Deep learning challenge to generate *apo* RNA conformations with cryptic binding site

Ikuo Kurisaki[1*], Michiaki Hamada[2,3,4*]

[1]Waseda Research Institute for Science and Engineering, Waseda University, 169-8555, Tokyo, Japan

[2]Faculty of Science and Engineering, Waseda University, 169-8555, Tokyo, Japan

[3]AIST-Waseda University Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), National Institute of Advanced Industrial Science and Technology, 169-8555, Tokyo, Japan

[4]Graduate School of Medicine, Nippon Medical School, 113-8602, Tokyo, Japan

[*]Ikuo Kurisaki:

E-mail: ikuo.kurisaki@aoni.waseda.jp

[*]Michiaki Hamada:

E-mail: mhamada@waseda.jp

Abstract

RNA plays vital roles in diverse biological processes, thus drawing much attention as potential therapeutic target. Notably, cryptic ligand binding sites can be promising targets due to high specificity in RNA-ligand interaction. Structure Based Drug Design (SBDD) for RNA is a practical approach to achieve successful drug candidates. However, RNA-targeted SBDD faces a technical challenge, a paucity of available tertiary structures of RNA-ligand complexes. To address the problem, we present a computational strategy to generate RNA conformations harboring cryptic binding sites from *apo* RNA conformations, by using Molearn, a generative deep learning (DL) model designed to explore high energy states along conformational transition pathways. Focusing on the paradigmatic RNA-ligand complex, HIV-1 Transactivation Response Element (TAR) binding to MV2003, we succeed to generate TAR conformations with a cryptic binding cavity for MV2003, without *a priori* knowledge of the cavity. These conformations were docked to MV2003, yielding poses with binding scores comparable to those calculated for experimentally resolved TAR-MV2003 complexes. Prior atomistic molecular dynamics and DL-based RNA 3D structure prediction studies on this target have not reported recovery of the cryptic site from *apo* conformations. Finally, we discuss the technical advantages, acknowledge current limitations, and outline concrete avenues for improvement. To our knowledge, this is the first report of theoretical generation of RNA conformations with a cryptic ligand binding site. This finding demonstrates a baseline for future studies in this direction and suggests a practical route to advance RNA-targeted SBDD.

Introduction

RNA molecules are involved in a wide variety of biological processes. [1-4] Their disfunctions bring about serious diseases and they have been regarded as important targets for drug discovery[5-8]. Recalling that Structure-Based Drug Design (SBDD) has drawn much attention for the monumental successes in protein-target drug discovery[9-13], utilizing tertiary structures of RNA is similarly a promising approach to systematically discover RNA-targeted drugs among abundant candidate compounds[14]. In the line of SBDD approach, if we could design drugs interacting with a cryptic RNA-ligand binding site, which only emerge in an RNA-ligand complex, such drugs should be underscored due to higher specificity to target RNA and weaker side effects for off-target RNA molecules. Meanwhile, although the number of RNA-ligand complex tertiary structures has increased, it remains insufficient, thus hindering development of RNA-targeting SBDD. RNA sequence information is available in the order of tens of million[15], whereas experimentally derived RNA-ligand complex structures remain limited to roughly 300, and RNAs harboring cryptic binding sites are even more sparsely understood. It thus remains utterly infeasible to routinely perform cryptic binding site-targeted SBDD for any RNA of interest.

Because of paucity of experimentally-resolved RNA structures, theoretical prediction of *apo* RNA harboring cryptic binding site is expected as the initial step to generate RNA-ligand docked poses. However, theoretical approaches have not achieved such tasks so far, even for a paradigmatic RNA to study interaction with ligands, HIV-1 TAR (Humam Immunodeficiency Virus type 1 Trans Activation Response element, referred to as TAR hereafter; **Fig. 1A** and **1B** for the second and tertiary structures, respectively) [16-23]. Among TAR-binding ligands, MV2003 (Arginine 4-methoxy-β-napthylamide) is known as a

small molecule to bind to TAR's cryptic binding cavity[21] (**Fig. 1C** and **1D** for MV2003-bound and *apo* TAR conformations, respectively). There have been several molecular dynamics (MD) simulation studies to try generating TAR conformations competent for MV2003 binding (referred to as 'MV2003-bindable TAR conformations' hereafter) from the *apo* forms. [24-26] However, none of these MD studies succeeded in generation of MV2003-bindable TAR conformations.

Such disappointing results may stem from limitations of fixed-charge (additive) force field (FF) parameters. As in the previous studies[24-26], MD simulations with fixed-charge FF parameters are still widely employed to sample RNA conformations, owing to their strong track record reproducing selected physicochemical observables[27-30], favorable computational cost, and still-limited maturity of polarized FF parameters[31] that have been developed with aim of improving accuracy for treatment of electronic interactions. Meanwhile, it is well appreciated that fixed-charge FF parameters for RNA can misestimate electrostatic contributions[28, 32] and may create an imbalance between base stacking and base-pair hydrogen-bonding interactions in RNA. As for the case of TAR, base pair stackings in neighboring to C30 and A35 may be overly stabilized, thereby impeding emergence of the cryptic binding cavity that requires disruption of the C30-A35 pair (*see* **Fig. 1A**, **Fig. 1C** and **1D**). Unfortunately, there is no clear evidence that merely extending MD-based sampling will, by itself, resolve the issues about delayed conformational changes. For several mainstream RNA force fields, long unbiased or enhanced-sampling MD has at times populated non-native conformations. [33] Because of these two technical problems, MD started from *apo* RNA conformations may not yet constitute a general standalone route to cryptic-site discovery and is likely to be most effective when combined with complementary methods.

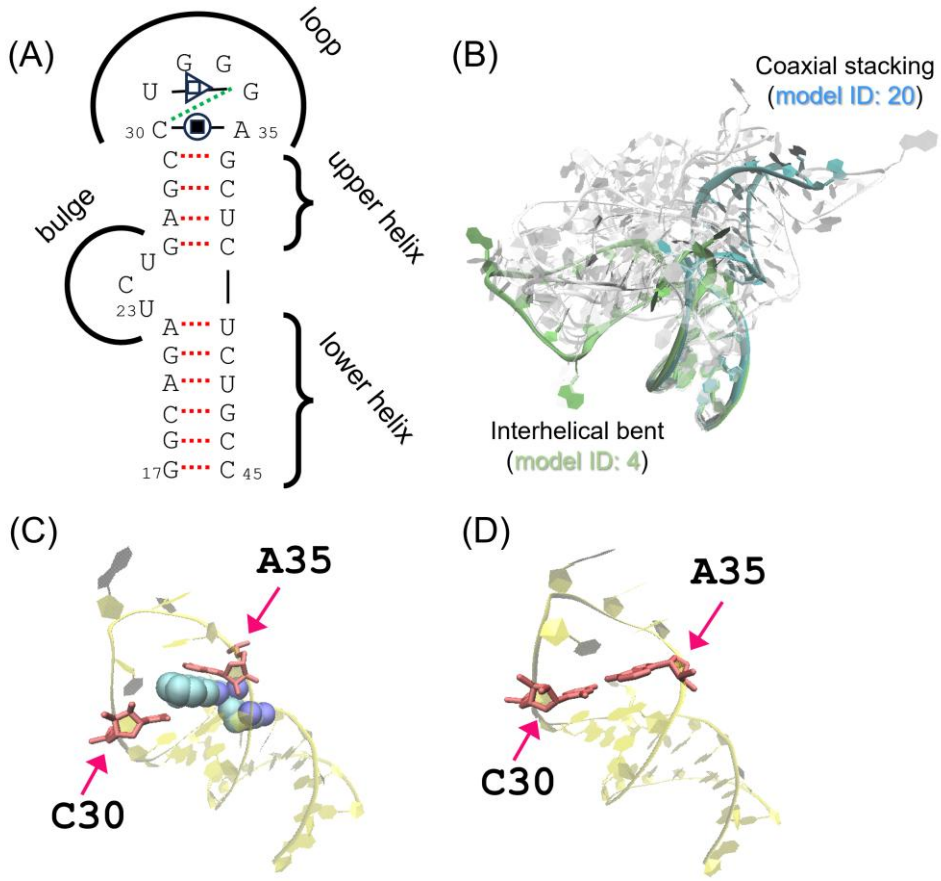

Figure 1. Structures of HIV-1 TAR. (A) Secondary structures and (B) Tertiary structure ensemble in *apo* form (PDB entry: 7JU1). (C) MV2003-bound structure (model No. 1 of PDB entry: 2L8H). (D) MV2003-inaccessible structure (model No. 1 of PDB entry: 7JU1). In panel A, models 4 and 20, a pair with the largest RMSd value of the all pairs ($_{20}C_{19}$ = 190) is simultaneously considered. Dotted lines are for Watson-Crick base pair (WC). WC in green only forms in the model No. 4. Closed square in circle and open square in triangle are for non-WC interactions. In panel B, models 4 and 20 are highlighted by green and blue, respectively. In panels C and D, MV2003 and the interacting residues are shown by van der Waals spheres and red sticks, respectively. HIV-1 TAR structure is illustrated by transparent yellow ribbon.

To overcome the inherent limitations of straightforward MD simulations, we employ a generative deep learning (DL) model, Molearn[34]. Degiacomi and colleagues proposed

this convolutional neural network model to systematically generate a set of intermediate protein conformations along the transition path, capturing the global structural reconfigurations that connect the two endpoint states (**Fig. 2** shows a technical outline; detailed explanation is given in **Materials and Methods**). Notably, the earlier Molearn studies succeeded in generating biologically relevant intermediate conformations of proteins and outperformed other conventional methods in terms of transition pathway prediction[34, 35].

From practical point of view, Molearn models are trained with MD-derived conformational ensembles at equilibrium states, and transition pathways are internally defined on the latent space (*see* **Fig. 2A** and **Fig. 2B**). These features could make Molearn a promising approach to systematically generate ligand-bindable *apo* RNA conformations at a realistic computational cost. A ligand-bindable RNA is expected to adopt relatively high-energy intermediate conformations, due to energetic trade-offs between intra- and inter-molecular interactions in order to minimize the total system energy. Such high-energy RNA conformations would often emerge through global conformational changes along transition pathways. Generally, such conformational changes progress within time period of several milliseconds or longer[36-39] (as for TAR, disruption of hydrogen bonds formed between C30 and A36 followed by rearrangement of the stem loop), thus being inaccessible even for maximally-long non-equilibrium MD simulations of today. By using Molearn, we could obtain transition pathways undergoing global conformational changes with routinely-performed equilibrium MD simulations reachable within realistic computational time.

In this study, we trained this DL model on MD-derived *apo* TAR conformations and sought to generate MV2003-bindable TAR conformations. The original Molearn study

considered protein backbone atoms for conformation generations. [34] Meanwhile, we trained Molearn with full-atomic RNA models (RNA backbone and bases atoms except for hydrogen atoms) with suitable modification of the distributed Molearn programs. Full-atomic RNA representation enables us to quantitatively test ligand-bindable RNA conformations by employing an RNA-ligand docking simulation and an RNA-ligand docking score estimation. In addition, we assessed scalability by applying this DL approach to a larger RNA molecule. Together, these results provide an initial step toward a practical framework to generate RNA conformations with cryptic ligand-binding sites without prior complex structures.

