Published online 26 November 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 4 **1**
https://doi.org/10.1093/nargab/lqab111

Structure-based deep learning for binding site detection in nucleic acid macromolecules

Igor Kozlovskii and Petr Popov *

iMolecule, Skolkovo Institute of Science and Technology, Moscow, 121205, Russia

Received August 19, 2021; Revised October 14, 2021; Editorial Decision November 08, 2021; Accepted November 09, 2021

ABSTRACT

Structure-based drug design (SBDD) targeting nucleic acid macromolecules, particularly RNA, is a gaining momentum research direction that already resulted in several FDA-approved compounds. Similar to proteins, one of the critical components in SBDD for RNA is the correct identification of the binding sites for putative drug candidates. RNAs share a common structural organization that, together with the dynamic nature of these molecules, makes it challenging to recognize binding sites for small molecules. Moreover, there is a need for structure-based approaches, as sequence information only does not consider conformation plasticity of nucleic acid macromolecules. Deep learning holds a great promise to resolve binding site detection problem, but requires a large amount of structural data, which is very limited for nucleic acids, compared to proteins. In this study we composed a set of ∼2000 nucleic acid-small molecule structures comprising ∼2500 binding sites, which is ∼40-times larger than previously used one, and demonstrated the first structure-based deep learning approach, BiteNet$_N$, to detect binding sites in nucleic acid structures. BiteNet$_N$ operates with arbitrary nucleic acid complexes, shows the state-of-the-art performance, and can be helpful in the analysis of different conformations and mutant variants, as we demonstrated for HIV-1 TAR RNA and ATP-aptamer case studies.

INTRODUCTION

RNA molecules are vital in many cellular processes, such as gene regulation and cell information transfer, thus, representing a promising class of pharmacological targets (1). RNA-targeting drug discovery campaigns explore various perspectives, including design of stabilizers of DNA G-quadruplex (2), riboswitch-targeting antibiotics (3), antisense RNA (4) and RNA-targeting antivirals, to name a few. RNA targets that expand druggable genome, including those linked to 'undruggable' protein targets or non-coding microRNAs, are of particular interest (5). However, RNA drug development is dotted with numerous obstacles (6), among others, related to the low chemical diversity and the dynamic nature of RNA structures. Similar to proteins, RNA molecules are highly structured to form binding sites, through which small molecules can modulate them (7). Therefore, there is a need for efficient, structure-specific RNA-small molecule ligand binding site detectors to advance RNA-targeting drug discovery.

Despite the abundance of protein-specific approaches, there is a very limited number of methods developed to predict RNA-small molecule interaction sites, and they can be roughly divided into knowledge-based, empirical, and machine learning approaches. Knowledge-based approaches, such as InfoRNA, mine RNA motifs in a database of known RNA–small molecule binding sites (8). Empirical approaches, such as Rsite (9), Rsite2 (10) or RBind (11), rely on simple geometric characteristics of RNA structures and look for the extremes of these characteristics as the indicators of a binding site. Most recently, a machine learning approach, RNAsite, was developed; it comprises a Random Forest model that operates with calculated RNA's structure-based and sequence-based features (12). Using deep learning is expected to improve the RNA binding site detectors; however, it is hampered due to the relatively small number of available RNA structures. Indeed, while the most recent deep learning approaches for a protein-small molecule or protein-peptide binding site detection rely on datasets of thousands of examples (13,14), the RNAsite model was trained on just 60 RNA-small molecule complexes (12).

In this study, we demonstrated the first structure-based deep learning approach for nucleic acid-small molecule ligand binding site prediction. To overcome the small dataset problem, we considered both RNA and DNA complexes, interaction interfaces formed with the crystallographic symmetry mates, NMR models, and data augmentation. We composed a dataset of ∼2000 nucleic acid-small molecule structures, comprising ∼2500 binding site interfaces retrieved from Protein Data Bank (PDB) (15). Next, we developed the voxel-based view of nucleic chain structures, such that each voxel represents a 1Å^3 cube in the physical space and stores eight channels corresponding to the atomic

*To whom correspondence should be addressed. Email: p.popov@skoltech.ru

© The Author(s) 2021. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

densities of a particular type. The voxelized representations are then fed to the 3D convolutional neural network that scores segments in nucleic acid structures concerning the binding sites. The obtained structure-based deep learning model, dubbed BiteNet$_N$, predicts the coordinates of binding site interface centers, the probability scores for each center, and scores for each nucleotide in a binding site. We observed the superior performance of BiteNet$_N$ compared to the other methods on the constructed test sets. To demonstrate the applicability of BiteNet$_N$ for relevant nucleic acids, we considered two pharmacologically-oriented case studies, including i) different structures of the HIV-1's transactivating response region bound to small molecules and ii) molecular dynamics trajectories of ATP-aptamers; we showed that BiteNet$_N$ is capable of correct identification of the conformation-specific binding sites.

