

# Mechanistic Studies of Small Molecule Ligands Selective to RNA Single G Bulges

Shalakha Hegde<sup>1,11</sup>, Sana Akhter<sup>2,11</sup>, Zhichao Tang<sup>1,11</sup>, Chang Qi<sup>3</sup>, Chenguang Yu<sup>4</sup>, Anna Lewicka<sup>5</sup>, Yu Liu<sup>6</sup>, Kushal Koirala<sup>2</sup>, Mikhail Reibarkh<sup>3</sup>, Kevin P. Battaile<sup>7</sup>, Anne Cooper<sup>8</sup>, Scott Lovell<sup>8</sup>, Erik D. Holmstrom<sup>9</sup>, Xiao Wang<sup>3</sup>, Joseph A. Piccirilli<sup>5,10</sup>, Qi Gao<sup>\*3</sup>, Yinglong Miao<sup>\*2</sup>, Jingxin Wang<sup>\*1</sup>.

<sup>1</sup> Section of Genetic Medicine, Department of Medicine, Biological Sciences Division, University of Chicago, Chicago, IL, USA

<sup>2</sup> Pharmacology and Computational Medicine Program, University of North Carolina, Chapel Hill, NC, USA

<sup>3</sup> Analytical Research & Development, Merck & Co., Inc, Rahway, NJ, USA

<sup>4</sup> Calibr-Skaggs Institute for Innovative Medicines, The Scripps Research Institute, La Jolla, CA, USA

<sup>5</sup> Department of Biochemistry and Molecular Biology, Biological Sciences Division, University of Chicago, Chicago, IL, USA

<sup>6</sup> Department of Chemistry, Rockhurst University, Kansas City, MO, USA

<sup>7</sup> New York Structural Biology Center, Upton, NY, USA

<sup>8</sup> Protein Structure and X-ray Crystallography Laboratory, University of Kansas, Lawrence, KS, USA

<sup>9</sup> Department of Molecular Biosciences, University of Kansas, Lawrence, KS, USA

<sup>10</sup> Department of Chemistry, Physical Sciences Division, University of Chicago, Chicago, IL, USA

<sup>11</sup> These authors contributed equally.

\* E-mail: wang.jingxin@uchicago.edu, yinglong\_miao@med.unc.edu, qi.gao1@merck.com

**Keywords:** RNA, small molecule, minor groove, G bulge, SOFAIR.

**Abstract:** Small-molecule RNA binders have emerged as an important pharmacological modality. A profound understanding of the ligand selectivity, binding mode, and influential factors governing ligand engagement with RNA targets is the foundation for rational ligand design. Here, we report a novel class of coumarin derivatives exhibiting selective binding affinity towards single G RNA bulges. Harnessing the computational power of all-atom Gaussian accelerated Molecular Dynamics (GaMD) simulations, we unveiled a rare minor groove binding mode of the ligand with a key interaction between the coumarin moiety and the G bulge. This predicted binding mode is consistent with results obtained from structure-activity-relationship (SAR) studies and transverse relaxation measurements by NMR spectroscopy. We further generated 444 molecular descriptors from 69 coumarin derivatives and identified key contributors to the binding events, such as charge state and planarity, by lasso (least absolute shrinkage and selection operator) regression. Strikingly, small structure perturbations on these key contributors, such as the addition of a methyl group that disrupts the planarity of the ligand resulted in > 100-fold reduction in the binding affinity. Our work deepened the understanding of RNA-small molecule interactions and integrated a new generalizable platform for the rational design of selective small-molecule RNA binders.

## Introduction

RNA plays critical roles in gene regulation and various cellular processes in almost all life forms, including transcription, translation, splicing, and epigenetic modifications.<sup>1,2</sup> Selective targeting of RNA structures using small molecules is an important pharmacological modality that complements traditional protein targeting approaches.<sup>3-9</sup> For example, bacteria ribosomal RNA (rRNA) is an important antibiotic target with numerous clinically validated drug classes, such as aminoglycoside, tetracycline, macrolide, lincosamide, and oxazolidinone.<sup>10</sup> Recently, two

synthetic compounds, risdiplam and branaplam, both targeting precursor messenger RNA (pre-mRNA)-U1 small nuclear ribonucleoprotein (snRNP) complex, attracted tremendous attention as RNA splicing modulators to treat genetic diseases, including spinal muscular atrophy<sup>11-17</sup> and Huntington's disease.<sup>18-20</sup> We previously demonstrated that a class of coumarin analogs of risdiplam can induce GA-rich sequences to form loop-like structures using molecular dynamics (MD) simulations and proposed that this interaction in cells provided additional selectivity of the coumarin derivatives to the GA-rich SMN2 gene.<sup>21</sup>

In addition to rRNA in bacteria and pre-mRNA in humans, several other classes of RNA have been targeted by chemical probes and drug candidates, including bacteria riboswitches,<sup>22-24</sup> yeast self-splicing introns,<sup>25</sup> microRNAs,<sup>26-31</sup> untranslated regions of mRNAs<sup>32-34</sup>, and long non-coding RNAs (lncRNAs).<sup>35-37</sup> In viruses, highly structured RNA regions have also been explored as targets for small molecules, such as an internal ribosome entry site (IRES) in the 5'-untranslated region (UTR) of the hepatitis C virus (HCV)<sup>38-41</sup> and a transactivation response (TAR) hairpin in human immunodeficiency virus 1 (HIV-1).<sup>42-44</sup> After the outbreak of SARS-CoV-2, we and others illustrated that the structural elements in the SARS-CoV-2 genome can also be targeted to suppress virus replication.<sup>45-50</sup> Specifically, we discovered that some coumarin derivatives (**Fig. 1**) can be "repurposed" to selectively bind to a single G bulge in 5' UTR of SARS-CoV-2 without retaining splicing modulatory activities or binding to GA-rich loops.<sup>46</sup> We further demonstrated that covalently linking a ribonuclease (RNase) L recruiter and the coumarin-based G bulge binder yielded an active ribonuclease targeting chimera (RIBOTAC), which is effective in targeting SARS-CoV-2-infected epithelial cells.<sup>45</sup>

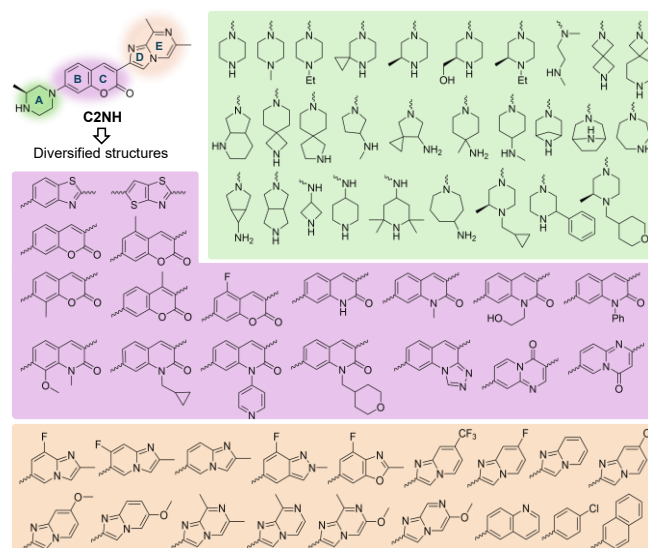
In general, attaining selective and effective targeting of RNA using small molecules is a challenging endeavor due to several factors, including its conformational flexibility/heterogeneity and its polyanionic backbone, which prevents the formation of deep hydrophobic binding pockets.<sup>8,51</sup> Cheminformatics work has uncovered key factors governing the activity and selectivity of RNA binders<sup>52–55</sup> and has been recently further advanced by machine learning approaches.<sup>56</sup> However, the lack of methods for mechanistic studies of flexible RNA–small molecule ligand interactions critically limits further optimization of RNA ligands. A powerful approach to studying RNA–small molecule interactions is to use MD simulations,<sup>21,57,58</sup> which are able to fully account for the RNA flexibility on an atomistic level. Here, we present an integrated platform that combines all-atom Gaussian accelerated MD (GaMD) simulations, which can rapidly predict ligand binding modes, with NMR transverse relaxation ( $R_2$ ) measurements and structure-activity relationship (SAR) studies that experimentally probe RNA–ligand interactions.<sup>21,59</sup> We envision that our new platform for mechanistic studies on RNA ligands can provide a systematic approach to improving binding affinity for other RNA targets.

## Results and Discussion

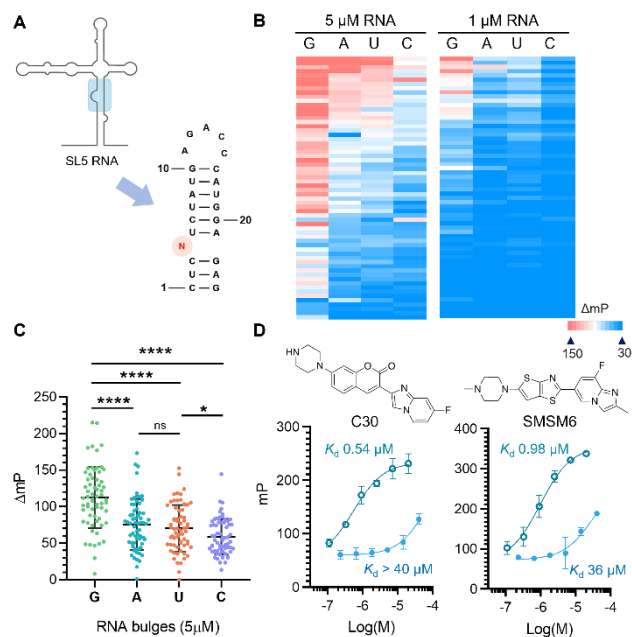
### Coumarin derivatives selectively bind to RNA G (1×0) bulge

The prototype coumarin derivative C2NH, which binds to RNA single G bulges (denoted as G 1×0 bulge) at a moderate binding affinity, contains five heterocyclic rings: piperazine (A), coumarin (BC), and a [5,6]-fused ring (DE) (**Fig. 1**). We previously reported C2NH as an active splicing modulator that can bind to a GA-rich loop within the SMN2 gene.<sup>21</sup> We modified the E ring to remove the splicing modulatory activity and repurposed the scaffold, resulting in a potent G 1×0 bulge binder that strongly associates with a structural motif in the RNA genome of SARS-CoV-2.<sup>21,46</sup> To further probe the mechanism of the coumarin derivative in RNA binding interaction, we synthesized a collection of 69 analogs of C2NH (**Fig. 1**). Each compound in this collection comprises at least one ring distinct from the parent compound. For instance, in Ring A, the piperazine was replaced by cyclic amines of varying sizes. In Ring BC, the coumarin was substituted by other heterocycles with various substituents. Similarly, the [5,6]-fused Ring DE was replaced by [6,5]- or [6,6]-fused rings (**Fig. 1**).