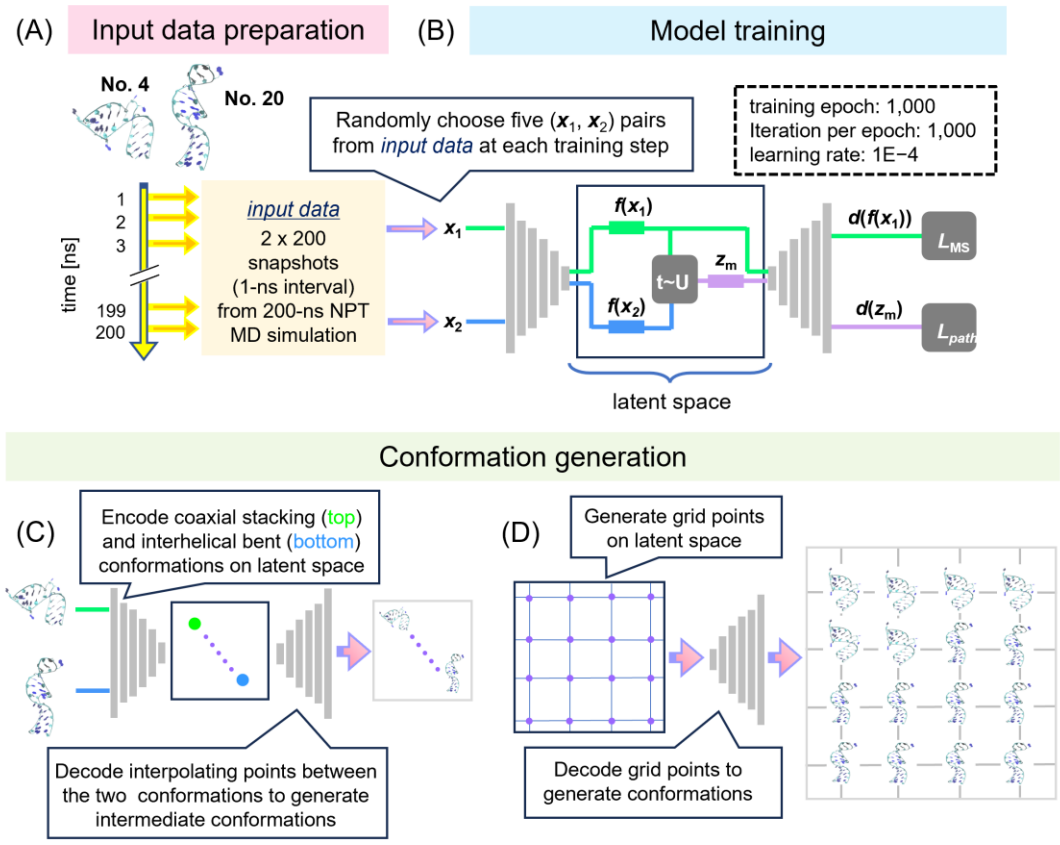

Figure 2. Procedure of Molearn model training and conformation generation. (A) Preparation of input dataset with MD simulations for models No. 4 and No. 20 (PDB entry: 7JU1). (B) Molearn model training. (C) Conformation generation by interpolating two representative conformations, a pair of MD snapshot structures with maximum RMSd value. (D) Conformation generation from grid points on Molearn's latent space. In panel B, gray rectangles denote one dimensional convolution layers. Atomic coordinates are represented by $\mathbf{x_1}$ and $\mathbf{x_2}$. Symbols *f* and *d* are for encoded and decoded atomic coordinates. L_{MSE} and L_{path} are mean square error-based and physics-based loss functions (*see* Materials and Methods for definitions). Latent space coordinate, z_m, is interpolation point defined by $tf(\boldsymbol{x}_1) + (1-t)f(\boldsymbol{x}_2)$, where *t* is a value randomly taken from [0, 1] shown as **U** in the panel.

Materials and Methods

System setup

We considered the NMR-derived *apo* TAR structures (Protein data bank (PDB) entry ID: 7JU1) as the conformation ensemble in the ground state[40] (*see* **SI-1** Supporting Information for comparison with other two *apo* structure ensembles). Among the 20 conformations, the models No. 4 and 20 are selected as the representative interhelical bent and the coaxial stacking conformations, respectively (*see* **Fig. 1B**) because they are the most structurally different pair among the 190 pairs (*i.e.*, ${}_{20}C_2$) in terms of Root Mean Square deviation (RMSd).

Each of the two *apo* TAR structures are solvated in the rectangular box with 22,580 water molecules, and electronically neutralized with 28 Na^+. To calculate the forces acting among atoms, AMBER FF 99 OL3 with backbone phosphate modification[41], TIP3P water model[42, 43], and Joung/Cheatham ion parameters adjusted for the TIP3P water model[44, 45] were applied for RNA residues, water molecules, and ions, respectively. In this study, the MD simulations are performed within a markedly short time period compared with time scale of RNA conformational transitions[46-49], then simply sampling conformations around each of the predefined, thermodynamically stable states, inter-helical bent and coaxial stacking states (*see* **Fig. 1B**). We thus suppose that choice of FF parameters is not be critical for our computational results.

Molecular Dynamics Simulations

First, atomic steric hindrances were removed by using molecular mechanics (MM) simulations, which consist of 1,000 steps of steepest descent method followed by 49,000 steps of conjugate gradient method. Then, by using molecular dynamics (MD) simulations, the system temperature and pressure were relaxed around 300 K and 1 bar,

respectively. The following 200-ns NPT MD simulations (300 K, 1 bar) were performed to prepare the training set for Molearn training (*cf.* **Fig. 2A**).

Each of MD simulations was performed under the periodic boundary condition. Electrostatic interaction was treated by the Particle Mesh Ewald method, where the real space cutoff was set to 9 Å. The vibrational motions associated with hydrogen atoms were frozen by the SHAKE algorithm. The translational center-of-mass motion of the whole system was removed by every 500 steps to keep the whole system around the origin, avoiding the overflow of coordinate information from the Amber MD trajectory format.

The system temperature and pressure were regulated with Berendsen thermostat[50] with a 5-ps of coupling constant and Monte Carlo barostat with volume change by each 100 steps, respectively. A set of initial atomic velocities were randomly assigned from the Maxwellian distribution at 0.001 K at the beginning of an MD simulation. The time step of integration was set to 2 fs. Each of MD simulations was performed using the Amber 22 [51] GPU-version PMEMD module based on SPFP algorism [52] with NVIDIA GeForce RTX 3090. Further details for simulation procedures are given in **SI-2** in Supporting Information.

Molearn model and the training procedure

We obtained the Molearn program suit from GitHub on October in 2022 from https://github.com/Degiacomi-Lab/molearn/tree/diffusion and modified a part of module programs to use the RNA FF parameter files[41]. The Molearn model for this RNA study consists of twelve layers (six encoder and six decorder layers) (**Fig. 2B**) [34]. The first layer has 3 input channels and 32 output channels and the subsequent layers with depth t have $\lfloor 32 \times 1.8^t \rfloor$ input where batch normalization and rectified linear unit activations are considered.

The training dataset is made from the 200-ns NPT MD trajectories by removing water and ion atoms from snapshot structures. Each snapshot structure is superposed on the models No. 1 of TAR (PDB entry: 7JU1) by using root-mean-square fitting algorithm with phosphate and C4' atoms in the stem region (residues number are from 17 to 21 and from 41 to 45 in biological annotation, *see* **Fig. 1A**). This treatment is indispensable due to that a Molearn model uses the cartesian coordinates as input information so far and is not equivalent to translation and rotation[34]. The model uses size 4 kernels with strides of 2 and padding of 1 in the convolution process. The total number of trainable network parameters is 5,414,899, which occupies disk space of 445 MB.

We considered all non-hydrogen atoms in RNA residue (27 atom types in total) to train a Molearn model. It is possible to execute the current version of Molearn program with fewer atom sets (linear, non-cyclic backbone trace, and minimum sidechain atom for coarse grained representation of nucleic acid bases; P, O3', O5', C4', C3', C5', O4', C1', N1, N9, C2, C4), while we here used a full set of non-hydrogen atoms to generate full atomic RNA structures.

We consider the following loss function to evaluate training processes

$$L = L_{MSE} + \alpha \cdot L_{path} \qquad \textbf{(Eq. 1)}$$

L_{MSE} is a mean square error (MSE) calculated for sets of atomic coordinates of the original and reconstructed snapshot structures (corresponding to $\boldsymbol{x}_1$ and $d(f(\boldsymbol{x}_1))$ in **Fig. 2B**, respectively). L_{path} is atomistic potential energies for intermediate structures generated by Molearn (represented by $\boldsymbol{d}(\mathbf{z_m})$ in **Fig. 2B**). The energy terms consists of atomic bonding, atomic angle, dihedral angle, and approximated non-bonded interaction defined by Coulomb interactions and 12 repulsive Lennard-Jones potential[34]. FF parameters are derived from AMBER FF 99 OL3 with backbone phosphate

modification[41].

At every optimization step, the scaling factor α is dynamically given by benchmarking an MSE value as follows. L_{path} is repeatedly multiplied by 0.1 until L_{path} becomes smaller than L_{MSE}. This dynamic assignment automatically adjusts the strength of L_{path} to L_{MSE} through all training period.

Adam algorism was employed to optimize Molearn's parameter set. The model training procedure consists of 1,000 epochs, and one epoch has 1,000 optimization iteration steps. The learning rate is set to $1e^{-4}$ and the weight decay is set to 0 as does the distributed Molearn running script. In each optimization step, 5 pairs of snapshot structures are randomly selected from the training dataset as a batch data. For each pair, either of the two is used to calculate L_{MSE}, while L_{path} is calculated from a set of Molearn-generated intermediate structures. A midpoint of the pair on the Molearn's latent space is randomly given and then decoded into a tertiary structure.

Molearn models were trained by using 5 nodes on CPU machines (AMD E7763) of Research Center for Computational Science (RCCS) and took around 6 days for a single 1,000 epoch training.

Each of Molearn-generated RNA structures was energetically relaxed by using QRNAS program[53]. An optimization step was repeated by 20,000 times with non-bonded interaction cutoff of 12Å under the implicit solvation model. A QRNAS computation for HIV-1 TAR takes about 36 minutes with single core of the RCCS's CPU machines. Since we finally obtain 10,201 HIV-1 TAR conformations from each trained model (meaning of the number will be explained in Results and Discussion section), completing all 10,201 QRNAS computations take 12 hours with 500 parallel computations in RCCS.

