

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

Samhitha Bodangi
Massachusetts Academy of Math and Science
STEM Project
Instructor: Kevin Crowthers, Ph.D.

Table of Contents:

GLP Record Keeping Contract.....	3
Logbook Etiquette Date: 09/26/2023.....	4
Brainstorming.....	5
Pie Diagrams:.....	5
Mindmaps.....	8
Fishbone Diagrams.....	11
Systems Map.....	14
Project Abstract:.....	15
Project Introduction:.....	16
Project Introduction References:.....	18
Professional Communication:.....	19
Hello Dr. Lanese,.....	19
Hello, Dr. Zhang,.....	20
Hello Dr. Chen,.....	21
Hello Dr. Martins,.....	22
Hello Dr. Movahedi,.....	23
Hello, Dr. Mullens,.....	24
Dear Charles River Laboratories,.....	25
Hello Dr. Keeler,.....	26
Hello Dr. Kent,.....	27
Hello Dr. Brehm,.....	28
Hello Professor Stern,.....	29
Hello Dr. Politz,.....	30
Hello, Dr. Brownell.....	31
Materials and Methods:.....	32
Materials List:.....	32
Procedure:.....	34
Background:.....	38
Background References:.....	40
Daily Entries:.....	43
Entry 1: MATLAB Training, 11/19/23,.....	43
Entry 2: MATLAB Training and Software Download, 11/23/23,.....	43
Entry 3: Beginning Model Development, 11/24/23,.....	44
Entry 4: Meeting with Demetri Maxim, 11/28/2023,.....	44
Entry 5: Machine Learning Research, 12/03/23,.....	45
Entry 6: Meeting with Dr. Keeler, 12/05/23,.....	45
Entry 7: Meeting with Dr. Stern, 12/06/23,.....	45
Entry 8: Algorithm Research, 12/07/23,.....	45
Entry 9: GEO Kidney Biopsy Gene Analysis, 12/11/23,.....	46

Entry 10: Machine Learning Model Practice Training, 12/21/23,	48
Entry 11: Machine Learning GEO Practice Model for Biomarker Hunt, 12/22/23,	48
Entry 12: UNOS Data Received, 12/23/23,	49
Entry 13: Random Forest Machine Learning Model GEO, 1/2/2023,	50
Entry 14: Support Vector Machine Learning Model GEO, 1/15/24,	51
Entry 15: STEM Update Meeting #6 Takeaways, 1/17/24,	51
Entry 15: K-Nearest Neighbor Machine Learning Model GEO, 1/17/24,	52
Entry 16: Feature Experimentation with Negative Control, 1/23/24,	53
Entry 17: MHC-Peptide Methodology Testing, 1/31/24,	54
Entry 18: Web Application Design in Figma and Visual Studio Code, 1/27/24,	55
Entry 19: Accessing fasta Files and Identifying Amino Acid Mismatches, 2/5/24,	56
Entry 20: IPD/IMGT-HLA Sequence fasta File Processing, 2/6/2024,	57
Entry 21: Aligning HLA Protein Sequences, 2/7/24,	58
Entry 22: Filtering Solvent-Accessible Mismatches with NetsurfP, 2/08/24,	59
Entry 23: Immune Epitope Database for Peptide Prediction, 2/09/24,	60
Entry 25: Complete Model Building, 2/09/24,	60
Entry 26: Linear Regression Model with HLA-Epi Compatibility Scores, 3/2/24,	61
Entry 27: Remaining Regression Model with HLA-Epi Scores, 3/3/24,	62
Entry 28: Rejection vs. No Rejection Scores, 3/7/24,	64
Entry 30: Feature Selection and Weightages for Ridge Regression Model, 3/14/24,	65
Entry 31: Web Application Building with Python and React, 4/1/24,	66
STEM Hours Time-Log:	68

GLP Record Keeping Contract

I, Samhitha Bodangi commit to record keeping in accordance with Good Laboratory Practices.

- My experiments and records will be reproducible, traceable, and reliable.
- I will NOT write my notes on scraps of paper, post-it notes, or other disposable items. My notes will go directly into my laboratory notebook.
- My data will be recorded in real-time. If I cannot record data in real-time, I will record raw data as soon as physically possible.
- I will record both qualitative and quantitative observations in my laboratory notebook and laboratory reports.
- My laboratory notebook will include information on the materials and instruments utilized during experimentation.
- I will initial and date over the edge of any material that is taped into my laboratory notebook.
- I will provide a real-time record of any analysis I perform.
- I will use blue or black pen to make entries in my laboratory notebook. I will NOT use pencil.
- I will define ALL abbreviations.
- If I make a mistake in my laboratory notebook, laboratory worksheets, or other written material, I will not obliterate or obscure the mistake. Instead, I will cross out the mistake using a single line. Any empty spaces in tables or partially used notebook pages will be crossed out using a single diagonal line.
- If I record information online (ex. In Google Drive), I will login so that my contributions are traceable.
- I will initial and date each page in my notebook and the front of each laboratory report.

Samhitha Bodangi



09/26/2023

A more detailed description of GLP is located here:

<https://docs.google.com/document/d/1zeYoNSniKTc7MlBgTG1SEnhJiCK3UimCvTcKPQcyHGw/edit?usp=sharing>

Logbook Etiquette

Date: 09/26/2023

For research and engineering purposes, a logbook is considered a legal document and will help in providing documentation for the origin of ideas.

1- When adding something written in Pen- Blue or Black ~~not a Pencil~~ (and DO NOT USE WHITEOUT- ~~mistakes~~ can be corrected by adding the information above the crossed out material and adding your initials and date

2- Don't worry about neatness- it is a living document but **should be legible but understandable**

3- Page Numbers should be consecutive and located on the top corner of the page- outer edge

4- Do not remove pages

5- Put a line through empty space

6- Neat handwriting

7- **Make an entry every time you work on your project**

8- Make sure your entries are verified by a mentor/ teacher signature and your signature

9- Organize your Notebook: Format

A: Table of Contents

B: Brainstorming and Topic Ideas

C: Project Introduction: Topic, Phrase 1 (Testable Question/Engineering Need/Mathematical Conjecture), Phrase 2 + Timeline

D: Communications (i.e. to corresponding authors, mentors, and expert consultation, etc)

E: Draft of Materials and Methods (this can be performed for daily entries if variations occur over the course of the project).

F: Background- ie. competitor/market analysis, criteria/constraints

G: Daily Entries (every time you complete work on the project)

1: Title and Date

2: Short Introduction (putting the experiment/observations into context/objectives)

3: Methods/Materials (if not included in the beginning of the notebook)

Materials become important when someone needs to repeat your experiments

4: Observations/Experimental Data (both RAW and ANALYZED)-

A: graphs/figures

B: data tables

C: pictures

D: sketches or proof of concepts and prototypes (with labels)

E: Decision matrices

E: Ethical responsibility

5: Calculations and Data Analysis (STATISTICS)

6: Final Concluding Remarks

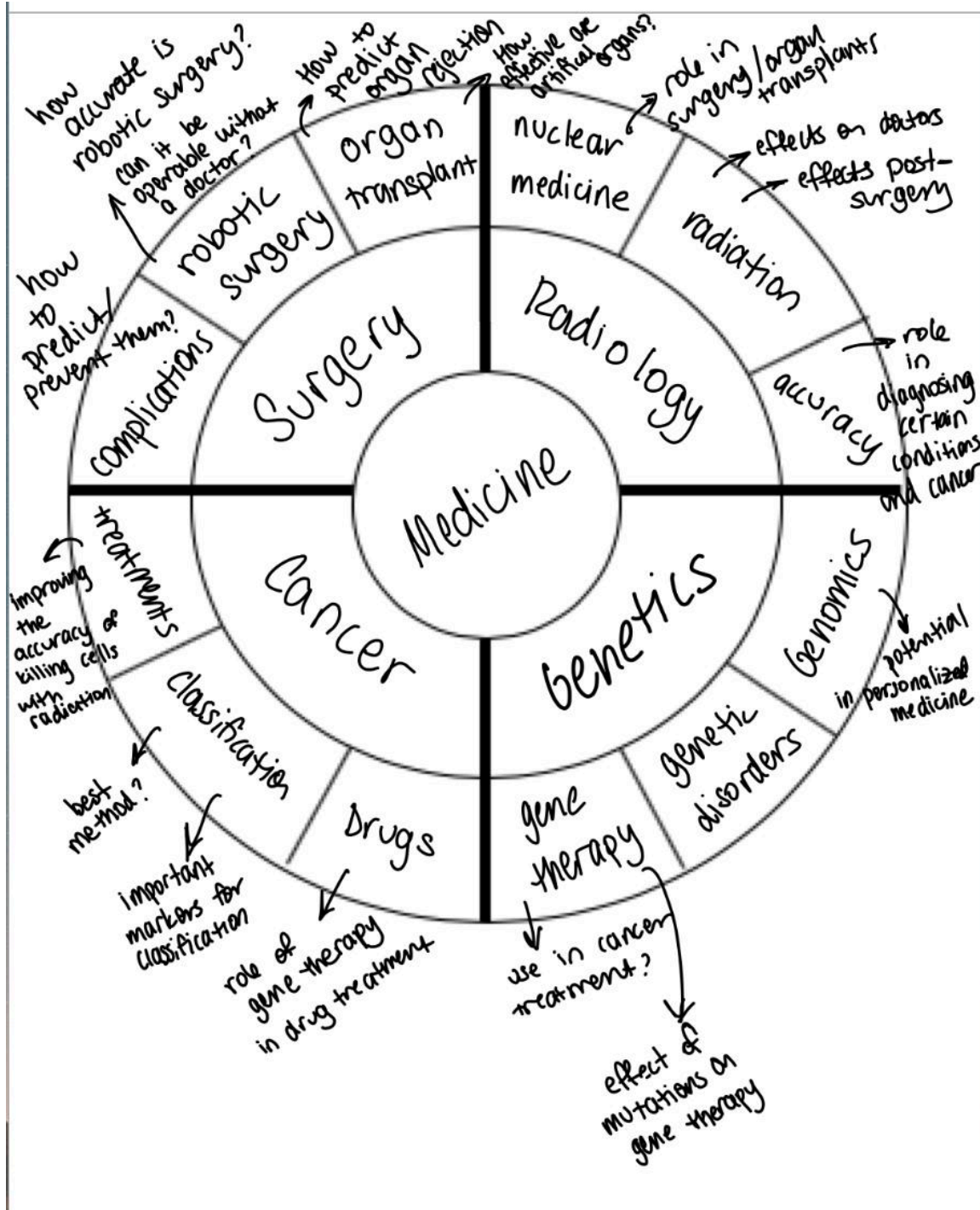
Things to keep in mind:

-You don't want to have too much blank space

-If you are adding a pre-printed graph or sketch, paste in and sign + date.

Brainstorming

Pie Diagrams:

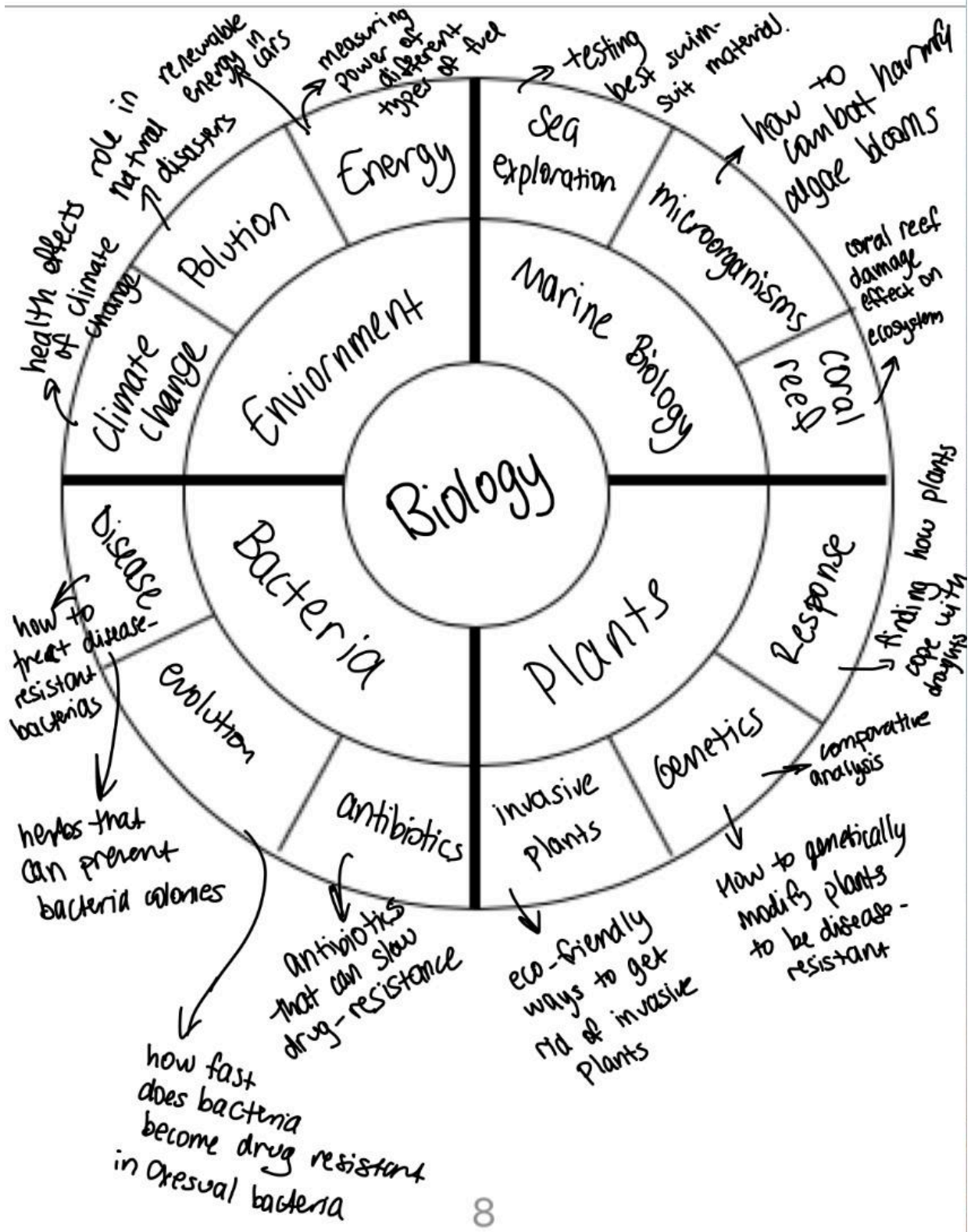


August 20, 2023

7:50pm

Samhitha Bodangi

Brainstorming Pie Diagram about Medicine





August 20, 2023

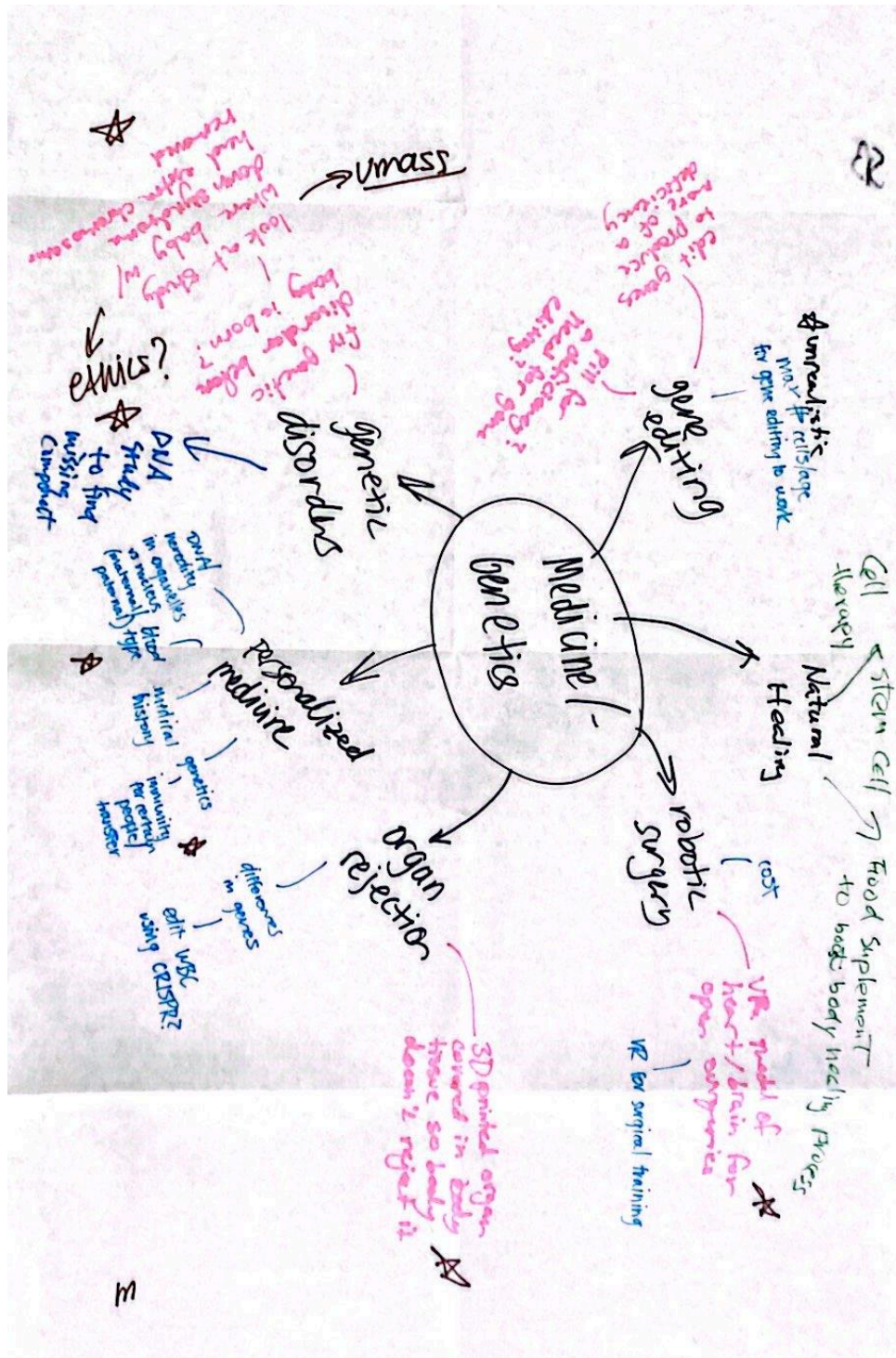
7:53pm

Samhitha Bodangi

Brainstroming Pie Diagram about Technology

These pie diagrams were done over the summer in preparation for the STEM project. Pie diagrams are a brainstorming method where the broad interest is placed in the center, surrounded with topics that are associated with the central idea. I chose to focus on medicine and technology. I used news stories and prior knowledge to come up with potential projects about each topic.

