Novel lead compounds that target the ribosomal peptidyl transferase center

Benjamin Tam[1], Shira Cohen[1], Dror Sherf[1], Sarah Adi Eisdorfer[1], Moshe Perez[1], Adam Soffer[2], Dan Vilenchik[2], Gerhard Wagner[3], and Barak Akabayov[1]

[1]Department of Chemistry and [2]Department of Communication Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel, [3]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA

Abstract

We have developed new lead compounds that target the ribosomal peptidyl transferase center (PTC) of *Mycobacterium tuberculosis*. For this purpose, we used a fragment-based virtual screening (FBVS) methodology, in which the first step was the novel exploitation of NMR T_2 relaxation to identify fragment molecules that bind specifically to RNA hairpin 91 in the ribosomal PTC of *M. tuberculosis*. This initial screening was followed by computational optimization of the fragment molecules into larger molecules with drug-like properties. Specifically, the virtual filtration followed by a high-throughput docking procedure that yielded drug-sized molecules for the ribosomal PTC. Using the data obtained from these experiments and simulations, we trained various machine-learning models for predicting the docking binding energy (ΔG) as a function of 33 geometric features extracted from each of the above molecules. From these models, we drew non-trivial conclusions regarding advantageous chemical modifications on the initial scaffold and guidelines to which geometric properties the structures that attach to these molecular locations are favorable. As superior inhibitors, the machine-learning model predicted two molecules with a phenylthiazole moiety, namely,1-[(2-methylthiazol-4-yl)methyl]-4-[(2-phenylthiazol-4-yl)methyl]piperazine and 1-ethyl-4-[(2-phenylthiazol-4-yl)methyl]piperazine. Remarkably, these two molecules exhibited IC_{50} values superior to that of chloramphenicol, an antibiotic drug that acts on the ribosomal PTC.

Introduction

A recent U.K. government review (https://amr-review.org/) on the global threat of anti-microbial resistance predicts that by 2050 ten million people will be at risk each year of dying from a drug-resistant infection, with about 25% of these deaths likely to be due to drug-resistant strains of *Mycobacterium tuberculosis*, the causative agent of tuberculosis (TB). There is thus an ongoing effort both in academia and on the part of drug companies to develop new, and better, antimicrobials against a range of drug-resistant pathogens [1]. Two crucial stages in this process – both of which are addressed in this study – are the selection of the drug target for the specific pathogen and the discovery of lead compounds (which are subsequently chemically refined and tested in vitro and in vivo for their antibacterial activity and toxicity). As the bactericidal target site in the development of a new type of lead compound for TB, we chose the ribosomal peptidyl transferase center (PTC), with the thinking underlying this choice being set out below.

Gene translation, an essential event in the cell cycle of every bacterium, is performed by ribosomes, multi-macromolecular machines composed of two subunits made up of RNA and proteins. A single bacterial cell may contain as many as 100,000 ribosomes during the log phase of growth. Although the structure of the *M. tuberculosis* ribosome remains to be elucidated, it is highly likely that it shares common features with other bacterial ribosomes, as has been shown for the three bacterial ribosomes whose structures are known—those of *Escherichia coli* [2], *Staphylococcus aureus* [3] and *M. smegmatis* [4]. The elucidation of the structures of these ribosomes has led to the understanding that bacterial ribosomes fulfill most of the requirements for potential drug targets in that: 1) they differ significantly from the ribosome of human cells and will hence constitute highly selective targets, and 2) their basic components are all essential for the viability of the bacterial cell. We are thus confident that the ribosomal PTC, being the heart of the gene translation machinery, constitutes a highly selective drug target module within the ribosome. Indeed, most of the available antibacterial compounds that target the large (50S) ribosome subunit of bacteria act on the PTC [5], a universally conserved ribonucleoside chain (rRNA) that catalyzes the formation of peptide bonds during peptide chain synthesis. The PTC forms a 'pocket' structure in the center of the ribosome, and it is this PTC pocket, being a hotspot for inhibition, that has been targeted in the development of antibacterial agents, such as pleuromutilins, lincosamides, phenicols, and

macrolides [6]. Currently the design and/or discovery of small molecules targeting the PTC pocket is thus one of the major concepts in the development of antibacterial agents.

Small molecule inhibitors that target the ribosome can be obtained by using high-throughput screening (HTS), fragment-based screening (FBS), or virtual screening of evolved libraries. Among these three options, we have chosen an FBS-virtual screening tandem by virtue of the two major advantages of FBS over HTS: FBS requires only a single technological method, usually NMR, to detect binding [7], and fragment-based libraries cover a larger chemical space than HTS-generated libraries [8]. The current methodology builds on that developed for our previous study using NMR-saturated transfer difference (STD) spectroscopy for screening prototype inhibitors of T7 DNA primase as a model compound [8,9]. The current study addresses the problem inherent in NMR-STD spectroscopy that the low distribution of protons in nucleic acids (compared to proteins) makes it difficult to excite only the macromolecule and hence to obtain effective transfer of magnetization to the bound fragment molecule. In this study, we therefore used T_2 relaxation NMR spectroscopy for initial screening of fragment molecules. This step was followed by virtual screening comprising virtual filtration to obtain fragment-containing drug-sized molecules and high-throughput docking to select the best binders to the RNA target. The docking was conducted in a feedback loop with machine learning algorithms for optimization of the inhibitors, thereby pinpointing drug-sized molecules with good binding to the PTC and hence inhibition of ribosomal peptidyl transfer.

Results

Our rational design comprises a combined screening approach, i.e., selection of fragment molecules by using T_2-relaxation NMR spectroscopy and growth of the fragments into potent drug-sized inhibitors by using computational chemistry (**Fig. 1**). The latter part is divided into three steps (**Fig. 1**, steps 3-5): 1) virtual filtration to select from the ZINC database [10] drug-sized molecules that contain the fragment molecule found by NMR; 2) docking of the molecules found in step 1 and ranking the docked molecules on the basis of their binding energy (docking score); and 3) using these results to train a machine-learning model to predict docking. Importantly, the final step comprises an *in-vitro* translation assay to select, from the bank of molecules pinpointed by the screening and docking steps, small molecules that inhibit mycobacterial ribosome activity, presumably through specific binding to the PTC.

NMR-fragment based and virtual screening (FBVS) – A combined screening approach

In this – first – step of screening for fragment molecules that bind ribosomal PTC, we used the 29 mer RNA sequence that corresponds to hairpin 91 in the PTC of *M. tuberculosis*. A model structure of hairpin 91 from the ribosomal PTC of *S. aureus* is shown in **Fig. 2**a-b (colored yellow). The sequence identity of hairpin 91 between *M. tuberculosis* and *S. aureus* is high (90%) and their corresponding RNA folds are identical, as reflected from their thermal stabilities obtained using circular dichroism spectroscopy (**Fig. S1**).

The target *M. tuberculosis* RNA was screened against fragments from the Maybridge Ro3 2500 Diversity Fragment Library (Maybridge, Inc.) by using T_2 relaxation NMR. In this type of NMR, a fragment molecule that binds to the larger RNA will adopt the relaxation properties of the complex [11,12], which, in turn, will tumble at slower pace than the free fragment molecule, leading to a faster relaxation time for the complex and thereby facilitating the identification of those fragment molecules that bind to the RNA [8]. Specifically, the nature of the binding is assessed by determining the magnetization decay rates, R_2, of T_2 relaxation-edited Carr–Purcell–Meiboom–Gill (CPMG) spectra of the fragment molecules in free form and bound to the RNA target, according to the procedure outlined by Hajduk et al. [13]. In addition, the differential line broadening (DLB) for the line width at half height of the NMR peak – a measure that is indirectly related to the relaxation times – also contributes to the identification of bound fragments.

