UNIVERSITY OF
CHEMISTRY AND TECHNOLOGY
PRAGUE

Faculty of Chemical Technology

Cheminformatics analysis of RNA-binding ligands

MASTER THESIS

Ing. Jozef Fülöp

Prague 2023

SUMMARY

Cheminformatics analysis of RNA-binding ligands

This master's thesis explores the classification and analysis of RNA-binding molecules utilizing chemoinformatics and sophisticated machine learning techniques. It conducts a comprehensive review of chemical databases to uncover unique structural features and properties that differentiate RNA ligands from protein ligands. Central to this analysis is the employment of advanced machine learning strategies, integrating ensemble methods with graph neural networks to enhance the identification and classification processes of RNA-binding molecules. Techniques like t-SNE and UMAP are utilized to visualize and interpret complex chemical spaces, providing deeper insights into molecular interactions with RNA. By leveraging tools such as the RDKit library and machine learning platforms including scikit-learn, PyTorch, and the Deep Graph Library (DGL), this project aims to develop robust models that predict and analyze the behavior of potential therapeutic agents. The ultimate goal of this research is to advance our understanding of small molecule interactions with RNA, which is crucial for innovating new treatments for RNA-related diseases. This thesis highlights the significant impact of combining cheminformatics with ensemble machine learning methods on the progress of medical science and pharmaceutical innovation.

SOUHRN

Cheminformatická analýza ligandů RNA

Tato magisterská práce se zabývá klasifikací a analýzou molekul vázajících RNA s využitím chemoinformatiky a sofistikovaných technik strojového učení. Provádí komplexní přehled chemických databází s cílem odhalit jedinečné strukturní rysy a vlastnosti, které odlišují RNA ligandy od proteinových ligandů. Ústředním bodem této analýzy je využití pokročilých strategií strojového učení, které integrují ansámblové metody s grafovými neuronovými sítěmi s cílem zlepšit procesy identifikace a klasifikace molekul vázajících RNA. Techniky jako t-SNE a UMAP jsou využívány k vizualizaci a interpretaci složitých chemických prostorů, což poskytuje hlubší vhled do molekulárních interakcí s RNA. S využitím nástrojů, jako je knihovna RDKit a platformy strojového učení včetně scikit-learn, PyTorch a Deep Graph Library (DGL), je cílem tohoto projektu vyvinout robustní modely, které předpovídají a analyzují chování potenciálních terapeutických látek. Konečným cílem tohoto výzkumu je zlepšit naše chápání interakcí malých molekul s RNA, což je zásadní pro inovaci nových způsobů léčby nemocí souvisejících s RNA. Tato práce poukazuje na významný dopad kombinace cheminformatiky s metodami ansámblového strojového učení na pokrok lékařské vědy a farmaceutické inovace.

ACKNOWLEDGEMENT

I would like to extend my deepest gratitude to my supervisor, Professor Daniel Svozil, and my consultant, Professor Andrea Brancale, for their expert guidance and invaluable advice throughout my research. My sincere thanks are also due to my colleagues from the Department of Informatics and Chemistry for their continual inspiration and support during this project. I must also acknowledge my family and closest ones for their unwavering patience and encouragement on this journey.

Computational Resources: This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254).

Contents

1 INTRODUCTION

RNA, or ribonucleic acid, plays a pivotal role in various biological processes, serving as a crucial intermediary in the translation of genetic information into functional proteins. Beyond this traditional role, RNA has become a significant target in drug discovery due to its involvement in numerous disease pathways [1]. The structural complexity of RNA, characterized by dynamic conformational changes and diverse secondary and tertiary structures, presents unique challenges in designing RNA-binding small molecules [1, 2].

The specificity of RNA-ligand interactions adds to the complexity of targeting RNA in drug discovery. The highly charged and flexible nature of RNA requires the development of small molecules with precise physicochemical properties for selective binding [3]. These characteristics require careful adjustment to align with the molecular variety and flexibility of RNA structures, posing significant challenges but also offering immense therapeutic potential [4, 5].

Cheminformatics, which combines chemistry, computer science, and information technology, has become an essential tool in addressing these challenges. It applies computational techniques to analyze chemical data, predict molecular properties, and design novel compounds [1]. In this field, cheminformatics aids in exploring chemical space, analyzing molecular properties, and predicting RNA-ligand interactions [5]. The integration of these methodologies enhances our capacity to discover and optimize novel RNA-targeted therapies efficiently.

This thesis utilizes comprehensive datasets from specialized libraries such as the Enamine Hit Locator Library and ChemDiv's miRNA-targeted Library, among others. These sources provide a rich array of compounds focused on RNA-binding, supporting cheminformatic analyses. Using these diverse resources, this work aim to deepen our understanding of the interaction dynamics between small molecules and RNA targets.

The research primarily describes the chemical space and properties of RNA-binding molecules and applies machine learning and deep learning models for their analysis. These techniques, including advanced predictive models and chemical space visualization methods like UMAP, t-SNE, and logistic PCA, help to investigate the chemical properties of RNA-binding ligands. By identifying patterns and trends, we can guide the design of novel therapeutics.

In conclusion, the methodologies and analyses presented advance our understanding of RNA-binding ligands as potential therapeutic agents. Through cheminformatics and machine learning, this research navigates the complexities of RNA structures and interactions, contributing significantly to drug discovery and the development of innovative therapies.

2 LITERATURE

2.1 Introduction to Cheminformatics in RNA Research

Cheminformatics merges computational techniques with chemical insights to revolutionize the approach to drug discovery, particularly targeting RNA molecules. The multidisciplinary field combines aspects of chemistry, computer science, and biology, enabling the systematic exploration and manipulation of chemical data to discover potential RNA-binding ligands. The use of advanced computational tools in cheminformatics facilitates the prediction of molecular behavior and interactions at an unprecedented scale and speed [6].

RNA molecules play vital roles in various cellular processes, including regulation of gene expression and protein synthesis. The structural diversity and functional importance of this macromolecule make it an attractive target for therapeutic intervention. However, targeting RNA presents unique challenges because of its dynamic nature and the structural complexity of its folded forms. Traditional drug discovery efforts have predominantly focused on proteins, but the increasing understanding of RNA's biological significance has shifted some focus to RNA [2].

The integration of artificial intelligence (AI) and machine learning (ML) technologies into cheminformatics has further propelled this shift. AI-driven tools in cheminformatics have streamlined the drug discovery process, enabling the rapid screening of vast chemical libraries and enhancing the prediction accuracy of RNA-ligand interactions. These technologies allow for a more nuanced understanding of the chemical space relevant to RNA-binding molecules, helping to identify potent and selective compounds [6].

Despite advances in the development of small molecules that effectively target RNA is still in its infancy. One of the primary hurdles is the lack of comprehensive understanding of the types of small molecules that can bind RNA specifically. The structural variability of RNA, from linear messenger RNAs to complex tertiary structures found in ribosomal RNAs, complicates the design of effective ligands. Moreover, the RNA-targeted drug discovery process often encounters issues with selectivity and efficacy due to the RNA's ability to adopt multiple conformations [7].

Efforts to overcome these challenges have been diverse. Computational drug discovery under RNA-focused conditions has seen the employment of structure-based approaches that leverage molecular dynamics simulations and virtual screening to predict how small molecules interact with RNA targets. This approach has identified novel small-molecule modulators that show promise in modulating RNA function with high specificity [2].

The computational chemists now face the task of extending traditional methodologies designed for protein targets to accommodate the specific characteristics of RNA structures. This includes addressing the high charge density and complex structural dynamics of RNA molecules, which require a rethinking of standard computational procedures to enhance the accuracy of predictions and the effectiveness of the designed molecules [2].

In summary, cheminformatics has become indispensable in the realm of RNA-targeted drug discovery. The field continues to evolve rapidly, driven by technological advances in computational methodologies and a deeper understanding of RNA biology. As cheminformatics tools become more sophisticated and our understanding of RNA structures improves, the potential to design effective RNA-targeting therapies becomes increasingly feasible, promising new avenues for treating diseases at the genetic and molecular levels.

2.2 Advancements in Cheminformatics Tools

The evolution of cheminformatics tools has revolutionized the field of RNA-targeted drug discovery, marked by significant advances in computational methods and the development of specialized databases. These innovations have broadened the scope of potential therapeutic targets, extending beyond traditional protein targets to encompass structurally complex and functionally diverse RNA molecules.

Computer-Aided Design of RNA-Targeted Small Molecules has emerged as a cornerstone in the design and optimization of RNA-binding ligands. The complexity of RNA structures, from their folding patterns to their involvement in critical cellular processes, poses unique challenges that require precise and sophisticated computational tools. Computer-aided design (CAD) systems enable detailed molecular modeling and simulation, providing insights that are crucial for the successful targeting of RNA. These tools facilitate the exploration of new compounds with potential therapeutic effects, significantly speeding up the drug discovery process by predicting how small molecules interact with RNA targets [1].

The Role of Cheminformatics in Enabling RNA Drug Discovery has expanded as tools and methodologies have advanced. Databases such as R-SIM and HARIBOSS exemplify key developments in the field, offering comprehensive data that support the rational design of RNA-targeted therapies. R-SIM provides a robust platform that includes binding affinities and interaction data, which are instrumental in understanding the selectivity and efficacy of RNA-binding molecules [8]. Similarly, HARIBOSS offers curated structures of RNA-small molecule complexes, aiding researchers in visualizing potential interactions and predicting the therapeutic potential of new compounds [9].

Contemporary Progress and Opportunities in RNA-targeted drug discovery have been influenced by these cheminformatics tools. The integration of high-throughput screening, molecular docking, and virtual screening has become standard in the industry, driven by the capabilities provided by advanced computational tools. These methods have not only improved the efficiency of identifying promising compounds but have also enhanced the accuracy of targeting specific RNA motifs, which is critical for the development of effective therapeutics [10].

In addition, continued updates and enhancements of RNA-targeted databases facilitate ongoing improvements in drug discovery processes. For example, databases such as R-BIND, which catalog bioactive RNA-targeting small molecules and their associated RNA secondary structures, offer new insights and tools to design small molecules based on physicochemical and spatial properties analyzed [4]. These resources are invaluable for researchers seeking to design and screen compounds in what is considered an 'RNA-privileged' chemical space, thereby supporting the targeted development of new RNA-focused therapies.

Furthermore, analysis from databases such as R-BIND highlights the physicochemical properties and spatial characteristics that distinguish RNA-binding molecules, providing invaluable insights that drive the rational design and optimization of these agents [4]. This analysis reveals that RNA binders often exhibit unique features like increased aromaticity, hydrogen bonding capacity, and specific spatial arrangements, which are crucial for engaging with RNA structures effectively. The detailed characterization of these molecules enriches our understanding of RNA-binding chemical properties and supports the hypothesis of a unique RNA-targeting chemical space [4].

In summary, advances in cheminformatics tools have not only improved our understanding of RNA as a drug target, but have also provided the technological foundation necessary for developing novel RNA-targeted therapeutics. As these tools continue to evolve, they promise to further accelerate the discovery and optimization of drugs that can modulate RNA functions, potentially transforming treatment paradigms for a variety of diseases.

2.3 AI and ML in Drug Discovery

The application of machine learning (ML) and artificial intelligence (AI) in RNA drug discovery has been transformative, particularly with the adoption of advanced modeling techniques such as Geometric Graph Neural Networks (GNNs) and ensemble learning algorithms like XGBoost. These technologies have significantly enhanced the ability to model and understand the complex three-dimensional structures of RNA, which are crucial for the development of RNA-targeted therapeutics.

2.3.1 Ensemble Learning

Ensemble methods, such as XGBoost, Random Forest, and LightGBM, enhance the predictive performance by combining the outputs of individual models, thus producing a more accurate prediction than any single model. XGBoost, or eXtreme Gradient Boosting, stands out by utilizing a gradient boosting framework at its core. This method involves constructing new models that predict the residuals or errors of prior models and then combining them to make the final prediction. The strength of XGBoost lies in its scalability and ability to handle sparse data efficiently.

XGBoost, a renowned scalable tree boosting system, is extensively utilized by data scientists to achieve state-of-the-art results in various machine learning challenges. Developed by Tianqi Chen and Carlos Guestrin, XGBoost improves upon traditional gradient boosting methods by introducing a novel sparsity-aware algorithm that handles sparse data effectively, and a weighted quantile sketch for approximate tree learning. These enhancements allow XGBoost to handle larger datasets more efficiently than traditional methods. Furthermore, XGBoost incorporates advanced features like cache-aware block structures for out-of-core tree learning and parallel and distributed computing, significantly speeding up the learning process and facilitating quicker model iterations. As such, XGBoost not only accelerates the training process but also scales effectively to billions of examples, making it a preferred choice in large-scale machine learning applications. [11].

Random forests (RF) are an influential ensemble learning method used extensively for classification and regression tasks, characterized by their robustness and high accuracy. As introduced by Leo Breiman, a random forest consists of numerous decision trees operating as a collective. Each tree contributes to the final decision, which makes the method less prone to overfitting and enhances its generalizability. In classification tasks, RF predicts the class that is the mode of the classes predicted by individual trees. For regression, it uses the mean prediction of the trees. This technique effectively reduces error by averaging the predictions of multiple deep decision trees, each trained on a different subset of the same data set. The final prediction is thus more accurate and stable than that of individual trees. Breiman's seminal work has underscored the strength of combining the simplicity of decision trees with powerful ensemble techniques to achieve superior accuracy across various applications [12].

LightGBM, a highly efficient Gradient Boosting Decision Tree (GBDT) implementation, stands out for its speed and accuracy in handling large datasets with high-dimensional features. Developed by Guolin Ke et al., LightGBM optimizes the traditional GBDT approaches by introducing two novel techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS improves information gain estimation by focusing on data instances with larger gradients, effectively prioritizing more significant instances without extensive data requirements. EFB reduces the number of features in sparse datasets by bundling mutually exclusive features, thus maintaining accuracy while significantly decreasing computational complexity. These innovations enable LightGBM to achieve remarkable speeds—up to 20 times faster than conventional GBDT—without compromising the accuracy, making it an excellent choice for real-world applications that require efficient processing of extensive data [13].

2.3.2 Graph Neural Networks

Graph Neural Networks (GNNs) represent a significant advancement in artificial intelligence, specifically designed to handle non-Euclidean data like graph-structured information, which is common in cheminformatics [14, 15]. Unlike traditional neural networks, GNNs are adept at managing complex relational data inherent to molecular structures, making them particularly valuable for modeling interactions within biological systems [16].

These GNN models are instrumental in cheminformatics, particularly in tasks such as predicting interactions between RNA molecules and potential ligands. The Graph Convolutional Network (GCN) continues to play a pivotal role by aggregating features from a node's neighbors to refine its representation through iterative processing [17]. Additionally, Gated Graph Neural Networks (GG-NNs) use Gated Recurrent Units to handle non-sequential outputs, enhancing the model's ability to manage complex molecular interactions without the need for parameter constraints to ensure convergence [18]. GraphSAGE extends the capabilities of GCNs by learning to aggregate feature information from the local neighborhood of nodes, utilizing standard backpropagation techniques[19]. The inclusion of Graph Attention Network (GAT), which employs an attention mechanism to prioritize critical node features, further underscores the sophistication of GNNs in discerning essential molecular features.[20].

2.4 Current Challenges and Future Directions in RNA-Targeted Drug Discovery

The landscape of RNA-targeted drug discovery is rapidly evolving, driven by significant scientific advancements and a deeper understanding of RNA's biological importance. However, numerous challenges inhibit progress, necessitating innovative approaches to realize the therapeutic potential of RNA.

Computational and Predictive Challenges: Computational models leveraging machine learning and deep learning have revolutionized RNA informatics, enhancing the accuracy of RNA structure and interaction predictions. Despite these advancements, the dynamic nature of RNA poses significant predictive challenges. RNA molecules exhibit multiple conformations influenced by their cellular environments, complicating accurate modeling. Addressing these complexities requires advanced computational techniques capable of adapting to the variable structural configurations of RNA under physiological conditions[21].

Moreover, the predictive models are often limited by the availability of comprehensive, annotated datasets necessary for training. The development of robust, generalizable models is constrained by the scarcity of high-quality RNA structural data. Expanding these datasets is crucial for advancing RNA informatics and enhancing the predictive performance of computational models[21, 22].

Biological and Chemical Intricacies: The inherent flexibility and transient configurations of RNA present substantial biological challenges in identifying stable, druggable targets. The effectiveness of potential therapeutics depends on stable and predictable RNA interactions, which are currently difficult to achieve due to RNA's structural variability[23].

From a chemical perspective, achieving specificity in RNA-binding interactions remains a significant challenge. Many RNA-targeted compounds exhibit off-target effects, leading to potential toxicity and reduced therapeutic efficacy. The development of highly specific RNA-binding ligands is essential for minimizing these effects and maximizing the therapeutic potential of RNA-targeted treatments[23].

