UNIVERSITÀ DEGLI STUDI DI MILANO

UNIVERSITÀ DEGLI STUDI DI MILANO

Doctoral school in pharmaceutical sciences - XXXIX Cycle

Department of Pharmaceutical Sciences

Thesis for the Degree of Doctor of Philosophy in

Development of Virtual Screening Methodologies for Identifying RNA-based Therapies

Parisa Aletayeb
Matricola: R14167

Tutor: Prof. Alessandro Pedretti

Co-tutor: Prof. Giulio Vistoli

Academic Year 2025/2026

Dedicated to my family

ACKNOWLEDGEMENTS

I would like to express my deepest and most sincere gratitude to my supervisor, **Prof. Alessandro Pedretti,** for giving me the opportunity to pursue my doctoral studies abroad. His trust and support marked a decisive turning point in my academic and personal journey. I am profoundly thankful, both for his scientific expertise and for his genuine care as a mentor and as a human being.

I am also grateful to my co-supervisor, **Prof. Giulio Vistoli**, for his valuable guidance, insightful discussions, and continuous support in addressing scientific challenges during my research.

My heartfelt thanks go to **Dr. Akash Deep Biswas** for his mentorship and for showing me the path of growth and progress. His support has extended far beyond scientific guidance; he has been a trusted advisor, a problem solver, and a true friend. I am deeply grateful for his constant presence and encouragement, at every step of this journey.

I am thankful to **Prof. Matteo Dal Peraro**, my supervisor during the PhD visiting period at EPFL, for giving me the opportunity to explore science from a different perspective and for his support through the scholarship that enabled this visit.

I owe my deepest gratitude to my family, to **my parents**, whose sacrifices, courage, and purest form of unconditional love have shaped who I am and given true meaning to my life; their unwavering faith in me has been my greatest strength throughout this journey. To my siblings, **Bahar, Davoud, and Arash**, whose constant presence, encouragement, and boundless support have been a source of resilience and motivation, without them, this path would not have been possible, nor as meaningful.

I would also like to thank my **Italian family**, Loredana, Marco, and Gaia, whose warmth, kindness, and open arms made me feel at home far from home. I am deeply grateful to **Loredana**, from whom I learned that kindness could transcend borders, languages, and cultures.

A special thank you goes to **Mahbubeh**, my comfort zone of entire life, whose presence, understanding, and support have been invaluable throughout this time.

I would like to thank **Angelica Mazzolari** for her kindness and for always being a supportive friend and colleague, regardless of the situation. I am thankful to **Leen Zaheer** for always being there when I needed her and for being all ears for me through both my good and difficult days.

Finally, I am thankful to my colleagues in the Pharmaceutical Science department of University of Milan, especially to **Serena Vittor**io and **Stefano Rocca** for being generous to help me.

ABSTRACT

This thesis focuses on the development of computational methodologies for predicting protein-RNA binding affinities and designing RNA sequences with specific functional outcomes. The research introduces PANTHER (Protein-Affinity for Nucleic Target-binding, Hybridization, and Energy Regression), a novel machine learning-based scoring function that predicts protein-RNA binding affinities through a local-to-global approach. The methodology integrates molecular dynamics simulations and machine learning models to estimate local interaction energies between amino acids and nucleotides, which are then aggregated to predict global binding affinities. A dedicated web service has also been developed to provide the public access of PANTHER score (https://nova.disfarm.unimi.it/panther/). In addition, the thesis extends the RISoTTo (RIbonucleic acid Sequence design from TerTiary structure) framework to handle conformationally flexible RNA molecules, enabling context-aware RNA sequence design. The integrated computational framework presented in this work aims to advance the understanding and practical prediction of protein-RNA interactions, supporting the development of RNA-targeted therapeutics and enhancing the field of computational structural biology.

Keywords: *protein-RNA interactions, binding affinity prediction, machine learning, molecular dynamics simulations, Random Forest regression, local-to-global approach, PANTHER score, web service, RNA sequence design, geometric transformers, conformational flexibility, RNA therapeutics*

Index

Chapter 1

Introduction

1.1 The Biological Significance of Protein-RNA Interactions

Protein-RNA interactions constitute one of the most fundamental molecular recognition processes in cellular biology, underlying a vast array of essential functions including regulation of gene expression, mRNA splicing, post-transcriptional modifications, RNA transport, and stability [1], [2]. These interactions are involved not only in fine-tuning steady-state levels of transcripts but also in responding to environmental signals, developmental cues, and stress by modulating RNA behavior at multiple levels [3]. The quantitative measure of protein-RNA interactions, namely binding affinity, serves as a critical determinant of biological function. High-affinity interactions often correlate with more robust gene regulation, greater stability of RNA molecules, and tighter control of post-transcriptional processes [4]. For example, engineered RNA-binding domains with enhanced affinity have been shown to increase regulatory impact without sacrificing specificity[5]. Studies of RNA-binding protein function also emphasize that biological outcomes, such as mRNA splicing, localization, translation, and decay are shaped not merely by the presence of binding RNA-protein, but by how strongly the binding occurs [6]. The biological importance of protein-RNA binding affinities extends beyond basic cellular processes to disease mechanisms and therapeutic interventions. Dysregulation of RNA-binding proteins (RBPs) has been increasingly recognized as a key factor in the onset and progression of numerous diseases, including neurodegeneration, cancer, and developmental disorders [7]. This recognition has positioned protein-RNA interactions as attractive therapeutic targets, particularly in the emerging field of RNA-targeted drug discovery [8],[9],[10]. Hence, investigating the affinity of the protein-nucleic acid complexes have a broad spectrum of applications, such as designing complexes with desired affinities, predictive methods for the target sites, and the quantitative simulation of gene regulation networks.

1.2 Protein-RNA Interactions in Drug Discovery

Understanding the binding affinity between proteins and RNA has become a cornerstone of modern drug discovery. While only a small fraction of the human genome encodes proteins, more than 70% is transcribed into diverse non-coding RNAs [9]. If these RNAs can be selectively targeted,

they represent a vast expansion of the druggable space, opening therapeutic opportunities well beyond traditional protein-centric approaches. RNA therapeutics provide versatile strategies to modulate protein-driven diseases, particularly where conventional modalities such as small molecules or antibodies are limited. Unlike protein-targeted drugs, RNA molecules can be engineered to recognize and engage biological systems through multiple mechanisms [10]. Small interfering RNAs (siRNAs) utilize the endogenous RNA-induced silencing complex (RISC) to degrade complementary mRNAs, thereby preventing the production of pathogenic proteins [11]. Their clinical success is demonstrated by several FDA-approved therapeutics, including patisiran [12], givosiran [13], lumasiran [14], inclisiran [15], [16], vutrisiran [17], and nedosiran [18]. Notably, many of these siRNAs preferentially target untranslated regions rather than coding sequences, reflecting the precision achievable in modulating protein–RNA interactions [19]. Antisense oligonucleotides (ASOs) act through complementary binding to RNA, where they recruit proteins such as RNase H or spliceosome components to induce transcript degradation or modulate splicing [20]. Clinically approved examples include nusinersen, which modulates SMN2 splicing for spinal muscular atrophy [21], eteplirsen, which promotes exon 51 skipping in the dystrophin transcript for Duchenne muscular dystrophy[22], and inotersen, which triggers RNase H–mediated degradation of transthyretin mRNA in hereditary transthyretin amyloidosis [23].

MicroRNAs further illustrate the complexity of RNA-based therapeutics, as they regulate gene expression post-transcriptionally by forming transient complexes with Argonaut proteins in RISC. A single microRNA can fine-tune the expression of hundreds of target transcripts, highlighting the breadth of protein–RNA regulatory networks [24]. RNA aptamers are single-stranded oligonucleotides that fold into three-dimensional structures capable of binding proteins with high affinity and specificity, often described as “chemical antibodies” [25]. They are identified through the SELEX process, an iterative in vitro selection strategy [25]. To date, two aptamer therapeutics have been approved by the FDA, such as pegaptanib [26], which targets VEGF165 in neovascular age-related macular degeneration [26], and avacincaptad pegol [27], which inhibits complement protein C5 in geographic atrophy secondary to AMD. Aptamer-protein complexes frequently achieve dissociation constants in the pico- to nanomolar range [28], and their thermodynamic stability can be described by binding free energy relationships [29]. The spectrum of RNA therapeutics extends further: ribozymes capable of catalyzing RNA cleavage[30], long non-coding RNAs that serve as scaffolds for protein recruitment [31], CRISPR guide RNAs that direct Cas proteins for programmable genome editing [32], and circular RNAs that act as molecular sponges for

microRNAs or RNA-binding proteins [33]. The programmable nature of RNA allows the creation of multifunctional therapeutic constructs that act simultaneously as targeting ligands and therapeutic payloads [34]. Aptamer-siRNA chimeras are a striking example, combining selective protein recognition with gene-silencing activity and illustrating how precise binding affinities can be harnessed in complex disease contexts [35]. By exploiting these properties, RNA therapeutics are being developed for diseases rooted in dysregulated protein-RNA interactions, including cancer, neurodegeneration, viral infections, and genetic disorders [36]. Through a precise understanding of protein-RNA binding affinities, integrated RNA systems can distinguish disease-associated proteins from off-targets and deliver therapeutic RNAs that disrupt pathological interactions directly to affected cells. However, the accurate quantification of protein–RNA binding affinities ultimately relies on experimental methods such as electrophoretic mobility shift assays (EMSA), surface plasmon resonance(SPR), and isothermal titration calorimetry (ITC), which remain essential for validating specificity and guiding the development of therapeutic protein-RNA complexes [37].[38], [39], [40],[41], [42], [43], [44].

1.3 Experimental Challenges in Protein-RNA Binding Affinity Determination

The experimental determination of protein-RNA binding affinities relies on several sophisticated biophysical techniques, each with distinct advantages and limitations. Traditional methods include fluorescence spectroscopy [38] surface plasmon resonance [39] electrophoretic mobility shift assays [9,10], isothermal titration calorimetry [40] and filter binding assays [41]. While these techniques provide accurate and reliable measurements, they are inherently time-consuming, labor-intensive, and costly procedures that require specialized equipment and expertise. [45][46][47][48]. More importantly, these experimental approaches face fundamental challenges when applied to protein-RNA systems. As RNA molecules exhibit significant conformational flexibility, making structural characterization difficult [42], the dynamic nature of protein-RNA interactions involves induced-fit mechanisms and allosteric effects, which further complicates experimental analysis [43]. Additionally, many protein-RNA complexes are transient or context-dependent, requiring specific cellular conditions that are difficult to replicate in vitro [44] For instance, RNA chaperones interact only fleetingly with their RNA substrates to promote proper folding without forming stable complexes [45]. Similarly, the assembly of nascent ribonucleoproteins during transcription involves a cascade of transient protein-RNA contacts that

are replaced as the RNA matures [46]. Context-specific assemblies such as stress granules further demonstrate this principle, as they form through dynamic protein-RNA interactions under stress conditions and rapidly dissolve once normal cellular states are restored [47], as mentioned in previously, regulatory RNAs also exemplify this dynamic behavior, microRNAs form transient complexes with Argonaut proteins within the RNA-induced silencing complex (RISC), and their activity depends on the presence of specific target transcripts and cellular conditions [48] The scalability limitations of experimental methods become particularly apparent in the context of modern drug discovery efforts, which increasingly focus on larger biomolecules such as RNA aptamers and riboswitches [49]These constraints have created an urgent need for reliable computational alternatives that can complement and, in some cases, replace experimental measurements.

1.4 Evolution of Computational Approaches

1.4.1 Early Sequence-Based Methods

The computational prediction of protein-RNA binding affinities has undergone significant evolution over the past two decades. Early approaches primarily relied on sequence-based methods that analyzed amino acid and nucleotide compositions, evolutionary conservation patterns, and basic physicochemical properties for binding site predictions. During mid-2006, BindN [50] employed amino acid properties such as charge and polarity, for predicting the binding site. Right after a year in mid-2007 RNABindR [51] was introduced which uses sequence conservation and composition to identify RNA-binding residues and identifies its binding sites which are critical to bind with a protein. [57] Right after these pieces of software, in the early-2008 PPRInt [52] used evolutionary profiles and support vector machines for prediction in this area of research. Although computationally efficient, these methods were limited by their inability to capture the three-dimensional aspects of molecular recognition that are crucial for accurate binding affinity prediction. The limitations of sequence-only approaches became increasingly apparent as structural biology advanced, and more protein-RNA complex structures became available through X-ray crystallography and NMR spectroscopy. The recognition that “structure matters” in protein-RNA interaction prediction marked a paradigm shift toward incorporating three-dimensional structural information into computational frameworks [53].

1.4.2 Structure-Based Methods

The incorporation of three-dimensional structural information into computational models marked a major advance over sequence-only methods for predicting protein-RNA binding affinity. Early structure-based approaches relied on docking simulations coupled with statistical scoring functions to evaluate protein-RNA interfaces. Among the pioneering efforts, in 2003 Prof. Bonvin and his team introduced HADDOCK [54], in this area and later extended its data-driven docking framework to protein-RNA complexes, integrating experimental restraints with energetic terms such as electrostatics, van der Waals, and desolvation energies. Similarly, Pérez-Cano, Solernou, Pons & Fernández-Recio [55] in 2010, developed a propensity-based statistical potential for protein-RNA docking, capturing residue-nucleotide interaction propensities at the interface to score rigid-body docking poses, which improved prediction accuracy. Other structural studies emphasized the importance of solvent accessibility and conservation of interaction surfaces in defining protein-RNA [56]. For a long time, docking and statistical approaches marked as a significant advancement. Clearly improved upon sequence-based predictors by incorporating structural determinants of binding. However, their utility remained constrained by the limited number of available experimental structures and by their reliance on static models that could not capture the flexibility of protein-RNA interfaces. As these limitations became increasingly apparent, the field gradually entered a new phase: the emergence of machine learning methods, which offered the potential to overcome these constraints and achieve further improvements in predictive accuracy.

1.4.3 Machine Learning and Deep Learning Methods

Although support vector machines and other statistical classifiers were already employed in some of the early sequence-based predictors [57], these models largely relied on limited handcrafted features and were constrained by data availability. The subsequent era marked a broader expansion of machine learning approaches, in which data-driven algorithms such as Random Forests and Gradient Boosted Trees were applied to larger and more diverse feature sets derived from both sequence and structure [58], [59], [60]. This shift represented a genuine methodological advance, moving from shallow classifiers applied to narrow feature representations toward more systematic and scalable ML frameworks for affinity prediction, which marked another significant milestone in the field's evolution. Building on these developments, deep learning methods have recently become highly influential in protein-RNA interaction prediction. While machine learning models such as Random Forests, GBoost and XGBoost offer important advantages for biological application**s**, including greater interpretability through feature importance analysis,

robustness on relatively small data sets, and reduced risk of overfitting, on the other hand deep learning approaches gained popularity due to their ability to efficiently process large-scale data and capture complex, nonlinear dependencies. Convolutional Neural Networks (CNNs) have been applied to detect local sequence or structural motifs relevant for binding [61], and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) units have been employed to model long-range dependencies in protein and RNA sequences [62]. Building on these advances in sequence and structure modeling, recent efforts have turned toward predicting binding affinities, where the goal is not only to recognize interactions but also to quantify their strength. This shift marks an important step toward developing more accurate and practical models of protein–RNA recognition through machine learning approaches.

1.5 State-of-the-Art in Binding Affinity Prediction

1.5.1 Advances in Machine Learning Approaches

Early ML approaches, such as PredPRBA introduced by Deng et al. in 2019 employed gradient boosted regression trees with diverse sequence and structural features, demonstrating the potential of data-driven methods to capture complex, non-linear relationships in protein-RNA binding that were difficult to model using traditional physics-based approaches. The PredPRBA methodology further incorporated different predictive models depending on the interacting RNA structure, recognizing that binding affinity prediction might require system-specific strategies. Further, Nithin et al**.** in 2019 developed a structure-informed ML model that integrated interface parameters such as hydrophobicity, contact surface, and hydration patterns, highlighting the value of combining geometric and physicochemical descriptors with statistical learning for affinity estimation [63]. These efforts underscored both the promise and the challenges of ML-based methods: while they achieved reasonable correlations with experimental data on training sets, their performance on diverse, independent test cases often revealed limitations in generalizability. Recently in early-2024 the development of PRA-Pred represented another significant advancement, utilizing features such as base parameters, interaction energies, number of contacts, and hydrogen bonds to predict ΔG values through a web-based interface[64]. While more accessible than earlier models, PRA-Pred's performance on diverse datasets again reflected the ongoing challenges in developing robust, generalizable prediction methods.

1.5.2 Advances in Deep Learning Approaches

1.5.2.1 Transformer Architectures and Language Models

Recent advances in deep learning have begun to reshape protein–RNA binding affinity prediction. The most notable development is CoPRA [65] which integrates pretrained protein language models (ESM-2) and RNA language models (RNA-FM) with structural interface features. CoPRA employs a transformer-based architecture (Co-Former) to fuse multimodal representations and was trained on a large, curated dataset (PRA310) of protein-RNA complexes. This approach achieved state-of-the-art performance in binding affinity prediction and mutation effect analysis, highlighting the promise of large-scale pretrained models and cross-modal fusion for modeling protein-RNA interactions. Nevertheless, the scarcity of high-quality protein-RNA binding affinity datasets remains a major bottleneck, often limiting the generalizability of machine and deep learning models. This challenge underscores the need for curated experimental data and innovative strategies such as data augmentation or transfer learning.

1.6 Limitations and Research Gaps

1.6.1 Data Scarcity and Quality Issues

One of the most significant challenges facing AI and ML approaches in protein-RNA interaction prediction is the limited availability of high-quality training data. The Protein Data Bank (PDB) contains relatively few protein-RNA complex structures compared to protein-only or protein-small molecule structures [66]. This scarcity of structural data limits the development and validation of machine learning models. Furthermore, experimental binding affinity data for protein-RNA complexes is even more limited. While databases like PDBbind [67], PRBAB v2 [68], and ProNAB [69], have made valuable contributions, the total number of protein-RNA complexes with experimentally determined ΔG values remains insufficient for training robust machine learning models without careful data augmentation strategies.

1.6.2 Structural Flexibility and Dynamics

RNA molecules exhibit markedly higher conformational flexibility than proteins, existing in dynamic equilibrium among multiple secondary and tertiary structures. This intrinsic flexibility complicates structure-based prediction methods that typically rely on a single static structure. Studies have shown that unbound RNA samples heterogeneous ensembles and that their conformational dynamics are directly linked to binding thermodynamics and recognition specificity [70]. Riboswitches provide a clear illustration of this principle, their ability to adopt alternative

structural states in response to small-molecule ligands underlies precise regulatory outcomes, and the energetic cost of switching between conformations is central to their binding behavior [71]. Similarly, in protein–RNA complexes, affinity is not determined solely by the static interface but also by the conformational adjustments of RNA upon binding, with mechanisms such as conformational selection and induced fit playing key roles [72].

As a result, accurate affinity prediction requires consideration of RNA dynamics and conformational ensembles rather than static structural snapshots, while most existing models for predicting protein-RNA affinity do not account for molecular dynamics [4] [73]

1.6.3 Generalization and Accuracy Challenges

A persistent challenge in computational binding affinity prediction is developing methods that generalize well across different protein families, RNA types, and biological systems. Many existing models achieve strong results on their training datasets but show substantial performance degradation when applied to structurally or evolutionarily distant complexes. This issue is well documented in blind tests of Protein-RNA binding affinity prediction, where methods that perform well on known complexes underperform on novel ones that differ in sequence or structure [4]. Reviews of the field also emphasize that the limited diversity of high-quality, experimentally annotated protein-RNA complexes constrains model [74], [75]. Moreover, increasing model complexity (e.g., adding many structural descriptors or complex features) often improves in-dataset performance but raises risks of overfitting, interpretability loss, and computational burden-factors that can reduce applicability in practical settings [76]. Moreover, experimental measurement of protein–RNA binding affinity is subject to systematic variability introduced by assay conditions such as temperature, pH, ionic strength, buffer composition, and the specific biophysical technique employed. These methodological factors can cause deviations in absolute values across laboratories, making direct numerical comparison of affinity measurements difficult [63], [77]. Consequently, training machine learning models to predict exact experimental values is inherently limited, since the reference data themselves are assay dependent.

1.7 Addressing Current Limitations: Contribution of the Thesis

1.7.1 Overcoming Data Scarcity by The Local-to-Global Strategy

The above-mentioned challenges, compounded by the scarcity of reliable experimental data, highlight the need for alternative strategies that move beyond direct prediction of binding affinities. Therefore, the development of Protein-Affinity for Nucleic Target-binding, Hybridization, and

Energy Regression (PANTHER) scoring function was motivated by the recognition that a fundamentally different approach was needed to overcome the critical limitations in protein-RNA binding affinity prediction. Rather than attempting to directly predict global binding affinities from limited experimental data, we implemented a local-to-global strategy that focuses on learning and predicting local interaction energies between individual amino acids and nucleotides. This approach addresses the scarcity of experimental ΔG data by decomposing each protein-RNA complex into numerous local pairwise interactions, effectively transforming the data scarcity problem into a data abundance opportunity. Instead of requiring extensive experimental binding affinity datasets, the local-to-global methodology relies on locally decomposed pairwise interaction energies between amino acids and nucleotide bases derived from molecular dynamics simulations. The PANTHER methodology represents an innovative integration of physics-based simulations with machine learning techniques. By using molecular dynamics simulations to generate local interaction energies and then training machine learning models to predict these energies from simple structural features, the approach combines the accuracy of physics-based methods with the efficiency and scalability of machine learning. This hybrid strategy addresses key limitations of both approaches: physics-based methods, while accurate, are computationally expensive, whereas machine learning methods, while efficient, often lack the physical grounding necessary for accurate prediction and interpretation.

1.7.2 Overcoming Structural Flexibility by Molecular Dynamics Simulations

The inherent flexibility of RNA molecules presents a significant challenge for structure-based prediction methods that typically rely on static crystal structures. To address this limitation, we employed molecular dynamics simulations to capture the dynamic behavior of protein-RNA interactions and account for conformational flexibility. Our approach utilized extensive MD simulations to generate time-averaged interaction energies that represent persistent binding modes while filtering out transient, non-specific interactions. By implementing a temporal sampling strategy that focuses on thermodynamically significant interactions, we ensured that machine learning models are trained on biologically relevant binding patterns rather than random fluctuations.

1.7.3 Overcoming Generalizability by Evaluating Models with Stress Set and Binding Affinity Scores vs. Absolute Binding Energies

1.7.3.1 Stress Set

To rigorously assess the generalizability and real-world applicability of our approach, we implemented a comprehensive evaluation strategy using an independent stress set specifically designed to challenge our models under realistic conditions. The stress set consisted of protein-RNA complexes that were deliberately left untreated, without structural curation or MD simulations, to evaluate the method's performance on raw structural data as would be encountered in practical applications. This evaluation approach provided a stringent test of model transferability and robustness across diverse protein-RNA systems. Comparison with existing methods on this challenging dataset demonstrated the superior performance and generalizability of our local-to-global approach, confirming its potential as a reliable tool for practical applications in structural biology and drug discovery research.

1.7.3.2 Binding Affinity Scores

A more pragmatic strategy is to develop binding affinity scoring functions that focus on reproducing relative rankings or comparative strengths of interactions. Such orderings tend to be more reproducible across methods and conditions than absolute magnitudes [78], [79]. By prioritizing correlation over absolute agreement, scoring functions acknowledge the constraints of experimental data while still providing biologically useful insights for ranking interactions and guiding experimental design. Such ranking-oriented approaches therefore provide a practical foundation for extending computational models toward broader applications in RNA biology.

While accurate estimation of protein–RNA binding affinity is important for understanding molecular stability and the energetics of complex formation, it represents only one dimension of protein-RNA recognition. Equally critical is the ability to predict which RNA sequences are preferentially bound by a given protein under physiological conditions. As already known proteins interact with RNA in a selective manner, exhibiting preferences for particular sequence or structural motifs that determine the specificity of binding and ultimately regulate gene expression, splicing, and RNA metabolism [80].Thus, complementing affinity prediction with sequence preference prediction provides a complete picture of protein-RNA interactions, the former quantifies the strength of association, whereas the latter identifies the potential RNA partner. Integrating these perspectives is essential to advance both mechanistic understanding and therapeutic applications, since effective modulation

of protein-RNA complexes requires knowledge of not only how strongly proteins bind RNA but also which RNA sequences they are likely to target.

1.8 RNA Sequence Prediction

Building on this foundation, attention has increasingly shifted from predicting interactions to the more ambitious task of designing RNA molecules with specific functional outcomes. Therefore, while we accurately predict binding affinity which represents as one of the critical aspects of understanding protein-RNA interactions, the inverse problem, designing RNA sequences that achieve desired interactions, presents equally formidable challenges. Traditionally, computational drug discovery has emphasized the development of small-molecule compounds and protein-based therapeutics, often targeting symptoms or late-stage manifestations of disease. More recently, however, there has been a surge of interest in designing RNA-based medicines capable of intervening earlier in disease pathways, thereby disrupting the transmission of pathogenic information within cells [81], [82]. Prominent examples already reshaping biotechnology include mRNA vaccines and CRISPR-driven genome editing approaches [83]. Within the context of structure-guided RNA design, particular attention has been directed toward ribozymes and riboswitches located in the untranslated regions of mRNAs [84], [85]. RNA molecules play central roles in nearly every aspect of cellular biology. The diverse functional group of RNAs is enabled by their structural versatility: a single RNA sequence can fold into multiple alternative conformations, allowing dynamic switching between functional states [70]. These properties make RNA an exceptionally powerful biomolecule, but also a challenging target for computational modeling and therapeutic design. In this context, RNA sequence prediction has emerged as a critical research frontier. Predicting or designing sequences that adopt desired structures and interactions under physiological conditions has broad implications for understanding RNA biology and for developing novel therapeutics. From a biomedical perspective, the ability to design RNA molecules with predictable structural and functional outcomes opens the door to new generations of RNA-based therapies. This is especially relevant given recent successes such as mRNA vaccines, which exploit synthetic messenger RNAs to direct the transient production of viral antigens and thereby elicit protective immune responses [86], and CRISPR–Cas systems, which use programmable guide RNAs to target specific nucleic acid sequences for editing, regulation, or degradation of genetic

material [87], [88]. These advances highlight the transformative potential of RNA as both a target and a therapeutic agent [89], [90] within a certain limitation.

1.8.1 Current Limitations in RNA Sequence Design

1.8.1.1 Secondary Structure-Based Design

Initial computational efforts in RNA design concentrated on secondary structure, a problem formally known to be NP-hard [91]. A landmark study by Hofacker et al. (1994) introduced a local search strategy that optimized candidate sequences using thermodynamic energy functions, laying the foundation for systematic approaches to RNA inverse folding [92]. Building on this foundation, many subsequent methods relied on energy minimization guided by experimentally derived thermodynamic parameters [93], [94]. Parallel to deterministic approaches, probabilistic models of sequence distributions were developed as alternative strategies [95], [96], [97]. In addition, heuristic optimization techniques, including genetic algorithms [98], constraint programming [99], and ant colony optimization [100], were applied to improve search efficiency and sequence diversity. Together, these approaches established the computational principles of RNA sequence design within the secondary structure framework. Recent years have seen the integration of machine learning into RNA inverse folding. Notable examples include LEARNA and Meta-LEARNA, which employ reinforcement learning to iteratively generate RNA sequences predicted to adopt specified secondary structures [101]. These data-driven models highlight the growing potential of machine learning to replace handcrafted optimization schemes by directly learning sequence–structure relationships from large datasets.

1.8.1.2 Emerging Efforts in Three-Dimensional RNA Design

In contrast to the well-studied field of 2D design, three-dimensional RNA sequence design has received comparatively limited attention. Early attempts, such as FARNA [102] and Rosetta fixed-backbone redesign [103] , introduced energy-based strategies remained constrained by computational cost and backbone rigidity. Moreover, recent deep learning approaches have begun to address these challenges. gRNAde applies a multi-state graph neural network with autoregressive decoding to design RNA sequences across multiple backbone conformations, explicitly incorporating structural flexibility [104] . RiboDiffusion frames the inverse folding problem as conditional sequence generation on a fixed 3D backbone, employing generative diffusion models to improve design accuracy [105]. Meanwhile, RhoDesign has been developed for

the de novo design of RNA aptamers, coupling structural prediction with generative modeling to guide sequence design [104]

1.8.1.3 Context-Aware Geometric Deep Learning for RNA Design

Despite the previous methods progress, existing 3D RNA design methods typically assume isolated RNA structures [104]and do not consider the broader molecular environments in which RNAs operate. In biological contexts, RNA structure and stability are often influenced or even determined by interactions with proteins [106], small molecules [107], ions [108], or DNA [109]. The development of RISoTTo (RIbonucleic acid Sequence design from TerTiary structure) addressed the context limitation by adapting the CARBonAra protein design framework for RNA applications [110]. RISoTTo introduced a parameter-free geometric deep learning approach that generates RNA sequences conditioned on both backbone scaffolds and surrounding molecular environments, including proteins, small molecules, DNA, and ions. The success of RISoTTo in achieving superior sequence recovery compared to existing methods (62% average recovery versus 45% for Rosetta and 43% for contemporary approaches) demonstrated the critical importance of molecular context in RNA design. The method's context-aware architecture enabled it to capture geometric and chemical constraints imposed by interacting partners, yielding biologically relevant sequences that maintained native-like folds when predicted using structural prediction tools.

1.8.2 Addressing Current Limitations in RNA Sequence Prediction

1.8.2.1 The Multi-State Extension Challenge

Despite its advances in context awareness, RISoTTo operated under the assumption of static structural scaffolds, treating each RNA backbone as a fixed geometric entity. This limitation became particularly apparent when attempting to design functional RNAs that require conformational flexibility for their biological roles. The recognition that many RNA functions emerge from the interplay between conformational dynamics and molecular context presented a new challenge that required extending beyond single-state design paradigms. The inspiration for addressing this challenge came from gRNAde's demonstration that multi-state modeling could improve sequence recovery for structurally flexible RNAs [104]. However, gRNAde's approach was limited to isolated RNA molecules and could not incorporate the contextual information that RISoTTo had shown to be crucial for functional design. This created an opportunity to develop the first context-aware, multi-state RNA sequence design framework.

While significant progress has been made in developing state-of-the-art methods for binding affinity prediction and in advancing strategies for rational RNA design, the ultimate impact of these innovations depends on their accessibility to the broader scientific community. Computational tools, no matter how powerful, must be delivered in a form that enables widespread adoption, reproducibility, and ease of integration into existing research pipelines. At present, the most effective means of achieving this dissemination is through the development of web-based platforms. Such platforms provide user-friendly interfaces, remove the need for extensive local computational resources, and facilitate community-driven benchmarking and iterative improvement. By lowering technical barriers, web-based implementations ensure that novel algorithms and design frameworks are not restricted to specialists in computational biology, but instead become broadly usable resources that can accelerate experimental discovery and translational applications.

1.9 Web-Based Platforms for Computational Biology

The development of web-accessible platforms has become a critical step in translating methodological advances in computational biology into tools that can be adopted by the broader research community. While standalone models and scripts often demonstrate strong predictive capabilities, their impact remains limited if accessibility requires specialized expertise or extensive computational resources. Web servers help overcome this barrier by providing user-friendly interfaces that enable non-specialists to apply advanced predictive methods to their own systems. In the specific field of protein-RNA binding affinity prediction, only a few dedicated web resources exist. The PredPRBA server [111] introduced a gradient boosted regression tree framework for predicting binding affinities, offering one of the earliest publicly available implementations in this domain. More recently**,** PRA-Pred [64] provided a structure-based web server that integrates base parameters, interaction energies, and hydrogen bond counts to estimate ΔG values. Both servers represented important steps toward practical deployment, yet their predictive accuracies and dataset coverage remain limited when applied to structurally diverse protein–RNA systems. To address these limitations, we developed the PANTHER web server, which extends our machine learning framework based on a novel local-to-global scoring methodology and Random Forest regression. By integrating amino acid–nucleotide interaction energies derived from molecular dynamics simulations into a global binding score, PANTHER demonstrates superior performance across benchmark datasets compared to existing tools. Importantly, the server was designed with accessibility in mind: it allows predictions from single PDB identifiers, batch submissions, or

uploaded structures, and provides both quantitative affinity scores (in kcal/mol) and thermodynamic categorizations. By making the method freely available through a web interface, PANTHER lowers the entry barrier for experimental and computational biologists, thereby facilitating applications in structural biology, functional genomics, and RNA-targeted drug discovery.

1.10 Thesis Contributions: An Integrated Framework

1.10.1 The Complete Research Cycle

. The progression from the development of the PANTHER methodology to its deployment as a web server exemplifies the complete research cycle in computational biology. The initial methodological innovation addressed fundamental limitations in binding affinity prediction through the local-to-global approach, the integration of machine learning techniques, and the evaluation of the model using uncurated stress set. The web server development transformed this innovation into a practical and accessible tool, while the extended validation studies demonstrated real-world performance and reliability.

This integrated approach provides multiple supporting contributions. The methodological development advances our fundamental understanding of protein-RNA recognition mechanisms and establishes new frameworks for biomolecular interaction prediction. The web server implementation ensures broad community access to the developed technology, while the extended validation provides confidence in the method's reliability across diverse applications. In conclusion, PANTHER web service demonstrates the importance of coupling methodological innovation with user’s accessibility into considerations.

1.10.2 Synergies with Multi-State RNA Design

The extension from Protein-RNA binding affinity prediction to multi-state sequence design further illustrates how fundamental insights from binding prediction can inform engineering approaches for designing functional biomolecules. The PANTHER web service also creates new possibilities for integrating binding affinity prediction with sequence design applications. The platform's ability to rapidly evaluate binding affinities for multiple structures enables systematic validation of sequences generated by Multi-State RISoTTo. This creates a feedback loop where designed sequences can be evaluated for their predicted binding properties, enabling iterative optimization of both binding affinity and conformational compatibility. The combination of accessible binding affinity prediction with context-aware, multi-state sequence design represents a comprehensive toolkit for protein-RNA interaction engineering. Researchers can use PANTHER to

evaluate existing complexes, identify optimization targets, and then employ Multi-State RISoTTo to generate improved sequences that maintain stability across conformational ensembles. Moreover, the binding affinity prediction and sequence design represents a fundamental duality in computational RNA biology. While binding affinity prediction seeks to understand and quantify existing interactions, sequence design aims to engineer new interactions with desired properties. Both problems share common underlying principles: they require understanding the molecular determinants of recognition, the role of structural complementarity, and the contribution of dynamic effects to binding thermodynamics. The local-to-global approach developed in PANTHER for binding affinity prediction revealed that protein-RNA interactions could be effectively decomposed into pairwise contributions between amino acids and nucleotides. This insight suggests that the reverse process, designing sequences to optimize local interaction patterns, might provide a pathway for rational RNA design. However, such an approach must account for the fact that functional RNAs often exist as ensembles of conformational states, each contributing to the overall binding thermodynamics.

1.11 Future Directions and Thesis Organization

1.11.1 Toward Unified Prediction and Design Platforms

The work presented in this thesis points toward future developments where binding affinity prediction, sequence design, and conformational analysis become unified within integrated computational platforms. Such integration could enable comprehensive workflows for protein-RNA system engineering, from initial complex analysis through optimized sequence design and final validation.

1.11.2 Thesis Structure

This thesis is organized to present both individual contributions and their integrated significance. Chapter 2 contains the details of methods used in this thesis, Chapter 3 details the PANTHER methodology development, including the local-to-global approach, machine learning model optimization, and validation studies. Chapter 4 presents the web service development, extended validation studies, and analysis of community impact. Chapter 5 covers the Multi-State RISoTTo extension, including context-aware geometric deep learning and multi-conformer fusion strategies. Chapter 6 provides an integrated analysis examining the complementary insights and combined implications of all contributions. Chapter 7 discusses broader impacts, limitations, and

future research directions that could further advance the field of computational protein-RNA interaction analysis and design. Finally, the conclusion of this thesis is reported in chapter 8. The comprehensive scope of this thesis, from fundamental method development through community deployment and design applications, provides both immediate contributions to computational biology and establishes foundations for future advances in understanding and engineering biomolecular interactions.

Chapter 2

Methods and Models

2.1 Molecular Modeling

Molecular modeling is a set of computational techniques used to represent, visualize, and engineer biomolecular structures at atomic resolution [112]. These methods constitute fundamental tools in structural biology and computational chemistry, providing essential capabilities for structural analysis, validation, and preparation prior to advanced computational studies such as molecular dynamics simulations or molecular docking experiments [113], [114]. Modern molecular visualization and modeling software packages provide sophisticated interactive environments for exploring biomolecular conformations and identifying structural inconsistencies. Among the most widely used tools are PyMOL [115], which offers advanced rendering capabilities, visualization, scripting functionality for engineering the molecules and offers rendering of high-quality molecular graphics; UCSF Chimera [116], which integrates seamlessly MODELLER [117] to support the visualization, structural analysis and homology modeling capabilities; and VEGA ZZ [118], which provides complete molecular editing and structure preparation utilities. Critical tasks in molecular modeling workflow typically include: (i) detection and correction of bond connectivity issues arising from resolution limitations or crystallographic artifacts; (ii) assignment and validation of appropriate atom types according to standard force field parameters; (iii) identification and proper representation of disulfide bonds, which are crucial for protein stability and function; (iv) addition of hydrogen atoms, which are sometimes not resolved in X-ray crystallographic structures but are essential for accurate computational modeling; (v) reconstruction of missing structural elements such as loop regions, terminal residues, or entire domains that may be absent due to disorder or experimental limitations; and (vi) correction or adjustment the protonation states [119], [120]. Structure preparation protocols often combine automated algorithms with manual validation to ensure chemical and physical reasonableness. For instance, missing loop regions can be modeled using homology-based approaches or ab initio prediction methods, while hydrogen atom placement typically follows standard geometric criteria based on hybridization states and local chemical environment [121], [122].

In this thesis, the integration of multiple computational tools ensures complete structure validation and optimization, guaranteeing that molecular models meet the stringent requirements necessary for reliable computational analysis. All complexes studied in this work were processed using molecular modeling techniques, employing VEGA ZZ and UCSF Chimera to prepare protein-RNA complexes for molecular dynamics simulations. The key tasks included detection and correction of broken bonds, validation of atom types, identification and proper representation of disulfide bridges (Cys–Cys bonds), addition of hydrogen atoms, and reconstruction of missing residues or structural elements. Notably, MODELLER's automated modules were used to model incomplete protein regions, especially when experimental structures contained missing loops or terminal residues. The terminal residues were modelled only if located within 9 Å of the protein-RNA binding sites. VEGA ZZ was further employed for molecular integrity verification, file format conversion, and construction of complete structural models suitable for subsequent molecular dynamics simulations (MD) and analysis.

2.2 Molecular dynamics

2.2.1 Theoretical Foundation

MD simulation provides a computational framework for investigating complex biomolecular systems that are too large for quantum mechanical treatment but can be accurately described through classical mechanics [123]. For systems that do not involve chemical reactions, simulated at low temperatures (where quantum effects dominate), or do not require a detailed hydrogen motion (quantum mechanical motion), the classical MD approach offers an effective and computationally efficient solution. This approach is governed by Newton's equations of motion (Equations 2.1 and 2.2):

$$\frac{d^2 r_i}{dt^2} = \frac{1}{m_i} F_i \tag{2.1}$$

$$\text{Where } F_i = -\frac{\partial V(r_1, r_2, \ldots, r_N)}{\partial r_i} \tag{2.2}$$

represents the force acting on atom *i* with mass m_i. $V(r_1, r_2, \ldots, r_N)$ is the force field operating on the system, which constitutes an effective interaction mechanism incorporating the average electronic effects. Standard molecular force fields employ a common functional form (Equation 2.3).

$$V = \sum_{\text{bonds}}(1/2)\ K_b\left(b - b_{\text{eq}}\right)^2 + \sum_{\text{angles}}\frac{1}{2}K_\theta\left(\theta - \theta_{\text{eq}}\right)^2 + \sum_{\text{pairs}} 4\ \varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{\text{pairs}}\frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \quad (2.3)$$

Where the bonded terms describe covalent interactions, including bond stretching, angle bending, torsional rotations (dihedrals), while non-bonded terms [124] account for van der Waals (Lennard-Jones) and electrostatic (Coulombic) interactions. Force field parameters are typically derived from quantum mechanical calculations and experimental data for specific molecular classes. There are different force field parametrization strategies. In some cases, parameters are derived from *ab initio* quantum chemical calculations on small molecular fragments. Alternatively, they may be optimized to reproduce experimental measurements such as crystallographic data, vibrational spectra, X-ray scattering profiles, liquid-phase properties (e.g., density, enthalpy of vaporization), solvation free energies, or NMR data. Importantly, each force field is calibrated for a particular class of molecules (such as inorganic systems, organic compounds, proteins, or nucleic acids). As a result, no single force field is universally applicable. The suitability of a force field depends on the biomolecular system under investigation as well as the thermodynamic and boundary conditions considered in the simulation.

For biomolecular simulations, widely adopted parameter sets include those from the Amber suite, such as ff14SB for proteins [124] , OL3 for RNA nucleotides and OL15 for DNA [125]. In GROMACS, commonly used force fields include CHARMM36 [126] and GROMOS54a7 [127], which are frequently applied in simulations of proteins, lipids, and carbohydrates. The careful choice of these parameter sets is essential, as it directly affects the accuracy of structural dynamics, interaction energies, and thermodynamic predictions generated during a MD simulation.

Newton's equations of motion are numerically integrated using algorithms, typically the Verlet [128] or leap-frog methods, with alternative approaches including Beeman [129] or Gear predictor-corrector [130] algorithms. The simulation time step (δt ≈ 1-2 fs) is kept significantly shorter than molecular vibrational periods [131] . Periodic boundary conditions (PBCs) [132] are applied to eliminate surface effects by surrounding the simulation box with infinite translated copies, ensuring bulk-like behavior throughout the system.

Electrostatic interactions require special treatment due to their long-range nature. The Particle Mesh Ewald (PME) method [133] efficiently computes these interactions by decomposing them into short-range real-space and long-range reciprocal-space components (Equation 2.4):

$$E = E_{\text{real}} + E_{\text{reciprocal}} + E_{\text{correction}} \tag{2.4}$$

This decomposition strategy transforms the conditionally convergent Coulombic summation into rapidly convergent series by introducing Gaussian screening functions. The real-space term E_{real} accounts for short-range interactions in direct space using complementary error functions that decay exponentially with distance, thereby effectively screening long-range contributions. The reciprocal-space term $E_{\text{reciprocal}}$ captures the long-range electrostatic interactions through Fourier transforms in k-space, where the periodic nature of the system is naturally accommodated. The correction term $E_{\text{correction}}$ removes false self-interactions that arise from the artificial Gaussian charge distributions introduced during the screening process. In terms of computational efficiency, bond constraints (SHAKE [134], LINCS [135]) eliminate high-frequency vibrations, particularly O-H and N-H stretches, allowing larger time steps. Additionally, virtual sites can replace hydrogen atoms [136], with their positions calculated geometrically from heavy atoms. This approach reduces computational overhead while preserving essential physical accuracy of the simulated system.

