biomolecules

MDPI

Review

Advances and Challenges in Scoring Functions for RNA–Protein Complex Structure Prediction

Chengwei Zeng †, Chen Zhuo †, Jiaming Gao, Haoquan Liu and Yunjie Zhao *

Institute of Biophysics and Department of Physics, Central China Normal University, Wuhan 430079, China; cwzengwuhan@mails.ccnu.edu.cn (C.Z.); chenzhuowh@mails.ccnu.edu.cn (C.Z.); jmgao@mails.ccnu.edu.cn (J.G.); liuhaoquan@mails.ccnu.edu.cn (H.L.)

* Correspondence: yjzhaowh@ccnu.edu.cn

† These authors contributed equally to this work.

Abstract: RNA–protein complexes play a crucial role in cellular functions, providing insights into cellular mechanisms and potential therapeutic targets. However, experimental determination of these complex structures is often time-consuming and resource-intensive, and it rarely yields high-resolution data. Many computational approaches have been developed to predict RNA–protein complex structures in recent years. Despite these advances, achieving accurate and high-resolution predictions remains a formidable challenge, primarily due to the limitations inherent in current RNA–protein scoring functions. These scoring functions are critical tools for evaluating and interpreting RNA–protein interactions. This review comprehensively explores the latest advancements in scoring functions for RNA–protein docking, delving into the fundamental principles underlying various approaches, including coarse-grained knowledge-based, all-atom knowledge-based, and machine-learning-based methods. We critically evaluate the strengths and limitations of existing scoring functions, providing a detailed performance assessment. Considering the significant progress demonstrated by machine learning techniques, we discuss emerging trends and propose future research directions to enhance the accuracy and efficiency of scoring functions in RNA–protein complex prediction. We aim to inspire the development of more sophisticated and reliable computational tools in this rapidly evolving field.

Keywords: RNA-protein complex; scoring function; machine learning; structure prediction; molecular docking

Citation: Zeng, C.; Zhuo, C.; Gao, J.; Liu, H.; Zhao, Y. Advances and Challenges in Scoring Functions for RNA–Protein Complex Structure Prediction. *Biomolecules* **2024**, *14*, 1245. https://doi.org/10.3390/biom14101245

Academic Editor: Kyungsook Han

Received: 4 September 2024
Revised: 24 September 2024
Accepted: 30 September 2024
Published: 1 October 2024

1. Introduction

RNA–protein complexes are vital for cellular functions, such as DNA repair, RNA splicing, and protein synthesis [1–3]. They play a crucial role in gene regulation and the maintenance of chromosome ends [4,5]. Disruptions in RNA–protein interactions are linked to various human diseases, including cancer [6], AIDS [7], and neurodegenerative disorders [8,9]. For example, HIV, a global retrovirus that attacks the human immune system, had caused approximately 39 million infections and 630,000 AIDS-related deaths by the end of 2022. The core mechanism of HIV infection involves an RNA–protein complex, in which the viral protein Tat takes over the host's positive transcription elongation factor b (P-TEFb) along with the cis-acting transactivation response element (TAR) RNA to regulate transcription elongation. It is crucial to understand the structure of these RNA–protein complexes [7,10]. However, the lack of crystal structures is a major obstacle to developing effective therapies. Therefore, understanding these complexes is crucial for cell biology and for developing targeted therapies [11–13].

Understanding RNA–protein interactions requires 3D structural information. However, experimental methods like X-ray crystallography, NMR, and cryo-electron microscopy are costly and time-consuming [14,15]. The high flexibility and complex interaction patterns of RNA make it challenging to determine their structures through experimental

methods [16,17]. For instance, it is difficult to achieve high-resolution crystal structures via X-ray crystallography [18], maintain RNA stability during cryo-electron microscopy [19], and obtain precise NMR spectroscopy data for larger RNA molecules [20]. As of 17 August 2024, the Protein Data Bank (PDB) contained 223,790 structures, but only 4888 were RNA–protein complexes [21]. When redundancies were removed, fewer than 400 high-resolution, unique, non-redundant RNA–protein complexes were available for analysis [22,23]. This number is significantly lower than the expected number of RNA–protein complex structures formed within cells.

Therefore, computational structure prediction has gained attention as a viable alternative. A platform for assessing advancements in computational structure prediction is the biennial Critical Assessment of protein Structure Prediction (CASP) competition [24], complemented by CAPRI for protein complexes [25] and RNA-Puzzles for RNA structures [26]. In recent years, CASP challenges have expanded beyond protein structure prediction to include RNA and RNA–protein complex prediction. CASP15 featured two RNA–protein complex targets, and CASP16 introduced additional targets, emphasizing the growing focus on accurately modeling these complex biological interactions. Reliable computational methods can bridge the gap between the scarcity of known RNA–protein structures and the biological processes they control. The demand for accurate theoretical methods to predict the structure of RNA–protein complexes is becoming increasingly urgent. From a computational perspective, RNA–protein complex structure prediction primarily relies on docking, which involves two main steps: conformational sampling and evaluation [27–30]. The flexibility of RNA and proteins leads to conformational changes, making it difficult to sample near-native structures adequately [31,32]. Sometimes, the sampling process may generate tens of thousands of possible structures, yet none closely resembling the native state [27,33]. Another main challenge lies in conformational evaluation, where a scoring function is used to rank and identify near-native structures among tens of thousands of possible structures [22]. This process is particularly challenging because it requires accurately distinguishing models between near-native structures and others. Existing scoring functions for RNA–protein complexes are based on various assumptions and constructed using different methodologies, each with strengths and limitations.

Several scoring functions have been developed for evaluating the structure of RNA–protein complexes, building on earlier advances in protein and RNA structure prediction [12,22,31]. The cornerstone of this field is knowledge-based scoring functions, which evaluate RNA–protein interactions as a weighted sum of pairwise statistical potentials [34,35]. These statistical potentials utilize formulas derived from the inverse Boltzmann principle. Knowledge-based methods can be categorized into coarse-grained and all-atom approaches. Coarse-grained potentials, such as DARS- RNP, QUASI-RNP [36], and 3dRPC-Score [37], rely on a statistical analysis of interface propensities between nucleotide–residue pairs, capturing pairwise interactions effectively. The coarse-grained representation allows for quicker computational speeds, making them suitable for high-throughput analysis, especially effective when complex formation induces minor structural alterations. In contrast, all-atom potentials, such as dRNA [12], ITScore-PR [22], and DITScore-PR [31], provide higher accuracy when dealing with near-native models among decoys due to their higher spatial resolution. These methods are particularly effective in bound–bound cases where fine structural details are crucial. However, they are less effective in handling unbound–unbound predictions as significant conformational changes occur upon binding.

Deep neural networks have shown potential in recent years across diverse fields, including biophysics in structure prediction [38–40]. Well-established methods are available for modeling the 3D structures of proteins [41–46], RNAs [47–54], and protein–protein complexes [55–57]. Recent studies have shown that machine learning is now excelling in evaluating RNA–protein complex structures. DRPScore, a 4D-CNN-based scoring function, has achieved success comparable to knowledge-based methods in bound–bound cases and has surpassed them in more challenging unbound–unbound cases [23]. These

advancements highlight the potential of deep neural networks in revolutionizing the evaluation of RNA–protein complex structures.

This review provides a comprehensive overview of scoring functions for RNA–protein structure prediction, presenting the latest advances in the field. We discuss the fundamental principles of various scoring functions, including knowledge-based approaches (coarse-grained and all-atom models) and machine-learning-based methods (Figure 1). Moreover, we assess the strengths and limitations of current scoring techniques, offering a detailed evaluation of their performance. Given the significant progress demonstrated by machine learning approaches, we propose future research directions that could further enhance the accuracy and efficiency of scoring functions. We aim to inspire the development of more robust and sophisticated scoring techniques to advance RNA–protein complex structure prediction.

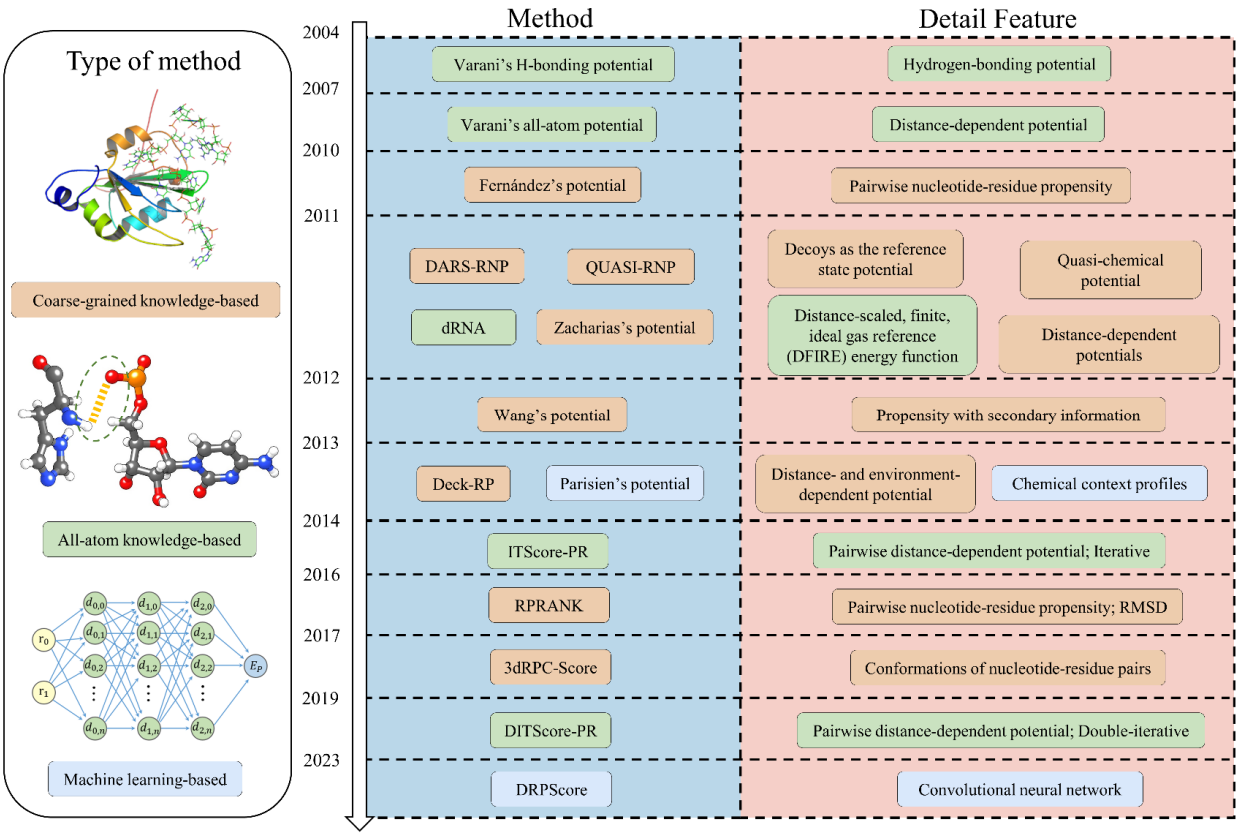

Figure 1. Timeline of the development of RNA–protein complex structure prediction. The most recent advancements in RNA–protein structure prediction encompass coarse-grained knowledge-based, all-atom knowledge-based, and machine-learning-based approaches.

2. Knowledge-Based Scoring Functions

Knowledge-based scoring functions are mathematical functions derived from statistical observations of interactions at the interfaces of known RNA–protein complexes. Figure 2 shows that these functions commonly use the inverse Boltzmann relationship to convert distance-dependent pairwise contact probability distributions into statistical potential functions. This means that specific nucleotide and residue interactions are observed at the RNA–protein interface with a higher frequency than would be expected by random chance [58–60].

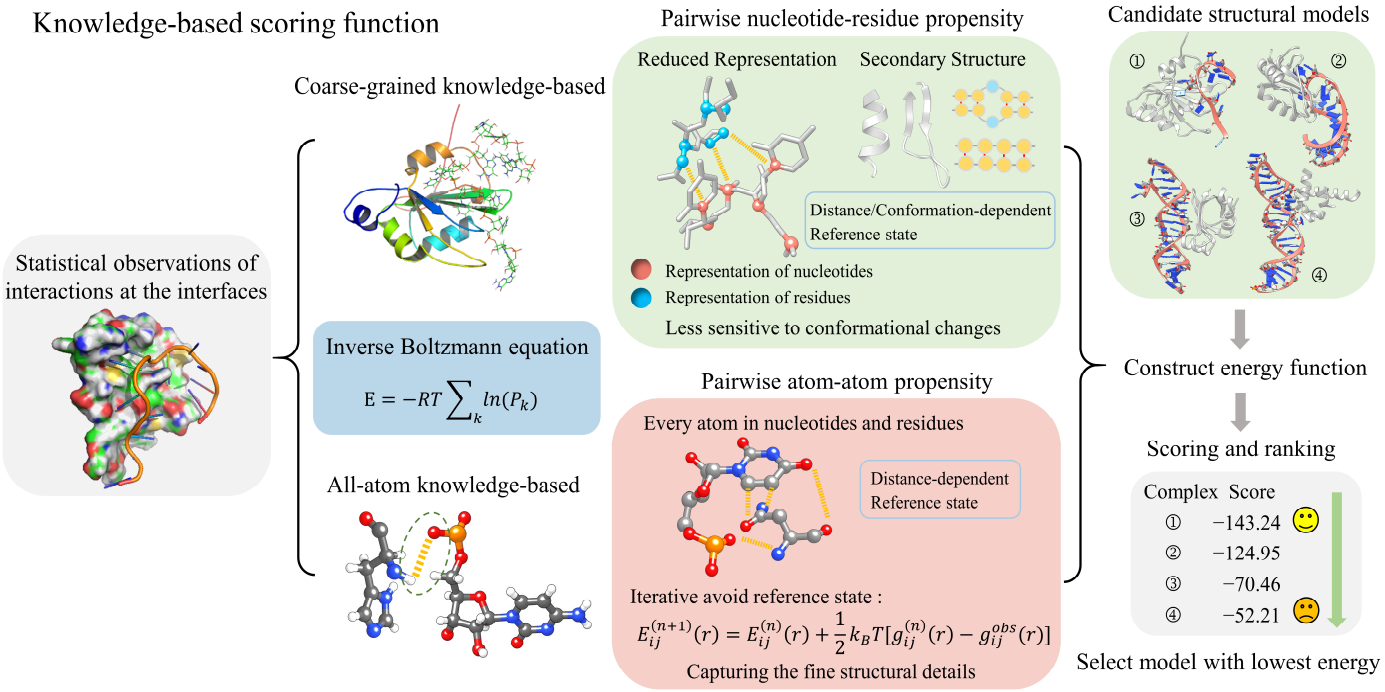

Figure 2. The process and principles of knowledge-based scoring functions in evaluating RNA–protein complexes. These scoring functions can be categorized into coarse-grained and all-atom models derived from the inverse Boltzmann equation. Coarse-grained scoring functions utilize a simplified representation, while all-atom scoring functions account for every atom within nucleotides and residues. Once the energy function is constructed, these scoring functions can evaluate and rank RNA–protein complexes, allowing the selection of structures with the lowest energy scores.