Technological Advancements and Future Prospects: Emerging technologies in structural biology and genomics, such as cryo-electron microscopy and next-generation sequencing, promise to overcome some of the existing barriers. These techniques are expected to reveal detailed insights into RNA structures and interactions, facilitating the discovery of novel RNA targets and the design of more precise therapies[24]. Additionally, the discovery of RNA-binding small molecules often involves the identification of ligands that can specifically interact through stacking and hydrogen-bonding interactions. This approach enhances the potential to stabilize complex RNA structures like G-quadruplexes or target specific RNA motifs through precise molecular recognition strategies[25].

In conclusion, the field of RNA-targeted drug discovery, while faced with formidable challenges, is positioned for significant breakthroughs because of technological and computational advancements. Continued research and the development of new technologies are imperative to advance this field and to achieve the full therapeutic potential of RNA.

3 METHODOLOGY

This thesis explores cheminformatics techniques to analyze RNA-binding ligands, aiming to enhance the predictive capabilities in drug discovery. The primary focus is on identifying the chemical properties that influence RNA binding and developing binary classification models to distinguish between RNA-binding and non-binding ligands. This research leverages extensive datasets and utilizes advanced machine learning algorithms, including Random Forest, XGBoost, and Graph Neural Networks, optimized for high accuracy in predictive performance. Through this approach, the work contributes to the targeted development of therapeutic agents, particularly in conditions where RNA plays a crucial role.

3.1 Analysis Workflow

This subsection provides a graphical representation of the workflow used in the cheminformatics analysis of RNA binding ligands. The workflow illustrates the sequential steps from data acquisition to the final analysis, highlighting the comprehensive approach taken in this research.

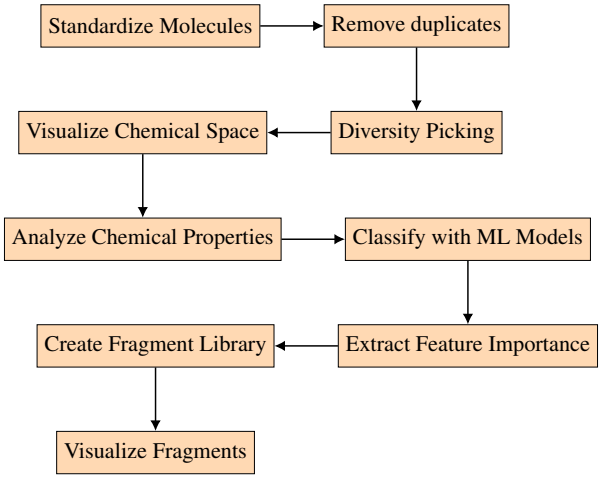


Figure 1: This figure displays a simplified diagram of the pipeline, illustrating the fundamental stages of the process.

3.2 Dataset Preparation and Standardization

The rationale behind using two distinct datasets lies in the fundamental differences between RNA binders and non-binders, and the diverse interactions they encompass. The first dataset consists of chemical libraries geared towards drug discovery, including compounds with potential to bind both proteins and RNA. This dataset provides broad coverage of chemical diversity, enabling the identification of novel binders based on structural motifs and properties. The second dataset, however, is explicitly curated with experimentally confirmed RNA binders and

non-binders, ensuring that models can accurately distinguish these two molecular classes based on empirical binding behavior. Differentiating between these groups is crucial for developing predictive models effective in identifying RNA-targeting compounds while avoiding misclassification.

Ensuring high-quality and standardized datasets is essential for meaningful cheminformatics analysis. To achieve this, a comprehensive standardization process was applied to both datasets, removing redundancies and ensuring consistency for accurate computational analysis. All chemical structures were converted to canonical SMILES using the RDKit library and standardized through the ChEMBL Structure Pipeline [26], ensuring a uniform format and facilitating the identification and removal of duplicate entries. Venn diagrams were instrumental in visualizing overlaps between datasets, assisting in precise curation.

A diverse range of machine learning models, including ensemble methods and graph neural networks, was then applied to the standardized datasets to analyze and predict RNA-binding activities. By leveraging the comprehensive chemical and functional diversity in both datasets, the models were developed to distinguish between RNA and protein binders, enriching predictive capabilities and addressing the complex nature of molecular functionalities.

First Dataset: $Set1\_Large$ The dataset $Set1\_Large$ comprises a curated collection of chemical libraries, each distinctly contributing to drug discovery and molecular interaction studies:

1. **Enamine Hit Locator Library (HLL):** Incorporates 460,160 compounds, primarily a protein-binding library with diverse chemical space. Stringent MedChem filters such as PAINS and smart clustering ensure novelty and diversity [27].

2. **ChemDiv miRNA-targeted Library:** Offers 20,000 small molecules tailored for miRNA modulation, enriched through meticulous substructure filtering [28].

3. **Enamine RNA Library:** Contains 15,520 compounds curated for RNA binding, targeting various RNA structures for therapeutic interventions [29].

4. **Life Chemicals RNA Focused and Targeted Libraries:** Provides 5,544 compounds using cheminformatics methods aimed at human RNA for novel drug discovery [30].

5. **ROBIN Database:** Contributes 2,003 RNA-binding molecules, propelling RNA-targeting algorithm development [31].

This combined dataset comprehensively represents diverse chemical libraries used in drug discovery, encompassing various molecular interactions. Using chemical space diversity and targeted substructure filtering to ensure robust and meaningful cheminformatics analysis.

Second Dataset: $Set2\_Small$ This dataset, known as $Set2\_Small$, includes carefully curated collections of compounds from four distinct sources, each enhancing training and validation in predictive models:

1. **RNA Binders** from the Repository Of Binders to Nucleic acids (ROBIN): 2,003 compounds specifically designed or identified to interact with RNA [31].

2. **RNA Non-Binders** from the ROBIN repository: 22,489 compounds serving as vital negative examples to refine model specificity [31].

3. **Protein Binders** from the Probes & Drugs database: 2,952 compounds with known interactions with protein targets. This library includes approved protein-binding drugs and was filtered to ensure affinity for protein targets, minimizing biases in machine learning models. Molecular weights range from 100 to 739, covering diverse drug types [32].

4. **Non-Binders** from ZINC15's "Dark Matter" section: 25,000 compounds with low predicted binding affinity, providing a broad control group [33].

This dataset represents a comprehensive collection encompassing both RNA and protein binders and non-binders, providing essential training and validation data for predictive algorithms.

The identification and removal of duplicates are crucial for maintaining the integrity of cheminformatics datasets, ensuring that subsequent analyses, such as machine learning modeling, are based on unique and representative molecular entities. This step helps avoid biases and errors in predictive modeling and statistical analyses, promoting more accurate and reliable results.

Given the extensive size of the Enamine Hit Locator Library, which is considered in this study to be a protein-binding library, a representative subset of 50,000 compounds was selected using the MaxMin diversity picking algorithm. This approach was designed to cover a broad chemical space while keeping the dataset size manageable for subsequent computational and analytical tasks, which is essential to ensure robustness of the machine learning models and validity of the analysis.

3.3 Visualization of Chemical Space

To explore the high-dimensional data inherent in cheminformatics, this study used a combination of unsupervised machine learning techniques to visualize the chemical space represented by our datasets. This involved using logistic PCA, t-SNE, and UMAP to project high-dimensional fingerprints onto a two-dimensional plane for easier interpretation and analysis.

Dimensionality Reduction Techniques

- **Logistic PCA:** Initially, logistic PCA was applied to reduce the dimensionality of our data while maintaining the binary nature of ECFP6 fingerprints. This method provided a linear projection that was particularly useful for identifying global structures within the datasets.

- **t-SNE and UMAP:** For more nuanced visualization that can capture non-linear relationships, t-SNE and UMAP were subsequently used. These techniques are well-suited for complex molecular datasets as they tend to preserve the local structure, making clusters more interpretable. t-SNE was particularly effective in revealing different groups within the data, suggesting functional similarities among compounds. UMAP, on the other hand, provided a more global view of the data's structure, balancing local and global aspects of the data which is crucial for understanding the overarching relationships among the datasets.

Visualization and Analysis The dimensionality reduction results were visualized using scatter plots, with points colored according to their originating data set to trace their properties and sources. This visualization not only facilitated the identification of inherent clusters and outliers but also enabled us to evaluate the effectiveness of our diversity picking strategy by observing the dispersion of points from the same source. The outputs of these visualizations were saved as high-resolution images, ensuring that detailed observations could be made from the visual data. These plots serve as a crucial exploratory tool to hypothesize about potential biological activities and properties of the compounds based on their clustering behavior.

3.4 Chemical Property Analysis

A thorough analysis of chemical properties was conducted to understand the physicochemical profiles of compounds within our datasets. This exploration is essential for identifying property trends that could influence biological activity and pharmacokinetic properties.

Data Preparation and Property Calculation Chemical properties such as molecular weight (MW), number of heavy atoms (#HeavyAtoms), counts of specific atom types (NumO, NumN, NumC, NumCl, NumF, NumS), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), number of rings (Rings), calculated octanol-water partition coefficient (ClogP), number of rotatable bonds (#RotBonds), topological polar surface area (TPSA), quantitative estimate of drug-likeness (QED), and calculated distribution coefficient (ClogD) were computed using RDKit [34] and a machine learning model from Pat Walters' repository [35, 36]. These descriptors are crucial for evaluating drug-likeness and potential biological efficacy.

Distribution Analysis The distributions of these properties were visualized using histograms to assess the diversity and typical property values within our chemical libraries. This analysis was facilitated by seaborn (version 0.12.2) and matplotlib (version 3.7.1) libraries in Python, providing insights into the physicochemical space covered by our datasets. Histograms were generated with bins optimized for each property to ensure detailed and informative visualizations.

Custom bin ranges were employed for properties with wide ranges or specific analytical interests, such as molecular weight and TPSA, to focus on regions of particular relevance to drug discovery. For example, molecular weight was examined within the range of 150 to 550 Da, a common range for drug-like molecules, to ensure the chemical libraries' relevance to therapeutic applications.

This detailed exploration of chemical properties helps to ensure that the compounds selected for further studies are not only chemically diverse but also possess desirable properties that increase their likelihood of being successful in therapeutic applications.

3.4.1 Scaffold Analysis

Following the chemical property analysis discussed in Section 3.4, the scaffold analysis represents a critical step in understanding the structural framework that underpins the biological activity of compounds in our datasets. This analysis seeks to identify common scaffolds—the core structural components—across different chemical datasets and assess their frequency and distribution.

Purpose of the Scaffold Analysis
The primary goal of this analysis is to uncover prevalent molecular frameworks within our chemical libraries. By identifying common scaffolds, we can infer potential structural features that may be responsible for particular biological activities or properties. This information is crucial for rational drug design and can guide the synthesis of new compounds with desired biological properties.

Methodology Scaffold analysis was conducted through the following detailed steps:

1. **Data**: Two separate datasets, each consisting of various chemical libraries, were analyzed independently to ensure a comprehensive evaluation.
2. **Scaffold Extraction**:
    - **Murcko Scaffold** (BM): These include the core rings and linkers within a molecule. Murcko scaffolds are critical for understanding the fundamental framework of bioactive compounds[37].
    - **Cyclic Skeletons (CSK)**: This form simplifies Murcko scaffolds by abstracting certain structural elements, retaining only the core topology of the molecule.

3. **Frequency Calculation**: I computed the frequency of each unique scaffold to measure its prevalence across various sources.

4. **Visualization**: The most common scaffolds were visualized to provide a graphical representation of their distribution within the datasets[38].

Expected Outcomes By analyzing the scaffold distributions, I expect to:

- Identify scaffold motifs that are over-represented in our datasets, which could indicate a higher relevance to biological activity.
- Understand the diversity of our chemical libraries, which is critical for enhancing the robustness of future high-throughput screening campaigns.
- Generate insights that can drive the design of new molecules based on the most promising scaffolds, potentially leading to the discovery of new therapeutic agents.

Implications for Further Research
The scaffold analysis establishes a basis for subsequent research that might explore correlations with biological activities or other chemical properties. This foundational understanding of prevalent scaffolds within our datasets can inform targeted research directions, such as the selection of scaffolds for synthesis and modification. Moreover, it provides a structural perspective that can be crucial for theoretical and computational models of drug interactions, thus contributing indirectly to the drug discovery process.

3.5 Machine Learning Models

Developing binary classification models is a critical component of this study, in order to distinguish between RNA-binding and non-RNA-binding molecules. Given the inherent challenges associated with molecular data, I carefully balanced the dataset to prevent any bias towards one class, ensuring fair training and testing conditions.

3.5.1 Data Balancing and Preprocessing

I balanced the datasets through under-sampling the majority class to ensure equal representation of both classes in a binary classification framework. Specifically, the $Set1\_Large$ was prepared with 77,420 molecules, divided equally between two classes. For $Set2\_Small$, the dataset was adjusted to a total of 3,922 molecules, maintaining a balanced representation between RNA binders and protein binders.

Features for machine learning models were derived from canonical SMILES, which were converted into Extended Connectivity Fingerprints (ECFP6). Each molecule in the dataset was represented by a fixed-length binary vector of 2048 bits. This method of feature extraction ensures that all models are trained on a consistent and comprehensive set of features, which is critical for maintaining the validity and reproducibility of the predictive models developed.

The models developed are:

- **Model 1: RNA-Binders vs. Non-RNA-Binders** aims to discern molecules interacting specifically with RNA from those that do not.
- **Model 2: RNA-Binders vs. Protein-Binders** focuses on differentiating RNA-binding molecules from protein-binding counterparts.
- **Model 3: Binders vs. Non-Binders** evaluates RNA binders against a comprehensive group of all non-binders to identify unique molecular signatures of RNA-binding activity.

3.5.2 Ensemble Model Selection and Hyperparameter Tuning

I selected three robust machine learning algorithms for this study:

XGBoost (Extreme Gradient Boosting) — A highly efficient and flexible machine learning system that excels in various data science challenges. XGBoost is particularly notable for its scalability and performance, which can process billions of examples with fewer resources than other systems. It uses a novel sparsity-aware algorithm that optimizes handling of sparse data, an important feature given the frequent occurrence of sparse formats in real-world datasets. Additionally, XGBoost uses a weighted quantile sketch to efficiently handle large datasets in approximate tree learning. These capabilities are crucial for effective learning where computational resources are limited or when working with very large datasets. The system's innovative design regarding cache access patterns, data compression, and data sharding makes it a state-of-the-art tool for building scalable tree boosting models. It has been recognized for its performance across various machine learning challenges, significantly outperforming other methods in terms of speed and scalability [11].

Random Forest Classifier — A robust ensemble machine learning technique that operates by constructing a multitude of decision trees during training time and outputting the mode of the classes (for classification) or mean prediction (for regression) of the individual trees. Random Forests effectively correct the habit of decision trees overfitting to their training set. An additional advantage of using Random Forests is their ability to rank the importance of features through a built-in mechanism. This feature importance is derived from how much each feature contributes to decreasing weighted impurity in a decision tree, which is valuable in understand-

ing which features of data contribute the most significantly to the prediction outcome[12]. This capability is particularly useful in contexts where identifying key features is crucial, such as the analysis of specific fragments for RNA affinity .

LightGBM (Light Gradient Boosting Machine) — A state-of-the-art machine learning framework that advances the field of gradient boosting decision trees (GBDT) by focusing on efficiency and scalability, particularly in scenarios with high-dimensional features and large data sizes. Unlike traditional GBDT implementations that require scanning all data instances to estimate information gain for each feature, LightGBM introduces two innovative techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS enhances the efficiency by focusing on instances with larger gradients, which are more significant for the calculation of information gain, thus reducing the number of data instances needed for accurate estimations. EFB effectively reduces the feature space by bundling mutually exclusive features (those that rarely take nonzero values simultaneously), thereby addressing the sparsity commonly found in high-dimensional data. These methods collectively enable LightGBM to achieve substantial speed improvements, up to 20 times faster than conventional GBDT, without compromising accuracy, making it an ideal choice for handling modern big data applications [13].

The hyperparameter tuning for each model was conducted using Optuna[39], an optimization framework that employs Bayesian optimization techniques to systematically explore the hyperparameter space. Optuna was configured to maximize the accuracy, which is particularly effective for binary classification tasks with balanced classes. This optimization process involved trials where different combinations of parameters were tested to find the configuration that yielded the best validation performance.

3.5.3 Evaluation Metrics

The models' performances were evaluated based on several metrics, including:

- **Accuracy** — The proportion of correctly predicted observations to the total observations. It is defined as:
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
- **Precision and Recall** — Precision is the ratio of correctly predicted positive observations to the total predicted positives. Recall (or Sensitivity) measures the ratio of correctly predicted positive observations to all observations in the actual class Yes. They are defined as:
$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1 Score** — The weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. It is especially useful when the class distribution is uneven. The formula for the F1 Score is:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **ROC-AUC** — The Receiver Operating Characteristic curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The Area Under the Curve (AUC) represents a measure of separability between the classes.

$$\text{FPR} = \frac{FP}{FP + TN}$$

The receiver operating characteristic (ROC) curve is generated by plotting the true positive rate (TPR), which is the same as recall, against the false positive rate (FPR) for various threshold values, providing a tool to assess the classifier's performance across different levels of sensitivity and specificity. This method is well discussed in the context of data classification evaluation [40].