Mindmaps

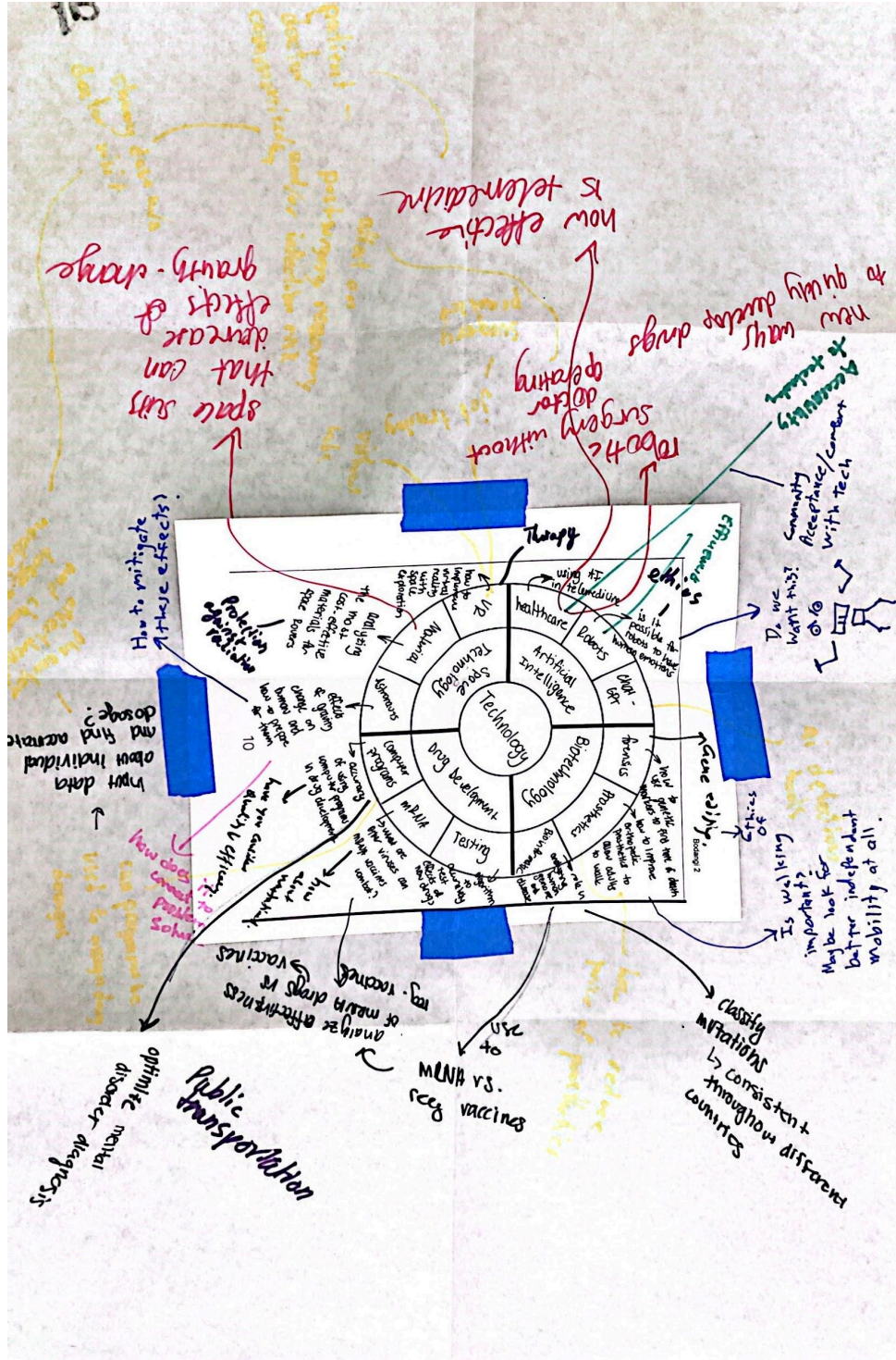


October 1, 2023

3:49pm

Samhitha Bodangi

Brainstorming Mindmap about Medicine/Genetics

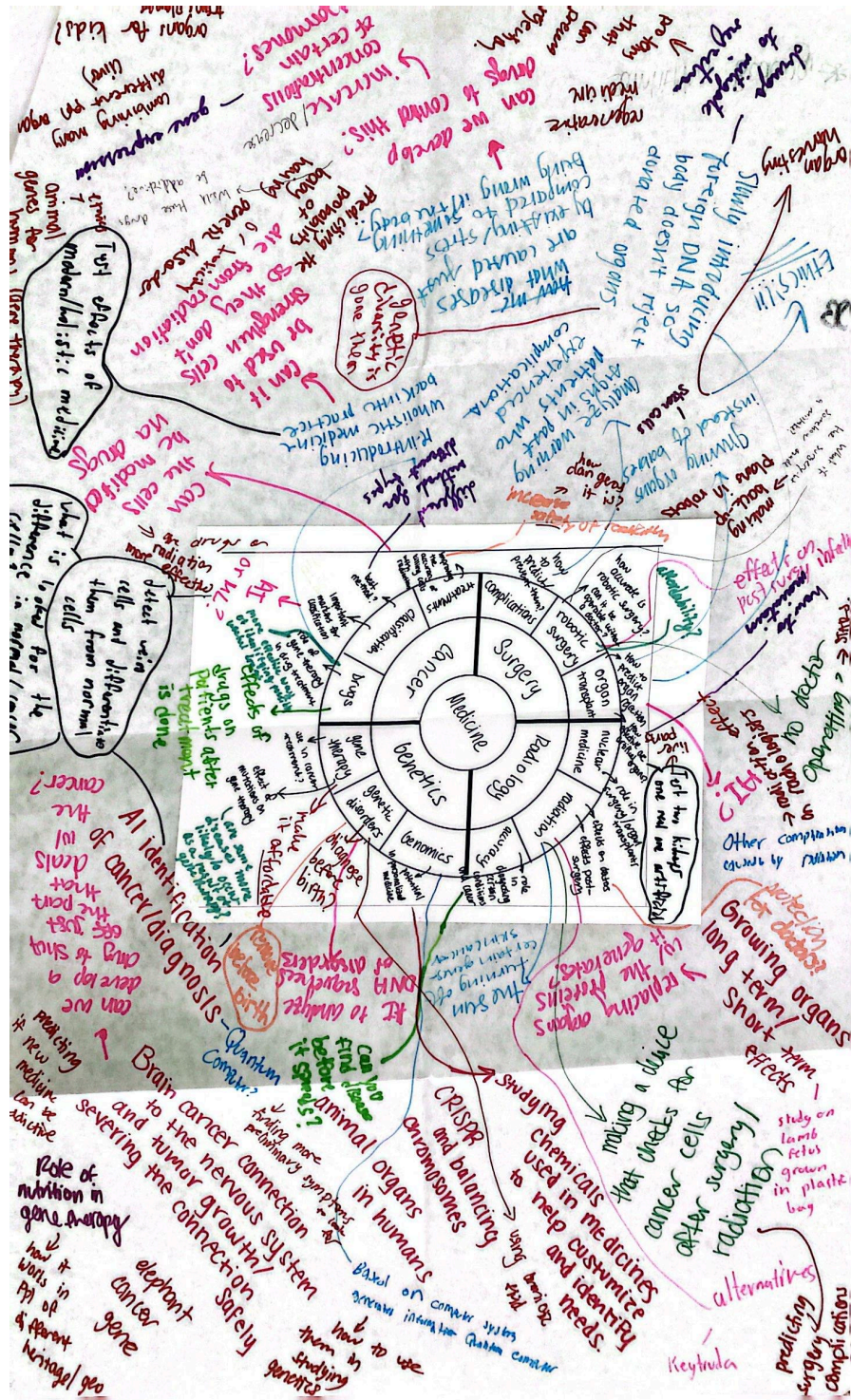


August 20, 2023

3:50pm

Samhitha Bodangi

Brainstorming Mindmap about Technology



October 1, 2023

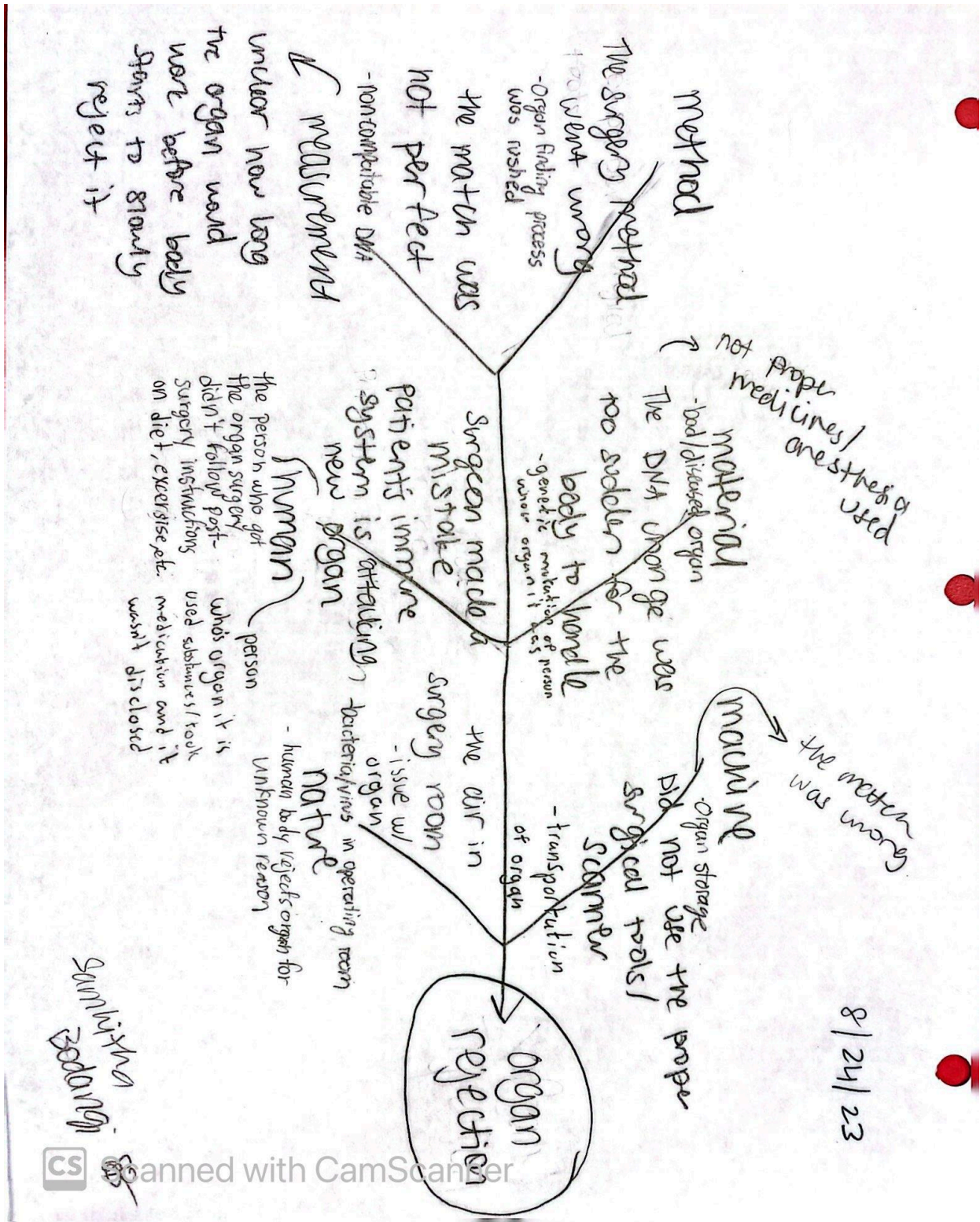
3:53pm

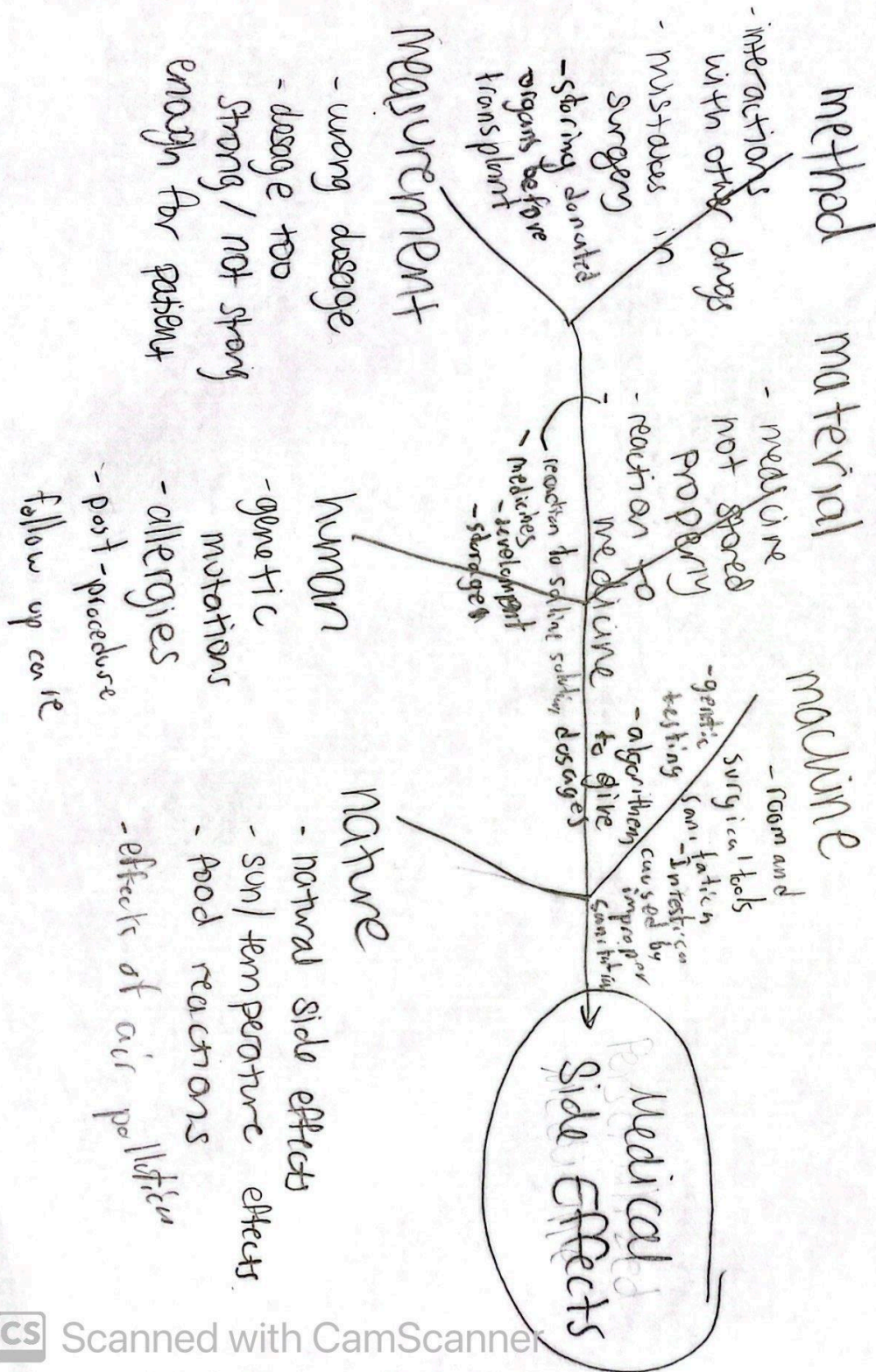
Samhitha Bodangi

Brainstorming Mindmap about Medicine

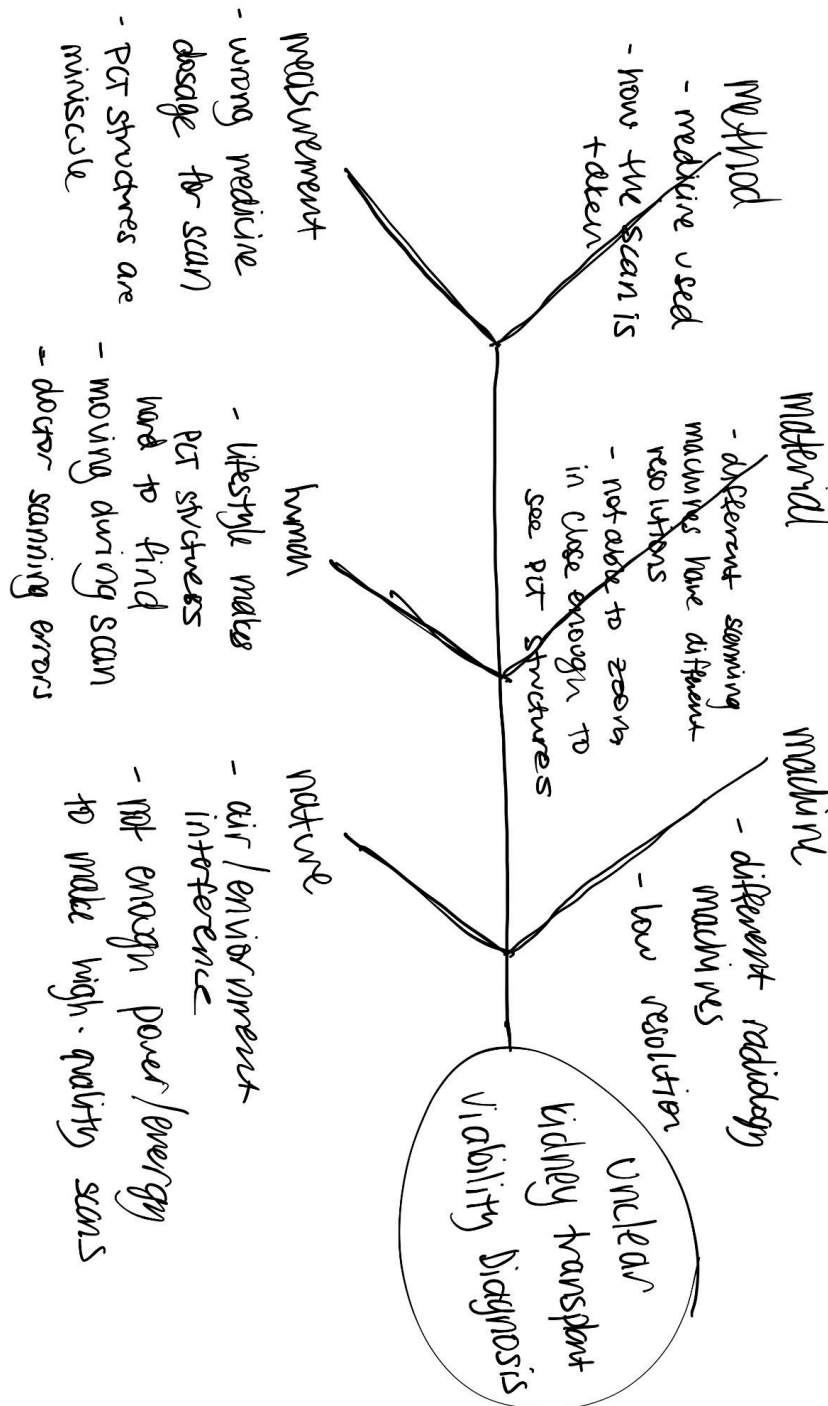
These mindmaps were done in school. After giving a quick pitch about my interests and potential topics, my classmates gave me ideas on project topics to pursue.

