Multi-state RNAs and the cocktail party problem

Pablo Cordero, Wipapat Kladwang, Jeehyung Lee, Adrien Treuille, and Rhiju Das

September 26, 2013

Abstract

The ability to change conformations is central to the functions of many non-coding RNAs, ranging from gene regulation to protein synthesis. Fully understanding these changes requires characterization of these molecules' structural ensembles at equilibrium, but this problem has remained experimentally and computationally daunting. To address this challenge, we present the RNA ensemble extraction from footprinting insights technique (REEFFIT), an algorithm that infers ensembles of secondary structures from multi-dimensional chemical mapping experiments, and its integration with Eterna, a cloud biochemistry platform for RNA science. Inspired by the successes in blind source separation (the "cocktail party problem") we developed REEFFIT as a factor analysis model. In a computational benchmark of 20 simulated, noisy data sets for RNAs with 2 to 7 major states, REEFIT accurately recovers within error the population weights, secondary structures, and mutation-induced local perturbations corresponding to each state's contact map, without a set of structures given *a priori*. Experimentally, we show that REEFIT enables the dissection of secondary structures and weights for four RNA systems with single, two, and multiple states. In one case, a bistable hairpin, the REEFIT analysis of new chemical mapping data recover the structures and populations weights inferred from previous 'gold-standard' NMR studies. These results establish REEEFIT and cloud-based multi-dimensional chemical mapping as a uniquely powerful and rapid approach to the RNA structural ensemble problem.

1 Author Summary

RNA is a versatile biological macromolecule that underlies important processes such as gene expression, gene regulation, and structural scaffolding throughout living systems. Most of these roles rely on the ability of RNA molecules to self-assemble into intricate structures. However, it is an oversimplification to think that an RNA molecule folds just into one structure. Intense theoretical and experimental work has established that RNA is a dynamic structural entity that can adopt multiple conformations: an ensemble of structures that are present at equilibrium and whose intercoversions are functionally critical. Although computational algorithms for predicting RNA structural ensembles are available, the verification and dissection of these ensembles experimentally is still limited due to the size, resolution, and infrastructure constraints of state-of-the-art methods such as magnetic resonance and X-ray scattering. We show here that coupling RNA chemical mapping, an already widely used technique to probe RNA structure, to systematic mutagenesis (the mutate-and-map or M^2 strategy) provides a powerful experimental tool to probe ensembles at the single-nucleotide level. We leveraged algorithmic frameworks for the source-separation problem in signal processing and aided by public chemical mapping data sources and new cloud-based technology for synthesis and probing RNA structure online to develop REEFFIT. This factor analysis technique calculates the secondary structural ensemble that most parsimoniously fits the observed patterns of a multi-dimensional mapping experiment. We expect REEFFIT to be the first of a new generation of tools that will take advantage of RNA chemical mapping data at its most fundamental level: by extracting the underlying structural principles governing the observed signals to produce accurate secondary, and perhaps even tertiary, structural hypotheses

2 Introduction

RNA is a versatile biological macromolecule that is heavily involved in gene expression, gene regulation, and structural scaffolding. Many of these roles rely on the ability of RNA molecules to self-assemble into intricate three-dimensional structures. Furthermore, intense theoretical and experimental work has established that interconversion between multiple structures is a pervasive

property of natural and engineered RNAs. Riboswitches and ribozyme can fold into several states to detect and respond to a small molecule [**?**, **?**]; catalyze self-splicing reactions [**?**, **?**, **?**]; carry out protein translation [**?**, **?**, **?**]; and activate logical circuits in cells [**?**, **?**]. Dissecting and re-engineering these transitions depends on modeling the multiple states of an RNA's structural ensemble and their respective equilibrium fractions [**?**]. Determination of structural ensembles, as opposed to single structures, is therefore critical for understanding the mechanisms of natural non-coding RNA [**?**, **?**]. However, data for these ensembles are missing for the vast majority of RNAs.

From a computational viewpoint, structural ensembles have been thoroughly studied using classic dynamic programming techniques devised in the past three decades [**?**, **?**]. However, the practical applications of these methods have been mainly limited to refining minimum free energy secondary structure predictions. In theory, bigger challenges such as riboswitch design and detection of disease-causing mutations can be addressed with these tools, but the lack of experimental data that can validate the estimated structural ensembles has hindered progress in RNA design. Indeed, there are few experimental techniques that can probe structural ensembles. State-of-the-art methods that can capture structural dynamics such as nuclear magnetic resonance (NMR) [**?**] and small angle X-ray scattering (SAXS) [**?**, **?**] require costly infrastructure and focused technical expertise, are limited to small RNA chains, or lack nucleotide-level resolution. Chemical mapping or footprinting is a widely-used biochemical technique in which the RNA is exposed to chemical agents that modify unstructured parts of nucleic acids [**?**, **?**]. These chemical modifications can be detected through a reverse transcription assay in which cDNA fragments can be mapped back to the RNA sequence using electrophoretic separation or deep sequencing [**?**, **?**]. Chemical mapping can probe molecules in solution at nucleotide resolution at least 200 nucleotides at a time, and has long been used to probe entire ribosomes and viral genomes [**?**, **?**, **?**]. The ability to perform this technique in high-throughput allowed us to couple it with systematic mutagenesis: a mutate-and-map (M^2) strategy that can be used to infer the base-paring partners of each nucleotide in an RNA [**?**, **?**]. In these data we have observed that some single-nucleotide mutants have dramatically different chemical mapping profiles in all RNAs studied by M^2 to date, suggesting that these alternative states are prevalent throughout nature.

After proper quantification, chemical mapping yields a reactivity profile, one value per nucleotide[**?**,

?]. If there are multiple conformations sampled by the RNA, the chemical reactivity profile will be the linear combination of the structures present in the RNA's structural ensemble (Figure ??A). All the information about the ensemble is projected into the one-dimensional profile. Due to this projection, recovering the individual reactivity profiles for each state in the ensemble is impossible without further information. However, if the structural ensemble could be perturbed and remeasured enough times, we would then have enough data to estimate it. Here we accomplish these perturbations through systematic mutagenesis. We recently introduced a mutate-and-map (M^2) strategy for RNA structure, whereby a library of single-nucleotide mutants of the RNA sequence of interest is chemically mapped. The signals of resulting from the release of base-pairing partners of the mutated residue in some of the mutants approximate the RNAs "contact map" and can be used to guide secondary structure algorithms. Other mutations induce drastic changes in secondary structure, stabilizing structures that could potentially be excited states in the wild type ensemble. Therefore, we can argue that each mutant's reactivity profile in M^2 experiments is the result of a linear combination of the same structures present in the wild type (Figure ??B) as well as mutant-specific structures.

The challenge to recover each state's reactivity profile and its weights can be cast as a source separation problem, a classical topic in signal processing [?, ?]. The source separation problem is often described by the "cocktail party" analogy [?]. In the analogy, we are placed in a cocktail party with several people speaking at the same time and are challenged with finding the waveforms corresponding to the speech audio of each of the attendees, as well as their relative volumes, given audio recordings by microphones located in different corners of the room. When the number of microphones matches or exceeds the number of attendees, the mixed sources can be recovered robustly using automated methods. Recovering RNA structural ensembles from M^2 data is analogous: each attendee corresponds to a particular structure and each microphone recording to the observed reactivity profile of each mutant. Using this insight, we developed a method that models M^2 data as a linear combination of structures to estimate the structural ensembles of the wild type and all the single nucleotide mutants in the experiment. This analysis, which we have named RNA ensemble extraction from footprinting insights technique (REEFFIT), takes advantage of the variation in the reactivity profiles of each mutant caused by the stabilization and destabilization of structures in

the ensembles induced by the mutation.

REEFFIT uses a factor analysis framework to model the reactivity profile of each mutant in terms of the structures in the ensemble: it finds the expected reactivity profile of each structure, the convex combination of weights that best model the data, and sequence-position-wise noise levels. To achieve this, we have generalized the standard, gaussian factor analysis to include heterogeneous, non-gaussian priors for the expected reactivities that depend on both structure and sequence position. Local perturbations due to mutations that cannot be modeled by large structural rearrangements, such as the release of base-pairs induced by the single nucleotide change, are also included in the statistical model. The resulting model is more difficult to fit to the data than gaussian factor analysis since it does not accept closed-form solutions for some of the steps in the fitting algorithm and a stochastic simulation bayesian inference method must be employed.

REEFFIT is furthermore instrumentated to handle cases in which the set of structures in the ensemble is not known *a priori*. REEFFIT forms an initial set of structures by clustering the suboptimal structures predicted for the mutants, picking the best representative of each cluster given the data, and subsequently adding more structures that improve the value of an information score which penalized the inclusion of too many structures.

[THIS NEEDS TO BE REVISITED: The purpose of this manuscript is to give a detailed description and validation of the statistical framework and technique. We report a proof of concept of the technique by validating our method using simulated and experimental data. For our experimentally probed systems, we analyze previously published model sequences that fold into stable and bistable hairpins, as well as newly designed systems with more heterogeneous ensembles containing multiple excited states. For the purpose of clarity and following the example of others when introducing new methodologies [**?**] [NEED OTHER CITATIONS, WHAT OTHER MODEL HAIRPIN PUBLICATIONS ARE THERE?], we use the thorough analyses on these small systems to present and discuss the technique in order to lay a foundation for further explorations of biological RNAs of interest. — NEED TO ADD CLOUD STUFF, WHERE DOES IT GO?]

3 Results

3.1 Simulated data evaluations: an *in silico* benchmark

Evaluating REEFFIT's performance is challenging since we have to simultaneously compare the weights of the predicted and data-generating structures and also, at the same time, to assess the model's goodness of fit to the data. We therefore chose to first carry out a benchmark to test REEFFIT on simulated data for which accuracy estimates would be unambigouus. This simulated benchmark consisted of representative RNA sequences for 20 Rfam families [**?**] that include microRNAs, small nucleolar RNAs, and viral sequences. We simulated M^2 chemical mapping data using sets of structures drawn from the suboptimal structures of each RNA (see Materials and Methods) The simulated datasets are shown in Figures S1B to S20B, one for each test cases. Here and throughout the article we show the quantified reactivity data for M^2 experiments. Each simulated dataset in our benchmark was designed to test different scenarios that REEFFIT may encounter in real datasets and included systems with various degrees of thermodynamic stability: from simple systems with ensembles dominated by one structure (e.g. the $3'$ conserved hairpin of the eel long interspersed element UnaL2 of family RF00436) to multi-stable RNAs containing multiple excited states (such as the small nucleolar RNA sR4 of family RF01125). Test cases with minor variations on one predominant structure across most mutants included the alfalfa mosaic virus RNA 1 $5'$ UTR stem-loop (RF00196), the $3'$ conserved hairpin of the eel long interspersed element UnaL2 (RF00436), and the microRNA mir-572 sequence (RF01020). Although we used 3 and 4 structures to simulate these datasets (for RF00196 and RF00436, RF01020 respectively), in every case all states are minor variations of one structured hairpin (see Figures X, Y , and Z) that is evident by the punctate "cross-diagonal" pattern induced by the mutations. The small cajal body specific RNA 16 (RF00424) and the leader sequence of the ribosomal protein L13 (RF00555), both simulated with 3 structures, provided test cases with a similarly predominant structure motif but with more pronunced structural variation, simulating a scenario where a simple one-helix system folds into a rich ensemble of structures with small differences. Extreme cases of this scenario were portrayed in the the 4-state microRNA mir-472 (RF01020) and the 3-state avian HBV RNA encapsidation signal epsilon (RF01313). Other test cases included systems in which the wild type

sequence folds into one predominant structure, but the single-nucleotide mutants alternate between 2 to 4 states, and were designed to test whether REEFFIT could accurately estimate the ensembles of the wild type sequence as well as all mutants. These test cases included the class I RNA from *D. discoideum* (RF01414, 2 states) and the small nucleolar sR40 (RF01297, 2 states), SNORD116 (RF00108, 3 states), SNORD64 (RF00570, 3 states), and snoR4a (RF01301, 4 states). Challenging test cases of RNAs with multiple excited states and very sensitive to mutations included U7 nuclear RNA (RF00066), and the hairpin ribozyme (RF00173), both with 5 structures, as well as the 7-state small nucleolar RNA sR4 (RF01125). Some of the datasets were also designed to test REEFFIT's modeling of the mutation-induced local perturbations. For example, in RF00436 the mutation perturbations can be clearly seen as punctate patterns in the simulated data. These local patterns disappear when most of the mutants change structure drastically, as in RF00066. As a first test of our ensemble extraction method, we used REEFFIT to fit the data to our statistical model with the structures that were used to generate the data (see Materials and Methods). Goodness of fit metrics for factor analysis have been extensively studied in psychometrics under the umbrella of structural equation modelling [**?**, **?**, **?**]. In line with these studies, we report χ^2/df and root mean square error of approximation (RMSEA) as goodness of fit metrics for the fitted models to each simulated dataset. The general low RMSEA values and χ^2/df indicated that the structures proposed by REEFFIT fit the data well: no test case had a χ^2/df value over 2 (see Table X).