For instance, the highly negative charge of RNA attracts positively charged amino acids, such as arginine (ARG), lysine (LYS), and histidine (HIS), which are prevalent at the RNA–protein interaction interface [58,61,62]. Electrostatic interactions are crucial for the stabilization of RNA–protein complexes [58,63,64]. Notably, guanine (G) is found at the interface more frequently than cytosine (C), adenine (A), or uracil (U), with its occurrence exceeding 30%. On the other hand, while hydrophobic residues, including aromatic residues, are the least favored at these interfaces, aromatic residues still play an important role in interacting with unpaired RNA bases [65].

Consequently, RNA–protein interactions are primarily mediated by electrostatic forces rather than the hydrophobic forces and desolvation effects that dominate protein–protein binding. It has been observed that the 'LYS-ARG' fragment in proteins is a highly favorable binding motif. In RNA, 'CG' and 'GG' are the preferred binding fragments among the nucleotides [62]. This preference is primarily due to guanine's unique double-ring structure, which facilitates extensive hydrogen bonding and stacking interactions. These enhance stability and versatility in interactions with amino acids. Knowledge-based scoring functions measure the tendency of these distance-dependent pairwise contacts to occur at the interface [34,35].

The Boltzmann distribution in statistical physics provides a foundation for understanding interaction probabilities in thermal equilibrium. The interaction energy ΔE_k correlates with the probability distribution P_k as follows:

$$P_k \propto e^{-\frac{\Delta E_k}{RT}} \quad (1)$$

where P_k is the probability of the system being in a state k where the interaction or distance occurs; the state can either be a specific nucleotide/residue or a specific nucleotide–residue

pair. ΔE_k is the interaction energy between states k, R is the gas constant, and T is the absolute temperature. In previous work, the value of RT was set at 0.59 kcal/mol.

Therefore, the inverse Boltzmann distribution can allow us to derive the interaction potential energy. By statistically analyzing the interaction frequencies from existing structural databases, corresponding potential energy functions can be inferred. For example, consider the distribution probability P_k of state k at the interface. According to the inverse Boltzmann principle, this observed probability distribution can be directly related to the interaction potential energy ΔE_k:

$$\Delta E_k = -RTln(P_k) \tag{2}$$

Therefore, negative statistical potential values indicate favorable binding energies. Thus, the total score ΔE for a given RNA–protein complex is obtained by summing the interaction energies for all specific states at the interface:

$$\Delta \mathrm{E} = \sum_k \Delta E_k \tag{3}$$

The next and most crucial step is to analyze the geometric information of interaction pairs at the interaction interfaces of RNA–protein complexes in known structural databases (such as RCSB PDB [21] and NDB [66,67]) to obtain the probability distribution P_k. In general, for each selected state k, the observed frequency N_k in all RNA–protein complexes is counted and then normalized to obtain the probability distribution P_k:

$$P_k = \frac{N_k}{\sum_k N_k} \tag{4}$$

where $\sum_k N_k$ is the total frequency observed across all states.

The definition of state k can differ among research groups. This state may be defined by a general distance range, such as 1 to 5 Å, or by specific criteria tailored to the type of interaction and the involved atomic groups. Several classical formulations exist for constructing the propensity distribution P_k. As illustrated in Equation (5), the propensity for a nucleotide or a residue of type k is classically defined as the ratio of the observed frequencies:

$$P_k = \frac{N_k^I / \sum_K N_k^I}{N_k^A / \sum_K N_k^A} \tag{5}$$

where N_k^I is the number of nucleotides/residues of type k involved in the interface, and $\sum_K N_k^I$ is the total number of interface nucleotides/residues. N_k^A is the total number of nucleotides/residues of type k, and $\sum_K N_k^A$ is the total number of nucleotides/residues.

The Fernández-Recio group adopted a different propensity definition, emphasizing the surface nucleotides and residues [68]. They defined propensity as

$$P_k = \frac{N_k^I / \sum_K N_k^I}{N_k^S / \sum_K N_k^S} \tag{6}$$

where N_k^I and $\sum_K N_k^I$ have the same definitions given in equation (5), N_k^I is the number of nucleotides/residues of type k involved in the interface, and $\sum_K N_k^I$ is the total number of interface nucleotides/residues. N_k^S is the total number of nucleotides/residues of type k on the surface, and $\sum_K N_k^S$ is the total number of nucleotides/residues on the surface.

Additionally, the propensity of nucleotide–residue pairs on the interfaces of RNA–protein complexes is used to develop propensity-based statistical potentials. The nucleotide–residue pairs are defined based on a cutoff distance between the nearest atoms [63]. Several statistical potentials based on propensity have been developed for evaluating RNA–protein complexes. These models consider factors such as secondary structure details, relative distances, and orientations between nucleotide–residue pairs. In the upcoming sections, we will provide a detailed analysis emphasizing the selection of state k at either the residue

level (coarse-grained) or the atomic level (all-atom). By taking this approach, we aim to assess the impact of these different resolutions on the accuracy and applicability of the statistical potentials used in modeling RNA–protein interactions.

2.1. Coarse-Grained Knowledge-Based Scoring Functions

Over the last decade, significant progress has been made in developing coarse-grained potentials for evaluating the structure of RNA–protein complexes (see Table 1). Starting with Fernández's work, which established the basis for distance-dependent pairwise nucleotide–residue scoring [63], subsequent models like DARS-RNP and QUASI-RNP improved on this approach by including more advanced reference states and better representations [36]. The Xiao group further developed this concept with Deck-RP [69], RPRANK [70], and 3dRPC-Score [37], incorporating secondary structure information and utilizing new statistical methods beyond simple distance dependence. These enhancements have broadened the usefulness of coarse-grained potentials, making them more effective in a broader range of structural evaluation situations. The following sections will provide a detailed overview of these methods.

Table 1. List of coarse-grained knowledge-based scoring methods for RNA–protein complex structure evaluation. This table includes the development time, the representation of RNA–protein molecules, the type and features of these methods, and their availability.

Name	Time	Feature	Availability as a Standalone Method	Reference
Fernández's potential	2010	Pairwise nucleotide–residue propensity	N/A	[63]
DARS-RNP	2011	Decoys as the reference state potential	https://genesilico.pl/software/stand-alone/statistical-potentials (accessed on 29 September 2024)	[36]
QUASI-RNP	2011	Quasi-chemical potential	https://genesilico.pl/software/stand-alone/statistical-potentials (accessed on 29 September 2024)	[36]
Zacharias's potential	2011	Distance-dependent potential	N/A	[71]
Wang's potential	2012	Pairwise nucleotide–residue propensity with secondary information	N/A	[72]
Deck-RP	2013	Distance- and environment-dependent potential	http://biophy.hust.edu.cn/new/3dRPC (accessed on 29 September 2024)	[69]
RPRANK	2016	Pairwise nucleotide–residue propensity; RMSD	http://biophy.hust.edu.cn/new/3dRPC (accessed on 29 September 2024)	[70]
3dRPC-Score	2017	Conformations of nucleotide–residue pairs	http://biophy.hust.edu.cn/new/3dRPC (accessed on 29 September 2024)	[37]

Fernández's potential: In 2010, the Fernández group developed a distance-dependent pairwise nucleotide–residue propensity to score the RNA–protein complexes [63]. In this approach, we can determine the propensities by comparing the observed frequencies of

specific nucleotide–residue pairs (where i = 1 to 4 for nucleotides and j = 1 to 20 for residues) at the RNA–protein interface with their expected frequencies.

$$P_{ij} = \frac{N_{ij}^I / \sum_{ij} N_{ij}^I}{N_i^S / \sum_i N_i^S \times N_j^S / \sum_j N_j^S} \quad (7)$$

where N_{ij}^I is the number of pairs between nucleotide type i and residue type j at the interface, $\sum_{ij} N_{ij}^I$ is the total number of nucleotide–residue pairs at the interface, and N_i^S and N_j^S are the number of nucleotides of type i and the number of residues of type j on the interface, respectively, while $\sum_i N_i^S$ and $\sum_j N_j^S$ are the total number of nucleotides and residues on the surface, respectively. These expected frequencies are based on the overall composition of RNA and protein surfaces. The nucleotide–residue pairs were defined by having at least one atom within a cutoff distance of 4 Å from each other, which serves as the distance threshold for defining a contact. Additionally, the surface nucleotide or residue was defined as that with an ASA (accessible surface area) > 0.1 Å^2. This distance-dependent potential uses a cutoff distance to evaluate contacts between nucleotide–residue pairs. While this approach is less sensitive to subtle differences in model structures, particularly those that share identical contact pairs, it offers the advantage of being more resilient to minor conformational changes. This tolerance makes it useful in scenarios where slight structural variations are expected. Still, this approach cannot fully capture the highly detailed interactions within the complex model. Unfortunately, this potential was designed to improve the discriminative power of the FTDock potential and is not available as a standalone program.

DARS-RNP and QUASI-RNP: In 2011, Tuszynska and Bujnicki introduced two medium-resolution, coarse-grained potentials—namely, the quasi-chemical potential (QUASI-RNP) and the Decoys As the Reference State potential (DARS-RNP)—to evaluate RNA–protein complex structures [36]. This coarse-grained methodology simplifies the all-atom representation of macromolecular structures into a reduced form based on nucleotide or residue type [73]. The backbone is represented by two united atoms for nucleotides: one for the phosphate group (P) and one for the ribose (RIB). Pyrimidines are modeled with a single atom, while purines are represented with two atoms. Conversely, residues are depicted using one to three united atoms, depending on their molecular size.

These two potentials, QUASI-RNP and DARS-RNP, use the same mathematical base but differ in their reference state:

$$E = E_d + E_a + E_s + E_p \quad (8)$$

where E is the total energy term, and E_d, E_a, E_s, and E_p are the distance-dependent, angular-dependent, site-dependent, and penalty terms for steric clashes, respectively. All four terms of the energy function are equally weighted for they exhibit comparable values. Among these energy terms, the interaction energies E_d, E_a, and E_s between the united atom type from the RNA i and the united atom type from the protein j are calculated by the same formula:

$$E(i,j,d) = -RTln\frac{N_{obs}(i,j,d)}{N_{exp}(i,j,d)} \quad (9)$$

where E represents E_d, E_a, or E_s, $N_{obs}(i,j,d)$ denotes the number of observed contacts between atom types i and j within a specific distance or angular bin d in the training set, and $N_{exp}(i,j,d)$ refers to the expected number of contacts within the same distance/angular bin in the reference state. The bin size is not standardized and is determined through empirical testing. There are two bin types: a distance bin of 1 Å used for the distance-dependent term E_d, and an angular bin of 20° used for the angular-dependent term E_a. The energy for each RNA–protein united atom pair is calculated when they are within 9 Å of each other. For the site-dependent term E_s, the parameter d represents one of three interaction types between

residues and nucleotide edges: Watson–Crick, Sugar, or Hoogsteen edges [74]. The penalty term E_p for steric clashes restricts united atom pairs, preventing them from approaching within a predefined cutoff distance.

It is essential to address the reference state problem when creating a distance-dependent statistical potential energy between pairs of particles. While calculating $N_{obs}(i, j, d)$, from a given training dataset, the observed number of contacts between atom types i and j with bin d is straightforward. Still, estimating the expected contact number, $N_{exp}(i, j, d)$, is more challenging. The expected distribution of paired nucleotide residues over distances can be adjusted using valid reference state definitions, including mean reference states, quasi-chemical approximate reference states, and finite ideal-gas reference states. QUASI-RNP and DARS-RNP employ distinct methods to determine the reference state in their calculations. For QUASI-RNP, molar fractions of residues are used to calculate $N_{exp}(i, j, d)$:

$$N_{exp}(i, j, d) = X_i * X_j * N_{obs}(d) \tag{10}$$

where X_i and X_j are the molar fractions of atom types i and j in the given training set, respectively, and $N_{obs}(d)$ is the total number of contacts in bin d irrespective of atom type. For DARS-RNP, $N_{exp}(i, j, d)$ is a normalized number of contacts between atom types i and j in bin d, calculated from 1000 decoys generated by the docking program GRAMM [75] for each RNA–protein complex in the training set. In both bound and unbound docking tests, DARS-RNP demonstrated a slightly stronger performance than QUASI-RNP to identify near-native structures. DARS-RNP is constructed from a much more extensive training set, providing a more realistic representation of "random" protein–RNA interactions. These two scoring functions are designed to be less affected by structural changes. As a result, they are expected to be more effective in distinguishing between different structures when complex formation causes only small changes. Furthermore, these functions provide a better spatial resolution and a more accurate representation of the reference state, leading to an improved accuracy in distinguishing between near-native structures and decoys compared to Fernandez's potential. In cases where molecules are already bound together, these functions were able to effectively differentiate between similar RNA–protein complex structures with small differences (RMSD < 10 Å). These functions showed a competitive discriminative ability in more challenging cases where the molecules were not initially bound together. The package of the model is freely available at https://genesilico.pl/software/stand-alone/statistical-potentials (accessed on 29 September 2024).

Zacharias's potential: In 2011, The Zacharias group developed a distance-dependent, coarse-grained force field for RNA–protein docking [71]. This potential enables fully systematic docking through energy minimization in the binding partners' rotational and translational degrees of freedom. In this coarse-grained approach, each residue is represented by up to four pseudo atoms (beads): two for the main chain nitrogen (N) and oxygen (O), and one or two for the short and long side chains, respectively. For nucleotides, three pseudo atoms represent the phosphate/ribose part, and three or four represent purine and pyrimidine bases. There are 31 pseudo-atom types for proteins and 17 pseudo-atom types for RNA.