Robustness and Reproducibility

To ensure the robustness and reproducibility of the findings, each machine learning model was computed ten times with different initialization seeds. This process allowed for a comprehensive assessment of the models' stability and variance under varying conditions. After these repeated runs, a detailed statistical analysis was performed on the collected metrics such as accuracy, precision, recall, F1 score, and ROC-AUC. This analysis was crucial in confirming the models' consistency and reliability in predicting RNA-binding and non-binding molecules, further strengthening the confidence in the experimental results. The insights drawn from these analyses also guided the final adjustments in model configurations, ensuring optimal performance.

3.5.4 Feature Importance Analysis

The feature importance analysis was conducted using the model that demonstrated the highest accuracy on multiple test runs. This approach ensured that the insights gained were based on the most reliable available predictive performance. The steps of the analysis included the following.

- Identifying and sorting each feature's importance as derived from the decision trees of the highest-performing model. This sorting highlighted which parts of the molecules, represented as bits in a 2048-bit vector from the Extended Connectivity Fingerprints (ECFP6) of canonical SMILES, were crucial for RNA-binding classification.
- Reverse engineering the top-ranked bits to molecular fragments and visualizing these on the molecules. This step pinpointed the specific substructures within the molecules that were most influential in the classification decisions.
- Saving the canonical SMILES representations of these key substructures for further analysis to explore their potential roles in RNA-binding interactions.
- Highlighting the most important features, or fragments, from the test set that were predicted with high probability to belong to the RNA-binding class, enhancing our understanding of molecular interactions with RNA.

This focused analysis on the highest-accuracy model provided a robust foundation for understanding the molecular dynamics at play in RNA-binding, contributing significantly to the development of potential therapeutic interventions.

3.6 Graph Neural Networks

Graph Neural Networks (GNNs), particularly those employing graph convolutional techniques, represent a class of deep learning models designed to capture complex non-linear relationships within data structured as graphs. These models are adept at leveraging both the structural and relational information of molecules, making them particularly suitable for tasks such as predicting RNA-binding propensities.

3.6.1 Preparing Molecular Graphs for GNN Training

Molecular graphs are fundamental for training Graph Neural Networks (GNNs) in cheminformatics applications. In this study, molecular graphs were constructed using the Deep Graph Library (DGL) in conjunction with RDKit, supported by DGL-LifeSci utilities. This approach facilitated the conversion of canonical SMILES strings into bi-directed graphs that accurately reflect the molecular structure.

Graphs were featurized using `CanonicalAtomFeaturizer` and `CanonicalBondFeaturizer` from DGL-LifeSci, which encode essential atomic and bonding properties into nodes and edges, respectively. The `mol_to_bigraph` function was employed to convert these featurized molecules into graphs suitable for graph convolutional processes.

Each graph was labeled according to its RNA-binding affinity to enable a supervised learning approach. The graphs and their labels were then serialized and stored, ensuring efficient handling during the training of GNN models. The integration of RDKit and DGL-LifeSci illustrates the use of advanced cheminformatics techniques to facilitate the application of graph-based deep learning models in molecular science.

3.6.2 Graph Neural Network Architectures

Deep learning models, particularly Graph Neural Networks (GNNs), are capable of capturing non-linear and complex patterns in data structured as graphs. In this study, three distinct GNN architectures were utilized to predict RNA-binding propensities, leveraging the inherent graph structure of molecular data.

GATv2 The Graph Attention Network version 2 (GATv2) improves upon the original GAT model by introducing a dynamic attention mechanism that allows each node to attend differently to its neighbors, depending on the query node itself. This method addresses the limitations of static attention, where the ranking of attention coefficients is unconditioned on the query node, thus enhancing the model's expressiveness and ability to capture more complex patterns. The GATv2 architecture is defined by:

$$h_i^{(l+1)} = \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} W_{right}^{(l)} h_j^{(l)} \tag{1}$$

where α_{ij} is the attention score between nodes, calculated as:

$$\alpha_{ij}^{(l)} = \text{softmax}_i(e_{ij}^{(l)}) \tag{2}$$

$$e_{ij}^{(l)} = \vec{a}^{T(l)} \text{LeakyReLU}\left(W_{left}^{(l)} h_i + W_{right}^{(l)} h_j\right) \tag{3}$$

This attention mechanism allows GATv2 to potentially outperform the original GAT across various benchmarks by enabling a more nuanced interaction between node features. This advancement makes GATv2 particularly effective in domains where understanding the inter-node relationship dynamics is crucial, such as in molecular interaction predictions.[41]

SAGEConv SAGEConv, or Sample and Aggregate Convolution, focuses on sampling neighbor nodes and aggregating their features to learn a representation that captures local graph topology effectively. This model is particularly useful for large graphs where full neighborhood aggregation is computationally expensive. It leverages a methodology that can selectively emphasize the most significant features from neighboring nodes without processing the entire neighborhood at once. The GraphSAGE framework extends this concept by using trainable ag-

gregation functions to generate node embeddings inductively, enabling the model to generalize to unseen nodes and graphs efficiently. This capability is critical for applications in dynamic environments, such as social networks or biological networks, where the graph is continually evolving.

The GraphSAGE method involves:

$$h_{\mathcal{N}(i)}^{(l+1)} = \text{aggregate}\left(\{h_j^l, \forall j \in \mathcal{N}(i)\}\right) \tag{4}$$

$$h_i^{(l+1)} = \sigma\left(W \cdot \text{concat}(h_i^l, h_{\mathcal{N}(i)}^{l+1})\right) \tag{5}$$

$$h_i^{(l+1)} = \text{norm}(h_i^{(l+1)}) \tag{6}$$

Where $h_{\mathcal{N}(i)}^{(l+1)}$ is the aggregated feature from the neighbors of node i, and $h_i^{(l+1)}$ is the updated feature of node i after combining its own features with those of its neighbors, followed by a non-linearity σ and normalization. This process allows for the generation of node embeddings that incorporate both the local neighborhood structure and node features, facilitating effective inductive learning.

The framework not only captures the local role of a node within the graph but also its global position, aligning newly observed subgraphs to the node embeddings optimized during training. By operationalizing this approach, SAGEConv effectively supports scenarios where the graph's structure and node attributes are predictive of node roles, significantly enhancing the performance of tasks like node classification, even in graphs that evolve over time.[42]

GatedGraphConv The GatedGraphConv architecture utilizes gated recurrent units (GRUs) to manage graph-structured data, integrating dynamic and temporal patterns effectively. This makes it particularly suitable for applications involving dynamic or evolving graphs such as social networks, recommendation systems, and biochemical structures. The model's ability to capture sequential and relational patterns allows it to adapt over time to changes in the graph structure, maintaining robust performance even in non-static environments.

The process in the GatedGraphConv layer can be described as follows:

$$\begin{aligned} h_i^0 &= [x_i \| \mathbf{0}] \\ a_i^t &= \sum_{j \in \mathcal{N}(i)} W_{e_{ij}} h_j^t \\ h_i^{t+1} &= \text{GRU}(a_i^t, h_i^t) \end{aligned} \tag{7}$$

Where h_i^0 is the initial state of node i, incorporating its features x_i padded with zeros. a_i^t represents the aggregated inputs from the neighbors of node i at time t, weighted by $W_{e_{ij}}$, the edge-specific weight matrix. h_i^{t+1} is the updated state of node i at time $t+1$, computed by a GRU cell, which processes the aggregated input and the previous state.

This model was developed as part of the broader exploration into Gated Graph Sequence Neural Networks, presented at ICLR 2016, which highlights its utility in various domains including chemistry and computer program verification.[43]

3.6.3 Data Split and Model Training Strategy

Dataset Splitting The dataset was divided into training, validation, and testing sets to enable robust training and evaluation. The data was initially split into an 80% training set and a 20% test set, with the training set further divided to create a 20% validation set for performance monitoring and overfitting prevention.

Hyperparameter Tuning and Training Phases Hyperparameter tuning utilized the Optuna framework, focusing on a subset (20%) of the training and validation sets for efficient optimization, just with $Set1\_Large$ dataset. This involved defining a search space for key hyperparameters such as learning rates, dropout rates, and the number of convolution layers. The process included:

- **Initial Training:** Conducting trials on the smaller subset to determine optimal hyperparameters.
- **Refinement Training:** Training on the full training and validation datasets using the best hyperparameters to finalize the optimal number of epochs via early stopping.
- **Final Training:** Employing the entire training dataset (combined training and validation sets) to train the models to the epochs determined previously. This phase maximized learning from all available training data.

Training Strategy and General Setup The training process included:

- Setting consistent random seeds at the start of experiments to ensure reproducibility.
- Employing DGL for efficient graph data manipulation and batch processing.
- Utilizing a collate function for batching, crucial for leveraging GPU acceleration.
- Implementing early stopping based on validation loss and accuracy to prevent overfitting and optimize computational resources.
- Using gradient scaling to support mixed precision training, enhancing training speed and memory efficiency.

Model Evaluation Models were rigorously evaluated using the test set. Performance metrics such as accuracy, ROC-AUC, precision, recall, and F1-score were calculated to assess the models' ability to predict RNA-binding propensities accurately. Outputs, including model weights and evaluation metrics, were serialized and stored for detailed analysis.

3.6.4 Implementation Details

Consistent random states were used for all data splits to ensure reproducibility across training runs. The architecture was specifically adapted to meet the requirements of processing graph-based molecular data. The Adam optimizer facilitated the training, with early stopping mechanisms strategically employed to curtail training based on real-time validation performance, ensuring models did not overfit. Final model weights were saved after the completion of training phases, and performance metrics were thoroughly analyzed to confirm the efficacy of the training approach.

3.7 Computational Environments for Model Training

This section details the computational setups used for training ensemble models and graph neural networks. It emphasizes the hardware configurations and specific software versions tailored to different computational needs.

3.7.1 Ensemble Model Computing Environment

- **CPU:** Intel Xeon Processor (Skylake, IBRS), 18 cores, optimized for multi-threading and complex computations.
- **System:** Ubuntu 22.04.4 LTS, providing a stable and modern Linux environment for scientific computing.
- **Python Version:** 3.10.10, within a Conda environment to ensure reproducibility and consistency across computational experiments.
- **Key Libraries:**
  - **XGBoost** (version 2.0.3), **LightGBM** (version 4.3.0) for gradient boosting frameworks.
  - **SciPy** (version 1.10.1), **NumPy** (version 1.23.5) for numerical operations.
  - **Pandas** (version 1.5.3) for data manipulation and analysis.
  - **Scikit-learn** (version 1.4.1.post1) for implementing various machine learning models.
  - **Matplotlib** (version 3.7.1), **Seaborn** (version 0.12.2) for data visualization.

3.7.2 Graph Neural Network Training Environment

- **CPU:** AMD EPYC Processor with 32 cores and advanced processing capabilities.
- **GPU:** NVIDIA Tesla T4 with 15360 MiB of memory, providing substantial computational power for deep learning.

- **System:** Debian GNU/Linux 10 (buster), chosen for its stability and performance in high-performance computing settings.
- **Python Version:** 3.8.18, within a Conda environment.
- **Key Software:**
  - **PyTorch** (version 2.1.2) and **DGL** (version 2.0.0) for constructing and training graph neural networks.
  - **DGL-LifeSci** (version 0.3.2) for cheminformatics-specific operations, crucial for preprocessing and managing molecular data.
  - **CUDA** (version 12.4) and **NVIDIA Driver** (version 550.54.14) for optimizing GPU utilization during the training of deep learning models.
  - Additional libraries: **RDKit** (version 2023.09.5), **NumPy** (version 1.24.3), **SciPy** (version 1.10.1), **Pandas** (version 2.0.3), and **Matplotlib** (version 3.7.2).

4 RESULTS AND DISCUSSION

4.1 Dataset Preparation and Standardization

The cheminformatics analysis conducted in this study required high-quality and standardized datasets to obtain meaningful insights. This section discusses the results of the preparation of the data set and the impacts of the standardization process on both sets of data used in the computational experiments.

4.1.1 Removing Duplicates

This section presents the results of duplicate analysis performed on two cheminformatics datasets: $Set1\_Large$ and $Set2\_Small$. Venn diagrams were utilized to visualize the overlap between different chemical libraries and subsets, facilitating the identification and removal of duplicate entries based on canonical SMILES representations.

Dataset $Set1\_Large$

The $Set1\_Large$ dataset comprised multiple libraries targeting RNA and protein interactions. Figure 2 illustrates unique and shared entries among these libraries.

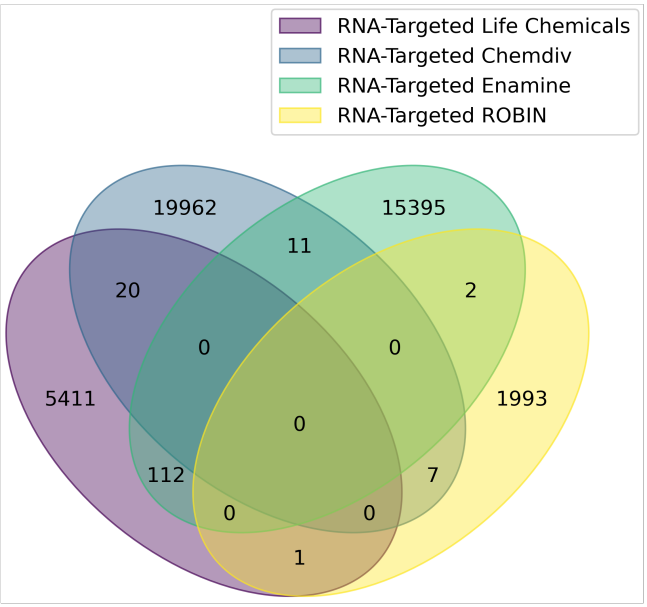


Figure 2: Overlap of RNA-targeted libraries in $Set1\_Large$.

Figures 3 and 4 show further details:

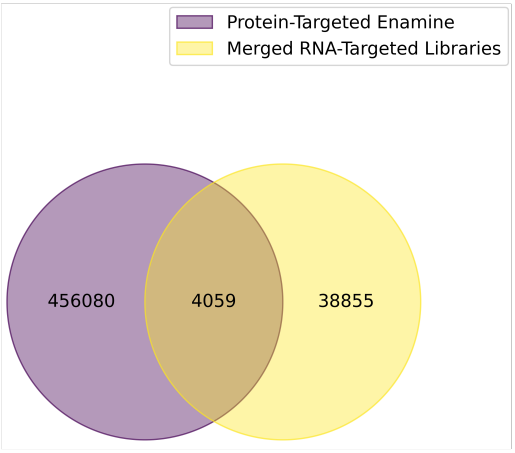


Figure 3: Overlap of protein-targeted and merged RNA-targeted libraries in $Set1\_Large$.

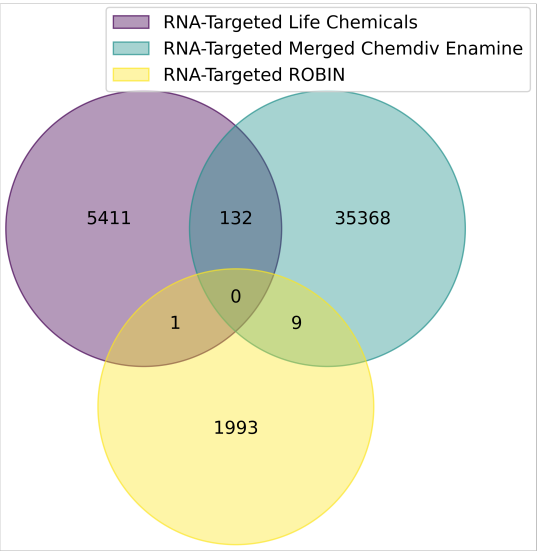


Figure 4: Detailed analysis of RNA-targeted libraries in $Set1\_Large$.

The analysis revealed a significant number of unique entries in conjunction with a smaller subset of common entries across the datasets, indicating a varied but partially overlapped chemical space.

Dataset $Set2\_Small$

For $Set2\_Small$, the focus was on RNA binders and non-binders. Figures 5 and 6 display the distribution of entries within and across these categories.