We then tested the full REEFFIT pipeline by running our structure set selection algorithm and fitted the resulting model to the data. Because REEFFIT uses a conservative information score to obtain a set of structures to model the data, in many cases it inferred less structures than those used to generate the data. In these cases (such as the RF01125 sequence, see Table **??** and Figure **??**) the generating structures are too similar to each other and can be accounted for by taking just one of the structures as a representative. Because REEFFIT selects the states in the ensemble from clusters of a multitude of suboptimal structures, we can compare the weights of these extra generating structures to our predicted ones by grouping them according to REEFFIT's clusters. Therefore, instead of comparing predicted and original structures, we compare predicted structures, that serve as representatives of each cluster, and clusters of original structures. The structure weights of each group of original structures is then defined to be the sum of the weights

of all the structures in each group. A similar evaluation scheme can be taken in the case when there are more predicted structures than the generating structures (such as RF00108). In these cases we compare predicted structures with clusters of generating structures. In the same way as before, the weight for each cluster of predicted structures is the sum of the weights of the generating structures in that cluster. When executing the full pipeline without the structures known *a priori* REEFFIT again yielded general low RMSEA values and χ^2/df: only the RF00108 sequence, the only case that was modeled with more structures than it was generated with, gave a χ^2/df value over 2. Using this validation scheme, REEFFIT was able to recover the ensemble weights within one standard deviation in systems with 2 to 5 structures on an average of 47% of the time (see Table **??**). Furthermore, all of REEFFIT's selected structures recovered the helix topologies of at least on of the original structures in 15 of the 20 test cases. For RF01414, a two-structured system whose mutants alternate between two states, REEFFIT was able to recover the structural ensemble at helix-wise accuracy exactly and their respective fractions match the original weights within one standard deviation 67% of the time. REEFFIT captured the main aspects of the RF01414 ensemble: it predicted the wild type sequence to fold predominantly in one state while the rest of the mutants switch back and forth between two structures. The local perturbations, apparent in the shared helices located at the 5' and 3' ends, are also adequately captured in the fit. The RF01274 test case provided a more complicated ensemble, comprised of 4 structures with two main helices shared in three of the states. REEFFIT modeled this ensemble using only two structures, which approximately match two of the original states. However, even without the two additional structures, REEFFIT was able to estimate the ensemble weights with high confidence (80% of the original weights, grouped into two clusters, were at one standard deviation from the predicted weights) and capture the mutation-induced switches and local perturbations of the helices shared by the states. An even more complex ensemble was given by RF01125. A total of seven structures form this ensemble: 4 of them are variations on a two helix system, while the remaining three have almost no helices in common (see Figure **??**A). REEFFIT modeled this dataset with 4 structures: three of which correspond, with minor variations, to the three disparate structures, and a remaining structure containing the two helices shard by 4 of the original states. Therefore, REEFFIT grouped 4 structures into one, and was able to recover the weights reasonably well (64% of the original

weights, grouped into two clusters, were at one standard deviation from the predicted weights) (Figure **??**B). Interestingly, the χ^2/df value for the fit is high (3.6), indicating that not all features were captured by the model. Indeed, the χ^2/df value serves a conservative indicator to detect when REEFFIT was unable to adequately select the the structures and/or fit the data properly. For example, for a two-structured system with one predominant structure such as RF01297, REEFFIT was able to recover the structures in the ensemble exactly at helix-wise accuracy, but failed to recover its weights robustly (only 9% of the original weights were at one standard deviation from the predicted weights). Conversely, in RF00066, REEFFIT was not able to recover any of the original structures helix-wise, but is evaluated favorably when assessing weight recovery accuracy (73% of the grouped weights match the original weights within one standard deviation). In this and other cases where REEFFIT failed either to confidently recover either the structures in the ensembles or their respective weights, high χ^2/df and RMSEA values served as indicators for model mis-fitting and helped guard against false ensemble inference (type I errors).

All of the analyses described above were performed using the RNAstructure package to obtain starting weights for the fits (see also Materials and Methods). REEFFIT, like other factor analysis methodologies, uses a numerical EM approach to maximize the likelihood function and is only guaranteed to find a local maximum. If the likelihood function that is highly non-convex with multiple local maximums, these methods are sensitive to initial conditions. Therefore, to assess robustness to initial conditions, we repeated our analyses on the Rfam benchmark using ViennaRNA package to calculate the initial weights for each structure. The fits given by ViennaRNA were in agreement with the RNAstructure results using the same structures to fit the data. We also tried using ViennaRNA to select the structures in each tested ensemble. [THIS KIND OF SOUNDS LIKE WE'RE CHEATING BY USING RNASTRUCTURE TO BOTH SIMULATE AND MODEL SELECT; PLUS, I DON'T KNOW WHAT TO MAKE OF IT] However, the correspondence of the selected structures compared to the original structures and the quality of the fits fell drastically.

3.2 Experimental evaluations

Encouraged by REEFFIT's performance in the *in silico* benchmark above, we proceeded to test REEFFIT with experimental data. Since an experimental "gold standard" for directly and confidently probing structural ensembles of RNA molecules remains elusive, we focused on analyzing simple systems in which the true nature of the ensemble can be easily rationalized. The model hairpin systems used here have been previously studies by us and others using the M^2 methodology and NMR chemical shifts, providing further experimental cross-checks. We tested our method on chemical mapping data using the two main high-throughput read outs for these assays: the capillary electrophoresis and deep sequencing platforms.

3.3 Evaluation using M^2-seq on the MedLoop hairpin systems

To test our method using deep sequencing, we obtained M^2-seq measurements for a benchmark of two RNA sequences of our own design using the Eterna expert interface (see Materials and Methods): the MedLoop, MedLoopΔ. We have previously studied the MedLoop RNA in a proof-of-concept test case for the M^2 methodology [**?**]. The wild type MedLoop forms a stable 10 base-pair hairpin. Most mutations either do not disrupt the derived chemical mapping profile or enhance the chemical reactivity of the mutated nucleotide and its base-paring partner, providing a "contact map" for the RNA. Some mutations that disrupt that helix force it into an alternative stem (see Figure **??**A). According to REEFFIT, the MedLoop M^2 data can be modeled with these two structures: a 10-base pair helix (which we will denote by mlp_1) and an alternative 3′ hairpin (mlp_2). As expected, mlp_1 (Figure **??**A, blue) is the predominant state in most of the constructs, including the wild type sequence, while mlp_2 (green) becomes dominant when the mlp_1 helix is destabilized by mutation. Interestingly, mutations in the loop of mlp_1, such G11C and A17U, seem to destabilize the structure, resulting in a mixture of the two states. Furthermore, mutating the 5′ part of the mlp_1 helix results in more destabilization of the structure than mutating the 3′ side (compare the weights of each structure e.g. A5U to U31A). This is expected since the helices of mlp_1 and mlp_2 share the 3′ range of nucleotides for that helix (G26 to G34); mutating these residues destabilizes mlp_1 but produces a weaker version of mlp_2, resulting in a mixture of states.

The MedLoopΔ construct differs from MedLoop in that it lacks five nucleotides in the $3'$ end, which abolishes the alternative stem. REEFFIT models the MedLoopΔ M^2 data with three structures: mlp_1 (blue in Figure **??**A, same structure as mlp_1 for the MedLoop RNA), mlp_1' (green), and mlp_3 (red). The mlp_1' structure is a variation of the mlp_1 structure of the MedLoop RNA but presenting an extra helix that is made possible by the 5-nucleotide deletion. In the wild type sequence, mlp_1 is the main structure with a predicted mean weight of 1. Across the rest of the mutants, mlp_1 and mlp_1' alternate as the dominant structure. When the main helix of mlp_1 and mlp_1' is destabilized (e.g. in the C3G mutant, see also the I and II features marked in Figure **??**), mlp_3 appears in the structural ensemble, but it is never stabilized beyond a weight of 0.5.

3.4 A 45-nucleotide RNA with several excited states

The final M^2-seq test case was the M-stable construct, an RNA whose minimum free energy structure is predicted to be a simple hairpin with a tetraloop (see orange structure, Figure **??**A, which we will denote as mst_1). The minimum free energy structure is deceptively simple: the structural ensemble of the construct is predicted to have at least 2 other structures 1 kcal away from the mst_1. This unstable ensemble is susceptible to change drastically by mutation and therefore presents a challenging case for REEFFIT. The M^2-seq measurements clearly indicate the presence of more than three structures when all the mutants are taken into account. REEFFIT models the M-stable data with 4 states (see Figure **??**A). Interestingly, each alternative structure (blue, green, and red; denoted as mst_2, mst_3, and mst_4, respectively) has at least one non-cannonical base pair – each of these structures were found in the set of suboptimal structures of a mutant. Nevertheless, REEFFIT can model the data well with these four states. Weights for the wild type include structures mst_2 and mst_3. In particular, mst_2 is predicted to be as stable as mst_1. These two structures share the same hairpin, but structure mst_2 has two helices surrounding it, separated by a G10 bulge. Evidence for mst_2 in the wild type profile can be seen in nucleotides A8 to C14 and G39 to A45, where the extra helices of mst_2 partially protect those residues. Structure mst_3 also appears in the wild type ensemble, but only in a 0.15 fration – the only evidence of its presence in the wild type profile being the increased reactivity at positions A30, A31 and U44, A45. In the rest of the

mutants, these structures change weights drastically, but mst_1 and mst_2 maintain dominance as made evident by the only punctate feature of the data: the cross-diagonal pattern of the shared helix of mst_1 and mst_2. We note that all structures are predicted to be in the wild type ensemble by RNAstructure, although mst_1 and mst_4 are predicted to be present in less than 10% of the total population.

3.5 Experimental evaluation using capillary electrophoresis

Previously, Hobartner and colleagues used imino proton NMR spectra to estimate the ensemble weights of two structures in a bistable RNA hairpin which we call BST (Figure **??**A) [**?**]. The simplicity of this system allowed the authors to decompose its NH...N 1H NMR spectrum into a weighted sum of the NMR spectra of each hairpin without the need of resonance assignment and characterize its ensemble. These gold standard measurements make the BST an excellent test case for REEFFIT. We synthesized this RNA and obtained M^2 measurements for this system using capillary electrophoresis as a read-out (see Materials and Methods). Visual inspection of the reactivity profile for the wild type and the global rearrangements occurring in most mutants hint of the bistable nature of this RNA (Figure **??**C). Unless perturbed, this bistable RNA folds predominantly into a haripin with its helix in the $3'$ end (blue structure in Figure **??**), which we will denote as bst_1. When the bst_1 helix is perturbed by mutation, the RNA switches to a $5'$ helix structure (green, denoted here as bst_2). Some constructs in this M^2 experiment, e.g. G8C, have a mixture of these two structures in their ensembles. As in the MedLoop M^2 data, the two states share a region of nucleotides in their respective helices (G11 to U13) and therefore the mutations have asymmetrical effects: mutations affecting this region have little effect in the ensemble since they destabilize both structures. In contrast, mutations in the $3'$ end destabilize bst_1 significantly, revealing bst_2. The wild type RNA is a mixture of these two states, bst_1 is present 70% $\pm X$ of the time and bst_2 30% $\pm X$. These estimated weights are in agreement with the previously reported fractions by Hobartner et al. (Figure **??**B). As a final test, combining the profiles of U4A and C23G, mutants that are predicted by REEFFIT to stabilize bst_1 and bst_2 respectively, result in a profile that follows closely the wild type reactivities (Figure **??**).

4 Discussion

Probing the structural ensembles of RNA molecules is a formidable task that is essential to understand their thermodynamics and can give important insights in their biological functions [**?**]. Current experimental techniques used to probe these ensembles require significant infrastructure investment, technical expertise, and are limited to small RNA sequences. We presented a novel strategy based on easily obtainable chemical mapping measurements that can be used to address this challenge. By taking advantage of the ensemble perturbations induced by single-nucleotide mutations in M^2 experiments, we have cast the RNA structural ensemble challenge as a source separation "cocktail party" problem [**?**]. The resulting algorithm that solves this problem, REEFFIT, extracts structural ensembles out of multi-dimensional chemical mapping experiments. Because of its underlying general factor analysis framework, REEFFIT is easily extensible and provides a basis for future modeling of chemical mapping data. Importantly, REEFFIT is not limited to M^2 experiments: the only requirements are that all reactivity profiles in the dataset are perturbed versions of approximately the same ensemble of structures and that the number of profiles is at least as large as the number of the most prevalent structures in the ensemble. It is therefore theoretically possible to use sources of perturbations other than mutations such as changing ionic concentrations [**?**], temperature, or concentration of small molecule binders [**?**], or macromolecule partners [**?**].

For M^2 experiments, REEFFIT can be used to tease out the observed changes due to structural ensemble perturbations and the local perturbations induced by the mutations themselves. This enables extracting additional insights from this information-rich strategy and helps deconvolute global structural rearrangements and the embedded "contact map" of the RNA that we have previously used to refine secondary structure predictions [**?**].

REEFFIT's statistical model can be easily extended to use different priors for the structure reactivity profiles due to the use of MCMC. For example, instead of classifying nucleotides in paired and unpaired states, we could group them by secondary structure element (hairpin, edge base-pair, bulges, interior loops, etc.) and assign different reactivity distributions for each element, analogous to what we have done to calculate $RMDB_U$ and $RMDB_P$. In fact, following this rationale, it would also be possible to separately model tertiary structures using an appropriate

structure space sampling scheme and known reactivity patterns of tertiary motifs such as kink turns and tetraloop/receptors [**?**, **?**, **?**, **?**]. This may become a reality in the near future as we obtain enough chemical mapping measurements in order to robustly estimate the prior distributions of each three-dimensional structure motif.