Fishbone Diagrams





CS Scanned with CamScanner



November 12, 2023

10:37 pm

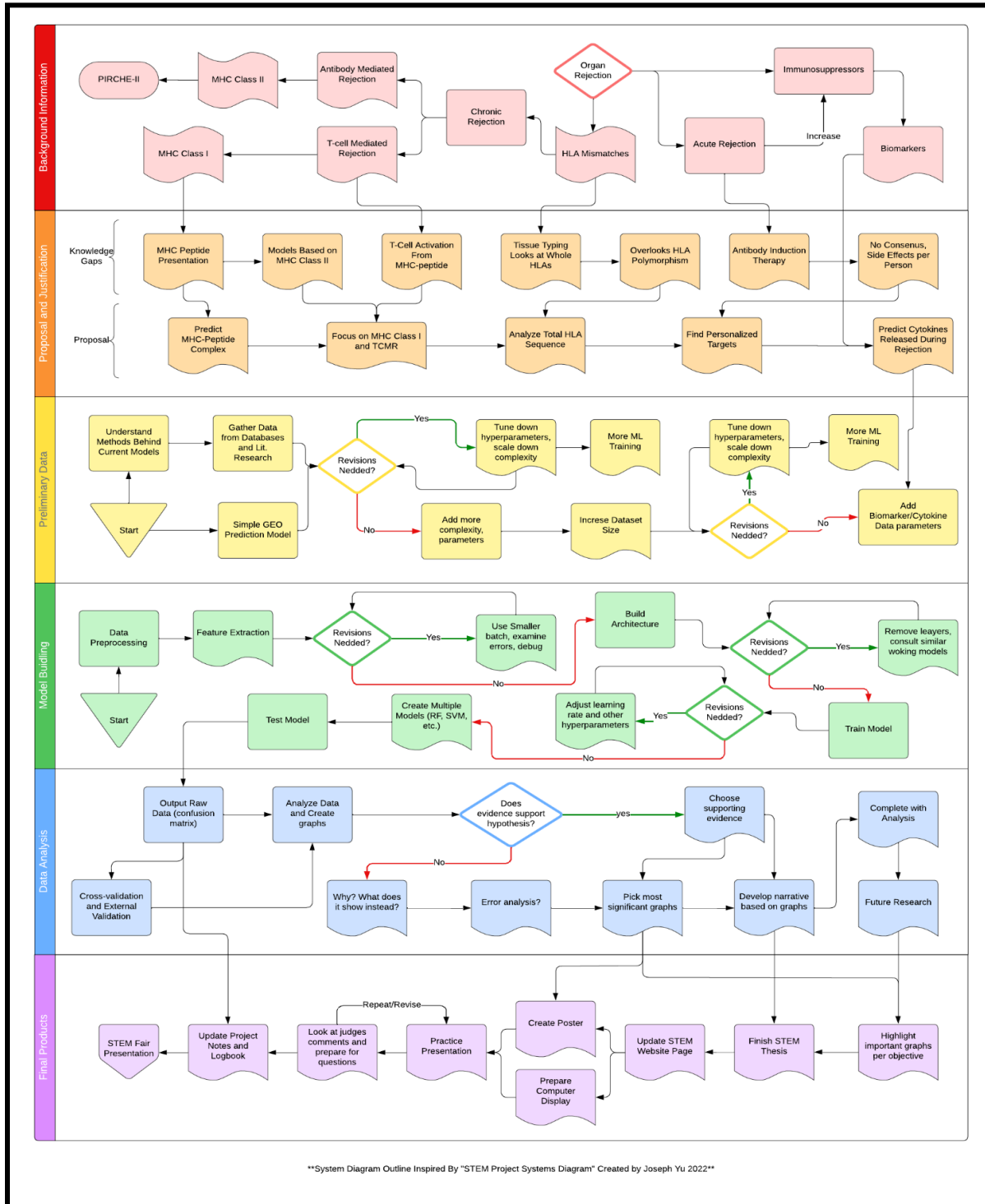
Samhitha Bodangi

Brainstorming FishBone Diagram about Diagnosing Kidney Transplant Viability

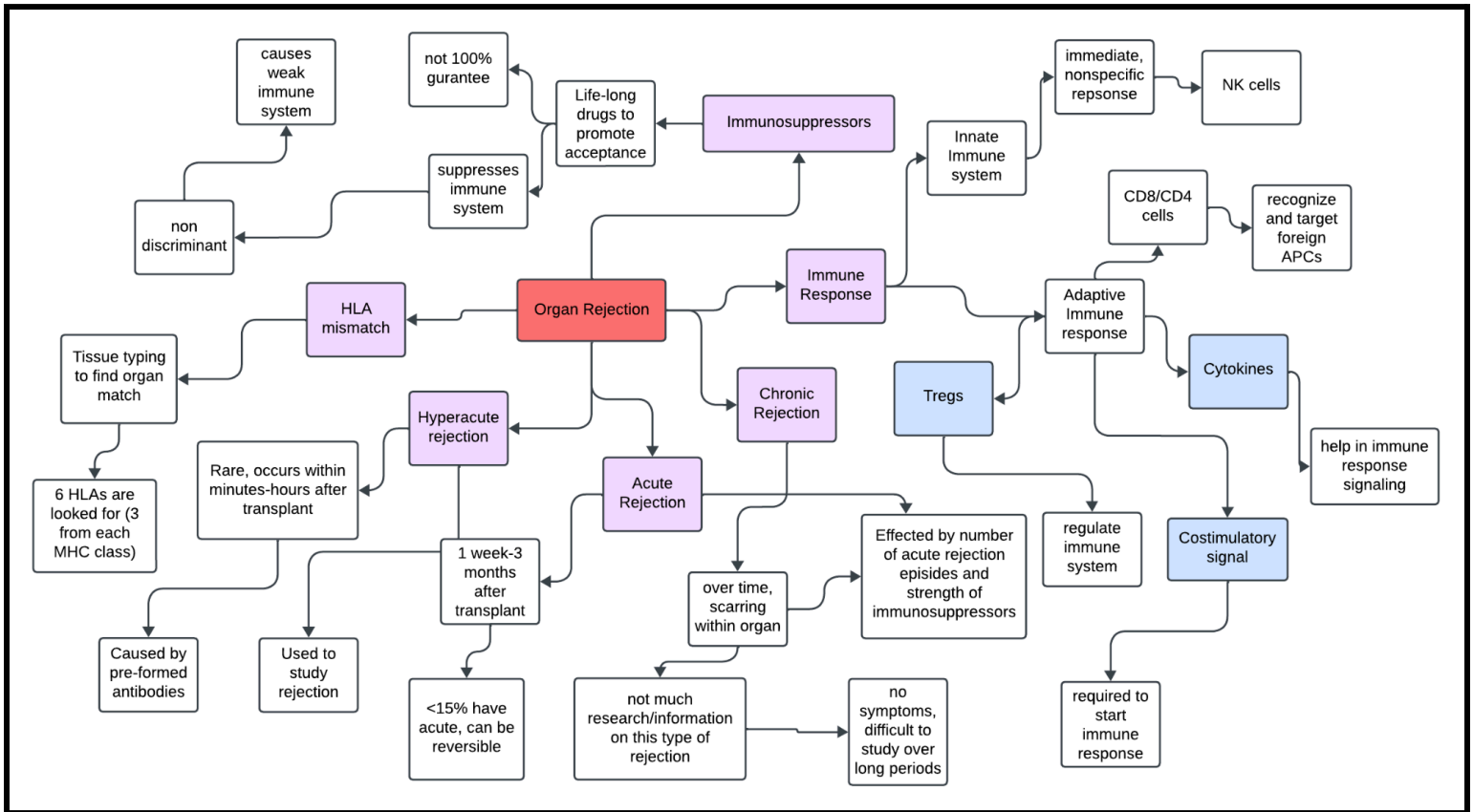
These are fishbone diagrams, a brainstorming method where we take a problem and brainstorm the different causes for that problem.

Systems Map

STEM I Project Outline Flowchart.pdf



Organ Rejection.png



November 19, 2023

1:28pm

Samhitha Bodangi

Organ Rejection Systems Diagram

These are systems diagrams of organ rejection and the overall research outline. System maps help visualize and understand the different components of a specific system.

Project Abstract:

Organ rejection is a dangerous medical complication that affects 50% of all kidney transplants five years post-transplant. Currently, all transplant patients are prescribed life-long immunosuppressors to decrease the risk of organ rejection. However, the side effects of 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 current HLA typing methods do not account for the diverse amino acid variations within each allele that can initiate rejection. By focusing on indirect allorecognition, one solution is to create a machine-learning model that can analyze donor and recipient HLA

alleles 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 HLA allele amino acid sequences and multiple servers to aid in the prediction methods. The result is that the model can accurately predict MHC-peptide complexes and rejection targets. In conclusion, focusing on indirect MHC-peptide presentation can account for HLA polymorphism, providing clinicians with greater insight into specific rejection pathways. Additionally, this data can be used to administer personalized and targeted immunosuppressors or decrease the need for broad immunosuppressors altogether. In the future, the model can be modified to support other organ transplants, positively contributing to the health of many future organ transplant recipients.

Keywords: Organ rejection, immune system, antibodies, cytokines, T cells, machine learning. T-cell mediated rejection, HLA class I, MHC-peptide complex

Project Introduction:

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.

Research Question: How can analyzing HLA sequences be used as a form of precision medicine to predict the risk of rejection and provide better targets for selective T-cell inhibition?

Engineering Objective: The objective is to make a machine learning model that can predict rejection, given donor and recipient HLA alleles. The model will work by identifying solvent-accessible amino acid mismatches. Then, the model will use these mismatches to predict donor-derived peptides that would bind to recipient MHC class II molecules. Using public databases and open-source servers, this model will predict the risk of rejection and provide information on targets for personalized immunosuppression.

Research Hypothesis: The model will be successful in predicting rejection as it focuses on the MHC-peptide complex on the donor organ that will initiate T-cell activation. By learning from current MHC-peptide predicting neural networks and public data, the model can accurately find HLA mismatches and potential immunosuppressive targets.

Brief Overview: Organ transplants are among the greatest advances in modern medicine, saving many lives every year. However, many medical complications may occur after the transplant,

such as organ rejection. Currently, all transplant patients are prescribed life-long immunosuppressors to decrease organ rejection. While these medications prevent organ rejection to an extent, about 10-20% of patients will still experience at least one episode of rejection. Additionally, they can also severely weaken the immune system, increasing the risk of cancer, infections, and other diseases. Rejection is primarily caused because of the Human Leukocyte Antigen (HLA) mismatches between the donor and the recipient. HLA genes are very polymorphic and classifying entire HLA mismatches does not account for the allele differences that can start rejection. The main objective is to understand how organ rejection can be decreased with selective T-cell inhibition by analyzing donor and recipient HLA sequences and predicting MHC-peptide complexes. Additionally, understanding the specific MHC-peptide complexes that will initiate rejection can provide greater insight into specific immunosuppressive targets. The model should accurately predict rejection and provide specific targets for precise immunosuppression and can be constructed within an open-source web application.

Research Outline

1. Collect data from publicly available databases, preferably in CSV or table-like format that can be imported as a dataframe
 - a. HLA allele typing data for donor and recipient and the status of rejection from U.N.O.S. to be used for testing and model validation
 - b. HLA amino acid sequence data from IPD-IMGT/HLA database
2. Align amino acid sequences for each HLA loci and find amino acid mismatched
 - a. Common algorithms such as Needleman-Wunsch Algorithm or BLAST
 - b. Any position where the amino acids differ between donor and recipient are identified as mismatches
3. Find solvent-accessible amino acids using NetSurfP server
4. Generate donor-derived peptides using solvent accessible amino acid mismatches from IEDB
 - a. Peptide length is 15-20 amino acids (for MHC class II)
5. Find most significant peptides using NetMHCIIpan for binding affinity
6. Develop scoring system with most significant peptide-binding affinities
7. Immunosuppressive targets are the most significant peptides that might cause rejection
8. Construct multiple models, such as Support Vector Machine, K-Nearest Neighbor, Random Forest, and Neural Networks
9. Test model using U.N.O.S. donor and recipient HLA typing samples to validate accuracy
10. Create user friendly, open-source web application that holds the model
 - a. Using Visual Studio Code and HTML to handle UI
11. Compare U.N.O.S. samples with competitor models (HLA-EMMA, PIRCHE-II, and HLAMatchmaker) to assess increased or decreased accuracy

Project Introduction References:

- Iglesias, M., Brennan, D. C., Larsen, C. P., & Raimondi, G. (2022, September 2). Frontiers | Targeting inflammation and immune activation to improve CTLA4-Ig-based modulation of transplant rejection. <https://www.frontiersin.org/articles/10.3389/fimmu.2022.926648/full>
- Kirk, A. D., Harlan, D. M., & Armstrong, N. N. (1997, August 5). CTLA4-Ig and anti-CD40 ligand prevent renal allograft rejection in primates | PNAS. <https://www.pnas.org/doi/full/10.1073/pnas.94.16.8789>
- Malhotra, P. (2023, July 11). *Immunology of Transplant Rejection: Overview, History, Types of Grafts*. Medscape. <https://emedicine.medscape.com/article/432209-overview?form=fpf>
- Matching and Compatibility | Transplant Center | UC Davis Health. (n.d.). Retrieved November 8, 2023, from <https://health.ucdavis.edu/transplant/livingkidneydonation/matching-and-compatibility.html>

Professional Communication:

Hello Dr. Lanese,

I am Samhitha Bodangi from the Massachusetts Academy of Math and Science at WPI. Currently, I am in the process of conducting a five-month research project regarding organ rejection. I have read many of your articles, and I found them very interesting, as they exposed me to some unique medical treatments. I read about the gene therapy eye drops, the new ultrasound for chemotherapy, and the gene therapy "syringes."

Your articles about organ rejection seemed very relevant to the goal of my project. I am researching to find the most optimal transplant method that can predict and prevent organ rejection. For example, your article about the three kidney transplants not needing immunosuppressors was very interesting. Additionally, I wanted to investigate more about the universal transplants you wrote about in another article.

Based on your experience, do you have any other specific transplant methods that can prevent organ rejection? I have read about gene targeting the organ, but I was wondering if it would be more efficient to target specific genes (potentially ones that make T-cells) or to focus on organ transplantation at the surgical level. I was also wondering what causes an organ match to be rejected. Based on your expertise, do you think there is a problem with the initial testing of the organ, or is it something more unpredictable? That way, I could focus more on either the diagnosis part or preventative measures, depending on where the problem stems from. Any information would be greatly appreciated in helping me get closer to achieving the goal of my project. Thank you; I look forward to hearing from you!

Thank you,
Samhitha Bodangi

Description: LIVE Science Author who wrote many articles about organ transplant and gene editing

Response: No response as of 11/4/2023

Hello, Dr. Zhang,

I am Samhitha Bodangi, and I am currently a junior at the Massachusetts Academy of Math and Science at WPI. Currently, I am in the process of conducting a 5-month-long research project related to organ rejection. I was extremely fascinated by your OCT device, which helps assess kidneys and their viability for organ transplants, and would love to learn more about it.

The long waitlists and the critical shortage of organs can be incredibly challenging and an emotionally taxing experience for many patients across the world. Yet, there is no guarantee that the patient will live a healthy life after transplantation, as many other health complications, such as organ rejection, can arise. I am trying to conduct a research project that will help decrease the risk of organ rejection or develop new tools to accurately predict the risk of organ rejection.

I recently read your abstract about using robotic-assisted OCT for pre-transplant kidney monitoring. I would love to learn more about your work and discuss some future advances I could potentially implement in my project. I would greatly appreciate the opportunity to work with you and use the resources at your lab to help progress my project.

I am passionate about medical research and finding better ways to treat different medical conditions. I have been wanting to gain experience and deeper knowledge in this specific area of study, and thus I am hoping to discuss my project with you.

Please let me know if we could discuss my potential involvement with your research and if there is any other information I should provide. I am readily available at this email or by phone at 508-667-9268. Thank you so much for your time and consideration. I look forward to hearing from you.

Sincerely,
Samhitha Bodangi

Description: Professor at WPI who made OCT to assess kidneys and their viability for transplantation

Response: Connected me with the post-doc student who worked on the device. I met with him and got a backup project idea/mentorship about PCT kidney structures and making an AI algorithm to detect those microstructures in a variety of focuses.

Hello Dr. Chen,

I am Samhitha Bodangi, and I am currently a junior at the Massachusetts Academy of Math and Science at WPI. Currently, I am in the process of conducting a 5-month-long research project related to organ rejection. I was extremely fascinated by your OCT device, which helps assess kidneys and their viability for organ transplants, and would love to learn more about it.

The long waitlists and the critical shortage of organs can be incredibly challenging and an emotionally taxing experience for many patients across the world. Yet, there is no guarantee that the patient will live a healthy life after transplantation, as many other health complications, such as organ rejection, can arise. I am trying to conduct a research project that will help decrease the risk of organ rejection or develop new tools to accurately predict the risk of organ rejection.

I recently read your abstract about using robotic-assisted OCT for pre-transplant kidney monitoring. I would love to learn more about your work and discuss some future advances I could potentially implement in my project. I would greatly appreciate the opportunity to work with you and use the resources at your lab to help progress my project.

I am passionate about medical research and finding better ways to treat different medical conditions. I have been wanting to gain experience and deeper knowledge in this specific area of study, and thus I am hoping to discuss my project with you.

Please let me know if we could discuss my potential involvement with your research and if there is any other information I should provide. I am readily available at this email or by phone at 508-667-9268. Thank you so much for your time and consideration. I look forward to hearing from you.

Sincerely,
Samhitha Bodangi

Description: Worked on the OCT device with Dr. Zhang

Response: Does not work with high school students

Hello Dr. Martins,

I am Samhitha Bodangi, and I am currently a junior at the Massachusetts Academy of Math and Science at WPI. Currently, I am in the process of conducting a 5-month-long research project related to organ rejection. I was extremely fascinated by publications related to liver transplants and would love to learn more about them.

The long waitlists and the critical shortage of organs can be incredibly challenging and an emotionally taxing experience for many patients across the world. Yet, there is no guarantee that the patient will live a healthy life after transplantation, as many other health complications, such as organ rejection, can arise. I am trying to conduct a research project that will help decrease the risk of organ rejection or develop new tools to accurately predict the risk of organ rejection.

I recently read your abstract about modifying organs with gene therapy. My project is very similar, and I would love to learn more about your work and discuss some future advances I could potentially implement in my project. I would greatly appreciate the opportunity to work with you and use the resources at your lab to help progress my project.

I am passionate about medical research and finding better ways to treat different medical conditions. I have been wanting to gain experience and deeper knowledge in this specific area of study, and thus I am hoping to discuss my project with you.

Please let me know if we could discuss my potential involvement with your research and if there is any other information I should provide. I am readily available at this email or by phone at 508-667-9268. Thank you so much for your time and consideration. I look forward to hearing from you.

Sincerely,
Samhitha Bodangi

Description: Professor of the Martins Lab at UMass, which specializes in transplantation immunology

Response: No response as of 10/31/23

Hello Dr. Movahedi,

I am Samhitha Bodangi, and I am currently a junior at the Massachusetts Academy of Math and Science at WPI. Currently, I am in the process of conducting a 5-month-long research project related to organ rejection. I was extremely fascinated by publications related to liver transplants and would love to learn more about them.

The long waitlists and the critical shortage of organs can be incredibly challenging and an emotionally taxing experience for many patients across the world. Yet, there is no guarantee that the patient will live a healthy life after transplantation, as many other health complications, such as organ rejection, can arise. I am trying to conduct a research project that will help decrease the risk of organ rejection or develop new tools to accurately predict the risk of organ rejection.

I recently read your abstract about graft-host disease in liver transplant recipients. My project is related to graft survival, and I would love to learn more about your work and discuss some future advances I could potentially implement in my project. I would greatly appreciate the opportunity to work with you and use the resources at your lab to help progress my project.

I am passionate about medical research and finding better ways to treat different medical conditions. I have been wanting to gain experience and deeper knowledge in this specific area of study, and thus I am hoping to discuss my project with you.

Please let me know if we could discuss my potential involvement with your research and if there is any other information I should provide. I am readily available at this email or by phone at 508-667-9268. Thank you so much for your time and consideration. I look forward to hearing from you.

Sincerely,
Samhitha Bodangi

Description: Professor of the Martins Lab at UMass, which specializes in transplantational immunology

Response: Responded after a follow-up email. Informed me that Martins Lab is shut down, and connected me with Dr.Mullens.

Hello, Dr. Mullens,

I am Samhitha Bodangi, and I met with Dr. Mohavedi a few days ago about my current independent research project. As he may have already told you, I am currently conducting an independent research project related to organ rejection.

Thank you for your email, and I understand that you must be very busy mentoring other students. However, do you have any availability for a quick phone call or meeting to discuss my research? I am currently in the process of forming the methodology for my project, and I would appreciate the opportunity to ask you for your advice and potential models I could study.

Again, I understand you may be very busy to give me a mentorship, but I am still interested in discussing my research if you are available.

I was interested in learning more about your liver organoid models, how you use them to study genes and how specific genetic changes affect the inflammation of the liver. These models seem to be very relevant to my project and what I would potentially monitor.

Please let me know if you have any availability for us to discuss my research and if there is any other information I should provide. I am readily available at this email or by phone at [508-667-9268](tel:508-667-9268). Thank you for your time, and I look forward to hearing from you.

Sincerely,
Samhitha Bodangi

Description: Dr. Mohavedi referred me to Dr. Mullens from the UMass Gastroenterology Mullen Lab. Unfortunately, Dr. Mullens is very busy, and this is a follow-up email to ask for a quick meeting

Dear Charles River Laboratories,

I am Samhitha Bodangi, and I am currently a junior at the Massachusetts Academy of Math and Science at WPI. Currently, I am in the process of conducting a 5-month-long research project related to organ rejection.

The long waitlists and the critical shortage of organs can be incredibly challenging and an emotionally taxing experience for many patients across the world. Yet, there is no guarantee that the patient will live a healthy life after transplantation, as many other health complications, such as organ rejection, can arise. I am trying to conduct a research project that will help decrease the risk of organ rejection or develop new tools to accurately predict the risk of organ rejection. I would greatly appreciate the opportunity to work with you and use the resources at your lab to help progress my project.

