**Structure Generating**

* Data Preprocessing: The code handles crucial initial steps like loading RNA sequence data, removing rows with missing sequences, and deduplicating entries to ensure that each sequence is unique. This preprocessing is fundamental because the dataset comprises experimentally validated interacting pairs of miRNA and lncRNA, where accurate structural information is critical for downstream analysis.
* Structure Prediction: The function predict\_structure\_with\_ipknot utilizes the IPknot tool to predict RNA secondary structures in dot-bracket notation. This is particularly significant for the project since one of the computational models being explored uses both sequence and structural information to predict miRNA-lncRNA interactions. Accurate structure prediction enables the Sequence+Structure Transformer Model to integrate structural details with sequence data, enhancing its ability to model RNA interactions based on both sequence identity and spatial conformation.
* Data Integration and Tracking: The script processes data through a structured workflow that tracks the progress and logs all significant steps and errors. This ensures transparency in the data handling process, which is crucial when managing large datasets in computational biology. This structured processing allows for the integration of newly predicted structural data into the dataset, which can then be used for training the sequence and structure-based computational models.
* Logging and Error Handling: Robust logging and error handling are built into the code to monitor the pipeline's progress and troubleshoot issues effectively. This aspect is vital in computational studies involving large datasets and complex models, as it helps in maintaining the integrity and reliability of the computational analysis.
* Efficiency and Automation: The script automates the process of appending structural data to the existing dataset, thereby saving time and reducing the risk of manual errors. This automation is essential for efficiently handling datasets with potentially thousands of RNA sequences, which is common in genomic studies.