RNA Structure Prediction

Name: TORRES MUNOZ HECTOR ID: H2410662
Name: INGA JUAN DIEGO ID: H2410581

Introduction

This project is based on RNA, a molecule that plays a vital role in biological systems, coming from DNA, is a direct transcript of our genetic information and allows protein synthesis in our cells. These proteins are fundamental for proper body function, making RNA an essential molecule for survival. When RNA molecules are synthesized, they are simple, linear chains without many functions, but quickly fold into complex tridimensional structures and acquire specific functions within the cell. Understanding RNA structure is crucial for understanding how biological systems work.

In this project we aim to predict the three-dimensional structure of RNA molecules using machine learning approaches. By analyzing the DNA sequences that the RNA comes from we can predict the final structure of the RNA molecule, and therefore, its function with further analysis. This has a high importance for a variety of biological processes, including gene regulation or protein synthesis mechanisms. Prediction of RNA is increasing its importance every day in the molecular biology and bioinformatics area and the advancements in computational methods and machine learning makes this task more feasible and accurate.

Materials and Methods

To predict RNA structures, we used a database consisting of a big amount of DNA sequences, from these sequences we can predict the RNA sequence and structure.

Different machine learning models were used for prediction: SVM, Random Forest and XGBoost, being the best and the final one, Random Forest. We measured the values of: Sensitivity, PPV, MCC and F1-score. These models are based on traditional machine learning approaches but we also worked with Eternafold, a specific machine learning model for predicting RNA molecule structures, these types of models are pretrained models used for RNA and other molecules.

Key Steps:

- **Sequence Input**: DNA sequences obtained from Kaggle.
- **Exploratory Data analysis**
- **Prediction Algorithm**: The computational tools Random Forest and Eternafold.

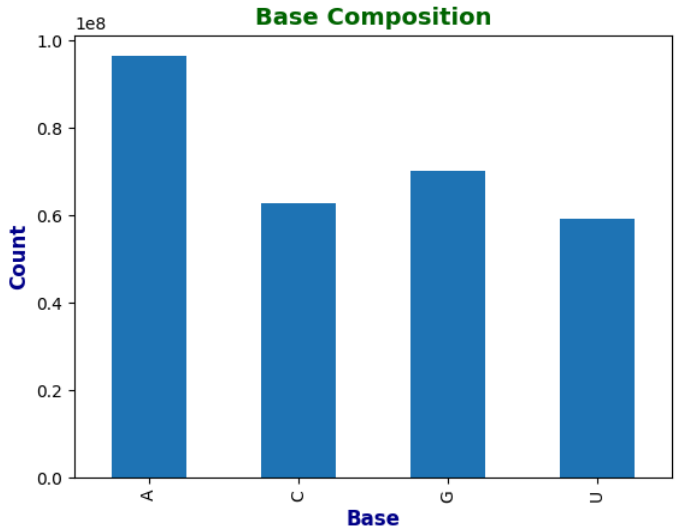
Used Parameters

The parameters used during the prediction process were:

1. **Sensitivity:** measures proportion of TP and FP + FN.
2. **PPV:** Positive Predictive Value measures proportion of TP out of all positive predictions made by the model.
3. **Matthews Correlation Coefficient MCC:** takes into account TP, TN, FP AND FN.
4. **F-measure (F1-score):** is the harmonic mean of precision and recall, in RNA structure prediction it helps evaluate the overall performance of the model.