I am passionate about medical research and finding better ways to treat different medical conditions. I have been wanting to gain experience and deeper knowledge in this specific area of study, and thus I am hoping to discuss my project with you.

Please let me know if we could discuss my potential involvement with your lab and if there is any other information I should provide. I am readily available at this email or by phone at [508-667-9268](tel:508-667-9268). Thank you so much for your time and consideration. I look forward to hearing from you.

Sincerely,
Samhitha Bodangi

Description: Community lab in Worcester, MA

Response: Met with him on 11/6/23, and informed me about the methodology and specific technology I would need to conduct my experiments.

Hello Dr. Keeler,

I am Samhitha Bodangi, and I am currently a junior at the Massachusetts Academy of Math and Science at WPI. Currently, I am in the process of conducting a 5-month-long research project related to organ rejection and T-cells. I was extremely fascinated by your publication related to gene editing CAR T cells and would love to learn more about it.

The long waitlists and the critical shortage of organs can be incredibly challenging and an emotionally taxing experience for many patients across the world. Yet, there is no guarantee that the patient will live a healthy life after transplantation, as many other health complications, such as organ rejection, can arise. I am trying to conduct a research project that will help decrease the risk of organ rejection by selectively inhibiting T-cell activation with a personalized medicine approach.

I recently read your abstract about modulating immune responses to AAV by polyclonal Tregs and CAR Treg cells. My project is related to precision medicine and T-cells, and I would love to learn more about your work and discuss some future advances I could potentially implement in my project. Even though your research is related to AAV, I have read multiple papers related to CAR T-cells and increasing Tregs to prevent rejection. I wanted to investigate whether CAR T-cells can be modified to be more resistant to antigens they should not attack to prevent organ rejection. Given your work in CAR T-cell research, I would greatly appreciate the opportunity to work with you and use the resources at your lab to help progress my project.

I am passionate about medical research and finding better ways to treat different medical conditions. I have been wanting to gain experience and deeper knowledge in this specific area of study, and thus I am hoping to discuss my project with you.

Please let me know if we could discuss my potential involvement with your research and if there is any other information I should provide. I am readily available at this email or by phone at 508-667-9268. Thank you so much for your time and consideration. I look forward to hearing from you.

Sincerely,
Samhitha Bodangi

Description: Dr. Keeler is the professor at the Keeler Lab at Umass. Conducts research related to CAR T-cells and immunotherapy in pediatrics.

Response:

Hello Dr. Kent,

I am Samhitha Bodangi, and I am currently a junior at the Massachusetts Academy of Math and Science at WPI. Currently, I am in the process of conducting a 5-month-long research project related to organ rejection and T-cells. I was extremely fascinated by your publication related to tolerogenic nanoparticles and T-cell immunotherapy and would love to learn more about it.

The long waitlists and the critical shortage of organs can be incredibly challenging and an emotionally taxing experience for many patients across the world. Yet, there is no guarantee that the patient will live a healthy life after transplantation, as many other health complications, such as organ rejection, can arise. I am trying to conduct a research project that will help decrease the risk of organ rejection by selectively inhibiting T-cell activation with a personalized medicine approach.

I recently read your abstract about the tolerogenic nanoparticles that inhibit T cell-mediated autoimmunity through SOCS2. My project is related to T-cell inhibition and the immune system, and I would love to learn more about your work and discuss some future advances I could potentially implement in my project. Even though your research is related to Type 1 diabetes, I have read multiple papers related to increasing Treg generation to prevent rejection. I wanted to investigate different ways to selectively inhibit the activation and proliferation of T cells. Given your work in T-cell inhibition research, I would greatly appreciate the opportunity to work with you and use the resources at your lab to help progress my project.

I am passionate about medical research and finding better ways to treat different medical conditions. I have been wanting to gain experience and deeper knowledge in this specific area of study, and thus I would like to discuss my project with you.

Please let me know if we could discuss my potential involvement with your research and if there is any other information I should provide. I am readily available at this email or by phone at 508-667-9268. Thank you so much for your time and consideration. I look forward to hearing from you.

Sincerely,
Samhitha Bodangi

Description: Dr. Kent is a professor at Umass who works with T-cell inhibition and downregulating the immune system. However, her research is on Type 1 diabetes

Response: Referred me to Dr. Harlen who gave an article related to costimulatory blockade and meeting on 11/13/23

Hello Dr. Brehm,

I am Samhitha Bodangi, and I am currently a junior at the Massachusetts Academy of Math and Science at WPI. Currently, I am in the process of conducting a 5-month-long research project related to organ rejection and T-cells. I was extremely fascinated by your publication related to the early attrition of T Cells and costimulation blockade and would love to learn more about it.

The long waitlists and the critical shortage of organs can be incredibly challenging and an emotionally taxing experience for many patients across the world. Yet, there is no guarantee that the patient will live a healthy life after transplantation, as many other health complications, such as organ rejection, can arise. I am trying to conduct a research project that will help decrease the risk of organ rejection by selectively inhibiting T-cell activation with a personalized medicine approach.

I recently read your abstract about the early attrition of memory T cells during inflammation and costimulation blockade regulated by the proteins Fas and Bim. My project is related to T-cell inhibition and the immune system, and I would love to learn more about your work and discuss some future advances I could potentially implement in my project. Even though your research is related to Type 1 diabetes, I noticed that this paper was in relation to allograft survival. I have read multiple papers blocking the costimulation signal to prevent rejection. I want to investigate different ways to selectively inhibit the activation and proliferation of T cells, specifically by costimulation blockade. Given your work in T-cell inhibition research, I would greatly appreciate the opportunity to work with you and use the resources at your lab to help progress my project.

I am passionate about medical research and finding better ways to treat different medical conditions. I have been wanting to gain experience and deeper knowledge in this specific area of study, and thus I am hoping to discuss my project with you.

Please let me know if we could discuss my potential involvement with your research and if there is any other information I should provide. I am readily available at this email or by phone at 508-667-9268. Thank you so much for your time and consideration. I look forward to hearing from you.

Sincerely,
Samhitha Bodangi

Description: Dr. Brehm is a professor at Umass who works with T-cell downregulation. His research is mainly on type 1 diabetes, but the referenced paper is about tissue survival.

Response: No response as of 12/12/23

Hello Professor Stern,

I am Samhitha Bodangi, and I am currently a junior at the Massachusetts Academy of Math and Science at WPI. Currently, I am in the process of conducting a 5-month-long research project related to organ rejection and T-cells. I was extremely fascinated by your publication related to MHC class II peptide loading to regulate Tregs and would love to learn more about it.

The long waitlists and the critical shortage of organs can be incredibly challenging and an emotionally taxing experience for many patients across the world. Yet, there is no guarantee that the patient will live a healthy life after transplantation, as many other health complications, such as organ rejection, can arise. I am trying to conduct a research project that will help decrease the risk of organ rejection by selectively inhibiting T-cell activation with a personalized medicine approach.

I recently read your abstract about MHC class II peptide loading and its role in regulating the selection and function of Tregs. My project is related to the MHC complex and the immune system, and I would love to learn more about your work and discuss some future advances I could potentially implement in my project. Even though your research is related to cancer cells, I have read many papers about increasing Treg populations to promote allograft survival. I want to investigate different ways to increase Tregs, specifically by using proteins or peptides based on the organ's MHC complex. Given your work in MHC research, I would greatly appreciate the opportunity to work with you and use the resources at your lab to help progress my project.

I am passionate about medical research and finding better ways to treat different medical conditions. I have been wanting to gain experience and deeper knowledge in this specific area of study, and thus I am hoping to discuss my project with you.

Please let me know if we could discuss my potential involvement with your research and if there is any other information I should provide. I am readily available at this email or by phone at 508-667-9268. Thank you so much for your time and consideration. I look forward to hearing from you.

Sincerely,
Samhitha Bodangi

Description: Dr. Stern is a professor at Umass who works with proteins and T cells. His work is mainly focused on cancer, but his research related to T cells is similar to organ rejection studies.

Response: Met with him on 12/6/23 and gave me databases for antigen peptide prediction

Hello Dr. Politz,

I am Samhitha Bodangi, and I am currently a junior at the Massachusetts Academy of Math and Science at WPI. Currently, I am in the process of conducting a 5-month-long research project related to organ rejection and T-cells. I was extremely fascinated by your publication related to nematomucin antigen expression using anti-peptide antibodies and would love to learn more about it.

The long waitlists and the critical shortage of organs can be incredibly challenging and an emotionally taxing experience for many patients across the world. Yet, there is no guarantee that the patient will live a healthy life after transplantation, as many other health complications, such as organ rejection, can arise. I am trying to conduct a research project that will help decrease the risk of organ rejection by selectively inhibiting T-cell activation with a personalized medicine approach.

I recently read your abstract about using anti-peptide antibodies to identify nematomucin antigen expression. My project is related to the immune system and antigen presentation, and I would love to learn more about your work and discuss some future advances I could potentially implement in my project. I want to give specific regimens based on organ antigen presentation to decrease rejection while maintaining the integrity of the whole system. Given your work in antigen research, I would greatly appreciate the opportunity to work with you and use the resources at your lab to help progress my project.

I am passionate about medical research and finding better ways to treat different medical conditions. I have been wanting to gain experience and deeper knowledge in this specific area of study, and thus I am hoping to discuss my project with you.

Please let me know if we could discuss my potential involvement with your research and if there is any other information I should provide. I am readily available at this email or by phone at 508-667-9268. Thank you so much for your time and consideration. I look forward to hearing from you.

Sincerely,
Samhitha Bodangi

Description: Dr. Politz is a professor at WPI.

Response: He is a retired professor

Hello, Dr. Brownwell

I am Samhitha Bodangi, and I am currently a junior at the Massachusetts Academy of Math and Science at WPI. Currently, I am in the process of conducting a 5-month-long research project related to organ rejection and T-cells. I was extremely fascinated by your lab research and would love to learn more about them.

The long waitlists and the critical shortage of organs can be incredibly challenging and an emotionally taxing experience for many patients across the world. Yet, there is no guarantee that the patient will live a healthy life after transplantation, as many other health complications, such as organ rejection, can arise. I am trying to conduct a research project that will help decrease the risk of organ rejection by creating proteins that can selectively inhibit T-cell activation with a personalized medicine approach.

I noticed that you research organic chemistry and creating new polymers. My project involves analyzing the antigens present in donor organ cells and potentially creating specific proteins that can block the costimulatory signal between the antigen and T cells. I am interested in learning about the specific software I can use to analyze and create these proteins.

I would greatly appreciate the opportunity to work with you and use the resources at your lab to help progress my project. I am passionate about medical research and finding better ways to treat different medical conditions. I have been wanting to gain experience and deeper knowledge in this specific area of study, and thus I am hoping to discuss my project with you.

Please let me know if we could discuss my potential involvement with your research and if there is any other information I should provide. I am readily available at this email or by phone at 508-667-9268. Thank you so much for your time and consideration. I look forward to hearing from you.

Sincerely,
Samhitha Bodangi

Description:

Response: Met on 11/19/23

Materials and Methods:

Materials List:

Software:

- Google Colaboratory (Python programming language) Colab Notebooks
- Rstudio (R programming language) <https://posit.co/download/rstudio-desktop/>
- GitHub <https://github.com/samhithabodangi/Organ-Rejection-Model>
- Microsoft Excel <https://www.office.com/launch/Excel?ui=en-US&rs=US>
- Statistical Analysis System (SAS) <https://welcome.oda.sas.com/>
- Visual Studio Code <https://code.visualstudio.com/>

Servers:

- NetMHCpan 4.0
- Immune Epitope Database (IEDB)
- NetSurfP
- HLA-EMMA
- HLAMatchmaker
- PIRCHE-II

Datasets:

- Gene Expression Omnibus (GEO)
- United Network for Organ Sharing (U.N.O.S.) STAR Files
- IPD-IMGT/HLA from the European Bioinformatics Institute (EBI)
- HLA-Epi Dataset for compatibility score output

Data Collection

1. Donor and recipient HLA type data (HLA-A, HLA-B, and HLA-C alleles)
 - a. From U.N.O.S. STAR Files (given as a link to a folder in <https://app.box.com/>)
 - b. Contains donor and recipient HLA alleles and rejection outcomes for KT
 - c. Has data available since 1987, and has HLA data with different resolutions
2. HLA-Epi Dataset
 - a. From the HLA-Epi repository: <https://gitlab.univ-nantes.fr/crtiteam5/easy-hla>
 - b. Has donor and recipient alleles and the compatibility score for both HLA-Epi and the PIRCHE-II model
 - i. Use the PIRCHE-II model score as the model also focuses on indirect allorecognition, and HLA-Epi focuses on direct recognition
3. Peptide-MHC Binding Affinity Predictions
 - a. In silico predictions of which peptides are likely to bind to specific MHC class I molecules

- b. NetMHCpan (<https://services.healthtech.dtu.dk/services/NetMHC-4.0/>)
 - i. Takes a peptide sequence as input and predicts how strongly it will bind to a specific MHC class I molecule. Based on Artificial Neural Networks
 - c. HLAMatchmaker (<http://www.epitopes.net/>)
 - i. Compares HLA sequences and calculates number of mismatched epitopes and their immunogenicity. Based on experimental data.
 - d. HLA-EMMA (<https://hla-emma.com/>)
 - i. Focuses on amino acid mismatches between donor and recipient HLA sequences.
4. HLA Sequence Data
- a. IPD-IMGT/HLA (<https://www.ebi.ac.uk/ipd/imgt/hla/>)
 - i. Provides HLA allele-specific information such as HLA sequences and known peptide-binding motifs
 - b. Immune Epitope Database (<https://www.iedb.org/>)
 - i. Contains a vast repository of experimentally validated T cell epitopes, including those associated with transplantation and TCMR.
 - c. NetSurfP (<https://services.healthtech.dtu.dk/services/NetSurfP-3.0/>)
 - i. Predicts the solvent accessibility of amino acids in an amino acid sequence.
5. Gene Expression data (will be used for external validation)
- a. Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>)
 - b. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE212160>
 - i. GSE212160 has 1395 samples of renal biopsies, with samples for no rejection, Antibody-mediated rejection, T-cell mediated rejection, and no rejection. 532 no rejection biopsies and 437 TCMR renal biopsies
 - ii. Platform: GPL30305 NanoString Human Organ Transplant Panel
 - c. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21374>
 - i. GSE21374 has 282 renal biopsy samples. 76 rejection samples and 206 no rejection samples. Has the time (in days) of rejection and of the biopsy
 - ii. Platform: GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
 - d. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE112927>
 - i. GSE112927 has 235 samples of PBMCs for acute rejection in kidney transplants.
 - ii. Platform: GPL20301 Illumina HiSeq 4000 (Homo sapiens)
 - e. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47755>
 - i. GSE47755 has 528 samples of PBMCs from kidney transplants. Has tolerant, stable, and chronic rejection samples.
 - ii. Platform: GPL8798 Human oligo array from MWG cancerochips_v2009

Procedure:

HLA Amino Acid Sequence Data Preprocessing:

1. Import the necessary libraries such as pandas, numpy and sklearn commands
2. Download HLA Sequence data from IPD-IMGT/HLA from the ftp server
 - a. HLA amino acid sequences will be extracted for HLA-A, -B, -C, -DRB1, -DRB3, -DRB4, -DRB5, -DQA1, -DQB1, -DPA1 and -DPB1 up to the 2-field resolution
3. Upload HLA sequence data as fasta file into Jupyter Notebook
4. Convert file as a pandas dataframe into Google Colab file using pandas library

Simple Model Building for Preliminary Data:

1. Test methodology by manually performing hypothesized procedure
2. Obtain sample donor and recipient HLA typing data file from HLA-EMMA
3. Using the IPD/IMGT-HLA database, find the corresponding amino acid sequences for the sample typing data
4. Google Colab to find mismatches by vertically aligning sequences and finding unique mismatches
5. Use NetSurfP to filter mismatches to the solvent-accessible ones
6. NetMHCIIpan to input donor sequence and recipient class II alleles
7. Filter strong binding peptides with ones containing solvent accessible mismatches
8. Analyze peptides
 - a. Find repetitive peptides → immunosuppressive targets
 - b. Number of strong peptides → risk of organ rejection

(Aim 1) Analyze HLA Sequence Mismatches Between Donor and Recipient:

1. Extract HLA Sequence Amino Acid Based on Donor and Recipient HLA alleles
 - a. Access IPD-IMGT/HLA database using the specific HLA alleles identified for both the donor and recipient.
 - b. Database provides the corresponding amino acid sequences for each of those alleles
2. Sequence Alignment for Amino Acid Comparison
 - a. Align donor and recipient amino acid sequences using common alignment algorithms (Needleman-Wunsch Algorithm)
 - b. Use Basic Local Alignment Search Tool (BLAST) method to compare nucleotide or protein sequences to sequences in a database. BLAST finds regions of similarity between sequences and calculates the statistical significance of matches.
 - c. Any positions where the amino acids differ between the donor and recipient are identified and stored as mismatches
3. Mismatch Feature Extraction and Solvent Accessibility

- a. Use NetSurfP to predict solvent accessibility for each amino acid residue in the sequence
- b. Identifies mismatches at highly exposed residues with greater potential for immune recognition and risk of rejection.
- c. Accessibility score categories: buried or exposed

(Aim 2) Use Donor-Recipient HLA Sequence to Predict MHC-Peptide Complex:

1. Generate all possible peptides from donor alleles using allele fasta sequence
 - a. Access NetMHCIIpan from the IEDB database
2. Find binding affinity and eluted ligand scores of peptides to recipient MHC class II allele
 - a. The Score threshold for strong binders is <1 , which is the commonly used Frank threshold by NetMHCIIpan
3. Filter peptides with ones classified as SB (strong binding)
4. Filter SB peptides with ones that have solvent-accessible mismatches
 - a. These peptides have the highest chance for immunogenicity

(Aim 3) Determine Optimal Targets for Immunosuppression and Predict Rejection:

1. Repeat Aim 1 and Aim 2 for all donor sequences
 - a. Count all peptides that have the highest chance of immunogenicity
2. Higher number of donor-derived peptides that could bind with recipient alleles means there is a higher chance of rejection
3. Create a regression or binary classification model that is trained on datasets that have compatibility scores or the outcome of rejection
4. Model will develop a scoring system based on the number of mismatched peptides and the rate of rejection
 - a. Use algorithms such as SHAP or LIME to predict which recipient HLA molecules have the most weight in predicting rejection.
 - b. The peptides that bind to those recipient molecules are the most significant peptide targets for precise immunosuppression

U.N.O.S. HLA Donor and Recipient Data Preprocessing:

1. U.N.O.S. STAR files must be opened in the Statistical Analysis System (SAS), as the files are in the .sas file type format
2. U.N.O.S. data has data on multiple different organ transplants. The data on *living* kidney transplants will be considered
 - a. Kidneys are the most transplanted organ, and living transplants are frequent
 - b. An attempt to eliminate the possibility of external factors contributing to rejection as much as possible
3. U.N.O.S. files will be split into smaller subsets of data with specific resolutions

- a. For example, subsets with 2-field resolution and 4-field resolution will be constructed
4. Patient samples with a substantial amount of HLA allele data missing will be omitted
 - a. However, samples with few missing cells will be considered. In some cases, the absence of an allele may indicate that the individual is homozygous for the other allele at that locus.
5. Data files will be converted into an Excel file format (.xlsx), to make it feasible to upload it to Google Colab as a panda's data frame

Model Training:

1. Using UNOS processed dataset, create regression models that predict rejection to assess the model's accuracy and find compatibility score
 - a. The compatibility score will be validated by using a PIRCHE-II score or HLA-Epi score dataset
2. Split data into 80% testing, 10% validation, and 10% testing
3. Models to construct: Support vector machine, random forest, K-Nearest Neighbor, logistic regression
4. Compare models with common accuracy metrics

Data Analysis:

1. Focus on sensitivity rather than specificity (use in conjunction with accuracy)
2. Decision Matrix comparing the different algorithms
 - a. Models will be assessed with accuracy, F1 score, precision, recall, and AUC score
3. Each model will have a confusion matrix, showing the true positives, true negatives, false positives, and false negatives of the model's predictions.
4. An ROC curve will be made and AUC score will be calculated for each model
5. Brier score measures mean squared difference between predicted probabilities and the actual outcomes
6. Conduct statistical testing for significant results (most likely a T-test)
 - a. Run the optimal model and the 2nd most optimal model 5 times, keeping track of the accuracy (n=5)
 - b. Find the means of both trials, and perform T-test with calculator