We expect REEFFIT to be used routinely to analyze multi-dimensional mapping experiments, not only to dissect the complex ensembles of various biologically-relevant molecules such as the structure fractions of riboswitches in the absence of ligand, but to detect systematic flaws in current secondary structure energy functions and provide insights for the design of RNA circuitry. The commoditization of sequencing technologies will further expand the use of RNA chemical mapping for structure probing and REEFFIT will help in the dissection of these data and push forward a view of that includes experimental probing of structure thermodynamics that goes beyond the implicit dogma of "one sequence, one structure".

5 Materials and Methods

5.1 The REEFFIT statistical model

Given a set $S = \{S_1, S_2, \ldots, S_t\}$ of secondary structures with reactivity profiles $D = \{D_1, D_2, ..., D_t\}$, REEFFIT models the observed reactivity profiles, D^{obs} as linear, convex combinations of the reactivity profiles of D plus white noise (see Figure **??**B). D is a set of hidden variables since the isolated reactivity profiles of each structure are not available. However, we can impose a prior on each of these hidden profiles depending on their corresponding secondary structure. Each D_s can be decomposed into a set of univariate random variables D_{si}, one for each nucleotide $i = 1, 2, ..., n$, where n is the number of nucleotides (here and throughout the text, D will be treated as an $m \times n$ matrix, with rows denoted as D_s and columns as D_i, where s is a structure index and i is a position index). Because low and high chemical reactivities are correlated with structured and unstructured states, a reasonable prior would force D_{si} to be small if i is paired in structure S_s and higher if it is unpaired. To get distributions of the reactivities of paired and unpaired nucleotides, we compiled all the SHAPE reactivity data in the RNA Mapping Database (RMDB) [**?**] of RNAs with known

crystallographic structure and modeled these distributions in terms of mixtures of gamma distributions as described previously [**?**]. Let $RMDB_U$ and $RMDB_P$ be the modeled RMDB unpaired and paired distributions respectively, we can then define a prior likelihood for D_s as:

$$\begin{aligned} RMDB_s(x) = \prod_i (&RMDB_U(x)\mathrm{I}\left[i \text{ is paired in } S_s\right] \\ &+ RMDB_P(x)\mathrm{I}\left[i \text{ is unpaired in } S_s\right]) \end{aligned} \tag{1}$$

where I is the indicator function (1 or 0 if the condition is satisfied or not, respectively). Because these distributions were estimated using SHAPE data, any statistical model based on them will only work properly in SHAPE-based chemical mapping experiments. To handle other modifiers, e.g. dimethyl sulfate, $RMDB_P$ and $RMDB_U$ can be replaced with the respective estimated distributions using unpaired and paired data of that modifier [**?**] and will be reported elsewhere. After defining our priors for each structure profile, we can write the REEFFIT model (see also Figure **??**):

$$\begin{aligned} D_{ji}^{obs} &= \sum_{s \in \text{structures}} W_{js} D_{si} + \epsilon_{ji} \\ D_s &\sim RMDB_s \\ \epsilon_{ji} &\sim \mathcal{N}(0, \Psi_i) \\ \sum_{s \in \text{structures}} W_{js} &= 1, \forall j = 1, \ldots, m \end{aligned} \tag{2}$$

Here, D_i^{obs} and D_i are the column vectors of D indexed by nucleotide position, W_{js} is the weight for the profile of structure s in the observed profile $j = 1, 2, ..., m$ with m the number of mapping profiles in the experiment. The column variable ϵ_i is centered white noise with a diagonal covariance matrix $\Psi_i I_m \in \Re^{m \times m}$ where Ψ_i is a scalar and I_m is the $m \times m$ identity matrix. Our model differs from the standard factor analysis model [**?**] in two ways. First, we have n rather than one ϵ variable to capture experimental noise. This derives from our observations that different sequence positions have different experimental error distributions depending on the intensity of the reactivity and proximity to the beginning and end of the sequence [**?**]. Second, different sequence positions

are assigned different prior distributions that depend on the pairing state of each structure in the model. It is important to note that here, as in standard factor analysis, all covariance matrices for the noise must be diagonal, that is, the measurements are independent of each other [**?**, **?**]. This assumption holds in the case of multiple chemical mapping measurements, since each measurement is carried out in different capillaries (in capillary electrophoresis) or based on poisson distributed counts derived from separate single molecules (in deep sequencing).

There is no closed form maximum likelihood estimation for this model and therefore we used an expectation maximization (EM) algorithm to obtain expected values for the hidden variables (in our case, D), and maximum likelihood estimates for W [**?**, **?**]. In standard factor analysis, the E-step can be obtained in closed form by calculating the sufficient statistics for the likelihood function, which happen to be the first two moments of the posterior distribution of D: $E[D|D^{obs}]$ and $E[DD^T|D^{obs}]$. Unfortunately, the non-gaussian form our priors for each D_s precludes us from having a closed form for these statistics and we must calculate them using Markov Chain Monte Carlo (MCMC) [see next section]. For the M-step, closed form solutions depending on these sufficient statistics exist for W. Additionally, we incorporate convex combination constraints on W by casting it as a quadratic problem and solving numerically (see Materials and Methods).

5.2 Maximum likelihood estimation of the REEFFIT model

Given the REFFIT factor analysis model (**??**) we want to calculate maximum likelihood estimates for W and each Ψ_i given the hidden variables D. We will use an expectation maximization (EM) algorithm to calculate these maximum likelihood estimates given expected values of the hidden variables. Let $R = -\frac{nm}{2}\log(2\pi)$, then the log-likelihood function L can be written as:

$$
\begin{aligned}
L(W,\Psi) &= \sum_i^n \log \frac{1}{(2\pi)^{m/2)^T}|\Psi|^{1/2}} \exp(-1/2(D_i^{obs} - WD_i)1/\Psi I_m (D_i^{obs} - WD_i)) \\
&= R - \frac{1}{2}\sum_i^n -\log|\Psi_i I_m| + ((D_i^{obs})^T 1/\Psi I_m D_i^{obs} - 2(D_i^{obs})^T 1/\Psi I_m W D_i + D_i^T W^T 1/\Psi_i I_m W D_i) \\
&= R - \frac{1}{2}\sum_i^n -m\log(\Psi_i) + ((D_i^{obs})^T 1/\Psi_i I_m D_i^{obs} - 2(D_i^{obs})^T 1/\Psi_i I_m W D_i + Tr[W^T 1/\Psi_i I_m W D_i D_i^T])
\end{aligned}
\tag{3}
$$

, where $Tr[.]$ is the trace operator. For the E-step, we take the expectation of L conditioning the hidden variables D with respect to the data:

$$
\begin{aligned}
E[L|D^{obs}] = R - \frac{1}{2}\sum_i^n &-m\log(\Psi_i) + ((D_i^{obs})^T 1/\Psi_i I_m D_i^{obs} - 2(D_i^{obs})^T 1/\Psi_i I_m W E[D_i|D_i^{obs}] \\
&+ Tr[W^T 1/\Psi_i I_m W E[D_i D_i^T|D_i^{obs}])
\end{aligned}
\tag{4}
$$

We thus have that the sufficient statistics of L are the first two moments of the posterior distribution of D: $E[D_S|D^{obs}]$ and $E[DD^T|D^{obs}]$. Because of the forms of our priors (mixtures of gamma distributions, see [**?**]), these sufficient statistics cannot be written in closed form. We therefore estimate them using a Markov Chain Monte Carlo algorithm with an adaptive step rule implemented in the pymc python library for bayesian statistics [**?**]. MCMC can be computationally expensive, especially when used in combination with EM. Because we are assuming no positional correlation between reactivities, the MCMC simulations can be performed independently. Therefore, in our implementation we parallelize these position-wise MCMC calculations using the python joblib parallelization library.

For the M-step, given the sufficient statistics above, we calculate W by maximizing L with respect each variable, enforcing the convex combination constraints in (**??**). We can cast the minimization of L with respect to W as a quadratic program; let W_j be the j-th column of W, indexed by measurement/mutant, then for each chemical mapping measurement j:

$$\begin{aligned} &\text{minimize } (\sum_{i}^{n} \frac{1}{2\Psi_i}) W_j^T E[D_S D_S^T | D^{obs}] W_j - (1/\Psi_i I_m (\sum_{i}^{n} D_{ji}^{obs} E[D_S | D^{obs}]_i))^T W_j \\ &\text{subject to } \sum_{s}^{t} W_{js} = 1, \forall j = 1, \ldots, m \end{aligned} \tag{5}$$

To calculate W, we solve the resulting quadratic programs using the CVXOPT python library [**?**]. In standard factor analysis the single noise covariance matrix is re-estimated in each EM iteration. However, because we have different covariance matrices for each sequence position, we are underpowered to estimate these from the data. We therefore do not re-estimate each Ψ_i and maintain its initial value.

where diag[.] is the diagonal operator to constrain Ψ to be a diagonal matrix. The final estimates of W and the expected reactivities $E[D_S|D^{obs}]$ are calculated until the value for L is stabilized and does not substantially improve.

Because of the hidden reactivities D_S, L need not be convex and may have multiple local maxima. The EM algorithm simply reports one of these local maxima and is therefore sensitive to initial conditions. For each Ψ_i we choose the empirical variance of position i across all chemical mapping measurements, consistent with the variance calculation performed when using M^2 z-scores as pseudo-energy bonuses for secondary structure prediction [**?**]. For W, we use the dynamic programming algorithm in RNAstructure to calculate the energies of each structure in each mutant. For mutant j, let ΔG_{js} be the RNAstructure-calculated energy for structure i, k_B the Boltzmann constant, and T the temperature at which the experiments were performed, then the initial value W_0 at position (j, s) is:

$$W_{0,js} = \frac{\exp(-\Delta G_{js}/(k_B T)}{\sum_{s'} \exp(\Delta G_{js'}/(k_B T))}$$

To calculate uncertainties for W, we use a Fisher's information matrix approach. Usually, the information matrix for W would take the form:

$$\mathcal{I}(W) = -E\left[\frac{\partial^2 L}{\partial W^2} \mid W\right]$$

For expectation maximization, in the case of hidden variables, it has been shown that $\mathcal{I}$ can be estimated as follows [**?**]:

$$\mathcal{I}(W) = -\frac{\partial^2 E[L|D^{obs}]}{\partial W^2} \mid_{W=W^*}$$

with W^* the maximum likelihood estimate obtained from the EM algorithm. Throughout this paper, we report uncertainties for W in the form of $\sigma_W = \frac{1}{\sqrt{\mathcal{I}(W)}}$. Uncertainties for the expected reactivity profiles, $E[D_S|D^{obs}]$, are reported as standard errors resulting from the MCMC simulations:

$$SE_{D_i} = \frac{\sigma_{D_i}}{\sqrt{nsim}}$$

Here, σ_{D_s} is the sample standard deviation from the MCMC trace for the reactivity values of s at sequence position i and $nsim$ is the number of samples used in the simulation calculations. Finally, uncertainties for the predicted data $W^*E[D|D^{obs}]$ are calculated by propagating the uncertainties $\mathcal{I}(W)$ and $SE^2_{D_i}$.

5.3 Handling local perturbations

Systematic experimental perturbations used to alter the RNA's structural ensemble may induce local changes that cannot be captured by the linear combination of the weights W and the expected reactivities $E[D|D^{obs}]$. This is the case in M^2 experiments, where mutations induce local perturbations in the underlying reactivities of each structure. To model these perturbations, we add a set of random variables to our model: $C = \{C_{sji}\}$ for all structure s, mutant j, and sequence position i where i is at most one nucleotide away from the site of a mutation in mutant j or from a base pair that would be disrupted in structure s in mutant j. These variables take the values of the change in reactivity of D_S at perturbed positions that are needed to account for the data. We include C in the MCMC calculations used to estimate the sufficient statistics, setting their prior distribution to the distribution of differences in reactivities in M^2 experiments available in the RMDB, which we call $RMDB_\Delta$ and that we model with a Laplace distribution. The MCMC simulations yield the

expected values of C given the data: $E[C|D^{obs}]$. In all likelihood calculations and in the quadratic problem for W (**??**) we incorporate $E[C|D^{obs}]$ by adding $E[C_{sji}|D^{obs}]$ to $E[D_{si}|D^{obs}]$ in the relevant operations involving mutant j.

5.4 Marking missed data points

To detect where our model fails to predict observed values we use the estimated noise levels in Ψ. For each data point D_{ij}^{obs} we can calculate its probability given our model:

$$P(D_{ij}^{obs} \mid W, \Psi, E[D_S|D^{obs}]) = \mathcal{N}(WE[D_S|D^{obs}]_{ij}, \Psi_{ii})$$

Any data point whose probability given our model is below 0.05 is marked as missed (see Figures **??**, **??**, and **??**, red squares).