In the first experiment, which was aimed to determine the optimal conditions for the T_2 relaxation screening, 7 out of the 30 fragments tested showed clear evidence of binding to the target RNA. **Fig. 3** illustrates the process of binding identification and pool deconvolution for the compound that presented the best binding, N-methyl-4-(1,3-thiazol-2-yl)benzylamine, and **Table S1** lists the values for the transverse relaxation rate (R_2) for this fragment. A full fragment screening experiment followed in which all the fragment molecules in the Maybridge Ro3 2500 collection were pooled into groups of 10 molecules each, and 250 samples were prepared both with and 250 without the RNA target. In all samples prepared for the screening (including the initial mini-screen), the RNA concentration was 15 μM, and the concentration of the fragment molecules was 300 μM. To reduce the effect of the amide proton in amides and carbines on the R_2 when measuring binding of a small molecule to RNA, deuterium labeling of the RNA was used. Similarly, solutions were prepared in deuterated phosphate buffered saline (PBS), pH 7.4. All measurements were carried out at 25 °C. The molecules identified in the full screening were ranked according to the binding

strength to the RNA target on the basis the NMR observables, i.e., DLB and CPMG. Some molecules gave changes in both DLB and CPMG, and others showed changes in only one observable (**Table S2**). Importantly, the hit fragment molecule that presented the best NMR observable values in the full screening (DLB and CPMG) was [2-(3-chlorophenyl)-1,3-thiazol-4-yl]methanamine (**Fig. 3**, molecular structure is marked in grey asterisk). The two best hits, both in the mini-screening and the full-screening, contain a phenylthiazole moiety. This moiety was therefore used as the scaffold for the subsequent optimization steps in the FBVS workflow (**Fig. 1**, steps 3-5).

Lead optimization and candidate selection

Since binding of small fragments is often weak, obtaining a lead compound from fragments traditionally requires fragment growing or linking [7,14,15]. However, the challenging task of linking the binding event to inhibition still remains the rate-limiting step in the field of FBS. To circumvent the tedious steps of growing and linking fragments, we applied computational optimization steps that include virtual filtration and high-throughput docking. In the virtual filtration step, we searched the rapidly growing ZINC database ([10]; http://zinc.docking.org/; which currently contains 230 million drug-sized compounds) for compounds with at least 70% similarity to the fragments identified by NMR. This step of virtual filtration yielded 919 phenylthiazole-containing compounds (**Fig. S2**, **Table S3**).

The next step in the FBVS workflow (**Fig. 1**) was to dock the compounds against the available crystal structure of the target molecule. Since an atomic resolution structure of the large ribosomal subunit (50S) of *M. tuberculosis* is yet to be elucidated and that of the closely related *M. smegmatis* was not available at the time of the docking, we exploited the available structure for the large ribosomal subunit of *S. aureus* ([3], **Fig. 2a**) to extract the PTC for docking; the validity of such an approach rests on the high sequence conservation between the PTCs of *M. tuberculosis* and *S. aureus* (>85%, **Fig. 2**), allowing us to use the ribosomal PTC of *S. aureus* as a proxy for the target in *M. tuberculosis*. This step yielded drug-like molecules containing the phenylthiazole moiety that were subsequently sorted on the basis of their binding energy to the ribosomal PTC. The ten drug-sized molecules with the highest binding efficiency (large -ΔG, **Table S3**) were tested for their ability to inhibit the peptidyl transfer capability of the *M. smegmatis* ribosome.

A machine-learning guided approach for augmented design principles of structure-activity relationships

The docking study described above yielded data that was subsequently used for establishing design principles for drug-sized molecules with increased binding efficiency to ribosomal PTC. The docking data yielded 811 molecules with a common phenylthiazole scaffold as described in *Methods*. A feature engineering step yielded three features for each of the 11 atom positions in the scaffold (33 features overall). Finally, two regression models (linear regression and random-forest regression) were trained using the information from the 811 molecules (**Fig. 4a**). The model features include: 1) MODi, for which each position i in the scaffold was designated true or false, depending on whether a chemical modification of any kind had occurred at that position, 2) DISTi, which represents pairwise distances within the molecule (where the pairwise distance is a proxy for the molecule's volume), and 3) VARi, an integer that represents the variance of DISTi. The variance is a proxy for the 'regularity' of the structure, i.e., the smaller the variance, the more regular the structure (as illustrated in **Fig. S3**). In addition to the 33 independent variables, the dependent variable for each molecule was BOND, which encompasses the calculated ΔG values given by the docking procedure described in the previous section. A data matrix of 811 by 34 (denoted by X) was computed, where each row corresponds to a certain molecule and each column to a certain feature. **Fig. 4b** provides a two-dimensional visualization of 400 molecules belonging to the first and fourth quartiles in terms of binding energy (ΔG). The 400 data points were projected on the leading two principal components (PCs) of the covariance matrix of X.

We trained two statistical regression models for BOND as a function of MODi, DISTi, and VARi – linear regression and random forest. From each model the most influential features for predicting BOND were extracted. A graphical priority map (**Fig. 4a**) that overlays the influence of each feature with its physical location on the scaffold shows that the most influential locations, i.e., those with the highest chances of accepting chemical modification were 6 and 8. Weaker tendencies for chemical modification were found for locations 5 and 9 and a small but still significantly important tendency for 11, with the other locations not exhibiting a significant tendency for modification. Importantly, in each of the influential locations, the most significant information was given by DISTi and VARi and not by MODi. The larger the structure at locations 5, 8 and 9 on the phenylthiazole scaffold, the stronger the binding to the RNA, whereas, at location 8 the more regular the structure the stronger the binding to the RNA. With regard to pairwise interactions between locations, a

significant link between the size of modification at location 8 and the regularity at location 6 was discovered: the larger the product $DIST_8$*VAR_6, the weaker the binding of the molecule to the RNA target. Another significant interaction was VAR_8*$DIST_8$. In this case, the larger the product, the stronger the binding. Therefore, improved binding can be obtained by modifying location 8 with a very large and a very irregular structure (quantification is presented in the *Methods* section). The pairwise interactions provide highly non-trivial constraints on the structure of the molecule, which would probably not have been revealed by "eyeballing" the data.

The random forest model was used to predict the binding energy of the two molecules with the best docking probabilities to the ribosomal PTC. Strikingly, among the 10 molecules that were tested in the *in-vitro* translation assay, the random forest algorithm predicted the binding energy of compound **2** to be in the 0.85 percentile, and that for compound **8**, in the 0.95 percentile.

Small molecules that inhibit *M. smegmatis* ribosomes

M. smegmatis, being a close but nonpathogenic relative of *M. tuberculosis*, is often used to test antitubercular agents due to the similar morphological and biochemical properties of these two species. Importantly, for the ribosomal PTC used in this study (2687-2850), there is 100% identity between *M. tuberculosis* and *M. smegmatis* (the identity of the whole 23S rRNA is 2873/3150, 91%). The inhibition effect of compounds **1-10** on *M. smegmatis* ribosomes was tested in a bacterial coupled transcription-translation assay (*see* Methods [16]) in which the expression of luciferase gene was measured. A reconstituted mycobacterial translation apparatus was obtained by mixing purified translation factors, ribosomes, tRNAs, and aminoacyl-tRNA synthetase mix, as described previously [17]. The full list of reagents and concentrations of components in our translation system is presented in the *Methods* section. The mycobacterial reconstituted translation system yielded 40% of the signal usually obtained for the *E. coli* translation system [18]. The results indicated effective inhibition of *M. smegmatis* ribosomes by molecules **2** and **8** (1-[(2-methylthiazol-4-yl)methyl]-4-[(2-phenylthiazol-4-yl)methyl]piperazine and 1-ethyl-4-[(2-phenylthiazol-4-yl)methyl]piperazine, respectively) (**Fig. 5**). Remarkably, the IC_{50} values for molecules **2** and **8** were 8.9 and 4.8 μM, respectively, appear to be superior to the IC_{50} for chloramphenicol (9.2 μM), a broad-spectrum antibiotic that binds the PTC, thereby inhibiting ribosome activity (**Fig. 5**).