When Newton's equations of motion are integrated without external constraints, the total energy of the system remains constant. Under these conditions, if the simulation volume is also fixed, the resulting trajectory corresponds to a microcanonical ensemble (NVE). However, microcanonical dynamics rarely reflect experimental conditions, where temperature and pressure are typically controlled. To address this limitation, several algorithms have been developed to regulate these parameters. One of the simplest and most widely used schemes is the Berendsen thermostat [137], which mimics weak coupling of the system to an external heat bath at a reference temperature T_0. In this method, the system velocities are periodically rescaled using a relaxation constant, gradually adjusting the instantaneous temperature toward T_0 according to an exponential relation. An alternative approach is the isothermal (or isokinetic/isogaussian) thermostat[138], which enforces strict temperature control by modifying the equations of motion with an additional friction-like term. This term dynamically adjusts particle velocities so that the kinetic energy remains constant throughout the simulation. While this method correctly reproduces the configurational distribution expected in the canonical (NVT) ensemble, it does not yield a truly canonical momentum distribution [139] More rigorous thermostats that correctly generate canonical (NVT) ensemble

distributions include the Nosé-Hoover thermostat [140], [141], [142], which extends the system by coupling it to a fictitious heat bath variable. This introduces a dynamic friction coefficient that adjusts particle velocities according to the temperature deviation from the target value, ensuring proper sampling of both configurational and momentum distributions in the canonical ensemble. For simulations requiring constant-pressure conditions, the isothermal–isobaric (NPT) ensemble maintains fixed particle number, temperature, and pressure by allowing the system volume to fluctuate. Pressure control can be implemented using barostat algorithms such as the Berendsen barostat [143], which efficiently equilibrates pressure through isotropic box scaling, or the Parrinello–Rahman barostat [144], which treats box vectors as dynamic variables to permit anisotropic deformations and rigorously sample the NPT ensemble.

In practice, molecular dynamics simulations often employ both ensembles, NVT for equilibration at constant volume and temperature, and NPT for production runs or conditions requiring pressure control to ensure thermodynamic stability and realistic representation of experimental environments.

2.2.2 Simulation Protocol

Standard MD protocols typically include: (i) system setup with appropriate solvation and neutralization, (ii) energy minimization to remove unfavorable steric contacts, (iii) equilibration with gradual heating and density adjustment, and (iv) production simulations of sufficient duration to ensure statistical convergence. Modern implementations achieve scalability through domain decomposition and parallel processing. In this thesis, MD simulations for Protein-RNA complexes were performed using the AMBER20 simulation suite in conjunction with the AMBER14SB force field and TIP3P water model in cubic periodic boundary conditions. All systems were neutralized by adding appropriate numbers of counterions to ensure electrical neutrality. A standardized four-step MD protocol was implemented: First, energy minimization was conducted through 30,000 steps of steepest descent optimization in three substages-hydrogen atoms were minimized initially, followed by water molecules, and finally the entire system. Second, gradual heating from 0 K to 300 K was performed, with temperature control maintained using the velocity-rescaling algorithm. Third, a comprehensive six-substage equilibration protocol was employed, beginning with 200 ps of NPT ensemble simulation at 1 bar pressure, followed by five consecutive 100 ps NPT runs to gradually adjust system volume and achieve water density of 1 g/cm^3, concluding with a 100 ps NVT ensemble simulation to relax all bonds before production. The LINCS algorithm was used to constrain bond

lengths, enabling a 2 fs integration timestep, while long-range electrostatic interactions were calculated using the particle mesh Ewald method with a 1.1 nm cutoff radius. Fourth, production simulations were executed using NVT ensemble conditions to generate 500 ns trajectories for each system. Trajectory analysis was performed using custom scripts within the Cpptraj program of the AMBER20 suite, including calculations of Root Mean Square Deviation (RMSD), Root Mean Square Fluctuation (RMSF), and secondary structure analysis using DSSP methodology.

2.3 Molecular-Dynamics-Derived Energy Decomposition

2.3.1 Theorical Framework

Energy decomposition analysis emerged as a fundamental computational method in the early development of molecular mechanics (MM) force fields. Pioneering work by Lifson and Warshel in the 1960s established the theoretical basis for partitioning molecular energies into physically meaningful components [145]. The extension of this approach to biomolecular simulations gained prominence in the 1980s with the work of Kollman and colleagues, who developed systematic strategies for analyzing pairwise interactions in proteins, nucleic acids, and protein–ligand complexes using classical force fields [146], [147]. These developments were pivotal, enabling researchers to dissect and rationalize the complex energy landscape of biomolecular recognition into interpretable physical terms. Subsequent studies by Massova and Kollman [148] demonstrated that energy decomposition could provide quantitative insights into binding energetics, often reproducing experimental affinities within chemical accuracy, thereby establishing its role as a robust and computationally efficient analysis technique. The theoretical foundation of MD-driven energy rests on the principle of energy additivity of classical force fields. Within this framework, the total system energy is expressed as a sum of bonded (bond, angle, and dihedral) and non-bonded (van der Waals and electrostatic) terms (Equation 2.5):

$$E_{\text{total}} = E_{\text{bonded}} + E_{\text{non-bonded}} \tag{2.5}$$

Because the non-bonded terms are pairwise additive, they can be partitioned into contributions between specific groups of atoms (e.g., an amino acid residue and a nucleotide base). This forms the basis of pairwise energy decomposition in molecular dynamics simulations.

2.3.2 Statical Analysis

Since MD trajectories represent ensembles sampling the canonical (NVT) or isothermal–isobaric (NPT) distribution, instantaneous pairwise energies fluctuate considerably due to thermal

motion. To obtain meaningful values, energies are averaged across many equilibrated trajectory frames (Equation 2.6):

$$\langle E_{ij} \rangle = \frac{1}{N} \sum_{t=1}^{N} E_{ij}(t) \tag{2.6}$$

Where *N* is number of frames and $E_{ij}(t)$ is the instantaneous interaction energy at time *t*. Ensemble averaging serves to reduce noise arising from short-term fluctuations and to produce statistically converged estimates of residue-nucleotide interaction strengths. However, convergence must be carefully assessed, as insufficient sampling can bias the results.

2.3.3 Computational Implementation

In practice, decomposition is carried out by post-processing MD trajectories. For each snapshot, the pairwise interaction energy E_{ij} between groups *i* and *j* and is computed using the same functional forms employed during the simulation. This ensures thermodynamic consistency between decomposition analysis and the underlying MD force field. The interaction energy is commonly decomposed into van der Waals and electrostatic contributions, following the Lennard-Jones and Coulomb potential forms, respectively. Pairwise energy decomposition is a computational strategy used to quantify the energetic contributions between specific groups of atoms, such as amino acid residues and nucleotide bases. The total interaction energy between a residue *i* and a nucleotide *j* is typically expressed as the sum of van der Waals (E_{vdw}) and electrostatic (E_{elec}) components [149], [150] (Equation 2.7):

$$E_{ij} = E_{vdw,ij} + E_{elec,ij} \tag{2.7}$$

The van der Waals interactions ($E_{vdw,ij}$) are calculated by the Lennard-Jones potential for all atom pairs (a ϵ *i*, b ϵ *j*) (Equation 2.8).

$$E_{vdw,ij} = \sum_{a \,\epsilon\, i} \sum_{b \,\epsilon\, j} 4\epsilon_{ab} \left[\left(\frac{\sigma_{ab}}{r_{ab}} \right)^{12} - \left(\frac{\sigma_{ab}}{r_{ab}} \right)^{6} \right] \tag{2.8}$$

where: ϵ_{ab} is the depth of the potential well for the interaction between atom a and atom b, σ_{ab} is the distance at which the potential between a and b is zero or the minimum, and r_{ab} is the distance between atom a (in group i) and atom b (in group j) [151]. The 12th term and the 6th term represent the repulsive and attractive terms, respectively. The electrostatic energy component is calculated by Coulomb's law (Equation 2.9).

$$E_{elec,ij} = \sum_{a \,\epsilon\, i} \sum_{b \,\epsilon\, j} \frac{q_a \, q_b}{4\pi\epsilon_0 r_{ab}} \tag{2.9}$$

where: q_a and q_b are the partial charges on atoms a (in group *i*) and atom b (in group *j*), respectively, ϵ_0 is the relative dielectric constant in implicit solvent, and r_{ab} is the distance between atom a and atom b. Electrostatic interactions are long-ranged but decay with distance, and their treatment must remain consistent with the simulation parameters, such as the cutoff scheme or use of Ewald summation.

In this thesis, to reduce computational costs and to eliminate noise from the data, a cutoff distance (r_{cut}) of 12 Å was applied for pairwise energy decomposition calculations. These calculations were conducted to analyze the interactions between amino acids and nucleotide pairs only if the distance between their respective centers of mass (COM) satisfied $r_{ab} \leq r_{cut}$. This threshold was chosen to be slightly larger than the 11 Å cutoff used with the Particle Mesh Ewald (PME) method [152]. The additional 1 Å margin ensures that we can capture all significant interactions that might occur at the boundary of the PME cutoff. The selection of 12 Å is also justified by the exponential decay of electrostatic and van der Waals interactions beyond this threshold, which renders the energetic contributions negligible. This approach encompasses both short-range and intermediate-range interactions, ensuring the inclusion of biologically relevant contacts while maintaining consistency with the overall simulation parameters.

2.4 Data Generating for Machine Learning

To consider the dynamic behavior of the interactions and to generate sufficiently large datasets for ML training, we adopted a time-averaged interaction energy approach. Specifically, the interaction energy between the amino acid *i* and the nucleotide base *j* was averaged over 10 consecutive trajectory frames (Equation 2.10).

$$\langle E_{ij} \rangle = \frac{1}{N} \sum_{t=1}^{N} E_{ij}^{(t)} \tag{2.10}$$

where N is the averaging window and is equal to 10, and $E_{ij}^{(t)}$ is the interaction energy between residue *i* and nucleotide *j* at the time frame *t*. After calculating the average interaction energy for a set of 10 consecutive frames, we skipped the following 40 frames and repeated the calculations on the next 10-frame window. This procedure was repeated throughout the entire 500 ns simulation, ensuring consistent sampling and reliable time-averaged interaction profiles. By applying this method for each interaction pair, we obtained a series of 10-frame averaged interaction energies. The choice of a sampling window comprising 10 frames, separated by 40 skipped frames represents an optimal balance between statistical robustness and computational efficiency.

Averaging over 10 frames provides sufficient sampling to smooth thermal fluctuations while maintaining the ability to capture relevant interactions. The 40 skipped frames minimize redundancy between computed averages, thereby maintaining statistical independence between consecutive data points. This approach filters out transient interactions by focusing on persistent binding modes and discarding those interactions that are present in only 20-30% of the trajectory, as these are likely to represent stochastic events rather than stable binding conformations. Such filtering strategy ensures that the ML models are trained exclusively on thermodynamically significant interactions, thus improving the biological relevance of their predictions. Along with the above-mentioned energetic averages, a set of structural descriptors was collected to be used as features to train the ML models. These included (i) pairwise distances, quantified as the Euclidean distance between the centers of mass (Å) of the interacting amino acid and nucleotide, and (ii) a set of additional interaction-specific features to better describe the relevant biophysical aspects of protein-RNA recognition and binding.

2.5 Machine Learning Models

2.5.1 Classification and Regression

Supervised machine learning tasks are broadly divided into classification and regression, depending on the nature of the target variable being predicted [153], [154]. Classification refers to predicting categorical outcomes, such as class labels, phenotypes, or molecular functional states. The model learns a decision boundary that separates data points belonging to different categories. The model performances are typically measured using metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Example applications in computational biology include the prediction of protein functional classes, RNA binding sites, and the discrimination between healthy and diseased samples. Regression, in contrast, predicts continuous numerical outcomes, such as binding affinity, expression level, or free energy change. The model performances are usually assessed by metrics such as mean squared error (MSE), root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). While classification and regression share common algorithmic foundations, their objectives and evaluation criteria differ fundamentally. In biomolecular modeling, regression methods are often crucial, because they allow the estimation of quantitative molecular properties (e.g., binding energy landscapes, folding stability changes) rather than only categorical outcomes. This justifies the in-depth focus on regression models in the following subsections.

2.5.2 Machine Learning Regression Models

Machine learning (ML) regression algorithms provide a sophisticated mathematical framework for predicting continuous variables by discovering complex statistical patterns from empirical data, transcending the constraints of predefined physical equations. In the context of biomolecular modeling and protein-RNA interaction analysis, regression methods prove particularly invaluable for capturing intricate, high-dimensional, and frequent nonlinear relationships between structural descriptors, energetic features, and molecular properties of biological significance. The fundamental objective of ML regression is to learn an optimal mapping function that accurately models the relationship between multi-dimensional input features and continuous target variables. This learning paradigm optimizes model parameters through the minimization of prediction errors while incorporating regularization techniques to prevent overfitting and ensure generalization to unseen data [154], [155]. The performance evaluation of regression models typically employs standardized metrics including the coefficient of determination, root mean square error, and mean absolute error, providing quantitative assessments of model accuracy and predictive capability. These metrics enable objective comparison between different algorithmic approaches and facilitate model selection based on specific performance criteria [153]. A comprehensive taxonomy of algorithms encompasses interpretable linear models, robust tree-based ensembles, and sophisticated neural architectures, each presenting distinct trade-offs between computational efficiency, model interpretability, robustness to noise, and predictive accuracy. Algorithm selection depends critically on dataset characteristics including dimensionality, sample size, signal-to-noise ratio, and the underlying complexity of the target function.

2.5.3 Linear Regression

Linear regression constitutes the foundational regression methodology in statistical learning. It models the target variable through a linear combination of input features, each weighted by a coefficient and accompanied by an interception term. The approach assumes that changes in the target variable are proportional to changes in the predictor variables, with the error term normally distributed, having zero mean and constant variance (Equation 2.11) [156].

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \tag{2.11}$$

where β_0 is the intercept, β_j are regression coefficients, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Parameter estimation in linear regression relies on the ordinary least squares (OLS) criterion, which minimizes the sum of squared residuals between predicted and observed values. This optimization

yields analytical solutions when the feature matrix is invertible, providing the best linear unbiased estimator, according to the *Gauss-Markov theorem*. The resulting coefficients represent the expected change in the target variable for unit changes in each predictor, enabling direct interpretation of feature contributions [157]. However, in high-dimensional biological datasets, where the number of features exceeds the number of samples or multicollinearity is present, regularized extensions become essential. *Ridge regression* incorporates penalty terms proportional to the sum of squared coefficients, shrinking parameter estimates toward zero while maintaining all features in the model. In contrast, *Lasso regression* employs penalty terms proportional to the sum of absolute coefficient values, providing automatic feature selection through sparsity induction by setting irrelevant coefficients to exactly zero and thereby performing automatic feature selection. *Elastic Net* combines both penalties, balancing feature selection and coefficient shrinkage to handle correlated feature groups effectively [158], [159]. Despite its computational simplicity and statistical interpretability, linear regression assumes linearity between features and target, feature independence, homoscedasticity of residuals, and absence of influential outliers. These assumptions are frequently violated in complex biomolecular systems exhibiting nonlinear interactions, cooperative binding effects, and intricate regulatory mechanisms that require more sophisticated modeling approaches.

2.5.4 Gradient Boosting Regression

Gradient boosting represents a powerful ensemble methodology that constructs strong predictive models by sequentially combining multiple weak learners, typically shallow decision trees, within an additive modeling framework. The algorithm follows the principle of functional gradient descent, iteratively fitting new models to the negative gradients of the loss function with respect to current predictions, thereby progressively correcting errors made by previous models [160]. Gradient boosting can be formally described as a stage-wise additive model. The process begins with an initial model $F_0(x)$ that minimizes the overall loss by predicting a constant value. At each iteration m, the algorithm computes pseudo-residuals r_{im}, which are the negative gradients of the loss function with respect to current predictions. A new weak learner $h_m(x)$ is then fitted to these pseudo-residuals, and the model is updated as what shows in Equation 2.12.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{2.12}$$

Where, γ_m is a step size determined through optimization or set as a fixed learning rate. This iterative procedure progressively reduces the loss function and allows the ensemble to capture

complex nonlinear relationships. Here, $h(x)$ is a weak learner trained to approximate the negative gradient of the loss function (the pseudo-residuals), while γ_m is a step size controlling how much the new learner contributes. The mathematical foundation involves building an ensemble through a sequential process where each iteration adds a new weak learner trained on pseudo-residuals representing the negative gradient of the loss function. The initial model provides a constant prediction minimizing the overall loss, followed by successive models that capture patterns in the residual errors. Each new model is scaled by an optimal step size, determined through line search optimization to minimize the overall loss function [160].

The method allows modeling of complex nonlinearities and interactions. Gradient boosting is highly flexible with respect to the choice of loss function: one can use squared error (for regression), absolute error, Huber loss, or other differentiable (or nearly differentiable) loss functions to trade off robustness and sensitivity to outliers.

2.5.5 XGBoost Regression

XGBoost (eXtreme Gradient Boosting) is an optimized implementation of gradient boosting that incorporates algorithmic enhancements and regularization. It introduces second-order Taylor expansion of the loss function for more accurate approximation, regularization terms on tree complexity (number of leaves, leaf weights), and efficient handling of sparse or missing data (Equation 2.13) [161].

$$\mathcal{L} = \sum_i L(y_i, \widehat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.13}$$

where ℓ is the loss function (e.g. squared error) and Ω(f) is the regularization term penalizing model complexity (e.g. $\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ for a tree with *T* leaves and weights w_j. The XGBoost objective function integrates both prediction error and model complexity through explicit regularization terms that penalize the number of leaves and the magnitude of leaf weights in each tree. This dual regularization approach prevents overfitting while maintaining model flexibility, enabling superior performance on both training and validation datasets. The key theoretical advancement lies in employing second-order Taylor expansion for loss function approximation, utilizing both first and second derivatives of the loss function with respect to predictions. XGBoost incorporates numerous practical enhancements including column-wise sparse data structures for efficient memory usage, cache-aware algorithms for improved computational performance, and parallel processing capabilities for scalability to large datasets. The implementation provides built-in

cross-validation, early stopping, and hyperparameter optimization features that streamline the model development process [161], [162].

2.5.6 Random Forest Regression

Random Forest constitutes a robust ensemble methodology that constructs multiple decision trees through bootstrap aggregating combined with random feature selection, effectively reducing model variance and mitigating overfitting compared to individual decision trees (Equation 2.14, 2.15). The algorithm implements two key randomization strategies [163]:

1. Bootstrap sampling of data (bagging) to generate varied training sets for each tree.
2. Random selection of a subset of features at each node split to decorrelate the trees.

$$\widehat{y_{RF}}(x) = \frac{1}{B}\sum_{\beta=1}^{B} T_{\beta}(x) \tag{2.14}$$

$$\text{Var}(\widehat{y_{RF}}) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{2.15}$$

where ρ = correlation between trees, σ² = individual tree variance

The algorithmic framework relies on generating multiple bootstrap samples from the original training set, each created by sampling with replacement to produce datasets of equal size but with different instance compositions. For every bootstrap sample, a decision tree is constructed, and at each split decision, only a randomly selected subset of features is considered, typically the square root of the total number of features for regression tasks. The final prediction aggregates individual tree predictions through simple averaging, which significantly reduces prediction variance through the ensemble effect. This averaging process leverages the statistical principle that the variance of averaged independent predictions decreases inversely with the number of predictors, while the bias remains unchanged. The random feature selection further reduces correlation between trees, maximizing the variance reduction achieved through aggregation. In addition, Random Forest provides valuable diagnostic capabilities through out-of-bag (OOB) error estimation, computed using predictions from trees that did not include specific samples during training. This internal validation mechanism eliminates the need for separate test sets in model evaluation and provides unbiased estimates of generalization performance during model development. Feature importance assessment employs two complementary metrics: mean decrease impurity quantifying average impurity reduction attributable to each feature across all trees, and permutation importance

measuring accuracy decrease when individual feature values are randomly shuffled. These measures provide insights into feature relevance and enable dimensionality reduction in high-dimensional datasets [163]. This method demonstrates excellent performance for nonlinear, high-dimensional biological data, provides robust handling of missing values through surrogate splits, and requires minimal hyperparameter tuning compared to other ensemble methods. Random Forest naturally handles mixed data types and provides reliable uncertainty estimates through prediction intervals. However, its extrapolation capability beyond training data ranges is limited, and may struggle with highly imbalanced datasets or datasets with very strong predictors that dominate tree construction [164], [165].

2.5.7 Neural Network Regression

Artificial neural networks represent a class of flexible computational models inspired by biological neural systems, capable of approximating arbitrary continuous functions through the universal approximation theorem. In regression applications, multilayer perceptron transforms input features through successive layers of linear transformations followed by nonlinear activation functions, enabling the representation of highly complex, nonlinear mappings between input and output spaces [166], [167]. The architectural foundation consists of interconnected layers of computational units called *neurons* or *nodes*, organized in a feedforward structure. The input layer receives raw features, hidden layers performing nonlinear transformations, and output layer generates the final predictions. Each connection between neurons has an associated weight parameter that determines the strength and direction of information flow, while each neuron applies an activation function to its weighted inputs plus a bias term.

Common activation functions include the *rectified linear unit* (ReLU) that outputs the maximum of zero and its input sigmoid functions that squash inputs to values between zero and one, *hyperbolic tangent functions* (Tanh) producing outputs between negative one and positive one, and leaky ReLU variants that allow small negative outputs. The ReLU activation has become particularly popular due to its computational efficiency and ability to mitigate vanishing gradient problems that plague deep networks [168]. Common activation functions are shown in equations 2.16, 2.17 and 2.18.

$$\text{ReLU: } f(x) = max(0, x) \tag{2.16}$$

$$\text{Sigmoid: f(x) = } f(x) = \frac{1}{1+e^{-x}} \tag{2.17}$$

$$\text{Tanh: f(x)} = f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2.18}$$

Parameter optimization in neural networks relies on the backpropagation algorithm, which computes gradients of the loss function with respect to all network parameters using reverse-mode automatic differentiation. The algorithm propagates error signals backward through the network, computing gradients layer by layer using the chain rule of calculus. These gradients are then used to update parameters through optimization algorithms such as *stochastic gradient descent* (SGD), *Adam*, or *RMSprop* [169]. Regularization techniques are crucial for neural network training, helping to prevent overfitting and improve generalization. Dropout mitigates overfitting by randomly setting a fraction of neuron outputs to zero during training, forcing the network to learn redundant representations and reducing co-adaptation between neurons. Weight decay adds penalty terms proportional to the sum of squared weights to the loss function, encouraging smaller parameter values. Batch normalization standardizes layer inputs to have zero mean and unit variance, accelerating training and improving numerical stability [170], [171].

2.5.8 Stacked Ensemble Regression

Stacked generalization (commonly known as *stacking*) is a sophisticated meta-learning approach that combines predictions from multiple heterogeneous base learners through a meta-learner, trained to optimize their integration. Unlike bagging methods that use identical algorithms on different data subsets or boosting methods that sequentially correct errors, stacking leverages the complementary strengths of diverse algorithmic approaches including linear models, tree-based ensembles, and neural networks [172], [173].

The architecture proceeds on two levels:

- Level 0 (base learners): multiple heterogeneous models are trained on the original data (or subsets).
- Level 1 (meta-learner): uses the base learners' predictions (often on validation splits) as features to learn how to combine them

Training employs cross-validation procedures to generate out-of-fold predictions from base learners, preventing information leakage that would occur if the same data used to train base learners were used to train the meta-learner. Each base learner is trained on *k-1* folds and makes predictions on the held-out fold, repeating this process across all folds to generate a complete set of out-of-fold predictions that serve as training data for the meta-learner. The meta-learner can

employ any regression algorithm, with popular choices including *linear regression* for simplicity and interpretability, *ridge regression* for handling correlated base predictions, *neural networks* for capturing complex nonlinear combinations, or *tree-based methods* for automatically discovering interaction effects between base learner outputs. The choice of an appropriate meta-learner depends on the diversity of base learners and the suspected complexity of their optimal combination. Some stacking systems use multiple hierarchical levels (multi-layer stacking), where outputs of one meta-level feed into a higher level. Stacking can outperform individual models or simple averaging by adaptively assigning weights depending on input region. However, this approach increases computational complexity (multiple model trainings, cross-validation), reduces interpretability, and requires careful validation to avoid meta-level overfitting. The success of stacking often hinges on diversity among base learners (i.e. differing inductive biases [174]. The concept of stacking was originally proposed by Wolpert (1992) [173], and further extended by Ting & Witten (1999) [172], who investigated model selection strategies and attributes for the meta-learner.

2.6 Deep Learning

Deep learning constitutes a computational framework employing multi-layered neural networks to learn hierarchical data representations through gradient-based optimization. The methodology differs fundamentally from traditional machine learning approaches by automatically extracting features from raw input data rather than relying on predetermined descriptors, enabling direct learning from high-dimensional biological datasets without manual feature engineering [169], [175]. The fundamental building block of deep learning is the artificial neuron, which computes a weighted sum of input features followed by a nonlinear activation function (Equation 2.19).

$$h = \sigma\left(\sum_{i=1}^{p} w_i x_i + b\right) \tag{2.19}$$

Where x_i are input features, w_i are learnable weights, b is a bias term and $\sigma(.)$ is a nonlinear activation function such as the rectified linear unit (ReLU) or sigmoid. Deep learning models are trained by minimizing a loss function, such as mean squared error for regression or cross-entropy for classification. The optimization is performed using *stochastic gradient descent* (SGD) or its adaptive variants (*Adam*, *RMSprop*), where gradients are computed via backpropagation. Training requires large-labeled datasets and significant computational resources, often accelerated by

graphic processing units (GPU) or tensor processing units (TPU). Due to their high capacity, deep learning models are prone to overfitting, especially in domains with limited data. Regularization strategies are critical for improving generalization:

- Dropout randomly deactivates neurons during training [170].
- Weight decay (L2 regularization) penalizes large weight values.
- Batch normalization stabilizes and accelerates learning by normalizing intermediate activations [171].
- Early stopping halts training once validation performance ceases to improve.

2.6.1 Network Architectures

Convolutional Neural Networks (CNNs) implement local connectivity patterns and parameter sharing through convolution operations, making them suitable for processing structured data with spatial or topological relationships [176]. The convolution operation computes feature maps by applying learnable filters across the input dimensions, followed by pooling operations for dimensionality reduction and translation invariance. *Recurrent Neural Networks* (RNNs) maintain hidden states to process sequential data. *Long Short-Term Memory* (LSTM) [177], and *Gated Recurrent Unit* (GRU) variants [178] address the vanishing gradient problems through gating mechanisms. These architectures enable the modeling of temporal dependencies and variable-length sequences common in biological data. Transformer architecture employs self-attention mechanisms to capture long-range dependencies without recurrent processing constraints [179]. The attention mechanism computes weighted combinations of input representations based on learned query, key, and value projections, enabling parallel processing and superior handling of long sequences.

2.6.2 Transformer-Based Geometric Deep Learning

Deep learning transformer models represent a class of neural network architectures originally developed for *Natural Language Processing* (NLP), where they revolutionized the field by introducing self-attention mechanisms capable of capturing long-range dependencies in sequential data [179]. This architecture later became the foundation for large-scale models such as *BERT* and *GPT*, which process text in parallel rather than sequentially like recurrent neural networks (RNNs), enabling both faster training and richer contextual representations [180], [181].

The success of transformer architecture in NLP has inspired their adaptation to structural biology, where molecules can be treated as structured sequences embedded in three-dimensional space. A major advancement in this direction is *geometric deep learning*, which generalizes neural networks to data defined on non-Euclidean domains such as graphs, manifolds, or point clouds [182]. By incorporating geometric invariances (e.g., translation, rotation, reflection), geometric transformers can directly process atomic structures, rather than relying exclusively on handcrafted descriptors.

The central computational principle of transformers is the attention mechanism [179], which computes a weighted combination of values V based on the similarity between queries Q and keys K (Equation 2.20).

$$Attention(Q, K, V) = softmax!\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.20}$$

This allows the model to dynamically focus on the most relevant parts of the input, enabling context-dependent representations. In practice, transformer architecture consists of multiple stacked layers of:

- Multi-head self-attention, where attention is computed in parallel across several subspaces.
- Feedforward neural networks, applied independently to each position.
- Residual connections and layer normalization, which stabilize and accelerate training.

When applied to molecular systems, transformers are adapted into geometric transformer layers, where each entity (e.g., atom, nucleotide, or residue) is described by scalar and vector features. These features are updated using attention-based message passing over local neighborhoods while incorporating relative positional encodings such as interatomic distances D_{ij} and displacement vectors R_{ij} [183], [184], [185]. A generic update rule can be expressed as what is shown in equation 21.

$$q_i^{(l+1)}, p_i^{(l+1)} = TransformerLayer\left(q_i^{(l)}, p_i^{(l)}, \{q_j^{(l)}, p_j^{(l)}, D_{ij}, R_{ij}\}_{j\in nn(i)}\right) \tag{2.21}$$

where q_i and p_i represent scalar and vector features of entity I, and $nn(i)$ denotes its neighbors. This formulation preserves rotational and translational equivariance, a critical property for modeling molecular geometry. Transformer-based deep learning models excel at capturing complex, high-dimensional, and context-dependent relationships within biological data. They can

integrate heterogeneous input types such as sequences, structures, and molecular environments, and automatically learn hierarchical feature representations without manual feature engineering. However, these advantages come with significant computational demands, particularly for large molecular systems or geometric variants, and limited interpretability compared to simpler models such as linear regression or random forests.

In this thesis, we used a parameter-free geometric transformer model for multi-state RNA sequence design conditioned on fixed tertiary structures. The model incorporates interactions with non-RNA entities including proteins, small molecules, ions, and DNA, enabling context-aware design. The model was trained and validated on RNA-containing tertiary structures from the PDB database, ensuring exposure to a diverse set of biologically relevant conformations and interaction contexts.

Chapter 3

PANTHER - Protein-Affinity for Nucleic Target binding, Hybridization, and Energy Regression

3.1. Introduction

Protein-RNA interactions play a central role in numerous key biological processes [186], from gene expression to cellular signaling and regulation. Experimentally determined binding affinities provide critical insights into functions and specificity and can be used to better evaluate protein-nucleic acid interactions. Binding affinity quantitatively describes the strength of interaction between a protein interacts and a nucleic acid and is essential for regulating processes like transcription, replication, and repair. Often a high binding affinity correlates with effective gene regulation and cellular function [187], [188]. Understanding binding affinities also provides mechanistic insights by elucidating the recognition mechanisms between proteins and nucleic acids and allowing a more precise understanding of the structural or sequence elements that can affect interactions [188], [189]. Experimental methods for determining the binding affinity of protein-nucleic acid complexes include several well-established techniques. These comprise fluorescence spectroscopy, as reported by Vivian and Callis in 2001; surface plasmon resonance, as discussed by Katsamba in 2002; electrophoretic mobility shift assay, as described by Hellman and Fried in 2007 and later by Ryder et al. in 2008; isothermal titration calorimetry, as highlighted by Feig in 2009; filter binding assay, as proposed by Rio in 2012. Although highly accurate, these experimental methods are both costly and labor-intensive. This limitation reduces their applicability in high-throughput screenings of large libraries as the field of drug discovery evolves to focus on increasingly larger biomolecules such as RNA-aptamers (Bell et al. 2020). As the demand for reliable techniques to investigate protein-nucleic acid interactions has rapidly grown, computational methods have gained popularity to overcome the limitations of these experimental techniques [189], [197], [198], [199]. Emerging computational methodologies have led to the widespread utilization of molecular docking software for predicting binding conformations and ranking them through scoring functions [200], [201]. Focusing on protein-RNA binding free energy, four notable protein-RNA web-based/standalone predictive tools are available: *PRA-Pred* [202], *PRdeltaGPred* [203], the structure-based model as proposed by Nithin et al. in 2019, and *PredPRBA* [205]. *PRdeltaGPred* is a standalone program that utilizes complex structural features, including non-interaction surfaces, desolvation

energy, hydrogen bond energy, and salt bridge energy, to predict experimental binding free energy. Similarly, PRA-Pred is a web-based tool that predicts ΔG value based on simple features such as base parameters, interaction energies, number of contacts, and hydrogen bonds. Next, Nithin et al. developed a structure-based model that predicts the ΔG value by considering interface parameters such as hydrophobicity, contact surface, and hydration pattern. Finally, PredPRBA [205] includes different predictive models depending on the interacting RNA structure. At present, the limited availability of tools capable of accurately calculating the experimental ΔG value for protein-RNA interactions highlights a critical gap in computational biology. Despite the availability of above-mentioned tools, the field continues to face an urgent need for more robust, accessible, and accurate methods. Developing new computational-based methods to predict experimental *binding free energy* (BFE) will significantly enhance protein-RNA research, advancing our understanding of their roles in biological systems and supporting drug discovery efforts. To address the need for specifically tailored approaches, we developed a machine learning (ML) model that predicts local energies between interacting amino acid-nucleotide base of a protein-RNA complex and later we integrate these local energy contributions through a scoring function, by a local-to-global approach. We refer to the entire process as the *Protein-Affinity for Nucleic Target-binding, Hybridization, and Energy Regression* score (PANTHER score), which reveals a strong and encouraging correlation with the experimentally determined binding free energy between proteins and nucleic acids. The graphical abstract illustrating this workflow is shown in figure 3-1. This approach focuses on modelling local pairwise interactions between amino acids and nucleotide bases and can be subdivided into three main parts: (i) A representative set of local interaction energies was derived from molecular dynamics (MD) simulations involving a training set composed by 46 curated protein-RNA complexes (training set 1 & training set 2); (ii) ML models were trained to predict these local interaction energies without performing MD runs but based on simple pairwise interaction features, such as amino acid type, nucleotide base type, the number of hydrogen bonds, and distances between interacting amino acids and nucleotide bases; (iii) the so-predicted local energies were aggregated through a local-to-global global integration scheme to calculate the model-specific PANTHER score. The ML models and the resulting model-specific PANTHER scores were initially tested and refined on a reduced test set composed of 7 protein-RNA complexes with known experimental ΔG values. Optimal performance was achieved through the application of *Random Forest Regression* (RFR). The resulting RFR model-specific PANTHER score demonstrated a Pearson correlation coefficient (r) of 0.80 and a mean absolute error of 1.63 kcal/mol. Furthermore, when

tested on a markedly more extended non-redundant protein-RNA dataset of 110 complexes, the RFR model-specific PANTHER score achieved an encouraging (r) of 0.64. These results underscore the potential of ML models for accurately predicting pairwise interaction energies and the reliability of the local-to-global approach in evaluating the protein-RNA binding energies. To further validate the model, we evaluated its performance on an independent dataset comprising 110 uncorrelated protein-RNA complexes and compared with existing prediction techniques. In detail, the comparison involved the PredPRBA and PRA-Pred, both of which are easily accessible through web services. The correlation coefficients (r) between experimental and predicted ΔG values were 0.64 for PANTHER score, 0.18 PredPRBA, and 0.18 PRA-Pred, respectively. This comparison underscores the potential of our model for evaluating protein-RNA interactions and demonstrates its robustness and superior accuracy when compared to currently available methods.

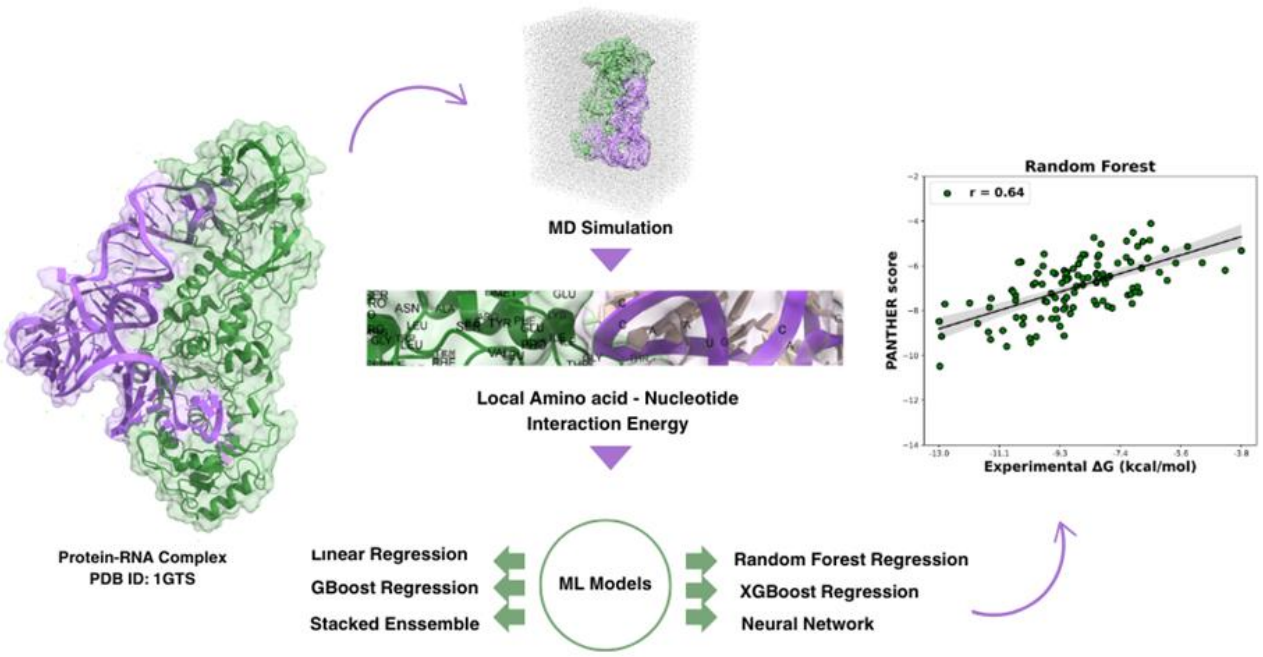

Figure 3-1. The graphical abstract.

3.2. Materials and Methods

3.2.1 PANTHER Score

In this study, we developed PANTHER score, a scoring system designed to estimate protein-RNA binding affinities through an integrated, multi-step computational strategy. The workflow involves several sequential steps. First, a curated dataset of RNA-protein complexes was selected to ensure the inclusion of systems with known experimental binding free energies (ΔG), strategically

increasing the number of complexes to tackle the lack of experimental data for developing an effective ML model. Next, MD simulations were performed on representative complexes to generate physically realistic conformations and extract energetic and structural data (further detailed in Methods sub-section “Extraction of Pairwise Local Energies and Noise Reduction Technique”). From these trajectories, pairwise decomposed interaction energies between amino acid residues and nucleotide bases were calculated, providing a detailed description of the local energetic landscape at the binding interface. These pairwise energies and structural descriptors were then used to train multiple regression machine learning (ML) models aimed at estimating the binding affinity of protein-RNA complexes. This was achieved by integrating ML-predicted local interaction energies with the proposed local-to-global aggregation framework. Accordingly, the workflow operates in two main stages:

- **Local Energy Prediction:** Multiple ML regression models, including Random Forest Regression, Gradient Boosting Regression, Extreme Gradient Boosting, Linear Regression, Stacked Ensemble, and Neural Networks were trained using amino acid-nucleotide pairwise local energy data derived from MD simulations. This training enables the prediction of local interaction energies for any given amino acid-nucleotide pair based on distance-weighted structural descriptors.
- **Local-to-Global Integration:** The predicted local energies are then aggregated through a distance-weighted summation scheme, in which each interaction contributes proportionally to its spatial proximity and biophysical relevance. This integration produces the PANTHER score (expressed in kcal/mol), representing the overall binding affinity of protein-RNA complex.

This local-to-global framework effectively combines atomistic simulation data with ML-derived predictions, providing an interpretable and generalizable measure of protein-RNA binding energetics. Notably, to define the final PANTHER score, we rigorously evaluated all the ML models through both a test set and a stress set, and we finally selected only one model (refer to Result and Discussion sub-section: “*ML model Assessment*” for further details) to ensure an interpretable and generalizable measure of protein-RNA binding energetics. The following sections provide a detailed description of each step involved in the PANTHER score workflow.

3.2.2. Datasets Selection and Preparation

In this study, a total of 163 RNA-protein complexes were retrieved from the Protein Data Bank (PDB) [206], analyzed (refer to tables 3-1, 3-2, 3-3 & 3-5), and subsequently subdivided into three distinct datasets for the training, testing, and stress testing phases of the ML model. The selection criteria for these complexes included: (i) three-dimensional structures resolved through X-ray crystallography or NMR with a resolution better than 3.0 Å, (ii) the inclusion of only single-stranded RNA, (iii) the exclusion of inhibitors and metal ions, (iv) the presence of standard residues and nucleotide bases exclusively, (v) the omission of zinc-finger proteins, (vi) an even distribution of the four nucleotide bases in RNA, and (vii) a diverse representation across various protein classifications and organisms. The selected complexes were then filtered by removing redundancy using *Clustal Omega* program [207], with a 70% sequence identity threshold to collect a total of 163 protein-RNA complexes. By comparing our dataset with existing resources, including PRBAB v2.0 [208], PDBbind [209], ProNAB [210], and the dataset from [211], we found experimental ΔG values for 130 (out of 163) complexes. To effectively divide the dataset for ML applications, we considered that the initial training of ML models is based only on MD simulations, while the following validation of the PANTHER score requires complexes with experimental ΔG values. Accordingly, the training set consisted of 33 complexes without experimental ΔG values (*Training Set 2*) and 13 complexes with experimental ΔG values (*Training Set 1*), selected to ensure comprehensive coverage of the protein-RNA interaction space, as schematized in Figure 3-2. Among the remaining 117 complexes with experimental ΔG values, we randomly selected 7 complexes that constitute the *Test Set* for preliminary tuning and validation of both ML models and ML-derived PANTHER scores. All 53 complexes included in training (Set 1 + Set2) and test sets underwent MD simulations. Notably, the 20 complexes (13 from Training Set 1 + 7 from Test Set) with known experimental ΔG values, which underwent MD runs, were also exploited to initially assess the reliability of the here-adopted local-to-global approach.

Finally, the remaining 110 complexes with known experimental ΔG values were assigned to the *Stress Set*, which was used to assess the predictive capability of each ML model-specific PANTHER score and to identify the most accurate model for final implementation. It should be noted that, unlike the train and test sets, the complexes of the stress set did not undergo MD simulations or post-processing of the extracted data. Instead, they were used to directly extract the interacting features required by the ML models. Importantly, the here-adopted local-to-global methodology and the so-organized datasets offer some relevant advantages. They allow: (i) combining complexes

with and without experimental ΔG values for the training phase, while complexes with experimental ΔG values were used only for the test and validation phases; (ii) minimizing the computational cost by carrying out MD runs only on 53 complexes and not on all 163 collected systems; (iii) augmenting the data to be used for training ML models by considering pairwise local interaction energies between amino acids and nucleotide bases as derived from MD simulations.

Prior to MD simulations, all complexes underwent a careful structural curation process to model missing amino acids or atoms using *VEGA ZZ* [212], *Modeler* [213], and *Chimera* [214] programs. The 53 curated structures of training and test sets underwent MD simulations to derive pairwise interaction energies (further described in sub-section "*Extraction of Pairwise Local Energies and Noise Reduction Technique*"). For training the ML models, a total of 87,117 pairwise interactions derived from MD simulations, each associated with its corresponding local energy value, were used as the training dataset. Once the models were trained, 18,971 additional MD-derived pairwise interactions were used to evaluate how accurately the ML-predicted local energies reproduced the original MD-derived values. These ML-derived local energies were then integrated using the local-to-global methodology to compute the model-specific PANTHER scores.

For the stress set, the procedure slightly differed. Instead of relying on expensive MD-derived time-averaged interaction energy approach data (described in sub-section "*Extraction of Pairwise Local Energies and Noise Reduction Technique*"), we used only the raw structural information from the PDB files as input. The input features consisted of the pairwise interaction descriptors, as illustrated in Table 3-10. The trained ML models were then used to predict the local interaction energies directly from these structural input features. The resulting local energies were subsequently aggregated through the local-to-global integration approach to generate the PANTHER scores for each model.

3.2.3. Molecular Dynamics Simulation

The curated 53 complexes subjected to MD simulations using the *Amber20* suite [215] with the AMBER14SB [216] (for proteins) and OL3 force-fields [217] (for RNA), the TIP3P water model [218], and embedding the system into a cubic box. The LEaP algorithm [215] was used to generate coordinate and topology files for AMBER. To neutralize the systems, Na^+ or Cl^- counterions were added as needed, ensuring that all ions remained in the solvent and at least 10 Å away from the protein-RNA binding interface. Hence, all complexes underwent an initial minimization, during which the hydrogen atoms were first minimized, then the whole complex was minimized, and finally the water molecules were minimized. Subsequently, each system were heated gradually from 0 K

to 300 K using a velocity-rescaling thermostat. Lastly, the equilibration phase was organized as follows: (i) each system underwent a 200 ps isobaric-isothermal ensemble (NPT ensemble) [219] at 1 bar of pressure, (ii) followed by 5 consequent 100 ps MD runs with the NPT ensemble to gradually adjust the volume of the system to attain water density of 1 gm/cm^3 within the water cubic box, and (iii) a short 100 ps MD run with the isothermal-isochoric ensemble (NVT ensemble) [220] was carried out to relax all the atoms before production. During the process, the bonds were kept fixed using the LINCS algorithm [221], thus allowing an integration step of 2 fs. The particle mesh Ewald method [152] was used to compute long-range interactions above a cutoff radius of 1.1 nm. The final step involved the production run of 500 ns with the NVT ensemble [220], generating the corresponding trajectories, in which one frame is saved every 50 ps for a total of 10,000 frame per trajectory.

