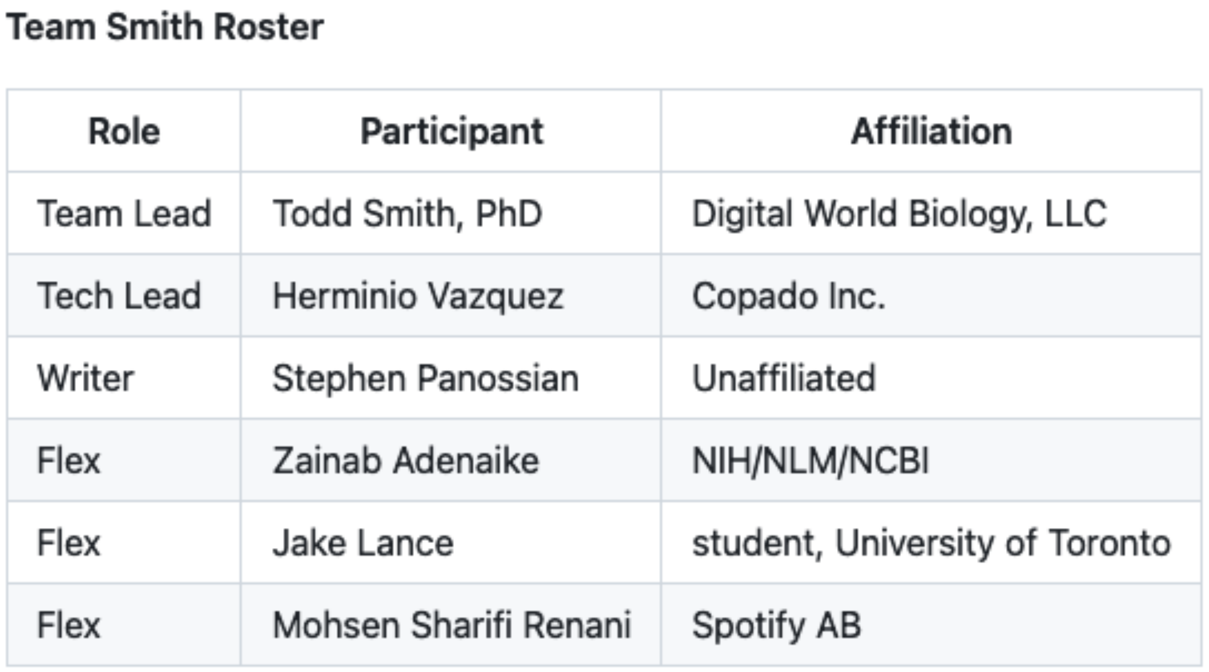
A gentle introduction to ML via antibody-engineering

NCBI: Building Transparent ML/AI Solutions to Advance Biological Research Virtual Codeathon Feb. 26 – Mar 1, 2024