5.5 Initial structure set selection procedure

In most realistic scenarios it is not known *a priori* what set of structures would best model the data. To select an initial set of structures, we calculate a set of suboptimal structures for each sequence in the multi-dimensional chemical mapping experiment; for M^2 experiments, the suboptimal structures of all mutants involved are taken into account. We then proceed to cluster the structures using hierarchical clustering with a probabilistic distance metric and a Calinski-Barabasz cutoff (see Materials and Methods and Figure **??**A). Taking the structures in ensembles of all mutants allows for a broad initial search of structures. We select structure medoids for each cluster by taking the one that is found to be most stabilized in the wild type sequence, according to the RNAstructure package. Finally, structures are scored by their maximum observed stability in all mutants and are added in a greedy manner by minimizing the Akaike information criterion which penalizes model complexity to avoid overfitting [**?**] (see below and Figure **??**B).

Theoretically, to find the best structures that model the data we could enumerate all possible combinations of structures, fit model (**??**) to each combination, calculate some penalized information score, and choose the combination with the best score. In practice, this method is intractable for a large set of structures due to the exponential number of combinations of struc-

tures. Furthermore, the quadratic nature of the likelihood function does not allow a nested decomposition required for search space reduction techniques such as dynamic programming. Instead, we choose to explore structures that are most representative of the combined set of suboptimal structures. To achieve this, we perform hierarchical clustering using a distance metric that takes into account the position-wise *a priori* reactivity distributions of each structure S_s (in our case, either $RMDB_U$ or $RMDB_P$). For a nucleotide position i, let the probability distribution $RMDB_{S_{is}}(x) = RMDB_U(x)I\left[i \text{ is paired in } S_s\right] + RMDB_P(x)I\left[i \text{ is unpaired in } S_s\right]$, and JS the Jensen-Shannon distance between distributions; we define the following distance metric:

$$d(S_s, S_r) = \sum_{i}^{n} JS(RMDB_{S_{si}}, RMDB_{S_{ri}})$$

Performing hierarchical clustering using d as our metric of choice clusters the structures given the nature of their position-wise prior information. Currently, because we are only considering paired and unpaired states, clustering using this metric yields the same results than clustering using the trivial metric that counts the number of positions where S_s and S_r differ. Nevertheless, we decided to use this metric in our implementation to allow for future extensions to the method, where more base pair states than only paired and unpaired are taken into account.

To pick a cutoff for the dendrogram resulting from the hierarchical clustering above to generate structure clusters we use the Calinski-Harabasz index (CH) that has been previously used to cluster RNA secondary structures [**?**]. We choose a dendrogram cutoff that results in a set of clusters H that maximize CH. Let $m_c, \forall c \in H$ be the medoids of each cluster, i.e. the structures with minimum average distance to all cluster members for each cluster, we select a set of structures using the following steps:

1. Structure set initialization: We start with the set $S = m_c \mid c \in H$.

2. Medoid swap step: We define a data score for each structure s as

$$score(s) = \max_{j=1,\ldots,m}(RMDB_s(D_j^{obs}))$$

For each cluster $c \in H$, we change its selected medoid m_c to the structure with maximum

score in the c. We then fit the model (**??**) to the new S.

3. Greedy structure addition step: Let S' be a sequence of all structures that are not in S, that is, that were not selected as medoids by the two steps above, sorted by *score*. Using the maximum likelihood estimates for W from the step above as starting values, we perform several EM iterations. In each EM iteration we incorporate the next structure given in the sequence S' and accept the new solution if it decreases the corrected Akaike information criterion $AICc = 2MC + \frac{2MC(MC+1)}{m-MC-1} - 2L$, where is the number of parameters in the model that reflects model complexity.

4. Model refitting: Using the resulting set of structures S from the above steps, we re-initialize W and re-fit the model.

The *score* function defined above intends to favor structures that are maximally supported by some chemical mapping measurement in the data. For example, a mutant in M^2 experiments may drastically stabilize a structure s not seen in other mutants. Its $score(s)$ will be high in this situation and we therefore expect for it to be included in the selected structures S that are needed to account for the data. We have observed that it becomes readily apparent if the structure set needs to be expanded in the first few iterations of the greedy structure addition step: in cases where more structures were needed, the first few iterations almost always see an increase in $AICc$, whereas if the set of structures contained enough structures to model the data the $AICc$ would significantly worsen because of the very small increase in likelihood. This makes REEFFIT a conservative algorithm when selecting the structure set: it always will prefer simpler models when possible, grouping together similar structures and just taking one as their respresentative. The clusters calculated by REEFFIT have thermodynamical significance, since usually the weights for the selected medoid of the cluster will correspond to the sum of weights of the structures in the cluster (see simulated data results above).

5.6 M^2 of the Hobartner bistable RNA

The Hobartner bistable RNA and its complementary single-nucleotide mutants were constructed using PCR assembly, in vitro transcription, and probed with 1M7 as described previously. Briefly,

an assembly consisting of 4 primers was designed to assemble the construct by PCR (see Table **??**). DNA was purified with AMPureXP beads (Agencourt, Beckman Coulter) and in vitro transcribed for 3 hours. The resulting RNA was purified with AMPureXP and folded in 50 mM NA-HEPES pH 8 and 10 mM $MgCl_2$ at room temperature for 1 hour. Because we wanted to probe the ensemble of the RNA with minimal interference from the $3'$ unpaired sequence that we use as the primer binding site, we folded the RNA in the presence of the fluorescent primer attached to the oligo(dT) beads (Ambion) that we regularly use for purification. Folding in this condition sequesters any additional single stranded regions that may interfere with our sequence of interest. The RNA was then subjected to 1M7 mapping (5 mM final concentration), purified with the oligo(dT) beads, and reverse transcribed for 30 minutes at 42°C. Umodified RNA controls were also included in the experiment. RNA was then degraded using alkaline hydrolysis and cDNA was purified, eluted in Hi-Di Formamide spiked with a fragment analysis ladder (ROX 350 standard, Applied Biosystems), and electrophoresed in an ABI 3150 capillary electrophoresis sequencer.

Electrophoretic traces were aligned, baseline subtracted, and normalized with the Hi-TRACE MATLAB toolkit. 1M7 modification traces were quantified, background subtracted, and corrected for attenuation using 50X dilutions, the unmodified controls, and the pentaloops added in the ends of the constructs as reference (Thomas Mann, personal communication, see Figure **??**).

5.7 M^2-seq of the MedLoop RNA

M^2 measurements were obtained by high-throughput sequencing through the EteRNA expert interface... **Should I write the lab's protocol for this or just say that we obtained this through EteRNA?**

5.8 Simulating M^2 data

We used the following procedure to simulate M^2 data for a sequence Seq. We start with the set of structures S that lie at most 1 kcal above of the minimum free energy wild type structure obtained by the AllSub RNAstructure program and generate mock reactivity profiles for each one by sampling $RMDB_U$ and $RMDB_P$ for unpaired and paired residues respectively. For each single-nucleotide

mutant, we calculate the free energies of the structures given the mutated sequence using the efn2 RNAstructure program. We then add white noise to the energies to simulate deviations from our incomplete understanding of RNA thermodynamics. Using these noisy energies, we calculate the Boltzmann weights of each structure for each mutant. The reactivity profile of each mutant is then the linear combination of the reactivity profiles of S weighted by the set of weights calculated previously. To simulate local perturbations due to mutations, we randomly added Laplace-distributed reactivity differences in sites at most one nucleotide away from the mutation position and any base-pair affected by the nucleotide change. These simulated datasets showed hallmarks of experimental M^2 data, such as global structure rearrangements and punctate mutation marks (see Figure **??**C).

6 Data set and software availability

REEFFIT has been integrated into the RNA dataset toolkit (RDATkit) and is available at http://simtk.org/ree alongside with software documentation and tutorials. M^2 data for the Hobartner bistable RNA has been deposited in the RMDB (RMDB ID BSTHPN_SHP_0001.rdat). M^2-seq data for the MedLoop, MedLoopΔ, and M-stable RNAs are part of the EteRNA cloud lab, rounds 72 and 73, and is also available at the RMDB (RMDB IDs ETERNA_ R72_ 0000 and ETERNA_ R73_ 0000). RDAT files for the simulated datasets are available on request.

References

[1] H Akaike. Factor analysis and AIC. *Psychometrika*, 1987.

[2] Hagai Attias. Inferring parameters and structure of latent variable models by variational bayes. pages 21–30, July 1999.

[3] MA Babyak and SB Green. Confirmatory factor analysis: an introduction for psychosomatic medicine researchers. *Psychosomatic medicine*, 2010.

[4] Jameson R Bothe, Evgenia N Nikolova, Catherine D Eichhorn, Jeetender Chugh, Alexandar L Hansen, and Hashim M Al-Hashimi. Characterizing RNA dynamics at atomic resolution using solution-state NMR spectroscopy. *Nature methods*, 8(11):919–31, December 2011.

[5] J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.

[6] P Cordero, JB Lucks, and R Das. An RNA Mapping Database for curating RNA structure mapping experiments. *Bioinformatics*, 2012.

[7] Pablo Cordero, Wipapat Kladwang, Christopher C VanLang, and Rhiju Das. Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry*, 51(36):7037–9, September 2012.

[8] J Dahl and L Vandenberghe. Cvxopt: A python package for convex optimization. *Proc. eur. conf. op. res*, 2006.

[9] Rhiju Das and David Baker. Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences of the United States of America*, 104(37):14664–9, September 2007.

[10] Rhiju Das, John Karanicolas, and David Baker. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature methods*, 7(4):291–4, April 2010.

[11] Ye Ding, Chi Yu Chan, and Charles E Lawrence. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA (New York, N.Y.)*, 11(8):1157–66, August 2005.

[12] Jes Frellsen, Ida Moltke, Martin Thiim, Kanti V Mardia, Jesper Ferkinghoff-Borg, and Thomas Hamelryck. A probabilistic model of RNA conformational space. *PLoS computational biology*, 5(6):e1000406, June 2009.

[13] PP Gardner, J Daub, and J Tate. Rfam: Wikipedia, clans and the "decimal" release. *Nucleic acids ...*, 2011.

[14] Z Ghahramani and GE Hinton. The EM algorithm for mixtures of factor analyzers. 1996.

[15] C Höbartner and R Micura. Bistable secondary structures of small RNAs and their structural probing by comparative imino proton NMR spectroscopy. *Journal of molecular biology*, 2003.

[16] Magdalena A Jonikas, Randall J Radmer, Alain Laederach, Rhiju Das, Samuel Pearlman, Daniel Herschlag, and Russ B Altman. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA (New York, N.Y.)*, 15(2):189–99, February 2009.

[17] W Kladwang, CC VanLang, P Cordero, and R Das. A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nature chemistry*, 2011.

[18] Wipapat Kladwang and Rhiju Das. A Mutate-and-Map Strategy for Inferring Base Pairs in Structured Nucleic Acids: Proof of Concept on a DNA/RNA Helix. *Biochemistry*, 49(35):7414–7416, September 2010.

[19] Wipapat Kladwang, Christopher C VanLang, Pablo Cordero, and Rhiju Das. Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry*, 50(37):8049–56, September 2011.

[20] Julius B Lucks, Stefanie A Mortimer, Cole Trapnell, Shujun Luo, Sharon Aviran, Gary P Schroth, Lior Pachter, Jennifer A Doudna, and Adam P Arkin. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences of the United States of America*, 108(27):11063–11068, 2011.

[21] Maumita Mandal and Ronald R Breaker. Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nature structural & molecular biology*, 11(1):29–35, January 2004.

[22] HW Marsh, JR Balla, and RP McDonald. Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological bulletin*, 1988.

[23] J S McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–19.

[24] Somdeb Mitra, Inna V Shcherbakova, Russ B Altman, Michael Brenowitz, and Alain Laederach. High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis. *Nucleic acids research*, 36(11):e63, June 2008.

[25] Stefanie A Mortimer and Kevin M Weeks. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *Journal of the American Chemical Society*, 129(14):4144–5, April 2007.

[26] D Oakes. Direct calculation of the information matrix via the EM. *Journal of the Royal Statistical Society: Series B (...*, 1999.

[27] A Patil, D Huard, and CJ Fonnesbeck. PyMC: Bayesian stochastic modelling in Python. *Journal of statistical software*, 2010.

[28] Anke Reining, Senada Nozinovic, Kai Schlepckow, Florian Buhr, Boris Fürtig, and Harald Schwalbe. Three-state mechanism couples ligand and temperature sensing in riboswitches. *Nature*, 499(7458):355–9, July 2013.

[29] DB Rubin and DT Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 1982.

[30] Rick Russell, Xiaowei Zhuang, Hazen P Babcock, Ian S Millett, Sebastian Doniach, Steven Chu, and Daniel Herschlag. Exploring the folding landscape of a structured RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 99(1):155–60, January 2002.

[31] JB Schreiber and A Nora. Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of ...*, 2006.

[32] Pilar Tijerina, Sabine Mohr, and Rick Russell. DMS footprinting of structured RNAs and RNA-protein complexes. *Nature Protocols*, 2(10):2608–2623, 2007.

[33] Sergio M Villordo, Diego E Alvarez, and Andrea V Gamarnik. A balance between circular and linear forms of the dengue virus genome is crucial for viral replication. *RNA (New York, N.Y.)*, 16(12):2325–35, December 2010.

