Deep learning generates *apo* RNA conformations with cryptic ligand 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 drug target. Structure Based Drug Design (SBDD) for RNA is a promising approach but available RNA-ligand complex tertiary structures are substantially limited so far. Then, theoretical RNA-ligand docking simulations play central roles. However, success of SBDD highly depends on use of RNA structure sufficiently close to ligand-boundable conformations, which do not necessarily appear in experimentally-resolved and theoretically sampled *apo* RNA conformations. To overcome such difficulty in SBDD for RNA, we leverage efficiently sampling ligand-boudable *apo* RNA conformations by using a generative deep learning model (DL). We succeeded to generate HIV-1 Transactivation Response Element (TAR) conformations with a cryptic MV2003 binding cavity without *a priori* knowledge of the cavity. These conformations bind to MV2003 with binding scores similar to those calculated for experimentally-resolved TAR-MV2003 complexes, illustrating new application of DL to promote SBDD for RNA.

Introduction

RNA molecules are involved in a wide variety of biological processes. [1-4] Since their disfunctions bring about serious diseases, they have been regarded as important targets for drug discovery[5-8]. Recalling monumental successes in protein-target drug discovery[9-13], Structure-Based Drug Design (SBDD) has drawn much attention as promising approaches to efficiently find RNA-targeting drugs among abundant candidate compounds[14]. Meanwhile, available RNA-ligand complex structures are substantially restricted so far, thus often hindering RNA-targeting SBDD. Comparing an enormous number of RNA sequence information (*e.g.*, 34 million non-coding RNAs in RNAcentral [15]), the number of experimentally-resolved RNA-ligand complex structures is around 300 so far, thus being utterly insufficient to routinely perform SBDD.

Under such a circumstance, theoretical prediction of RNA-ligand complexes plays important roles as technical counterparts. However, such a theoretical prediction may be sometimes unsuccessful in evaluating druggability of candidate compounds, because we usually prepare RNA-ligand complexes by using ligand-free, *apo* RNA structures. They do not necessarily form proper ligand-binding sites, which are often arranged through ligand-binding processes[16-18]. If we use such ligand-binding-incompetent RNA structures for the theoretical prediction, we would obtain incorrect RNA-ligand bound pose,

wrongly estimate its RNA-ligand binding affinity and finally overlook promising drug candidates.

We could solve the above problem if we could systematically generate ligand-binding RNA conformations from *apo* RNA structures. Since RNA forms atomistic interaction with a ligand in expense of intra-RNA interaction, a ligand-binding RNA conformations should take energetically high conformations in energy landscape of *apo* RNA. This physicochemical observation lets us conjecture that they emerge in RNA conformation transition pathways connecting basins in an RNA energy landscape. We can test the conjecture by analyzing transition pathways of RNA conformations.

Such transition pathways could be examined by using atomistic molecular dynamics (MD) simulations, theoretically[19, 20]. Nonetheless, MD simulations do not necessarily work as a standalone methodology when we consider complex molecular systems such as biomacromolecules. Functionally important conformational changes of biomacromolecules often progress in the timescale of several-millisecond or longer[21-24], which is still difficult to access with MD simulations with available computational resources of today. Besides, the multiple nature of biological conformation transition pathways under thermal fluctuation[25-28], concerning us that MD simulations cannot necessarily find pathways involved in ligand-binding conformations with realistic

computational time.

Thus, we need to leverage available information obtained from MD simulations to more extensively search conformational space of *apo* RNA with aim of sampling ligand-binding RNA conformations. We here consider this problem by employing a generative deep learning model designed to systematically predict transition pathways of biomacromolecules, referred to as Molearn[29], a convolutional neural network model (**Fig. 1**). The earlier studies succeeded to generate biologically relevant intermediate conformations of proteins by training a Molearn model with MD-derived conformations[29, 30]. Besides the practical utility as generative model, it is worthwhile noting that Molearn reduces a complex conformational space to a finite area in a two-dimensional space by mapping atomic conformations on the latent space. Assuming that Molearn's latent spaces give simplified representations of complex molecules, we could explore conformation transition pathways on the low dimension spaces.

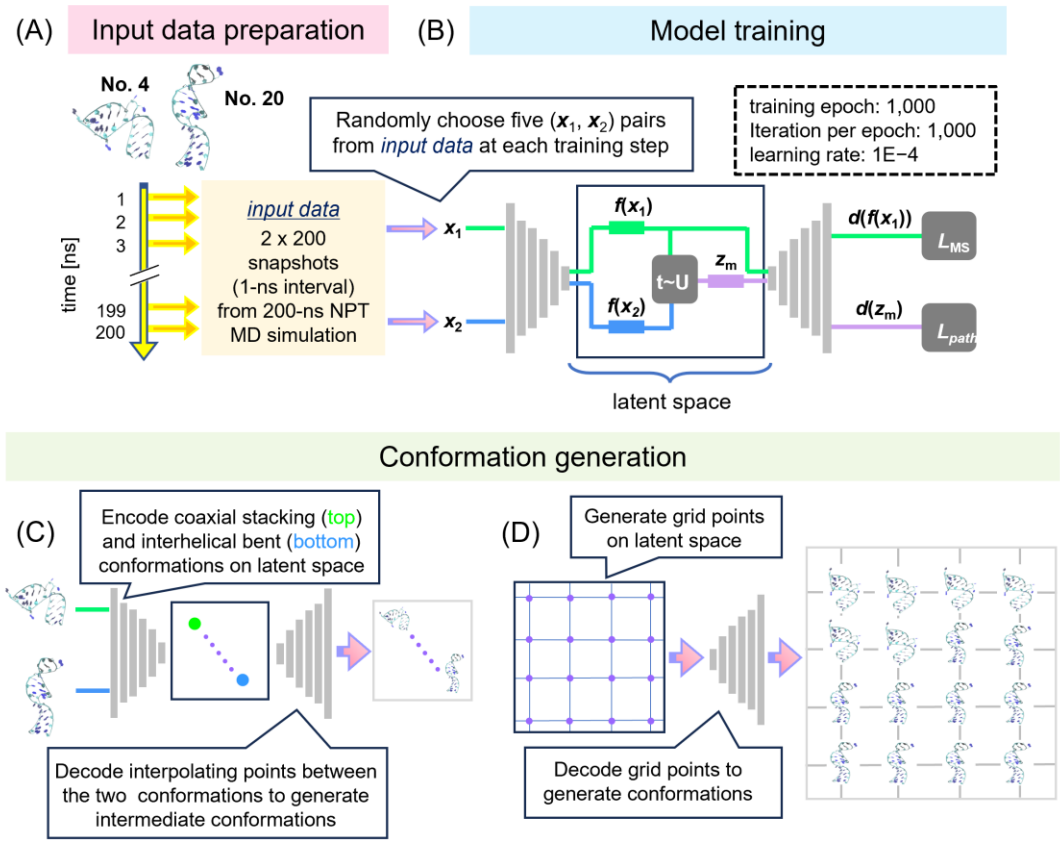

Figure 1. Procedure of Molearn model training and conformation generation. (A) Preparation of input dataset with MD simulations for models No. 4 and No. 20 (PDB entry: 7JU1). (B) Molearn model training. (C) Conformation generation by interpolating two representative conformations, a pair of MD snapshot 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.