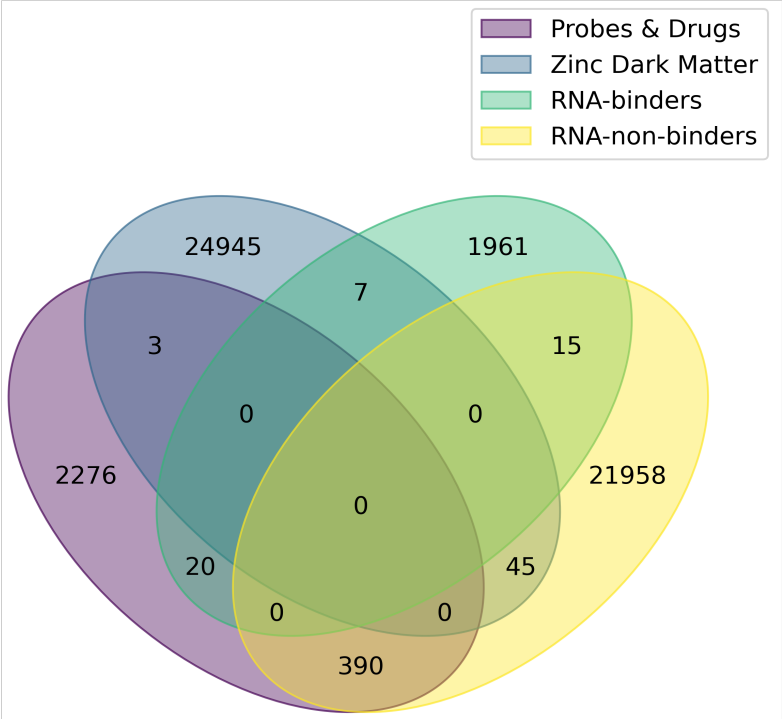


Figure 5: Interactions between various libraries within $Set2\_Small$.

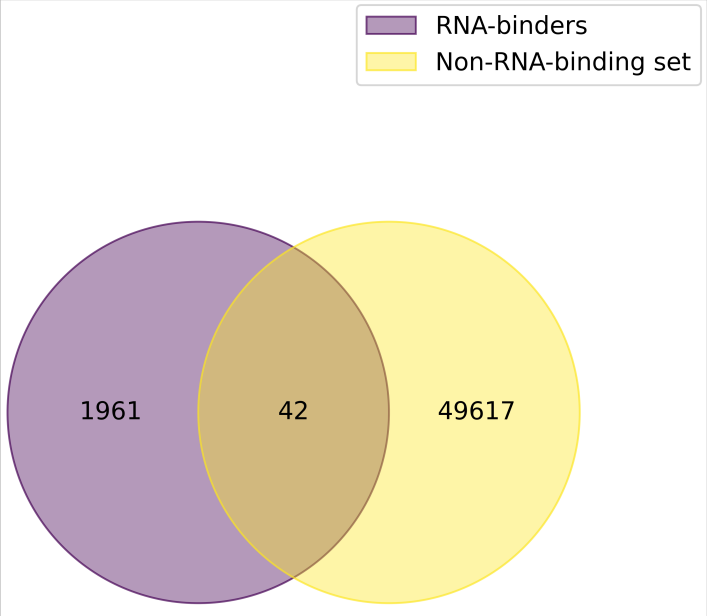


Figure 6: Distribution of RNA binders versus non-binders within $Set2\_Small$.

The Venn diagrams demonstrated a clear delineation between RNA binders and non-binders, with minimal overlap, highlighting the effectiveness of the initial dataset curation.

4.1.2 Data Integrity and Quality Assurance

Post-standardization, several measures were implemented to ensure the integrity of the data:

- Re-evaluation of each molecule's SMILES string to confirm the absence of errors such as residual salts or incomplete fragment removal.
- Reassessment of the datasets to ensure no duplicate or erroneous entries remained.
- Systematic reconversion to molecular objects (RDKit mol objects) to verify that all SMILES strings were correctly standardized and could be reliably converted back into molecular form.

These preparations ensure that the data used in the study is of the highest quality and consistency, laying a robust foundation for the cheminformatics analyses and machine learning models to follow.

4.2 Visualization of Chemical Space Across Datasets

The extensive exploration of the chemical space through advanced visualization techniques provides invaluable information on the structural diversity and properties of ligands. Using logistic PCA, t-SNE, and UMAP, this study unveils the intricate patterns within two chemically diverse datasets, shedding light on the complex interactions between RNA-binding and protein-binding molecules.

4.2.1 Set1_Large Dataset Analysis

The $Set1\_Large$ dataset comprises a compilation of compounds known for their RNA-binding and protein-binding properties, originating from various chemical libraries. The visualization techniques applied to this dataset revealed a multifaceted chemical space with unique and shared molecular features.

Figure 7: Logistic PCA of ECFP6 Fingerprints for the $Set1\_Large$.

Logistic PCA provided an initial linear transformation, revealing an overlap in the RNA and protein-binding ligands' chemical space, suggestive of shared ECFP6 fingerprints among the compounds.

Figure 8: t-SNE Visualization of ECFP6 Molecular Fingerprints for the $Set1\_Large$.

t-SNE, a non-linear technique, delineated more granular clusters, highlighting potential specificity in molecular interactions.

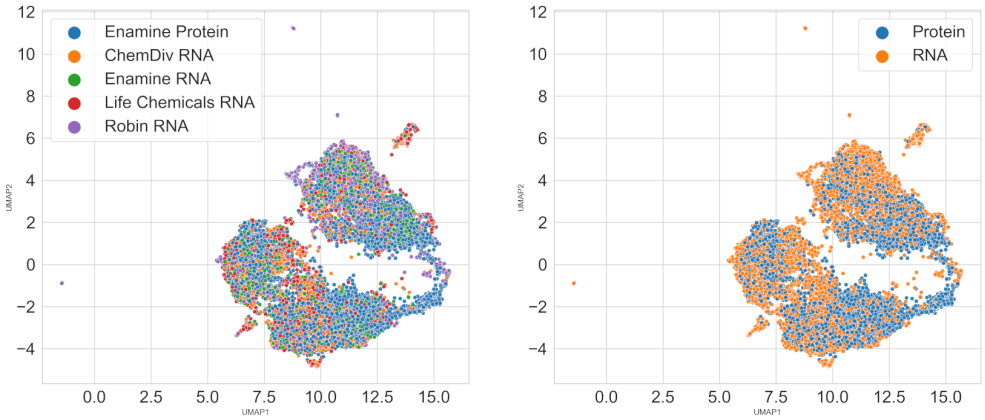


Figure 9: UMAP Visualization of ECFP6 Molecular Fingerprints for the $Set1\_Large$.

UMAP offered the most distinct clustering of the datasets, demonstrating its effectiveness in preserving local and global structures within the data.

4.2.2 Set2_Small Dataset Analysis

The $Set2\_Small$ includes an alternative assemblage of compounds, enhancing the breadth of the comparative analysis with distinct chemical entities.

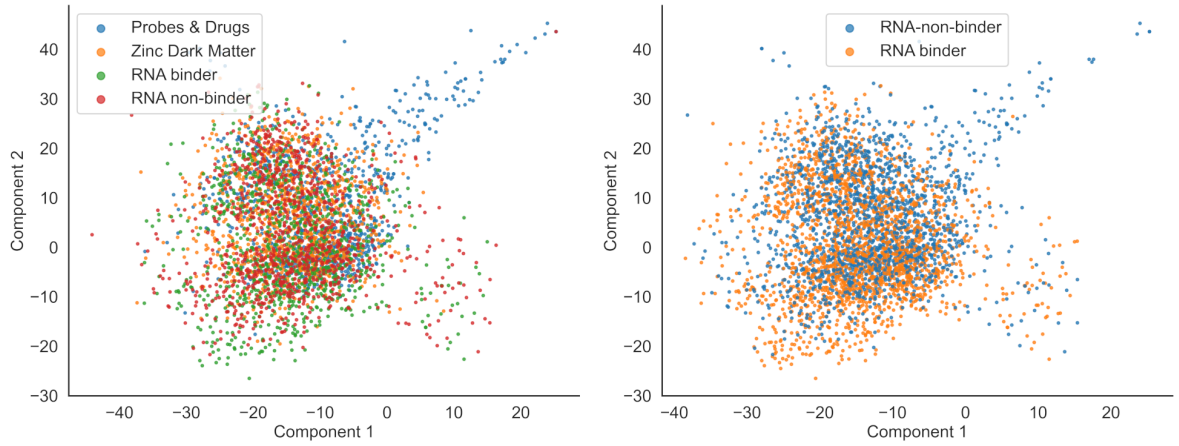


Figure 10: Logistic PCA of ECFP6 Fingerprints for the $Set2\_Small$.

Logistic PCA for the $Set2\_Small$ mirrored the first, with notable overlaps, yet presented unique peripheral clustering.

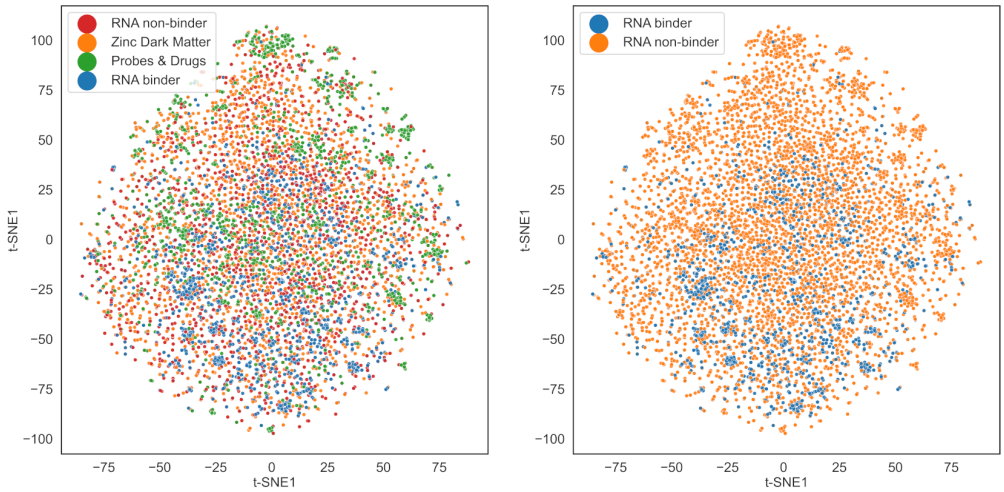


Figure 11: t-SNE Visualization of ECFP6 Molecular Fingerprints for the $Set2\_Small$.

t-SNE revealed an intricate tapestry of intermixed RNA binders and non-binders, stressing the complexity of chemical space segregation.

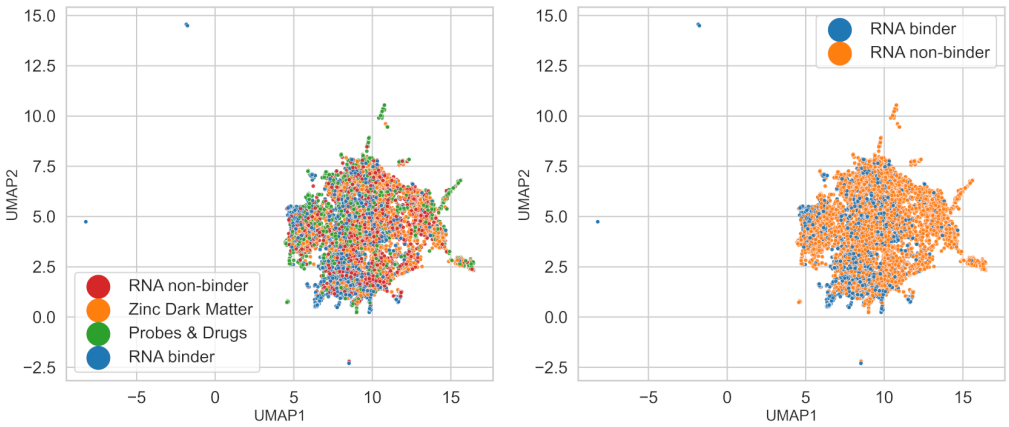


Figure 12: UMAP Visualization of ECFP6 Molecular Fingerprints for the $Set2\_Small$.

UMAP emphasized the potential for identifying nuanced distinctions within the molecular fingerprints and binding affinities.

4.2.3 Comparative Analysis

The comparative interpretation of the dimensionality reduction outputs underscores both the shared and distinct molecular features of RNA-binding versus protein-binding ligands across datasets. This comparative study facilitates an understanding of the convoluted nature of the chemical space and underscores the complexities of classifying binding properties.

Logistic PCA

Comparing the logistic PCA results from both datasets highlighted common structural motifs and underscored the necessity of nuanced analysis tools to distinguish between binder types.

t-SNE

The t-SNE findings illuminated the intricate molecular patterns unique to RNA-binding ligands, suggesting a complex interplay of structural features dictating binding specificity.

UMAP

UMAP findings revealed the inherent diversity and specificity within the datasets, suggesting a stratified chemical space that aligns with biological activity.

4.2.4 Discussion

The cheminformatics exploration of RNA and protein-binding ligands via dimensionality reduction visualizations has yielded significant insights into their chemical space. The observed overlap between these ligands' molecular fingerprints indicates shared structural motifs, yet it underscores the necessity of advanced analysis to decipher subtle distinctions that may influence binding properties. Logistic PCA's broad strokes, t-SNE's detailed clustering, and UMAP's nuanced segregation each contribute to a composite picture of ligand interaction potential, suggesting the need for integrative, multi-faceted approaches in predictive modeling and therapeutic design.

Particularly, t-SNE's capability to reveal distinct clusters suggests the presence of specific binding modalities, meriting further investigation into their biological functions. UMAP's delineation of densely populated regions within the chemical space implies the existence of 'privileged structures', which could be pivotal in identifying new RNA-binding ligands and enhancing the efficacy of RNA-targeted therapies.

The intermingling of binders and non-binders challenges the cheminformatics community to refine screening methods to accurately distinguish true binding activity. This is critical to minimize false positives and negatives in high-throughput screening endeavors.

In summary, the dimensionality reduction techniques employed here reveal the complexities and subtleties of the chemical space, with each method providing a unique perspective. Logistic PCA indicated the shared features among binders, t-SNE offered more detailed cluster formation, and UMAP distinguished the datasets with a higher degree of separation, presenting a promising route for future predictive modeling. These visualizations are instrumental for understanding the underlying structures that could facilitate the discovery of novel therapeutic molecules with specific binding affinities.

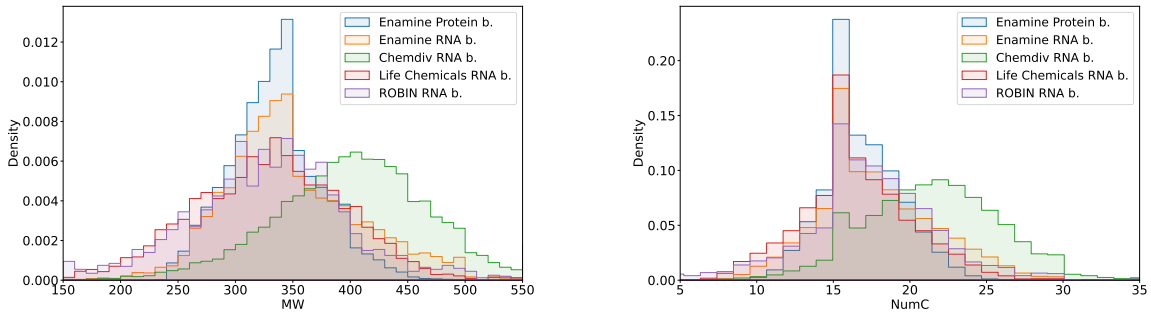
4.3 Chemical Property Analysis

This section discusses the results from the chemical property analysis for both sets of data.

Analysis of Molecular Properties from Set1_Large

Molecular Weight Distribution and Carbon Atom Count

The analysis of molecular weight (MW) across different datasets revealed distinct distributions that reflect the structural diversity within RNA-binding and protein-binding ligands. Meanwhile, the count of carbon atoms varies significantly among the datasets, demonstrating complexity in molecular frameworks necessary for effective RNA binding.



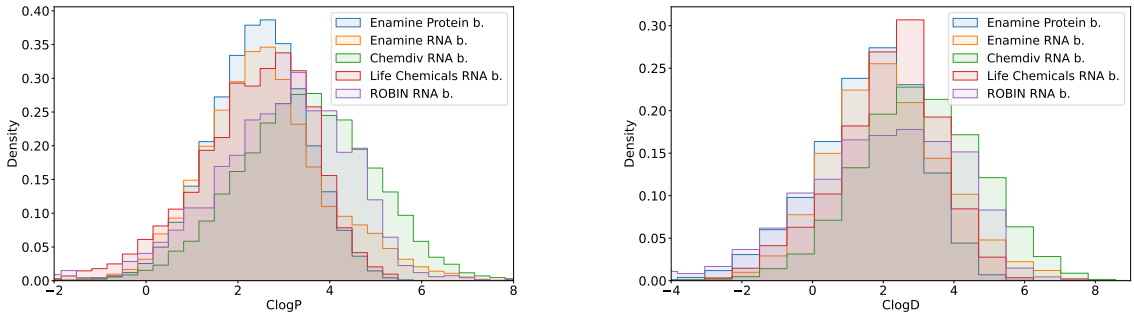
(a) Molecular weight distribution

(b) Number of carbon atoms

Figure 13: Distribution of molecular weight and carbon atoms across different datasets.

