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. Cryptic RNA-ligand binding sites can be promising targets due to high specificity in RNA-ligand interaction and Structure Based Drug Design (SBDD) for RNA is a practical approach to achieve successful drug candidates. However, SBDD for RNA faces a technical challenge, a paucity of available tertiary structures of RNA-ligand complexes. We addressed this problem by theoretically generating RNA conformations with cryptic binding sites from *apo* RNA conformations. Recalling that molecular dynamic simulation studies have not resulted in RNA conformations with the cryptic binding site, even for the paradigmatic RNA-ligand complex, HIV-1 Transactivation Response Element (TAR) binding to MV-2003, we leveraged sampling performance of *apo* RNA conformations by using a generative deep learning model (DL), Molearn. This model is designed to extensively search high energy states emerging in a conformational transition pathway of biomacromolecules. We succeeded to generate TAR conformations with a cryptic binding cavity for MV-2003, 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. This is the first report to theoretically generate RNA conformations with a cryptic ligand binding site, then illustrating a new application of DL to promote SBDD for RNA and the current status.

Introduction

RNA molecules are involved in a wide variety of biological processes. [1, 2, 3, 4] Their disfunctions bring about serious diseases and they have been regarded as important targets for drug discovery[5, 6, 7, 8]. Recalling that Structure-Based Drug Design (SBDD) has drawn much attention for the monumental successes in protein-target drug discovery[9, 10, 11, 12, 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 order of tens of million[15], while the number of experimentally resolved RNA-ligand complex structures is around 300 so far, thus being utterly insufficient to routinely perform cryptic binding site-targeted SBDD for RNA of interest.

Under this circumstance, theoretical prediction of RNA-ligand complexes is expected as useful counterparts. However, such a theoretical approach would not necessarily succeed, 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, 17, 18, 19, 20, 21, 22, 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 studies to try generating *MV2003-bindable* TAR conformations from the *apo* forms by employing molecular dynamics (MD) simulations. [24, 25, 26] However, none of these studies succeeded in generation of MV2003-bindable TAR conformations. Such disappointing results may stem from limitation of fix-charge empirical force field parameters which often overestimate Coulomb interactions[27]. Individual RNA hydrogen bonds in TAR are artificially stabilized in MD simulations, then preventing disruption of non-Watson Click base pair between C30 and A35 (*see* **Fig. 1C** and **1D**) essential to emergence of the cryptic binding cavity, within a realistic computational time.

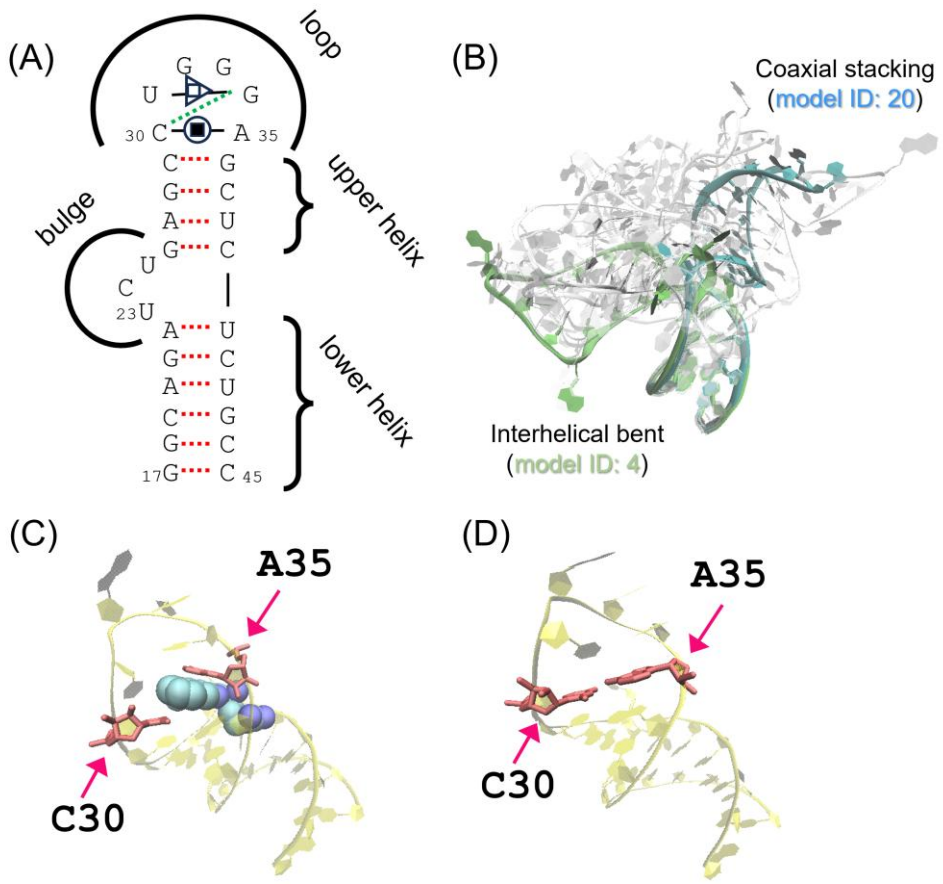

Figure 1. Structures of HIV-1 TAR. (A) Secondary structures and (B) Tertiary structure ensemble in *apo* form (PDB entry: 7JU1). (C) MV2003-bound structure (model No. 1 of PDB entry: 2L8H). (D) MV2003-inaccessible structure (model No. 1 of PDB entry: 7JU1). In panel A, models 4 and 20, a pair with the 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.