We verify the above research strategy by considering a cryptic binding cavity appearing in *apo* HIV-1 TAR (Humam Immunodeficiency Virus type 1 Trans Activation Response

element, referred to as TAR, hereafter). TAR is the paradigmatic RNA to study RNA-ligand complex formation[31-38] (**Fig. 2A** and **2B** for the second and tertiary structures, respectively). Among several TAR-binding ligands, MV2003 (Arginine 4-methoxy-β-napthylamide) binds to TAR's cryptic binding cavity[36], whose formation accompanies a remarkable conformational change of TAR (**Fig. 2C** and **2D** for MV2003-bound and *apo* TAR conformations, respectively). As the term of *cryptic* suggests, the cavity does not appear in any of experimentally resolved *apo* TAR, and it is not shared with other ligand-binding TAR[36]. Besides, there are no atomistic simulation studies to observe the formation of the cryptic binding cavity[39-41], to best our knowledge. Therefore, our present study addresses to generate *unprecedent* MV2003-binding TAR conformations from *apo* TAR ones, and to systematically identify MV2003-boundable conformations among whole of generated conformations. We evaluate whether such a MV2003-boundable conformation has a cryptic cavity to binds MV2003 and shows reasonable ligand-docking score calculated through RNA-ligand docking simulations. This challenge would be a paradigm to practical way to generate ligand-binding RNA conformations without *a priori knowledge* of ligand-bound ones.

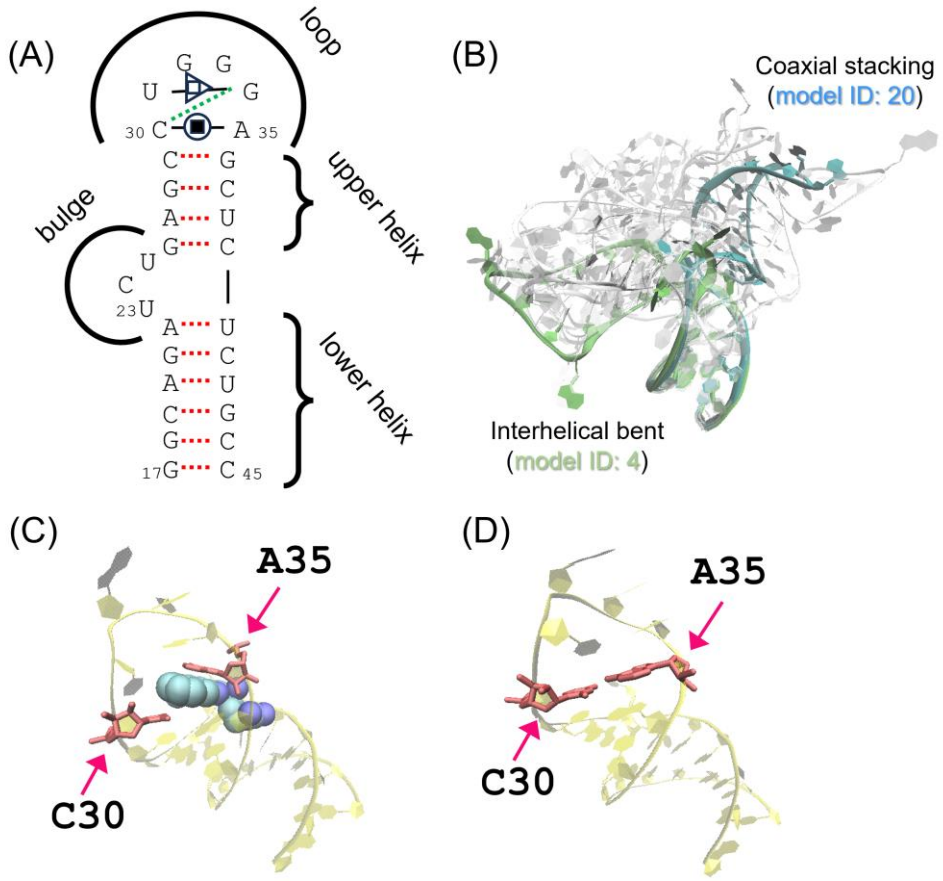

Figure 2. Structures of HIV-1 TAR. (A) Secondary structures and (B) Tertiary structure ensemble in *apo* form (PDB entry: 7JU1). (C) MV2003-binding structure (model No. 1 of PDB entry: 2L8H). (D) MV2003-inaccessible structure (model No. 1 of PDB entry: 7JU1). In panel A, models 4 and 20, a pair with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

Methods

System setup

We considered the NMR-resolved *apo* TAR structures (Protein data bank (PDB) entry ID: 7JU1) as the conformation ensemble in the ground state[42] (*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. 2B**) 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[43], TIP3P water model[44, 45], and Joung/Cheatham ion parameters adjusted for the TIP3P water model[46, 47] were applied for RNA residues, water molecules, and ions, respectively.

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. 1A**).

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[48] 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[49] GPU-version PMEMD module based on SPFP algorism[50] with NVIDIA GeForce RTX 3090. Further details for simulation procedures are given in **SI-1** 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[43]. The Molearn model for this RNA study consists of twelve layers (six encoder and six decorder layers) (**Figure 1B**) [29]. 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* **Figure 2A**). 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[29].

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 **Figure 1B**, respectively). L_{path} is atomistic potential energies for intermediate structures generated by Molearn (represented by $\boldsymbol{d}(\mathbf{z_m})$ in **Figure 1B**). The energy function consists of atomic bonding, atomic angle, dihedral angle, and approximated non-bonded interaction defined by Coulomb interactions and 12 repulsive Lennard-Jones potential[29]. Empirical force field parameters are derived from AMBER force field 99 OL3 with backbone phosphate modification[43].