Structure analyses

Analyses of RMSd and calculation of geometrical properties (bond length and bond angle involved in RNA) were performed by using Cpptraj module of Amber22[51]. We evaluate chemical accuracy of atomic RNA structures by using $\%_{TotalErr}$.

$$\%_{TotalErr} = 100\left\{\frac{1}{N_{bonds}+N_{angles}}\left(\sum_{bonds}|r-r_0|/r_0 + \sum_{angles}|\theta-\theta_0|/\theta_0\right)\right\} \quad \textbf{(E.q. 2)}$$

This is an average of deviation of atomic bond length (r) and bond angle (θ) from their equilibrium values (r_0 and θ_0, respectively). N_{bonds} and N_{angles} denote the numbers of bonds and angle values, respectively.

A single point energy of RNA is calculated by using the BRiQ potential[54] to examine an energy landscape of *apo* TAR on Molearn's latent space. POVME2 is used to calculate atomic volume of cavity in the vicinity of C30 and A35[55, 56]. Molecular structures are illustrated by using Visual Molecular Dynamics (VMD) [57]. RNA-Ligand docking simulations are performed by using RLDOCK[58] and evaluate docking poses by using PoseBusters[59] and AnapuRNA scoring function[60]. Technical details for POMVE2, RLDOCK and PoseBusters are given in **SI-3**, **SI-4**, and **SI-5** in Supporting Information, respectively. In-house python scripts for data analyses were prepared with support of chatGPT-4o.

Results and Discussion

Physicochemical characterization of TAR in MV2003-bound state

We examined two physicochemical properties of NMR-derived TAR conformations, energetic stability of TAR conformations and geometrical condition to form the MV2003-binding cavities. These information are indispensable to progress this study, although the earlier studies have not reported it. [21]

By examining energetic stability of TAR in *apo* and MV2003-bound states, we tested whether TAR adopts high-energy conformations in the ligand-bound states. As for this case study, we expect that *apo* TAR forms an MV2003-binding cavity with expense of energetic stability of RNA. To confirm the promise, we calculated the state-of-the-art RNA-specific knowledge-based potential, BRiQ potential[54]. The BRiQ energy scores of TAR in the MV2003-bound state (red squares in **Fig. 3A**) always overtake those in *apo* state (green squares in **Fig. 3A**). It can be said that the TAR conformations in the MV2003-bound state appear in relatively high energy region in the conformation space of *apo* TAR. This result supports our research strategy that we find MV2003-bindable TAR conformation in high-energy states by using Molearn.

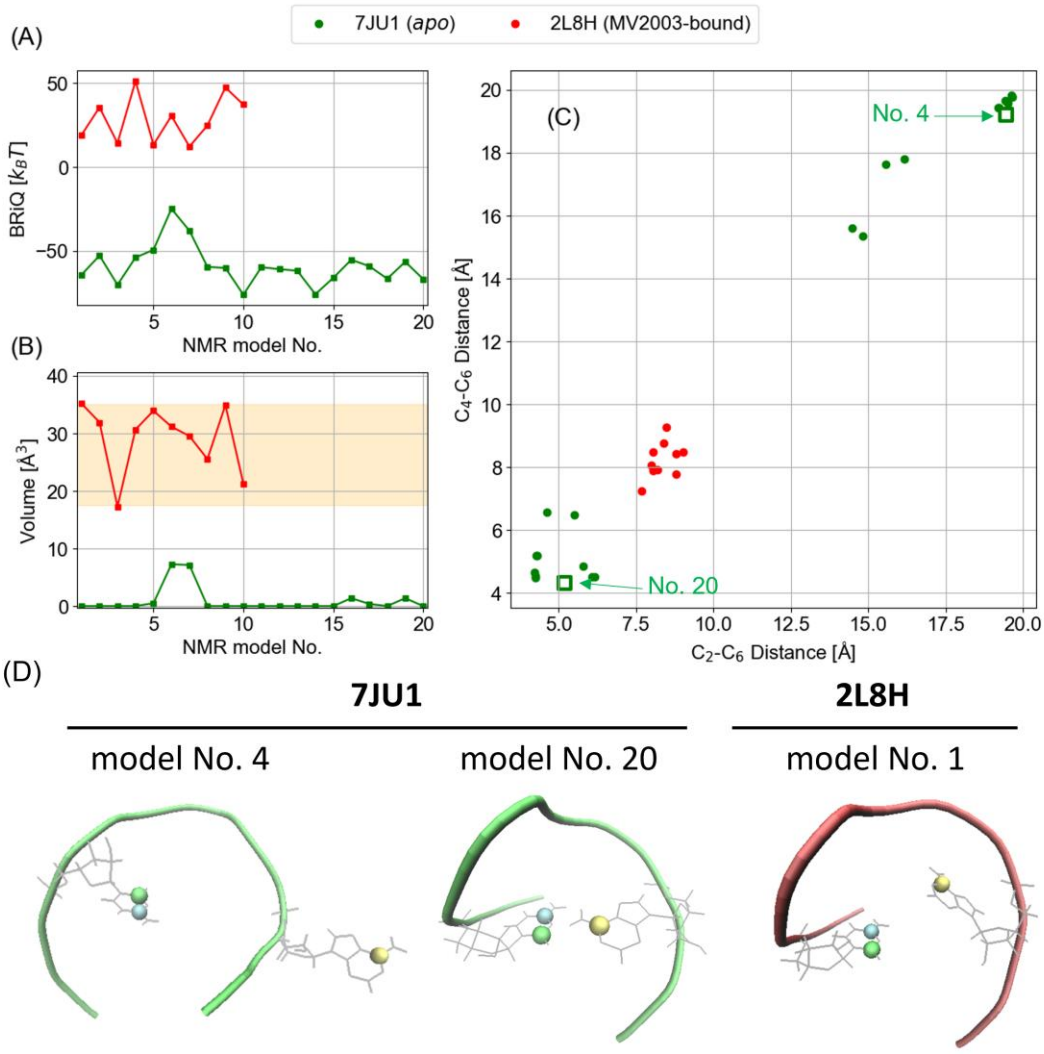

Figure 3. Physicochemical characterization of NMR-derived HIV-1 TAR conformations (*apo* and MV2003-bound forms annotated by PDB entries of 7JU1 and 2L8H, respectively). (A) BRiQ energy score (BRiQ). (B) Volume of (potential) MV2003-binding cavity. (C) Interatomic distances characterizing positional relationship between C30 and A35. (D) Location of C30 and A35 in *apo* and MV2003-bound forms. In panels A and B, orange belts denote value ranges for volume size of MV2003-binding cavity and BRiQ for HIV-1 TAR conformations in the MV2003-bound state. In panels C, green open circles at top-right and bottom-left are for interhelical bent and coaxial stacking conformations (*c.f.*, **Fig. 2B**), respectively. In panel D, blue, green and yellow balls represent C30's C_2, C30's C_4 and A35's C_6 atoms, respectively.

Next, we consider geometrical conditions satisfied by TAR conformations in the MV2003-bound state. In the following sections, we will systematically generate TAR conformations with Molearn models and then comprehensively search MV2003-bindable ones, which have a cavity accessible for MV2003. However, there have been no earlier studies to quantitatively characterize the cavity with geometrical metrics, to our best knowledge. Aiming to perform comprehensive search for TAR conformations of interest, we address to give geometrical conditions to form the MV2003-binding cavities.

We analyzed volume of cavity formed between the Cyt30 and Ade35 for each of *apo* and MV2003-bound TAR conformations (**Fig. 3B**). Each of MV2003-bound TAR conformations has an MV2003-binding cavity, and the values of cavity volume range from 17 to 35 $Å^3$ (*see* red squares in **Fig. 3B**, where the value range is highlighted by the orange belt). Meanwhile, the *apo* TAR conformations do not form sufficiently large cavities (*see* green squares in **Fig. 3B**). The models No. 6 and No 7 of *apo* TAR show relatively large volume values, although the values are around 8 $Å^3$, thus being incapable of binding MV2003. It is noted that the energy values of No. 6 and No. 7 are relatively high compared with the others (*see* **Fig 3A**). It could be said that such a cavity formation results in energetic destabilization of *apo* TAR, then supporting our motivation to use for generation of the cryptic binding cavity.

A relative position between MV2003-interacting RNA bases is also important metrics for the cryptic MV2003-binding cavity. The disruption of the RNA base pair results in the formation of the MV2003-binding cavity (**Fig. 1C** and **1D**). The methoxy naphthalene group of MV2003 is bound to TAR by a pseudo-stacking with Cys30 and Ade35, which are found below the apical loop (*see* **Fig. 1C** for the tertiary MV2003-bound structure). Then, we characterized a positional relationship between these RNA bases by using a pair

of inter-atomic distances. One is distance between C_2 of Cyt30 and C_6 of Ade35, and the other is that between C_4 of Cyt30 and C_6 of Ade35 (hereafter, the former and later are referred to as $d_{2\text{-}6}$ and $d_{4\text{-}6}$, respectively). We can find that the ($d_{2\text{-}6}$, $d_{4\text{-}6}$) distribution for *apo* conformations and that for MV2003-bound conformations is mutually exclusive (**Fig. 3C**). The MV2003-bound conformations are localized in the ($d_{2\text{-}6}$, $d_{4\text{-}6}$) space and the distribution appears to be unimodal (a representative case is illustrated in the right panel in **Fig. 3D**). The value ranges of $d_{2\text{-}6}$ and $d_{4\text{-}6}$ are [7.7, 9.0] and [7.2, 9.3] (Unit: Å), respectively, which is a necessary condition for TAR to form pseud-stacking with MV2003.

Meanwhile, the distribution for *apo* TAR conformations appears to be bimodal and is located outside of that for the MV2003-bound TAR conformations. These observations denote that any of the NMR-derived *apo* TAR conformations does not take MV2003-bindable configurations for Cyt30 and Ade35. For example, the representative interhelical bent and coaxial stacking conformations (the models No. 4 and No. 20 in **Fig 1B**, highlighted by open green squares in **Fig 3C**) are different in positional relationship of the RNA bases. Ade35 is flipped out in the model No. 4, while Cys30 and Ade35 make contacts in the model No. 20 (*see* the left and center panels in **Fig. 3D**). These arrangements of the bases inhibit pseud-stacking formation with a methoxy naphthalene group of MV2003.

According to the above geometrical analyses, we suppose that predicted MV2003-bindabld conformations should satisfy the three geometrical conditions at the same time: (i) the cavity volume ranges from 17.5 to 35 $Å^3$; (ii) $d_{2\text{-}6}$ is found between 7.7 and 9.0 Å; (iii) $d_{4\text{-}6}$ is found between 7.2 and 9.3 Å. This geometrical condition is employed to analyze MD-derived and Molearn generated conformations in the following section.

Molearn can generate RNA intermediate conformations

As in the case of the original Molearn study, we trained Molearn models by using a set of snapshot structures obtained from atomistic MD simulations. We performed the two 200-ns NPT MD simulations for TAR. One and the other start from the NMR-derived TAR No. 4 and 20 models (**Fig. 1B**), respectively. Each of the two MD trajectories exhibits equilibrated regions corresponding to representative TAR conformational states: the inter-helical bend and coaxial stacking conformations (*see* **SI-2** and **Fig S2** for related discussion). We then picked up 200 snapshot structures from each of the MD trajectories with 1-ns interval. These two sets of 200 snapshot structures are integrated to use as the training dataset (*cf.* **Fig. 2A**).