With regard to challenge to discover such cryptic binding sites by using *apo* RNA conformations, MD simulations do not work as a standalone, general methodology so far.

To overcome the inherent limitations of straightforward MD simulations, we employ a generative deep learning (DL) model, Molearn. Degiacomi and colleagues developed a convolutional neural network model[28] to systematically generate conformational conversion pathways, along which biomacromolecules pass through energetically unstable intermediate states (**Fig. 2** shows a technical outline; detailed explanation is given in **Materials and Methods**). This feature could make Molearn a promising approach to systematically generate ligand-bindable *apo* RNA conformations at a realistic computational cost. First, a ligand-bindable RNA is expected to adopt higher-energy, relatively unstable intermediate conformations in the *apo* RNA free energy landscape, due to energetic trade-offs between intra- and inter-molecular interactions in order to minimize the total system energy. Second, conformational conversions of biomacromolecules, including TAR, often take several milliseconds or longer[29, 30, 31, 32] but such chemical processes cannot be captured by MD simulations given current computational resources. By using Molearn, one could capture large conformational changes of RNA without performing long-timescale MD simulations, because its training dataset consists of MD-derived conformational ensembles at equilibrium states. Notably, the earlier Molearn studies succeeded in generating biologically relevant intermediate conformations of proteins[28, 33].

In this study, training this DL mode with MD-derived *apo* TAR conformations, we challenged to generate MV2003-bindable TAR conformations. We furthermore performed RNA-ligand docking simulations to test whether such MV2003-bindable conformations can actually take MV2003-binding poses with reasonable ligand-docking score. This deep learning challenge would be a new paradigm to practical way to generate RNA conformations with cryptic ligand-binding sites without *a priori knowledge* of

RNA-ligand complex structures.

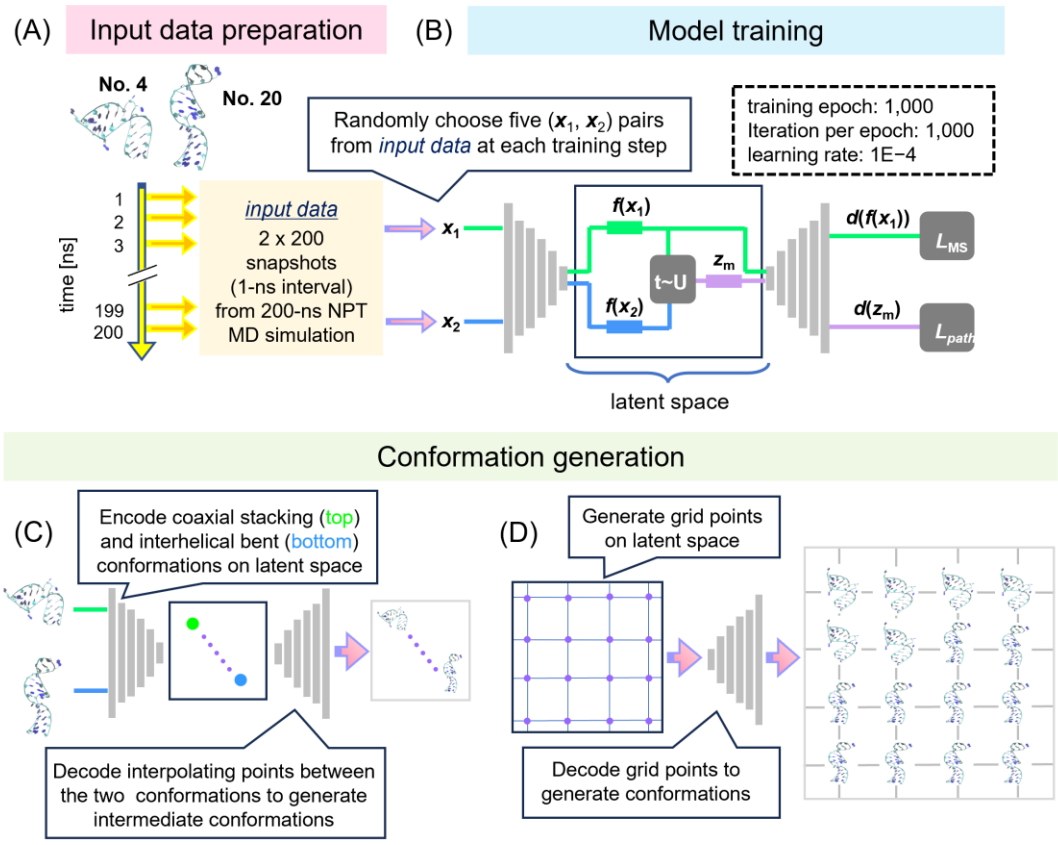

Figure 2. Procedure of Molearn model training and conformation generation. (A) Preparation of input dataset with MD simulations for models No. 4 and No. 20 (PDB entry: 7JU1). (B) Molearn model training. (C) Conformation generation by interpolating two representative conformations, a pair of MD snapshot 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.