The selection of AMBER14SB along with OL3 combination was made because it is a well-established and widely used setup for protein-RNA simulations, with OL3 being a standard AMBER RNA parametrization and ff14SB+OL3 already adopted in multiple published protein-RNA studies spanning viral nucleoprotein-RNA, splicing factor-RNA, zinc-finger/RNA, and RNA-editing complexes [222], [223].

3.2.4. Extraction of Pairwise Local Energies

The details of the Extraction Pairwise local energy is already explained in chapter 2 (*Methods and Models*, section: *Molecular-Dynamics-Derived Energy Decomposition*). In brief, we calculated pairwise energy between amino acids-nucleotides within the protein-RNA complexes using the MD-driven decomposition energy. Using the time-averaged interaction energy approach, we extracted a total of 331,744 amino acid–nucleotide pair interactions and their corresponding local energies from the MD trajectories of 46 protein-RNA complexes (training set 1 + training set 2). For the test set, an additional 58,000 pairwise interactions were obtained from the MD trajectories of 7 protein-RNA complexes. To minimize noise and ensure data quality, only those interactions that remained stable for at least 70% of the simulation time were retained, thereby focusing on persistent and meaningful amino acid-nucleotide contacts. After this refinement, the datasets consisted of 87,117 and 18,971 interaction pairs for the training and test sets, respectively, each associated with its corresponding MD-derived local energy. These refined, time-averaged interaction data were then used to train multiple ML models. Once trained, the ML models were applied to the test set to calculate local interaction energies to evaluate their predictive performances. The resulting 18,971

predicted local energy values were then grouped according to their corresponding PDB IDs. This process first enabled verification of the ML-derived data and subsequently allowed the application of the local-to-global methodology to integrate the predicted local energies, thereby generating model-specific PANTHER scores for each of the 7 protein-RNA complexes of the test set. This workflow ensured the inclusion of diverse yet reliable interaction patterns, while minimizing noise and avoiding oversampling from closely correlated trajectory frames.

In addition to the energetic averages described above, a set of structural descriptors was collected to be used as features to train the ML models. They included: (i) pairwise distances, quantified as the Euclidean distance between the centers of mass (Å) of the interacting amino acid and nucleotide, and (ii) various features for each pairwise interaction to better describe the relevant biophysical aspects of protein-RNA recognition and binding (as detailed in Table 3-10).

3.2.5 Development of Prediction Models

To predict protein-RNA local interaction energies, we explored a diverse range of regression models, starting with linear regression as a baseline and progressing to more complex approaches, including tree-based ensemble methods such as random forest regression (RFR) [224], boosting-based ensembles such as gradient boosting regression (GBR) [225] and extreme gradient boosting regression (XGBoost) [226], as well as a stacked ensemble model [227]. The stacked ensemble model combines the predictions of RFR, GBR, and XGBoost using a direct aggregation strategy. Additionally, a neural network model was implemented to further enhance predictive performance. These models were chosen to balance interpretability, computational efficiency, and predictive accuracy. A standardized ML pipeline was developed in Python 3.9 using *Scikit-Learn* 0.24.1 [228], *Keras* version 3.3.3, and *TensorFlow* version 2.16.1 libraries. Feature preprocessing was performed using Scikit-Learn's *ColumnTransformer*, ensuring consistency across all models. Numerical features (e.g., distance and number of hydrogen bonds) were standardized using *StandardScaler*, while categorical features (e.g., amino acid and nucleotide types) were one-hot encoded via *OneHotEncoder*. The dataset was initially partitioned into training (80%) and validation (20%) subsets using *train, test, and split*. To further evaluate model generalization, we applied a 10-fold cross-validation strategy [229] exclusively on the training set, where in each fold, 90% of the training data was used for model fitting and 10% for validation. Hyperparameter tuning was performed via grid search [230] over an extensive parameter space to optimize each model's configuration. Model performance was evaluated using Pearson's correlation coefficient (r), mean squared error (MSE), Spearman's rank correlation coefficient (r_s), and the statistical significance (p-value) [231]. Furthermore, both

learning and loss curves were monitored to ensure stable training and convergence. The neural network architecture was implemented using Keras and TensorFlow, incorporating ReLU activation functions, L2 regularization, and the Adam optimizer to enhance model robustness and generalization.

3.2.6 Local-to-Global Prediction

Our approach to predict protein-RNA binding affinities employs a local-to-global methodology, moving from local interaction energies to global energy predictions, which we call the PANTHER score. This process consists of several key steps that utilize molecular features to compute a comprehensive energy landscape. The workflow begins with identifying interacting amino acid-nucleotide pairs within a given protein-RNA complex using an in-house interaction pattern detection script, which analyzes the structural data to detect significant interactions based on distance thresholds and binding patterns. Specifically, interacting amino acid-nucleotide pairs are identified by considering pairs of amino acid residues and nucleotide bases characterized by a distance between their centers of mass (COM) within a cutoff threshold of 12 Å. For each identified interaction pair, we extract features including amino acid type, nucleotide base identity, center-of-mass distance, and number of H-bonds (Table 3-10). These features serve as input to ML models to predict the corresponding local pairwise interaction energies. The PANTHER score (expressed in kcal/mol) is then determined through an integrative approach that aggregates these local interaction energies according to equation 3.1.

$$\mathrm{PANTHER}_{\mathrm{score}} = \int_{\Omega} \boldsymbol{\rho}(\boldsymbol{r})\boldsymbol{E}(\boldsymbol{r})\, \boldsymbol{dr} \tag{3.1}$$

where $\rho(\mathbf{r})$ represents the spatial density of interaction pairs at position r, $E(\mathbf{r})$ is the local energy contribution at position r, and Ω is the binding interface (all areas where interactions occur). We sum up energy contributions from all interaction pairs while considering their distribution in space. This integral can be discretized for computational implementation for energy prediction as is shown in equation 3.2.

$$\mathrm{PANTHER}_{\mathrm{score}} = \sum_{i=1}^{n} \omega_i \, E_i \tag{3.2}$$

where ω_i represents a weighting factor that adjusts each interaction's contribution, and E_i is the predicted energy for the i^{th} amino acid-nucleotide pair. To ensure that closer interactions have greater contribution, ω_{i} is based on an exponential decay function in equation 3.3.

$$\omega_i = \frac{exp(^{-r_i}/_{r_0})}{\sum_{j=1}^{N} exp(^{-r_i}/_{r_0})} \tag{3.3}$$

where, r_i is the distance of the interaction pair, and r_0 is a characteristic length scale of 9 Å. The negative sign ensures that the function decreases as distance increases. By applying this weighting function, interactions occurring at shorter distances (particularly below 9 Å) contribute more significantly to the total binding energy. Finally, we refine our PANTHER score by using a distance-dependent function in equation 3.4.

$$\mathrm{PANTHER}_{\mathrm{score}} = \frac{\sum_{i=1}^{n} \alpha(r_i) E_i}{n} \tag{3.4}$$

Where $\alpha(r_i)$ is a distance-dependent weighting function that typically decreases with increasing distance, reflecting the diminishing contribution of more distant interactions to the binding free energy. The resulting PANTHER score, expressed in kcal/mol, offers an interpretable and accurate estimation of protein-RNA binding energetics, making it a valuable tool for understanding biomolecular recognition processes.

3.2.7 Large-scale Application

After completing the workflow to calculate the PANTHER score, we aimed to automate its application to the protein-RNA complexes of interest. To achieve this, we developed two in-house Python scripts. The former incorporates a parsing algorithm designed to read and extract data from PDB files. It also detects the hydrogen bonds by evaluating both the distance and angle to evaluate their stability. In detail, a hydrogen bond was defined based on two geometric criteria: the distance between the donor hydrogen atom and the acceptor atom had to be less than 3.5 Å, and the donor–hydrogen–acceptor angle was required to fall within the range of 120° to 180°. The position of the hydrogen atom was determined by the *Bond Vector Method*, which assumes that the hydrogen atom lies along the bond vector extending from the donor atom towards the acceptor atom at a typical bond length (approximately 1 Å for N-H or O-H bonds). The second script utilizes the input features derived from a raw PDB file (e.g., center of mass distance, amino acid and nucleotide base types, hydrogen bond count) as input for the ML model to predict pairwise local interaction energies without performing MD simulations. The predicted interaction terms are then aggregated to estimate the global PANTHER score for a given protein-RNA complex. This workflow was designed to enable efficient large-scale predictions and was applied to the stress set of 110 structures, achieving a processing speed of approximately 55 structures per minute on a server equipped with an 11th Generation Intel® Core™ i5-11400 CPU (2.60 GHz, 6 cores, 12 threads) and 8 GB RAM. Performance may vary depending on the hardware configuration. To evaluate the performance of our models, we conducted a comparative analysis with the web-based tools PredPRBA [205] and

PRA-Pred [202] on the stress set consisting of 110 protein-RNA complexes. In detail, we compared the linear correlation between the experimental and predicted ΔG values for the three methods.

3.3. Results and Discussion

3.3.1 Evaluation of Local-to-Global Scoring Methodology

The proposed method is based on the reliability of the local-to-global approach, which implies an additive nature of the ΔG values where an interaction score (kcal/mol) can be computed by integrating the contributions of the local interaction energies of amino acid-nucleotide bases, that here we refer to as MD-derived local-to-global score. These local contributions can be obtained either from MD simulations or predicted by ML models. Hence, the preliminary but crucial step of this study is assessing the reliability of this fragmental approach, at least when integrating the local energies as derived from MD runs. Since the MD-derived local energies will later serve as training data for various ML models, it is essential to perform an initial evaluation to verify that the local-to-global methodology exhibits a robust correlation with experimental data. Accordingly, out of the 130 complexes with experimental ΔG, initially we kept away 110 complexes for the stress set and out of remaining 20 complexes, 13 complexes were considered for training set 1 and 7 complexes for test set (Figure 3-2, Tables 3-1 & 3-2). The MD simulation for both training set 1 and test set were carried out and the local interaction energy between amino-acid and nucleotide for each complex was extracted. Subsequently, the local interactions energies were integrated to the local-to-global score (methodology described in sub-section "*Local-to-Global Integration*") and the local-to-global scores were correlated with the experimental ΔG values of the simulated complexes. The complete workflow is shown in Figure 3-3.

The obtained results are shown in Figure 3-4, which plots MD-derived local-to-global score versus the experimental ΔG and reveals an encouraging correlation coefficient (r) of 0.60 obtained from 20 protein-RNA complexes. Furthermore, the correlation plots for training set 1 (r = 0.75) and test set (r = 0.59) are shown separately in Figure 3-5 and 3-6, respectively, providing supporting evidence that the local-to-global scoring system is reliable. As shown in Figure 3-4, 3-5 and 6, the slopes of the regression lines deviate significantly from the ideal 45° reference line, indicating that, although strong correlations are observed, the predicted ΔG values do not perfectly match experimental values. However, the primary goal of this study is to develop a scoring function that exhibits a consistent correlation with experimental binding affinities, thereby enabling the relative

comparison of BFEs across different RNA-protein complexes rather than the exact prediction of absolute ΔG values.

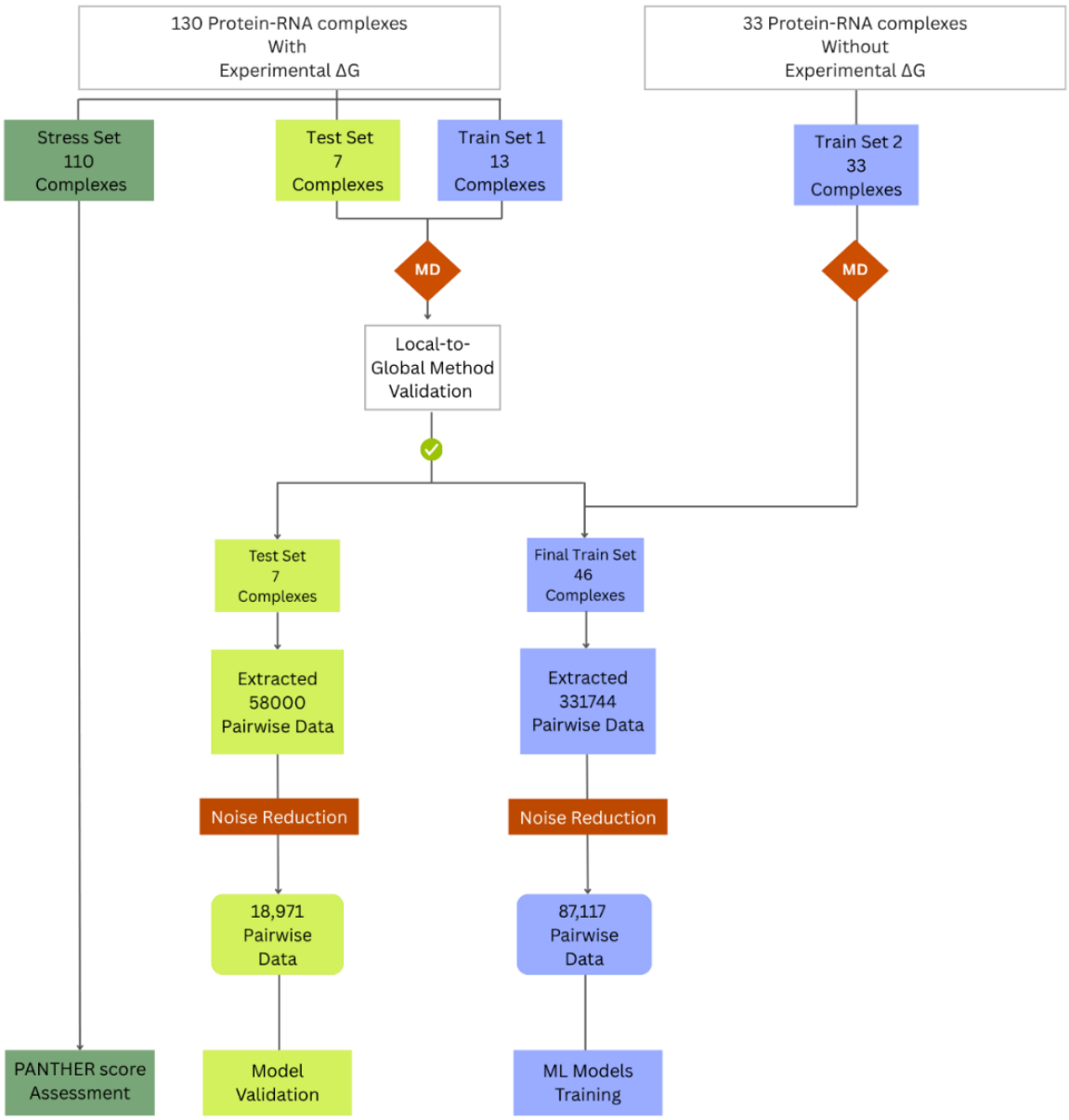

Figure 3-2. Schematic representation of the dataset preparation and validation workflow. Protein-RNA complexes with and without experimental ΔG values were subjected to MD simulations, followed by the local-to-global energy strategy to compute the MD-derived PANTHER scores. The validated scores were used to construct the final training and test sets for ML model development and validation.

Even though the primary objective of this study is the development of an innovative protein-RNA binding affinity score that does not require MD simulations, the preliminary results support the additive nature of binding energy. This finding indicates that global binding affinities can be

estimated by integrating the contribution of the local interactions and suggests that such a local-to-global approach can be clearly exploited when using MD-derived local energies to obtain reliable and interpretable predictions of MD-derived PANTHER score. This data-driven approach has the potential to substantially reduce computational time while maintaining satisfactory level of predictive accuracy. Together, this two-tiered strategy, starting with physics-based calculations and moving toward ML-based predictions, sets up a pipeline for a scalable and efficient framework to estimate binding affinities in protein-RNA systems.

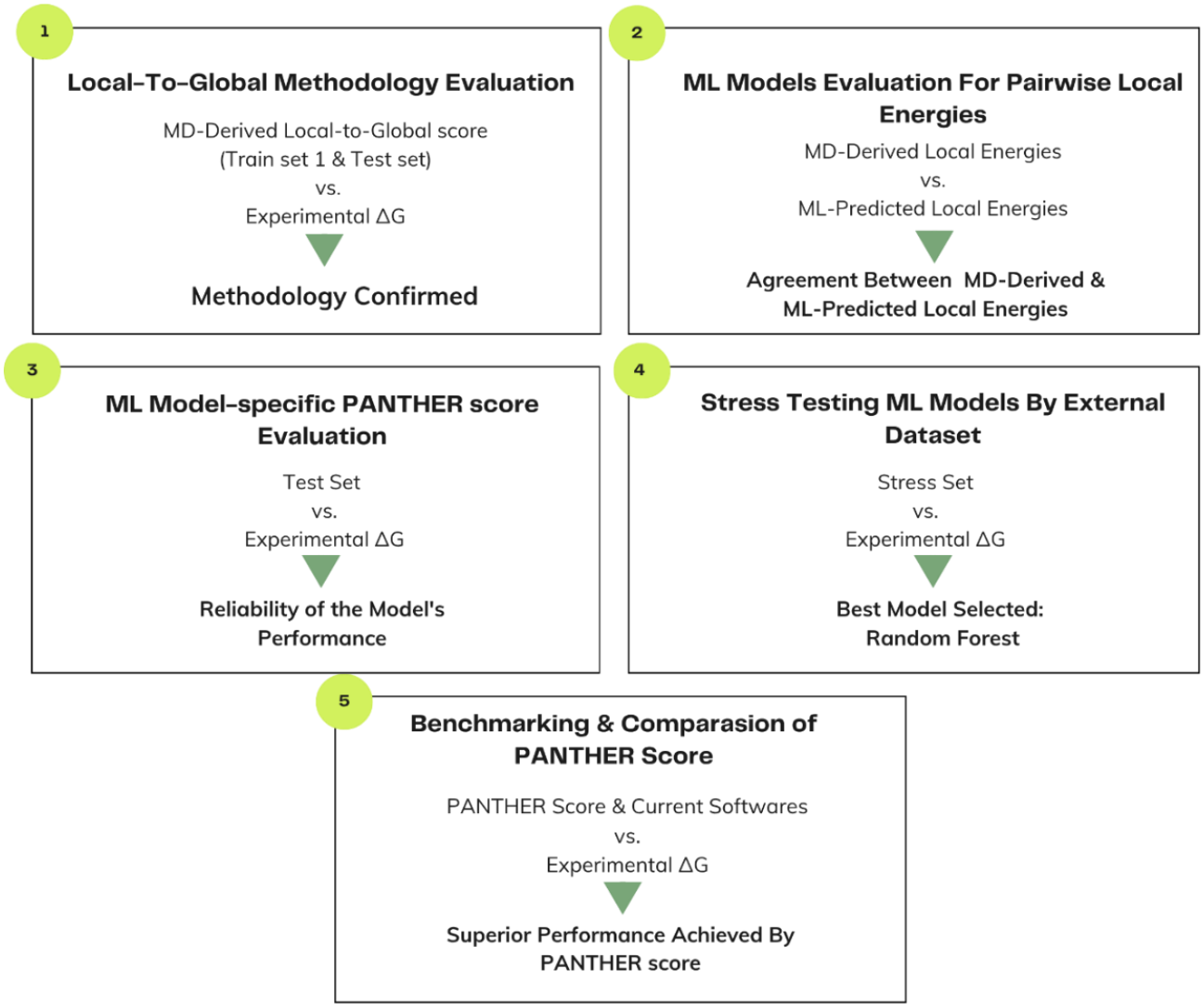

Figure 3-3. Workflow followed for developing the PANTHER score.

After validating the reliability of the approach, we expanded the training set by including 33 additional protein-RNA complexes as training set 2 (Figure 3-2). At this stage, we included complexes

without experimental ΔG values, as method's performance and reliability had already been validated. We performed MD simulation for all the three sets (training set 1, 2 and test set). From the MD simulations, we extracted local interaction energies for each amino acid-nucleotide pair across all protein-RNA trajectories. This process yielded 87,117 noise free data points for the combined training set (set 1 + set 2) and 18,971 noise free data points for the independent test set (Figure 3-2). The methodology used for noise reduction is described in detail in the Method sub-section “Extraction of Pairwise Local Energies and Noise Reduction Technique”.

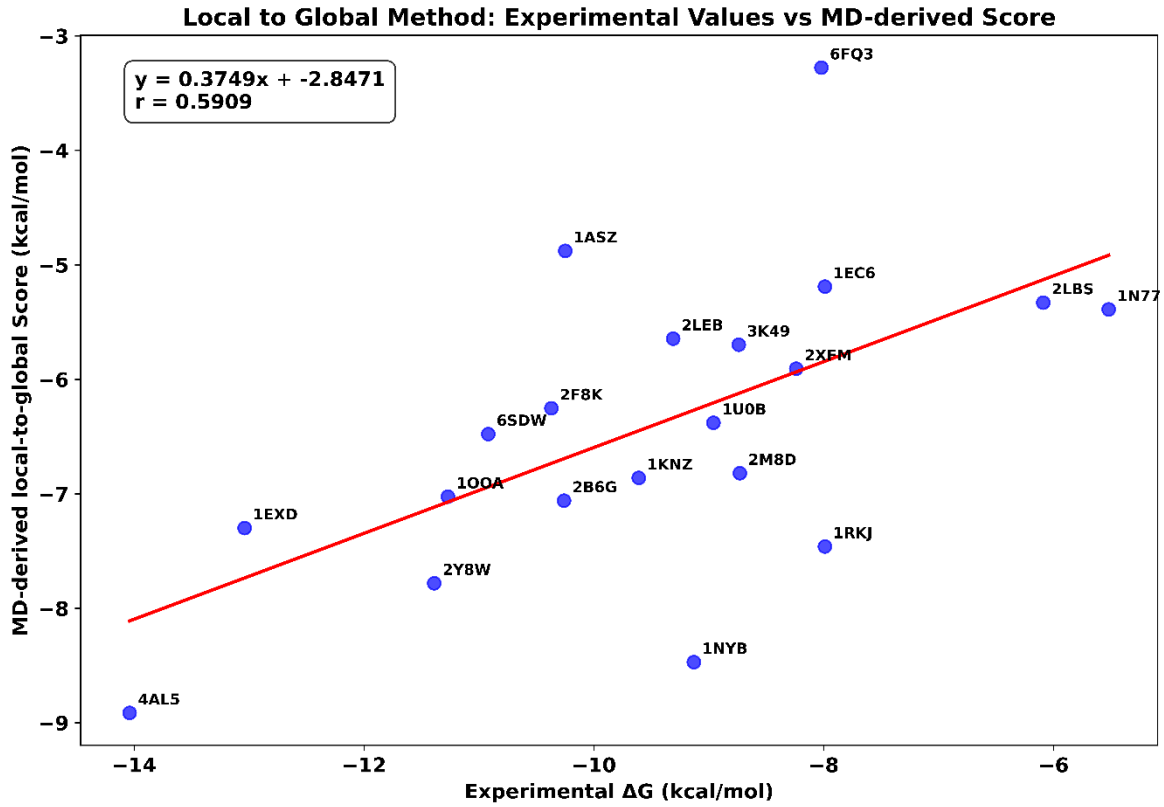

Figure 3-4. Correlation between the MD-based scores and experimental ΔG values for 20 protein-RNA complexes (13 training set + 7 test set). The Pearson correlation coefficient (r) of 0.6 confirms the reliability of fragmental approach.

In the present work, however, MD was used primarily to generate a large and diverse training set of local amino acid–nucleotide interaction energies for the subsequent machine-learning stage, rather than to derive final free energies for each system. For this reason, we prioritized coverage of many different protein-RNA complexes (53 systems, 500 ns each) over replicating trajectories of fewer complexes. Moreover, we applied several procedures to reduce the dependence on transient fluctuations and initial conditions: interaction energies were averaged

over consecutive frames, sampled with frame skipping to reduce temporal correlation, and only interactions persisting for at least 70% of the trajectory were retained. Finally, the resulting model was not judged only on the MD-derived training data, but was validated on an independent test set and, more importantly, on an external stress set of 110 untreated complexes. Together, these steps provide an orthogonal assessment of robustness that partly compensates for the absence of replica trajectories.

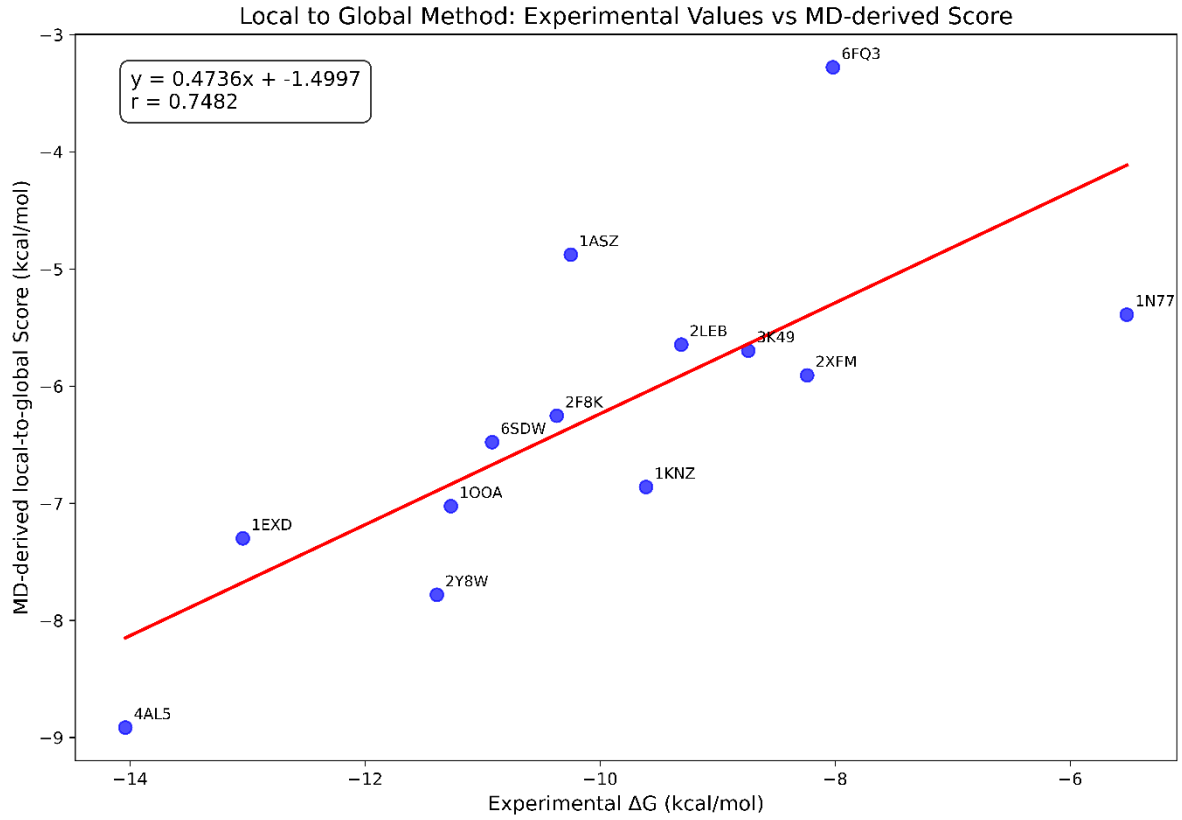

Figure 3-5. Correlation between the MD-based score and experimental ΔG values for 13 protein-RNA complexes of training set. The Pearson correlation coefficient (r) of 0.75 confirms the reliability of the fragmental approach.

3.3.2 ML Models & Performance on Test Set

3.3.2.1 Datasets Preparation

The performance of any ML model is intrinsically linked to the quality of the dataset used for training. To ensure robust predictions, a comprehensive structural analysis was performed for each protein-RNA complex, validating the residues and bonds in accordance with procedures described in the Methods section (see *Datasets Selection and Preparation Section*). We then performed MD simulations on 53 complexes (including training set 1, training set 2 and test set), from which we extracted MD-derived local interacting energies between each interacting amino acid and

nucleotide base (Figure 3-2). From this initial pool of 53 complexes, 7 structures were designated as the test set, thereby defining curated training (Table 3-1) and test datasets (Table 3-2).

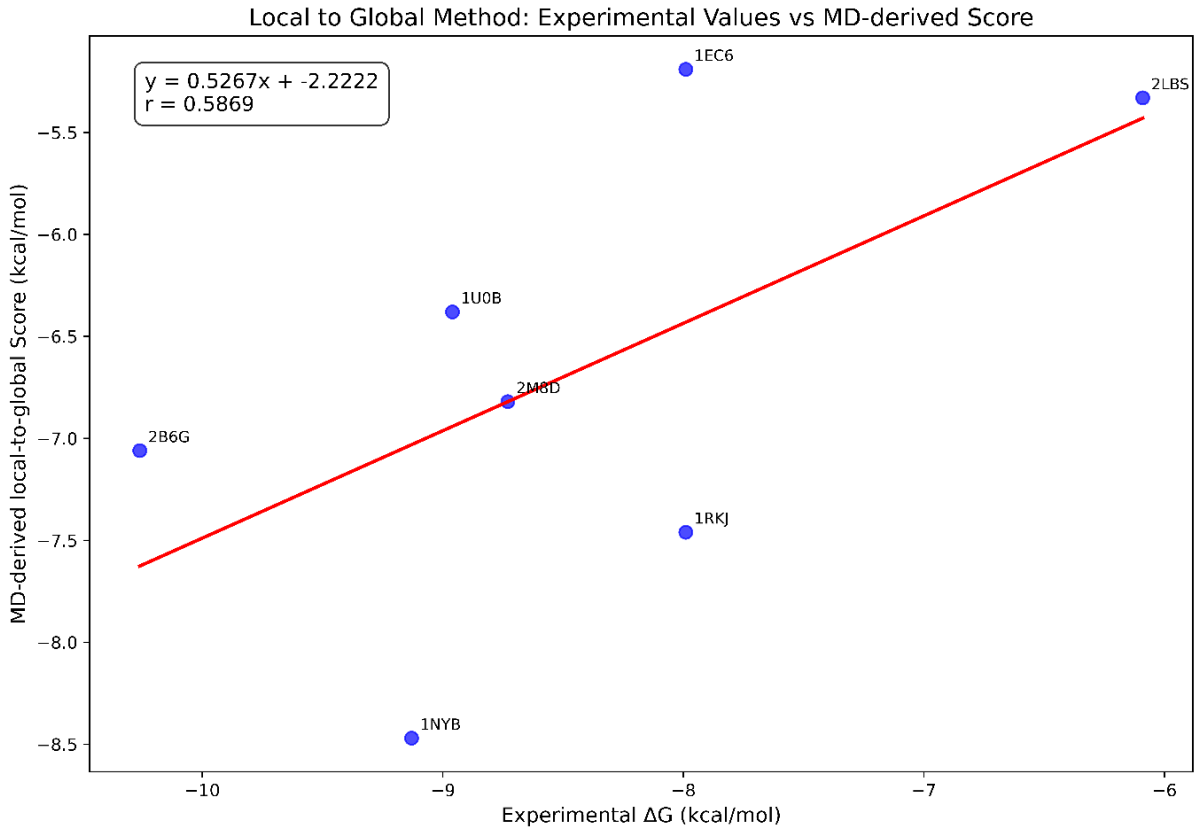

Figure 3-6. Correlation between the MD-based score and experimental ΔG values for 7 protein-RNA complexes of test set. The Pearson correlation coefficient (r) of 0.59 confirms the reliability of the fragmental approach.

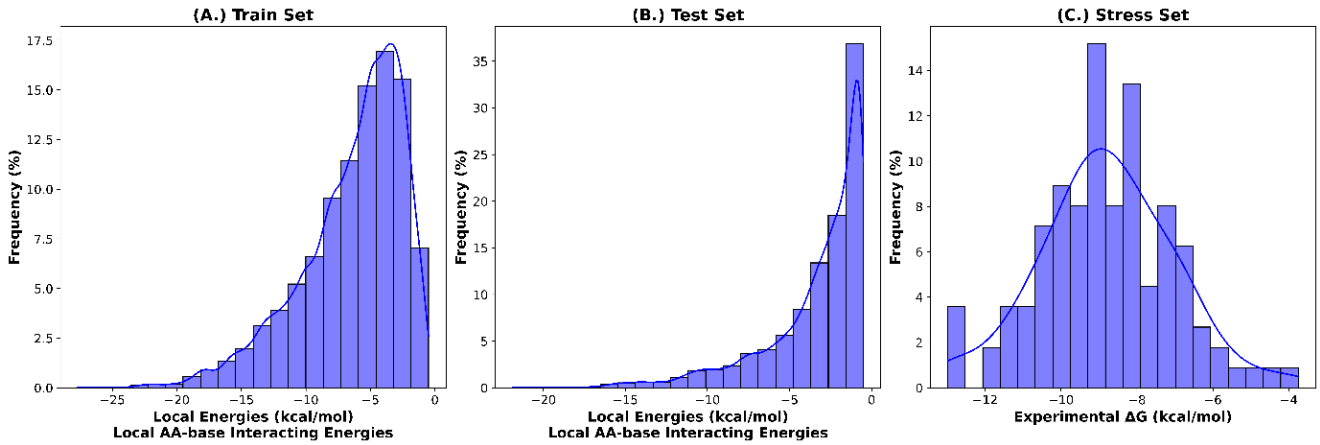

Figure 3-7. Distribution of local MD derived energies of the interacting amino acid-nucleotide bases for the (A) training set and (B) test set, alongside the experimental free energies for the stress set (C). The distributions demonstrate consistent energetic patterns across all datasets.

Table 3-1. Training Set and Experimental ΔG, Data not Available (DNoA)

Training Set	MD-derived local-to-global score (kcal/mol)	Experimental ΔG (kcal/mol)
1ASZ	-4.88	-10.25
1EXD	-7.30	-13.04
1KNZ	-6.86	-9.61
1JBT	-	DNoA[1]
1KUQ	-	DNoA
1N77	-5.39	-5.52
1OOA	-7.02	-11.27
1SJ4	-	DNoA
1ZL3	-	DNoA
2AB4	-	DNoA
2ASB	-	DNoA
2ATW	-	DNoA
2JLW	-	DNoA
2LEB	-5.65	-9.31
2UMW	-	DNoA
2XFM	-5.91	-8.24
2XLK	-	DNoA
2Y8W	-7.78	-11.39
2F8K	-6.25	-10.37
2ZUE	-	DNoA
3K49	-5.70	-8.74
3T5N	-	DNoA
3UMY	-	DNoA
4AL5	-8.91	-14.04
4M6D	-	DNoA
4ZLD	-	DNoA
5GJB	-	DNoA
5HO4	-	DNoA
5MFX	-	DNoA
5WT1	-	DNoA
6DU4	-	DNoA
6F4H	-	DNoA
6FQ3	3.28	-8.02

[1] DNoA: Data Not Available

Training Set	MD-derived local-to-global score (kcal/mol)	Experimental ΔG (kcal/mol)
6IV9	-	DNoA
6KYV	-	DNoA
6LAX	-	DNoA
6SDW	-6.48	-10.92
6ZDQ	-	DNoA
7A9W	-	DNoA
7EIU	-	DNoA
7OZQ	-	DNoA
7P0V	-	DNoA
7v2Z	-	DNoA
7VKL	-	DNoA
7XJZ	-	DNoA
7ZEX	-	DNoA

Furthermore, as detailed in the Methods sub-section “Development of Prediction Models”, we trained ML models to predict the local energies derived from MD trajectories using simple descriptors, such as amino acid residues, nucleotide bases, local interaction energies, the number of hydrogen bonds formed, and interatomic distances. To optimize predictive performances, we explored a broad range of regression algorithms, starting with linear regression as a baseline and progressing to more advanced algorithms capable of handling complex data. These included RFR, Extreme Gradient Boosting (XGBoost), Gradient Boosting (GBoost), Stacking Regression, and Neural Networks. Moreover, we analyzed the diversity of local interaction energies as input data in the training set, which is illustrated in Figure 3-7 A. This figure shows a wide distribution of interaction strengths, ranging from weak to strong. To ensure that the ML models can accurately predict local energies across diverse interaction strengths, we evaluated them on a test set containing MD-derived local energies with an equally broad distribution (Figure 3-7 B). This wide range challenges the models to generalize across both weak and strong local interactions. The similarity between training and test set distributions (Figures 3-7 A-B) indicates that model evaluation is unbiased. In contrast, Figure 3-7 C reports the distribution of the experimental ΔG values for the 110 complexes, which exhibits a symmetric, approximately Gaussian distribution with fewer extreme values. In this study, these experimental binding affinities serve as an important validation benchmark for overall model performance and to guide the selection process of the optimal ML model for PANTHER score prediction. This validation step is essential to confirm that integration of local energetic

contributions accurately reproduces a binding affinity score, which exhibits a strong correlation with the experimental ΔG values. The predictive performance of the models was evaluated in two stages: first, using the test set (Table 3-4, see the "Model Evaluation" sub-section), and then using the larger stress dataset (see the "Scoring Function Assessment" sub-section).

3.3.2.2 Agreement between MD- and ML-Derived Local Energies

Having established the distribution of MD-derived local energies used as input, the next step is to determine whether ML models trained on MD-derived descriptors can correctly predict the local energy values. Only if the ML models can successfully recover the MD signal at the local level, the integration to perform local-to-global score will be reliable. Such an agreement was quantified using Pearson correlation coefficients (*r*) calculated between MD-derived and ML-predicted local energies, computed for each complex and summarized across models (Table 3-2). While showing expected variability, the analysis of the individual model performances (see Table 3-2) reveals overall satisfactory results with (r) almost always exceeding 0.5. Interestingly, the (r) mean value of each protein unravels a marked variability, suggesting that some complexes perform clearly better. These differences can be ascribed to the structure complexity, which is, in turn, related to the number of local energies to be predicted. Taken together, the high average correlation values across diverse protein systems and machine learning approaches (overall (r) mean = 0.72) indicate that the developed ML models can effectively capture the patterns in local interaction energies originally derived from computationally expensive MD simulations.

Considering only one exemplificative case, Figure 3-8 shows a detailed visualization of the correlations between MD-derived and ML-derived local energy values for the 1EC6 protein system as obtained by six different ML approaches. The scatter plots demonstrate that all models capture the general trends in local energies, with points colored according to their distance from the trend line to highlight prediction accuracy. From Table 3-2, we observe that for the 1EC6 complex, the ML-predicted local energies achieved (r) values of 0.80, 0.82, 0.84, 0.82, 0.84 and 0.76 for RFR, GBoost, Neural Network, Stacked Ensemble, XGBoost, and Linear Regression respectively. Overall, when the correlations are calculated for all the 7 PDB IDs of the test set, their results are encouraging, with mean (r) typically ≥ 0.7, for most complexes. Therefore, the performance demonstrated by the ML models provides compelling evidence that machine learning can serve as a viable alternative to MD simulations for predicting the interacting local energies, provided that the training dataset is of high quality and carefully curated. All ML approaches tested show good correlations with MD data, confirming that the local energies of protein-RNA complexes can be successfully calculated by using

these ML models. This finding is particularly significant given the computational resources typically required for MD simulations, while ML models might yield comparable results at a fraction of the computational cost and time. As the ML models can reliably reproduce the MD-derived local energies, we next assess how these models perform by comparing their predictions, integrated as model-specific PANTHER scores, with the corresponding experimental ΔG values in the test set.

Table 3-2. Pearson correlation coefficient (r) between raw MD-derived predicted local energies and raw ML-derived predicted local energies for various models for each PDB IDs considered in the test set. Refer to Figure 3-3 for an example of 1EC6. This correlation analysis was conducted to evaluate how accurately the ML models predict the MD-derived local energies. The data from the test set used to predict the local energies through ML models are derived from the MD simulations.

Test Set	(r) Random Forest prediction vs MD-derived local energies	(r) GBoost prediction vs MD-derived local energies	(r) Neural Network prediction vs MD-derived local energies	(r) Stacked Ensemble prediction vs MD-derived local energies	(r) XGBoost prediction vs MD-derived local energies	(r) Linear Regression prediction vs MD-derived local energies	Mean (r)
1U0B	0.74	0.76	0.78	0.75	0.77	0.72	0.75
1RKJ	0.65	0.67	0.68	0.67	0.68	0.67	0.67
1NYB	0.78	0.79	0.83	0.79	0.80	0.76	0.79
1EC6	0.80	0.82	0.84	0.82	0.84	0.76	0.81
2B6G	0.63	0.64	0.74	0.65	0.67	0.77	0.68
2LBS	0.46	0.50	0.68	0.57	0.60	0.60	0.57
2M8D	0.74	0.73	0.81	0.74	0.79	0.77	0.76
Mean	0.69	0.70	0.77	0.71	0.74	0.72	0.72

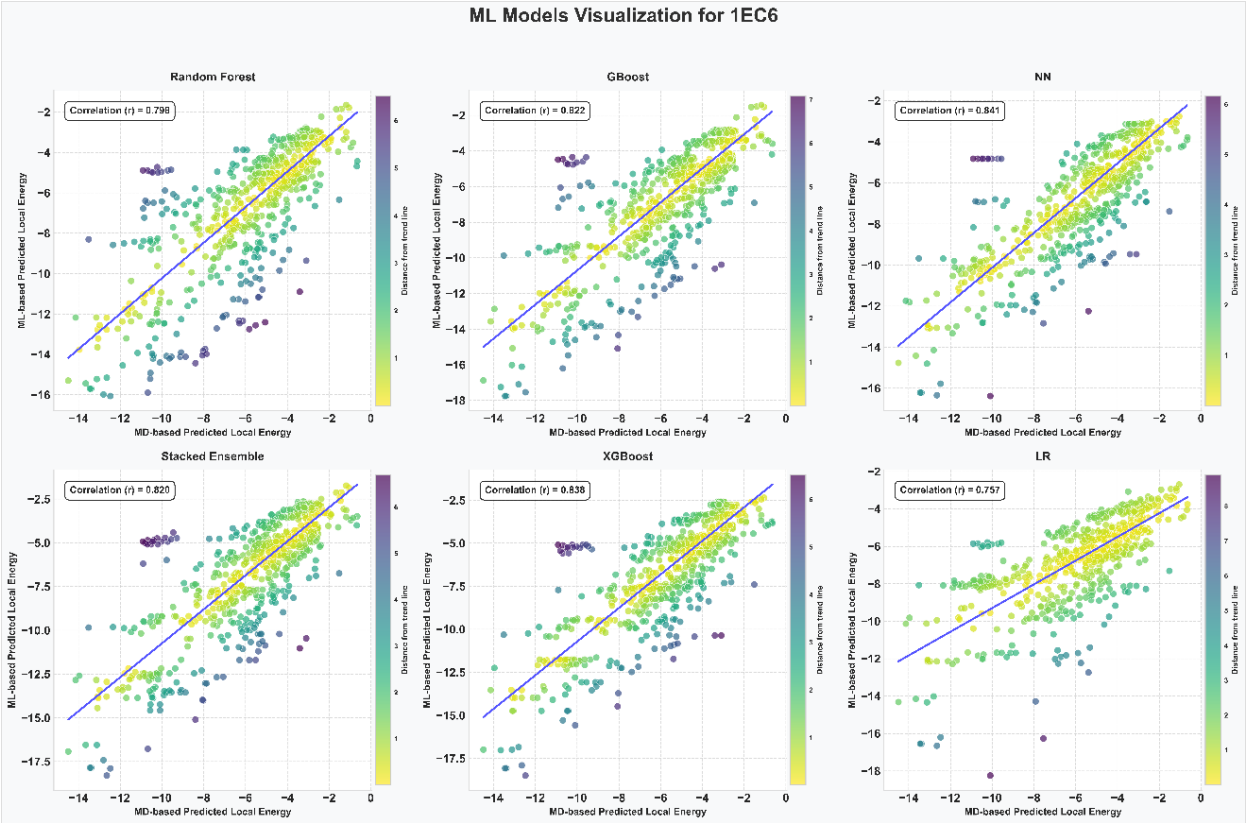

Figure 3-8. Comparison of MD-based and ML-based predicted local energy across six ML models. Points are colored by their distance from the trend line. Similar colors indicate proximity to the correlation line. The correlation coefficient (r) quantifies the agreement between the two prediction methods for each model.

