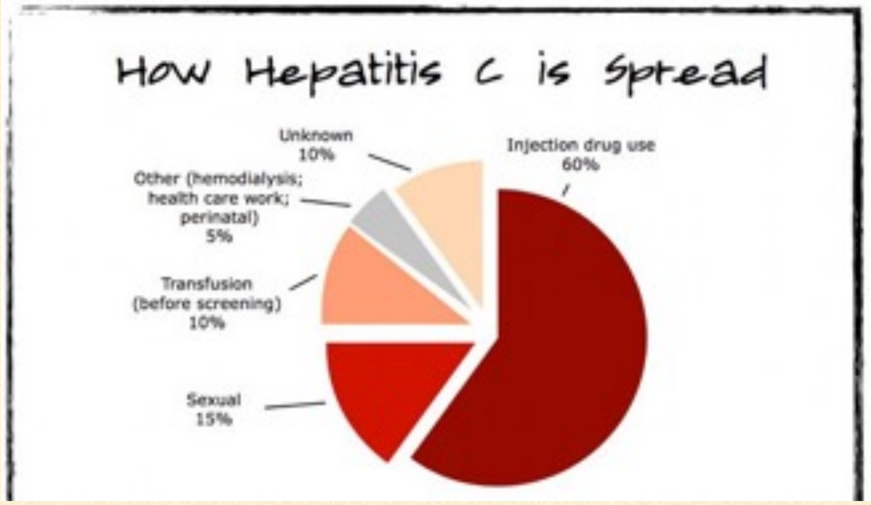
INTRO TO PREDICTIVE MODELING FOR HEPATITIS-C

COBB2010: Foundations of Computational Biology
Name: Kamaru-deen Lawal
Email: kal246@pitt.edu

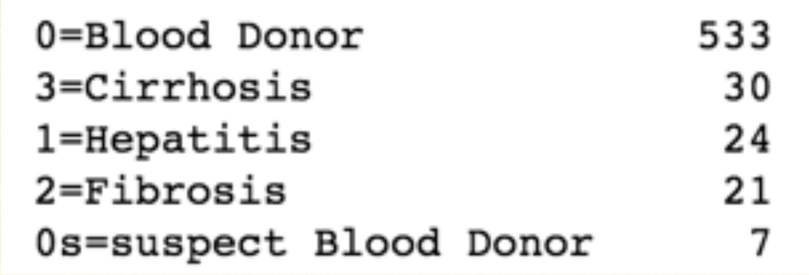
University of Pittsburgh
Department of Medicine

INTRODUCTION

- Hepatitis C is a liver infection spread through contact with blood from an infected person
- Most people become infected by sharing needles or other equipment used to prepare and inject drugs. It can also be transmitted sexually
- The disease can also be transmitted via unprotected sex (especially when blood is present)
- For some, Hepatitis C causes nothing more than a short illness, but for others it can result in serious problems such as cirrhosis and liver cancer
- Goal of project is to build interpretable models that can aid in predicting the presence of Hepatitis C



DATA BACKGROUND

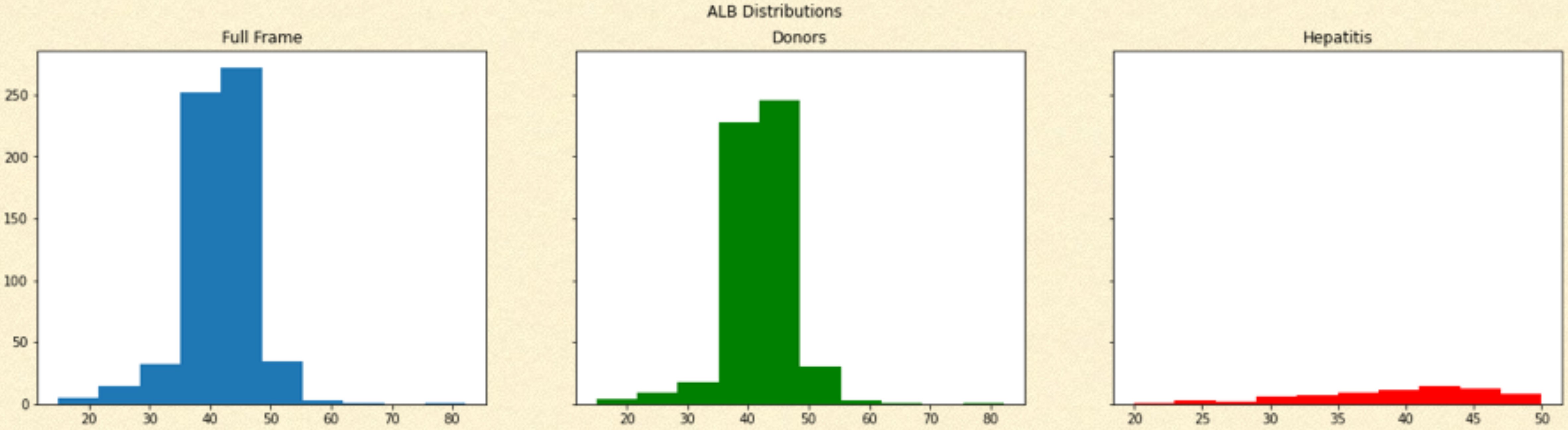


```
0=Blood Donor              533
3=Cirrhosis                 30
1=Hepatitis                 24
2=Fibrosis                  21
0s=suspect Blood Donor       7
```

- Dataset has twelve features
    - 11 Quantitative (Age, ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, PROT)
    - 1 Categorical (Gender)
- Dataset has one multi-class target representing level of Hepatitis C present for patient
- Each patient can have a Hepatitis C rating of 0 (blood donor), 0s (suspected blood donor), 1 (Hepatitis), 2 (Fibrosis), or 3 (Cirrhosis)
- The dataset was extremely imbalanced. Some levels occurred fewer than 10 times in the dataset. As a result we examine a binarized classification problem instead of the original multi-class problem
- In the following slides we examine some of the key features we will use to build our model

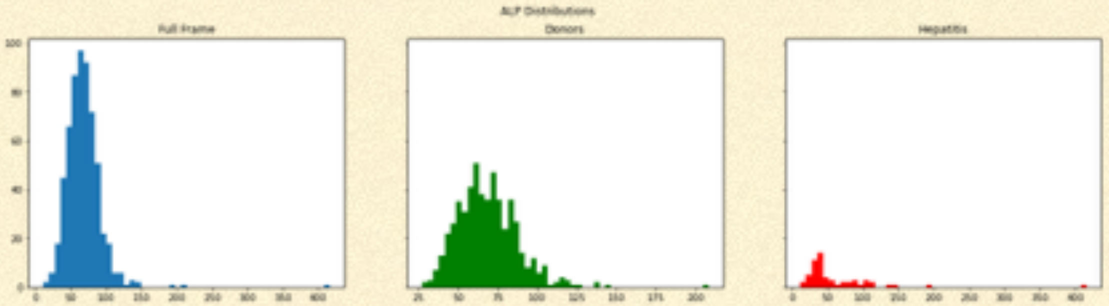
ALBUMIN (ALB)

- Globular proteins commonly found in blood plasma
- They differ from other proteins in that they are not glycosylated
- Conclusion: Blood Donors have a few more outliers for albumin. It also seems like Hepatitis C carriers have more uniformly distributed albumin values

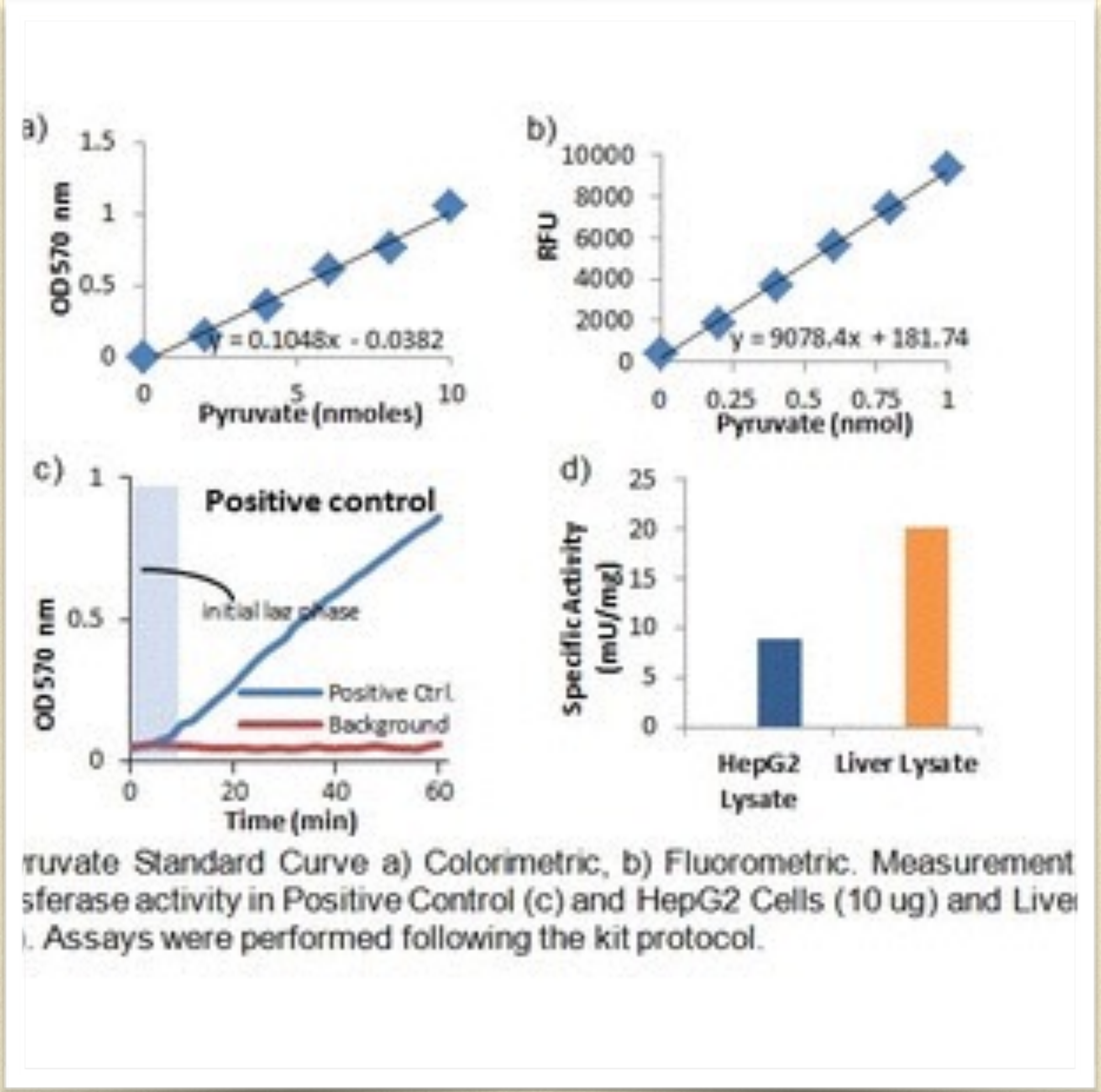


ALKALINE PHOSPHATASE (ALP)

- In humans, it is present in all tissues throughout the body
- Diagnosticians commonly use it as a biomarker to search for the presence of Hepatitis and Osteomalacia
- Conclusion: There is a higher degree of variability for ALP amongst Hepatitis C carriers



ALANINE TRANSAMINASE (ALT)



ruvate Standard Curve a) Colorimetric, b) Fluorometric. Measurement
sferase activity in Positive Control (c) and HepG2 Cells (10 ug) and Live
. Assays were performed following the kit protocol.