Analysis of inhibitor-RNA interaction

In line probing analysis was used to show specific binding of the small molecules in hairpin 91 of the ribosomal PTC. In-line probing assays have been used to elucidate the secondary structure and binding properties to riboswitch aptamers of variety of metabolites and small molecules [19]. This assay utilizes a slow “in-line” nucleophilic attack on a phosphodiester bond in RNA by the 2' oxygen atom of a spatially adjacent nucleoside. This selective cleavage reveals secondary structural features of the backbone RNA when products of the cleavage are examined using gel electrophoresis. Binding of a small molecule to the RNA “locks" the RNA in an in-line rigid conformation and promotes cleavage of the linkage involved. The outcome is elevated intensities of the gel bands, indicating hotspot linkage in the RNA target. Rigidity of some linkages along the RNA made others less likely to adopt an in-line conformation, which is indicated by reduced gel bands attributed to the cleavage sites. Such change in the pattern of the RNA cleavage is observed in an in-line probing gel upon titration of the small molecules inhibitors, **2** and **8**, found as described in the previous section (**Fig. 6**).

Discussion

Screening for molecules that bind the ribosomal PTC

This study shows that T_2 relaxation is an extremely efficient and sensitive technique for screening for small molecules that bind RNA. As the NMR-fragment screening approach for finding inhibitors that target the PTC relies on a binding assay rather than a functional assay, it is not necessary to retain the peptidyl transfer activity in the PTC fragment used for the binding assay: The effect of the binding on the peptidyl transfer activity is measured at later stages of the development when drug-sized molecules have been identified.

Implications of FBVS for the field of drug design

A critical step in fragment-based drug discovery is to optimize the hits found during the screening process to develop compounds with better binding properties than the small fragments. Typically, this step is accomplished by using medicinal chemistry aided by NMR, but this technique is very time and resource consuming [8]. The *in-silico* FBVS approach enables us to reduce the number of steps in lead compound optimization [8,9]. Similarly, the virtual filtration followed by automated docking enabled us to select compounds that bind the

main cleft in the PTC structure. This step provided a fast, economic, and efficient route to two drug-sized molecules that inhibit *M. smegmatis* ribosomes. Importantly, this study thus provides the basis for a complete design workflow that can be used for different RNA targets.

Implications of machine learning for drug discovery

The machine learning modus operandi for further optimization of inhibitors that was developed here constitutes a new paradigm in drug discovery. The data obtained from docking of the drug-sized molecules to the RNA target was used to train machine learningmodels, from which advantageous physicochemical characteristics of the molecule scaffold were deduced. Identifying the propensity for binding produced useful information [through investigation of many structure-activity relationships (SAR)] that could be applied for optimization of the compounds into a potent lead molecules.

In this study, we described how machine learning can be exploited in the design of inhibitors with increased efficiency. The success of our model in predicting the two best inhibitors for the ribosomal PTC provides an important step in rational drug design, where machine learning can be used as an integral part of the optimization phase, by indicating *in advance* which type of molecules to consider for testing and *in hindsight*, after fragment screening, for analyzing the obtained data and drawing further conclusions

Implications of our study for antimicrobials and antimicrobial resistance

TB, the disease caused by *M. tuberculosis*, is associated with significant morbidity and mortality, mainly due to the lack of suitable treatment regimes for drug-resistant strains. However, TB is not the only disease whose treatment is affected by the emergence of antibiotic-resistant pathogens, and the workflow for lead compound discovery elaborated in this study can certainly be applied for drug discovery for treating other infectious diseases. By developing molecules that target the translation of mRNA by bacterial ribosomes to create proteins, we aim to inhibit the bacterial translational machinery and hence to eliminate the pathogenic bacteria. Our method, which combines experimental and computational techniques, provides proof of concept for the development of inhibitors that target the ribosomes of mycobacterial cells. The two lead compounds that we discovered could potentially be passed on to the pharmaceutical industry for testing and development into new drugs for treating active TB. The broader implications of our work are that our innovative scientific methodology will be applicable to other antibiotic-resistant microorganisms, with the attendant medical, social and economic advantages.

Methods

All chemicals were obtained from Sigma, unless otherwise mentioned. Creatine phosphate and creatine kinase were obtained from Roche Molecular Biochemicals. Small molecule inhibitors were purchased from UORSY, Ukraine. RNase T1 and cAMP were from Thermo Scientific. Quick calf intestinal alkaline phosphatase and T4 polynucleotide kinase were purchase from New England Biolabs, and [γ–^{32}P] ATP (800 Ci mmol^{-1}), from Perkin Elmer. The luciferase assay kit was purchased from Promega.

RNA preparation

RNA (GCAUCCUGGGGCUGGAGCAGGUCCCAAGG) representing the sequence of hairpin 91 in the 23S rRNA of *M. tuberculosis* (G2754-G2782) for NMR screening was obtained from Integrated DNA technology. RNA for further *in-vitro* experiments was produced in high amounts and purity as follows. The full-length rRNA of *M. tuberculosis* 23S rRNA of *M. tuberculosis* (H37Rv ATCC 27294) was cloned in a pLitmus28 vector, and a fragment of the T7 promoter–PTC(164nt, U2686-C2850) – BSPQI (naagatc) inserted to pUC18 vector. The restriction enzyme BSPQI was used to linearize the plasmid DNA. Preparation of the DNA template for transcription, followed by *in-vitro* 'run-off' transcription and RNA purification was performed as previously described [20,21]. RNA *in vitro* transcription was carried out in a reaction mixture containing 8 µg of linearized plasmid, 18.5 µM T7 RNA polymerase, 1 mM for ATP, CTP, GTP and UTP, 10 units of RNase inhibitor, in an appropriate reaction buffer [80 mM HEPES-KOH pH 7.5, 10 mM spermidine, 40 mM dithiothreitol (DTT), 25 mM $MgCl_2$]. Reaction mixtures (100 µL) were incubated at 37ºC for 4 h.

T_2-relaxation experiment

In the initial screening 10 fragment molecules from the Maybridge Ro3 2500 Diversity Fragment Library (Maybridge, Inc.) were pooled, mixed with the above-described 29mer RNA molecule that represents hairpin 91 of the *M. tuberculosis* PTC (2543-2571: GCAUCCUGGGGCUGGAGCAGGUCCCAAGG), 10 µM, and subjected to NMR (T_2 relaxation) screening. The RNA used for screening was dissolved in proton-free phosphate buffer (PBS).

The Carr Purcell Maiboom Gill (CPMG) sequence was used to determine the relaxation time T_2 of the small molecules in their free form and in the presence of the biomolecule. The large difference in relaxation times between the bound and unbound forms was used to filter out resonances originating from the biomolecule and the complex by using a long echo time in the CPMG experiment.

In silico molecular docking

Molecular docking simulations were performed using AutoDock 4.2 [22] to estimate the binding free energy (ΔG_{bind}) and the poses of the investigated compounds in relation to the receptor. The PTC receptor for simulations was derived from the Cartesian coordinates of the large ribosomal SA50S originating from *S. aureus* (PDBID 4WCE) [3]. Cognate ligands chosen for this purpose were imported from the ZINC database [10]; all demonstrated at least 70% similarity with the putative inhibitor molecules suggested by the NMR experiments. The virtual screening protocol was conducted through the Raccoon implementation (http://autodock.scripps.edu/resources/raccoon), in which the receptor and ligand molecules were preprocessed for docking. The docking grid was set with 126 points in each dimension and the default spacing was 0.375 Å. The obtained grid map was centered with respect to the receptor. Free energy calculations and conformational sampling of the ligands were then carried out using the Lamarckian genetic algorithm (LGA), with an initial population size of 150 individuals, 2,500,000 free energy evaluations and 27,000 LGA generations. Clustering of the results was performed by root mean-square deviation (RMSD) calculations for the obtained poses of the ligands, with a constant tolerance of 2.0 Å. Further analyses of the results were performed using default AutoDock VS tutorial scripts [23] along with several in-house written scripts.