All compounds in this collection are fluorescent with an excitation/emission wavelength at ~400/480 nm, which allowed us to use fluorescence polarization (FP) assay to rapidly determine their binding affinity to the bulge G RNA. Using a bulged G RNA segment from SARS-CoV-2 SL5 RNA (RNA1) as a model, we extensively profiled this 69-compound library against all four 1×0 RNA bulge variants (RNA1–4) for binding affinities (**Fig. 2A**, **Supplementary Fig. 1**, and **2**; **Supplementary Table 1**). Almost all binding molecules showed superior selectivity for the G bulge compared to other RNA bulges (bulged A, U, and C), as judged by the polarization change ( $\Delta$ mP) at two concentrations (1 and 5  $\mu$ M) (**Fig. 2B**).



**Fig. 1.** Molecular diversity of coumarin derivative analogs designed to bind bulged G RNA. The A, B/C, and D/E rings were replaced by the shaded green, purple, and orange structures, respectively (wavy lines = connecting bonds).



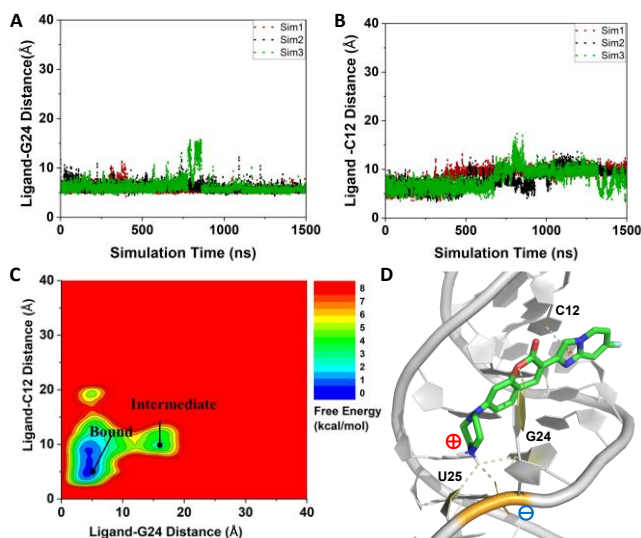
**Fig. 2.** (A) RNA structures of the 1×0 RNA bulges used for in vitro binding profiling. N = G, A, U, or C (RNA1–4). (B) Heatmap profile of the  $\Delta$ mP = (Polarization<sup>RNA-ligand</sup> – Polarization<sup>ligand</sup>)×1000 for RNA binders in the presence of [RNA] = 5 or 1  $\mu$ M (red = high polarization, blue = low polarization). (C)  $\Delta$ mP of RNA-ligand complex for RNA ligands at 5  $\mu$ M. Each data point represents a measurement of a ligand in the 69-compound collection. \*\*\*\* indicates  $p < 0.0001$ . (D) Dose-response curves for compounds selectively (SMSM6, C30) binding to bulged G RNA (RNA1) compared to an 11-nucleotide GA-rich sequence that would form a double loop-like RNA structure.

Statistical analysis of all 69 compounds revealed that the binding affinity for different RNA bulges followed the following trend: G >>

$A \approx U > C$  (**Fig. 2C**). Since C2NH can also bind to GA-rich RNA loops<sup>21</sup> via an induced-fit mechanism, resulting in the formation of a double loop-like structure, we tested the binding of coumarin derivatives to both a bulged G and a flexible GA-rich RNA (5'-U(GAAG)<sub>2</sub>GU). Interestingly, certain compounds, such as C30 and SMSM6, demonstrated > 35-fold selectivity towards bulged G over GA-rich RNA (**Fig. 2D**), while few compounds bind to both structures with comparable binding affinities (**Supplementary Fig. 3**). This suggests that coumarin derivatives may employ a distinct binding mechanism to selectively target RNA G bulges.

### GaMD simulations captured spontaneous minor groove binding of coumarin derivatives to RNA G (1x0) bulges

To explore the binding of specific RNA G bulge ligands, we performed all-atom simulations using the GaMD-enhanced sampling method.<sup>59</sup> GaMD works by adding a harmonic boost potential to smooth the potential energy surface and reduce system energy barriers.<sup>59</sup> GaMD has been shown to accelerate biomolecular simulations by orders of magnitude.<sup>60,61</sup> For our GaMD simulations, we used the G bulge binder C30, which was among the most specific binders, and a model RNA hairpin with a G (1x0) bulge (RNA5: 5'-AAGAUGGAGAGCGAAACACACUCGUCUAUCUU; see **Extended Data Fig. 1A** for its secondary structure).



**Fig. 3. GaMD simulations captured stable binding of coumarin derivative C30 to RNA5.** (A) Time courses of the center-of-mass distance between heavy atoms of the coumarin core in the ligand and the RNA bulge G24 calculated from three 1500 ns GaMD production simulations. (B) The center-of-mass distance between heavy atoms of the fused D/E ring of C30 and RNA nucleotide C12 plotted as a function of simulation time. (C) 2D free energy profile calculated with all three GaMD simulations combined, showing two distinct low-energy states, namely the “Bound” and “Intermediate”. (D) Representative conformation of RNA-C30 complex in the Bound state (grey dash line =  $\pi$ - $\pi$  stacking, yellow dash line = hydrogen bonding, orange dash line = ionic interaction). The “Intermediate” conformation is shown in Extended Data Fig. 2.

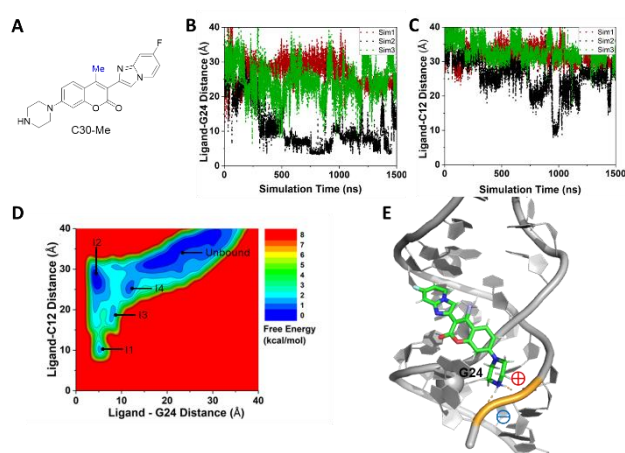
We found that C30 bound spontaneously to the G bulge and minor groove of RNA5 during the GaMD equilibration (**Extended Data Fig. 1B**) and three independent 1500 ns GaMD production simulations (**Fig. 3**). Upon binding to the minor groove of RNA5, the distance between the coumarin core of the ligand and the bulged G at position 24 (G24) was 3.5–5 Å, within a polar interaction range (**Fig. 3A**). Moreover, a  $\pi$ - $\pi$  stacking interaction was observed between C12 and the fused D/E ring of the C30 ligand in simulations (**Fig. 3B**). We used these distances as reaction coordinates to further calculate a 2D free energy profile of C30 binding to RNA5, which showed two low-energy states, designated as “Bound” (more stable) and “Intermediate” states (**Fig. 3C**; Bound state structure was deposited in Model Archive Project ma-q6hl4). To experimentally probe the minor groove binding mechanism that we observed in the GaMD simulations, we conducted additional FP binding assays using C30 and various DNA versions of the RNA G bulge sequences (same sequences as RNA1 and RNA5). Our results show that the deoxyribose modification gives rise to a > 13-fold decrease in binding affinity (**Supplementary Fig. 4**). This result differed from what we observed with GA-rich loop binders, where the DNA aptamers bind to the ligands better than the RNA aptamers with the same sequences.<sup>21</sup> Given that double-stranded (ds) DNA has a narrower minor groove than dsRNA,<sup>62</sup> this result indicated that the groove region in dsDNA may not have sufficient space for C30 binding, supporting the minor groove binding mechanism.

In the simulation predicted “Bound” state, C30 formed three primary interactions within the minor groove of RNA5 (**Fig. 3D**). (1) The bulged G (G24) contributed to a hydrogen bond via its N1 position to the coumarin lactone moiety in C30. (2) A phosphate group in the RNA backbone was involved in an ionic interaction with the protonated  $\text{NH}_2^+$  group in the piperazine ring of C30. (3) Nucleotide C12 formed  $\pi$ - $\pi$  stacking interactions with the ligand C30 in the RNA minor groove. In the transient “Intermediate” state, C30 was located at a much larger distance from the G24 nucleotide and did not insert into the RNA minor groove (**Fig. 3C** and **Extended Data Fig. 2**).

We performed further GaMD simulations on two inactive analogs of C30, namely C30-Me (**Fig. 4A**) and SMSM64 (**Extended Data Fig. 3A**). C30-Me merely has an additional methyl group on the C ring compared to C30, which would break the planarity of Rings B/C and D/E in the compound (see discussions below). On the other hand, SMSM64 has an N-pyridinyl quinolone replacing the coumarin core, whose bulkiness might block the polar interaction with the RNA G bulge. In experiments, both compounds exhibited > 100-fold reduced binding affinities toward RNA5, with a dissociation constant ( $K_d$ ) of > 50  $\mu\text{M}$  for both of the compounds, in comparison to C30, which has a  $K_d$  of  $0.27 \pm 0.01 \mu\text{M}$  to RNA5 (**Supplementary Fig. 5**). Similar binding affinities were observed for these compounds when binding to RNA1 (see **Supplementary Table 1**).

In all three 1500 ns GaMD simulations for C30-Me, the ligand seldom reached the target site in the minor groove of RNA5 (**Fig. 4**). In the situation where C30-Me transiently interacted with G24 nucleotide (“Sim2” in **Fig. 4B**), the ligand remained out of the RNA minor groove with a distance > 10 Å from nucleotide C12 (**Fig.**

**4C**). Altogether, four transient binding states were identified from the free energy profile, designated as "Intermediate" states I1–I4, as well as the "Unbound" state, where the ligand dissociated from the RNA (**Fig. 4D** and **Extended Data Fig. 4**). The presence of multiple intermediate states suggested that the ligand explored various binding positions but was unable to achieve stable insertion into the minor groove. These intermediate conformations all maintained ionic interactions between the positively charged piperazine ring on the ligand and at least one phosphate group on the RNA backbone but could not fit into the minor groove (**Fig. 4E**). For SSM64, the ligand remained mostly more than ~15 Å away from key nucleotides G24 and C12 throughout the 1500 ns GaMD simulations (**Extended Data Fig. 3B** and **3C**). The resulting free energy profile showed only an "Unbound" state (**Extended Data Fig. 3D** and **3E**).