3.3.3 Model Evaluation

As described in the Methods section (see "*PANTHER score*" and "*Local-to-Global Integration*" sections), the PANTHER score represents a descriptor of binding affinity for protein-RNA complexes. This score is derived by integrating ML-predicted local interactions energies, specifically those involving amino acid-nucleotide base pairs identified within a range of 9 Å distance from their center of mass. This integration of these local energetic contributions provides a measure of each complex's strength of binding affinity. To ensure methodological accuracy, each ML model, which includes RFR, Gradient Boosting, Neural Network, Stacked Ensemble, XGBoost, and Linear Regression, were trained independently to predict local energies. As a result, each model produces its own prediction of the overall PANTHER score for a given complex. Throughout this framework, these results are referred to as model-specific PANTHER scores until a particular ML model is finally selected as the optimal solution to predict the PANTHER score (see Figure 3-3 for the workflow).

However, the predicted accuracy of the approach is assessed by comparing these model-specific PANTHER scores to the experimental ΔG values. This comparison enables evaluation of each model's ability to predict binding affinities and guides the selection process for the optimal ML methodology within the framework. As shown in Figure 3-9, the x-axis represents the PDB IDs of the test set complexes, while the y-axis shows the corresponding experimental ΔG values (grey line) and the PANTHER scores predicted by the different ML models. The experimental ΔG values vary substantially (from approximately −6 to −12 kcal/mol), reflecting diverse binding affinities among the complexes. All ML-derived PANTHER predictions closely follow the experimental trend, exhibiting high consistency across models. The correlation coefficients ($r > 0.7$; see Table 3-4) indicate good predictive performance, with only minor differences among models ($\Delta r < 0.15$). The RFR and Gradient Boosting models achieve the highest correlations ($r \approx 0.8$), although the performance of the other ML models was comparable. To further evaluate model robustness, additional statistical metrics were calculated, including the Spearman correlation coefficient (r_s), Mean Absolute Error (MAE), and ρ-values to evaluate statistical significance (data shown in Table 3-3).

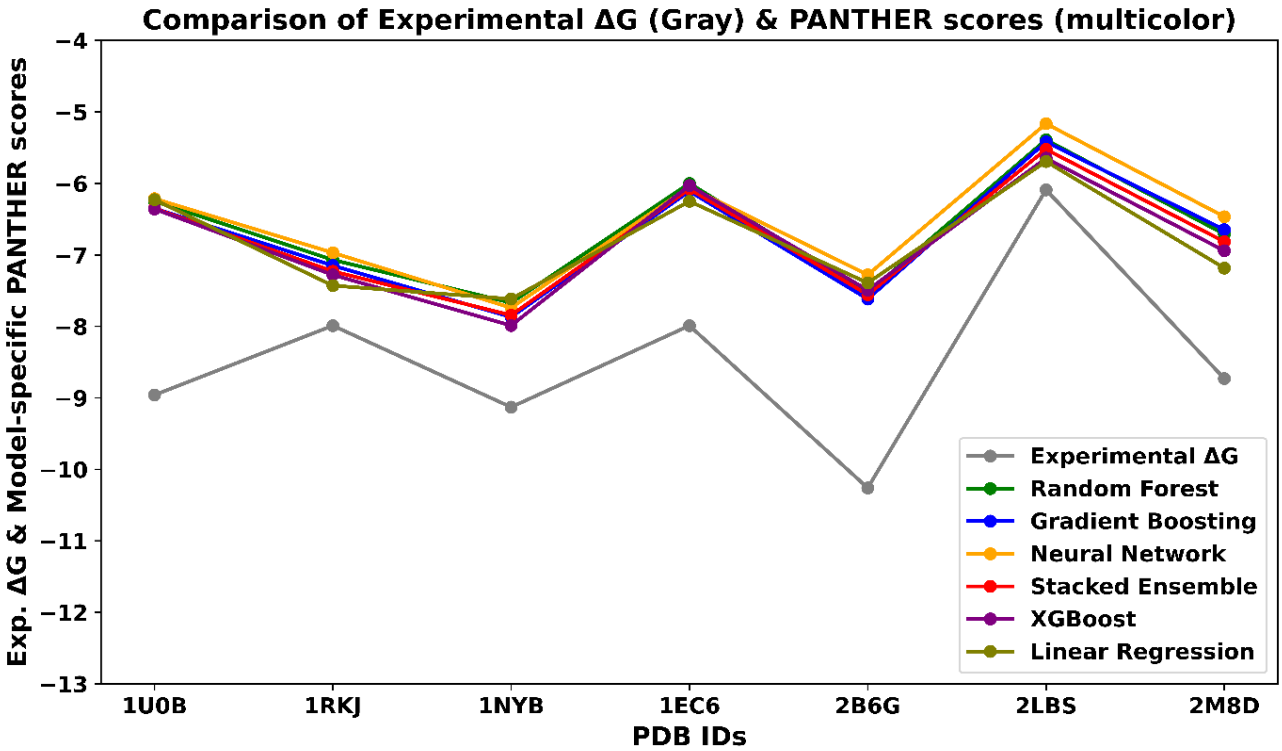

Figure 3-9. Comparison between PANTHER scores and experimental binding free energy (ΔG) values for the seven complexes of the test set, obtained using various machine learning approaches.

Table 3-3. Experimental ΔG (kcal/mol) and predicted model-specific PANTHER score (kcal/mol) for the test set, obtained using various ML models. The data from the test set used to predict the local energies through ML models are derived from the MD simulations.

Test Set	Exp. ΔG (kcal/mol)	MD-derived local-to-global Score (kcal/mol)	Model-specific Random Forest PANTHER score (kcal/mol)	Model-specific GBoost PANTHER score (kcal/mol)	Model-specific Neural Network PANTHER score (kcal/mol)	Model-specific Stacked Ensemble PANTHER score (kcal/mol)	Model-specific XGBoost PANTHER score (kcal/mol)	Model-specific Linear Regression PANTHER score (kcal/mol)	PredPRBA (kcal/mol)	PRA-Pred (kcal/mol)
1U0B	-8.96	-6.38	-6.25	-6.35	-6.32	-6.35	-6.36	-6.23	-13.92	-11.07
1RKJ	-7.99	-7.46	-7.07	-7.15	-7.33	-7.23	-7.28	-7.43	-11.17	-10.48
1NYB	-9.13	-8.47	-7.67	-7.87	-8.05	-7.84	-7.99	-7.62	-10.31	-9.58
1EC6	-7.99	-5.19	-6.00	-6.11	-6.06	-6.08	-6.03	-6.25	-8.68	-8.48
2B6G	-10.26	-7.06	-7.53	-7.62	-7.38	-7.56	-7.49	-7.39	-10.85	-8.37
2LBS	-6.09	-5.33	-5.39	-5.42	-5.70	-5.52	-5.65	-5.69	-12.10	-8.98
2M8D	-8.73	-6.82	-6.72	-6.65	-7.17	-6.82	-6.94	-7.18	-9.14	-7.17
(r)		0.59	0.80	0.79	0.68	0.77	0.72	0.69	-0.10	-0.06

Table 3-4. Performance metrics of various machine learning models on the test set for PANTHER score predictions. Metrics include correlation coefficient (r), mean absolute error (MAE), Spearman's rank correlation coefficient (rs), and p-value.

	Test Set			
Model	**r^2**	**MAE[3] kcal/mol**	**[4]r_s**	**ρ-value[5]**
Random Forest Regression	0.80	1.79	0.77	0.04
GBoosting Regression	0.79	1.71	0.77	0.04
Neural Network	0.68	1.59	0.77	0.04
Stacked Ensemble	0.77	1.68	0.77	0.04
XGBoosting Regression	0.72	1.63	0.77	0.04
Linear Regression	0.69	1.62	0.52	0.22

[2] r: Pearson Correlation
[3] MAE: MAE: Mean Absolute Error
[4] r_s: Spearman Correlation
[5] ρ value: Spearman's Rho

Figure 3-10 and Table 3-4, we summarize the obtained results (linear correlation between model-specific PANTHER scores vs Experimental ΔG, data shown Table 3-3) and further highlight the similarity in the performances reached by the tested ML approaches through the test set. In more detail, with these additional parameters we confirm few points: (1) the nearly identical and rather satisfactory performances of both RFR and GBoost methods with (r) of 0.80 and 0.79, with ρ-value of 0.04 respectively; (2) the notably encouraging performances by neural networks showing the lowest MAE value of 1.59 kcal/mol with the lowest (r) of 0.68; and (3) the relatively poor performances of linear regression as evidenced by its (r) of 0.69, r_s of 0.52 and ρ-value of 0.22. Overall, the analysis of mean absolute errors indicates low and comparable values across models, effectively ruling out markedly inaccurate predictions. It should be noted, however, that MAE values have limited relevance since our proposed methodology was developed to reach a convenient relationship between PANTHER score and experimental ΔG values which does not necessarily imply a numerical agreement between them. In other words, this study is not focused on predicting the absolute ΔG values; rather, we introduce a new fragmental methodology (local-to-global integration) to calculate a score (in kcal/mol) which satisfactorily aligns with the experimental values. Interestingly, the spearman correlation for all the models except for Linear Regression (lowest R_s = 0.52) has high correlation of ranks with the value 0.77. This shows that almost all the models are consistent in re-producing promising local energy values. Taken together, the model-specific PANTHER scores and statistical analyses (ρ-value < 0.04) demonstrate that the developed ML model performances are reliable; however, the optimal ML model for final implementation into the PANTHER scoring system remains to be identified. The next section focuses on the ML model selection process by considering the performance on stress set.

Figure 3-10. Performance comparison of machine learning models based on Pearson correlation (r), Spearman correlation (r_s), and Mean Absolute Error (MAE) for the test set. Random Forest demonstrates the highest predictive accuracy, while Linear Regression shows the lowest performance.

3.3.4. ML model Assessment

To better evaluate the obtained models and considering the very similar performances reached when analyzing the test set, we collected an extended set of 110 protein-RNA complexes that meet the predefined criteria (detailed in Methods sub-section "*Data Selection and Preparation*"). These complexes were utilized to calculate the corresponding PANTHER scores, which were compared with the respective experimental ΔG values. Notably, unlike the training and test datasets, no structural curation or molecular-dynamics (MD) simulations were performed on the complexes included in the additional dataset, referred to as the stress set (Table 3-5). This stress set was specifically designed to challenge our models and facilitate an unbiased evaluation under realistic conditions, ensuring that our assessment reflects true predictive performances in practical applications. This challenging stress test revealed clear performance differences among models, allowing the identification and selection of the best-performing one. Additionally, testing on untreated structure allows evaluation of the dependence of the computed score on structure

preparation and provides a clear assessment of the score's transferability to novel protein-RNA systems.

Table 3-5. Experimental ΔG (kcal/mol) and Predicted model-specific PANTHER score (kcal/mol) from various models, and Data not Available (DNoA) for the stress data set.

Stress Data Set	Exp. ΔG kcal/mol	Random Forest kcal/mol	GBoost kcal/mol	Neural Network kcal/mol	Stacked Ensemble kcal/mol	XGBoost kcal/mol	Linear Regression kcal/mol	PredPRBA kcal/mol	PRA-Pred kcal/mol
1C9S	-12.91	-10.10	-9.44	-24.10	-12.34	-11.03	-52.35	-14.83	DNoA[6]
1CVJ	-10.92	-10.34	-9.88	-19.39	-12.93	-11.77	-34.63	-8.21	-8.74
1IL2	-7.53	-6.13	-5.80	-6.74	-6.06	-6.14	-7.03	DNoA	-8.77
1JBS	-8.18	-8.44	-7.86	-9.42	-8.48	-8.48	-10.67	-11.63	-8.34
1K1G	-8.18	-7.05	-6.82	-8.72	-7.58	-7.23	-9.69	-10.27	-9.33
1K8W	-8.73	-6.06	-5.88	-6.36	-5.98	-6.01	-6.84	-8.75	-9.66
1KQ2	-9.89	-8.58	-8.49	-13.55	-9.98	-9.86	-17.93	-9.86	-10.71
1MMS	-9.44	-6.70	-6.38	-8.37	-7.00	-6.96	-9.69	-13.45	-10.27
1QTQ	-9.13	-6.66	-6.51	-6.94	-6.59	-6.67	-7.03	-14.30	-11.44
1RLG	-7.99	-7.35	-6.90	-10.14	-7.85	-7.67	-12.82	-10.42	-9.22
1SDS	-9.03	-7.68	-6.55	-11.48	-8.38	-8.07	-14.19	-11.78	-9.23
1UTD	-11.81	-9.60	-9.34	-17.42	-11.37	-10.51	-32.24	DNoA	DNoA
1WSU	-12.97	-10.84	-9.99	-13.57	-11.40	-10.68	-16.45	-7.94	-10.87
1WWD	-8.18	-7.87	-6.54	-9.34	-7.60	-7.32	-10.49	-7.81	-6.36
1YTU	-6.87	-8.06	-7.38	-9.41	-8.23	-8.50	-9.81	-7.96	-12.37
1YTY	-9.16	-6.61	-6.27	-8.39	-6.85	-6.79	-9.46	-8.96	-8.68
2A8V	-7.22	-7.30	-6.83	-10.73	-6.69	-6.69	-11.87	-9.08	-8.09
2C4R	-9.50	-6.67	-6.29	-8.19	-6.52	-6.67	-8.40	-10.68	-10.55
2DRB	-9.78	-5.64	-5.48	-5.80	-5.35	-5.56	-6.13	-12.30	-9.73
2I91	-8.86	-7.91	-7.29	-8.90	-7.92	-7.91	-10.60	-11.06	-12.31
2IX1	-10.48	-6.27	-5.87	-6.87	-6.34	-6.26	-7.14	-8.41	-9.95
2JEA	-7.98	-6.52	-6.23	-8.95	-6.81	-6.90	-8.99	-9.81	-10.75
2KX5	-11.41	-9.78	-9.54	-11.21	-9.90	-10.14	-11.22	-11.66	-7.74
2L41	-4.25	-6.79	-5.97	-8.89	-7.19	-6.87	-9.85	-8.82	-8.89
2LA5	-11.48	-10.19	-9.15	-11.05	-10.03	-10.18	-11.31	-13.45	-7.65
2LEC	-9.44	-8.19	-7.36	-10.38	-8.23	-7.93	-10.81	-7.96	-6.80
2MB0	-6.80	-6.91	-6.46	-7.67	-7.07	-7.04	-7.99	-7.81	-6.83
2PJP	-12.98	-8.63	-8.64	-8.81	-8.54	-8.38	-8.34	-12.35	-10.59
2UWM	-7.09	-7.91	-7.86	-9.34	-8.32	-8.40	-10.33	-12.16	-10.43

[6] DnoA: Data Not Avaialble

Stress Data Set	Exp. ΔG kcal/mol	Random Forest kcal/mol	GBoost kcal/mol	Neural Network kcal/mol	Stacked Ensemble kcal/mol	XGBoost kcal/mol	Linear Regression kcal/mol	PredPRBA kcal/mol	PRA-Pred kcal/mol
2X1A	-7.13	-6.75	-7.51	-8.72	-7.06	-7.38	-9.04	-5.87	-7.24
2XC7	-7.07	-7.06	-7.55	-8.09	-7.56	-7.76	-8.37	-8.05	-8.55
2XGJ	-7.70	-9.00	-9.08	-10.16	-9.91	-9.81	-10.98	-10.09	-13.25
2XNR	-5.40	-5.27	-5.05	-6.40	-5.50	-5.43	-6.81	-9.25	-7.42
2XS2	-10.11	-7.42	-7.11	-9.61	-7.57	-7.40	-10.60	-8.92	-8.18
2ZZN	-9.24	-7.34	-6.94	-8.17	-7.29	-7.20	-9.04	DNoA	-9.48
3ADB	-9.81	-6.14	-5.77	-6.88	-5.86	-6.09	-7.49	-14.00	-17.12
3BX2	-9.97	-7.23	-5.95	-10.27	-7.32	-6.85	-12.74	-9.96	-9.37
3D2S	-9.38	-8.82	-8.73	-12.09	-9.07	-9.18	-15.66	-9.02	DNoA
3IEV	-10.76	-7.47	-7.03	-7.41	-6.96	-7.14	-7.56	-9.63	-9.01
3K61	-8.89	-6.44	-5.97	-7.14	-6.37	-6.36	-7.22	-8.94	-10.37
3K62	-8.74	-6.67	-6.15	-7.45	-6.75	-6.55	-7.74	-8.79	-9.85
3L25	-8.59	-7.35	-6.53	-8.74	-6.95	-7.11	-10.06	-9.65	-11.54
3MDG	-7.27	-5.45	-5.73	-5.75	-5.41	-5.56	-6.13	-7.33	-8.71
3NMR	-8.44	-6.63	-6.68	-7.23	-6.50	-6.77	-7.31	-8.44	-7.66
3O3I	-6.52	-3.98	-3.92	-4.56	-3.93	-3.99	-5.61	-10.03	-7.87
3O8C	-8.26	-5.09	-4.84	-5.48	-4.93	-4.93	-5.86	-10.09	-12.48
3PF5	-8.53	-5.27	-5.58	-6.76	-5.44	-5.73	-7.76	-7.83	-6.97
3Q0P	-12.81	-8.44	-7.46	-11.76	-8.54	-8.54	-14.77	-8.90	-10.21
3QG9	-10.56	-6.16	-5.63	-6.87	-6.21	-6.17	-7.13	-9.28	-9.91
3R2C	-10.63	-9.30	-8.97	-12.38	-9.46	-9.43	-15.65	-9.50	-10.05
3SIU	-9.17	-7.94	-7.46	-9.71	-8.18	-8.13	-11.61	-14.01	-10.02
3TRZ	-11.04	-10.46	-9.59	-17.44	-11.63	-10.92	-28.53	-10.55	-9.36
4ALP	-10.02	-10.51	-9.39	-12.15	-8.42	-9.55	-15.32	-8.08	DNoA
4B8T	-7.13	-6.26	-5.57	-7.46	-5.73	-5.89	-7.52	-9.21	-8.62
4C7O	-8.59	-5.76	-5.57	-6.04	-5.46	-5.67	-6.35	DNoA	-12.43
4CQN	-5.76	-6.03	-5.73	-7.01	-5.91	-6.10	-7.99	-13.68	-15.29
4CSF	-8.98	-9.54	-8.01	-24.35	-11.43	-10.08	-51.52	-9.88	-16.34
4ED5	-9.08	-7.72	-6.52	-9.87	-7.80	-7.52	-12.27	-9.63	-8.41
4ERD	-9.54	-8.31	-8.52	-10.40	-9.00	-9.02	-11.37	-9.47	-9.46
4G0A	-10.18	-9.37	-9.17	-12.67	-10.38	-10.11	-14.07	-9.94	-12.69
4GHA	-6.92	-5.32	-5.29	-6.37	-5.60	-5.43	-7.19	-10.38	-10.80
4GHL	-7.06	-4.78	-4.52	-5.25	-4.38	-4.31	-5.87	-11.89	DNoA
4H5P	-9.56	-8.86	-6.61	-14.32	-9.42	-8.65	-19.54	-9.55	-11.57
4I67	-6.76	-5.06	-5.27	-5.54	-4.95	-5.42	-5.65	-8.63	-7.93

Stress Data Set	Exp. ΔG kcal/mol	Random Forest kcal/mol	GBoost kcal/mol	Neural Network kcal/mol	Stacked Ensemble kcal/mol	XGBoost kcal/mol	Linear Regression kcal/mol	PredPRBA kcal/mol	PRA-Pred kcal/mol
4J1G	-9.09	-8.15	-7.07	-12.47	-8.81	-8.19	-17.11	-11.80	-18.30
4JVY	-9.76	-9.14	-8.08	-10.15	-8.75	-8.64	-11.25	-8.99	-11.23
4JXX	-8.14	-5.99	-6.05	-6.50	-6.04	-6.24	-6.42	-14.30	-12.02
4JXZ	-8.14	-5.58	-5.61	-6.05	-5.69	-5.76	-5.93	-14.30	-11.57
4LG2	-8.89	-7.96	-8.26	-8.81	-8.03	-8.22	-8.78	-9.33	-10.95
4LJ0	-8.59	-7.80	-7.18	-9.54	-7.21	-7.45	-10.96	-6.28	-8.17
4M7A	-9.70	-8.43	-7.41	-10.99	-8.12	-8.04	-14.48	-9.16	-8.48
4N2Q	-11.45	-8.54	-8.06	-8.57	-8.40	-8.32	-8.62	-10.09	-11.31
4O26	-7.91	-6.82	-6.63	-8.11	-6.99	-6.86	-8.87	-14.83	-10.39
4QEI	-6.55	-4.95	-4.83	-5.11	-4.78	-4.88	-5.60	-13.61	-6.75
4QI2	-9.17	-10.61	-9.48	-15.71	-11.79	-11.51	-23.36	-11.60	-11.32
4R3I	-7.77	-7.29	-7.11	-6.87	-7.18	-6.91	-6.52	-8.72	-9.06
4RCJ	-8.18	-6.32	-7.05	-7.19	-6.41	-6.87	-7.60	-9.45	-8.95
4TUX	-10.41	-8.71	-8.23	-10.13	-8.86	-9.11	-11.35	-10.42	-11.07
4U8T	-9.13	-8.21	-8.34	-13.14	-10.01	-9.20	-16.33	-8.58	-8.24
4WZM	-3.75	-5.30	-5.11	-5.84	-5.13	-5.28	-6.29	-9.74	-10.07
4YVI	-8.67	-5.86	-5.81	-6.29	-5.91	-6.16	-6.29	-13.52	-10.37
4Z31	-8.79	-7.45	-6.98	-8.29	-7.60	-7.52	-9.49	-11.91	-16.22
5DET	-8.11	-5.81	-5.29	-7.31	-5.16	-5.83	-8.65	-9.24	-8.28
5DNO	-7.83	-6.55	-6.22	-6.64	-6.11	-6.25	-7.29	-8.77	-7.31
5ELR	-6.25	-6.94	-7.01	-7.74	-6.95	-7.17	-7.67	-9.25	-9.02
5F5H	-9.18	-8.14	-8.05	-9.58	-9.02	-8.18	-10.73	-9.04	-8.79
5H1K	-7.98	-8.28	-7.74	-9.90	-8.24	-8.01	-10.52	-10.90	-10.26
5H1L	-8.18	-6.98	-6.52	-7.63	-7.52	-7.36	-7.82	-10.47	-10.19
5UDZ	-10.65	-7.84	-7.44	-9.51	-7.99	-8.05	-11.23	-11.95	-8.29
5V7C	-6.99	-7.16	-7.29	-7.63	-7.20	-7.36	-7.31	-9.86	-9.62
5WWW	-6.61	-6.04	-5.94	-6.49	-5.99	-6.13	-6.41	-7.81	-8.36
5WWX	-9.08	-6.43	-6.45	-7.16	-6.46	-6.62	-6.89	-8.47	-8.28
5WZG	-10.05	-7.21	-6.87	-7.82	-7.10	-7.09	-7.95	-9.40	-9.99
5WZK	-9.11	-6.67	-6.37	-7.39	-6.61	-6.47	-7.68	-8.83	-10.17
5X6B	-9.84	-7.48	-7.02	-7.72	-7.19	-7.28	-9.13	-11.89	-16.57
5XJ2	-6.02	-7.04	-7.44	-7.09	-7.07	-7.14	-6.92	DNoA	-9.53
5Y58	-9.42	-8.52	-7.66	-10.97	-8.85	-8.54	-12.75	-15.02	-14.99
5YTV	-8.04	-4.84	-5.11	-5.24	-4.83	-4.97	-5.11	-8.83	-5.93
5Z9X	-4.95	-5.90	-5.90	-6.90	-6.13	-6.49	-7.02	-10.26	-9.49

Stress Data Set	Exp. ΔG kcal/mol	Random Forest kcal/mol	GBoost kcal/mol	Neural Network kcal/mol	Stacked Ensemble kcal/mol	XGBoost kcal/mol	Linear Regression kcal/mol	PredPRBA kcal/mol	PRA-Pred kcal/mol
6CMN	-7.83	-8.02	-7.55	-9.12	-7.65	-7.88	-9.10	-11.66	-7.55
6D12	-9.30	-9.06	-8.79	-11.22	-9.19	-9.23	-12.20	-12.20	-7.85
6DB9	-8.25	-7.66	-7.53	-8.06	-7.57	-7.42	-8.34	-14.90	-10.48
6DCL	-10.65	-8.50	-7.59	-9.35	-7.98	-8.12	-9.75	-9.60	-9.13
6FPQ	-6.93	-7.05	-6.75	-7.26	-7.02	-6.93	-7.99	-8.95	-8.36
6FPX	-9.29	-9.83	-8.88	-14.06	-11.03	-10.51	-18.76	-11.48	-8.49
6FQR	-6.07	-5.04	-5.37	-5.38	-5.22	-5.21	-5.95	-6.46	-8.17
6GX6	-7.26	-4.50	-4.75	-4.79	-4.70	-4.67	-5.21	-7.96	-8.21
6MCE	-10.43	-9.29	-8.56	-10.65	-8.68	-8.94	-11.37	-13.50	-8.38
6MCF	-10.22	-9.87	-9.21	-10.74	-9.46	-9.35	-11.62	-13.50	-10.55
6O16	-10.72	-7.73	-6.82	-10.57	-7.48	-7.14	-12.72	DNoA	-11.95
Total Sample	112	112	112	112	112	112	112	106	107
(r)[7]		**0.63**	**0.59**	**0.53**	**0.59**	**0.59**	**0.44**	**0.18**	**0.18**

During stress assessment, we applied the same evaluation framework as done for the test set to an independent stress set with 110 complexes, which contained a larger and more diverse collection of molecular complexes. Figure 3-11 and Table 3-6 summarize the results obtained for the stress set and confirm, on average, satisfactory results with differences between the performances of the models in agreement with those monitored for the test set. In detail, the RFR model once again demonstrated the best performances in terms of both (r)=0.64 with the lowest MAE of 1.63, and rS (0.64). These results are further supported by the lowest ρ-value (6.02×10^{-14}), indicating that the model's predictive capabilities are highly significant and unlikely to have occurred by chance.

Table 3-6. Performance metrics of various machine learning models on the stress set, correlating the PANTHER scores with experimental ΔG values. Metrics include correlation coefficient (r), mean absolute error (MAE), Spearman's rank correlation coefficient (r_s), and ρ value. Random Forest Regression demonstrates the highest predictive accuracy on both datasets. The table also includes the performances achieved by the other two tested methods (i.e., PredPRBA and PRA-Pred), which are discussed in the following section.

[7] (r): Pearson Correlation value

	Stress Set			
Model	**r[8]**	**MAE[9] kcal/mol**	**r_s[10]**	**ρ value[11]**
Random Forest Regression	0.64	1.98	0.61	9.19×10^{-13}
GBoosting Regression	0.59	1.97	0.56	1.34×10^{-10}
Stacked Ensemble	0.59	1.68	0.59	3.02×10^{-12}
XGBoosting Regression	0.59	1.67	0.59	8.54×10^{-12}
Linear Regression	0.44	3.58	0.57	3.02×10^{-12}
PredPRBA	0.18	2.16	0.19	0.05
PRA-Pred	0.18	2.10	0.25	9.19×10^{-3}

Figure 3-12 shows the plots of the model-specific PANTHER score (kcal/mol) versus the experimental ΔG values (kcal/mol) and highlights the satisfactory performances reached by RFR, followed by the very similar results yielded by Gradient Boost, XGBoost and Stacking ensemble, while neural network and linear regression confirm their lower performances with some complexes which behave as clear outliers. Similar performances observed for RFR, Gradient Boost, XGBoost and stacking ensemble can be explained by considering that the computed scores are highly intercorrelated as can be seen from Table 3-5. In contrast, these interrelations decrease when considering the scores from neural network and linear regression. All the linear equations for the correlations reported in Figure 3-12 show slopes far from being equal to 45° and intercept $\neq 0$ (data not shown) but, as discussed above, the relevance of the PANTHER score is its encouraging correlation with experimental ΔG values regardless of the numerical agreement between these values.

[8] r: Pearson Correlation
[9] MAE: Mean Absolute Error
[10] r_s: Spearman Correlation
[11] ρ value: Spearman's Rho

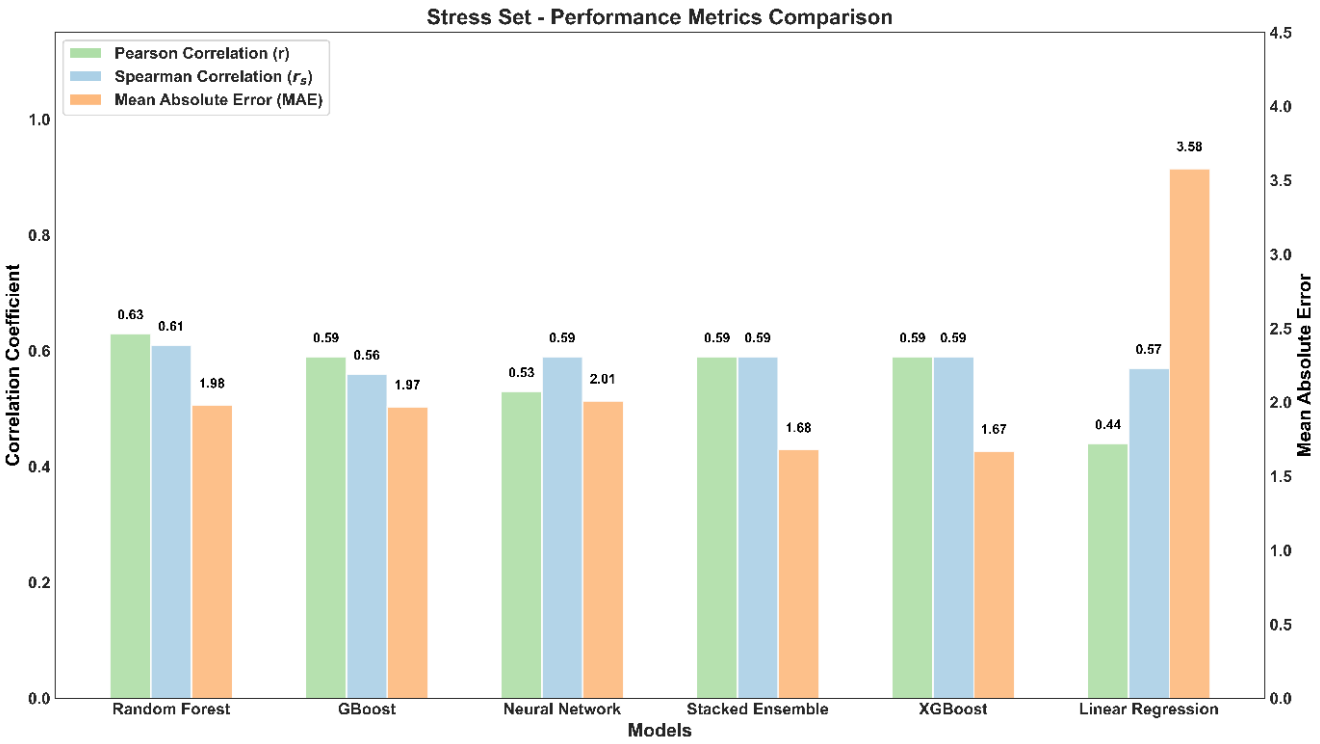

Figure 3-11. Performance comparison of machine learning models based on Pearson correlation (r), Spearman correlation (r_s), and Mean Absolute Error (MAE) for the stress set. Random Forest Regression (RFR) demonstrates the highest predictive accuracy, whereas Linear Regression shows poor correlations and higher MAE compared to the other models.

The performance discussed of the various methods find relevant confirmations also when analyzing the resulting plots of the ranks of model-specific PANTHER scores versus the experimental ΔG ranks. As evidenced by the corresponding Spearman rank correlation values, the RFR model is confirmed to be the best performing one with r_s (0.64), followed by Gradient Boost, XGBoost, neural network and Stacking ensemble which reveal r_s value (0.59) as can be observed in Figure 3-13. Notably, when analyzing the rank correlation, the linear correlation also provides similar performances (r_s = 0.57). These results indicate positive strength and direction of association between experimental ΔG and score values, a very important outcome since PANTHER score will be primarily used to rank and prioritize the analyzed RNA-protein complexes. Finally, all machine learning models demonstrated highly significant correlations with experimental ΔG values, exhibiting ρ-values < 10^{-10}. This result emphasizes that these performances cannot occur by chance. While the correlation and significance values of several models specifically RFR, GBoost, Stacked Ensemble, and XGBoost appear to be similar, this convergence reflects the robustness of the dataset and the comparable learning capacity of tree-based algorithms applied to the same descriptors.

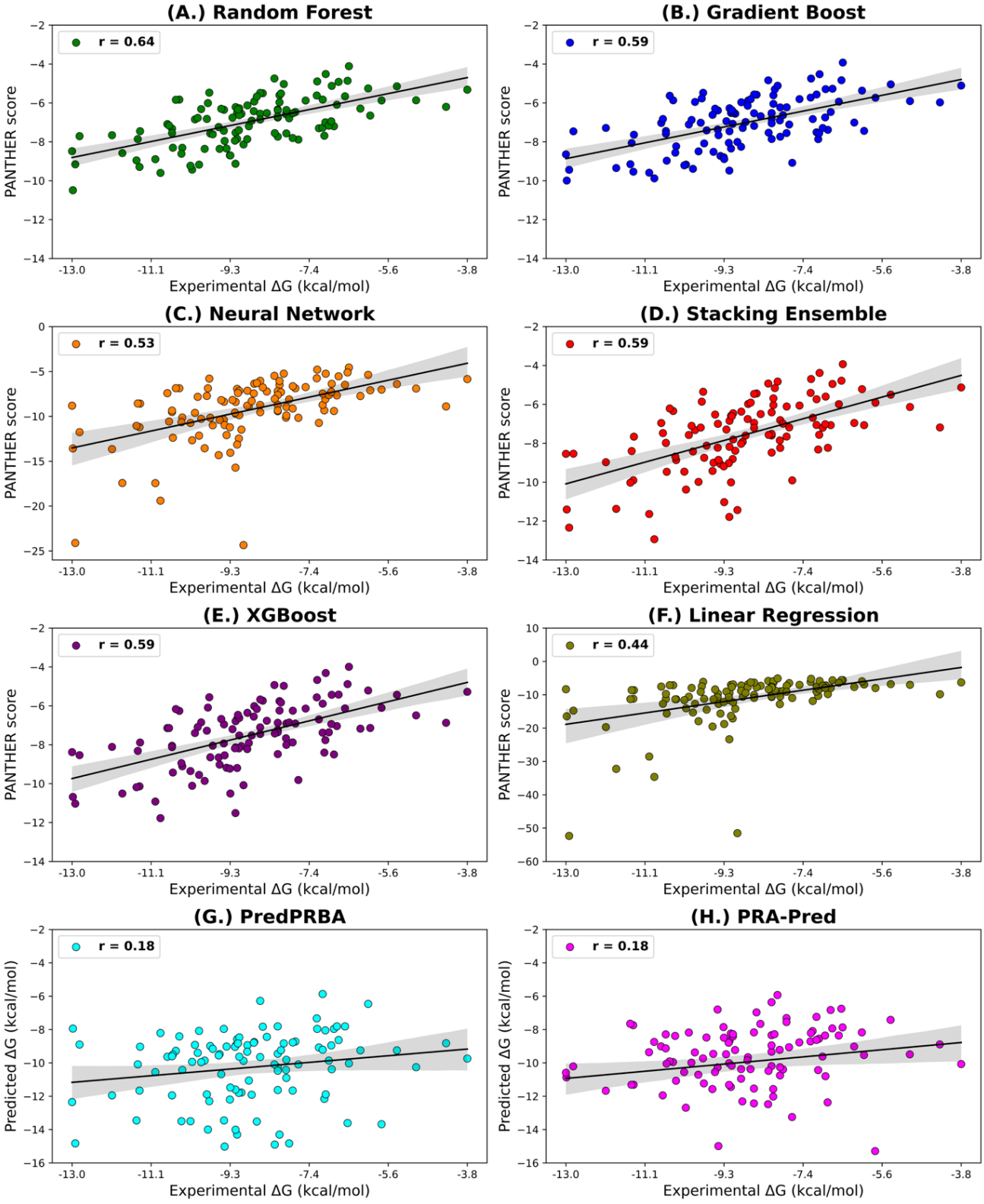

Figure 3-12. Scatter plots showing the linear correlations (as expressed by r values) between experimental ΔG values and computed PANTHER scores for various models: Random Forest, Gradient Boost, Neural Network, Stacking Ensemble, XGBoost Regression, Linear Regression.

Nevertheless, Figure 3-12, Figure 3-13 and Table 3-6 reveals that RFR displayed the best overall performance, with the highest correlation coefficients (r = 0.64, r_s = 0.64) and the lowest

MAE (1.63 kcal/mol), outperforming the next-best GBoost model (r = 0.59, r_s = 0.56, MAE = 1.97 kcal/mol) by approximately 8–10% in correlation and 17% in prediction accuracy. Furthermore, its highly significant ρ-value (6.02×10^{-14}) supports the reliability of this result. Overall, these findings indicate that, although several ensemble models perform comparably, RFR emerges as the most accurate and stable predictor and was therefore selected for the final implementation in the PANTHER scoring system. Now onwards, the term *PANTHER score* refers to the value obtained by combining the local interaction energies between amino acid-nucleotide base pairs, extracted from the here developed RFR ML model, and later processed by the local-to-global integration approach. In this workflow, the local interaction energies are first predicted by the ML model and then integrated using the local-to-global method to compute the final PANTHER score. This whole process is discussed again with an example in following sub-section "*Demonstration of the Random Forest Regression predictions for PANTHER Score Calculation*".

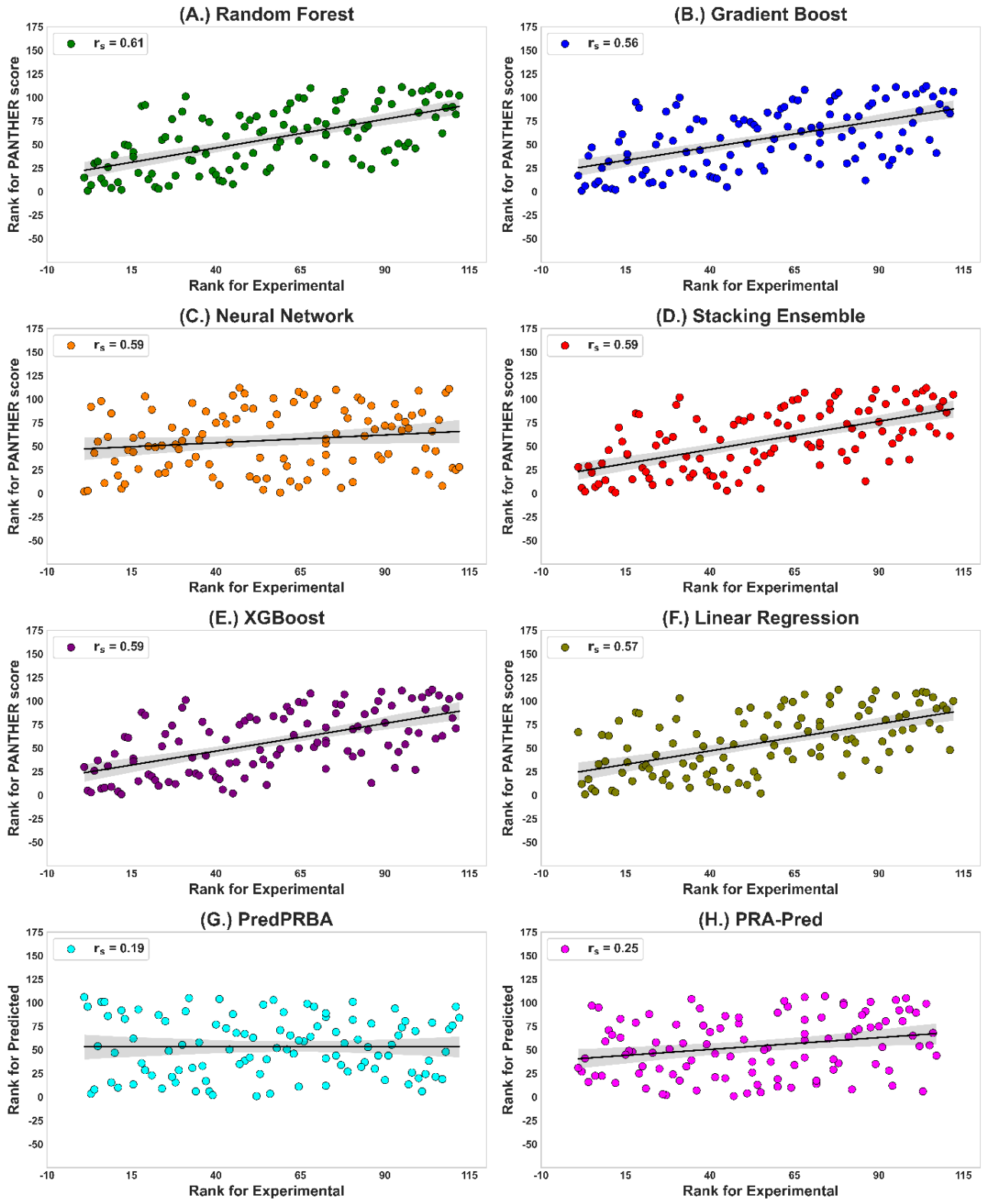

Figure 3-13. Scatter plots showing the Spearman rank correlation (r_s) between experimental ΔG ranks and ranks of PANTHER score for the models. This figure also includes the corresponding plots for the other two tested methods (PredPRBA and PRA-Pred).

3.3.5. Permutation Feature Importance Analysis

To assess whether model predictions are driven by meaningful features, we performed permutation importance analysis on the test set for the two numerical features: distance and number of hydrogen bonds (Figures 3-14 and 3-15, Tables 3-7 and 3-8). For each feature, we conducted 30 independent permutations where feature values were randomly shuffled across all pairwise interactions, and the pairwise local energies were predicted using the RFR model and converted to PANTHER score using our local-to-global method. Performance degradation was assessed relative to baseline using Root Mean Square Error (RMSE) and Pearson correlation coefficient (r). Distance emerged as key feature: permuting it increased RMSE by 0.140 ± 0.014 kcal/mol (+7.6%) and reduced Pearson (r) by 0.135 ± 0.013 (reported relative to (r) ≈ 0.80, ~16.9% loss; both $\rho < 10^{-6}$, one-sided t-test, mean ± SD over 30 permutations). The number of hydrogen bonds had an even larger impact on absolute error, increasing RMSE by 0.259 ± 0.007 kcal/mol (+14.0%), and reduced Pearson (r) by 0.083 ± 0.011 (~10.4% loss; $p < 10^{-6}$). These results align with molecular recognition: distance captures the distance-dependent nature of intermolecular forces and contributes most to rank-order discrimination across complexes (larger Δr), while number of hydrogen bonds reflects hydrogen-bonding networks and contributes to the magnitude of predicted PANTHER score (larger ΔRMSE). Together, these features capture complementary aspects of protein-RNA binding thermodynamics.

