AUGMENTED BASE PAIRING NETWORKS ENCODE RNA-SMALL MOLECULE BINDING PREFERENCES

CARLOS OLIVER, VINCENT MALLET, ROMAN SARRAZIN GENDRON, VLADIMIR REINHARZ, WILLIAM L. HAMILTON, NICOLAS MOITESSIER, JÉRÔME WALDISPÜHL

S1. Data Preparation

S1.1. **Binding Site extraction.** A crucial step in our learning pipeline is the construction of ABPNs from crystal structure data. Once all crystal structures are acquired, we consider spheres of varying radii around the ligand to define a binding site. We studied two parameter choices which affect the number and quality of the extracted binding sites: radius and protein vs. RNA content. As the radius increases, we obtain a larger number of binding sites. However, since the crystal structures often contain proteins, we increase the probability that the binding site will be dominated by protein residues. We tehrefore compute the ratio of RNA to Protein residues in the binding site. The resulting counts for binding sites are shown in **Fig. S1.1**. From this data we choose a minimum RNA concentration of 0.6 for our training model. If a PDB contains multiple binding events of the same ligand, we keep only one at random to reduce redundancies. At this stage, we have identified a set of atomic coordinates which correspond to binding sites. Since we apply a hard distance cutoff in the crystal structure, the resulting graphs often have chain discontinuities. To address this issue, we add all 1-neighbour breadth first nodes to the original graph, as well as remove any disconnected components with fewer than 4 nodes.

S1.2. **Fingerprint representation.** We use the MACCS fingerpints to represent the chemical space of small molecules. A projection of this representation as well as the space occupied by RNA ligand is depcited in **Fig. S2**

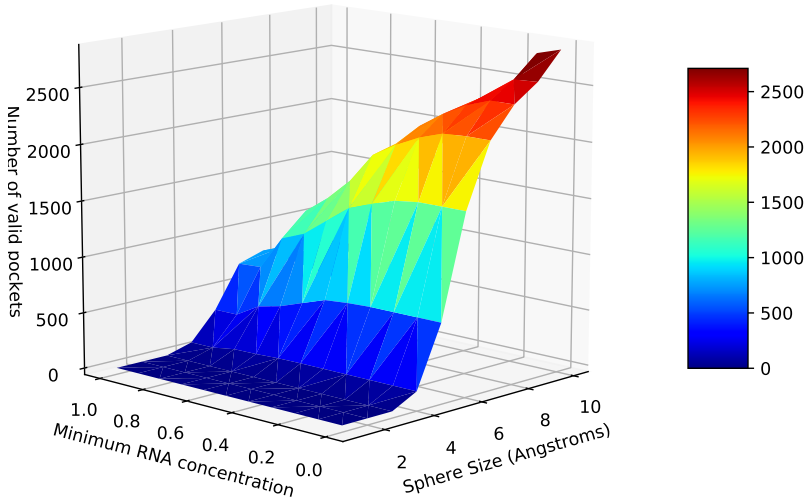

FIGURE S1. Number of binding sites retrieved versus distance threshold and RNA concentration threshold

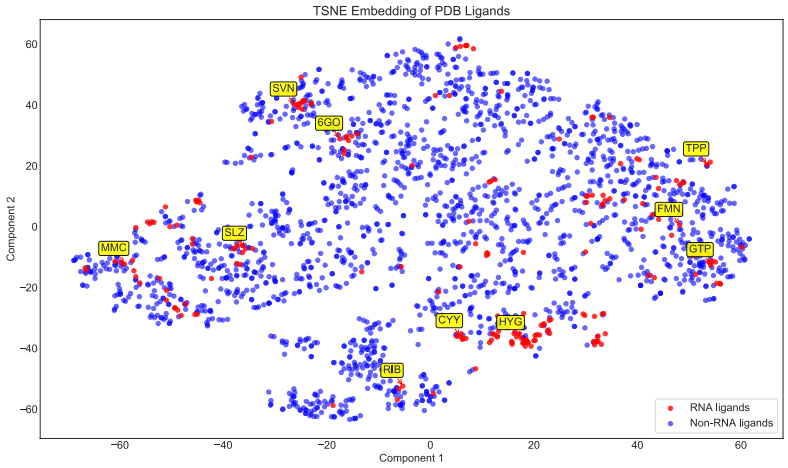

FIGURE S2. Two dimensional TSNE [1] embeddings of chemical fingerprints sampled from the PDB databank (RNA and protein binding). RNA ligands are highlighted in red and protein ligands in blue. We label a few interesting ligands such as 'KAN' and 'FMN' which correspond to well-known RNA binding classes known as aminoglycosides and riboswich-binding amines respectively.