- Commonly measured clinically as part of liver function tests
- Significantly elevated levels often indicate some sort of medical problem
- For years, the Red Cross used ALT testing to ensure the safety of its bloody supply
- Conclusion: Both Blood Donors and Hepatitis C carriers have outliers

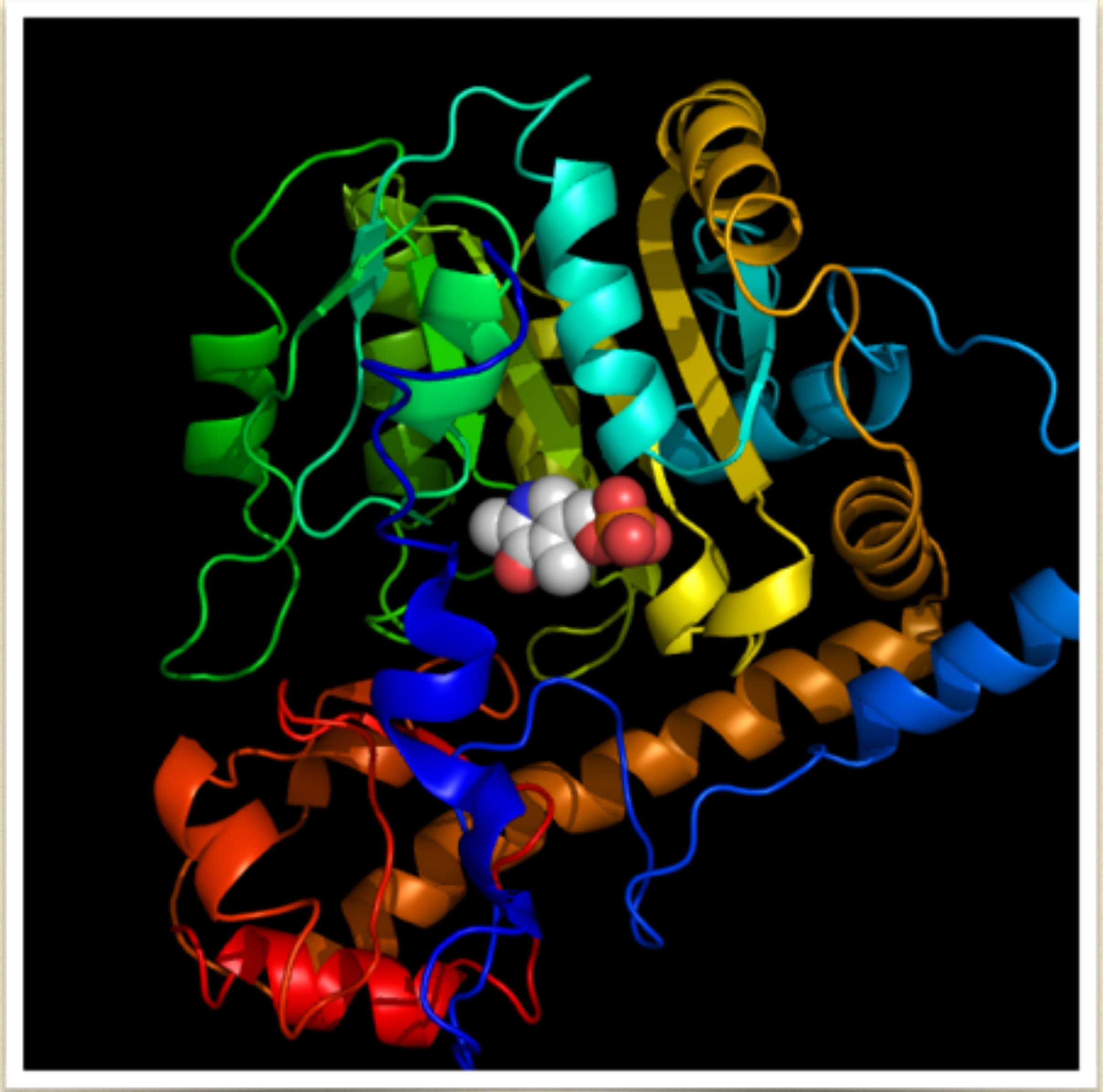
ASPARTATE TRANSAMINASE(AST)

- Similar to ALT in that both enzymes are associated with liver parenchymal cells

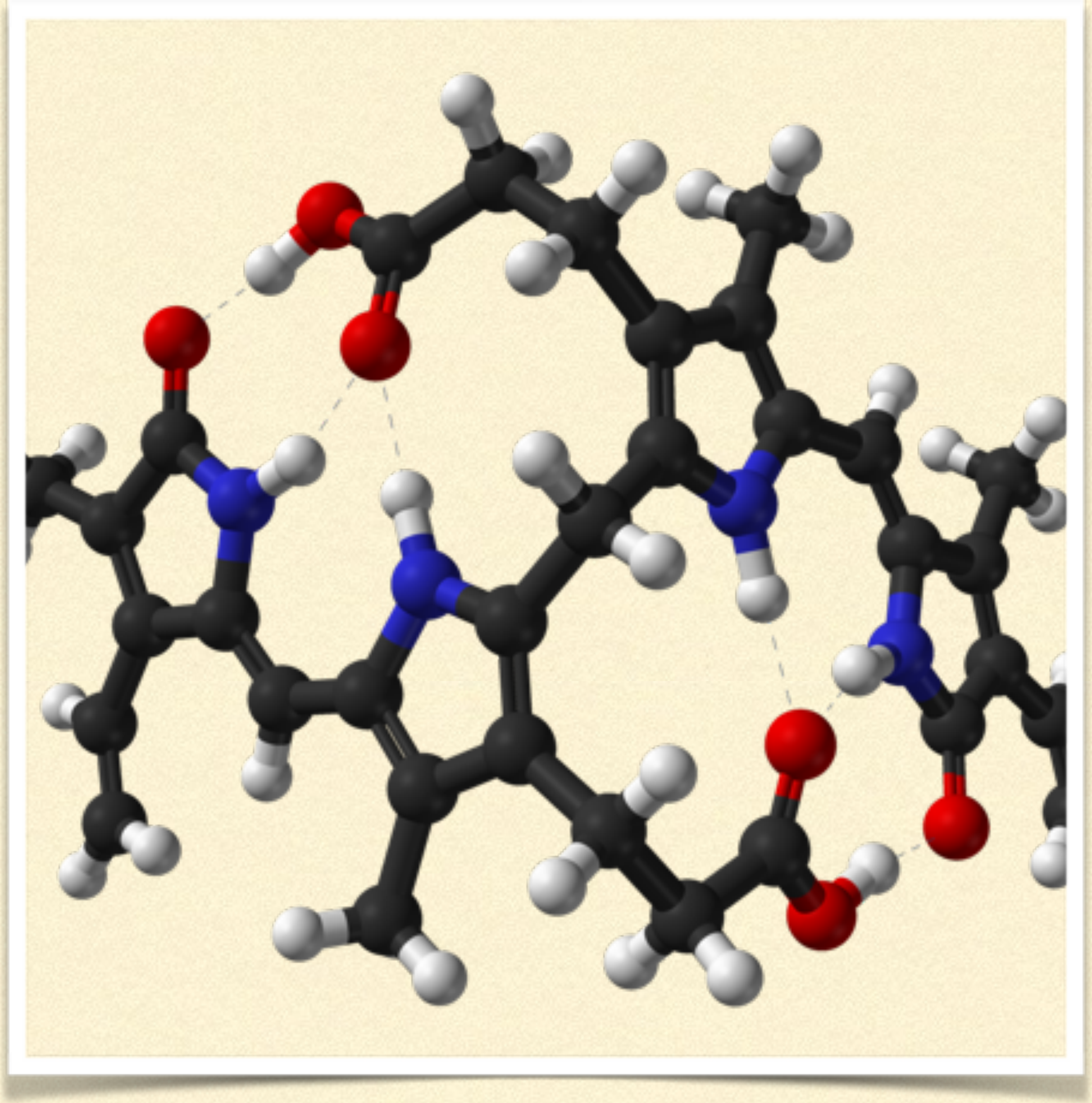
- Difference is ALT is found mainly in the liver, while AST is found in the liver, heart, skeletal muscles, kidneys, brain, and red blood cells

- Important to note source of AST may reflect pathology in organs other than the liver

- Conclusion: Similarly to ALP, there is a significantly higher degree of variability for AST amongst Hepatitis C carriers



