Analyzing HLA Sequences to Predict Organ Rejection and Find Optimal Targets for Precise Immunosuppression

Samhitha Bodangi

Massachusetts Academy of Math and Science

Advanced STEM with Scientific and Technical Writing

Instructor: Kevin Crowthers, Ph.D.

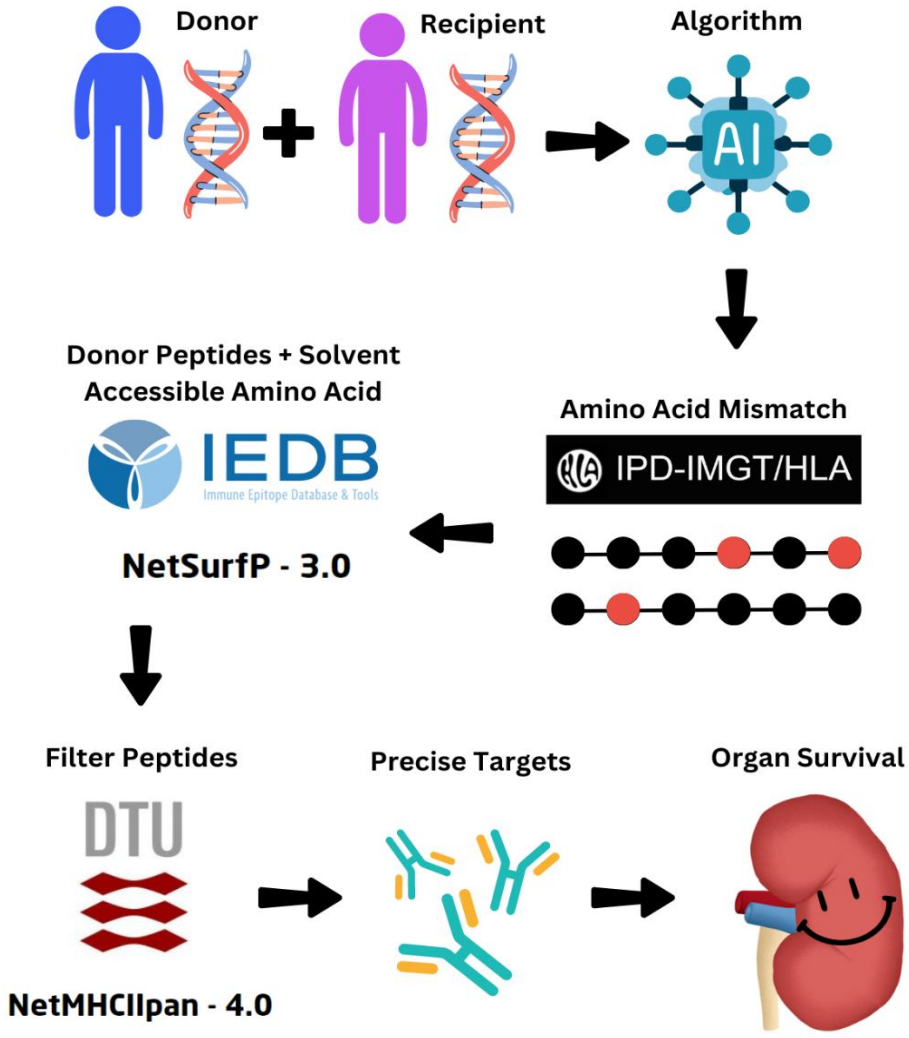
Worcester, MA. 01605

Abstract

Organ rejection is a dangerous medical complication that can occur after an organ transplant. Currently, all transplant patients are prescribed life-long immunosuppressors to decrease the risk of organ rejection. However, these medications can increase the susceptibility to other infections and cancers. Human leukocyte antigen (HLA) mismatches between donors and recipients can initiate T-cell activation, which is known to be the primary mediator of organ rejection. However, HLA genes are very polymorphic, and classifying "whole" HLA mismatches does not account for the minor amino acid differences that can start rejection. One solution is to create a machine-learning model that can analyze donor and recipient HLA sequences to predict MHC-peptide complexes, which are the molecules that T-cells recognize to start an immune response. This information can be used to predict rejection and find precise targets for immunosuppression. The project used datasets with MHC class I-peptide binding information to analyze donor and recipient HLA sequences. The result is that the model can accurately predict MHC-peptide complexes and rejection targets, with an R^2 value of 0.723. In conclusion, focusing on MHC-peptide presentation can account for HLA polymorphism and is more accurate in predicting organ rejection. Additionally, this data can be used to administer personalized and targeted immunosuppressors or decrease the need for broad immunosuppressors altogether. In the future, a similar model can be developed to predict antibody-mediated rejection (AMR) using MHC-class II datasets and be modified to support other organ transplants.

Keywords: Organ rejection, immune system, antibodies, cytokines, T cells, machine learning

Graphical Abstract



Acknowledgements

I sincerely thank Dr. David Harlan of UMass Diabetes Center of Excellence, Demetri Maxim of Nephrogen, Dr. Floyd E. Brownewell of Worcester Polytechnic Institute, and Dr. Lawrence Stern of UMass Chan Medical School for their support and input during the development of my project. Finally, I extend my greatest appreciation to Dr. Kevin Crowthers for his support and feedback during the development of my project and this thesis.

Selective T-Cell Inhibition Using Precision Medicine to Prevent Organ Rejection

Organ transplants are among the greatest advances in modern medicine, saving tens of thousands of lives every year. By increasing life expectancies and improving the quality of life, they remain the best therapy for terminal and irreversible organ failure (Grinyó, 2013). However, there is currently a major problem in the organ transplant industry: the demand is vastly greater than the supply. Due to a lack of organ donations, about seventeen people die each day while waiting for an organ transplant (*Organ, Eye and Tissue Donation Statistics*, n.d.). The immense demand emphasizes that every donated organ has the potential to change lives, and it is crucial to maintain the long-term health of each organ for the sake of the patient and the organ as well.