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[34] (*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 force field 99 OL3 with backbone phosphate modification[35], TIP3P water model[36, 37], and Joung/Cheatham ion parameters adjusted for the TIP3P water model[38, 39] 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[40, 41, 42, 43], 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 force field 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[44] 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[45] GPU-version PMEMD module based on SPFP algorism[46] 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 force field parameter files[35]. The Molearn model for this RNA study consists of twelve layers (six encoder and six decorder layers) (**Fig. 2B**)[28]. 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[28]. This 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 (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 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[28]. Empirical force field parameters are derived from AMBER force field 99 OL3 with backbone

phosphate modification[35].

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[47]. 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[45]. 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 the RNA-BRiQ potential[48] 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[49, 50]. Molecular structures are illustrated by using Visual Molecular Dynamics (VMD) [51]. RNA-Ligand docking simulations are performed by using RLDOCK[52] and evaluate docking poses by using PoseBusters [53] and AnapuRNA scoring function[54]. 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 for this study, although the earlier studies have not reported it. [21]

First, we tested the conjecture that RNA 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 statistical potential energy (BRiQ potential energy) [48]. The energies of TAR conformations in the MV2003-bound state (red squares in **Fig. 3A**) always overtake those of *apo* conformations (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.

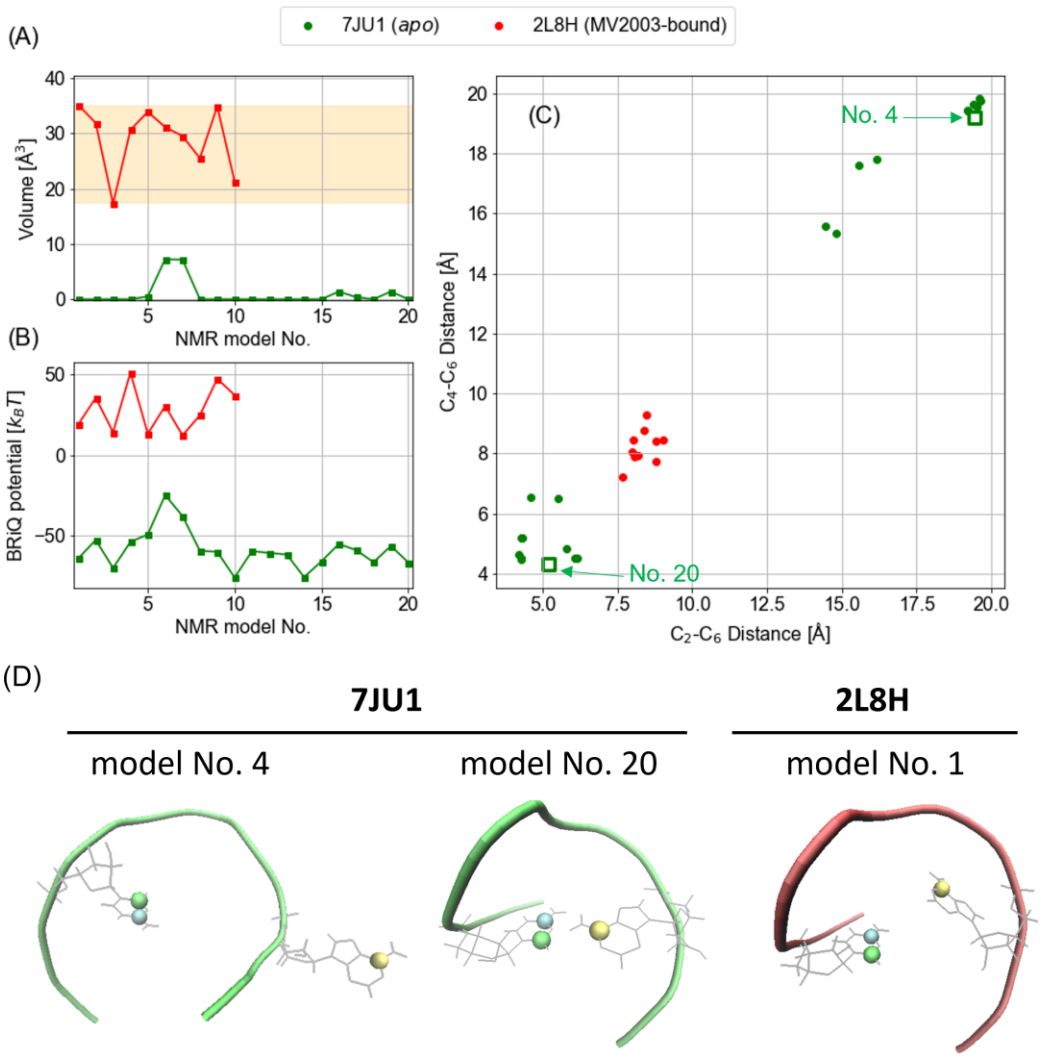

Figure 3. Physicochemical characterization of NMR-derived HIV-1 TAR conformations (*apo* and MV2003-bound forms annotated by PDB entries of 7JU1 and 2L8H, respectively). (A) RNA-BRiQ statistical potential energy. (B) Volume of (potential) MV2003-binding cavity. (C) Interatomic distances characterizing 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 of MV2003-binding cavity and BRiQ potential energy 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 comprehensively generate TAR conformations with Molearn models and then 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 systematic 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 conjecture for formation of the cryptic binding cavity.