[34] Jinbu Wang, Xiaobing Zuo, Ping Yu, Huan Xu, Mary R Starich, David M Tiede, Bruce A Shapiro, Charles D Schwieters, and Yun-Xing Wang. A method for helical RNA global structure determination in solution using small-angle x-ray scattering and NMR measurements. *Journal of molecular biology*, 393(3):717–34, October 2009.

[35] JM Watts, KK Dang, and RJ Gorelick. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, 2009.

[36] KA Wilkinson, RJ Gorelick, SM Vasa, and N Guex. High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS biology*, 2008.

[37] Wade C Winkler and Ronald R Breaker. Genetic control by metabolite-binding riboswitches. *Chembiochem : a European journal of chemical biology*, 4(10):1024–32, October 2003.

[38] S. Yoon, J. Kim, J. Hum, H. Kim, S. Park, W. Kladwang, and R. Das. HiTRACE: high-throughput robust analysis for capillary electrophoresis. *Bioinformatics*, 27(13):1798–1805, May 2011.

Table 1: Results for the *in silico* simulated M^2 benchmark.

Name	Sequence length	Number of structures	Number of predicted structures	Number of correct structures	χ^2/df	RMSEA	Pe of co pr
RF01092;GP knot2	61	3	2	2	1.35	0.08	20
RF01020;mir-572	56	4	2	2	1.32	0.08	84
RF01313;HHBV epsilon	57	3	2	2	1.34	0.08	84
RF00066;U7	61	5	3	0	3.45	0.20	73
RF01300;snoU49	58	4	3	2	2.10	0.14	55
RF01139;sR2	55	3	2	2	0.90	0.00	19
RF01274;sR45	55	4	2	2	1.88	0.13	80
RF01297;sR40	61	2	2	2	2.46	0.16	9.6
RF00555;L13 leader	56	3	2	2	1.16	0.05	24
RF00775;mir-432	53	4	3	3	1.37	0.08	39
RF00173;Hairpin	46	5	3	2	2.41	0.18	19
RF00196;AMV RNA1 SL	40	3	3	3	3.10	0.23	13
RF00424;SCARNA16	53	3	2	2	2.18	0.15	70
RF00108;SNORD116	36	3	4	1	2.41	0.20	45
RF00570;SNORD64	58	3	2	2	1.09	0.04	66
RF01301;snoR4a	53	4	2	2	1.80	0.12	51
RF01125;sR4	58	7	4	4	3.68	0.22	64
RF00436;UnaL2	32	4	2	2	2.05	0.18	18
RF01151;snoU82P	49	4	2	2	1.95	0.14	69
RF01414;class I RNA	58	2	2	2	1.69	0.11	67

Table 2: Primer sequences for the Hobartner bistable sequence PCR assembly

Primer 1	TTCTAATACGACTCACTATAGGTGGC
Primer 2	TCCGGTACATAAGGCTTCTACTCGAAGCCACCTATAGTGAGTCGT
Primer 3	AGAAGCCTTATGTACCGGAAGGTGCGAATCTTCCGAAGGATCCGAGTAGGATCCAAA
Primer 4	GTTGTTGTTGTTGTTTCTTTTTGGATCCTACTCGGATCCTTCGGAAGATTCGCA

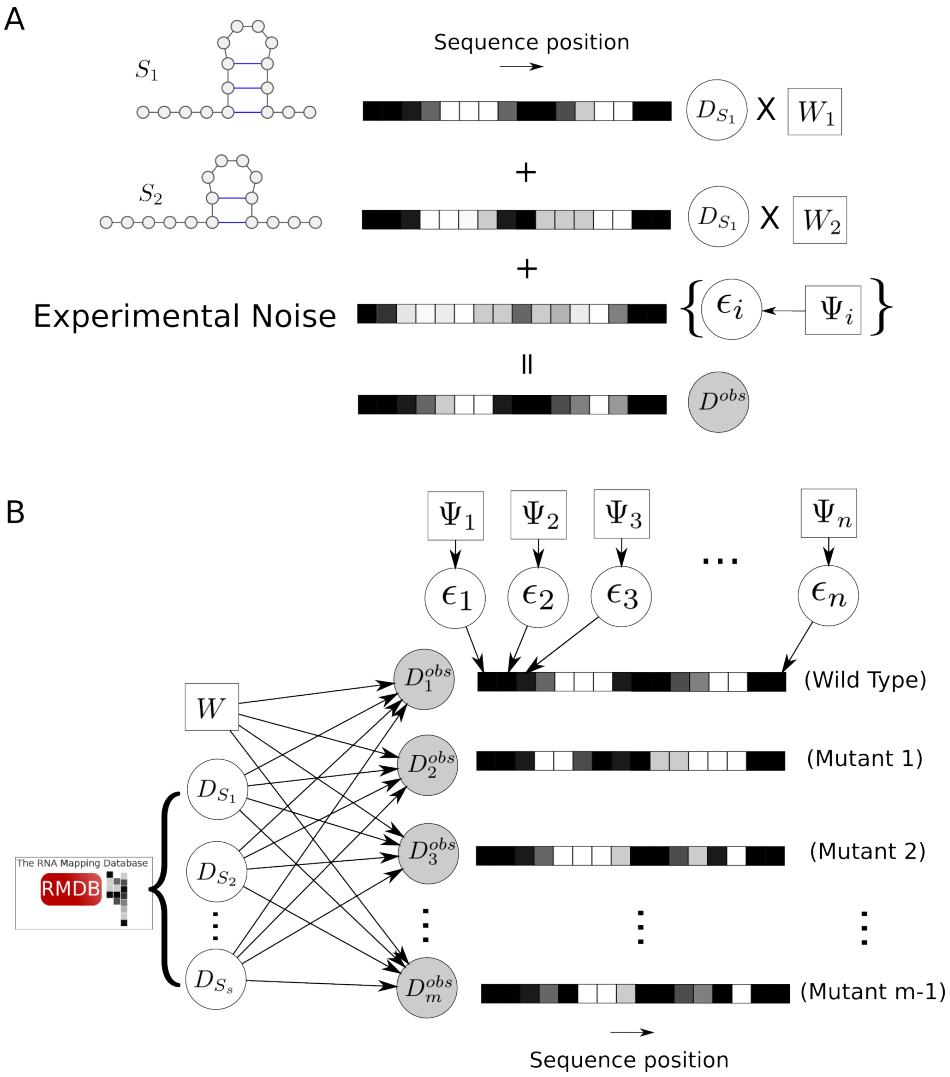

Figure 1: A. Chemical mapping or footprinting experiments for probing RNA structure can be conceptualized as linear combinations of the underlying structures in the RNA's structural ensemble. (A) The chemical mapping profiles of an ensemble of two structures, S_1 and S_2, represented as as one-dimensional heat maps, are scaled by their respective Boltzmann weights, W_1 and W_2, and added together with experimental noise to form the observed chemical mapping profile of an RNA. (B) Graphical model representation of the REEFFIT statistical model. Multi-dimensional mapping experiments experiments provide chemical mapping measurements of the structural ensemble altered by systematic perturbations (mutations, in the case of M^2 experiments) and can therefore be described in a factor analysis framework. White cirlces represent hidden random variables (in this model, each structure's individual chemical mapping profile and the measurement-wise experimental noise); white squares show model parameters to be estimated (the weights per structure W and the experimetnal noise covariance matrix Ψ; shaded circles are the observed values, (e.g. the chemical mapping profiles of each mutant in M^2 data). Structure-dependant prior distributions for the hidden chemical mapping profiles are modeled from observed reactivity values of paired and un-paired nucleotides in the RMDB.

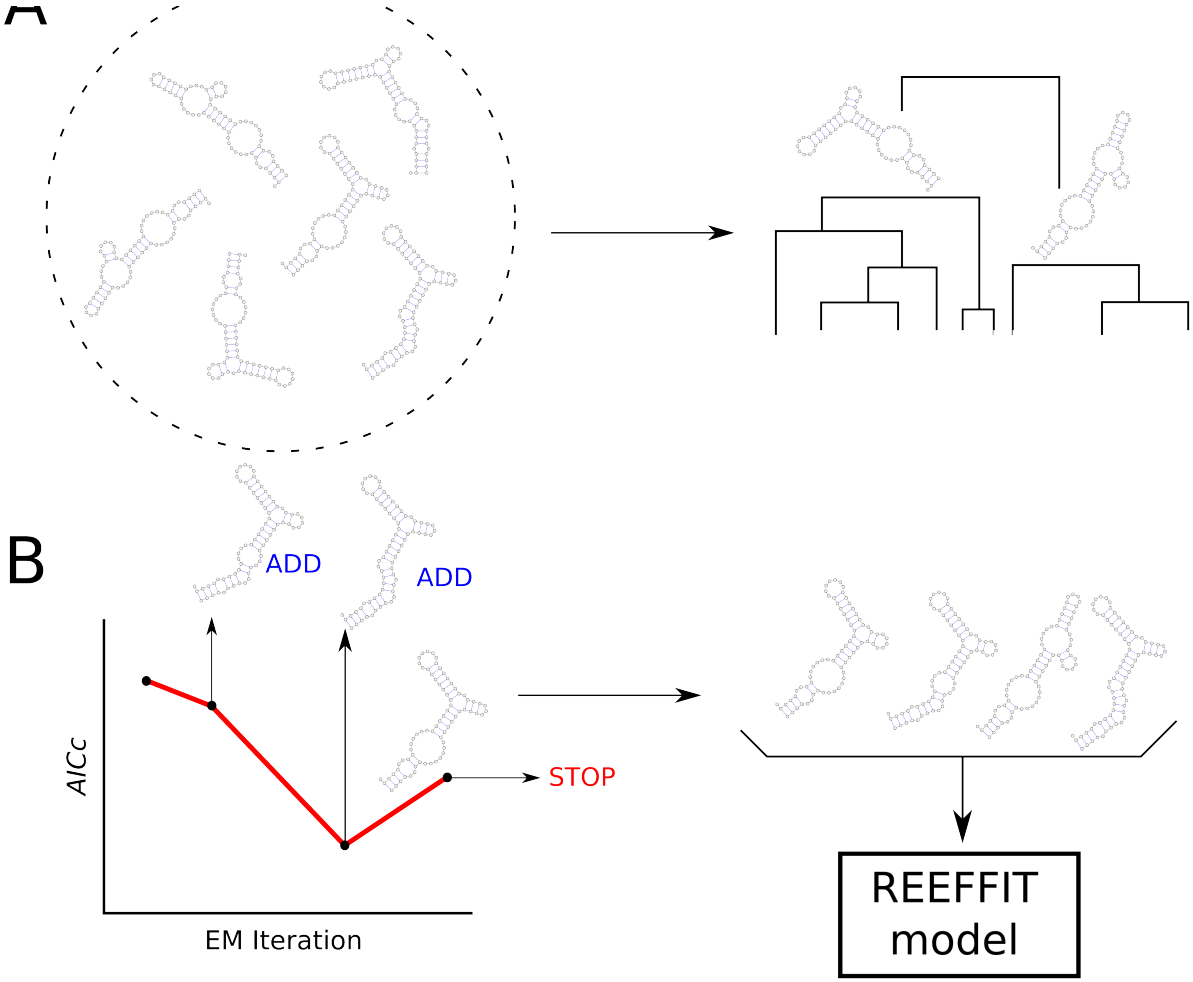

Figure 2: Selection algorithm for obtaining the structure set used to model the data when no structural ensemble is given *a priori*. (A) An initial set of structures is calculated by clustering the suboptimal structures for all sequences in the experiment. Initial structure medoids are those that maximize a *score* function that favors structures that are seen to be most stabilized in some measurement/mutant. (B) Additional structures from a sequence of *score*-sorted structures are added until the *AICc* score, calculated within EM iterations, does not decrease further. The selected structures are finally re-fitted to the data using re-initialized variables in the REEFFIT statistical model model.