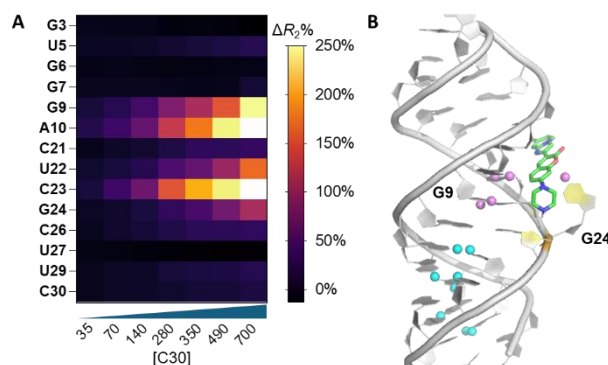
**Fig. 4. GaMD simulations captured transient binding of ligand C30-Me to RNA5:** (A) Chemical structure of C30-Me. (B) Time courses of the center-of-mass distance between heavy atoms of the coumarin core in the ligand and the RNA bulge G24 calculated from three independent 1500 ns GaMD simulations. (C) The center-of-mass distance between the heavy atoms of the fused D/E ring of C30-Me and RNA nucleotide C12 plotted as a function of simulation time. (D) 2D free energy profile calculated with all three GaMD simulations combined, showing five low-energy states, namely the "I1", "I2", "I3", "I4" and "Unbound". (E) Representative conformation of C30-Me–RNA5 complex in the I1 state (orange dash line = ionic interaction).

Interestingly, during the simulation studies, we observed a highly flexible G24 nucleotide within the RNA5 when the ligand was not bound to the RNA. Experimentally, we screened ~20 crystal structures of RNA1 obtained using fragment antigen-binding region (Fab) chaperon-assisted crystallography (for a representative structure, see Protein Data Bank with accession code 9DN4) and observed dynamic conformations of the G bulge nucleotide,<sup>63</sup> whereas other nucleotides remained relatively static (**Extended Data Fig. 5** and **Supplementary Table 2**). This result is also consistent with our chemical probing results in SARS-CoV-2 RNA, where a high SHAPE (selective 2' hydroxyl acylation analyzed by primer extension) signal was observed with high concentrations of acylating agents (e.g., 10 mM FAI-N3).<sup>46</sup> Importantly, GaMD results demonstrated that the ligand with high affinity to RNA5 would bind and stabilize the flexible G bulge. Expectedly, we compared root-mean-square fluctuations (RMSFs) of each nucleotide in RNA5 across the simulated systems of C30, C30-Me, and SSM64, and found that the

interaction with C30 resulted in the lowest nucleotide G24 fluctuation throughout the simulation time course (**Extended Data Fig. 6**).

### NMR validation of the minor groove binding mode

Next, we used NMR experiments to validate the predicted binding mode between coumarin analog and bulged G RNA. First, we assigned imino protons and some other protons on the nucleobases in <sup>1</sup>H NMR using a reported assignment that contains the segment of RNA5.<sup>64</sup> The assigned peaks were distinguishable ones within 0.15 ppm from the reported <sup>1</sup>H NMR chemical shifts (**Supplementary Table 3**). Next, we applied a recently published NMR method, <sup>1</sup>H SOFAIR (band-Selective Optimized Flip-Angle Internally-encoded Relaxation),<sup>65</sup> to quantify  $R_2$  relaxation rate of the receptor signals in order to characterize ligand-receptor interactions.  $R_2$  relaxation reflects on dynamics and motion changes of molecules, which is sensitive to weak binding ( $K_d \sim \mu\text{M}$ ), and has been widely utilized as an NMR approach for identifying the binding sites of biomolecules<sup>66,67</sup>. Here, the SOFAIR pulse sequence<sup>65</sup> was utilized to facilitate signal acquisition with high sensitivity of an RNA sample at mM concentration. Notably, SOFAIR was specifically designed to speed up data acquisition, and in this instance, led to a reduction in acquisition time from several hours, characteristic of conventional proton  $R_2$  measurements using Carr-Purcell-Meiboom-Gill (CPMG) type of methods,<sup>68,69</sup> to ~20 minutes.



**Fig. 5. NMR relaxometry validation of the minor groove binding mode:** (A) Normalized  $R_2$  relaxation rate percentage changes obtained from each assigned proton of RNA nucleotide in the absence and presence of C30 ligand. The colored columns in the bar plot represented  $R_2$  measured at different ligand concentrations. Normalized  $R_2$  values were calculated using the equation  $\Delta R_2^{\%} = \frac{R_2^{\text{ligand}^+} - R_2^{\text{ligand}^-}}{R_2^{\text{ligand}^-}} \times 100\%$ . (B) Overlay of the simulation-predicted "Bound" state of C30-RNA5 and all identifiable protons in <sup>1</sup>H SOFAIR (pink = increased  $\Delta R_2^{\%}$ ; blue = unchanged or decreased  $\Delta R_2^{\%}$ ).

The  $R_2$  relaxation rates were obtained from RNA nucleobases during the titration of the C30 ligand (**Supplementary Table 4**). As shown in **Fig. 5A**, the titration of C30 induced an overall  $R_2$  change, indicating binding between coumarin analog and bulged G RNA. The most pronounced increase in  $R_2$  was observed in G9, A10, U22, C23, and G24, implying the direct involvement of these nucleotides in binding. In contrast, the relaxation rates observed from G3 to G7 and C26 to C30 exhibited much smaller increases, or even negative changes upon ligand addition, suggesting that these regions of the RNA are not directly involved in binding. These findings from NMR experiments regarding the bound and unbound RNA nucleotides are consistent with those obtained from the GaMD simulations (**Fig. 5B**). Interestingly, the putative binding location is selective to one side of the G bulge (U22–G24),

implying sequence selectivity in the minor groove. It is worth noting that only a few small molecular ligands have been reported as minor groove binders (e.g., PDB 1QD3),<sup>70</sup> likely because the minor groove is wide and shallow in A-form dsRNAs. In our study, the NMR data strongly supported a minor groove binding mechanism for C30, as the ligand is unlikely to bind to the major groove of the RNA given the observed  $R_2$  relaxation changes. This finding further highlights the critical role of the bulged G in ligand interactions within this unusual minor groove binding mechanism.

### Molecular features on the ligands for RNA binding

We also determined how the molecular characteristics of these ligands contribute to their efficacy in RNA G bulge binding. Our approach involved a quantitative structure-activity-relationship (QSAR) investigation based on in vitro binding affinity data. Given the similarities in shape, size, and molecular scaffolds of our 69-compound library, we expected the QSAR analysis to offer detailed molecular insight into the specific structural and electronic properties responsible for their potency.

We used Molecular Operating Environment (MOE) software to individually predict the most likely protonation state based on the 2D structures of each molecule. The majority of molecules were found to be mono-protonated at the aliphatic cyclic amine (A ring), with a few exceptions that carried two positive charges (Supplementary Data). We then optimized the 3-dimensional (3D) structure of each molecule using ab initio density-functional theory (DFT) calculation with B3LYP 6-31G(d) basis set (Supplementary Data). Using these 3D structures as input, we generated 443 molecular descriptors using MOE software (Supplementary Data). To account for the planarity of the coumarin derivatives, we introduced a new dihedral descriptor between the plane B/C and D/E based on the most stable conformer predicted by DFT calculations. We then used the least absolute shrinkage and selection operator (lasso) regression technique to identify the important electronic and structural features among these molecular descriptors using a modified analytical pipeline.<sup>56,71</sup> Lasso regression is a linear regression approach used for feature selection, which effectively eliminates unimportant variables. This process resulted in 16 molecular descriptors that significantly contributed to the binding affinity (Supplementary Table 5), of which eight molecular features are related to the charge and shape of the RNA ligands (Table 1).

**Table 1.** Molecular descriptors selected by lasso regression for RNA binding.<sup>[a]</sup>

Molecular descriptor	Description <sup>[a]</sup>	Class/ Impact <sup>[b]</sup>
FCharge	Total formal charge of the molecule.	charge +/-
a_base	Number of basic atoms.	charge +/-
PEOE_VSA_FPOS	Fractional positive VDWSA.	charge +/-
PEOE_VSA_FNEG <sup>[c]</sup>	Fractional negative VDWSA.	charge /--
PEOE_VSA_NEG	Total negative VDWSA.	charge /--

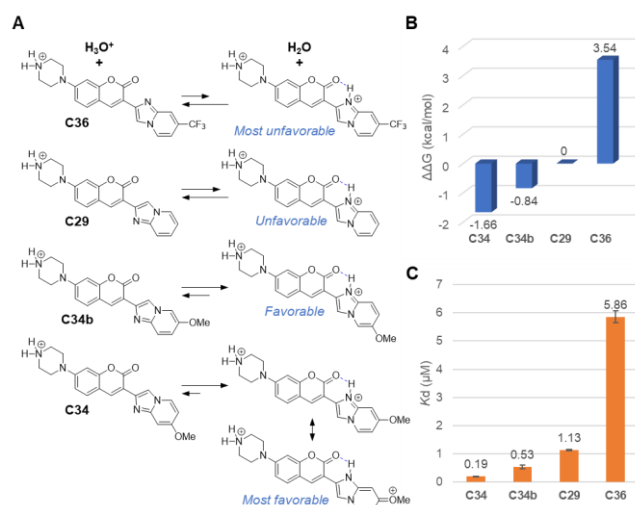
NPR1	PMI <sup>[d]</sup> ratio: PMI1/PMI3	shape /--
std_dim1	The square root of the largest eigenvalue of the covariance matrix of atomic coordinates.	shape +/-
$\omega$ (BC-DE)	Dihedral between BC and DE rings in the optimal structure.	shape /--

[a] VDWSA = van der Waals surface area. [b] + and – signs indicate variables positively or negatively correlated to the binding affinity. [c] PEOE\_VSA\_FNEG = –PEOE\_VSA\_FPOS. [d] PMI = Normalized principal moments of inertia.

The five charge-related molecular descriptors were based on the total formal charge of the molecule (FCharge), the number of basic atoms that can be potentially protonated in physiological pH (A\_base), the fractional positive (PEOE\_VSA\_FPOS) and negative (PEOE\_VSA\_FNEG) charges per unit area, and the total negative charge per unit area (PEOE\_VSA\_NEG). Since RNA is densely negatively charged, it is reasonable that positive charges would significantly contribute to RNA binding due to charge attraction. In the GaMD simulations with C30, intermolecular ionic interactions between the positive charge on the piperazine ring of C30 and phosphate groups on the RNA backbone were critical in maintaining the stability of the RNA-ligand complex. When we acetylated the piperazine ring of C30 at the N4 position (C30-Ac) to prevent protonation, the binding affinity decreased by a factor of > 5, highlighting the importance of electrostatic interaction between the ligand and RNA (Supplementary Fig. 6). We further hypothesized that the ligand used the positive charge on Ring A to explore suitable binding pockets at the early stage of the binding process. This hypothesis was supported by GaMD simulations, in which we observed all identifiable transient binding states (“Intermediate” states) of C30–RNA5 and C30–Me–RNA5 complexes retained an ionic interaction with RNA backbone phosphates (Extended Data Fig. 2 and Extended Data Fig. 4; Supplementary Movie 1).