Overview of Organ Rejection

Even if a patient is successful in receiving an organ transplant, many medical complications may occur after the transplant, the most common being organ rejection. The immune system is a body system that destroys foreign cells to protect the body from harm. In the case of organ rejection, the immune system recognizes the transplanted organ as foreign and attempts to attack it by producing cells or antibodies that invade the organ (*Understanding Transplant Rejection | Stony Brook Medicin*e, n.d.). Currently, all transplant patients are prescribed immunosuppressors to decrease the risk of organ rejection. However, recipients must take immunosuppressive drugs for their entire lives for their bodies to accept a donated organ. While these medications prevent organ rejection to an extent, about 10-20% of patients will still experience at least one episode of rejection within the first three months to one year after a transplant (*Organ Rejection after Renal Transplant | Columbia Surgery*, n.d.). Additionally, they can also severely weaken the immune system, increasing the risk of cancer, infections, and other diseases (Kelly, 2022). New treatments are necessary to prevent organ rejection without using broad immunosuppressors that weaken the entire immune system.

Chronic Rejection

Depending on the mechanisms and timeframe of the rejection episode, rejection can be categorized into many different types. Acute and chronic rejection are categorized based on the time rejection occurred after the transplant. Acute rejection occurs within the first three months to a year after the transplant, while chronic rejection can occur after the first year of the transplant. Chronic rejection is often irreversible and can lead to graft failure or death (Hunt & Saab, 2012). Immunosuppressors are effective in decreasing the risk of acute rejection but not against chronic rejection. By five years post-transplant, chronic rejection affects up to 50% of kidney transplants (Gautreaux, 2017). Since chronic rejection is often asymptomatic and occurs over an extended period, there is currently no medicine to date that can treat chronic rejection symptoms (*Understanding Transplant Rejection | Stony Brook Medicine,* n.d.). The common treatment method is to increase the dosage of immunosuppressive drugs, which can exacerbate the dangerous side effects. Therefore, it is imperative to understand and target the mechanisms involved in chronic rejection to maintain long-term allograft health.

MHC-Peptide Presentation

Early chronic organ rejection is primarily caused by T-cell-mediated rejection (Chong, 2020). T-cells are a type of immune cell that plays a crucial role in identifying and eliminating foreign cells. When T-cells misinterpret donated organ cells as foreign, it can lead to T-cell activation and an attack on the transplanted organ. MHC peptide presentation plays a vital role in T-cell activation and can lead to developing strategies to prevent transplant rejection. The major histocompatibility complex (MHC) is a group of genes that code for MHC molecules found on the surface of cells. These molecules play a vital role in the immune system's ability to distinguish between "self" and "non-self" (King, 2007). There are two main types of MHC molecules: MHC class I and MHC class II molecules. While MHC class I molecules

are found on the surface of all nucleated cells, MHC class II molecules are only present on antigen-presenting cells (Lakna, 2018). Nonetheless, the main function of all MHC molecules is to bind peptide fragments derived from pathogens (or donor cells) and display them on the cell surface for recognition by the appropriate T cells (Hewitt, 2003). If T-cell receptors (TCRs) recognize a peptide from the transplanted organ on an MHC molecule, it activates, starting the immune response against the transplanted organ and initiating rejection.

Indirect Allorecognition

Antigen presentation can occur through direct or indirect pathways. However, chronic rejection is primarily mediated by the indirect pathway (Siu et al., 2018). As donor organ cells die and are replenished, the damaged donor cells shed MHC molecules. The MHC molecules are taken up by the recipient antigen-presenting cells (APCS), which break down donor MHC molecules into smaller peptide fragments (Mak et al., 2014). These peptides are loaded onto recipient MHC class II molecules and are presented on the surface of recipient APCs (SITNFlash, 2012). If there is a significant mismatch in the peptides displayed and the recipient's MHC molecules, naïve T-cells may recognize the peptide complex displayed on APCs as foreign, starting an immune attack against the donor organ (Mak et al., 2014).

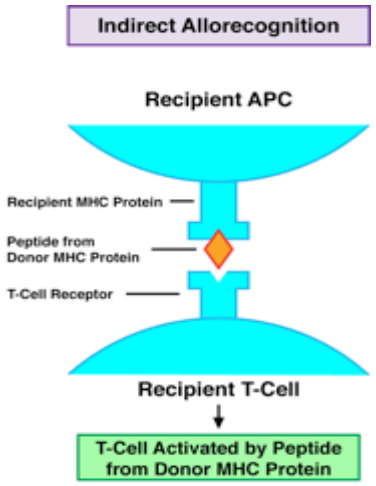


Figure 1: Damaged donor cells shed MHC molecules into the graft and surrounding tissue. These molecules are broken down into donor-derived peptides, which are taken up by recipient antigen-presenting cells (APCs) and are presented in MHC class II. Because the peptide is derived from a molecule that is not expressed in the recipient, the MHC-peptide on the surfaces of these APCs is seen as "non-self" by recipient T-cells. These T cells activate and attack the graft (Mak et al., 2014; SITNFlash, 2012).

Tissue Typing and Immune Profiling

When looking for organ matches, doctors perform Human Leukocyte Antigen (HLA) typing to understand the similarity in antigens between the donor and the recipient. The HLA is a group of genes that provide instructions to make antigens present on the surface of cells (Manski et al., 2019). Six

specific HLAs are looked for, and a higher similarity results in a likely chance of an organ match (*Matching and Compatibility | Transplant Center | UC Davis Health*, n.d.). However, HLA genes are the most polymorphic genes in the human genome. This means that HLAs have many different allele combinations, and their variant alleles have high degrees of sequence similarity. The similarity can be difficult to establish with current serological and low-resolution tests (Dasgupta, 2016). Therefore, understanding the exact differences in HLAs between the donor and recipient can result in a better treatment method that is personalized and accurate for the recipient. One way to do so is with machine learning.

Benefits of Machine Learning

Machine learning is a subset of artificial intelligence that uses statistical techniques that allow computer systems to automatically learn and develop from experience without being explicitly programmed (Costa, 2019). Previous studies have employed machine learning techniques to sift through massive datasets of gene expression data. Machine learning algorithms can analyze data to identify patterns and establish relationships from complex datasets. For this project, machine learning would allow HLA sequence data to be used to make a prediction model. By training the model on datasets of HLA sequences and peptide binding affinities, the algorithm can predict these complexes with high accuracy, paving the way for personalized and targeted immunosuppression. There have been many studies that employ machine learning to predict organ rejection. However, those models focus on "whole" HLA mismatches, which do not account for HLA polymorphism or the peptide sequences. Therefore, by focusing on HLA sequences and peptides, a more accurate and robust model can be created to prevent organ rejection. This way, we can protect the patient and the organ from harm.

