**Trends in Characterization and Binding of DNA, RNA, and Protein Sequences**

**Team Members:** Micah, Ali, Shruti, and Jennifer

DNA and RNA research is a relatively new and exciting area that has garnered a lot of interest, especially after the success of RNA vaccines in saving humanity from the crippling pandemic stemming from the SARS-COV-2 virus. Much is unknown about DNA and RNA’s ability to act as binders and potential drugs by acting as inhibitors. There are a variety of new drugs in clinical trials right now, which are DNA and RNA-based sequences, but it is challenging to design these drugs in the first place as there is no definitive starting point like other research areas. There is a lot of data available about the structure and binding of DNA and RNA sequences, and data science can help us analyze this information and identify some key trends that can be used in making better DNA and RNA-based drugs.

In this project, we aim to look at the dataset provided on Protein Databank (PDB), which contains information about many proteins, DNA, RNA, and DNA-RNA hybrid sequences. We aim to focus on some key questions, such as identifying trends in the reported size, structure resolution, identification methods, and binding affinity of reported ligands for DNA and RNA sequences. We aim to look at the size distribution of the reported DNA and RNA sequences and sort them by the date they were first reported. We also want to look at the binding affinity of the DNA and RNA sequences to their reported ligands and compare them with each other as well the proteins as a whole as well as proteins with the same ligands.