We also verified the impact of local positive charges on coumarin derivatives on in vitro binding by selecting four compounds, C29, C36, C34, and C34b, which only differ in the structures of the E ring. These compounds have two potential protonation sites: a piperazine A ring and an imidazole D ring. The second protonation site on the D ring can be partially stabilized by the coumarin moiety by forming an internal hydrogen bond. We speculated that the propensity of imidazolium formation significantly depends on the substituents on the E ring (Fig. 6A). For example, substituting the E ring with a trifluoromethyl group makes the molecule less amenable to protonation due to the electron-withdrawing effects. In contrast, the presence of an electron-donating methoxy group in compound C34b enhances the favorability of imidazolium formation. When the methoxy group is positioned at the 4' location (C34), the existence of a resonance structure further contributes to stabilizing the positive charge (Fig. 6A). We verified the protonation energy of the four compounds relative to C29 using DFT calculations and compared it with the in vitro binding data (Fig. 6B and 6C). The dissociation constants for these four compounds exhibit a consistent trend with respect to protonation

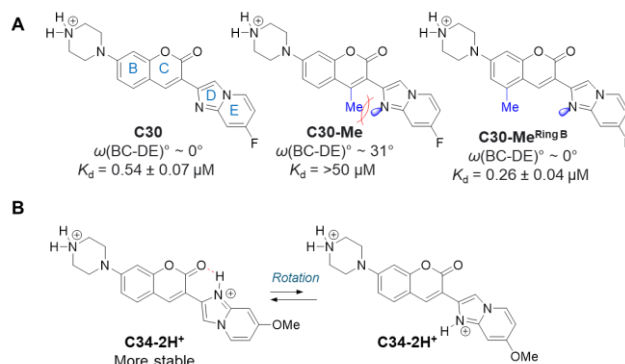
energy, providing compelling evidence that local positive charges significantly contribute to the binding affinity.



**Fig. 6.** (A) Equilibria for the protonation reactions of four coumarin derivatives. (B) Protonation energy (relative to C29) was calculated using DFT with B3LYP 6-31G(d) basis set. (C) Observed binding affinity of the four compounds.

The 3D shape descriptors also strongly correlate with the binding affinity (Table 1). For example, NPR1 and NPR2 are numeric shape descriptors with values between 0 and 1 that characterize the general three-dimensional geometries of molecules.<sup>72</sup> All compounds in our compound collection exhibit a small NPR1 value (< 0.2) and a large NPR2 value (> 0.85), indicating rod-like molecular structures. This observation is consistent with a prior cheminformatic analysis of diverse RNA-binding molecules.<sup>53</sup> In addition, the positive contribution of the shape descriptor *std\_dim1* indicates that a longer molecule makes the ligand more favorable for binding, which is consistent with our expectations for groove binders.

Finally, we observed a positive correlation between planarity and binding affinity, as indicated by the inverse relationship between the dihedral angle of rings BC and DE ( $\omega(\text{BC-DE})$ ) and the natural logarithm of the binding constant ( $\text{Ln}K_d$ ). In C30, the dihedral angle between the BC-DE ring is  $\sim 0$ , making it a planar molecule (Extended Data Fig. 7), which facilitates groove binding. However, adding a methyl group on ring C of C30 (C30-Me) causes steric hindrance between the methyl group and the lone pair electron of the imidazole nitrogen, disrupting the planarity of the molecule, rendering it a poor binder (Fig. 7A and Extended Data Fig. 7). We also tested the role of this methyl group on ring B (C30-Me<sup>RingB</sup>), where it no longer sterically clashes with the imidazole ring. As expected, C30-Me<sup>RingB</sup> is planar in its most favorable conformation (Extended Data Fig. 7), and the binding affinity was comparable to that of C30 (Fig. 7A). Planarity might also contribute to the high binding affinity of C34 ( $K_d = 0.10 \pm 0.01 \mu\text{M}$  to RNA1). In the second protonation site of C34, the imidazole ring can form an internal hydrogen bonding with the coumarin lactone, further stabilizing the planar conformation (Fig. 7B and Extended Data Fig. 7).



**Fig. 7.** (A) Electronic clashes can be avoided by flipping the DE ring or forming an internal hydrogen bond between Rings C and D to maintain coplanarity. (B) C34 maintains planarity, favouring the in vitro interaction with Bulge G RNA.

## Conclusion

In this study, we have reported a new group of coumarin derivatives that exhibit selective binding to bulge G RNA. Using all-atom GaMD simulations, NMR, and QSAR studies, we have identified critical interactions that permit minor groove binding as well as crucial molecular properties of the ligands that significantly contribute to their binding affinity to bulge G RNAs. These factors include the shape and charge of the molecules. The minor groove ligand-RNA binding interface was validated by <sup>1</sup>H SOFAIR NMR experiments that can rapidly characterize the RNA binding behavior. Our research represents a significant advancement in the understanding of RNA-small molecule interactions and offers a versatile platform for developing RNA binders with low nanomolar affinity.

## References

- Butler, A. A., Webb, W. M. & Lubin, F. D. Regulatory RNAs and control of epigenetic mechanisms: Expectations for cognition and cognitive dysfunction. *Epigenomics* vol. 8 135–151 at <https://doi.org/10.2217/epi.15.79> (2016).
- Wilkinson, E., Cui, Y. H. & He, Y. Y. Roles of RNA Modifications in Diverse Cellular Functions. *Frontiers in Cell and Developmental Biology* vol. 10 at <https://doi.org/10.3389/fcell.2022.828683> (2022).
- Meyer, S. M. *et al.* Small molecule recognition of disease-relevant RNA structures. *Chemical Society Reviews* vol. 49 7167–7199 at <https://doi.org/10.1039/d0cs00560f> (2020).
- Ursu, A. *et al.* Design of small molecules targeting RNA structure from sequence. *Chemical Society Reviews* vol. 49 7252–7270 at <https://doi.org/10.1039/d0cs00455c> (2020).
- Childs-Disney, J. L. *et al.* Targeting RNA structures with small molecules. *Nat. Rev. Drug Discov.* **21**, 736–762 (2022).
- Donlic, A. & Hargrove, A. E. Targeting RNA in mammalian systems with small molecules. *Wiley Interdiscip. Rev. RNA* **9**, e1477 (2018).
- Hargrove, A. E. Small molecule-RNA targeting: Starting with the fundamentals. *Chem. Commun.* **56**, 14744–14756 (2020).
- Connelly, C. M., Moon, M. H. & Schneekloth, J. S. The Emerging

- Role of RNA as a Therapeutic Target for Small Molecules. *Cell Chem. Biol.* **23**, 1077–1090 (2016).
9. Warner, K. D., Hajdin, C. E. & Weeks, K. M. Principles for targeting RNA with drug-like small molecules. *Nat. Rev. Drug Discov.* **17**, 547–558 (2018).
10. Walsh, C. & Wencewicz, T. *Antibiotics: Challenges, Mechanisms, Opportunities (Edition 2)*. (Wiley, 2016).
11. Naryshkin, N. A. *et al.* SMN2 splicing modifiers improve motor function and longevity in mice with spinal muscular atrophy. *Science* **345**, 688–693 (2014).
12. Palacino, J. *et al.* SMN2 splice modulators enhance U1-pre-mRNA association and rescue SMA mice. *Nat. Chem. Biol.* **11**, 511–517 (2015).
13. Woll, M. G. *et al.* Discovery and Optimization of Small Molecule Splicing Modifiers of Survival Motor Neuron 2 as a Treatment for Spinal Muscular Atrophy. *J. Med. Chem.* **59**, 6070–6085 (2016).
14. Ratni, H. *et al.* 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.* **59**, 6086–6100 (2016).
15. Pinard, E. *et al.* Discovery of a Novel Class of Survival Motor Neuron 2 Splicing Modifiers for the Treatment of Spinal Muscular Atrophy. *J. Med. Chem.* **60**, 4444–4457 (2017).
16. Ratni, H. *et al.* Discovery of Risdiplam, a Selective Survival of Motor Neuron-2 ( SMN2 ) Gene Splicing Modifier for the Treatment of Spinal Muscular Atrophy (SMA). *J. Med. Chem.* **61**, 6501–6517 (2018).
17. Cheung, A. K. *et al.* Discovery of Small Molecule Splicing Modulators of Survival Motor Neuron-2 (SMN2) for the Treatment of Spinal Muscular Atrophy (SMA). *J. Med. Chem.* **61**, 11021–11036 (2018).
18. Krach, F. *et al.* An alternative splicing modulator decreases mutant HTT and improves the molecular fingerprint in Huntington's disease patient neurons. *Nat. Commun.* **13**, 6797 (2022).
19. Bhattacharyya, A. *et al.* Small molecule splicing modifiers with systemic HTT-lowering activity. *Nat. Commun.* **12**, 7299 (2021).
20. Keller, C. G. *et al.* An orally available, brain penetrant, small molecule lowers huntingtin levels by enhancing pseudoexon inclusion. *Nat. Commun.* **13**, 1150 (2022).
21. Tang, Z. *et al.* Recognition of single-stranded nucleic acids by small-molecule splicing modulators. *Nucleic Acids Res.* **49**, 7870–7883. PMC8373063 (2021).
22. Blount, K. F. & Breaker, R. R. Riboswitches as antibacterial drug targets. *Nat. Biotechnol.* **24**, 1558–64 (2006).
23. Howe, J. A. *et al.* Selective small-molecule inhibition of an RNA structural element. *Nature* **526**, 672–677 (2015).
24. Tran, B. *et al.* Parallel Discovery Strategies Provide a Basis for Riboswitch Ligand Design. *Cell Chem. Biol.* **27**, 1241–1249.e4 (2020).
25. Fedorova, O. *et al.* Small molecules that target group II introns are potent antifungal agents. *Nat. Chem. Biol.* **14**, 1073–1078 (2018).
26. Gumireddy, K. *et al.* Small-molecule inhibitors of microRNA miR-21 function. *Angew. Chem. Int. Ed. Engl.* **47**, 7482–4 (2008).
27. Suresh, B. M. *et al.* A general fragment-based approach to identify and optimize bioactive ligands targeting RNA. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 33197–33203 (2020).
28. Li, Y. & Disney, M. D. Precise Small Molecule Degradation of a Noncoding RNA Identifies Cellular Binding Sites and Modulates an Oncogenic Phenotype. *ACS Chem. Biol.* **13**, 3065–3071 (2018).
29. Tong, Y. *et al.* Programming inactive RNA-binding small molecules into bioactive degraders. *Nature* **618**, 169–179 (2023).
30. Lorenz, D. A. *et al.* Expansion of cat-ELCCA for the Discovery of Small Molecule Inhibitors of the Pre-let-7-Lin28 RNA-Protein Interaction. *ACS Med. Chem. Lett.* **9**, 517–521 (2018).
31. Borgelt, L. *et al.* Trisubstituted Pyrrolinones as Small-Molecule Inhibitors Disrupting the Protein-RNA Interaction of LIN28 and Let-7. *ACS Med. Chem. Lett.* **12**, 893–898 (2021).
32. Zhang, P. *et al.* Translation of the intrinsically disordered protein  $\alpha$ -synuclein is inhibited by a small molecule targeting its structured mRNA. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 1457–1467 (2020).
33. Angelbello, A. J. *et al.* Precise small-molecule cleavage of an r(CUG) repeat expansion in a myotonic dystrophy mouse model. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 7799–7804 (2019).
34. Bush, J. A. *et al.* Ribonuclease recruitment using a small molecule reduced c9ALS/FTD r(G4C2) repeat expansion in vitro and in vivo ALS models. *Sci. Transl. Med.* **13**, eabd5991 (2021).
35. Donlic, A. *et al.* Discovery of Small Molecule Ligands for MALAT1 by Tuning an RNA-Binding Scaffold. *Angew. Chem. Int. Ed. Engl.* **57**, 13242–13247 (2018).
36. Donlic, A., Zafferani, M., Padroni, G., Puri, M. & Hargrove, A. E. Regulation of MALAT1 triple helix stability and in vitro degradation by diphenylfurans. *Nucleic Acids Res.* **48**, 7653–7664 (2020).
37. Rakheja, I., Ansari, A. H., Ray, A., Chandra Joshi, D. & Maiti, S. Small molecule quercetin binds MALAT1 triplex and modulates its cellular function. *Mol. Ther. Nucleic Acids* **30**, 241–256 (2022).
38. Seth, P. P. *et al.* SAR by MS: Discovery of a New Class of RNA-Binding Small Molecules for the Hepatitis C Virus: Internal Ribosome Entry Site IIA Subdomain. *J. Med. Chem.* **48**, 7099–7102 (2005).
39. Parsons, J. *et al.* Conformational inhibition of the hepatitis C virus internal ribosome entry site RNA. *Nat. Chem. Biol.* **5**, 823–5 (2009).
40. Wang, W. *et al.* Hepatitis C viral IRES inhibition by phenazine and phenazine-like molecules. *Bioorg. Med. Chem. Lett.* **10**, 1151–4 (2000).
41. Dibrov, S. M. *et al.* Hepatitis C Virus Translation Inhibitors Targeting the Internal Ribosomal Entry Site. *J. Med. Chem.* **57**, 1694–1707 (2014).
42. Sztuba-Solinska, J. *et al.* Identification of biologically active, HIV TAR RNA-binding small molecules using small molecule microarrays. *J. Am. Chem. Soc.* **136**, 8402–10 (2014).
43. Abulwerdi, F. A. *et al.* Development of Small Molecules with a Noncanonical Binding Mode to HIV-1 Trans Activation Response (TAR) RNA. *J. Med. Chem.* **59**, 11148–11160 (2016).
44. Davidson, A. *et al.* Simultaneous recognition of HIV-1 TAR RNA bulge and loop sequences by cyclic peptide mimics of Tat protein. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 11931–6 (2009).
45. Zhao, J., Qiu, J., Aryal, S., Hackett, J. L. & Wang, J. The RNA Architecture of the SARS-CoV-2 3'-Untranslated Region. *Viruses* **12**, 1473 (2020).
46. Tang, Z. *et al.* Chemical-guided SHAPE sequencing (cgSHAPE-seq) informs the binding site of RNA-degrading chimeras targeting SARS-CoV-2 5' untranslated region. *bioRxiv Prepr.* (2023) doi:10.1101/2023.04.03.535453.
47. Haniff, H. S. *et al.* Targeting the SARS-CoV-2 RNA Genome with