This potential assumed pairwise additive interactions between protein and RNA beads, which are described by a distance-dependent potential with two forms, corresponding to attractive and repulsive interactions. The attractive potential is of the Lennard-Jones type:

$$U_{ij}^{attr} = \epsilon_{ij}\left(\frac{\sigma_{ij}}{r^8} - \frac{\sigma_{ij}}{r^6}\right) \tag{11}$$

The repulsive potential is

$$U_{ij}^{rep}(r) = \begin{cases} U_{ij}^{attr}(r) + 2U_{ij}^{m}, & r \leq r_{ij}^{m} \\ -U_{ij}^{attr}(r), & r > r_{ij}^{m} \end{cases} \tag{12}$$

Two pairwise-specific parameters σ_{ij} and ϵ_{ij} describe the interaction of each pair ij of RNA and protein beads, governing the interaction range and strength, respectively. r_{ij}^{m} and U_{ij}^{m} correspond to the position and minimum value of U_{ij}^{attr}. Thus, there are, in total, 1054 parameters ($31 \times 17 \times 2$) that need to be derived in a knowledge-based manner. Similar to DARS-RNP and QUASI-RNP [36], the distance-dependent statistical potentials $E(i, j, d)$ were constructed for each bead pair by a set of RNA–protein complexes determined through Equation (9). Then, the initial values of σ and ϵ parameters were obtained by fitting the attractive potential in Equation (11) and repulsive potential in Equation (12) to $E(i, j, d)$. The parameter values were subsequently adjusted to optimize docking results, aiming to find the correct (close to native) binding mode and achieve appropriate scoring. These potentials can accommodate moderate conformational changes but cannot be used without the ATTRACT docking protocol.

Wang's potential: In 2012, the Wang group developed four pairwise nucleotide–residue propensity potentials from a given training set, depending on whether the secondary structure element (SSE) information of RNA and proteins was considered [72]. Based on the propensity values of protein SSEs, eight types of SSEs calculated by the DSSP program were categorized into three classes: X (π-helix "I", 3_{10}-helix "G", and bend "S", whose $p > 1$), Y (β-sheet "E", β-bridge "B", turn "T", and unclassified, whose $p \approx 1$), and Z (α-helix "H", whose $p < 1$). Similarly, three types of nucleotides calculated by the X3DNA program were categorized into two classes: NP (unpaired and non-WC paired nucleotides, whose $p > 1$) and P (WC paired nucleotides, whose $p < 1$). Therefore, similar to Equation (9), the propensity can be calculated from the observed probability of the specific residue–nucleotide pair of type $ai - bj$ (where $a = 1...20$ for residues, i = X, Y, Z for protein secondary structure classes, $b = 1...4$ for nucleotides, j = P, NP for RNA secondary structure classes) at interfaces, divided by the expected probability. The authors concluded that the RNA secondary structure information plays a more significant role than the protein secondary structure in accurately discriminating the RNA–protein complex structures. Unfortunately, this potential is not available as a standalone program.

Deck-RP: In 2013, the Xiao group developed Deck-RP, a distance- and environment-dependent potential specifically designed for RNA–protein complexes generated by RPDOCK [69]. Deck-RP merges the strengths of both Wang's potential and DARS-RNP by incorporating an enhanced reference state that accounts for propensities, secondary structure states, and interface preferences of nucleotides and residues. The reference state in Deck-RP is a hybrid, composed of a decoy-based component and a molar-fraction-corrected component. The decoy-based component takes account of all decoys in the training set as the reference state, while the molar-fraction-corrected component takes account of the interface concentration or specific preferences of nucleotides and residues. As a result, similar to Equation (9), the propensity of residue–nucleotide pairs can be derived from their observed probabilities. The model considers 168 unique nucleotide–residue pairs, encompassing four nucleotide types across two secondary structure states and seven residue types across three secondary structure states. The 3dRPC protocol includes the docking program RPDOCK and the scoring program Deck-RP, which has been developed into a user-friendly web-server version at http://biophy.hust.edu.cn/new/3dRPC (accessed on 29 September 2024), while the package is freely available at http://biophy.hust.edu.cn/new/resources/3dRPC (accessed on 29 September 2024).

RPRANK: In 2016, the Xiao group developed a new knowledge-based potential, RPRANK, using root mean square deviation (RMSD) as a measure [70]. Unlike the previous statistical potential, RPRANK does not use distance to classify the residue-base pairs directly. The conformational differences between nucleotide–residue pairs from decoys and standard pairs from native structures were used to calculate the statistical potential. The nucleotide–residue pairs are clustered based on the RMSD between each other. Then, the energies of the nucleotide–residue pair clusters are decided by a statistical method based on the number of pairs in each cluster. The 3dRPC protocol includes the docking program RPDOCK and the scoring program RPRANK, which has been developed into a user-friendly

webserver version at http://biophy.hust.edu.cn/new/3dRPC (accessed on 29 September 2024). The package is freely available at http://biophy.hust.edu.cn/new/resources/3dRPC (accessed on 29 September 2024).

3dRPC-Score: In 2017, the Xiao group introduced a new statistical potential called 3dRPC-Score [37]. Unlike the commonly used distance-dependent statistical potential, this method considers the conformations of nucleotide–residue pairs as statistical variables. The group proposed that accurately defining the energy of a nucleotide–residue pair requires considering not only the relative distance between the partners but also their relative distance and orientation. They classified the nucleotide–residue pairs into 10 classes based on the relative root mean square deviation (RMSD) between their conformations. This classification allows pairs with similar conformations to be considered to have the same energy. Therefore, the statistical potential $E_{ij}(C)$ could be calculated:

$$E_{ij}(C) = -ln\left(\frac{P_{ij}(C)}{P_iP_j * P_v}\right) \tag{13}$$

where $P_{ij}(C)$ is the occurrence probability of the pair of i-type nucleotide and *j*-type residue in class *C*, P_i and P_j are the probabilities of nucleotide *i* and residue *j* at the interface, respectively, and P_v is the probability of class *C* in the whole conformational space of nucleotide–residue pairs in an ideal state. In an ideal state, each class of nucleotide–residue pairs has the same probability in conformational space. Thus,

$$E_{ij}(C) = -ln\left(\frac{P_{ij}(C)}{P_iP_j}\right) + constant \tag{14}$$

where the constant = lnP_v. The scoring function performs best when the constant is set as −4. The 3dRPC webserver is available at http://biophy.hust.edu.cn/new/3dRPC (accessed on 29 September 2024). The package can be downloaded at http://biophy.hust.edu.cn/new/resources/3dRPC (accessed on 29 September 2024).

Coarse-grained representations are less sensitive to conformational changes, which makes them suitable for high-throughput scenarios and situations involving minor to moderate conformational changes. These representations are expected to have greater discriminatory power when complex formation induces only minor structural alterations in its components. However, they may struggle to capture fine structural details and complex molecular interactions, especially when dealing with significant conformational changes. In such cases, these methods may need to be supplemented with higher-resolution techniques for tasks requiring detailed structural analysis or when addressing more complex challenges.

2.2. All-Atom Knowledge-Based Scoring Functions

The development of all-atom knowledge-based scoring functions for evaluating RNA–protein complexes has progressed from basic models to more advanced techniques (Table 2). The Varani group initially created a hydrogen-bonding potential to lay the foundation for specific recognition between proteins and RNA based on the sequence [11]. Later, realizing that hydrogen bonds only represent a fraction of the interactions at RNA–protein interfaces, they introduced an all-atom, distance-dependent potential to improve the accuracy of structural predictions [76]. Building on this work, the Zhou group developed dRNA, using a carefully constructed reference state to enhance the accuracy of pairwise potentials [12]. Subsequently, the Zou and Huang groups advanced the field with ITScore-PR [22] and DITScore-PR [31], both employing iterative processes to improve potentials and effectively eliminate the need for a predefined reference state, thus enhancing accuracy and applicability, especially in the consideration of fine structural details in RNA–protein interactions. In the following sections, we will provide a detailed overview of each of these methods.

Table 2. List of all-atom knowledge-based scoring methods for RNA–protein complex structure evaluation. This table includes the development time, the representation of RNA–protein molecules, the type and features of these methods, and their availability.

Name	Time	Feature	Availability as a Standalone Method	Reference
Varani's H-bonding potential	2004	Hydrogen-bonding potential	N/A	[11]
Varani's all-atom potential	2007	Distance-dependent potential	N/A	[76]
dRNA	2011	Volume-fraction corrected distance-scaled, finite, ideal gas reference (DFIRE) energy function	N/A	[12]
ITScore-PR	2014	Pairwise distance-dependent potential; iterative	https://zoulab.dalton.missouri.edu/resources_itscorepr.html (accessed on 29 September 2024)	[22]
DITScore-PR	2019	Pairwise distance-dependent potential; double-iterative	http://huanglab.phys.hust.edu.cn/mprdock/ (accessed on 29 September 2024)	[31]

Varani's H-bonding potential: In 2004, the Varani group developed an atomic-level, distance- and orientation-dependent hydrogen-bonding (H-bond) potential [11]. This hydrogen-bonding potential consists of a distance-dependent energy term [$E(\delta_{HA})$] and three angular-dependent energy components: $E(\Theta)$ (the angle at the hydrogen atom), $E(\Psi)$ (the angle at the acceptor atom), and $E(X)$ (the dihedral angle of the hydrogen bond). The total hydrogen-bond energy (E_{HB}) is then derived as a linear combination of these four distance- and orientational-dependent terms under the assumption that they are independent of each other:

$$E_{HB} = E(\delta_{HA}) + E(\Theta) + E(\Psi) + E(X) \quad (15)$$

However, hydrogen bonds represent only approximately 25% of the contacts at RNA–protein complex interfaces. A more comprehensive approach is needed to effectively describe the full types of interactions occurring at these interfaces. Unfortunately, this potential is not available as a standalone program.

Varani's all-atom potential: In 2007, the Varani group developed a distance-dependent statistical potential for predicting sequence-specific recognition between proteins and RNA, building upon the previous H-bonding potential, which represents only one aspect of the complex interactions at play [76]. This all-atom potential treats every atom, in every nucleotide and residue, as a unique type (e.g., Ala C_β and Arg C_β are considered unique atom types under this scheme), resulting in a total of 158 protein and 81 RNA atom types. Chemically similar atoms were grouped together based on the CHARMM atom definitions, allowing interactions between these atoms to be treated consistently. This potential is useful for distinguishing between RNA–protein complex models similar to the native structure, particularly those with a root mean square deviation (RMSD) of less than 5 Å. However, in practical unbound–unbound cases, obtaining a significant number of decoys with RMSD < 5 Å is challenging. This potential is not available as a standalone program.

dRNA: In 2011, Zhou group developed dRNA, a volume-fraction-corrected, distance-scaled, finite, ideal gas reference (DFIRE) statistical energy function and a measure of relative structural similarity by Z-score for RNA–protein complex interactions [12]. The definition of the reference state is critical for developing distance-dependent potentials accurately. The reference state serves as the basis for comparing observed interactions, and its precise formulation is essential for the reliability of the statistical potential. However,

accurately determining the functional form of the reference state remains a significant challenge. Errors in the reference state can result in inaccuracies in the calculated potentials, especially when pairwise interactions are inaccurately represented or abrupt truncations are applied [77,78]. The final statistical energy $E_{ij}(r)$ could be calculated as follows:

$$E_{ij}(r) = \begin{cases} -\eta ln \dfrac{N_{obs}(i,j,r)}{\left(\frac{f_i^v(r) f_j^v(r)}{f_i^v(r_{cut}) f_j^v(r_{cut})}\right)^{\beta} \frac{r^{\alpha}\Delta r}{r_{cut}^{\alpha}\Delta r_{cut}} N_{obs}^{lc}(i,j,r_{cut})}, & r < r_{cut} \\ 0, & r \geq r_{cut} \end{cases} \quad (16)$$

where the volume-fraction factor $f_i^v(r)$ is

$$f_i^v(r) = \frac{\sum_j N_{obs}^{RNA-protein}(i,j,r)}{\sum_j N_{obs}^{All}(i,j,r)} \quad (17)$$

where $N_{obs}(i,j,r)$ is the number of pairs of atom i and atom j within the spherical shell at distance r observed in a given RNA–protein complex structure database, and the interaction cutoff distance r_{cut} is 15 Å. Δr_{cut}, the bin width at r_{cut}, is 0.5 Å. The value of α is set to 1.61, determined by the best fit of r^{α} to the actual distance-dependent number of ideal-gas points in finite protein-sized spheres. The value of β for volume correction is set to 0.5. The factor η is 0.01 to control the magnitude of the energy score. Similar to Varani's all-atom potential, this all-atom potential also treats every atom, in every nucleotide and residue, as a unique type, resulting in a total of 167 protein and 86 RNA atom types. dRNA offers an improved accuracy in distinguishing low-RMSD, near-native models from thousands of decoys compared to coarse-grained scoring functions, primarily due to its higher spatial resolution inherent in the energy function. Unfortunately, this potential is not available as a standalone program.

ITScore-PR: In 2014, the Zou group developed a pairwise distance-dependent atomic interaction potential, ITScore-PR, using a statistical mechanics-based iterative method [22]. ITScore-PR addresses the reference state problem by iteratively improving the interatomic pair potentials. This is achieved by comparing RNA–protein complexes' predicted pair distribution functions with the experimentally observed pair distribution functions of native crystal structures in a specific training set. The potential $E_{ij}(r)$ over all atom pairs ij in the RNA and protein is determined through an iterative formula:

$$E_{ij}^{(n+1)}(r) = E_{ij}^{(n)}(r) + \Delta E_{ij}^{n}(r) \quad (18)$$

$$\Delta E_{ij}^{n}(r) = \frac{1}{2} k_B T \left[g_{ij}^{(n)}(r) - g_{ij}^{obs}(r) \right] \quad (19)$$

where n denotes the iterative step, and $E_{ij}^{(n+1)}(r)$ are the improved potentials from $E_{ij}^{(n)}(r)$ after correction, used in the next iterative step. The separations r between atom i and atom j are divided into bins of 0.2 Å with a maximum cutoff value of 10 Å. $g_{ij}^{(n)}(r)$ and $g_{ij}^{obs}(r)$ stand for the pair distribution functions for atom pair ij, calculated according to $E_{ij}^{(n)}(r)$ and calculated from the native crystal structures in the training set, respectively. $g_{ij}^{obs}(r)$ is calculated by the following:

$$g_{ij}^{obs}(r) = \frac{1}{k} \sum_{k=1}^{K} g_{ij}^{k*}(r) \quad (20)$$

where K is the total number of the RNA–protein complexes in the training set, and $g_{ij}^{k*}(r)$ is the pair distribution function of the k-th native complex structure. $g_{ij}^{(n)}(r)$ is the pair distribution function calculated from the ensemble of the binding modes according to the

binding score-dependent Boltzmann probabilities P_k^l obtained from the potential $E_{ij}^{(n)}(r)$ at the n-th step.

$$g_{ij}^{(n)}(r) = \frac{1}{k}\sum_{k=1}^{K}\sum_{l=0}^{L} P_k^l g_{ij}^{kl}(r) \tag{21}$$

where $g_{ij}^{kl}(r)$ is the pair distribution function for atom pair ij observed in the l-th binding state of the k-th complex. Thus, for a given set of initial potentials $E_{ij}^{(0)}(r)$,

$$E_{ij}^{(0)}(r) = \begin{cases} w_{ij}(r) & \text{, for hydrogen bond pairs} \\ \frac{v_{ij}(r)e^{-v_{ij}(r)} + w_{ij}(r)e^{-w_{ij}(r)}}{e^{-v_{ij}(r)} + e^{-w_{ij}(r)}} & \text{, otherwise} \end{cases} \tag{22}$$

where $v_{ij}(r)$ is the van der Waals (VDW) potential by ZDOCK 2.1, and $w_{ij}(r) = -k_B T \ln g_{ij}^{obs}(r)$ is the potential of mean force. The iteration continues through Equations (18)–(21) until all native structures in the training set can be discriminated from decoys by the current potentials. 12 RNA atom types and 20 protein atom types are used in this statistical potential. ITScore-PR clearly outperforms other scoring functions using detailed all-atom representation and an iterative processing approach, especially in bound–bound cases. The package is available at https://zoulab.dalton.missouri.edu/resources_itscorepr.html (accessed on 29 September 2024).