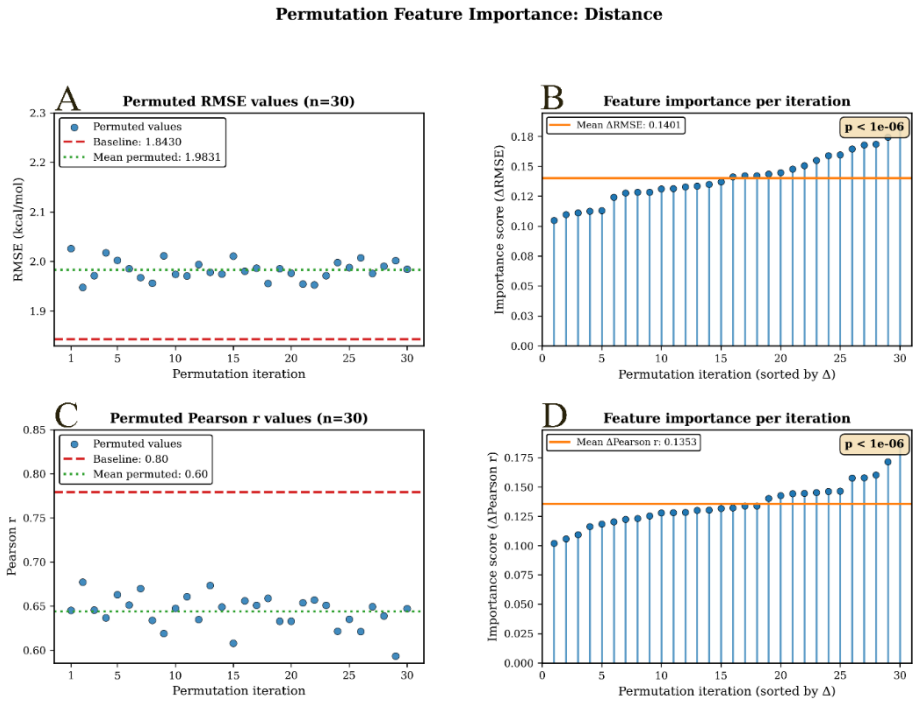

Figure 3-14. Permutation feature importance analysis for distance (n = 30 shuffles). A, C: Individual permutation results (blue circles) with the baseline using intact features (red dashed line) and the

mean across permutations (green dotted line). B, D: Importance scores with runs sorted by magnitude; the orange line marks the mean. We define *ΔRMSE = RMSE_perm − RMSE_base* and *Δr = r_base − r_perm* so larger values indicate greater degradation. Top row: RMSE (kcal/mol). Bottom row: Pearson correlation coefficient. Permuting Distance significantly degrades performance (Δ RMSE = 0.140 ± 0.014 kcal/mol; Δ Pearson r = 0.135 ± 0.013; both $p < 10^{-6}$, one-sided t-test; mean ± SD), indicating that intermolecular distance is a critical driver of the model's predictions, consistent with fundamental principles of molecular interaction.

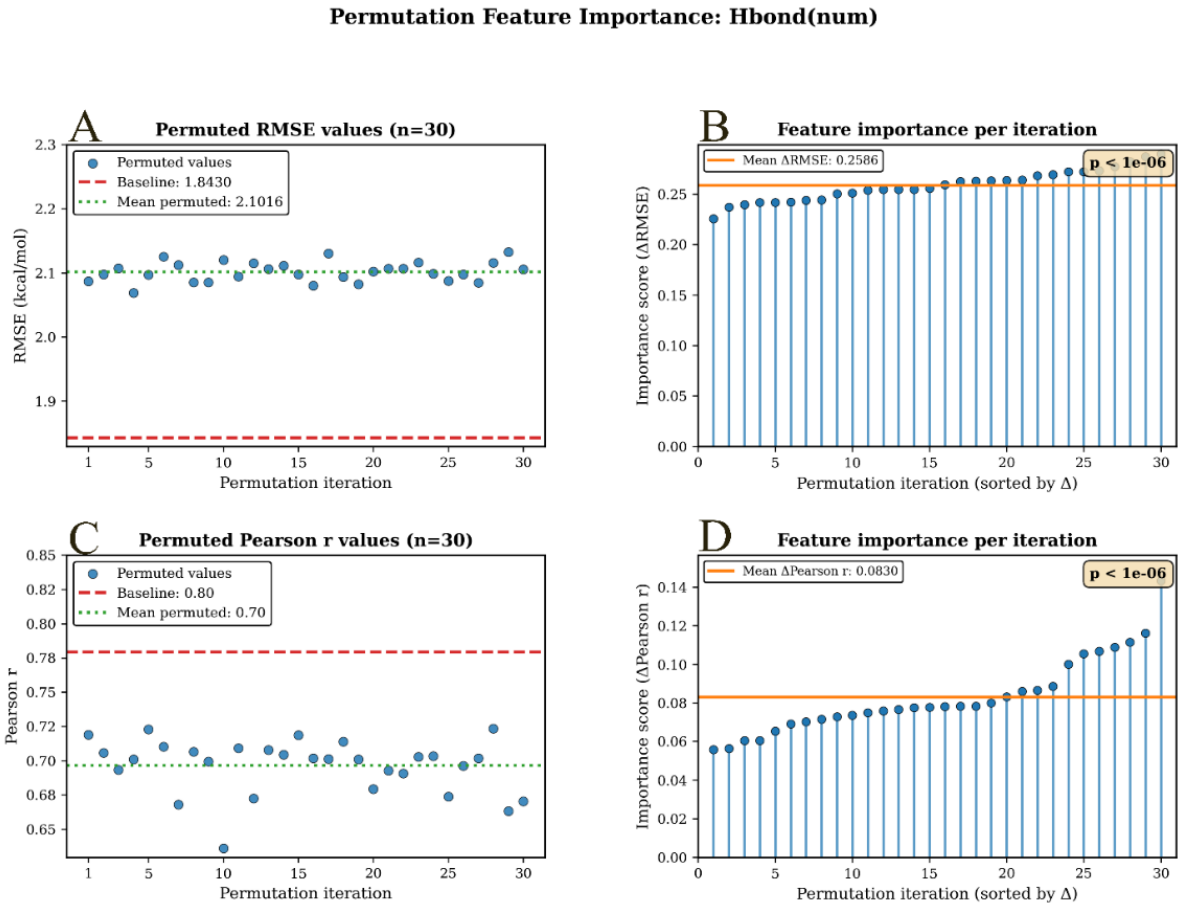

Figure 3-15. Permutation feature importance analysis for Hbond (num) (n = 30 shuffles), following the same methodology as Figure 3-1. A, C: Individual permutation results (blue circles) with the baseline using intact features (red dashed) and the mean across permutations (green dotted). B, D: Importance scores with runs sorted by magnitude; the orange line marks the mean. We define *ΔRMSE = RMSE_perm − RMSE_base* and *Δr = r_base − r_perm* so larger values indicate greater degradation. Permuting Hbond (num) increased error by ΔRMSE = 0.259 ± 0.007 kcal/mol and reduced correlation by Δ Pearson r = 0.083 ± 0.011 (both $p < 10^{-6}$, one-sided t-test; mean ± SD).

Table 3-7. ML-predicted local-to-global score predictions (kcal/mol) for the test set, obtained across 30 permutations in which distance feature was randomly shuffled. Columns: PDB code, experimental ΔG, baseline prediction (original features), and 30 permuted predictions (Perm 01 to Perm 30). Each permuted prediction corresponds the local-to-global decomposed energies after distance perturbation.

TestSet	Exp. ΔG	Baseline-Prediction	Perm01[12]	Perm02	Perm03	Perm04	Perm05	Perm06	Perm07	Perm08	Perm09	Perm10
1EC6	-7.99	-6.03	-5.34	-5.35	-5.43	-5.39	-5.40	-5.40	-5.40	-5.46	-5.37	-5.38
1NYB	-9.13	-7.91	-8.20	-8.22	-8.37	-8.25	-8.27	-8.24	-8.21	-8.45	-8.45	-8.32
1RKJ	-7.99	-7.13	-7.23	-7.27	-7.15	-7.18	-7.16	-7.30	-7.24	-7.24	-7.08	-7.27
1U0B	-8.96	-6.31	-6.33	-6.39	-6.37	-6.37	-6.32	-6.31	-6.36	-6.41	-6.36	-6.37
2B6G	-10.26	-7.59	-7.40	-7.55	-7.48	-7.36	-7.44	-7.46	-7.55	-7.47	-7.39	-7.55
2LBS	-6.09	-5.50	-5.54	-5.54	-5.65	-5.60	-5.52	-5.53	-5.57	-5.64	-5.67	-5.60
2M8D	-8.73	-6.94	-6.53	-6.72	-6.55	-6.51	-6.56	-6.62	-6.58	-6.52	-6.49	-6.50
TestSet	**Exp. ΔG**	**Baseline-Prediction**	**Perm11**	**Perm12**	**Perm13**	**Perm14**	**Perm15**	**Perm16**	**Perm17**	**Perm18**	**Perm19**	**Perm20**
1EC6	-7.99	-6.03	-5.35	-5.35	-5.31	-5.40	-5.31	-5.40	-5.41	-5.42	-5.39	-5.37
1NYB	-9.13	-7.91	-8.28	-8.26	-8.32	-8.20	-8.38	-8.31	-8.28	-8.34	-8.29	-8.29
1RKJ	-7.99	-7.13	-7.21	-7.32	-7.15	-7.25	-7.27	-7.29	-7.23	-7.16	-7.24	-7.29
1U0B	-8.96	-6.31	-6.37	-6.43	-6.38	-6.37	-6.41	-6.36	-6.41	-6.35	-6.36	-6.40
2B6G	-10.26	-7.59	-7.52	-7.41	-7.54	-7.45	-7.35	-7.48	-7.44	-7.51	-7.46	-7.40
2LBS	-6.09	-5.50	-5.58	-5.58	-5.53	-5.63	-5.63	-5.51	-5.58	-5.61	-5.66	-5.61
2M8D	-8.73	-6.94	-6.64	-6.52	-6.61	-6.64	-6.52	-6.55	-6.51	-6.67	-6.55	-6.65
TestSet	**Exp. ΔG**	**Baseline-Prediction**	**Perm21**	**Perm22**	**Perm23**	**Perm24**	**Perm25**	**Perm26**	**Perm27**	**Perm28**	**Perm29**	**Perm30**
1EC6	-7.99	-6.03	-5.37	-5.39	-5.42	-5.28	-5.38	-5.29	-5.30	-5.36	-5.40	-5.46
1NYB	-9.13	-7.91	-8.21	-8.42	-8.27	-8.21	-8.34	-8.28	-8.35	-8.26	-8.29	-8.33
1RKJ	-7.99	-7.13	-7.26	-7.22	-7.19	-7.25	-7.25	-7.26	-7.31	-7.28	-7.22	-7.24
1U0B	-8.96	-6.31	-6.36	-6.30	-6.33	-6.40	-6.36	-6.37	-6.39	-6.34	-6.37	-6.35
2B6G	-10.26	-7.59	-7.54	-7.57	-7.48	-7.46	-7.45	-7.40	-7.57	-7.47	-7.30	-7.39
2LBS	-6.09	-5.50	-5.65	-5.57	-5.63	-5.71	-5.61	-5.64	-5.56	-5.60	-5.76	-5.53
2M8D	-8.73	-6.94	-6.69	-6.66	-6.64	-6.58	-6.55	-6.57	-6.51	-6.57	-6.63	-6.59

[12] Perm*: Permutation

Table 3-8. ML-predicted local-to-global score predictions (kcal/mol) for the test set, obtained across 30 permutations, in which number of Hbond was randomly shuffled. Columns: PDB code, experimental ΔG, baseline prediction (original features), and 30 permuted predictions (Perm 01 to Perm 30). Each permuted prediction represents the local-to-global decomposed energies after Hbond perturbation.

TestSet	Exp. ΔG	Baseline-Prediction	Perm01 [13]	Perm02	Perm03	Perm04	Perm05	Perm06	Perm07	Perm08	Perm09	Perm10
1EC6	-7.99	-6.03	-6.20	-6.20	-6.18	-6.23	-6.20	-6.18	-6.23	-6.25	-6.18	-6.20
1NYB	-9.13	-7.91	-7.67	-7.69	-7.67	-7.71	-7.57	-7.60	-7.72	-7.68	-7.71	-7.78
1RKJ	-7.99	-7.13	-6.77	-6.79	-6.74	-6.83	-6.74	-6.74	-6.70	-6.79	-6.70	-6.83
1U0B	-8.96	-6.31	-6.17	-6.18	-6.21	-6.17	-6.16	-6.21	-6.18	-6.21	-6.17	-6.13
2B6G	-10.26	-7.59	-7.06	-7.07	-6.99	-7.02	-7.01	-6.99	-6.98	-7.07	-7.01	-6.96
2LBS	-6.09	-5.50	-5.58	-5.57	-5.65	-5.58	-5.63	-5.60	-5.69	-5.60	-5.66	-5.64
2M8D	-8.73	-6.94	-6.41	-6.29	-6.35	-6.50	-6.51	-6.30	-6.31	-6.30	-6.51	-6.28
TestSet	**Exp. ΔG**	**Baseline-Prediction**	**Perm11**	**Perm12**	**Perm13**	**Perm14**	**Perm15**	**Perm16**	**Perm17**	**Perm18**	**Perm19**	**Perm20**
1EC6	-7.99	-6.03	-6.14	-6.17	-6.18	-6.13	-6.23	-6.19	-6.15	-6.20	-6.22	-6.22
1NYB	-9.13	-7.91	-7.65	-7.67	-7.66	-7.66	-7.57	-7.63	-7.58	-7.65	-7.67	-7.66
1RKJ	-7.99	-7.13	-6.80	-6.78	-6.79	-6.75	-6.77	-6.80	-6.79	-6.80	-6.78	-6.81
1U0B	-8.96	-6.31	-6.21	-6.19	-6.16	-6.18	-6.19	-6.17	-6.17	-6.22	-6.14	-6.20
2B6G	-10.26	-7.59	-7.04	-7.02	-7.01	-7.05	-7.08	-7.06	-7.01	-7.06	-7.04	-7.01
2LBS	-6.09	-5.50	-5.61	-5.68	-5.55	-5.61	-5.64	-5.68	-5.60	-5.58	-5.63	-5.66
2M8D	-8.73	-6.94	-6.38	-6.26	-6.39	-6.33	-6.33	-6.46	-6.29	-6.31	-6.46	-6.30
TestSet	**Exp. ΔG**	**Baseline-Prediction**	**Perm21**	**Perm22**	**Perm23**	**Perm24**	**Perm25**	**Perm26**	**Perm27**	**Perm28**	**Perm29**	**Perm30**
1EC6	-7.99	-6.03	-6.19	-6.18	-6.09	-6.20	-6.24	-6.14	-6.22	-6.14	-6.18	-6.17
1NYB	-9.13	-7.91	-7.64	-7.54	-7.70	-7.65	-7.65	-7.66	-7.65	-7.54	-7.69	-7.80
1RKJ	-7.99	-7.13	-6.77	-6.78	-6.77	-6.80	-6.78	-6.74	-6.78	-6.71	-6.80	-6.78
1U0B	-8.96	-6.31	-6.17	-6.19	-6.23	-6.18	-6.17	-6.23	-6.20	-6.22	-6.20	-6.15
2B6G	-10.26	-7.59	-6.98	-6.98	-6.98	-7.06	-7.01	-7.02	-7.04	-7.07	-6.88	-7.02
2LBS	-6.09	-5.50	-5.63	-5.74	-5.55	-5.60	-5.74	-5.70	-5.66	-5.69	-5.59	-5.60
2M8D	-8.73	-6.94	-6.42	-6.44	-6.33	-6.32	-6.41	-6.36	-6.38	-6.30	-6.33	-6.31

[13] Perm*: Permutation

3.3.6 Comparison of PANTHER Score with Existing Functional Software

The performances of all the model-specific PANTHER scores were compared with those obtained using two similarly structured, publicly available methods known as PredPRBA and PRA-Pred. These two tools were chosen because they represent recent and accessible we-based applications for protein-RNA binding affinity prediction. Notably, both methods accept the PDB ID as input; therefore, the 110 complexes of the stress set were directly submitted without requiring any preparation. The results obtained for these two methods, reported in Table 3-5, Table 3-6, Figures 3-11 and 3-12, reveal significantly poorer performance compared to that achieved by the PANTHER score. In detail, PredPRBA and PRA-Pred exhibit very low (r) (both 0.18) and r_s (0.19 and 0.25, respectively) values, indicating a lack of reliable correlations between the experimental ΔG values, and the scores predicted by both methods when analyzing the stress set of 110 untreated complexes spanning various categories. In addition, the mean absolute errors (MAE) exhibit rather high values (2.16 and 2.10 kcal/mol for PredPRBA and PRA-Pred respectively) and the ρ-values (> 10^{-3}) suggest a limited significance of these correlations. Collectively, these comparative analyses further underline the notable performance achieved by the PANTHER score since the proposed fragmental approach outperforms the other available tested tools. To understand the substantial performance gap between PANTHER and existing methods, we analyzed the key differences in their underlying methodologies and training strategies. PredPRBA trained Gradient Boosted Regression Tree (GBRT) models on 103 non-redundant protein-RNA complexes using 37 features. Although these features are structure-based, they represent global or interface-averaged properties, such as secondary structure content, solvent-accessible surface area, RNA base-pair frequencies, and predicted folding energies, rather than explicit residue-to-nucleotide interactions or pairwise energetics. While the GBRT framework effectively captures statistical correlations between these global descriptors and binding affinity, it lacks the capability to resolve atom-level energetic contributions as well as to account for conformational dynamics. This limitation constrains generalization to structurally diverse or untreated complexes. On the other hand, PRA-Pred advanced structure-based modeling by assembling 217 non-redundant complexes and defining 17 structural descriptors. However, these descriptors are aggregated across entire binding interfaces rather than being resolved for individual residue-nucleotide pairs. Furthermore, the multiple linear regression (MLR) framework employs only 4-8 interface-averaged variables per model and assumes simple linear additivity, and therefore cannot represent the critical non-linear, distance-dependent effects in RNA-protein binding. Although both methods achieved high performance on their

respective benchmarking datasets, their accuracy declined substantially when applied to the independent stress set. In contrast, PANTHER introduces a two-tier, physics-aware framework that fundamentally differs from both approaches. At the local level, MD simulations mitigate data scarcity and quality limitations by generating time-averaged pairwise interaction energies. Specifically, only interactions persisting for more than 70% of simulation time are retained, effectively filtering transient and non-specific contacts. Each persistent residue-nucleotide pair is then decomposed into van der Waals and electrostatic components, providing physically meaningful, atom-level training data. Additionally, PANTHER employs RFR to predict local energies from structural features. In contrast to the linear MLR framework used in PRA-Pred, RFR can capture complex non-linear relationships and feature interactions inherent in protein-RNA binding, thus leading to improved predictive accuracy for local interaction energies. Then, predicted local energies are integrated to estimate overall binding affinity with our local-to-global approach. This design offers three key advantages. First, it captures explicit residue-nucleotide energetics rather than interface-averaged properties, thereby preserving spatial resolution. Second, it leverages MD-derived dynamic data, enabling the model to learn physically derived relationships between pairwise features and interaction energies that would be inaccessible from static structures alone. Third, the two-stage framework of local energy prediction followed by local-to-global integration facilitates generalization across diverse RNA and protein types, as local interactions represent transferable building blocks for binding affinity predictions.

3.3.7 Demonstration of the Random Forest Regression predictions for PANTHER Score Calculation

To illustrate the local-to-global approach of the PANTHER score, we analyzed the structure of S.pombe Mmi1 in complex with 7-mer RNA complex (PDB ID: 6FPQ, experimental ΔG = -6.93 kcal/mol). Figure 3-16 shows the 3D structure of the complex, highlighting the key local interactions between amino acid and nucleotide bases (e.g., ASN-336 with A-5, ARG-338 with C-6). Each local interaction energy, as predicted by our RFR model, is listed in Table 3-9, while an input example for RFR model has been detailed in Table 3-10. Notably, the RFR model accurately estimates these local energies without requiring computationally expensive MD simulations, demonstrating the efficiency of the here developed approach. For instance, ARG-488 and U-1 exhibit a strong local energy of -18.95 kcal/mol, while other pairs like TYR-352 and A-3 contribute -5.69 kcal/mol. By averaging these local energies, we obtained a PANTHER score of -7.05 kcal/mol, closely matching the experimental ΔG value (difference: -0.12 kcal/mol). This close agreement is a clear example of

the predictive power of the approach, where ML-predicted local energy contributions effectively replace MD-derived local energies maintaining satisfactory predictive accuracy. The graphical representation in Figure 3-16, together with the predicted energy values reported in Table 3-9, highlights the explainability of the PANTHER score, which enables the evaluating protein-RNA binding recognition processes through a computationally efficient framework.

Table 3-9. Example of protein-RNA complex 6FPQ (Experimental ΔG = -6.93) to predict the Local energies by the Random Forest model and integrated with local-to-global method to be called as PANTHER score of -7.05.

Amino Acid Number	Amino Acid Name	Nucleotide Base Number	Nucleotide Base Name	Local Energy kcal/mol
336	ASN[14]	5	A[15]	-4.60
336	ASN	6	C[16]	-3.55
338	ARG[17]	6	C	-17.67
352	TYR[18]	3	A	-5.69
392	TYR	4	A	-3.39
406	TYR	3	A	-5.15
436	LYS[19]	3	A	-8.63
437	THR[20]	1	U[21]	-8.04
466	TYR	4	A	-3.37
470	SER[22]	3	A	-3.40
477	ASN	3	A	-5.10
487	ASP[23]	1	U	-4.11
488	ARG	1	U	-18.95
Average				**-7.05**

[14] ASN: Asparagine
[15] A: Adenine
[16] C: Cytosine
[17] ARG: Arginine
[18] TYR : Tyrosine
[19] LYS: Lysine
[20] THR: Threonine
[21] U: Uracil
[22] SER: Serine
[23] ASP: Asparagine

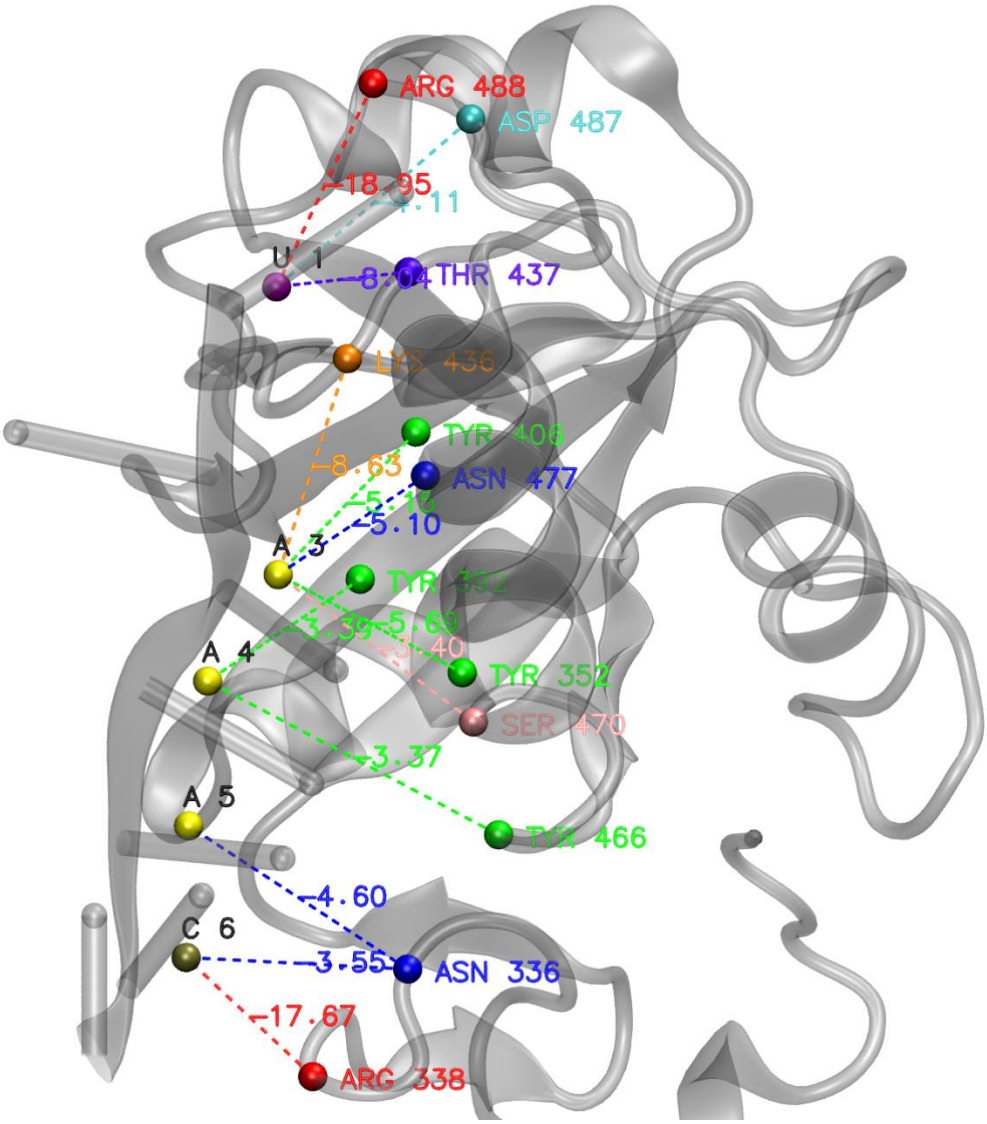

Figure 3-16. Visualization of protein-nucleotide interactions for the Mmi1 protein in complex with 7-mer RNA (PDB Id 6FPQ) showing residue pairs with their local binding energies (kcal/mol). Residue labels are positioned close to their corresponding spheres, with energy values displayed along dashed connection lines. These spheres are the center of mass of each interacting residue. By averaging these local energies obtained from Random Forest Regression models we finally achieved the PANTHER score of -7.05 kcal/mol whereas the experimental value of 6FPQ is -6.93 kcal/mol. The interacting groups are coloured according to their stabilizing energies by a colour ramp ranging from red (strongly interacting groups) to blue (poorly interacting groups) colours. The data is shown in Table 3-9.

Table 3-10. Example of feature inputs used to train the ML models in this study.

Amino Acid ID	Nucleotide Base ID	Distance (Å)	Local Interaction Energy (kcal/mol)	Hbond [24]Number
ASN[25]	G[26]	9.77	-3.99	2
LYS[27]	A[28]	10.71	-7.62	2
ARG[29]	U[30]	7.76	-8.03	2
GLN[31]	C[32]	8.67	-4.63	3

3.4. Conclusion

In this study, we introduced the PANTHER scoring function, a machine learning-based model developed to calculate scores specially tailored for protein-RNA interactions with high accuracy and reliability. By leveraging pairwise interaction features derived from MD simulations and a diverse dataset of protein-RNA complexes, PANTHER score demonstrated superior performance compared to existing methods, including PredPRBA [205], and PRA-Pred [202], as evidenced by its robust correlation coefficients and reduced prediction errors.

Among the models tested, those generated by RFR emerged as the most reliable ones, achieving the highest model-specific PANTHER score vs experimental ΔG Pearson correlation coefficient (r = 0.80) and a mean absolute error (MAE = 1.79 kcal/mol) on the test set. Even under the challenging conditions of the stress set, RFR maintained its robustness, with (r) = 0.64 and MAE = 1.63 kcal/mol, highlighting its ability to handle unseen and untreated data. These findings underscore the strength of the developed RFR model in capturing complex, non-linear relationships inherent in protein-RNA interactions.

Thus, our approach demonstrates the value of employing a local-to-global methodology to generate high-quality data for model training. By calculating local energies purposely extracted from MD simulations and filtered by implementing noise reduction techniques (as detailed in Methods sub-section; "*Extraction of Pairwise Local Energies and Noise Reduction Technique*"), we produced

[24] Hbond: Hydrogen Bond
[25] ASN: Asparagine
[26] G: Guanine
[27] LYS: Lysine
[28] A: Adenine
[29] ARG: Arginine
[30] U: Uracil
[31] GLU: Glutamine
[32] C: Cytosine

a large and reliable dataset of intermolecular interactions to effectively train the models. Such an approach shows several advantages, which can be summarized as follows. First, to reduce the high-frequency oscillations typically observed in MD simulations, interaction energies are averaged over 10 consecutive frames and only those interactions which persist during 70% of the simulation period were considered. This lifetime cutoff and averaging process reduces random or negligible interactions and, enhances the reliability of the local dataset by focusing on the most stable interaction energy profiles. Second, to ensure comprehensive temporal coverage of the simulation data, a frame-skip interval of 40 frames is implemented between the 10-frame analyzed windows. This strategy effectively reduces the risk of oversampling from highly correlated adjacent frames, which could lead to biased results. By maintaining a balanced sampling approach, the interaction data reflects a broader spectrum of the system's dynamics throughout the entire simulation time. Third, the combined methodology of averaging interaction energies and implementing frame-skipping effectively captures the harmonic fluctuations inherent in the system. This dual approach strikes a balance between preserving detailed local interaction data and encompassing broader temporal dynamics. Consequently, it enhances the dataset's capacity to reveal significant trends and patterns in molecular interactions over time. Fourth, the resulting dataset, characterized by average and temporally spaced interaction energies, is particularly suited for training machine learning models. The balanced structure of the dataset improves the generalizability of the trained models by exposing them to a diverse array of interaction patterns across various timeframes. Finally, the 12 Å cutoff distance is chosen to include biologically relevant short-range interactions, such as ion-pairs, hydrogen bonding, and van der Waals forces, while effectively excluding long-range interactions that contribute negligibly to pairwise energy calculations. This criterion ensures that the analysis remains focused on interactions that are significant from a biophysical perspective. Altogether, this approach allowed us to overcome limitations posed by the scarcity of experimental ΔG values and enhance the predictive reliability of PANTHER score.

Remarkably, such a fragmental approach can have a general role and can find fruitful application to analyze the intermolecular interactions in protein-protein complexes or even in ligand-protein complexes. Moving forward, the proposed approach is flexible and can be further expanded by incorporating additional protein-RNA complexes to enhance the predictive power of the PANTHER score by exploring more interaction features. These efforts will continue to refine the predictive capabilities of the PANTHER score, broadening its applicability to other nucleic acid-protein systems and deepening our understanding of these essential biological interactions.

Chapter 4

PANTHER score Web Service: An Accessible Platform for Predicting Protein-RNA Binding Affinities Using Machine Learning

4.1. Introduction

Protein-RNA interactions are essential for cellular processes including gene expression, RNA processing, and post-transcriptional regulation [186]. Accurate prediction of binding affinities between proteins and RNA molecules is crucial for understanding these mechanisms and advancing therapeutic applications targeting RNA-protein complexes [232], [233]. Current computational approaches for predicting protein-RNA binding affinities still face significant limitations: a) Molecular docking method, while widely used, often lacks the accuracy required for reliable affinity prediction [234]; b) Physics-based approaches using molecular dynamics simulations provide detailed insights but require substantial computational resources, limiting their accessibility for routine applications and is very expensive for large-scale *in silico* screening [201]. Thus, it can be inferred that accuracy is expensive, but Machine learning approaches have shown great potential in addressing these challenges by learning complex patterns from experimental and simulation data [111] [235]. However, most existing ML models for protein-RNA interactions focus on binding site prediction rather than quantitative affinity estimation, although few others are available as user-friendly web interfaces. Existing web-based tools such as *PredPRBA* [111] and *PRA-Pred* [64] offer limited accuracy and lack comprehensive validation on diverse protein-RNA systems. We previously developed *PANTHER* score (Protein Affinity for Nucleic Targeting, Hybridization, and Energy Regression), a machine learning model trained on data derived from molecular dynamics stimulations, designed to mimic observation. This model predicts protein-RNA binding affinities using a novel local-to-global approach, as described in Chapter 3. This method combines carefully curated data from molecular dynamics simulations with Random Forest Regression to achieve superior predictive performance compared to the existing tools. Here, we present the PANTHER web service, which makes this powerful predictive capability freely accessible to the broader scientific community. A graphical overview of the web service architecture is shown in Figure 4-1.

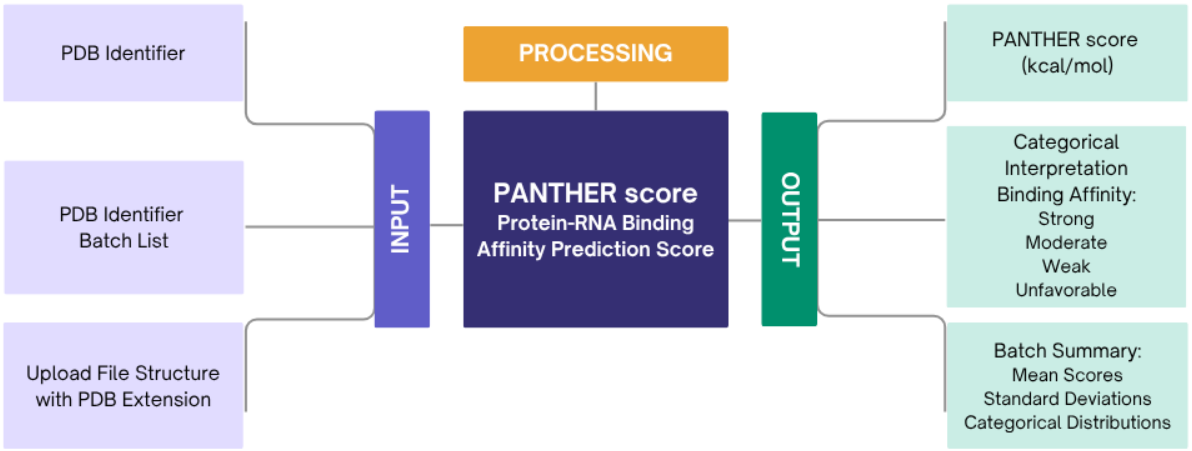

Figure 4-1. PANTHER web service architecture.

4.2. Materials and Methods

4.2.1. Workflow

The PANTHER web service provides a streamlined interface to evaluate protein–RNA binding interactions through computation of the PANTHER score. Upon submission, the input protein–RNA complex can be specified either by providing a PDB identifier, uploading a batch list containing up to 10 PDB IDs, or providing a custom PDB file. The service automatically retrieves (if necessary) and parses structural files, identifies protein residues and RNA bases, and extracts relevant interaction features including pairwise distances, hydrogen bonds, and residue identities. These features are processed by a pre-trained Random Forest regression model, which predicts the local interaction energies between amino acid–nucleotide pairs. Local predictions within a 9 Å cutoff are then integrated into a global binding energy to estimate the PANTHER score (in kcal/mol). Results are presented as overall binding scores, detailed tables of residue-base interaction energies, and thermodynamic categorizations (strong, moderate, weak, or unfavorable binding). For batch submissions, additional statistical summaries (mean scores, standard deviations, and category distributions) are also provided. All outputs are available through interactive visualizations directly on the web interface and can also download in JSON and CSV formats for downstream analysis. The scientific and mathematical methodology underlying the PANTHER score has been described in

detail in our previous chapter (Chapter 3). However, we briefly summarize the key methodological aspects of the web service workflow (Figure 4-2).

4.2.2. Dataset and Model Training

The PANTHER score model was trained on well curated 87,117 points, which are amino acid-nucleotide pairwise interactions extracted from 46 protein-RNA complexes. These 46 protein-RNA complexes were extracted from the Protein Data Bank under stringent quality criteria and were curated as described in detail in our previous chapter. The initial data were generated through molecular dynamics (MD) simulations performed with the AMBER20 package, using amber99SB force field for proteins and OL3 for RNA. From the MD simulation trajectories, the pairwise interaction energies between amino acids and nucleotide bases were further computed. These local energetic contributions combined with basic geometric descriptors such as pairwise distances between the interacting partners, their chemical properties (amino acid and nucleotide types), and structural parameters (hydrogen bond counts) formed the basis features for machine learning model training. Several regression models were evaluated, including Random Forest, Gradient Boosting, Stacked Ensemble, linear regression and neural networks. Among them, Random Forest regression achieved the best compromise between accuracy and robustness for predicting the local interaction energy between the amino acid-nucleotide base of a protein-RNA complex when provided as input. These local predictions under a cutoff distance of 9 Å radius distance calculated between the center-of-mass of interacting amino acid-nucleotide base further integrates into a global scoring framework. Thus, in short, the local-to-global approach of amino acid-nucleotide interaction energies aggregates into a single binding energy estimate, named as the PANTHER score (represented in kcal/mol).

4.2.3. Service Architecture and Implementation

PANTHER score website is designed for accessibility, robustness, and transparency. The service operates on the University of Milan's Nova web server (*https://nova.disfarm.unimi.it/panther/*), where a *Python CGI* script handles backend processing, and an *HTML5* & *JavaScript* interface provides a modern, responsive user-friendly experience.

Frontend: The user interface provides three input options: single PDB ID entry, multiple PDB IDs batch submission (maximum up to 10 structures), and user's complex uploads in PDB file format.

The interface implements real-time validation for PDB identifiers (4-character alphanumeric format), displays progress indicators during processing, and provides comprehensive error handling with informative feedback messages.

Backend: The service backend is implemented in Python 3.8+ with *BioPython* [236] for structural parsing and *scikit-learn* [237] for machine learning predictions. The system validates input, downloads PDB files from RCSB servers when needed, processes structures to identify protein-RNA interactions, extracts feature for ML prediction and generates results in JSON format for seamless frontend integration.

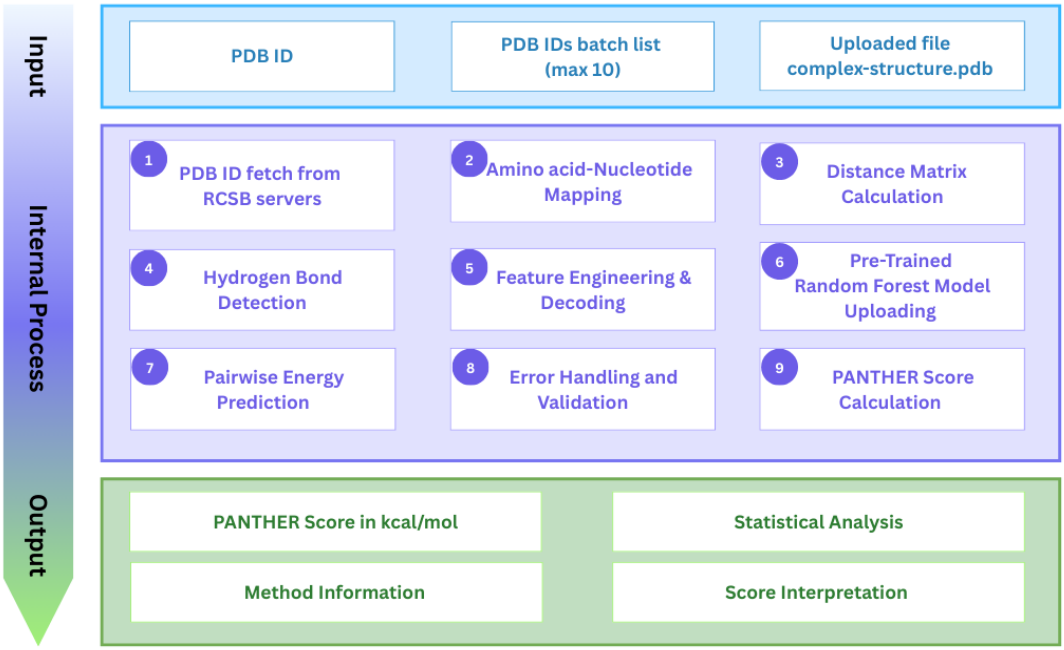

Figure 4-2. PANTHER Web Server workflow.

4.2.4. Input Processing Workflow

The PANTHER web service executes a computational workflow designed for robust processing of protein-RNA complexes across diverse structural contexts (Figure 4-2). The pipeline begins with input handling that distinguishes between PDB identifiers, batch submission lists and uploaded structure files. For PDB identifiers, the system establishes the connections to the RCSB Protein Data Bank servers downloading and caching structure files to minimize redundant network operations. Structure parsing employs PDB coordinate analysis that processes ATOM records and processes only the first structural model for simplicity when multiple models are present in PDB

files. The parser systematically identifies protein residues by screening for the twenty standard amino acids and RNA components by detecting adenine, uracil, guanine and cytosine. It should be noted that complexes containing non-standard amino acids, DNA chains, and ligands within the binding-site may lead to false interpretation by the web service. Furthermore, in the pipeline the chain identification and residue numbering preservation maintain structural context throughout the analysis pipeline. Geometric analysis proceeds through center of mass calculations for each residue and nucleotide using atomic coordinates weighted by standard atomic masses. Distance measurements between all protein-RNA pairs employ three-dimensional Euclidean geometry with a 12 Å cutoff threshold that balances computational efficiency with biological relevance. The built-in hydrogen bond detection tool implements the bond-vector method to estimate hydrogen positions from donor atoms, applying geometric criteria that define a maximum distance of 3.5 Å between donor and acceptor atoms. Feature engineering aggregates interacting data by summarizing multiple contacts between identical residue pairs, calculates average distances to account for conformational flexibility, and encodes chemical identities. The preprocessing pipeline applies trained column transformers to normalize numerical features and encode categorical variables using one-hot representation schemes compatible with the Random Forest architecture. Prediction generation applies the pre-trained Random Forest ensemble to the processed features, yielding local interaction energies that reflect the thermodynamic contributions of individual amino acid-nucleotide pairs. The final score calculation integrates these local predictions within a cutoff distance of 9 Å (center-of-mass distances between two amino acid-nucleotide base pairs) using distance-weighted averaging to produce the global PANTHER score, expressed in kcal/mol units, which further facilitates the thermodynamic interpretation.

4.2.5. Error Handling and Validation

The service implements complete validation through multiple integrated layers designed to ensure reliable operation across diverse input conditions. Input validation encompasses PDB identifier format verification using regular expressions to confirm four-character alphanumeric structure, duplicate detection to prevent redundant processing, and file size limitations to maintain system performance. Structure validation involves examining the coordinates to ensure compliance with the PDB format guidelines, verifying the completeness of atomic coordinates, and confirming standard residue composition by checking for the twenty canonical amino acids and four standard RNA bases. Processing validation monitors feature extraction success by confirming the

identification of protein-RNA interactions, evaluates prediction confidence through ensemble variance analysis, and tracks computational resource utilization to prevent system overload. Output validation ensures result range verification by checking that predicted scores fall within thermodynamically reasonable bounds, performs statistical consistency assessment across multiple structures in batch queries (maximum PDB IDs input is limited to 10), and validates file format integrity for downloadable results.

4.2.6. Output Processing Workflow

Upon submission, the service processes the structure/s as described above and returns detailed results including overall binding assessment and statistical summaries (the latter are available only for batch queries; all other queries, such as single PDB submissions or user-uploaded files, will display null values in this field) as shown in results section. The output presentation focuses on PANTHER scores, expressed in kcal/mol units with thermodynamic significance, accompanied by binding affinity categorization based on established thermodynamic thresholds. Strong favorable binding corresponds to scores below -10 kcal/mol, indicating high-affinity interactions likely to form stable complexes under physiological conditions. Moderate favorable binding encompasses scores between -10 and -5 kcal/mol, representing interactions of biological relevance that may require specific cellular contexts for stability. Weak or neutral binding includes scores between -5 and -2 kcal/mol, suggesting marginal interactions that likely require cofactors or specific environmental conditions. Unfavorable binding covers scores above -2 kcal/mol, indicating interactions unlikely to form stable complexes under normal physiological conditions. Statistical summaries provide comprehensive analysis including mean PANTHER scores across all PDB IDs queried in a batch of uu to 10 structures, standard deviations indicating score variability, and binding category distributions, showing the proportion of structures in each thermodynamic class mentioned above.

