DrugPred_RNA – A tool for structure-based druggability predictions for RNA binding sites

*Illimar Hugo Rekand and Ruth Brenk**

Department of Biomedicine, University of Bergen, Jonas Lies Vei, 5020 Bergen, Norway

*corresponding author: ruth.brenk@uib.no

ABSTRACT

RNA is an emerging target for drug discovery. However, like for proteins, not all RNA binding sites are equally suited to be addressed with conventional drug-like ligands. To this end, we have developed the structure-based druggability predicator DrugPred_RNA to identify druggable RNA binding sites. Due to the paucity of annotated RNA binding sites, the predictor was trained on protein pockets, albeit using only descriptors that can be calculated for both, RNA and protein binding sites. DrugPred_RNA performed well in discriminating druggable from less druggable binding sites for the protein set and delivered sensible predictions for selected RNA binding sites. In addition, the majority of drug-like ligands contained in a data set of RNA pockets were found in pockets predicted to be druggable, further adding confidence to the performance of DrugPred_RNA. The method is robust against conformational changes in the binding site and can contribute to direct drug discovery efforts for RNA targets.

INTRODUCTION

The vast majority of targets for approved drugs are proteins.[1,2] However, in recent years it has been increasingly realized that also RNAs constitute promising drug targets as they play a key role in many biological processes, can fold into diverse 3D structures, and specifically recognize small molecules.[3–6] By targeting RNA the functions of currently undruggable protein-mediated pathways and the non-coding transcriptome can be modulated and thus the size of the druggable genome can be increased considerably.[3] A prime example of an RNA drug target is the bacterial ribosome, where protein synthesis is inhibited through binding of small molecules.[7] This is illustrated by linezolid, an FDA-approved antibiotic, which acts by binding to ribosomal RNA (Figure 1).[8] Another active research area are RNA-binding splicing modifiers for the treatment of spinal muscular atrophy with several compounds in clinical trials.[9,10] Riboswitches, which are non-coding RNA structures in the 5'untranslated region and regulate gene expression through metabolite binding are new RNA drug targets for antibiotics.[11,12] For example, compounds binding to the flavin mononucleotide (FMN) riboswitch, e. g. ribocil and 5FDQD, have been shown to kill bacteria (Figure 1).[13,14] Riboflavin is known to bind to both the FMN riboswitch and riboflavin kinase. In both binding sites, the ligand is recognized in a similar way forming hydrophobic contacts and hydrogen bonds between the surrounding residues and the pteridine ring system, the dimethylbenzene ring, and the ribose chain. This fact nicely illustrates the capability of RNA to make specific molecular interactions with a wide variety of functional groups and ligand surfaces.[3]

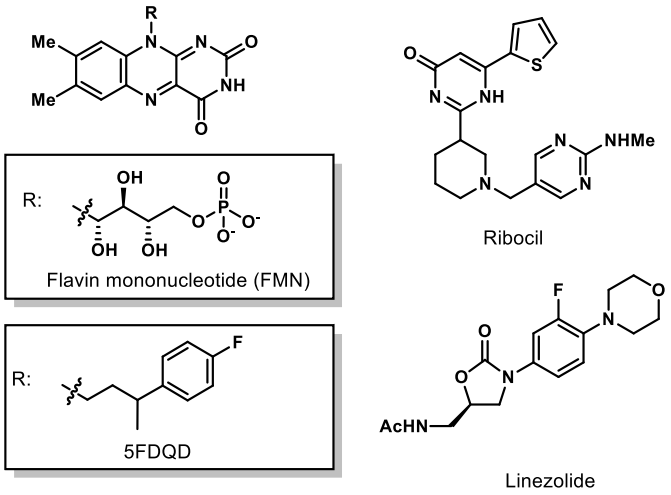


Figure 1. Examples of RNA-binding small molecules.

When targeting RNA, the question arises which targets are best suited for drug discovery and where in chemical space to look for potent ligands. Analysis of RNA-binding small molecules has revealed that some RNA ligands have drug-like properties comparable to FDA approved drugs while others lie outside this space.[4,15] Warner et al. have argued that RNA targets that bind such drug-like molecules and are thus deemed to be "ligandable" hold the greatest promise.[3] Consequently, tools are needed to identify such targets.

Targets are commonly considered to be "ligandable" or "druggable" if they possess binding sites that allow them to bind orally bioavailable drugs with high affinity.[16,17] The terms to name such pockets are hotly debated and several alternative terms such as "bindability", "tractability" or "chemical tractability" have been proposed.[17] We will use the term "druggability" throughout this manuscript because it is the prevalent term used in literature. Druggability is not an absolute property and for other pockets potent drugs can be developed, albeit larger efforts might be required. According, we will label pockets that are not classified to be druggable as "less-druggable".

Over the last few years, several methods have been reported that are able to segregate druggable pockets from less-druggable ones based on the 3D structure of the binding site.[17] Typically, these methods use descriptors describing the hydrophobicity, size and shape of the pockets to classify the them using machine learning methods. As training and validation sets, protein pockets that have been assigned to either category are used. One of these methods, the DLID (drug-like density) measure,[18] has also been applied to analyse RNA pockets. DLID uses PocketFinder[19] to identify potential binding sites and the descriptors volume, buriedness and hydrophobicity to estimate how likely a pocket is to bind a drug-like molecule. Warner et al. used this approach to illustrate the diversity of selected RNA binding sites.[3] Hewitt et al. conducted a comprehensive analysis of RNA structures in the PDB using the same method and concluded that many RNAs contain pockets that are likely suitable for small molecule binding.[20] However, they did not distinguish between the binding of drug-like ligands and other molecules.

In our group, we have developed DrugPred as a structure-based druggability prediction method.[21,22] DrugPred describes the size and shape of the binding site using a "superligand" as a negative print, which is obtained by merging predicted binding modes of drug molecules that were docked into the pocket using only steric constraints. Descriptors encoding polarity and size of the pocket are subsequently calculated based on the superligand and used to predict the druggability of the binding site. DrugPred was trained and validated on a set of non-redundant druggable and less druggable protein binding sites (NRDLD) which has become a standard in the field. In comparison studies, DrugPred performed at least equally well than other methods and achieved an accuracy of about 90%.[21,23,24]

Here, we adopted DrugPred for druggability predictions of RNA binding sites. Two of the original DrugPred descriptors could only be calculated for amino acids (the hydrophobicity indices of

amino acids and the relative occurrence of hydrophobic amino acids in the pockets).[21] Therefore, we have implemented alternative descriptors and thus made a prediction software which is applicable to both, protein and RNA binding sites. Compared to the protein field, there is very little data about ligands binding to RNA, and even less data that can be accessed in an efficient way. In the Protein Data Bank (PDB),[25] only 43 crystal structures containing only RNA as macromolecule are annotated with affinity data from PDBbind[26] mapping to about 20 unique sequences. The NALDB and SMMRNA databases contain affinities of small molecules binding to RNA extracted from the literature.[27,28] However, it is not possible to download the data for further processing. The R-BIND database links binding data to RNA crystal structures, but for only five of the ligands in this database a complex structure is available in the PDB.[29] Due to the paucity of suitable RNA data, we opted to train our modified DrugPred model, which we termed DrugPred_RNA, on protein data. Subsequently, DrugPred_RNA was used for druggability predictions of RNA structures including the ribosome. In the following, we present the construction of DrugPred_RNA together with its validation on protein and RNA binding sites.

METHODS

Scripts to download crystal structures from the PDB, process them and to calculate ligand and binding site descriptors were written using Python 3.6.8. with the Biopython (1.73) and RDKit (2019.09.1) libraries.[30,31]

NRDLD set for training and validation

As training and test set, our NRDLD set with the most recent modifications was used.[21,22] The binding sites and surrounding residues were carved out of the cif-files downloaded from the PDB by keeping all residues with an atom within 15 Å of the ligand. The isolated part of the structures

together with co-factors and metal ions if present were saved in the PDB format and used for descriptor calculation as described below.

Descriptor calculation

A superligand as a negative print of the binding site was obtained as done previously with minor modifications.[21] In brief, a set of approved drug molecules was docked into the pocket using DOCK 3.6.[32] Since the aim of docking was solely to obtain information about the shape and the volume of the binding site, all receptor atoms were set to carbon atoms and assigned a partial charge of 0. Subsequently, compounds for which a docking pose were obtained and for which the ratio of van der Waals (VDW) score to number of heavy atoms was ≤ -1.3 were merged into a superligand. However, during this process only atoms adhering to all of the following criteria were retained: 1) the atom had to be a non-hydrogen atom 2) at least two atoms coming from different docked compounds had to be closer than 1.2 Å 3) only one of the atoms within 1.2 Å from other atoms was kept in the final superligand. If no docked ligands passed these filters, the ligand contained in the original complex structure was used as superligand.

Based on the superligand, binding site and buried superligand atoms were determined. For that purpose, using FreeSASA[33] as implemented in RDKit, the solvent accessible surface area (SASA) of each receptor and superligand atom in the superligand-bound and superligand-unbound state was calculated using a 1.0 Å probe radius and ProtOr radii[34]. All receptor atoms for which the SASA differed between superligand-bound and unbound state were assigned as being binding site atoms. Likewise, all superligand atoms for which the SASA changed between the free and complexed state were assigned as buried superligand atoms and all superligand atoms with SASA> 0 Å in the unbound state were assigned as superligand surface atoms.

Using superligand and binding site atoms as input, descriptors describing the size, shape and polarity of the pocket were calculated (Table S1). For shape descriptors that are not based on the surface area or the number of receptor or superligand atoms, the Descriptors3D module of RDKit was used. For calculating polarity descriptors, we considered all carbon, phosphor, and sulphur atoms in addition to nitrogen atoms of the RNA bases that are bound to the ribose to be hydrophobic and all oxygen atoms of amino acids, ribose sugars and phosphate groups in addition to non-aromatic nitrogen atoms of amino acids to be polar. The SASA values of these atoms were calculated with FreeSASA using the same settings as described above. The side chains of histidine and tryptophane residues as well as the RNA bases are known to form hydrogen bonds in the plane of the heterocycles while parallel to this plane they engage in pi-stacking interactions which are more hydrophobic in nature. To account for this ambivalent behaviour, the SASA of endocyclic aromatic nitrogen atoms of bases and amino acid side chains and exocyclic oxygen and nitrogen atoms of the bases was split into a hydrophobic and a polar contribution in the following way. The SASA of these atoms was calculated both in the absence (SASA_total) and the presence (SASA_pol) of two blocking carbon atoms which were placed perpendicular to the plane of the aromatic ring with a 1.70 Å distance from the atom of interest. The area SASA_pol was considered to belong to a polar atom while the difference SASA_total – SASA_pol was considered to belong to a hydrophobic atom. Similarly, if more than half of these atom's SASA was deemed to be hydrophobic, the atom was included in the hydrophobic binding site atom count.

Training the predictive model using decision trees

Machine learning was carried using the decision tree algorithm eXtreme Gradient Boosting package (XGBoost)[35] in R[36]. As learning objective, logistic regression for binary classification with output probability was used. Thus, all binding sites obtained a score between 0.0 and 1.0,

whereas pockets with a score ≥ 0.5 were labelled druggable and pockets with a score < 0.5 were labelled as less druggable. Divergent from the default settings, the following parameters were used for training the model:

- *Max_depth* = 3 (Maximum depth of trees)
- *Scale_pos_weight* = 0.59 (Adjusts for the skewness between training and testing set)
- *Early_stopping_rounds* = 20 (Validation metric needs to improve at least once in every 20 rounds to continue training.).

The influence of the descriptors on the model was evaluated with the help of Shapley Additive Explanation (SHAP) values as implemented in the SHAPforxgboost package.[37,38] The same package was also used to make Figure 2 and Figure S3. Descriptors included in the final model were chosen by iteratively removing the least impactful descriptors until the predictive performance of the model was negatively affected. To further assess the robustness of the final model (called DrugPred_RNA), leave-one-out-cross validation was carried out yielding a training and testing error of 0.00342 and 0.127, respectively.