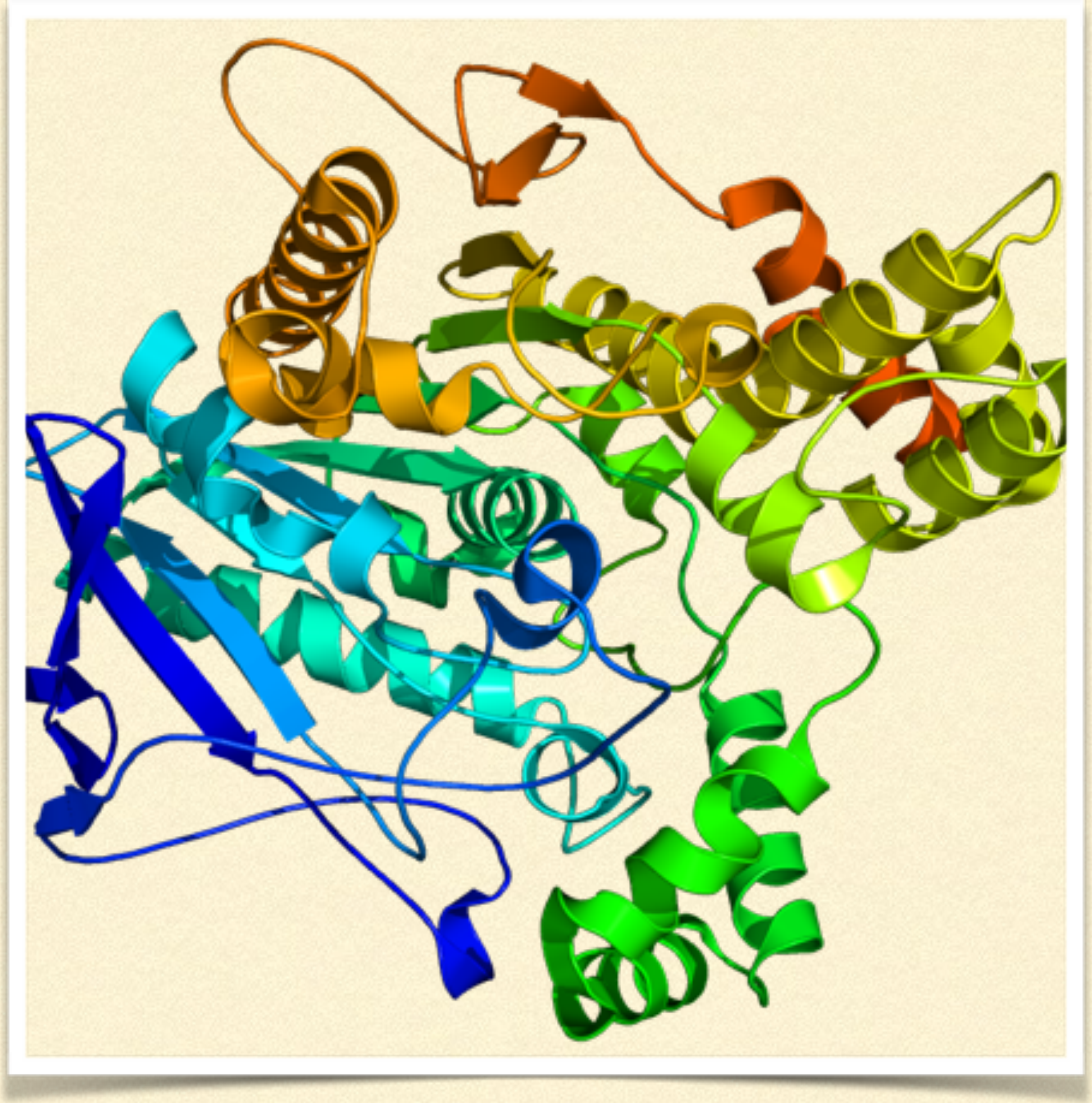
BILIRUBIN (BIL)



- BIL levels in the body represent the balance between production and excretion
- Not usually detected in the urine of healthy people. If blood level of conjugated bilirubin becomes elevated, excess is excreted in the blood
- Conclusion: Mean BIL is 282% higher for Hepatitis C carriers. An indicator variable based on BIL could be useful if it can be constructed in a way that doesn't result in a near zero variance feature (see appendix for more information on NZV features)

CHOLINESTERASE (CHE)

- Levels may be reduced in patients with advanced liver disease
- Conclusion: Both Blood Donors and Hepatitis C carriers follow roughly the same distribution for CHE



CHOLESTEROL (CHOL)

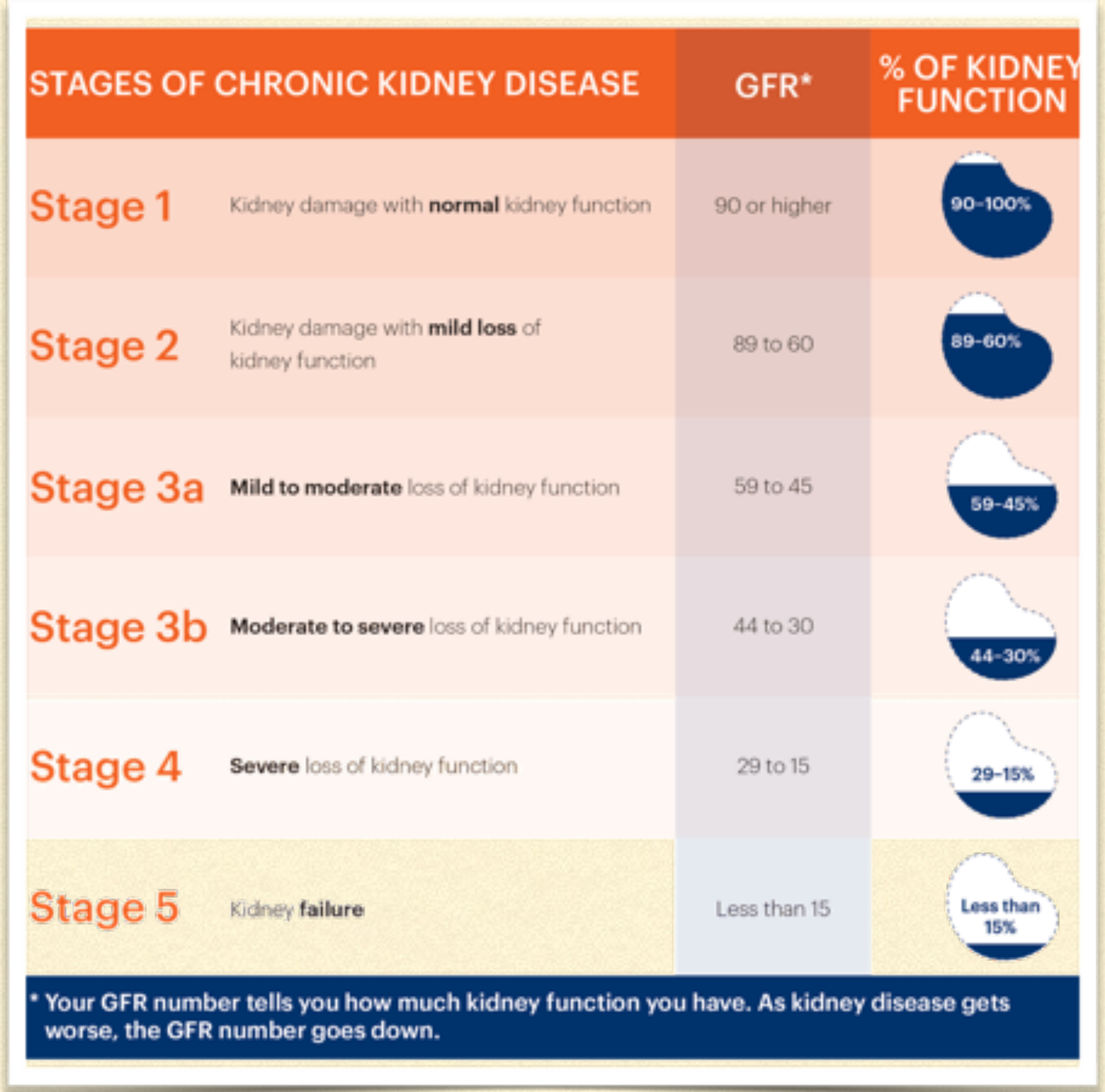
- Lipid profiles are a panel of blood tests commonly used as screening tools for abnormalities in lipids, such as cholesterol and triglycerides

- Results can identify genetic diseases along with approximate risks for cardiovascular diseases and and certain forms of pancreatic diseases

- Conclusion: Both Blood Donors and Hepatitis C carriers follow roughly the same distribution for CHOL.

CREATININE (CREA)