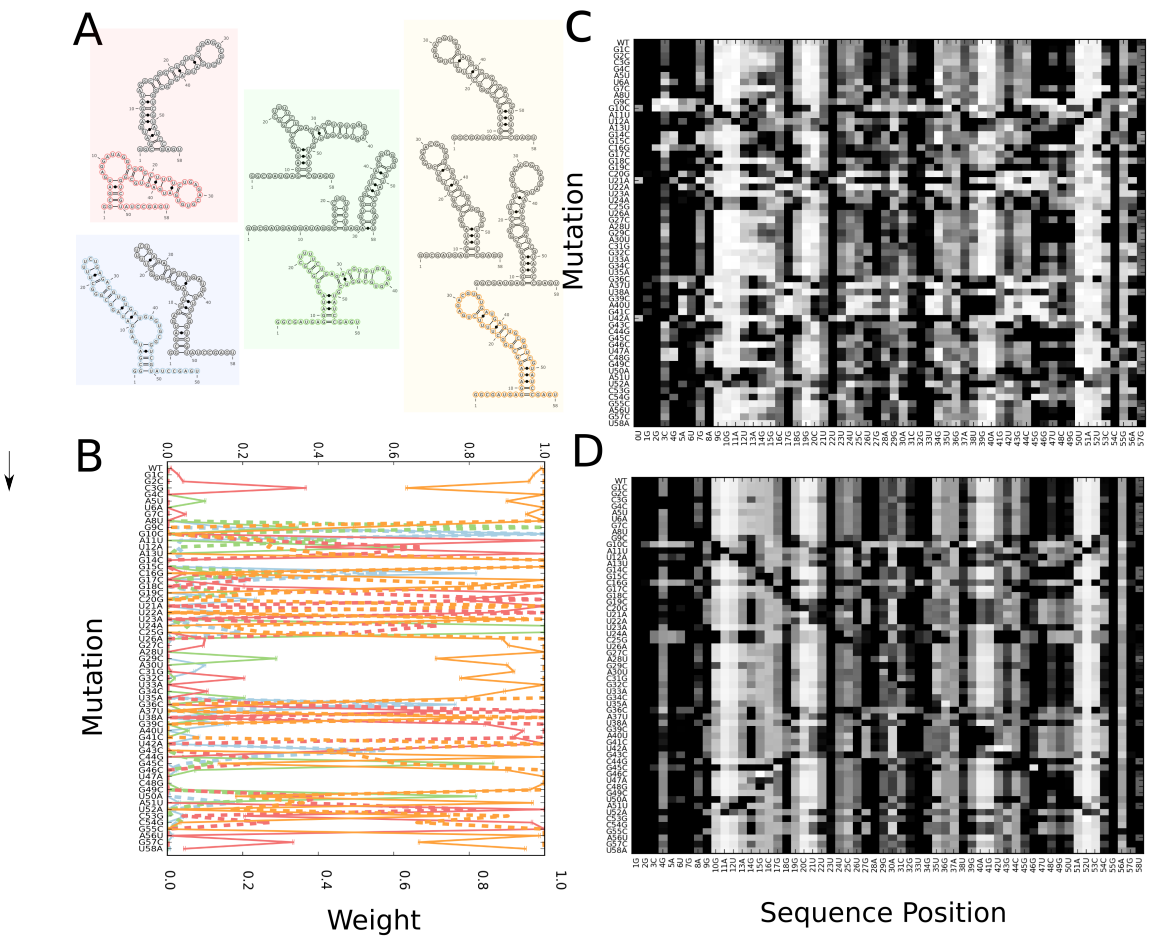

Figure 3: Evaluation of the REEFIT method with simulated data. Here, we show the simulated data and REEFFIT analysis for an representative sequence of the RF01125 family. (A) Structures used for generating the data and those selected by REEFFIT to model the data. REEFFIT selected 4 structures, effectively collapsing several structures into different groups. (B) REEFFIT's predicted weights of each structure (solid lines with error bars, color-coded by structure color as in (A)) compared to the sums of true weights of structures per clusters defined by the two structure medoids (dotted lines, color-coded by structure medoid in each cluster as in (A)). Even though only two structures were used to model the data, the weights capture the underlying thermodynamic partition of the ensemble, assigning them correct weights. (C) Simulated M^2 data of the RF01125 representative, with noisy energies. (D) Predicted data using REEFFIT's estimation for the structure weights, the expected values for the hidden chemical mapping profiles of each structure, and local mutation-induced perturbations.

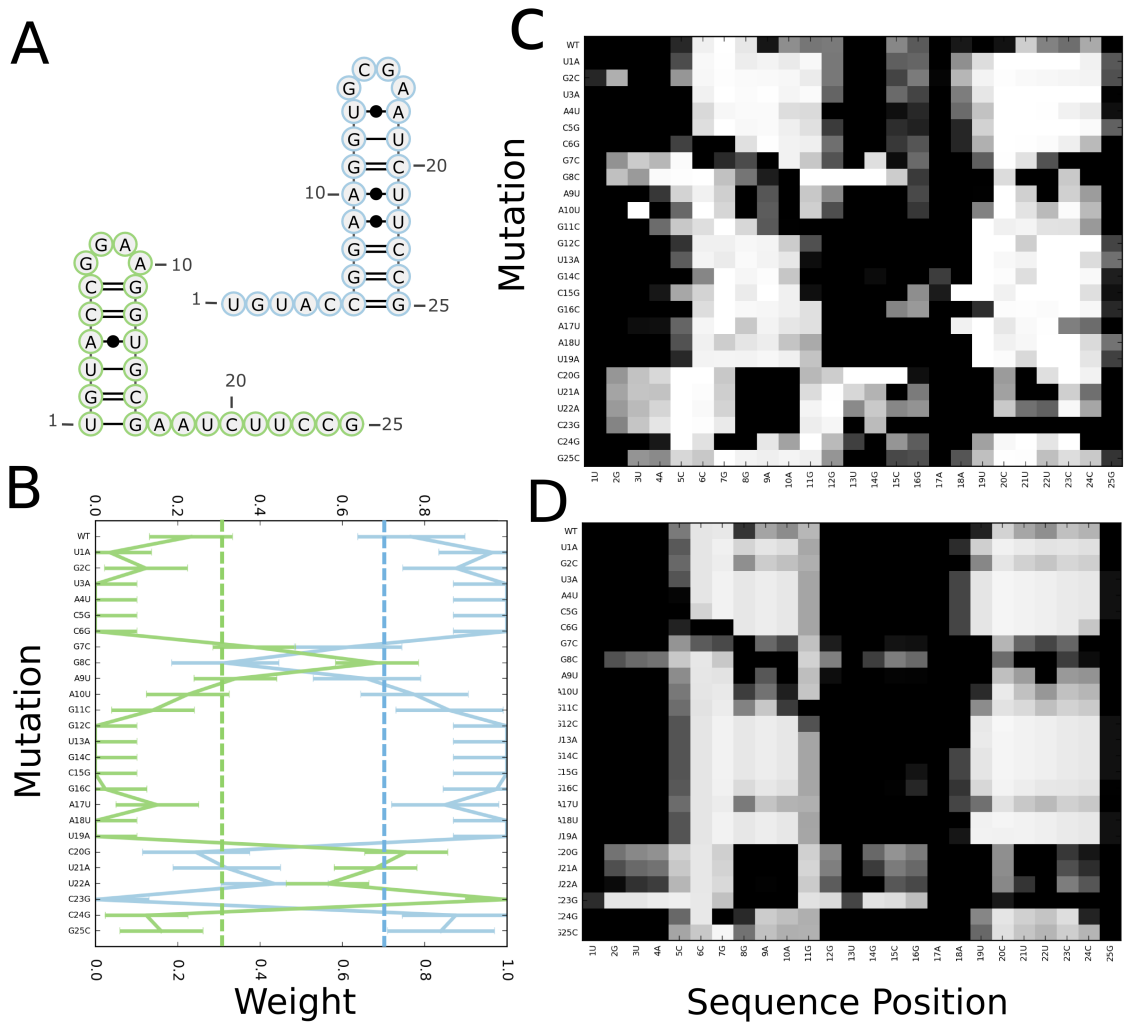

Figure 4: M^2 measurements and predicted weights of the Hobartner bistable RNA structural ensemble. (A) The Hobartner bistable sequence folds into two different hairpins. (B) Weights calculated by REEFFIT for the wild type are within error from the NMR estimates (dotted horizontal lines).(C) M^2 measurements displayed in the same format as the simulated datasets in Figure **??**. (D) Data predicted by the REEFFIT statistical model. Annotations are analogous to Figure **??**.

.

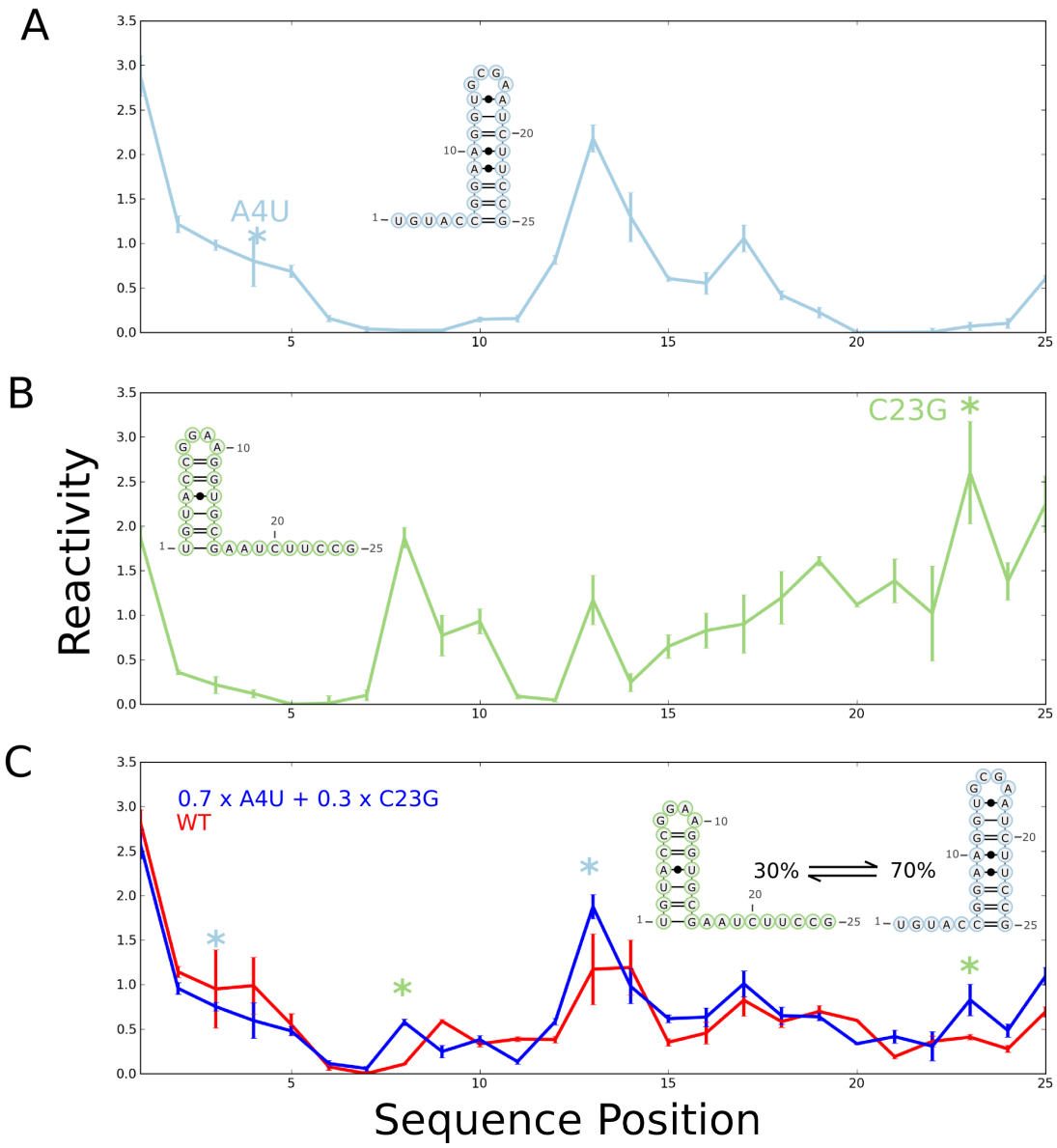

Figure 5: Reactivities of the wild type Hobartner bistable RNA as a function of two mutants that are predicted by REEFFIT to fully stabilize each structure in the structural ensemble. (A) A4U is predicted to fold predominantly in structure hob_1 (blue); (B) C23G is most of the time in state hob_2 (green). (C) Combining their profiles with weights 0.7 for A4U and 0.3 for G23C (blue) results in reactivities similar to the wild type profile (red). The differences between the weighted combination of mutants and the wild type reactivities can be explained by their respective mutations and their effects on their helix partners in each state (blue and green stars, for mutants A4U and C23G, respectively).

.