DITScore-PR: In 2019, building on ITScore-PR [22], the Huang group developed a set of effective pair potentials, DITScore-PR, for protein–RNA interactions using a double-iterative method [31]. This algorithm circumvents the reference state problem by updating the potentials until they can effectively distinguish native structures from binding decoys. It overcomes the decoy-dependent limitation by iteratively constructing the binding decoys. Similar to ITScore-PR but with a distinct approach, DITScore-PR consists of an inner loop and an outer loop for the two iteration processes:

$$E_{ij}^{(n,k+1)}(r) = E_{ij}^{(n,k)}(r) + \frac{1}{2}k_B T\left[g_{ij}^{(n,k)}(r) - g_{ij}^{obs}(r)\right] \tag{23}$$

where n and k stand for the iterative indices of the outer and inner loops, i and j represent the types of protein and RNA atoms, $g_{ij}^{(n,k)}(r)$ is the predicted pair distribution function by the current potentials $E_{ij}^{(n,k)}(r)$ at the k-th inner iteration cycle for a fixed n-th iterative cycle, and $g_{ij}^{obs}(r)$ is the experimentally observed pair distribution function in the native complex structures of a training set. The outer loop conducts one iteration cycle when the inner loop is completed once. The outer loop repeats until the iterative steps reach a set number or the inner loop is converged. The definition for RNA and protein atom types is the same as in ITScore-PR, which gave 12 RNA atom types and 20 protein atom types. The separations r between atom i and atom j are divided into bins of 0.2 Å with a maximum cutoff value of 9 Å. DITScore-PR demonstrates a higher accuracy than coarse-grained potentials in bound–bound cases, achieving a success rate of approximately 80%. However, while outperforming other approaches in the more challenging unbound–unbound cases, the success rate still requires improvement. In true flexible docking processes, where binding partners can adapt to each other, the performance of DITScore-PR remains limited in handling significant conformational changes. The package of the model is freely available at http://huanglab.phys.hust.edu.cn/mprdock/ (accessed on 29 September 2024).

Overall, all-atom potentials provide a superior accuracy and resolution in capturing the detailed atomic interactions of RNA–protein complexes compared to coarse-grained potentials. These methods are highly effective in bound–bound cases, outperforming coarse-grained potentials by offering more precise discrimination of near-native structures. However, the enhanced resolution of all-atom methods comes with increased computational complexity and a reduced performance in facing complex conformational changes. This limitation is particularly pronounced in true flexible docking scenarios, where binding

partners must adapt to substantial conformational changes [79]. Therefore, while all-atom potentials offer significant advantages in high-precision applications, their limitations in flexibility suggest that they may need to be supplemented with other techniques, especially in situations involving substantial conformational changes.

3. Machine-Learning-Based Scoring Functions

In recent years, rapid advancements in artificial intelligence have had a profound impact on science and technology [80–86]. One breakthrough example is AlphaFold [41–43], a machine-learning-based approach that has revolutionized protein structure prediction with remarkable accuracy. While methods for predicting and modeling the 3D structures of proteins [41–43,87], RNAs [47,48,50], and protein–protein complexes [55,88] have made significant progress, emerging research highlights the growing efficacy of machine learning in evaluating RNA–protein complex structures, as summarized in Figure 3 and Table 3. The following sections will provide a detailed overview of these cutting-edge methods, showcasing how machine learning is reshaping our understanding and evaluation of RNA–protein interactions.

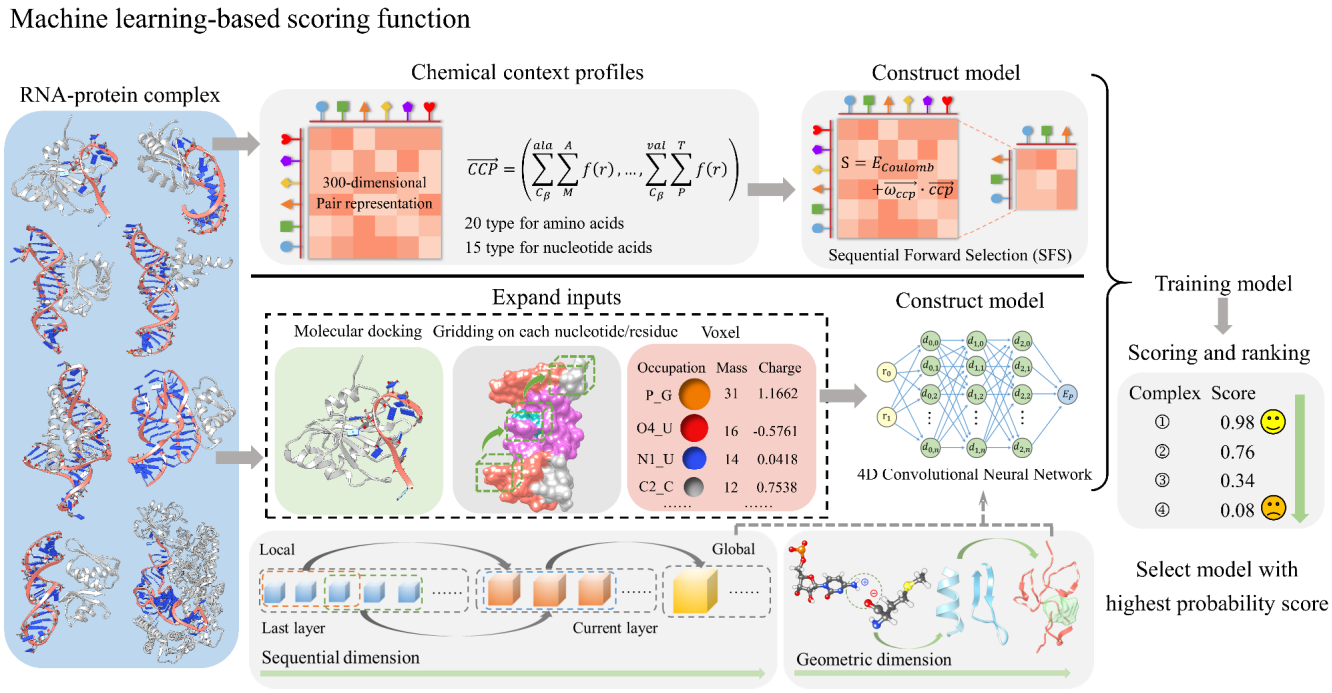

Figure 3. The process and principles of machine-learning-based scoring functions for evaluating RNA–protein complexes. The top approach in the figure employs chemical context profiles to represent RNA–protein complexes, followed by Sequential Forward Selection (SFS) to build a machine-learning model that reduces the initial 300-dimensional pair representation to a lower-dimensional space. The bottom approach first involves molecular docking, gridding on each nucleotide and residue, and using voxels containing atomic occupancy, mass, and charge to expand the input features. A 4D convolutional neural network is then employed to construct a machine learning model that integrates sequential and geometric dimensions. After model training, both models can score and rank RNA–protein complexes, facilitating the selection of structures with the highest probability scores.

Parisien's potential: In 2013, the Parisien group developed a machine-learning-based scoring function by utilizing the interface's CCP (chemical context profile) from known

RNA–protein complex structures [89] (Table 3). Specifically, the *CCP* is defined as a 300-dimensional vector:

$$\overrightarrow{CCP} = \left(\sum_{C_\beta}^{ala} \sum_{M}^{A} f(r), \sum_{C_\beta}^{ala} \sum_{m}^{A} f(r), \sum_{C_\beta}^{ala} \sum_{P}^{A} f(r), \ldots, \sum_{C_\beta}^{val} \sum_{P}^{T} f(r) \right) \tag{24}$$

Each double sum of *CCP* is the summation of the interaction strengths over a given pair. For example, the first interaction term involves the C_β of alanine (Ala) and the major groove of adenosine (A), with the interaction strength set to be inversely proportional to the distance between these pairs. This scoring function employs a simplified representation of both RNA and protein structures. Specifically, the 300 dimensions are derived from 20 amino acid types and 15 nucleic acid types. For nucleic acids, a heavy atom in the major groove (M), one in the minor groove (m), and a phosphate group (P) are selected, respectively, as the interaction centers for the five nucleotides [A, C, G, U, and T], covering both RNA and DNA. In the context of RNA, entries associated with thymines (T) have a CCP value of zero. For proteins, the C_β carbon atom servers as the interaction center for each residue, simplifying the complex interactions into a manageable framework while retaining essential biochemical information. $f(r)$ is the distance-dependent energy function assigned to each interaction pair:

$$f(r) = \frac{1}{\max\left(3.5\text{Å}, r - \hat{e}\right)} \tag{25}$$

where r is the distance between the interaction centers, and $\hat{e}$ is the average distance between C_β and its partner interaction center. Any RNA–protein complex is represented with the CCP vector. The similarity of the decoy and the native complexes can be obtained by computing the angle between their CCP vectors. This angle, or CCD (chemical context discrepancy), is defined as a relation in terms of two arbitrary vectors $\overrightarrow{CCP_1}$ and $\overrightarrow{CCP_2}$:

$$\cos(CCD) = \frac{\left(\overrightarrow{CCP_1} \cdot \overrightarrow{CCP_2}\right)}{\left(\left|\overrightarrow{\mathrm{CCP}_1}\right| \times \left|\overrightarrow{\mathrm{CCP}_2}\right|\right)} \tag{26}$$

Table 3. List of machine-learning-based scoring methods for RNA–protein complex structure evaluation. This table includes the development time, the representation of RNA–protein molecules, the type and features of these methods, and their availability.

Name	Time	Representation	Feature	Availability as a Standalone Method	Reference
Parisien's potential	2013	Coarse-grained	Chemical context profiles	N/A	[89]
DRPScore	2023	All-atom	Convolutional neural network	https://github.com/Zhaolab-GitHub/DRPScore_v1.0 (accessed on 29 September 2024)	[23]

The more different the CCPs, which represent the chemical properties of the RNA–protein complex interface, the greater the angle. Then, the CCP-based scoring function S is designed by weighting the entries of a CCP to identify near-native structures with low *CCD* values to the native structure:

$$S = \text{Coulomb} + \overrightarrow{\omega_{ccp}} \cdot \overrightarrow{ccp} \tag{27}$$

where Coulomb is the generic electrostatic energy term, and $\overrightarrow{\omega_{ccp}}$ is a vector enabling the weighted sum of *CCP* components. The forward version of the sequential feature selection

(SFS) approach, a machine-learning-based method, is used to identify the most important interacting pairs among all possible ones, reducing the nonzero entries in $\overrightarrow{\omega_{ccp}}$ to 12 from the original 300 dimensions [90]. After training, these dimensions are further reduced to 12 for scoring tRNA–protein complexes and 6 for scoring other RNA–protein complexes. This potential is not available as a standalone program.

DRPScore: In 2023, the Zhao group developed a deep-learning-based scoring function, DRPScore, to better account for the structural flexibility of RNA–protein complexes [23]. Specifically, DRPScore utilizes a 4D convolutional neural network to train models that can effectively identify near-native RNA–protein structures. To overcome the limitations of scarce data, DRPScore utilizes physics-based simulations targeting RNA–protein interfaces, generating 500 decoy structures for each RNA–protein complex in the initial process. This approach enabled the creation of a training dataset with over 100,000 structures, a significant improvement over the typical dataset size of fewer than 300 structures in traditional knowledge-based methods. The input for DRPScore includes nucleotides and residues at the RNA–protein interaction interface within a 6 Å distance. The model accurately describes molecular systems at the atomic level, classifying 85 atom types for RNA nucleotides and 225 atom types for protein residues. It assigns accurate mass and charge values to each atom through detailed feature processing. The 4D convolutional approach includes an additional operation along the sequential dimension, preserving critical information about nucleotide and residue interactions.

During preprocessing, each RNA–protein complex is represented as a tensor with the dimensions $1 \times 3 \times L \times (H \times W \times D)$. Here, the value 3 corresponds to the three captured features: the accumulations of the occupation number, mass, and charge of the atoms within each grid box. The parameter L = 128 represents the maximum allowable length for RNA–protein complex sequences, while H, W, and D define the height, width, and depth of a 3D cube that represents each nucleotide and residue within the RNA–protein complex, with each dimension set to 32 units. This grid structure effectively captures the spatial arrangement and intricate interactions at the atomic level, providing a detailed representation of the molecular architecture and facilitating accurate modeling of RNA–protein interactions.

DRPScore comprises six layers, with the final layer being a fully connected layer for classification. Each of the first five layers includes a Conv4d module, an optional BatchNorm module, and a MaxPooling module. In these Conv4d modules, the channel numbers progressively change: 64, 128, 256, 512, and 512. The strides applied in each module are set to 2, 2, 2, 1, and 1, respectively. This design effectively halves the feature length of the RNA–protein complex in the first three blocks while maintaining it in the last two blocks. All MaxPooling layers have a kernel size and stride of 2, halving each pooling module's height, width, and depth dimensions. The final representation of the RNA–protein complex is obtained through global average pooling, resulting in an 8192-dimensional feature vector that encapsulates the complex's characteristics. Finally, after applying a 4D convolution in the last layer, an adaptive spatial pooling for the final RNA–protein complex representation $O_{overall}$ is utilized:

$$O_{overall} = \frac{1}{H}\frac{1}{W}\frac{1}{D}\sum_{i\in H}\sum_{i\in W}\sum_{i\in D} O_{LN}\,[i,j,k] \tag{28}$$

After adding a linear classification layer to the model, probability scores can be generated to evaluate and select near-native RNA–protein complex structures. The representations learned by DRPScore effectively capture intra- (local) and inter-nucleotide/residue (global) information. This is accomplished by integrating convolutional layers along the sequence dimension, while also expanding on the spatial dimension. Each layer progressively models a wider range of interactions between nucleotides and residues. It has been extensively evaluated for its ability to identify near-native RNA–protein structures across diverse cases. Although DRPScore achieves comparable success in bound–bound cases and outperforms knowledge-based methods in more challenging unbound–unbound

cases, its success rate in these unbound–unbound cases still requires substantial improvement. Recently, this method has also been successfully extended to evaluate the structure of DNA–protein complexes [91]. The package of the model is freely available at https://github.com/Zhaolab-GitHub/DRPScore_v1.0 (accessed on 29 September 2024).