MATERIALS AND METHODS

Training and test sets

Despite constant growth in deposited molecular structures to PDB, the number of nucleic-ligand molecular complexes is very limited. For example, RNAsite was trained and tested on datasets of just 60 (TR60) and 18 (TE18) single RNA chain structures in complex with small molecules or ions (12). Apparently, TR60 and TE18 datasets are not sufficient to derive robust deep learning models; therefore, we composed a larger dataset of 1933 nucleic acid-small molecule complexes, as it follows. Firstly, we retrieved from PDB (15) structures containing RNA or DNA but not protein chains with resolution $\leq$3 Å or solved by NMR. We filtered out structures exceeding 200 Å along the first principal axes (e.g. ribosomal complexes with the minimal low ratio of binding nucleotides in the given structure) and considered only complexes with small molecules of at least 10 heavy atoms surrounded by at least 15 nucleotides' heavy atoms within 4Å, resulting in total of 780 structures. In many cases, we observed intermolecular interfaces formed by the asymmetric unit and the symmetry mates (see Supplementary Figure S1). More precisely, 20% (235 out of 1010) small molecules observed in 634 X-ray structures form strong interactions with $\geq$15 heavy atoms of a symmetry mate, and Supplementary Figure S2 shows the distribution of the binding site interface sizes formed by the asymmetric unit and symmetry mates. In contrast, the corresponding ratio in the protein-small molecule binding site dataset (13) is only $\sim$ 2%, emphasizing the differences between interfaces formed with the symmetry mates. To take such interfaces into account, we built symmetry mates within 64 Å of the asymmetric unit and calculated the size of each nucleic-ligand and nucleic-nucleic interface. The latter is done to avoid potentially false negative examples from the dataset corresponding to the interaction interfaces between complementary or consecutive nucleic acid chains, which can be also observed in the asymmetric units. Indeed, 130 and 184 nucleic acid chains (out of 1162) in 634 asymmetric units have at least one chain from a symmetry mate with number of interacting nucleotides $\geq$40 or with ratio of interacting nucleotides $\geq$0.5, respectively. Supplementary Figure S3 shows the distributions of the interface sizes between nucleic acid chains from asymmetric unit and symmetry mates, and Supplementary Figure S4 demonstrates examples of such interacting nucleic acid chains. Then, for each structure, we form 'parts' in the following way: starting from the asymmetric unit, a 'part' corresponds to a single nucleic acid chain. Next, we extend each 'part' by adding small molecules and nucleic acids chains interacting with it. Finally, we merge 'parts' that share small molecules or nucleic acids chains. We considered interacting small molecules with the corresponding single chain binding interface of at least 5 heavy atoms, and interacting nucleic acid chains with the binding interface of $\geq$40 heavy atoms in any chain (consecutive chains), or the ratio of interacting nucleotides with respect to the chain's size of $\geq$0.5 (complementary chains). We continued the extension procedure until convergence, that is (i) 'part"s composition does not change in the next iteration, (ii) a 'part' contains at most five nucleic acid chains and (iii) a 'part' does not exceed 200 Å along the first principal axis. Consequently, we filtered out complexes, for which the extension procedure did not converge. In several cases, the obtained X-ray complexes consist of non-interacting parts, that we split into 471 different structures; as for the NMR structures, we considered each model as a separate complex, resulting in 1462 structures. Supplementary Figure S5 shows the distributions of the binding interfaces per single chain as well as per complex, and Supplementary Figure S6 shows the size of the resulting training set with respect to the threshold for the nucleic chain-chain interfaces. Finally, we refined the obtained structures by adding missing atoms and restoring hydrogens using ICM-Pro (www.molsoft.com). In total, we constructed the BN$_N$1933 dataset of 471 + 1462 = 1933 nucleic chain-small molecule structures. It is important to emphasize that some structures were constructed from the same experimental complexes (for example, multiple non-interacting complexes from a single crystallographic structure or multiple models in NMR structures). Therefore, a careful train-validation-test split is required, as a random split would likely result in over-estimated performance.

To split BN$_N$1933 into training, validation, and test subsets, we computed the sequence identity and structural similarity for each pair of 1933 complexes. The sequence identity was calculated as the number of identical nucleotides divided by the average length of two sequences aligned with Biopython (16), that scores identical nucleotides with +1, different nucleotides with 0, gap opening with $-$1, and gap extending with 0. The structural similarity was calculated as the averaged similarity score obtained with RNAalign (17). We considered the maximum sequence identity and structure similarity when several chains present in a structure. Then, we clustered the complexes using 0.7 and 0.8 threshold values for the sequence identity and structure similarity, respectively, resulting in 116 clusters, such that *all* pairs of complexes with sequence identity or structure similarity exceeding the thresholds are in the same cluster. For rigorous comparison, we constructed 10 cluster-based train-validation-test splits as it follows. For each split, we assigned 40 clusters to the test partition, comprising all 7 clusters that share at least one structure similar to RB19 (11) or TE18 (12); and 33 clusters randomly chosen from 78 clusters that do not share similar structures with TR60 and are not be-

long to the top-5 largest clusters. Next, we identified a subset of complexes containing a single RNA chain for each cluster in the test partition. To construct the $BN_N^{TE}40_i$ test set, from each cluster, we chose random complex from the single RNA chain subset, if it is not empty, or random complex, otherwise. As existing tools operate with single RNA chains only, in addition, we composed $BN_N^{TE}40_i^{SUB}$ solely from the single RNA chain-ligand complexes from $BN_N^{TE}40_i$. The remaining $116 - 40 = 76$ clusters were assigned into the train-validation partition. Therefore, train-validation sets do not contain any complexes similar to TE18 or RB19, as all 7 clusters sharing similar structures with these test sets are belong to $BN_N^{TE}40_i$ test sets, $i = 1 \ldots 10$. We used five-fold cross-validation to check model robustness as well as to tune parameters for the nucleotide-based scoring (see Section 'Model'). Top-4 largest clusters were assigned to the train partition, and the remaining 72 clusters were split into five folds using grouped splitting. Therefore, for each of 10 training-validation-test splits one has the training-validation set $BN_{N\ i}^{TR}$, the test set $BN_N^{TE}40_i$, and the test set $BN_N^{TE}40_i^{SUB}$ corresponding to the single RNA chain-ligand complexes. Supplementary Figure S7 schematically illustrates the construction of train-validation-test splits. Supplementary Table S1 provides detailed information about the number of clusters and structures in each split, Supplementary Table S2 lists the clusters in the test set for each split, and Supplementary Table S3 lists numbers of all and single chain RNA complexes in each cluster. For comparison with the other methods, we also considered the RB19 (11) and TE18 (12) test sets; from TE18, we discarded six complexes with only non-relevant ligands (ions or covalently bound modified nucleotide residues), resulting in the TE12 test set.

Model

We represented nucleic-ligand structures as the 4D tensors, where the first three dimensions correspond to x, y, z with discretization of 1.0 Å, and the fourth dimension corresponds to the eight channels that store atomic densities for different atom types (see Equation 1).

$$\rho(r) = \begin{cases} e^{-r^2/2}, & \text{if } r \leq r_{cutoff} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Here we used the hybridization types of C, O, N, and P for the different channels as well as non-standard atoms observed in the modified nucleotides (e.g. B, Br, Cl, F, I, Pt, Te, V, S and Se), as a separate channel. The atom types for the non-standard and modified nucleotides were taken from the SYBYL atom typing in the MOL2 files (see Table 1). As the resulting voxel grids vary in size with respect to the input structure, we split the obtained 4D tensors into the 4D sub-tensors of a fixed size ($64 \times 64 \times 64 \times 8$), which are fed into the 3D convolutional neural network. We used the BiteNet's 3D convolutional neural network architecture (13), that consists of ten 3D convolutional layers with kernel size of 3 and filter sizes of 32, 32, 32, 32, 32, 64, 64, 64, 128 and 4, respectively; all convolutions except the last one are followed by the Batch Normalization layer and ReLU activation function, and the sigmoid function is applied to on the last layer. The model outputs probability scores and coordinates of the predicted binding sites for each of 512 cells