Problem Statement

Chronic organ rejection affects about 50% of kidney transplants five years post-transplant. Due to chronic rejection occurring over a long period of time, there are limited methods to diagnose and treat chronic rejection. Even though Human Leukocyte Antigen (HLA) mismatches are the primary cause of rejection, HLA genes are very polymorphic, and current HLA typing methods do not account for the diverse amino acid variations within each allele that can initiate rejection.

Objective

The objective is to make a machine learning model that can predict rejection and provide specific targets that will cause rejection, given donor and recipient HLA sequences. The model will work by predicting the MHC-peptide complex on the donor organ by focusing on the specific HLA allele mismatches. Ideally, this model will use mismatches to provide information on targets for personalized immunosuppression.

Obj. 1a: Access the amino acid sequence of each HLA allele and align sequences to identify amino acid mismatches between given donor and recipient alleles at a specific locus.

Obj. 1b: Finding solvent-accessible amino acids to filter amino acid mismatches to ones that have the highest probability of immunogenicity.

Obj. 2: Generate donor-derived peptides based on the solvent-accessible amino acid mismatches as 15 amino acid chain peptide fragments.

Obj. 3: Filter peptides by considering the ones with the highest binding affinity to recipient HLA class II alleles and ones.

Obj. 4: Validate the model's accuracy by testing the algorithm with HLA typing data from known rejection and non-rejection donor and recipient samples.

Obj 5: Create a binary classification machine-learning model that is trained on rejection and no-rejection samples to find significant gene features.

Hypothesis

Based on previous organ rejection prediction models, it is hypothesized that the proposed methodology will be successful in predicting rejection as it focuses on specific amino acids mismatched to predict MHC-peptide complexes that may initiate T-cell activation. By learning from current MHC-peptide-predicting data sources and public repositories, the methodology results in accurate HLA mismatches and potential immunosuppressive targets.

Section II: Methodology

Role of Student vs. Mentor

I (the student) conducted all project development, research, testing, and analysis. My mentor guided me with the structure of written proposals and gave substantial feedback on technical documents and presentations. This project has had significant work contributed to it over the course of six months.

Equipment and Materials

Data Collection and Preprocessing:

HLA Protein Sequences. Various tools were used to obtain the data used in the machine-learning model and the data analysis methods. The Immuno-Polymorphism Database (IPD-IMGT/HLA) version 3.55.0 from the European Bioinformatics Institute (EBI) was accessed through the database's public FTP site hosted by the EBI. The database provides a central repository for sequences of HLA alleles, including the protein sequences in the FASTA format. HLA allele sequences were filtered to only include the commonly typed HLA loci: HLA-A, -B, -C, -DRB1, -DRB3, -DRB4, -DRB5, –DQA1, -DQB1, -DPA1 and -DPB1 (Hamed et al., 2018). The alleles were converted into field type two resolution, and duplicates were removed as higher resolution typing does not affect the amino acid sequence of the protein (Kramer et al., 2020).

Study Cohorts. To validate the model, two study cohorts were used. The STAR files were obtained by the United Network for Organ Sharing (U.N.O.S.), which include donor and recipient transplant data dated back to 1987. The large dataset was preprocessed, resulting in a small, manageable dataset with living kidney transplantations. The dataset contains past donor and recipient HLA alleles along with the rejection outcome. Chronic rejection was defined as rejection episodes that occur at least one year after the transplant (Justiz Vaillant & Mohseni, 2023). The inclusion and exclusion processing can be seen in Figure 2.

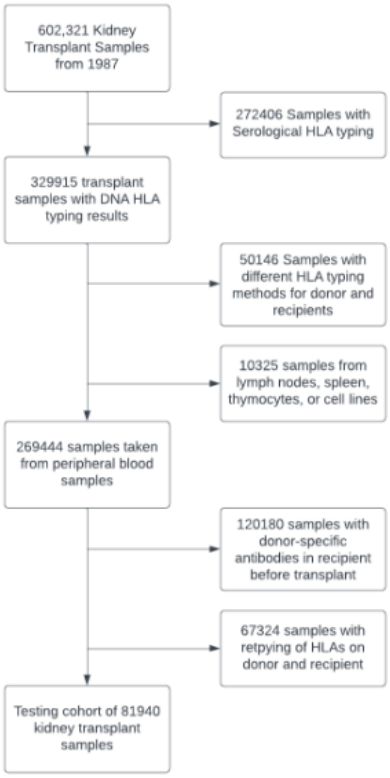


Figure 2: U.N.O.S. Dataset Inclusion and Exclusion Algorithm

HLA-Epi is another model that calculated the epitopic mismatch load between potential recipient-donor pairs. The HLA-Epi dataset contains donor and recipient HLA alleles along with their calculated compatibility scores (Geffard et al., 2022). Even though the model focuses on direct allorecognition, the compatibility scores can be used to validate the proposed model's performance through regression models. Additionally, they have scores calculated by the PIRCHE-II model for the same donor and recipient alleles. The PIRCHE-II model is another algorithm to predict indirectly recognizable HLA epitopes Geneugelijk & Spierings, 2020). The PIRCHE-II model does not consider solvent-accessible mismatches. Therefore, the scores in the HLA-Epi dataset can be used to compare the performance of the proposed model with competitor models.

Bioinformatic Servers:

Bioinformatic analysis servers were used to analyze and compare the amino acid sequences of donor and recipient HLA alleles. NetSurfP version 3.0 from the Danmarks Tekniske Universitet (DTU

Health Tech) was used to predict the surface accessibility of individual amino acids in an amino acid sequence. Additionally, NetMHCIIpan version 4.1 from DTU Health Tech was used to predict the binding affinity and eluted ligand of donor HLA peptides to recipient HLA class II alleles.

Software and Software Packages:

Google Colaboratory was used to code the machine learning models, as it is a hosted Jupyter Notebook to write and execute Python code through the browser. Microsoft Excel was used to format the data in a table format to make it easier to upload as a data frame into Google Collab. The HLA Epitope Mismatch Algorithm (HLA-EMMA) was used to validate amino acid mismatch results. Python libraries such as “Pandas” were used to import Excel data files, and “NumPy” was used to support the large arrays in the data files. Additionally, “MatplotLib” was used to visualize data, and “Seaborn” was used to create a confusion matrix. Lastly, the Statistical Analysis System (SAS) software will be used to convert the U.N.O.S. data files into a readable Excel file.