Creating an Open-Source Web Application:

1. Visual Studio Code will be used to create a web application that can output the rejection predictions based on user input of HLA alleles
2. Page layout will be designed in Figma
 - a. Figma dev mode will be used to convert Figma designs into HTML and CSS code using Figma extension in Visual Studio Code
3. HTML and CSS will be used to create the page layout and UI as designed in Figma

- a. Allele inputs for the donor and recipient for commonly typed HLAs: HLA-A, HLA-B, HLA-C, DPA1, DPB1, DQA1, DQB1, DRB1, DRB3, DRB4, DRB5
- b. Output display, with elements that display prediction and allow for more information about the prediction
4. Add code to handle user input (HLA alleles) and pass to ML
5. Set up database using Django with allele sequences from IPD-IMGT/HLA and peptide affinities from NetMHCpan
 - a. Django is a free and open-source Python web framework designed to develop database-driven websites
6. Embed best ML model based on decision matrix into web app
7. Write code to handle user input (HLA alleles) and pass to ML
8. Deploy web application by uploading code into GitHub

External Validation:

1. Test web app using individual U.N.O.S. samples to validate its function as equivalent to model (approximately 80)
2. Import same U.N.O.S. samples into competitor models to compare results
 - a. HLA-EMMA gives amino acid mismatches
 - b. PIRCHE-II model gives a score based on mismatched epitopes
 - c. HLAMatchmaker finds eplet mismatches (for AMR)
3. Create histogram graphs to compare results of each model

Background:

Supply vs. Demand

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 Medicine*, 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

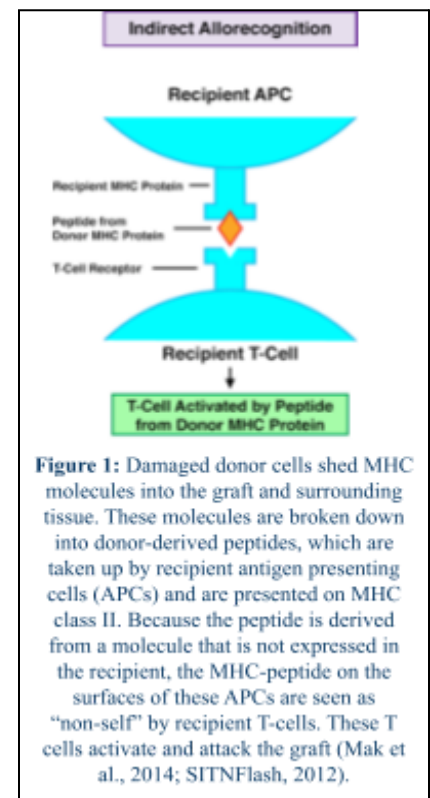
Early chronic organ rejection is primarily caused by T-cell-mediated rejection (Chong, 2020). T-cells are a type of immune cell that play a crucial role in identifying and eliminating foreign cells. When T-cells misinterpret donated organ cells as foreign, it can lead to T-cell activation and an attack 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" (*General, Non-Specific Defenses Against Infection*, n.d.). There are two main types of MHC molecules: MHC class I and MHC class II molecules. While MHC class I molecules are found on 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.

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).

Tissue Typing and Immune Profiling

When looking for organ matches, doctors perform Human Leukocyte Antigen ¹ (HLA) typing to understand the similarity in



¹ The human leukocyte antigen (HLA) complex is synonymous with the human MHC. The main difference is that MHC is found in all vertebrates, while HLA is only found in humans (Viatte, 2023).

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 high 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 more personalized and accurate.

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.

Background References:

- Chong, A. S. (2020). B cells as antigen-presenting cells in transplantation rejection and tolerance. *Cellular Immunology*, 349, 104061. <https://doi.org/10.1016/j.cellimm.2020.104061>
- Costa, C. D. (2019, August 26). What Is Machine Learning & Deep Learning? *Medium*. <https://medium.com/@clairedigitalogy/what-is-machine-learning-deep-learning-7788604004d>
- 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>

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>

General, Non-specific Defenses Against Infection. (n.d.). Defense Mechanism. Retrieved January 26, 2024, from

https://sphweb.bumc.bu.edu/otlt/mph-modules/ph/ph709_defenses/ph709_defenses_print.html

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>

Hewitt, E. W. (2003). The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology*, 110(2), 163–169.

<https://doi.org/10.1046/j.1365-2567.2003.01738.x>

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>

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>

Lakna. (2018, February 13). *Difference Between MHC Class 1 and 2 | Definition, Structure, Antigen Presentation, Similarities and Differences*. Pediaa.Com.

<https://pediaa.com/difference-between-mhc-class-1-and-2/>

Mak, T. W., Saunders, M. E., & Jett, B. D. (Eds.). (2014). Chapter 17 - Transplantation. In *Primer to the Immune Response (Second Edition)* (pp. 457–486). Academic Cell.

<https://doi.org/10.1016/B978-0-12-385245-8.00017-0>

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>

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>

SITNFlash. (2012, April 16). Trials and Tribulations of a Transplant. *Science in the News*.

<https://sitn.hms.harvard.edu/flash/2012/issue116/>

Siu, J. H. Y., Surendrakumar, V., Richards, J. A., & Pettigrew, G. J. (2018). T cell

Allorecognition Pathways in Solid Organ Transplantation. *Frontiers in Immunology*, *9*, 2548.

<https://doi.org/10.3389/fimmu.2018.02548>

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

Daily Entries:

Entry 1: MATLAB Training, 11/19/23, ~~SB~~

Process: Using the *MATLAB Essentials* course from edx, I learned basic data analysis in MATLAB by analyzing and creating figures for a given dataset. Additionally, I looked at the Medical Imaging toolbox resource from MATLAB to gain a better understanding of the different types of data analysis.

Resources Used:

- Edx course: https://learning.edx.org/course/course-v1:MathWorks+intro_matlab+2T2021
- MATLAB Medical Imaging Toolbox: <https://www.mathworks.com/products/medical>

Outcome: Begin to learn and understand the format for MATLAB. Found the specific package I need to create the image analysis algorithm.

Reflection: The next steps include downloading the software with the respective packages and getting images of kidney OCT scans from previous literature research or from Dr. Xihan.

Entry 2: MATLAB Training and Software Download, 11/23/23, ~~SB~~

Process: Continued MATLAB training through an informational session from MATLAB. Downloaded MATLAB onto my computer along with the AI (Deep Learning) and medical image analysis package.

Resources Used:

- Tutorial: <https://www.mathworks.com/videos/medical-image-processing-with-matlab>
- Get Started: <https://www.mathworks.com/help/medical-imaging/ug/get-started-with-med>
- MATLAB download: <https://matlab.mathworks.com/>

Outcome: The packages and software were successfully downloaded. I learned more about image analysis and how to measure and train the model for radiology scans.

Reflection More training is needed to becoming more familiar with the software. Kidney scans must be obtained for image analysis.

Entry 3: Beginning Model Development, 11/24/23, SB

Build: Found OCT kidney images from previous literature papers. Started measuring PCT structures in MATLAB session. Pictures were in jpg format, so I had to convert it into a dicom format for the MATLAB software to accept it.

Resouces Used:

- Kidney scans: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6485011/>
- Kidney scans: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5745648/>
- MicroDicom: <https://www.microdicom.com/>

Outcome: Measured and highlighted the PCT structures in the kidney photos:

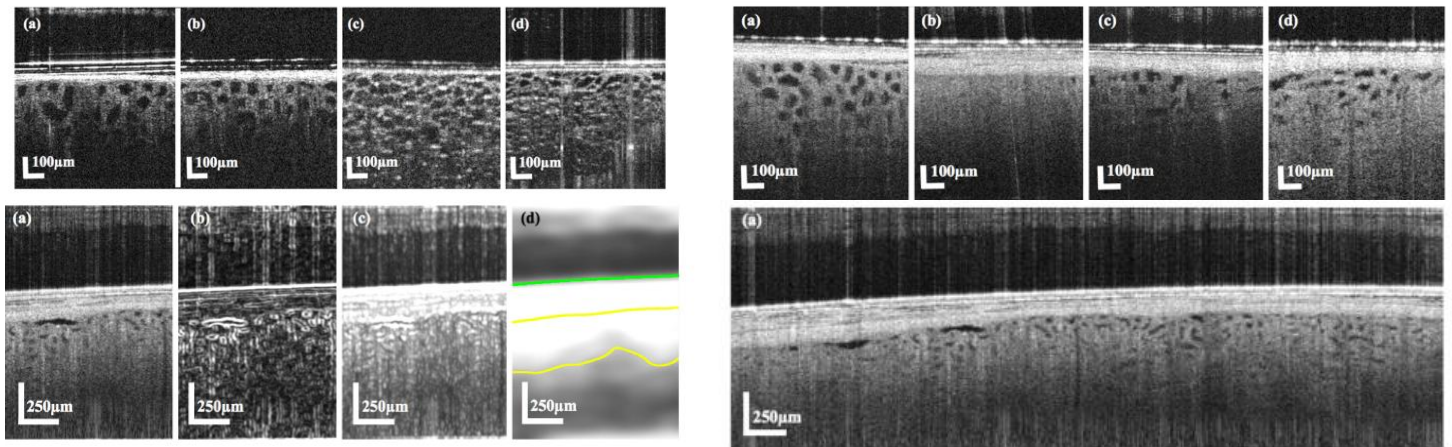


Figure 1: Images of kidneys under an optical coherence tomography (OCT) scan.

November 11, 2023

1:12 pm

Samhitha Bodangi

Reflection: After doing more research to get the OCT imaging scans, Kenkel et al., 2019 has already made a MATLAB model for kidney OCT scans. As there isn't much improvements I can make, deciding to change the project.

Entry 4: Meeting with Demetri Maxim, 11/28/2023, SB

Discussion: Met with Mr. Maxim from Nephrogen who went to ISEF 2015 for his project/patent related to biomarker identification of vascular endothelial growth factor-C (VEGF-C) an indicator of chronic transplant rejection.

Notes: Demetri Meeting

- ISEF Patent: <https://patents.google.com/patent/US10908156B2>

Outcome: Clarified some doubts about HLAs, biomarkers and how to use them to predict rejection. Helped me get some contacts to reach out to for data (Dr. Melissa Yeung, who is a nephrologist from Brigham and Women's Hospital myeung@bwh.harvard.edu).

Reflection: Next steps would be to reach out to Dr. Yeung for transplant data

Update: As of 1/17/2024, Dr. Yeung has not responded with the data. Will use publicly available datasets for model instead.

Entry 5: Machine Learning Research, 12/03/23, SB

Process: Researched different AI models to be constructed along with a list of criteria to evaluate the models after they are made.

Resources Used:

- AI in precision medicine:
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10233311/>
- Predicting drug dosage with AI: <https://arxiv.org/pdf/2308.11167.pdf>

Outcome: The models that will be made are Random Forest, Support Vector Machine, K-Nearest Neighbor, Format Concept Analysis, and Naive Bayes. The evaluation criteria included accuracy, precision, recall, ROC Curve and AUC, and external/cross validation.

Reflection: Next steps would include downloading Python and the necessary softwares.

Entry 6: Meeting with Dr. Keeler, 12/05/23, SB

Discussion: Met with Dr. Keeler from UMass who researches CAR T cell therapy. She works with pediatricians and currently does clinical trials for CAR T cell therapy.

Outcome: Learned about CAR T cells and how they can potentially be used to prevent organ rejection.

Reflection: Not my area of research, but something to look at if interested.

Entry 7: Meeting with Dr. Stern, 12/06/23, SB

Discussion: Met with Dr. Stern from UMass who researches MHC molecules and peptides. He works with T cells and potentially T regs.

Notes:

- [IEDB.org: Free epitope database and prediction resource](https://www.iedb.org/) (epitope presentation)
- Netmhcpan and netmhc2pan (for class I and II MHC) and mhc allusion

Outcome: Referred me to multiple databases that could help me make a model with peptide prediction. Explained the antigen presentation process.

Reflection: Narrowed research topic and will collect data soon.

Entry 8: Algorithm Research, 12/07/23, SB

Process: Researched different AI models to be constructed along with a list of criteria to evaluate the models after they are made.

Resources Used:

- AI in precision medicine:
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10233311/>
- Predicting drug dosage with AI: <https://arxiv.org/pdf/2308.11167.pdf>

Outcome: The models that will be made are Random Forest, Support Vector Machine, K-Nearest Neighbor, Format Concept Analysis, and Naive Bayes. The evaluation criteria included accuracy, precision, recall, ROC Curve and AUC, and external/cross validation.

Reflection: Next steps would include downloading Python and the necessary softwares.

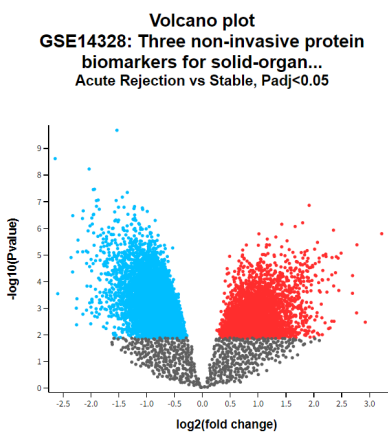
Entry 9: GEO Kidney Biopsy Gene Analysis, 12/11/23, ~~SB~~

Process: Searched up datasets in GEO related to gene expression in organ rejection. Used GEO2R software to find most significant genes that compared rejection biopsies vs. stable biopsies. The top upregulated and downregulated genes were searched for to analyze function and role in the immune response.

Resources Used:

- GEO Dataset: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14328>
- Excel for graphs and table
- Google for Protein biomarker lookup

Graphs and Data Analysis:



ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
210163_at	0.001179	1.59e-06	5.657466	5.106796	3.2180556	CXCL11	C-X-C motif chemokine ligand 11
224321_at	0.0218054	3.32e-03	3.129994	-1.91384	2.9231111	TMEFF2	transmembrane protein with EGF like and two follistatin like domains 2
211122_s_at	0.0015506	4.15e-06	5.35504	4.217932	2.7726667	CXCL11	C-X-C motif chemokine ligand 11
214768_x_at	0.0135148	1.49e-03	3.420387	-1.19027	2.7667222	IGKC	immunoglobulin kappa constant
206134_at	0.0033994	5.93e-05	4.506835	1.760568	2.6950556	ADAMDEC1	ADAM like dectsin 1
229543_at	0.0058754	2.73e-04	4.003683	0.354901	2.6926667	FAM26F	family with sequence similarity 26 member F
229391_s_at	0.0020105	8.54e-06	5.127202	3.551073	2.4895	FAM26F	family with sequence similarity 26 member F
217028_at	0.0052757	2.14e-04	4.085556	0.579226	2.4477222	CXCR4	C-X-C motif chemokine receptor 4
205114_s_at	0.0022204	1.32e-05	4.939931	3.151138	2.4358111	CCL3L3	C-C motif chemokine ligand 3
219620_x_at	0.0021291	1.16e-05	5.029875	3.267341	2.4037778	TOR4A	torsin family 4 member A
202988_s_at	0.0031102	4.62e-05	4.587363	1.9901	2.3888333	RGSI	regulator of G-protein signaling 1
231697_s_at	0.0207624	3.07e-03	3.158241	-1.84502	2.3622778	MIR21//NMVP1	microRNA 21//vacuole membrane protein 1
203915_at	0.0046283	1.45e-04	4.213909	0.935571	2.3560556	CXCL9	C-X-C motif chemokine ligand 9
M97935_s_at	0.0010321	1.17e-06	5.753168	5.388494	2.3518889	STAT1	signal transducer and activator of transcription 1
229625_at	0.002172	1.24e-05	5.00794	3.203511	2.3259444	GBPS	guanylate binding protein 5
229390_at	0.0032819	5.54e-05	4.528727	1.822863	2.3229444	FAM26F	family with sequence similarity 26 member F
205242_at	0.0106828	3.06e-03	3.160243	-1.84013	2.3217222	CXCL13	C-X-C motif chemokine ligand 13
217157_x_at	0.0293734	5.44e-03	2.94491	-2.35571	2.2912778	IGK//IGKC	immunoglobulin kappa locus//immunoglobulin kappa constant
214244_s_at	0.0306279	5.80e-03	2.920659	-2.41239	2.2651667	ATP9V0E1	ATPase H+ transporting V0 subunit e1
209795_at	0.0037699	8.29e-05	4.397754	1.451422	2.2573333	CD69	CD69 molecule
233589_x_at	0.0056391	2.47e-04	4.037396	0.447295	2.2352222	TOR4A	torsin family 4 member A
204286_s_at	0.0021186	1.14e-05	5.034659	3.281271	2.2187778	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1
204232_at	0.0020465	9.28e-06	5.100704	3.473747	2.2130556	FCER1G	Fc fragment of IgE receptor Ig
214000_s_at	0.0225341	3.51e-03	3.109578	-1.96335	2.2013889	RGSI10	regulator of G-protein signaling 10
206785_s_at	0.0020983	1.05e-05	5.061908	3.360637	2.1665556	KLRC2//KLRCL1	killer cell lectin like receptor C2//killer cell lectin like receptor C1
204533_at	0.0029139	3.81e-05	4.649761	2.168855	2.1607778	CXCL10	C-X-C motif chemokine ligand 10
213566_at	0.0038698	8.84e-05	4.377076	1.393065	2.1561667	RNASE5	ribonuclease A family member k5
208262_x_at	0.0418734	9.75e-03	2.718873	-2.87245	2.1516667	MEFV	Mediterranean fever
225655_at	0.0038056	8.46e-05	4.391443	1.433604	2.1495	UHRF1	ubiquitin like with PHD and ring finger domains 1
212592_at	0.0173312	2.27e-03	3.268818	-1.57229	2.1437222	JCHAIN	joining chain of multimeric IgA and IgM

Figure 2: Volcano plot of Acute rejection vs. Stable rejection biopsies gene expression. Red shows upregulated genes, blue shows downregulated genes, grey shows genes that were not significantly different

Figure 3: Sample table of gene expression between acute and stable rejection biopsies

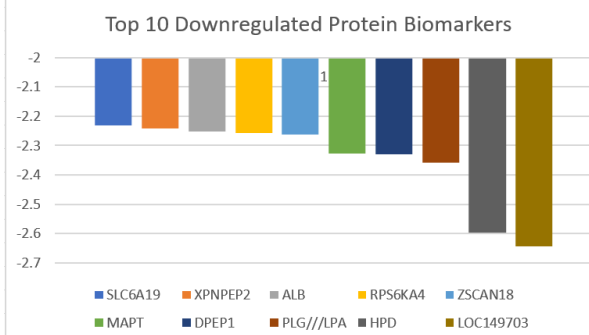
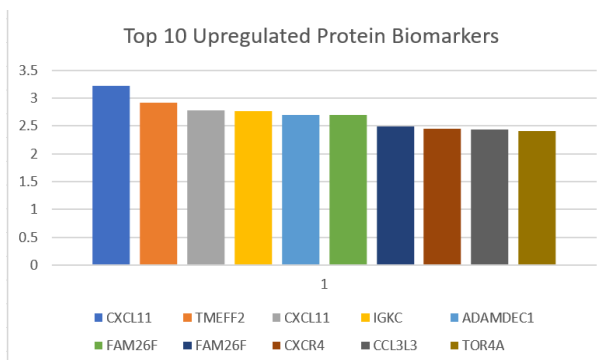


Figure 4: Top 10 upregulated genes in acute rejection
 December 11, 2023

Figure 5: Top 10 downregulated genes in acute rejection
 8:37pm
 Samhitha Bodangi

Results: Significant genes were looked up for function and role in the rejection response:

Top 10 Upregulated Protein Biomarkers

- *CXCL11* = gene that codes for cytokine that belongs to CXC chemokine family
- *TMEFF2* = gene that is involved in metabolism and endocrine function
- *IGKC* = encodes chains for antibodies. Serves as receptors that trigger clonal expansion and differentiation of B cells
- *ADAMDEC1* = expressed by macrophages and dendritic cells. Regulates immune system
- *FAM26F* = Creates synapses between immune cells and regulates the induction of inflammatory cytokines
- *CXCR4* = receptor proteins that span the outer membrane of cells. Expressed on the cell surface of most leukocytes
- *CCL3L3* = cytokine gene
- *TOR4A* = involved in the response to elevated platelet cytosolic Ca²⁺. Plays key role in initiating the innate immune system

Top 10 Downregulated Protein Biomarkers:

- *SLC6A19* = encodes a system transporter proteins that transport amino acids to epithelial cells in the proximal tubule of kidney
- *XPNPEP2* = involved in protein metabolism and targets bonds found in cytokines
- *RPS6KA4* = Involved in the IL-1 mediated signaling pathway
- *ZSCAN18* = expression negatively correlated with infiltration of B cells
- *MAPT* = makes tau proteins that stabilizes neuronal microtubules
- *DPEP1* = participates in leukotriene metabolism. Highly expressed in proximal tubular cells and peritubular capillaries of the kidney
- *PLG* = provides instructions for making the plasminogen protein, which plays a role in innate immunity. Produces cytokines and regulates macrophage phagocytosis
- *HPD* = provides instructions for making the enzyme 4-hydroxyphenylpyruvate dioxygenase. Important for tyrosine catabolism



Conclusion: Many of the up regulated genes were for chemokines, which are a specific class of cytokines that help guide T-cells to the organ site. This provides evidence for cytokines and chemokines being an important rejection biomarker. Additionally, many of the genes were kidney-specific. This shows that the methods for the project must be focused on one organ, as the genes/pathways for each organ response may be different. Many of the down-regulated genes target cytokines and regulate the immune system.