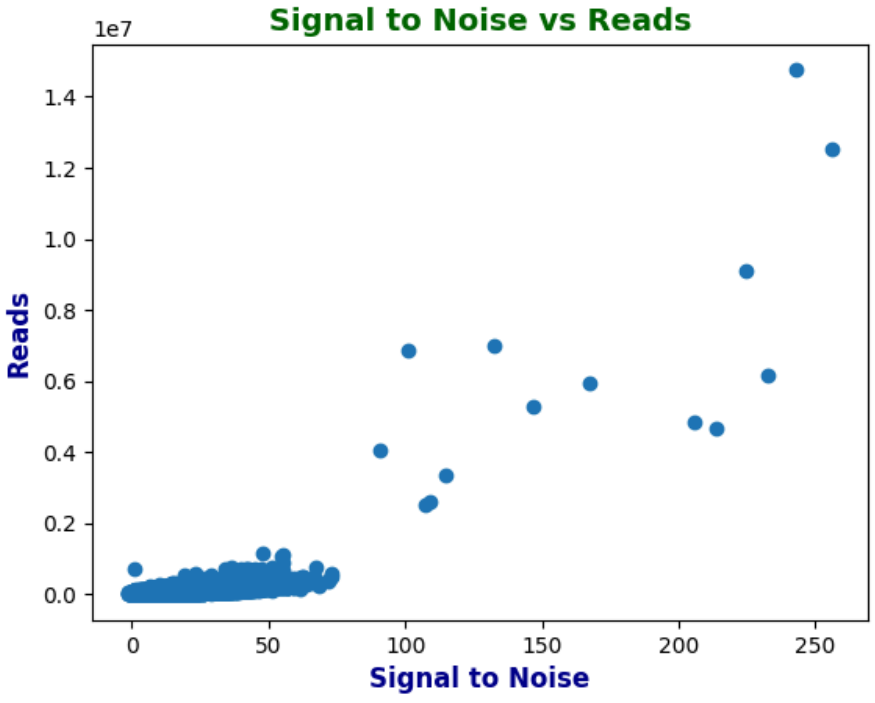
Exploratory Data Analysis

- Base composition of the sequences: bases are the components of each sequence of DNA and RNA.



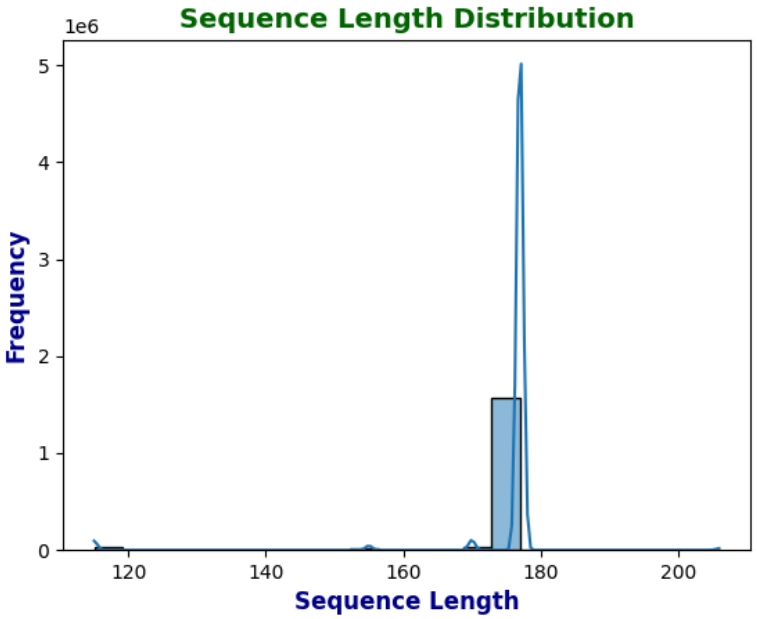
As we can see there is a higher amount of A (Adenine) in the sequences, however this imbalance is not big enough to worry about, if we had bigger imbalances like a big shortage of one of the bases, for example C, this could negatively affect our results