4. Benchmarks and Datasets for Assessing Scoring Functions

Evaluating scoring functions for RNA–protein complexes requires rigorous testing on a 3D RNA–protein complex structural benchmark. Consequently, various benchmarks and datasets have been established to assess performance. This section discusses the multiple benchmarks curated by different groups (Table 4). Due to the inherent flexibility of both RNA and proteins, significant conformational changes can be induced during the docking process. The RNA–protein complexes in these datasets can be broadly categorized into three types: (1) bound–bound cases, (2) bound–unbound cases, and (3) unbound–unbound cases [92–94].

Table 4. List of RNA–protein complex docking benchmarks. The time and number of total, bound–unbound, unbound–unbound, easy, medium, and difficult cases are listed in this table.

Benchmark	Development	Time	Total Number of Cases	Number of Cases						Availability
				Bound–Unbound	Unbound–Unbound	Easy	Medium	Difficult	Reference	
Benchmark I	Zou group	2013	72	20	52	49	16	7	[95]	https://zoulab.dalton.missouri.edu/RNAbenchmark/index.htm (accessed on 29 September 2024)
Benchmark II	Fernández-Recio group	2012	106 *	62	9	64	24	18	[93]	https://life.bsc.es/pid/protein-rna-benchmark/ (accessed on 29 September 2024)
Benchmark III	Bahadur group	2012/2016	126	105	21	72	25	19	[92,94]	N/A

* Contains 35 homology-modeled cases.

In bound–bound cases, no conformational changes occur in the RNA or the protein during the docking process. This means that both monomers involved in the docking come from the same complex. Bound–unbound cases unequivocally involve conformational changes in the RNA or the protein, where one may exist in an unbound state or come from a different complex. Unbound–unbound cases explicitly refer to scenarios where both the RNA and protein are in unbound conformations or originate from two distinct complexes. These scenarios undeniably make the docking process particularly challenging due to the complex conformational shifts involved.

For a robust performance evaluation, benchmark datasets must possess three key characteristics: (1) Diversity of targets. The benchmarks must include a wide range of targets to effectively test the robustness of different molecular docking algorithms. (2) Experimentally resolved structures: It is crucial to use datasets derived from experimentally resolved structures to avoid introducing computational biases or errors. (3) Bound and unbound conformations: The benchmarks must contain both bound and unbound conformations of the individual monomers, allowing for the assessment of conformational changes upon complex formation [96,97].

Benchmark I constructed by Zou et al. comprises 72 RNA–protein complexes, among which 52 are unbound–unbound cases and 20 are bound–unbound cases [95]. Based on the degree of conformational change observed in unbound structures upon binding, these 72 RNA–protein complexes can be further categorized into 49 easy ($I_{rmsd} \leq 1.5$ Å or $f_{nat} \geq 0.8$), 16 medium (1.5 Å $< I_{rmsd} \leq 4.0$ Å and $0.4 < f_{nat} \leq 0.8$), and 7 difficult targets ($I_{rmsd} > 4.0$ Å or $f_{nat} < 0.4$). The interface root mean square deviation (I_{rmsd}) is defined as the RMSD of the interaction interface region after optimal superimposition of the bound and unbound conformations. The fraction of native contacts (f_{nat}) is defined as the proportion of native nucleotide–residue pairs in the unbound conformation. Specifically, it is the ratio of the number of native nucleotide–residue pairs in the optimally superimposed unbound conformation to the total number of nucleotide–residue pairs in the bound conformation. The RNA–protein complex benchmark can be accessed at https://zoulab.dalton.missouri.edu/RNAbenchmark/index.htm (accessed on 29 September 2024).

Benchmark II constructed by Fernandez-Recio et al. comprises 106 RNA–protein complexes [93]. Among these cases, 71 cases were taken from crystallography or NMR experiments, while 35 cases were built using homology modeling. Of the 71 experimental RNA–protein complexes, 9 unbound–unbound cases and 62 bound–unbound cases exist. The 35 homology-modeled cases consist of 13 unbound–model, 19 bound–model, and 3 model–model RNA–protein complexes. In unbound–model cases, the RNA or protein exists in an unbound state or comes from a different complex, while the other is a homology-based prediction structure. In bound–model cases, one molecule is in a bound state, and the other is a homology-based prediction structure. In model–model cases, both the RNA and protein are homology-based prediction structures, with no native complex involved. Based on the degree of conformational change observed in unbound structures upon binding, these 106 RNA–protein complexes can also be further categorized into 64 easy ($0 \leq I_{rmsd} < 2.5$ Å), 24 medium ($2.5 \leq I_{rmsd} \leq 5.0$ Å), and 18 difficult ($I_{rmsd} > 5.0$ Å) targets. The RNA–protein complexes benchmark can be accessed at https://life.bsc.es/pid/protein-rna-benchmark/ (accessed on 29 September 2024).

Benchmark III constructed by Bahadur et al. comprises 45 RNA–protein complexes, among which 9 are unbound–unbound cases and 36 are bound–unbound cases [92]. Based on the degree of conformational change observed in unbound structures upon binding, these 45 RNA–protein complexes can also be further categorized into 34 easy ($0 \leq I_{rmsd} < 1.5$ Å), 8 medium ($1.5 \leq I_{rmsd} < 3.0$ Å), and 3 difficult ($I_{rmsd} \geq 3.0$ Å). Later, Bahadur et al. developed an extended version of benchmark III [94]. The non-redundant RNA–protein complex benchmark contains 126 RNA–protein complexes, a three-fold increase in the number of structures compared to the previously proposed RNA–protein complex benchmark III. Among these cases, 21 are unbound–unbound cases and 105 are bound–unbound cases. Also, based on the degree of conformational change observed in unbound structures upon binding, these 126 RNA–protein complexes can also be further categorized into 72 easy ($0 \leq I_{rmsd} < 1.5$ Å), 25 medium ($1.5 \leq I_{rmsd} < 3.0$ Å), and 19 difficult ($I_{rmsd} \geq 3.0$ Å).

5. Criteria and Assessment of the Prediction Quality

The quality of RNA–protein complex predictions is evaluated using the CAPRI criteria [98,99], focusing on two primary metrics: interface root mean square deviation (I_{rmsd}) and ligand root mean square deviation (L_{rmsd}). I_{rmsd} measures the deviation at the interface between native and predicted structures after protein superposition. L_{rmsd} quantifies the displacement of the RNA under the same conditions. Specifically, interface residues and nucleotides are extracted from both the native RNA–protein complexes and the decoys, and superposition is performed to calculate I_{rmsd}. Similarly, all nucleotides from native and decoy structures are extracted and superposed to calculate L_{rmsd}. This ensures an accurate assessment of deviations at the interaction interface and the overall RNA conformation [100]. Typically, a decoy is classified as a near-native structure if its I_{rmsd} relative to the native complex is ≤4.0 Å, or if its L_{rmsd} is ≤10.0 Å. A scoring function is successful if it

ranks near-native structures among the top N decoys [22]. The root mean square deviation (including I_{rmsd} and L_{rmsd}) is defined as

$$RMSD = \sqrt{\frac{1}{N}\sum_i \left(\left| \overrightarrow{X_{A_i}} - \overrightarrow{X_{B_i}} \right|^2 + \left| \overrightarrow{Y_{A_i}} - \overrightarrow{Y_{B_i}} \right|^2 + \left| \overrightarrow{Z_{A_i}} - \overrightarrow{Z_{B_i}} \right|^2 \right)} \quad (29)$$

where $\overrightarrow{X}$, $\overrightarrow{Y}$, and $\overrightarrow{Z}$ represent the coordinates of the native and predicted structures. N is the total number of atoms.

The evaluation of RNA–protein structures is an area of research that has not been extensively explored. Previous studies have mainly focused on rigid-body docking, overlooking the structural flexibility inherent in RNA–protein interactions. Although scoring functions in rigid-body docking have achieved success rates of about 80%, they still need significant improvement for flexible docking scenarios, especially in fully flexible unbound–unbound docking. One major challenge in this field is accurately sampling the dynamic conformations that RNAs and proteins adopt during their interactions. In fully flexible unbound–unbound docking, the interaction interface can change dramatically, which presents a significant challenge to scoring functions that have not previously encountered such diverse structures.

As shown in Figure 4A, we evaluated the performance of various scoring functions on the most challenging unbound cases from RNA–protein complex benchmark I. The benchmark comprises 57 unbound cases, selected using a 0.95 sequence similarity cutoff, using CD-HIT to avoid redundancy [101–103]. The assessment focused on three types of scoring functions: coarse-grained knowledge-based methods (e.g., DAR-RNP and 3dRPC-Score), all-atom knowledge-based methods (e.g., ITScore-PR), and recent machine-learning-based methods (e.g., DRPScore). Among these, the machine-learning-based DRPScore consistently outperformed the traditional scoring functions. Specifically, DRPScore achieved the highest success rates across all prediction categories, reaching a peak of 57.89% for the top 10, 20, 30, 40, and 50 predictions. This result underscores the superiority of machine learning approaches in evaluating RNA–protein complex structures more accurately than conventional methods. However, despite the advancements represented by these scoring functions, the overall success rates across all methods remain below 60%, averaging 52.19% for the top 50 predictions. This limitation highlights the need for significant refinements in current approaches to achieve a higher accuracy in predicting RNA–protein interactions. These findings also indicate a need for developing more sophisticated models or integrating additional biological data to improve the accuracy of these tools, especially in complex conformational change docking scenarios. The detailed performance and corresponding PDB IDs are provided in Supplementary Tables S1 and S2.

Moreover, we further evaluated the performance of various scoring functions on the unbound cases with different interface interactions from RNA–protein complex benchmark I. We initially calculated the interaction interface of the RNA–protein complex using a 6 Å distance cutoff. The number of nucleotides and amino acids at the interaction interface ranged from a minimum of 23 to a maximum of 199, with an average of 75. Therefore, we categorized interactions based on the number of amino acids and nucleotides at the interface, using 70 as the threshold to distinguish between relatively small and large interface interactions. Figure 4B shows that the machine-learning-based DRPScore consistently outperformed traditional scoring functions for cases involving relatively small interface interactions. Moreover, Figure 4C shows that the success rates were slightly improved compared to those in small interface interactions for cases involving relatively large interface interactions. Even in the top 10 predictions, the average success rate for these methods reached 58.33%. However, the overall success rates across all methods remained relatively low, with the highest average success rate in the top 50 being just 63.54%. The detailed performance and corresponding PDB IDs are provided in Supplementary Tables S3–S6.

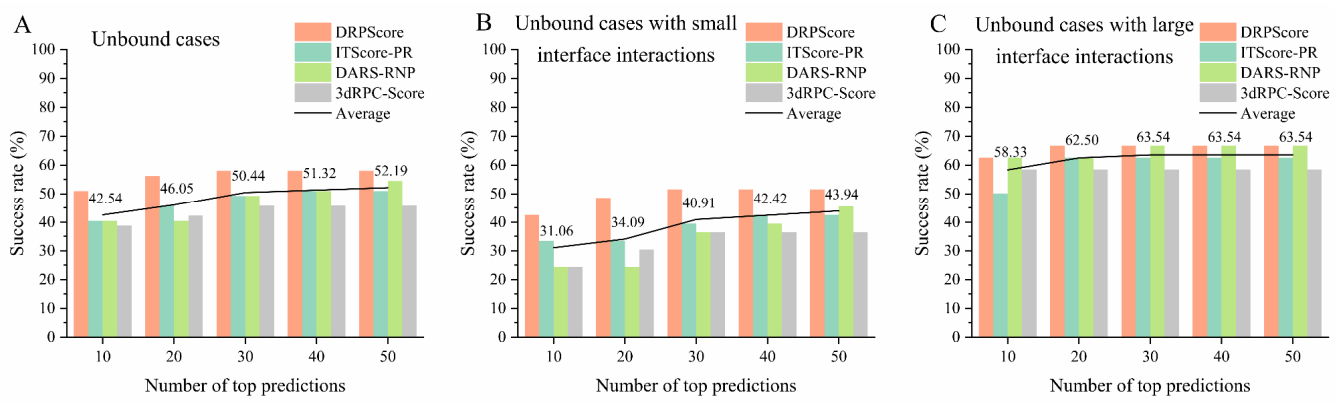

Figure 4. The performance of various scoring functions on the unbound cases from RNA–protein complex benchmark I. The success rate of DRPScore (orange bar), ITScore-PR (blue bar), DARS-RNP (green bar), 3dRPC-Score (gray bar), and the average (black line) on the (**A**) unbound cases, (**B**) unbound cases with small interface interactions, and (**C**) unbound cases with extensive interface interactions from RNA–protein complex benchmark I.

The results highlight the importance of interaction interface size in the performance of scoring functions for predicting RNA–protein complexes. While machine-learning-based approaches like DRPScore outperform traditional scoring methods in smaller interaction interfaces, there is still room for improvement in overall success rates. The superior performance of machine-learning-based methods is due to their ability to capture complex multi-body interactions at the interface, which traditional methods often miss. However, the limited information within smaller interfaces restricts the model's ability to learn and score interactions accurately. Additionally, smaller interfaces often have more structural flexibility and involve complex non-specific interactions, making precise modeling and prediction more challenging. Success rates improve with more extensive interaction interfaces. This is due to the increased contact points and features within these interfaces, allowing models to capture key interaction patterns more effectively and enhance prediction accuracy.

We further evaluated the performance of various scoring functions on unbound cases involving either single-stranded or double-stranded RNA partners from RNA–protein complex benchmark I. As shown in Figure 5A, the machine-learning-based DRPScore consistently outperformed traditional scoring functions in cases with single-stranded RNA partners. Figure 5B demonstrates a slight improvement in success rates compared to cases with double-stranded RNA partners. Overall, the performance of all scoring functions was lower in cases involving double-stranded RNA partners. In the top 10 predictions, the average success rate of each scoring function for single-stranded RNA cases was 43.89%, compared to 37.50% for double-stranded RNA cases. This discrepancy may be attributed to the increased complexity of multi-body interactions at the RNA–protein interface. In the case of double-stranded RNA, interactions involve both the protein and the inter-chain interactions between the two RNA strands. These additional layers of complexity make it more challenging for scoring functions to model the binding interface accurately. The detailed performance and corresponding PDB IDs are provided in Supplementary Tables S7–S10.