Reflection: Next steps would to follow a similar procedure for chronic rejection as that is the objective of this project.

Entry 10: Machine Learning Model Practice Training, 12/21/23, ~~SB~~

Process: Watched youtube video by *Bioinformatics Coach* and made a practice machine learning model that predicts cancer using Gene Expression Data (solely for practice)

Resources Used:

-  Machine Learning for Bioinformatics | Cancer Prediction using Gene Expr...
- Google Colab Link:  cancerPrediction.ipynb
 - Google Colab is a hosted Jupyter Notebook to write and execute Python code through the browser

Outcome: Learned about the basic code structure of machine learning model. While this model is for cancer prediction, a simple flowchart was developed to understand the machine learning model building process. This flowchart is in relation to the cancerPrediction Jupyter Notebook, but was made in the organ rejection setting:

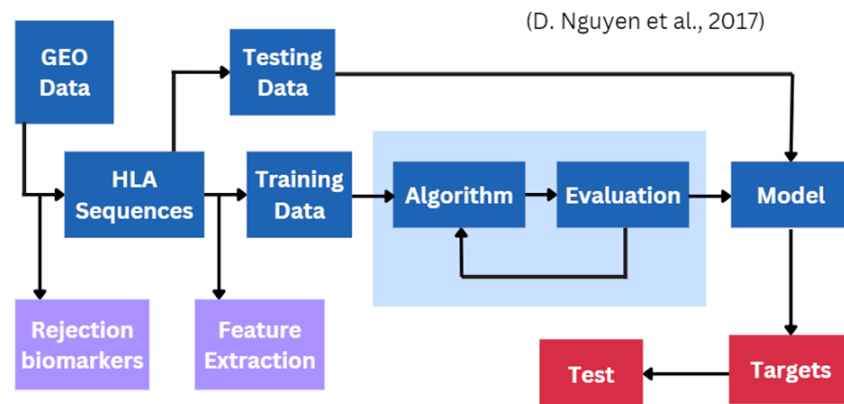


Figure 6: Flowchart of machine learning model that takes GEO data to predict rejection.

Flowchart is inspired by D. Nguyen et al., 2017

December 21, 2023

11:47pm

Samhitha Bodangi

Reflection: While this model is or cancer prediction, a similar model can be built using Gene Expression Omnibus data to predict organ rejection. Next steps wold include creating the simple model, making sure to keep the code similar to this flowchart.

Entry 11: Machine Learning GEO Practice Model for Biomarker Hunt, 12/22/23, ~~SB~~

Process: Used similar code framework as cancerPrediciton.ipynb and used Gene Expression Omnibus dataset to make basic organ rejection prediction model. Purpose is for practice and validation. Detailed procedure is below:

1. Download GEO dataset into Excel and transpose the data to fit the cancerPrediction notebook data format

- a. In the GEO dataset, the genes are along the rows while the samples are along the columns. Transposing the datafile made it so the samples were along the rows and the genes were along the columns
2. Copying the same code framework, the data was uploaded into a new Google Colab Notebook and the same steps were followed. The dataset was split into 80% training data and 20% testing data
3. The confusion matrix and ROC curve were graphed to analyze model accuracy

Resources Used:

- [Machine Learning for Bioinformatics | Cancer Prediction using Gene Expr...](#)
- Google Colab Framework Link: [cancerPrediction.ipynb](#)
- GEO Dataset(s): [GSE14328](#) → dataset was used during December fair
- Google Colab with GEO data: [gse14328Model.ipnb](#)

Outcome: The dataset is very small, with only 40 samples. This model was purely done for practice, as a much bigger dataset must be acquired for the final model.

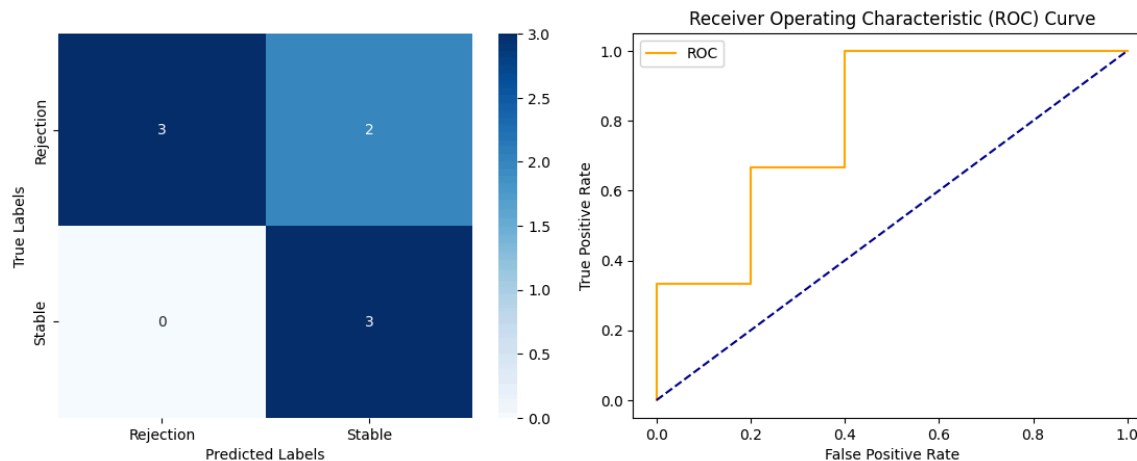


Figure 7: GSE14328 Random Forest Model Confusion Matrix

Figure 8: GSE14328 Random Forest Model Receiver Operating Characteristic (ROC) Curve
December 22, 2023 4:49pm Samhitha Bodangi

Conclusion: Accuracy of 75% and an AUC score of 0.8

Reflection: Using similar framework, a bigger dataset must be obtained to have better, more relevant results. Need to do more research to understand the math and specific statistical techniques used in machine learning to prepare for future questions.

Entry 12: UNOS Data Received, 12/23/23, SB

Process: Requested transplant data from United Network for Organ Sharing.

Resources Used:

- Data form: <https://optn.transplant.hrsa.gov/data/view-data-reports/request-data/data-request-instructions/>

- The STAR_SAS dataset was given as a folder in <https://app.box.com/>. However, I am not permitted to share the link with other people
- SAS software: <https://welcome.oda.sas.com/>

Outcome: The STAR files were download from U.N.O.S. and saved into a folder of my computer. The files are in SAS format, which requires a separate development software called SAS. An account was made and the data was loaded into a SAS project.

Reflection: Next steps would include analyzing the U.N.O.S. data and converting the SAS file into an excel file so it can be loaded onto a Google Colab notebook.

Entry 13: Random Forest Machine Learning Model GEO, 1/2/2023, ~~SB~~

Process: Similar process was used as in Entry 11 to make Random forest model for bigger dataset. This dataset had over 1300 sample, with AMR, TCMR, Mixed rejection, and No rejection. As this project will focus on TCMR, 532 no rejection biopsies and 437 TCMR samples were used. The data was transposed and a model was constructed.

Resources Used:

- Google Colab Framework Link: [cancerPrediction.ipynb](#)
- GEO Dataset(s): [GSE212160](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE212160)
- Google Colab link: [gse212160\(RF\).ipynb](#)

Outcome:

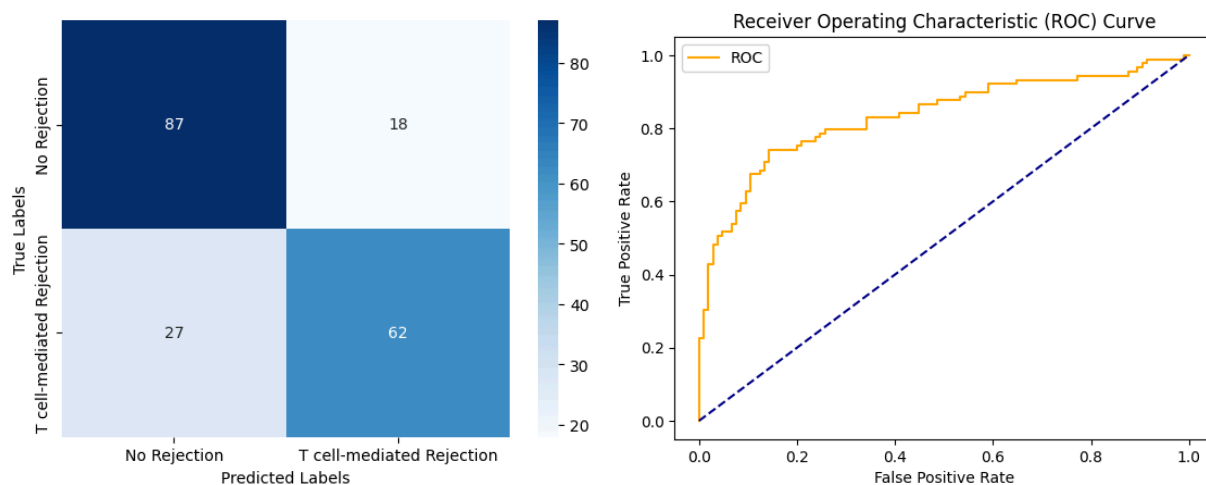


Figure 9: GSE212160 Random Forest Model Confusion Matrix

Figure 10: GSE212160 Random Forest Model Receiver Operating Characteristic (ROC) Curve

January 2, 2024

4:32pm

Samhitha Bodangi




Conclusion: Accuracy of 76.8% and an AUC score of 0.8334

Reflection: Use more datasets to increase accuracy of model. All datasets must be from the same platform (NanoString Human Organ Transplant Panel) or must be normalized to prevent any biases or inaccuracies from arising. Similar code can be used to develop another type of model (SVM, KNN, etc) to compare accuracies among models.

Entry 14: Support Vector Machine Learning Model GEO, 1/15/24, *SB*

Process: Using same dataset as Entry 13, a machine learning model was made with the Support Vector Machine (SVM) algorithm. There are many different algorithms that have different statistical techniques, and it is important to compare results among different models. Based on a tutorial, the code was modified for an SVM model, but maintained a similar coding structure and design.

Resources Used:

-  Hands on Session | Heart Attack Analysis Using AI | Python | Google Colab...
- Google Colab Framework Link:  cancerPrediction.ipynb
- GEO Dataset(s): [GSE212160](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE212160)
- Google Colab Link:  gse212160(SVM).ipynb

Outcome:

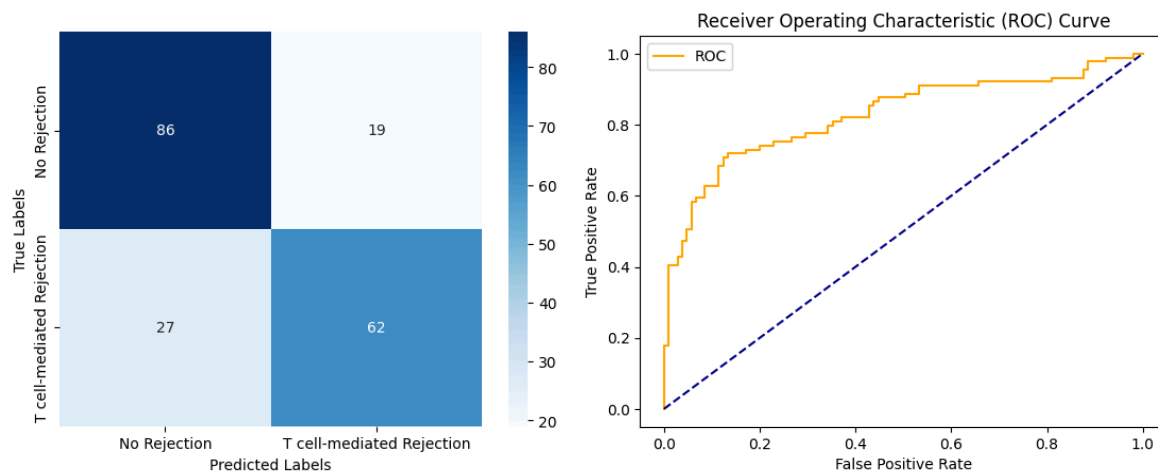


Figure 11: GSE212160 Support Vector Machine Model Confusion Matrix

Figure 12: GSE212160 Support Vector Machine Model Receiver Operating Characteristic (ROC) Curve

January 15, 2024

6:58pm

Samhitha Bodangi

Conclusion: Accuracy of 75.7% and an AUC score of 0.8256

Reflection: Accuracy can be increased by using other datasets based on the same gene expression platform as this GEO dataset. Compared to the RF model, this SVM model has a slightly lower accuracy. Additionally, it has a slightly lower AUC value, which correlates to the lower accuracy. A KNN model will be developed to compare accuracies as well.

Entry 15: STEM Update Meeting #6 Takeaways, 1/17/24, *SB*

- Experiment with different feature numbers
- Use downregulated features as -ve control
- Increase confidence of focusing on highest upregulated genes by having feature selection into specific ranges (high upreg, med upreg, low upreg, downreg, etc)

- More MHC-peptide research
 - what data will be inputted into model to predict this
 - Where the peptide comes from (antigen from HLA, or another surface antigen)
- Understand PIRCHE-II model methods and how it predicts MHC peptides
- Idea: Use specific differences in proteins and combine with mismatch data for more target information

Entry 15: K-Nearest Neighbor Machine Learning Model GEO, 1/17/24, SB

Process: Using same dataset as Entry 13 and 14, a machine learning model was made with the K-Nearest Neighbor (KNN) algorithm. Based on a tutorial, the code was modified for an KNN model, but maintained a similar coding structure and design.

Resources Used:

- [AI & Machine Learning Made Simple Coding 8: KNN Algorithm w Python...](#)
- Google Colab Framework Link: [cancerPrediction.ipynb](#)
- GEO Dataset(s): [GSE212160](#)
- Google Colab Link: [gse212160\(KNN\).ipynb](#)

Outcome:

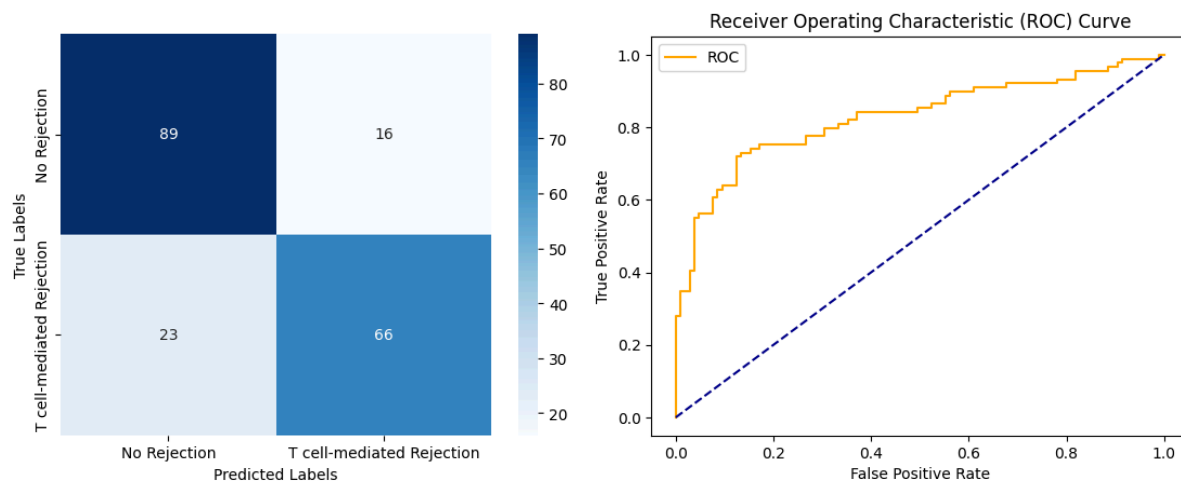


Figure 13: GSE212160 K-Nearest Neighbor Model Confusion Matrix

Figure 14: GSE212160 K-Nearest Neighbor Model Receiver Operating Characteristic (ROC) Curve
January 17, 2023 6:22pm Samhitha Bodangi

Conclusion: Accuracy of 79.5% and an AUC score of 0.8378

Reflection: Accuracy can be increased by using other datasets based on the same gene expression platform as this GEO dataset. Compared to the RF and SVM model, this KNN model has the highest accuracy. Additionally, it has a slightly higher AUC value, which correlates to the higher accuracy. A final decision matrix was made comparing these models:

Criteria	Random Forest (RF)	Support Vector Machine (SVM)	K-Nearest Neighbor (K-NN)
Accuracy	76.26%	75.78%	79.46%
ROC AUC Score	83.34%	82.56%	83.78%
Precision	76.86%	76.31%	79.93%
Recall	76.80%	76.29%	79.90%
F1 Score	76.66%	76.17%	79.81%

Figure 15: Decision Matrix of all three models (Random Forest, Support Vector Machine, and K-Nearest Neighbor) for the GSE212160 data set. The criteria equations are defined below:

January 17, 2023

7:46pm

Samhitha Bodangi

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{F1 Score} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$

Discussion: In all categories, the K-Nearest Neighbor had the highest scores and had the highest accuracy. Next steps would be to use the KNN model to identify significant genes to provide a rationale for this project to focus on the MHC-peptide complex for rejection targets.

Entry 16: Feature Experimentation with Negative Control, 1/23/24, ~~SB~~

Process: As the KNN model was chosen as the optimal model for the GSE21260 dataset, we attempted to increase the accuracy by changing the number of features the KNN model considered. Additionally, KNN models were constructed using downregulated gene features or a negative control.

Resources Used:

- Google Colab Link: [gse212160\(KNN\).ipynb](#)
- `n_features=500`
- `selected_scores_indices = np.argsort(MI)[:n_features]`

Outcome:

Number of Features	Accuracy with Upregulated Genes	Accuracy with Downregulated Genes
100	75.77%	76.80%
300 (original)	79.90%	68.04%
500	79.90%	78.35%



Conclusion: Because of the decreased accuracy with downregulated genes, it can be concluded that upregulated genes have a greater weight in determining the rejection sample.

Reflection: Provides evidence for upregulated genes being more significant features in predicting rejection. Next steps would be to investigate PECAM1 role in immune response.

Entry 17: MHC-Peptide Methodology Testing, 1/31/24, ~~SB~~

Process: The model methodology was tested using sample donor and recipient alleles in the B locus by manually using the servers and database to find amino acid mismatches and find peptides. The results were validated using the results from HLA-EMMA.