4.3. Results and Discussion

The underlying PANTHER score machine learning model was previously evaluated using a stress set of 110 non-redundant protein-RNA complexes which contain experimentally determined binding affinities (see Chapter 3). This stress set, assembled from diverse literature sources, represented a rigorous challenge for model generalization across varied structural and functional classes of protein-RNA interactions. The benchmark results with the stress set demonstrated PANTHER's better predictive accuracy compared to existing web-accessible tools. The PANTHER score achieved a Pearson correlation coefficient (r) of 0.64 ($\rho = 6.02\times10^{-14}$) when compared with the

respective experimental binding free energies, substantially outperforming PredPRBA (r = 0.18, ρ = 0.05) and PRA-Pred (r = 0.18, ρ = 9.19×10^{-3}) . Moreover, the Spearman rank correlation analysis also yielded consistent results, PANTHER score achieved r_s = 0.61 compared to r_s = 0.19 and r_s = 0.25 for PredPRBA and PRA-Pred respectively. Mean absolute error also confirmed PANTHER score's enhanced accuracy (MAE = 1.98 kcal/mol) compared to PredPRBA (MAE = 2.16 kcal/mol) and PRA-Pred (MAE = 2.10 kcal/mol). Thus, PANTHER score represents a 3.5-fold improvement in predictive. Furthermore, we expanded the stress set from 110 to 155 complexes.

4.3.1. Benchmarking

4.3.1.1. Dataset Preparation and Computational Pipeline

We used the PANTHER score web service to make predictions on protein complexes. We refer to the original 110 complexes as the stress set. We then added 45 new complexes (Table 4-1), which we call the extended stress set. Altogether, the combined 155 complexes are referred to as the total stress set. The extended dataset was assembled from the PDBbind v. 2020 release, which includes 1,030 protein-nucleic acid complexes with experimentally determined binding affinities. To prepare these data for analysis, we implemented a computational pipeline in Python (v3.8+) incorporating BioPython (v1.79) for structural parsing. Protein and RNA components were systematically separated. Proteins were restricted to the 20 standard amino acids and RNA chains limited to the ribonucleotides (A, U, G, C).

Applying the above pipeline mentioned above to the initial set of 1030 complexes we restricted protein-DNA complexes (492, 47.7%), mixed complexes containing multiple nucleic acid types (314, 30.4%) were also excluded. We retained protein–RNA complexes (200, 19.4%) for further analysis. Among these protein–RNA complexes those complexes containing ligands (14, 1.4%) were conditionally accepted only if the ligands were not located within the protein–RNA binding interface. Additional exclusions included peptide complexes (7, 0.7%) with insufficient protein content, nucleic acid–only structures (3, 0.3%) lacking a protein partner, and one unclassified entry (0.1%) removed due to structural ambiguity, finally considering 214 complexes. Furthermore, out of the 214 structurally validated protein–RNA complexes, quantitative binding affinity data were available for 187 (87.4%) eliminating further 27 complexes. To prevent data overlapping and to ensure independent validation, 48 complexes (25.7%) that overlapped with the already studied stress set were removed. From the remaining 139 complexes, we applied the seven selection criteria to assemble a final extended stress set of 45 protein–RNA complexes (Table 4-1). The final criteria ensured both structural quality and biochemical consistency: (i) structures

determined by X-ray crystallography with resolution better than 3 Å (12, 26.7%) or by NMR spectroscopy (33, 73.3%); (ii) inclusion of only single-stranded RNA to avoid complications from double-stranded regions; (iii) exclusion of inhibitors and metal ions to focus solely on protein–RNA interactions; (iv) restriction to standard amino acids and nucleotides to maintain uniform chemical environments; (v) omission of zinc-finger proteins to eliminate metal-dependent binding artifacts; (vi) approximately balanced representation of the four RNA bases (A, U, G, C) to reduce sequence composition bias; and (vii) broad coverage of RNA-binding protein folds (12 distinct domain families) and taxonomic diversity to enhance the general applicability of the dataset.

Table 4-1. Extended Stress Data Set (45 new complexes added), showing the experimental binding free energies (ΔG, kcal/mol) and the predicted PANTHER scores (kcal/mol), along with the corresponding values achieved from PredPRBA and PRA-Pred web services.

Extended Stress Set	Experimental ΔG kcal/mol	PANTHER score kcal/mol	PredPRBA kcal/mol	PRA-Pred kcal/mol
1FJE	-11.87	-9.96	-8.82	-6.87
1G59	-7.27	-8.07	-13.98	-10.01
1QFQ	-10.50	-9.55	-9.12	-9.08
1S03	-9.14	-6.67	-14.12	-8.52
1U6P	-9.72	-8.14	-13.47	-7.55
1WWE	-9.23	-8.06	-8.15	-6.41
1WWG	-9.58	-7.49	-7.81	-6.61
2JPP	-9.65	-9.19	-10.52	-9.00
2KFY	-8.70	-8.34	-8.74	-7.85
2KG0	-7.28	-5.91	-8.74	-8.22
2KG1	-8.70	-7.50	-8.73	-7.75
2KXN	-7.70	-7.09	-8.03	-6.73
2L2K	-8.77	-7.16	-11.66	-9.64
2MFC	-7.80	-8.82	-8.78	-10.18
2MFE	-9.18	-9.61	-8.78	-10.16
2MFF	-7.94	-8.13	-8.65	-8.76
2MFG	-7.55	-8.26	-8.65	-9.46
2MFH	-7.44	-9.12	-9.20	-7.85
2MJH	-9.84	-7.66	-9.85	-8.45
2MKI	-6.55	-5.90	-9.26	-7.34
2MKK	-7.91	-7.67	-9.25	-8.39
2MQV	-10.21	-8.86	-13.45	-7.67
2MTV	-9.53	-8.74	-9.31	-9.37
2MXY	-7.77	-6.94	-8.15	-6.48
2MZ1	-6.76	-7.83	-8.47	-6.18
2RQC	-7.80	-7.99	-7.96	-8.32

2RU3	-8.52	-7.91	-7.81	-8.65
2YH1	-8.02	-7.07	-8.99	-7.83
Extended Stress Set	**Experimental ΔG kcal/mol**	**PANTHER score kcal/mol**	**PredPRBA kcal/mol**	**PRA-Pred kcal/mol**
3BX3	-9.88	-7.71	-9.86	-9.39
3K4E	-11.63	-8.97	-10.6	-11.37
3QSU	-10.52	-8.15	-9.56	-8.57
4BS2	-10.33	-7.91	-9.46	-8.30
4YHW	-9.49	-8.46	-14.13	-10.32
5HSW	-5.88	-6.04	-12.10	-9.85
5MPG	-8.91	-7.63	-8.15	-6.99
5MPL	-8.55	-9.77	-7.87	-9.18
5U9B	-9.39	-6.81	-9.29	-9.55
5WWE	-8.56	-6.33	-9.34	-7.63
5WWF	-9.04	-8.42	-9.28	-7.35
5YTT	-7.40	-4.78	-7.81	-7.28
6G99	-6.18	-7.34	-8.47	-7.54
6GBM	-5.56	-7.19	-9.77	-7.95
6HYU	-9.41	-7.66	-9.60	-11.44
6NOD	-10.29	-8.29	-8.70	-10.84
6SO9	-7.15	-6.96	-9.21	-7.95
Total	**45**	**45**	**45**	**45**
r		**0.54**	**0.14**	**0.24**

4.3.1.2. Extended stress set characteristics

The newly added 45 complexes represent diverse biological systems from multiple domains, including viral complexes (Moloney murine leukemia virus, bacteriophages), bacterial systems (Thermus thermophilus), and eukaryotic assemblies (human, plant). The RNA targets include transfer RNA (tRNA), ribosomal RNA (rRNA), messenger RNA elements such as splicing enhancers, viral RNA packaging signals, and regulatory motifs like boxB or G-tracts. Experimental binding affinities span a physiologically relevant range with K_D values from 2.0 nM to 84.0 µM. The dataset comprises 44 complexes (97.8%) with K_D values and 1 complex (2.2%) with the IC_{50} value. Converting to free energy units, the strongest binding exhibits ΔG = -12.98 kcal/mol and the weakest binding shows ΔG = -3.75 kcal/mol. This dynamic range captures the full spectrum of protein-RNA binding strengths from very tight complexes (nanomolar) to moderate interactions (micromolar). Analysis of predictive performance on this extended set (Figure 4-3, 4-4, and Table 4-2) further demonstrates that while all three models capture general binding trends, they differ in accuracy and robustness. The PANTHER score shows the strongest correlation with experimental ΔG values ($r = 0.53$, $r_s = 0.50$, $\rho = 4.07 \times 10^{-4}$), indicating reliable predictive power across diverse systems. PRA-Pred exhibits

moderate performance (r = 0.23), while PredPRBA shows weaker correlations (r = 0.14) and larger prediction errors, suggesting limited generalizability to more structurally diverse complexes. The mean absolute errors (MAE) support this trend, with PANTHER score (MAE = 1.27 kcal/mol) outperforming PredPRBA (MAE = 1.59 kcal/mol) and PRA-Pred (MAE = 1.32 kcal/mol). These results highlight the strengths of PANTHER score in handling complex and heterogeneous RNA-binding scenarios, while also revealing areas where other predictors may underperform when faced with biologically diverse datasets.

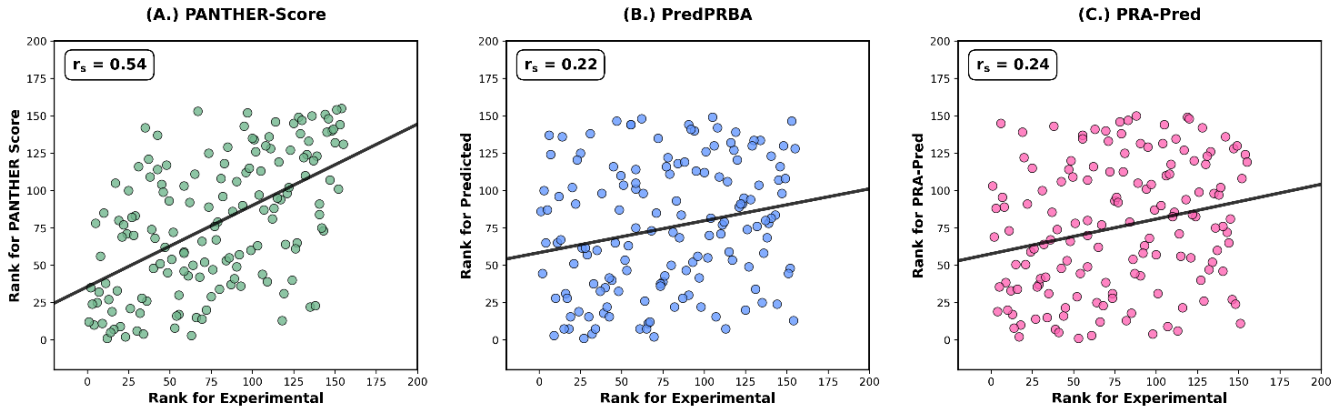

Figure 4-3. Scatter plots showing the linear correlations (as expressed by r values) for 45 complexes, between experimental ΔG values and computed PANTHER scores, PRA-Pred and PredPRBA respectively.

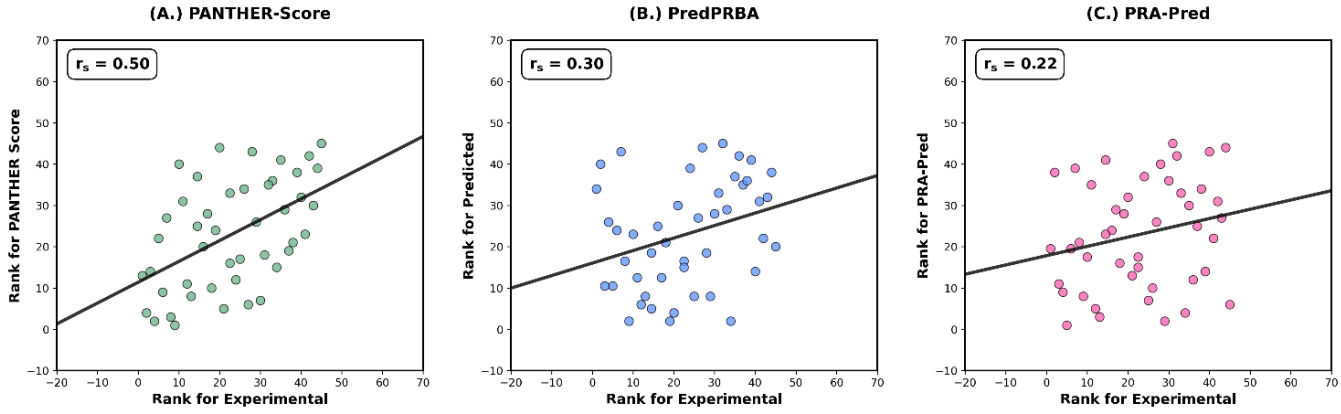

Figure 4-4. Scatter plots show the Spearman rank correlation (r_s) for 45 complexes, between experimental ΔG ranks and ranks of PANTHER score, PRA-Pred and PredPRBA respectively.

Table 4-2. Performance metrics of correlating for extended stress set of the PANTHER scores, PredPRBA and PRA-Pred with known experimental ΔG values. Metrics include correlation coefficient (r), mean absolute error (MAE), Spearman's rank correlation coefficient (r_s), and ρ-value. PANTHER Score demonstrates the highest predictive accuracy on total stress set.

	Extended Stress Set			
Model	**r**	**MAE**	**r_s**	**ρ Value**
PANTHER Score	0.53	1.27	0.50	4.07×10^{-4}
PredPRBA	0.14	1.59	0.30	4.30×10^{-2}
PRA-Pred	0.23	1.32	0.22	1.38×10^{-1}

4.3.2. Performance Evaluation and Statistical Analysis

We evaluated the PANTHER score's web service's output against the existing protein-RNA binding affinity prediction web services for PredPRBA and PRA-Pred, using the total stress set containing 155 diverse protein-RNA complexes. Data shown in Table 4-3 represents the comparative performance metrics for all three methods evaluated here.

Table 4-3. Performance metrics of correlating the PANTHER scores, PredPRBA and PRA-Pred with known experimental ΔG values. Metrics include correlation coefficient (r), mean absolute error (MAE), Spearman's rank correlation coefficient (r_s), and ρ-value. PANTHER Score demonstrates the highest predictive accuracy on total stress set.

	Total Stress Set			
Model	**r**	**MAE**	**r_s**	**ρ Value**
PANTHER Score	0.57	1.75	0.54	2.45×10^{-13}
PredPRBA	0.18	2.01	0.22	6.91×10^{-3}
PRA-Pred	0.16	1.89	0.24	3.39×10^{-3}

PANTHER score demonstrated better performance across multiple metrics. The method achieved a Pearson correlation coefficient of (r) of 0.57, representing a 3.2-fold improvement over PredPRBA (r of 0.18) and 3.6-fold improvement over PRA-Pred (r of 0.16). PANTHER score also exhibited the lowest mean absolute error (MAE of 1.75 kcal/mol) compared to PredPRBA (2.01 kcal/mol) and PRA-Pred (1.89 kcal/mol). The Spearman correlation derived from the data shown in Table 4-2 analysis also demonstrates PANTHER score to have achieved better results with respect

to it's counterparts with r_s of 0.54, outperforming both PredPRBA (r_s = 0.22) and PRA-Pred (r_s = 0.24). Notably, PANTHER score exhibited better statistical significance ($\rho = 2.45 \times 10^{-13}$), approximately 10 orders of magnitude more significant than the competitors having 6.91 x 10^{-3} and 3.39 x 10^{-3} for PredPRBA and PRA-Pred respectively (data shown in Table 4-1). Data plotted for the total stress set (155 complexes; see Table 4-1 for the extended stress set from this article and Table 3-5 of previous chapter for PANTHER score values), as shown in Figure 4-5, highlights clear differences in linear correlation performance. PANTHER score predictions show a noticeable linear relationship with experimental energies (r = 0.57), with tighter clustering around the regression line across the full ΔG range (-13 to -4 kcal/mol), with relatively tight clustering around the regression line and minimal systematic bias (r =0.57). In contrast, both PredPRBA (r = 0.18) and PRA-Pred (r = 0.16) exhibit substantial scatter with poor correlation with experimental values. PredPRBA predictions span roughly -13 to -4 kcal/mol but are broadly dispersed, with many points around -9 kcal/mol regardless of experimental binding strength. PRA-Pred predictions also range from -13 to -4 kcal/mol, but the data exhibits wide scatter, with no distinct clustering pattern and little apparent relationship to experimental affinities. Figure 4-6, represents the Spearman rank correlation analysis, re-confirms PANTHER score's ability to correctly order protein-RNA complexes by binding strength (r_s = 0.54). The competing methods show poor ranking performance, with PredPRBA (r_s = 0.22) and PRA-Pred (r_s = 0.24), demonstrating limited ability to distinguish between strong and weak binders. A case study with 6FPQ as an example to use the model through the web service is presented in the next section.

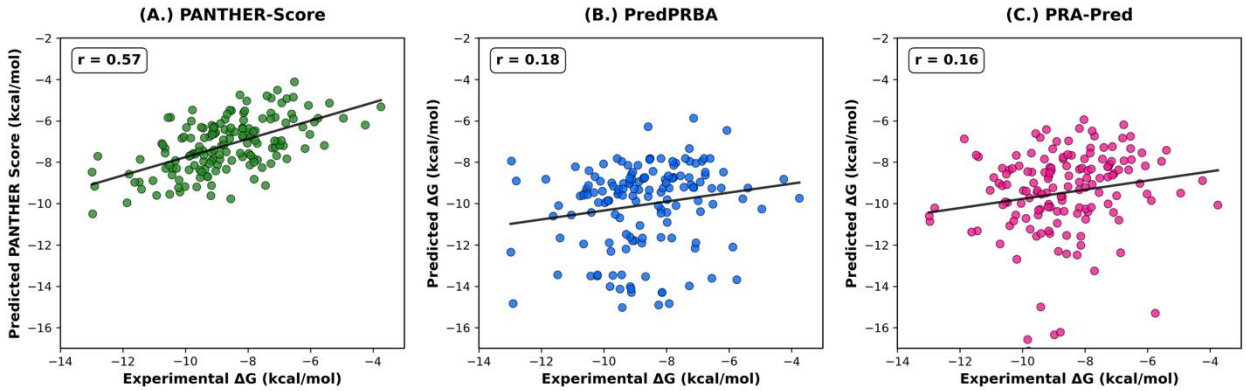

Figure 4-5. Scatter plots showing the linear correlations for total stress set of 155 complexes, between experimental ΔG values and computed PANTHER Score, PredPRBA and PRA-Pred respectively.

Comparing PANTHER score with existing methods demonstrates an improvement in predicting protein-RNA binding affinities, corresponding to a 3.2-3.6-fold improvement in linear correlation relative to alternative approaches. The statistical significance of this correlation ($\rho = 2.45 \times 10^{-13}$) indicates the robustness of the predictions, effectively ruling out the possibility of false associations. This enhanced performance can be attributed to PANTHER score's local-to-global learning strategy, which decomposes binding free energy into amino acid-nucleotide interaction features derived from molecular dynamics simulations and integrates them into a predictive framework. To rigorously evaluate model generalization, we implement multi-tiered validation strategy (test set → stress set → extended stress set → total stress set). The total stress set, composed of non-curated structures without manual refinement of the crystal data, represents real-world conditions and demonstrates the robustness of our approach in handling structural heterogeneity and variations in data quality. Validation on the total stress set of 155 complexes, including both crystallographic and NMR structures, confirms broad applicability across different experimental conditions. Importantly, the consistent predictive accuracy observed across a resolution range of 1.5-3.0 Å indicates that PANTHER score is not sensitive to coordinate precision, thereby reinforcing its robustness and utility for diverse structural datasets.

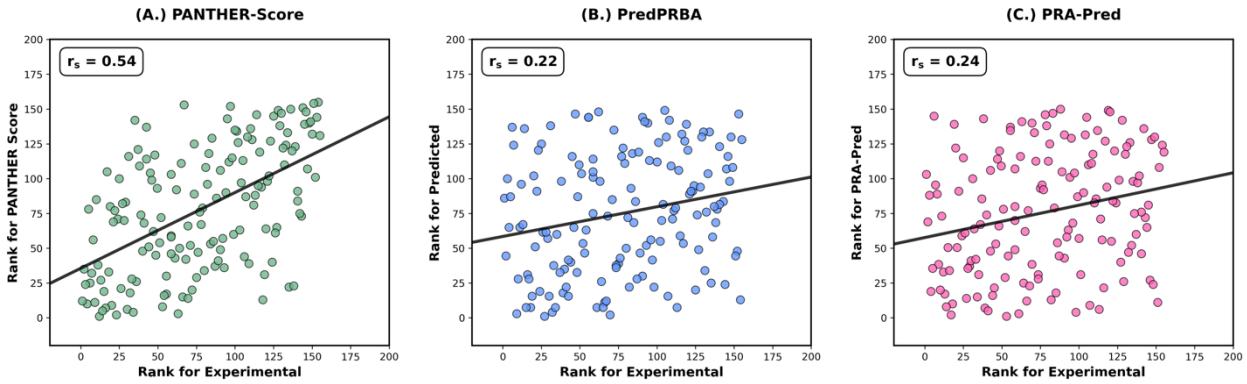

Figure 4-6. Scatter plots showing the Spearman rank correlation (rs) for total stress set of 155 complexes, between experimental ΔG ranks and ranks of PANTHER Score, PredPRBA and PRA-Pred respectively.

Residual analysis further supports PANTHER score's performance. Prediction errors were normally distributed and centered near zero, indicating unbiased estimates across binding affinity ranges. Studying the total stress set, we observe that PANTHER score maintained consistent accuracy for both strong binders (ΔG < -10 kcal/mol, 21 complexes, 13.5% of the dataset) and moderate binders (-10 ≤ ΔG ≤ -5 kcal/mol, 120 complexes 77.4%). The method also performed

reliably for weak binders (–5 < ΔG ≤ –2 kcal/mol; 14 complexes, 9.03%), without evidence of systematic over- or underestimation. These results demonstrate that PANTHER score achieves balanced predictive performance across the full experimental ΔG spectrum. In contrast, PredPRBA exhibited the opposite trend, consistently predicting overly favorable binding (mean -9.59 kcal/mol), resulting in significant bias. Whereas, PRA-Pred systematically overestimated binding strength, with an average predicted ΔG of -8.46 kcal/mol compared to the experimental mean of -8.65 kcal/mol. The 2.3-fold improvement in correlation coefficient achieved by PANTHER score relative to Pred-PRBA & PRA-Pred demonstrates its reliability and robustness in capturing the thermodynamic diversity of protein-RNA recognition.

The observed performance differences have important implications for the choice of prediction method. Based on our analyses, the PANTHER score appears particularly well-suited for applications requiring reliable estimates of protein-RNA binding affinities. For example, in Drug Discovery, the method may assist in ranking RNA-targeting compounds to support lead optimization. In Structural Biology, it could guide mutagenesis experiments and aid in the interpretation of binding data. Within biology systems, the approach offers the potential to enable large-scale analyses of RNA-protein interaction networks. Finally, in synthetic biology, the PANTHER score may provide a useful tool for designing orthogonal RNA-protein pairs to support cellular engineering efforts.

PANTHER scores validate the local-to-global strategy for biomolecular interaction prediction. The approach addresses the challenge of limited experimental training data by generating training data through molecular dynamics simulations to generate additional input and by learning transferable amino acid–nucleotide interaction patterns that generalize across diverse complexes. The 10-fold difference in statistical significance between PANTHER score and competing methods highlights the importance of rigorous validation frameworks and physically grounded methodologies in computational biology. Future developments could focus on incorporating dynamic effects, allosteric regulations, and cooperative binding phenomena, all of which are known to significantly influence biomolecular recognition. The substantial performance improvement demonstrated in this study not only establishes a new benchmark for protein-RNA binding prediction but also provides a conceptual framework for the development of next-generation predictors of biomolecular interactions across a wide range of systems.

4.3.3. Interface with Case Study

Figure 4-7 illustrates the graphical user interface and the output of the PANTHER web service, using *6FPQ* as an input example. On the left panel (Figure 4-7, A) users can submit a structure either by entering a single PDB ID (e.g., 6FPQ), providing multiple PDB IDs (up to 10 per run), or uploading a custom PDB file. Once the structure is submitted, the service processes the input, generates binding affinity predictions, and produces the PANTHER score, as shown in the right panel of Figure 4-7, B. The right panel displays the output summary, which includes the mean binding free energy score (ΔG), the number of favorable versus unfavorable predictions, and statistical measures such as standard deviation, best, and worst scores. It is to be noted that the statistical measures are only useful when a user provides a list of PDB IDs to calculate in batch. For PDB ID 6FPQ, the predicted binding free energy was –7.05 kcal/mol, classified as strong favorable binding (ΔG < –5 kcal/mol). Additional details include the method description, highlighting the use of a random forest regression model with 100 decision trees trained on protein–RNA complexes, and the feature set comprising distance metrics, amino acid types, hydrogen bonds, and nucleotide types. Finally, a score interpretation guide provides an intuitive scale to classify predictions into strong, moderate, weak/neutral, or unfavorable binding affinities. In addition, Figures 4-8, 4-9, and 4-10 show the step-by-step workflow that guides users through the web service.

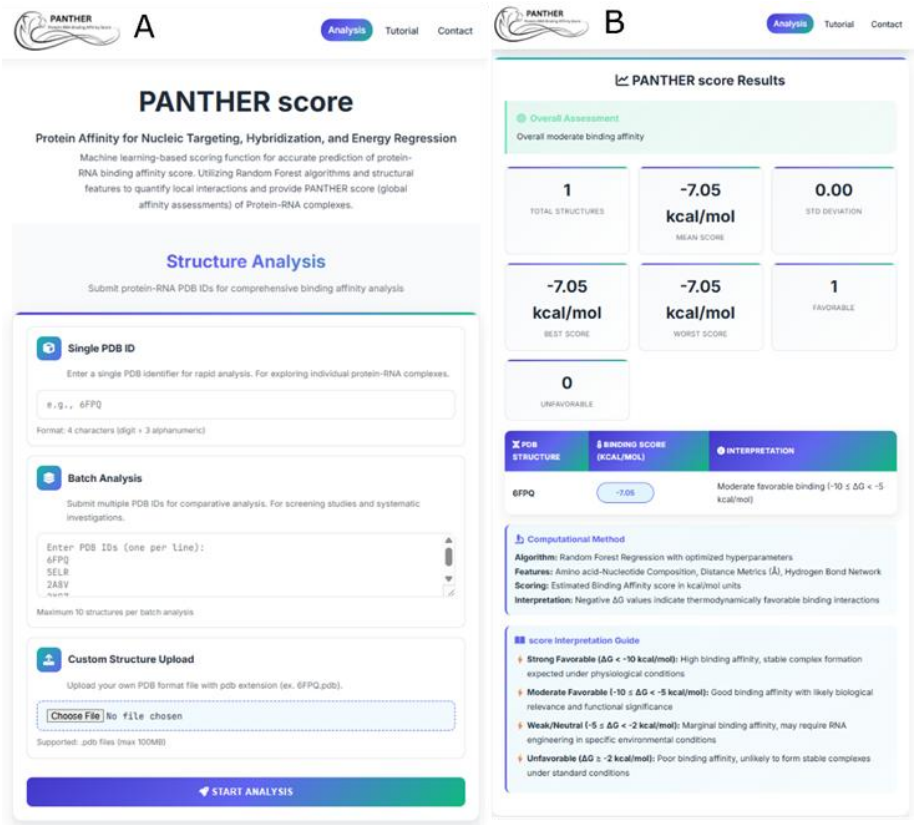

Figure 4-7. Interface and case study example of the PANTHE Score website.

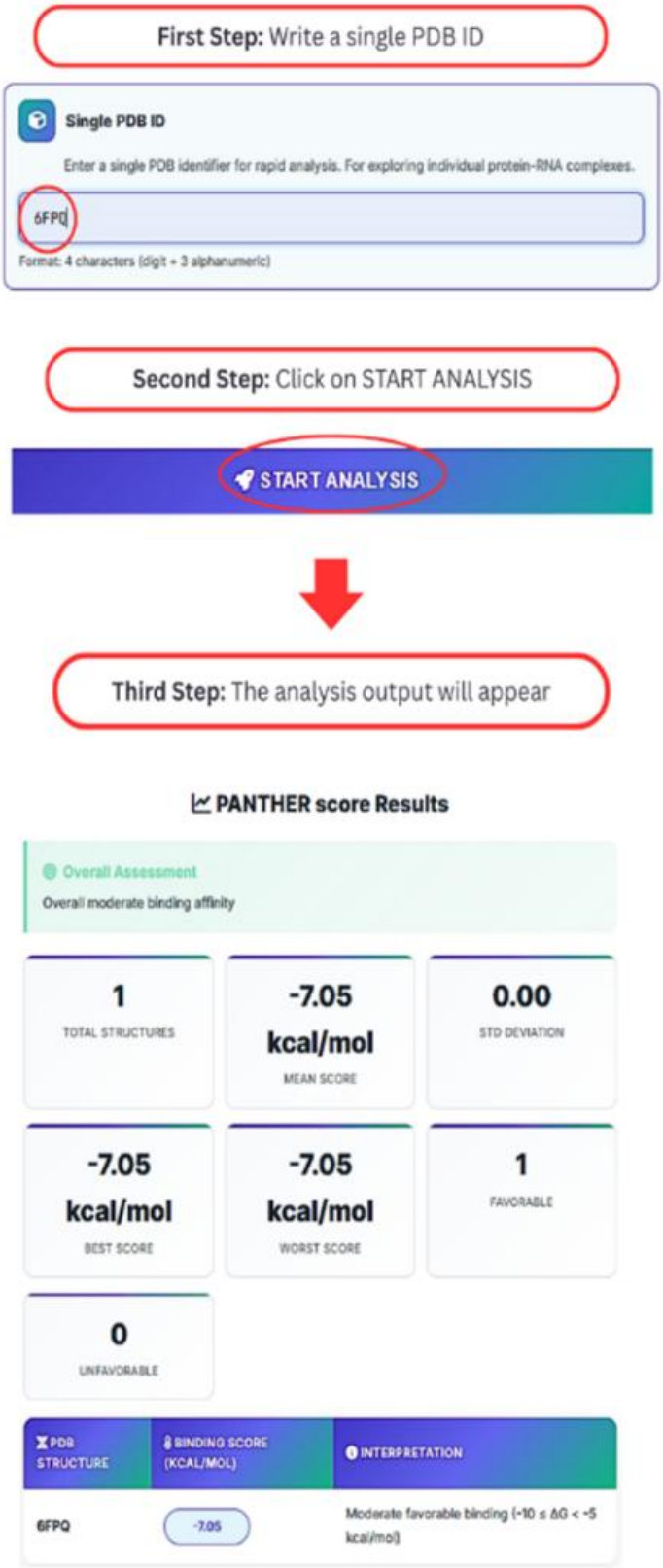

Figure 4-8. Step-by-step guide for submitting a single PDB ID to the PANTHER Score web service.

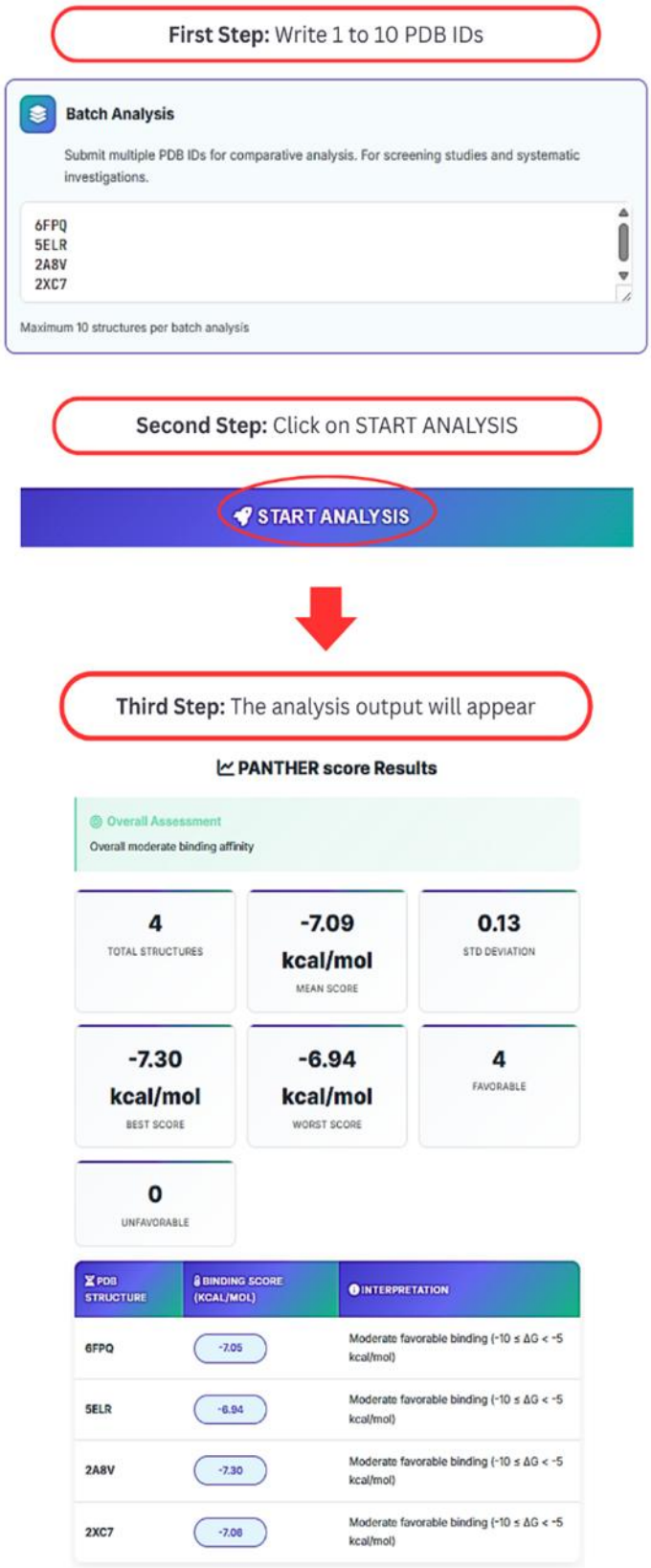

Figure 4-9. Instructions for batch processing multiple PDB IDs (up to 10) using the PANTHER Score web service.

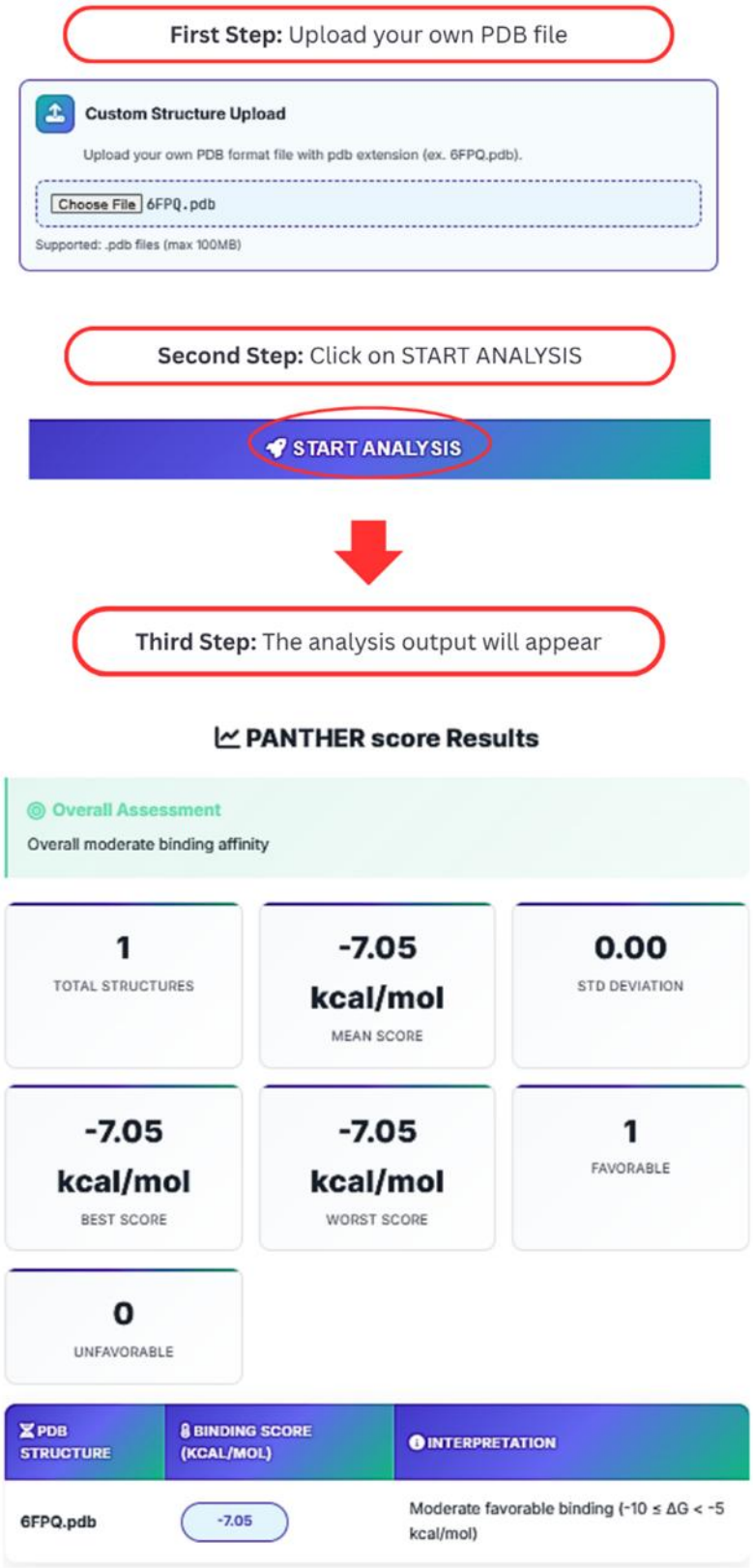

Figure 4-10. Guide for analyzing user-uploaded protein-RNA complex files.

Chapter 5

Multi-State RISoTTo: Context-Aware RNA Sequence Design Across Conformational Ensembles

5.1. Introduction

RNA design has emerged as a critical discipline in synthetic biology, enabling the development of functional RNA molecules for therapeutic applications, biosensors, and gene regulation systems [238], [239].The fundamental challenge lies in solving the inverse folding problem: designing RNA sequences that fold into desired three-dimensional structures and maintain biological function within their native molecular environments. Traditional RNA design approaches have predominantly focused on secondary structure prediction, often neglecting tertiary structural constraints and the broader molecular context in which RNAs operate. Early methods, such as *ViennaRNA* [240] and thermodynamic optimization approaches [241], while foundational, are limited by their reliance on two-dimensional base-pairing patterns and their inability to account for three-dimensional geometric constraints or molecular interactions. Physics-based methods like *Rosetta's fixed-backbone redesign protocol* [242], [243] represent significant advances in incorporating tertiary structure information yet remain computationally expensive and limited in their ability to account for dynamic conformational behavior. Recent advances in geometric deep learning have revolutionized structure-based RNA design [244]. *RISoTTo* (RIbonucleic acid Sequence design from TerTiary structure) [245] represents a significant advancement in this field, introducing a context-aware geometric transformer model that conditions sequence prediction on both RNA backbone geometry and surrounding molecular partners, including proteins, small molecules, DNA, and ions. Building upon the *CARBonAra* protein design framework [246], RISoTTo demonstrated superior performance in native sequence recovery compared to existing methods, achieving 62% average recovery on benchmark RNA structures, outperforming Rosetta (45%) and contemporary deep learning approaches like *RDesign* (43%) [247]. This performance advantage is especially evident at molecular interfaces, where contextual information is critical for functional RNA design. However, RISoTTo operates under the assumption of static structural scaffolds, treating each RNA backbone as a fixed geometric entity. However, this approach overlooks a fundamental characteristic of RNA molecules: their intrinsic conformational flexibility [248], [249]. Many functional RNAs, including riboswitches, ribozymes, and regulatory elements, adopt multiple distinct conformational states to

execute their biological roles. Riboswitches, for example, undergo ligand-induced conformational transitions between apo and holo states [250], while ribozymes exhibit dynamic folding pathways essential for catalytic activity. The importance of multi-state modeling has been demonstrated by *gRNAde* [251], which introduced a multi-state graph neural network architecture capable of processing conformational ensembles. By representing multiple backbone conformations as geometric multi-graphs and employing conformational state order-invariant pooling through Deep Set operations [252], gRNAde showed that explicit ensemble modeling improves sequence recovery by 3-5% for structurally flexible RNAs compared to single-state variants. The model demonstrated effectiveness particularly for surface nucleotides undergoing positional or secondary structural changes across conformational states. However, gRNAde's approach is limited to isolated RNA molecules and does not incorporate the contextual molecular information that RISoTTo has shown to be crucial for accurate design. Contemporary methods such as *RiboDiffusion* [253] and *RhoDesign* [254] have explored alternative generative approaches for RNA design, employing diffusion models and structure-guided generation respectively. While these methods offer novel perspectives on the inverse folding problem, they similarly focus on single-state design scenarios and do not address the dual requirements of conformational flexibility and molecular context awareness. This limitation presents a significant gap in current RNA design methodologies: while RISoTTo excels at context-aware design for single conformations and gRNAde addresses multi-state flexibility, no existing framework combines both capabilities. Given that functional RNAs in biological systems often exhibit conformational flexibility while simultaneously interacting with diverse molecular partners, an integrated approach is essential for advancing the field toward biologically relevant design scenarios. To address this limitation, we introduce a multi-state extension of the RISoTTo framework that preserves its context-aware geometric transformer architecture while incorporating conformational ensemble handling. Drawing inspiration from gRNAde's multi-state graph neural network approach and ProteinMPNN's multi-backbone protein design strategies [255], our method introduces two distinct fusion strategies for multi-state information integration. Feature-level fusion enforces conformation-invariant embeddings by averaging geometric features across conformational states before sequence prediction, capturing conserved structural motifs essential for function. Logit-level fusion independently processes each conformational state through the geometric transformer and combines prediction logits, allowing the model to weight individual conformations based on their relevance to sequence compatibility. This dual-strategy approach

provides both a technical advancement in multi-state RNA design and establishes a foundation for context-aware design of conformationally flexible RNA molecules.

5.2. Materials and Methods

5.2.1. Overview of RISoTTo Architecture

RISoTTo (RIbonucleic acid Sequence design from TerTiary structure) [245] implements a context-aware geometric transformer operating on atomic point clouds. Each RNA backbone is represented using a coarse-grained four-bead model corresponding to atoms P, C4', O2', and N1 (for pyrimidines) or N9 (for purines). This representation captures RNA backbone conformation efficiently while maintaining physical interpretability. Surrounding non-RNA atoms, including those from proteins, DNA, small molecules, and ions, are included to provide contextual cues essential for accurate sequence prediction. Atomic element types are one-hot encoded and projected into a 64-dimensional scalar feature space using a three-layer perceptron. These features are processed through 20 geometric transformer (GT) layers, each performing message passing among progressively expanding neighborhoods (8, 16, 32, and 64 nearest neighbors). Each GT layer updates scalar and vector features using rotation-equivariant attention operations features using rotation-equivariant attention operations as shown in equation 5-1 [245].

$$q_i^{l+1}, p_i^{l+1} = \mathrm{GT}\left(q_i^l, p_i^l, \{q_j^l, p_j^l, D_{ij}, R_{ij}\}_{j \in \mathrm{nn}(i)}\right) \tag{5.1}$$

where D_{ij} denotes the Euclidean distance and R_{ij} the normalized displacement vector between atoms i and j. SE(3)-equivariance is preserved throughout by maintaining vector attention consistency across rotations.

Following geometric encoding, residue pooling aggregates atomic embedding into nucleotide-level representations via a four-headed self-attention mechanism. The pooled embeddings are decoded through a multilayer perceptron (MLP) into position weight matrices (PWMs), which define per-nucleotide probability distributions over A, U, G, and C. The overall architecture is illustrated in Figure 5-1.