S2. RGCN

We use a Relational Graph Convolutional Network (RGCN) [2] as the core of the fingerprint prediction model. An RGCN is a function that associates real vectors of size d for each node of a graph, known as node embeddings. Given initial node embeddings h_u^0 for each node u, an activation function σ and a graph structure that induces $\mathcal{R}$ edge-types and neighboring structure for each node $\mathcal{N}_u^r$, we can then use learnable matrices W that yield other node embeddings, according to the formula :

$$h_u^{l+1} = \sigma\left(W_{r_0}^l h_u^l + \sum_{r\in\mathcal{R}}\sum_{v\in\mathcal{N}_u^r}\frac{1}{c_{u,r}}W_r^l h_v^l\right)$$

Successive embeddings for each node are obtained by repeatedly using this process, until each node is attributed a final embedding $h_L(u)$. For this work, we consider the base pairing types to be distinct edge-typess. We believe this is a fair approximation given the results of isostericity comparisons showing that computing the geometric discrepancy between of all pairs of edge types yields close to a diagonal matrix [3].

Once node embeddings are computed, we concatenate the resulting embedding matrix with a one-hot encoding of the input graph's nucleotides (`A, U, C, G`). Next, graph-level representation is obtained by applying a widely used trainable Graph Attention Pooling layer [4], to map the node embeddings to a single vector $e_f \in \mathbb{R}^d$. Finally, we feed e_f through a Multi Layer Perceptron which yields probabilities for each index of the fingerprint $\hat{y}$.

$$\begin{aligned} e_f &= \text{GAT(Graph, final node embeddings)} \\ \hat{y} &= \text{MLP}(e_f) \end{aligned}$$

We supervise this process using the binary cross entropy $\mathcal{L}_{fp}$ between the predicted fingerprint and the observed one y over all dimensions i, and train the model by minimizing this loss over the training data.

$$\mathcal{L}_{fp} = \sum_{i=0}^{k}[y_i \cdot \log\hat{y}_i + (1-y_i)\cdot\log(1-\hat{y}_i)]$$

S3. Unsupervised Pre-Training

As described in the main text, we wish to define similarity functions K that take two nodes u and v and return a number close to one if they have similar neighborhoods and close to zero for dissimilar ones. In this paper we construct such a measure K from another measure d, that is a function which compares the sets of edges at a distance l, from u and v, denoted R_u^l, R_v^l. We then aggregate the results of the comparison of these sets for an increasing distance, according to the formula :

$$k_L(u,v) := N^{-1} \sum_{l=0}^{L-1} \lambda^l d(R_u^l, R_v^l)$$

The λ^l is a decay term which allows us to attend more to structural information close to the root nodes and we set $\lambda = 0.5$. We use N as a normalization constant ($\mathrm{N} = \frac{1-\lambda}{1-\lambda^L}$) to ensure the sum saturates at 1. d is defined to be a simple overlap measure on the histograms of base-pairing edge types f_R, and f'_R (i.e $f_R(i)$ stores the number of times edge type i is observed).

$$d(R, R') := \frac{|f_R \cap f'_R|}{|f_R \cup f'_R|}$$

To compensate for the over-representation of a few edge types such as backbones and Watson-Crick edges, we scale the d value with the commonly-used Inverse-Document Frequency (IDF) factor [5]. We show in **Fig. S3** an example of a pair of nodes that obtained a high similarity score K after embeddings were computed.

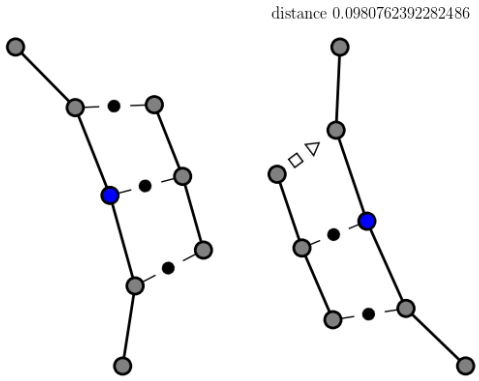

FIGURE S3. Example of pair of nodes given similar embeddings $\phi(u), \phi(v)$. The central pair of nodes which were used to make the comparison are colored in blue.

We then use the same RGCN function on the graph as before to annotate the nodes and get for each node, its vector embedding $h_L(u)$ that depend on the parameters W. Finally, as in the supervised setting, we define a loss function $\mathcal{L}_{rep}$ to minimize, that makes our network learn to approximate the dissimilarity function :

$$\mathcal{L}_{rep} = \|K(u,v) - cosine(h_L(u), h_L(v))\|_2^2$$

$$\text{rank}_{\mathcal{C}}(y, \hat{y}) = 1 - \frac{\rho_{y,\hat{y},\mathcal{C}}}{|\mathcal{C}|} \tag{1}$$

S4. Model Architecture and Hyperparameters

Hyperparameter	**Value**
RGCN Layers Dimensions	16, 16, 16
RGCN Number of Relations	13
RGCN Basis Sharing	None
RGCN Activation	ReLu
RGCN Dropout Probability	0.5
GAT Layer	Default
Fully Connected Dimensions	16 166

Table S1. Hyperparameter choices for learning pipeline. RGCN parameters are identical for the unsupervised pre-training and the fingerprint prediction networks.