Accessing Amino Acid Sequences

A sample donor and recipient file was downloaded from the HLA-EMMA software. The donor and recipient alleles were reported in the second field typing resolution. However, the IPD/IMGT-HLA database reports most HLA alleles in the fourth field typing resolution (Casey, 2023). Therefore, if an allele was A*01:01, the FASTA sequence for A*01:01:01:01 was used, as they are an equivalent representation of the same allele located on the HLA-A locus. The FASTA sequence was input into a Google Colaboratory file that would find the amino acid mismatches.

Modified Needleman-Wunsch Algorithm

The IPD/IMGT-HLA database has allele sequences in different lengths. However, to find the amino acid mismatches, the sequences must be of equal length to be vertically aligned. Therefore, a

modified Needleman-Wunsch algorithm was used to make the sequences have the same lengths, allowing for mismatches to be found. The Needleman-Wunsch algorithm is a common global alignment method that uses a scoring matrix and dynamic programming to find the optimal alignment between two sequences (Mittal, 2024). The traditional algorithm adds gaps between the protein sequences, representing the evolutionary changes between the two sequences. As seen in Figure 3, the gaps attempt to optimize the alignment score and reveal any mutations, insertions, or deletions that may have occurred over time (NandiniUmbarkar, 2020). However, to find the amino acid mismatches between the donor and recipient FASTA sequences, there should not be any additional modifications to the sequence. Therefore, the model uses a similar scoring system but has a very high gap penalty. The gap penalty is a negative score that is added to the score any time a gap is inserted in the sequences (Mount, 2008). By having a high negative gap penalty, the overall score will significantly decrease. To have a high alignment score, the sequences will not be modified.

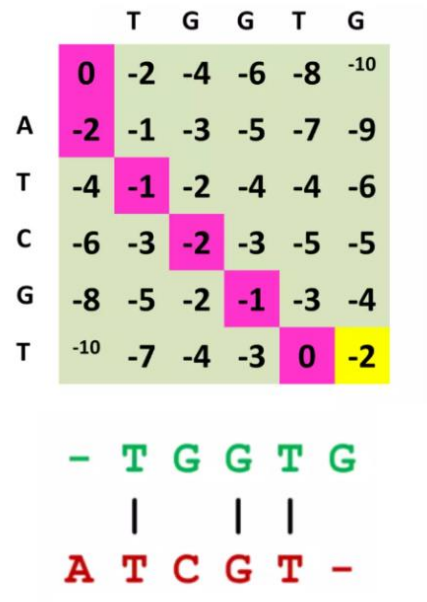


Figure 3: The Needleman-Wunsch algorithm creates a scoring matrix that determines which alignment has the greatest score. The optimal alignment contains gaps on the ends of the sequence, representing potential evolutionary mutations between the sequences. (Coghlan, 2013)

Aligning Amino Acid Sequences and Identifying Mismatches

The Python programming software was used to compare a single donor allele sequence to the respective recipient allele sequence using Google Colaboratory. This way, amino acid mismatches were donor amino acids that are not present in either recipient alleles, as they have a lower chance of initiating an immune response. The program goes through each position to store the mismatches. The

mismatches were validated with HLA-EMMA, and the code was modified to confirm the results. HLA-EMMA is a software that finds mismatches between donor and recipient alleles.

Finding Solvent-Accessible Amino Acid Mismatches

Solvent-accessible amino acids are amino acids in a protein that are exposed to the solvent surrounding the protein. Solvent accessibility is an important structural property of proteins because active sites are often located on the surfaces of proteins (Savojardo et al., 2021). These amino acids have a much higher chance of being recognized by T-cells. Therefore, focusing on solvent-accessible amino acid mismatches can provide specific peptides that have a greater chance of causing rejection. NetSurfP was used to predict the solvent accessibility for each amino acid in the donor alleles, and the solvent-accessible amino acids, which were also amino acid mismatches were stored for peptide analysis.

Generating Donor-Derived Peptide Chains

NetMHCIIpan is a server that generates peptides and predicts the binding strength of those peptides to an MHC-II molecule. The donor alleles were used for peptide sequence generation, and the molecules were input as the recipient MHC class II molecules. Because MHC class II molecules bind to peptides that have a length of 15 amino acids, the donor-derived peptides were 15 amino acids in length (Schafer et al., 1995). The binding affinity and the eluted ligand were found for all generated peptides.

Filtering Peptides With Binding Affinity and Eluted Ligand

The eluted ligand score is the likelihood of a peptide being an MHC ligand, while binding affinity is the strength of attraction between the peptide and the molecule (Wongklaew et al., 2024). NetMHCIIpan reports the strongest binding peptide sequences to each MHC class II molecule. Out of

those, the peptides containing the solvent-accessible amino acid mismatches were stored as the most significant peptides that may cause rejection.

Machine-Learning Model Training and Testing:

After the model is completed, the HLA-Epi data will be used to create regression models between the predicted compatibility score and the true compatibility score. The donor and recipient samples be run through the model, and the predicted scores will be recorded. Then, the true scores of the respective samples will be matched with the predicted score from the model. Regression models will be made to validate the model's ability to accurately predict a score for a sample on a scale. The model will be improved until it reaches an accuracy of at least 70% or greater. If needed, feature selection algorithms such as random forest will be used to find the most influential HLA alleles, which can improve the accuracy of the regression models.

Two Samples and Paired T-tests

In this case, a 2-sample t-test can measure the difference between the means of two different groups. Thus, a t-test was used to measure the difference in the means of the rejection and non-rejection scores. Performing this test yields a t-statistic of 6.269 and a p-value of 1.646e-9. This is below the commonly accepted threshold $\alpha \leq 0.05$, thus it can be concluded that there is a statistically significant difference in mean scores between the rejection and non-rejection groups.

Additionally, a paired t-test can be used to measure the mean difference between paired data samples. A paired t-test was used to measure the difference between the predicted and actual compatibility scores in the regression models. Performing this test yields a t-statistic of 1.918 and a p-value of 0.057. The p-value is above the commonly accepted threshold. Thus, it can be concluded that there is no statistically significant difference between the predicted scores and the actual scores.