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 $1e^{-4}$. 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[51]. An optimization step was repeated by 20,000 times with non-bonded interaction cutoff of 12Å under the implicit solvation model.

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[49]. 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).

A single point energy of RNA is calculated by the RNA-BRiQ potential[52] to examine an energy landscape of *apo* TAR on Molearn's latent space. A minimum energy pathway on Molearn's latent space is found by using Dijkstra's algorism. POVME2 is used to calculate atomic volume of cavity in the vicinity of C30 and A35[53, 54]. Molecular structures are illustrated by using Visual Molecular Dynamics (VMD) [55]. Ligand-docking simulations are performed by using RLDOCK[56] and evaluate docking poses by using AnapuRNA scoring function[57]. Technical details for POMVE2 and RLDOCK are given in **SI-3** and **SI-4** in Supporting Information, respectively.

Results and Discussion

Physicochemical characterization of MV2003-binding TAR conformation

We start to examine physicochemical properties of NMR-resolved TAR conformations, energetic stability of TAR conformations and geometrical condition to form MV2003-binding cavity. This information is indispensable to show validity of our research strategy and to systematically find MV2003-binding TAR conformations among Molearn-generated ones, although the earlier study has not reported it. [36]

First, we analyze energetic stability of TAR conformations to rationalize the conjecture that TAR takes the MV2003-binding conformations during transitions between two stable states. This conjecture has such a promise that TAR forms an MV2003-binding cavity with expense of energetic stability as *apo* TAR. Then, we confirm the promise by calculating the state-of-the-art RNA-specific statistical potential energy (BRiQ potential energy) [52] for each of *apo* and MV2003-bound conformations (**Fig. 3A**). We can find that the energies of MV2003-bound conformations (red squares in **Fig. 3A**) always overtake those of *apo* conformations (green squares in **Fig. 3A**). It can be said that the MV2003-bound conformations appear in relatively high energy region in the conformation space of *apo* TAR. This result makes our conjecture worthwhile to systematically search MV2003-binding TAR conformation with Molearn models.

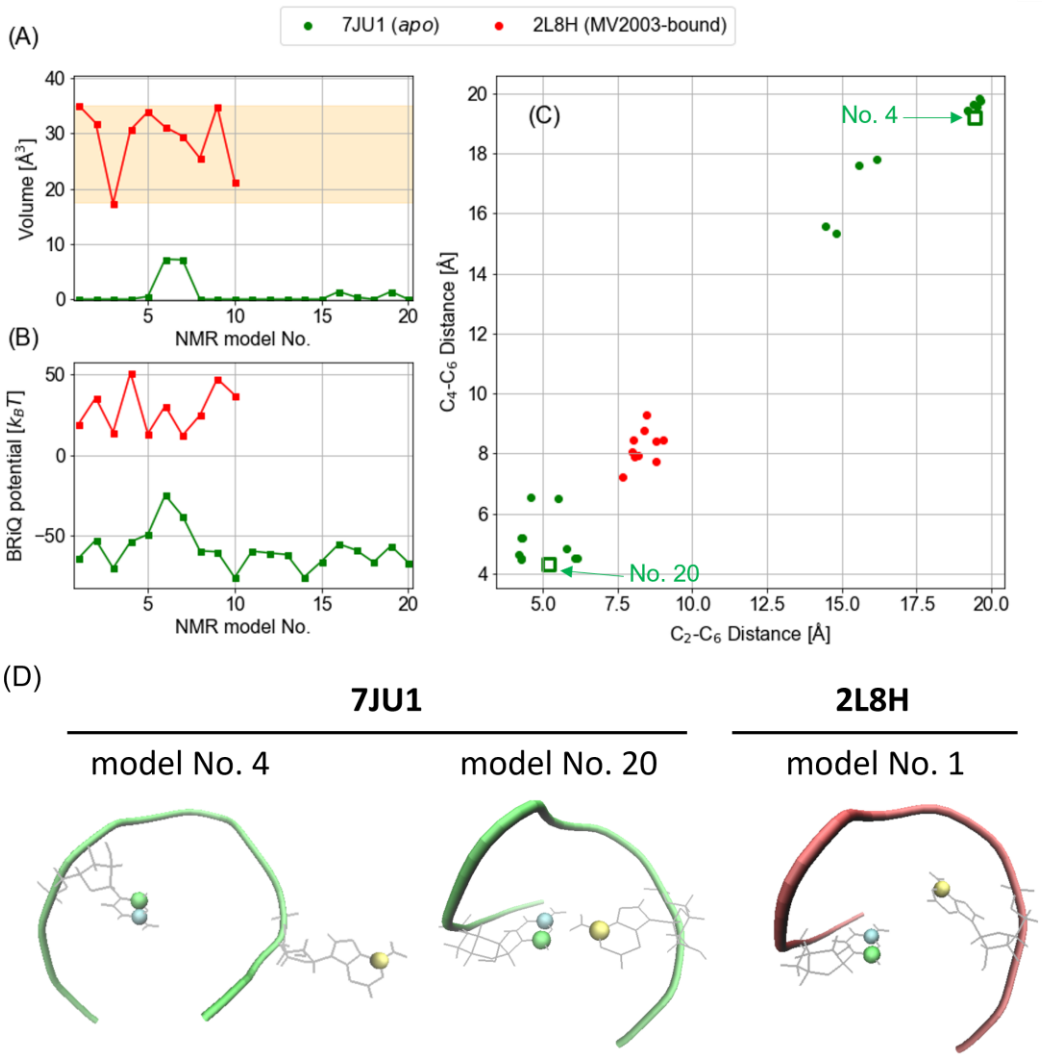

Figure 3. Physicochemical characterization of NMR-resolved HIV-1 TAR conformations (*apo* and *liganded* forms annotated by PDB entries of 7JU1 and 2L8H, respectively). (A) RNA-BRiQ statistical potential energy. (B) Volume of (potential) MV2003-binding cavity. (C) Interatomic distances characterizing positional relationship between C30 and A35. (D) Location of C30 and A35 in *apo* and *liganded* forms. In panels A and B, orange belts denote value ranges for volume of MV2003-binding cavity and BRiQ potential energy for MV2003-bound HIV-1 TAR conformations. 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 MV2003-bound TAR conformations. From viewpoint of structural biology, it is known that the MV2003-binding cavity is a cryptic, ligand-binding induced structure[36, 41]. Meanwhile, there have been no earlier studies to quantitatively characterize the cavity with geometrical metrics, to our best knowledge. In the following sections, we will comprehensively generate TAR conformations with Molearn models and then search MV2003-*boundable* ones, which have a cavity accessible for MV2003. Aiming to perform systematic search for TAR conformations of interest, we address to give geometrical conditions to form MV2003-binding cavity.