A machine learning guided approach to inhibitor design

Data preparation: The data contained 811 molecules with a common scaffold. The remaining 107 molecules from the initial number of molecules used (919) were aligned in another coordinate system, and were therefore removed. Each molecule was composed of 20 to 51 atoms, 33 on average, 11 of which belonged to the scaffold. The raw data contained the coordinates of each atom relative to the center of the molecule. The coordinates were used to compute a total of 33 features per molecule, which are explained in **Table S4**. The atoms of the scaffold were numbered from 1 to 11, and for every location i three features were computed (33 overall). The feature DISTi is a proxy for the volume of the structure

connected at location i – the larger the average distance, the bigger the structure. The feature VARi is a proxy for the measure of regularity or smoothness of the structure – the smaller the variance the more regular the structure (if, for example, the variance is 0 it means that the distance between every two pairs of atoms is exactly the same; in contrast, a spiky structure results in a large variance, **Fig. S3**). In addition to the 33 independent variables, the dependent variable for each molecule was BOND, which holds the binding energy given by the docking results.

Summary statistics: At least one of locations 1, 2, 4, 5, 6, 8, and 9 was modified in 98% to 100% of the molecules. Locations 3, 7 and 10 were never modified. Location 11 was modified in 1% of molecules. The binding energy BOND ranged from -15.680 (strongest) to -4.77 (weakest). The average was -9.85 with standard deviation of 1.89.

Machine learning algorithms used: a. Random forest. Random forests constitute a well-known and widely used ensemble learning method for classification or regression that operates by constructing a multitude of regression trees at training time and outputting a value that is the average of the predictions of the individual trees. Random forests are widely used in a multitude of machine learning tasks and are considered to be among the top methods in terms of predictive power. The main advantage of random forests over linear regression is the fact that the regression surface is not linear.

We trained a random forest with 50 trees, with no restriction on the tree-depth or the number of features to be considered for each node split. The RMSE (root mean squared error) of the model prediction was 0.56 (less than third of the standard deviation of BOND). We trained a second random forest model, this time using only the 11 MODi features. The resulting RMSE was 1.86, more than three times larger. Alongside the RMSE, a useful and commonly used statistic of the random forest is called *feature importance*. The feature importance is computed as follows: for every feature, all the splits where this feature was used as the splitting criterion are examined, and the average decrease in variance is computed (variance of the node before the split minus the pooled variance of the two children nodes after the split). The larger the difference, the more important is the feature. The features were sorted according to their feature importance value, and the values are presented in **Table S5**: location 8 was the most important one, then – by an order of magnitude – locations 6, 9 and 5. The remaining locations were at least another order of magnitude lower.

b. Linear regression. A regression model was trained, and a forward-backward stepwise procedure was used to select significant features. The significant features (with *p*-value

smaller than 0.05) were (in order of significance) DIST*8*, MOD*8*, VAR*8*, DIST*5*, DIST*9* and VAR6. We then regressed BOND on those features and obtained a regression model with an adjusted R^2 value of 0.23; F(6,804) was 41.96 and *p*-value:< $2.2e^{-16}$. The regression coefficients are shown in **Table S6**.

The fact that the coefficient of all the DIST variables is negative implies that increasing their value will decrease (strengthen) BOND. Similarly the fact that VAR*8* is positive implies that increasing VAR*8* increases (weakens) BOND. Next, we regressed BOND against the six variables in the above table, and in addition all the pairwise interaction variables (e.g., VAR*6**DIST*5*). The regression adjusted R^2 increased to 0.25 and F(15,795)=19.11 giving p-value below 2.2e-16. Interestingly, VAR*6* became insignificant, and the product VAR*6**DIST*8* became significant. The second significant interaction was VAR*8**DIST*8* with a negative coefficient -0.9. VAR*8* and DIST*8* remained significant with coefficients 5.88 and -4.25, respectively. This means that to have a strong bonding energy, i.e., negative BOND, the structure attached at location 8 needs to be either very regular and large, or irregular and large so that the product -4.25*VAR*8**DIST*8* is much larger than 5.88*VAR*8*-4.25*DIST*8*. This is feasible since the product grows faster than the linear terms.

In-vitro translation

The inhibition effect of compounds **1-10** and chloramphenicol as a reference compound on *M. smegmatis* ribosomes was tested in a bacterial coupled transcription/translation assay, in which the expression of luciferase gene was measured. The luciferase gene was inserted into the plasmid downstream from the T7 RNA polymerase promotor. The reaction mixture contained: 160 mM HEPES-KOH (pH 7.5), 6.5% polyethylene glycol 8000, 0.074 mg/ml tyrosine, 1.3 mM ATP, 0.86 mM for CTP, GTP and UTP, 208 mM potassium glutamate, 83 mM creatine phosphate, 28 mM ammonium acetate, 0.663 mM cAMP, 1.8 mM DTT, 0.036 mg/ml folinic acid, 0.174 mg/ml *E. coli* tRNA mix, 1 mM of each amino acid, 0.25 mg/ml creatine kinase, 0.044 mg/ml T7 RNA polymerase, *E. coli* cell-free extract, 0.003 g/l luciferase-encoding plasmid and compounds **1-10** in concentrations ranging from 152 nM to 1 mM. The effect of these compounds, at a final concentration of 300 nM, was also tested against *M. smegmatis* ribosomes. The reaction mixture was incubated at 37 ℃ for 1 h and terminated by the addition of erythromycin at a final concentration of 8 µM. To quantify the reaction products, luciferin assay reagent (LAR, Promega) was added at 5:3 (luciferase:

reaction mix) volume ratio, and luminescence was measured. The results were plotted (compound concentration vs. luminescence intensity) and IC_{50} values were calculated.

In line probing

in the in-line probing assay was performed as described previously [17]. An RNA construct was prepared using T7 RNA polymerase (*see* RNA preparation). The RNA 5' phosphate was removed by quick dephosphorylation kit (New England biolabs) and phosphorylated by T4 polynucleotide kinase (New England biolabs) and [γ -^{32}P]ATP (3000 Ci/mmol) (Perkin Elmer). The samples were loaded onto 10% polyacrylamide denaturing gel (8 M urea). The gel bands were detected by autoradiography, and slices containing monodispersed RNA were excised, recovered, and dissolved in ultrapure water. Labeled RNAs were incubated at 25 °C either with or without the compounds (1 mM, 250 μM, 62 μM, 16 μM) for 50 h in 10 μl reactions containing 20 mM $MgCl_2$, 100 mM KCl, and 50 mM Tris-HCl. The incubation times with RNase T1 and alkaline buffer were 2 and 5 min, respectively. The samples were dried, dissolved in 5 M urea, and loaded on 10% polyacrylamide gel. The gels were visualized using autoradiography.

Figures

Figure 1. Plan for development of small molecule inhibitors for ribosomal PTC. The approach combines NMR-fragment based screening with virtual screening. Using NMR (T_2 relaxation) and a fragment library, we identified scaffolds that bind hairpin 91 in the ribosomal PTC of *M. tuberculosis*. These scaffolds were used to filter larger compounds containing the fragment molecules from the ZINC database [10]. Nearly 1000 compounds containing phenylthiazole were docked, using Autodock software [10,24] into the PTC of the available crystal structure of a bacterial ribosome, and hits were ranked on the basis of the binding energy. Conclusions as to structure activity relationships were drawn using machine learning algorithms, and ten compounds were selected and tested for their ability to inhibit translation in *M. tuberculosis*.

Figure 2. RNA molecules used in this study. a, The 50S ribosome subunit of *S. aureus* (PDBID: 4WCE, [3]) contains two RNA molecules (white) and 34 proteins (blue). b, The sequence of RNA hairpin (hairpin 91) used for T_2 relaxation screening. The modified nucleotides in the sequence for *M. tuberculosis* are marked in arrows. The structure of the

hairpin 91 RNA construct used in this study is presented as a cartoon. A superposed hairpin 91strucuture of the ribosomal PTC of *M. smegmatis* [4] is shown in grey. Root mean square (RMS) values for PTC and hairpin 91 of *S. aureus* and *M. smegmatis* are 0.858 and 0.788, respectively. c, Thermal stability obtained from a CD spectrum of *S. aureus* (Sa) PTC-RNA compared to that of *M. tuberculosis* (*Mtb*). The full spectrum is presented in **Fig. S1**.