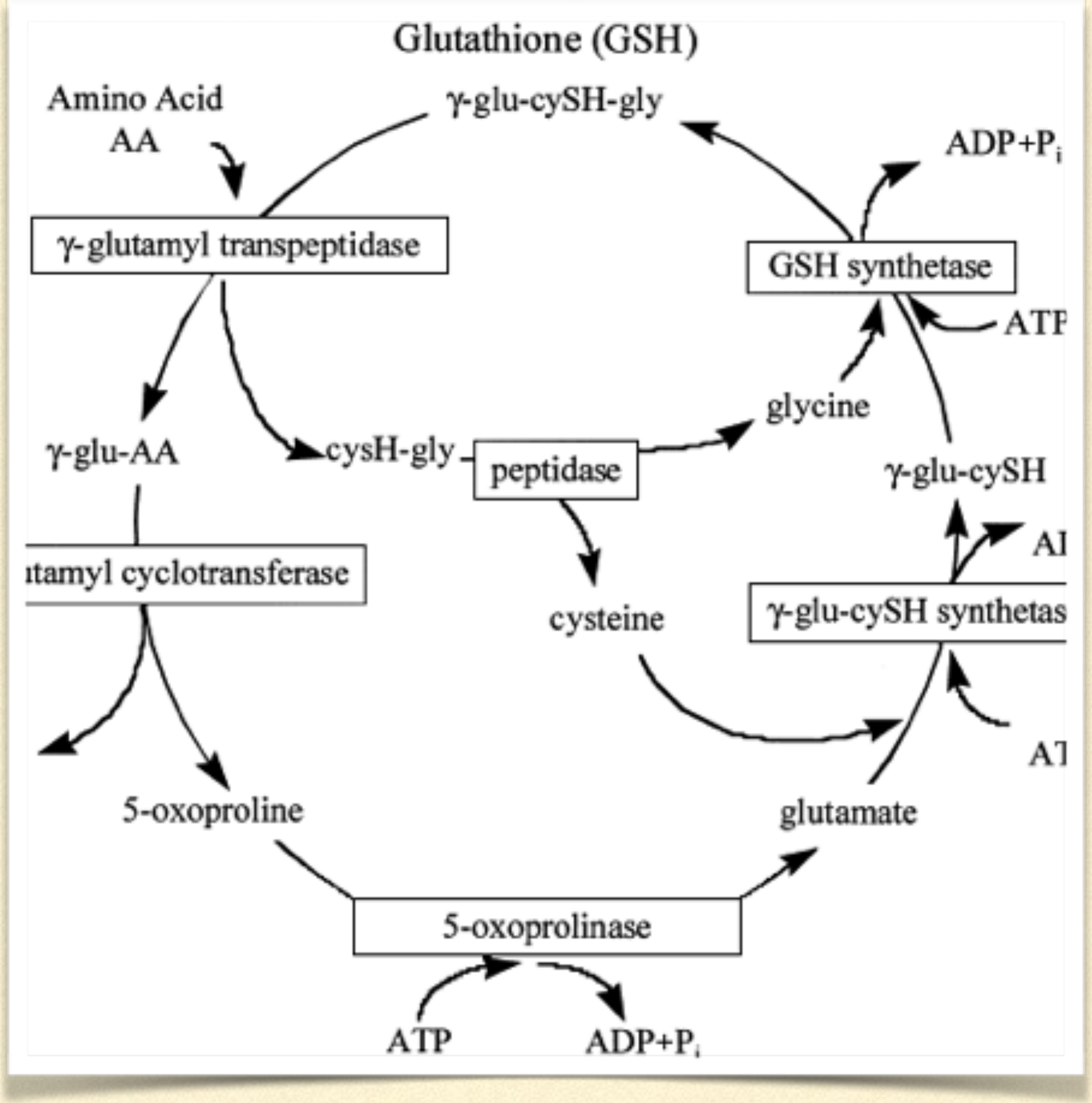
- Most commonly used indicator of renal function
- Elevated creatinine is not always representative of a true reduction in GFR. A high reading may be due to increased production not due to decreased kidney function
- Worked on a project at Enterprises where Creatinine was the key featured used to create a target for the dataset
- Conclusion: Hepatitis C carriers have a significantly higher mean and variance for CREA.



STAGES OF CHRONIC KIDNEY DISEASE		GFR*	% OF KIDNEY FUNCTION
Stage 1	Kidney damage with **normal** kidney function	90 or higher	90-100%
Stage 2	Kidney damage with **mild loss** of kidney function	89 to 60	89-60%
Stage 3a	**Mild to moderate** loss of kidney function	59 to 45	59-45%
Stage 3b	**Moderate to severe** loss of kidney function	44 to 30	44-30%
Stage 4	**Severe** loss of kidney function	29 to 15	29-15%
Stage 5	Kidney **failure**	Less than 15	Less than 15%

* Your GFR number tells you how much kidney function you have. As kidney disease gets worse, the GFR number goes down.

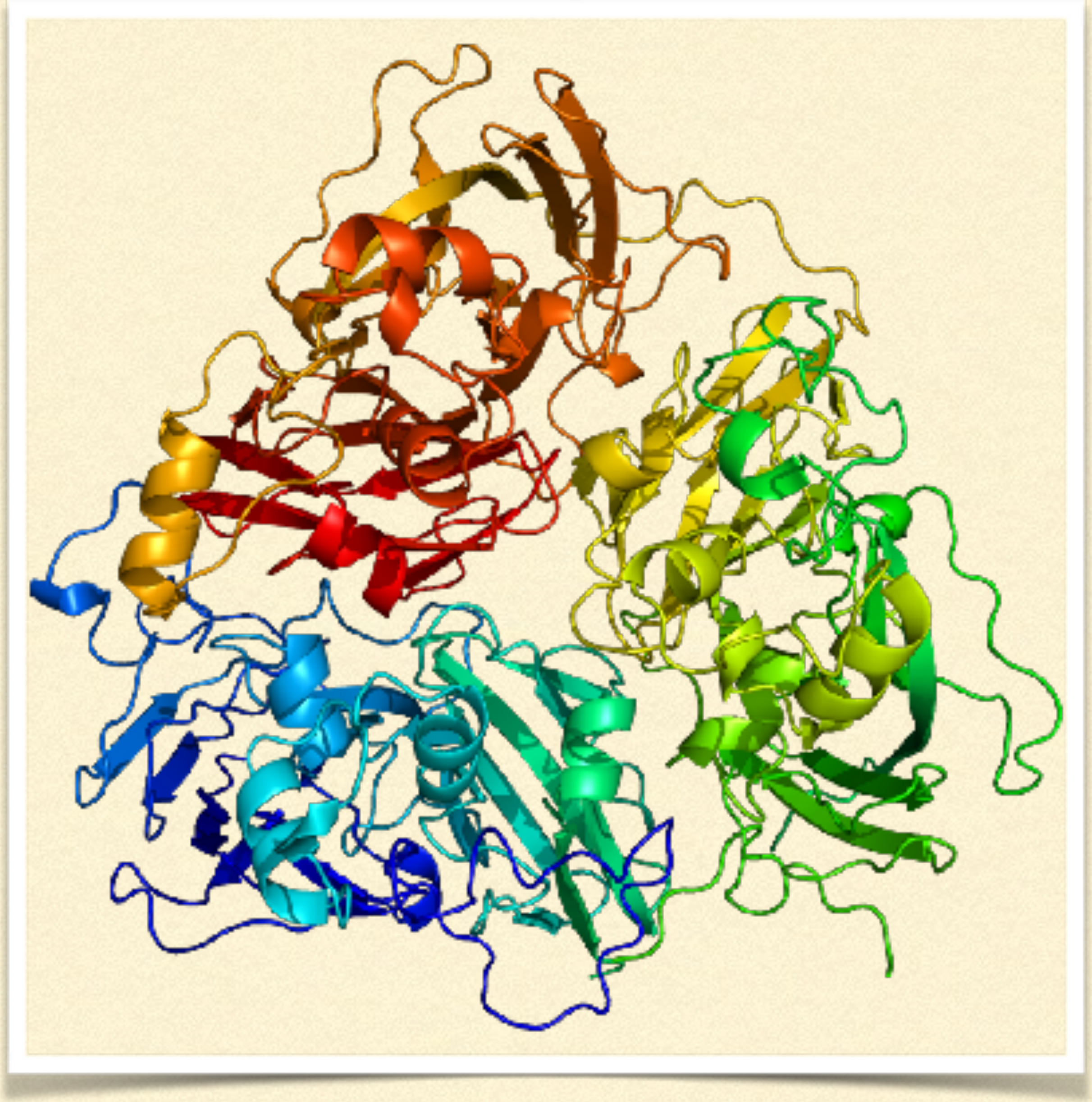
GAMMA-GLUTAMYLTRANSFERASE (GGT)



- Predominately used as a diagnostic marker for liver disease
- Latent elevations in GGT are typically seen in patients with chronic viral hepatitis
- Elevated levels can also be found in diseases of the biliary system and pancreas
- Conclusion: Hepatitis C carriers have a significanlty higher mean and variance for GGT.

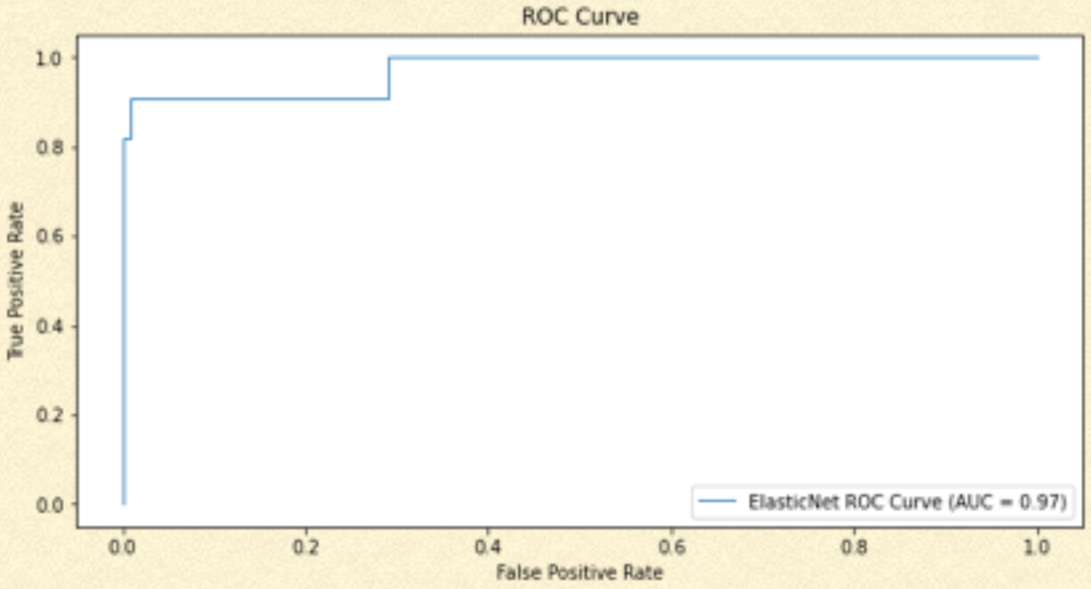
SERUM TOTAL PROTEIN (PROT)

