenine; in bacteria, riboswitches regulate gene expression at the levels of transcription or translation through conformational changes

Biomolecule: A molecule produced by a living organism; technically nucleic acids, proteins, lipids, and carbohydrates are all classes of biomolecules, but here the term refers specifically to DNA, RNA, and proteins

Conformation: A specific shape or spatial arrangement; here used to refer to an individual structure in a biomolecule's ensemble

Contact epistasis: Direct mechanism for epistatic coupling; mutations make direct contact with each other or are connected by a network of communicating sites

Ensemble: A group of items; here used to describe the group of interconverting structures specified by the sequence of a biomolecule

Epistasis: Phenomenon in which the effect of a mutation depends on the presence of other mutations; can be thought of as coupling between paired mutations

Ensemble epistasis: Indirect mechanism for epistatic coupling; mutations with different energetic effects on different conformations produce nonlinear ensemble redistribution when combined; only applies to biomolecules with three or more populated conformations

Mutation: Sequence change in a biomolecule

Wildtype: natural or reference sequence of a biomolecule; mutations are defined relative to this sequence

2AP: 2-aminopurine; a small fluorescent molecule that can bind to the adenine riboswitch in place of adenine

Bibliography

Leipply, D., & Draper, D. E. (2010). The dependence of RNA tertiary structure stability on Mg2+ concentration: Interpretation of the Hill equation and coefficient. *Biochemistry*, *49*(9), 1843–1853. https://doi.org/10.1021/bi902036j

Leipply, D., & Draper, D. E. (2011). Effects of Mg^{2+} on the Free Energy Landscape for Folding a Purine Riboswitch RNA. *Biochemistry*, *50*(14), 2790–2799. https://doi.org/10.1021/bi101948k

Lemay, J.-F., Penedo, J. C., Tremblay, R., Lilley, D. M. J., & Lafontaine, D. A. (2006). Folding of the Adenine Riboswitch. *Chemistry & Biology*, *13*(8), 857–868. https://doi.org/10.1016/j.chembiol.2006.06.010

Morrison, A. J., Wonderlick, D. R., & Harms, M. J. (2020). Ensemble epistasis: Thermodynamic origins of non-additivity between mutations. *BioRxiv*, 2020.10.14.339671. https://doi.org/10.1101/2020.10.14.339671

Nussinov, R., Tsai, C.-J., & Jang, H. (2019). Protein ensembles link genotype to phenotype. *PLOS Computational Biology*, *15*(6), e1006648. https://doi.org/10.1371/journal.pcbi.1006648

Porter, E. B., Marcano-Velázquez, J. G., & Batey, R. T. (2014). The purine riboswitch as a model system for exploring RNA biology and chemistry. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1839*(10), 919–930. https://doi.org/10.1016/j.bbagrm.2014.02.014

Starr, T. N., & Thornton, J. W. (2016). Epistasis in protein evolution. *Protein Science*, *25*(7), 1204–1218. https://doi.org/10.1002/pro.2897

Stoddard, C. D., Widmann, J., Trausch, J. J., Marcano-Velázquez, J. G., Knight, R., & Batey, R. T. (2013). Nucleotides Adjacent to the Ligand-Binding Pocket are Linked to Activity Tuning in the Purine Riboswitch. *Journal of Molecular Biology*, *425*(10), 1596–1611. https://doi.org/10.1016/j.jmb.2013.02.023

Tyrrell, J., McGinnis, J. L., Weeks, K. M., & Pielak, G. J. (2013). The Cellular Environment Stabilizes Adenine Riboswitch RNA Structure. *Biochemistry*, *52*(48), 8777–8785. https://doi.org/10.1021/bi401207q