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* 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 insufficient. It is noted that their energy values 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 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. 2C** for 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 conformations. These observations denote that any of the NMR-resolved *apo* TAR conformations does not take MV2003-*boundable* configurations for Cyt30 and Ade35. For example, the representative interhelical bent and

coaxial stacking conformations (the models No. 4 and No. 20 in **Fig 2B**, 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 TAR with a potential MV2003-binding cavity, referred to as MV2003-boundable conformation hereafter, satisfies 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. 2B**), respectively. We 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. 1A**).

It is noted that any of MD-derived TAR conformations does not have 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 MV2003-bound conformations (the value range for MV2003-bound 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 MV2003-bound conformations (red squares in **Fig. 4B**). Therefore, we trained Molearn model without *a priori* knowledge for the cryptic MV2003-binding cavity.

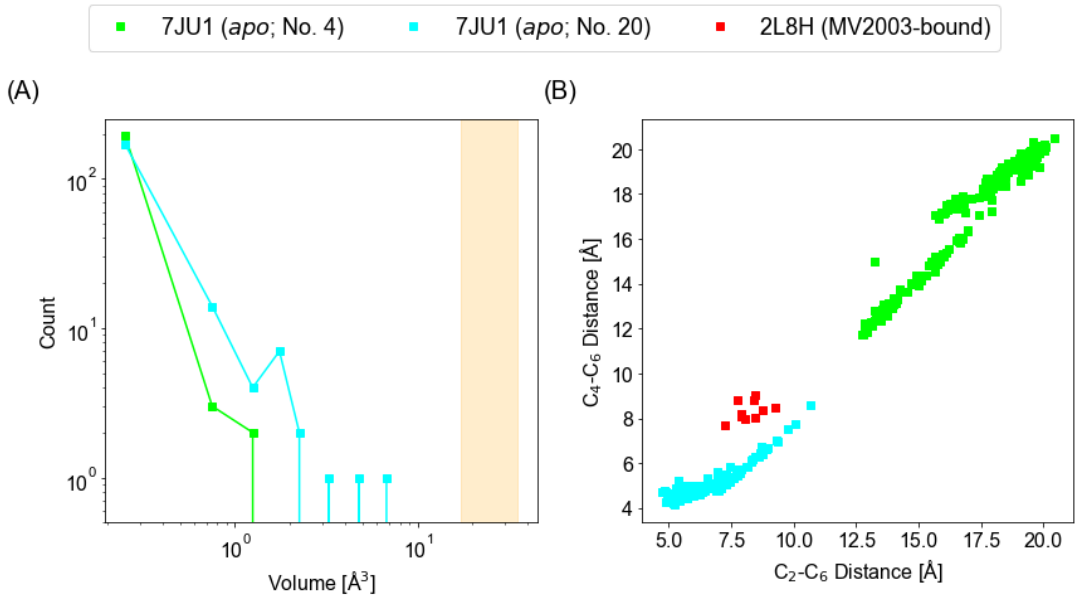

Figure 4. Geometrical properties of MD derived-Molearn input data. (A) Volume of (potential) MV2003-binding cavity. (B) A pair of interatomic distances characterizing positional relationship between C30 and A35. Green and blue squares are for snapshot structures obtained from MD simulations starting 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 MV2003-bound HIV-1 TAR conformations (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 1000th 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**).

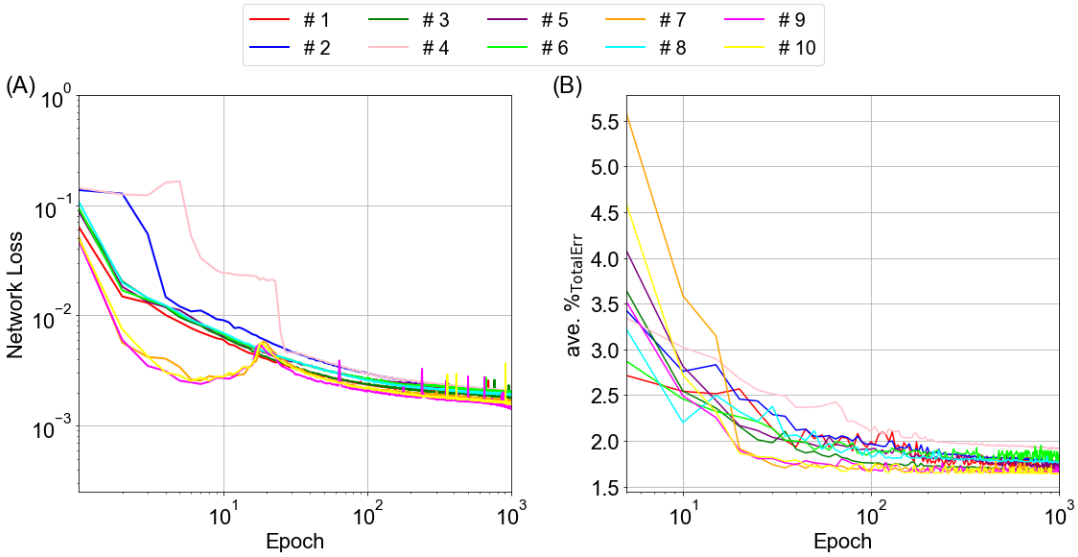

Figure 5. Evaluation of trained Molean model. (A) Network loss along training epoch. Value at each epoch is an average over 1,000 training steps in the epoch. (B) Atomic precision of Molearn-generated conformations along training epoch. #1 to #10 in the legend box denote the ten independently trained Molearn models. With regard to (B), $\%_{TotalErr}$ 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 $\%_{TotalErr}$ 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 $\%_{TotalErr}$ (**Eq. 2**). Intermediate structures were generated by interpolating the representative interhelical bent and coaxial stacking conformations (a pair with the maximum RMSd value picked up from the training dataset, which conformationally resembles the No. 4 and 20 models shown in **Fig 2B**, respectively) on the latent space given from every Molearn model (*cf.* **Fig. 1C**). We used 48 points between the two input conformations to generate intermediate 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 1000th epoch is worthwhile employing as a representative for conformation generations.