- SNR and Reads



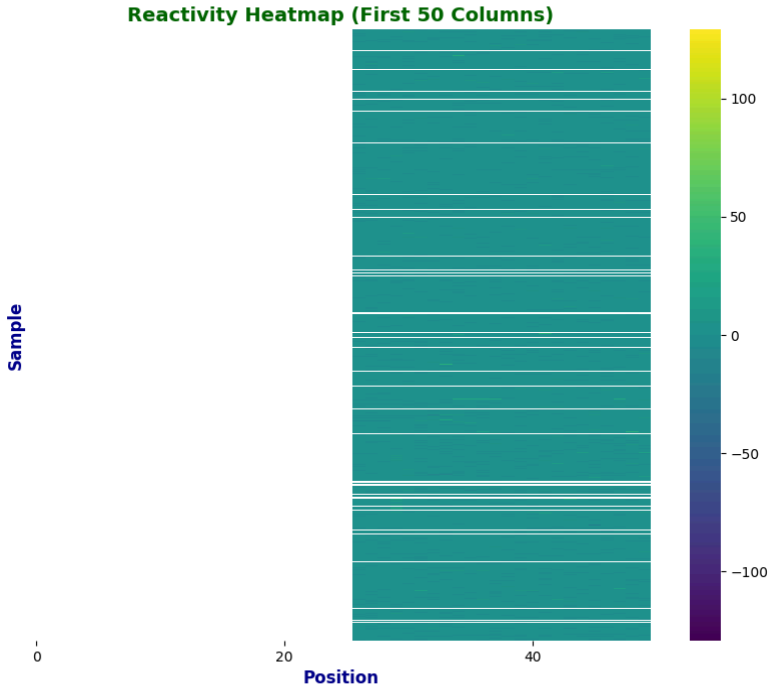
There is a big density of data with low values for Reads and Signal to noise, this could mean that the sequences are more inconsistent. Regardless of this, a big part of the DNA is made of sequences that do not code for anything or they are overlapped sequences for different RNA, so that's why we are not filtering this part.

- Sequence length distribution



There is almost no variability in the length differences in sequences. These are good results with high homogeneity, simplifying the preparation of the results. Because all the sequences are the same length, it is not necessary to align them, a step that would be important in case that the sequences were very different length wise.

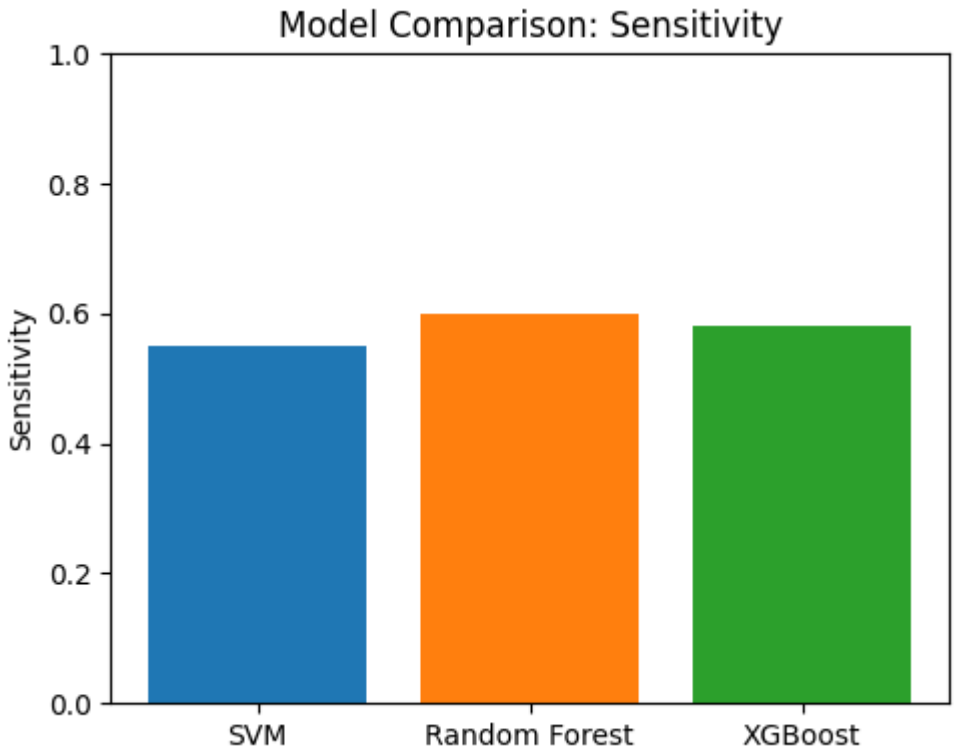
- Reactivity Heatmap



The heatmap was done only on 50 columns because there is a huge quantity of reads in the database. Here we can see the reactivity values of different regions of the sequences, this is important to know if the sequences have actual possibilities of generating a functional RNA.