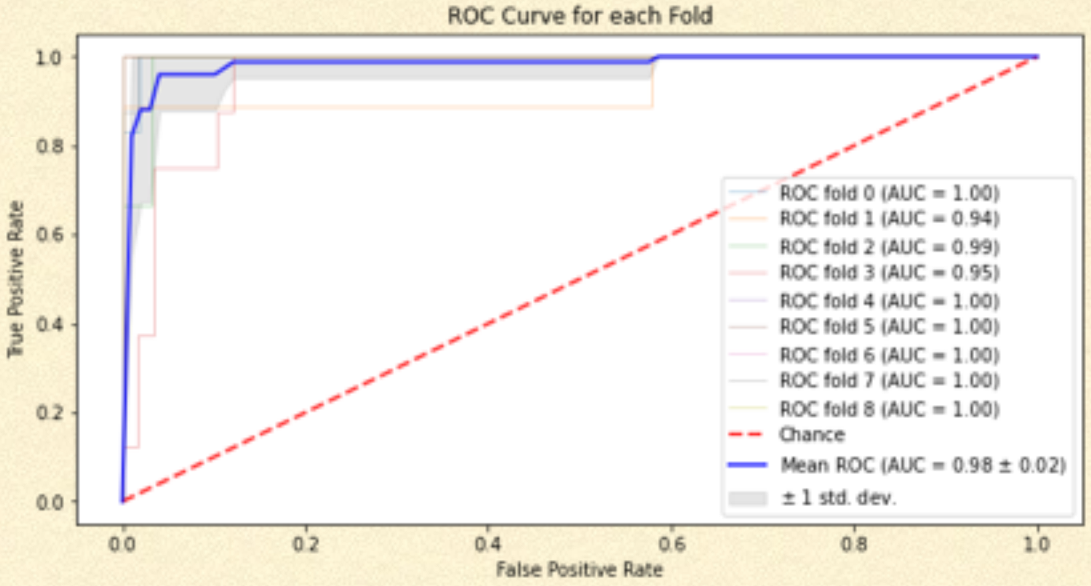
- Concentrations below the reference range usually reflect low albumin concentration
- Concentrations above usually reflect leukemia
- Conclusion: Both Blood Donors and Hepatitis C carriers follow roughly the same distribution for PROT



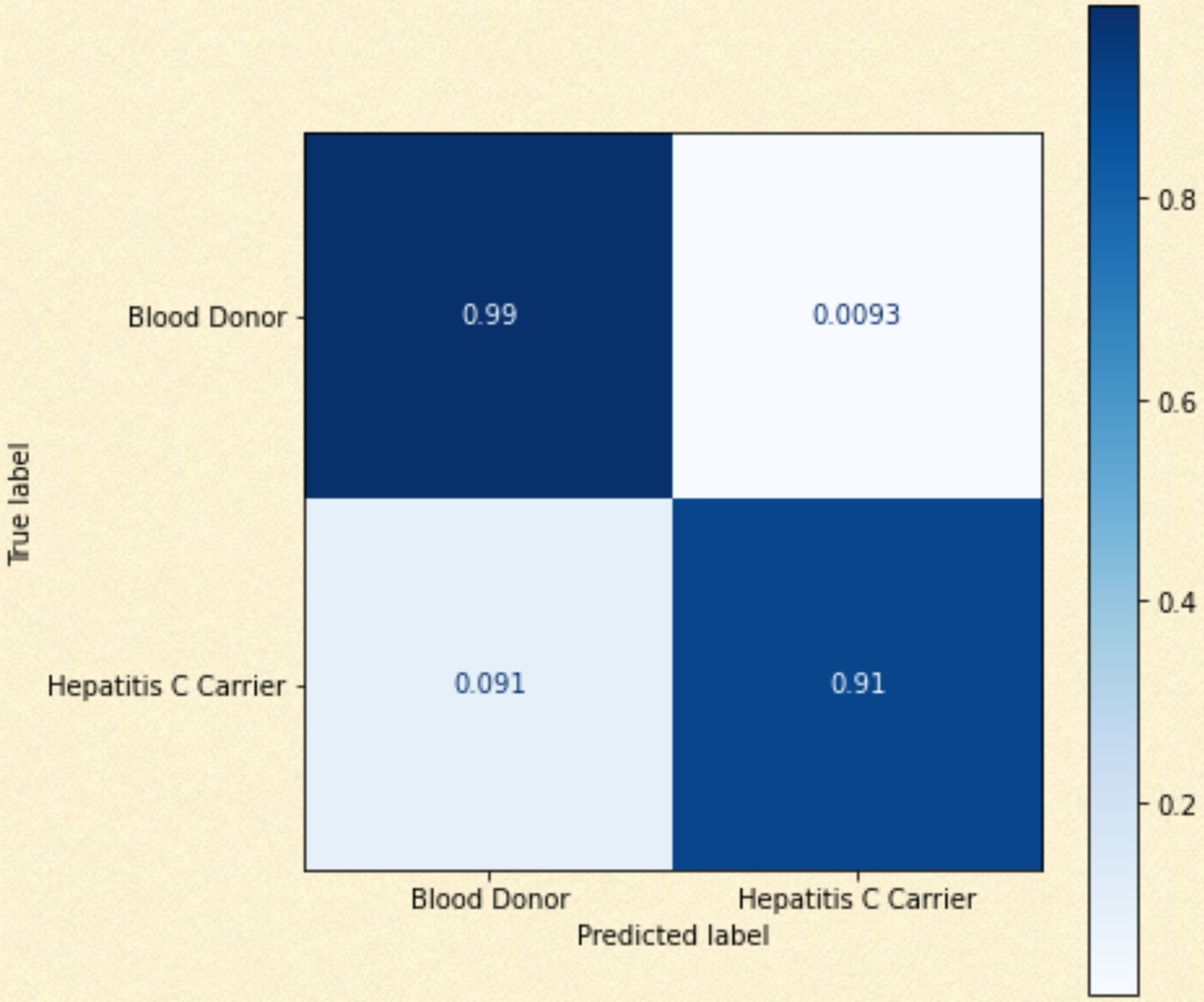
MODELING/EVALUATION TECHNIQUES

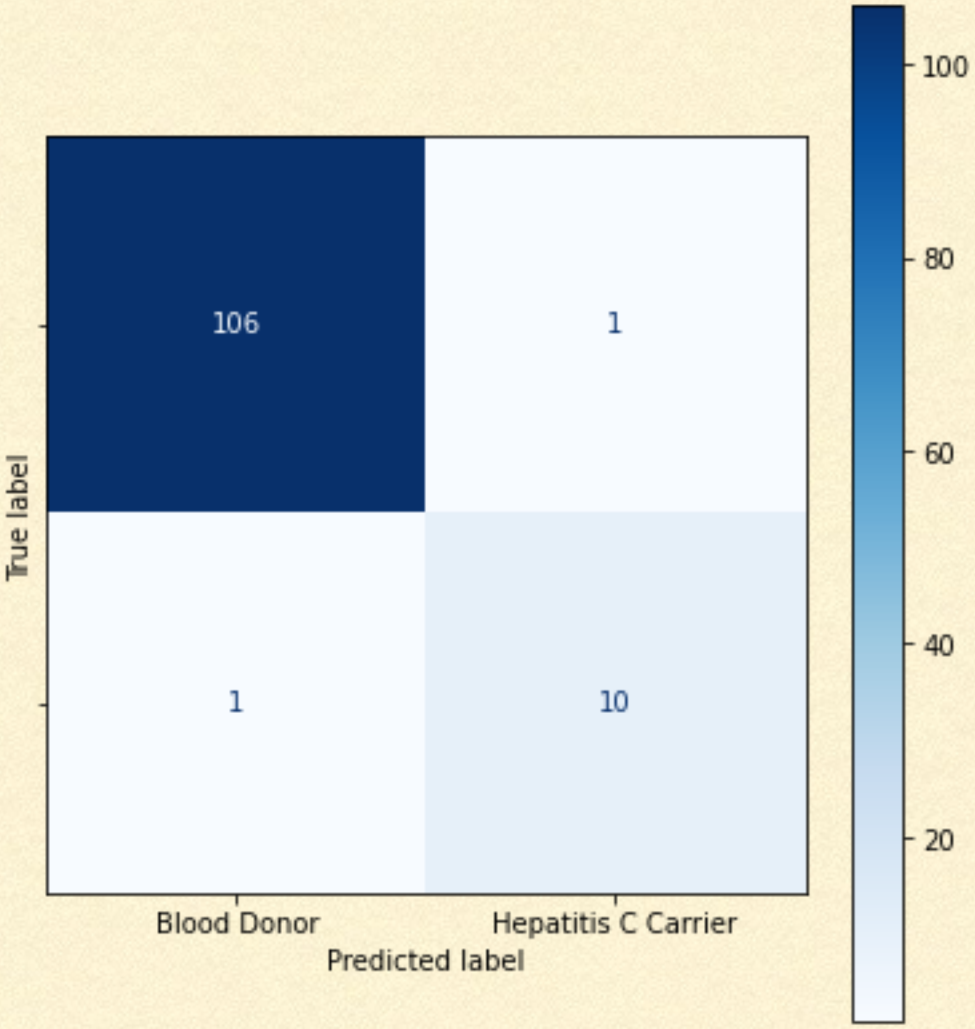
- Assessed performance of model using two different evaluation methods
  - 80/20 (Train/Test split)
  - k-Fold Cross Validation (k=9)
- Model: LogisticRegression w/ an ElasticNet penalty
- Achieved AUC of .97 for the 80/20 split
- Achieved Mean AUC of .98 for the k-Fold Cross Validation