Section III: Results

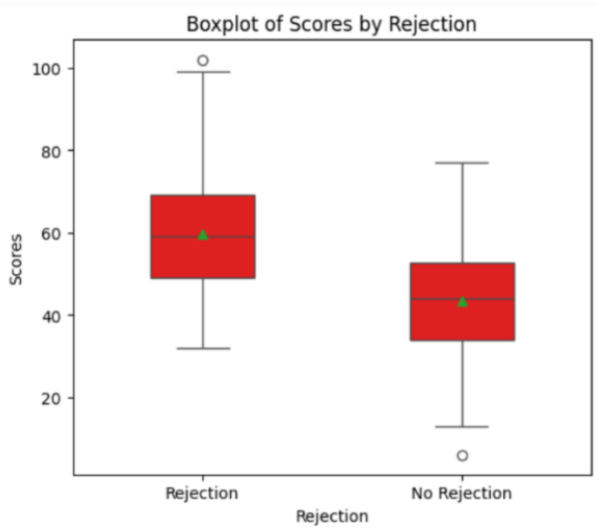
The MHC-peptide prediction methodology was carried out using HLA-B locus alleles. Each classification model was tested on an unseen dataset, from which certain accuracy metrics were obtained.

U.N.O.S. Dataset and Scores

After the UNOS dataset was processed, a random sample of 250 donor and recipient HLA allele combinations was split into rejection and no-rejection groups. Chronic rejection was classified as rejection that occurred at least one year after the initial transplant. The samples were run through the

Figure 4: Box and Whisker plots for predicted total peptide score from model in rejection and no-rejection groups. Green symbol represents the mean score of both groups.

model and the output score was recorded. Box and whisker plots were created to compare the scores between the rejection and no-rejection groups, which can be seen in Figure 4.



MHC-Peptide Prediction Results

Following a similar format as HLA-EMMA, each donor allele was compared to both recipient alleles. For the sample test, the recipient had HLA-B locus alleles B*08:01 and B*40:02, while the donor had alleles B*07:02 and B*35:03. As seen in Table 1, there are multiple mismatches between both donor alleles and recipient alleles. However, only some were predicted to be solvent accessible by NetSurfP. The B*35:03 donor allele had many more amino acid mismatches compared to the B*07:02 allele.

Table 1: Amino acid mismatches for donor HLA-B locus allele B*07:02 (top) and donor allele B*35:03 (bottom) and recipient alleles B*08:01 and B*40:02. Amino acid sequences were extracted in FASTA format from the IPD/IMGT-HLA database. Python code in Google Colaboratory found amino acid mismatches (yellow) and their positions in the allele sequence (red). NetSurfP was used to predict the amino acid mismatches which would be solvent-accessible (green).

Info	Allele	33	91	93	94	95	138	176	180	202
Recipient	B*08:01	D	F	T	N	T	N	V	D	T
Recipient	B*40:02	H	S	T	N	T	N	V	L	T
Donor	B*07:02	Y	Y	A	Q	A	D	E	R	K
Total AA MM	9	Y	Y	A	Q	A	D	E	R	K
Solvent Accessible MM	3	---	---	A	Q	---	---	E	---	K

Info	Allele	33	48	69	118	119	121	127	138	140	155	187	218	306	329	349
Recipient	B*08:01	D	S	E	T	L	S	V	N	Y	R	T	I	V	A	C
Recipient	B*40:02	H	T	K	T	L	S	V	N	Y	R	E	I	V	A	C
Donor	B*35:03	Y	A	T	I	I	R	L	D	F	S	L	V	I	T	S
Total AA MM	14	Y	A	T	I	I	R	L	D	F	S	L	V	I	T	S
Solvent Accessible MM	7	---	---	T	---	---	---	---	---	---	S	L	V	I	T	S

After noting down the solvent-accessible amino acid mismatches, the donor allele B*07:02 sequence was inputted into NetMHCIIpan, and the recipient MHC class II alleles were entered as the MHC molecules to calculate the binding affinity and eluted ligand scores. A higher binding affinity and eluted ligand score means the peptide has a higher chance of binding to the MHC molecule. The recipient's HLA class II alleles were DRB1*11:01, DRB1*13:01, DRB3*02:01, DQB1*03:01, DQB1*06:03, DQA1*01:03, DQA1*05:05, DPB1*04:01, DPB1*105:01, and DPA1*01:03. Four peptides containing the solvent-accessible mismatches were predicted to bind to the DQA1*01:03-DQB1*03:01 molecule, five peptides were predicted to bind to the DQA1*05:05-DQB1*06:03 molecule.

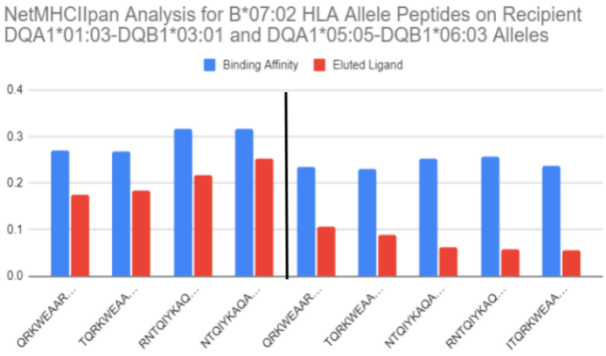


Figure 5: Binding affinity and eluted ligand scores of strong peptide sequences containing solvent-accessible mismatched amino acids. Peptides were generated from HLA-B locus donor allele B*07:02. Left of the chart is peptide data for HLA class II recipient DQA1*01:03-DQB1*03:01 molecules, and right is peptide data for binding on HLA class II recipient DQA1*05:05-DQB1*06:03 molecules.

A similar procedure was applied to the second donor allele. However, as Figures 6 and 7 show, there were many more peptides that contained solvent-accessible mismatches for the B*35:03 allele compared to the B*07:02 allele.

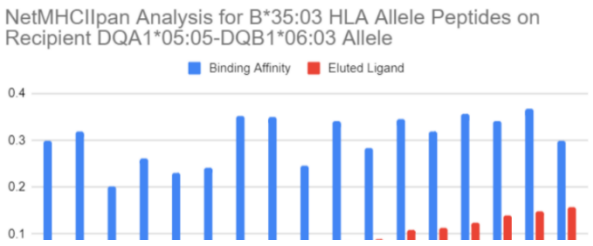


Figure 6: Binding affinity and eluted ligand scores of strong peptide sequences containing solvent-accessible mismatched amino acids. Peptides were generated from HLA-B locus donor allele B*35:03. Peptide data for binding to HLA class II recipient DQA1*01:03-DQB1*03:01 molecules.