Assembly of data set with RNA binding sites

We selected RNA structures for druggability assessment by querying the PDB for structures containing only RNA and ligands (accessed November 2019). In addition, the PDB was searched for entries containing ligands and the keyword riboswitch to include structures which were excluded in the first query due to the presence of proteins. In total, this yielded 1084 structures. Subsequently, all structures that contained only ligands that were detergents, buffer salts or crystallization components were filtered out reducing the data set to 427 unique entries (Table 1, see supplementary material for three letter codes of rejected ligands). If a crystal structure

contained several instances of the same ligand, only the first instance was retained. In addition, all metal ions and water molecules were deleted (for a list of metal abbreviations see supplementary material). This resulted in 465 distinct binding sites spanning 224 unique ligands. A second variant of this set was also prepared. In this variant, only pockets with metal ions which were not more than 5 Å away from a ligand atom were retained. If a binding site contained several metal ions, several copies of the binding sites each of them containing one of the metal ions were prepared. This variant contained 343 entries. In the following, the first variant is called the metal-free and the second variant the metal-containing set. Further, a data set containing ligand binding sites in ribosome crystal structures was compiled by querying the PDB for structures that contained "ribosome" as keyword. These structures were treated as described above. In addition, the ligands were visually inspected to remove buffer components that had slipped the filter rules. This resulted in 731 binding sites in the metal-free ribosome set and 732 in the metal-containing set.

Table 1. Data sets of RNA and ribosomal binding sites for assessing DrugPred_RNA.

	RNA set (metal-free/metal-containing set)	**Ribosome set (metal-free/metal-containing set)**
Unique PDB IDs	427	590
Binding sites containing small molecule ligands	465/343	713/732
Unique ligands	224	247
Druggable entries	172/126	215/141

The binding site regions were carved out of the original cif-files by keeping all RNA residues with at least one atom within 15 Å of the ligand and potentially metal ions as described for the NRDLD set and subjected to descriptor calculations.

Determination of binding site similarity and consensus scoring

To investigate the robustness of DrugPred_RNA, binding sites were grouped based on binding site similarity. First, the binding site sequence of each pocket was generated by including all residues that contained at least one binding site atom (identified as described above) in ascending order while for modified nucleic residues the name of the corresponding unmodified residue was used (see supplementary material for a list of residue IDs for modified residues). Subsequently, all binding site sequences were pairwise aligned using BioPython and the global alignment similarity was calculated. If this value was > 85%, the pockets were assigned to the same family. As done previously, the consensus of the druggability predictions within each family (C) was calculated using the formula

$$C = \frac{|n_d - n_{ld}|}{N} \times 100\%$$

where n_d is the number of druggable binding sites within the family, n_{ld} is the number of less druggable binding sites and N is the total number of family members.[22]

RESULTS AND DISCUSSION

Construction of DrugPred_RNA

Compared to protein data, there is very little data about ligands binding to RNA and a data set of sufficient size composed of druggable or less druggable RNA bindings sites to train a druggability predictor could not be compiled. Therefore, we opted to predict the druggability of RNA binding sites by training a descriptor on protein binding sites and to subsequently apply it to the prediction

of RNA pockets. This approach required that only descriptors that can be calculated both for protein and RNA binding sites were used. This was not the case for our previously derived DrugPred model, as it contained the two descriptors "relative occurrence of hydrophobic amino acid" and "hydrophobicity indices of the amino acids".[21] Thus, a modified DrugPred model, termed DrugPred_RNA was derived. As training and test set, our NRDLD set of druggable and less druggable binding sites with the most recent modifications was used.[21,22] For all 110 binding sites in the NRDLD, 23 descriptors describing the size, shape and polarity were calculated (Table S1). Subsequently, the data set was divided into a training and test set as done previously[22] to train and evaluate a predictor. For DrugPred and DrugPred 2.0, partial least squares-discriminant analysis (PLS-DA) was used to model the data. However, using only protein-independent descriptors with PLS-DA resulted in worse predictions (data not shown). Therefore, we retreated to decision tree modelling based on XGBoost.[35] To avoid overfitting, the maximum depth of trees was limited to 2 and the early stopping option was used (Figure S1). In an iterative process, weak descriptors were removed until the predictive performance of the model was negatively affected. With the final model, termed DrugPred_RNA, of the 75 binding sites in the training set, 1 druggable pocket was misclassified as less druggable, and of the 35 binding sites in the validation set, 4 were misclassified (2 false positives and 2 false negatives) leading to accuracy, precision and recall values between 0.86 and 1.00 (Table 2 and Figure S2). With DrugPred_RNA, the error in the test set is slightly better than with DrugPred 2.0 while the error for the training set is slightly worse.

Table 2. Performance of DrugPred_RNA and DrugPred 2.0 on the training and test set of the NRDLD.

	Training set [druggable / less druggable]		Test set [druggable / less druggable]	
	DrugPred_RNA	**DrugPred 2.0**	**DrugPred_RNA**	**DrugPred 2.0**
Accuracy	0.99	0.91	0.91	0.94
Precision	1.00 / 0.97	0.92 / 0.89	0.95 / 0.86	0.95 / 0.93
Recall	0.98 / 1.00	0.94 / 0.86	0.91 / 0.92	0.95 / 0.93

The influence of the descriptors on the model output was evaluated with the help of Shapley Additive Explanation (SHAP) values which describe the importance of each descriptor for the model output taking into account the interactions with other descriptors.[37,39] Each descriptor for each data point (here, a particular binding site) is assigned a positive or negative SHAP value describing the contribution of the descriptor to the model output (here, druggable or less druggable) for that data point. The mean SHAP value formed by all SHAP values for a descriptor for the entire data set indicates the importance of the descriptor for the model (the larger the absolute mean SHAP value, the more important the descriptor, Figure 2A). For DrugPred_RNA, positive SHAP values imply high druggability probability, while negative SHAP values imply low druggability probability. Further, by plotting the individual SHAP values for a descriptor against the descriptor values, it becomes evident which descriptor values contribute positively or negatively to the model (Figure 2B). The sum of the SHAP values of all descriptors for a single data point indicates the direction of the prediction for that data point.

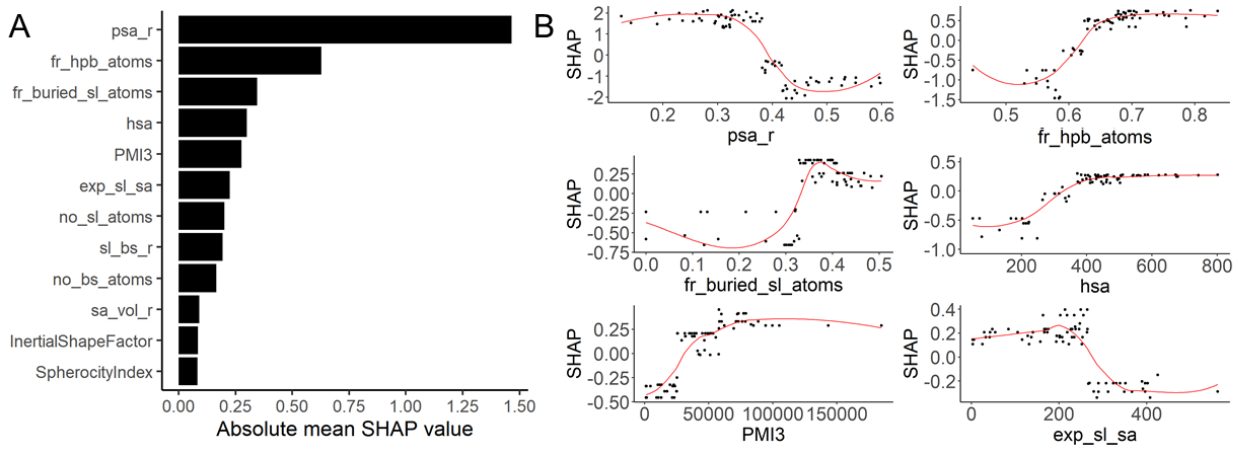


Figure 2. SHAP values for the DrugPred_RNA model. **A)** Absolute mean SHAP values for each descriptor ranked from highest to lowest impact on the model output. **B)** Individual SHAP values for each pocket in the training set for the top six descriptors in the model plotted against the descriptor values. Locally estimated scatterplot smoothing (LOESS) curves are overlaid on the descriptor observations (black dots). The midpoint in each curve indicates the cut-off value from where the prediction changes the direction. Positive SHAP values are associated with druggable and negative SHAP values with less druggable binding sites. The plots for the reaming descriptors are displayed in Figure S3.

The final DrugPred_RNA predictor was based on 12 descriptors (Figure 2A, Table S1). According to the SHAP values, the two most important descriptors were the relative polar surface area (*psa_r*, absolute mean SHAP value = 1.46) and the fraction of hydrophobic binding site atoms (*fr_hpb_atoms,* absolute mean SHAP value = 0.63), which both describe the polarity of the binding site. As expected, druggable binding sites were less polar than less druggable sites (Figure 2B and Figure S3). Both the high-ranking descriptor *fr_buried_sl_atoms* (absolute mean SHAP

value = 0.34) and the less important descriptor *sa_vol_*r (absolute mean SHAP value = 0.09) encode how compact a pocket is with less druggable pockets being more shallow (lower descriptor values for *fr_buried_sl_atoms* and higher values for *sa_vol_r*) than druggable ones. Further, two descriptors for the solvent accessibility of the pocket (*exp_sl_sa*, absolute mean SHAP value = 0.22 and *sl_bs_r,* 0.19) were included in the final model. Here, it was found that druggable binding sites were less solvent accessible than less druggable ones. The descriptor *hsa* was also found to be among the more important ones (absolute mean SHAP value = 0.30). This descriptor describes size of the surface area of hydrophobic binding site atoms and correlates roughly with the size of the pocket. Other descriptors describing the size of the pocket were also included in the model but had less influence on the predictions (*no_bs_atoms*, absolute mean SHAP value = 0.17, and *no_sl_atoms*, 0.20). In agreement with previous findings, druggable pockets were larger and more hydrophobic than less druggable ones. The descriptors *InertialShapeFactor*, *SpherocityIndex* and *PMI3* describing the shape of the superligand as a negative print of the binding site were also included in the final model. Pockets with a superligands with a larger third moment of inertia (*PMI3*, absolute mean SHAP value = 0.27) and which were less spherical (*SpherocityIndex*, absolute mean SHAP value = 0.08, *InertialShapeFactor*, absolute mean SHAP value = 0.08) were more likely to be assessed as druggable, albeit the latter two descriptors were determined to be less important.

Druggability predictions for RNA binding sites

Encouraged by the good performance of DrugPred_RNA on the NRDLD, we proceeded with druggability predictions for RNA and ribosomal binding sites. No benchmark set for the evaluation of RNA druggability predictions is available in the public domain. Therefore, using the PDB, we compiled two data sets for this purpose, one containing RNA only binding sites and one with

ribosome binding sites which in addition to ribosomal RNA could also contain ribosomal proteins. As binding sites, we considered all pockets that contained a ligand that is not a common crystallization buffer component. If a binding site contained metal ions within 5 Å of the ligand, several copies of the binding sites each of them containing one of the metal ions in addition to the metal-free pocket were prepared. In total, the RNA binding site set was composed of 427 unique PDB IDs spanning 465 binding sites in the metal-free and 343 in the metal-containing subset (Table 1). 224 different ligands were found in these pockets. The binding sites were grouped into different families whereas family members were required to have a binding site sequence similarity of > 85%. This resulted in 46 different families in the RNA set in addition to 234 singletons in the metal-free and in the 164 metal-containing set (Table S2). The ribosomal binding site set was prepared in a similar fashion resulting in 590 unique PDB IDs with 731 pockets in the metal-free and 732 in the metal-containing subset. 247 different ligands were bound to these pockets which were grouped into 52 different families while 440 singletons remained in the metal-free and 407 in the metal-containing set.

Subsequently, the druggability of the pockets in all sets was predicted. In the RNA data set, 36% of the binding pockets (metal-containing and metal-free combined) were predicted to be druggable while in the ribosomal data set 24% of the pockets were predicted to be druggable (see supplementary material for individual predictions for all pockets).

To assess the impact of metal ions on the druggability prediction, we compared the predictions of metal-free and metal-containing versions of same parent pocket. In both sets, for the majority of the cases (90% in the RNA set and 83% in the ribosome set) no change in the prediction outcome was found. Accordingly, metal ions had only a minor influence on the predictions. In the following, we therefore only present data for pockets which were stripped of metal ions.