- Small Molecule Binders and Ribonuclease Targeting Chimera (RIBOTAC) Degraders. *ACS Cent. Sci.* **6**, 1713–1721 (2020).
48. Sun, Y. *et al.* Restriction of SARS-CoV-2 replication by targeting programmed -1 ribosomal frameshifting. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
49. Zafferani, M. *et al.* Amilorides inhibit SARS-CoV-2 replication in vitro by targeting RNA structures. *Sci. Adv.* **7**, eabl6096 (2021).
50. Mikutis, S. *et al.* Proximity-Induced Nucleic Acid Degradation (PINAD) Approach to Targeted RNA Degradation Using Small Molecules. *ACS Cent. Sci.* **9**, 892–904 (2023).
51. Bell. 2018\_Pnas\_Si\_Spe. *Proc. Natl. Acad. Sci.* **120**, 2017 (2017).
52. Velagapudi, S. P., Seedhouse, S. J., French, J. & Disney, M. D. Defining the RNA internal loops preferred by benzimidazole derivatives via 2D combinatorial screening and computational analysis. *J. Am. Chem. Soc.* **133**, 10111–8 (2011).
53. 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. Chemie Int. Ed.* **56**, 13498–13502 (2017).
54. Eubanks, C. S., Forte, J. E., Kapral, G. J. & Hargrove, A. E. Small Molecule-Based Pattern Recognition To Classify RNA Structure. *J. Am. Chem. Soc.* **139**, 409–416 (2017).
55. Allen, T. E. H. *et al.* Physicochemical Principles Driving Small Molecule Binding to RNA. *bioRxiv* 2024.01.31.578268 (2024).
56. Yazdani, K. *et al.* Machine Learning Informs RNA-Binding Chemical Space. *Angew. Chem. Int. Ed. Engl.* **62**, e202211358 (2023).
57. Wang, Y., Parmar, S., Schneekloth, J. S. & Tiwary, P. Interrogating RNA-Small Molecule Interactions with Structure Probing and Artificial Intelligence-Augmented Molecular Simulations. *ACS Cent. Sci.* **8**, 741–748 (2022).
58. Akhter, S. *et al.* Mechanism of Ligand Binding to Theophylline RNA Aptamer. *J. Chem. Inf. Model.* (2024) doi:10.1021/acs.jcim.3c01454.
59. Miao, Y., Feher, V. A. & McCammon, J. A. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *J. Chem. Theory Comput.* **11**, 3584–3595 (2015).
60. Wang, J. *et al.* Gaussian accelerated molecular dynamics: Principles and applications. *WIREs Comput. Mol. Sci.* e1521 (2021) doi:10.1002/wcms.1521.
61. Miao, Y. & McCammon, J. A. Gaussian Accelerated Molecular Dynamics: Theory, Implementation, and Applications. *Annu. Rep. Comput. Chem.* **13**, 231–278 (2017).
62. Vuković, L. V., Koh, H. R., Myong, S. & Schulten, K. Substrate Recognition and Specificity of Double-Stranded RNA Binding Proteins. (2014) doi:10.1021/bi500352s.
63. Koirala, D. *et al.* Affinity maturation of a portable Fab-RNA module for chaperone-assisted RNA crystallography. *Nucleic Acids Res.* **46**, 2624–2635 (2018).
64. Wacker, A. *et al.* Secondary structure determination of conserved SARS-CoV-2 RNA elements by NMR spectroscopy. *Nucleic Acids Res.* **48**, 12415–12435 (2020).
65. Xu, X., Dal Poggetto, G., McCoy, M., Reibarkh, M. & Mourino, P. T. Rapid Characterization of Structural and Behavioral Changes of Therapeutic Proteins by Relaxation and Diffusion 1H-SOFAST NMR Experiments. *Anal. Chem.* **In press**, (2024).
66. Vise, P. D., Baral, B., Latos, A. J. & Daughdrill, G. W. NMR chemical shift and relaxation measurements provide evidence for the coupled folding and binding of the p53 transactivation domain. *Nucleic Acids Res.* **33**, 2061–2077 (2005).
67. Walinda, E., Morimoto, D. & Sugase, K. Resolving biomolecular motion and interactions by R2 and R1ρ relaxation dispersion NMR. *Methods* **148**, 28–38 (2018).
68. Carr, H. Y. & Purcell, E. M. Effects of Diffusion on Free Precession in Nuclear Magnetic Resonance Experiments. *Phys. Rev.* **94**, 630–638 (1954).
69. Meiboom, S. & Gill, D. Modified Spin-Echo Method for Measuring Nuclear Relaxation Times. *Rev. Sci. Instrum.* **29**, 688–691 (1958).
70. Faber, C., Sticht, H., Schweimer, K. & Rösch, P. Structural rearrangements of HIV-1 Tat-responsive RNA upon binding of neomycin B. *J. Biol. Chem.* **275**, 20660–6 (2000).
71. Cai, Z., Zafferani, M., Akande, O. M. & Hargrove, A. E. Quantitative Structure-Activity Relationship (QSAR) Study Predicts Small-Molecule Binding to RNA Structure. *J. Med. Chem.* **65**, 7262–7277 (2022).
72. Sauer, W. H. B. & Schwarz, M. K. Molecular shape diversity of combinatorial libraries: A prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci.* **43**, 987–1003 (2003).
73. Xu, X., Poggetto, G. D., McCoy, M., Reibarkh, M. & Trigo, P. Rapid Characterization of Structural and Behavioral Changes of Therapeutic Proteins by Relaxation and Diffusion 1H-SOFAST NMR Experiments. (2024).

## Methods

### Fluorescence polarization assay

Synthetic RNA oligomers were procured from GenScript (Piscataway, NJ, USA) and reconstituted in nuclease-free water. Compounds were prepared at a concentration of 50 μM in DMSO and were diluted in 2× assay buffer (50 mM MES, 100 mM NaCl, 0.004% TritonX, pH 6.1) to a concentration of 0.2 μM. A 1:3 dilution series (6 points) of each RNA was then prepared in 20–30 μL water, resulting in concentrations ranging from 0.1–20 μM. Subsequently, 20 μL of 2× working solution containing the assay buffer and small-molecule ligand was added to each RNA sample in a 1:1 (v/v) ratio and mixed by pipetting. To measure fluorescence polarization, 8 μL of the resulting 1× sample solution was transferred into a 384-well, black, flat-bottom microplate (Greiner, #784076) in duplicates or triplicates. The plate was equilibrated at room temperature for 5 minutes prior to data acquisition using BioTek Synergy H1 (Winooski, VT, USA) with an excitation/emission of 360/460 nm at 25 °C. Experimental data were analyzed using the Prism 8.0 software package (GraphPad, San Diego, USA). A nonlinear curve fitting was employed to calculate the dissociation constant ( $K_d$ ), reported with a 95% confidence interval.

### Lasso regression

Lasso regression. The structure of each of the 69 compounds was individually optimized by DFT calculation with B3LYP/6-31G(d) basis set (ground state). The protonation state was predicted by MOE 2022 software (Chemical Computing Group, Montreal, Canada) at pH 7.0. The structures of all compounds were loaded on MOE 2022 and 443 molecular descriptors were generated



using MOE 2022. Two molecular descriptors, h\_pKa and h\_pKb were excluded because the protonation states of the compound library are different. The dihedrals between BC and DE rings within the optimized 3-dimensional structure in the unit of degree (°) were added as a new molecular descriptor. In the lasso regression analysis, the natural logarithm of the dissociation constant ( $\ln(K_d)$ ,  $K_d$  in molar unit), expressed in molar units, was utilized as the dependent variable (y-value). (for final values of the molecular descriptors used for Lasso regression, see MolecularDescriptors.csv). Lasso regression (L1 regularization) was performed in R (4.2.2) according to the reported protocol.<sup>71</sup> The non-zero coefficients were determined as described in **Supplementary Table 5**.