Notably, none of the 400 MD-derived TAR conformations exhibits a potential MV2003-binding cavity (**Fig. 4**). They show apparently smaller volume of cavity formed between Cyt30 and Ade35 (green and blue graphs in **Fig. 4A**) than the NMR-derived TAR conformations in the MV2003-bound state (the value range for such conformations is given orange belt in **Fig. 4A**). Although their ($d_{2\text{-}6}$, $d_{4\text{-}6}$) distributions (green and blue squares in **Fig. 4B**) is spatially expanded by the influence of thermal fluctuation, they are still exclusive to that for the NMR-derived TAR conformations in the MV2003-bound state (red squares in **Fig. 4B**). Therefore, we trained Molearn model without *a priori* knowledge for the cryptic MV2003-binding cavity.

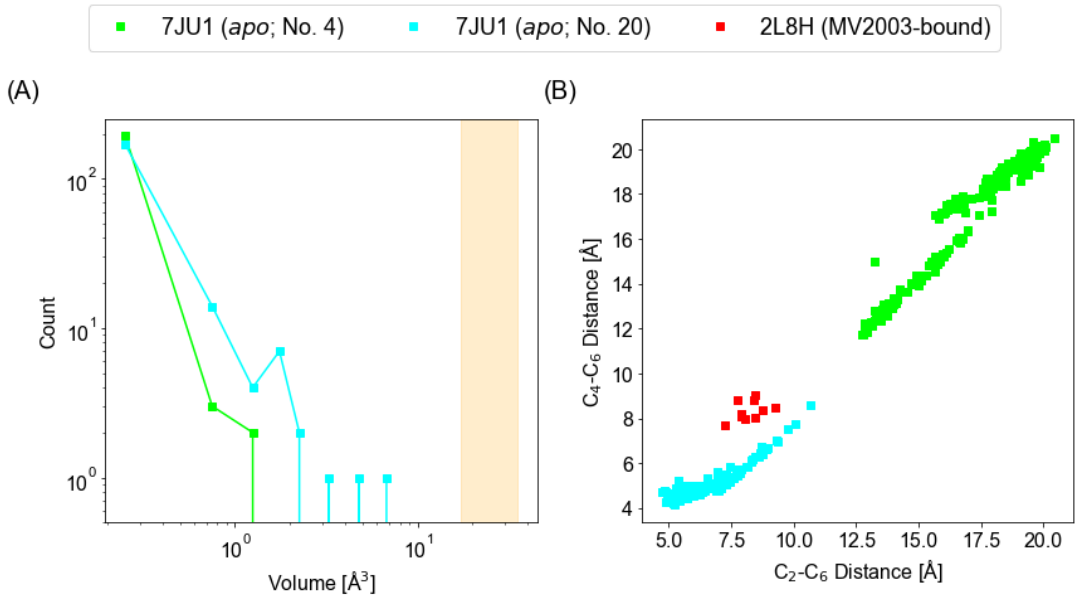

Figure 4. Geometrical properties of MD derived-Molearn input data. (A) Volume of (potential) MV2003-binding cavity. (B) A pair of interatomic distances characterizing positional relationship between C30 and A35. Green and blue squares are for snapshot structures obtained from MD simulations starting from *apo* HIV-1 TAR conformations annotated by No. 4 and No. 20 (PDB entry: 7JU1). In panel A, orange belt denotes value ranges for volume of MV2003-binding cavity for NMR-derived HIV-1 TAR conformations in the MV2003-bound state (PDB entry: 2L8H). In panel B, red squares are for the NMR derived conformations.

By using the training dataset, we independently trained Molearn models ten times. Each of trained Molearn models shows decreasing network-loss through the training epochs and the values appear to converge by 1,000th epoch (**Fig. 5A**). We assume that each of Molearn models are finely trained to reproduce the training data and to generate intermediate structures with chemically reasonable geometry (*cf.* Molearn's loss function defined in **Eq. 1**).

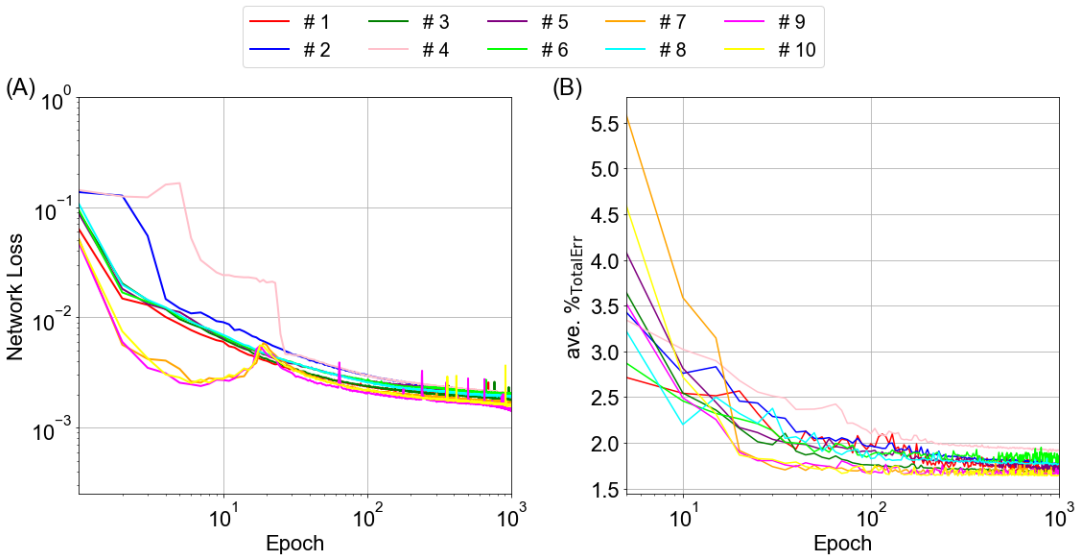

Figure 5. Evaluation of trained Molean model. (A) Network loss along training epoch. Value at each epoch is an average over 1,000 training steps in the epoch. (B) Atomic precision of Molearn-generated conformations along training epoch. #1 to #10 in the legend box denote the ten independently trained Molearn models. With regard to (B), $\%_{\mathrm{TotalErr}}$ values are calculated for 50 RNA conformations generated by interpolating the interhelical bent and coaxial stacking forms on Molearn's latent space (*c.f.* **Fig. 1C** and **Fig. 2B**) at each epoch. A set of 50 $\%_{\mathrm{TotalErr}}$ values are averaged at each epoch to show the graph.

We confirmed the assumption by calculating the chemical accuracy of Molearn-generated intermediate conformations with $\%_{\mathrm{TotalErr}}$ (**Eq. 2**). Intermediate structures were generated by interpolating the representative conformations (*cf.* **Fig. 2C**), the interhelical bent and coaxial stacking ones (a pair with the maximum RMSd value picked up from the training dataset, which conformationally resembles the No. 4 and 20 models in **Fig 2B**, respectively) on the latent space given from every Molearn model. We used 48 latent space coordinates connecting between the two input conformations to generate intermediate TAR structures. They are combined with the two representative conformations, which are used as the terminals of interpolation on the latent space, for

the following analyses.

For each of the ten Molearn models, the average of $\%_{TotalErr}$ decreases to 2.0 or smaller by 1,000th epoch (**Fig. 5B**). Meanwhile, the average of $\%_{TotalErr}$ for the MD-derived training dataset is 2.3, then being subtly larger than that for the Molearn-generated conformations. This may be due to that the training dataset is obtained under thermal fluctuation. Atomic bonds and angles should be stretched from equilibrium ones by kinetic energies at 300 K. As remarked above, the Molearn-generated conformations are energetically optimized on potential energy surface at 0 K to remove potential steric hindrance of atoms, then resulting in relatively small $\%_{TotalErr}$ values. Recalling these $\%_{TotalErr}$ values, we suppose that our Molearn-generated RNA conformations have sufficient chemical accuracy. Considering the chemical accuracy of the Molearn-generated conformations, we then suppose that the parameter sets at 1,000th epoch is worthwhile employing as a representative for conformation generations. Although we tested each of the 10 Molearn models, we could not find MV2003-bindable conformations among those generated along the transition pathways (details are given in **SI-7** in **Supporting Information**,). We will discuss this unexpected result in the last subsection in **Results and Discussion** in details. Then, from practical viewpoint, we examine performance of Molearn as full atomistic RNA conformational generator hereafter.

Molearn generates MV2003-bindable conformations without *a prior* knowledge

We comprehensively generated TAR conformations from a set of grid points on the Molearn's latent space (*cf.* **Fig. 2D**). The latent space of Molearn is a finite two-dimensional square region, whose x and y coordinates range from 0 to 1. A space between

nearest grid points is set to 0.01 for both x and y directions. We consider 10,201 (= 101 x 101) grid points on the latent space to decode into atomic structures for each Molearn model.

We confirmed the presence of MV2003-bindable conformations for each set of the Molearn generated conformations. We calculated the volume of cavity formed between Cys30 and Ade35 and picked up those satisfying the volume value criteria (*cf.* **Fig 3B**). We further screened them by using the criteria for a pair of interatomic distances, ($d_{2\text{-}6}$, $d_{4\text{-}6}$) discussed in **Fig 3C**. Finally, seven of the ten Molearn models succeeded in generating TAR conformations harboring MV2003-bindable cavity (**Table 1**). In total, we obtained 61 MV2003-bindable TAR conformations. As remarked above, the training dataset does not have any conformations with the cavity in the sufficient size (*cf.* **Fig. 4A**), so that it could be said that Molearn has a potential to predict unprecedent, important conformations such as TAR with the cryptic binding cavity, without *a priori* knowledge of the conformations. Of the 61 conformations, we visually selected one generated conformation whose arrangement of Cys30 and Ade35 is the most similar to that of the NMR-derived TAR conformations in the MV2003-bound state (this point will be discussed in the next subsection in details; the other conformations are shown in **Fig. S4** in **Supporting Information**). This result suggests that the Molearn approach outperforms the early MD studies[24-26] in generating MV2003-bindable TAR conformations harboring the cryptic binding cavity (a more detailed discussion is provided in the next-after subsection in **Results and Discussion**). Nonetheless, it is noted as technical issue for the present Molearn model that the generative performance of conformation is not deterministic, but rather probabilistic. **Table 1** shows variations of the number of MV2003-bindable TAR conformations and structural similarity to the NMR-derived

conformations in terms of RMSd. Training models to stably achieve specific functions is a future challenge to make Molearn general method to generate unprecedent, important RNA conformation such as TAR with the cryptic binding cavity.