S5. Results

Experiment	**Ranks**		**L2**	
	DecoyFinder	RNA	*DecoyFinder*	RNA
random	0.265880	0.276721	0.0384392	0.038299
majority	0.320012	0.269375	0.073969	0.074892
swap	0.319836	0.269233	0.071212	0.071308
no-label	0.317259	0.272816	0.072830	0.073768
primary	0.323843	0.064917	0.181	0.066853
secondary	0.318527	0.299428	0.074738	0.076667
ABPN	0.322124	0.301635	0.091479	0.092328
ABPN + unsup.	0.303712	0.294309	0.093006	0.095090

Table S2. Standard deviation on ligand screen ranks and L2 distance achieved on held-out binding sites for each condition on both decoy sets.

method_2 method_1	ABPN	secondary	primary	no-label	majority	swap	random
ABPN + unsup -	2.9-06	5.0e-26	1.4-22	2.0e-21	9.3e-25	7.1-26	2.3e-18
ABPN	-	1.6e-11	5.6e-11	1.4e-08	4.2e-10	6.3e-12	2.0e-08
secondary		-	3.2e-01	7.6e-01	1.2e-01	2.8e-02	1.7e-01
primary			-	4.2e-01	2.7e-01	2.3e-02	3.1e-01
no-label				-	5.5e-01	1.5e-02	1.7e-01
majority					-	3.6e-01	3.3e-01
swap						-	5.4e-01

TABLE S3. Pairwise Wilcoxon test for the DecoyFinder decoy set over the ligand ranks.

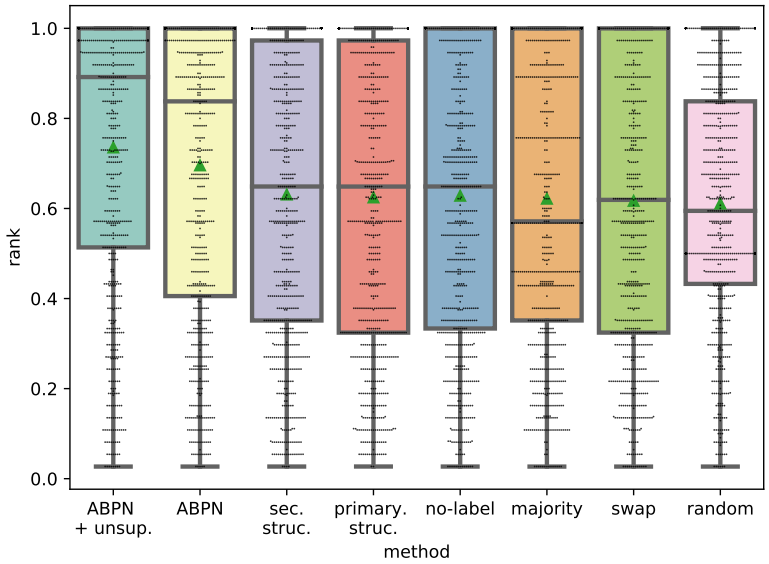

FIGURE S4. Ranks achieved by RNAmigos against the DecoyFinder screen.

Table S4: Details for each ligand in the dataset. PDB codes and full names are in the first two columns, followed by the number of occurrences. The `inforna` and `RNAmigos` columns contain the score achieved on average by each tool on the given ligand.

Ligand	**Name**	**Count**	`Inforna`	`RNAmigos`
SPD	spermidine	11		0.74

Continued on next page

Table S4: Details for each ligand in the dataset. PDB codes and full names are in the first two columns, followed by the number of occurrences. The `inforna` and `RNAmigos` columns contain the score achieved on average by each tool on the given ligand.

Ligand	**Name**	**Count**	`Inforna`	`RNAmigos`
NMY	neomycin	11	0.95	0.94
FME	n-formylmethionine	11		0.95
SPM	spermine	11	0.57	**0.82**
SAM	s-adenosylmethionine	11	0.60	**0.82**
PAR	paromomycin	11	0.97	**0.99**
LYS	lysine	10	0.75	**0.91**
GAI	guanidine	10	0.24	**0.65**
FMN	flavin	9	**0.48**	0.38
VAL	valine	9		0.81
PRF	7-deaza-7-aminomethyl-guanine	8	0.47	**0.78**
HPA	hypoxanthine	8	0.28	**0.94**
2BA	(2r,3r,3as,5r,7ar,9r,10r,10as,12r,14ar)-2,9-bi...	8	0.54	0.54
ADE	adenine	8	0.33	**0.89**
ACY	acetic	8		0.88
EDO	1,2-ethanediol	8		0.40
GLY	glycine	8		0.76
ARG	arginine	8	**0.72**	0.69
EOH	ethanol	7		0.60
PGE	triethylene	7		0.67
PPU	puromycin-5'-monophosphate	7		0.51
GOL	glycerol	7	0.69	**0.88**
PUT	1,4-diaminobutane	7		0.94
C2E	9,9'-[(2r,3r,3as,5s,7ar,9r,10r,10as,12s,14ar)-...	7		0.79
GUN	guanine	7	0.26	**0.97**
PEG	di(hydroxyethyl)ether	7		0.80
GET	geneticin	6		0.83
SPS	sparsomycin	6	0.32	0.34
SRY	streptomycin	6		0.58
TRP	tryptophan	6		0.88
LLL	(2r,3r,4r,5r)-2-((1s,2s,3r,4s,6r)-4,6-diamino-...	6		0.70
GNG	2'-deoxy-guanosine	5		0.98
TPP	thiamine	5		0.16
GLP	glucosamine	5		0.70

Continued on next page

Table S4: Details for each ligand in the dataset. PDB codes and full names are in the first two columns, followed by the number of occurrences. The `inforna` and `RNAmigos` columns contain the score achieved on average by each tool on the given ligand.