Resources Used:

- IPD/IMGT-HLA: <https://github.com/ANHIG/IMGTHLA>
- NetSurfP: <https://services.healthtech.dtu.dk/services/NetSurfP-3.0/>
- NetMHCIIpan: <https://services.healthtech.dtu.dk/services/NetMHCIIpan-4.1/>
- Google Sheets File:  HLA MM PreLim
- Google Colab file:  hlaMismatchPreLim.ipynb

Outcome:

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

Figure 16: General amino acid mismatches (yellow) and solvent-accessible mismatches (green) for sample donor allele B*07:02 and recipient alleles B*08:01 and B*40:02

January 28, 2023

10:34pm

Samhitha Bodangi

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

Figure 17: General amino acid mismatches (yellow) and solvent-accessible mismatches (green) for sample donor allele B*35:03 and recipient alleles B*08:01 and B*40:02

January 28, 2023

10:36pm

Samhitha Bodangi

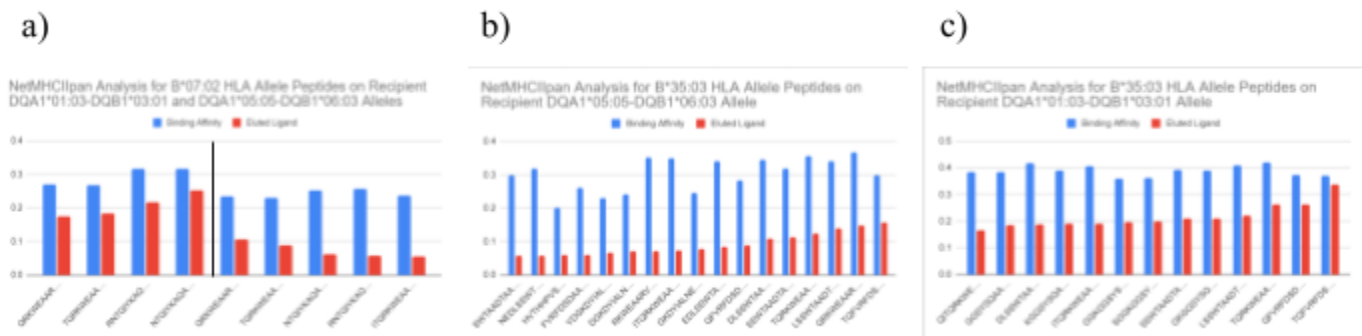


Figure 18: Graph a shows the binding affinity and eluted ligand scores from NetMHCIIpan for donor-derived peptides on donor allele B*07:02. Graph B and C shows the strongest peptides from donor allele B*35:03 to recipient alleles.

January 28, 2023

11:01 pm

Samhitha Bodangi


Conclusion: Amino acid mismatches and the solvent accessible mismatches were similar to the ones found in HLA-EMMA. Shows evidence for using NetSurfP as an accurate prediction server to find solvent-accessible regions in the HLA sequence. The low mismatches from both HLA-EMMA and the algorithm show evidence for this sample HLA-B locus alleles have a low immunogenicity, or a low risk for rejection. After using NetMHCIIpan, there was a low number of strongly binding peptides that contained solvent-accessible mismatches. The strongest ones are depicted in the graphs in Figure 18. Additionally, many of the peptides repeat, showing evidence for those peptides likely being the ones that would be recognized by T-cells, if rejection were to occur. As HLA-EMMA also predicted a low chance for rejection, these results provide a rationale for moving forward with the methodology.

Reflection: Next steps would include creating a model that would automatically perform the methodology, given a donor and recipient sequence.

Entry 18: Web Application Design in Figma and Visual Studio Code, 1/27/24, SB

Process: With the finished model, a web application was made to hold the model and handle a user-friendly interface (UI). HTML and CSS were used to handle the UI in Visual Studio Code

Resources Used:

- Figma file: <https://www.figma.com/file/mnWf3qlnamdCt7VmJzxCg/STEM-I-Web-App-Design?type=design&node-id=0%3A1&mode=design&t=2idHJWqAL9L0A3JR-1>
- PDF file:  STEM I Web App Design (2).pdf
- Visual Studio Code: <https://code.visualstudio.com/>
- Create React App Configuration: <https://github.com/facebook/create-react-app>

Outcome: HTML code for the user interface is created. However, the web application is not formatted to support smaller screen windows such as iPads or iPhones. If time permits, the

web application will be formatted, but may have limited function as the holding the machine learning model required a local database to hold the data and server downloads.

Conclusion: Web application has a friendly UI

Reflection: Next steps for the web application would be to create a local database in python using Django or similar program.

Entry 19: Accessing fasta Files and Identifying Amino Acid Mismatches, 2/5/24, ~~SP~~

Procedure: Using the IPD/IMGT-HLA database, an initial program was constructed to access the amino acid sequences depending on the given HLA alleles. The program was constructed for the HLA-A locus, as the sequence fasta files were separated based on HLA locus. The fasta file was converted into a pandas dataframe, and the sample donor and recipient sequences from HLA-EMMA were used to test the program. Mismatches were only included if the donor amino acid is not equal to both the recipient alleles (XOR). After finding and displaying the mismatches, the same alleles were put into HLA-EMMA to validate the mismatches.

Resources Used:

- Google Colab file: [🔗 A_locus_mm-02-05-24.ipynb](#)
- HLA-EMMA Template: [📄 HLA-EMMA-Template \(1\)](#)
- Reading Fasta files:
https://github.com/lanadominkovic/rosalind/blob/main/reading_sequence_files/fasta.ipynb
- Converting Fasta file to df:
<https://gist.github.com/fomightez/8cd6d9ba88f975b64e43eba562894dec>
- Converting 2-field resolution to 4-field resolution:
<https://thesequencingcenter.com/what-are-the-differences-between-2-field-and-4-field-reporting-for-hla-typing/>
- Fasta files: <https://github.com/ANHIG/IMGTHLA>

Outcome:

Details		Residue Properties													
Sequence Overview															
<input type="radio"/> Full <input checked="" type="radio"/> Compact															
Info	HLA Allele	62	66	74	76	77	95	97	105	107	114	116	127	142	145
Recipient	A*01:01	Q	N	D	A	N	I	I	P	G	R	D	N	I	R
Recipient	A*32:01	Q	N	D	E	S	I	M	P	G	Q	D	N	I	R
Donor	A*02:01	G	K	H	V	D	V	R	S	W	H	Y	K	T	H
Total AA Mismatches	14	G	K	H	V	D	V	R	S	W	H	Y	K	T	H
Solvent Accessible AA Mismatches	12	G	K	-	V	D	-	R	S	W	H	Y	K	T	H

Figure 19: Amino acid mismatches from HLA-EMMA for donor allele A*02:01 and recipient alleles A*01:01 and A*32:01. Yellow represents mismatches, while red represents solvent-accessible mismatches

February 6, 2024

8:05am

Samhitha Bodangi



Unique Amino Acid Mismatches:

Position 10: Donor 'V' vs. Recipient1 'L' and Recipient2 'L'
 Position 86: Donor 'G' vs. Recipient1 'Q' and Recipient2 'Q'
 Position 90: Donor 'K' vs. Recipient1 'N' and Recipient2 'N'
 Position 98: Donor 'H' vs. Recipient1 'D' and Recipient2 'D'
 Position 100: Donor 'V' vs. Recipient1 'A' and Recipient2 'E'
 Position 101: Donor 'D' vs. Recipient1 'N' and Recipient2 'S'
 Position 119: Donor 'V' vs. Recipient1 'I' and Recipient2 'I'
 Position 121: Donor 'R' vs. Recipient1 'I' and Recipient2 'M'
 Position 129: Donor 'S' vs. Recipient1 'P' and Recipient2 'P'
 Position 131: Donor 'W' vs. Recipient1 'G' and Recipient2 'G'
 Position 138: Donor 'H' vs. Recipient1 'R' and Recipient2 'Q'
 Position 140: Donor 'Y' vs. Recipient1 'D' and Recipient2 'D'
 Position 151: Donor 'K' vs. Recipient1 'N' and Recipient2 'N'
 Position 166: Donor 'T' vs. Recipient1 'I' and Recipient2 'I'
 Position 169: Donor 'H' vs. Recipient1 'R' and Recipient2 'R'

Figure 20: Amino acid mismatches from the constructed programm for donor allele A*02:01 and recipient alleles A*01:01 and A*32:01.

February 6, 2024

8:10am

Samhitha Bodangi

Conclusion: HLA-EMMA does not consider the N-terminal (amino acids 1-24) or the C-terminal (amino acids 300-365) as they are not generally used for HLA typing. However, as the constructed algorithm includes mismatches in the N- and C-terminal, there is an extra mismatch detected by the algorithm (position 10). However, the other mismatches are identical to the mismatches detected by HLA-EMMA.

Reflection: A similar program can be constructed for the other HLA locus alleles by uploading the respective fasta file from the IPD/IMGT-HLA database. Next steps including connecting the mismatches to NetSurfP to find solvent accessible mismatches.

Entry 20: IPD/IMGT-HLA Sequence fasta File Processing, 2/6/2024, *SB*

Process: The IPD/IMGT-HLA database can be accessed through the ftp server to make the model more efficient as the large file does not have to be downloaded into the Google Colab notebook. After converting the file into pandas data frame, the data was processed by deleting all HLA alleles that were not the 11 commonly typed alleles. Then, duplicate sequences were deleted and alleles in field 3 or 4 were converted to field 2. After conversion, the duplicates were removed and the final data frame was uploaded as a csv file.

Resources Used:

- Google Colab file:  fastaProcessing.ipynb
- Processed csv file:  processed_sequences
- Opening ftp server directory:
https://colab.research.google.com/github/astg606/py_materials/blob/master/data_retrieval/introduction_data_retrieval.ipynb
- Converting pandas dataframe into csv:
<https://saturncloud.io/blog/exporting-dataframe-as-csv-file-from-google-colab-to-google-drive/>
- Processing guidelines:
 - <https://www.sciencedirect.com/science/article/pii/S0198885921001154>
 - <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7317360/>
 - <https://hla.alleles.org/nomenclature/naming.html>


Outcome: The file is processed and contains only the HLA allele sequences relevant to organ transplant tissue typing. All of the alleles are in the same field-type format, allowing for more standardization during the testing of the model.

Reflection: Next steps would include using the processed data file to obtain amino acid sequences from donor and recipient HLA alleles and start building the model.

Entry 21: Aligning HLA Protein Sequences, 2/7/24, *SB*

Process: The sequences in the IPD/IMGT-HLA database have different sequence lengths. In order to find the mismatches, they must be properly aligned. Using the Needleman-Wunsch algorithm, the sequences will be aligned to the 01:01 allele of that same locus.

Resources Used:

-  organRejectionv2.ipynb
- Algorithm:
[https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_\(Kellis_et_al.\)/02%3A_Sequence_Alignment_and_Dynamic_Programming/2.05%3A_The_Needleman-Wunsch_Algorithm](https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.)/02%3A_Sequence_Alignment_and_Dynamic_Programming/2.05%3A_The_Needleman-Wunsch_Algorithm)
- Code:
<https://medium.com/@nandiniumbarkar/needleman-wunsch-algorithm-7bba68b510db>
- <https://gist.github.com/slowkow/06c6dba9180d013dfd82bec217d22eb5>
- <https://pypi.org/project/minineedle/>
- Troubleshoot:
<https://stackoverflow.com/questions/63120727/needleman-wunsch-algorithm-for-two-sequences-of-different-length>
- IPD/IMGT-HLA Sequence Alignment Tool:
<https://www.ebi.ac.uk/ipd/imgt/hla/alignment/>



Outcome: The code works by finding the highest degree of similarity between the two sequences. Then, it uses gap characters from the longer sequence to substitute the gaps in the shorter sequence after obtaining a high similarity. Then, the aligned sequences are stored and can be used to find mismatches.

Reflection: Initially difficult to create the code inserting gap characters within the sequence. However, now, the alignment sequences can be used to find accurate mismatches. Next steps include using NetSurfP to find solvent accessible mismatches.

Entry 22: Filtering Solvent-Accessible Mismatches with NetsurfP, 2/08/24, *SB*

Process: All amino acid mismatches were filtered to include only the solvent-accessible mismatches. This way, the number of features will decrease to the ones that have the highest chance of immunogenicity. Using the biolib package, netsurfp was installed locally and was called to find solvent-accessible amino acids in a given donor sequence. The result was visualized and stored as a pandas data frame.

Resources Used:

-  organRejectionv3.ipynb
- Server: <https://services.healthtech.dtu.dk/services/NetSurfP-3.0/>
- Server Methods: <https://academic.oup.com/nar/article/50/W1/W510/6596854>
- Running Server: <https://dtu.biolib.com/NetSurfP-3/>
- Demo:  NSP3_demo.ipynb

Outcome:

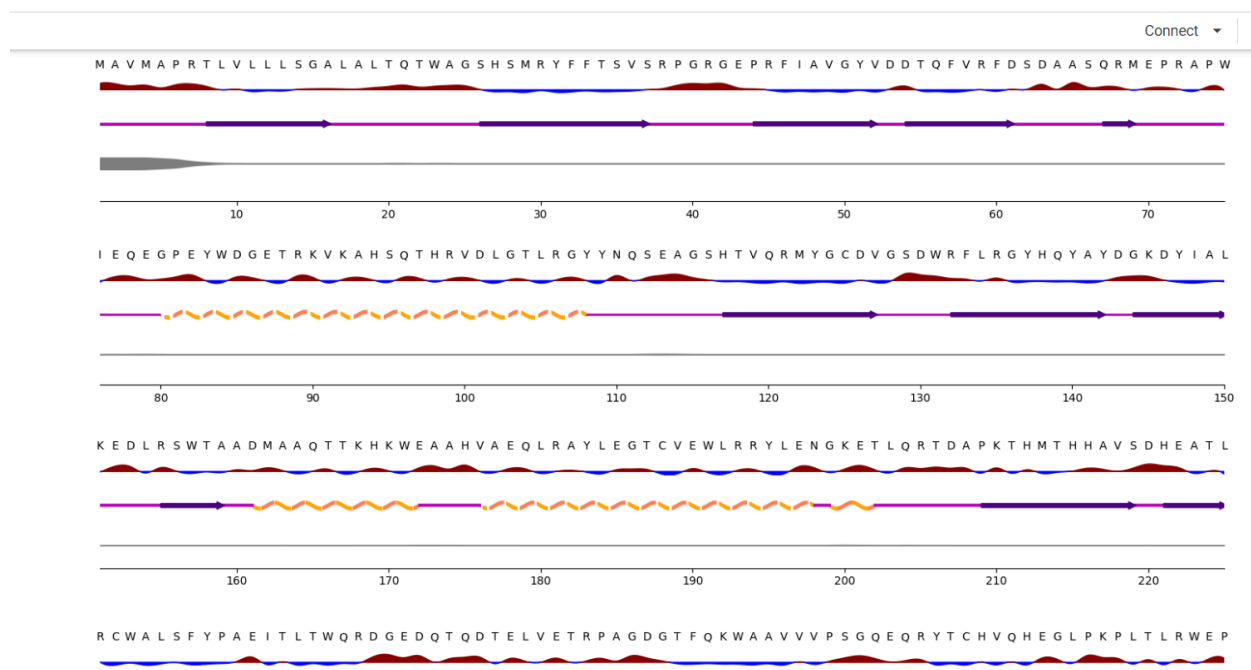


Figure 21: Example Visualization output for solvent-accessible amino acid from the constructed program for donor allele A*02:01



Conclusion: Results were stored as a dataframe and the solvent-accessible sequences were used to filter all of the initial mismatches.

Reflection: If necessary, attempt to download NetSurfP server locally onto Google Colab to make queries faster. Next steps include finding all possible donor-derived peptides that contain the solvent-accessible mismatches and predicting the binding affinity.

Entry 23: Immune Epitope Database for Peptide Prediction, 2/09/24, ~~SB~~

Process: Using iedb python package, donor-derived peptides were generated and the score was calculated. NetMHCIIpan can be accessed through IEDB by specifying the method of peptide prediction. Donor sequences and recipient alleles were input to predict and filter peptides. Additionally, the tools api was tested to see which package would be most efficient.

Resources Used:

-  organRejectionv4.ipynb
-  organRejectionv5.ipynb
- <https://github.com/mattfemia/iedb-python/tree/master>
- <http://tools.iedb.org/main/tools-api/>

Outcome: Peptides were generated and eluted ligand score was calculated. The threshold for strong binders was <1, which is the commonly used Frank threshold by NetMHCIIpan.

Strong Bindings:

	allele	seq_num	start	end	length	core_peptide	peptide	score	rank
0	HLA-DRB1*11:01	1	30	44	15	FTSVSRPGR	RYFFTSVSRPGRGEP	0.8764	0.56
1	HLA-DRB1*11:01	1	29	43	15	FTSVSRPGR	MRYYFTSVSRPGRGE	0.8707	0.58

Figure 22: Example Output of Strong peptides using the IEDB database with NetMHCIIpan. Sequence is donor sequence A*01:01, and binding predictions were for recipient allele DRB1*11:01

February 12, 2024

11:23 am

Samhitha Bodangi

Conclusion: The IEDB tools API is more efficient in generating all possible donor-derived peptides that can strongly bind to recipient class II molecules.

Reflection: Next steps include completing the entire model for all HLA alleles

Entry 25: Complete Model Building, 2/09/24, ~~SB~~

Process: Now that the methodology and packages are downloaded, the complete model can be built. The packages were tested with one HLA alleles, but can be replicated for all alleles.

Resources Used:

-  organRejectionv6.ipynb

Outcome: All HLA alleles are in the model, and the peptides can be predicted.

Reflection: Next steps would include testing the model.

Entry 26: Linear Regression Model with HLA-Epi Compatibility Scores, 3/2/24, *SB*

Process: Using the compatibility scores from HLA-Epi, regression models can be made to see how well my model's scores and peptide target numbers compare to the compatibility scores.

Resources Used:

- Google Colab File: [🔗 hlaEpiLinearRegression.ipynb](#)
- Data: [📄 hla-epi-EUROPEAN_distri_output](#)
- Video Tutorial: [🔗 practice.ipynb](#)
 - [📺 Machine Learning in Python: Building a Linear Regression Model](#)
- https://github.com/dataprofessor/code/blob/master/python/linear_regression.ipynb
- <https://gitlab.univ-nantes.fr/crtiteam5/easy-hla/-/tree/main>
- Color change: https://www.practicalpythonfordatascience.com/ap_seaborn_palette
- <https://stackoverflow.com/questions/17197492/is-there-a-library-function-for-root-mean-square-error-rmse-in-python>

Outcome:

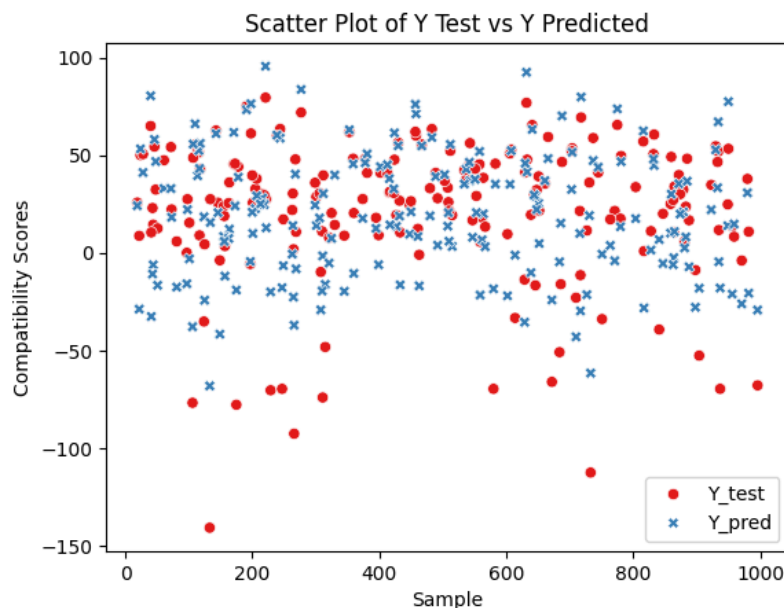


Figure 23: Scatterplot of Linear Regression Model With HLA-Epi Dataset

February 12, 2024

11:23 am

Samhitha Bodangi








Conclusion: The regression model has a coefficient of determination (R^2) value of 0.624

Reflection: The scores from my model do not predict the outlier compatibility scores in the HLA-Epi dataset. This could potentially be because some HLA loci have a greater impact on the outcome of rejection than others. Adding weights to number of peptides in each loci could potentially help make the model more accurate. Next steps would include creating more regression models of different algorithms and compare the accuracies between the models.