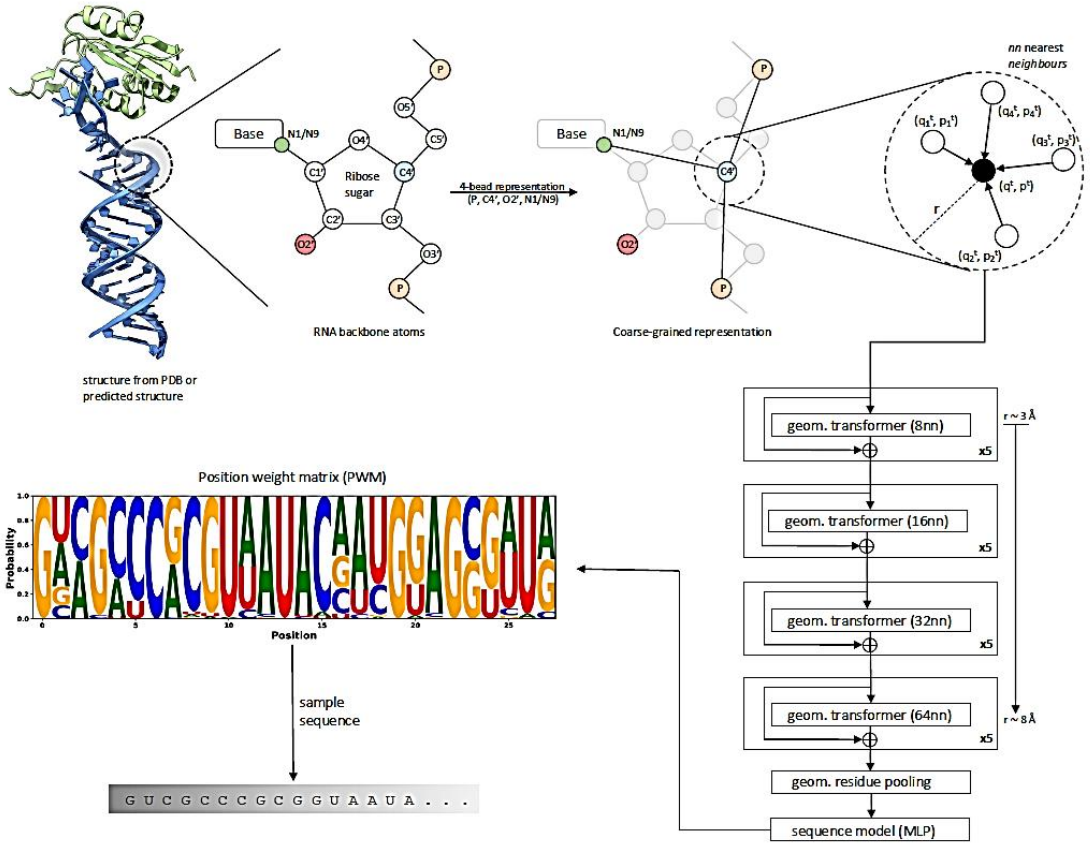

Figure 5-1. Overview of RISoTTo workflow for context-aware RNA design. The RNA backbone is described using a 4-bead coarse-grained representation, capturing the coordinates of four atoms per nucleotide: P, C4', O2', and either N1 (for pyrimidines) or N9 (for purines). The model comprises 20 layers of geometric transformers with residual connections, progressively expanding the neighborhood size from 8 to 64 nearest neighbors. Structural information is aggregated into a residue-level representation through transformer-based geometric pooling. The residue-level representations are aggregated and finally passed through a multilayer perceptron (MLP) to generate the final position weight matrix (PWM) [238]

5.2.2. Multi-State Dataset Preparation

5.2.2.1 Base RISoTTo Dataset

The dataset of RISoTTo comprises RNA-containing tertiary structures from the Protein Data Bank (PDB)[256], and predicted structures generated by *trRosettaRNA* [257]. The dataset includes standalone RNAs, RNA–protein complexes, and RNA–DNA hybrids, totaling approximately 14,000 RNA subunits for training, 693 for validation, and 379 for testing. For RNAs unavailable in legacy PDB format, structures were retrieved from the *RNASolo* database [258]. Dataset splitting followed the protocol described by Joshi et al. [251], ensuring structural dissimilarity between sets via *TM-score* thresholds computed with *US-align* [259], a universal alignment tool that compares 3D structures

of nucleic acids and proteins. Structures with TM-score ≥ 0.45 between training and test sets were excluded to prevent data leakage and overestimation of generalization.

5.2.3 Construction of Multi-State Ensembles

To adapt the RISoTTo dataset for multi-state modeling, we adopted a protocol analogous to that used in gRNAde [251], ensuring consistency in structural diversity and non-redundancy across conformational ensembles. First, all RNA-containing chains from the combined PDB and trRosettaRNA sources were clustered by sequence identity using *CD-HIT* [260] at a threshold of ≥99%, grouping conformers sharing identical primary sequences into putative structural ensembles. Each cluster thus represents multiple experimentally resolved or predicted conformations of the same RNA sequence, including apo/holo riboswitch pairs and catalytic ribozyme intermediates. Within each cluster, conformers were aligned in all-versus-all mode using *US-align* [259], and the resulting pairwise TM-score matrix was used to quantify structural heterogeneity. Ensemble diversity was defined by the median intra-cluster TM-score, and highly redundant ensembles (mean TM > 0.9) were down-sampled to avoid overrepresentation of rigid RNAs. Conversely, clusters with TM < 0.45 were excluded to ensure sufficient structural correspondence between conformers for learning meaningful geometric correlations. Following alignment, each RNA ensemble was assigned to the train, validation, or test partitions using structure-level clustering, ensuring no overlapping folds occurred between dataset splits. Specifically, ensembles sharing any conformer with TM-score ≥ 0.45 to a member of another partition were excluded from that set, preventing structural leakage and overestimation of generalization. The final dataset preserved the proportional split of RISoTTo (approximately 14,000 training, 693 validation, and 379 test subunits), while guaranteeing that no conformational variants or near-duplicate backbones occurred across partitions. To maintain computational feasibility, complexes with >12,288 atoms or incomplete RNA segments (<20 nucleotides) were removed, and only structures resolved at ≤4.0 Å were retained. This pipeline, adapted from gRNAde [251], thus produced a curated and conformationally diverse dataset suitable for evaluating ensemble-aware RNA design models while maintaining full compatibility with the RISoTTo architecture and data handling framework.

5.2.4. Multi-State Architectural and theoretical foundations

The central contribution of this work is the extension of RISoTTo [245], to process conformational ensembles rather than single backbones. We developed and compared two

complementary fusion strategies: feature-level fusion, logit-level fusion and Independent Conformer Processing with Deep Set Pooling [252], which is shown in figure 5-2.

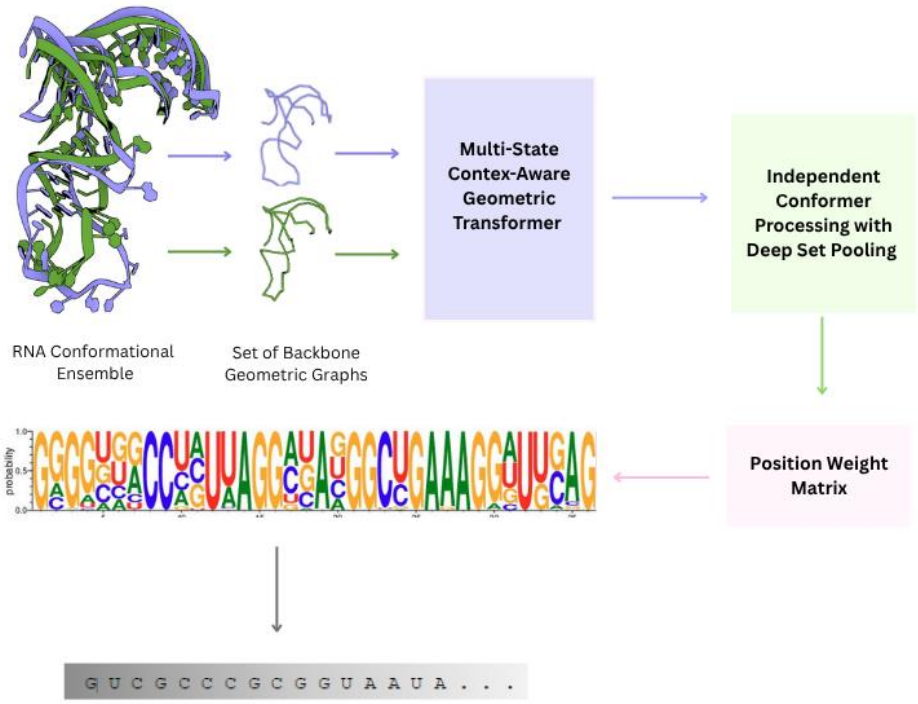

Figure 5-2. Multi-state RNA sequence prediction, with independent conformer processing and deep-Set pooling method.

5.2.4.1. Feature-Level Fusion

Feature-level fusion [255], integrates conformational information early in the network, before sequence decoding. Each conformer C_i is first passed independently through RISoTTo's embedding and geometric transformer layers to produce scalar and vector features F_i, describing the atomic geometry of that state as is in equation 5.2 [246].

$$F_i = \mathrm{GT}(C_i) \tag{5.2}$$

These per-conformer features are then averaged across the ensemble to obtain a single conformation-invariant representation (equation 5.3).

$$F_{\text{fused}} = \frac{1}{K}\sum_{i=1}^{K} F_i \tag{5.3}$$

This arithmetic mean pooling enforces invariance to the ordering of conformers and emphasizes geometric features that remain consistent across all states, such as conserved helices or tertiary

contacts. The fused representation F_{fused} is subsequently passed through the standard RISoTTo residue-pooling and decoding layers to output one position-weight matrix (PWM) per RNA sequence. While conceptually simple, this approach mixes features that may not be perfectly aligned across conformers. When local neighborhoods differ strongly in topology (for example, flexible loops shifting between states), arithmetic averaging can blur informative structural signals, introducing noise into the embedding and slightly reducing recovery accuracy.

5.2.4.2 Logit-Level Fusion

Logit-level fusion instead combines information late in the network, after each conformer has been fully processed and decoded. Each conformer C_i is independently passed through the complete RISoTTo pipeline to produce per-residue logits L_i, which encode the unnormalized probability scores for each nucleotide (A, U, G, C) at each position. To obtain a single ensemble-level prediction, these logits are then merged across conformers using one of two pooling function (equations 2.4 & 2.5).

- **Arithmetic Mean Pooling**

This simple averaging assumes that all conformers contribute equally to the overall design and smooths out minor structural differences (equation 5.4) [255].

$$L_{fused} = \frac{1}{k}\sum_{i=1}^{k} L_i \tag{5.4}$$

- **Log-Sum-Exp Pooling**

This function is a smooth approximation of the maximum, giving proportionally higher weight to conformers with stronger (more confident) nucleotide predictions. In statistical physics terms, it behaves similarly to a Boltzmann-weighted average, where conformers with lower “energy” (higher model confidence) dominate the ensemble contribution (equation 5.5) [255].

$$L_{\text{fused}} = \log\left(\sum_{i=1}^{k} \exp(L_i)\right) \tag{5.5}$$

Compared to feature-level fusion, logit-level fusion preserves each conformer’s full geometric processing path and only combines predictions at the final decision stage. This strategy therefore retains conformer-specific nuances while still producing a single consensus sequence distribution. Both fusion schemes were implemented with minimal changes to RISoTTo’s core pipeline and share the same training configuration and loss functions.

5.2.4.3. Independent Conformer Processing with Deep Set Pooling

While both feature-level and logit-level fusion provide mechanisms to integrate conformational information, they represent two extremes in the fusion hierarchy early and late aggregation. To balance information preservation and ensemble invariance, we implemented a third strategy, *Independent Conformer Processing with Deep Set Pooling* (ICDP–DSP), which treats each conformer as an independent observation processed through the full RISoTTo encoder before performing a mathematically defined, permutation-invariant aggregation. In this configuration, every conformer C_i in the ensemble is processed separately through all geometric transformer layers to obtain residue-level scalar and vector embeddings $(q_a^{(i)}, p_a^{(i)})$ for each nucleotide position a. Let K denote the number of conformers in the ensemble. The Deep-Set pooling operation aggregates these representations across conformers using arithmetic mean pooling (equation 5.6) [252].

$$(q_a, p_a)_{\text{fused}} = \frac{1}{K}\sum_{i=1}^{K}\left(q_a^{(i)}, p_a^{(i)}\right) \tag{5.6}$$

This operation enforces permutation invariance with respect to the order of conformers in the ensemble, a key theoretical requirement of Deep-Set formulations [14]. The fused representation $(q_a, p_a)_{\text{fused}}$ is then passed through RISoTTo's standard residue-pooling and decoding layers to produce the nucleotide probability distribution $P(a \mid \text{ensemble})$. In contrast to feature-level fusion (Eq. 2–3), which mixes embeddings before they are contextually stabilized, and logit-level fusion (Eq. 4–5), which merges predictions post-decoding, ICDP–DSP performs aggregation at the intermediate representation level**,** after complete conformer encoding but before decoding. This allows each conformer's geometry to be fully interpreted by the network, preserving conformer-specific details while still enforcing ensemble consistency. Conceptually, the mean operation in equation 5.7 [261] acts as a statistical ensemble average, analogous to computing an expectation value over structural states.

$$E_C[(q_a, p_a)] = \int_C (q_a, p_a)\ p(C)\, dC \approx \frac{1}{K}\sum_{i=1}^{K}\left(q_a^{(i)}, p_a^{(i)}\right) \tag{5.7}$$

where $p(C)$ denotes the empirical distribution of conformers in the dataset. This probabilistic view links the Deep-Set aggregation to a Boltzmann-like ensemble average over conformations. Practically, the ICDP–DSP approach introduces minimal additional computation and retains compatibility with the RISoTTo training pipeline. It was ultimately selected as the primary

multi-state configuration due to its stable optimization behavior, interpretability, and superior recovery consistency across flexible RNA families. This design philosophy draws from recent advances in set-based neural representations, multi-state graph modeling in gRNAde [251] and multi-backbone protein design approaches such as *ProteinMPNN* [255]. By combining independent geometric encoding with permutation-invariant aggregation, the model captures ensemble-level sequence constraints while remaining grounded in the context-aware transformer architecture introduced in CARBonAra [246]and RISoTTo[245].

5.2.5 Training Configuration and Evaluation Protocol

All models were implemented in *PyTorch* [262], and trained on NVIDIA A100 GPUs (80 GB) [263]. Training hyperparameters matched the single-state configuration:

- Optimizer: Adam [264],
- Learning rate [265]: 1×10^{-4},
- Gradient clipping [265]: 1.0,
- Scheduler: ReduceLROnPlateau [266] (patience = 5 epochs),
- Loss: Binary cross-entropy (BCE) [267] with nucleotide class weighting.

Atomic coordinates were centered, and Gaussian noise [268] (σ = 0.1 Å) was added during training to improve robustness to crystallographic perturbations. Training excluded complexes exceeding memory thresholds or incomplete RNAs (<10 nucleotides).

5.3. Results and Discussion

5.3.1 Overview of Model Evaluation

To evaluate the performance of the multi-state RISoTTo framework, we focused on the Independent Conformer Processing with Deep Set Pooling strategy, which demonstrated the most consistent and biologically meaningful results among all tested configurations. This variant preserves the original RISoTTo geometric transformer pipeline for each conformer while introducing a permutation-invariant aggregation mechanism inspired by Deep-Sets [252]. Each conformer is processed independently through all transformer layers and the decoding head, producing atomic- and residue-level embeddings (q_a^i, p_a^i). These embeddings are then fused via arithmetic mean pooling across conformers as shown in equation 5.8.

$$(q_a, p_a)_{fused} = \frac{1}{k}\sum_{i=1}^{k}(q_{a_i}, p_{a_i}) \quad (5.8)$$

This operation enforces ensemble invariance, meaning that the fused representation does not depend on the order of conformers in the input cluster, aligning with the theoretical guarantees of Deep Sets. Importantly, because each conformer is processed independently, local geometric variations are preserved within each state before aggregation, avoiding the distortion that occurs in direct feature averaging methods. The model achieved a mean sequence recovery accuracy of 57.3% ± 0.02 across the test set, with a median recovery of 58.3%. These results confirm that the multi-state framework maintains the high predictive quality of the original context-aware architecture while extending its applicability to conformationally flexible RNA molecules. The consistent recovery values demonstrate that integrating ensemble information preserves model accuracy and generalization, providing a robust foundation for context-aware design across structural ensembles. A detailed breakdown of recovery rates across molecular categories is provided in Table 5-1.

Table 5-1. Mean native sequence recovery across RNA complex types for multi-state RISoTTo with Deep Set pooling.

RNA Type	Number of Complexes	Mean Recovery (%)	Std. Dev.
Standalone RNA	94	58.9	2.4
RNA-Protein Complexes	152	57.1	2.1
RNA-DNA Hybrids	48	56.7	1.9
RNA-Small Molecule Complexes	85	56.5	2.7
Overall Mean	**379**	**57.3**	**±0.02**

5.3.2 Riboswitch benchmark

Riboswitches represent the most stringent test for ensemble-based design due to their ligand-induced conformational switching. To specifically evaluate the model's ability to capture such flexibility, we extracted a subset of 162 riboswitch sequences from the Ribocentre database [269], containing paired apo and holo structures. For this benchmark, each riboswitch pair was processed as a two-conformer ensemble, and recovery was measured separately for both states as well as for the fused ensemble prediction. The ensemble-predicted sequences achieved an average recovery rate of 42.1% ± 0.03, compared to 40.0% and 38.7% for apo-only and holo-only predictions,

respectively. This result indicates that the ensemble model learns a shared sequence signature consistent with both conformational states.

5.3.3 Biological Interpretation

The performance of the Deep-Set-based ensemble fusion highlights a fundamental principle of RNA design: sequence constraints are ensemble-defined rather than conformation-specific. In biological systems, RNAs fluctuate between multiple substates, but their functional elements, base-pair stems, conserved tertiary contacts, and active-site motifs, remain stable across conformations. By averaging embeddings across conformers, the model effectively identifies these invariant features and encodes them into the predicted sequence distribution. The multi-state RISoTTo can thus be viewed as approximating a Boltzmann-weighted sequence optimization, where conformers contribute to the prediction proportionally to their frequency or stability. This concept is consistent with experimental observations indicating that functional RNAs exist as dynamic ensembles governed by thermodynamic equilibria [270]. In this context, Deep- Set pooling functions act as a statistical integrator merging state-specific constraints into a thermodynamically averaged design.

5.3.4 Error Analysis

Error inspection revealed that the most frequent sequence mismatches occurred in loop regions with large positional variability (>4 Å RMSD across conformers). In such cases, the mean aggregation may suppress state-specific interactions, leading to less accurate predictions. Additionally, missing residues in some PDB entries required masking during training, which occasionally disrupted message-passing continuity. To mitigate these issues, future implementations could incorporate mask-aware attention mechanisms or quality-weighted pooling, where conformers with better completeness and compactness contribute more strongly to the ensemble representation. These mechanisms could build upon the consensus-topology approach introduced in preliminary experiments.

5.4. Conclusion

This work demonstrates that independent conformer processing, combined with Deep Set aggregation, provides an effective framework for ensemble-aware RNA design. By preserving the local geometric integrity of each conformer and fusing learned representations in a permutation-invariant manner, the model captures the essential statistical features of RNA conformational ensembles. The resulting multi-state RISoTTo architecture enables context-aware RNA sequence

design that accounts for flexibility, offering a foundation for applications in riboswitch engineering, ribozyme design, and protein-RNA interface optimization.

The Multi-State RISoTTo framework is particularly suited for practical RNA design scenarios where function depends on structural adaptability, such as aptamers and riboswitches, which often operate through conformational switching upon ligand or protein binding. By explicitly accounting for multiple conformational states during sequence optimization, the method enables the design of RNA sequences that preserve functionality across structurally distinct ensembles, rather than stabilizing a single minimum-energy conformation. This is especially relevant for systems where binding-competent and inactive states coexist in equilibrium. However, multi-state design introduces important considerations for downstream experimental validation: sequences optimized to balance multiple states may exhibit reduced stability of individual conformations, altered folding kinetics, or population shifts that depend sensitively on environmental conditions (e.g., ion concentration, temperature, or binding partner availability). As a result, experimental validation may require ensemble-sensitive techniques (such as SHAPE probing or single-molecule assays) rather than single-structure characterization. Despite these challenges, multi-state design provides a more realistic representation of RNA behavior in biological contexts and can improve the success rate of functional RNA engineering when conformational plasticity is essential.

Chapter 6

Conclusion

In this thesis, we have explored the development and validation of the PANTHER scoring function, a novel computational framework designed to predict protein-RNA binding affinities. This research began by the identification of the limitations in existing experimental and computational methods for determining binding affinities, thereby highlighting the need for more accurate, accessible, and biologically relevant prediction tools.

Chapter 3 detailed the development of the PANTHER methodology, which integrates machine learning (ML) models with molecular dynamics (MD) simulations to predict local interaction energies between amino acids and nucleotide bases. By leveraging a local-to-global approach, the PANTHER score aggregates these local energies to estimate the overall binding affinity of protein-RNA complexes. The validation of this approach demonstrated a strong correlation between the MD-derived local-to-global scores and experimental ΔG values, confirming the reliability and the predictive accuracy of the fragmental methodology.

In Chapter 4, we extended the PANTHER framework by developing a web-based service, thereby making the predictive capability freely accessible to the broader scientific community. The service architecture implements a streamlined computational pipeline that accepts PDB identifiers or custom structural files, automatically extracts interaction features, predicts local amino acid-nucleotide energies using pre-trained ML models, and integrates these predictions to generate global PANTHER scores. This democratization of access enables researchers without computational expertise to perform sophisticated analyses, thereby accelerating experimental discovery and translational applications. After developing a robust scoring function to understand which RNA strand can be energetically favorable for binding to a protein structure of interest, the next step was to create dedicated software to design stable RNA structure.

The applicability domain of PANTHER is focused on structured protein–RNA complexes for which reliable atomic models are available. The method is particularly suitable for systems where binding is governed by stable, persistent intermolecular interactions, as these are effectively captured through the MD-derived energy features used for training. A key strength of the approach is its integration of physics-based interaction energies with machine-learning models, enabling generalization across diverse protein–RNA interfaces. Its robustness is further supported by

validation on both an independent test set and a larger external stress set comprising 110 complexes not used during training.

However, the method also has limitations. It relies on MD sampling from a single trajectory per system, which may not fully capture rare or transient conformational states. The approach assumes that persistent interactions dominate binding, making it less accurate for highly dynamic or transient interfaces. Performance may also decline for systems outside the training distribution, such as complexes with substantial disorder, non-canonical nucleic acids or modified residues absent from the training data, or cases involving large conformational rearrangements upon binding.

Chapter 5 introduced the Multi-State RISoTTo extension, addressing the challenge of RNA sequence design under conformational flexibility. By incorporating deep learning using geometry of the system and multi-conformer fusion strategies, this extension enhances the context-aware design of RNA molecules, ensuring that they maintain functionality across multiple conformational states. The integration of binding affinity prediction with sequence design applications creates a comprehensive toolkit for protein-RNA interaction engineering, enabling systematic validation and iterative optimization of both binding affinity and conformational compatibility.

The contributions of this thesis advance the field of computational structural biology by providing accurate, accessible, and biologically relevant prediction tools. The PANTHER scoring function, with its superior performance over existing methods, demonstrates that machine learning models trained on high-quality MD-derived data can capture the complex, non-linear relationships underlying protein-RNA recognition. The development of the PANTHER web service ensures broad community access, while the Multi-State RISoTTo extension establishes a foundation for biologically relevant RNA design scenarios.

In conclusion, this integrated approach not only supports the development of RNA therapeutics but also provides mechanistic insights into the structural and energetic determinants of binding specificity. Future directions include expanding the PANTHER training dataset, exploring deep learning architectures for local energy prediction, and integrating explicit solvent effects and conformational dynamics. The demonstrated success of combining physics-based molecular simulations with machine learning prediction establishes a powerful paradigm for addressing complex problems in structural biology, paving the way for future advances in understanding and engineering biomolecular interactions.

List of Publications

- Aletayeb, P., Biswas, A. D., Rocca, S., Talarico, C., Vistoli, G., & Pedretti, A. (2026). PANTHER Score: Protein-Affinity for Nucleic Target-binding, Hybridization, and Energy Regression. *RNA*, *32*(2), 131-149.

- PANTHER score Web Server: An Accessible Platform for Predicting Protein-RNA Binding Affinities Using Machine Learning (Under-Review)

Data Availability and Reproducibility

All the data produced are either shared in the main manuscript or in the Supplemental Material of the published paper (doi/10.1261/rna.080646.125), which makes it easier for reproducibility and applicability with other methods.

References

[1] T. Glisovic, J. L. Bachorik, J. Yong, and G. Dreyfuss, “RNA-binding proteins and post-transcriptional gene regulation,” *FEBS Lett.*, vol. 582, no. 14, pp. 1977–1986, Jun. 2008, doi: 10.1016/J.FEBSLET.2008.03.004.

[2] M. Khoroshkin *et al.*, “Systematic identification of post-transcriptional regulatory modules,” *Nature Communications* , vol. 15, no. 1, pp. 1–21, Dec. 2024, doi: 10.1038/S41467-024-52215-7;TECHMETA.

[3] S. Rehman, S. Bahadur, W. Xia, C. Runan, M. Ali, and Z. Maqbool, “From genes to traits: Trends in RNA-binding proteins and their role in plant trait development: A review,” *Int. J. Biol. Macromol.*, vol. 282, no. Pt 4, Dec. 2024, doi: 10.1016/J.IJBIOMAC.2024.136753.

[4] K. Kappel, I. Jarmoskaite, P. P. Vaidyanathan, W. J. Greenleaf, D. Herschlag, and R. Das, “Blind tests of RNA–protein binding affinity prediction,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 17, pp. 8336–8341, Apr. 2019, doi: 10.1073/PNAS.1819047116.

[5] B. Chaves-Arquero *et al.*, “Affinity-enhanced RNA-binding domains as tools to understand RNA recognition,” *Cell Reports Methods*, vol. 3, no. 6, Jun. 2023, doi: 10.1016/j.crmeth.2023.100508.

[6] A. Armaos, E. Zacco, N. Sanchez de Groot, and G. G. Tartaglia, “RNA-protein interactions: Central players in coordination of regulatory networks,” *Bioessays*, vol. 43, no. 2, Feb. 2021, doi: 10.1002/BIES.202000118.

[7] D. Kang, Y. Lee, and J. S. Lee, “RNA-Binding Proteins in Cancer: Functional and Therapeutic Perspectives,” *Cancers (Basel).*, vol. 12, no. 9, pp. 1–33, Sep. 2020, doi: 10.3390/CANCERS12092699.

[8] M. D. Disney, “Targeting RNA with Small Molecules To Capture Opportunities at the Intersection of Chemistry, Biology, and Medicine,” *J. Am. Chem. Soc.*, vol. 141, no. 17, pp. 6776–6790, May 2019, doi: 10.1021/JACS.8B13419.

[9] I. Dunham *et al.*, “An integrated encyclopedia of DNA elements in the human genome,” *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012, doi: 10.1038/NATURE11247;SUBJMETA.

[10] J. A. Kulkarni *et al.*, “The current landscape of nucleic acid therapeutics,” *Nat. Nanotechnol.*, vol. 16, no. 6, pp. 630–643, Jun. 2021, doi: 10.1038/S41565-021-00898-0;SUBJMETA.

[11] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello, “Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans,” *Nature*, vol. 391, no. 6669, pp. 806–811, Feb. 1998, doi: 10.1038/35888;KWRD.

[12] D. Adams *et al.*, “Patisiran, an RNAi Therapeutic, for Hereditary Transthyretin Amyloidosis,” *N. Engl. J. Med.*, vol. 379, no. 1, pp. 11–21, Jul. 2018, doi: 10.1056/NEJMOA1716153.

[13] M. Balwani *et al.*, “Phase 3 Trial of RNAi Therapeutic Givosiran for Acute Intermittent Porphyria,” *N. Engl. J. Med.*, vol. 382, no. 24, pp. 2289–2301, Jun. 2020, doi: 10.1056/NEJMOA1913147.

[14] S. F. Garrelfs *et al.*, “Lumasiran, an RNAi Therapeutic for Primary Hyperoxaluria Type 1,” *N. Engl. J. Med.*, vol. 384, no. 13, pp. 1216–1226, Apr. 2021, doi: 10.1056/NEJMOA2021712.

[15] K. K. Ray *et al.*, “Inclisiran in Patients at High Cardiovascular Risk with Elevated LDL Cholesterol,” *N. Engl. J. Med.*, vol. 376, no. 15, pp. 1430–1440, Apr. 2017, doi: 10.1056/NEJMOA1615758.

[16] “FDA approves add-on therapy to lower cholesterol among certain high-risk adults | FDA.” Accessed: Sep. 26, 2025. [Online]. Available: https://www.fda.gov/drugs/news-events-human-drugs/fda-approves-add-therapy-lower-cholesterol-among-certain-high-risk-adults

[17] D. Adams *et al.*, “Efficacy and safety of vutrisiran for patients with hereditary transthyretin-mediated amyloidosis with polyneuropathy: a randomized clinical trial,” *Amyloid*, vol. 30, no. 1, pp. 18–26, 2023, doi: 10.1080/13506129.2022.2091985.

[18] D. S. Goldfarb *et al.*, “Nedosiran in primary hyperoxaluria subtype 3: results from a phase I, single-dose study (PHYOX4),” *Urolithiasis*, vol. 51, no. 1, Dec. 2023, doi: 10.1007/S00240-023-01453-3.

[19] G. M. Traber and A. M. Yu, “The Growing Class of Novel RNAi Therapeutics,” *Mol. Pharmacol.*, vol. 106, no. 1, pp. 13–20, Jul. 2024, doi: 10.1124/MOLPHARM.124.000895.

[20] C. F. Bennett, “Therapeutic Antisense Oligonucleotides Are Coming of Age,” *Annu. Rev. Med.*, vol. 70, pp. 307–321, Jan. 2019, doi: 10.1146/ANNUREV-MED-041217-010829.

[21] R. S. Finkel *et al.*, “Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy,” *N. Engl. J. Med.*, vol. 377, no. 18, pp. 1723–1732, Nov. 2017, doi: 10.1056/NEJMOA1702752.

[22] J. R. Mendell *et al.*, “Eteplirsen for the treatment of Duchenne muscular dystrophy,” *Ann. Neurol.*, vol. 74, no. 5, pp. 637–647, Nov. 2013, doi: 10.1002/ANA.23982.

[23] M. D. Benson *et al.*, “Inotersen Treatment for Patients with Hereditary Transthyretin Amyloidosis,” *N. Engl. J. Med.*, vol. 379, no. 1, pp. 22–31, Jul. 2018, doi: 10.1056/NEJMOA1716793.

[24] D. P. Bartel, “MicroRNAs: genomics, biogenesis, mechanism, and function,” *Cell*, vol. 116, no. 2, pp. 281–297, Jan. 2004, doi: 10.1016/S0092-8674(04)00045-5.

[25] A. D. Ellington and J. W. Szostak, “In vitro selection of RNA molecules that bind specific ligands,” *Nature*, vol. 346, no. 6287, pp. 818–822, 1990, doi: 10.1038/346818A0;KWRD.

[26] E. S. Gragoudas, A. P. Adamis, E. T. Cunningham, M. Feinsod, and D. R. Guyer, “Pegaptanib for neovascular age-related macular degeneration,” *N. Engl. J. Med.*, vol. 351, no. 27, pp. 2805–2816, Dec. 2004, doi: 10.1056/NEJMOA042760.

[27] C. Kang, “Avacincaptad Pegol: First Approval,” *Drugs*, vol. 83, no. 15, pp. 1447–1453, Oct. 2023, doi: 10.1007/S40265-023-01948-8.

[28] A. D. Keefe, S. Pai, and A. Ellington, “Aptamers as therapeutics,” *Nat. Rev. Drug Discov.*, vol. 9, no. 7, pp. 537–550, Jul. 2010, doi: 10.1038/NRD3141.

[29] M. Famulok and G. Mayer, “Aptamers and SELEX in Chemistry & Biology,” *Chem. Biol.*, vol. 21, no. 9, pp. 1055–1058, Sep. 2014, doi: 10.1016/J.CHEMBIOL.2014.08.003.

[30] J. A. Doudna and T. R. Cech, “The chemical repertoire of natural ribozymes,” *Nature*, vol. 418, no. 6894, pp. 222–228, Jul. 2002, doi: 10.1038/418222A.

[31] J. L. Rinn and H. Y. Chang, “Genome regulation by long noncoding RNAs,” *Annu. Rev. Biochem.*, vol. 81, pp. 145–166, Jul. 2012, doi: 10.1146/ANNUREV-BIOCHEM-051410-092902.

[32] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, “A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity,” *Science*, vol. 337, no. 6096, pp. 816–821, Aug. 2012, doi: 10.1126/SCIENCE.1225829.

[33] T. B. Hansen *et al.*, “Natural RNA circles function as efficient microRNA sponges,” *Nature*, vol. 495, no. 7441, pp. 384–388, Mar. 2013, doi: 10.1038/NATURE11993;SUBJMETA.

[34] E. Karzbrun, A. M. Tayar, V. Noireaux, and R. H. Bar-Ziv, “Synthetic biology. Programmable on-chip DNA compartments as artificial cells,” *Science*, vol. 345, no. 6198, pp. 829–832, Aug. 2014, doi: 10.1126/SCIENCE.1255550.

[35] K. V. Morris and J. S. Mattick, “The rise of regulatory RNA,” *Nat. Rev. Genet.*, vol. 15, no. 6, pp. 423–437, 2014, doi: 10.1038/NRG3722.

[36] A. Serganov and E. Nudler, “A decade of riboswitches,” *Cell*, vol. 152, no. 1–2, pp. 17–24, Jan. 2013, doi: 10.1016/J.CELL.2012.12.024.

[37] I. Jarmoskaite, I. Alsadhan, P. P. Vaidyanathan, and D. Herschlag, “How to measure and evaluate binding affinities,” *Elife*, vol. 9, pp. 1–34, Aug. 2020, doi: 10.7554/ELIFE.57264.

[38] J. T. Vivian and P. R. Callis, “Mechanisms of tryptophan fluorescence shifts in proteins,” *Biophys. J.*, vol. 80, no. 5, pp. 2093–2109, May 2001, doi: 10.1016/S0006-3495(01)76183-8.

[39] P. S. Katsamba, S. Park, and I. A. Laird-Offringa, “Kinetic studies of RNA–protein interactions using surface plasmon resonance,” *Methods*, vol. 26, no. 2, pp. 95–104, Feb. 2002, doi: 10.1016/S1046-2023(02)00012-9.

[40] A. L. Feig, “Studying RNA–RNA and RNA–Protein Interactions by Isothermal Titration Calorimetry,” *Methods Enzymol.*, vol. 468, pp. 409–422, Jan. 2009, doi: 10.1016/S0076-6879(09)68019-8.

[41] D. C. Rio, “Filter-Binding Assay for Analysis of RNA–Protein Interactions,” *Cold Spring Harb. Protoc.*, vol. 2012, no. 10, p. pdb.prot071449, Oct. 2012, doi: 10.1101/PDB.PROT071449.

[42] C. Nithin, S. Mukherjee, and R. P. Bahadur, “A structure-based model for the prediction of protein–RNA binding affinity,” *RNA*, vol. 25, no. 12, pp. 1628–1645, Dec. 2019, doi: 10.1261/RNA.071779.119.

[43] M. L.- Molecules and undefined 2023, “Recent advances in deep learning for protein-protein interaction analysis: A comprehensive review,” *mdpi.comM LeeMolecules, 2023•mdpi.com*, 2023, doi: 10.3390/molecules28135169.

[44] D. Li *et al.*, “RNA-Protein Interaction Prediction Based on Deep Learning: A Comprehensive Survey,” *Journal of the ACM*, vol. 37, no. 4, p. 34, Sep. 2024, doi: XXXXXXX.XXXXXXX.

[45] L. Rajkowitsch *et al.*, “RNA chaperones, RNA annealers and RNA helicases,” *RNA Biol.*, vol. 4, no. 3, pp. 118–130, 2007, doi: 10.4161/RNA.4.3.5445.

[46] G. Singh, G. Pratt, G. W. Yeo, and M. J. Moore, “The Clothes Make the mRNA: Past and Present Trends in mRNP Fashion,” *Annu. Rev. Biochem.*, vol. 84, pp. 325–354, Jun. 2015, doi: 10.1146/ANNUREV-BIOCHEM-080111-092106.

[47] D. S. W. Protter and R. Parker, “Principles and Properties of Stress Granules,” *Trends Cell Biol.*, vol. 26, no. 9, pp. 668–679, Sep. 2016, doi: 10.1016/j.tcb.2016.05.004.

[48] S. Jonas and E. Izaurralde, “Towards a molecular understanding of microRNA-mediated gene silencing,” *Nat. Rev. Genet.*, vol. 16, no. 7, pp. 421–433, Jul. 2015, doi: 10.1038/NRG3965;SUBJMETA.

[49] D. R. Bell, J. K. Weber, W. Yin, T. Huynh, W. Duan, and R. Zhou, “In silico design and validation of high-affinity RNA aptamers targeting epithelial cellular adhesion molecule dimers,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 15, pp. 8486–8493, Apr. 2020, doi: 10.1073/PNAS.1913242117/-/DCSUPPLEMENTAL.

[50] L. Wang and S. J. Brown, “BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences,” *Nucleic Acids Res.*, vol. 34, no. Web Server issue, Jul. 2006, doi: 10.1093/NAR/GKL298.

[51] M. Terribilini *et al.*, “RNABindR: a server for analyzing and predicting RNA-binding sites in proteins,” *Nucleic Acids Res.*, vol. 35, no. Web Server issue, Jul. 2007, doi: 10.1093/NAR/GKM294.

[52] M. Kumar, M. M. Gromiha, and G. P. S. Raghava, “Prediction of RNA binding sites in a protein using SVM and PSSM profile,” *Proteins*, vol. 71, no. 1, pp. 189–194, Apr. 2008, doi: 10.1002/PROT.21677.

[53] D. Marchese, N. S. de Groot, N. Lorenzo Gotor, C. M. Livi, and G. G. Tartaglia, “Advances in the characterization of RNA-binding proteins,” *Wiley Interdiscip. Rev. RNA*, vol. 7, no. 6, pp. 793–810, Nov. 2016, doi: 10.1002/WRNA.1378.

[54] C. Dominguez, R. Boelens, and A. M. J. J. Bonvin, “HADDOCK: a protein-protein docking approach based on biochemical or biophysical information,” *J. Am. Chem. Soc.*, vol. 125, no. 7, pp. 1731–1737, Feb. 2003, doi: 10.1021/JA026939X.

[55] L. Pérez-Cano, A. Solernou, C. Pons, and J. Fernández-Recio, “Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials,” *Pac. Symp. Biocomput.*, pp. 293–301, 2010, doi: 10.1142/9789814295291_0031.

[56] C. A. Sotriffer, P. Sanschagrin, H. Matter, and G. Klebe, “SFCscore: scoring functions for affinity prediction of protein-ligand complexes,” *Proteins*, vol. 73, no. 2, pp. 395–419, Nov. 2008, doi: 10.1002/PROT.22058.

[57] M. Kumar, M. M. Gromiha, and G. P. S. Raghava, “SVM based prediction of RNA-binding proteins using binding residues and evolutionary information,” *J. Mol. Recognit.*, vol. 24, no. 2, pp. 303–313, Mar. 2011, doi: 10.1002/JMR.1061.

[58] D. S. Jain, S. R. Gupte, and R. Aduri, “A Data Driven Model for Predicting RNA-Protein Interactions based on Gradient Boosting Machine,” *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, Dec. 2018, doi: 10.1038/S41598-018-27814-2;SUBJMETA.

[59] L. Deng, Y. Sui, and J. Zhang, “XGBPRH: Prediction of Binding Hot Spots at Protein–RNA Interfaces Utilizing Extreme Gradient Boosting,” *Genes (Basel).*, vol. 10, no. 3, p. 242, Mar. 2019, doi: 10.3390/GENES10030242.

[60] A. Agarwal, K. Singh, S. Kant, and R. P. Bahadur, “A comparative analysis of machine learning classifiers for predicting protein-binding nucleotides in RNA sequences,” *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 3195–3207, Jan. 2022, doi: 10.1016/J.CSBJ.2022.06.036.

[61] X. Pan, P. Rijnbeek, J. Yan, and H. Bin Shen, “Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks,” *BMC Genomics*, vol. 19, no. 1, Jul. 2018, doi: 10.1186/S12864-018-4889-1.

[62] S. Zhang *et al.*, “A deep learning framework for modeling structural features of RNA-binding protein targets,” *Nucleic Acids Res.*, vol. 44, no. 4, Sep. 2016, doi: 10.1093/NAR/GKV1025.

[63] C. Nithin, S. Mukherjee, and R. P. Bahadur, “A structure-based model for the prediction of protein-RNA binding affinity,” *RNA*, vol. 25, no. 12, pp. 1628–1645, 2019, doi: 10.1261/rna.071779.119.

[64] K. Harini, M. Sekijima, and M. M. Gromiha, “PRA-Pred: Structure-based prediction of protein-RNA binding affinity,” *Int. J. Biol. Macromol.*, vol. 259, Feb. 2024, doi: 10.1016/j.ijbiomac.2024.129490.

[65] R. Han *et al.*, “CoPRA: Bridging Cross-domain Pretrained Sequence Models with Complex Structures for Protein-RNA Binding Affinity Prediction,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 1, pp. 246–254, Apr. 2025, doi: 10.1609/AAAI.V39I1.32001.

[66] C. Zeng, C. Zhuo, J. Gao, H. Liu, and Y. Zhao, “Advances and Challenges in Scoring Functions for RNA–Protein Complex Structure Prediction,” *Biomolecules*, vol. 14, no. 10, p. 1245, Oct. 2024, doi: 10.3390/BIOM14101245/S1.

[67] R. Wang, X. Fang, Y. Lu, C. Y. Yang, and S. Wang, “The PDBbind database: methodologies and updates,” *J. Med. Chem.*, vol. 48, no. 12, pp. 4111–4119, Jun. 2005, doi: 10.1021/JM048957Q.

[68] X. Hong *et al.*, “An updated dataset and a structure-based prediction model for protein–RNA binding affinity,” *Proteins: Structure, Function, and Bioinformatics*, vol. 91, no. 9, pp. 1245–1253, Sep. 2023, doi: 10.1002/PROT.26503.

[69] K. Harini, A. Srivastava, A. Kulandaisamy, and M. M. Gromiha, “ProNAB: database for binding affinities of protein-nucleic acid complexes and their mutants,” *Nucleic Acids Res.*, vol. 50, no. D1, pp. D1528–D1534, Jan. 2022, doi: 10.1093/NAR/GKAB848.

[70] L. R. Ganser, M. L. Kelly, D. Herschlag, and H. M. Al-Hashimi, “The roles of structural dynamics in the cellular functions of RNAs,” *Nat. Rev. Mol. Cell Biol.*, vol. 20, no. 8, pp. 474–489, Aug. 2019, doi: 10.1038/S41580-019-0136-0.

[71] A. Serganov and E. Nudler, “A decade of riboswitches,” *Cell*, vol. 152, no. 1–2, pp. 17–24, Jan. 2013, doi: 10.1016/J.CELL.2012.12.024.

[72] R. J. Loureiro, S. Maiti, K. Mondal, S. Mukherjee, and J. M. Bujnicki, “Modeling flexible RNA 3D structures and RNA-protein complexes,” *Curr. Opin. Struct. Biol.*, vol. 94, Oct. 2025, doi: 10.1016/J.SBI.2025.103137.

[73] A. Sabei, C. Hognon, J. Martin, and E. Frezza, “Dynamics of Protein–RNA Interfaces Using All-Atom Molecular Dynamics Simulations,” *J. Phys. Chem. B*, vol. 128, no. 20, pp. 4865–4886, May 2024, doi: 10.1021/ACS.JPCB.3C07698.

[74] J. Wei, S. Chen, L. Zong, X. Gao, and Y. Li, “Protein–RNA interaction prediction with deep learning: structure matters,” *Brief. Bioinform.*, vol. 23, no. 1, Jan. 2022, doi: 10.1093/BIB/BBAB540.