Figure 3. Identification of N-methyl-4-(1,3-thiazol-2-yl)benzylamine from a T_2-edited CPMG mini-screen. Top: CPMG T_2-edited spectra with CPMG delay times of pool number 2 of three pools consisting of 10 fragments each, with (red) and without (blue) the target RNA. All spectra are expanded to show the aromatic region. Middle: The N-methyl-4-(1,3-thiazol-2-yl)benzylamine fragment that shows evidence of binding to the target RNA. Bottom: Singleton proton spectra of the constituent fragments from pool number 2 of the mini-screen. Fragment number 2, second from the top, is identified as a binder, as indicated on the figure.

Figure 4. Machine learning guided structure-activity relationships. Each location is numbered, and its influence on the value of the bonding energy is marked with a color: red – very significant, yellow – medium significance, teal – low significance and grey not significant. b. Projection of 400 33-dimensional data points on the top two principal components of the data matrix. The 400 data points correspond to the first and fourth quartiles in terms of bonding energy. Blue points have the strongest bonding energy, and red the weakest. Clear separation can be observed between most of the blue and red points.

Figure 5. Effect of compounds on protein synthesis measured by in vitro translation reaction. a, Structures of ten phenylthiazol-containing molecules with the best docking binding energy (ΔG). b, The luciferase gene was inserted into a plasmid downstream from the T7 RNA polymerase promotor. The translation reaction contained: 160 mM HEPES-KOH (pH 7.5), 6.5% polyethylene glycol 8000, 0.074 mg/ml tyrosine, 1.3 mM ATP, 0.86 mM U/G/CTP, 208 mM potassium glutamate, 83 mM creatine phosphate, 28 mM ammonium acetate, 0.663 mM cAMP, 1.8 mM DTT, 0.036 mg/ml folinic acid, 0.174 mg/ml *E. coli* tRNA mix, 1 mM of each amino acid, 0.25 mg/ml creatine kinase, 0.044 mg/ml T7 RNA polymerase, *E. coli* cell-free extract, 0.003 g/l luciferase-encoding plasmid and the compound to be tested (from compounds 1-10) in concentrations ranging from 152 nM to 1 mM. The reaction mixture was incubated at 37 ℃ for 1 h and terminated by the addition of excess erythromycin. The effect of the compounds was tested against M. *smegmatis* ribosomes. Inhibition of mycobacterial growth by compound **2**, 1-[(2-methylthiazol-4-yl)methyl]-4-[(2-phenylthiazol-4-yl)methyl]piperazine, and compound **8**, 1-ethyl-4-[(2-phenylthiazol-4-yl)methyl]piperazine),

in the indicated concentrations is presented. The IC_{50} values for molecules **2** and **8** are 8.9 and 4.8 μM, respectively. The IC_{50} value of chloramphenicol is 9.2 μM.

Figure 6. In-line probing gel image of ribosomal PTC of *M. tuberculosis*. The RNA segment used for this experiment (2687-2850) was in-vitro transcribed and purified (**Fig. S4**). The reaction contained 5 nM RNA labeled with ^{32}P at its 5′ terminus and in-line reaction buffer (50 mM Tris HCl, pH 8.3, 20 mM $MgCl_2$, and 100 mM KCl). After incubation at 25 °C for 50 h, the RNA product was analyzed by electrophoresis through 10% 29:1 polyacrylamide gel containing 8 M urea and visualized using autoradiography. Right: Lane 1-4 contains the untreated RNA (C1), partially digested using T1 RNase (T1), alkaline digested (-OH, C2). Lane 4-5 shows cleavage pattern of the RNA in the absence of the small molecules (C3) and in the presence of DMSO (D). Lanes 6-9 and lanes 10-13 contain compound **2** and compound **8**, respectively, in the indicated concentrations. Left: Structural model of the RNA shows hotspots of scission of PTC in the presence of the compounds **2** and **8**.

Acknowledgments

This research was supported by Grant No. 2016142 from the United States-Israel Binational Science Foundation (BSF) and the IMTI (TAMAT)/Israel Ministry of Industry–KAMIN Program, Grant No. 59081.

References

1. Verma, M., Furin, J., Langer, R. & Traverso, G. Making the case: developing innovative adherence solutions for the treatment of tuberculosis. *BMJ Glob Health* **4**, e001323 (2019).
2. Zhang, W., Dunkle, J.A. & Cate, J.H. Structures of the ribosome in intermediate states of ratcheting. *Science* **325**, 1014-7 (2009).
3. Eyal, Z. et al. Structural insights into species-specific features of the ribosome from the pathogen Staphylococcus aureus. *Proc Natl Acad Sci U S A* **112**, E5805-14 (2015).
4. Hentschel, J. et al. The Complete Structure of the Mycobacterium smegmatis 70S Ribosome. *Cell Rep* **20**, 149-160 (2017).
5. Polacek, N. & Mankin, A.S. The ribosomal peptidyl transferase center: structure, function, evolution, inhibition. *Crit Rev Biochem Mol Biol* **40**, 285-311 (2005).
6. Auerbach-Nevo, T. et al. Ribosomal Antibiotics: Contemporary Challenges. *Antibiotics (Basel)* **5**(2016).

7. Baker, M. Fragment-based lead discovery grows up. *Nat Rev Drug Discov* **12**, 5-7 (2013).
8. Singh, M., Tam, B. & Akabayov, B. NMR-Fragment Based Virtual Screening: A Brief Overview. *Molecules* **23**(2018).
9. Ilic, S. et al. Identification of DNA primase inhibitors via a combined fragment-based and virtual screening. *Sci Rep* **6**, 36322 (2016).
10. Irwin, J.J. & Shoichet, B.K. ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model* **45**, 177-82 (2005).
11. Pervushin, K., Riek, R., Wider, G. & Wuthrich, K. Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc Natl Acad Sci U S A* **94**, 12366-71 (1997).
12. Wu, B. et al. High-Throughput Screening by Nuclear Magnetic Resonance (HTS by NMR) for the Identification of PPIs Antagonists. *Curr Top Med Chem* **15**, 2032-42 (2015).
13. Hajduk, P.J., Olejniczak, E.T. & Fesik, S.W. One-Dimensional Relaxation- and Diffusion-Edited NMR Methods for Screening Compounds That Bind to Macromolecules. *J Am Chem Soc* **119**, 12257-12261 (1997).
14. Congreve, M., Carr, R., Murray, C. & Jhoti, H. A 'rule of three' for fragment-based lead discovery? *Drug Discov Today* **8**, 876-7 (2003).
15. Jhoti, H. Fragment-based drug discovery using rational design. *Ernst Schering Found Symp Proc*, 169-85 (2007).
16. Matzov, D. et al. Structural insights of lincosamides targeting the ribosome of Staphylococcus aureus. *Nucleic Acids Res* **45**, 10284-10292 (2017).
17. Regulski, E.E. & Breaker, R.R. In-line probing analysis of riboswitches. *Methods Mol Biol* **419**, 53-67 (2008).
18. Srivastava, A. et al. Reconstitution of Protein Translation of Mycobacterium Reveals Functional Conservation and Divergence with the Gram-Negative Bacterium Escherichia coli. *PLoS One* **11**, e0162020 (2016).
19. Soukup, G.A. & Breaker, R.R. Relationship between internucleotide linkage geometry and the stability of RNA. *RNA* **5**, 1308-25 (1999).
20. Akabayov, B. Weizmann Institute of Science (2007).
21. McKenna, S.A. et al. Purification and characterization of transcribed RNAs using gel filtration chromatography. *Nat Protoc* **2**, 3270-7 (2007).
22. Morris, G.M. et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* **30**, 2785-91 (2009).
23. Cosconati, S. et al. Virtual Screening with AutoDock: Theory and Practice. *Expert Opin Drug Discov* **5**, 597-607 (2010).
24. Goodsell, D.S., Morris, G.M. & Olson, A.J. Automated docking of flexible ligands: applications of AutoDock. *J Mol Recognit* **9**, 1-5 (1996).