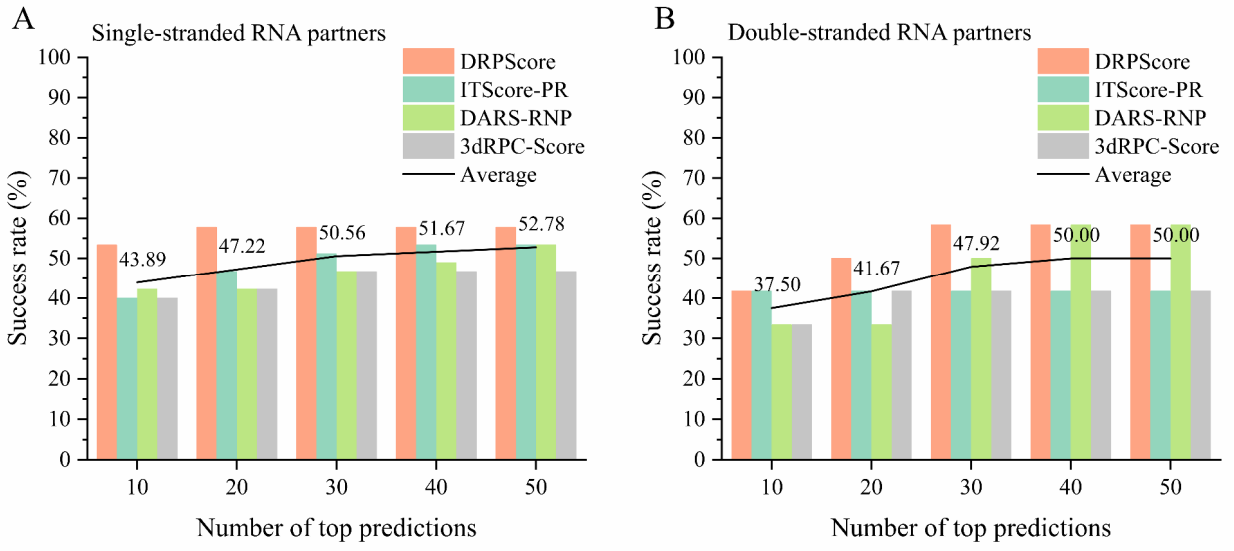

Figure 5. The performance of various scoring functions on the unbound cases from RNA–protein complex benchmark I. The success rate of DRPScore (orange bar), ITScore-PR (blue bar), DARS-RNP (green bar), 3dRPC-Score (gray bar), and average (black line) on the (**A**) single-stranded RNA partners, and (**B**) double-stranded RNA partners from RNA–protein complex benchmark I.

6. Discussion and Future Directions

Predicting the structure of RNA–protein complexes is essential for understanding biological processes and developing new treatments. Several factors influence local interactions in these complexes. Figure 6 shows the differences in structural selections made by each scoring function across three examples, highlighting their respective abilities to capture local interaction features. The analysis focuses on nucleotide–residue pairs within a cutoff distance of 6 Å compared to the native structures. The red and black dots represent nucleotide–residue pairs that are added or reduced relative to the native RNA–protein complex structure. Overall, each scoring function captures native interactions to different extents, with changes primarily occurring in localized regions. DDPScore captures interactions closely aligned with the native contacts, indicating minimal disruption. Similarly, ITScore-PR also captures interactions near the native positions, albeit to a slightly lesser degree. In contrast, DARS-RNP and 3dRPC-Score identify interactions that deviate further from the native contacts. This discrepancy may be due to the reliance of DARS-RNP and 3dRPC-Score on a coarse-grained representation, which simplifies molecular details and omits crucial side-chain interaction information. Since side-chain interactions play a crucial role in determining the specificity and strength of RNA–protein binding, these models may struggle to capture the nuanced geometric and energetic properties necessary for precise structural predictions. On the other hand, ITScore-PR, which counts atom–atom pairs, can capture more precise atom-level interactions. The machine-learning-based scoring function DRPScore efficiently recognizes local features of RNA–protein complex structures. This capability enables it to accurately capture binding patterns and effectively distinguish structures resembling the native state.

Specifically, we utilized methionyl-tRNAfMet formyltransferase complexed with formyl-methionyl-tRNAfMet (PDB ID: 2FMT) to analyze the electrostatic interactions between RNA and protein using PyMOL (version 1.8.0.3). Figure 7 shows the lowest interface RMSD model among the top 10 models selected by each scoring function. The lowest I_{rmsd} for the DRPScore-selected model is 3.73 Å, compared to 8.52 Å for ITScore-PR, 11.27 Å for DARS-RNP, and 16.13 Å for 3dRPC-Score. Overall, the RNA in the selected structures tends to bind to similar regions on the protein, likely due to intrinsic properties of the protein surface, such as electrostatic potential and hydrophobicity, which naturally favor specific binding sites. However, the varying RMSD values suggest geometric matching and positioning accuracy differences. DRPScore effectively captures the interface interaction

patterns and achieves a lower I_{rmsd}. In contrast, scoring functions like DARS-RNP and 3dRPC-Score show higher deviations, possibly due to their reliance on coarse-grained representations, which may lack the precision to accurately model spatial alignment and side-chain interactions essential for RNA–protein binding. ITScore-PR, with its moderate performance, balances these aspects but still falls short in capturing the intricate details of the RNA–protein interface. Since RNA carries a strong negative charge, it is expected to bind preferentially to the positively charged regions of proteins. The DRPScore-selected model shows RNA bound to a positively charged protein region. The model selected by ITScore-PR primarily binds to a positively charged region, with only minor deviations from favorable electrostatic interactions. In contrast, structures selected by DARS-RNP and 3dRPC-Score often bind to negatively charged regions, indicating less accurate electrostatic complementarity. This discrepancy may be attributed to DRPScore explicitly incorporating atomic charge information as input features during training, enabling it to capture electrostatic interactions precisely. Since the net charge of a protein is primarily distributed on its side chains, all-atom knowledge-based scoring functions can capture more detailed interaction features compared to coarse-grained scoring functions. Considering these findings, future advancements in scoring functions should focus on developing methods tailored to the specific characteristics of various interaction interfaces. A particular emphasis should be placed on accurately modeling complex multi-body interactions, to enhance prediction robustness and precision.

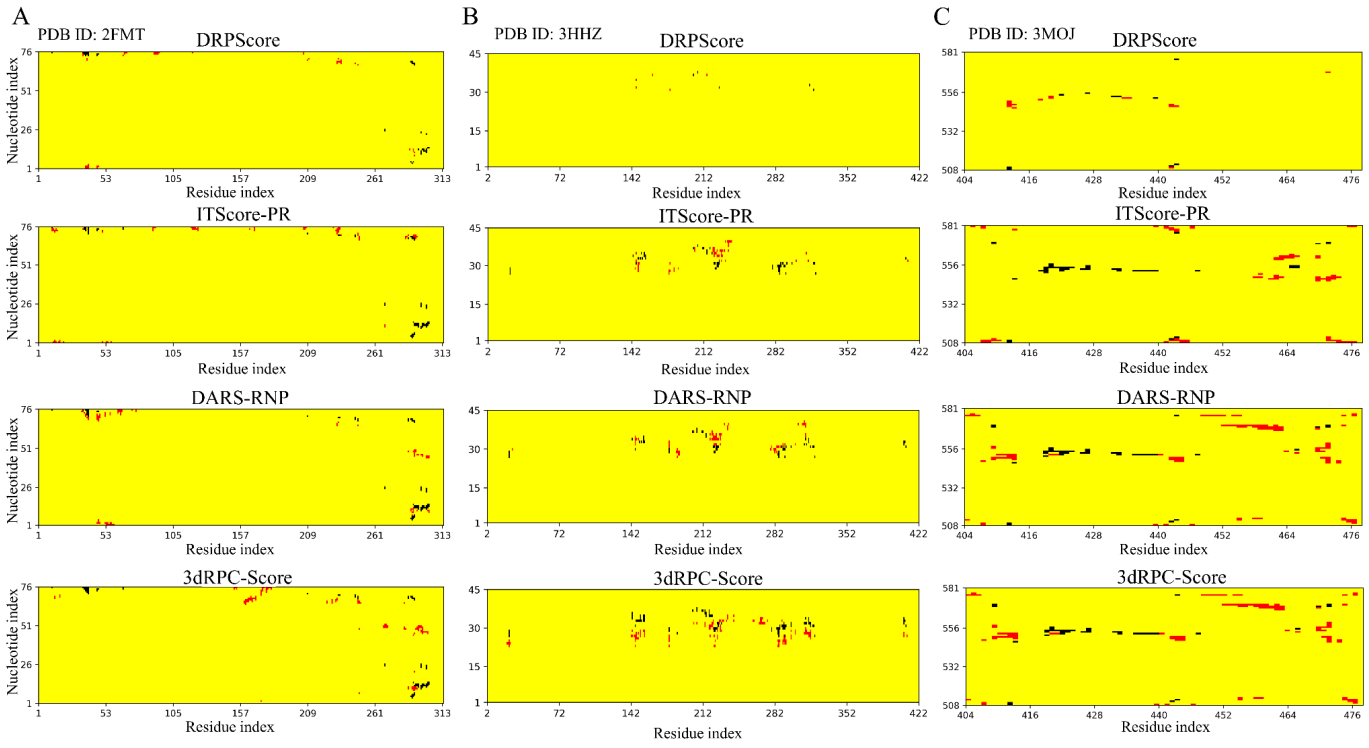

Figure 6. Contact distributions for three unbound docking examples. The contact maps (from top to bottom) show interactions between nucleotides and residues within a 6 Å range in the lowest RMSD model among the top 10 models selected by DRPScore, ITScore-PR, DARS-RNP, and 3dRPC-Score for (**A**) PDB ID: 2FMT, (**B**) PDB ID: 3HHZ, and (**C**) PDB ID: 3MOJ. Red and black dots indicate nucleotide–residue pairs that are added or reduced in the models selected by each scoring function compared to the native RNA–protein complex structure.

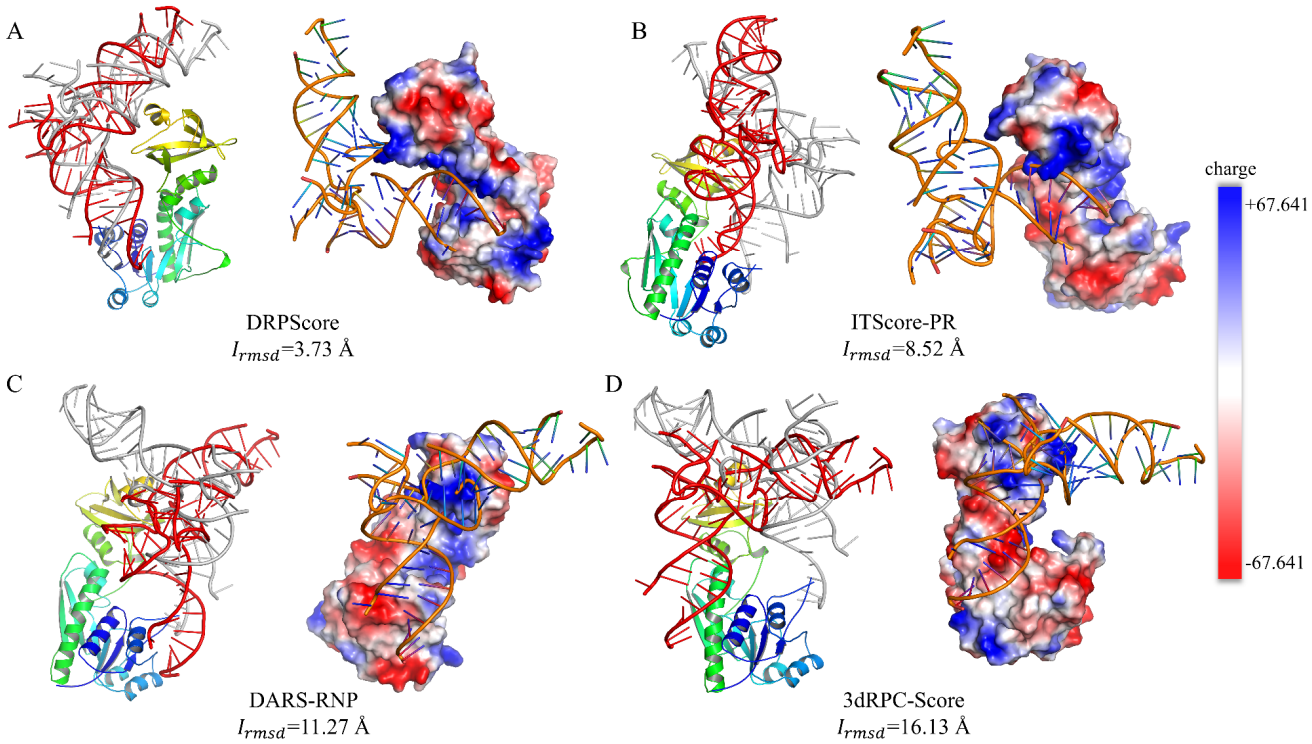

Figure 7. The lowest interface RMSD model and electrostatic interaction distribution. The lowest interface RMSD model and the corresponding electrostatic interaction distribution among the top 10 models (RNA in red) selected by (**A**) DRPScore, (**B**) ITScore-PR, (**C**) DARS-RNP, and (**D**) 3dRPC-Score, compared to the native RNA–protein complex (RNA in gray). Positively charged regions are shown in blue, while negatively charged regions are shown in red.

Recently, there has been a growing focus on predicting the structures of RNA–protein complexes in biomolecular research. This interest has been fueled by advancements in machine learning techniques, which have significantly improved structure prediction [39,43,104]. The increased interest in this area has led to significant progress in predicting the structures of individual structures and complex predictions like protein–protein and RNA–protein complexes.

This review has compared two main scoring functions for predicting RNA–protein complexes: knowledge-based (coarse-grained and all-atom) and machine-learning-based approaches. While each scoring function has its advantages, both types of scoring functions share a common limitation: they lack a strong theoretical foundation in physics. For example, knowledge-based scoring functions often assess RNA–protein interactions using a weighted sum of statistical potentials. However, the exact relationship between these scores and the system's free energy needs to be defined. Applying machine-learning-based scoring functions in evaluating RNA–protein complexes has shown promise. These functions outperform traditional methods by capturing complex, multibody interactions in RNA–protein binding. They can learn from vast amounts of data and represent intricate interaction patterns that are difficult for traditional methods. However, they have a relatively low success rate in challenging unbound–unbound cases, typically below 60%. These models struggle with structural diversity within training datasets, leading to potential overfitting. Integrating knowledge-based models could help mitigate these issues and enhance prediction accuracy.