### Gaussian accelerated Molecular Dynamics (GaMD) Methodology

GaMD is an enhanced sampling technique in which a harmonic boost potential is added to smooth the potential energy surface and reduce the system energy barriers.<sup>59</sup> GaMD can accelerate biomolecular simulations by order of magnitude and eliminates the need for predefined collective variables. Moreover, because GaMD boosts potential following a Gaussian distribution, it helps to properly recover biomolecular free energy profiles through cumulant expansion to the second order.<sup>59</sup> A brief description of the method is described here.

Given a system with  $N$  atoms positioned at a specific location,  $r \equiv \{\vec{r}_1, \dots, \vec{r}_N\}$ , a boost potential  $\Delta V(\vec{r})$  is added when the system potential energy  $V(\vec{r})$  drops below the threshold energy  $E$ , the modified system potential  $V^*(\vec{r})$  is calculated as:

$$V^*(\vec{r}) = V(\vec{r}) + \Delta V(\vec{r}), V(\vec{r}) < E \quad (1)$$

$$\Delta V(\vec{r}) = \frac{1}{2}k(E - V(\vec{r}))^2, V(\vec{r}) < E \quad (2)$$

Where variable  $k$  represents the harmonic force constant. Two criteria must be met by the boost potential  $\Delta V(\vec{r})$ . First, for any two arbitrary potential values  $V_1(\vec{r})$  and  $V_2(\vec{r})$  found on the original energy surface, if  $V_1(\vec{r}) < V_2(\vec{r})$ ,  $\Delta V$  should be a monotonic function that does not change the relative order of the biased potential values ( $V_1^*(\vec{r}) < V_2^*(\vec{r})$ ). Second, if  $V_1(\vec{r}) < V_2(\vec{r})$ , the potential difference observed on the smoothed energy surface should be smaller than that of the original ( $V_2^*(\vec{r}) - V_1^*(\vec{r}) < V_2(\vec{r}) - V_1(\vec{r})$ ). By combining the first two criteria and plugging in 1 and 2, we obtain:

$$V_{max} \leq E \leq V_{min} + \frac{1}{k} \quad (3)$$

where  $V_{min}$  and  $V_{max}$  are the system minimum and maximum potential energies and  $k$  satisfies:  $k \leq \frac{1}{V_{max} - V_{min}}$ . If we define  $k \equiv \frac{k_0}{V_{max} - V_{min}}$ , then  $0 < k_0 \leq 1$ . The greater the  $k_0$  value is, the higher the boost potential  $\Delta V(\vec{r})$  is added to the potential energy surface. Third, the standard deviation of  $\Delta V$  should be small (i.e., narrow distribution) for accurate reweighting using cumulant expansion to the second order:

$$\sigma_{\Delta V} = k(E - V_{av})\sigma_V \leq \sigma_0 \quad (4)$$

in which  $V_{av}$  and  $\sigma_V$  are the average and standard deviation of the system potential energies and  $\sigma_{\Delta V}$  is the standard deviation of  $\Delta V$  with  $\sigma_0$  as a user-specified upper limit ( $10k_B T$ ) for precise reweighting. Eq. 3 states that, when  $E$  is set to the lower bound  $E = V_{max}$ ,  $k_0$  can be calculated as:

$$k_0 = \min(1.0, k'_0) = \min\left(1.0, \frac{\sigma_0}{\sigma_V} \cdot \frac{V_{max} - V_{min}}{V_{max} - V_{av}}\right) \quad (5)$$

On the other hand, when the threshold energy  $E$  is set to its upper bound  $= V_{min} + \frac{1}{k}$ ,  $k_0$  is set to:

$$k_0 = k_0'' \equiv \left(1.0 - \frac{\sigma_0}{\sigma_V}\right) \cdot \frac{V_{max} - V_{min}}{V_{av} - V_{min}} \quad (6)$$

if  $k_0''$  is calculated between 0 and 1. Otherwise,  $k_0$  is calculated using Eq 5.

The GaMD method provides options to add only the total potential boost  $\Delta V_p$ , only the dihedral potential boost  $\Delta V_D$ , or the dual boost potential (both  $\Delta V_p$  and  $\Delta V_D$ ). Among these, the dual-boost GaMD (GaMD\_Dual) mode provides the highest acceleration for enhanced sampling of simulations.

The simulation parameters comprise the threshold energy  $E$  for applying boost potential and the effective harmonic force constants,  $k_{0P}$  and  $k_{0D}$  for the total and dihedral boost potential, respectively.

Example input parameters used in dual-boost GaMD simulations include the following in addition to those used in conventional MD: igamd = 3, iE = 1, irst\_gamd = 0, ntcmd= 1500000, nteb = 30000000, ntave = 300000, ntcmdprep = 600000, ntebprep = 600000, sigma0P = 6.0, sigma0D = 6.0

### Energetic Reweighting of GaMD simulations

In biomolecular systems, the probability distribution along a selected reaction coordinate  $A(r)$  is represented as  $p^*(A)$ , where  $r$  is representative of the atomic locations. The canonical ensemble distribution,  $p(A)$ , can be recovered by reweighting  $p^*(A)$  using each frame's boost potential,  $\Delta V(r)$ .

$$p(A_j) = p^*(A_j) \frac{\langle e^{\beta \Delta V(\vec{r})} \rangle_j}{\sum_{i=1}^M \langle e^{\beta \Delta V(\vec{r})} \rangle_i} \quad (7)$$

where  $M$  is the number of bins,  $\beta = k_B T$  and  $\langle e^{\beta \Delta V(\vec{r})} \rangle_j$  is the ensemble-averaged Boltzmann factor of  $\Delta V(\vec{r})$  for the simulation frames found in the  $j^{\text{th}}$  bin. To approximate the ensemble-averaged reweighting factor, one can use the cumulant expansion method:

$$\langle e^{\beta \Delta V(\vec{r})} \rangle = \exp\left\{\sum_{k=1}^{\infty} \frac{\beta^k}{k!} C_k\right\}, \quad (8)$$

where the first two cumulants are given by:

$$C_1 = \langle \Delta V \rangle, \quad (9)$$

$$C_2 = \langle \Delta V^2 \rangle - \langle \Delta V \rangle^2 = \sigma_{\Delta V}^2.$$

The boost potential obtained from GaMD simulations shows a near-Gaussian distribution. Cumulant expansion to the second order thus provides a good approximation for computing the reweighting factor. The reweighted free energy  $F(A) = -k_B T \ln p(A)$  is calculated as:

$$F(A) = F^*(A) - \sum_{k=1}^2 \frac{\beta^k}{k!} C_k + F_c, \quad (10)$$

where  $F^*(A) = -k_B T \ln p^*(A)$  is the modified free energy obtained from GaMD simulation and  $F_c$  is a constant.

### NMR experiments

An 84.1 nmol unlabeled RNA5 sample (Sigma-Aldrich) was dissolved in 135  $\mu\text{L}$  of potassium buffer (25 mM potassium

phosphate buffer, 50 mM potassium chloride, pH 6.2, 10% D<sub>2</sub>O) to prepare an ~600  $\mu$ M NMR sample. A similar sample condition was used in a previous report,<sup>64</sup> where NMR assignment was determined for a <sup>15</sup>N-labeled RNA containing the segment of RNA5.

All NMR spectra were acquired using a Bruker 800 MHz Ascend spectrometer equipped with a TCI cryoprobe at 298 K (**Extended Data Fig. 8**). RNA proton peak assignment was performed by comparing the measured <sup>1</sup>H chemical shifts with literature values (**Supplementary Table 3**).<sup>64</sup> Proton peaks were assigned if the chemical shift difference was less than 0.15 ppm. All NH protons of U and G residues that showed peaks in the 10-14 ppm region were assigned accordingly, except for the one from G24, where the literature assignment was missing. Two proton peaks observed in this chemical shift region that were not previously assigned should correspond to the NH of G24 as well as that of G13. G13 is a part of the linker that differs from the RNA sequence used in the literature. The proton peaks of G7-NH and U25-NH did not appear until the addition of a 70  $\mu$ M ligand. All the resonances that were unambiguously assigned are summarized in **Supplementary Table 3**.

The ligand stock solution prepared for the titration experiment contained 10 mM C30 in DMSO-d<sub>6</sub>. 0.47  $\mu$ L, 0.47  $\mu$ L, 0.94  $\mu$ L, 1.88  $\mu$ L, 0.94  $\mu$ L, 1.88  $\mu$ L, and 2.82  $\mu$ L of the stock solution were titrated into the RNA NMR sample to achieve final ligand concentrations of 35  $\mu$ M, 70  $\mu$ M, 140  $\mu$ M, 280  $\mu$ M, 350  $\mu$ M, 490  $\mu$ M, and 700  $\mu$ M, respectively. Proton transverse relaxation rate  $R_2$  was measured from SOFAIR (band-selective optimized flip-angle internally-encoded relaxation) (**Extended Data Fig. 9**).<sup>73</sup> The band-selective excitation pulse p39 was centered at 12.1 ppm with a bandwidth of 5.2 ppm for NH region, and was centered at 7.6 ppm with a bandwidth of 3 ppm for NH<sub>2</sub>/aromatic region. Transverse relaxation was encoded through the incrementation of a delay  $t$  flanking the refocusing pulse p40. The delay time was set to between 0 to 0.4 s with a total of 12 increments (0, 0.002, 0.005, 0.010, 0.015, 0.020, 0.025, 0.030, 0.050, 0.1, 0.2, 0.4 s). The duration of each experiment is about 18.5 minutes. Data were processed and analyzed using MestReNova.  $R_2$  of each resonance was determined through area integration and fitting the integrals to the following equation:  $I = I_0 e^{-tR_2} + B$ , where  $t$  is the delay time,  $R_2$  is the transverse relaxation rate, and  $B$  is a constant to account for any baseline differences between experiments.

#### Data Availability Statement

The representative C30-RNA5 bound conformation generated by GaMD simulations is available in PDB format in the Model Archive repository (<https://modelarchive.org>) under project ma-q6hl4. The Representative apo RNA crystal structure of RNA1 is available in PDB under accession number 9DN4.

#### Supporting Information

Supplementary figures and tables, experimental and computation methods, and compound characterization data are available in the Supporting Information file. Calculated molecular descriptors (csv format), the code used for lasso regression (R markdown file), optimized structures for the 69-compound collection (compressed

mol2 file), and simulated C30-RNA5 binding pathway (Supplementary Movie 1, mp4 movie) are also available in the Supporting Information.