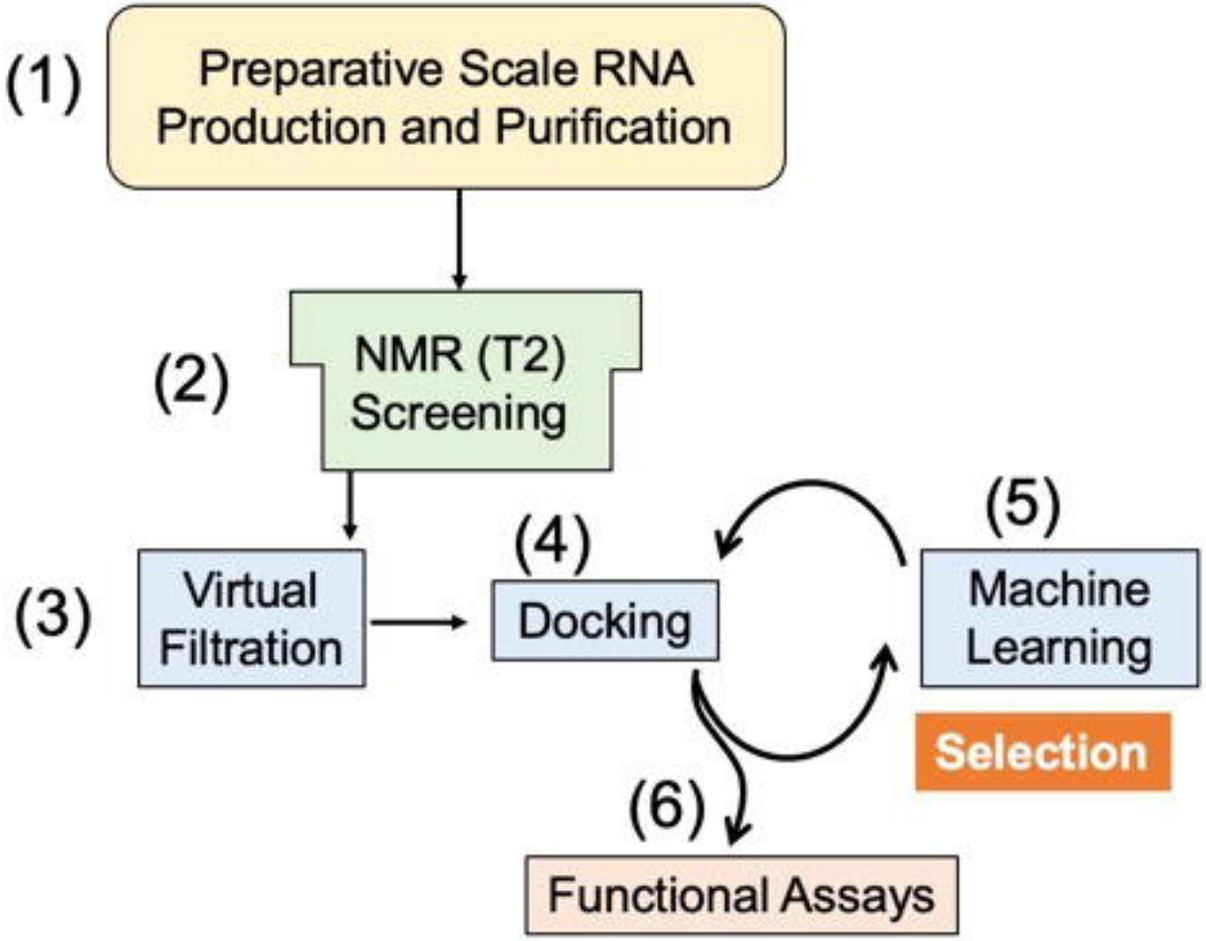

(1)
Preparative Scale RNA Production and Purification
(2)
NMR (T2) Screening
(3)
Virtual Filtration
(4)
Docking
(5)
Machine Learning
Selection
(6)
Functional Assays

Figure 1

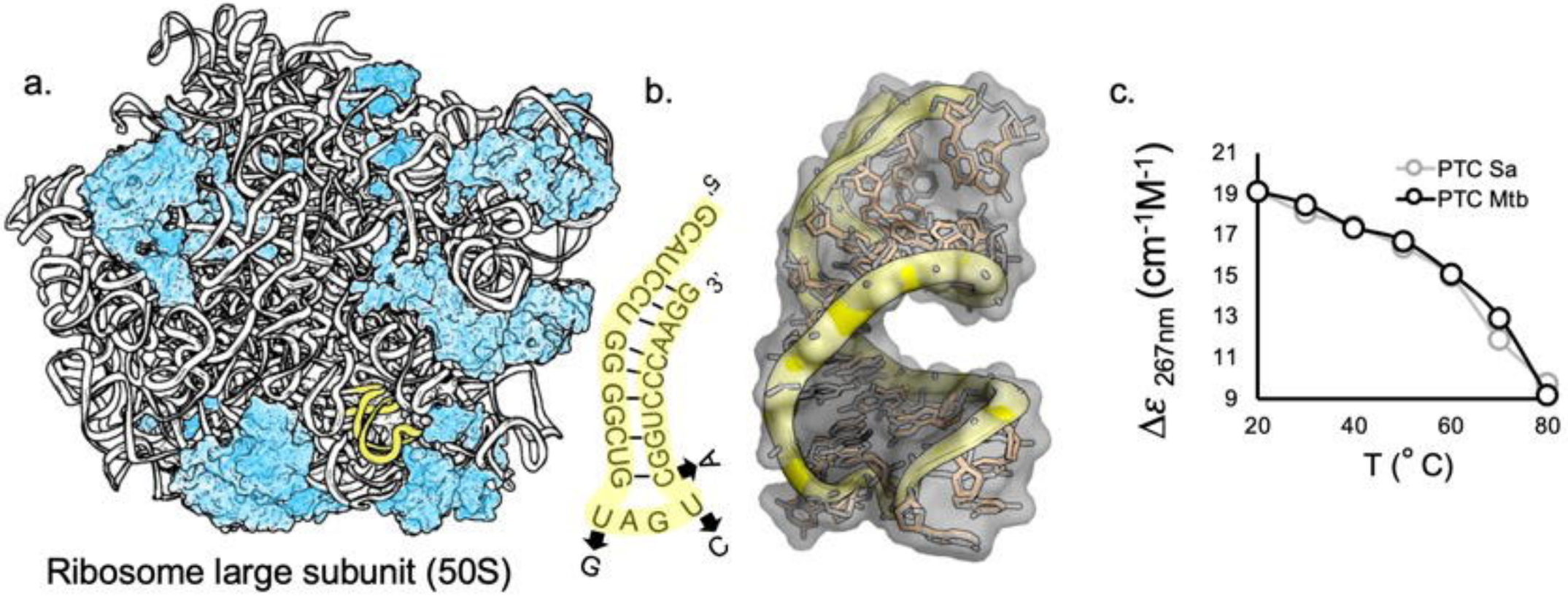

a.
Ribosome large subunit (50S)
b.
5'
3'
GCAUCCU
GG GGCUG
CGGUCCCAAGG
U A G U
A
C
G
c.
PTC Sa
PTC Mtb
Δε 267nm (cm-1M-1)
21
19
17
15
13
11
9
20
40
60
80
T (° C)