One of the major challenges in predicting RNA–protein interactions is the conformational flexibility of both RNA and protein components upon binding. For example, in the NF-κB dimer system, the RMSD of RNA before and after binding can be as high as 5.4 Å [68]. Although extensive docking simulations can generate large datasets to address data scarcity, accurately modeling the loose atomic packing and unique interactions re-

mains a challenge. A combination of various docking and scoring methodologies can be used to develop consensus models, clustering predictions based on their scoring outcomes to enhance reliability. In cases where a consensus is not achieved, the top-scoring models from different methods could be proposed as alternative solutions. Knowledge-based statistical potentials are effective for rigid structures and large interaction interfaces. However, for cases with small interaction interfaces or complex flexible structures, deep learning approaches may be required to accurately capture the intricate multi-body interactions.

Current machine-learning-based scoring functions are hindered by the lack of three-dimensional structural data and the highly variable nature of RNA–protein interfaces. A promising method to enhance RNA–protein complex prediction involves integrating multi-scale modeling techniques, which combine coarse-grained and all-atom models to address the diverse nature of RNA–protein interfaces at different resolutions. This multi-stage approach allows for the rapid identification of potential conformations using coarse-grained models, then refined and precisely scored with all-atom models. The approach achieves detailed structural information and accommodates conformational changes. Moreover, developing dynamic scoring functions that adjust weights based on the local environment of RNA and proteins could provide greater flexibility in handling conformational changes, especially in regions with loosely packed atoms at the interface. Such approaches leverage the strengths of different scales, capturing relationships and features that single-modal methods might miss. Additionally, integrating high-resolution structural data from crystallography or NMR, low-resolution information from cryo-electron microscopy, or other experimental techniques could enhance the robustness of RNA–protein complex predictions [88,105–107].

RNA–protein complex prediction remains a challenge in the fields of soft matter physics and biophysics. With advancing computational techniques, structure prediction becomes a potent complement to experimental methods, offering fresh insights into the RNA–protein mechanism and downstream applications. Despite persistent challenges, particularly in flexible docking and complex assembly, the ongoing advancements in experimental and computational approaches are poised to drive transformative breakthroughs imminently.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/biom14101245/s1. Table S1: The performance of DRPScore, ITScore-PR, DARS-RNP, and 3dRPC-Score on unbound cases in RNA-protein complex benchmark I; Table S2: The PDB IDs of the unbound cases in RNA-protein complex benchmark I for the performance of DRPScore, ITScore-PR, DARS-RNP, and 3dRPC-Score; Table S3: The performance of DRPScore, ITScore-PR, DARS-RNP, and 3dRPC-Score on unbound cases with relatively small interface interactions in RNA-protein complex benchmark I; Table S4: The PDB IDs of the unbound cases with relatively small interface interactions in RNA-protein complex benchmark I for the performance of DRPScore, ITScore-PR, DARS-RNP, and 3dRPC-Score; Table S5: The performance of DRPScore, ITScore-PR, DARS-RNP, and 3dRPC-Score on unbound cases with relatively large interface interactions in RNA-protein complex benchmark I; Table S6: The PDB IDs of the unbound cases with relatively large interface interactions in RNA-protein complex benchmark I for the performance of DRPScore, ITScore-PR, DARS-RNP, and 3dRPC-Score; Table S7: The performance of DRPScore, ITScore-PR, DARS-RNP, and 3dRPC-Score on unbound cases with single-stranded RNA partners in RNA-protein complex benchmark I; Table S8: The PDB IDs of the unbound cases with relatively large interface interactions in RNA-protein complex benchmark I for the performance of DRPScore, ITScore-PR, DARS-RNP, and 3dRPC-Score; Table S9: The performance of DRPScore, ITScore-PR, DARS-RNP, and 3dRPC-Score on unbound cases with double-stranded RNA partners in RNA-protein complex benchmark I; Table S10: The PDB IDs of the unbound cases with relatively large interface interactions in RNA-protein complex benchmark I for the performance of DRPScore, ITScore-PR, DARS-RNP, and 3dRPC-Score.

Author Contributions: C.Z. (Chengwei Zeng) and C.Z. (Chen Zhuo) collected the information on the computational models and wrote the manuscript; J.G. and H.L. assisted with discussion and

information collection; Y.Z. designed the project and supervised the overall study. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (grant no. 12175081); the Science Fund for Distinguished Young Scholars of Hubei Province (grant no. 2024AFA077); the Fundamental Research Funds for the Central Universities (grant nos. CCNU22QN004, CCNU24JCPT011, and KJ02502022-0450); and Central China Normal University's excellent postgraduate education innovation funding project (grant no. 2024CXZZ146).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chung, C.S.; Tseng, C.K.; Lai, Y.H.; Wang, H.F.; Newman, A.J.; Cheng, S.C. Dynamic protein-RNA interactions in mediating splicing catalysis. *Nucleic Acids Res.* **2019**, *47*, 899–910. [CrossRef]
2. Licatalosi, D.D.; Darnell, R.B. RNA processing and its regulation: Global insights into biological networks. *Nat. Rev. Genet.* **2010**, *11*, 75–87. [CrossRef]
3. Lunde, B.M.; Moore, C.; Varani, G. RNA-binding proteins: Modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 479–490. [CrossRef]
4. Mittal, N.; Roy, N.; Babu, M.M.; Janga, S.C. Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 20300–20305. [CrossRef]
5. Muller-McNicoll, M.; Neugebauer, K.M. How cells get the message: Dynamic assembly and function of mRNA-protein complexes. *Nat. Rev. Genet.* **2013**, *14*, 275–287. [CrossRef]
6. Khalil, A.M.; Rinn, J.L. RNA-protein interactions in human health and disease. *Semin. Cell Dev. Biol.* **2011**, *22*, 359–365. [CrossRef]
7. Ning, S.; Zeng, C.; Zeng, C.; Zhao, Y. The TAR binding dynamics and its implication in Tat degradation mechanism. *Biophys. J.* **2021**, *120*, 5158–5168. [CrossRef]
8. Modic, M.; Ule, J.; Sibley, C.R. CLIPing the brain: Studies of protein-RNA interactions important for neurodegenerative disorders. *Mol. Cell. Neurosci.* **2013**, *56*, 429–435. [CrossRef]
9. De Conti, L.; Baralle, M.; Buratti, E. Neurodegeneration and RNA-binding proteins. *Wiley Interdiscip. Rev. RNA* **2017**, *8*, e1394. [CrossRef]
10. Khatkar, P.; Mensah, G.; Ning, S.B.; Cowen, M.; Kim, Y.; Williams, A.; Abulwerdi, F.A.; Zhao, Y.J.; Zeng, C.; Le Grice, S.F.J.; et al. HIV-1 Transcription Inhibition Using Small RNA-Binding Molecules. *Pharmaceuticals* **2024**, *17*, 33. [CrossRef]
11. Chen, Y.; Kortemme, T.; Robertson, T.; Baker, D.; Varani, G. A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys. *Nucleic Acids Res.* **2004**, *32*, 5147–5162. [CrossRef] [PubMed]
12. Zhao, H.; Yang, Y.; Zhou, Y. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. *Nucleic Acids Res.* **2011**, *39*, 3017–3025. [CrossRef]
13. Zhao, H.; Yang, Y.; Zhou, Y. Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. *RNA Biol.* **2011**, *8*, 988–996. [CrossRef]
14. Wu, J.; Niu, S.S.; Tan, M.; Huang, C.H.; Li, M.Y.; Song, Y.; Wang, Q.M.; Chen, J.; Shi, S.H.; Lan, P.F.; et al. Cryo-EM Structure of the Human Ribonuclease P Holoenzyme. *Cell* **2018**, *175*, 1393–1404. [CrossRef]
15. Khatter, H.; Myasnikov, A.G.; Natchiar, S.K.; Klaholz, B.P. Structure of the human 80S ribosome. *Nature* **2015**, *520*, 640–645. [CrossRef] [PubMed]
16. Zhang, J.; Fei, Y.; Sun, L.; Zhang, Q.C. Advances and opportunities in RNA structure experimental determination and computational modeling. *Nat. Methods* **2022**, *19*, 1193–1207. [CrossRef]
17. Schneider, B.; Sweeney, B.A.; Bateman, A.; Cerny, J.; Zok, T.; Szachniuk, M. When Will RNA Get Its AlphaFold Moment? *Nucleic Acids Res.* **2023**, *51*, 9522–9532. [CrossRef]
18. Turnbull, A.P.; Wu, X. Studying RNA-Protein Complexes Using X-ray Crystallography. *Methods Mol. Biol.* **2021**, *2263*, 423–446. [CrossRef]
19. Brito Querido, J.; Sokabe, M.; Kraatz, S.; Gordiyenko, Y.; Skehel, J.M.; Fraser, C.S.; Ramakrishnan, V. Structure of a human 48S translational initiation complex. *Science* **2020**, *369*, 1220–1227. [CrossRef]
20. Bothe, J.R.; Nikolova, E.N.; Eichhorn, C.D.; Chugh, J.; Hansen, A.L.; Al-Hashimi, H.M. Characterizing RNA dynamics at atomic resolution using solution-state NMR spectroscopy. *Nat. Methods* **2011**, *8*, 919–931. [CrossRef]
21. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef] [PubMed]
22. Huang, S.Y.; Zou, X. A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. *Nucleic Acids Res.* **2014**, *42*, e55. [CrossRef] [PubMed]
23. Zeng, C.; Jian, Y.; Vosoughi, S.; Zeng, C.; Zhao, Y. Evaluating native-like structures of RNA-protein complexes through the deep learning method. *Nat. Commun.* **2023**, *14*, 1060. [CrossRef]
24. Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. Critical assessment of methods of protein structure prediction (CASP)-Round XV. *Proteins Struct. Funct. Bioinform.* **2023**, *91*, 1539–1549. [CrossRef]

25. Lensink, M.F.; Brysbaert, G.; Raouraoua, N.; Bates, P.A.; Giulini, M.; Honorato, R.V.; van Noort, C.; Teixeira, J.M.C.; Bonvin, A.M.J.J.; Kong, R.; et al. Impact of AlphaFold on structure prediction of protein complexes: The CASP15-CAPRI experiment. *Proteins Struct. Funct. Bioinform.* **2023**, *91*, 1658–1683. [CrossRef]
26. Cruz, J.A.; Blanchet, M.F.; Boniecki, M.; Bujnicki, J.M.; Chen, S.J.; Cao, S.; Das, R.; Ding, F.; Dokholyan, N.V.; Flores, S.C.; et al. A CASP-like evaluation of RNA three-dimensional structure prediction. *RNA* **2012**, *18*, 610–625. [CrossRef]
27. Yan, Y.; Zhang, D.; Zhou, P.; Li, B.; Huang, S.Y. HDOCK: A web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res.* **2017**, *45*, W365–W373. [CrossRef]
28. Tuszynska, I.; Magnus, M.; Jonak, K.; Dawson, W.; Bujnicki, J.M. NPDock: A web server for protein-nucleic acid docking. *Nucleic Acids Res.* **2015**, *43*, W425–W430. [CrossRef]
29. Van Zundert, G.C.P.; Rodrigues, J.; Trellet, M.; Schmitz, C.; Kastritis, P.L.; Karaca, E.; Melquiond, A.S.J.; van Dijk, M.; de Vries, S.J.; Bonvin, A. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **2016**, *428*, 720–725. [CrossRef]
30. Zeng, C.W.; Zhao, Y.J. Advances in RNA-protein structure prediction. *Sci. Sin.-Phys. Mech. Astron.* **2023**, *53*, 290018. [CrossRef]
31. He, J.; Tao, H.; Huang, S.Y. Protein-ensemble-RNA docking by efficient consideration of protein flexibility through homology models. *Bioinformatics* **2019**, *35*, 4994–5002. [CrossRef] [PubMed]
32. Kappel, K.; Das, R. Sampling Native-like Structures of RNA-Protein Complexes through Rosetta Folding and Docking. *Structure* **2019**, *27*, 140–151. [CrossRef] [PubMed]
33. Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H.J. PatchDock and SymmDock: Servers for rigid and symmetric docking. *Nucleic Acids Res.* **2005**, *33*, W363–W367. [CrossRef]
34. Qiu, L.; Zou, X. Scoring Functions for Protein-RNA Complex Structure Prediction: Advances, Applications, and Future Directions. *Commun. Inf. Syst.* **2020**, *20*, 1–22. [CrossRef]
35. Nithin, C.; Ghosh, P.; Bujnicki, J.M. Bioinformatics Tools and Benchmarks for Computational Docking and 3D Structure Prediction of RNA-Protein Complexes. *Genes* **2018**, *9*, 432. [CrossRef] [PubMed]
36. Tuszynska, I.; Bujnicki, J.M. DARS-RNP and QUASI-RNP: New statistical potentials for protein-RNA docking. *BMC Bioinform.* **2011**, *12*, 348. [CrossRef] [PubMed]
37. Li, H.; Huang, Y.; Xiao, Y. A pair-conformation-dependent scoring function for evaluating 3D RNA-protein complex structures. *PLoS ONE* **2017**, *12*, e0174662. [CrossRef]
38. Wang, K.; Jian, Y.; Wang, H.; Zeng, C.; Zhao, Y. RBind: Computational network method to predict RNA binding sites. *Bioinformatics* **2018**, *34*, 3131–3136. [CrossRef]
39. Baek, M.; McHugh, R.; Anishchenko, I.; Jiang, H.; Baker, D.; DiMaio, F. Accurate prediction of protein-nucleic acid complexes using RoseTTAFoldNA. *Nat. Methods* **2024**, *21*, 117–121. [CrossRef]
40. Yao, J.M.; Liang, W.; Zheng, Z.M.; Ouyang, Y.L.; Liao, C.Y. Research on maintenance cycle prediction for energy equipment with limited and sensitive data. *Eng. Fail. Anal.* **2024**, *164*, 108696. [CrossRef]
41. Senior, A.W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Zidek, A.; Nelson, A.W.R.; Bridgland, A.; et al. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706–710. [CrossRef] [PubMed]
42. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef] [PubMed]
43. Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A.J.; Bambrick, J.; et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493–500. [CrossRef] [PubMed]
44. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G.R.; Wang, J.; Cong, Q.; Kinch, L.N.; Schaeffer, R.D.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876. [CrossRef]
45. Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; Baker, D. Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 1496–1503. [CrossRef]
46. Chowdhury, R.; Bouatta, N.; Biswas, S.; Floristean, C.; Kharkare, A.; Roye, K.; Rochereau, C.; Ahdritz, G.; Zhang, J.; Church, G.M.; et al. Single-sequence protein structure prediction using a language model and deep learning. *Nat. Biotechnol.* **2022**, *40*, 1617–1623. [CrossRef]
47. Townshend, R.J.L.; Eismann, S.; Watkins, A.M.; Rangan, R.; Karelina, M.; Das, R.; Dror, R.O. Geometric deep learning of RNA structure. *Science* **2021**, *373*, 1047–1051. [CrossRef] [PubMed]
48. Wang, W.; Feng, C.; Han, R.; Wang, Z.; Ye, L.; Du, Z.; Wei, H.; Zhang, F.; Peng, Z.; Yang, J. trRosettaRNA: Automated prediction of RNA 3D structure with transformer network. *Nat. Commun.* **2023**, *14*, 7266. [CrossRef] [PubMed]
49. Sha, C.M.; Wang, J.; Dokholyan, N.V. Predicting 3D RNA structure from the nucleotide sequence using Euclidean neural networks. *Biophys. J.* **2023**, *17*, 2671–2681. [CrossRef]
50. Li, J.; Zhu, W.; Wang, J.; Li, W.; Gong, S.; Zhang, J.; Wang, W. RNA3DCNN: Local and global quality assessments of RNA 3D structures using 3D deep convolutional neural networks. *PLoS Comput. Biol.* **2018**, *14*, e1006514. [CrossRef]
51. Chen, K.; Zhou, Y.Q.; Wang, S.; Xiong, P. RNA tertiary structure modeling with BRiQ potential in CASP15. *Proteins Struct. Funct. Bioinform.* **2023**, *91*, 1771–1778. [CrossRef]
52. Xu, X.J.; Zhao, P.N.; Chen, S.J. Vfold: A Web Server for RNA Structure and Folding Thermodynamics Prediction. *PLoS ONE* **2014**, *9*, e107504. [CrossRef]