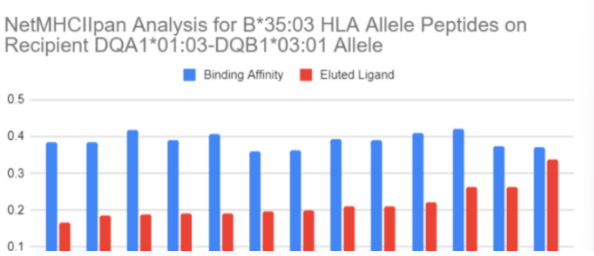


Figure 7: Binding affinity and eluted ligand scores of strong peptide sequences containing solvent-accessible mismatched amino acids. Peptides were generated from HLA-B locus donor allele B*35:03. Peptide data for binding to HLA class II recipient DQA1*05:05-DQB1*06:03 molecules.

Regression Model Scatterplot

Regression models were constructed based on the predicted and true scores of the samples in the HLA-Epi dataset. Ridge regression, random forest regression, linear regression, lasso regression, and polynomial regression were created. The performance of each regression model can be seen in Appendix C. As regression models cannot be measured in terms of "accuracy," the R^2 value was used to gauge how well the predicted scores could match the true scores. As the initial R^2 values were well below the desired 70%, a random forest feature selection algorithm was used to identify the importance of specific HLA allele types. Weightages were assigned based on their importance, with a greater weight being assigned to more influential allele types, based on the results of the feature selection algorithm. The final R^2 value of the Ridge Regression model was 0.723.

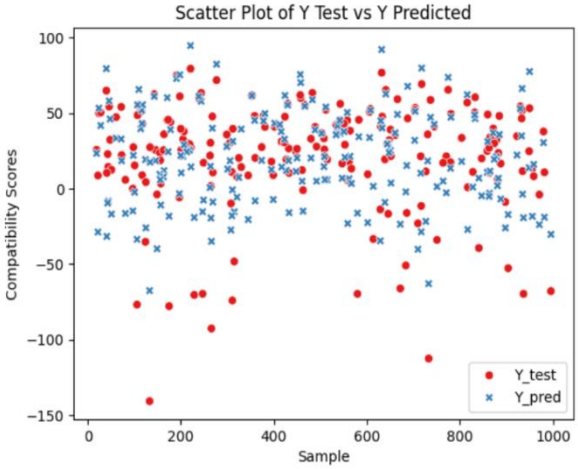


Figure 8: Scatter Plot of the ridge regression model before adding weightages to allele types. The scatter plot had an R^2 value of 0.626.

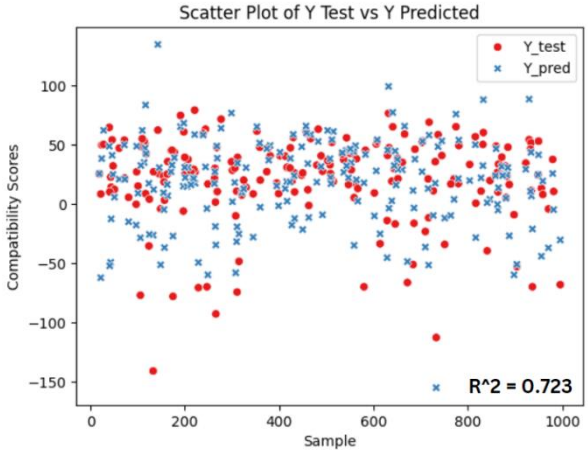


Figure 9: Scatter Plot of the ridge regression model after conducting random forest feature selection. The new scatter plot has an R^2 value of 0.723.

Section IV: Discussion

As chronic organ rejection is a dangerous and prevalent medical condition after a transplant, a model was created that could identify minute differences between donor and recipient HLA alleles to

predict rejection and find targets for precise immunosuppression. The amino acid differences that cause rejection are different for every transplant, and so should the medications.

Based on these findings, it can be determined that predicting MHC peptide complexes can be used to predict rejection. Focusing on amino acid differences between donor and recipient sequences provided a more accurate understanding of the specific peptides that had a higher chance of immunogenicity. Additionally, keeping unique mismatches was important in reducing the number of features the model would use. For example, all the mismatches from Table 1 contained the amino acid mismatches in the donor that were not present in either of the recipient alleles, which allowed only significant mismatches to surface, which influenced the peptide selection. While NetMHCIIpan showed multiple strong binding peptides, only the ones that contained solvent-accessible mismatches were stored. Out of all the string peptides, most of them contained solvent-accessible peptides. Additionally, multiple strong peptides contained many of the same amino acid positions. For example, in Figure 2, QRKWEAAREAEQRRA started at position 87, and TQRKWEAAREAEQRR started at position 86. After looking at the solvent-accessible amino acid mismatches, it was evident that these two peptides contained two solvent-accessible mismatches. Specifically, at positions 93 and 94. Similarly, peptide RNTQIYKAQAQTDRE begins at position 167, and peptide NTQIYKAQAQTDRES begins at position 168, and contains a solvent-accessible mismatch at position 176. A similar result was seen in the donor-derived peptides in donor allele B*35:03. However, this allele had many more strong peptides that contained solvent-accessible mismatches. It is probable that because of the higher mismatches, there was a higher number of strong peptides, showing evidence for this allele being more immunogenic than the second donor allele B*07:02. In both cases, many of the peptides repeated for both the recipient alleles, which again shows evidence for using peptides to find immunosuppressive targets as repeated peptides have a higher chance of initiating an immune response.

Additionally, the results from the UNOS dataset shows a clear correlation between higher scores corresponding to rejection samples, and lower scores corresponding to non-rejection samples. Additionally, performing a two-sample t-test resulted in the difference in mean scores between the groups as statistically significant. The significance reinforces the model's ability to present different scores based on the rejection outcome. As peptides are counted for the targets, a greater number of peptide possibilities corresponds to a greater chance of rejection. However, as everyone has at least some difference in their DNA, it is more beneficial to understand the compatibility score of a specific recipient and donor combination, as the box and whisker plots have a significant overlap in scores. Therefore, comparing the model's scores to already tested compatibility scores can give us more insight into the model's accuracy.

Regression models were created and analyzed to find the correlation between the model's scores and true compatibility scores. The ridge regression model had performed the best, with an R^2 value of 0.626. However, to get the desired accuracy, finding the most influential HLA alleles can aid in making the model more accurate. By finding the most influential HLA allele types, a greater weight can be added to those alleles. After conducting a random forest feature selection, HLA-A and HLA-B were found to be the most important HLA types. By giving those alleles the greatest weightages, the R^2 value increased to 0.723. The increase in accuracy shows evidence for those alleles being the most influential in the rejection outcome, and clinicians should make an effort to match donor and recipients with a high similarity in those alleles.