Ligand	**Name**	**Count**	`Inforna`	`RNAmigos`
SAH	s-adenosyl-l-homocysteine	5	0.73	**0.78**
5GP	guanosine-5'-monophosphate	5		0.23
CLM	chloramphenicol	5		0.45
NH4	ammonium	5	0.21	0.53
MES	2-(n-morpholino)-ethanesulfonic	5	0.21	0.21
ACT	acetate	5	0.23	0.42
ERY	erythromycin	5		0.42
HYG	hygromycin	5		0.81
PRP	alpha-phosphoribosylpyrophosphoric	5	**0.64**	0.35
S9L	2-[2-(2-hydroxyethoxy)ethoxy]ethyl	4		0.79
T1C	tigecycline	4		0.37
BLS	blasticidin	4		0.31
PRO	proline	4		0.97
G4P	guanosine-5',3'-tetraphosphate	4	0.45	**0.76**
AMZ	aminoimidazole	4	**0.53**	0.49
SIS	(1s,2s,3r,4s,6r)-4,6-diamino-3-{[(2s,3r)-3-ami...	4		0.89
3HE	4-{(2r)-2-[(1s,3s,5s)-3,5-dimethyl-2-oxocycloh...	4		0.81
AM2	apramycin	3	0.89	**0.95**
PO4	phosphate	3		0.10
TAC	tetracycline	3		0.35
VIR	virginiamycin	3		0.48
1PE	pentaethylene	3		0.92
EKJ	4-[(3-{2-[(2-methoxyethyl)amino]-2-oxoethyl}-1...	3		0.91
HGR	hygromycin	3		0.68
ANM	anisomycin	3		0.76
SCM	spectinomycin	3		0.76
8UZ	tc007	3		0.97
6HS	(1s,2s,3r,4s,6r)-4,6-diamino-3-{[(2s,3r)-3-ami...	3		0.88
XXX	(2r,3s,4r,5r,6r)-6-((1r,2r,3s,4r,6s)-4,6-diami...	3		0.98
CLY	clindamycin	3		0.54
NEG	negamycin	3		0.70
DOL	5-(2-diethylamino-ethanesulfonyl)-21-hydroxy-1...	3		0.04
B6M	(1r,2s,3s,4r,6r)-4,6-diamino-2-{[3-o-(2,6-diam...	3		1.00

Continued on next page

Table S4: Details for each ligand in the dataset. PDB codes and full names are in the first two columns, followed by the number of occurrences. The `inforna` and `RNAmigos` columns contain the score achieved on average by each tool on the given ligand.

Ligand	**Name**	**Count**	`Inforna`	`RNAmigos`
ZLD	n-{[(5s)-3-(3-fluoro-4-morpholin-4-ylphenyl)-2...	3		0.52
IDG	o-2,6-diamino-2,6-dideoxy-beta-l-idopyranose	3		0.93
GLN	glutamine	3		0.99
AG2	agmatine	2		0.70
3CO	cobalt	2		0.25
6GU	6-chloroguanine	2	0.15	**0.97**
8OS	5'-o-[(s)-hydroxy(4-methyl-1h-imidazol-5-yl)ph...	2	**0.40**	0.06
EGD	n-ethylguanidine	2	0.37	**0.64**
UAM	amicoumacin	2		0.79
RIO	ribostamycin	2	**0.96**	0.99
4BW	2-amino-9-[(2r,3r,3as,5r,7ar,9r,10r,10as,12r,1...	2		0.61
KSG	(1s,2r,3s,4r,5s,6s)-2,3,4,5,6-pentahydroxycycl...	2		0.82
SPK	spermine	2		0.50
6AP	9h-purine-2,6-diamine	2	0.27	**0.82**
SE4	selenate	2		0.06
GTP	guanosine-5'-triphosphate	2	**0.51**	0.34
ACA	6-aminohexanoic	2		0.98
747	(5z)-5-[(3,5-difluoro-4-hydroxyphenyl)methylid...	2		0.59
6GO	6-o-methylguanine	2	0.30	**0.94**
HMT	(3beta)-o~3~-[(2r)-2,6-dihydroxy-2-(2-methoxy-...	2		0.19
MLI	malonate	2		0.47
BDG	o-2,6-diamino-2,6-dideoxy-alpha-d-glucopyranose	2		0.94
CYY	2-deoxystreptamine	2		0.92
G6P	alpha-d-glucose-6-phosphate	2		0.84
PHA	phenylalaninal	2		0.87
TOC	2,3,6-trideoxy-2,6-diamino	2		0.91
PCY	pactamycin	2		0.68
SIN	succinic	2		0.89
PHE	phenylalanine	2		0.95
VIF	flopristin	2		0.13
TOA	3-deoxy-3-amino	2	0.86	**0.91**
TFX	2-[4-(dimethylamino)phenyl]-3,6-dimethyl-1,3-b...	2	0.01	**0.70**
BDR	beta-d-ribofuranosyl	2	0.77	0.77

Continued on next page

Table S4: Details for each ligand in the dataset. PDB codes and full names are in the first two columns, followed by the number of occurrences. The `inforna` and `RNAmigos` columns contain the score achieved on average by each tool on the given ligand.