Table 1. Nucleic acid atom types and corresponding tensor channels

Channel	Description	SYBYL atom type
1	carbon sp, sp2, aromatic	C.1, C.2, C.ar, C.cat
2	carbon sp3	C.3
3	nitrogen sp, sp2, amide, aromatic, trigonal	N.1, N.2, N.ar, N.am, N.pl3
4	nitrogen sp3, quaternary	N.3, N.4
5	oxygen sp2	O.2
6	oxygen sp3	O.3
7	phosphorus sp3	P.3
8	other	S.2, S.3, S.o2, Se, B, Br, Cl, F, I, Pt, Te, V

of size $8 \times 8 \times 8 \times 8$ constituting a single sub-tensor. The non-max-suppression with the distance threshold of 8 Å is applied to get the final predictions of the binding sites' centers: on the i-th step one chooses the top-i scored prediction and filters out all the predictions within the distance threshold of 8Å from it. Therefore, after the non-max suppression, all the predictions are at least 8 Å apart from each other. During training, we used an implicit data augmentation by a random rotation of structures in each epoch. The loss function to minimize consists of three terms: (i) the cross-entropy term for the probability score, (ii) the mean squared error for cells with the true binding site's center and (iii) the L_2 regularization term:

$$\begin{aligned} \text{Loss} = & \sum_{i=1}^{N_{\text{cells}}} (-s_i \cdot \log(\hat{s}_i) - (1 - s_i) \cdot \log(1 - \hat{s}_i)) \\ & + \lambda_{\text{coord}} \cdot \sum_{i=1}^{N_{\text{cells}}} s_i \cdot ((x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + (z_i - \hat{z}_i)^2) \\ & + \gamma L_2, \end{aligned} \quad (2)$$

where N_{cells} is the number of cells in a single sub-tensor, s_i and $\hat{s}_i$ are the true (0 or 1 for true binding site center absent or present in the cell, respectively) and predicted probability scores for the cell, x_i, y_i, \$z\$$_i$ and $\hat{x}_i$, $\hat{y}_i$, $\hat{z}_i$ are the true and predicted coordinates for i-th cell, respectively, L_2 is the Euclidean norm of model's weights, $\lambda_{\text{coord}} = 5.0$ is the coefficient for the coordinate loss term, and $\gamma = 1e{-}5$ is the coefficient for regularization term. We trained models for 40 000 steps with the Adam optimizer (18) using batch size of 16 and the learning rate decay from 1e−3 to 1e−5. In the inference mode, we averaged results obtained for 50 replicas of the input structure obtained by rotation about ten different axes corresponding to the centroids of the icosahedron facets (19) by $\pi/3$, $2\pi/3$, π, $4\pi/3$ and $5\pi/3$ angles.

To convert the BiteNet$_N$ predictions to the nucleotide-based probability scores, we firstly scored each atom a within the distance threshold d from a prediction p:

$$s_a = \max_{p, ||\mathbf{r}_p - \mathbf{r}_a|| \leq d} s_p \times e^{-\frac{||\mathbf{r}_p - \mathbf{r}_a||^2}{2r_{norm}}} \quad (3)$$

where s_p is the probability score of the prediction p, $\mathbf{r}_p$ and $\mathbf{r}_a$ are the coordinates of p and a, r_{norm} is the normalization coefficient, respectively. In case there are several predictions within d from a, the highest s_a was taken. Then, the nucleotide's score was calculated as the maximum of its heavy atom scores. Finally, we defined nucleotides with scores higher than the score threshold s_r as belonging to the

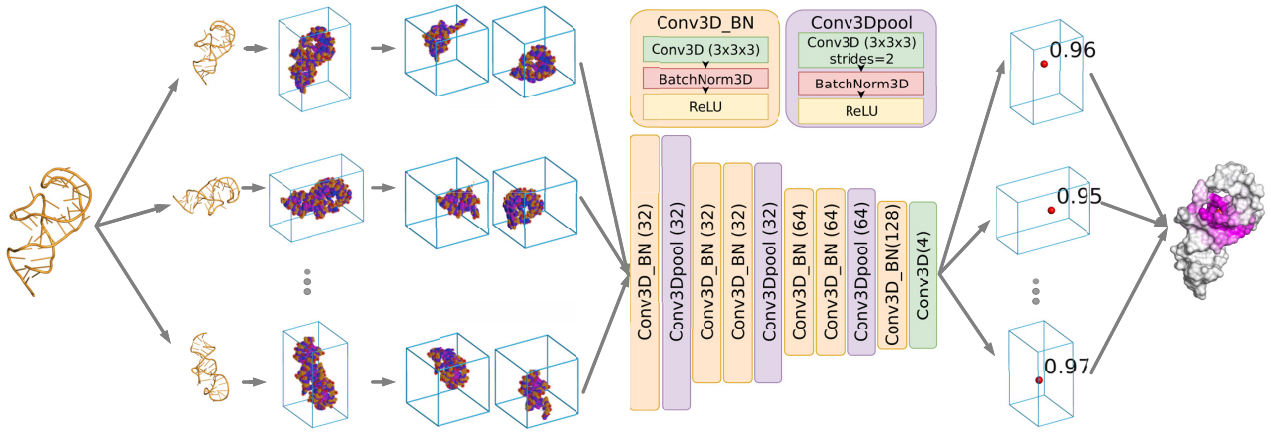

Figure 1. Illustration of BiteNet$_N$'s workflow. Multiple orientations of the input nucleic acid structure are voxelized into fixed-size 4D tensors. The tensors are split into the set of cubic voxel grids of fixed size ($64 \times 64 \times 64$), which are fed to 3D CNN that outputs probability scores and coordinates of putative binding site centers. Finally, each nucleotide is scored according with the mapping function. Nucleic acid structure is represented with yellow cartoon, voxels with non-zero values are colored with respect to the channel type, predictions are shown with red spheres, and the nucleotides in the final output are colored with magenta with respect to the predicted probability score.

binding site. A single complex from each cluster was sampled during a training epoch to avoid over-fitting towards the most populated cluster. We determined the best parameters d, r_{norm} and s_r based on the highest averaged nucleotide-based AP and MCC (see section 'Metrics') on the five-fold cross-validation (see Section 'Training and test sets').