Final Presentation

Digital World Biology, LLC

Team

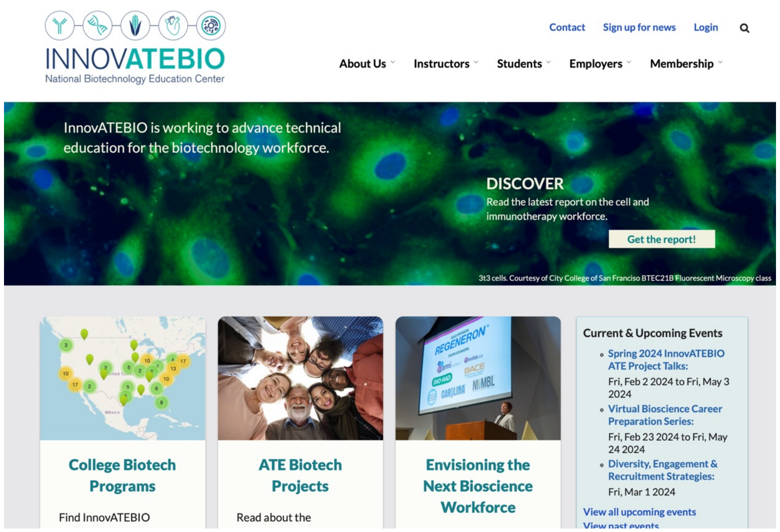


Team Smith Roster

Role	Participant	Affiliation
Team Lead	Todd Smith, PhD	Digital World Biology, LLC
Tech Lead	Herminio Vazquez	Copado Inc.
Writer	Stephen Panossian	Unaffiliated
Flex	Zainab Adenaike	NIH/NLM/NCBI
Flex	Jake Lance	student, University of Toronto
Flex	Mohsen Sharifi Renani	Spotify AB

Background

Project based on Digital World Biology’s work in community college biotech workforce education



InnovATEBIO is working to advance technical education for the biotechnology workforce.

DISCOVER

Read the latest report on the cell and immunotherapy workforce.

3t3 cells. Courtesy of City College of San Franciso BTEC21B Fluorescent Microscopy class

College Biotech Programs

Find InnovATEBIO

ATE Biotech Projects

Read about the

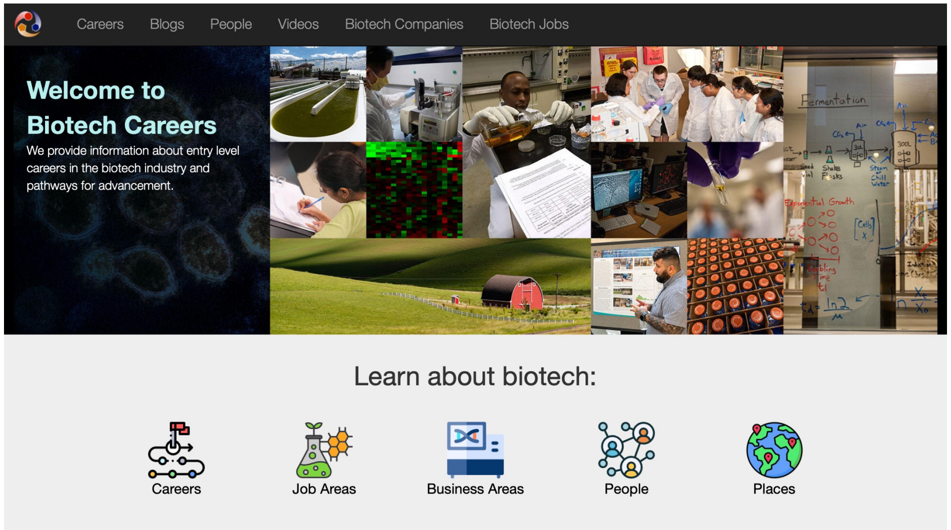
Envisioning the Next Bioscience Workforce

Current & Upcoming Events

- Spring 2024 InnovATEBIO ATE Project Talks:
  Fri, Feb 2 2024 to Fri, May 3 2024
- Virtual Bioscience Career Preparation Series:
  Fri, Feb 23 2024 to Fri, May 24 2024
- Diversity, Engagement & Recruitment Strategies:
  Fri, Mar 1 2024

View all upcoming events
View past events

InnovATEBIO.org
DUE 1901984



Welcome to Biotech Careers

We provide information about entry level careers in the biotech industry and pathways for advancement.