Lipophilicity: ClogP and ClogD

Lipophilicity, measured via ClogP and ClogD, was assessed to understand the balance of hydrophobic and hydrophilic characteristics essential for drug-likeness.



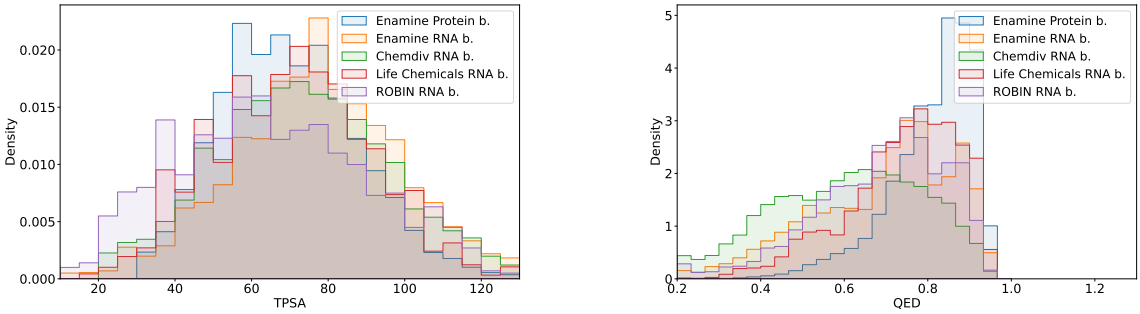
(a) ClogP distribution

(b) ClogD distribution

Figure 14: Histograms of lipophilicity metrics (ClogP and ClogD) across different datasets.

Topological Polar Surface Area (TPSA) and Quantitative Estimate of Drug-likeness (QED)

The TPSA values highlighted a trend where RNA binders often presented higher areas, while QED assessments underscored a promising range of drug-like qualities.



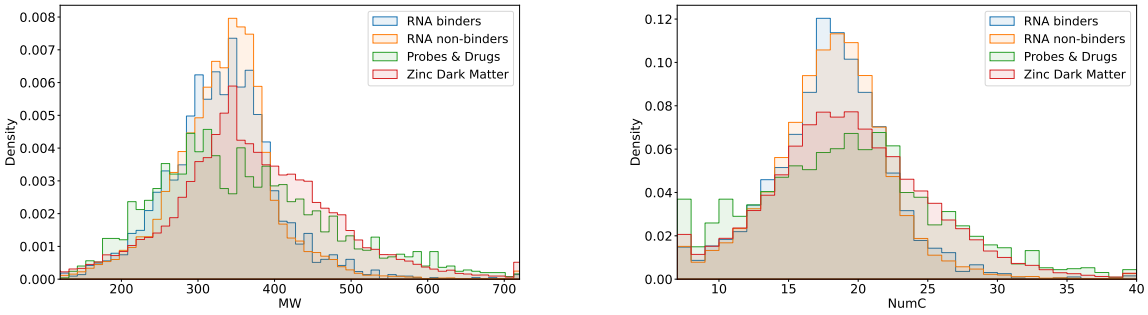
(a) TPSA distribution **(b)** QED distribution

Figure 15: Histograms of TPSA and QED across different datasets, indicating molecular interaction potential and drug-likeness.

Analysis of Molecular Properties from Set2_Small

Molecular Weight Distribution and Carbon Atom Count

The molecular weight and carbon atom counts for the second set of data revealed unique distributions that underscore the specificity of the chemical libraries involved. RNA binders showed peaks that suggest a preference for certain molecular sizes tailored for RNA interaction, while Zinc Dark Matter exhibited broader and higher ranges, indicating diversity in function and structure.

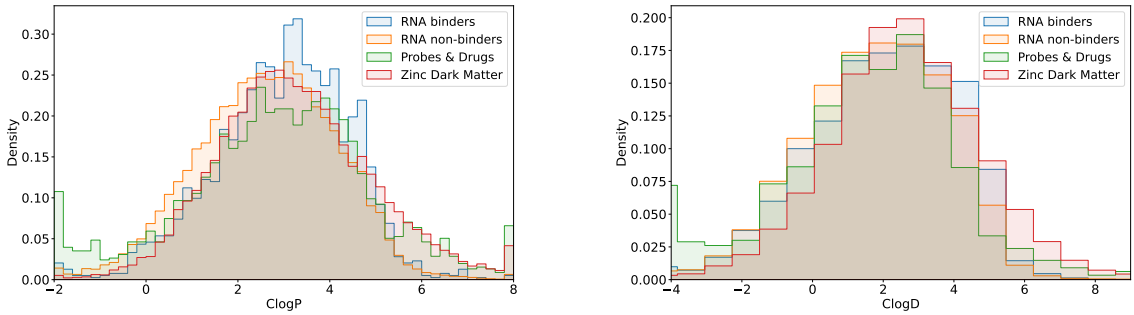


(a) Molecular weight distribution **(b)** Number of carbon atoms

Figure 16: Distribution of molecular weight and carbon atoms across different datasets in Set2_Small.

Lipophilicity: ClogP and ClogD

The analysis of ClogP and ClogD for Set2_Small shows differences in lipophilicity that may impact the biological function of these molecules. Probes & Drugs displayed a peak in ClogD around 3, aligning with their enhanced cell permeability.



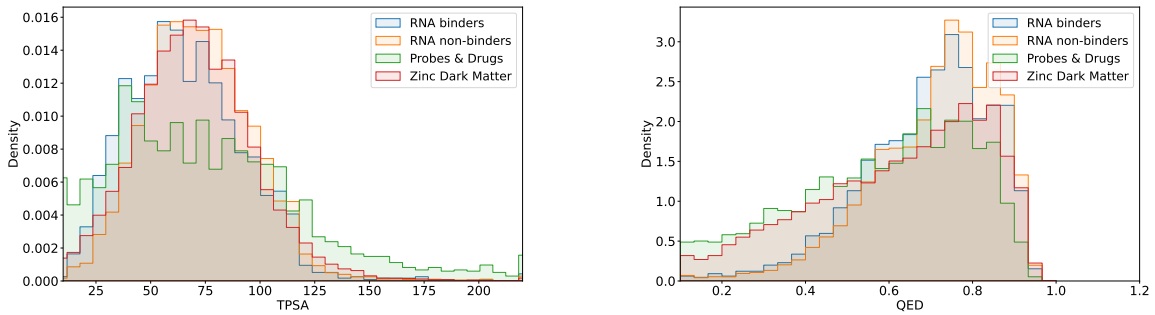
(a) ClogP distribution

(b) ClogD distribution

Figure 17: Histograms of lipophilicity metrics (ClogP and ClogD) across different datasets in Set2_Small.

Topological Polar Surface Area (TPSA) and Quantitative Estimate of Drug-likeness (QED)

TPSA and QED values indicate that RNA binders in Set2_Small possess favorable properties for potential therapeutic application. The TPSA values suggest effective interactions with RNA through hydrogen bonds, while the QED values highlight their drug-likeness.



(a) TPSA distribution

(b) QED distribution

Figure 18: Histograms of TPSA and QED across different datasets in Set2_Small, indicating molecular interaction potential and drug-likeness.

Discussion

In the exploration of chemical properties across Set1_Large and Set2_Small, distinct trends reflect their targeted biological applications.

Set1_Large exhibits a wide range of molecular weights and carbon atom counts, indicative of its broad chemical diversity. This diversity supports its utility in probing a variety of protein and RNA interactions, essential for multifaceted drug discovery efforts. The variability in lipophilicity, as measured by ClogP and ClogD, further underscores its capacity to address different biological pathways, enhancing the dataset's versatility in developing compounds with balanced pharmacokinetic properties.

In contrast, Set2_Small shows more focused chemical distributions, particularly in molecular weights and TPSA values tailored towards specific RNA binding. This concentration aligns with the dataset's role in targeted therapeutic strategies, especially in treatments that require precise molecular interactions with RNA structures. The narrow lipophilicity ranges seen in Set2_Small suggest optimization towards effective cellular uptake and interaction specificity, crucial for therapeutic efficacy.

Overall, these observations highlight Set1_Large's role in broad-spectrum molecular exploration and Set2_Small's specialized application in RNA-targeted therapy development. The distinct chemical profiles of each dataset are instrumental in guiding cheminformatics research towards optimizing ligand design for specific biological functions.

4.4 Scaffold Analysis and Implications for RNA and Protein Binder Design

The comprehensive analysis of molecular scaffolds derived from the Murcko and CSK methodologies revealed distinct structural motifs prevalent in RNA-binding ligands compared to protein-binding ligands and non-binders. Data from both $Set1\_Large$ and $Set2\_Small$ were utilized, encompassing a variety of sources, including commercial libraries and curated datasets.

Murcko Scaffolds

The Murcko scaffold analysis for $Set1\_Large$ indicated a dominance of simpler aromatic systems in protein binders, while RNA binders showed a preference for more complex heteroaromatic and polycyclic systems. The detailed structures and their respective occurrences are presented in Figure 19.

Table 1: Top 5 Murcko Scaffolds in $Set1\_Large$

Murcko Scaffold	**Count**	**Source**	**Ratio**
c1ccccc1	5490	Protein Binder Enamine	1.20%
O=S(=O)(Nc1ccccc1)c1ccccc1	2090	Protein Binder Enamine	0.46%
O=C(Nc1ccccc1)c1ccccc1	1571	Protein Binder Enamine	0.34%
O=C(c1ccccc1)N1CCCCC1	1165	Protein Binder Enamine	0.26%
O=S(=O)(c1ccccc1)N1CCCCC1	997	Protein Binder Enamine	0.22%
O=C(Nc1ccccc1)c1ccccc1	209	Chemdiv RNA	1.05%
c1ccc(CN2CCN(Cc3ccccc3)CC2)cc1	89	Chemdiv RNA	0.45%
O=C(c1ccccc1)N1CCN(Cc2ccccc2)CC1	75	Chemdiv RNA	0.38%
c1ccccc1	69	Chemdiv RNA	0.35%
c1ccc(C2=NN[C@H](c3ccccc3)C2)cc1	39	Chemdiv RNA	0.20%
O=c1[nH]c(=O)c2[nH]c(N3CCN(Cc4ccccc4)CC3)nc2[nH]1	117	Enamine RNA	1.02%
O=c1[nH]c2nc3n(c2c(=O)n1Cc1ccccc1)CCN3c1ccccc1	113	Enamine RNA	0.98%
O=c1[nH]c(=O)c2c(nc(N3CCNCC3)n2Cc2ccccc2)[nH]1	96	Enamine RNA	0.83%
O=c1[nH]c(=O)c2c(nc(N3CCN(Cc4ccccc4)CC3)n2Cc2c...)	80	Enamine RNA	0.70%
O=c1[nH]c(=O)c2c(nc3n(-c4ccccc4)c(-c4ccccc4)cn...)	76	Enamine RNA	0.66%
O=c1[nH]c(=O)c2[nH]cnc2[nH]1	144	Life Chemicals RNA	2.71%
O=c1[nH]c(=O)c2c(nc3[nH]ccn32)[nH]1	143	Life Chemicals RNA	2.69%
O=c1[nH]c(=O)c2c(ncn2Cc2ccccc2)[nH]1	122	Life Chemicals RNA	2.30%
c1ccc2scnc2c1	76	Life Chemicals RNA	1.43%
O=C(Nc1cc2c(s1)CNCC2)c1ccccc1	71	Life Chemicals RNA	1.34%
c1ccccc1	41	ROBIN RNA	2.06%
c1ccc(CNc2ncc(-c3ccccc3)[nH]2)cc1	36	ROBIN RNA	1.81%
c1ccc(CNCc2ccccc2)cc1	26	ROBIN RNA	1.31%
c1ccc(CCNCc2ccccc2)cc1	24	ROBIN RNA	1.20%
c1ccc(-c2nc3ccccc3o2)cc1	22	ROBIN RNA	1.10%

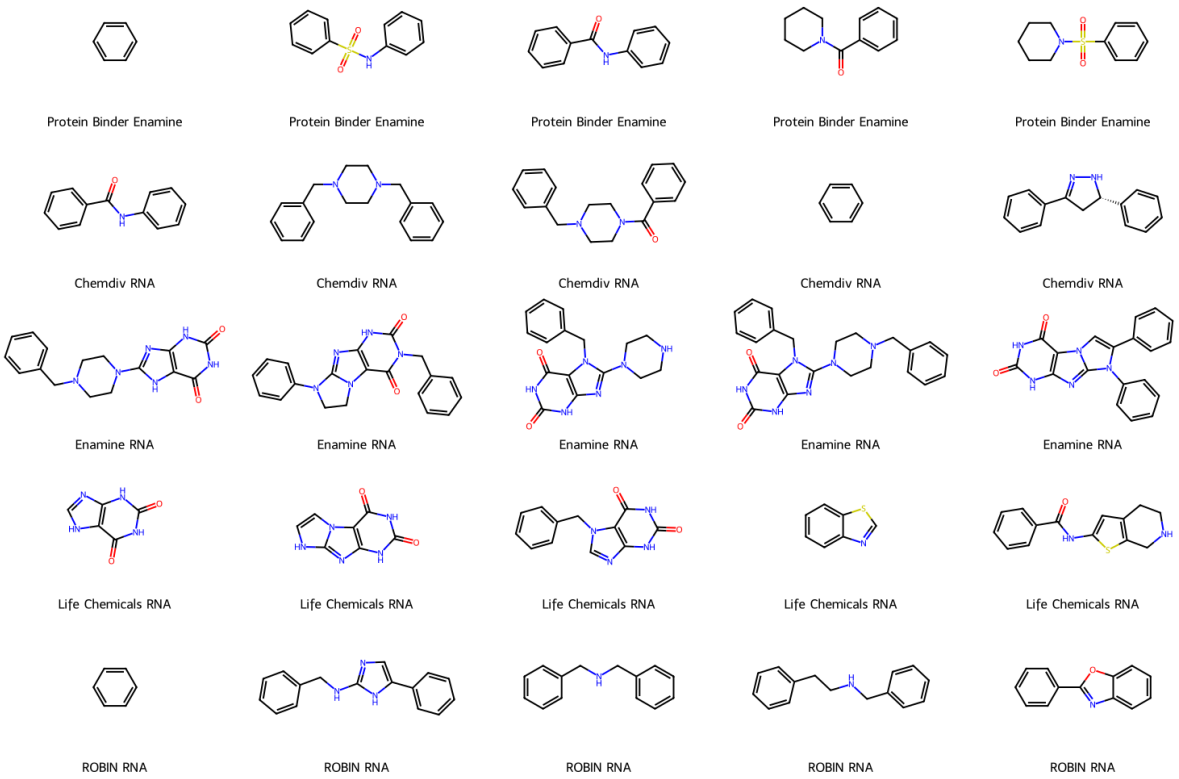


Figure 19: Visual representation of the predominant Murcko scaffolds identified in RNA and protein binders for $Set1\_Large$.

In $Set2\_Small$, similar trends were observed with RNA binders featuring specific functionalized aromatic rings conducive to RNA interaction. The Murcko scaffold distribution for this set is detailed in Figure 20.