Table 1. RMSd of Molearn generated conformations with potential MV2003-binding cavity to NMR-derived conformations in the MV2003-bound state (PDB entry: 2L8H), where C30 and A35 are considered [Unit: Å]

Molearn model #	the number of conformation	RMSd of ave. with CI95 †	min. RMSd value
1	8	2.68±0.27	1.73
2	19	3.01±0.06	2.76
3	2	2.93±0.49	2.28
4	19	2.11±0.07	1.78
5	1	2.67±0.19	2.57
6	10	2.91±0.24	2.11
7	2	3.19±0.17	2.95

†For a Molearn generated conformation, each of the NMR-derived 10 TAR conformations in the MV2003-bound state are used to calculate RMSd, then giving 10 RMSd values per one conformation.

Molearn models generate TAR conformation capable of binding MV2003.

As yet, it is unclear whether each of the 61 Molearn-generated TAR conformations is actually MV2003-bindable. We then verified this point by docking a given TAR conformation, one by one, against each of the ten NMR-derived MV2003 conformations

in the TAR-bound state (PDB entry: 2L8H)[21]. From each pairing of a TAR conformation with an MV2003 conformation, docking generated a set of poses (the docking procedure is illustrated in **Fig. S5**). These were ranked by the AnnapuRNA scoring function[60], and the top-ranked pose was selected as the representative. We finally obtained 610 docked TAR-MV2003 poses. Furthermore, they are evaluated by PoseBusters [59] to check unphysical inter-atomic contacts between TAR and MV2003. The 484 of 610 poses satisfy the geometrical criteria, then being considered for the subsequent analyses (*see* **SI-5** for technical details).

Fig. 6A shows distributions of AnnapuRNA scores calculated for the 484 docked poses. The orange belt indicates the value range of AnnapuRNA score calculated for NMR-derived TAR-MV2003 complexes (the maximum and minimum values are −121 and −180, respectively). 86% of docking poses appear within the value range, meaning that they earn energetically stability at the binding site as well as the NMR-derived TAR-MV2003 complexes. Furthermore, the 11% of docked poses show lower AnnapuRNA scores (left to the orange belt in **Fig. 6A**). These docked poses show stronger RNA-ligand interactions than the NMR-derived complexes. In total, we can highly frequently obtain TAR-MV2003 docked poses comparative to the experimentally resolved ones in terms of AnnapRNA score. This result suggests that the Molearn could accomplish to generate MV2003-bindable cavities, which are comparable to those in the NMR-derived TAR conformations.

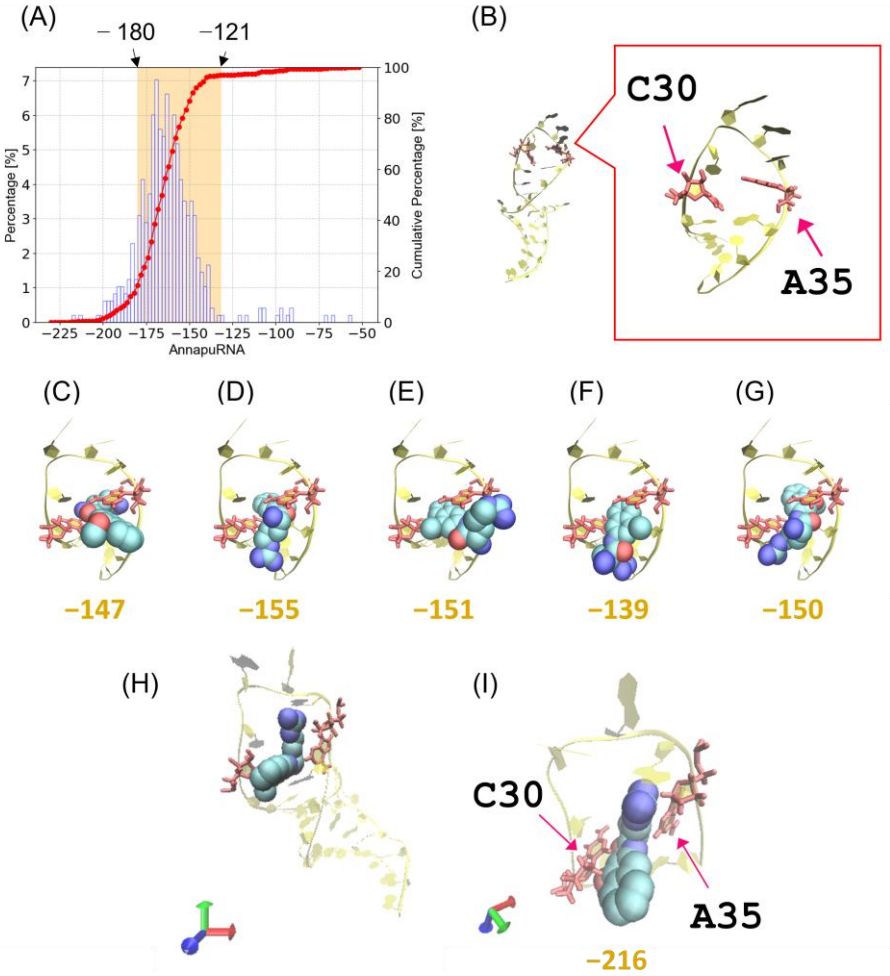

Figure 6. Analyses of docked TAR-MV2003 pose. (A) RNA-ligand binding scores calculated by AnnapuRNA for all docked poses; the orange belt for the value range for NMR-derived TAR-MV2003 complexes. (B) Molearn generated TAR most similar to NMR-derived TAR in the MV2003-bound state. MV2003 interaction site is magnified in the red frame. (C)-(G) docked poses given from the conformation shown in (B). (H) Docked pose with the minimum AnnapuRNA score, obtained from other Molearn generated TAR conformation. (I) Closing view of TAR-MV2003 conformation shown in (G). In (A), blue open bars and red closed circles are for percentages and cumulated percentage of the histogram, respectively. MV2003 and the interacting residues are shown by van der Waals spheres and red sticks, respectively. HIV-1 TAR structure is illustrated by transparent yellow ribbon, as in the case of **Fig. 1C**. In (C)-(G) and (I), values, colored in orange and shown below closing of TAR-MV2003 docked poses, denote the corresponding Annapurna scores.

As remarked above, we found one Molearn-generated MV2003-bindable conformation that most closely resembled the arrangement of Cys30 and Ade35 observed in the NMR-derived TAR conformations in the MV2003-bound state (**Fig. 6B**). In this conformation, a cavity is present between Ade35 and Cys30 that can accommodate intercalation of the methoxy naphthalene group of MV2003, and Ade35 is positioned above Cys30 (*see* **Fig. 1C** for comparison). Docking simulations of this TAR conformation with MV2003 yielded five valid poses (the other five were excluded by PoseBusters because of steric clashes between RNA and ligand). In these poses, the methoxy naphthalene groups of MV2003 were intercalated between Cys30 and Ade35 (**Fig. 6C-G**), and the pose in **Fig. 6C** most closely resembled the NMR-derived TAR-MV2003 complex at the cryptic binding cavity (*see* **Fig. 1C**). Each of these poses received an AnnapuRNA score within the range observed for NMR-derived TAR–MV2003 complexes (−147, −155, −151, −139, and −150 in the order of panels in **Fig. 6**). Although none of these poses corresponds to the absolute lowest-scoring conformation (**Fig. 6H, I**)—likely because the scoring function emphasizes TAR-MV2003 interactions without fully accounting for internal energies of TAR and MV2003 (*see* **SI-6** for further discussion)—they nonetheless demonstrate that Molearn-generated TAR conformations can form stable cryptic binding cavities highly consistent with the NMR-derived TAR-MV2003 complexes.

Technical advantage of Molearn in comparison with other computational approaches

Molearn successfully generates TAR conformations with cryptic MV2003 binding cavity. It is then worthwhile to compare our Molearn approach with those by earlier computational studies. Studies to predict cryptic binding sites are classified into two

classes, atomistic MD simulations and DL prediction models, in general. [61] Atomistic MD simulations have been employed to search cryptic binding sites inside proteins[62-64], while there are few examples to straightforwardly apply them into RNA's cryptic binding site[26]. This may be due to that the presence of RNA's cryptic binding site has not been generally recognized yet. Recently, Panei and colleagues applies atomistic MD simulations to investigate potential ligand-binding sites of RNA, including cryptic binding site of TAR. [26] Their MD-based approach suggested a part of MV2003 binding site in TAR, molecular surface interacting with a portion of MV2003. However, their approach could not generate MV2003-bindable tertiary conformations harboring the cryptic binding cavity. Meanwhile, we obtained MV2003-bindable TAR conformations with the Molearn approach. Therefore, it is said that our approach outperforms their MD approach with regard to generation of MV2003-bindable TAR conformations.

As it has been discussed long time, performance of MD simulations strongly depends on FF parameters used to simulate atomic dynamics[31, 65-70]. If we use MD simulations to sample RNA conformations, computational results may sensitively depend on combinations of FF parameters for RNA, water molecules and ions (additionally probe ligands to promote formation of ligand binding sites as in the case of Panei's study[26]). In particularly, it is an unresolve concern of MD simulations for RNA that conventional charge-fixed FF parameters misestimate stability of interaction acting between RNA residues[28]. Such an imbalance would impede rearrangements of intra-RNA interaction and then formation of cryptic binding sites, as discussed in **Introduction**. Even if we may find it a practical option to test all possible combinations for an optimal combination of FF parameters[33, 71, 72], it is routinely unrealistic to complete such a task within a realistic time. Because of technical immaturities of FF parameters and limited computational

resources, MD-derived approaches may have fails to reproduce TAR conformations harboring the cryptic binding cavity, probably due to that the cavity itself is not formed through their MD simulations.

We also employed MD simulations to prepare Molearn training dataset, although our MD simulations are designed to sample conformations around equilibrium states, obtained from the experimentally resolved structures, with relatively short simulation length. [31, 70] This is because we do not have to consider remarkable conformational changes through MD simulations. It is a distinguishable advantage of Molearn that the we could generate ligand-bindable RNA conformations without careful adjustments of empirical FF combinations and a potential concern about falling into non-native RNA folds through long MD simulations. We could suppose that Molearn may work as tools to compensate MD simulations to inspect rarely-emerging ligand-bindable RNA conformations.

From viewpoints of inspecting cryptic binding site, it is worthwhile to note the earlier DL models, which are designed for proteins to predict a set of residues contributing to formation of cryptic binding sites. [73-76] To our best knowledge, there are no counterparts to predict cryptic binding sites in RNA so far. Developing an RNA-targeted DL model may be hampered by shortage of RNA tertiary structure information, which occupies 1% of PDB entries (<2,000 RNA entries, including about 300 liganded-RNA ones on 5th April, 2025). Fundamentally, cryptic binding sites in RNA molecules are not well recognized in the research field of RNA-targeted drug yet. Under this circumstance, our Molearn approach appears to be practical to discover cryptic binding site in RNA in comparison with previous methods.