Results

Sensitivity:

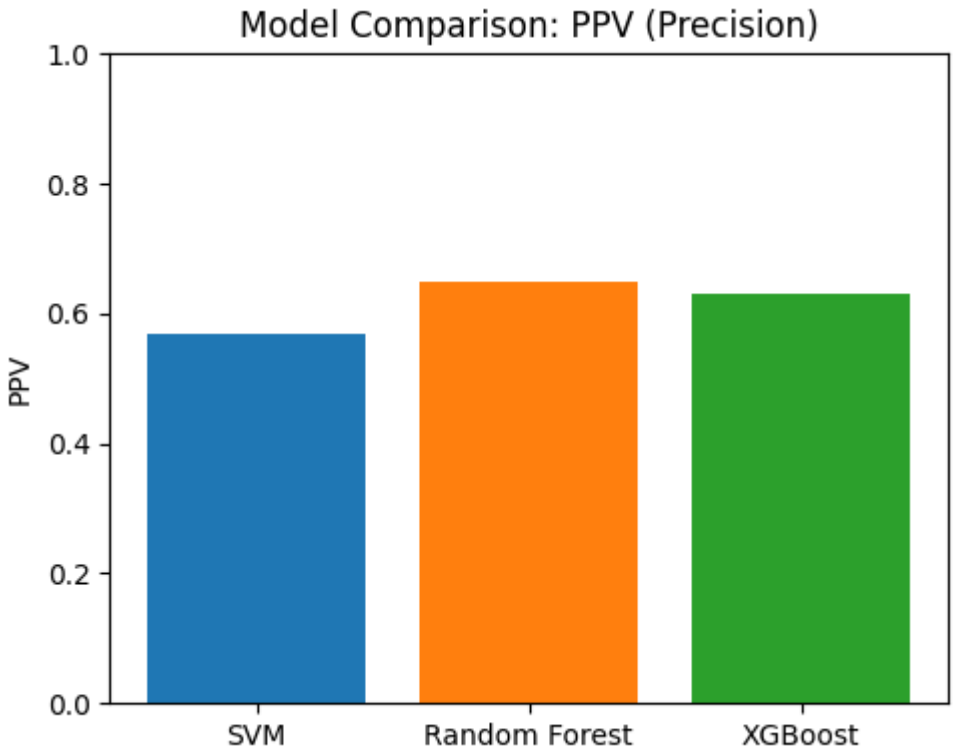


- **Highest**: Random Forest (0.60)
- SVM: 0.55
- XGBoost: 0.58

Plausible reason:

- A higher Sensitivity (or Recall) means the model correctly identifies more true positives out of all actual positives. Random Forest's highest sensitivity indicates it is somewhat more effective in capturing positive instances (e.g., correctly identifying base pairs).SVM, having the lowest sensitivity, may be missing more positive cases than expected.

<u>Positive Predictive Value (PPV) or Precision</u>:

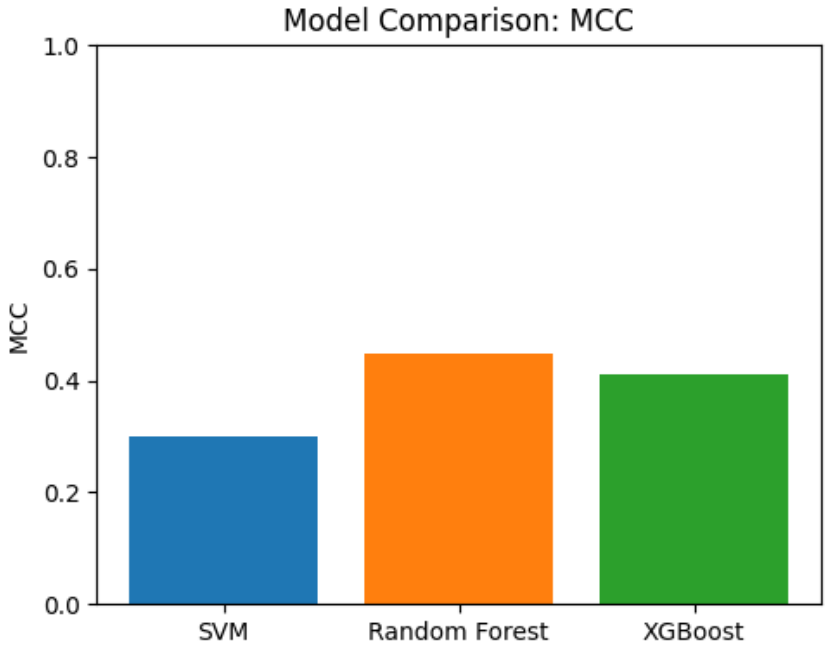


- **Highest:** Random Forest (0.65)
- SVM: 0.57
- XGBoost: 0.63

Plausible reason:

- Random Forest also excels here, which means when it predicts a positive, it is more likely to be correct. SVM has the lowest precision, so when SVM predicts something as positive, it's more often wrong compared to the other two models (i.e., more false positives).

MCC (Matthews Correlation Coefficient):

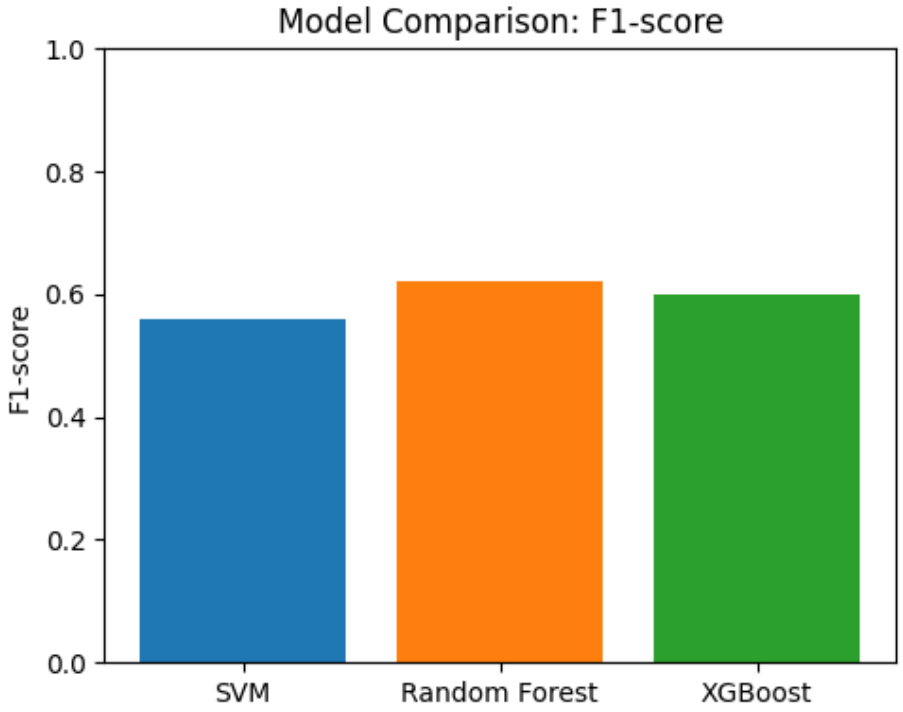


- **Highest**: Random Forest (0.45)
- SVM: 0.30
- XGBoost: 0.41

Plausible reason:

- MCC takes into account **all** confusion matrix elements (TP, TN, FP, FN). A higher MCC means a more balanced and robust performance. Random Forest again leads, with XGBoost second, and SVM in third place.SVM's lower MCC suggests it might be skewed in terms of false positives or false negatives.