Overall, we obtained 50 models for the cross-validation stage, 10 models for the test stage, single models for the HIV-1 TAR RNA and ATP-aptamer DNA case studies, and the final model trained on the full BN$_N$1933 dataset incorporated into the web-server https://sites.skoltech.ru/imolecule/tools/bitenet/. We used the Zhores supercomputer (20) to train the models.

Metrics

We calculated AP (average precision; area under the precision-recall curve) as the area under the precision-recall curve, ROC AUC (area under the receiver operating curve) as the area under the true positive rate versus the false positive rate curve and the MCC (Matthews correlation coefficient) metric defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \quad (4)$$

for nucleotides classified as either belonging to the nucleic-ligand binding site or not, where TP and FP are the numbers of binding and non-binding nucleotides classified as binding, respectively; TN and FN are the numbers of non-binding and binding nucleotide classified as non-binding, respectively; precision $= \frac{TP}{TP+FP}$, and recall $= \frac{TP}{TP+FN}$. The nucleotide is defined as binding if it contains at least one heavy atom within 4 Å from any heavy atom of the ligand. To take into account different binding site sizes across the complexes, we also weighted each nucleotide with the inverse number of nucleotides in the structure, thus, calculating the weighted performance metrics. During cross-validation, we optimized the weighted AP additionally balanced to the cluster size to consider both the number of similar structures in the training set and the size of the complexes: the weight of a nucleotide is the inverse of product of the number of nucleotides in a structure and the size of the corresponding cluster.

RESULTS AND DISCUSSION

BiteNet$_N$

Dataset. To train the BiteNet$_N$ deep learning models, we constructed a large dataset of 1933 nucleic acid-ligand complexes, including 1065 DNA and 886 RNA structures (18 structures contain both DNA and RNA) of different types. Namely, there are 865, 575, 51, 217, 358 and 20 complexes corresponding to the A-form double, B-form double, Z-form double, triple, quadruple helixes and undefined structural types, respectively, as retrieved from the NDB database (21). The dataset contains DNA and RNA complexes with the average number of nucleotides and chains per complex equal 34 and 1.5, respectively, 2469 binding sites occupied with small molecules with the average number of heavy atoms equals 35.5, the average interaction interface size of 45.8 heavy atoms, and the average binding site solvent accessible surface area (SASA) of 245.7Å^2 (calculated with FreeSasa (22)). It is worth noting that there are 938 complexes formed by several nucleic acid chains, and the average interaction interface size per single nucleic acid chain is 29.8, emphasizing the difference between mono- and oligo- nucleic-ligand complexes (see Supplementary Figure S5). The BN$_N$1933 dataset is imbalanced in terms of the number of binding ($\sim$ 30%, 19182 nucleotides) and non-binding ($\sim$ 70%, 46 459 nucleotides) nucleotides. Interestingly, distributions of the number of heavy atoms in nucleic-small molecule interfaces resemble those for protein-small molecule complexes (31.5 and 46.1 for small molecule and interaction interface sizes, respectively) (13).

Model. We trained BiteNet$_N$ on the curated nucleic acid structures using 3D CNN architecture, proven to be top-performing for protein-small molecule molecule

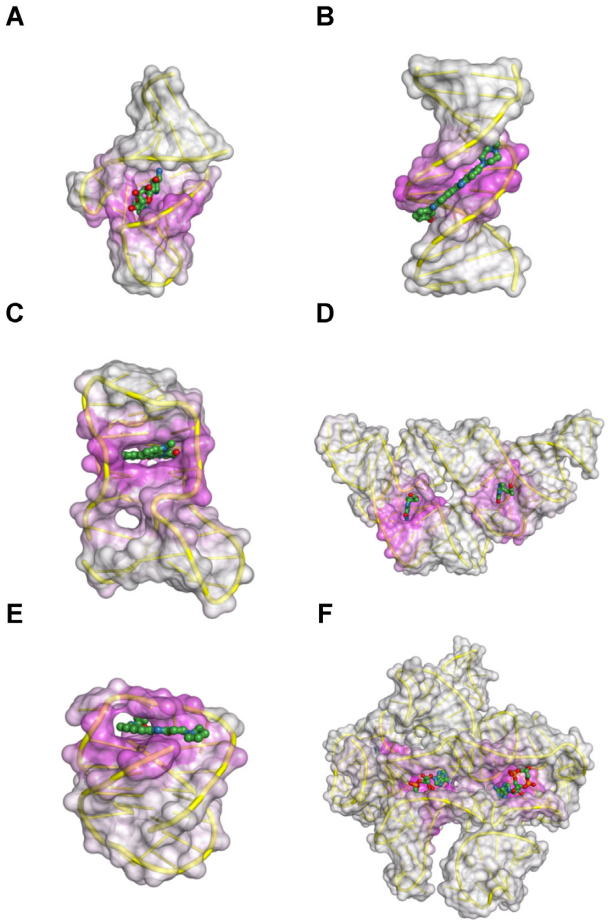

Figure 2. Demonstration of the BiteNet$_N$'s applicability to the different types of DNA or RNA structures. (**A**) A-form double helix RNA complex (PDB ID: 2KXM); (**B**) B-form double helix DNA complex (PDB ID: 302D); (**C**) Z-form double helix DNA complex (PDB ID: 6SX3); (**D**) triple helix RNA complex (PDB ID: 4LVX); (**E**) quadruple helix DNA complex (PDB ID: 1L1H); and (**F**) unspecified RNA (PDB ID: 6N5L). Nucleic acid chains are represented as yellow ribbon and white transparent surface, small molecules are shown with sticks and spheres, and nucleic acid surface is colored with magenta according to the predicted binding scores for nucleotides.