Criteria for the assessment of druggability predictions for RNA binding sites

Next, the quality of the predictions of DrugPred_RNA on RNA binding sites was evaluated. The following aspects were considered for the evaluation 1) the agreement of the predictions with anecdotal examples, 2) the extent to which binding sites which are known to efficiently bind drug-like ligands were predicted to be druggable, and 3) the robustness of predictions with respect to substitutions and conformational changes in the binding sites.

For this assessment, the drug-likeness of the ligands was predicted using the quantitative estimate of drug-likeness (QED) score.[40] This score weighs multiple molecular features (e. g. molecular weight, number of hydrogen bond donors or acceptors, polar surface area, presence of unwanted functionalities) into one single unitless score, which ranges from 0 (undesirable) to 1 (desirable). Although this metric does not provide a clear cut-off to distinguish "desirable" from "undesirable" compounds, the authors denoted a mean score of 0.67 for attractive compounds, 0.49 for less attractive and 0.34 for too complex and unattractive compounds. Accordingly, in the following we classified compounds with a QED score $\geq$0.67 as drug-like, with a QED score $\leq$0.49 as less drug-like and compounds with a score in between as moderate drug-like.

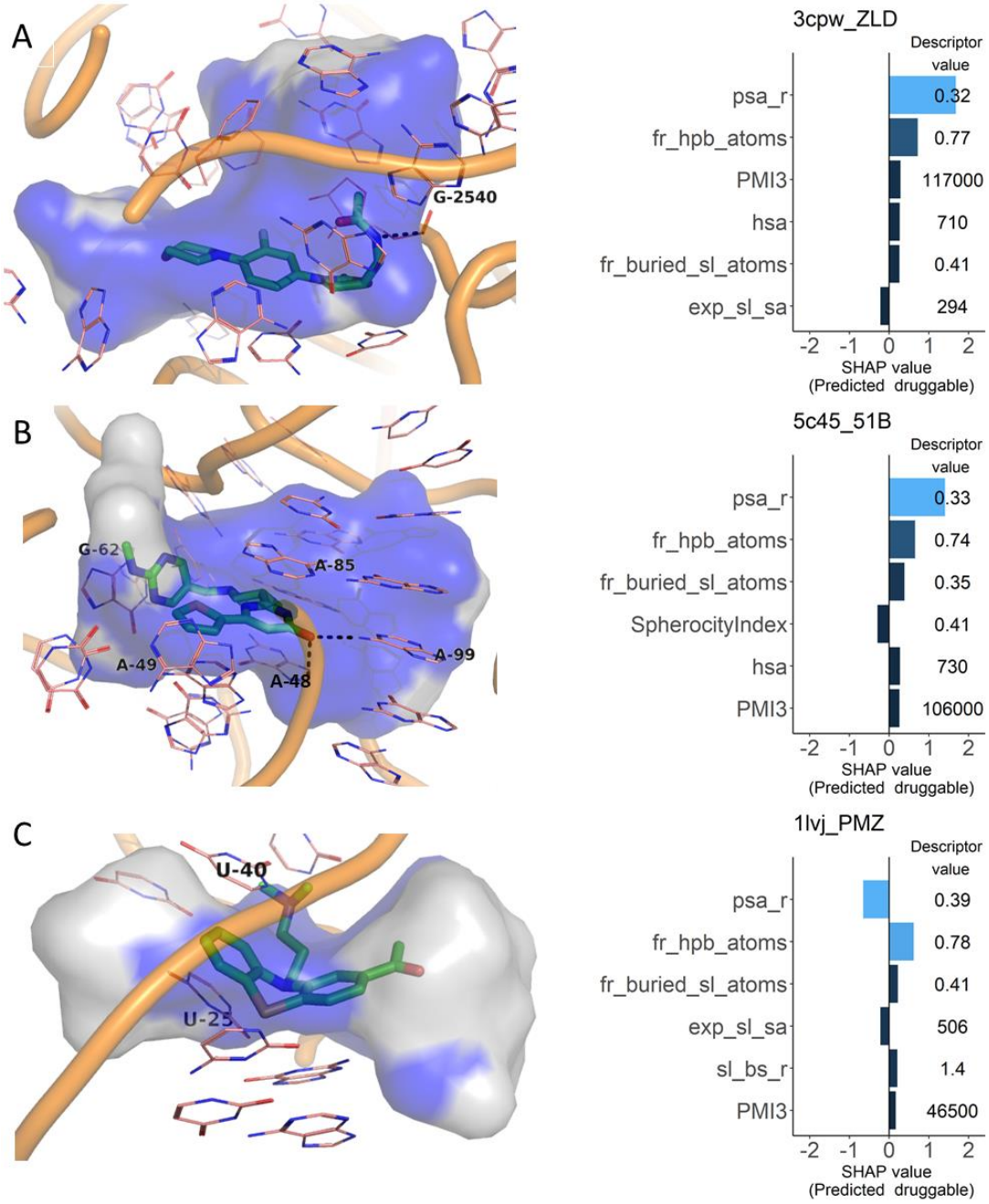
Further, for a binding site to be druggable, it needs to bind drug-like ligands potently.[16,17] Thus, if a potent drug-like ligand is known, one can with certainty say that a binding site is druggable. However, the absence of a potent drug-like ligand does not necessarily imply that a binding site is less druggable as there is always the possibility that a ligand can be optimized to increase its binding affinity. To take this into account, we used ligand efficiency (the binding energy normalized by the number of heavy atoms, LE) instead of binding affinity as a measure to judge if a ligand binds potently to its target.[41] Under the assumption that the ligand efficiency stays at its best constand during optimization, we considered ligands to bind potentilly to their target if they

had a LE of around 0.30 kcal·mol^{-1}·heavy atom^{-1} which translates to low nanomolar binding affinities of compounds with a molecular weight of maximum 500 Da. If no such ligand was known, we abstained from classifying a binding site on a general basis.

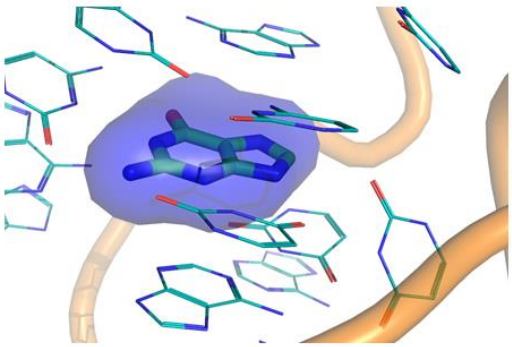
Evaluation of the performance of DrugPred_RNA based on anecdotal examples

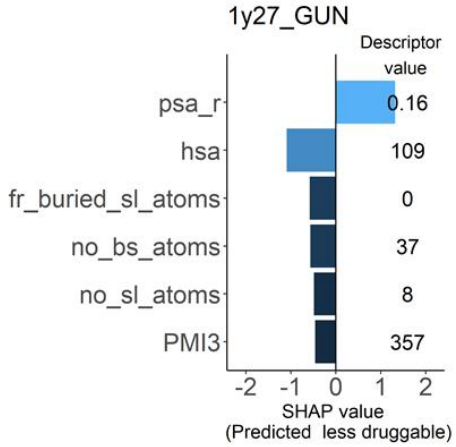
In the absence of a benchmarking set to assess the performance of RNA druggability predictions, we chose a few examples from the literature for a first validation of the predictions.

Linezolid is an FDA approved antibiotic targeting the 50S ribosomal subunit (Figure 1).[8] Based on its QED score of 0.89, it is highly drug-like. Its modest affinity of 20 µM translates to a LE of 0.27 kcal·mol^{-1}·heavy atom^{-1}. Linezolid is deeply buried in the pocket and forms mainly hydrophobic contacts in addition to a hydrogen bond to the ribose backbone of G2540 (Figure 3A). DrugPred_RNA predicted this pocket to be druggable. According to the individual SHAP values of the descriptor, the druggability was driven by the hydrophobicity of the pocket (*psa_r* = 0.32, *fr_hpb_atoms* = 0.77, *hsa* = 710 Å^2) and its shape (*PMI3* = 1,17 x 10^5, *fr_buried_sl_atoms* = 0.41) albeit the exposed surface area of the superligand being in a range that was more favourable for less druggable pockets (*exp_sl_sa* = 294 Å^2). Based on the binding mode of linezolid and the fact that linezolid is a drug-like ligand the prediction that this binding pocket is druggable appears to be sensible, despite the ligand not binding as potently as expected for a drug.

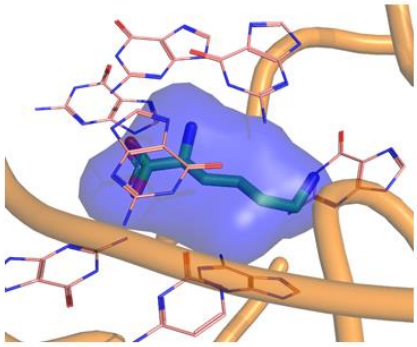

A
G-2540
3cpw_ZLD
Descriptor value
psa_r 0.32
fr_hpb_atoms 0.77
PMI3 117000
hsa 710
fr_buried_sl_atoms 0.41
exp_sl_sa 294
-2 -1 0 1 2
SHAP value (Predicted druggable)
B
G-62
A-85
A-49
A-48
A-99
5c45_51B
Descriptor value
psa_r 0.33
fr_hpb_atoms 0.74
fr_buried_sl_atoms 0.35
SpherocityIndex 0.41
hsa 730
PMI3 106000
-2 -1 0 1 2
SHAP value (Predicted druggable)
C
U-40
U-25
1lvj_PMZ
Descriptor value
psa_r 0.39
fr_hpb_atoms 0.78
fr_buried_sl_atoms 0.41
exp_sl_sa 506
sl_bs_r 1.4
PMI3 46500
-2 -1 0 1 2
SHAP value (Predicted druggable)

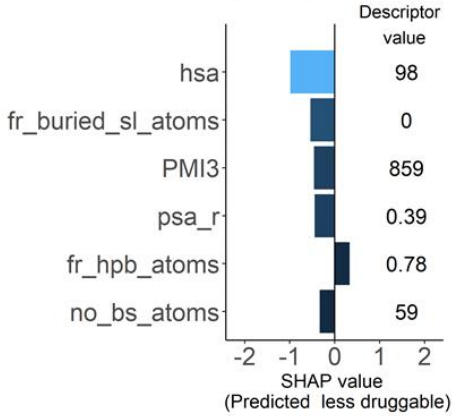
D



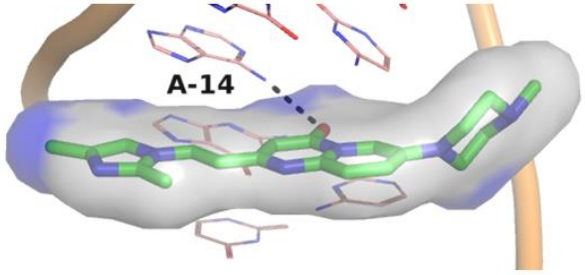

1y27_GUN
Descriptor value
psa_r
hsa
fr_buried_sl_atoms
no_bs_atoms
no_sl_atoms
PMI3
0.16
109
0
37
8
357
-2 -1 0 1 2
SHAP value
(Predicted less druggable)

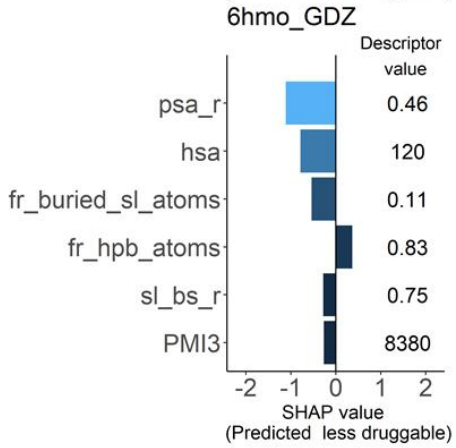
E




3d0u_LYS
Descriptor value
hsa
fr_buried_sl_atoms
PMI3
psa_r
fr_hpb_atoms
no_bs_atoms
98
0
859
0.39
0.78
59
-2 -1 0 1 2
SHAP value
(Predicted less druggable)