As for generative DL model for RNA, it is worthwhile to note DynaRNA[77] introduced

by Li and colleagues, a pioneering denoising diffusion probability model (DDPM) for generating RNA conformational ensembles introduced. This model successfully captured rarely-populated TAR conformations but did not yet remark reproduction of MV2003-bindable conformations to our knowledge. [77] Their study primarily addressed intrinsic RNA dynamics rather than ligand-induced conformational changes; thus, future extensions incorporating small-molecule binding states—such as those observed for MV2003-bound TAR—may further enhance its applicability to RNA–ligand complex modeling. Meanwhile, DynaRNA employs an E(3)-equivariant architecture, possibly enabling it to learn atomic geometries in RNA more accurately than the one dimensional convolutional neural network (1D CNN)–based architecture used in Molearn. [78] Incorporating a diffusion model such as DDPM, or other advanced deep-learning frameworks, in combination with an E(3)-equivariant network could further enhance the geometric realism of RNA conformations, thereby reducing the computational cost required for refining Molearn-generated structures (estimation of the computational costs are found in **Material Methods** and **SI-11** in **Supporting Information**).

Current status of computational prediction of cryptic binding site with Molearn approach

According to these achievements, it can be said that we accomplished the main purpose of this study, generation of TAR conformations with cryptic MV2003-binding cavity without *a priori* information. With regard to the challenge to generate an unprecedent, important RNA conformation, our Molearn approach first solve it and may outperform other MD and DL approaches in the scope of this study on TAR. Recalling such a milestone which we reached in this study, we remark remaining challenges of Molearn-

based approach to predict functional RNA conformations.

Notably, we obtained TAR conformations with the cryptic binding site, forming local interaction with MV2003 as in the case of the NMR-derived TAR-MV2003 complexes. It is reported that DL-based RNA 3D structure predictions are often less accurate to describe local interactions[79], so that successful generation of the cryptic binding cavity of TAR implies the technical advantage of Molearn in capturing local structures (as for TAR, disruption of base pair between Cyt30 and Ade35, shown in **Fig. 1C** and **1D**). Meanwhile, the present version of Molearn seemingly has limited performance to globally sample conformation space. Each of the 61 Molearn-generated conformations with cryptic binding cavities appear to take coaxial stacking conformation (*c.f.* conformation in blue in **Fig.1B**), while an NMR-derived TAR in complex with MV2003 takes those similar to interhelical bent conformation (*c.f.* conformation in green in **Fig.1B**). Although the training dataset equally includes both of these conformations (*see* **Materials and Methods**), the present version of Molearn could not generate such global RNA conformations harboring the cryptic binding cavity. Replacing the current 1D-CNN with a geometry-aware diffusion (or flow-matching) model could help Molearn capture long-range atomic relationships[80] and maintain the correct rotational and translational behavior of RNA structures, as demonstrated by recent generative models[77, 81].

It is also important, when we consider Molearn as practical applications, that we can systematically detect functional RNA conformations from whole of Molearn-generated ones in the framework of Molearn. However, such a systematic detection function remains unachieved so far (*see* details given in **SI-7** in Supporting Information). Despite our expectation, TAR conformations with the cryptic binding site does not appear near the conformational transition pathways. The present Molearn's latent space may not

necessarily work as a counterpart of physicochemical free energy landscape. From practical viewpoint, the above problem of the present Molearn could be complemented by employing ligand binding site finders such as fpocketR[82]. We could pick up ligand bindable conformations among Molearn-generated RNA conformations. Nonetheless, future refinement of Molearn looks a worthwhile task. If we can detect RNA-bindable conformations among a Molearn-generated RNA conformational ensemble by analyzing Molearn's latent space, we could obtain profound insights into essential descriptor characterizing physicochemical processes of RNA functional expression, simultaneously.

Using the paradigmatic RNA, TAR, as benchmark, we addressed the challenge and provided a partial but the first solution. It is the next step to consider scalability of Molearn, a key factor for DL methods. We applied Molearn to a larger RNA molecule, Enterovirus internal ribosome entry site (IRES) and obtained conformations with the cryptic ligand binding cavity, which can be accessed by the experimentally examined ligand molecule, DMA-135 with comparative Annapurna scores (*see* **SI-10** and **SI-11**, also **Fig**. **S11** and **Fig. S15**). Recently, Nithin and colleagues comprehensively examined whether six RNA 3D structure prediction methods, two DL models and four atomistic simulation-based methods, can generate RNA conformations in the ligand-bound states. [79] The test dataset comprised 139 complexes with biologically relevant ligands but the TAR-MV2003 complex was excluded in the test, because of that MV2003 is a synthetic chemical prove. Meanwhile, IRES is one of the RNA-ligand complexes which they tested so that we tested the cryptic binding sites by using all IRES conformations with they shared. Notably, any prediction methods which they tested did not recover IRES conformations featuring the cryptic binding site (*see* **SI-10**, **SI-12**, and **SI-13**; **Fig. S14**). It is noted that they filtered predicted structures up to 26 IRES models, although our

Molearn approach only reproduced IRES conformations harboring the cryptic binding pocket so far. This suggests that Molearn approach may overcome practical limitation of RNA structure prediction methods of today, with regard to generating cryptic binding cavities.

This observation may encourage us generally employ Molearn to generate RNA conformations with cryptic binding sites by only using *apo* RNA conformations. Whereas, we could not obtain IRES conformations globally similar to the NMR-derived conformations in the DMA-135-bound state, as in the case of TAR (*see* detailed discussion for the case study of IRES in **SI-11**). Besides, local interactions between DMA-135 and RNA residues in IRES are partially different to those of NMR-derived IRES conformation in the bound state. Therefore, the present Molearn model requires technical refinement to generate more accurate RNA conformations than those obtained in this study (related discussion is given in the later part of **SI-11** in **Supporting Information**).

Considering success of advantageous framework in the field of DL-based biomolecule's conformation prediction[83, 84], such as DynaRNA[77], introducing a state-of-the-art DL framework into Molearn, with keeping the essential idea, transition path generation, would result in more powerful generative DL models than the model at present. Molearn may achieve higher accuracy in describing both local molecular geometries, *e.g.* (cryptic) ligand-binding sites, and global molecular architectures, *e.g.* folds of RNA, thereby promoting SBDD for RNA.

Concluding Remarks

Cryptic binding sites of RNA are promising drug targets due to high specificity and low off-target effects. In aiming to discover such drugs, SBDD approach is an

unquestionably pivotal but technically unfeasible yet, due to paucity of tertiary RNA structures and theoretical difficulty of predicting cryptic binding sites. Under the circumstance, we addressed to theoretically predict RNA conformations harboring cryptic binding sites. By extending applicability of the generative DL model, Molearn, we first succeeded to capture cryptic binding cavities of TAR and IRES, without *a priori* knowledge for their cryptic cavities. To our knowledge, earlier standalone MD and DL studies did not capture these cryptic binding cavities [24-26, 77, 79]. Thus, this study provides the first illustration that a hybrid approach can generate RNA harboring cryptic-binding sites by only using *apo* RNA conformational ensembles, thereby providing a baseline for future studies in this direction. Indeed, key challenges remain for routinely generating RNA conformations that recover both cryptic binding sites and global folds across RNA of varying sequence length. Thus, the next step is to update the Molearn framework to address these challenges, thereby establishing it as a powerful deep learning model for RNA-targeted structure-based drug design (SBDD).

Acknowledgements

This research was supported by the Core Research for Evolutional Science and Technology (CREST) under Grant Number JPMJCR21F1, Japan Agency for Medical Research and Development (AMED) under Grant Number 23ae0121049h0003, and Grant-in-Aid for Scientific Research (A) under Grant Number 23H00509. The computation associated with Molearn was performed using Research Center for Computational Science, Okazaki, Japan (Project: 23-IMS-C129; 24-IMS-C123; 25-IMS-C122).

Author contributions

M.H. supervised the project. I.K. conceived and designed the research, performed computational experiments and analyzed all data. I.K. and M.H. interpreted and discussed the results, wrote the manuscript, and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Supplementary Information

The supplementary information is available free of charge. Further details for atomistic MD simulations, additional physicochemical analyses for NMR-derived and Molearn-generated conformations, supporting Tables and Figures.

Data Availability

The MD simulation dataset, Molearn-related data and cavity analyses for predicted RNA

structures in the earlier study[79] that support the findings of this study are obtained from https://github.com/hmdlab/molearn4rna and https://waseda.app.box.com/folder/300778804035?v=data-molearn4rna.

Code Availability

Codes needed to run Molearn for RNA are obtained from https://github.com/hmdlab/molearn4rna.

References

(1) Ganser, L. R.; Kelly, M. L.; Herschlag, D.; Al-Hashimi, H. M. The roles of structural dynamics in the cellular functions of RNAs. *Nat. Rev. Mol. Cell Bio.* **2019**, *20* (8), 474-489. DOI: https://doi.org/10.1038/s41580-019-0136-0.

(2) Hirose, T.; Ninomiya, K.; Nakagawa, S.; Yamazaki, T. A guide to membraneless organelles and their various roles in gene regulation. *Nat. Rev. Mol. Cell Bio.* **2023**, *24* (4), 288-304. DOI: https://doi.org/10.1038/s41580-022-00558-8.

(3) Chekulaeva, M. Mechanistic insights into the basis of widespread RNA localization. *Nat. Cell Biol.* **2024**, *26* (7), 1037-1046. DOI: https://doi.org/10.1038/s41556-024-01444-5.

(4) Zacco, E.; Broglia, L.; Kurihara, M.; Monti, M.; Gustincich, S.; Pastore, A.; Plath, K.; Nagakawa, S.; Cerase, A.; de Groot, N. S.; et al. RNA: The unsuspected conductor in the orchestra of macromolecular crowding. *Chem. Rev.* **2024**, *124* (8), 4734-4777. DOI: https://doi.org/10.1021/acs.chemrev.3c00575.

(5) Osborne, R. J.; Thornton, C. A. RNA-dominant diseases. *Hum. Mol. Genet.* **2006**, *15*, R162-R169. DOI: https://doi.org/10.1093/hmg/ddl181.

(6) Cooper, T. A.; Wan, L.; Dreyfuss, G. RNA and disease. *Cell* **2009**, *136* (4), 777-793. DOI: https://doi.org/10.1016/j.cell.2009.02.011.

(7) Wahlestedt, C. Targeting long non-coding RNA to therapeutically upregulate gene expression (vol 12, pg 433, 2013). *Nat. Rev. Drug. Discov.* **2014**, *13* (1). DOI: https://doi.org/10.1038/nrd4203.

(8) Tao, Y. N.; Zhang, Q.; Wang, H. Y.; Yang, X. Y.; Mu, H. R. Alternative splicing and related RNA binding proteins in human health and disease. *Signal Transduct. Tar.* **2024**, *9* (1). DOI: https://doi.org/10.1038/s41392-024-01734-2.

(9) Anderson, A. C. The process of structure-based drug design. *Chem. Biol.* **2003**, *10* (9), 787-797. DOI: https://doi.org/10.1016/j.chembiol.2003.09.002.

(10) von Itzstein, M. The war against influenza: discovery and development of sialidase inhibitors. *Nat. Rev. Drug. Discov.* **2007**, *6* (12), 967-974. DOI: https://doi.org/10.1038/nrd2400.