A relative position between MV2003-interacting RNA bases is also important for the binding to MV2003. The disruption of the RNA base pair results in the MV2003-binding cavities (**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.

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.

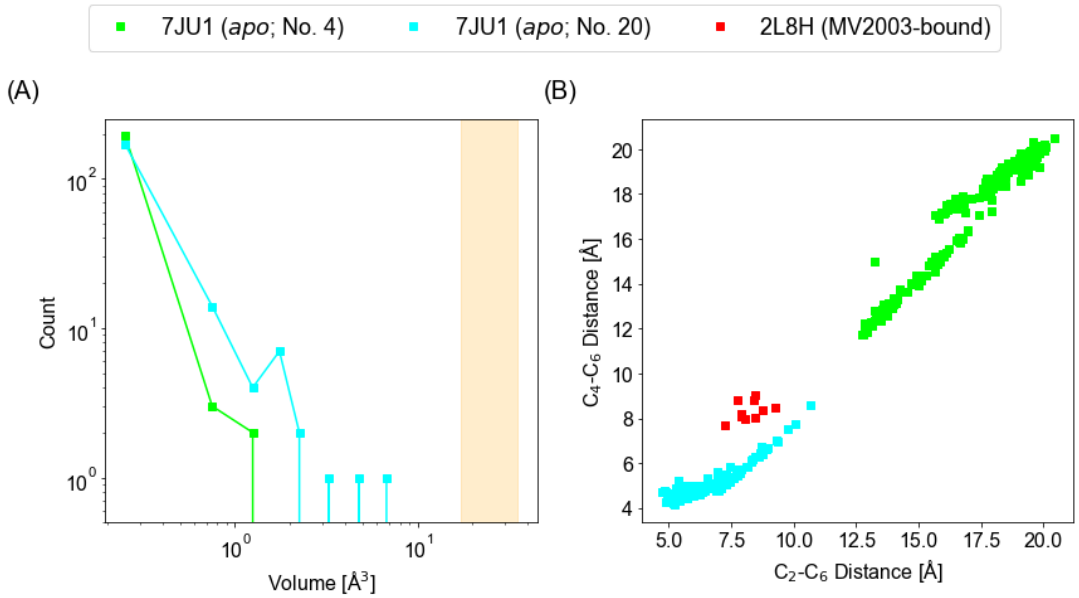

Figure 4. Geometrical properties of MD derived-Molearn input data. (A) Volume of (potential) MV2003-binding cavity. (B) A pair of interatomic distances characterizing positional relationship between C30 and A35. Green and blue squares are for snapshot structures obtained from MD simulations starting 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**).

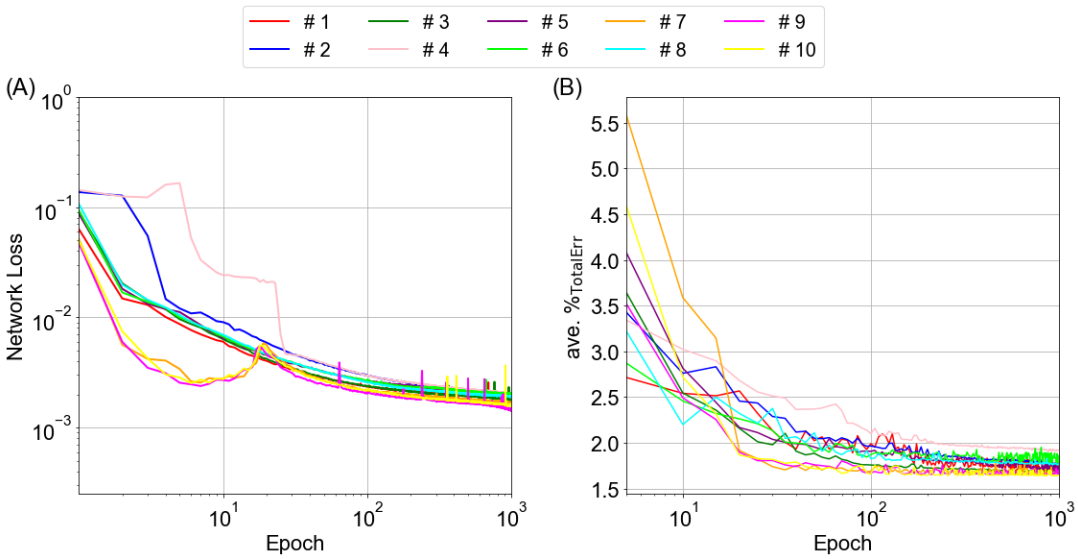

Figure 5. Evaluation of trained Molean model. (A) Network loss along training epoch. Value at each epoch is an average over 1,000 training steps in the epoch. (B) Atomic precision of Molearn-generated conformations along training epoch. #1 to #10 in the legend box denote the ten independently trained 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.

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 with potential MV2003-binding cavity (**Table 1**) and we obtained 61 MV2003-bindable TAR conformations in total. As remarked above, the training dataset does not have any conformations with the cavity (*cf.* **Fig. 4A**), so that it could be said that Molearn has a potential to predict *unprecedent* conformations such as TAR with the cryptic binding cavity. Of the 61 conformations, we found a 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).

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	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.