By considering the Molearn models at 1,000th epoch, we calculated RNA-BRiQ potential energy profiles along the transition pathways (**Fig. 6**). Each of the ten models often generates higher energy conformations at the interpolated points. It could be said that these Molearn models generate intermediate conformations on a transition pathway.

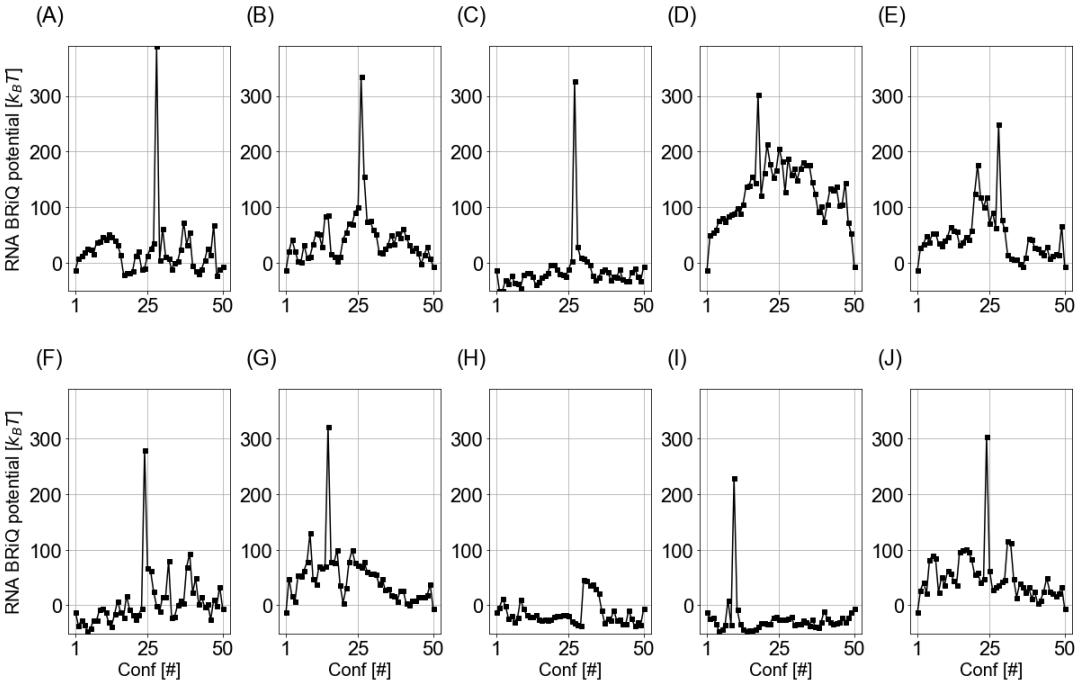

Figure 6. RNA-BRiQ potential energy for Molearn-generated conformations at the 1000th training epoch [Unit: k_BT]. Panels A to I correspond to each of the ten Molearn models, numbered from 1 to 10 (*cf.*, **Fig. 5**). Conf. [#] denotes index of TAR conformation on transition pathway: 1 and 50 are for the interhelical bent and coaxial stacking conformations (a pair with the maximum RMSd value, which is picked up from the training dataset and similar to models No. 4 and 20 shown in **Fig 2B**) and the others are for those generated by interpolating the two conformations.

Compared to those for a protein discussed in the original Molearn study[29], shapes of energy profiles are substantially jagged and volatile. This may be due to that we consider all heavy atoms in RNA to train Molearn models and that TAR undergoes global conformational changes (*cf.* **Fig. 2B**) plus remarkable rearrangements of RNA base configurations along transition pathways. Meanwhile, the earlier study only considered the four backbone atoms and one sidechain atom (C_α, C, N, O, C_β) for any amino acid

residue and relatively simple domain motions, then possibly resulting in smooth energy profiles. It is supposed that biomacromolecules inherently have rough energy profiles on their conformational transition pathways[58-60]. A complicated atomic description of biomacromolecules may give these TAR energy profiles rather than the simplified one employed in the original Molearn study[29].

Meanwhile, we found spiky changes for each of RNA-BRiQ potential energy profiles and relatively large energy values at the points (quantitatively characterized in **Table S1** in Supporting Information). This is probably due to that interpolated points are away from minimum energy pathways. The original Molearn study[29] employed a similar interpolation protocol to generate transition pathways, which may be sufficient to evaluate chemical accuracy. However, we had better to consider minimum energy pathways as transition pathways when we discuss functionally important intermediate conformations.

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

With aiming to obtain a minimum energy pathway connecting between inter helical and coaxial stacking conformations of TAR, we comprehensively generated TAR conformations from a set of grid points on the Molearn's latent space (*cf.* **Fig. 1D**). 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.

First, we confirmed the presence of MV2003-boundable 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**). As remarked above, the training dataset does not have any conformations with the cavity (*cf.* **Fig. 4A**), so that it is said that Molearn has a potential to predict unprecedent conformations such as TAR with a cryptic binding cavity.

Table 1. RMSd of Molearn generated conformations with potential MV2003-binding cavity to NMR-resolved MV2003-bound conformations (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 10 MV2003-bound conformations are used to calculate RMSd, then giving 10 RMSd values per one conformation.

Meanwhile, the performance for conformation generation is not similar among the seven models. The number of MV2003-boundable conformations varies from 1 to 19, while the abundance of interesting conformations appears to be irrelevant to structural similarity to the NMR-derived MV2003-bound conformations: RMSd does not correlate with the number of generated structures (correlation coefficient between the number of conformation and minimum RMSd: −0.31). Such variance of Molearn model's

performance might appear due to probabilistic nature of Molearn's training protocols. For example, a batch data is randomly selected from the input dataset at each training step (*cf.* **Fig 1B**). There are rooms to improve robustness of Molearn model performance. Nonetheless, addressing this issue is beyond the scope of this study, thus being left for future work. Employing multiple trained models is a practical way to generate functionally interesting conformations, so far.

We hereafter focus on the seven Molearn models. Recalling to test our conjecture for appearance of MV2003-boundable conformations in the *apo* TAR conformation space, we examined locations of MD-derived and Molearn-generated MV2003-boundable TAR conformations on the latent space. We first encoded the MD-derived TAR conformations, the training dataset, on the Molearn's latent space and analyzed spatial characteristics of their distribution. They appear in a subregion of the Molearn's latent space, which is a convex polygon enclosing the training dataset (dotted line in **Fig. 7**). The interhelical bent and coaxial stacking conformations (blue and green circles in Panels A to F in **Fig. 7**, respectively) separately distribute on the latent space. It could be said that two characteristic TAR conformations are distinguishable in our Molearn latent space.