Table 2: Top 5 Murcko Scaffolds in $Set2\_Small$

Murcko Scaffold	Count	Source	Ratio
c1ccccc1	40	RNA binder ROBIN	2.04%
c1ccc(CNc2ncc(-c3ccccc3)[nH]2)cc1	36	RNA binder ROBIN	1.84%
c1ccc(CNCc2ccccc2)cc1	27	RNA binder ROBIN	1.38%
c1ccc(CCNCc2ccccc2)cc1	24	RNA binder ROBIN	1.22%
c1ccc(-c2nc3ccccc3o2)cc1	22	RNA binder ROBIN	1.12%
c1ccccc1	319	RNA non-binder ROBIN	1.46%
c1ccc(Nc2ccnc3ccccc23)cc1	117	RNA non-binder ROBIN	0.53%
c1ccc(OCCCn2cnc3ccccc32)cc1	117	RNA non-binder ROBIN	0.53%
c1ccc(-c2noc(-c3cn(-c4ccccc4)nn3)n2)cc1	83	RNA non-binder ROBIN	0.38%
O=C1Nc2ccccc2C12C=COC=C2	83	RNA non-binder ROBIN	0.38%
c1ccccc1	163	Protein Binder P&B	7.16%
O=C1C=CC2C(=C1)CC[C@@H]1C2CCC2CCC[C@H]21	26	Protein Binder P&B	1.14%
O=C1C=C2CC[C@H]3[C@@H]4CCCC4CC[C@@H]3C2CC1	24	Protein Binder P&B	1.05%
c1ccc(Cc2ccccc2)cc1	23	Protein Binder P&B	1.01%
O=C1CN=C(c2ccccc2)c2ccccc2N1	21	Protein Binder P&B	0.92%
c1ccccc1	668	Non-Binder ZINC	2.68%
c1ccncc1	129	Non-Binder ZINC	0.52%
O=C(CNc1ccccc1)NCc1ccccc1	98	Non-Binder ZINC	0.39%
C1CCNCC1	63	Non-Binder ZINC	0.25%
c1cn[nH]c1	61	Non-Binder ZINC	0.24%

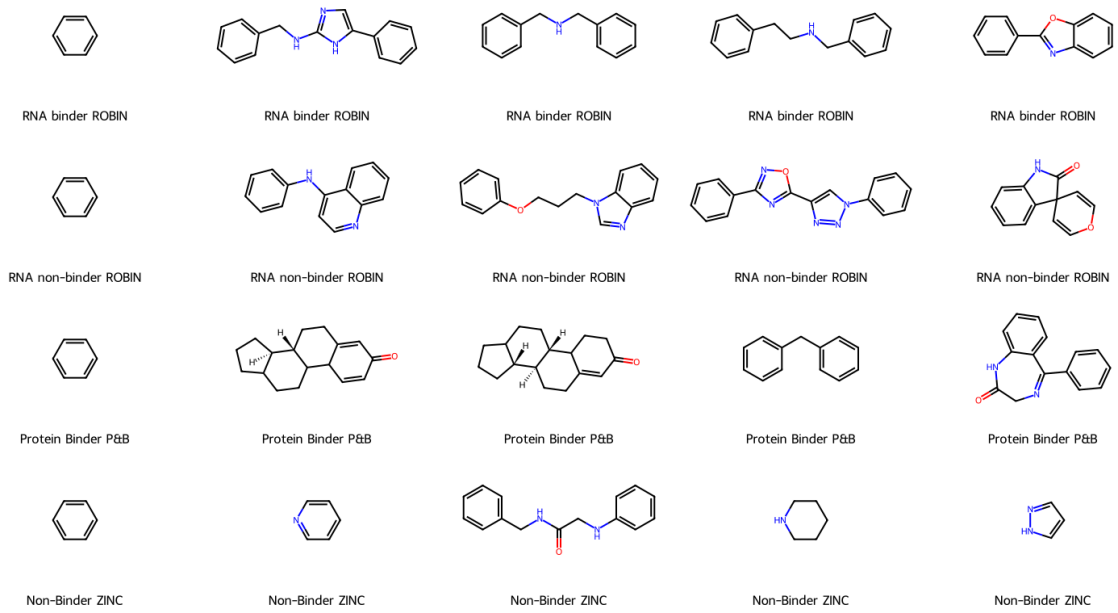


Figure 20: Visual representation of the predominant Murcko scaffolds identified in RNA and protein binders for $Set2\_Small$.

CSK Scaffolds

The CSK scaffold analysis presented a different perspective, with a clear delineation in the complexity and type of cyclic structures between the datasets. Protein binders exhibited larger and more complex cyclic structures, likely reflecting the structural requirements for protein interaction sites. The distribution of CSK scaffolds for both datasets is shown in Tables 3 and 4.

Table 3: Top 5 CSK Scaffolds in $Set1\_Large$

CSK Scaffold	Count	Source	Ratio
C1CCC(CCC2CCCCC2)CC1	12107	Protein Binder Enamine	2.65%
C1CCC(CCCC2CCCC2)CC1	10055	Protein Binder Enamine	2.20%
C1CCC(CCCC2CCCCC2)CC1	9741	Protein Binder Enamine	2.14%
C1CCC(CCC2CCCC2)CC1	9652	Protein Binder Enamine	2.12%
C1CCC(CC2CCCCC2)CC1	7673	Protein Binder Enamine	1.68%
C1CCC(CCC2CCCCC2)CC1	297	Chemdiv RNA	1.49%
C1CCC(CC2CCC(CC3CCCCC3)CC2)CC1	268	Chemdiv RNA	1.35%
C1CCC(CCCCC2CCCCC2)CC1	143	Chemdiv RNA	0.72%
C1CCC(CCCC2CCCCC2)CC1	143	Chemdiv RNA	0.72%
C1CCC(CC2CCCCC2)CC1	127	Chemdiv RNA	0.64%
C1CCC(CCCC2CCCC2)CC1	188	Enamine RNA	1.63%
C1CCC(CCCCC2CCCC2)CC1	185	Enamine RNA	1.61%
C1CCC(CCCC2CCCCC2)CC1	142	Enamine RNA	1.23%
C1CCC(CCCCCC2CCCC2)CC1	142	Enamine RNA	1.23%
C1CCC(CCC2CCCC2)CC1	132	Enamine RNA	1.15%
C1CCC2CCCC2C1	402	Life Chemicals RNA	7.57%
C1CCC(CCC2CC3CCCCC3C2)CC1	257	Life Chemicals RNA	4.84%
C1CCC(C2CC3CCCCC3C2)CC1	238	Life Chemicals RNA	4.48%
C1CCC(C2CCCC2)CC1	214	Life Chemicals RNA	4.03%
C1CCC(C2CCCCC2)CC1	162	Life Chemicals RNA	3.05%
C1CCCCC1	67	ROBIN RNA	3.36%
C1CCC(C2CC3CCCCC3C2)CC1	50	ROBIN RNA	2.51%
C1CCC(CC2CCC(C3CCCCC3)C2)CC1	49	ROBIN RNA	2.46%
C1CCC(CCC2CCC(C3CCCCC3)C2)CC1	45	ROBIN RNA	2.26%
C1CCC(CCCCC2CCCCC2)CC1	44	ROBIN RNA	2.21%

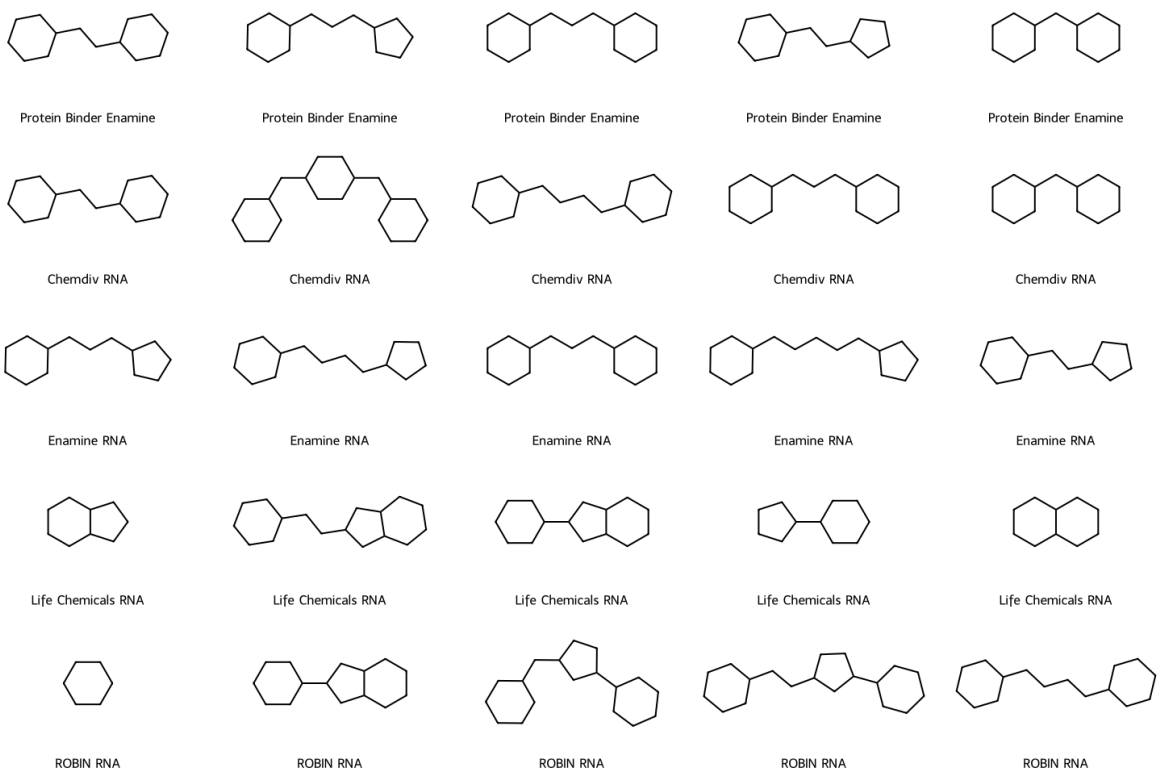


Figure 21: Visual representation of the predominant CSK scaffolds identified in RNA and protein binders for $Set1\_Large$.

Table 4: Top 5 CSK Scaffolds in $Set2\_Small$

CSK Scaffold	Count	Source	Ratio
C1CCCCC1	63	RNA binder ROBIN	3.21%
C1CCC(C2CC3CCCCC3C2)CC1	50	RNA binder ROBIN	2.55%
C1CCC(CC2CCC(C3CCCCC3)C2)CC1	49	RNA binder ROBIN	2.50%
C1CCC(CCC2CCC(C3CCCCC3)C2)CC1	45	RNA binder ROBIN	2.29%
C1CCC(CCCCC2CCCCC2)CC1	44	RNA binder ROBIN	2.24%
C1CCCCC1	611	RNA non-binder ROBIN	2.79%
C1CCC(C2CCCC2)CC1	490	RNA non-binder ROBIN	2.24%
C1CCC2CCCC2C1	434	RNA non-binder ROBIN	1.98%
C1CCC(CC2CCCCC2)CC1	349	RNA non-binder ROBIN	1.59%
C1CCC(CCC2CCCCC2)CC1	305	RNA non-binder ROBIN	1.39%
C1CCCCC1	219	Protein Binder P&B	9.62%
C1CCC2C(C1)CCC1C3CCCC3CCC21	155	Protein Binder P&B	6.81%
C1CCC(CC2CCCCC2)CC1	90	Protein Binder P&B	3.95%
C1CCC(CCC2CCCCC2)CC1	66	Protein Binder P&B	2.90%
C1CCC2CCCCC2C1	54	Protein Binder P&B	2.37%
C1CCCCC1	1130	Non-Binder ZINC	4.53%
C1CCCC1	417	Non-Binder ZINC	1.67%
C1CCC(C2CCCC2)CC1	410	Non-Binder ZINC	1.64%
C1CCC2CCCC2C1	386	Non-Binder ZINC	1.55%
C1CCC(CCC2CCCCC2)CC1	352	Non-Binder ZINC	1.41%

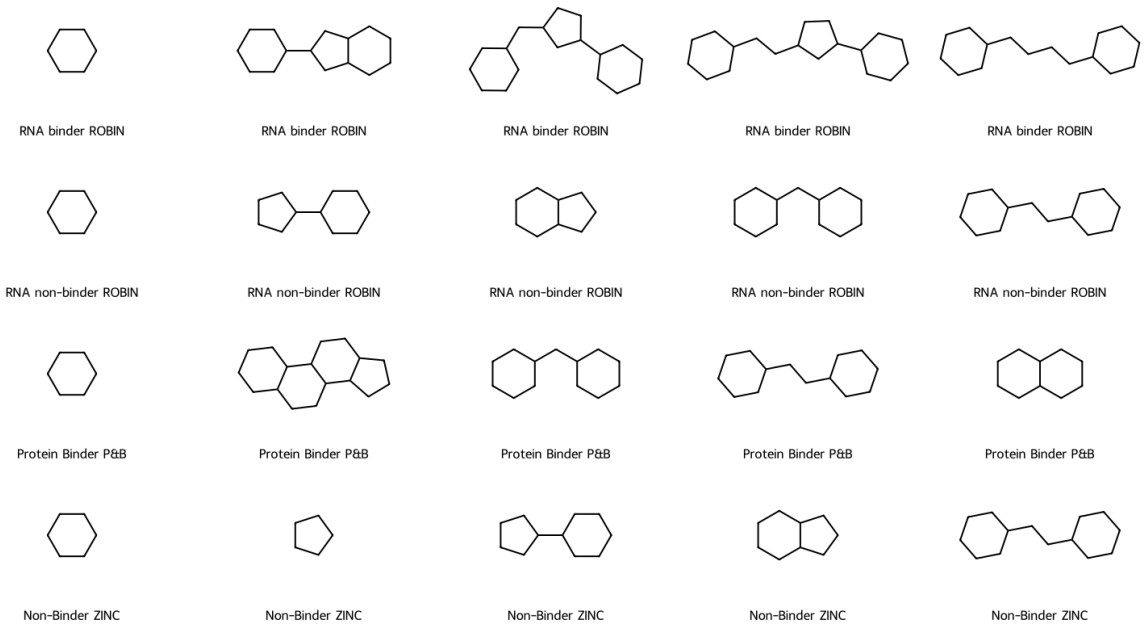


Figure 22: Visual representation of the predominant CSK scaffolds identified in RNA and protein binders for $Set2\_Small$.

4.4.1 Discussion

The differences in scaffold characteristics between RNA and protein binders suggest that ligand design can be tailored based on the target molecule's nature. For RNA targets, enhancing planarity and optimizing stacking interactions should be prioritized, as these features facilitate precise molecular interactions through stacking and potential hydrogen bonding with RNA bases. Conversely, protein targets often benefit from more substantial, complex scaffolds that can engage multiple interaction points within protein pockets. These scaffolds typically exhibit larger and more complex cyclic structures, reflecting the structural requirements for specific and strong interactions with protein binding sites.

The empirical data underscore that RNA-binding scaffolds are often simpler and less bulky, which may allow them to flexibly fit into the RNA structures and engage in essential interactions without the steric obstacles that could hinder binding. In contrast, protein-binding ligands often incorporate bulkier and more rigid frameworks to fit tightly into specific protein pockets, suggesting a need for a high degree of specificity and strong interaction to achieve effective binding.

These insights are invaluable for the targeted design of therapeutics with high specificity and efficacy. Further studies and iterative design based on these scaffold insights will enable the development of more effective and selective therapeutic agents, potentially leading to breakthroughs in drug design and development. This approach not only enhances our understanding of molecular interactions but also supports the advancement of cheminformatics as a crucial tool in drug discovery.

4.5 Machine Learning

4.5.1 Performance of Ensemble Models

$Set1\_Large$ Dataset: RNA Binders vs Protein Binders

Table 5: Summary of ensemble model performances (mean ± std from 10 runs) on the $Set1\_Large$ dataset, detailing accuracy, precision, recall, F1 score, and ROC AUC, with the best metrics in bold.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
XGBoost	**88.47 ± 0.32%**	**89.11 ± 0.39%**	87.64 ± 0.50%	**88.37 ± 0.34%**	**95.38 ± 0.21%**
LightGBM	88.36 ± 0.32%	88.81 ± 0.34%	**87.79 ± 0.58%**	88.30 ± 0.34%	95.26 ± 0.26%
Random Forest	86.61 ± 0.25%	88.12 ± 0.52%	84.63 ± 0.80%	86.34 ± 0.30%	94.15 ± 0.20%

The ensemble machine learning models, including XGBoost, LightGBM, and Random Forest, were applied to the large dataset comprising RNA binders vs protein binders. These models demonstrated high accuracy, with XGBoost showing slightly superior performance across most metrics, highlighting its robustness in handling complex datasets.

$Set2\_Small$ Dataset: Detailed Analysis

The $Set2\_Small$ dataset was further divided into three different model comparisons. RNA-Binders vs Protein-Binders, RNA-Binders vs Non-RNA-Binders and RNA-Binders vs Non-Binders. The performance of each model is detailed below:

Table 6: Performance comparison (mean ± std from 10 runs) of ensemble models on RNA-Binders vs Protein-Binders within the $Set2\_Small$ dataset, highlighting the top metrics in bold.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
XGBoost	87.12 ± 0.42%	88.23 ± 0.49%	85.64 ± 0.66%	86.91 ± 0.44%	94.41 ± 0.24%
LightGBM	**87.69 ± 0.60%**	**88.65 ± 0.70%**	86.43 ± 0.81%	**87.52 ± 0.61%**	94.47 ± 0.27%
Random Forest	86.34 ± 0.45%	85.78 ± 0.51%	**87.09 ± 0.51%**	86.43 ± 0.44%	**94.75 ± 0.16%**

Table 7: Evaluation of ensemble models on RNA Binders vs RNA Non-Binders in $Set2\_Small$ (mean ± std from 10 runs), with the highest scores bolded.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
XGBoost	60.64 ± 0.92%	61.53 ± 1.04%	56.51 ± 1.04%	58.91 ± 0.96%	65.49 ± 0.68%
LightGBM	59.83 ± 1.28%	60.60 ± 1.50%	55.99 ± 1.08%	58.20 ± 1.19%	64.05 ± 1.32%
Random Forest	**61.90 ± 0.58%**	**63.24 ± 0.71%**	**56.61 ± 1.17%**	**59.73 ± 0.75%**	**66.90 ± 0.25%**

Table 8: Analysis of ensemble models differentiating RNA-Binders from Non-Binders in $Set2\_Small$ (mean ± std from 10 runs), with best performance metrics bolded.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
XGBoost	72.65 ± 0.61%	71.76 ± 0.71%	**74.59 ± 0.89%**	73.15 ± 0.60%	**80.05 ± 0.28%**
LightGBM	71.82 ± 0.88%	70.80 ± 0.98%	74.18 ± 1.54%	72.44 ± 0.91%	79.25 ± 0.83%
Random Forest	**73.59 ± 0.89%**	**73.23 ± 0.52%**	74.26 ± 1.88%	**73.73 ± 1.14%**	79.96 ± 0.27%

These tables illustrate the nuanced capabilities of each model in distinguishing between closely related molecular interactions, with notable variations in performance metrics indicating the complexity of RNA-ligand interaction prediction.