Before we test the MV2003-binding capability for the 61 Molearn-generated conformations in the following subsection, it is worth noting an issue for the present Molearn from practical point of view. As suggested in **Table 1**, the performance of conformation generation is not deterministic, but instead variable across trained model. Among the seven Molearn models, the number of MV2003-bindable conformations varies from 1 to 19. However, this metric appears to be irrelevant to structural similarity to the NMR-derived conformations in the MV2003-bound state: RMSd does not correlate with the number of generated structures (correlation coefficient between the number of

conformations and minimum RMSd: −0.31). This variance in Molearn performance may partially result from the probabilistic nature of its training protocols. For example, optimizing batch sizes might be worthwhile to make Molearn model performance more deterministic than the models trained here. Nonetheless, the primary interest of this study is to generate the cryptic MV2003 binding pocket of TAR. Addressing this issue lies beyond the scope of this study and is left for future work.

In summary, the current version of Molearn has the potential to predict important RNA conformations such as MV2003-bindable TAR conformations. This result suggests that the Molearn approach outperforms the early MD studies[24, 25, 26] in generating MV2003-bindable TAR conformations with the cryptic binding cavity (a more detailed discussion is provided later). Meanwhile, it is an unresolved concern that the models are not deterministic. For a similar research purpose, employing multiple trained models could be a practical way to obtain RNA conformations of interest.

Molearn models generate TAR conformation capable of binding MV2003.

It is elusive yet whether each of the 61 molearn-generated TAR conformations (*see* **Table 1**) is actually MV2003-bindable. We then verified this 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. These were ranked by the AnnapuRNA scoring function[54], and the top-ranked pose was selected as the representative. We finally obtained 610 docked TAR-MV2003 poses. Furthermore, they are evaluated by PoseBusters [53] 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 are similarly stable compared with 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 docked TAR-MV2003 poses comparative to the experimentally resolved ones in terms of AnnapRNA score.