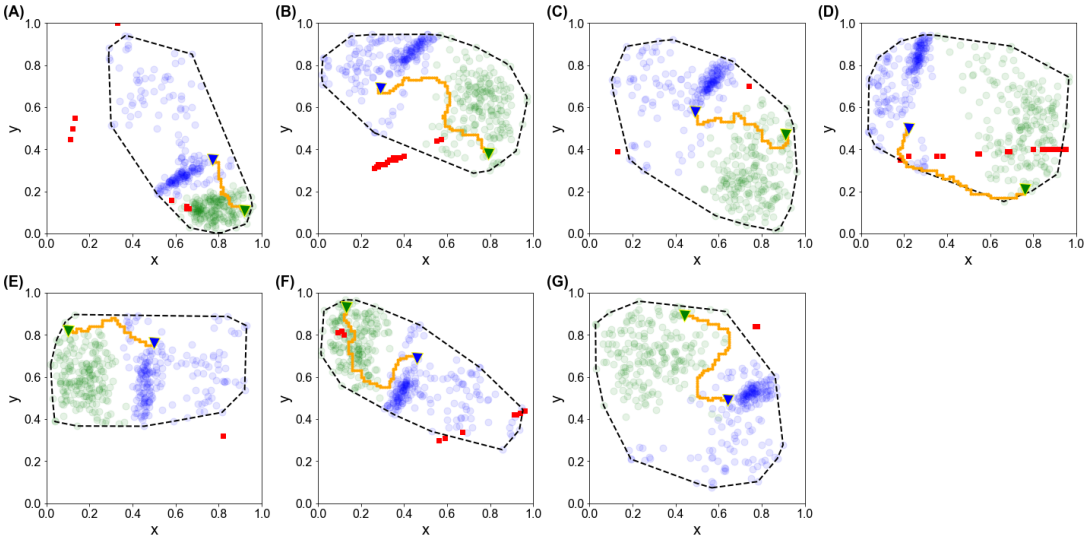

Figure 7. Mapping HIV-1 TAR conformations on Molearn's latent space derived from each of seven trained models. (A) to (G) correspond to #1 to #7 in **Fig. 5**. Red squares denote Molearn generated conformations with potential MV2003-binding cavities. Transparent bule and green circles are snapshot structures obtained from MD simulation for interhelical bent and coaxial stacking conformations (*cf.* **Fig. 2B**), respectively. Blue and green inversed triangle are for a pair of MD snapshots with the maximum RMSd value. Minimum RNA-BRiQ potential energy pathways connecting the inverse triangles are shown by orange lines. Black dotted line denotes a convex polygon covering MD-derived snapshot structures.

Our conjecture says that MV2003-boundable conformations appear in transition pathways connecting two stable conformational states. We here consider a minimum energy pathway as a transition pathway on Molearn's latent spaces. Each minimum energy pathway is generated by connecting the representative interhelical bent and coaxial stacking conformations which are a pair of MD-derived conformation with the maximum RMSd value (orange trace in each panel in **Fig. 7**).

Any of pathways is found inside the convex polygon (dotted line in **Fig. 7**). This may be explained by considering the formula of loss function (**Eq. 1**), which has both reproduction of the input dataset and physicochemical accuracy of interpolated conformations (*see* **Fig 1C**). Actually, conformations inside the convex polygon have relatively small values of $\%_{\mathrm{TotalErr}}$ and are relatively energetically stabilized in terms of RNA BRiQ potential (*see* **Tables S2** and **S3**). Compared with the interpolation-generation pathways, maximum value and spikiness of RNA-BRiQ potential energy profile are apparently decreased in these pathways (*see* **Fig. S4**, **Table S1** and **SI-5** for related discussion in Supporting Information). Then using these pathways appears to be better to test our conjecture than using the linearly interpolated pathways.

Although we expect that MV2003-boundable conformations (red squares in **Fig. 7**) appear through conformational transition between the two representative interhelical bent and coaxial stacking conformations, they are not necessarily found near these minimum energy pathways. As for the models shown in **Fig. 7B, 7D** and **7F**, a part of conformations seems to match our expectation, although the relative positions of these conformations to the corresponding minimum energy pathways are different among the three models. Thus, appearance of MV2003-boundable conformations near the minimum energy pathways is accidental and our Molearn latent spaces may be impractical to systematically discover

such conformations based on our conjecture, so far. This situation was unchanged even if we used a different pair of starting and ending points to make transition pathways (*see* **Fig S3** and **S4** and the related discussion in **SI-6** in Supporting Information). We may need to technically improve Molearn model and/or training protocols to obtain reliable free energy landscapes on the latent space.

In summary, the current version of Molearn has potential to predict unprecedent conformations such as MV2003-boundable TAR ones. Meanwhile it is still unfeasible to systematically discover such conformations by analyzing an energy landscape of *apo* TAR on the latent space. It is pursued but elusive to construct free energy landscapes of interest with biomolecular structure-generative models[61]. Expected improvement of Molearn model may be categorized as this kind of problem, which we would address to solve in future.

Molearn models generate MV2003-binding conformation.

We evaluate whether the 61 molearn-generated TAR conformations (*see* **Table 1**) can make MV2003-binding poses by using RNA-ligand docking simulations. We generate TAR-MV2003 docking poses by using each of ten sets of atomic coordinates of MV2003 in the NMR-solved complexes[36]. A set of docking poses simulated from a pair of TAR

and MV2003 was ranked by AnnapuRNA scoring function[57] and the top ranked pose is collected as a representative of the set. We finally obtained 610 TAR-MV2003 docking poses.

Fig. 8A shows distributions of AnnapuRNA scores calculated for the representative docking poses. The orange belt indicates the value range of AnnapuRNA score calculated for NMR-resolved TAR-MV2003 complexes. 86% of docking poses appear within the value range, meaning that they are similarly stable compared with the NMR-resolved TAR-MV2003 complexes. Furthermore, the 11% of docking poses show lower AnnapuRNA scores (left to the orange belt in **Fig. 8A**). These docking poses are more stable than the experimentally resolved complexes in terms of AnnapRNA score. In total, we can highly frequently obtain sufficiently stable MV2003-binding poses.