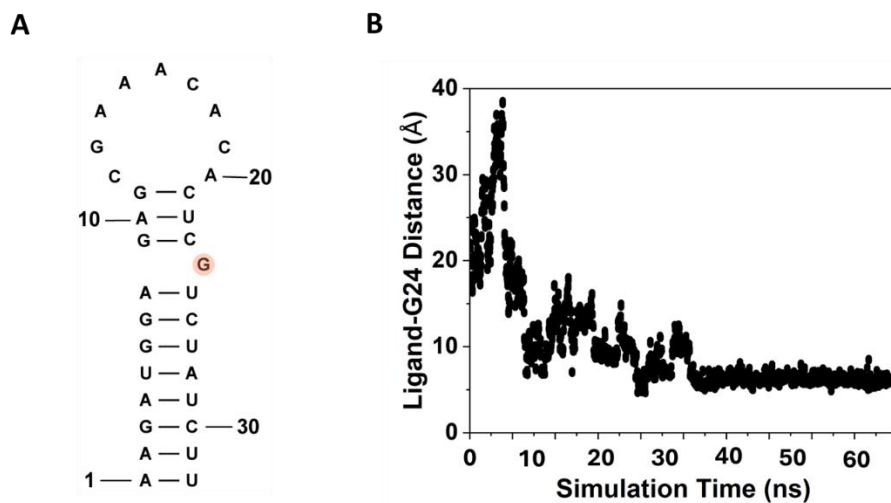
#### Acknowledgements

Research reported in this article was supported by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) under award numbers R35GM147498 (to J.W.) and start-up project 27110 at the University of North Carolina – Chapel Hill (to Y.M.). The authors acknowledge support from the MRL Postdoctoral Research Program, and Xingjian Xu at Merck & Co., Inc., Rahway, NJ, USA, for discussions of NMR measurements. This research used the NYX beamline 19-ID, supported by the New York Structural Biology Center, at the National Synchrotron Light Source II, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory under Contract No. DE-SC0012704. The NYX detector instrumentation was supported by grant S10OD030394 through the Office of the Director of the National Institutes of Health. The Center for Bio-Molecular Structure (CBMS) is primarily supported by the NIH-NIGMS through a Center Core P30 Grant (P30GM133893), and by the DOE Office of Biological and Environmental Research (KP1607011). NSLS2 is a U.S. DOE Office of Science User Facility operated under Contract No. DE-SC0012704. This publication resulted from the data collected using the beamtime obtained through NECAT BAG proposal # 311950.

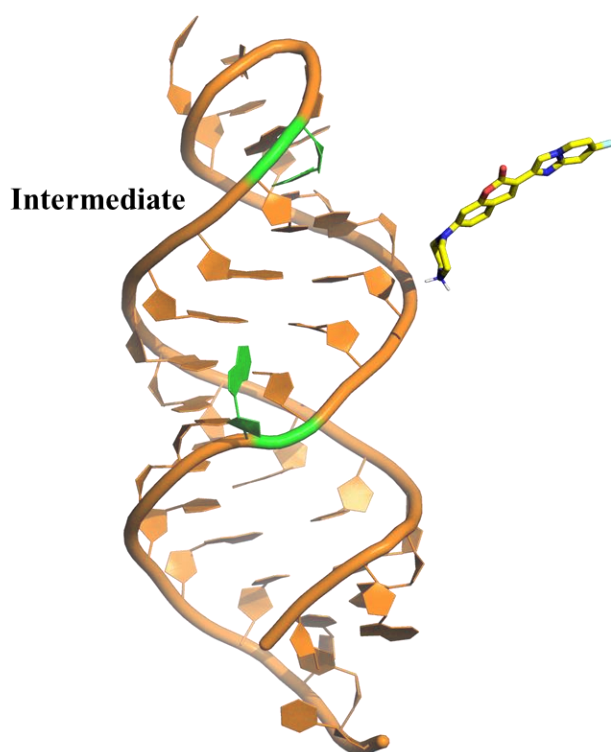
#### Competing interest

The authors declare no competing interests.

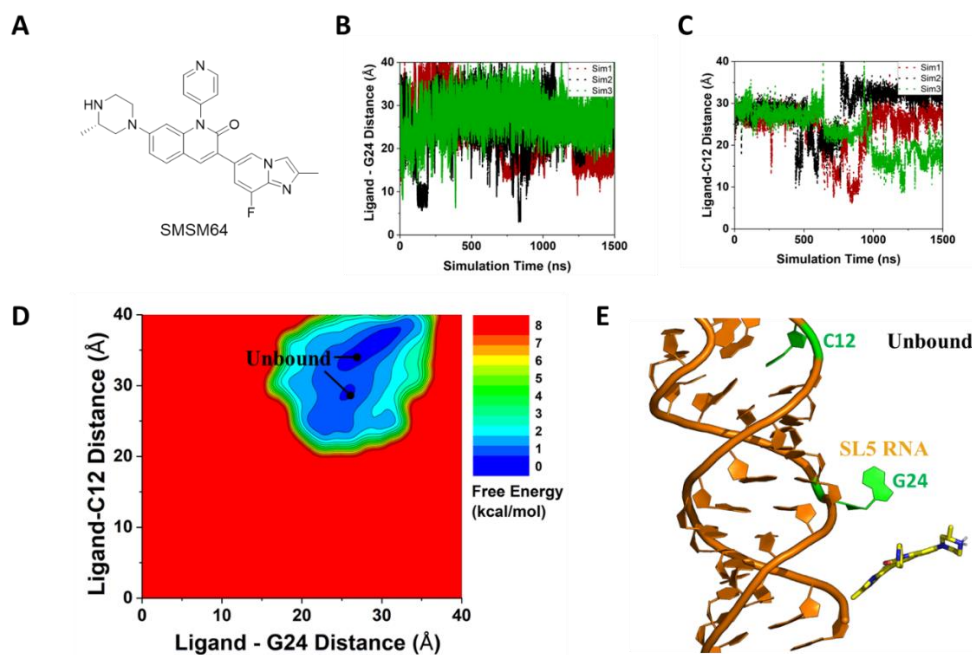
Extended data



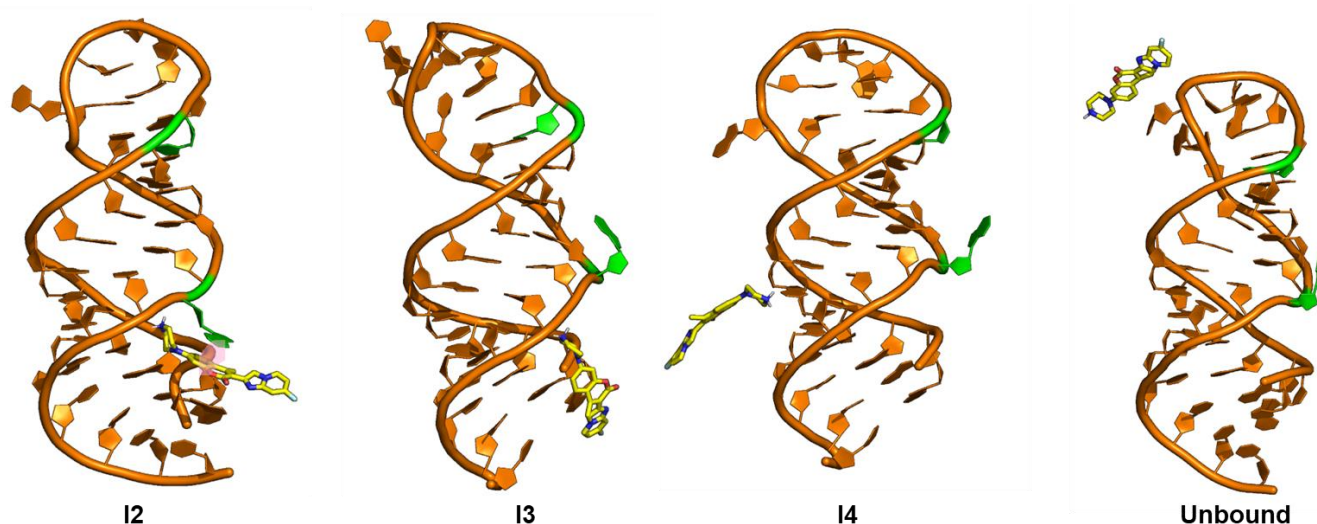
**Extended Data Fig 1: Binding of C30 with RNA5:** (A) Structure of RNA5. (B) Time courses of the center-of-mass distance between the ligand-heavy atoms in the coumarin core to the RNA bulge G (RNA5) calculated from 63 ns GaMD equilibration simulation.



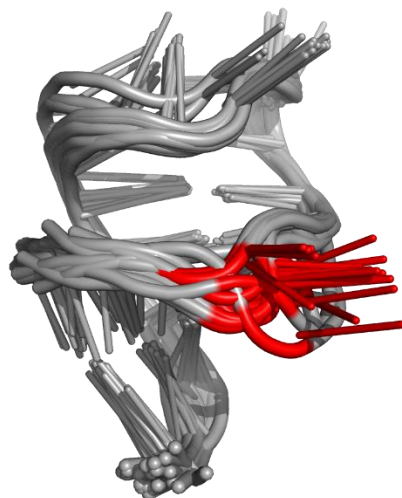
**Extended Data Fig. 2:** The “Intermediate” state of RNA5–C30 binding obtained from the GaMD simulations.



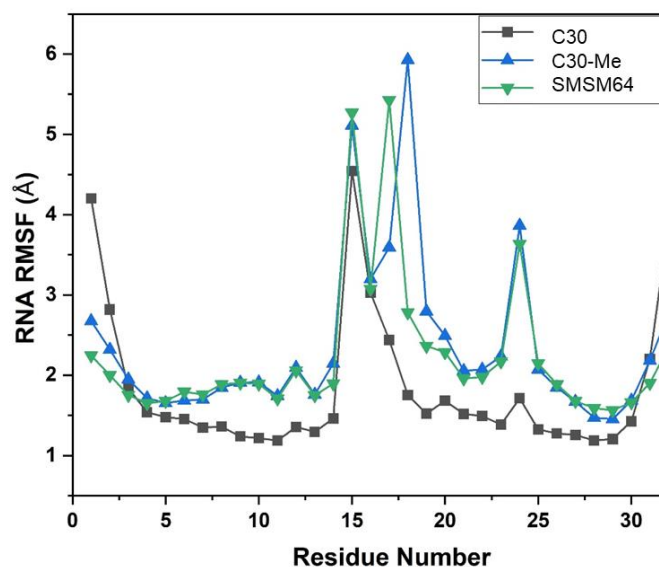
**Extended Data Fig. 3: GaMD simulations captured spontaneous binding of coumarin derivative SMSM64 to RNA5:** (A) Time courses of the center-of-mass distance between ligand heavy atoms in the coumarin core to the RNA bulge G24 calculated from three independent 1.5  $\mu$ s GaMD simulations. (B)  $\pi$ - $\pi$  stacking interaction distance between the fused (D, E) ring and nucleotide C12 tracked as a function of simulation time. (C) 2D free energy profile calculated with all three GaMD simulations combined, in which a single distinct low-energy state was identified, referred as the “Unbound” state. (D) Representative conformation of RNA not bound to SMSM64.