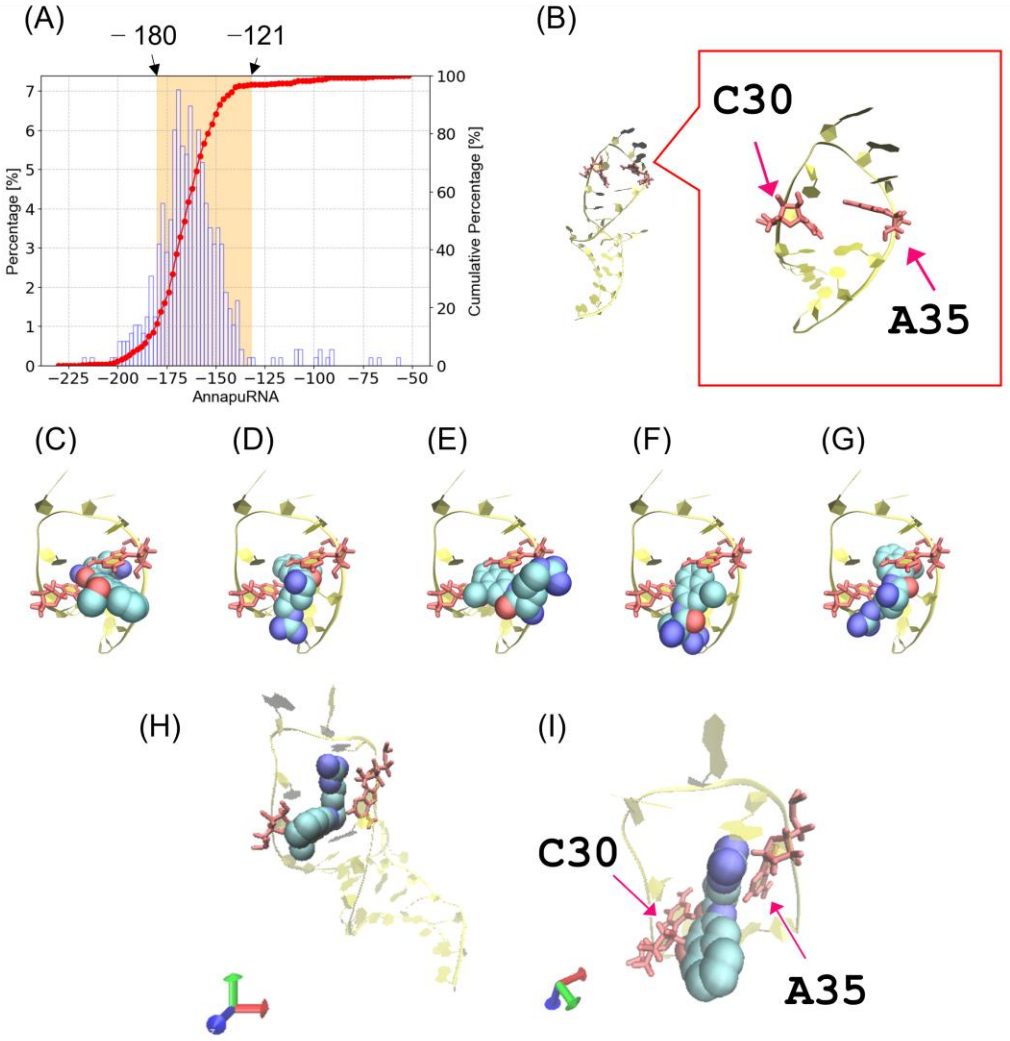

Figure 6. Analyses of docked TAR-MV2003 pose. (A) RNA-ligand binding scores calculated by AnnapuRNA for all docked poses. (B) Molearn generated TAR most similar to NMR-derived TAR in the MV2003-bound state. MV-2003 interaction site is magnified in the red frame. (C)-(G) docked poses given from the conformation shown in (A). (H) Global conformation of binding pose. (I) Closing view of TAR-MV2003 binding site. In panel A, the orange belt denotes the value range for NMR-derived TAR-MV2003 complexes. Blue open bars and red closed circles are for percentages and cumulated percentage of the histogram, respectively. In panels B and C, MV2003-binding pose with the minimum AnnapuRNA is illustrated. 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**.

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 space 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 due to steric clashes). 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-8** 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. [55] Atomistic MD simulations have been employed to search cryptic binding sites inside proteins[56, 57, 58], 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, they could not generate MV2003-bindable tertiary conformations with the cryptic binding cavity by their MD-based sampling. Meanwhile, we obtained MV2003-bindable TAR conformations with Molearn approach and then suppose that our approach outperforms their MD approach with regard to generation of TAR conformations with the cryptic binding site.

Notably, Molearn is designed to train with a set of MD-derived snapshot structures, sampled around equilibrium states. It is thus expected that Molearn approach is relatively insensitive to selection of empirical force field compared with Panai's approach. As it has been discussed long time, performance of MD simulations strongly depends on empirical force field parameters used to simulate atomic dynamics[59, 60, 61, 62, 63, 64, 65]. In Panei's study, RNA conformational searches possibly depend on combinations of force field parameters for RNA, probe ligands to promoting formation of ligand binding sites, water molecules and ions, while it is uneasy to test all possible combinations to find optimal simulation conditions. It is a further essential concern that conventional charge-fixed force field parameters artificially stabilize individual RNA base pairs, thereby preventing formation of cryptic binding sites, as discussed in **Introduction**. 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. [59, 65] This is because we do not have to consider remarkable conformational changes in the framework of MD simulations. Thus, we suppose that the Molearn approach is not markedly influenced by selected empirical force field parameters, thus possibly being available to generate functional RNA conformations without careful adjustments of empirical force field combinations.

From wider viewpoints of study on 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. [66, 67, 68, 69] 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 ca 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.

Besides site prediction approach with DL, there is also a straightforward DL study to predict ligand-binding RNA structures, which aims to ligand binding structures and is not restricted to discover cryptic binding site. Nithin and colleagues comprehensively examined performance of DL-based RNA 3D structure prediction methods[70]. They reported that several predicted RNA conformations are similar to those experimentally resolved in terms of RMSd and these conformations could bind to respective ligand molecules. However, they do not consider the TAR in the MV2003-bound state. Even the state-of-the-art RNA structure prediction methods could not accurately predict RNA

conformations with cryptic binding sites. At this point, our Molearn approach could have technical advantage, overcoming practical limitation of RNA structure prediction methods of today.

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* RNA conformation, our Molearn approach first solve it and may outperform other MD and DL approaches. Recalling such a milestone which we reached in this study, we remark remaining challenges of DL-based generation of functional RNA conformations.

Notably, we succeeded to generate TAR conformations with the cryptic binding site and to reproduce local interaction acting between TAR and MV2003. DL approaches are often less accurate to describe local interactions[70], so that generation of the cryptic binding cavity shows technical advantage of Molearn. 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 (colored by blue in **Fig.1B**), while an NMR-derived TAR in complex with MV2003 takes those similar to interhelical bent conformation (colored by green in **Fig. 1B**; the large RMSd values discussed in **Table 1** may be derived from difference in the global conformations). 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 with the cryptic binding

cavity. This technical problem may be due to the present DL framework. Meanwhile, it is also important as practical applications that we can systematically detect functional RNA conformations from whole of Molearn-generated ones in the framework of Molearn, although this remains to be an unsolved challenge so far (see **SI-9** in Supporting Information). These observations encourage us to use NL frameworks with more predictive power in future improvement of Molearn.

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 (*see* **SI-10** and **SI-11**, also **Fig**. **S10A**). As well as TAR's cryptic binding cavity, IRES' one is not considered in the earlier MD and NL studies by Panai and Nithin, respectively [26, 70]. This observation encourages us to generally use Molearn to generate RNA conformations with cryptic binding sites. Meanwhile, we could not obtain IRES conformations globally similar to the NMR-derived conformations in the DMA-135-bound state (*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.

Recent successful generative models targeting biomolecules employ the denoising diffusion probabilistic model (DDPM) combined with an SE(3)-equivariant architecture. [71, 72] We expect that the advantageous framework would achieve higher accuracy in describing both local molecular geometries, *e.g.* (cryptic) ligand-binding sites, and global molecular architectures, *e.g.* folds of RNA, due to its symmetry-preserving formulation.

By replacing the original variational autoencoder (VAE) framework, which employs the 1D CNN architecture, with such a state-of-the-art method, Molearn would become a more powerful generative DL model than at present. With appropriate technical updates to the Molearn model, we expect that the Molearn-based approach would become a routinely useful DL method to generate RNA conformations of arbitrary size with diverse ligand-binding sites, thereby promoting SBDD for RNA.