F


A-14


6hmo_GDZ
Descriptor value
psa_r
hsa
fr_buried_sl_atoms
fr_hpb_atoms
sl_bs_r
PMI3
0.46
120
0.11
0.83
0.75
8380
-2 -1 0 1 2
SHAP value
(Predicted less druggable)

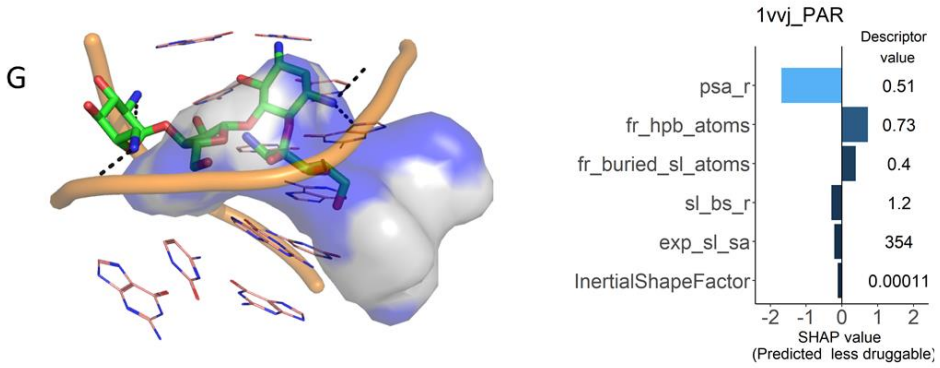


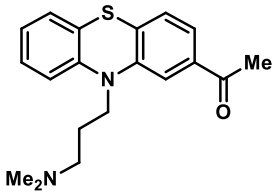
Figure 3. Evaluation of the performance of DrugPred_RNA based on anecdotal examples. The RNA backbones are shown as orange tubes, nucleobases as thin sticks with carbon atoms coloured pink and ligands as thick sticks with carbon atoms in green. The surface of the superligand created by DrugPred_RNA as a negative print of the pocket is shown as blobs with the solvent exposed surface area coloured grey and the remaining surface area coloured blue. Hydrogen bonds are indicated as dotted black lines. For each pocket, the individual SHAP values for the six most important descriptors together with the descriptor values are also displayed. The SHAP value plots are labelled with the PDB IDs of the receptors and the three letter codes of the ligands found in each pocket. A) The binding site of linezolid in the 50S ribosomal subunit. B) Ribocil bound to the FMN riboswitch. C) TAR RNA complexed with acetylpromazine. D) Guanine bound to the guanine riboswitch. E) Lysine in the binding site of the lysine riboswitch. F) Splicing site complexed with a splicing site modifier. G) Paromomycin bound to a bacterial ribosome site.

The FMN riboswitch has been validated as a target for the antibiotic compound ribocil, a drug-like small molecule (QED score = 0.71, Figure 1).[13] The affinity for ribocil (K_D = 13 nM) is driven by hydrogen bonding with the base of A99 and the ribose group of A48 as well as stacking interactions with A85, A49 and, G62 (Figure 3B).[42] The binding site was rather deep (*fr_buried_sl_atoms* =

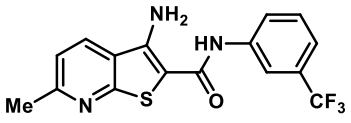
0.35) and characterized by a low relative polar surface area (*psa_r* = 0.33), a large the fraction of hydrophobic atoms (*fr_hpb_atoms* = 0.74), a rather large size of the hydrophobic contact surface area (*hsa* = 730 Å^2) and a large third principal moment of inertia (*PMI3* = 1.06 x 10^5). These values drove the site to be predicted as druggable despite its spherocity index lying in the less druggable range (*SpherocityIndex* = 0.41). The prediction agrees with the site binding drug-like ligands like ribocil with high ligand efficiency (LE = 0.41 kcal·mol^{-1}·heavy atom^{-1}).

A known ligand for the HIV-1 trans activating region (TAR) RNA is the drug acetylpromazine (QED = 0.85, Figure 4). Developed for a different target, the compounds binds only with moderate affinity and efficiency to TAR RNA (K_D = 270 µM, LE = 0.22 kcal·mol^{-1}·heavy atom^{-1}).[43] In the structure of the complex, the ligand forms stacking interactions with U25 and U40 (Figure 3C). DrugPred_RNA predicted the ligand binding site to be druggable. As with the examples above, the classification was driven by a large fraction of hydrophobic atoms (*fr_hpb_atoms* = 0.78), the depth of the pocket (*fr_buried_sl_atoms* = 0.41), the high ratio of the superligand atoms to binding site atoms (*sl_bs_r* = 1.4) and the large third moment of inertia (*PMI3* = 4.65 x 10^4). These properties overcame the high solvent accessibility *(exp_sl_sa* = 506 Å^2), and the relative high polarity of the binding site (*psa_r* = 0.39). More potent ligands for HIV TAR RNA are also known albeit structural information about their binding modes is lacking. Examples are a drug-like screening hit (QED = 0.72) and furimidazoline (QED = 0.72) which have affinities of 230 nM and

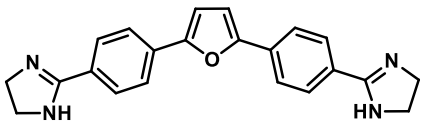
1 μM, resp. translating to LEs of 0.33 and 0.31 kcal·mol^{-1}·heavy atom^{-1} (Figure 4).[44,45] Assuming that these ligands bind into the same pocket as acetylpromazine, the prediction that this pocket is druggable appears to be reasonable.



Acetylpromazine
QED = 0.85
K_D = 270 μM
LE = 0.22 kcal · mol^{-1} · heavy atom^{-1}



Screening hit
QED = 0.72
K_D = 230 nM
LE = 0.33 kcal · mol^{-1} · heavy atom^{-1}



Furimidazoline
QED = 0.72
IC_{50} = 1 μM
LE = 0.33 kcal · mol^{-1} · heavy atom^{-1}

Figure 4. Ligands of HIV-1 TAR RNA.

Ligands binding to the guanine and lysine riboswitch have been shown to act as antibiotics.[46,47] In both cases, the pockets are rather small and almost fully enclose the natural ligands (Figure 3D and E). Structure-activity relationships (SAR) are very tight and only small modifications of the ligands are possible without losing binding affinity. DrugPred_RNA predicted these pockets to be less druggable which agrees with the SAR data. The predictions of the pockets were driven by their low relative polar surface areas (*psa_r* = 0.16 and 0.39, resp.), their lack of a sufficiently large hydrophobic surface area (*hsa* =109 Å^2 and 98 Å^2, resp.), small third principal moments of inertia (PMI3 = 357 and 859, resp.), their shallowness (*fr_buried_sl_atoms* = 0.0 in both cases), and their small size (*no_bs_atoms* = 37 and 59, *no_sl_atoms* = 8, 13, resp.).

Splicing modifiers for the treatment of spinal muscular atrophy are currently in clinical trials.[9,10] In our data set, the ligand SMN-C5 was included (Figure 3F). This ligand is moderate drug-like (QED = 0.55) and has a binding affinity of 28 μM translating to a LE of 0.22 kcal·mol^{-1}·heavy atom^{-1} for its target RNA. In the NMR structure, the flat ligand is lying in a highly solvent exposed binding site. DrugPred_RNA predicted this binding site to be less druggable. The prediction was

due to the pocket being polar (*psa_r* = 0.46, *hsa* = 120 Å^2), shallow (*fr_buried_sl_atoms* = 0.11), and having an undesirable shape (*sl_bs_r* = 0.75, *PMI3* = 8.38 x 10^3). The druggability prediction appears to be reasonable considering the binding mode of the ligand, but not the fact that splicing modifiers are currently in clinical trials. This discrepancy is probably caused by the compounds binding *in vivo* to a ribonucleoprotein-RNA complex with a still unknown structure.[48] Thus, the biological relevant pocket of this type of compounds was not included in our study.

One class of FDA-approved ribosome binding antibiotics are aminoglycosides. One example of an aminoglycoside is paromomycin which acts by binding to the 16S ribosomal RNA (Figure 3G). Its low QED score of 0.11 is in agreement with the poor bioavailability of this compound class and the fact that aminoglycosides get into the bacteria by active transport.[49] In the complex of paromomycin bound to the ribosome of *T. thermophilus*, the ligand forms several hydrogen bonds with surrounding binding site residues and water molecules (not shown), with little hydrophobic interactions. The terminal sugar ring in this ligand is located outside of the superligand created by DrugPred_RNA, suggesting that this area is a less optimal for ligand binding. The SHAP values suggested that despite the depth of the pocket (*fr_buried_sl_atoms* = 0.4) and the fraction of hydrophobic atoms (*fr_hpb_atoms* = 0.73) being in a range beneficial for druggable sites, the large polar surface area (*psa_r* = 0.51), the solvent-exposure (*exp_sl_sa* = 354 Å^2) combined with a less ideal shape (*InertialShapeFactor* = 1.10x10^{-4}, *sl_bs_r* = 1.2) contributed to the pocket being predicted as less druggable. This prediction agrees with the nature of the known ligands.

Overall, the results for the selected anecdotal examples looked very promising. As expected, pockets predicted to be druggable were generally larger and more hydrophobic while the less druggable sites among the selected examples were more polar and solvent exposed. The

predictions of DrugPred_RNA generally agreed with what one would await based on the nature of the pockets and the bound ligands.

Druggability predictions of RNA pockets binding to drug-like ligands

In the next step, we investigated if the drug-like ligands contained in our RNA test sets bound to pockets predicted to be druggable. In total, the sets contained 331 unique ligands which 22 of them having a QED score ≥ 0.67. Four of these ligands were found in the binding site of the preQ1 riboswitch. Upon closer inspection of these pockets it became evident that some of the bases in these structures were not resolved. These pockets were therefore not further considered. Out of the remaining ligands, 12 (67%) were found in binding sites assessed by DrugPred_RNA as druggable (Table 3) and 6 (23%) in binding sites assessed to be less druggable (Table 4). As only 16 % of all metal-free binding sites were predicted to be druggable, the drug-like ligands were clearly enriched in druggable binding sites.

Table 3. Drug-like ligands (QED ≥ 0.67) found in RNA binding sites predicted to be druggable.

Ligand ID	PDB ID	Receptor name	QED score	K_d [nM]	LE [kcal·mol^{-1}·heavy atom^{-1}]
RNA data set					
MGR	1q8n	Malachite green aptamer	0.76	800[50]	0.34
6YG	5kx9	FMN riboswitch	0.69	13.4[42]	0.41
L8H	2l8h	HIV-1 TAR RNA	0.67	NA#[51]	-
PMZ	1lvj	HIV-1 TAR RNA	0.85	27,000[44]	0.22
Ribosomal data set					
917	5v7q	50S ribosomal subunit	0.94	700[52]	0.39
ZLD	3cpw	50S ribosomal subunit	0.89	20,000[53]	0.27
G6M	6ddg	50S ribosomal subunit	0.79	2,600[54,55]	0.31
3HE	4u3u	80S ribosome	0.76	140[56]	0.48
G6V	6ddd	50S ribosomal subunit	0.76	2,600[54]	0.30
ANM	3cc4	50s ribosomal subunit	0.78	20,000[57]	0.34
HN8	5on6	80S ribosome	0.71	NA#	-
3K8	4u55	80S ribosome	0.71	39	0.32

binding affinity unknown

Table 4. Drug-like ligands (QED ≥ 0.67) found in RNA binding sites predicted to be less druggable.

Ligand ID	PDB code	Receptor name	QED	K_D [nM]	LE [kcal·mol^{-1}·heavy atom^{-1}]
RNA data set					
VIB	4nyg	TPP riboswitch	0.79	1,500[58]	0.45
2QC	4nyb	TPP riboswitch	0.77	103,000[58]	0.43
0EC	2lwk	Influenza A	0.86	50,000[59]	0.29
1TU	5ob3	Spinach aptamer	0.85	530[60]	0.49
218	2hop	TPP riboswitch	0.77	6,000[61]	0.38
Ribosomal data set					
TRP	4v6o	Tryptophan-sensing ribosomal site	0.67	NA#	-

binding affinity unknown

For 10 out of the 12 drug-like ligands binding to pockets predicted to be druggable, we could find binding data in the literature (Table 3). Based on this data, 8 ligands bind efficiently to their target with LEs > 0.30 kcal·mol^{-1}·heavy atom^{-1} hinting that these pockets are indeed druggable. The two remaining ligands were linezolid with the 50S ribosomal subunit as target and acetylpromazine binding to HIV-1 TAR RNA. For the reasons discussed above, these pockets also appear to be druggable. Thus, all druggability predictions for the pockets binding the 10 drug-like ligands with known binding data are sensible.

