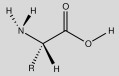
Merck: NLP on Protein Sequences

The Data Mine Corporate Partners Symposium 2024

Introduction

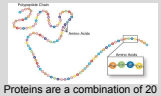


Description: Researched the descriptors of 137 antibodies in advanced clinical stages of drug development through data analysis, bioinformatics and libraries/packages.

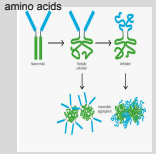
Goal: To find correlation between 1200+ protein descriptor columns and to stabilize proteins and extend shelf life.

Vision: Leverage the latest in data analytics to inform/predict risk levels associated with biotherapeutic manufacturability.

Scientific Background

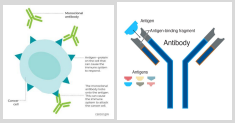


Proteins are a combination of 20 amino acids

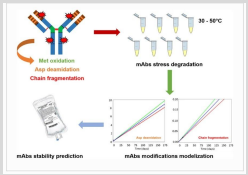


Antibody Aggregation can-

- Compromise Biological function
- Induce immune responses
- Evoke antibody clearance machinery in vivo



Antibodies are one of the proteins produced by the body. Used to treat different diseases.



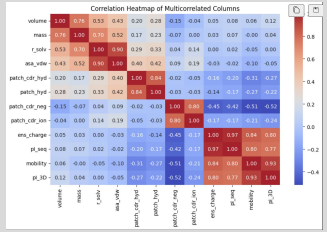
Reasons for aggregation:

- pH, temperature, storage buffer
- Protein sequences
- Hydrophobicity

Data Preprocessing and Multicollinearity

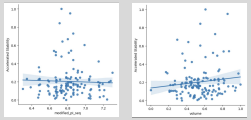
Prior to the implementation to Regression and Modeling, we had to:

- Remove Rilotumumab and other antibody outliers that are outside of our range (implemented Tukey's Fence)
- Used AbLang to extract over 800 descriptors that could be used for stability correlation analysis.
- Scaled the data using Min-Max Scaler through the sklearn preprocessing module
- Changed data types of the columns to fit appropriate scientific context
- Removed all NaN and insignificant columns to our goals
- Removed Variables that are Multicollinear variables
- Running a Correlation Matrix of all the columns



Modeling

Noted Linear Regression was underfitting, so we decided to explore more complex models

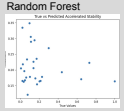


We implemented a Cross Validation Test on models we believed would best represent the data

```
Random Forest: Mean RMSE: 0.1826, Standard Deviation RMSE: 0.0347
XGBoost: Mean RMSE: 0.1969, Standard Deviation RMSE: 0.0499
Support Vector Machines: Mean RMSE: 0.1085, Standard Deviation RMSE: 0.0334
Neural Network: Mean RMSE: 110.7940, Standard Deviation RMSE: 29.1517
```

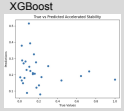
In addition we conducted all these tests individually and got their corresponding r squared as well as a scatter plot.

Random Forest



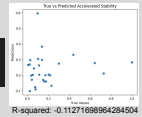
R-squared: -0.148743952327

XGBoost



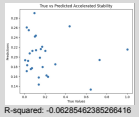
R-squared: -0.343586798031737 1

SVM



R-squared: -0.112716989964284504

Lasso Regression



R-squared: -0.0628546238526641 6

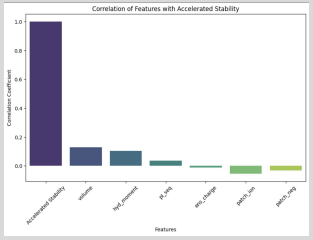
Conclusion:

The findings lay a solid foundation for further research and development efforts. Made significant progress towards impacting drug development by identifying meaningful relationships between specific variables and descriptors

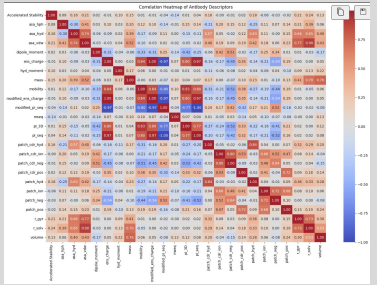
Future Goals:

- Find a correlation after the pre processing
- Make progress to impacting drug development
- Find meaningful relationships between specific variables and descriptors

Visualizations



Accelerated Stability in comparison with Target Columns



Complete Correlation Matrix derived from the MOE dataset

Sneha Subramanian, Shreya Ramamurthy, Snehal Adabala, Kihunn Anderson, Molik Mishra

Acknowledgements: Dr. Daniel Skomski, Dr. Rojan Shrestha, Dr. Deeptak Verma, Dr. Ruomeng Qiu, Dr. Katherine Williams
Data Mine: Nicholas Rosenorn, Dr. Mark Daniel Ward