Concluding Remarks

Since RNA plays the vital roles in diverse biological processes[1, 2, 3, 4], it is an important challenge to develop RNA-targeted drugs with high specificity and low side effect. Upon tackling such a problem, cryptic binding sites of RNA should be promising drug targets. For that purpose, SBDD approach for RNA is a definitely pivotal but technically unfeasible approach yet, due to insufficient information of experimentally resolved tertiary RNA structures. In the present study, we challenged to theoretically predict RNA conformations with cryptic binding sites. By considering a paradigmatic RNA for study on RNA-ligand interaction, TAR, we succeeded to generate TAR conformations with the cryptic binding cavity with deep learning model, Molearn[28]. These TAR conformations can be accessed by the small ligand MV2003[21] and RNA-ligand complex scores are comparative with those for experimentally resolved TAR-MV2003 complexes. There are no earlier MD simulation and DL studies to theoretically observe the formation of TAR's cryptic binding site[24, 25, 26]. Our DL approach combined with MD simulations overcomes the technical limitation of standalone MD simulation approach to achieve extended conformational spaces of RNA, then being the first example to remark promising potential of DL models to generate cryptic ligand-bindable RNA conformations by only

using *apo* RNA ones. The results obtained from this challenge encourages us to develop common-use MD-DL combined approaches to discover functional RNA conformations toward future success of RNA-targeted SDBB.

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 and Molearn-related data 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 Biol.* **20**, 474–489 (2019).

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 Biol.* **24**, 288–304 (2023).

3. Chekulaeva, M. Mechanistic insights into the basis of widespread RNA localization. *Nat. Cell Biol.* **26**, 1037–1046 (2024).

4. Zacco, E. *et al.* RNA: the unsuspected conductor in the orchestra of macromolecular crowding. *Chem. Rev.* **124**, 4734–4777 (2024).

5. Osborne, R. J. & Thornton, C. A. RNA-dominant diseases. *Hum. Mol. Genet.* **15**, R162–R169 (2006).

6. Cooper, T. A., Wan, L. & Dreyfuss, G. RNA and disease. *Cell* **136**, 777–793 (2009).

7. Wahlestedt, C. Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nat. Rev. Drug Discov.* **12**, 433–446 (2013); correction **13**, (2014).

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. Target. Ther.* **9**, (2024).

9. Anderson, A. C. The process of structure-based drug design. *Chem. Biol.* **10**, 787–797 (2003).

10. von Itzstein, M. The war against influenza: discovery and development of sialidase inhibitors. *Nat. Rev. Drug Discov.* **6**, 967–974 (2007).

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* **126**, 6599–6607 (2022).

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* **43**, 2229–2243 (2024).

13. Weller, J. A. & Rohs, R. Structure-based drug design with a deep hierarchical generative model. *J. Chem. Inf. Model.* **64**, 6450–6463 (2024).

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.* **18**, 5672–5691 (2022).

15. RNAcentral Consortium. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.* **49**, D212–D220 (2021).

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. *Nucleic Acids Res.* **24**, 4598 (1996).

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.* **275**, 20660–20666 (2000).

18. Du, Z. H., Lind, K. E. & James, T. L. Structure of TAR RNA complexed with a Tat–TAR interaction nanomolar inhibitor identified by computational screening. *Chem. Biol.* **9**, 707–712 (2002).

19. Davis, B. *et al.* Rational design of inhibitors of HIV-1 TAR RNA through the stabilisation of electrostatic hot spots. *J. Mol. Biol.* **336**, 343–356 (2004).

20. Murchie, A. I. H. *et al.* Structure-based drug design targeting an inactive RNA conformation: exploiting the flexibility of HIV-1 TAR RNA. *J. Mol. Biol.* **336**, 625–638 (2004).

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.* **410**, 984–996 (2011).

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.* **39**, 248–256 (2011).

23. Shortridge, M. D. *et al.* An ultra-high affinity ligand of HIV-1 TAR reveals the RNA structure recognized by P-TEFb. *Nucleic Acids Res.* **47**, 1523–1531 (2019).

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.* **50**, 1489–1501 (2010).

25. Levintov, L. & Vashisth, H. Role of conformational heterogeneity in ligand recognition by viral RNA molecules. *Phys. Chem. Chem. Phys.* **23**, 11211–11223 (2021).

26. Panei, F. P., Gkeka, P. & Bonomi, M. Identifying small-molecules binding sites in

RNA conformational ensembles with SHAMAN. *Nat. Commun.* **15**, (2024).

27. Duboué-Dijon, E., Javanainen, M., Delcroix, P., Jungwirth, P. & Martinez-Seara, H. A practical guide to biologically relevant molecular simulations with charge scaling for electronic polarization. *J. Chem. Phys.* **153**, 050901 (2020).

28. 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* **11**, 031052 (2021).

29. Frank, J. The ribosome—a macromolecular machine par excellence. *Chem. Biol.* **7**, R133–R141 (2000).

30. Ueno, T., Taguchi, H., Tadakuma, H., Yoshida, M. & Funatsu, T. GroEL mediates protein folding with a two successive timer mechanism. *Mol. Cell* **14**, 423–434 (2004).

31. 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.* **7**, 11770 (2016).