F1-Score:



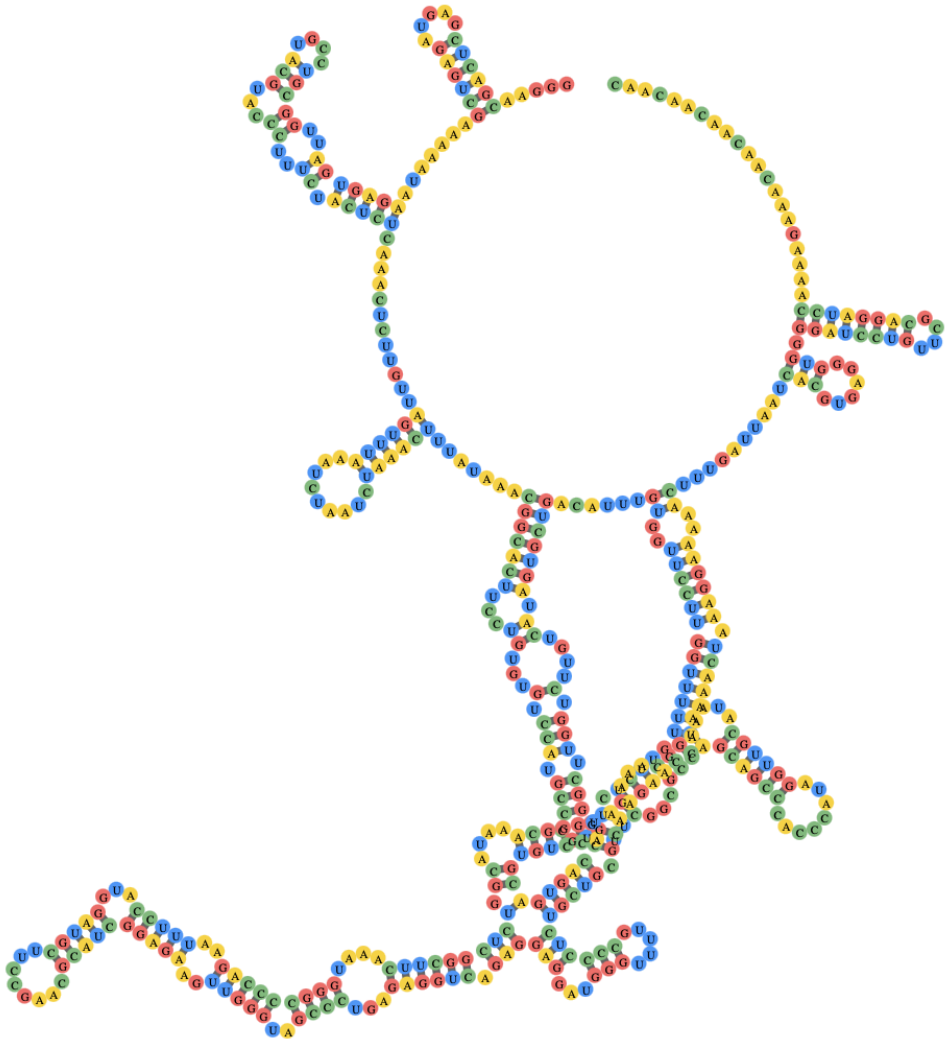
- **Highest**: Random Forest (0.62)
- SVM: 0.56
- XGBoost: 0.60

Plausible reason:

- F1-score is the harmonic mean of Precision and Recall (Sensitivity). A higher F1 indicates a good balance between catching positive cases (recall) and being correct about those predictions (precision). Random Forest's higher F1-score suggests that it maintains a **good trade-off** between Sensitivity and Precision.

Eternafold

EternaFold is a computational tool designed for RNA secondary structure prediction. It is a machine learning-based algorithm that uses neural networks to learn patterns in RNA folding.



As we said before, the first structure of RNA is a linear simple chain, and with eternafold we can predict the three dimensional structure of the molecule.