In the end, all the objectives were accomplished, as the result presented peptides, suggesting that they can be used as immunosuppressive targets. Potential limitations would include testing the model clinically. However, validating our results with current models, such as HLA-EMMA can provide more confidence in our methods and results.

Future Research

Future research would include creating models that could support other organ transplants, such as heart, lung, or liver. Additionally, a similar model could be created by focusing on the direct pathway or antibody-mediated rejection. There is also work that can be done to optimize the machine learning algorithms, including adding more features or testing the model with external datasets. Similarly, training the model with more patient information, such as age, weight, and family history, could potentially improve the model by using more patient features. These studies could improve donor selection and decrease the need for immunosuppressors. In short, the endless future research opportunities have the potential to revolutionize the healthcare industry from its current state today.

Section V: Conclusion

The ultimate objective of this project was to create a machine-learning model that can predict the risk of rejection, given donor and recipient HLA sequences, by finding the most significant peptides. Amino acid sequence data was obtained from the IPD/IMGT-HLA database, and a sample donor and recipient HLA sequence file was obtained from HLA-EMMA. Using Google Colab, the amino acid mismatches were identified, and NetSurf P was used to find the solvent-accessible mismatches. These mismatches have a higher chance of being recognized by recipient T-cells because they are exposed to the solvent in the peptide. Then, NetMHCIIpan was used to generate donor-derived peptides and calculate the binding affinity to the recipient alleles. Strong binding peptides that contained the solvent-accessible peptides were stored as the peptides that have the highest chance of being immunogenic. After analyzing the results, it was evident that there was a correlation between solvent-accessible mismatches and the number of strong peptides that were present, with a greater number of strong peptides correlating with a higher chance for rejection.

Additionally, many of the peptide sequences had overlapping positions or were in the sequence region with multiple amino acid mismatches. For example, in the donor allele sequence B*35:03, the peptide sequence TQFVRFDSDAASPRT was predicted to strongly bind to multiple recipient MHC molecules. Additionally, all of the peptides in the donor allele sequence B*07:02 predicted to bind to the recipient were extremely similar, having moved one or two amino acid positions in the sequence. This provides evidence to support the conclusion that similar peptide sequences are likely to cause rejection, as they have a higher chance of binding to multiple recipient alleles. Validating the results with HLA-EMMA supports the proposed methodology, and the model can be improved in the future by including more features. The amino acid differences that cause rejection are different for every transplant, and so should the medications. With this model, we can not only keep the organ safe, but keep the patient healthy throughout their life.

Section VI: References

Azzi, J. R., Sayegh, M. H., & Mallat, S. G. (2013). Calcineurin Inhibitors: 40 Years Later, Can't Live Without …. *The Journal of Immunology*, *191*(12), 5785–5791. https://doi.org/10.4049/jimmunol.1390055

Buhl, N. (2023, August 8). Mitigating Model Bias in Machine Learning | Encord. Encord.Com. https://encord.com/blog/reducing-bias-machine-learning/

Chen, H., Yang, J., Zhang, S., Qin, X., Jin, W., Sun, L., Li, F., & Cheng, Y. (2019). Serological cytokine profiles of cardiac rejection and lung infection after heart transplantation in rats. *Journal of Cardiothoracic Surgery*, *14*(1), 26. https://doi.org/10.1186/s13019-019-0839-5

Chouhan, K. K., & Zhang, R. (2012). Antibody induction therapy in adult kidney transplantation: A controversy continues. *World Journal of Transplantation*, *2*(2), 19–26. https://doi.org/10.5500/wjt.v2.i2.19

Costa, C. D. (2019, August 26). What Is Machine Learning & Deep Learning? Medium. https://medium.com/@clairedigitalogy/what-is-machine-learning-deep-learning-7788604004da

Dasgupta, A. (2016). Chapter 2 - Limitations of immunoassays used for therapeutic drug monitoring of immunosuppressants. In M. Oellerich & A. Dasgupta (Eds.), *Personalized Immunosuppression in Transplantation* (pp. 29–56). Elsevier. https://doi.org/10.1016/B978-0-12-800885-0.00002-3

Ding, M., He, Y., Zhang, S., & Guo, W. (2021). Recent Advances in Costimulatory Blockade to Induce Immune Tolerance in Liver Transplantation. *Frontiers in Immunology*, *12*. https://doi.org/10.3389/fimmu.2021.537079

Gautreaux, M. D. (2017). Chapter 17 - Histocompatibility Testing in the Transplant Setting. In G. Orlando, G. Remuzzi, & D. F. Williams (Eds.), *Kidney Transplantation, Bioengineering and Regeneration* (pp. 223–234). Academic Press. https://doi.org/10.1016/B978-0-12-801734-0.00017-5

Grinyo, J. M. (2013). Why Is Organ Transplantation Clinically Important? *Cold Spring Harbor Perspectives in Medicine*, *13*(11). https://doi.org/10.1101/cshperspect.a014985

Hamed, C. T., Meiloud, G., Veten, F., Hadrami, M., Ghaber, S. M., Boussaty, E. C., Habti, N., & Houmeida, A. (2018). HLA class I (-A, -B, -C) and class II (-DR, -DQ) polymorphism in the Mauritanian population. *BMC Medical Genetics*, *19*, 2. https://doi.org/10.1186/s12881-017-0514-4

Harlan, D. M., & Kirk, A. D. (1999). The Future of Organ and Tissue Transplantation: Can T-Cell Costimulatory Pathway Modifiers Revolutionize the Prevention of Graft Rejection? *JAMA*, *282*(11), 1076–1082. https://doi.org/10.1001/jama.282.11.1076

Hunt, D., & Saab, S. (2012). Post–Liver Transplantation Management - ScienceDirect. In *Zakim and Boyer's Hepatology* (Sixth, pp. 869–882). https://www.sciencedirect.com/science/article/abs/pii/B9781437708813000498

Iglesias, M., Brennan, D. C., Larsen, C. P., & Raimondi, G. (2022). Targeting inflammation and immune activation to improve CTLA4-Ig-based modulation of transplant rejection. *Frontiers in Immunology*, *13*. https://doi.org/10.3389/fimmu.2022.926648

Ingulli, E. (2010). Mechanism of cellular rejection in transplantation *Pediatric Nephrology*, *25*. https://doi.org/10.1007/s00467-008-1020-x