Ligand	**Name**	**Count**	`Inforna`	`RNAmigos`
8AN	3'-amino-3'-deoxyadenosine	2		0.23
38E	(5z)-5-(3,5-difluoro-4-hydroxybenzylidene)-2,3...	2		0.96
TEP	theophylline	2	0.08	0.39
EUS	n-[(1r,2s,3s,4r,5s)-5-amino-4-{[(2s,3r)-3-amin...	2		0.71
TRS	2-amino-2-hydroxymethyl-propane-1,3-diol	2		0.81
CPT	cisplatin	2		0.68
EDE	edeine	2		0.71
AMP	adenosine	2	**0.61**	0.33
SUC	sucrose	2		0.63
1TU	4-(3,5-difluoro-4-hydroxybenzyl)-1,2-dimethyl-...	2	0.13	**0.98**
DAI	(3as,9as)-2-pentyl-4-hydroxymethyl-3a,4,9,9a-t...	2		0.25
LC2	n-[(1s,2r,3e,5e,7s,9e,11e,13s,15r,19r)-7,13-di...	2		0.82
SPE	thermine	2		0.61
AB9	(2r)-4-amino-n-{(1r,2s,3r,4r,5s)-5-amino-2-{2-...	2		0.89
TOY	tobramycin	2		0.98
MGX	1-methylguanidine	1	0.19	**0.93**
6MN	2-amino-2-deoxy-6-o-phosphono-alpha-d-mannopyr...	1		0.86
GE2	3,5-diamino-cyclohexanol	1		0.98
2QB	5-(azidomethyl)-2-methylpyrimidin-4-amine	1		0.25
GCP	phosphomethylphosphonic	1		0.45
GE1	3,4-dideoxy-2,6-amino-alpha-d	1	**0.82**	0.72
IPA	isopropyl	1		0.53
NMZ	(2s)-4-amino-n-{(1r,2s,3r,4r,5s)-5-amino-3-{[3...	1		0.87
ZZR	3,6-diamino-1,5-dihydro[1,2,4]triazolo[4,3-b][...	1		0.93
GMP	guanosine	1		0.98
GZ4	7,8-dimethyl-2,4-dioxo-10-(3-phenylpropyl)-1,2...	1		0.79
3LK	bc-3205	1		0.36
3TS	(2s,3s,4r,5r,6r)-2-(aminomethyl)-5-azanyl-6-[(...	1		0.92
P12	4-[amino(imino)methyl]-1-[2-(3-ammoniopropoxy)...	1	**0.62**	0.41
ZZS	1,3,5-triazine-2,4-diamine	1		0.94
EZP	n-[(1r,2r)-1,3-dihydroxy-1-(4-nitrophenyl)prop...	1		0.46
RPO	(1r,2r,3s,4r,6s)-4,6-diamino-2-{[3-o-(2,6-diam...	1		0.96
GZ7	10-(6-carboxyhexyl)-8-(cyclopentylamino)-2,4-d...	1		0.98

Continued on next page

Table S4: Details for each ligand in the dataset. PDB codes and full names are in the first two columns, followed by the number of occurrences. The `inforna` and `RNAmigos` columns contain the score achieved on average by each tool on the given ligand.