Figure 2

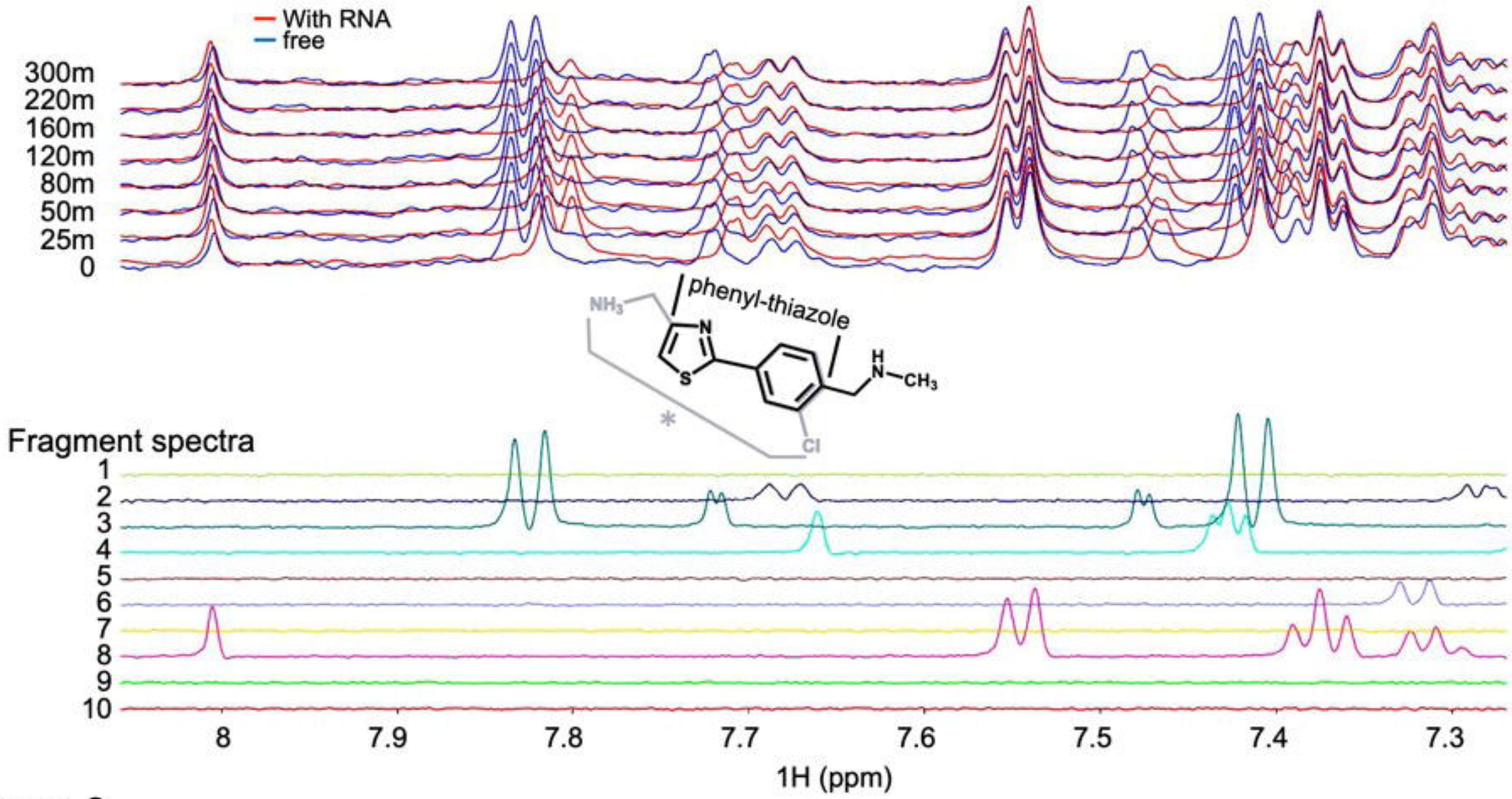
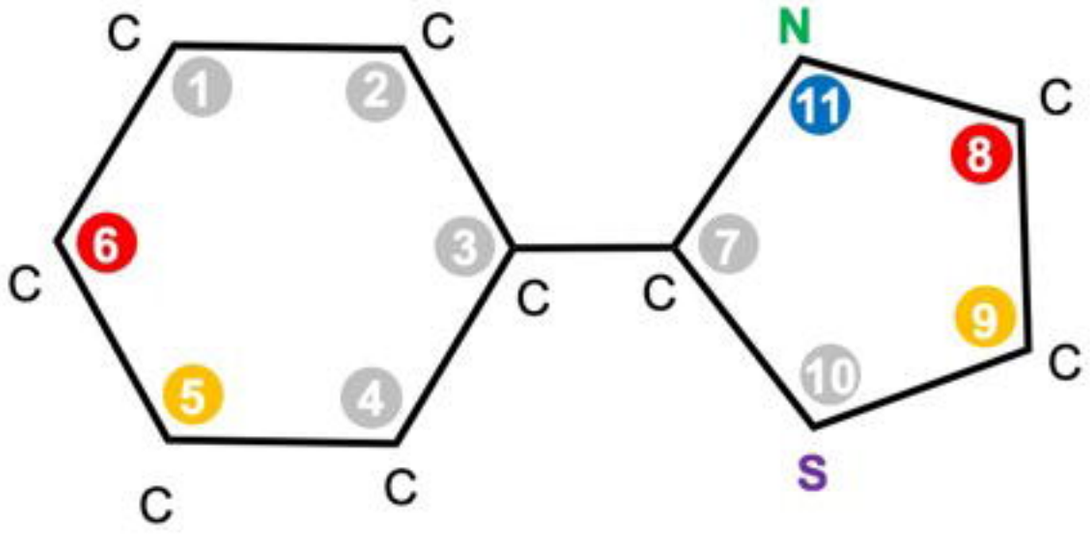

With RNA
free
300m
220m
160m
120m
80m
50m
25m
0
phenyl-thiazole
NH3
N
S
H
N
CH3
Cl
*
Fragment spectra
1
2
3
4
5
6
7
8
9
10
8
7.9
7.8
7.7
7.6
7.5
7.4
7.3
1H (ppm)

Figure 3

a.



b.

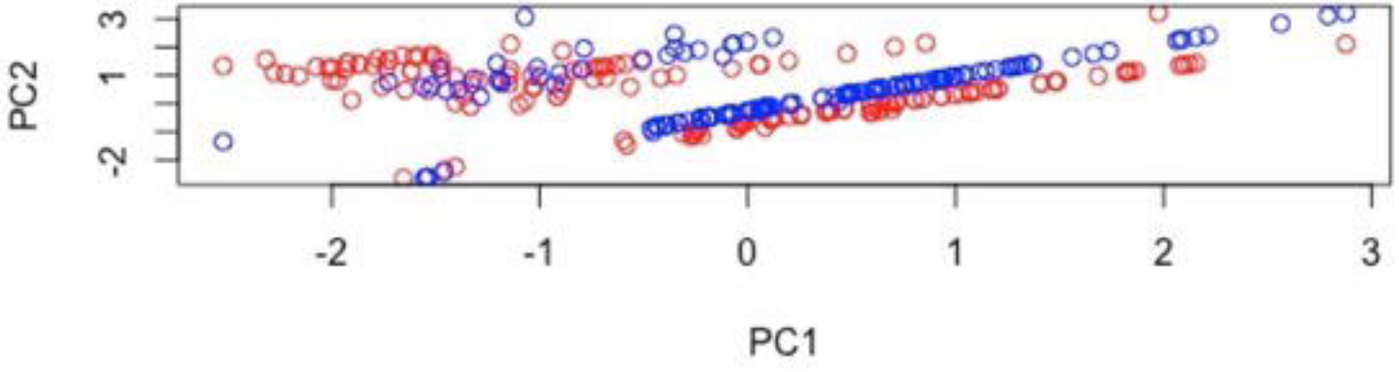
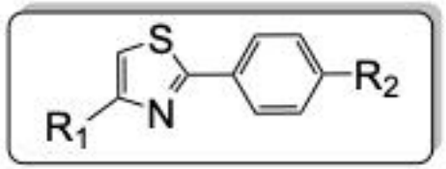
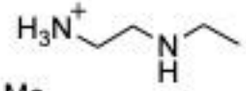
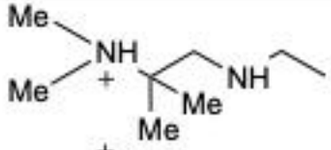
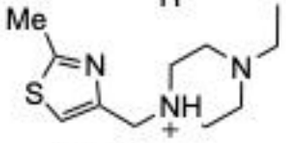
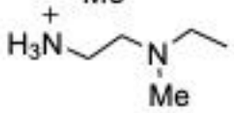
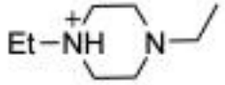
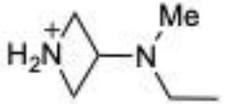
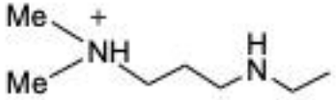


Figure 4

a.