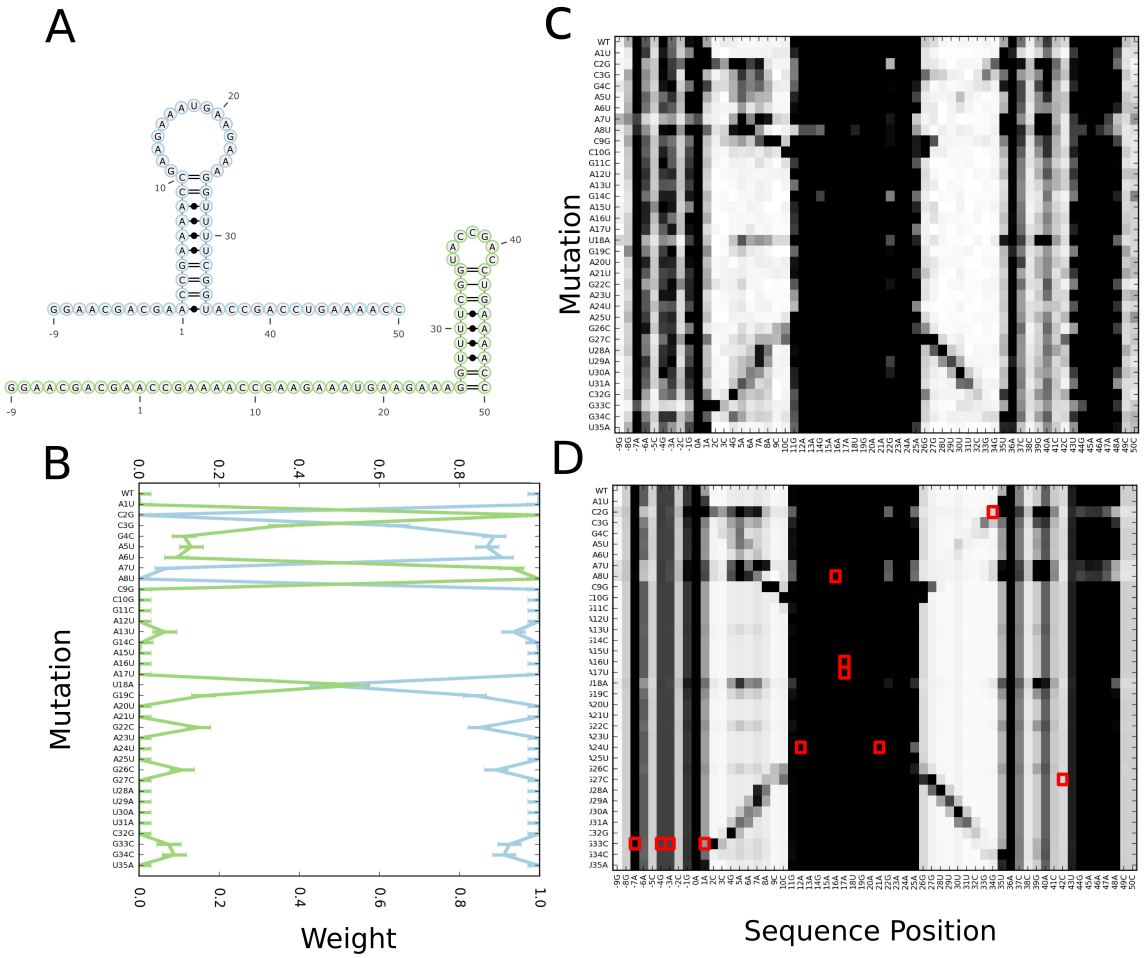

Figure 6: M^2-seq measurements and predicted weights of the MedLoop motif. (A) Structures selected by REEFFIT to model the data. One structure (blue) is predominant in most mutants; mutations that severly disrupt it (e.g. G4C) form an alternative structure (green) (B) M^2 measurements displayed in the same format as the simulated datasets in Figure **??**. (C) Data predicted by the REEFFIT statistical model. Annotations are analogous to Figure **??**. (D) Comparison of observed reactivities for the wild type MedLoop sequence to the predicted profile of REEFFIT.

.

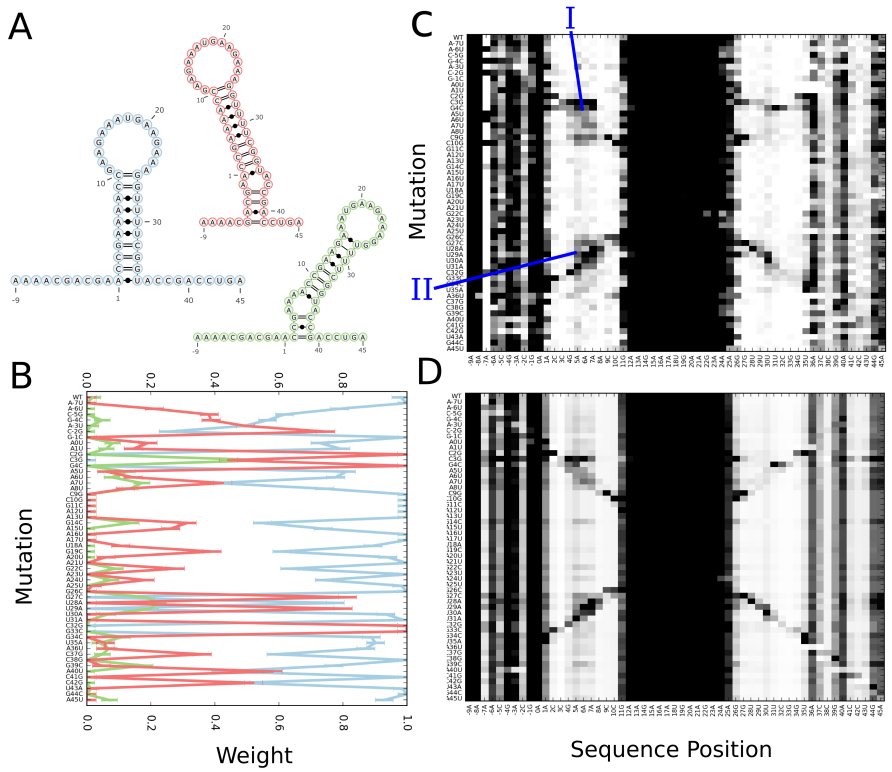

Figure 7: M^2-seq measurements and predicted weights of the MedLoopΔ construct. (A) Structures selected by REEFFIT to model the data. MedLoopΔ shares one structure with the MedLoop RNA (blue), but lacks the nucleotides for the alternative structure. Because of this deletion, another two alternative structures are stabilized by some of the mutants (green and red) (B) M^2 measurements displayed in the same format as the simulated datasets in Figure **??** and **??**. (C) Data predicted by the REEFFIT statistical model. Annotations are analogous to Figure **??**.

.

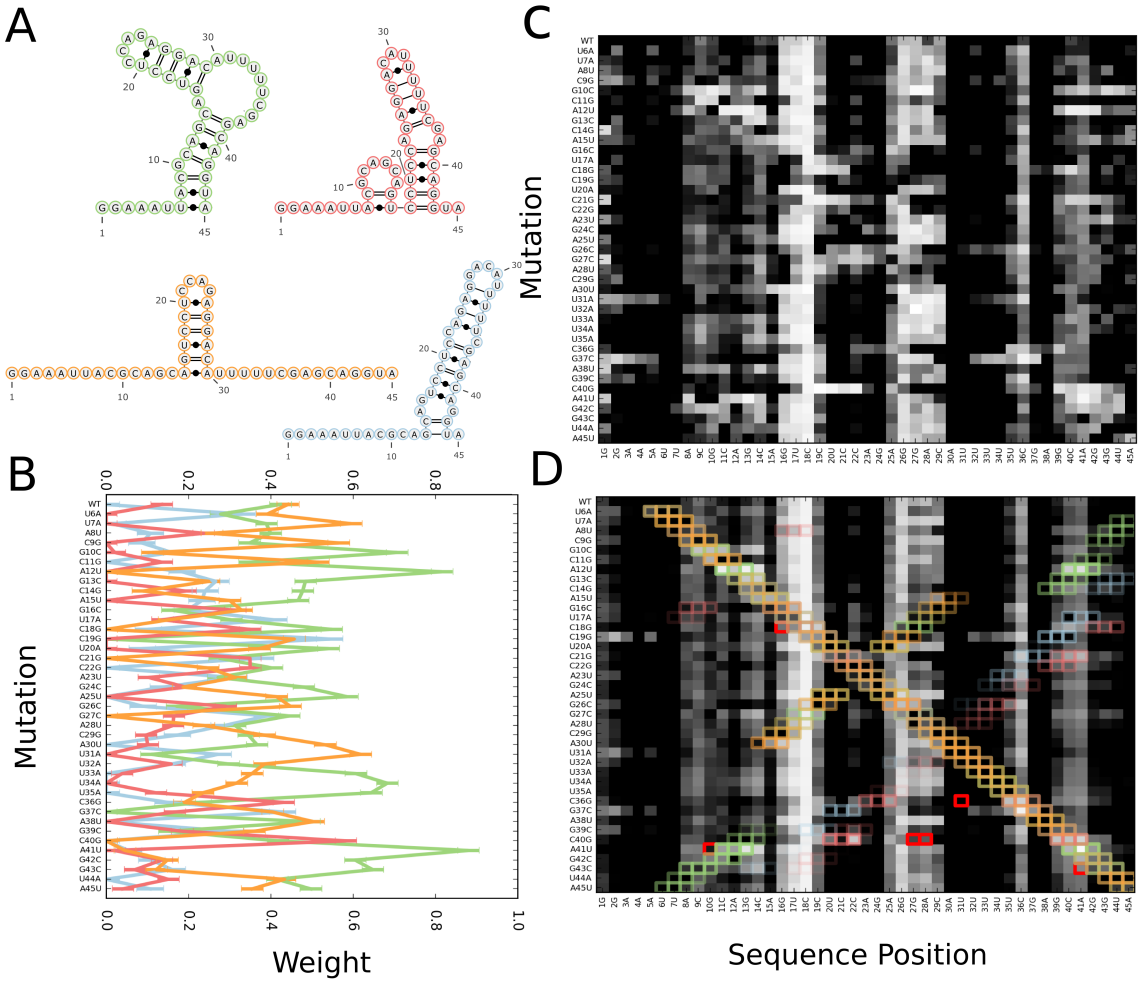

Figure 8: M^2-seq measurements and predicted weights of the M-stable RNA. (A) Structures selected by REEFFIT to model the data. MedLoopΔ shares one structure with the MedLoop RNA (blue), but lacks the nucleotides for the alternative structure. Because of this deletion, another two alternative structures are stabilized by some of the mutants (green and red) (B) M^2 measurements displayed in the same format as the simulated datasets in Figure **??**. (C) Data predicted by the REEFFIT statistical model. Annotations are analogous to Figure **??**.

.

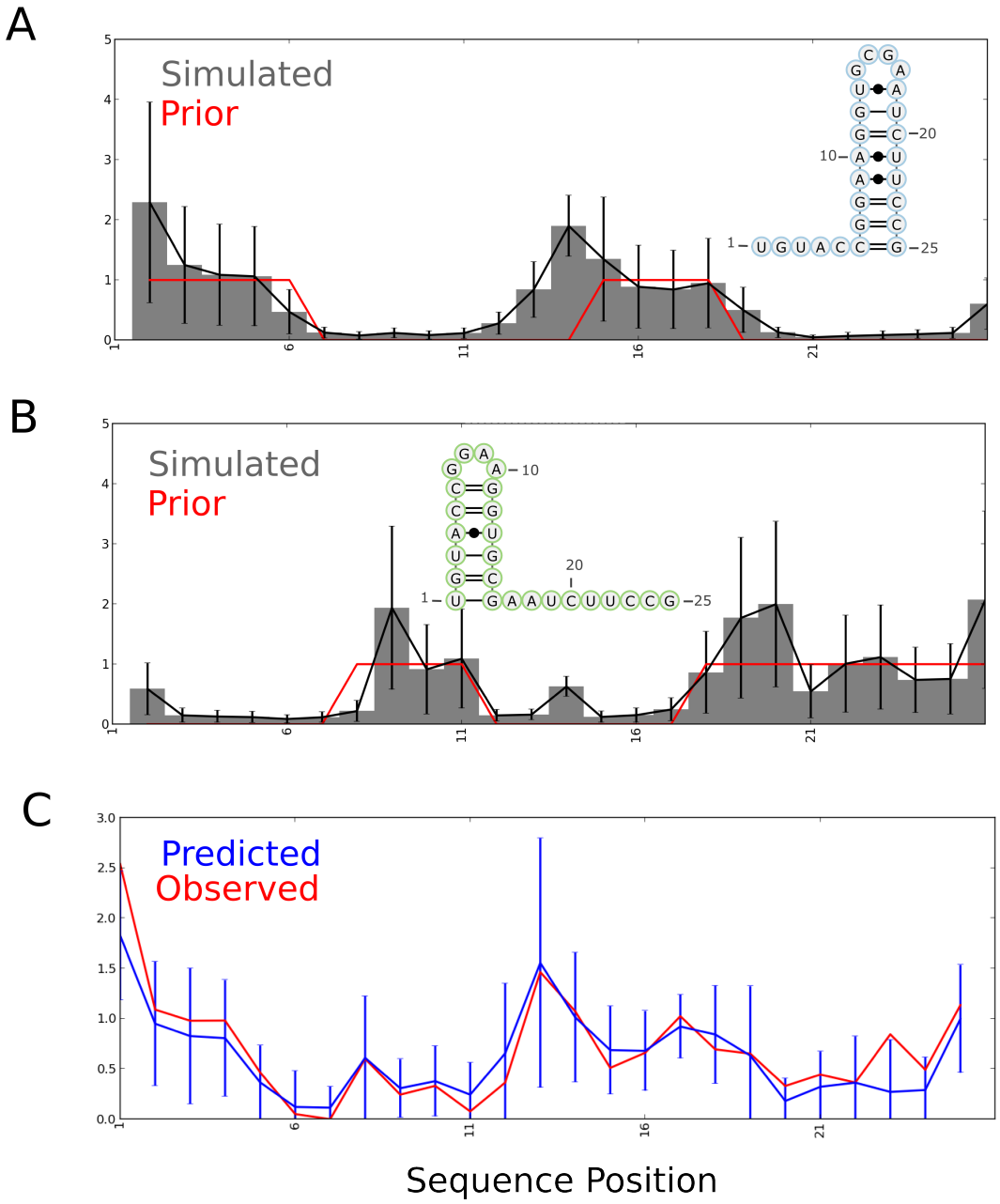

Figure 9: Expected and predicted reactivities for the two structures in the Hobartner bistable RNA ensemble. (A,B) Expected reactivities of the calculated by MCMC simulations. The solid red lines indicates our prior for each sequence position in each sructure (zero or one, for paired and unpaired residues in that structure, respectively)(C) Comparison of observed reactivities for the wild type Hobartner sequence to the predicted profile of REEFFIT. Error bars are the estimated value for Ψ_i at each position.