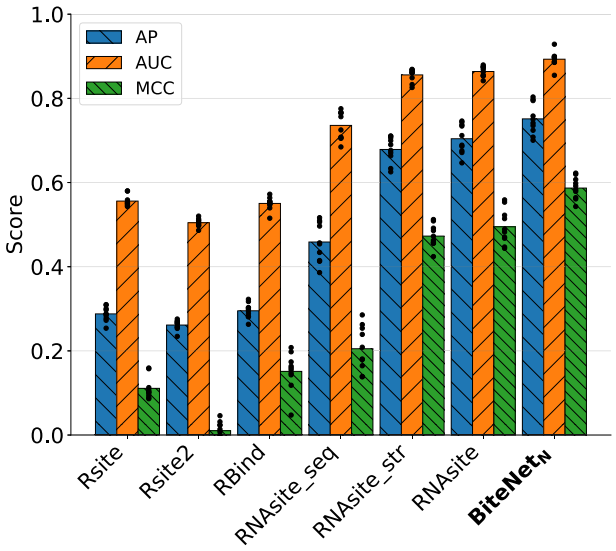

Figure 3. Weighted AP, ROC AUC and MCC performance metrics on the $BN_N^{TE}40_i^{SUB}$ test sets, $i = 1...10$. Bars correspond to the performance metrics averaged over the ten test sets, and black dots correspond to the individual performance metrics for each test set.

and protein-peptide binding site detection (13,14) (see the 'Methods' section), and Figure 1 illustrates the BiteNet$_N$ workflow. In the first step, the input nucleic acid structure is voxelized and splitted into the cubic voxel grids, resulting in the set of fixed-size 3D images ($64 \times 64 \times 64$) representing 64 Å^3 spatial cube, and each voxel (1 Å^3) contains eight channels corresponding to the atomic densities of a specific type. In the second step, the 3D images are fed into 3D CNN to output tensors of size $8 \times 8 \times 8 \times 4$, where the first three dimensions correspond to the cell coordinates relatively to the 3D images (regions of $8 \times 8 \times 8$ voxels), and the four scalars of the last dimension correspond to the probability score of a binding site center being in the cell and its 3D coordinates. Next, the obtained tensors are processed to output the most relevant ligand binding site predictions. Finally, each nucleotide is scored with respect to the obtained predictions according to Equation (3). We used 5-fold cross-validation to determine the best parameters for function mapping model predictions into nucleotide scores and found $d = 12$Å, $r_{norm} = 4$ and $s_r = 0.1$ to be optimal for nucleotide-based performance metrics (averaged AP and MCC). Overall, the input to BiteNet$_N$ is the spatial structure of nucleic acid, and the output is the scored centers of the predicted binding sites along with the scored nucleotides associated with each center. Figure 2 shows application of BiteNet$_N$ to the nucleic acid structures of different types.

Comparison with other methods. To compare BiteNet$_N$'s performance with other methods, we obtained the binding site predictions of four different approaches: Rsite (9), Rsite2 (10), RBind (11), RNAsite (15) for the ten test sets $BN_N^{TE}40_i^{SUB}$, $i = 1...10$. We calculated the weighted AP, ROC AUC, and MCC performance metrics for the existing methods and ten BiteNet$_N$'s models trained on the $BN_N^{TR}{}_i$ sets, $i = 1...10$. Figure 3 demonstrate the obtained results, and Supplementary Tables S4–S5 and Supplementary Figure S8 show more detailed BiteNet$_N$'s performance on the cross-validation and test sets for each of the splits, as well as on the RB19 and TE12 benchmarks. Overall, BiteNet$_N$ models outperform the other methods on the constructed test sets, achieving the average weighted AP, ROC AUC and MCC scores of 0.75, 0.89 and 0.59, respectively. It is interesting to note that empirical methods (Rsite, Rsite2 and RBind) demonstrated poor performance on all the test sets, emphasizing the complexity of the binding site detection problem and the anticipated improvements that machine learning methods bring. When applied BiteNet$_N$ to the modelled structures of RB19 retrieved from (12), we observed negative correlation between the performance metrics and the RMSD with respect to the experimental structures, emphasizing importance of the quality of the input structure for the binding site detection (see Supplementary Figure S9).

Case studies

A binding site is the structural and dynamic property of a macromolecule; therefore, a method to predict binding sites should distinguish conformations with open and collapsed binding sites and be applicable for the analysis of conformational ensembles. To demonstrate the use of BiteNet$_N$ for relevant nucleic-ligand binding site detection problems, we considered i) the transactivating response region of HIV-1 and ii) the ATP-aptamer.

HIV-1 TAR RNA. The transactivating response region (TAR) is a part of HIV-1's RNA, that promotes transcription of viral genome via binding with the transactivator Tat protein and the host cofactor cyclin T1. TAR is considered as a relevant pharmacological target, and there are numerous structural studies of TAR in complex with small molecules, peptides, or cations (23,24). To demonstrate the conformation sensitivity of BiteNet$_N$, we retrieved seven different structures of the TAR RNA bound to small molecules (PDB IDs: 1ARJ (25), 1LVJ (26), 1QD3 (27), 1UTS (28), 1UUD (29), 1UUI (29), 2L8H (30)). Although the nucleotide sequence is the same, these structures are different: the pair-wise root-mean-squared-deviation (RMSD) varies from 1.9Å to 10.7Å with the median value of 6.5Å, and the binding site sizes vary from 4 to 12 nucleotides. Moreover, the binding sites span almost the entire sequence of the TAR RNA (23 of 29 nucleotides correspond to at least one binding site) with only two binding nucleotides (A22 and U23) shared between all the structures (see Supplementary Figure S10). To exclude possible bias, we retrained BiteNet$_N$ model on a training subset without similar to TAR RNA complexes: from BN$_N$1933 we filtered out all clusters containing at least one structure similar to any of seven TAR-ligand structures, resulting in 115 clusters with 1831 complexes (a single cluster of 102 complexes was filtered out), followed by training of the BiteNet$_N$ model. We used the same procedures as described in Sections 'Training and test sets' for filtering and 'Model' for training. We then applied the BiteNet$_N$ model for each structure and measured the weighted AP, ROC AUC, and MCC performance metrics; for the NMR structures we considered all the models and averaged the results. BiteNet$_N$ achieves weighted AP, ROC AUC and MCC metrics of 0.727, 0.872 and 0.510, respectively, on the seven TAR-ligand structures (see Figure 4), demonstrating the ability to correctly identify binding nucleotides in the sequence with respect to the different conformations (see Figure 5 and Supplementary Figure S10). We want to emphasize that sequence-based methods alone inevitably fail on such case studies as their predictions do not depend on spatial information. It is also important to note, that there are six TAR structures bound to peptides (PDB ID: 2KDQ (31), 2KX5 (32), 5J0M (33), 5J1O (33), 5J2W (33), 6D2U (34)). We observed much lower performance metrics when applied BiteNet$_N$ on these complexes (AP, ROC AUC and MCC scores of 0.727, 0.872, 0.510 compared to 0.619, 0.679, 0.239 for small molecule-bound and peptide-bound complexes, respectively), suggesting the need for specific models for the peptide-based ligands (14).

