**Prediction of protein-to-ligand binding using Machine Learning**

**Drug Discovery Technical Challenge**

**Author:** Anatoly Buchin

**Introduction:** Protein-ligand binding prediction is a foundational component of drug discovery, enabling estimation of potential complex formation between proteins and ligands based on their molecular structures. In this project, we use the *Deloitte Drug Discovery Dataset* (Deloitte\_Drug\_Discovery\_dataset.csv), containing pairs of proteins from [UniProt](https://www.uniprot.org/) and ligands from [PubChem](https://pubchem.ncbi.nlm.nih.gov/), along with experimentally measured KIBA scores [[Yamanishi et al., 2010](https://www.google.com/url?q=https://academic.oup.com/bioinformatics/article/26/12/i246/282050?login%3Dfalse&sa=D&source=docs&ust=1731291103917879&usg=AOvVaw2-AZbaZRQIi1tvRK6MaC53)]. The KIBA score serves as an integrative metric of binding affinity, combining inhibition constant, dissociation constant, and half-maximal inhibitory concentration. The objective of this project is to classify the binding potential of protein-ligand pairs based on their structural information and corresponding KIBA scores measured experimentally.

**Methods:** Data was curated from UniProt and PubChem identifiers for each protein-ligand pair. A foundational model approach was applied to describe proteins, using the [Protein-BERT](https://huggingface.co/Rostlab/prot_bert) model to extract 1024-dimensional embeddings for proteins, representing key structural and functional features. Ligand data from PubChem included [molecular fingerprints](https://www.sciencedirect.com/science/article/abs/pii/S135964462200349X#:~:text=Molecular%20fingerprints%20are%20used%20to,a%20low%20computational%20cost%20way.) [Yang et al., 2022] and chemical properties such as molecular weight and polar surface area. The dataset was further augmented with synthetic protein-ligand pairs through shuffling to enhance classification accuracy. These features were combined into feature vectors for model training and testing. Several machine learning algorithms were assessed, including Random Forest, XGBoost, Logistic Regression, and a Fully Connected Neural Network.

**Results:** Extensive experiments were conducted to identify the model with optimal predictive performance. Among the evaluated models, Random Forest demonstrated the highest predictive accuracy, with an average precision of 93% and a recall score of 93%. The second best model was XGBoost with 81% of average accuracy on the test set. Random Forest was selected as the best predictor for binding affinity in this study. Performance metrics from all models were recorded to compare the model between each other. When examining Random Forest weights we found that chemical properties of ligands were more predictive than protein properties generated by Protein-Bert embeddings.

**Conclusions:** The Random Forest model trained on protein embeddings and ligand representations is the most robust approach for binding affinity prediction, outperforming the other models. We found that Random Forest and XGBoost are able to predict binding between given proteins and ligands. However the other algorithms, such as Logistic Regression and 2 layer Neural Network were not capable of predicting the binding affinity higher than random chance. We suggest that data pre-processing (scaling) and improvements in regularizations could further improve the model performance.

**Future directions:**

This study identified several areas for improvement in the modeling process:

1. Implementing flexible data and batch processing for model training could allow processing of the entire dataset (~2 million samples) without excessive memory use to improve the speed of training.
2. Exploration of feature engineering and interaction analysis between protein and ligand descriptors could help clarify which molecular features most strongly influence binding properties. Experiments: removing / adding features and analysis of model performance.
3. Optimizing the inference speed of foundational protein models, such as Prot-BERT, could streamline binding affinity prediction for novel proteins.
4. Extending model training duration and dataset size may further improve model performance.