[75] C. Zeng, C. Zhuo, J. Gao, H. Liu, and Y. Zhao, “Advances and Challenges in Scoring Functions for RNA–Protein Complex Structure Prediction,” *Biomolecules*, vol. 14, no. 10, p. 1245, Oct. 2024, doi: 10.3390/BIOM14101245/S1.

[76] T. Harren, T. Gutermuth, C. Grebner, G. Hessler, and M. Rarey, “Modern machine-learning for binding affinity estimation of protein–ligand complexes: Progress, opportunities, and challenges,” *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 14, no. 3, p. e1716, May 2024, doi: 10.1002/WCMS.1716.

[77] H. Gohlke and G. Klebe, “Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors,” Aug. 02, 2002, *Wiley-VCH Verlag*. doi: 10.1002/1521-3773(20020802)41:15<2644::AID-ANIE2644>3.0.CO;2-O.

[78] J. D. Dunitz, “Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions,” *Chem. Biol.*, vol. 2, no. 11, pp. 709–712, 1995, doi: 10.1016/1074-5521(95)90097-7.

[79] T. Harren, T. Gutermuth, C. Grebner, G. Hessler, and M. Rarey, “Modern machine-learning for binding affinity estimation of protein–ligand complexes: Progress, opportunities, and challenges,” *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 14, no. 3, p. e1716, May 2024, doi: 10.1002/WCMS.1716.

[80] D. Dominguez *et al.*, “Sequence, Structure, and Context Preferences of Human RNA Binding Proteins,” *Mol. Cell*, vol. 70, no. 5, pp. 854-867.e9, Jun. 2018, doi: 10.1016/J.MOLCEL.2018.05.001.

[81] Y. Liu, Y. Ou, and L. Hou, “Advances in RNA-Based Therapeutics: Challenges and Innovations in RNA Delivery Systems,” *Current Issues in Molecular Biology 2025, Vol. 47, Page 22*, vol. 47, no. 1, p. 22, Dec. 2024, doi: 10.3390/CIMB47010022.

[82] C. M. Connelly, M. H. Moon, and J. S. Schneekloth, “The Emerging Role of RNA as a Therapeutic Target for Small Molecules,” *Cell Chem. Biol.*, vol. 23, no. 9, p. 1077, Sep. 2016, doi: 10.1016/J.CHEMBIOL.2016.05.021.

[83] J. A. Doudna and E. Charpentier, “The new frontier of genome engineering with CRISPR-Cas9,” *Science (1979).*, vol. 346, no. 6213, Nov. 2014, doi: 10.1126/SCIENCE.1258096.

[84] K. Leppek, R. Das, and M. Barna, “Functional 5’ UTR mRNA structures in eukaryotic translation regulation and how to find them,” *Nat. Rev. Mol. Cell Biol.*, vol. 19, no. 3, pp. 158–174, Mar. 2018, doi: 10.1038/NRM.2017.103.

[85] M. Mandal and R. R. Breaker, “Gene regulation by riboswitches,” *Nat. Rev. Mol. Cell Biol.*, vol. 5, no. 6, pp. 451–463, Jun. 2004, doi: 10.1038/NRM1403;KWRD.

[86] N. Pardi, M. J. Hogan, F. W. Porter, and D. Weissman, “mRNA vaccines - a new era in vaccinology,” *Nat. Rev. Drug Discov.*, vol. 17, no. 4, pp. 261–279, Mar. 2018, doi: 10.1038/NRD.2017.243.

[87] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, “A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity,” *Science*, vol. 337, no. 6096, pp. 816–821, Aug. 2012, doi: 10.1126/SCIENCE.1225829.

[88] R. Barrangou and J. A. Doudna, “Applications of CRISPR technologies in research and beyond,” *Nat. Biotechnol.*, vol. 34, no. 9, pp. 933–941, Sep. 2016, doi: 10.1038/NBT.3659.

[89] Y. Zhu, L. Zhu, X. Wang, and H. Jin, “RNA-based therapeutics: an overview and prospectus,” *Cell Death Dis.*, vol. 13, no. 7, pp. 1–15, Jul. 2022, doi: 10.1038/S41419-022-05075-2;TECHMETA.

[90] T. R. Damase, R. Sukhovershin, C. Boada, F. Taraballi, R. I. Pettigrew, and J. P. Cooke, “The Limitless Future of RNA Therapeutics,” *Front. Bioeng. Biotechnol.*, vol. 9, Mar. 2021, doi: 10.3389/FBIOE.2021.628137.

[91] É. Bonnet, P. Rzazewski, and F. Sikora, “Designing RNA Secondary Structures Is Hard,” *J. Comput. Biol.*, vol. 27, no. 3, pp. 302–316, Mar. 2020, doi: 10.1089/CMB.2019.0420.

[92] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, “Fast folding and comparison of RNA secondary structures,” *Monatshefte für Chemie Chemical Monthly*, vol. 125, no. 2, pp. 167–188, 1994, doi: 10.1007/BF00818163/METRICS.

[93] D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner, “Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure,” *Proceedings of the National Academy of Sciences*, vol. 101, no. 19, pp. 7287–7292, May 2004, doi: 10.1073/PNAS.0401799101.

[94] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, “Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure,” *J. Mol. Biol.*, vol. 288, no. 5, pp. 911–940, May 1999, doi: 10.1006/JMBI.1999.2700.

[95] B. Knudsen and J. Hein, “RNA secondary structure prediction using stochastic context-free grammars and evolutionary history,” *Bioinformatics*, vol. 15, no. 6, pp. 446–454, 1999, doi: 10.1093/BIOINFORMATICS/15.6.446.

[96] K. Sato, M. Hamada, K. Asai, and T. Mituyama, “CENTROIDFOLD: a web server for RNA secondary structure prediction,” *Nucleic Acids Res.*, vol. 37, no. Web Server issue, 2009, doi: 10.1093/NAR/GKP367.

[97] T. Puton, L. P. Kozlowski, K. M. Rother, and J. M. Bujnicki, “CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction,” *Nucleic Acids Res.*, vol. 41, no. 7, pp. 4307–4323, Apr. 2013, doi: 10.1093/NAR/GKT101.

[98] N. Dromi, A. Avihoo, and D. Barash, “Reconstruction of natural RNA sequences from RNA shape, thermodynamic stability, mutational robustness, and linguistic complexity by evolutionary computation,” *J. Biomol. Struct. Dyn.*, vol. 26, no. 1, pp. 147–161, 2008, doi: 10.1080/07391102.2008.10507231.

[99] A. Esmaili-Taheri, M. Ganjtabesh, and M. Mohammad-Noori, “Evolutionary solution for the RNA design problem,” *Bioinformatics*, vol. 30, no. 9, pp. 1250–1258, May 2014, doi: 10.1093/BIOINFORMATICS/BTU001.

[100] R. Kleinkauf, M. Mann, and R. Backofen, “antaRNA: ant colony-based RNA sequence design,” *Bioinformatics*, vol. 31, no. 19, pp. 3114–3121, Oct. 2015, doi: 10.1093/BIOINFORMATICS/BTV319.

[101] F. Runge, D. Stoll, S. Falkner, and F. Hutter, “Learning to Design RNA”.

[102] R. Das and D. Baker, “Automated de novo prediction of native-like RNA tertiary structures,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 37, pp. 14664–14669, Sep. 2007, doi: 10.1073/PNAS.0703836104.

[103] R. Das, J. Karanicolas, and D. Baker, “Atomic accuracy in predicting and designing noncanonical RNA structure,” *Nat. Methods*, vol. 7, no. 4, pp. 291–294, 2010, doi: 10.1038/NMETH.1433.

[104] C. K. Joshi *et al.*, “gRNAde: Geometric Deep Learning for 3D RNA inverse design,” *13th International Conference on Learning Representations, ICLR 2025*, pp. 56390–56415, May 2023, Accessed: Sep. 22, 2025. [Online]. Available: https://arxiv.org/pdf/2305.14749

[105] H. Huang, Z. Lin, D. He, L. Hong, and Y. Li, “RiboDiffusion: tertiary structure-based RNA inverse folding with generative diffusion models,” *Bioinformatics*, vol. 40, no. Supplement_1, pp. i347–i356, Jun. 2024, doi: 10.1093/BIOINFORMATICS/BTAE259.

[106] M. Ramanathan, D. F. Porter, and P. A. Khavari, “Methods to study RNA-protein interactions,” *Nat. Methods*, vol. 16, no. 3, pp. 225–234, Mar. 2019, doi: 10.1038/S41592-019-0330-1.

[107] Y. Tong *et al.*, “Programming inactive RNA-binding small molecules into bioactive degraders,” *Nature*, vol. 618, no. 7963, pp. 169–179, Jun. 2023, doi: 10.1038/S41586-023-06091-8.

[108] I. V. Kornienko, O. Y. Aramova, A. A. Tishchenko, D. V. Rudoy, and M. L. Chikindas, “RNA Stability: A Review of the Role of Structural Features and Environmental Conditions,” *Molecules*, vol. 29, no. 24, Dec. 2024, doi: 10.3390/MOLECULES29245978.

[109] F. Cozzolino, I. Iacobucci, V. Monaco, and M. Monti, “Protein-DNA/RNA Interactions: An Overview of Investigation Methods in the -Omics Era,” *J. Proteome Res.*, vol. 20, no. 6, pp. 3018–3030, Jun. 2021, doi: 10.1021/ACS.JPROTEOME.1C00074.

[110] L. F. Krapp *et al.*, “Context-aware geometric deep learning for protein sequence design,” *Nat. Commun.*, vol. 15, no. 1, pp. 1–10, Dec. 2024, doi: 10.1038/S41467-024-50571-Y;TECHMETA.

[111] L. Deng, W. Yang, and H. Liu, “PredPRBA: Prediction of Protein-RNA Binding Affinity Using Gradient Boosted Regression Trees,” *Front. Genet.*, vol. 10, Jan. 2019, doi: 10.3389/fgene.2019.00637.

[112] S. Genheden, A. Reymer, P. Saenz-Méndez, and L. A. Eriksson, “Computational Chemistry and Molecular Modelling Basics,” *Computational Tools for Chemical Biology*, pp. 1–38, Oct. 2017, doi: 10.1039/9781788010139-00001.

[113] G. M. Morris and M. Lim-Wilby, “Molecular docking,” *Methods in Molecular Biology*, vol. 443, pp. 365–382, 2008, doi: 10.1007/978-1-59745-177-2_19/TABLES/1.

[114] A. Hospital, J. R. Goñi, M. Orozco, and J. L. Gelpí, “Molecular dynamics simulations: advances and applications,” *Advances and Applications in Bioinformatics and Chemistry*, vol. 8, no. 1, pp. 37–47, Nov. 2015, doi: 10.2147/AABC.S70333.

[115] W. L. Delano, “PyMOL: An Open-Source Molecular Graphics Tool”.

[116] E. F. Pettersen *et al.*, “UCSF Chimera--a visualization system for exploratory research and analysis,” *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, Oct. 2004, doi: 10.1002/JCC.20084.

[117] Z. Yang *et al.*, “UCSF Chimera, MODELLER, and IMP: an Integrated Modeling System,” *J. Struct. Biol.*, vol. 179, no. 3, p. 269, Sep. 2011, doi: 10.1016/J.JSB.2011.09.006.

[118] A. Pedretti, L. Villa, and G. Vistoli, “VEGA--an open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming,” *J. Comput. Aided. Mol. Des.*, vol. 18, no. 3, pp. 167–173, 2004, doi: 10.1023/B:JCAM.0000035186.90683.F2.

[119] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson, “Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation,” *J. Mol. Biol.*, vol. 285, no. 4, pp. 1735–1747, Jan. 1999, doi: 10.1006/jmbi.1998.2401.

[120] A. Fiser, R. K. G. Do, and A. Šali, “Modeling of loops in protein structures,” *Protein Sci.*, vol. 9, no. 9, pp. 1753–1773, Jan. 2000, doi: 10.1110/PS.9.9.1753.

[121] C. I. Bayly *et al.*, “A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules,” *J. Am. Chem. Soc.*, vol. 117, no. 19, pp. 5179–5197, 2002, doi: 10.1021/JA00124A002.

[122] D. A. Case *et al.*, “The Amber biomolecular simulation programs,” *J. Comput. Chem.*, vol. 26, no. 16, pp. 1668–1688, Dec. 2005, doi: 10.1002/JCC.20290.

[123] C. L. Brooks, “Computer simulation of liquids,” 1989, *Springer*.

[124] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, “ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB,” *J. Chem. Theory Comput.*, vol. 11, no. 8, pp. 3696–3713, Jul. 2015, doi: 10.1021/ACS.JCTC.5B00255.

[125] M. Zgarbová, J. Šponer, M. Otyepka, T. E. Cheatham, R. Galindo-Murillo, and P. Jurečka, “Refinement of the Sugar–Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA,” *J. Chem. Theory Comput.*, vol. 11, no. 12, pp. 5723–5736, Nov. 2015, doi: 10.1021/ACS.JCTC.5B00716.

[126] J. Huang and A. D. Mackerell, “CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data,” *J. Comput. Chem.*, vol. 34, no. 25, pp. 2135–2145, Sep. 2013, doi: 10.1002/JCC.23354.

[127] N. Schmid *et al.*, “Definition and testing of the GROMOS force-field versions 54A7 and 54B7,” *Eur. Biophys. J.*, vol. 40, no. 7, pp. 843–856, Jul. 2011, doi: 10.1007/S00249-011-0700-9.

[128] L. Verlet, “Computer ‘Experiments’ on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules,” *Physical Review*, vol. 159, no. 1, p. 98, Jul. 1967, doi: 10.1103/PhysRev.159.98.

[129] D. Beeman, “Some multistep methods for use in molecular dynamics calculations,” *J. Comput. Phys.*, vol. 20, no. 2, pp. 130–139, 1976, doi: 10.1016/0021-9991(76)90059-0.

[130] C. W. Gear, “Numerical Initial Value Problems in Ordinary Differential Equations (Automatic Computation),” p. 253, 1971, Accessed: Sep. 25, 2025. [Online]. Available: http://www.amazon.com/Numerical-Differential-Equations-Automatic-Computation/dp/0136266061

[131] T. Schlick, E. Barth, and M. Mandziuk, “Biomolecular dynamics at long timesteps: bridging the timescale gap between simulation and experimentation,” *Annu. Rev. Biophys. Biomol. Struct.*, vol. 26, pp. 181–222, 1997, doi: 10.1146/ANNUREV.BIOPHYS.26.1.181.

[132] “COMPUTER SIMULATION OF LIQUIDS”.

[133] T. Darden, D. York, and L. Pedersen, “Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems,” *J. Chem. Phys.*, vol. 98, no. 12, pp. 10089–10092, Jun. 1993, doi: 10.1063/1.464397.

[134] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, “Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes,” *J. Comput. Phys.*, vol. 23, no. 3, pp. 327–341, Mar. 1977, doi: 10.1016/0021-9991(77)90098-5.

[135] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, “LINCS: A Linear Constraint Solver for Molecular Simulations,” *J Comput Chem*, vol. 18, p. 14631472, 1997, doi: 10.1002/(SICI)1096-987X(199709)18:12.

[136] “Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems - PubMed.” Accessed: Sep. 25, 2025. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35619462/

[137] H. J. C. Berendsen *et al.*, “Molecular dynamics with coupling to an external bath,” *JChPh*, vol. 81, no. 8, pp. 3684–3690, 1984, doi: 10.1063/1.448118.

[138] D. M. Heyes, M. Barber, and J. H. R. Clarke, “Molecular dynamics computer simulation of surface properties of crystalline potassium chloride,” *Journal of the Chemical Society, Faraday Transactions 2: Molecular and Chemical Physics*, vol. 73, no. 7, pp. 1485–1496, Jan. 1977, doi: 10.1039/F29777301485.

[139] A. Amadei, G. Chillemi, M. A. Ceruso, A. Grottesi, and A. Di Nola, “Molecular dynamics simulations with constrained roto-translational motions: Theoretical basis and statistical mechanical consistency,” *J. Chem. Phys.*, vol. 112, no. 1, pp. 9–23, Jan. 2000, doi: 10.1063/1.480557.

[140] S. Nosé, “A unified formulation of the constant temperature molecular dynamics methods,” *J. Chem. Phys.*, vol. 81, no. 1, pp. 511–519, Jul. 1984, doi: 10.1063/1.447334.

[141] S. Nosé and M. L. Klein, “Constant pressure molecular dynamics for molecular systems,” *Mol. Phys.*, vol. 50, no. 5, pp. 1055–1076, Dec. 1983, doi: 10.1080/00268978300102851.

[142] W. G. Hoover, “Canonical dynamics: Equilibrium phase-space distributions,” *Phys. Rev. A (Coll. Park).*, vol. 31, no. 3, pp. 1695–1697, 1985, doi: 10.1103/PHYSREVA.31.1695.

[143] H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gunsteren, A. Dinola, and J. R. Haak, “Molecular dynamics with coupling to an external bath,” *J. Chem. Phys.*, vol. 81, no. 8, pp. 3684–3690, Oct. 1984, doi: 10.1063/1.448118.

[144] M. Parrinello and A. Rahman, “Polymorphic transitions in single crystals: A new molecular dynamics method,” *J. Appl. Phys.*, vol. 52, no. 12, pp. 7182–7190, Dec. 1981, doi: 10.1063/1.328693.

[145] S. Lifson, ; A Warshel, S. Urson And, and A. W. Arshel, “Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and n-Alkane Molecules,” *J. Chem. Phys.*, vol. 49, no. 11, pp. 5116–5129, Dec. 1968, doi: 10.1063/1.1670007.

[146] P. Kollman, “Free Energy Calculations: Applications to Chemical and Biochemical Phenomena,” *Chem. Rev*, vol. 93, pp. 2395–2417, 1993.

[147] U. C. Singh and P. A. Kollman, “An approach to computing electrostatic charges for molecules,” *J. Comput. Chem.*, vol. 5, no. 2, pp. 129–145, Apr. 1984, doi: 10.1002/JCC.540050204.

[148] I. Massova and P. A. Kollman, “Combined molecular mechanical and continuum solvent approach (MM- PBSA/GBSA) to predict ligand binding,” *Perspectives in Drug Discovery and Design*, vol. 18, no. 1, pp. 113–135, 2000, doi: 10.1023/A:1008763014207/METRICS.

[149] K. Raha *et al.*, “Pairwise decomposition of residue interaction energies using semiempirical quantum mechanical methods in studies of protein-ligand interaction,” *J. Am. Chem. Soc.*, vol. 127, no. 18, pp. 6583–6594, May 2005, doi: 10.1021/JA042666P.

[150] O. Serçinoğlu and P. Ozbek, “gRINN: a tool for calculation of residue interaction energies and protein energy network analysis of molecular dynamics simulations,” *Nucleic Acids Res.*, vol. 46, no. W1, pp. W554–W562, Jul. 2018, doi: 10.1093/NAR/GKY381.

[151] “Molecular Modelling Principles and Applications | PDF.” Accessed: Sep. 24, 2025. [Online]. Available: https://www.scribd.com/document/565769262/Molecular-Modelling-Principles-and-Applications

[152] T. Darden, D. York, and L. Pedersen, “Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems,” *J. Chem. Phys.*, vol. 98, no. 12, pp. 10089–10092, Jun. 1993, doi: 10.1063/1.464397.

[153] “Pattern Recognition and Machine Learning | SpringerLink.” Accessed: Sep. 25, 2025. [Online]. Available: https://link.springer.com/book/9780387310732

[154] T. Hastie, R. Tibshirani, and J. Friedman, “Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction”.

[155] V. N. Vapnik, “The Nature of Statistical Learning Theory,” *The Nature of Statistical Learning Theory*, 2000, doi: 10.1007/978-1-4757-3264-1.

[156] G. A. F. Seber and A. J. Lee, “Linear Regression Analysis,” *Linear Regression Analysis*, pp. 1–572, Jan. 2012, doi: 10.1002/9780471722199.

[157] G. James, D. Witten, T. Hastie, and R. Tibshirani, “An Introduction to Statistical Learning,” 2021, doi: 10.1007/978-1-0716-1418-1.

[158] R. Tibshirani, “Regression Shrinkage and Selection Via the Lasso,” *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996, doi: 10.1111/J.2517-6161.1996.TB02080.X.

[159] H. Zou and T. Hastie, “Regularization and Variable Selection Via the Elastic Net,” *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005, doi: 10.1111/J.1467-9868.2005.00503.X.

[160] J. H. Friedman, “Greedy function approximation: A gradient boosting machine.,” *https://doi.org/10.1214/aos/1013203451*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/AOS/1013203451.

[161] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785/SUPPL_FILE/KDD2016_CHEN_BOOSTING_SYSTEM_01-ACM.MP4.

[162] D. Nielsen, “Tree Boosting With XGBoost - Why Does XGBoost Win ‘Every’ Machine Learning Competition?,” 2016, Accessed: Sep. 25, 2025. [Online]. Available: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2433761

[163] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.

[164] G. Biau and E. Scornet, “A random forest guided tour,” *Test*, vol. 25, no. 2, pp. 197–227, Jun. 2016, doi: 10.1007/S11749-016-0481-7/FIGURES/4.

[165] D. R. Cutler *et al.*, “Random forests for classification in ecology,” *Ecology*, vol. 88, no. 11, pp. 2783–2792, Nov. 2007, doi: 10.1890/07-0539.1.

[166] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/NATURE14539.

[167] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural Networks*, vol. 2, no. 5, pp. 359–366, Jan. 1989, doi: 10.1016/0893-6080(89)90020-8.

[168] V. Nair and G. E. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines”.

[169] “Deep Learning.” Accessed: Sep. 25, 2025. [Online]. Available: https://www.deeplearningbook.org/

[170] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[171] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” *32nd International Conference on Machine Learning, ICML 2015*, vol. 1, pp. 448–456, Feb. 2015, Accessed: Sep. 25, 2025. [Online]. Available: https://arxiv.org/pdf/1502.03167

[172] K. M. Ting and I. H. Witten, “Issues in Stacked Generalization,” *Journal of Artificial Intelligence Research*, vol. 10, pp. 271–289, May 1999, doi: 10.1613/JAIR.594.

[173] D. H. Wolpert, “Stacked generalization,” *Neural Networks*, vol. 5, no. 2, pp. 241–259, Jan. 1992, doi: 10.1016/S0893-6080(05)80023-1.

[174] S. Džeroski and B. Ženko, “Is combining classifiers with stacking better than selecting the best one?,” *Mach. Learn.*, vol. 54, no. 3, pp. 255–273, Mar. 2004, doi: 10.1023/B:MACH.0000015881.36452.6E/METRICS.

[175] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/NATURE14539;SUBJMETA.

[176] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998, doi: 10.1109/5.726791.

[177] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/NECO.1997.9.8.1735.

[178] K. Cho *et al.*, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1724–1734, Jun. 2014, doi: 10.3115/v1/d14-1179.

[179] A. Vaswani *et al.*, “Attention Is All You Need,” p. 1, Jun. 2017, Accessed: Sep. 25, 2025. [Online]. Available: https://arxiv.org/pdf/1706.03762

[180] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Sep. 25, 2025. [Online]. Available: https://arxiv.org/pdf/1810.04805

[181] T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” *Adv. Neural Inf. Process. Syst.*, vol. 2020-December, May 2020, Accessed: Sep. 25, 2025. [Online]. Available: https://arxiv.org/pdf/2005.14165

[182] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, “Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges,” Apr. 2021, Accessed: Sep. 25, 2025. [Online]. Available: https://arxiv.org/pdf/2104.13478

[183] F. B. Fuchs, D. E. Worrall, V. Fischer, and M. Welling, “SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks,” *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1970–1981, 2020, Accessed: Sep. 25, 2025. [Online]. Available: https://github.com/FabianFuchsML/se3-transformer-public

[184] V. G. Satorras, E. Hoogeboom, and M. Welling, “E(n) Equivariant Graph Neural Networks,” *Proc. Mach. Learn. Res.*, vol. 139, pp. 9323–9332, Feb. 2021, Accessed: Sep. 25, 2025. [Online]. Available: https://arxiv.org/pdf/2102.09844

[185] J. Jumper *et al.*, “Highly accurate protein structure prediction with AlphaFold,” *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/S41586-021-03819-2;TECHMETA.

[186] S. Jones, “Protein-RNA interactions: structural biology and computational modeling techniques,” *Biophys. Rev.*, vol. 8, no. 4, pp. 359–367, Dec. 2016, doi: 10.1007/S12551-016-0223-9.

[187] B. M. Lunde, C. Moore, and G. Varani, “RNA-binding proteins: modular design for efficient function,” *Nat. Rev. Mol. Cell Biol.*, vol. 8, no. 6, pp. 479–490, Jun. 2007, doi: 10.1038/nrm2178.

[188] K. Harini, D. Kihara, and M. Michael Gromiha, “PDA-Pred: Predicting the binding affinity of protein-DNA complexes using machine learning techniques and structural features.,” *Methods*, vol. 213, pp. 10–17, May 2023, doi: 10.1016/j.ymeth.2023.03.002.

[189] W. Yang and L. Deng, “PreDBA: A heterogeneous ensemble approach for predicting protein-DNA binding affinity,” *Sci. Rep.*, vol. 10, no. 1, p. 1278, Jan. 2020, doi: 10.1038/s41598-020-57778-1.

[190] J. T. Vivian and P. R. Callis, “Mechanisms of Tryptophan Fluorescence Shifts in Proteins,” *Biophys. J.*, vol. 80, no. 5, pp. 2093–2109, May 2001, doi: 10.1016/S0006-3495(01)76183-8.

[191] P. Katsamba, “Kinetic studies of RNA–protein interactions using surface plasmon resonance,” *Methods*, vol. 26, no. 2, pp. 95–104, Feb. 2002, doi: 10.1016/S1046-2023(02)00012-9.

[192] L. M. Hellman and M. G. Fried, “Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions,” *Nat. Protoc.*, vol. 2, no. 8, pp. 1849–1861, Aug. 2007, doi: 10.1038/nprot.2007.249.

[193] S. P. Ryder, M. I. Recht, and J. R. Williamson, “Quantitative Analysis of Protein-RNA Interactions by Gel Mobility Shift,” 2008, pp. 99–115. doi: 10.1007/978-1-60327-475-3_7.

[194] A. L. Feig, “Studying RNA-RNA and RNA-protein interactions by isothermal titration calorimetry.,” *Methods Enzymol.*, vol. 468, pp. 409–22, 2009, doi: 10.1016/S0076-6879(09)68019-8.

[195] D. C. Rio, “Filter-Binding Assay for Analysis of RNA–Protein Interactions,” *Cold Spring Harb. Protoc.*, vol. 2012, no. 10, p. pdb.prot071449, Oct. 2012, doi: 10.1101/pdb.prot071449.

[196] D. R. Bell, J. K. Weber, W. Yin, T. Huynh, W. Duan, and R. Zhou, “In silico design and validation of high-affinity RNA aptamers targeting epithelial cellular adhesion molecule dimers,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 15, pp. 8486–8493, Apr. 2020, doi: 10.1073/pnas.1913242117.

[197] S. Bheemireddy, S. Sandhya, N. Srinivasan, and R. Sowdhamini, “Computational tools to study RNA-protein complexes,” *Front. Mol. Biosci.*, vol. 9, Oct. 2022, doi: 10.3389/fmolb.2022.954926.

[198] M. Baek, R. McHugh, I. Anishchenko, H. Jiang, D. Baker, and F. DiMaio, “Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA,” *Nat. Methods*, vol. 21, no. 1, pp. 117–121, Jan. 2024, doi: 10.1038/s41592-023-02086-5.

[199] Z. Ding and D. Kihara, “Computational Methods for Predicting Protein-Protein Interactions Using Various Protein Features.,” *Curr. Protoc. Protein Sci.*, vol. 93, no. 1, p. e62, Aug. 2018, doi: 10.1002/cpps.62.

[200] C. Pons, D. Talavera, X. De La Cruz, M. Orozco, and J. Fernandez-Recio, “Scoring by intermolecular pairwise propensities of exposed residues (SIPPER): A new efficient potential for protein-protein docking,” *J. Chem. Inf. Model.*, vol. 51, no. 2, pp. 370–377, Feb. 2011, doi: 10.1021/CI100353E/SUPPL_FILE/CI100353E_SI_001.PDF.

[201] Y. Li, J. Shen, X. Sun, W. Li, G. Liu, and Y. Tang, “Accuracy assessment of protein-based docking programs against RNA targets,” *J. Chem. Inf. Model.*, vol. 50, no. 6, pp. 1134–1146, Jun. 2010, doi: 10.1021/CI9004157/SUPPL_FILE/CI9004157_SI_001.PDF.

[202] K. Harini, M. Sekijima, and M. M. Gromiha, “PRA-Pred: Structure-based prediction of protein-RNA binding affinity,” *Int. J. Biol. Macromol.*, vol. 259, p. 129490, Feb. 2024, doi: 10.1016/j.ijbiomac.2024.129490.

[203] X. Hong *et al.*, “An updated dataset and a structure-based prediction model for protein–<scp>RNA</scp> binding affinity,” *Proteins: Structure, Function, and Bioinformatics*, vol. 91, no. 9, pp. 1245–1253, Sep. 2023, doi: 10.1002/prot.26503.

[204] C. Nithin, S. Mukherjee, and R. P. Bahadur, “A structure-based model for the prediction of protein–RNA binding affinity,” *RNA*, vol. 25, no. 12, pp. 1628–1645, Dec. 2019, doi: 10.1261/rna.071779.119.

[205] L. Deng, W. Yang, and H. Liu, “PredPRBA: Prediction of Protein-RNA Binding Affinity Using Gradient Boosted Regression Trees,” *Front. Genet.*, vol. 10, Aug. 2019, doi: 10.3389/fgene.2019.00637.

[206] H. M. Berman *et al.*, “The Protein Data Bank,” 2000. [Online]. Available: http://www.rcsb.org/pdb/status.html

[207] F. Sievers *et al.*, “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega,” *Mol. Syst. Biol.*, vol. 7, no. 1, Jan. 2011, doi: 10.1038/msb.2011.75.

[208] X. Hong *et al.*, “An updated dataset and a structure-based prediction model for protein–RNA binding affinity,” *Proteins: Structure, Function and Bioinformatics*, vol. 91, no. 9, pp. 1245–1253, Sep. 2023, doi: 10.1002/prot.26503.

[209] R. Wang, X. Fang, Y. Lu, and S. Wang, “The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures,” *J. Med. Chem.*, vol. 47, no. 12, pp. 2977–2980, Jun. 2004, doi: 10.1021/jm030580l.

[210] K. Harini, A. Srivastava, A. Kulandaisamy, and M. M. Gromiha, “ProNAB: database for binding affinities of protein–nucleic acid complexes and their mutants,” *Nucleic Acids Res.*, vol. 50, no. D1, pp. D1528–D1534, Jan. 2022, doi: 10.1093/nar/gkab848.

[211] C. Nithin, S. Mukherjee, and R. P. Bahadur, “A structure-based model for the prediction of protein-RNA binding affinity,” 2019, doi: 10.1261/rna.

[212] A. Pedretti, L. Villa, and G. Vistoli, “VEGA–an open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming,” *J. Comput. Aided. Mol. Des.*, vol. 18, no. 3, pp. 167–173, 2004.

[213] A. Fiser, R. K. G. Do, and A. Šali, “Modeling of loops in protein structures,” *Protein Science*, vol. 9, no. 9, pp. 1753–1773, Jan. 2000, doi: 10.1110/ps.9.9.1753.

[214] E. F. Pettersen *et al.*, “UCSF Chimera—A visualization system for exploratory research and analysis,” *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, Oct. 2004, doi: 10.1002/jcc.20084.

[215] D. A. Case *et al.*, “AmberTools,” *J. Chem. Inf. Model.*, vol. 63, no. 20, pp. 6183–6191, Oct. 2023, doi: 10.1021/acs.jcim.3c01153.

[216] C. Tian *et al.*, “ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution,” *J. Chem. Theory Comput.*, vol. 16, no. 1, pp. 528–552, Jan. 2020, doi: 10.1021/acs.jctc.9b00591.

[217] M. Zgarbová *et al.*, “Refinement of the Cornell et al. Nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles,” *J. Chem. Theory Comput.*, vol. 7, no. 9, pp. 2886–2902, Sep. 2011, doi: 10.1021/ct200162x.

[218] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, “Refined TIP3P model for water,” *J Chem Phys*, vol. 79, pp. 926–935, 1983.

[219] H. C. Andersen, “Molecular dynamics simulations at constant pressure and/or temperature,” *J. Chem. Phys.*, vol. 72, no. 4, pp. 2384–2393, Feb. 1980, doi: 10.1063/1.439486.

[220] B. J. Alder and T. E. Wainwright, “Studies in Molecular Dynamics. I. General Method,” *J. Chem. Phys.*, vol. 31, no. 2, pp. 459–466, Aug. 1959, doi: 10.1063/1.1730376.

[221] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, “LINCS: A linear constraint solver for molecular simulations,” *J. Comput. Chem.*, vol. 18, no. 12, pp. 1463–1472, Sep. 1997, doi: https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

[222] K. L. Rallapalli, B. L. Ranzau, K. R. Ganapathy, F. Paesani, and A. C. Komor, “Combined Theoretical, Bioinformatic, and Biochemical Analyses of RNA Editing by Adenine Base Editors,” *CRISPR J.*, vol. 5, no. 2, pp. 294–310, Apr. 2022, doi: 10.1089/crispr.2021.0131.

[223] T. Ando, “Molecular Dynamics Simulations of RNA Stem-Loop Folding Using an Atomistic Force Field and a Generalized Born Implicit Solvent,” *ACS Omega*, Oct. 2025, doi: 10.1021/acsomega.5c05377.

[224] T. K. Ho, “Random decision forests,” in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995, pp. 278–282 vol.1. doi: 10.1109/ICDAR.1995.598994.

[225] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001, [Online]. Available: http://www.jstor.org/stable/2699986

[226] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[227] D. H. Wolpert, “Stacked generalization,” *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992, doi: https://doi.org/10.1016/S0893-6080(05)80023-1.

[228] F. Pedregosa FABIANPEDREGOSA *et al.*, “Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot,” 2011. [Online]. Available: http://scikit-learn.sourceforge.net.

[229] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, in IJCAI’95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143.

[230] J. Bergstra, J. B. Ca, and Y. B. Ca, “Random Search for Hyper-Parameter Optimization Yoshua Bengio,” 2012. [Online]. Available: http://scikit-learn.sourceforge.net.

[231] J. H. Zar, “Significance Testing of the Spearman Rank Correlation Coefficient,” *J. Am. Stat. Assoc.*, vol. 67, no. 339, pp. 578–580, Sep. 1972, doi: 10.1080/01621459.1972.10481251.

[232] B. M. Lunde, C. Moore, and G. Varani, “RNA-binding proteins: modular design for efficient function,” *Nat. Rev. Mol. Cell Biol.*, vol. 8, no. 6, pp. 479–490, Jan. 2007, doi: 10.1038/nrm2178.

[233] K. Harini, D. Kihara, and M. Michael Gromiha, “PDA-Pred: Predicting the binding affinity of protein-DNA complexes using machine learning techniques and structural features,” *Methods*, vol. 213, pp. 10–17, May 2023, doi: 10.1016/j.ymeth.2023.03.002.

[234] W. Yang and L. Deng, “PreDBA: A heterogeneous ensemble approach for predicting protein-DNA binding affinity,” *Sci. Rep.*, vol. 10, no. 1, p. 1278, Jan. 2020, doi: 10.1038/s41598-020-57778-1.

[235] X. Pan and H. Bin Shen, “Predicting RNA-protein binding sites and motifs through combining local and global deep convolutional neural networks,” *Bioinformatics*, vol. 34, no. 20, pp. 3427–3436, Oct. 2018, doi: 10.1093/bioinformatics/bty364.

[236] P. J. A. Cock *et al.*, “Biopython: freely available Python tools for computational molecular biology and bioinformatics,” *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, Jun. 2009, doi: 10.1093/bioinformatics/btp163.

[237] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[238] T. R. Damase, R. Sukhovershin, C. Boada, F. Taraballi, R. I. Pettigrew, and J. P. Cooke, “The Limitless Future of RNA Therapeutics,” *Front. Bioeng. Biotechnol.*, vol. 9, Mar. 2021, doi: 10.3389/FBIOE.2021.628137.

[239] Y. Zhu, L. Zhu, X. Wang, and H. Jin, “RNA-based therapeutics: an overview and prospectus,” *Cell Death & Disease 2022 13:7*, vol. 13, no. 7, pp. 1–15, Jul. 2022, doi: 10.1038/s41419-022-05075-2.

[240] I. L. Hofacker, “Vienna RNA secondary structure server,” *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3429–3431, Jul. 2003, doi: 10.1093/NAR/GKG599.

[241] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, “Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure,” *J. Mol. Biol.*, vol. 288, no. 5, pp. 911–940, May 1999, doi: 10.1006/jmbi.1999.2700.

[242] R. Das, J. Karanicolas, and D. Baker, “Atomic accuracy in predicting and designing noncanonical RNA structure,” *Nat. Methods*, vol. 7, no. 4, pp. 291–294, 2010, doi: 10.1038/NMETH.1433.

[243] J. K. Leman *et al.*, “Macromolecular modeling and design in Rosetta: recent methods and frameworks,” *Nat. Methods*, vol. 17, no. 7, pp. 665–680, Jul. 2020, doi: 10.1038/S41592-020-0848-2.

[244] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, “Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges,” Apr. 2021, Accessed: Oct. 31, 2025. [Online]. Available: https://arxiv.org/pdf/2104.13478

[245] P. Bibekar, L. F. Krapp, and M. D. Peraro, “Context-aware geometric deep learning for RNA sequence design,” *bioRxiv*, p. 2025.06.21.660801, Jun. 2025, doi: 10.1101/2025.06.21.660801.

[246] L. F. Krapp *et al.*, “Context-aware geometric deep learning for protein sequence design,” *Nature Communications 2024 15:1*, vol. 15, no. 1, pp. 1–10, Jul. 2024, doi: 10.1038/s41467-024-50571-y.

[247] C. Tan *et al.*, “RDesign: Hierarchical Data-efficient Representation Learning for Tertiary Structure-based RNA Design,” *12th International Conference on Learning Representations, ICLR 2024*, Jan. 2023, Accessed: Oct. 31, 2025. [Online]. Available: https://arxiv.org/pdf/2301.10774

[248] L. R. Ganser, M. L. Kelly, D. Herschlag, and H. M. Al-Hashimi, “The roles of structural dynamics in the cellular functions of RNAs,” *Nat. Rev. Mol. Cell Biol.*, vol. 20, no. 8, pp. 474–489, Aug. 2019, doi: 10.1038/S41580-019-0136-0.

[249] M. L. Ken *et al.*, “RNA conformational propensities determine cellular activity,” *Nature*, vol. 617, no. 7962, pp. 835–841, May 2023, doi: 10.1038/S41586-023-06080-X.

[250] J. R. Stagno *et al.*, “Structures of riboswitch RNA reaction states by mix-and-inject XFEL serial crystallography,” *Nature*, vol. 541, no. 7636, pp. 242–246, Jan. 2017, doi: 10.1038/NATURE20599.

[251] C. K. Joshi *et al.*, “gRNAde: Geometric Deep Learning for 3D RNA inverse design,” *13th International Conference on Learning Representations, ICLR 2025*, pp. 56390–56415, May 2023, Accessed: Oct. 31, 2025. [Online]. Available: https://arxiv.org/pdf/2305.14749

[252] M. Zaheer, S. Kottur, S. Ravanbhakhsh, B. Póczos, R. Salakhutdinov, and A. J. Smola, “Deep Sets,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-December, pp. 3392–3402, Mar. 2017, Accessed: Oct. 31, 2025. [Online]. Available: https://arxiv.org/pdf/1703.06114

[253] H. Huang, Z. Lin, D. He, L. Hong, and Y. Li, “RiboDiffusion: tertiary structure-based RNA inverse folding with generative diffusion models,” *Bioinformatics*, vol. 40, no. Suppl 1, pp. i347–i356, Jul. 2024, doi: 10.1093/BIOINFORMATICS/BTAE259.

[254] F. Wong *et al.*, “Deep generative design of RNA aptamers using structural predictions,” *Nat. Comput. Sci.*, vol. 4, no. 11, pp. 829–839, Nov. 2024, doi: 10.1038/S43588-024-00720-6.

[255] J. Dauparas *et al.*, “Robust deep learning-based protein sequence design using ProteinMPNN,” *Science*, vol. 378, no. 6615, pp. 49–56, Oct. 2022, doi: 10.1126/SCIENCE.ADD2187.

[256] R. Wang, X. Fang, Y. Lu, C. Y. Yang, and S. Wang, “The PDBbind database: methodologies and updates,” *J. Med. Chem.*, vol. 48, no. 12, pp. 4111–4119, Jun. 2005, doi: 10.1021/JM048957Q.

[257] W. Wang *et al.*, “trRosettaRNA: automated prediction of RNA 3D structure with transformer network,” *Nature Communications 2023 14:1*, vol. 14, no. 1, pp. 1–13, Nov. 2023, doi: 10.1038/s41467-023-42528-4.

[258] B. Adamczyk, M. Antczak, and M. Szachniuk, “RNAsolo: a repository of cleaned PDB-derived RNA 3D structures,” *Bioinformatics*, vol. 38, no. 14, pp. 3668–3670, Jul. 2022, doi: 10.1093/BIOINFORMATICS/BTAC386.

[259] C. Zhang, M. Shine, A. M. Pyle, and Y. Zhang, “US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes,” *Nat. Methods*, vol. 19, no. 9, pp. 1109–1115, Sep. 2022, doi: 10.1038/S41592-022-01585-1.

[260] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, “CD-HIT: accelerated for clustering the next-generation sequencing data,” *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012, doi: 10.1093/BIOINFORMATICS/BTS565.

[261] L. R. Ganser, M. L. Kelly, D. Herschlag, and H. M. Al-Hashimi, “The roles of structural dynamics in the cellular functions of RNAs,” *Nat. Rev. Mol. Cell Biol.*, vol. 20, no. 8, pp. 474–489, Aug. 2019, doi: 10.1038/S41580-019-0136-0.

[262] A. Paszke *et al.*, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” *Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, Accessed: Oct. 31, 2025. [Online]. Available: https://arxiv.org/pdf/1912.01703

[263] “NVIDIA Announces Financial Results for Fourth Quarter and Fiscal 2020 | NVIDIA Newsroom.” Accessed: Oct. 31, 2025. [Online]. Available: https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-fourth-quarter-and-fiscal-2020

[264] D. P. Kingma and J. L. Ba, “Adam: A Method for Stochastic Optimization,” *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Dec. 2014, Accessed: Oct. 31, 2025. [Online]. Available: https://arxiv.org/pdf/1412.6980

[265] I. Loshchilov and F. Hutter, “SGDR: Stochastic Gradient Descent with Warm Restarts,” *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, Aug. 2016, Accessed: Oct. 31, 2025. [Online]. Available: https://arxiv.org/pdf/1608.03983

[266] R. Azad *et al.*, “Loss Functions in the Era of Semantic Segmentation: A Survey and Outlook,” Dec. 2023, Accessed: Oct. 31, 2025. [Online]. Available: http://arxiv.org/abs/2312.05391

[267] “Deep Learning.” Accessed: Oct. 31, 2025. [Online]. Available: https://www.deeplearningbook.org/

[268] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, “Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges,” Apr. 2021, Accessed: Oct. 31, 2025. [Online]. Available: https://arxiv.org/pdf/2104.13478

[269] Z. Lu *et al.*, “Ribocentre-aptamer: an integrative, structure-focused database for RNA aptamers,” *Nucleic Acids Res.*, vol. 1, no. 1256879, pp. 13–14, 2013, doi: 10.1093/NAR/GKAF1016.

[270] M. L. Ken *et al.*, “RNA conformational propensities determine cellular activity,” *Nature*, vol. 617, no. 7962, pp. 835–841, May 2023, doi: 10.1038/S41586-023-06080-X.