(11) Aplin, C.; Milano, S. K.; Zielinski, K. A.; Pollack, L.; Cerione, R. A. Evolving experimental techniques for structure-based drug design. *J. Phys. Chem. B* **2022**, *126* (35), 6599-6607. DOI: https://doi.org/10.1021/acs.jpcb.2c04344.

(12) Burley, S. K.; Wu-Wu, A.; Dutta, S.; Ganesan, S.; Zheng, S. X. F. Impact of structural biology and the protein data bank on us fda new drug approvals of low molecular weight antineoplastic agents 2019-2023. *Oncogene* **2024**, *43* (29), 2229-2243. DOI: https://doi.org/10.1038/s41388-024-03077-2.

(13) Weller, J. A.; Rohs, R. Structure-based drug design with a deep hierarchical generative model. *J. Chem. Inf. Model.* **2024**, *64* (16), 6450-6463. DOI: https://doi.org/10.1021/acs.jcim.4c01193.

(14) Kognole, A. A.; Hazel, A.; MacKerell, A. D. SILCS-RNA: Toward a structure-based

drug design approach for targeting RNAs with small molecules. *J. Chem. Theory Comput.* **2022**, *18* (9), 5672-5691. DOI: https://doi.org/10.1021/acs.jctc.2c00381.

(15) Consortium, R. N. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.* **2021**, *49* (D1), D212-D220. https://doi.org/10.1093/nar/gkaa921.

(16) Aboulela, G.; Karn, J.; Varani, G. Structure of HIV-1 TAR RNA in the absence of ligands reveals a novel conformation of the trinucleotide bulge (vol 24, pg 3974, 1996). *Nucleic Acids Res.* **1996**, *24* (22), 4598-4598.

(17) Faber, C.; Sticht, H.; Schweimer, K.; Rösch, P. Structural rearrangements of HIV-1 Tat-responsive RNA upon binding of neomycin B. *J. Biol. Chem.* **2000**, *275* (27), 20660-20666. DOI: https://doi.org/10.1074/jbc.M000920200.

(18) Du, Z. H.; Lind, K. E.; James, T. L. Structure of TAR RNA complexed with a Tat-TAR interaction nanomolar inhibitor that was identified by computational screening. *Chem. Biol.* **2002**, *9* (6), 707-712. Doi: https://doi.org/10.1016/S1074-5521(02)00151-5.

(19) Davis, B.; Afshar, M.; Varani, G.; Murchie, A. I. H.; Karn, J.; Lentzen, G.; Drysdale, M.; Bower, J.; Potter, A. J.; Starkey, I. D.; et al. Rational design of inhibitors of HIV-1 TAR RNA through the stabilisation of electrostatic "hot spots". *J. Mol. Biol.* **2004**, *336* (2), 343-356. DOI: https://doi.org/10.1016/j.jmb.2003.12.046.

(20) Murchie, A. I. H.; Davis, B.; Isel, C.; Afshar, M.; Drysdale, M. J.; Bower, J.; Potter, A. J.; Starkey, I. D.; Swarbrick, T. M.; Mirza, S.; et al. Structure-based drug design targeting an inactive RNA conformation: Exploiting the flexibility of HIV-1 TAR RNA. *J. Mol. Biol.* **2004**, *336* (3), 625-638. DOI: https://doi.org/10.1016/j.jmb.2003.12.028.

(21) Davidson, A.; Begley, D. W.; Lau, C.; Varani, G. 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. DOI: https://doi.org/10.1016/j.jmb.2011.03.039.

(22) Davidson, A.; Patora-Komisarska, K.; Robinson, J. A.; Varani, G. Essential structural requirements for specific recognition of HIV TAR RNA by peptide mimetics of Tat protein. *Nucleic Acids Res.* **2011**, *39* (1), 248-256. DOI: https://doi.org/10.1093/nar/gkq713.

(23) Shortridge, M. D.; Wille, P. T.; Jones, A. N.; Davidson, A.; Bogdanovic, J.; Arts, E.; Karn, J.; Robinson, J. A.; Varani, G. An ultra-high affinity ligand of HIV-1 TAR reveals the RNA structure recognized by P-TEFb. *Nucleic Acids Res.* **2019**, *47* (3), 1523-1531. DOI: https://doi.org/10.1093/nar/gky1197.

(24) Fulle, S.; Christ, N. A.; Kestner, E.; Gohlke, H. HIV-1 TAR RNA Spontaneously undergoes relevant apo-to-holo conformational transitions in molecular dynamics and constrained geometrical simulations. *J. Chem. Inf. Model.* **2010**, *50* (8), 1489-1501. DOI: https://doi.org/10.1021/ci100101w.

(25) Levintov, L.; Vashisth, H. Role of conformational heterogeneity in ligand recognition by viral RNA molecules. *Phys. Chem. Chem. Phys.* **2021**, *23* (19), 11211-11223. DOI: https://doi.org/10.1039/d1cp00679g.

(26) Panei, F. P.; Gkeka, P.; Bonomi, M. Identifying small-molecules binding sites in RNA conformational ensembles with SHAMAN. *Nat. Commun.* **2024**, *15* (1). DOI: https://doi.org/10.1038/s41467-024-49638-7.

(27) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E., 3rd; Laughton, C. A.; Orozco, M. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **2007**, *92* (11), 3817-3829. DOI: https://doi.org/10.1529/biophysj.106.097782.

(28) Tan, D.; Piana, S.; Dirks, R. M.; Shaw, D. E. RNA force field with accuracy

comparable to state-of-the-art protein force fields. *P. Natl. Acad. Sci. USA* **2018**, *115* (7), E1346-E1355. DOI: https://doi.org/10.1073/pnas.1713027115.

(29) Röder, K.; Pasquali, S. Assessing RNA atomistic force fields via energy landscape explorations in implicit solvent. *Biophys. Rev-Ger.* **2024**, *16* (3), 285-295. DOI: https://doi.org/10.1007/s12551-024-01202-9.

(30) Nithin, C.; Pilla, S. P.; Kmiecik, S. When does molecular dynamics improve RNA models? Insights from CASP15 and practical guidelines. *Comput. Struct. Biotec.* **2025**, *27*, 4201-4211. DOI: https://doi.org/10.1016/j.csbj.2025.10.003.

(31) Liebl, K.; Zacharias, M. The development of nucleic acids force fields: From an unchallenged past to a competitive future. *Biophys. J.* **2023**, *122* (14), 2841-2851. DOI: https://doi.org/10.1016/j.bpj.2022.12.022.

(32) Condon, D. E.; Kennedy, S. D.; Mort, B. C.; Kierzek, R.; Yildirim, I.; Turner, D. H. Stacking in RNA: NMR of four tetramers benchmark molecular dynamics. *J. Chem. Theory Comput.* **2015**, *11* (6), 2729-2742. DOI: https://doi.org/10.1021/ct501025q.

(33) Mlynsky, V.; Kührová, P.; Pykal, M.; Krepl, M.; Stadlbauer, P.; Otyepka, M.; Banás, P.; Sponer, J. Can we ever develop an ideal RNA force field? lessons learned from simulations of the UUCG RNA tetraloop and other systems. *J. Chem. Theory Comput.* **2025**, *21* (8), 4183-4202. DOI: https://doi.org/10.1021/acs.jctc.4c01357.

(34) Ramaswamy, V. K.; Musson, S. C.; Willcocks, C. G.; Degiacomi, M. T. Deep learning protein conformational space with convolutions and latent interpolations. *Phys. Rev. X* **2021**, *11* (1). DOI: https://doi.org/10.1103/PhysRevX.11.011052.

(35) Kacher, J.; Sokolova, O. S.; Tarek, M. A Deep Learning Approach to Uncover Voltage-Gated Ion Channels' Intermediate States. *J. Phys. Chem. B* **2024**, *128* (36), 8724-8736. DOI: https://doi.org/10.1021/acs.jpcb.4c03182.

(36) Frank, J. The ribosome - a macromolecular machine par excellence. *Chem. Biol.* **2000**, *7* (6), R133-R141. DOI: https://doi.org/10.1016/S1074-5521(00)00127-7.

(37) Ueno, T.; Taguchi, H.; Tadakuma, H.; Yoshida, M.; Funatsu, T. GroEL mediates protein folding with a two successive timer mechanism. *Mol. Cell.* **2004**, *14* (4), 423-434. DOI: https://doi.org/10.1016/S1097-2765(04)00261-8.

(38) Xue, Y.; Gracia, B.; Herschlag, D.; Russell, R.; Al-Hashimi, H. M. Visualizing the formation of an RNA folding intermediate through a fast highly modular secondary structure switch. *Nat. Commun.* **2016**, *7*. DOI: https://doi.org/10.1038/ncomms11768.

(39) Catterall, W. A.; Wisedchaisri, G.; Zheng, N. The conformational cycle of a prototypical voltage-gated sodium channel. *Nat. Chem. Biol.* **2020**, *16* (12), 1314-1320. DOI: https://doi.org/10.1038/s41589-020-0644-4.

(40) Shi, H. L.; Rangadurai, A.; Abou Assi, H.; Roy, R.; Case, D. A.; Herschlag, D.; Yesselman, J. D.; Al-Hashimi, H. M. Rapid and accurate determination of atomistic RNA dynamic ensemble models using NMR and structure prediction. *Nat. Commun.* **2020**, *11* (1). DOI: https://doi.org/10.1038/s41467-020-19371-y.

(41) Steinbrecher, T.; Latzer, J.; Case, D. A. Revised AMBER parameters for bioorganic phosphates. *J. Chem. Theory Comput.* **2012**, *8* (11), 4405-4412. DOI: https://doi.org/10.1021/ct300613v.

(42) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple Potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79* (2), 926-935, Article. DOI: https://doi.org/10.1063/1.445869.

(43) Kusalik, P. G.; Svishchev, I. M. The spatial structure in liquid water. *Science* **1994**, *265* (5176), 1219-1221. DOI: https://doi.org/10.1126/science.265.5176.1219.

(44) Joung, I. S.; Cheatham, T. E. Determination of alkali and halide monovalent ion

parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **2008**, *112* (30), 9020-9041, Review. DOI: https://doi.org/10.1021/jp8001614.

(45) Joung, I. S.; Cheatham, T. E., III. Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters. *J. Phys. Chem. B* **2009**, *113* (40), 13279-13290. DOI: 10.1021/jp902584c.

(46) Dethoff, E. A.; Petzold, K.; Chugh, J.; Casiano-Negroni, A.; Al-Hashimi, H. M. Visualizing transient low-populated structures of RNA. *Nature* **2012**, *491* (7426), 724-+. DOI: https://doi.org/10.1038/nature11498.

(47) Lee, J.; Dethoff, E. A.; Al-Hashimi, H. M. Invisible RNA state dynamically couples distant motifs. *P. Natl. Acad. Sci. USA* **2014**, *111* (26), 9485-9490. DOI: https://doi.org/10.1073/pnas.1407969111.