Analysis

In our evaluation, we compared three popular machine learning models—SVM, Random Forest, and XGBoost—to see which performs best for predicting RNA structures. We used four key metrics: Sensitivity, PPV (Precision), MCC, and F1-score. The results showed that Random Forest obtained the highest Sensitivity (0.60), indicating that it correctly captures a larger portion of actual positives (such as true base pairs) than the other two models. SVM, in comparison, had a Sensitivity of 0.55, potentially missing more positives, while XGBoost achieved 0.58, sitting between the other two.

Looking at PPV (Precision), Random Forest again led with 0.65, closely followed by XGBoost at 0.63, and then SVM at 0.57. In practical terms, when Random Forest predicts a base pair, it is correct more often than SVM or XGBoost, implying fewer false positives. The MCC values further highlight the performance balance across true positives, true negatives, false positives, and false negatives: Random Forest topped the chart at 0.45, with XGBoost at 0.41, and SVM at 0.30. Because MCC is sensitive to both classes, higher values typically suggest more consistent predictions across the board, underscoring Random Forest's stronger all-around performance.

Finally, in terms of the **F1-score**, Random Forest once again emerged as the leader with 0.62. Given that F1 is the harmonic mean of Precision and Recall (Sensitivity), this higher score indicates a solid balance between identifying positives and minimizing false alarms. XGBoost showed a slightly lower F1-score (0.60) but remained competitive, while SVM lagged behind at 0.56. This difference suggests that although SVM can identify true positives to some extent, it struggles to match the precision or balance of the other models.

It is worth noting that models can sometimes exhibit higher Sensitivity at the expense of Precision (or vice versa), which affects the F1-score. A model with very high Sensitivity but numerous false positives will see its F1-score drop. Conversely, a model with high Precision but missed positives will also have a lower F1-score. In this scenario, Random Forest maintained a favorable trade-off between these two factors—leading to strong overall performance in every metric.

Conclusions

In light of the results across Sensitivity, PPV (Precision), MCC, and F1-score, it is evident that all three models—SVM, Random Forest, and XGBoost—bring different strengths to the task of RNA structure prediction. In many cases, achieving high Sensitivity indicates that a model can successfully capture true positives (e.g., correctly identified base pairs). However, such a model might generate an abundance of false positives, which lowers its Precision and, ultimately, its F1-score. On the other hand, a model that focuses on minimizing false positives tends to have stronger Precision but risks missing actual positives, reducing its Sensitivity. The F1-score effectively balances these two contrasting elements, providing a more holistic view of a model's predictive ability.

Our findings show that Random Forest consistently demonstrates the best equilibrium among these metrics, particularly by maintaining higher Precision without sacrificing

Sensitivity. The model's highest F1-score confirms its strong overall performance in detecting true positives while simultaneously limiting inaccurate predictions. Furthermore, Random Forest's lead in MCC underscores that it handles both positive and negative classifications reliably, indicating a well-rounded robustness across all aspects of the confusion matrix.

In contrast, SVM achieved lower results in multiple metrics, suggesting it may require more extensive hyperparameter tuning or feature engineering designed for RNA sequences. Meanwhile, XGBoost ranks second in most metrics and remains competitive, indicating that additional optimization of parameters (e.g., learning rate or max depth) could potentially allow it to challenge or match Random Forest's performance.

Given the balancing act between detecting as many true positives as possible (high Sensitivity) and avoiding overprediction (maintaining high Precision), Random Forest emerges as the most reliable choice. It offers a strong trade-off across the board, making it the top model for RNA structure prediction based on the data and experimental conditions evaluated.

Bibliography

1. Wayment-Steele, H. K., Kladwang, W., Strom, A. I., Lee, J., Treuille, A., Becka, A., & Das, R. (2022). RNA secondary structure packages evaluated and improved by high-throughput experiments. Nature Methods, 19(10), 1234–1242. https://doi.org/10.1038/s41592-022-01605-0
2. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). From DNA to RNA. Molecular Biology of the Cell - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK26887/