On the other hand, 6 drug-like ligands were found in pockets predicted to be less druggable (Table 4). For 5 of them we could retrieve affinity data in the literature and all of these bind rather efficiently to their targets (LE ≥ 0.29 kcal·mol^{-1}·heavy atom^{-1}). Three of these ligands are

fragments binding to the TPP riboswitch, one is a ligand binding the influenza A virus promoter region, and one a ligand of the Spinach aptamer. Several examples of the TPP riboswitch binding site were contained in the RNA set (Figure 5). The pockets differ mainly in the conformation of G72 (Figure 5c) but in all cases the pocket is rather large and partially buried (Figure 5a and b). The pockets with G72 in one of the conformations were predicted to be druggable while pockets with G72 in the alternative conformation were predicted to be less druggable, discussed in more detail below. Based on the structures, it is not obvious why the TPP riboswitch binding site conformation binding efficiently the drug-like fragments should be less druggable. This prediction can therefore be considered as false negative. The drug-like ligand of the influenza A promoter region sits on the surface of the RNA molecule and is almost entirely solvent exposed (Figure 6). It is highly unusual that a ligand with such a binding mode binds that efficiently (LE = 0.29 kcal·mol^{-1}·heavy atom^{-1}). However, the structure of the complex has been determined by NMR and it is possible that the resolution of the structure is not accurate enough to reveal the details of the binding mode.[59] The Spinach aptamer binds a small molecule dye, DFHBI, which forms hydrogen bonding and pi-stacking interactions in the binding site (Figure 7). The top six SHAP values for this entry showed that while the fraction of hydrophobic atoms (*fr_hpb_atoms* = 0.8) and the hydrophobic surface area value (*hsa* = 409 Å^2) were in a range that is favorable for druggable binding sites, the shallow shape of the pocket (*fr_buried_sl_atoms* = 0.17, *sl_bs_r* = 0.79), combined with the solvent exposure (*exp_sl_sa* = 376 Å^2) and the high relative polar surface area (*psa_r* = 0.38) drove the site to be predicted as less druggable. Considering the drug-likeness of the ligand together with its efficient binding, this prediction can be considered to be a misclassification by DrugPred_RNA.

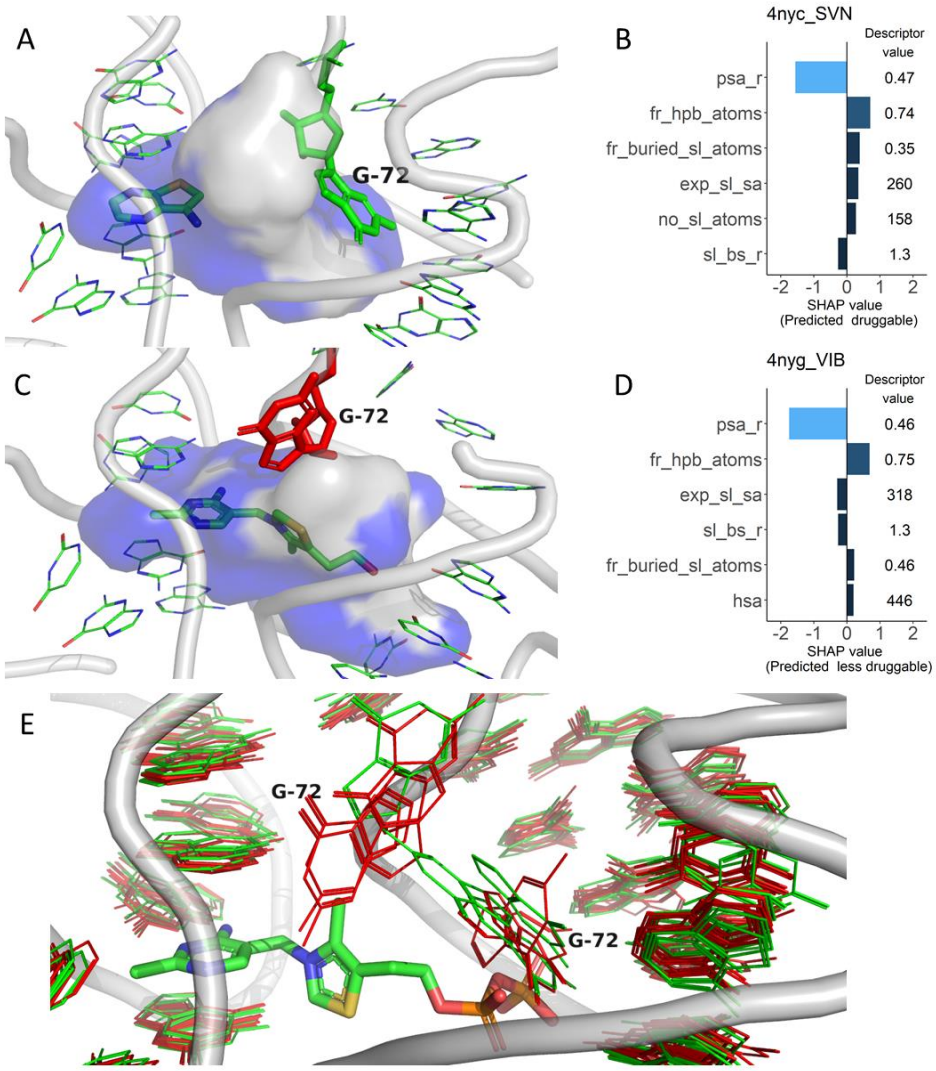


Figure 5. Druggability predictions for TPP riboswitch binding sites, with the flexible residue G72 highlighted. The surface of the superligand created by DrugPred_RNA as a negative print of the pocket is shown as a blob with the solvent exposed surface area coloured grey and the remaining area blue. For the pockets shown in A and B, the individual SHAP values for the six most important descriptors are shown together with their descriptor values. The SHAP plots are labelled with the PDB IDs of the receptors and three letter codes of the ligands found in each pocket (B, D). A) TPP riboswitch binding site (PDB ID 4nyc) in complex with a fragment screening hit (green sticks). C) TPP riboswitch binding site (PDB ID 4nyg) in complex with thiamine. E) Superposition of all *E.*

coli TPP riboswitch binding sites the RNA set. Entries predicted to be druggable are coloured green and to be less druggable red. For clarity, only the backbone (grey tube) from PDB entry 4nyc is shown. The conformation of the residue G72 influences the prediction.

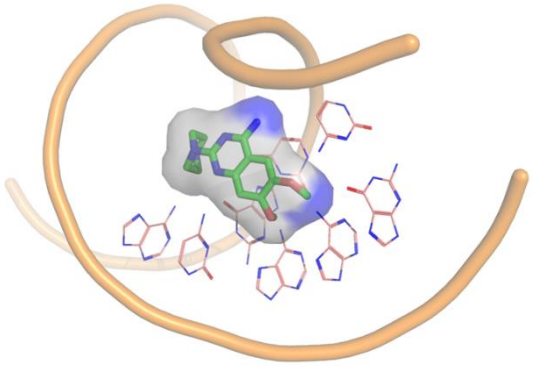


Figure 6. Binder of influenza A promoter region (PDB ID 2lwk). The surface of the superligand created by DrugPred_RNA as a negative print of the pocket is shown as a blob with the solvent exposed surface area coloured grey and the remaining surface area coloured blue.

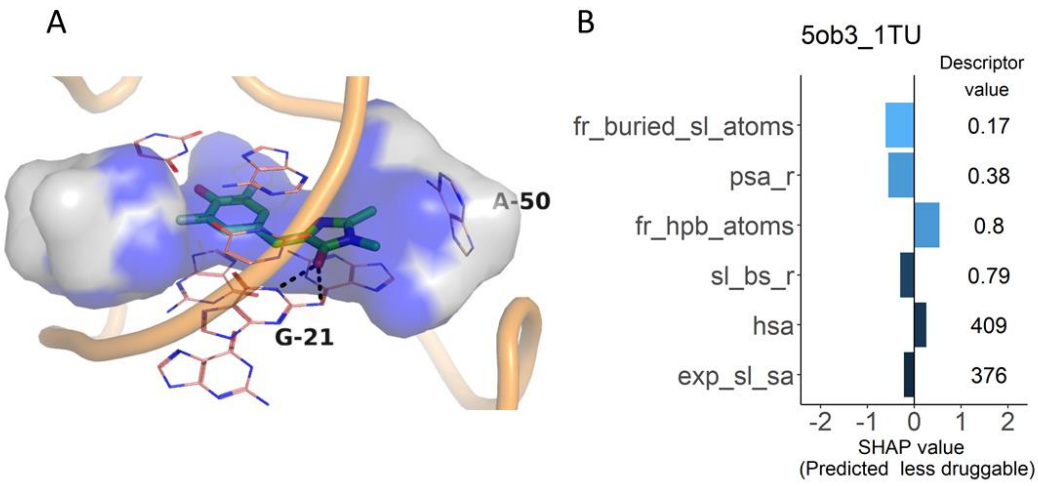


Figure 7. A) The Spinach aptamer (PDB ID 5ob3, orange thick and thin lines) bound to the dye DFHBI (green sticks). The surface of the superligand created by DrugPred_RNA as a negative print of the pocket is shown as a blob with the solvent exposed surface area coloured grey and the remaining surface area coloured blue. B) Individual SHAP values for the six most important descriptors together with the descriptor values obtained by DrugPred_RNA.

Taken together, the druggability predictions for the TPP riboswitch pockets binding the fragments and the Spinach aptamer were likely false negatives, while the predictions for the influence A promotor region and all predictions for the pockets predicted to be druggable appeared to be correct. This could confirm that DrugPred_RNA has a larger tendency to misclassify druggable binding sites as less druggable than vice versa, as already observed for the NRDLD test set (precision = 0.95 for druggable pockets vs. 0.86 for less druggable pockets, Table 2). However, the investigated data set was too small to conclude firmly on this.

Assessment of the robustness of the druggability predictions

Finally, we assessed the robustness of the predictions with respect to small changes of the conformation or base composition of the binding sites. For that purpose, all binding sites with a sequence similarity >85% were grouped together resulting in 46 families in the RNA binding site set and 52 in the ribosome set in addition to 674 singletons. The families spanned between 2 and 23 members in the RNA dataset (Table S2) and 2 and 56 in the ribosomal set (Table S3). Subsequently, the consensus of the predictions for each family was calculated. The consensus was defined as (100*|#druggable binding sites - #less-druggable binding sites|)/(total number of predictions).[22] Thus, 100% consensus would be obtained if all pockets in one family were predicted to belong to the same class (druggable or less druggable) and 0% if one half of the pockets was predicted to belong to one class and the other half to the other class. In the RNA set, for 34 of the 46 families (74%) a consensus of 100% was obtained and in the ribosome set for 39 of the 52 families (75%). Thus, in most cases using different crystal structures of the same or a related pocket did not change the outcome of the prediction.

In the RNA and ribosome sets, 6 families each had a consensus score of 33% or less (Table S2 and Table S3). Three of them, the families containing the TPP and ZTP riboswitches as well as the neomycin binding site of bacterial ribosome, are discussed below. The other families were not further considered as either their structures lacked side chains (preQ1-I riboswitch family), contained mainly less accurate NMR structures (HCV IRES family), were less interesting from a drug discovery point of few (Mango and Corn aptamer), were misclassified in the wrong set (a synthetic rRNA construct contained in the ribosome set) or contained only two members (the remaining low consensus pockets in the ribosome set).