(48) Geng, A.; Ganser, L.; Roy, R.; Shi, H. L.; Pratihar, S.; Case, D. A.; Al-Hashimi, H. M. An RNA excited conformational state at atomic resolution. *Nat. Commun.* **2023**, *14* (1). DOI: https://doi.org/10.1038/s41467-023-43673-6.

(49) Roy, R.; Geng, A. A.; Shi, H. L.; Merriman, D. K.; Dethoff, E. A.; Salmon, L.; Al-Hashimi, H. M. Kinetic Resolution of the Atomic 3D Structures Formed by Ground and Excited Conformational States in an RNA Dynamic Ensemble. *J. Am. Chem. Soc.* **2023**, *145* (42), 22964-22978. DOI: https://doi.org/10.1021/jacs.3c04614.

(50) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81* (8), 3684-3690. DOI: https://doi.org/10.1063/1.448118.

(51) *Amber 22*; University of California, San Francisco, 2022. http://ambermd.org/.

(52) Le Grand, S.; Gotz, A. W.; Walker, R. C. SPFP: Speed without compromise-A mixed precision model for GPU accelerated molecular dynamics simulations. *Comput. Phys.*

Commun. **2013**, *184* (2), 374-380. DOI: https://doi.org/10.1016/j.cpc.2012.09.022.

(53) Stasiewicz, J.; Mukherjee, S.; Nithin, C.; Bujnicki, J. M. QRNAS: software tool for refinement of nucleic acid structures. *BMC. Struct. Biol.* **2019**, *19*. DOI: https://doi.org/10.1186/s12900-019-0103-1.

(54) Xiong, P.; Wu, R. B.; Zhan, J.; Zhou, Y. Q. Pairing a high-resolution statistical potential with a nucleobase-centric sampling algorithm for improving RNA model refinement. *Nat. Commun.* **2021**, *12* (1). DOI: https://doi.org/10.1038/s41467-021-23100-4.

(55) Durrant, J. D.; de Oliveira, C. A. F.; McCammon, J. A. POVME: An algorithm for measuring binding-pocket volumes. *J. Mol. Graph. Model.* **2011**, *29* (5), 773-776. DOI: https://doi.org/10.1016/j.jmgm.2010.10.007.

(56) Durrant, J. D.; Votapka, L.; Sorensen, J.; Amaro, R. E. POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J. Chem. Theory. Comput.* **2014**, *10* (11), 5047-5056. DOI: https://doi.org/10.1021/ct500381c.

(57) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J Mol Graph Model* **1996**, *14* (1), 33-38. DOI: https://doi.org/10.1016/0263-7855(96)00018-5.

(58) Sun, L. Z.; Jiang, Y. W.; Zhou, Y. Z.; Chen, S. J. RLDOCK: A New Method for Predicting RNA-Ligand Interactions. *J. Chem. Theory. Comput.* **2020**, *16* (11), 7173-7183. DOI: https://doi.org/10.1021/acs.jctc.0c00798.

(59) Buttenschoen, M.; Morris, G. M.; Deane, C. M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem. Sci.* **2024**, *15* (9), 3130-3139. DOI: https://doi.org/10.1039/d3sc04185a.

(60) Stefaniak, F.; Bujnicki, J. M. AnnapuRNA: A scoring function for predicting RNA-small molecule binding poses. *Plos. Comput. Biol.* **2021**, *17* (2). DOI:

https://doi.org/10.1371/journal.pcbi.1008309.

(61) Gasparíková, D.; Chikhale, R.; Cole, J.; Pohl, E. Recent computational advances in the identification of cryptic binding sites for drug discovery. *Bioinform. Adv.* **2025**, *5* (1). https://doi.org/10.1093/bioadv/vbaf156.

(62) Oleinikovas, V.; Saladino, G.; Cossins, B. P.; Gervasio, F. L. Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. *J. Am. Chem. Soc.* **2016**, *138* (43), 14257-14263. DOI: https://doi.org/10.1021/jacs.6b05425.

(63) Kuzmanic, A.; Bowman, G. R.; Juarez-Jimenez, J.; Michel, J.; Gervasio, F. L. Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. *Acc. Chem. Res.* **2020**, *53* (3), 654-661. DOI: https://doi.org/10.1021/acs.accounts.9b00613.

(64) Smith, R. D.; Carlson, H. A. Identification of Cryptic Binding Sites Using MixMD with Standard and Accelerated Molecular Dynamics. *J. Chem. Inf. Model.* **2021**, *61* (3), 1287-1299. DOI: https://doi.org/10.1021/acs.jcim.0c01002.

(65) Kretsch, R. C.; Andersen, E. S.; Bujnicki, J. M.; Chiu, W.; Das, R.; Luo, B.; Masquida, B.; McRae, E. K. S.; Schroeder, G. M.; Su, Z.; et al. RNA target highlights in CASP15: Evaluation of predicted models by structure providers. *Proteins* **2023**, *91* (12), 1600-1615. DOI: https://doi.org/10.1002/prot.26550.

(66) Li, Z.; Mu, J.; Chen, J.; Chen, H. F. Base-specific RNA force field improving the dynamics conformation of nucleotide. *Int. J. Biol. Macromol.* **2022**, *222* (Pt A), 680-690. DOI: https://doi.org/10.1016/j.ijbiomac.2022.09.183.

(67) He, W.; Naleem, N.; Kleiman, D.; Kirmizialtin, S. Refining the RNA Force Field with Small-Angle X-ray Scattering of Helix-Junction-Helix RNA. *J. Phys. Chem. Lett.* **2022**, *13* (15), 3400-3408. DOI: https://doi.org/10.1021/acs.jpclett.2c00359.

(68) Yuan, Y.; Fu, S.; Huo, D.; Su, W.; Zhang, R.; Wei, J. Multipolar electrostatics for

hairpin and pseudoknots in RNA: Improving the accuracy of force field potential energy function. *J. Comput. Chem.* **2021**, *42* (11), 771-786. DOI: https://doi.org/10.1002/jcc.26497.

(69) Salsbury, A. M.; Lemkul, J. A. Recent developments in empirical atomistic force fields for nucleic acids and applications to studies of folding and dynamics. *Curr. Opin. Struct. Biol.* **2021**, *67*, 9-17. DOI: https://doi.org/10.1016/j.sbi.2020.08.003 From NLM Medline.

(70) Bergonzo, C.; Henriksen, N. M.; Roe, D. R.; Cheatham, T. E., 3rd. Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields. *RNA* **2015**, *21* (9), 1578-1590. DOI: https://doi.org/10.1261/rna.051102.115.

(71) Kührová, P.; Mlynsky, V.; Zgarbová, M.; Krepl, M.; Bussi, G.; Best, R. B.; Otyepka, M.; Sponer, J.; Banás, P. Improving the Performance of the Amber RNA Force Field by Tuning the Hydrogen-Bonding Interactions. *J. Chem. Theory Comput.* **2019**, *15* (5), 3288-3305. DOI: https://doi.org/10.1021/acs.jctc.8b00955.

(72) Mlynsky, V.; Kührová, P.; Kühr, T.; Otyepka, M.; Bussi, G.; Banás, P.; Sponer, J. Fine-Tuning of the AMBER RNA Force Field with a New Term Adjusting Interactions of Terminal Nucleotides. *J. Chem. Theory Comput.* **2020**, *16* (6), 3936-3946. DOI: https://doi.org/10.1021/acs.jctc.0c00228.

(73) Cimermancic, P.; Weinkam, P.; Rettenmaier, T. J.; Bichmann, L.; Keedy, D. A.; Woldeyes, R. A.; Schneidman-Duhovny, D.; Demerdash, O. N.; Mitchell, J. C.; Wells, J. A.; et al. CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *J. Mol. Biol.* **2016**, *428* (4), 709-719. DOI: https://doi.org/10.1016/j.jmb.2016.01.029.

(74) Meller, A.; Bhakat, S.; Solieva, S.; Bowman, G. R. Accelerating Cryptic Pocket

Discovery Using AlphaFold. *J. Chem. Theory Comput.* **2023**, *19* (14), 4355-4363. DOI: https://doi.org/10.1021/acs.jctc.2c01189.

(75) Meller, A.; Ward, M.; Borowsky, J.; Kshirsagar, M.; Lotthammer, J. M.; Oviedo, F.; Ferres, J. L.; Bowman, G. R. Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. *Nat. Commun.* **2023**, *14* (1), 1177. DOI: https://doi.org/10.1038/s41467-023-36699-3.

(76) Skrhák, V.; Novotny, M.; Feidakis, C. P.; Krivák, R.; Hoksza, D. CryptoBench: cryptic protein-ligand binding sites dataset and benchmark. *Bioinformatics* **2025**, *41* (1). https://doi.org/10.1093/bioinformatics/btae745.

(77) Li, Z. X.; Zhu, J. J.; Hong, X. K.; Mu, J. X.; Zheng, Z. Q.; Cui, T. Y.; Sun, Y. T.; Wei, T.; Chen, H. F. DynaRNA: accurate dynamic RNA conformation ensemble generation with diffusion model. *Commun. Biol.* **2025**, *8* (1). https://doi.org/10.1038/s42003-025-08875-2.

(78) Zhang, H.; Liu, Y.; Liu, X. Y.; Wang, C.; Guo, M. Z. Equivariant score-based generative diffusion framework for 3D molecules. BMC Bioinformatics 2024, 25 (1), 203. https://doi.org/10.1186/s12859-024-05810-w.

(79) Nithin, C.; Kmiecik, S.; Blaszczyk, R.; Nowicka, J.; Tuszynska, I. Comparative analysis of RNA 3D structure prediction methods: towards enhanced modeling of RNA-ligand interactions. *Nucleic Acids Res.* **2024**. DOI: https://doi.org/10.1093/nar/gkae541.

(80) Xu, M.; Yu, L.; Song, Y.; Shi, C.; Ermon, S.; Tang, J. GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation. In International Conference on Learning Representations, 2022.

(81) Schneuing, A.; Harris, C.; Du, Y. Q.; Didi, K.; Jamasb, A.; Igashov, I.; Du, W. T.; Gomes, C.; Blundell, T. L.; Lio, P.; et al. Structure-based drug design with equivariant

diffusion models. *Nat. Comput. Sci.* **2024**, *4* (12). DOI: https:// doi.org/10.1038/s43588-024-00737-x.

(82) Veenbaas, S. D.; Koehn, J. T.; Irving, P. S.; Lama, N. N.; Weeks, K. M. Ligand-binding pockets in RNA and where to find them. *Proc. Natl. Acad. Sci*. U. S. A. **2025**, 122 (17), e2422346122. DOI: https://doi.org/10.1073/pnas.2422346122.

(83) Zhu, J. T.; Gu, Z. H.; Pei, J. F.; Lai, L. H. DiffBindFR: an SE(3) equivariant network for flexible protein-ligand docking. *Chem. Sci.* **2024**, *15* (21). DOI: https://doi.org/10.1039/d3sc06803j.

(84) Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; et al. De novo design of protein structure and function with RFdiffusion. *Nature* **2023**, *620* (7976), 1089-1100. DOI: https://doi.org/10.1038/s41586-023-06415-8.