Kelly, J. (2022, April 27). *End of anti-rejection transplant drugs? A clinical trial at Hume-Lee hopes so.* VCU Health. https://www.vcuhealth.org/news/end-of-anti-rejection-transplant-drugs-a-clinical-trial-at-hume-lee-hopes-so

King, T. C. (2007). 2 - Inflammation, Inflammatory Mediators, and Immune-Mediated Disease. In T. C. King (Ed.), *Elsevier's Integrated Pathology* (pp. 21–57). Mosby. https://doi.org/10.1016/B978-0-323-04328-1.50008-5

Kirk, A. D., Harlan, D. M., Armstrong, N. N., Davis, T. A., Dong, Y., Gray, G. S., Hong, X., Thomas, D., Fechner, J. H., & Knechtle, S. J. (1997). CTLA4-Ig and anti-CD40 ligand prevent renal allograft

rejection in primates. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(16), 8789–8794. https://doi.org/10.1073/pnas.94.16.8789

Lewis, A., Koukoura, A., & Tsianos, Georgios-Ioannis. (2021). Organ donation in the US and Europe: The supply vs demand imbalance - ScienceDirect. *Transplantation Reviews*, *35*(2). https://doi.org/10.1016/j.trre.2020.100585

Mahmud, N., Klipa, D., & Ahsan, N. (2010). Antibody immunosuppressive therapy in solid-organ transplant. *MAbs*, *2*(2), 148–156. https://doi.org/10.4161/mabs.2.2.11159

Manski, C. F., Tambur, A. R., & Gmeiner, M. (2019). Predicting kidney transplant outcomes with partial knowledge of HLA mismatch. *Proceedings of the National Academy of Sciences*, *116*(41), 20339–20345. https://doi.org/10.1073/pnas.1911281116

Matching and Compatibility. (n.d.). UC Davis Health. Retrieved November 8, 2023, from https://health.ucdavis.edu/transplant/livingkidneydonation/matching-and-compatibility.html

Mittal, A. (2024, March 12). Sequence Alignment and the Needleman-Wunsch Algorithm. Analytics Vidhya. https://medium.com/analytics-vidhya/sequence-alignment-and-the-needleman-wunsch-algorithm-710c7b1a23a

Mota, A. P. L., Vilaça, S. S., das Mercês, F. L., de Barros Pinheiro, M., Teixeira-Carvalho, A., Silveira, A. C. O., Martins-Filho, O. A., Gomes, K. B., & Dusse, L. M. (2013). Cytokines signatures in short and long-term stable renal transplanted patients. *Cytokine*, *62*(2), 302–309. https://doi.org/10.1016/j.cyto.2013.03.001

Mount, D. W. (2008). Using gaps and gap penalties to optimize pairwise sequence alignments. *CSH Protocols*, *2008*, pdb.top40. https://doi.org/10.1101/pdb.top40

NandiniUmbarkar. (2020, October 12). Needleman-Wunsch Algorithm. *Medium*. https://medium.com/@nandiniumbarkar/needleman-wunsch-algorithm-7bba68b510db

Organ, Eye and Tissue Donation Statistics. (n.d.). Donate Life America. Retrieved November 8, 2023, from https://donatelife.net/donation/statistics/

Organ Rejection after Renal Transplant. (n.d.). Columbia Surgery. Retrieved November 8, 2023, from https://columbiasurgery.org/kidney-transplant/organ-rejection-after-renal-transplant

Reits, E., & Neefjes, J. (2022). HLA molecules in transplantation, autoimmunity and infection control: A comic book adventure. *Hla*, *100*(4), 301–311. https://doi.org/10.1111/tan.14626

Roberts, J. (2023, April 28). *Transplant rejection*. MedlinePlus Medical Encyclopedia. https://medlineplus.gov/ency/article/000815.htm

Signs of Kidney Transplant Rejection. (n.d.). Cleveland Clinic. Retrieved November 25, 2023, from https://my.clevelandclinic.org/health/diseases/21134-kidney-transplant-rejection

Trambley, J., Bingaman, A. W., & Lin, A. (1999). Asialo GM1+ CD8+ T cells play a critical role in costimulation blockade–resistant allograft rejection - PMC. *The Journal of Clinical Investigation*, *104*(12). https://doi.org/10.1172/JCI8082

Understanding Transplant Rejection. (n.d.). Stony Brook Medicine. Retrieved November 8, 2023, from https://www.stonybrookmedicine.edu/patientcare/transplant/rejection

Viatte, S. (2023, September 8). Human leukocyte antigens (HLA): A roadmap - UpToDate. Uptodate. https://www.uptodate.com/contents/human-leukocyte-antigens-hla-a-roadmap/print

Vijayan, S., Sidiq, T., Yousuf, S., van den Elsen, P. J., & Kobayashi, K. S. (2019). Class I transactivator, NLRC5: a central player in the MHC class I pathway and cancer immune surveillance. Immunogenetics, 71(3), 273–282. https://doi.org/10.1007/s00251-019-01106-z

Why Is Organ Donation Important? (2022, March 29). INTEGRIS Health. https://integrisok.com/resources/on-your-health/2022/march/why-is-organ-donation-important

Section VII: Appendices

Appendix A: Limitations and Assumptions

Limitations:

1. Classification models were tested and trained using the same dataset as data was limited.
2. The researcher could not legally test or confirm the results in a clinical setting.

Assumptions:

1. The IPD/IMGT-HLA database is accurate.
2. The HLA-EMMA sample donor and recipient HLA sequence is accurate.
3. The trends observed are predictive of the future.

Appendix B: Model Github Repository and Web Application Links

The full source code for the model and related files can be found at the GitHub repository: https://github.com/samhithabodangi/Organ-Rejection-Model. The final web application for PIPSA can be found at the GitHub repository: https://github.com/samhithabodangi/PIPSA-Model

Appendix C: Decision Matrix Comparing Compatibility Score Regression Models

Criteria	Linear Regression	Ridge Regression	Random Forest	Lasso Regression	Polynomial Regression
Mean Squared Error	487.682	485.305	531.052	488.244	485.564
Mean Absolute Error	17.028	16.782	15.068	16.967	16.886
Root Mean Squared Error	22.084	22.030	23.045	22.096	22.036
R Squared Value	0.624	0.626	0.590	0.623	0.625

Appendix D: Compatibility Score Regression Flow Chart