ATP-aptamer. Aptamers are oligonucleotides designed for selective binding to specific small molecule targets

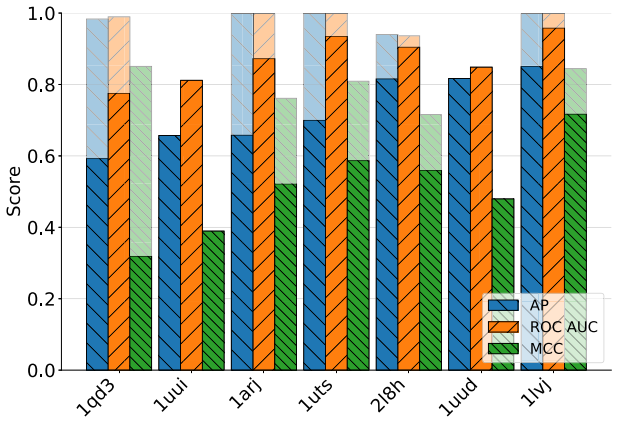

Figure 4. The AP (blue), ROC AUC (orange), and MCC (green) performance metrics for seven structures of TAR RNA bound to small molecules. Bars in pale colors correspond to the top-scored NMR conformation in terms of AP.

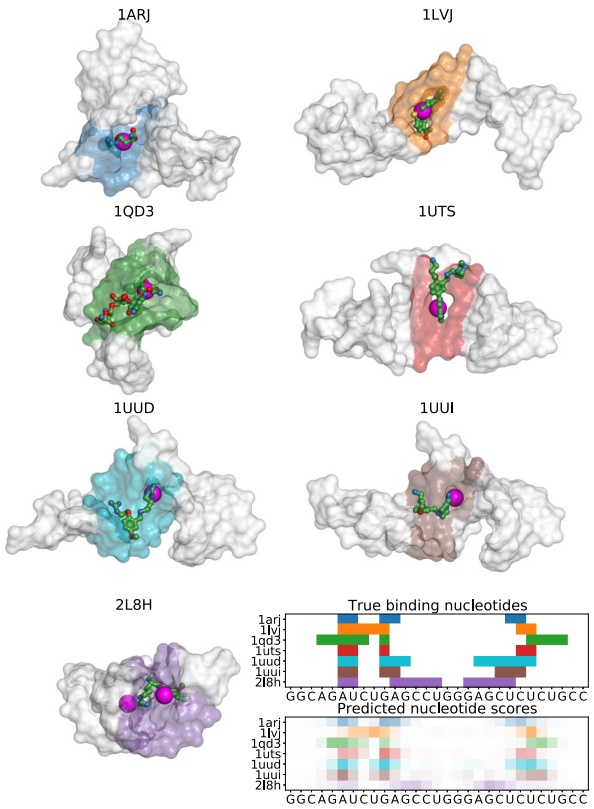

Figure 5. Predictions for seven structures of HIV-1 TAR RNA complexes with small molecule ligands. RNA and small molecules are shown with surface and sticks, respectively. The predicted binding site centers are shown with magenta spheres and the true binding site residues are highlighted with colors. For each PDB ID, NMR conformation with the highest AP score is shown.

(35,36), and they can be used in biosensors and drug discovery (37,38). To demonstrate the applicability of BiteNet$_N$ for the large-scale analysis, we considered molecular dynamics trajectories of ATP-aptamers. ATP-aptamer is a nucleic acid of 27 nucleotides (39,40); the NMR structure of ATP-aptamer has two binding sites with AMP molecules bound to it (PDB ID: 1AW4 (41)), corresponding to G6–G22–A23

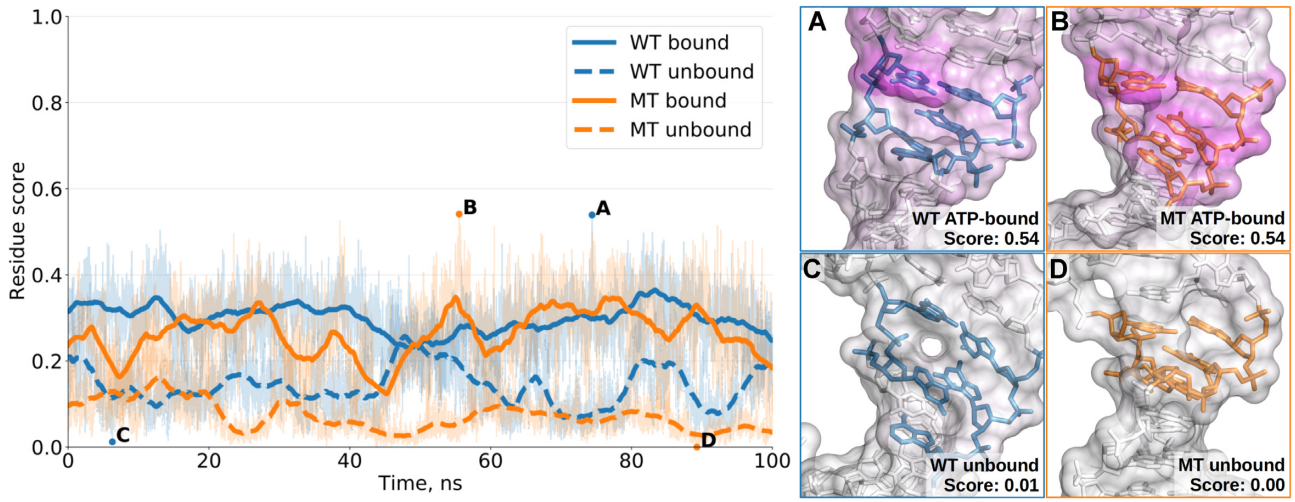

Figure 6. Binding site scores calculated over the ATP-bound and ATP-unbound MD trajectories of the wild-type ATP aptamer (WT, blue) and its G6A mutant (MT, orange). Bold solid (dashed) lines correspond to the moving average over 2 ns of the binding site scores for the ATP-bound (-unbound) trajectories. Molecular dynamics snapshots corresponding to the highest (lowest) scores for the ATP-bound (-unbound) trajectories and for both WT and MT simulations are shown in boxes **A** and **B** (**C** and **D**), respectively. RNA and its binding site are represented with surface and sticks, respectively; the surface is colored with magenta with respect to the nucleotide scores. ATP molecule is not shown for clarity.