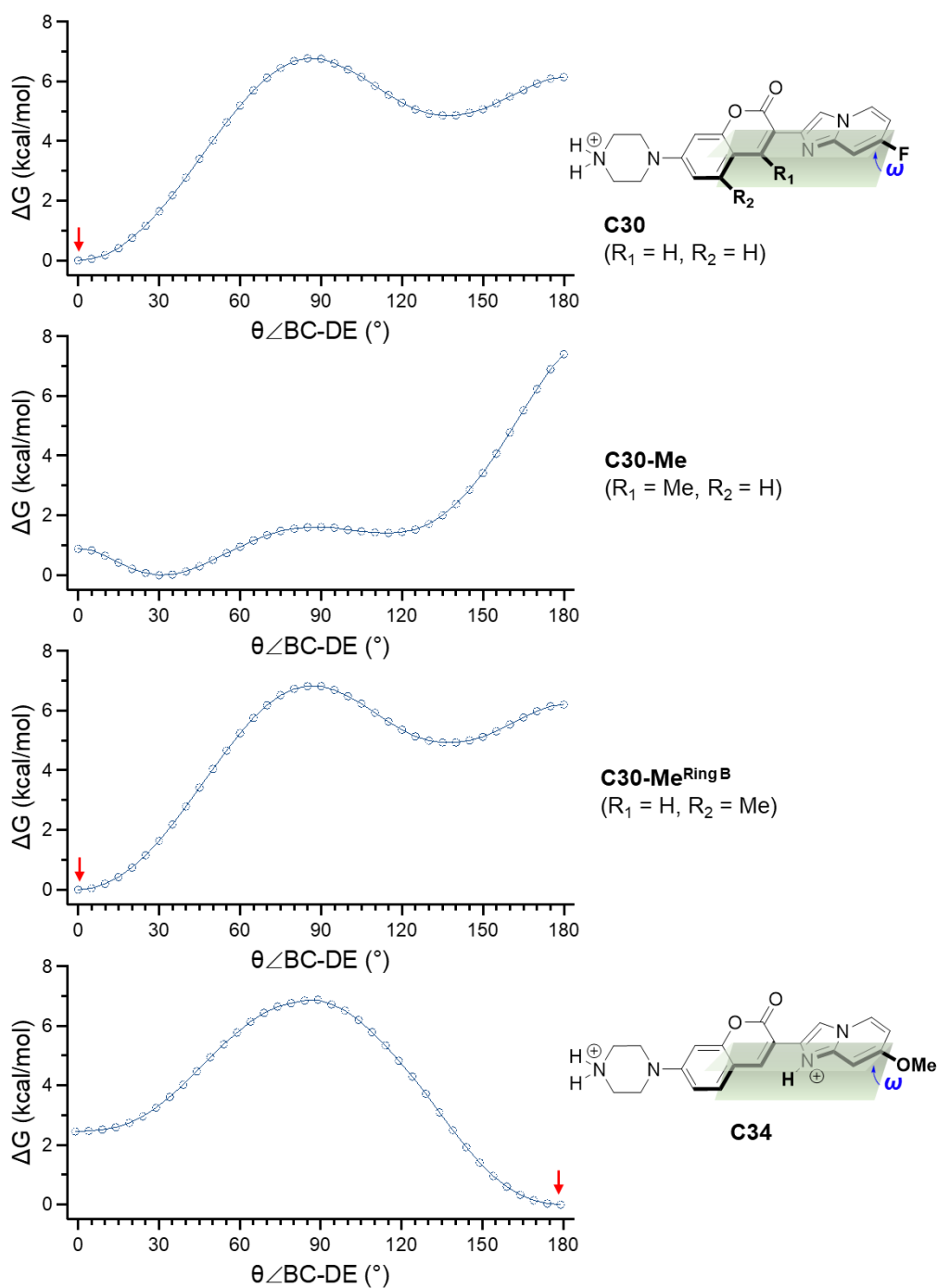
**Extended Data Fig. 4: Intermediate conformational states of C30-Me binding to the RNA5.** Intermediate states as identified from the 2D free energy profile from GaMD simulations, “I1”, “I2”, “I3”, “I4” and “Unbound” states. The RNA is in orange cartoons, and the ligand is in sticks with the C atoms, which are colored yellow. The unpaired G24 and C12 nucleotides are highlighted in light green.



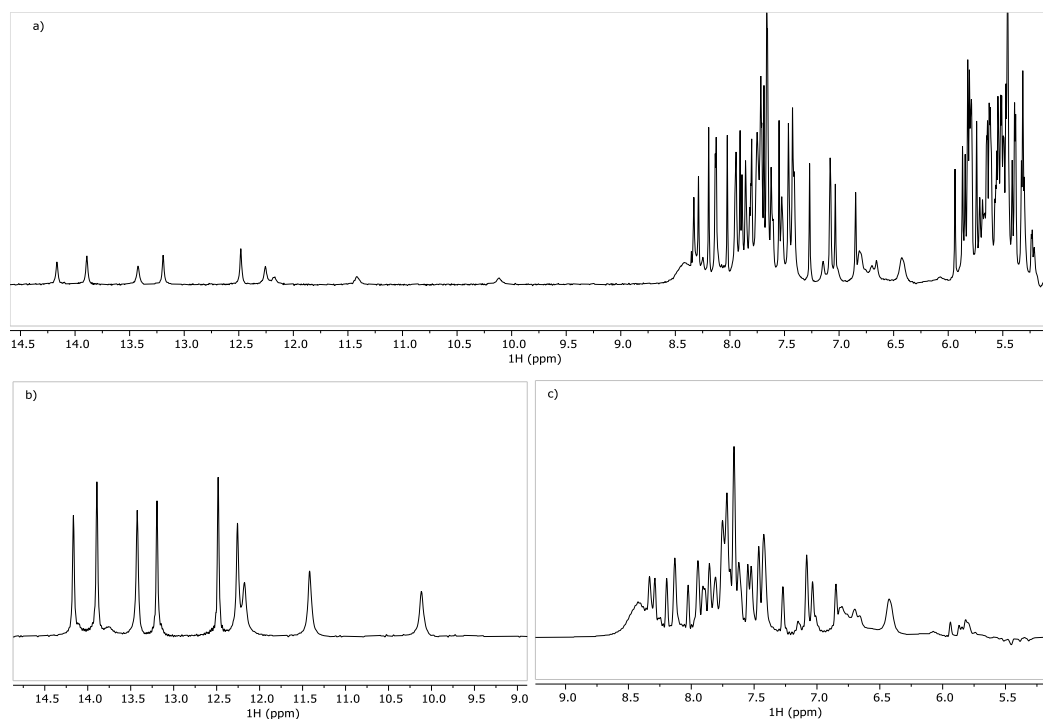
**Extended Data Fig. 5: Chaperone-assisted crystal structure of RNA 1.** Structural superposition of 21 RNA molecules derived from 12 datasets. Among aligned structures, 9 datasets consist of two molecules, and 3 datasets consist of one molecule. All models were refined at least 3 times. Resolution of collected datasets ranges from 1.39 Å to 2.45 Å; completeness from 93 % to 100 %. Structures were solved in triclinic, monoclinic and orthorhombic crystal systems.



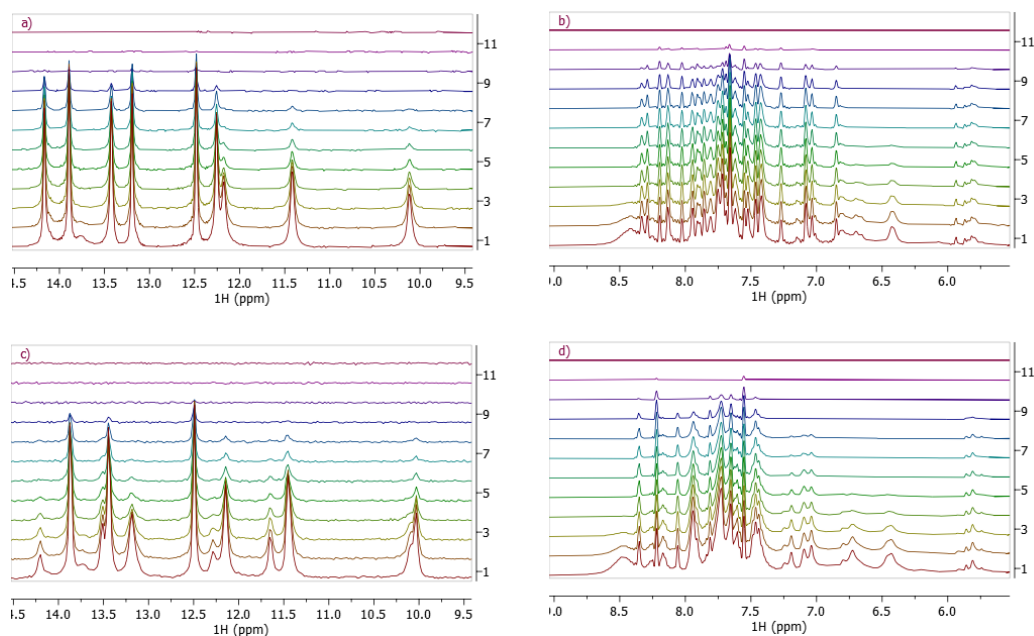
**Extended Data Fig. 6: Fluctuation of the RNA nucleotides across GaMD production simulation in the presence of the coumarin derivatives.** RMSF = Root Mean Square Fluctuation.



**Extended Data Fig. S7: Calculated total energy of coumarin derivatives with a scan of  $\omega(BC-DE)$  dihedral (0–180°).** The calculation was performed using DFT with a B3YLP 6-31(d) basis set. The red arrow indicates the most stable planar structures.



**Extended Data Fig. S8: NMR spectra of RNA5. (A)** 1D  $^1\text{H}$  experiment with solvent suppression. **(B)** The first increment of SOFAIR experiment with a center at 12.1 ppm. **(C)** The first increment of SOFAIR experiment with chemical shift center at 7.6 ppm.



**Extended Data Fig. S9: R2 relaxation measurement of RNA5 using SOFAIR experiments. (A) and (C)** measured R2 relaxation rates in the range of 10-14 ppm without and with 700  $\mu\text{M}$  C30 ligand. **(B) and (D)** measured R2 relaxation rates in the range of 6-8 ppm without and with 700  $\mu\text{M}$  C30 ligand. Vertical spectra were obtained with spin echo intervals of 0, 0.002, 0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.05, 0.1, 0.2, and 0.4 s from bottom to top.