Ligand	**Name**	**Count**	`Inforna`	`RNAmigos`
ACE	acetyl	1		0.07
RIB	ribose	1	0.77	**0.84**
3K8	(14ar)-2,3,6-trimethoxy-11,12,13,14,14a,15-hex...	1		0.35
JS5	(2s,3s,4r,5r,6r)-5-amino-2-(aminomethyl)-6-((2...	1		0.91
SLZ	l-thialysine	1		0.78
THF	5-hydroxymethylene-6-hydrofolic	1		0.36
51B	2-[(3s)-1-{[2-(methylamino)pyrimidin-5-yl]meth...	1		0.07
218	1-[(4-amino-2-methylpyrimidin-5-yl)methyl]-3-(...	1		0.45
JS4	(2s,3s,4r,5r,6r)-5-amino-2-(aminomethyl)-6-((2...	1		0.93
62B	lefamulin	1		0.51
H4B	5,6,7,8-tetrahydrobiopterin	1	0.66	**0.70**
SLD	(3z)-n-[(4e)-5-(4-{(5s)-5-[(acetylamino)methyl...	1		0.68
EZM	n-[(1r,2r)-1,3-dihydroxy-1-(4-nitrophenyl)prop...	1		0.44
6NO	avilamycin	1		0.10
3J2	nagilactone	1		0.28
HEZ	hexane-1,6-diol	1		0.98
SJP	(2r,3r)-4-amino-n-[(1r,2s,3r,4r,5s)-5-amino-4-...	1		0.87
JS6	(1r,2r,3s,4r,6s)-4,6-diamino-2-{[3-o-(2,6-diam...	1		0.92
3L2	(4s,5r,10e,12z,16r,16as,17s,18r,19ar,23ar)-4-h...	1		0.11
ISH	(7r)-7-[(dimethylamino)methyl]-1-[3-(dimethyla...	1	0.16	0.22
P14	n-[2-(2-{[(4-{[amino(imino)methyl]amino}butyl)...	1	0.69	0.66
G34	(3as,4r,5s,6s,8r,9r,9ar,10r)-5-hydroxy-4,6,9,1...	1		0.17
P13	n-[2-(3-aminopropoxy)-5-(1h-indol-5-yl)benzyl]...	1	0.55	**0.61**
KAN	kanamycin	1		1.00
7DG	7-deazaguanine	1	0.24	0.27
ON0	(1r,2r,3s,4r,6s)-4,6-diamino-2-{[3-o-(2,6-diam...	1		0.97
AB6	(2r)-4-amino-n-((1r,2s,3r,4r,5s)-5-amino-4-[(2...	1		0.89
GZG	4-{benzyl[2-(7,8-dimethyl-2,4-dioxo-3,4-dihydr...	1		0.53
G80	(3as,4r,5s,6s,8r,9r,9ar,10r)-5-hydroxy-4,6,9,1...	1		0.50
T8B	thermorubin	1		0.66
2TB	1,3-diamino-4,5,6-trihydroxy-cyclohexane	1		0.89
3AW	6-methyl-1,3,5-triazine-2,4-diamine	1		0.95
2BP	9h-purin-2-amine	1	0.27	**1.00**

Continued on next page

Table S4: Details for each ligand in the dataset. PDB codes and full names are in the first two columns, followed by the number of occurrences. The `inforna` and `RNAmigos` columns contain the score achieved on average by each tool on the given ligand.

Ligand	**Name**	**Count**	`Inforna`	`RNAmigos`
EZG	n-[(1r,2r)-1,3-dihydroxy-1-(4-nitrophenyl)prop...	1		0.50
ARF	formamide	1		0.32
5CR	n-acetyl-l-phenylalanine	1		0.99
3KF	(2s,3r,4s,4ar)-2,3,4,7-tetrahydroxy-3,4,4a,5-t...	1		0.50
ZBA	12,13-epoxytrichothec-9-ene-3,4,8,15-tetrol-4,...	1		0.24
N6M	n-methyl-9h-purin-6-amine	1	0.23	**0.99**
4M2	3'-deoxy-3'-{[(2e)-3-(4-{[(4z)-6-o-(6-deoxy-3,...	1		0.23
N30	(1r,2r,3s,4r,6s)-4,6-diamino-2-[(5-amino-5-deo...	1		0.99
2QC	1-[4-(1,2,3-thiadiazol-4-yl)phenyl]methanamine	1	0.35	0.44
2HP	dihydrogenphosphate	1		0.65
0EC	6,7-dimethoxy-2-(piperazin-1-yl)quinazolin-4-a...	1	0.43	0.31
RBF	riboflavin	1		0.96
EKM	1-methyl-4-[(1e)-3-(3-methyl-1,3-benzothiazol-...	1		0.88
MMC	methyl	1		0.18
EVN	(2r,3r,4r,6s)-6-{[(2r,3ar,4r,4'r,5's,6s,6'r,7s...	1		0.03
SFG	sinefungin	1	0.73	**0.99**
EEM	[(3s)-3-amino-4-hydroxy-4-oxo-butyl]-[[(2s,3s,...	1		0.23
CIR	citrulline	1		0.70
AKN	(2s)-n-[(1r,2s,3s,4r,5s)-4-[(2r,3r,4s,5s,6r)-6...	1		0.95
34G	emetine	1		0.53
MT9	(3r,4s,5s,7r,9e,11s,12r)-12-ethyl-11-hydroxy-3...	1		0.91
RAP	rapamycin	1	0.50	**0.66**
S81	(1r,2r,3s,4r,6s)-4,6-diamino-2,3-dihydroxycycl...	1		0.94
ATP	adenosine-5'-triphosphate	1		0.08
GND	2-amino-5-guanidino-pentanoic	1		0.95
PRL	proflavin	1		0.56
G0B	(1s,2r,3s,4r,6s)-4,6-bis{[amino(iminio)methyl]...	1	**0.94**	0.90
PMZ	1-[10-(3-dimethylamino-propyl)-10h-phenothiazi...	1	0.03	0.05
TPS	thiamin	1		0.11
PDI	phosphoric	1		0.60
A2F	2-fluoroadenine	1	0.17	**0.94**
CNY	13,15-diamino-2-(aminomethyl)-3,4,9,12-tetrahy...	1		0.95
LEU	leucine	1		0.97

Continued on next page

Table S4: Details for each ligand in the dataset. PDB codes and full names are in the first two columns, followed by the number of occurrences. The `inforna` and `RNAmigos` columns contain the score achieved on average by each tool on the given ligand.