and G9–A10–G19 nucleotides, respectively. It is also known that mutations of G6 and G22 reduce or abolish the binding efficiency to ATP, as measured with elution chromatography (39), and a recent molecular dynamics study attempted to characterize such effect on the structural level (42). We retrieved molecular dynamics trajectories for apo, and ATP-bound complexes of the wild-type ATP-aptamer (WT) and its G6A mutant (MT) from (42). As in the TAR RNA case study, we trained the BiteNet$_N$'s model on the dataset with no complexes similar to ATP-aptamer, comprising 115 clusters of 1926 complexes. Then we obtained the BiteNet$_N$'s model predictions for each frame of the trajectories. For each frame, we calculated the binding site score as the average score of the true ATP-binding site, defined as a set of five nucleotides: G5, G(A)6, G21, G22 and A23 (42). Figure 6 demonstrates the obtained binding site scores over the molecular dynamics trajectories. As expected, we observed that the trajectory of the wild-type ATP-bound aptamer demonstrates the highest binding site prediction score. The average scores of the binding site in the ATP-bound molecular dynamics trajectories are higher than the unbound ones: 0.30 versus 0.14 and 0.26 versus 0.07 for the wild-type and mutant-type, respectively. Interestingly, the binding site score dropped almost to zero in the ATP-unbound molecular dynamics trajectory of the G6A mutant while reaching 0.4 in the ATP-bound molecular dynamics trajectory. A closer look into the high- and low-scored RNA conformations revealed a pattern of stacking interactions formed by the binding site nucleotides suitable for the ATP binding, which is present in the high-scored conformations but not in the low-scored conformations. These observations suggest that the mutant-type binding site, though suitable for ATP binding, is likely collapsed for a larger amount of time than the wild-type binding site. This, in turn, correlates with the experimental data, showing $20-80\%$ binding efficiency for the mutant-type with respect to the wild-type aptamer (39).

CONCLUSION

To conclude we would like to emphasize, that nucleic acid structures differs from protein ones both in the atomic composition and structural folds, making difficult a direct application of protein binding site detection methods. Here we designed a specific typization for nucleic-acid structures that covers various nucleotides and suitable for both DNA and RNA, as well as their multiple chain complexes. We developed a 3D convolutional neural network, BiteNet$_N$, to identify small molecule binding sites in nucleic acid structures. To train BiteNet$_N$, we constructed a large dataset of ∼2000 nucleic acid–small molecule complexes and performed rigorous cross-validation using sequence- and structure-based splits to circumvent over-fitting. BiteNet$_N$ consistently outperformed the other methods on the constructed test sets. BiteNet$_N$ is conformation-specific, as we demonstrated by analyzing seven different HIV-1 TAR RNA structures bound small molecules. It is helpful for large-scale analysis, such as conformational ensemble or mutant variant analysis, as demonstrated in the ATP-aptamer case study. Finally, BiteNet$_N$ can operate with both RNA and DNA complexes, including multiple chains, and we made it publicly available at https://sites.skoltech.ru/imolecule/tools/bitenet/.

DATA AVAILABILITY

BiteNet$_N$ and the constructed datasets are publicly available at https://sites.skoltech.ru/imolecule/tools/bitenet.

SUPPLEMENTARY DATA

Supplementary data are available in NARGAB online.

ACKNOWLEDGEMENTS

We acknowledge the Zhores supercomputer (20) used to train BiteNet$_N$.

FUNDING

No external funding.
Conflict of interest statement. None declared.