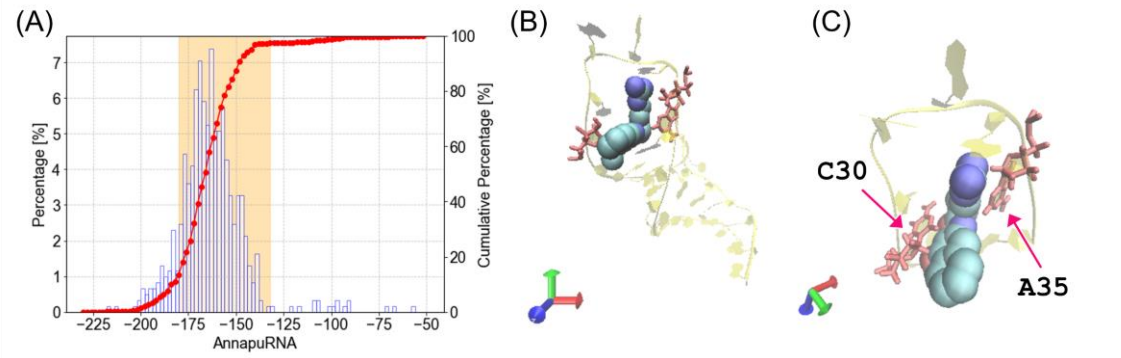

Figure 8. Predicted TAR-MV2003 binding pose. (A) RNA-ligand binding scores calculated by AnnapuRNA. (B) Global conformation of binding pose. (C) Closing view of TAR-MV2003 binding site. In panel A, the orange belt denotes the value range for NMR-resolved TAR-MV2003 complexes. Blue open bars and red closed 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. 2C**.

We focus on the predicted TAR-MV2003 binding pose with the minimum AnnapuRNA score as a representative. The TAR conformation shown in **Fig. 8B** appears to take coaxial stacking conformation, while an NMR-resolved TAR in complex with MV2003 takes interhelical bent conformation (*see* **Fig. 2B**). Besides, although the methoxy naphthalene group of MV2003 are intercalated between Cyt30 and Ade35, the pseudo-stacking formation is not the same as that of the NMR-resolved MV2003-bound complexes (compare **Fig. 8C** and **Fig. 2C**). We assume that the Molearn-generated conformation is not at the ground state of TAR-MV2003 complex but rather an intermediate appearing

during MV2003-binding (or releasing) process.

It is worthwhile to note that we obtained a possible TAR-MV2003 binding-poses by starting from *apo* RNA conformations without *a prioi* knowledge. Examining RNA's ligand-binding process[62] would be utilized to understand the physicochemical mechanisms of RNA-ligand molecular recognition. Such mechanical insights are indispensable for effective design of RNA-binding ligand with expected binding affinity during expected time periods. Recalling interest in RNA-ligand binding processes and the achievements of the present study, we suppose that Molearn has further potential to work as conformation generator which extend conformational spaces of RNA and proteins, which would promote drug-discovery with SBDD.

Concluding Remarks

Due to the vital roles in diverse biological processes[1-4], RNA has drawn much attention as target of drug discovery[5-8]. SBDD for RNA is a promising approach but still challenging so far, because available information of RNA-ligand complex structures is still restricted. Under the circumstances, theoretical prediction of RNA-ligand complex is feasible counterpart when we apply an SBDD approach to RNA targets. Meanwhile, prediction quality highly depends on how close RNA conformation are to their ligand-bound conformations, which could not necessarily appear among *apo* conformations and could not be necessarily observed during atomistic MD simulations with realistic computational resources of today.

Aiming to solve the above technical difficulty in SBDD approach for RNA targets, we addressed to generate ligand-boundable conformations without *a priori* knowledge of binding site structure by employing deep learning techniques. By using Molearn models[29], we succeeded to generate HIV-1 TAR conformations with the cryptic binding cavity for the small ligand, MV2003[36].

It is noted that a cryptic binding site is not observed in *apo* conformation ensemble and only formed under the influence of ligands, recently drawing attention as new druggable targets for conventional undruggable proteins[63, 64]. As for a cryptic MV2003-binding

cavity, there are no earlier studies to theoretically observe the formation[39-41], to our best knowledge. Then, our achievements denote promising potential of Molearn to generate unprecedent RNA conformations with specific functions. Meanwhile, Molearn's latent space is not practical yet to systematically discover ligand-boundable conformations generated by Molearn models. Future improvement of Molearn's framework and/or training protocols might solve the technical problem, then being indispensable methods to promote SBDD for RNA.

Acknowledgements

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

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-resolved and Molearn-

generated conformations, supporting Tables and Figures.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Ganser LR, Kelly ML, Herschlag D, Al-Hashimi HM. The roles of structural dynamics in the cellular functions of RNAs. *Nat. Rev. Mol. Cell Bio.* **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 Bio.* **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 RJ, Thornton CA. RNA-dominant diseases. *Hum. Mol. Genet.* **15**, R162-

R169 (2006).

6. Cooper TA, 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.* **13**, 433-446 (2014).

8. Tao YN, Zhang Q, Wang HY, Yang XY, Mu HR. Alternative splicing and related RNA binding proteins in human health and disease. *Signal Transduct. Tar.* **9**, (2024).

9. Anderson AC. 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 SK, Zielinski KA, Pollack L, Cerione RA. Evolving Experimental Techniques for Structure-Based Drug Design. *J. Phys. Chem. B* **126**, 6599-6607 (2022).

12. Burley SK, Wu-Wu A, Dutta S, Ganesan S, Zheng SXF. 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 JA, Rohs R. Structure-Based Drug Design with a Deep Hierarchical

Generative Model. *J. Chem. Inf. Model.* **64**, 6450-6463 (2024).

14. Kognole AA, Hazel A, MacKerell AD. SILCS-RNA: Toward a Structure-Based Drug Design Approach for Targeting RNAs with Small Molecules. *J. Chem. Theory. Comput.* **18**, 5672-5691 (2022).

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

16. Winkler WC, Cohen-Chalamish S, Breaker RR. An mRNA structure that controls gene expression by binding FMN. *P. Natl. Acad. Sci. USA* **99**, 15908-15913 (2002).

17. Vicens Q, Mondragón E, Batey RT. Molecular sensing by the aptamer domain of the FMN riboswitch: a general model for ligand binding by conformational selection. *Nucleic Acids Res.* **39**, 8586-8598 (2011).

18. Schroeder GM*, et al.* Analysis of a preQ-I riboswitch in effector-free and bound states reveals a metabolite-programmed nucleobase-stacking spine that controls gene regulation. *Nucleic Acids Res.* **48**, 8146-8164 (2020).