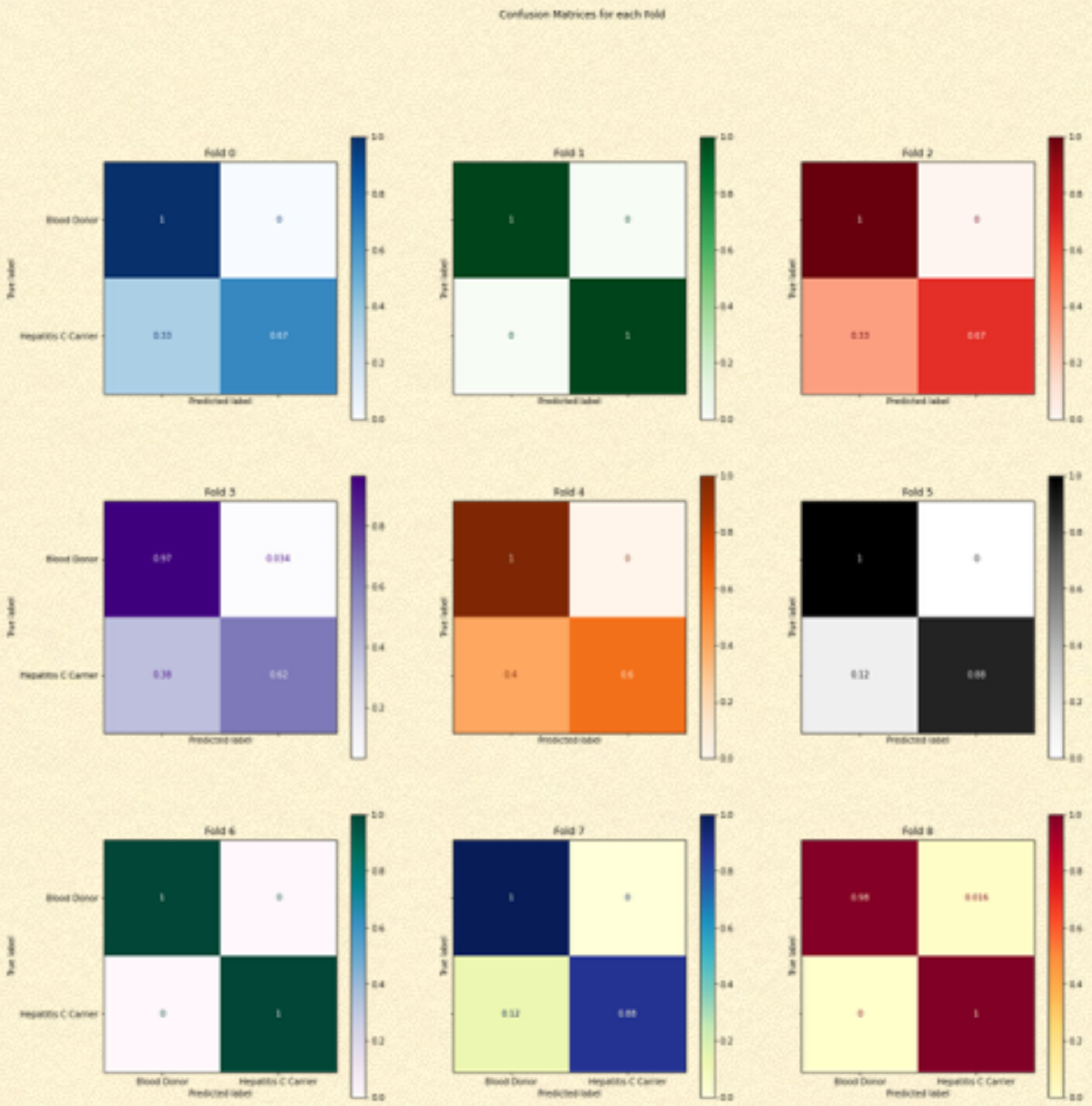


80/20 MODEL EVALUATION

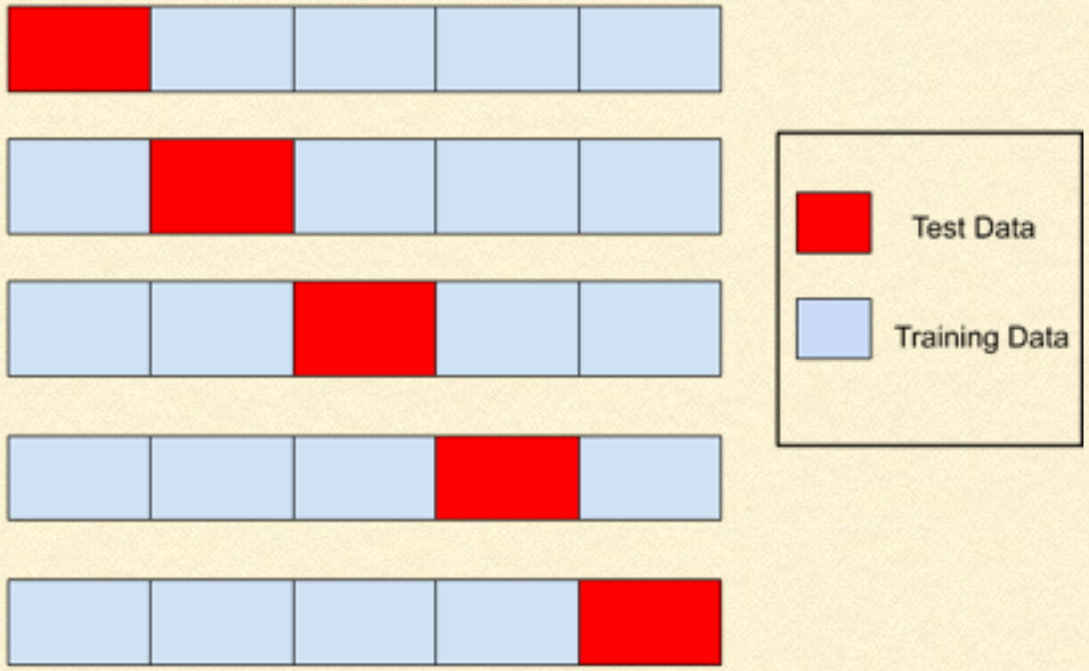




K-FOLD CV MODEL EVALUATION



K-FOLD CV EXPLAINED



- Used to select a model and give an approximation of the test error

- Randomly divide the data into K equal-sized parts. Train the model on K-1 of the parts, and use the K-th part to validate the model

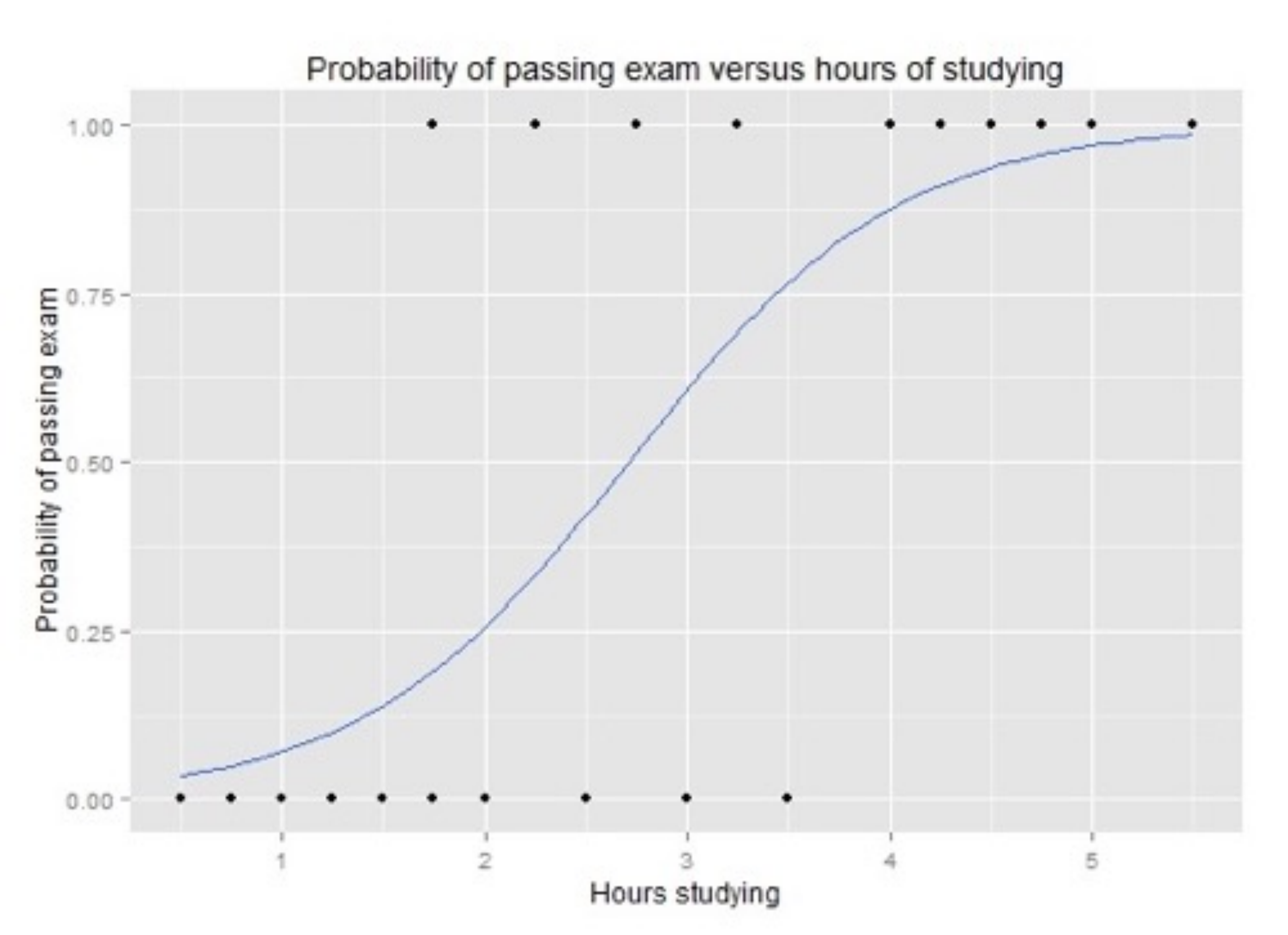
- We do this for each k = 1, …, K and then combine the results

- Method is particularly useful when data is limited

- Worth noting that choosing the right value for K is a bit of an art form as the bias-variance tradeoff ends up coming into play

LOGISTIC REGRESSION EXPLAINED

- Used to predict the probability of a binary event



ELASTIC-NET REGRESSION EXPLAINED

L1 Regularization

$$\text{Cost} = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2 + \lambda \sum_{j=0}^{M} |W_j|$$

L2 Regularization

$$\text{Cost} = \underbrace{\sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2}_{\text{Loss function}} + \underbrace{\lambda \sum_{j=0}^{M} W_j^2}_{\text{Regularization Term}}$$

- Despite being somewhat overly simplistic, linear models are often the model of choice because of their interpretability

- Shrinkage methods can be used to zero out weak predictors while in reducing variance

- Idea is to fit a model containing all features using a techniques that regularizes the coefficient estimates for each feature

- As with straight least squares, the goal of Elastic-Net Regression is to minimize RSS. But the summands with lambdas in front of them are minimized when the coefficients are close to 0

- Cross validation is often used to determine lambda

BIBLIOGRAPHY

- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.

- Kuhn, Max., and Kjell Johnson. Applied Predictive Modeling. New York: Springer, 2013.