The TPP riboswitch family which contained pockets from 16 distinct PDB entries obtained a low consensus score of 12.5% with the majority of the pockets predicted as less druggable (Table S2). Superimposing the pockets, it became evident that there is some plasticity in the binding sites (Figure 5c). One guanine residue (G72 in the *E. coli* TPP riboswitch) can adopt several conformations depending on the bound ligand leading to considerably different superligands (Figure 5a and b). Consequently, the pockets differ in compactness (*fr_buried_sl_atoms*, *sl_bs_r*) and solvent exposure (*exp_sl_sa*) leading to different prediction outcomes. However, based on the structures and the affinity of the bound ligands, both binding sites appear to be druggable.

Another family with a low consensus is the ZTP riboswitch (33.3%) with the majority of the pockets predicted to be less druggable. The three entries in the family are all bound to the same ligand, ZMP (aminoimidazole 4-carboxamide ribonucleotide), which is poorly drug-like (QED = 0.39). Superposition of the druggable pocket with the less druggable pockets revealed that one of the less druggable pockets has a clearly different conformation of the residue A60 resulting in very different superligands for the druggable and one of the less druggable pockets and thus different predictions (Figure 8A, C, and D). The second less druggable pocket has nearly the same conformation as the druggable pocket (Figure 8B). In this case, settle conformational changes were enough to obtain a slightly different superligand which in turn resulted in a switch of the prediction despite the descriptors with top six highest SHAP values being almost identical (Figure 8D and E).

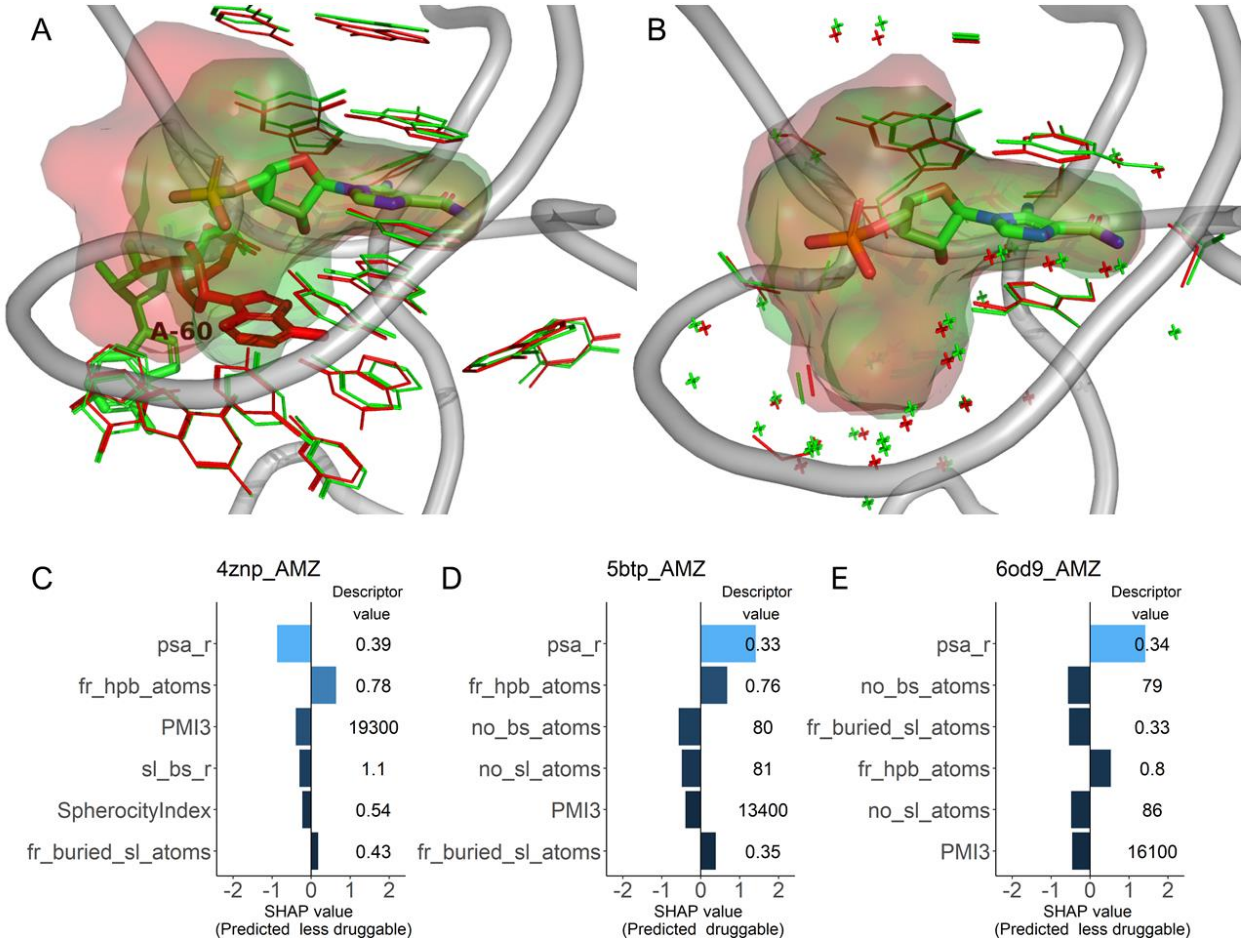


Figure 8. Superposition of the ZNP riboswitch binding site bound to ZNP (thick sticks with green carbon atoms). The superligands created by DrugPred_RNA are shown as blobs. For clarity only the backbone from 5btp is shown. A) Superposition of the pockets of the structures with the PDB IDs 4znp (red, less druggable) and 5btp (green, druggable). The entire residues forming the binding sites are shown. B) Superposition of the pockets of the structures with the PDB IDs 5btp (green, druggable) and 6od9 (red, less druggable). For clarity, only the atoms that DrugPred_RNA predicted to be in contact with the superligand are shown (thin sticks/crosses). C, D, E) Individual SHAP values for the six most important descriptors for the displayed binding sites together with the descriptor values.

The family containing the neomycin binding site of bacterial ribosome obtained a consensus score of 0%. The two druggable entries in this family were bound to neomycin (PDB IDs 4v52 and 4v57), while the two less druggable entries were bound to paromomycin (4woi) and gentamicin. (4v55). Compared to the neomycin-containing structures, A1913 is rotated in 4woi leading to a very different shape and size of the pocket with a different prediction outcome (Figure 9A, C and D). The structural differences between the pocket in 4v55 and the druggable sites are less pronounced but nevertheless sufficient to make the pocket more polar and thus less druggable (Figure 9B, E).

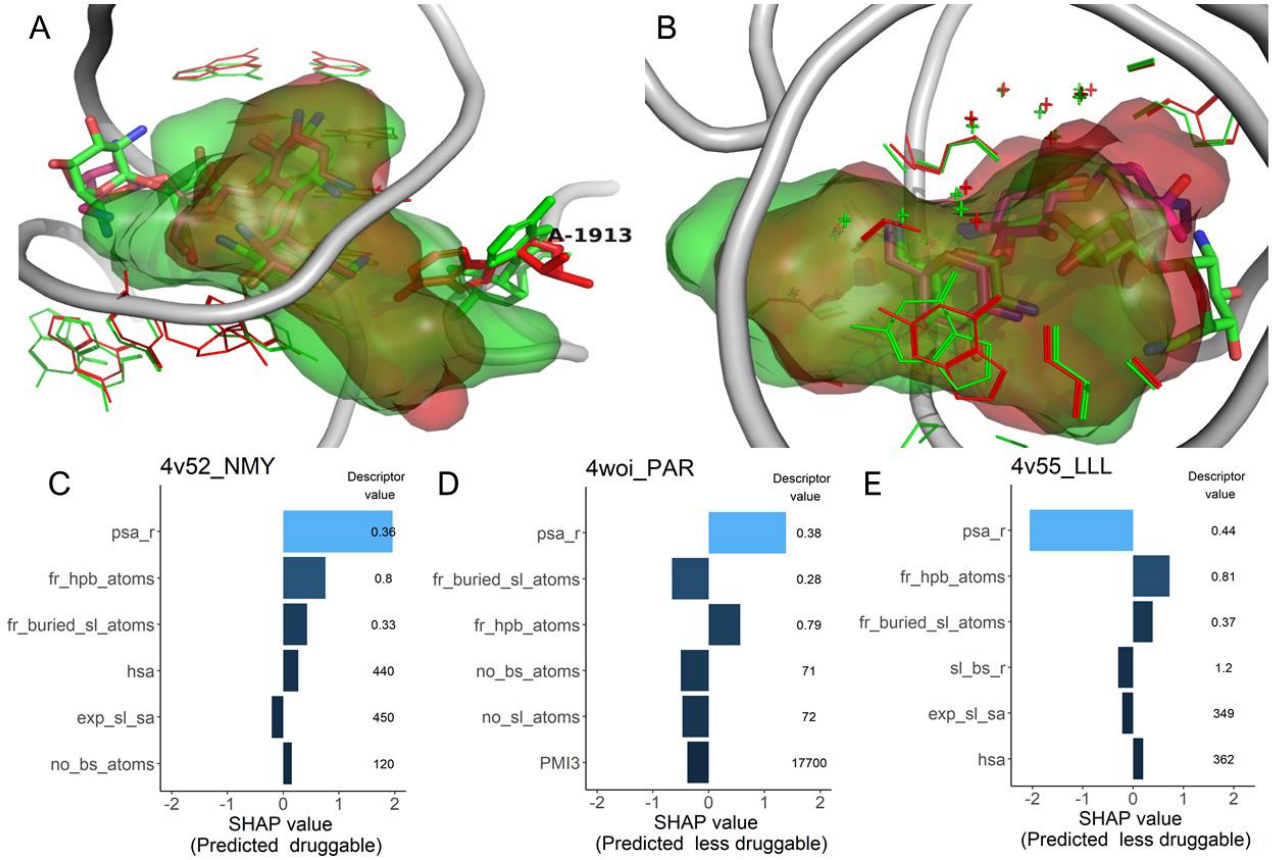


Figure 9. A) Superposition of the neomycin- (PDB ID 4v52, green) and paromomycin- (PDB ID 4woi, magenta and red) containing ribosomal binding sites. The backbone (taken from PDB ID 4v52) is shown as thick grey tube and the superligands created by DrugPred_RNA as blobs (green: 4v52, red: 4woi). A1913 is highlighted with thick lines. B) Superposition of the neomycin (green, thick sticks) and gentamicin (magenta, thick sticks)-containing binding sites (PDB IDs 4v52, 4v55), showing only atoms (thin lines, crosses) in direct contact with the superligands (green blob, 4v52, red blob 4v55) C, D, E) Individual SHAP values for the six most important descriptors together with the descriptor values. The label denotes the PDB ID of the structure followed by the three-letter code of the ligand.

In summary, while mostly using pockets arising from related structures led to the same prediction outcome, there were also examples as discussed above where this was not the case. In some of the illustrated examples a conformational change of a residue in the binding site led to a clearly differently shaped pocket and it was easily comprehensible how this could influence the predictions. In other examples the conformational changes that led to altered predictions were more settle. Thus, it appears to be advisable to score more than one example of a binding site if available to obtain reliable results.

CONCLUSION

RNA is an emerging target for drug discovery.[3–6] However, like for proteins, not all RNA binding sites are equally suited to be addressed with conventional drug-like ligands. We have developed the structure-based druggability predictor DrugPred_RNA to identify pockets that are primed to potently bind such ligands. Due to the paucity of annotated RNA binding sites, the predictor was trained on a set of protein pockets, albeit containing only descriptors that can be calculated for both, RNA and protein binding sites. DrugPred_RNA performed comparable on the protein binding site set as our previous DrugPred 2.0 predictor trained with slightly different descriptors (Table 2). In addition, druggability predictions of DrugPred_RNA on selected anecdotal examples appeared to be sensible (Figure 3). Likewise, the majority of the drug-like ligands contained in our RNA binding site sets were found in pockets predicted to be druggable, further adding confidence to the DrugPred_RNA predictions (Table 3 and Table 4). As observed before,[21,62] using different conformations of a binding site could result in opposing druggability predictions (Table S2 and Table S3). However, for the majority of cases consistent predictions were obtained indicating that DrugPred_RNA is generally robust towards small changes in binding site conformations.