19. Chong LT, Saglam AS, Zuckerman DM. Path-sampling strategies for simulating rare events in biomolecular systems. *Curr. Opin. Struc. Biol.* **43**, 88-94 (2017).

20. Noé F, Clementi C. Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Curr. Opin. Struc. Biol.* **43**, 141-147 (2017).

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

22. 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).

23. Xue Y, Gracia B, Herschlag D, Russell R, Al-Hashimi HM. Visualizing the formation of an RNA folding intermediate through a fast highly modular secondary structure switch. *Nat. Commun.* **7**, (2016).

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

25. Goldbeck RA, Thomas YG, Chen EF, Esquerra RM, Kliger DS. Multiple pathways on a protein-folding energy landscape: Kinetic evidence. *P. Natl. Acad. Sci. USA* **96**, 2782-2787 (1999).

26. Reddish MJ, Callender R, Dyer RB. Resolution of Submillisecond Kinetics of Multiple Reaction Pathways for Lactate Dehydrogenase. *Biophys. J.* **112**, 1852-

1862 (2017).

27. Liu SC, Ying YL, Li WH, Wan YJ, Long YT. Snapshotting the transient conformations and tracing the multiple pathways of single peptide folding using a solid-state nanopore. *Chem Sci* **12**, 3282-3289 (2021).

28. Shinobu A, Kobayashi C, Matsunaga Y, Sugita Y. Coarse-Grained Modeling of Multiple Pathways in Conformational Transitions of Multi-Domain Proteins. *J. Chem. Inf. Model.* **61**, 2427-2443 (2021).

29. Ramaswamy VK, Musson SC, Willcocks CG, Degiacomi MT. Deep Learning Protein Conformational Space with Convolutions and Latent Interpolations. *Phys. Rev. X* **11**, (2021).

30. Kacher J, Sokolova OS, Tarek M. A Deep Learning Approach to Uncover Voltage-Gated Ion Channels' Intermediate States. *J. Phys. Chem. B* **128**, 8724-8736 (2024).

31. 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**, 3974-3981 (1996).

32. 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).

33. Du ZH, Lind KE, James TL. Structure of TAR RNA complexed with a Tat-TAR interaction nanomolar inhibitor that was identified by computational screening. *Chem. Biol.* **9**, 707-712 (2002).

34. 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).

35. Murchie AIH*, 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).

36. Davidson A, Begley DW, 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).

37. Davidson A, Patora-Komisarska K, Robinson JA, 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).

38. Shortridge MD*, 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).

39. Fulle S, Christ NA, 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).

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

41. Panei FP, Gkeka P, Bonomi M. Identifying small-molecules binding sites in RNA conformational ensembles with SHAMAN. *Nat. Commun.* **15**, (2024).

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

43. Steinbrecher T, Latzer J, Case DA. Revised AMBER Parameters for Bioorganic Phosphates. *J. Chem. Theory. Comput.* **8**, 4405-4412 (2012).

44. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **79**, 926-935 (1983).

45. Kusalik PG, Svishchev IM. The Spatial Structure in Liquid Water. *Science* **265**, 1219-1221 (1994).

46. Joung IS, Cheatham TE. Determination of alkali and halide monovalent ion

parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **112**, 9020-9041 (2008).

47. Joung IS, Cheatham TE, 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).

48. Berendsen HJC, Postma JPM, Vangunsteren WF, Dinola A, Haak JR. Molecular Dynamics With Coupling to an External Bath. *J. Chem. Phys.* **81**, 3684-3690 (1984).

49. Case, D. A. et al. AMBER 2022. Assisted Model Building with Energy Refinement. University of California, San Francisco. https://ambermd.org (Accessed 6 Jul 2022).

50. Le Grand S, Gotz AW, Walker RC. SPFP: Speed without compromise-A mixed precision model for GPU accelerated molecular dynamics simulations. *Comput Phys Commun* **184**, 374-380 (2013).

51. Stasiewicz J, Mukherjee S, Nithin C, Bujnicki JM. QRNAS: software tool for refinement of nucleic acid structures. *Bmc. Struct. Biol.* **19**, (2019).

52. Xiong P, Wu RB, Zhan J, Zhou YQ. Pairing a high-resolution statistical potential with a nucleobase-centric sampling algorithm for improving RNA model

refinement. *Nat. Commun.* **12**, (2021).

53. Durrant JD, de Oliveira CAF, McCammon JA. POVME: An algorithm for measuring binding-pocket volumes. *J. Mol. Graph. Model.* **29**, 773-776 (2011).

54. Durrant JD, Votapka L, Sorensen J, Amaro RE. POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J. Chem. Theory. Comput.* **10**, 5047-5056 (2014).

55. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. *J. Mol. Graph. Model.* **14**, 33-38 (1996).

56. Sun LZ, Jiang YW, Zhou YZ, Chen SJ. RLDOCK: A New Method for Predicting RNA-Ligand Interactions. *J. Chem. Theory. Comput.* **16**, 7173-7183 (2020).

57. Stefaniak F, Bujnicki JM. AnnapuRNA: A scoring function for predicting RNA-small molecule binding poses. *Plos Comput. Biol.* **17**, (2021).

58. Jangid P, Chaudhury S. Transition Path Dynamics of Non-Markovian Systems across a Rough Potential Barrier. *J. Phys. Chem. A* **128**, 10041-10052 (2024).

59. Janovjak H, Knaus H, Muller DJ. Transmembrane helices have rough energy surfaces. *J. Am. Chem. Soc.* **129**, 246-247 (2007).

60. Hyeon C, Thirumalai D. Measuring the energy landscape roughness and the transition state location of biomolecules using single molecule mechanical

unfolding experiments. *J. Phys.-Condens. Mat.* **19**, (2007).

61. Riccabona JR*, et al.* Assessing AF2's ability to predict structural ensembles of proteins. *Structure*, (2024).

62. Borkar AN, Bardaro MF, Camilloni C, Aprile FA, Varani G, Vendruscolo M. Structure of a low-population binding intermediate in protein-RNA recognition. *P. Natl. Acad. Sci. USA* **113**, 7171-7176 (2016).

63. Vajda S, Beglov D, Wakefield AE, Egbert M, Whitty A. Cryptic binding sites on proteins: definition, detection, and druggability. *Curr. Opin. Chem. Biol.* **44**, 1-8 (2018).

64. Lazou M, Kozakov D, Joseph-McCarthy D, Vajda S. Which cryptic sites are feasible drug targets? *Drug Discov. Today* **29**, 104197 (2024).