32. Catterall, W. A., Wisedchaisri, G. & Zheng, N. The conformational cycle of a prototypical voltage-gated sodium channel. *Nat. Chem. Biol.* **16**, 1314–1320 (2020).

33. Kacher, J., Sokolova, O. S. & Tarek, M. A deep learning approach to uncover voltage-gated ion channels' intermediate states. *J. Phys. Chem. B* **128**, 8724–8736 (2024).

34. Shi, H. L. *et al.* Rapid and accurate determination of atomistic RNA dynamic ensemble models using NMR and structure prediction. *Nat. Commun.* **11**, 5531 (2020).

35. Steinbrecher, T., Latzer, J. & Case, D. A. Revised AMBER parameters for bioorganic phosphates. *J. Chem. Theory Comput.* **8**, 4405–4412 (2012).

36. 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.* **79**,

926–935 (1983).

37. Kusalik, P. G. & Svishchev, I. M. The spatial structure in liquid water. *Science* **265**, 1219–1221 (1994).

38. 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* **112**, 9020–9041 (2008).

39. 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* **113**, 13279–13290 (2009).

40. Dethoff, E. A., Petzold, K., Chugh, J., Casiano-Negroni, A. & Al-Hashimi, H. M. Visualizing transient low-populated structures of RNA. *Nature* **491**, 724–730 (2012).

41. Lee, J., Dethoff, E. A. & Al-Hashimi, H. M. Invisible RNA state dynamically couples distant motifs. *Proc. Natl Acad. Sci. USA* **111**, 9485–9490 (2014).

42. Geng, A. *et al.* An RNA excited conformational state at atomic resolution. *Nat. Commun.* **14**, (2023).

43. Roy, R. *et al.* Kinetic resolution of the atomic 3D structures formed by ground and excited conformational states in an RNA dynamic ensemble. *J. Am. Chem. Soc.* **145**, 22964–22978 (2023).

44. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).

45. Case, D. A. *et al. AMBER 2022*. (University of California, San Francisco, 2022).

46. 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. **184**, 374–380 (2013).

47. Stasiewicz, J., Mukherjee, S., Nithin, C. & Bujnicki, J. M. QRNAS: software tool for refinement of nucleic acid structures. *BMC Struct. Biol.* **19**, 5 (2019).

48. 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.* **12**, 5908 (2021).

49. Durrant, J. D., de Oliveira, C. A. F. & McCammon, J. A. POVME: an algorithm for measuring binding-pocket volumes. *J. Mol. Graph. Model.* **29**, 773–776 (2011).

50. 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.* **10**, 5047–5056 (2014).

51. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph. Model.* **14**, 33–38 (1996).

52. 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.* **16**, 7173–7183 (2020).

53. 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.* **15**, 3130–3139 (2024).

54. Stefaniak, F. & Bujnicki, J. M. AnnapuRNA: a scoring function for predicting RNA–small molecule binding poses. *PLoS Comput. Biol.* **17**, e1009015 (2021).

55. Gasparíková, D., Chikhale, R., Cole, J. & Pohl, E. Recent computational advances in the identification of cryptic binding sites for drug discovery. *Bioinform. Adv.* **5**, (2025).

56. 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. **138**, 14257–14263 (2016).

57. Kuzmanic, A., Bowman, G. R., Juarez-Jimenez, J., Michel, J. & Gervasio, F. L. Investigating cryptic binding sites by molecular dynamics simulations. *Acc. Chem. Res.* **53**, 654–661 (2020).

58. Smith, R. D. & Carlson, H. A. Identification of cryptic binding sites using MixMD with standard and accelerated molecular dynamics. *J. Chem. Inf. Model.* **61**, 1287–1299 (2021).

59. Liebl, K. & Zacharias, M. The development of nucleic acids force fields: from an unchallenged past to a competitive future. *Biophys. J.* **122**, 2841–2851 (2023).

60. Kretsch, R. C. *et al.* RNA target highlights in CASP15: evaluation of predicted models by structure providers. *Proteins* **91**, 1600–1615 (2023).

61. Li, Z., Mu, J., Chen, J. & Chen, H. F. Base-specific RNA force field improving the dynamics conformation of nucleotide. *Int. J. Biol. Macromol.* **222**, 680–690 (2022).

62. 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.* **13**, 3400–3408 (2022).

63. 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.* **42**, 771–786 (2021).

64. 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.* **67**, 9–17 (2021).

65. 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*

21, 1578–1590 (2015).

66. Cimermancic, P. *et al.* CryptoSite: expanding the druggable proteome by characterization and prediction of cryptic binding sites. *J. Mol. Biol.* **428**, 709–719 (2016).

67. Meller, A., Bhakat, S., Solieva, S. & Bowman, G. R. Accelerating cryptic pocket discovery using AlphaFold. *J. Chem. Theory Comput.* **19**, 4355–4363 (2023).

68. Meller, A. *et al.* Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. *Nat. Commun.* **14**, 1177 (2023).

69. Skrhák, V., Novotny, M., Feidakis, C. P., Krivák, R. & Hoksza, D. CryptoBench: cryptic protein–ligand binding sites dataset and benchmark. *Bioinformatics* **41**, (2025).

70. 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.* **52**, (2024).

71. 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.* **15**, (2024).

72. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).