53. Sarzynska, J.; Popenda, M.; Antczak, M.; Szachniuk, M. RNA tertiary structure prediction using RNAComposer in CASP15. *Proteins Struct. Funct. Bioinform.* **2023**, *91*, 1790–1799. [CrossRef] [PubMed]
54. Moafinejad, S.N.; de Aquino, B.R.H.; Boniecki, M.J.; Jeyeram, I.P.N.P.; Nikolaev, G.; Magnus, M.; Farsani, M.A.; Badepally, N.G.; Wirecki, T.K.; Stefaniak, F.; et al. SimRNAweb v2.0: A web server for RNA folding simulations and 3D structure modeling, with optional restraints and enhanced analysis of folding trajectories (May, 10.1093/nar/gkae356, 2024). *Nucleic Acids Res.* **2024**, *52*, W368–W373. [CrossRef] [PubMed]
55. Bryant, P.; Pozzati, G.; Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* **2022**, *13*, 1265. [CrossRef] [PubMed]
56. Soleymani, F.; Paquet, E.; Viktor, H.; Michalowski, W.; Spinello, D. Protein-protein interaction prediction with deep learning: A comprehensive review. *Comput. Struct. Biotechnol.J.* **2022**, *20*, 5316–5341. [CrossRef]
57. Feng, S.H.; Chen, Z.Y.; Zhang, C.W.; Xie, Y.H.; Ovchinnikov, S.; Gao, Y.Q.; Liu, S.R. Integrated structure prediction of protein-protein docking with experimental restraints using ColabDock. *Nat. Mach. Intell.* **2024**, *6*, 924–935. [CrossRef]
58. Jones, S.; Daley, D.T.; Luscombe, N.M.; Berman, H.M.; Thornton, J.M. Protein-RNA interactions: A structural analysis. *Nucleic Acids Res.* **2001**, *29*, 943–954. [CrossRef]
59. Jeong, E.; Kim, H.; Lee, S.W.; Han, K. Discovering the interaction propensities of amino acids and nucleotides from protein-RNA complexes. *Mol. Cells* **2003**, *16*, 161–167. [CrossRef]
60. Kim, O.T.; Yura, K.; Go, N. Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res.* **2006**, *34*, 6450–6460. [CrossRef]
61. Lejeune, D.; Delsaux, N.; Charloteaux, B.; Thomas, A.; Brasseur, R. Protein-nucleic acid recognition: Statistical analysis of atomic interactions and influence of DNA structure. *Proteins Struct. Funct. Bioinform.* **2005**, *61*, 258–271. [CrossRef] [PubMed]
62. Yang, R.; Liu, H.; Yang, L.; Zhou, T.; Li, X.; Zhao, Y. RPpocket: An RNA-Protein Intuitive Database with RNA Pocket Topology Resources. *Int. J. Mol. Sci.* **2022**, *23*, 6903. [CrossRef] [PubMed]
63. Perez-Cano, L.; Solernou, A.; Pons, C.; Fernandez-Recio, J. Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials. In *Biocomputing 2010—Proceedings of the Pacific Symposium*; World Scientific Publishing Company: Singapore, 2010; pp. 293–301. [CrossRef]
64. Bahadur, R.P.; Zacharias, M.; Janin, J. Dissecting protein-RNA recognition sites. *Nucleic Acids Res.* **2008**, *36*, 2705–2716. [CrossRef]
65. Iwakiri, J.; Tateishi, H.; Chakraborty, A.; Patil, P.; Kenmochi, N. Dissecting the protein-RNA interface: The role of protein surface shapes and RNA secondary structures in protein-RNA recognition. *Nucleic Acids Res.* **2012**, *40*, 3299–3306. [CrossRef]
66. Berman, H.M.; Olson, W.K.; Beveridge, D.L.; Westbrook, J.; Gelbin, A.; Demeny, T.; Hsieh, S.H.; Srinivasan, A.R.; Schneider, B. The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.* **1992**, *63*, 751–759. [CrossRef]
67. Coimbatore Narayanan, B.; Westbrook, J.; Ghosh, S.; Petrov, A.I.; Sweeney, B.; Zirbel, C.L.; Leontis, N.B.; Berman, H.M. The Nucleic Acid Database: New features and capabilities. *Nucleic Acids Res.* **2014**, *42*, D114–D122. [CrossRef]
68. Perez-Cano, L.; Fernandez-Recio, J. Optimal protein-RNA area, OPRA: A propensity-based method to identify RNA-binding sites on proteins. *Proteins Struct. Funct. Bioinform.* **2010**, *78*, 25–35. [CrossRef]
69. Huang, Y.; Liu, S.; Guo, D.; Li, L.; Xiao, Y. A novel protocol for three-dimensional structure prediction of RNA-protein complexes. *Sci. Rep.* **2013**, *3*, 1887. [CrossRef]
70. Huang, Y.; Li, H.; Xiao, Y. Using 3dRPC for RNA-protein complex structure prediction. *Biophys. Rep.* **2016**, *2*, 95–99. [CrossRef] [PubMed]
71. Setny, P.; Zacharias, M. A coarse-grained force field for Protein-RNA docking. *Nucleic Acids Res.* **2011**, *39*, 9118–9129. [CrossRef] [PubMed]
72. Li, C.H.; Cao, L.B.; Su, J.G.; Yang, Y.X.; Wang, C.X. A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys. *Proteins Struct. Funct. Bioinform.* **2012**, *80*, 14–24. [CrossRef]
73. Malolepsza, E.; Boniecki, M.; Kolinski, A.; Piela, L. Theoretical model of prion propagation: A misfolded protein induces misfolding. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 7835–7840. [CrossRef]
74. Leontis, N.B.; Westhof, E. Geometric nomenclature and classification of RNA base pairs. *RNA* **2001**, *7*, 499–512. [CrossRef]
75. Vakser, I.A.; Aflalo, C. Hydrophobic docking: A proposed enhancement to molecular recognition techniques. *Proteins Struct. Funct. Bioinform.* **1994**, *20*, 320–329. [CrossRef]
76. Zheng, S.; Robertson, T.A.; Varani, G. A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins. *FEBS J.* **2007**, *274*, 6378–6391. [CrossRef]
77. Chuang, G.Y.; Kozakov, D.; Brenke, R.; Comeau, S.R.; Vajda, S. DARS (Decoys As the Reference State) Potentials for Protein-Protein Docking. *Biophys. J.* **2008**, *95*, 4217–4227. [CrossRef]
78. Huang, S.Y.; Zou, X. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. *J. Comput. Chem.* **2006**, *27*, 1866–1875. [CrossRef]
79. Zhuo, C.; Zeng, C.W.; Yang, R.; Liu, H.Q.; Zhao, Y.J. RPflex: A Coarse-Grained Network Model for RNA Pocket Flexibility Study. *Int. J. Mol. Sci.* **2023**, *24*, 5497. [CrossRef]
80. Liu, H.; Jian, Y.; Hou, J.; Zeng, C.; Zhao, Y. RNet: A network strategy to predict RNA binding preferences. *Brief. Bioinform.* **2023**, *25*, bbad482. [CrossRef] [PubMed]

81. Wang, H.; Liu, H.; Ning, S.; Zeng, C.; Zhao, Y. DLSSAffinity: Protein-ligand binding affinity prediction via a deep learning model. *Phys. Chem. Chem. Phys.* **2022**, *24*, 10124–10133. [CrossRef] [PubMed]
82. He, X.; Zhao, L.; Tian, Y.; Li, R.; Chu, Q.; Gu, Z.; Zheng, M.; Wang, Y.; Li, S.; Jiang, H.; et al. Highly accurate carbohydrate-binding site prediction with DeepGlycanSite. *Nat. Commun.* **2024**, *15*, 5163. [CrossRef] [PubMed]
83. Zheng, W.; Wuyun, Q.; Li, Y.; Zhang, C.; Freddolino, P.L.; Zhang, Y. Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. *Nat. Methods* **2024**, *21*, 279–289. [CrossRef] [PubMed]
84. Qiao, Z.R.; Nie, W.L.; Vahdat, A.; Miller, T.F.; Anandkumar, A. State-specific protein-ligand complex structure prediction with a multiscale deep generative model. *Nat. Mach. Intell.* **2024**, *6*, 195–208. [CrossRef]
85. Liu, H.Q.; Zhao, Y.J. Integrated modeling of protein and RNA. *Brief. Bioinform.* **2024**, *25*, bbae139. [CrossRef] [PubMed]
86. Liu, H.Q.; Gong, Z.; Zhao, Y.J. Methods and Applications in Proteins and RNAs. *Life* **2023**, *13*, 672. [CrossRef]
87. Lotthammer, J.M.; Ginell, G.M.; Griffith, D.; Emenecker, R.; Holehouse, A.S. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Biophys. J.* **2024**, *123*, 43a. [CrossRef]
88. He, J.; Lin, P.; Chen, J.; Cao, H.; Huang, S.Y. Model building of protein complexes from intermediate-resolution cryo-EM maps with deep learning-guided automatic assembly. *Nat. Commun.* **2022**, *13*, 4066. [CrossRef]
89. Parisien, M.; Wang, X.; Perdrizet, G., 2nd; Lamphear, C.; Fierke, C.A.; Maheshwari, K.C.; Wilde, M.J.; Sosnick, T.R.; Pan, T. Discovering RNA-protein interactome by using chemical context profiling of the RNA-protein interface. *Cell Rep.* **2013**, *3*, 1703–1713. [CrossRef]
90. Romero, E.; Sopena, J.M. Performing feature selection with multilayer perceptrons. *IEEE Trans. Neural Netw.* **2008**, *19*, 431–441. [CrossRef]
91. Zeng, C.; Jian, Y.; Zhuo, C.; Li, A.; Zeng, C.; Zhao, Y. Evaluation of DNA-protein complex structures using the deep learning method. *Phys. Chem. Chem. Phys.* **2023**, *26*, 130–143. [CrossRef]
92. Barik, A.; C, N.; P, M.; Bahadur, R.P. A protein-RNA docking benchmark (I): Nonredundant cases. *Proteins Struct. Funct. Bioinform.* **2012**, *80*, 1866–1871. [CrossRef] [PubMed]
93. Perez-Cano, L.; Jimenez-Garcia, B.; Fernandez-Recio, J. A protein-RNA docking benchmark (II): Extended set from experimental and homology modeling data. *Proteins Struct. Funct. Bioinform.* **2012**, *80*, 1872–1882. [CrossRef]
94. Nithin, C.; Mukherjee, S.; Bahadur, R.P. A non-redundant protein-RNA docking benchmark version 2.0. *Proteins Struct. Funct. Bioinform.* **2017**, *85*, 256–267. [CrossRef]
95. Huang, S.Y.; Zou, X. A nonredundant structure dataset for benchmarking protein-RNA computational docking. *J. Comput. Chem.* **2013**, *34*, 311–318. [CrossRef]
96. Iwakiri, J.; Kameda, T.; Asai, K.; Hamada, M. Analysis of base-pairing probabilities of RNA molecules involved in protein-RNA interactions. *Bioinformatics* **2013**, *29*, 2524–2528. [CrossRef] [PubMed]
97. Barik, A.; C, N.; Pilla, S.P.; Bahadur, R.P. Molecular architecture of protein-RNA recognition sites. *J. Biomol. Struct. Dyn.* **2015**, *33*, 2738–2751. [CrossRef]
98. Janin, J.; Henrick, K.; Moult, J.; Eyck, L.T.; Sternberg, M.J.; Vajda, S.; Vakser, I.; Wodak, S.J. CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins Struct. Funct. Bioinform.* **2003**, *52*, 2–9. [CrossRef]
99. Mendez, R.; Leplae, R.; Lensink, M.F.; Wodak, S.J. Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. *Proteins Struct. Funct. Bioinform.* **2005**, *60*, 150–169. [CrossRef]
100. Nithin, C.; Kmiecik, S.; Blaszczyk, R.; Nowicka, J.; Tuszynska, I. Comparative analysis of RNA 3D structure prediction methods: Towards enhanced modeling of RNA-ligand interactions. *Nucleic Acids Res.* **2024**, *52*, 7465–7486. [CrossRef] [PubMed]
101. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef] [PubMed]
102. Li, W.; Jaroszewski, L.; Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **2001**, *17*, 282–283. [CrossRef] [PubMed]
103. Li, W.; Jaroszewski, L.; Godzik, A. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* **2002**, *18*, 77–82. [CrossRef]
104. Zhu, H.R.; Yang, Y.N.; Wang, Y.H.; Wang, F.Z.; Huang, Y.J.; Chang, Y.; Wong, K.C.; Li, X.T. Dynamic characterization and interpretation for protein-RNA interactions across diverse cellular conditions using HDRNet. *Nat. Commun.* **2023**, *14*, 6824. [CrossRef] [PubMed]
105. Li, T.; He, J.; Cao, H.; Zhang, Y.; Chen, J.; Xiao, Y.; Huang, S.Y. All-atom RNA structure determination from cryo-EM maps. *Nat. Biotechnol.* **2024**, 1–9. [CrossRef]
106. He, J.; Li, T.; Huang, S.Y. Improvement of cryo-EM maps by simultaneous local and non-local deep learning. *Nat. Commun.* **2023**, *14*, 3217. [CrossRef]
107. Song, X.T.; Bao, L.; Feng, C.J.; Huang, Q.; Zhang, F.; Gao, X.; Han, R.M. Accurate Prediction of Protein Structural Flexibility by Deep Learning Integrating Intricate Atomic Structures and Cryo-EM Density Information. *Nat. Commun.* **2024**, *15*, 5538. [CrossRef]