Ligand	**Name**	**Count**	`Inforna`	`RNAmigos`
3QB	lincomycin	1		0.71
7AL	chlorolissoclimide	1		0.33
29G	pyrimido[4,5-d]pyrimidine-2,4-diamine	1	0.33	**0.96**
D2X	3-[(4-hydroxy-2-methylpyrimidin-5-yl)methyl]-5...	1		0.36
HN8	haemanthamine	1		0.65
95H	˜{n}-[(1˜{r},2˜{r})-1-[(2˜{r},3˜{r},4˜{s},5˜{r...	1		0.26
MYC	3,5,7-trihydroxy-2-(3,4,5-trihydroxyphenyl)-4h...	1		0.53
AU3	gold	1		0.21
DKM	5-[(3s,4s)-3-(dimethylamino)-4-hydroxypyrrolid...	1		0.74
D2C	(2s,4s,4ar,5as,6s,11r,11as,12r,12ar)-7-chloro-...	1		0.37
6UQ	(2r,3s,4r,6s)-4-hydroxy-6-{[(2r,3ar,4r,4'r,5's...	1		0.09
3KD	(1s,2s,12bs,12cs)-2,4,5,7,12b,12c-hexahydro-1h...	1		0.36
B1Z	adenosylcobalamin	1		0.11
OLZ	o-(2-aminoethyl)-l-serine	1		0.81
ISI	(7s)-7-[(dimethylamino)methyl]-1-[3-(dimethyla...	1	0.16	0.18
AGU	aminoguanidine	1	0.25	0.82
FFO	n-[4-({[(6s)-2-amino-5-formyl-4-oxo-3,4,5,6,7,...	1	**0.66**	0.30
HRG	l-homoarginine	1		0.72
3J6	(3beta,7alpha)-3,7,15-trihydroxy-12,13-epoxytr...	1		0.41
CTC	7-chlorotetracycline	1		0.48
DX4	2-amino-1,9-dihydro-6h-purine-6-thione	1		0.94
L94	n'-{(z)-amino[4-(amino{[3-(dimethylammonio)pro...	1	0.30	0.27
NME	methylamine	1		0.88
PA1	2-amino-2-deoxy-alpha-d-glucopyranose	1	0.85	0.72
B12	cobalamin	1		0.01
L8H	4-methoxynaphthalen-2-amine	1		0.47
BFT	s-benzoylthiamine	1		0.26
ROS	n,n'-tetramethyl-rosamine	1		0.30
MIX	1,4-dihydroxy-5,8-bis({2-[(2-hydroxyethyl)amin...	1	0.73	0.78
NEB	2-deoxy-d-streptamine	1		0.81
M5Z	(1r,2r,3s,4r,6s)-4,6-diamino-2-{[3-o-(2,6-diam...	1		0.94
PQ0	2-amino-4-oxo-4,7-dihydro-3h-pyrrolo[2,3-d]pyr...	1	0.22	0.94
TOB	1,3-diamino-5,6-dihydroxycyclohexane	1		0.84

Continued on next page

Table S4: Details for each ligand in the dataset. PDB codes and full names are in the first two columns, followed by the number of occurrences. The `inforna` and `RNAmigos` columns contain the score achieved on average by each tool on the given ligand.

Ligand	**Name**	**Count**	`Inforna`	`RNAmigos`
V71	(1r,2r,3s,4r,6s)-4,6-diamino-2,3-dihydroxycycl...	1		0.94
3AY	pyrimidine-2,4,6-triamine	1		0.98
6O1	evernimicin	1		0.03
SVN	thieno[2,3-b]pyrazin-7-amine	1	0.15	0.94
29H	2-aminopyrimido[4,5-d]pyrimidin-4(3h)-one	1	0.27	0.98
6YG	2-[(3˜{s})-1-[(2-methoxypyrimidin-5-yl)methyl]...	1		0.59
GE3	5-methyl-4-methylamino-tetrahydro-pyran-2,3,5-...	1		0.73
MGR	malachite	1	0.05	0.34
IEL	n˜6˜-[(1z)-ethanimidoyl]-l-lysine	1		0.67
VIB	3-(4-amino-2-methyl-pyrimidin-5-ylmethyl)-5-(2...	1	0.30	**0.87**
BME	beta-mercaptoethanol	1		0.72
N33	(2s,3r,4r,5s,6r)-3-amino-4-({[(2s,3r,4r,5s,6r)...	1		0.99
ZIT	azithromycin	1		0.37
DGP	2'-deoxyguanosine-5'-monophosphate	1		0.94
3V6	bactobolin	1		0.49
IR3	iridium	1		0.20
EMK	(2r,3s,4r,5r,8r,10r,11r,12s,13s,14r)-2-ethyl-3...	1		0.04
G19	(2s,3ar,4r,5s,6s,8r,9r,9ar,10r)-2,5-dihydroxy-...	1		0.17
MUL	tiamulin	1		0.15
3H3	4-{(2r,5s,6e)-2-hydroxy-5-methyl-7-[(2r,3s,4e,...	1		0.63
7MB	agelastatin	1		0.97
917	n-({(5s)-2-oxo-3-[4-(1,3-thiazol-5-yl)phenyl]-...	1		0.36
RS3	1-deoxy-1-[8-(dimethylamino)-7-methyl-2,4-diox...	1		0.96