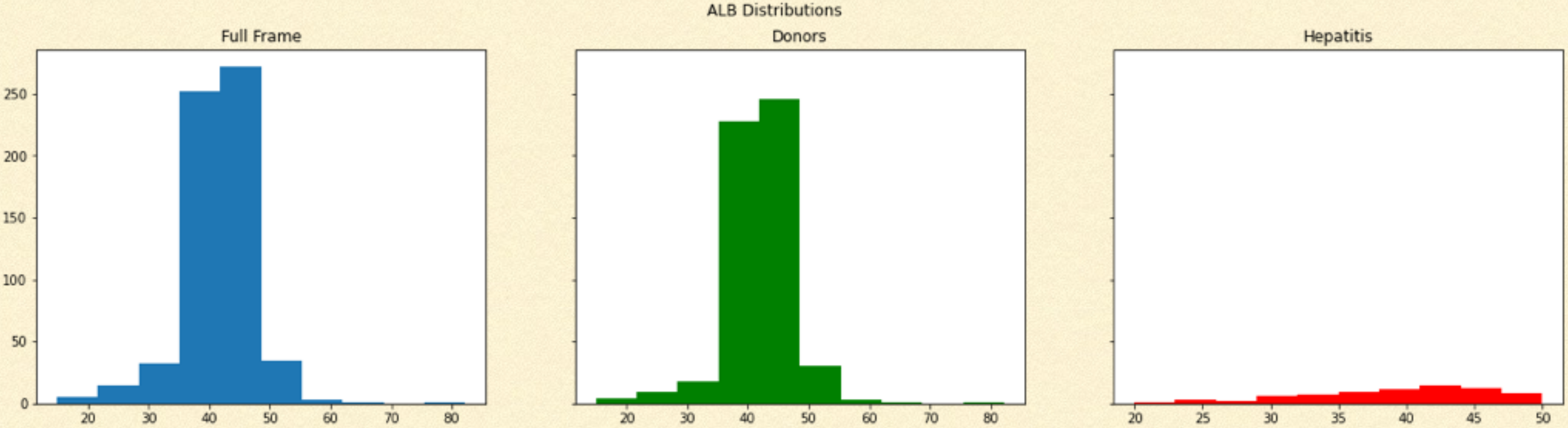
- Various images online for Data description

Appendix

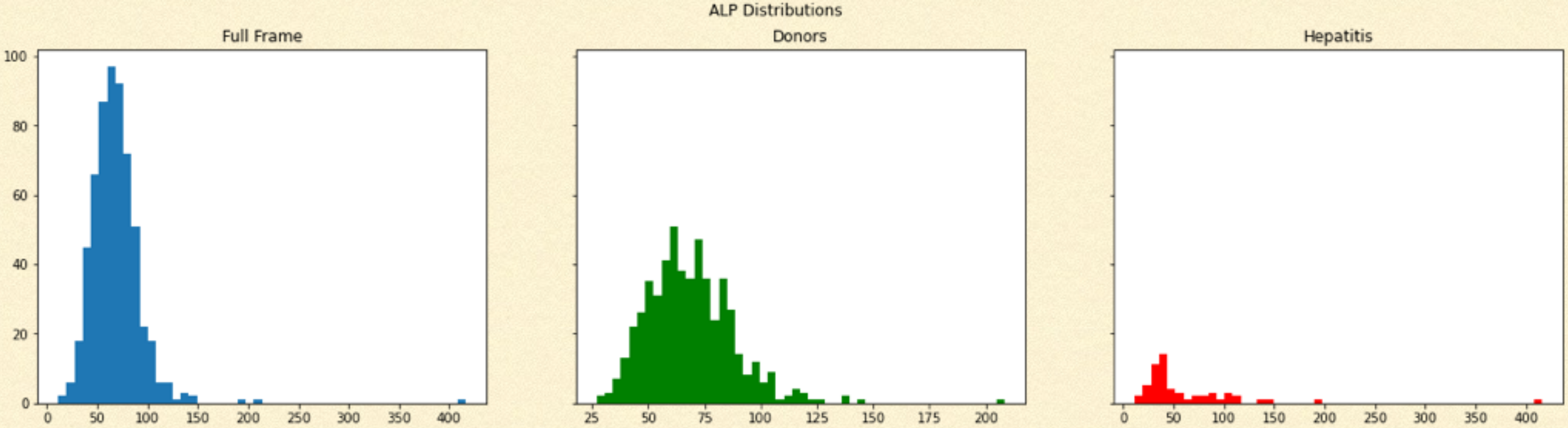
NEAR ZERO VARIANCE FEATURES

- Lots of potential advantages to removing unnecessary predictors
  - Fewer predictors => Lower computation time and complexity
  - Two predictors correlated => measuring same underlying property, removing one might make model more interpretable
- Good rule of thumb: if the fraction of unique values over the sample is low (<10%) AND the ratio of the 1st to 2nd most common value for the feature is large (>= 20), then the feature is near zero variance