Interestingly, many riboswitches were found among the binding site families that were predicted to be druggable (Table S2). This finding underlines the notation that these promising targets for new antibiotics could be addressed with drug-like ligands.[3,12,20] Further, also in the ribosomal binding site set druggable pockets were contained (Table S3). These predictions can help to direct efforts when targeting the ribosome for the development of drugs to overcome the looming antibiotic crisis.[7,55]

Notably, as DrugPred_RNA was trained with descriptors that can be calculated for both, RNA and protein binding sites, it can also be used to score pockets that are formed by both types of macromolecules. An example is a pocket in the protozoal 80S ribosomal site which highly efficiently (LE = 0.41 kcal·mol^{-1}·heavy atom^{-1}) binds to the drug-like molecule mefloquine (QED = 0.79) and was predicted to be druggable (Figure 10).[63]

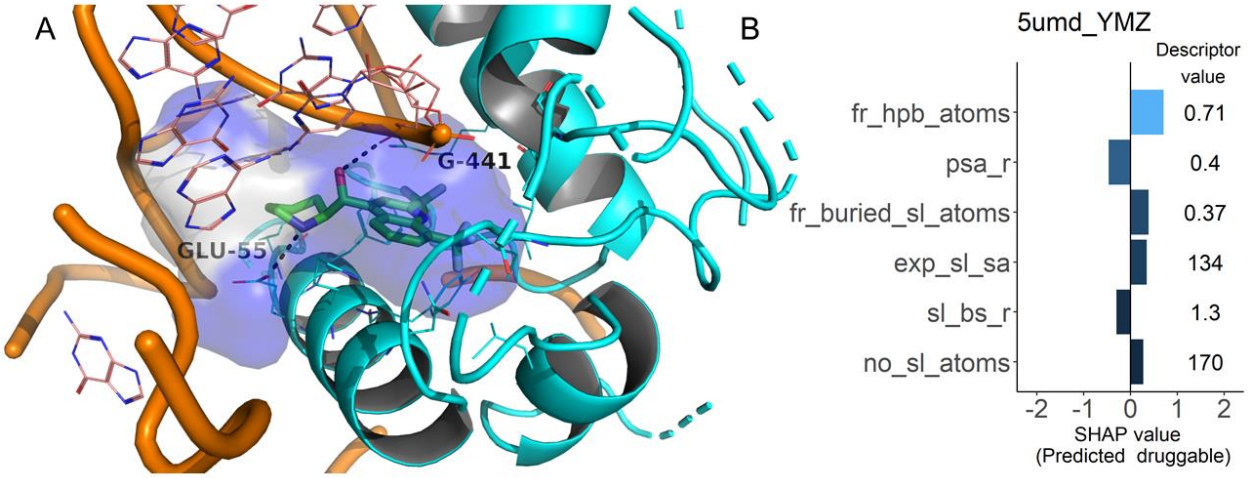


Figure 10. A) Ribosomal binding site of mefloquine which is formed by amino acids (cyan) and bases (orange, PDB ID 5umd). The ligand mefloquine is shown as green sticks, while the surface of the superligand created by DrugPred_RNA as a negative print of the pocket is shown as a blob with the solvent exposed surface area coloured grey and the remaining surface area coloured blue. B) Individual SHAP values for the six most important descriptors together with the descriptor values obtained by DrugPred_RNA

To conclude, DrugPred_RNA is a promising tool for structure-based druggability predictions of RNA binding sites that can be used to prioritize targets and to decide if a target can be addressed with drug-like ligands are another area of chemical space has to be searched for potent ligands.

ACKNOWLEDGMENT

We thank Sturla M. Grøndal for help with R. Docking and descriptor calculations were performed on resources provided by UNINETT Sigma2—the National Infrastructure for High Performance Computing and Data Storage in Norway.

SUPPORTING INFORMATION

- A file with tables with more information about the descriptors (Table S1), and overviews of the binding site families and consensus scoring results (Table S**2** and Table S3) as well as figures displaying the training and testing set accuracy error during the construction of DrugPred_RNA (Figure S1), druggability predictions with DrugPred_RNA for the NRDLD training and test set (Figure S2), and individual SHAP values for each pocket in the training set for all descriptors in the model plotted against the descriptor values (Figure S3)

- A file with three letter codes of ligands that were treated as buffer components.

- A file with metal abbreviations

- Two files with a list of commonly modified RNA residues

- A file with druggability predictions for all pockets in the ribosome and RNA set

Scripts to predict the druggability of binding sites with DrugPred_RNA and instructions on how to use them can be found on https://github.com/ruthbrenk/DrugPred_RNA.

REFERENCES

(1) Imming, P.; Sinning, C.; Meyer, A. Drugs, Their Targets and the Nature and Number of Drug Targets. *Nat. Rev. Drug Discov.* **2006**, *5* (10), 821–834. https://doi.org/10.1038/nrd2132.

(2) Santos, R.; Ursu, O.; Gaulton, A.; Bento, A. P.; Donadi, R. S.; Bologa, C. G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T. I.; Overington, J. P. A Comprehensive Map of Molecular Drug Targets. *Nat. Rev. Drug Discov.* **2017**, *16* (1), 19–34. https://doi.org/10.1038/nrd.2016.230.

(3) Warner, K. D.; Hajdin, C. E.; Weeks, K. M. Principles for Targeting RNA with Drug-like Small Molecules. *Nat. Rev. Drug Discov.* **2018**. https://doi.org/10.1038/nrd.2018.93.

(4) Rizvi, N. F.; Smith, G. F. RNA as a Small Molecule Druggable Target. *Bioorg. Med. Chem. Lett.* **2017**. https://doi.org/10.1016/j.bmcl.2017.10.052.

(5) Matsui, M.; Corey, D. R. Non-Coding RNAs as Drug Targets. *Nat. Rev. Drug Discov.* **2017**, *16* (3), 167–179. https://doi.org/10.1038/nrd.2016.117.

(6) Ursu, A.; Childs-Disney, J. L.; Andrews, R. J.; O'Leary, C. A.; Meyer, S. M.; Angelbello, A. J.; Moss, W. N.; Disney, M. D. Design of Small Molecules Targeting RNA Structure from Sequence. *Chem. Soc. Rev.* **2020**, *49* (20), 7252–7270. https://doi.org/10.1039/D0CS00455C.

(7) Wilson, D. N. Ribosome-Targeting Antibiotics and Mechanisms of Bacterial Resistance. *Nat. Rev. Microbiol.* **2013**, *12* (1), 35–48. https://doi.org/10.1038/nrmicro3155.

(8) Ippolito, J. A.; Kanyo, Z. F.; Wang, D.; Franceschi, F. J.; Moore, P. B.; Steitz, T. A.; Duffy, E. M. Crystal Structure of the Oxazolidinone Antibiotic Linezolid Bound to the 50S Ribosomal Subunit. *J. Med. Chem.* **2008**, *51* (12), 3353–3356. https://doi.org/10.1021/jm800379d.

(9) Calder, A. N.; Androphy, E. J.; Hodgetts, K. J. Small Molecules in Development for the Treatment of Spinal Muscular Atrophy. *J. Med. Chem.* **2016**, *59* (22), 10067–10083. https://doi.org/10.1021/acs.jmedchem.6b00670.

(10) Ratni, H.; Karp, G. M.; Weetall, M.; Naryshkin, N. A.; Paushkin, S. V.; Chen, K. S.; McCarthy, K. D.; Qi, H.; Turpoff, A.; Woll, M. G.; Zhang, X.; Zhang, N.; Yang, T.; Dakka, A.; Vazirani, P.; Zhao, X.; Pinard, E.; Green, L.; David-Pierson, P.; Tuerck, D.; Poirier, A.; Muster, W.; Kirchner, S.; Mueller, L.; Gerlach, I.; Metzger, F. Specific Correction of Alternative Survival Motor Neuron 2 Splicing by Small Molecules: Discovery of a Potential Novel Medicine To Treat Spinal Muscular Atrophy. *J. Med. Chem.* **2016**, *59* (13), 6086–6100. https://doi.org/10.1021/acs.jmedchem.6b00459.

(11) Rekand, I.; Brenk, R. Design of Riboswitch Ligands, an Emerging Target Class for Novel Antibiotics. *Future Med Chem* **2017**, *9* (14), 1649–1662. https://doi.org/10.4155/fmc-2017-0063.

(12) Panchal, V.; Brenk, R. Riboswitches as Drug Targets for Antibiotics. *Antibiotics* **2021**, *10* (1), 45. https://doi.org/10.3390/antibiotics10010045.

(13) Howe, J. A.; Wang, H.; Fischmann, T. O.; Balibar, C. J.; Xiao, L.; Galgoci, A. M.; Malinverni, J. C.; Mayhood, T.; Villafania, A.; Nahvi, A.; Murgolo, N.; Barbieri, C. M.; Mann, P. A.; Carr, D.; Xia, E.; Zuck, P.; Riley, D.; Painter, R. E.; Walker, S. S.; Sherborne, B.; de Jesus, R.; Pan, W.; Plotkin, M. A.; Wu, J.; Rindgen, D.; Cummings, J.; Garlisi, C. G.; Zhang, R.; Sheth, P.

R.; Gill, C. J.; Tang, H.; Roemer, T. Selective Small-Molecule Inhibition of an RNA Structural Element. *Nature* **2015**, *526* (7575), 672–677. https://doi.org/10.1038/nature15542.

(14) Blount, K. F.; Megyola, C.; Plummer, M.; Osterman, D.; O'Connell, T.; Aristoff, P.; Quinn, C.; Chrusciel, R. A.; Poel, T. J.; Schostarez, H. J.; Stewart, C. A.; Walker, D. P.; Wuts, P. G. M.; Breaker, R. R. Novel Riboswitch-Binding Flavin Analog That Protects Mice against *Clostridium Difficile* Infection without Inhibiting Cecal Flora. *Antimicrob. Agents Chemother.* **2015**, *59* (9), 5736–46. https://doi.org/10.1128/AAC.01282-15.

(15) Morgan, B. S.; Forte, J. E.; Culver, R. N.; Zhang, Y.; Hargrove, A. E. Discovery of Key Physicochemical, Structural, and Spatial Properties of RNA-Targeted Bioactive Ligands. *Angew. Chem. Int. Ed Engl.* **2017**, *56* (43), 13498–13502. https://doi.org/10.1002/anie.201707641.

(16) Edfeldt, F. N.; Folmer, R. H.; Breeze, A. L. Fragment Screening to Predict Druggability (Ligandability) and Lead Discovery Success. *Drug Discov. Today* **2011**. https://doi.org/10.1016/j.drudis.2011.02.002.

(17) Abi Hussein, H.; Geneix, C.; Petitjean, M.; Borrel, A.; Flatters, D.; Camproux, A.-C. Global Vision of Druggability Issues: Applications and Perspectives. *Drug Discov. Today* **2017**, *22* (2), 404–415. https://doi.org/10.1016/j.drudis.2016.11.021.

(18) Sheridan, R. P.; Maiorov, V. N.; Holloway, M. K.; Cornell, W. D.; Gao, Y. D. Drug-like Density: A Method of Quantifying the "Bindability" of a Protein Target Based on a Very Large Set of Pockets and Drug-like Ligands from the Protein Data Bank. *J. Chem. Inf. Model.* **2010**, *50* (11), 2029–2040. https://doi.org/10.1021/ci100312t.

(19) An, J.; Totrov, M.; Abagyan, R. Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes. *Mol. Cell. Proteomics* **2005**, *4* (6), 752–761. https://doi.org/10.1074/mcp.M400159-MCP200.

(20) Hewitt, W. M.; Calabrese, D. R.; Schneekloth, J. S. Evidence for Ligandable Sites in Structured RNA throughout the Protein Data Bank. *Bioorg. Med. Chem.* **2019**, *27* (11), 2253–2260. https://doi.org/10.1016/j.bmc.2019.04.010.

(21) Krasowski, A.; Muthas, D.; Sarkar, A.; Schmitt, S.; Brenk, R. DrugPred: A Structure-Based Approach to Predict Protein Druggability Developed Using an Extensive Nonredundant Data Set. *J. Chem. Inf. Model.* **2011**, *51* (11), 2829–2842. https://doi.org/10.1021/ci200266d.

(22) Sarkar, A.; Brenk, R. To Hit or Not to Hit, That Is the Question - Genome-Wide Structure-Based Druggability Predictions for Pseudomonas Aeruginosa Proteins. *PloS One* **2015**, *10* (9), e0137279. https://doi.org/10.1371/journal.pone.0137279.