4.6 Performance of GNN Models

The GNN models, including GatedGraphConv, GATv2Conv, and SageConv, were evaluated under similar conditions but with an emphasis on capturing the structural and relational nuances of molecules.

Table 9: GNN model performances on $Set1\_Large$ (mean ± std from 10 runs), showcasing accuracy, precision, recall, F1 score, and ROC AUC with best results in bold.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
GatedGraphConv	**87.33 ± 0.97%**	**87.52 ± 1.44%**	87.13 ± 2.62%	**87.29 ± 1.10%**	**87.33 ± 0.97%**
GATv2Conv	82.98 ± 1.88%	82.18 ± 2.72%	84.34 ± 1.29%	83.23 ± 1.65%	82.98 ± 1.88%
SageConv	86.53 ± 0.64%	85.96 ± 1.40%	**87.37 ± 1.56%**	86.64 ± 0.62%	86.53 ± 0.64%

Table 10: Performance of GNN models on RNA Binders vs Protein Binders in $Set2\_Small$ (mean ± std from 10 runs), highlighting top metrics in bold.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
GatedGraphConv	85.78 ± 5.25%	**87.82 ± 6.27%**	83.56 ± 7.18%	85.41 ± 5.26%	85.78 ± 5.25%
GATv2Conv	83.96 ± 3.74%	83.20 ± 4.67%	85.35 ± 3.65%	84.20 ± 3.53%	83.96 ± 3.74%
SageConv	**87.19 ± 1.60%**	87.70 ± 2.02%	**86.61 ± 3.70%**	**87.09 ± 1.82%**	**87.19 ± 1.60%**

Table 11: Operational efficiency of GNN models on RNA Binders vs RNA Non-Binders in $Set2\_Small$ (mean ± std from 10 runs), with best scores bolded.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
GatedGraphConv	60.98 ± 2.00%	61.80 ± 3.92%	60.91 ± 10.69%	60.61 ± 3.40%	60.98 ± 2.00%
GATv2Conv	60.13 ± 2.92%	60.88 ± 5.71%	**61.62 ± 8.74%**	60.51 ± 2.55%	60.13 ± 2.91%
SageConv	**61.78 ± 2.32%**	**62.70 ± 3.38%**	59.40 ± 6.98%	**60.69 ± 3.31%**	**61.78 ± 2.32%**

Table 12: GNN model results for distinguishing RNA-Binders from Non-Binders in $Set2\_Small$ (mean ± std from 10 runs), with optimal metrics in bold.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
GatedGraphConv	71.17 ± 8.64%	68.37 ± 7.60%	**83.13 ± 8.87%**	74.48 ± 5.23%	71.18 ± 8.63%
GATv2Conv	72.15 ± 4.32%	69.85 ± 3.84%	78.03 ± 5.48%	73.66 ± 4.17%	72.15 ± 4.32%
SageConv	**74.68 ± 1.20%**	**71.98 ± 1.21%**	80.97 ± 5.01%	**76.12 ± 1.86%**	**74.68 ± 1.20%**

The evaluation of Graph Neural Network (GNN) models across different datasets and classification tasks highlights their robustness and adaptability. In the $Set1\_Large$ dataset, the GatedGraphConv model demonstrated superior performance with an accuracy of 87.33% ± 0.97% and consistent high scores across precision, recall, F1 Score, and ROC AUC, suggesting effective handling of large-scale data. Comparatively, on the more focused $Set2\_Small$ dataset, while the SageConv model outperformed others with an accuracy of 87.19% ± 1.60% when differentiating between RNA binders and protein binders, its efficacy slightly dropped in scenarios involving RNA binders vs. RNA non-binders, reflecting the challenging nature of fine-grained classification tasks. However, its recovery in performance was notable in differentiating RNA binders from non-binders with an improved accuracy of 74.68% ± 1.20%, underscoring its strength in discerning subtle differences between closely related classes. These results collectively illustrate that while all GNN models show competence, certain architectures like SageConv may offer advantages in specific scenarios, which is crucial for targeted applications in RNA-binding ligand identification.

4.6.1 Discussion

In the $Set1\_Large$ dataset, ensemble models, particularly XGBoost, demonstrated strong accuracy with XGBoost leading at 88.47%. This indicates its effectiveness in handling datasets with complex RNA and protein interactions. In contrast, the GNN models, such as SageConv, performed optimally in the $Set2\_Small$ dataset, particularly notable was its accuracy rate of 87.19% when distinguishing between RNA binders and protein binders. This superior performance illustrates GNNs' capability to exploit molecular structural nuances essential for accurate classification in chemically detailed contexts.

4.7 Feature Importance Analysis and Fragment Visualization

This section outlines the results from feature importance analyses conducted with two distinct ensemble machine learning models: XGBoost and LightGBM. Each model was trained on a unique dataset, which provided valuable insights into the molecular determinants that underpin RNA-ligand interactions. A synthesis of empirical data with machine learning interpretations has yielded a set of molecular fragments that are postulated to be critical for RNA-binding affinity.

XGBoost Analysis on the Large Merged Dataset

The XGBoost model applied to the large merged dataset, consisting of RNA and protein binders, has also surfaced a cohort of molecular fragments with pronounced predictive power.

Physicochemical Traits of Identified Fragments The fragments identified through the XGBoost model highlight the physicochemical diversity necessary for effective RNA interaction. Recent analyses show that RNA-binding molecules typically exhibit increased counts of hydrogen bond donors and acceptors, along with a preference for aromatic structural features, which enhance molecular interaction with RNA through π-π stacking and hydrogen bonds [5].

Table 13: Top 20 Molecular Fragments Identified from the XGBoost Model, based on occurrences. Their visualization is given in Fig. 23

Rank	Fragment SMILES	Occurrences	Frequency	Predicted Proba
1	cc(n)=O	1109	0.1253	0.9899
2	CC(N)=O	946	0.1069	0.9843
3	O=S	542	0.0612	0.9481
4	cc1ncnc1n(c)C	491	0.0555	0.9960
5	ccc(cc)C(N)=O	395	0.0446	0.9827
6	cc(=O)n(C)c(n)=O	376	0.0425	0.9936
7	cc(O)cc(c)O	314	0.0355	0.9943
8	ccncc	313	0.0354	0.9846
9	CN(C)Cc1ccccc1	277	0.0313	0.9905
10	COc1cccc(OC)c1	268	0.0303	0.9941
11	c1ccsc1	186	0.0210	0.9711
12	cc(c)CN1CCNCC1	185	0.0209	0.9899
13	Cc1cccc(N)c1	124	0.0140	0.9913
14	CCCC(N)=O	123	0.0139	0.9884
15	Cc1ccc(N)s1	119	0.0134	0.9913
16	CN(C)Cc(c)c	91	0.0103	0.9831
17	ccc(cc)C(F)(F)F	76	0.0086	0.9835
18	C[C@H]1CCcc(s)C1	75	0.0085	0.9952
19	cc1cC[C@@H](C)CC1	75	0.0085	0.9952
20	COc1ccc(C)cc1OC	68	0.0077	0.9842

XGBoost Model Fragment Library

Figure 23 displays the molecular fragments identified as important by the XGBoost model. These fragments represent the substructural features most associated with RNA-binding molecules within the dataset used for the larger merged dataset analysis.

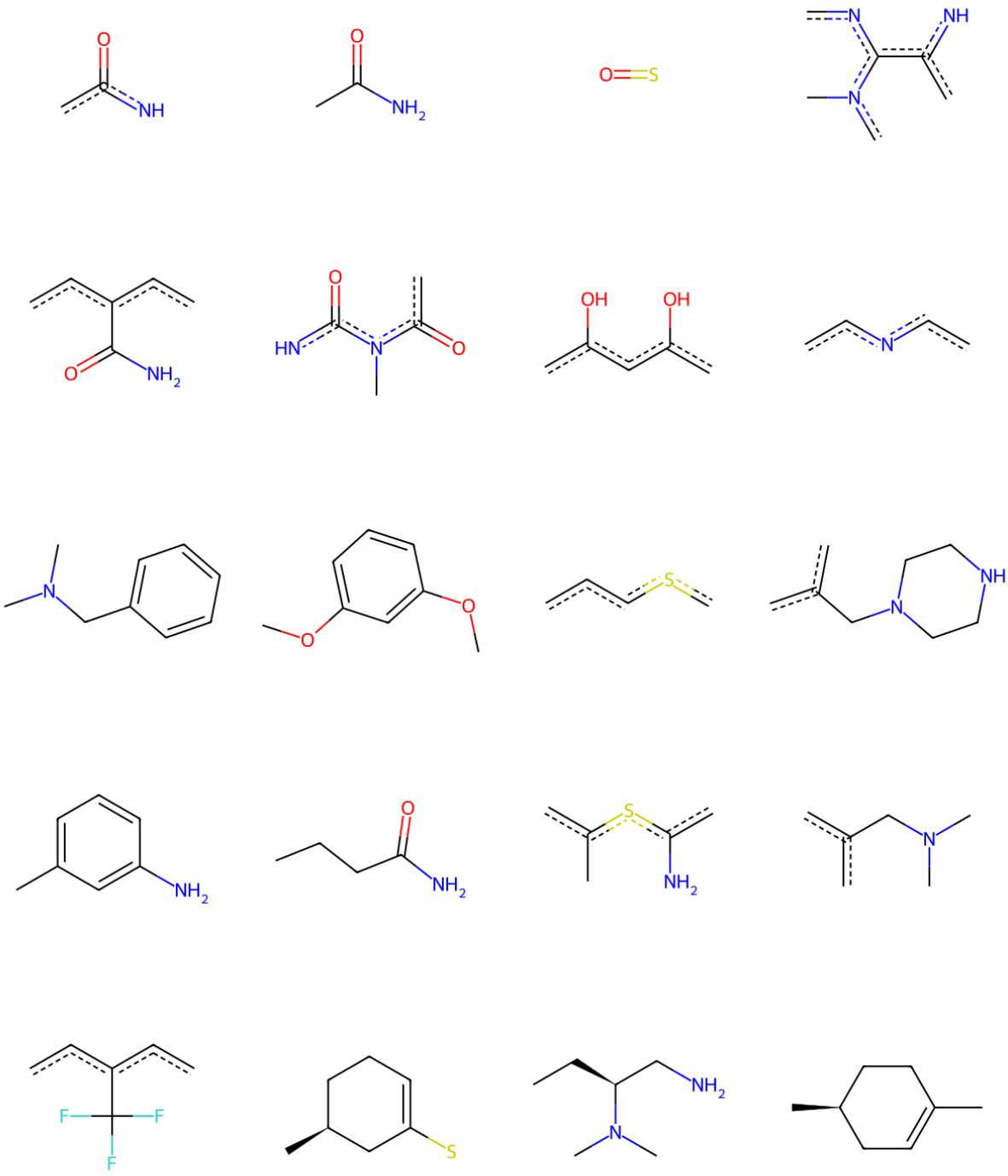


Figure 23: Visualization of the most important molecular fragments as identified by the XGBoost model.

LightGBM Analysis on the $Set2\_Small$

Analysis Overview The LightGBM model, utilizing a smaller dataset comprised of RNA-binding ligands from the ROBIN database and protein binders from Probes & Drugs, revealed several key molecular fragments. These findings substantiate our understanding of the structural features that contribute to RNA-binding specificity.

Table 14: Top 20 Molecular Fragments Identified from the LightGBM Model, based on occurrences. Their visualzation is given in Fig. 24

Rank	Fragment SMILES	Occurrences	Frequency	Predicted Proba
1	ccc	2031	0.1171	0.9899
2	ccccc	479	0.0276	0.9910
3	CCN	429	0.0247	0.9881
4	CCC	387	0.0223	0.9885
5	cc(c)C	313	0.0180	0.9906
6	cnc	295	0.0170	0.9902
7	cccc(c)C	260	0.0150	0.9904
8	cC	180	0.0104	0.9940
9	c-c(c)ccc	173	0.0100	0.9937
10	CO	169	0.0097	0.9889
11	cccc(c)N	162	0.0093	0.9869
12	cc(c)N	149	0.0086	0.9881
13	C=O	149	0.0086	0.9835
14	CC	145	0.0084	0.9882
15	CCO	144	0.0083	0.9901
16	cOC	142	0.0082	0.9863
17	cc(c)O	142	0.0082	0.9852
18	c-c(c)c	135	0.0078	0.9937
19	cCN	122	0.0070	0.9939
20	ccn	119	0.0069	0.9906

Visualization of Feature Importance and Molecular Fragments

The importance of aromatic rings and hydrogen bond donors and acceptors in the identified fragments can be directly linked to their roles in RNA-binding. Aromatic compounds often facilitate stacking interactions, a critical mechanism for binding to RNA structures [25]. Additionally, the presence of multiple hydrogen bond donors and acceptors in these fragments supports their high RNA-binding efficacy, as demonstrated in recent cheminformatics studies [3].

LightGBM Model Fragment Library

Figure 24 illustrates the molecular fragments as determined by the LightGBM model, trained on the smaller dataset. The occurrences of these fragments underscore their relevance in the context of RNA-binding potential.

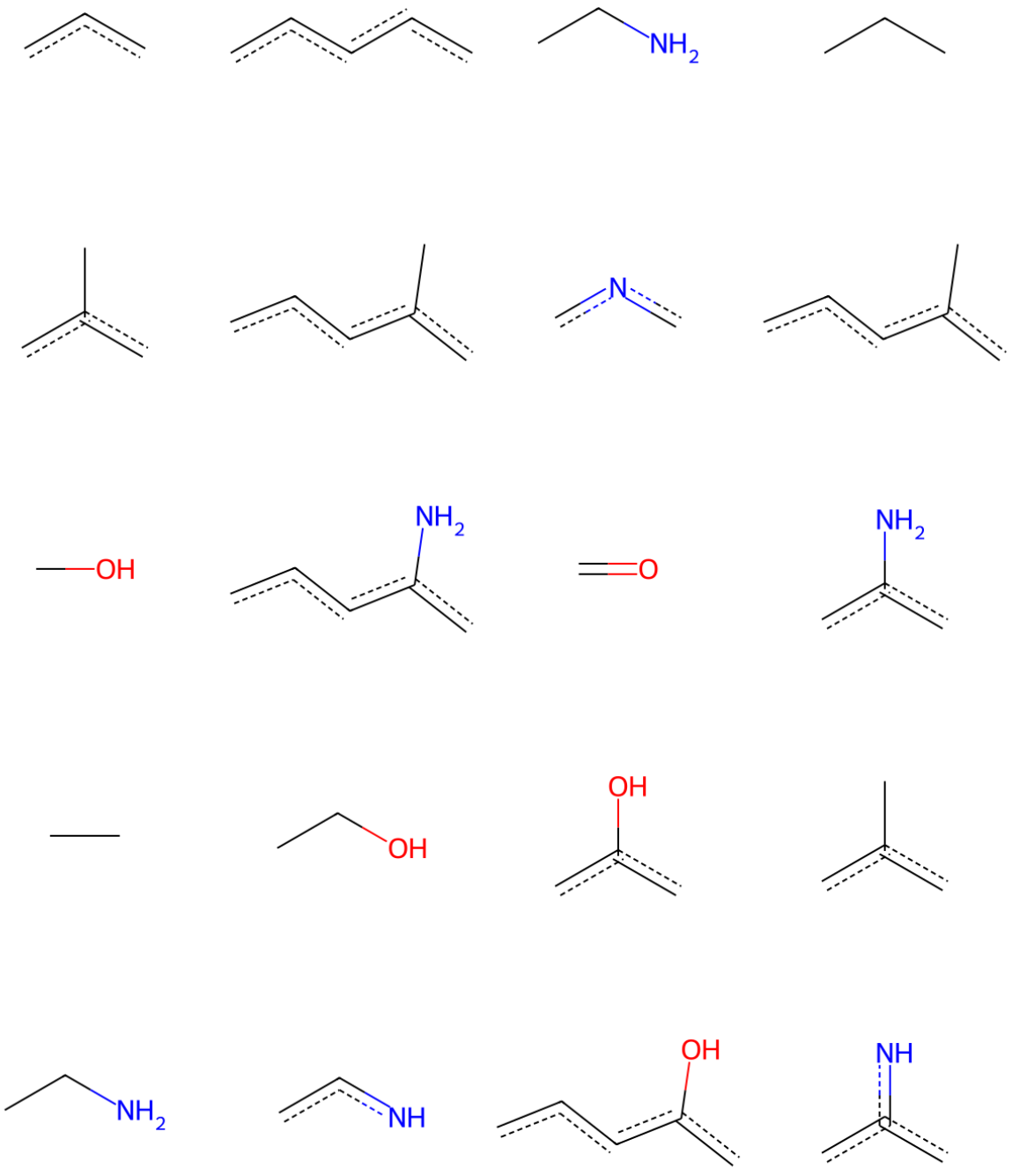


Figure 24: Most occurring molecular fragments identified by the LightGBM model.