REFERENCES

1. Warner,K.D., Hajdin,C.E. and Weeks,K.M. (2018) Principles for targeting RNA with drug-like small molecules. *Nat. Rev. Drug. Discov.*, **17**, 547–558.
2. Ortiz de Luzuriaga,I., Lopez,X. and Gil,A. (2021) Learning to model G-quadruplexes: current methods and perspectives. *Ann. Rev. Biophys.*, **50**, 209–243.
3. Panchal,V. and Brenk,R. (2021) Riboswitches as drug targets for antibiotics. *Antibiotics*, **10**, 45.
4. McClorey,G. and Wood,M.J. (2015) An overview of the clinical application of antisense oligonucleotides for RNA-targeting therapies. *Curr. Opin. Pharmacol.*, **24**, 52–58.
5. Matsui,M. and Corey,D.R. (2017) Non-coding RNAs as drug targets. *Nat. Rev. Drug. Discov.*, **16**, 167–179.
6. Falese,J.P., Donlic,A. and Hargrove,A.E. (2021) Targeting RNA with small molecules: from fundamental principles towards the clinic. *Chem. Soc. Rev.*, **50**, 2224–2243.
7. Yu,A.-M., Choi,Y.H. and Tu,M.-J. (2020) RNA drugs and RNA targets for small molecules: Principles, progress, and challenges. *Pharmacol. Rev.*, **72**, 862–898.
8. Disney,M.D., Winkelsas,A.M., Velagapudi,S.P., Southern,M., Fallahi,M. and Childs-Disney,J.L. (2016) Inforna 2.0: a platform for the sequence-based design of small molecules targeting structured RNAs. *ACS Chem. Biol.*, **11**, 1720–1728.
9. Zeng,P., Li,J., Ma,W. and Cui,Q. (2015) Rsite: a computational method to identify the functional sites of noncoding RNAs. *Sci. Rep.-UK*, **5**, 1–5.
10. Zeng,P. and Cui,Q. (2016) Rsite2: an efficient computational method to predict the functional sites of noncoding RNAs. *Sci. Rep.-UK*, **6**, 1–9.
11. Wang,K., Jian,Y., Wang,H., Zeng,C. and Zhao,Y. (2018) RBind: computational network method to predict RNA binding sites. *Bioinformatics*, **34**, 3131–3136.
12. Su,H., Peng,Z. and Yang,J. (2021) Recognition of small molecule-RNA binding sites using RNA sequence and structure. *Bioinformatics*. **37**, 36–42.
13. Kozlovskii,I. and Popov,P. (2020) Spatiotemporal identification of druggable binding sites using deep learning. *Commun. Biol.*, **3**, 1–12.
14. Kozlovskii,I. and Popov,P. (2021) Protein–peptide binding site detection using 3D convolutional neural networks. *J. Chem. Inform. Model*. **61**, 3814–3823.
15. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
16. Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
17. Gong,S., Zhang,C. and Zhang,Y. (2019) RNA-align: quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA. *Bioinformatics*, **35**, 4459–4461.
18. Kingma,D.P. and Ba,J. (2015) Adam: a method for stochastic optimization. *3rd International Conference on Learning Representations, (ICLR), May 7-9, Conference Track Proceedings*. San Diego, CA, USA.
19. Popov,P. and Grudinin,S. (2018) Eurecon: equidistant uniform rigid-body ensemble constructor. *J. Mol. Graph. Model*, **80**, 313–319.
20. Zacharov,I., Arslanov,R., Gunin,M., Stefonishin,D., Bykov,A., Pavlov,S., Panarin,O., Maliutin,A., Rykovanov,S. and Fedorov,M. (2019) 'Zhores'—Petaflops supercomputer for data-driven modeling, machine learning and artificial intelligence installed in Skolkovo Institute of Science and Technology. *Open Engineering*, **9**, 512–520.
21. Coimbatore Narayanan,B., Westbrook,J., Ghosh,S., Petrov,A.I., Sweeney,B., Zirbel,C.L., Leontis,N.B. and Berman,H.M. (2014) The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Res.*, **42**, D114–D122.
22. Mitternacht,S. (2016) FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research*, **5**, 189.
23. Bannwarth,S. and Gatignol,A. (2005) HIV-1 TAR RNA: the target of molecular interactions between the virus and its host. *Curr. HIV Res.*, **3**, 61–71.
24. Abulwerdi,F.A. and Le Grice,S.F. (2017) Recent advances in targeting the HIV-1 Tat/TAR complex. *Curr. Pharm. Design*, **23**, 4112–4121.
25. Aboul-ela,F., Karn,J. and Varani,G. (1995) The structure of the human immunodeficiency virus type-1 TAR RNA reveals principles of RNA recognition by Tat protein. *J. Mol. Biol.*, **253**, 313–332.
26. Du,Z., Lind,K.E. and James,T.L. (2002) Structure of TAR RNA complexed with a Tat-TAR interaction nanomolar inhibitor that was identified by computational screening. *Chem. Biol.*, **9**, 707–712.
27. Faber,C., Sticht,H., Schweimer,K. and Rösch,P. (2000) Structural rearrangements of HIV-1 Tat-responsive RNA upon binding of neomycin B. *J. Biol. Chem.*, **275**, 20660–20666.
28. Murchie,A.I., Davis,B., Isel,C., Afshar,M., Drysdale,M.J., Bower,J., Potter,A.J., Starkey,I.D., Swarbrick,T.M., Mirza,S. *et al.* (2004) Structure-based drug design targeting an inactive RNA conformation: exploiting the flexibility of HIV-1 TAR RNA. *J. Mol. Biol.*, **336**, 625–638.
29. Davis,B., Afshar,M., Varani,G., Murchie,A.I., Karn,J., Lentzen,G., Drysdale,M., Bower,J., Potter,A.J., Starkey,I.D. *et al.* (2004) Rational design of inhibitors of HIV-1 TAR RNA through the stabilisation of electrostatic 'hot spots'. *J. Mol. Biol.*, **336**, 343–356.
30. Davidson,A., Begley,D.W., Lau,C. and Varani,G. (2011) A small-molecule probe induces a conformation in HIV TAR RNA capable of binding drug-like fragments. *J. Mol. Biol.*, **410**, 984–996.
31. Davidson,A., Leeper,T.C., Athanassiou,Z., Patora-Komisarska,K., Karn,J., Robinson,J.A. and Varani,G. (2009) Simultaneous recognition of HIV-1 TAR RNA bulge and loop sequences by cyclic peptide mimics of Tat protein. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 11931–11936.
32. Davidson,A., Patora-Komisarska,K., Robinson,J.A. and Varani,G. (2011) Essential structural requirements for specific recognition of HIV TAR RNA by peptide mimetics of Tat protein. *Nucleic Acids Res.*, **39**, 248–256.
33. Borkar,A.N., Bardaro,M.F., Camilloni,C., Aprile,F.A., Varani,G. and Vendruscolo,M. (2016) Structure of a low-population binding intermediate in protein-RNA recognition. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 7171–7176.
34. Shortridge,M.D., Wille,P.T., Jones,A.N., Davidson,A., Bogdanovic,J., Arts,E., Karn,J., Robinson,J.A. and Varani,G. (2019) An ultra-high affinity ligand of HIV-1 TAR reveals the RNA structure recognized by P-TEFb. *Nucleic Acids Res.*, **47**, 1523–1531.
35. Dunn,M.R., Jimenez,R.M. and Chaput,J.C. (2017) Analysis of aptamer discovery and technology. *Nat. Rev. Chem.*, **1**, 1–16.
36. Röthlisberger,P. and Hollenstein,M. (2018) Aptamer chemistry. *Adv. Drug. Deliver. Rev.*, **134**, 3–21.
37. Kim,Y.S., Raston,N. H.A. and Gu,M.B. (2016) Aptamer-based nanobiosensors. *Biosens. Bioelectron.*, **76**, 2–19.
38. Zhu,G. and Chen,X. (2018) Aptamer-based targeted therapy. *Adv. Drug. Deliver. Rev.*, **134**, 65–78.
39. Huizenga,D.E. and Szostak,J.W. (1995) A DNA aptamer that binds adenosine and ATP. *Biochemistry*, **34**, 656–665.
40. Biniuri,Y., Luo,G.-F., Fadeev,M., Wulf,V. and Willner,I. (2019) Redox-switchable binding properties of the ATP–aptamer. *J. Am. Chem. Soc.*, **141**, 15567–15576.
41. Lin,C.H. and Patei,D.J. (1997) Structural basis of DNA folding and recognition in an AMP-DNA aptamer complex: distinct architectures but common recognition motifs for DNA and RNA aptamers complexed to AMP. *Chem. Biol.*, **4**, 817–832.
42. Xie,Y.-C., Eriksson,L.A. and Zhang,R.-B. (2020) Molecular dynamics study of the recognition of ATP by nucleic acid aptamers. *Nucleic Acids Res.*, **48**, 6471–6480.