(23) Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability Prediction of Protein-Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* **2012**, *52* (8), 2287–2299. https://doi.org/10.1021/ci300184x.

(24) Volkamer, A.; Rarey, M. Exploiting Structural Information for Drug-Target Assessment. *Future Med. Chem.* **2014**, *6* (3), 319–331. https://doi.org/10.4155/fmc.14.3.

(25) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

(26) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-Wide Collection of Binding Data: Current Status of the PDBbind Database. *Bioinformatics* **2015**, *31* (3), 405–412. https://doi.org/10.1093/bioinformatics/btu626.

(27) Kumar Mishra, S.; Kumar, A. NALDB: Nucleic Acid Ligand Database for Small Molecules Targeting Nucleic Acid. *Database J. Biol. Databases Curation* **2016**, *2016*. https://doi.org/10.1093/database/baw002.

(28) Mehta, A.; Sonam, S.; Gouri, I.; Loharch, S.; Sharma, D. K.; Parkesh, R. SMMRNA: A Database of Small Molecule Modulators of RNA. *Nucleic Acids Res.* **2014**, *42* (Database issue), D132-141. https://doi.org/10.1093/nar/gkt976.

(29) Morgan, B. S.; Sanaba, B. G.; Donlic, A.; Karloff, D. B.; Forte, J. E.; Zhang, Y.; Hargrove, A. E. R-BIND: An Interactive Database for Exploring and Developing RNA-Targeted Chemical Probes. *ACS Chem. Biol.* **2019**, *14* (12), 2691–2700. https://doi.org/10.1021/acschembio.9b00631.

(30) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.

(31) Landrum, G. RDKit: Open-source cheminformatics http://www.rdkit.org.

(32) Lorber, D. M.; Shoichet, B. K. Flexible Ligand Docking Using Conformational Ensembles. *Protein Sci.* **1998**, *7* (4), 938–50. https://doi.org/10.1002/pro.5560070411.

(33) Mitternacht, S. FreeSASA: An Open Source C Library for Solvent Accessible Surface Area Calculations. *F1000Research* **2016**, *5*. https://doi.org/10.12688/f1000research.7931.1.

(34) Tsai, J.; Taylor, R.; Chothia, C.; Gerstein, M. The Packing Density in Proteins: Standard Radii and Volumes11Edited by J. M. Thornton. *J. Mol. Biol.* **1999**, *290* (1), 253–266. https://doi.org/10.1006/jmbi.1999.2829.

(35) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; KDD '16; ACM: New York, NY, USA, 2016; pp 785–794. https://doi.org/10.1145/2939672.2939785.

(36) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.

(37) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; pp 4765–4774.

(38) Liu, Y.; Just, A. *SHAPforxgboost: SHAP Plots for "XGBoost";* manual; 2019.

(39) Rodríguez-Pérez, R.; Bajorath, J. Interpretation of Machine Learning Models Using Shapley Values: Application to Compound Potency and Multi-Target Activity Predictions. *J. Comput. Aided Mol. Des.* **2020**, *34* (10), 1013–1026. https://doi.org/10.1007/s10822-020-00314-0.

(40) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4* (2), 90–98. https://doi.org/10.1038/nchem.1243.

(41) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand Efficiency: A Useful Metric for Lead Selection. *Drug Discov. Today* **2004**, *9* (10), 430–431. https://doi.org/10.1016/S1359-6446(04)03069-7.

(42) Howe, J. A.; Xiao, L.; Fischmann, T. O.; Wang, H.; Tang, H.; Villafania, A.; Zhang, R.; Barbieri, C. M.; Roemer, T. Atomic Resolution Mechanistic Studies of Ribocil: A Highly Selective Unnatural Ligand Mimic of the E. Coli FMN Riboswitch. *RNA Biol.* **2016**, *13* (10), 946–954. https://doi.org/10.1080/15476286.2016.1216304.

(43) Du Z; Lind KE; James Tl. Structure of TAR RNA Complexed with a Tat-TAR Interaction Nanomolar Inhibitor That Was Identified by Computational Screening. *Chem. Biol.* **2002**. https://doi.org/10.1016/s1074-5521(02)00151-5.

(44) Sztuba-Solinska, J.; Shenoy, S. R.; Gareiss, P.; Krumpe, L. R. H.; Le Grice, S. F. J.; O'Keefe, B. R.; Schneekloth, J. S. Identification of Biologically Active, HIV TAR RNA-Binding Small Molecules Using Small Molecule Microarrays. *J. Am. Chem. Soc.* **2014**, *136* (23), 8402–8410. https://doi.org/10.1021/ja502754f.

(45) Gelus, N.; Bailly, C.; Hamy, F.; Klimkait, T.; Wilson, W. D.; Boykin, D. W. Inhibition of HIV-1 Tat-TAR Interaction by Diphenylfuran Derivatives: Effects of the Terminal Basic Side Chains. *Bioorg. Med. Chem.* **1999**, *7* (6), 1089–1096. https://doi.org/10.1016/S0968-0896(99)00041-3.

(46) Blount, K. F.; Wang, J. X.; Lim, J.; Sudarsan, N.; Breaker, R. R. Antibacterial Lysine Analogs That Target Lysine Riboswitches. *Nat. Chem. Biol.* **2007**, *3* (1), 44–49. https://doi.org/10.1038/nchembio842.

(47) Yan, L.-H.; Le Roux, A.; Boyapelly, K.; Lamontagne, A.-M.; Archambault, M.-A.; Picard-Jean, F.; Lalonde-Seguin, D.; St-Pierre, E.; Najmanovich, R. J.; Fortier, L.-C.; Lafontaine, D.; Marsault, É. Purine Analogs Targeting the Guanine Riboswitch as Potential Antibiotics against Clostridioides Difficile. *Eur. J. Med. Chem.* **2018**, *143*, 755–768. https://doi.org/10.1016/j.ejmech.2017.11.079.

(48) Sivaramakrishnan, M.; McCarthy, K. D.; Campagne, S.; Huber, S.; Meier, S.; Augustin, A.; Heckel, T.; Meistermann, H.; Hug, M. N.; Birrer, P.; Moursy, A.; Khawaja, S.; Schmucki, R.; Berntenis, N.; Giroud, N.; Golling, S.; Tzouros, M.; Banfai, B.; Duran-Pacheco, G.; Lamerz, J.; Hsiu Liu, Y.; Luebbers, T.; Ratni, H.; Ebeling, M.; Cléry, A.; Paushkin, S.; Krainer, A. R.; Allain, F. H.-T.; Metzger, F. Binding to SMN2 Pre-MRNA-Protein Complex Elicits Specificity for Small Molecule Splicing Modifiers. *Nat. Commun.* **2017**, *8* (1), 1476. https://doi.org/10.1038/s41467-017-01559-4.

(49) Krause, K. M.; Serio, A. W.; Kane, T. R.; Connolly, L. E. Aminoglycosides: An Overview. *Cold Spring Harb. Perspect. Med.* **2016**, *6* (6). https://doi.org/10.1101/cshperspect.a027029.

(50) Flinders, J.; DeFina, S. C.; Brackett, D. M.; Baugh, C.; Wilson, C.; Dieckmann, T. Recognition of Planar and Nonplanar Ligands in the Malachite Green–RNA Aptamer Complex. *ChemBioChem* **2004**, *5* (1), 62–72. https://doi.org/10.1002/cbic.200300701.

(51) A Small-Molecule Probe Induces a Conformation in HIV TAR RNA Capable of Binding Drug-Like Fragments. *J. Mol. Biol.* **2011**, *410* (5), 984–996. https://doi.org/10.1016/j.jmb.2011.03.039.

(52) Yang K; Chang Jy; Cui Z; Li X; Meng R; Duan L; Thongchol J; Jakana J; Huwe Cm; Sacchettini Jc; Zhang J. Structural Insights into Species-Specific Features of the Ribosome from the Human Pathogen Mycobacterium Tuberculosis. *Nucleic Acids Res.* **2017**. https://doi.org/10.1093/nar/gkx785.

(53) Lin, A. H.; Murray, R. W.; Vidmar, T. J.; Marotti, K. R. The Oxazolidinone Eperezolid Binds to the 50S Ribosomal Subunit and Competes with Binding of Chloramphenicol and Lincomycin. *Antimicrob. Agents Chemother.* **1997**, *41* (10), 2127–2131. https://doi.org/10.1128/AAC.41.10.2127.

(54) Belousoff, M. J.; Venugopal, H.; Wright, A.; Seoner, S.; Stuart, I.; Stubenrauch, C.; Bamert, R. S.; Lupton, D. W.; Lithgow, T. CryoEM-Guided Development of Antibiotics for Drug-Resistant Bacteria. *ChemMedChem* **2019**, *14* (5), 527–531. https://doi.org/10.1002/cmdc.201900042.

(55) Belousoff Mj; Wright A; Seoner S; Stuart I; Stubenrauch C; Bamert Rs; Lupton Dw; Lithgow T; T, L. cryoEM-Guided Development of Antibiotics for Drug-Resistant Bacteria https://pubmed.ncbi.nlm.nih.gov/30667174/. https://doi.org/10.1002/cmdc.201900042.

(56) Loubresse, N. G. de; Prokhorova, I.; Holtkamp, W.; Rodnina, M. V.; Yusupova, G.; Yusupov, M. Structural Basis for the Inhibition of the Eukaryotic Ribosome. *Nature* **2014**, *513* (7519), 517–522. https://doi.org/10.1038/nature13737.

(57) Blaha, G.; Gürel, G.; Schroeder, S. J.; Moore, P. B.; Steitz, T. A. Mutations Outside the Anisomycin-Binding Site Can Make Ribosomes Drug-Resistant. *J. Mol. Biol.* **2008**, *379* (3), 505–519. https://doi.org/10.1016/j.jmb.2008.03.075.

(58) Warner, K. D.; Homan, P.; Weeks, K. M.; Smith, A. G.; Abell, C.; Ferré-D'Amaré, A. R. Validating Fragment-Based Drug Discovery for Biological RNAs: Lead Fragments Bind and Remodel the TPP Riboswitch Specifically. *Chem. Biol.* **2014**, *21* (5), 591–595. https://doi.org/10.1016/j.chembiol.2014.03.007.

(59) Lee, M.-K.; Bottini, A.; Kim, M.; Bardaro, M. F.; Zhang, Z.; Pellecchia, M.; Choi, B.-S.; Varani, G. A Novel Small-Molecule Binds to the Influenza A Virus RNA Promoter and Inhibits Viral Replication. *Chem. Commun.* **2013**, *50* (3), 368–370. https://doi.org/10.1039/C3CC46973E.

(60) Warner, K. D.; Chen, M. C.; Song, W.; Strack, R. L.; Thorn, A.; Jaffrey, S. R.; Ferré-D'Amaré, A. R. Structural Basis for Activity of Highly Efficient RNA Mimics of Green Fluorescent Protein. *Nat. Struct. Mol. Biol.* **2014**, *21* (8), 658–663. https://doi.org/10.1038/nsmb.2865.

(61) Edwards, T. E.; Ferré-D'Amaré, A. R. Crystal Structures of the Thi-Box Riboswitch Bound to Thiamine Pyrophosphate Analogs Reveal Adaptive RNA-Small Molecule Recognition. *Structure* **2006**, *14* (9), 1459–1468. https://doi.org/10.1016/j.str.2006.07.008.

(62) Wehrhan, L.; Hillisch, A.; Mundt, S.; Tersteegen, A.; Meier, K. Druggability Assessment for Selected Serine Proteases in a Pharmaceutical Industry Setting. *ChemMedChem* **2020**, *15* (21), 2010–2018. https://doi.org/10.1002/cmdc.202000425.

(63) Wong, W.; Bai, X.-C.; Sleebs, B. E.; Triglia, T.; Brown, A.; Thompson, J. K.; Jackson, K. E.; Hanssen, E.; Marapana, D. S.; Fernandez, I. S.; Ralph, S. A.; Cowman, A. F.; Scheres, S. H. W.; Baum, J. Mefloquine Targets the Plasmodium Falciparum 80S Ribosome to Inhibit Protein Synthesis. *Nat. Microbiol.* **2017**, *2* (6), 1–9. https://doi.org/10.1038/nmicrobiol.2017.31.