Feature Importance in ML Models and Substructure Identification

In the analysis of RNA-binding and protein-binding ligands, the feature importance generated by machine learning models such as LightGBM (LGBM) and XGBoost (XGB) provided intriguing insights into the molecular features crucial for binding activity. In particular, LGBM tended to highlight entire molecules as important features for classification tasks, whereas XGB often pinpointed specific substructures or fragments within the molecules. This discrepancy can be attributed to inherent differences in the way these models handle feature interactions and dependencies. LGBM, with its gradient-based one-sided sampling and exclusive feature bundling, tends to generalize more, capturing broader patterns across entire molecules, which could indicate a holistic approach to identifying molecule-wide properties relevant to binding. In contrast, XGB, which utilizes a more precise and often more fragmented approach in tree construction, focuses on specific interactions and is better at identifying discrete substructures that directly contribute to the activity. Chemically, this means that XGB is effective at detecting critical functional groups or ring systems within a scaffold that are key for molecular interaction, such as hydrogen bond donors or acceptors, and aromatic rings involved in pi-stacking interactions with biological targets. This selective identification aligns with the chemical theory that specific atomic or group features within a molecule can substantially influence its binding affinity and specificity. Thus, while both models achieve similar accuracy, their interpretations of feature importance offer complementary views on the chemical features significant for ligand binding.

Implications for Drug Discovery

The identification of these fragments lays the foundation for future studies of the structure-activity relationship (SAR). They present a potential pathway for the exploration and development of novel RNA-targeted therapeutics, driving the cheminformatics field forward with data-driven insights.

Design and Optimization of RNA-Targeted Therapeutics

The recognition of specific fragment properties, such as increased aromaticity and hydrogen bonding potential, guides the rational design of RNA-targeted therapeutics. The insights derived from our machine learning analyses, corroborated by empirical data from recent studies, suggest a nuanced approach to developing molecules that can effectively target RNA with high specificity and efficacy [3].

5 CONCLUSIONS

In conclusion, this study provides an in-depth analysis of RNA and protein-binding ligands using sophisticated cheminformatics methodologies, including logistic PCA, t-SNE, UMAP, and advanced machine learning models. The systematic removal of duplicates from the datasets $Set1\_Large$ and $Set2\_Small$ refined the data curation process, offering insights into the chemical diversity and overlap between RNA and protein-binding libraries. These efforts, as outlined in the Venn diagrams, have underscored the critical role of initial data curation in enhancing the reliability and validity of conclusions drawn from subsequent analyses.

The comparative analysis of logistic PCA, t-SNE, and UMAP yielded valuable insights into the shared and distinct molecular features of RNA-binding versus protein-binding ligands. While logistic PCA provided broad structural motifs, t-SNE identified intricate clustering patterns that suggest distinct binding modalities (refer to Figure 8 and Figure 11), and UMAP offered a nuanced segregation of ligand chemical space (refer to Figure 9 and Figure 12), revealing privileged structures with high specificity. The nuanced analysis using these dimensionality reduction techniques underscores the importance of integrative approaches in predictive modeling and therapeutic design.

The ensemble and GNN models for binary classification both demonstrated robust capabilities in distinguishing between RNA-binding and protein-binding ligands. The ensemble models, particularly XGBoost, showed excellent performance on the $Set1\_Large$ dataset with an accuracy of 88.47% $\pm$ 0.32% (Table 5), reflecting their strength in handling diverse and complex datasets through feature engineering and traditional machine learning techniques. In contrast, the GNN models, such as GatedGraphConv and SageConv, excelled on structural data representations, with GatedGraphConv achieving an accuracy of 87.33% $\pm$ 0.97% on the $Set1\_Large$ dataset and SageConv reaching 87.19% $\pm$ 1.60% on the $Set2\_Small$ dataset (Table 10). These performances highlight GNNs' ability to capture intricate molecular interactions and subtle distinctions between ligand types, leveraging the relational data of molecules more effectively than traditional models.

The feature importance analysis using XGBoost and LightGBM provided valuable information about key molecular fragments influencing RNA-binding affinity (see Table 13 and Table 14). Visualizations of these fragments (see Figures 23 and 24) highlight the importance of structural motifs, such as aromatic rings and hydrogen-bond donors and acceptors, in ligand interaction potential. The fragment libraries identified in this study can guide future design and synthesis efforts to develop more targeted and effective RNA-binding therapeutics.

Overall, the results underscore the critical role of scaffold architecture in defining the binding affinity and specificity of ligands toward RNA and protein targets. Using insights from both $Set1\_Large$ and $Set2\_Small$, the study provides practical guidance for scaffold design, focusing on planarity and aromatic nature for RNA-binding ligands and complexity and rigidity for protein-binding ligands. These insights will help cheminformatics researchers in crafting better predictive models, facilitating the synthesis of novel compounds with promising biological activities.

The data and workflow utilized in this study can be accessed in the GitHub repository: `https://github.com/fulopjoz/diploma_thesis`.

Future work should expand the analysis to include more molecular descriptors and leverage identified patterns to guide the synthesis of new ligands. This comprehensive approach to feature importance and machine learning modeling will facilitate advancements in predictive modeling and pharmaceutical design.

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
CAD	Computer-Aided Design
ClogD	Calculated Distribution Coefficient
ClogP	Calculated LogP (Partition Coefficient)
CSK	Cyclic Skeleton
DGL	Deep Graph Library
ECFP6	Extended Connectivity Fingerprints, size 6
EFB	Exclusive Feature Bundling
FP	False Positive
FN	False Negative
FPR	False Positive Rate
GAT	Graph Attention Network
GATv2	Graph Attention Network version 2
GCN	Graph Convolutional Network
GG-NN	Gated Graph Neural Networks
GNN	Graph Neural Network
GOSS	Gradient-based One-Side Sampling
GRU	Gated Recurrent Unit
HARIBOSS	Harnessing RIBOnucleic acid-Small molecule Structures
LGBM	Light Gradient Boosting Machine
ML	Machine Learning
PCA	Principal Component Analysis
P&B	Probes & Drugs
QED	Quantitative Estimate of Drug-likeness
R-BIND	RNA-targeted BIoactive ligaNd Database (R-BIND)
RDKit	Cheminformatics software toolkit
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
ROC-AUC	Area Under the Receiver Operating Characteristic Curve
R-SIM	RNA-Small molecule Interaction Miner (R-SIM)
SAGEConv	Sample and Aggregate Convolution
SAR	Structure-Activity Relationship
SDF	Structure-Data File
SMILES	Simplified Molecular Input Line Entry System
TP	True Positive
TN	True Negative
TPR	True Positive Rate

TPSA	Topological Polar Surface Area
t-SNE	t-Distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
XGB	eXtreme Gradient Boosting

List of Figures

List of Tables

6 DECLARATION

Declaration of AI Utilization I hereby affirm that this thesis was independently prepared, utilizing artificial intelligence tools to support the writing and coding processes. I have thoroughly examined all AI-generated content to ensure accuracy and address any potential discrepancies.

I take full responsibility for the content of this thesis, affirming its accuracy and reliability. The use of AI technologies was conducted in strict adherence to ethical standards and principles of academic integrity, ensuring the originality and authenticity of the work.

List of AI Tools and Their Purposes

ChatGPT-4	`https://chat.openai.com/`
Usage:	Generation of outlines, drafts, rephrasing, and resource gathering
GitHub Copilot	`https://copilot.github.com`
Usage:	Code completion and correction
Grammarly	`https://www.grammarly.com`
Usage:	Text corrections
Writefull	`https://www.writefull.com/`
Usage:	Text corrections, rephrasing

SUPPLEMENTARY MATERIAL

A Tables

Model and Dataset	Best Hyperparameters
XGBoost (Set1: RNA Binders vs Protein Binders)	Learning rate: 0.2304, Max depth: 9, Estimators: 616, Subsample: 0.8946, Colsample bytree: 0.5081, Reg alpha: 0.0835, Reg lambda: 0.0109
Random Forest (Set2: RNA Binders vs RNA Non-Binders)	Max depth: 8, Estimators: 226, Min child weight: 6, Subsample: 0.7995, Colsample bytree: 0.7070, Reg alpha: 0.0107, Reg lambda: 2.1435
LightGBM (Set2: RNA Binders vs Protein Binders)	Learning rate: 0.0854, Max depth: 9, Estimators: 425, Min child weight: 1, Subsample: 0.8160, Colsample bytree: 0.6588, Reg alpha: 0.4880, Reg lambda: 3.8034
Random Forest (Set2: RNA-Binders vs Non-Binders)	Max depth: 8, Estimators: 165, Min child weight: 1, Subsample: 0.6349, Colsample bytree: 0.6276, Reg alpha: 0.0749, Reg lambda: 1.2649
SageConv (Set1: RNA Binders vs Protein Binders)	Hidden dim: 254, Aggregator type: mean, Dropout rate: 0.0357, Learning rate: 0.00203, Batch size: 128
SageConv (Set2: RNA Binders vs RNA Non-Binders)	Hidden dim: 160, Aggregator type: lstm, Dropout rate: 0.00826, Learning rate: 0.000675, Batch size: 256
GatedGraphConv (Set2: RNA Binders vs Protein Binders)	Steps: 2, Hidden dim: 187, Learning rate: 0.000876, Batch size: 128, Dropout rate: 0.365
SageConv (Set2: RNA-Binders vs Non-Binders)	Hidden dim: 124, Aggregator type: mean, Dropout rate: 0.0752, Learning rate: 0.00137, Batch size: 128

Table 16: Summary of the best hyperparameters for GNN and Ensemble models evaluated in the RNA-binding cheminformatics study. If not explicitly listed, default hyperparameters were used.

BIBLIOGRAPHY

[1] Manigrasso, J.; Marcia, M.; Vivo, M. D. Computer-aided design of RNA-targeted small molecules: A growing need in drug discovery. *Chem* **2021**, *7*, 2965–2988.

[2] Bernetti, M.; Aguti, R.; Bosio, S.; Recanatini, M.; Masetti, M.; Cavalli, A. Computational drug discovery under RNA times. *QRB Discovery* **2022**, *3*.

[3] Costales, M. G.; Childs-Disney, J. L.; Haniff, H. S.; Disney, M. D. How We Think about Targeting RNA with Small Molecules. *Journal of Medicinal Chemistry* **2020**, *63*, 8880–8900.

[4] Donlic, A.; Swanson, E. G.; Chiu, L.-Y.; Wicks, S. L.; Juru, A. U.; Cai, Z.; Kassam, K.; Laudeman, C.; Sanaba, B. G.; Sugarman, A.; Han, E.; Tolbert, B. S.; Hargrove, A. E. R-BIND 2.0: An Updated Database of Bioactive RNA-Targeting Small Molecules and Associated RNA Secondary Structures. *ACS Chemical Biology* **2022**, *17*, 1556–1566, PMID: 35594415.

[5] Disney, M. D. Targeting RNA with Small Molecules To Capture Opportunities at the Intersection of Chemistry, Biology, and Medicine. *Journal of the American Chemical Society* **2019**, *141*, 6776–6790.

[6] Lu, M. et al. Artificial Intelligence in Pharmaceutical Sciences. *Engineering* **2023**, *27*, 37–69.

[7] Guan, L.; Disney, M. D. Recent advances in developing small molecules targeting RNA. *ACS Chemical Biology* **2012**, *7*, 73–86.

[8] Krishnan, S. R.; Roy, A.; Gromiha, M. M. R-SIM: A Database of Binding Affinities for RNA-small Molecule Interactions. *Journal of Molecular Biology* **2023**, *435*.

[9] Panei, F. P.; Torchet, R.; Menager, H.; Gkeka, P.; Bonomi, M. HARIBOSS: a curated database of RNA-small molecules structures to aid rational drug design. *Bioinformatics* **2022**, *38*, 4185–4193.

[10] Garner, A. L. Contemporary Progress and Opportunities in RNA-Targeted Drug Discovery. *ACS Medicinal Chemistry Letters* **2023**, *14*, 251–259.

[11] Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**, *13-17-August-2016*, 785–794.

[12] Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5–32.

[13] Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree.

[14] Koren, Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA, 2008; p 426–434.

[15] Zhang, Y.; Sun, F.; Yang, X.; Xu, C.; Ou, W.; Zhang, Y. Graph-based Regularization on Embedding Layers for Recommendation. *ACM Trans. Inf. Syst.* **2020**, *39*.

[16] Wang, X.; He, X.; Wang, M.; Feng, F.; Chua, T.-S. Neural Graph Collaborative Filtering. Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019.

[17] Hu, X.; Liu, D.; Zhang, J.; Fan, Y.; Ouyang, T.; Luo, Y.; Zhang, Y.; Deng, L. A comprehensive review and evaluation of graph neural networks for non-coding RNA and complex disease associations. *Briefings in Bioinformatics* **2023**, *24*, bbad410.

[18] Li, Y.; Zemel, R.; Brockschmidt, M.; Tarlow, D. Gated Graph Sequence Neural Networks. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings* **2015**,

[19] Hamilton, W. L.; Ying, R.; Leskovec, J. Inductive Representation Learning on Large Graphs. *Advances in Neural Information Processing Systems* **2017**, *2017-December*, 1025–1035.

[20] Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. 2018.

[21] Sato, K.; Hamada, M.; Asahi-cho, S. Recent trends in RNA informatics: a review of machine learning and deep learning for RNA secondary structure prediction and RNA drug discovery. *Briefings in Bioinformatics* **2023**, *24*, 1–13.

[22] Schneider, B.; Sweeney, B. A.; Bateman, A.; Cerny, J.; Zok, T.; Szachniuk, M. When will RNA get its AlphaFold moment? *Nucleic Acids Research* **2023**, *51*, 9522–9532.

[23] Drugging the RNA World. *Cold Spring Harbor Perspectives in Biology* **2018**, *10*, a034769.

[24] Grover, N., Ed. *Fundamentals of RNA Structure and Function*, 1st ed.; Learning Materials in Biosciences; Springer Cham: Cham, 2022; pp XVI, 246, eBook Packages: Biomedical and Life Sciences, Biomedical and Life Sciences (R0).

[25] Falese, J. P.; Donlic, A.; Hargrove, A. E. Targeting RNA with small molecules: From fundamental principles towards the clinic. *Chemical Society Reviews* **2021**, *50*, 2224–2243.

[26] Bento, A. P.; Hersey, A.; Félix, E.; others An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics* **2020**, *12*, Available at https://github.com/chembl/ChEMBL_Structure_Pipeline.

[27] Enamine Hit Locator Library - 460. https://enamine.net/compound-libraries/diversity-libraries/hit-locator-library-460, 2024.

[28] ChemDiv miRNA Library. https://www.chemdiv.com/catalog/focused-and-targeted-libraries/mirna-library/, 2024.

[29] Enamine RNA Library. https://enamine.net/compound-libraries/targeted-libraries/rna, 2024.

[30] Life Chemicals RNA Focused Library. https://lifechemicals.com/screening-libraries/targeted-and-focused-screening-libraries/rna-focused-library, 2024.

[31] Yazdani et al. Machine Learning Informs RNA-Binding Chemical Space. https://figshare.com/articles/dataset/Machine_Learning_Informs_RNA-Binding_Chemical_Space/20401974, 2024.

[32] Probes & Drugs Database. https://www.probes-drugs.org/home, 2024.

[33] ZINC Database Dark Matter. https://files.docking.org/dark-matter/, 2024.

[34] RDKit Community Getting Started in Python - List of Available Descriptors. 2021; https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors, Accessed: 2024-05-02.

[35] Hill, A. P.; Young, R. J. Getting physical in drug discovery: a contemporary perspective on solubility and hydrophobicity. *Drug Discovery Today* **2010**, *15*, 648–655.

[36] Walters, P. Solubility Forecast Index (SFI) Calculation Code. 2021; https://github.com/PatWalters/sfi, Accessed: 2024-05-02.

[37] Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **1996**, *39*, 2887–2893, PMID: 8709122.

[38] Xu, Y.-j.; Johnson, M. Algorithm for Naming Molecular Equivalence Classes Represented by Labeled Pseudographs. *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 181–185, PMID: 11206372.

[39] Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019.

[40] M., H.; M.N, S. A Review On Evaluation Metrics For Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process* **2015**, *5*, 01–11.

[41] Brody, S.; Alon, U.; Yahav, E. How Attentive are Graph Attention Networks? 2022.

[42] Hamilton, W. L.; Ying, R.; Leskovec, J. Inductive Representation Learning on Large Graphs. 2018.

[43] Li, Y.; Tarlow, D.; Brockschmidt, M.; Zemel, R. Gated Graph Sequence Neural Networks. 2017.