Entry 27: Remaining Regression Model with HLA-Epi Scores, 3/3/24,

Process: Using the compatibility scores from HLA-Epi, a random forest regressor was made to compare with the linear regression model.

Resources Used

- Random Forest Regression:
 -  hlaEpiRFRegression.ipynb
 -  Random Forest Regression | Python
 - <https://levelup.gitconnected.com/random-forest-regression-209c0f354c84>
 - <https://www.geeksforgeeks.org/random-forest-hyperparameter-tuning-in-python/>
- Lasso Regression Model:
 -  hlaEpiLassoRegression.ipynb
 -  Lasso Regression | Machine Learning | Python
 - <https://alfurka.github.io/2018-11-18-grid-search/>
- Polynomial Regression Model:
 -  hlaEpiPolynomialRegression.ipynb
 -  Polynomial Regression in Python - sklearn
 - <https://stackoverflow.com/questions/47414819/gridsearchcv-for-polynomial-regression>
- Ridge Regression Model:
 -  hlaEpiRidgeRegression.ipynb
 -  Ridge Regression | Machine Learning | Python
 - <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingRegressor.html>
 - <https://alfurka.github.io/2018-11-18-grid-search/>

Outcome:



Figure 24: Scatterplot of Random Forest Regression Model with HLA-Epi Dataset

March 3, 2024 2:49pm Samhitha Bodangi

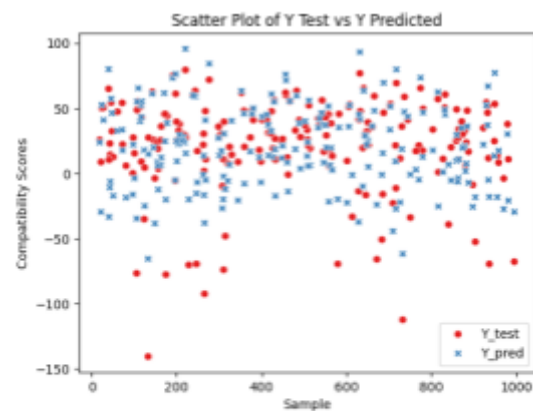


Figure 25: Scatterplot of Lasso Regression Model with HLA-Epi Dataset

March 3, 2024 2:53pm Samhitha Bodangi

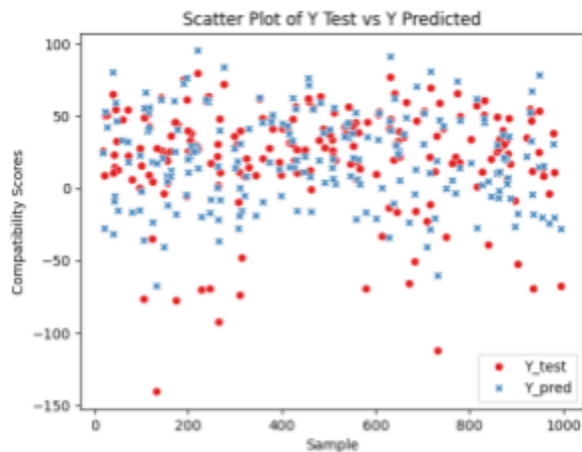


Figure 26: Scatterplot for Polynomial Regression Model with HLA-Epi Dataset

March 3, 2024 2:58pm Samhitha Bodangi

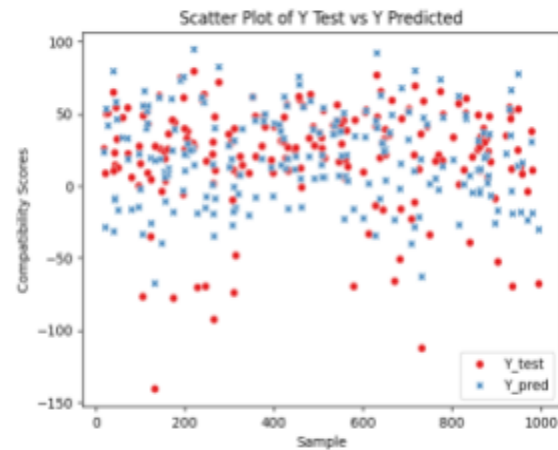


Figure 27: Scatterplot for Ridge Regression Model with HLA-Epi Dataset

March 3, 2024 3:04pm Samhitha Bodangi

Conclusion: The Random forest model has an R^2 value of 0.590. The Lasso model has an R^2 value of 0.623. The polynomial regressor has an R^2 value of 0.625. The degree is 1. The ridge regressor has an R^2 value of 0.626.

Reflection: Similar to the linear regression model. These models do not predict the outliers as well. Again, more research may need to be done to see which HLA loci have the greatest contribution to rejection.

For the polynomial regressor, even though having a degree of 1 makes this model similar to a linear regression model, it still has a slightly higher accuracy than the linear model. This could potentially be because of the scaling that occurred to the values. However, similar to the linear regression model. The model does not predict the outliers as well.



The random forest model had the lowest performance compared to the other models. This may be due the correlation between the different alleles. In the future, more tests and feature selection models can be run to see the inter-connected pattern of different alleles.

The ridge regression model performed the best out of all the other models. However, the accuracies are very close to each other. However, more research may need to be done to see which HLA loci have the greatest contribution to rejection. The next steps would include creating a final decision matrix comparing the models.

Entry 28: Rejection vs. No Rejection Scores, 3/7/24,

Process: Using a sample dataset from UNOS, the scores for rejection and not rejection were found and compared. Rejection was classified as rejection episodes that occurred at least one year after the kidney transplant.

Resources Used:

- Data processing:  unos.ipynb
- Score Comparison:  SampleScores.ipynb
- The STAR_SAS dataset was given as a folder in <https://app.box.com/>. However, I am not permitted to share the link with other people
- SAS software: <https://welcome.oda.sas.com/>
- SAS workbook:
https://odamid-usw2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT-04%253A00&ticket=ST-77426-x7ov4GQYCrYV4pMgoFoy-cas

Outcome:

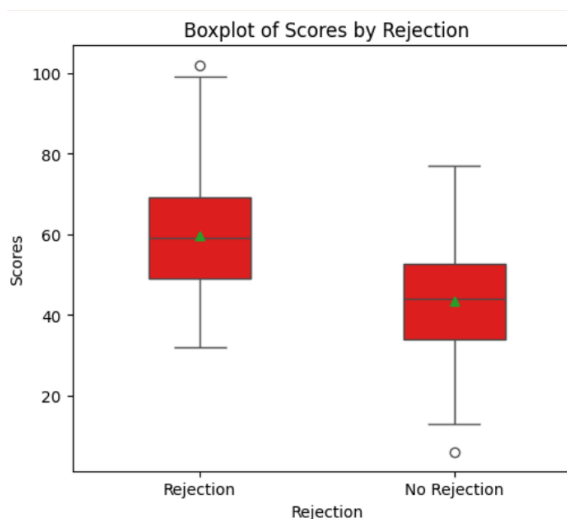


Figure 28: Box and Whisker Plots for the scores in the rejection and no-rejection groups. The Green symbol represents the average score for each group.

March 7, 2024

3:23 pm

Samhitha Bodangi


Conclusion: The means are statistically significant, with a t-statistic of 6.269 and a p-value of 1.646e-9.

Reflection: There is a clear correlation between peptide targets and the rejection outcome. A greater number of peptide targets with mismatches have a higher chance of causing rejection. However, there is still overlap between the ranges of scores, which is likely because of the fact that rejection is inevitable. Therefore, rather than classifying the groups in categories, it may be more beneficial to use compatibility scores instead. Next steps include conducting further statistical tests and improving the regression models.

Entry 30: Feature Selection and Weightages for Ridge Regression Model, 3/14/24,

Process: In order to improve the performance of the regression models, a random forest feature selection algorithm was used to identify important features. After the feature weights were found, they were implemented into the regression model by multiplying the coefficients with the weightages and testing the model to see if the performance improved.

Resources Used:

- Feature selection:
 - <https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/>
 - <https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/>
- Random forest feature selection:
 - <https://www.yourdatateacher.com/2021/10/11/feature-selection-with-random-forest/>
 - <https://www.kaggle.com/code/prashant111/random-forest-classifier-feature-importance>
 - <https://medium.com/@prasannarghattikar/using-random-forest-for-feature-importance-118462c40189>
-  PCAhlaEpiRidgeRegression.ipynb

Outcome:

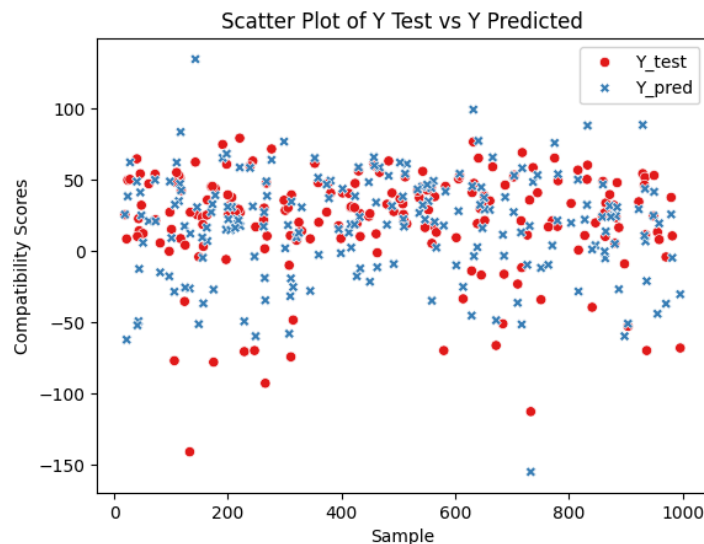


Figure 29: Box and Whisker Plots for the scores in the rejection and no-rejection groups. The Green symbol represents the average score for each group.

March 7, 2024

3:23 pm

Samhitha Bodangi

Conclusion: The new regression model performed better, with an R^2 value of 0.723. The feature weights for each HLA allele from the feature selection:

Epitopic mmA, Importance: 0.45528069104129437
 Epitopic mmB, Importance: 0.1724906487503538
 Epitopic mmC, Importance: 0.10886305038088126
 Epitopic mmDRB1, Importance: 0.1765105683784967
 Epitopic mmDQB1, Importance: 0.08685504144897378

Reflection: The model's performance increased from an R^2 value of 0.626 to 0.723, accounting for a 15.5% increase in performance. This increase shows that the feature importances that were found were accurate. From the random forest feature selection, it can be seen that mismatches in HLA A, DRB1 and B had the most contribution to the compatibility score. As these alleles have the most importance, clinicians should focus on matching donor and recipients with the highest similarity in those alleles to have a better match.

Entry 31: Web Application Building with Python and React, 4/1/24, *SB*

Process: Using the Visual Studio Code IDE, a web application was made with the React and Vite bundler. The developed Figma plan was used to design and build the web application using React and Javascript. Example inputs were made for easy display.

Resources:

- Installing pip: [How to install Python 3.9.2 and PIP on Windows 10](#)
- Table: [Table in React Js || Create Table from Array of Objects in React Js](#)
- Drop down: [React Navbar Dropdown Menu Responsive | How to create React Navbar ...](#)
- Installing python: <https://www.python.org/downloads/>
- Github Repo: <https://github.com/samhithabodangi/PIPSA-Model>

Outcome:

The application interface is titled "PIPSA Predicting Indirect Peptides with Solvent Accessibility". The main heading is "Predict The Risk of Organ Rejection in Kidney Transplant Recipients". Below this, it asks the user to "Please Enter The Donor and Recipient HLA Alleles in a 2-Field Format".

The input form is divided into two columns: "Recipient" and "Donor". Each column has a grid of input fields for HLA alleles (A, B, C) and their subtypes (DRB1, DRB3, DPA1, DQB1, DQA1, DPB1). Buttons for "Predict", "Example", and "Clear" are at the bottom.

The output screen displays the results:

- PIPSA Score:** 100
- Outcome:** Low Rejection Risk
- Significant Targets:** A bar chart showing the importance of various HLA alleles, with DRB1 and DRB3 being the most significant.
- Significant Targets Table:**

Info	HLA Allele	39	40	42	62
Recipient	DRB1*11:01	Y	S	S	N
Recipient	DRB1*13:01	Y	S	S	N
Recipient	DRB3*02:01	L	L	S	N
Recipient	DRB3*03:26	L	L	S	N
Donor	DRB1*04:01	Q	V	H	H
All AA Mismatches	4	Q	V	H	H
Solvent-Accessible AA Mismatches	4	Q	V	H	H

A note at the bottom states: "No Amino Acid Mismatches for Donor Alleles DRB1*11:04, DRB3*02:01, or DRB3*03:26".

Figure 32: Box and Whisker Plots for the scores in the rejection and no-rejection groups. The Green symbol represents the average score for each group.

March 7, 2024







3:23 pm

Samhitha Bodangi







Conclusion The web application currently works for the example, which was taken from HLA-EMMA. It is able to display the mismatches and the score for the specific donor and recipient HLA alleles.

Reflection: In the future, the python model will be embedded into the web application. In order to do so, the NetMHCIIpan and NetsurfP must be downloaded as packages and embedded into the web application to make it run smoothly.





STEM Hours Time-Log:

Date	Hours	What was done	Signature
August 20	1 hour	Reading and taking notes on general articles 2 and 3	
August 22	1 hour	Adding Summer reading articles	
August 25	2 hours	Brainstorming	
August 28	2 hours	Abstracts and presentations	
September 1	1 hour	Brainstorming	
September 3	3 hours	Reading and taking notes on general articles 6, 7, and 8	





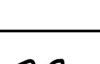

----- 10 hours for the weeks August 20 - September 3 -----

Date	Hours	What was done	Signature
September 8	1 hour	Brainstorming and making elevator pitch at Bourndale	
September 13	1 hour	Wrote and emailed Dr. Lanese	
September 14	1 hour	Started a second email to Dr. Gaudette and made a long summary on the Chili article	
September 16	1 hour	Found more articles and put links under my project notes to read in the future	
September 17	2 hours	Reading and taking notes on article 9 and Journal article 10	
September 18	4 hours	Preparing for STEM meeting Read and started to take notes on articles 11-13	







----- 10 hours for the weeks September 3 - September 18 -----

Date	Hours	What was done	Signature
September 20	3 hours	Read articles 16-18	
September 23	2 hours	Read 3 Wikipedia articles	
October 1	4 hours	Emailed 6 Professors, Read patents and articles	
October 2	8 hours	Prepared for STEM meeting	








----- 17 hours for the weeks September 18 - October 2 -----

Date	Hours	What was done	
October 7	4 hours	Read articles	
October 8	4 hours	Read articles	
October 12	1 hour	Watched videos related to T-cell activation	
October 13	3 hours	Added brainstorming pictures to Project LogBook. Read more articles.	
October 14	4 hours	Read more articles and emailed 2 more professors at UMass	
October 15	6 hours	Read articles and prepared for STEM meeting	

----- 22 hours for the weeks October 2 - October 16 -----

Date	Hours	What was done	
October 18	2 hours	Read articles and emailed a community lab	
October 20	2 hours	Emailed professors	
October 21	1 hour	Read articles	
October 25	2 hours	Read articles	
October 26	2 hours	Emailed professors	
October 31	1 hour	Met with professor and emailed more professors	

----- 10 hours for the weeks October 16 - October 31 -----

Date	Hours	What was done	
November 3	3 hours	Worked on MSEF proposal Intro - posted in this document	
November 4	5 hours	Emailed professors and made Mindmap for Grant Proposal	
November 5	4 hours	Worked on Grant Proposal Section 1 and emailed professors.	
November 6	3 hours	Read articles and met with Charles River Laboratories	
November 7	3 hours	Emailed professors and worked on Grant proposal Checkpoint 1	
November 12	4 hours	Watched video for backup, started MATLAB course for backup, emailed more professors	
November 13	5 hours	Met with Dr.Harlan, emailed more professors, and prepped for STEM meeting 4	

----- 27 hours for the weeks November 1 - November 13 -----

Date	Hours	What was done	
November 19	5 hours	Emails, MATLAB training, and created a systems map	SB
November 23	3 hours	MATLAB training, downloaded MATLAB package and microDicom	SB
November 24	3 hours	Started PCT structure model in MATLAB with kidney images from previous literature papers	SB
November 25	3 hours	Read more articles and patents	SB
November 26	4 hours	Updated logbook, worked on grant proposal, and read articles	SB
November 27	2 hours	Worked on Grant Proposal	SB

----- 20 hours for the weeks November 13 - November 27 -----

Date	Hours	What was done	
December 3	4 hours	Looked at GEO and EBI for datasets to use in Preliminary data	SB
December 5	1 hour	Met with Dr.Keeler and discussed CAR T ceel therapy	SB
December 6	2 hours	Met with Dr.Stern and received databases to find peptide bindings	SB
December 7	4 hours	Worked on and completed December Fair Poster	SB
December 10	3 hours	Prepared speech for December Fair	SB
December 11	5 hours	Prepared for December Fair	SB











----- 19 hours for the weeks November 28 - December 12 -----

Date	Hours	What was done	
December 15	2 hours	Preliminary python and R studio training and research	SB
December 18	1 hour	Looked over judges comments and assessed what to fix	SB
December 20	3 hours	Updated Project Logbook and worked on Grant Proposal	SB
December 21	2 hours	Preliminary Machine Learning research and training and model building	SB
December 22	1 hour	Used small GEO dataset to create simple RF model	SB
December 23	1 hour	Entry 12 (U.N.O.S. data and SAS software)	SB




----- 10 hours for the weeks December 12 - December 26 -----





Date	Hours	What was done	
December 28	1 hour	Looked through GEO for bigger TCMR datasets	SB
January 2	2 hours	Entry 13 (Created Random forest model)	SB
January 4	4 hours	Researched current models related to MHC-peptide complex	SB
January 7	3 hours	Downloaded and learned to make graphs from GEO data in R studio	SB
January 8	3 hours	Continued to experiment with R studio and pheatmaps	SB
January 9	2 hours	Worked on Thesis introduction	SB

----- 15 hours for the weeks December 26 - January 9 -----








Date	Hours	What was done	
January 12	2 hours	Worked on thesis introduction and abstract based on feedback	
January 14	1 hour	Created a research project outline systems diagram	
January 15	3 hours	Updated logbook, worked on grant proposal, and made SVM model	
January 16	6 hours	Worked on Grant proposal, prepared for STEM meeting 6	
January 17	4 hours	Updated logbook, created a KNN model and a decision matrix	
January 18	2 hours	Research on MHC peptide, worked on grant proposal, STEM website	
January 19	5 hours	Worked on Methodology and Research outline	
January 21	3 hours	Researched MHC peptide, worked on thesis Intro	
January 23	6 hours	Prepared for STEM meeting 7, finished thesis methodology draft	
January 24	2 hours	Entry 16, creating Figma design for web application	

----- 34 hours for the weeks January 10 - January 24 -----

Date	Hours	What was done	
January 27	2 hours	Web application home page was coded in Visual studio code	
January 29	2 hours	Simple model building, coding research and data download	
January 31	4 hours	Simple model building, Finished JSHS thesis and submitted	

February 3	2 hours	Coded STEM I computer science webpage with updated info	
February 5	1 hour	Model coding, started with A-locus using HLA EMMA sample	
February 6	3 hours	Processed IPD/IMGT-HLA sequence files and downloaded as csv	
February 7	4 hours	Updated poster, logbook, and NetSurfP python research	

----- 18 hours for the weeks January 24 - February 7 -----

Date	Hours	What was done	
February 8	3 hours	Create infographics for STEM website	
February 9	2 hours	Updated thesis discussion and conclusion section	
February 10	3 hours	Editing and submitting four STEM documents into google drive	
February 11	3 hours	Continued with model building and data collection	
February 12	6 hours	Finished model building and data collection	
February 13	6 hours	Finished coding and deploying web application	
February 14	6 hours	Practiced for Feb fair, and prepared poster and supporting materials	

----- 29 hours for the weeks February 8 - February 14 -----