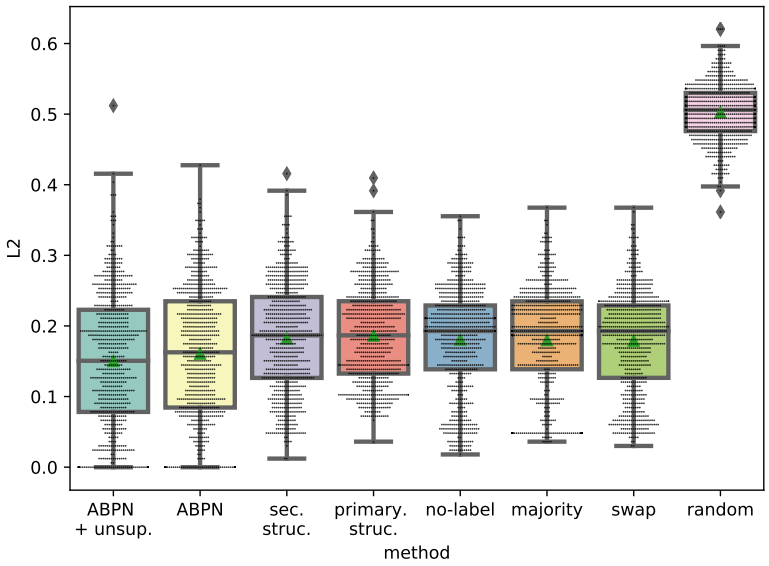

FIGURE S5. L2 distance from the native ligand achieved with RNAmigos.

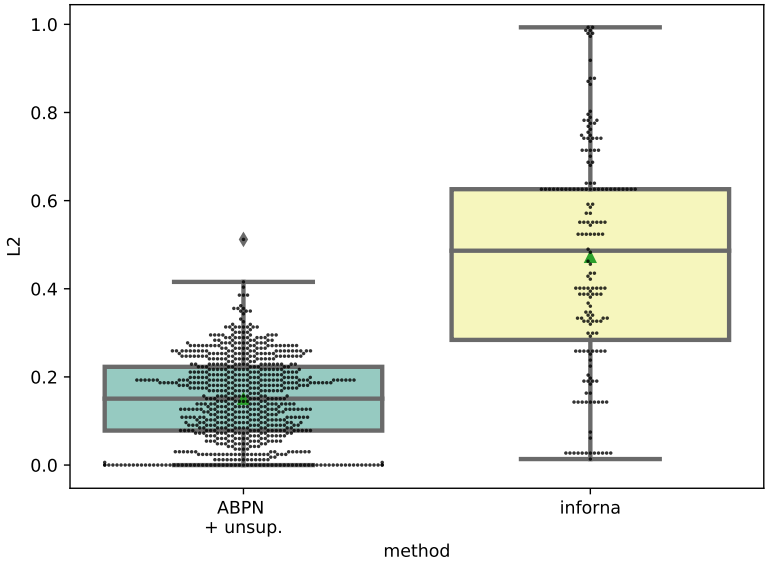

FIGURE S6. L2 distance from the native ligand achieved with Infornа

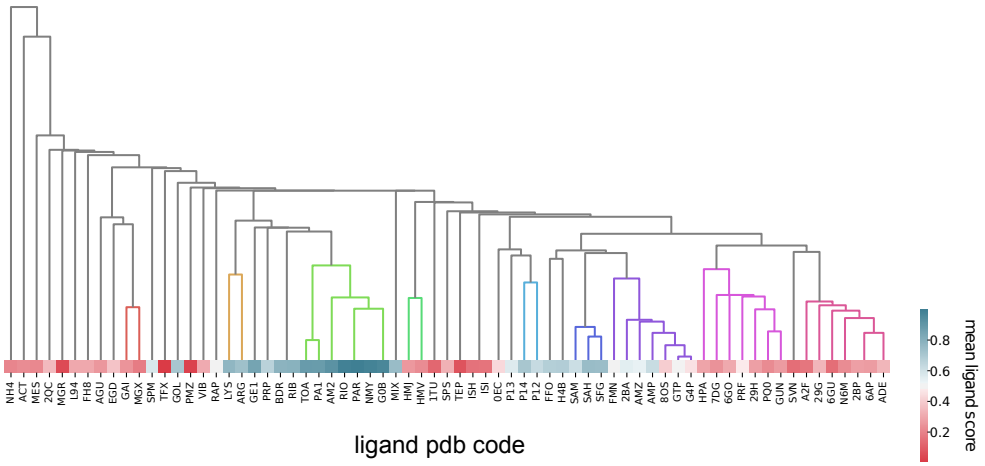

FIGURE S7. Performance per ligand type with Inforna software. A dendrogram is drawn to illustrate families of similar ligands.

References

[1] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[2] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer, 2018.

[3] Jesse Stombaugh, Craig L Zirbel, Eric Westhof, and Neocles B Leontis. Frequency and isostericity of rna base pairs. *Nucleic acids research*, 37(7):2294–2312, 2009.

[4] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[5] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142. Piscataway, NJ, 2003.