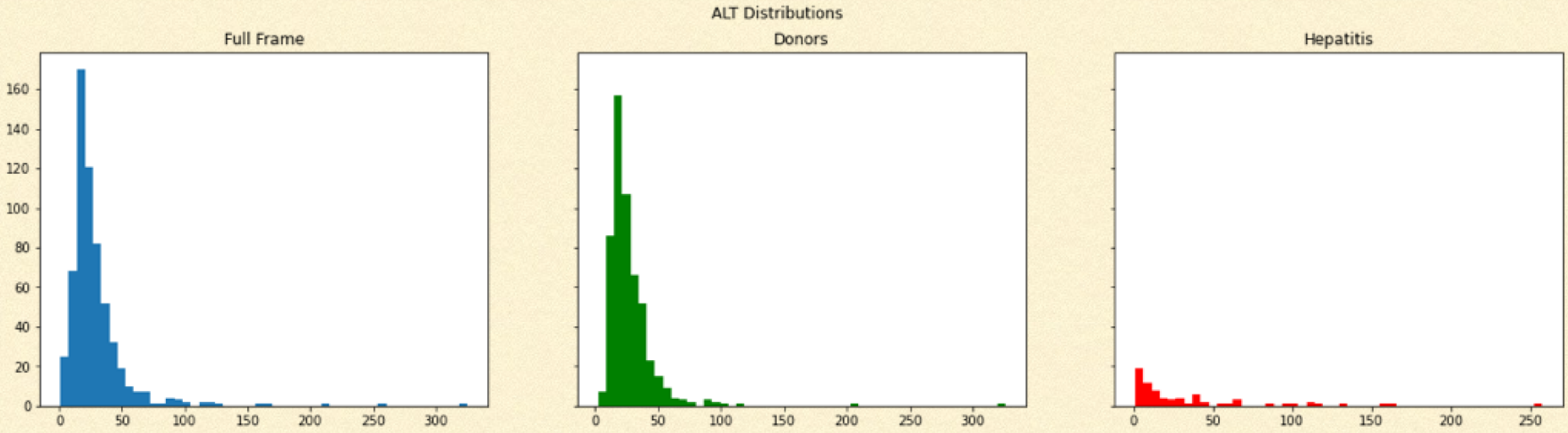
ALB DISTRIBUTION



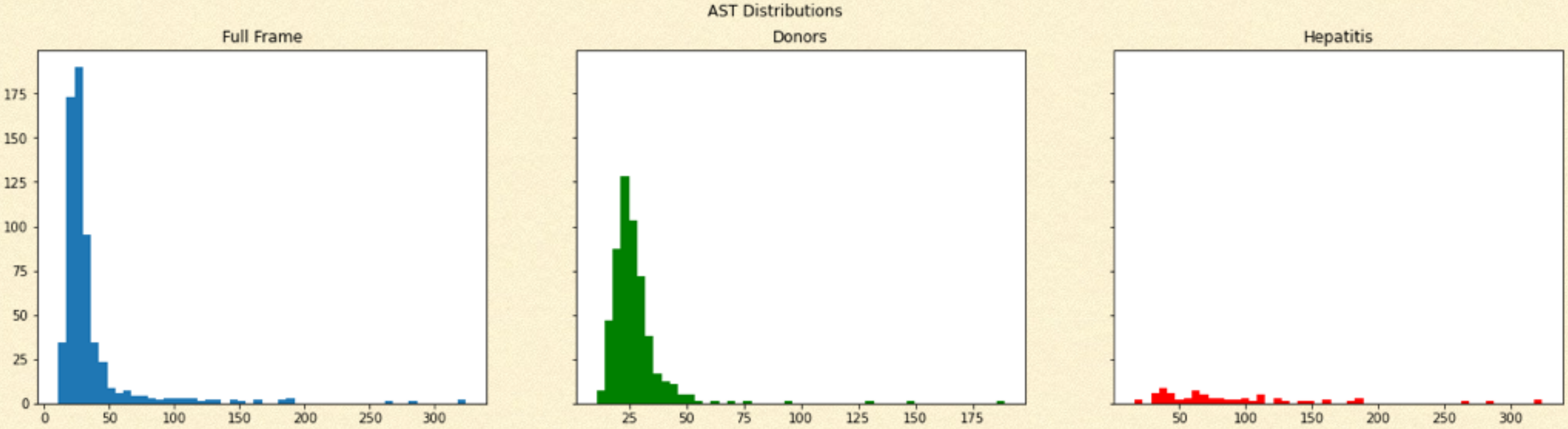
ALP DISTRIBUTION



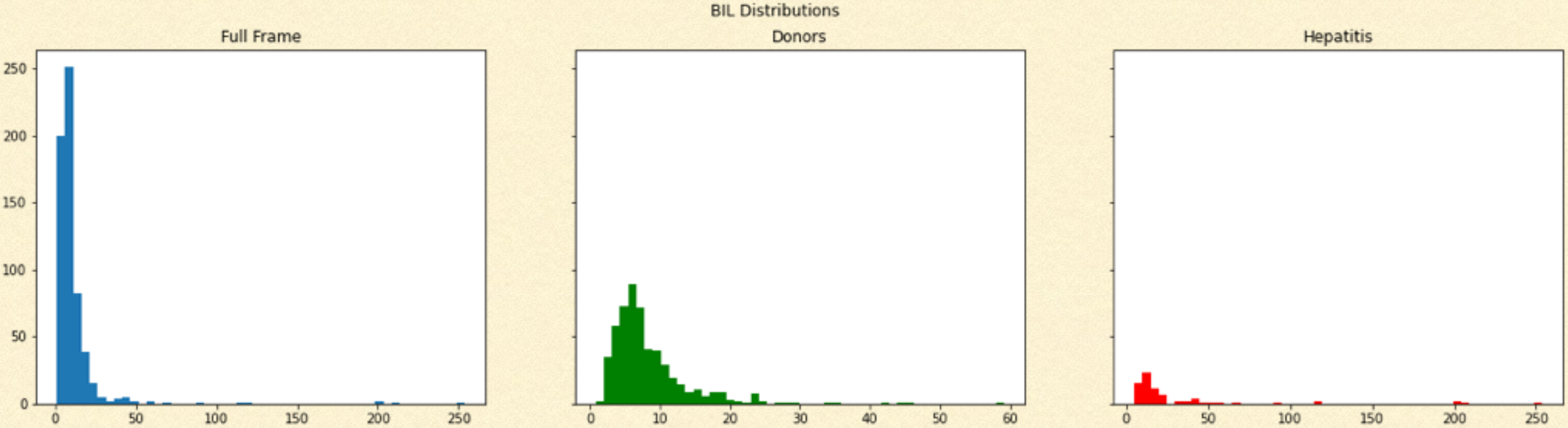
ALT DISTRIBUTION



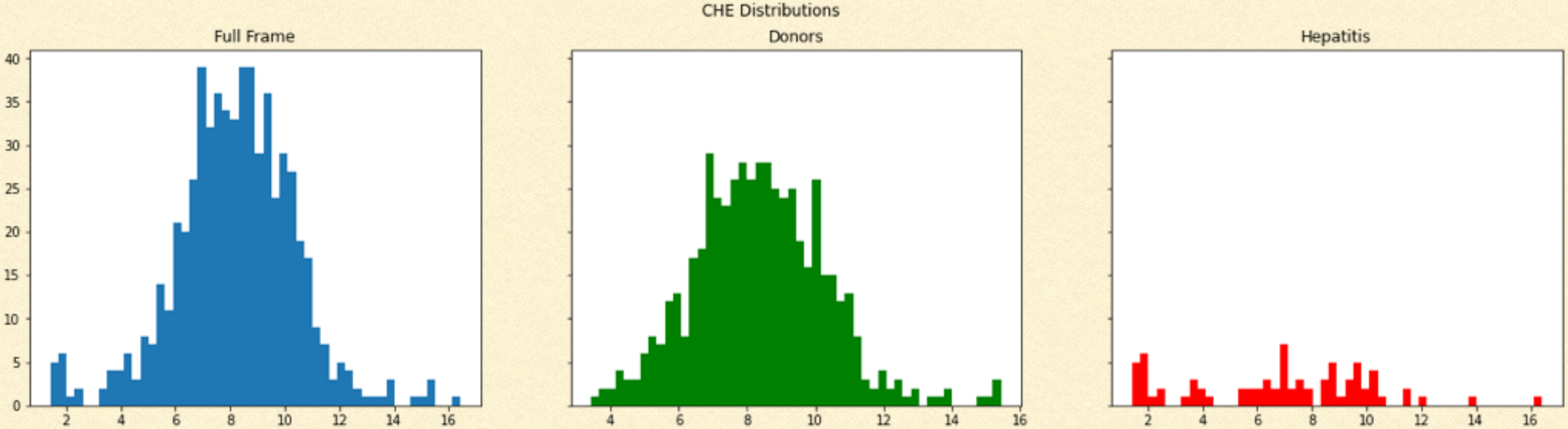
AST DISTRIBUTION



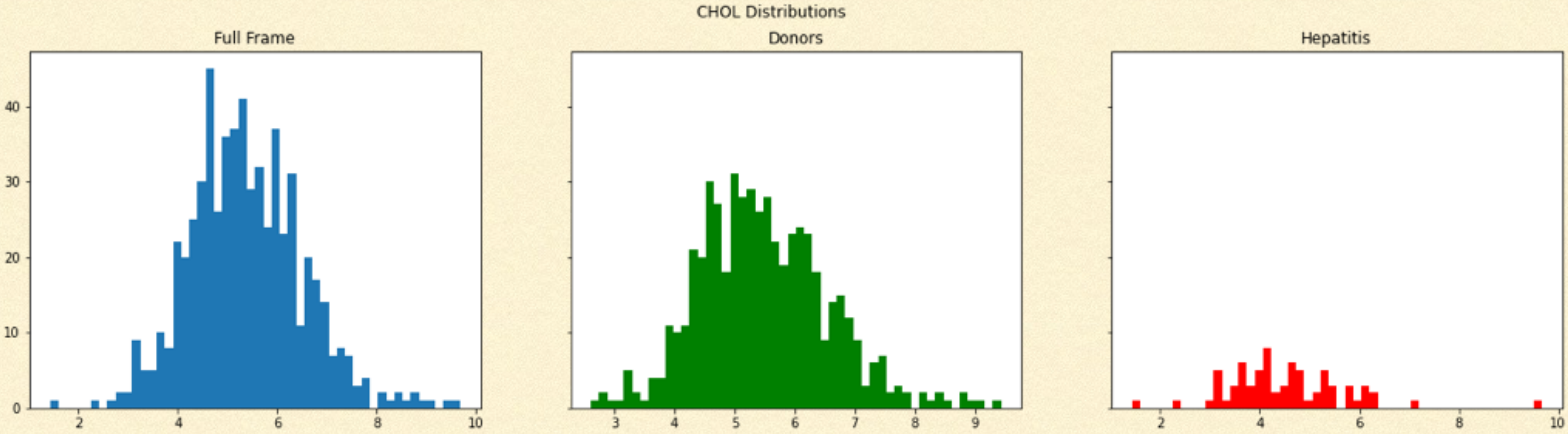
BIL DISTRIBUTION



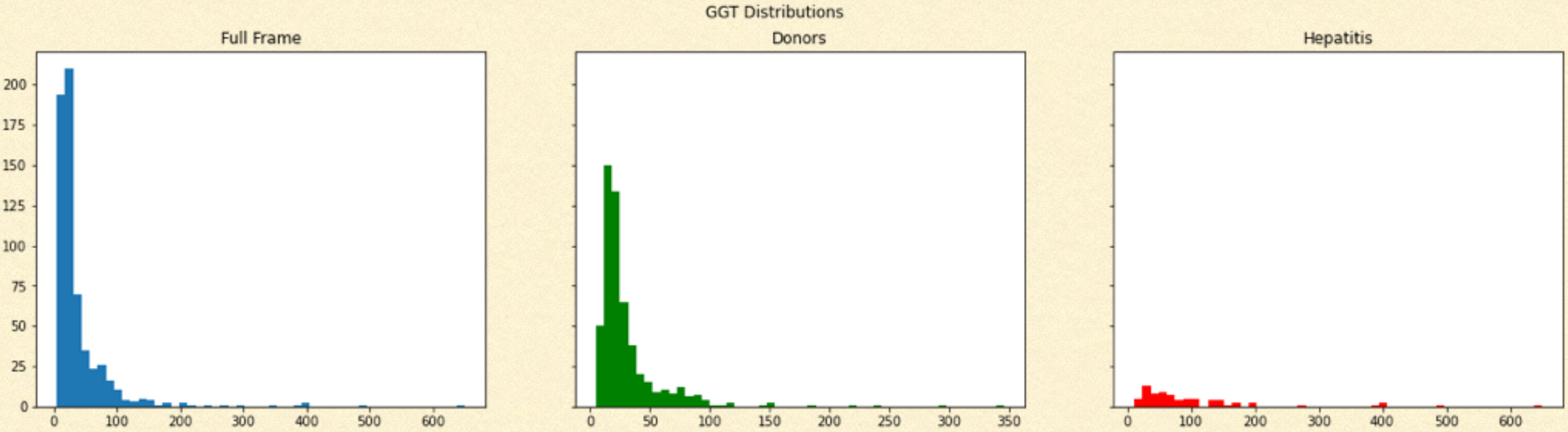
CHE DISTRIBUTION



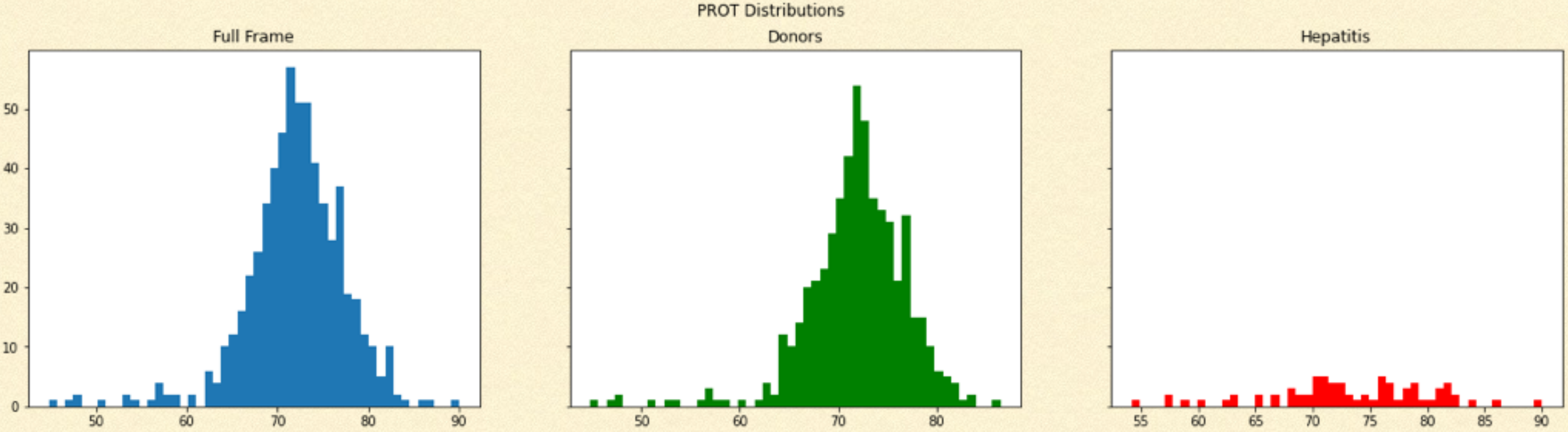
CHOL DISTRIBUTION



GGT DISTRIBUTION



PROT DISTRIBUTION


PROT Distributions
Full Frame
Donors
Hepatitis