.

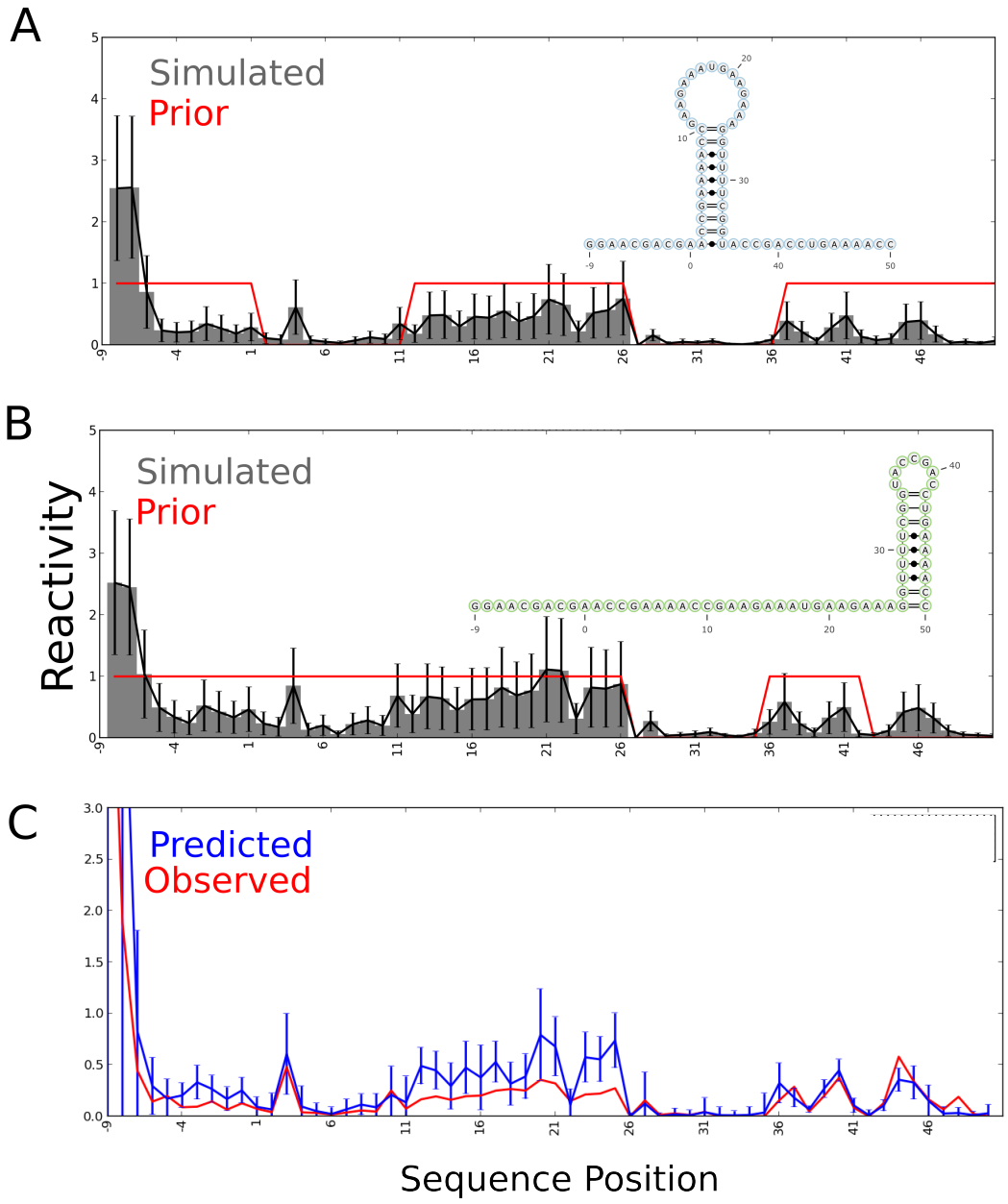

Figure 10: Expected and predicted reactivities for the two structures in the MedLoop RNA ensemble. (A,B) Expected reactivities of the two structures calculated by MCMC simulations. The solid red lines indicates our prior for each sequence position in each sructure (zero or one, for paired and unpaired residues in that structure, respectively)(C) Comparison of observed reactivities for the wild type MedLoop sequence to the predicted profile of REEFFIT. Error bars are the estimated value for Ψ_i at each position.

.

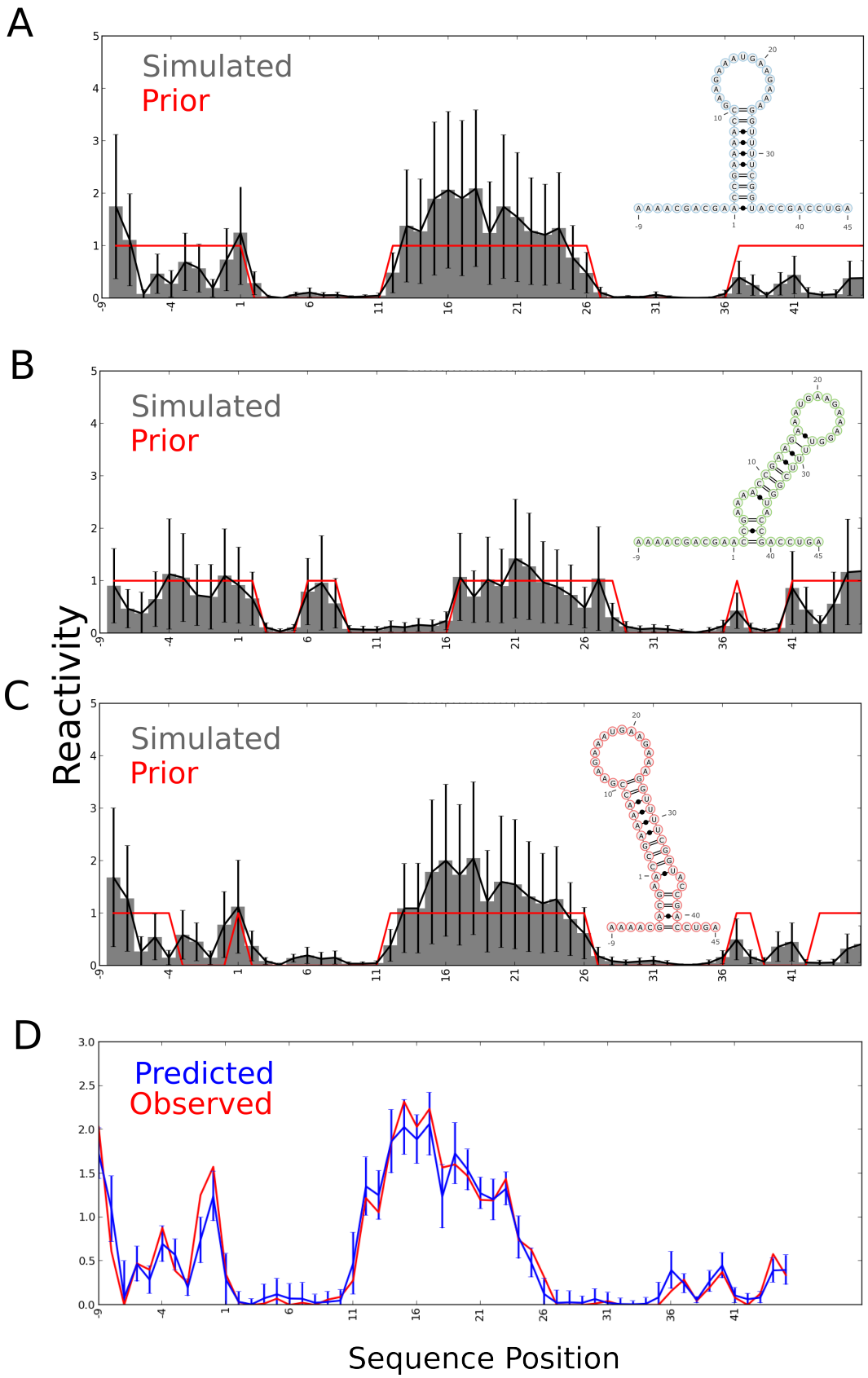

Figure 11: Expected and predicted reactivities for the three structures in the MedLoopΔ RNA ensemble. (A-C) Expected reactivities of the three structures calculated by MCMC simulations. The solid red lines indicates our prior for each sequence position in each sructure (zero or one, for paired and unpaired residues in that structure, respectively)(D) Comparison of observed reactivities for the wild type MedLoopΔ sequence to the predicted profile of REEFFIT. Error bars are the estimated value for Ψ_i at each position.

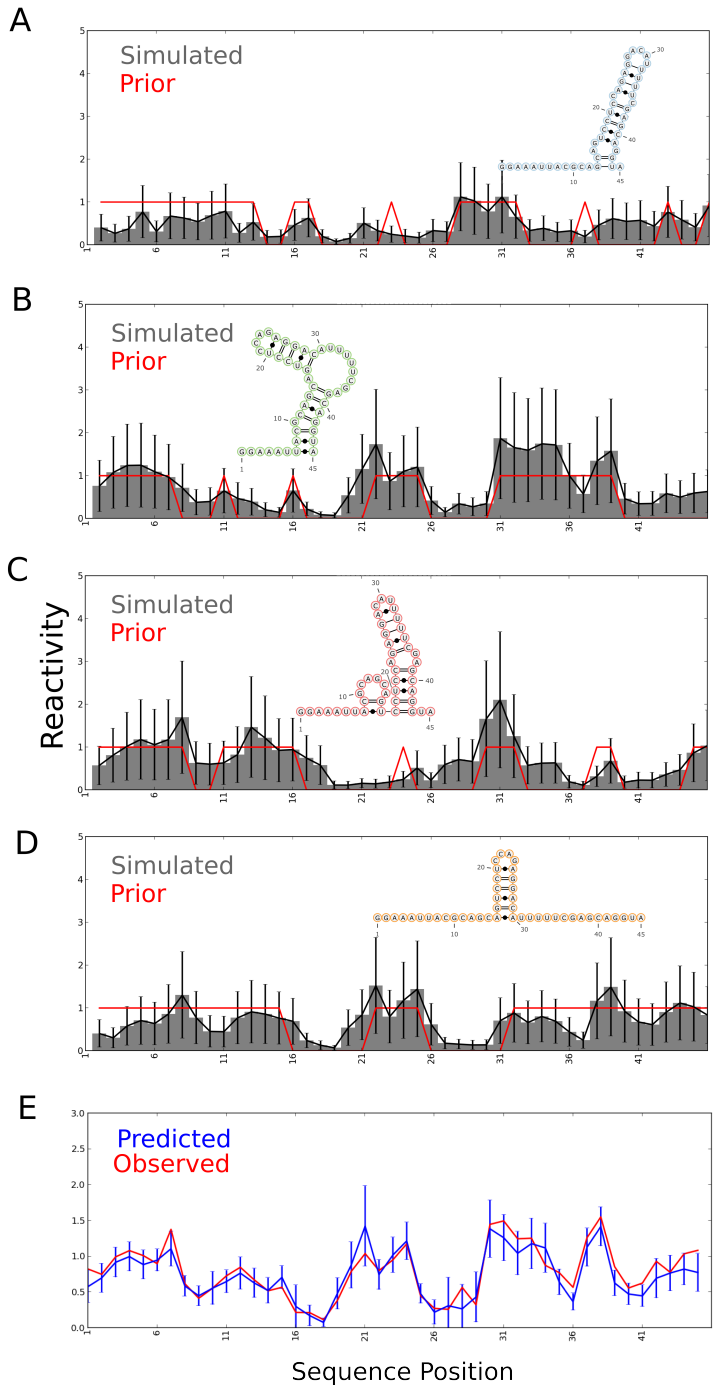

Figure 12: Expected and predicted reactivities for the three structures in the M-stable RNA ensemble. (A-D) Expected reactivities of the three structures calculated by MCMC simulations. The solid red lines indicates our prior for each sequence position in each sructure (zero or one, for paired and unpaired residues in that structure, respectively)(E) Comparison of observed reactivities for the wild type M-stable sequence to the predicted profile of REEFFIT. Error bars are the estimated value for Ψ_i at each position.

.

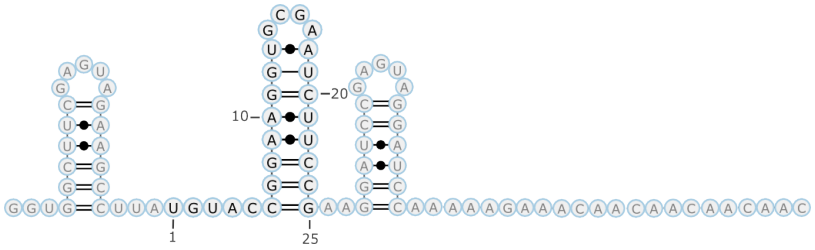

Figure 13: The full sequence of the Hobartner bistable RNA (the hob_1 structure is shown here). Auxiliary nucleotides are marked in gray.

.