	R1	R2		R1	R2
(1)		F	(6)		Me
(2)		H	(7)		Me
(3)		Me	(8)		H
(4)		H	(9)		F
(5)		F	(10)		F

b.

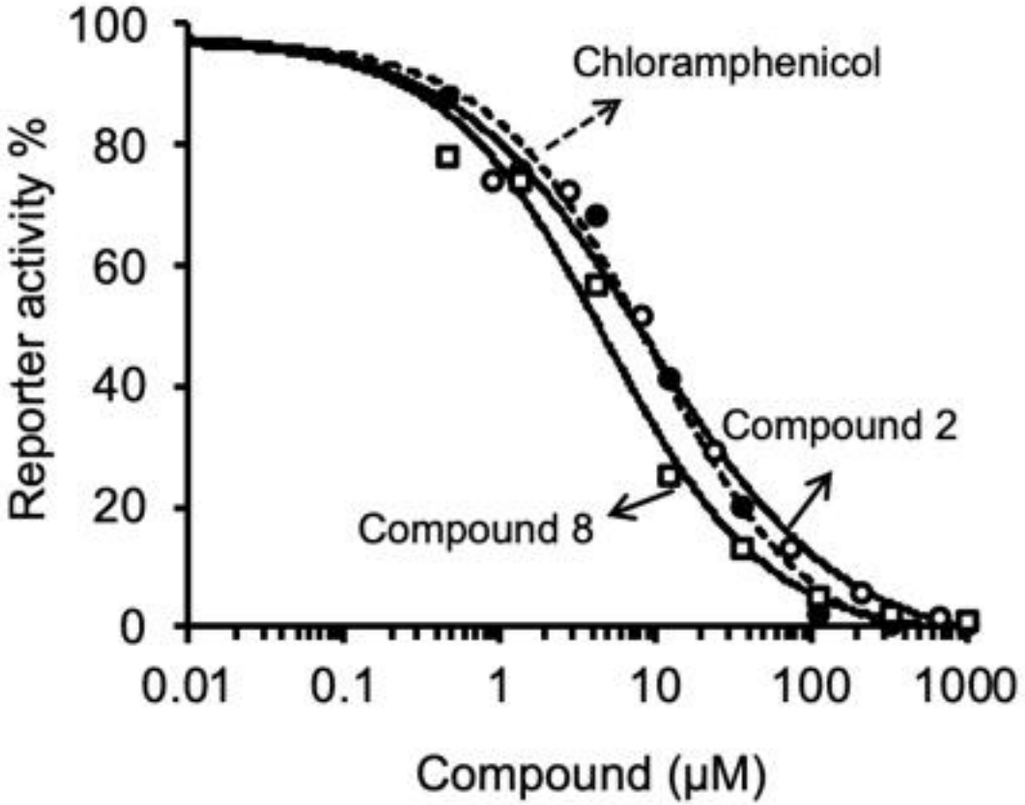


Figure 5

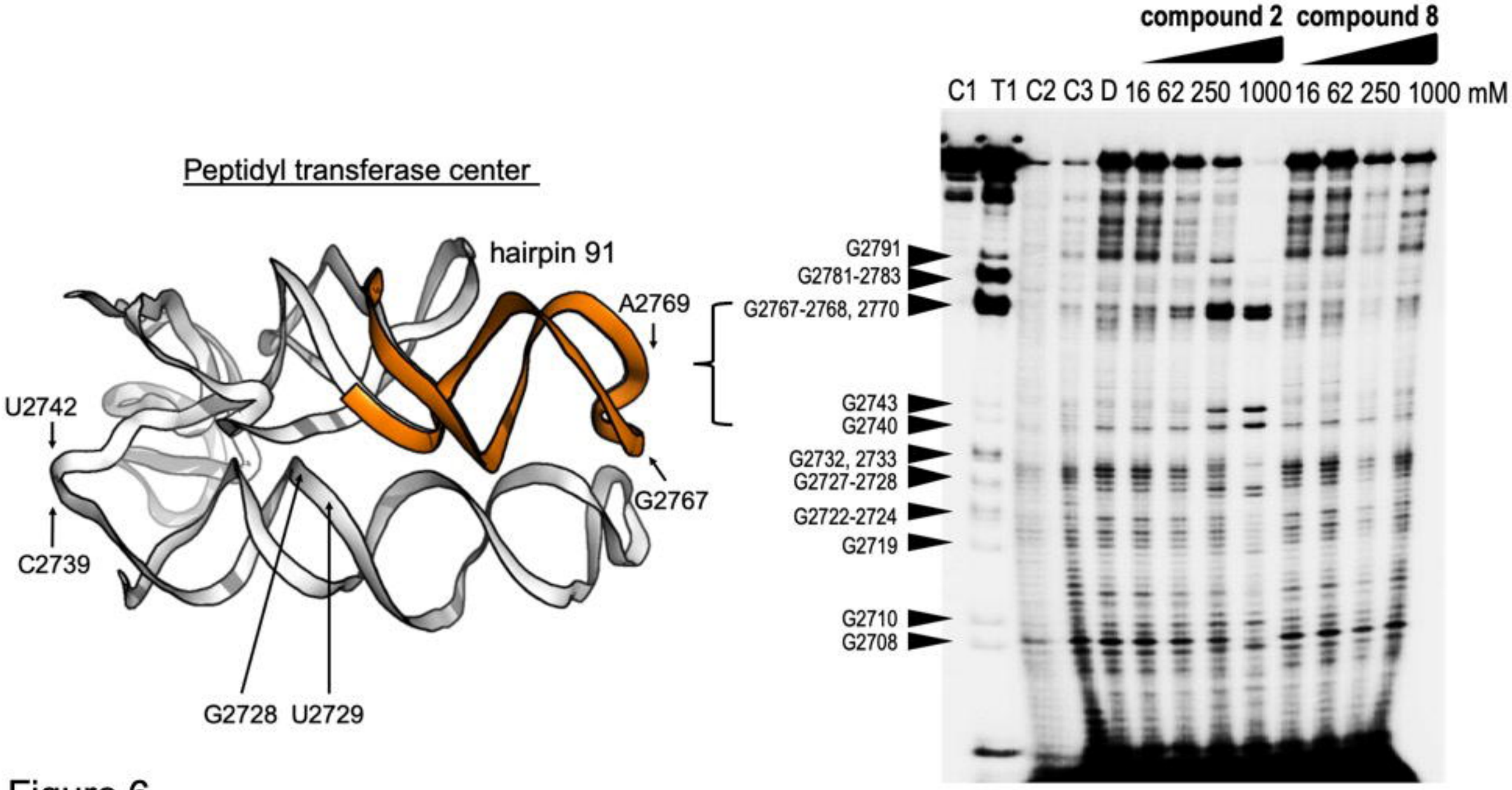

Peptidyl transferase center
hairpin 91
A2769
U2742
G2767
C2739
G2728
U2729
compound 2
compound 8
C1 T1 C2 C3 D 16 62 250 1000 16 62 250 1000 mM
G2791
G2781-2783
G2767-2768, 2770
G2743
G2740
G2732, 2733
G2727-2728
G2722-2724
G2719
G2710
G2708

Figure 6