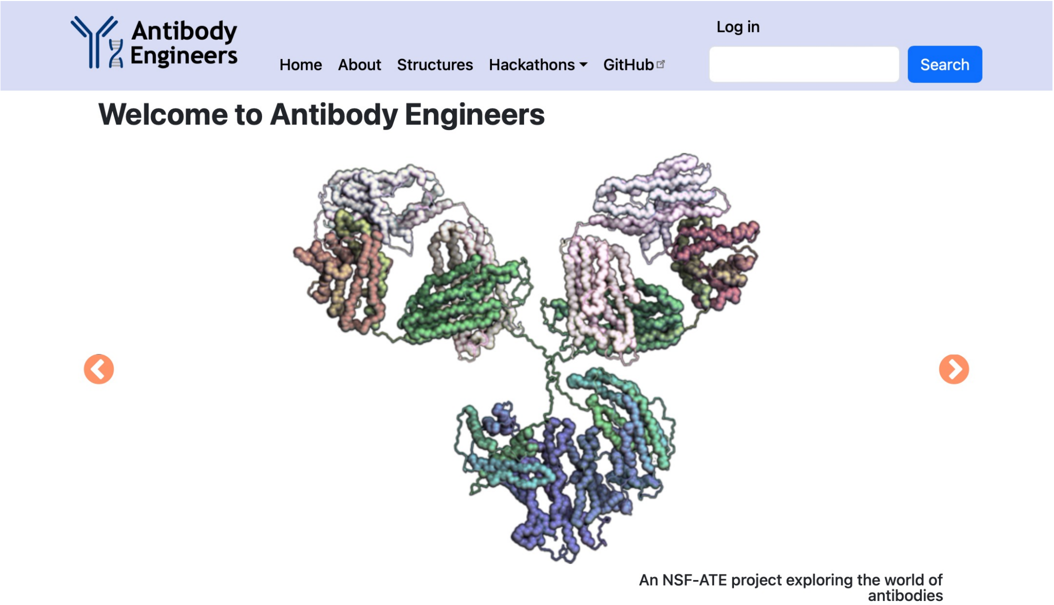
Learn about biotech:

Biotech-Careers.org
DUE 1764225

NSF Advanced Technological Education (ATE)

Supports the education of technicians for the high-technology fields that drive our nation's economy

Antibodies are major biotech products

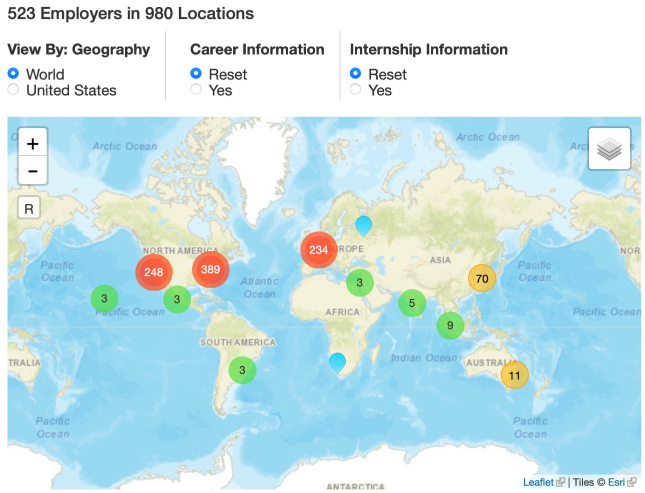


Welcome to Antibody Engineers

An NSF-ATE project exploring the world of antibodies

Antibody-Engineers.org

>500 Companies develop antibody-based products



523 Employers in 980 Locations

View By: Geography	Career Information	Internship Information
● World	● Reset	● Reset
○ United States	○ Yes	○ Yes

https://www.biotech-careers.org/company-core-activity/antibodies

Antibody Engineers

Project aims 1) develop modules to support **course-based undergraduate research experiences**. (CUREs); 2) **investigate hackathons** as a novel strategy for engaging participants in collaborative curriculum development.

Antibody Engineers is funded in part by the National Science Foundation DUE 2055036

Machine learning and antibodies

De novo antibody design

Immunotherapy

Humanizing mouse monoclonals | Improve stability/solubility

Tune binding affinities (specificity) | Convert Fab to VHH

CAR-T | Multivariate

Other applications

Diagnostic reagents | Flow cytometry | Staining …
Detect proteins in non-model organisms

image: Flaticon.com

Project goals

Motivation

- ML/AI is hot
- Antibodies are important
- Antibodies are used heavily in community college workforce education
- We get requests for ways to teach ML

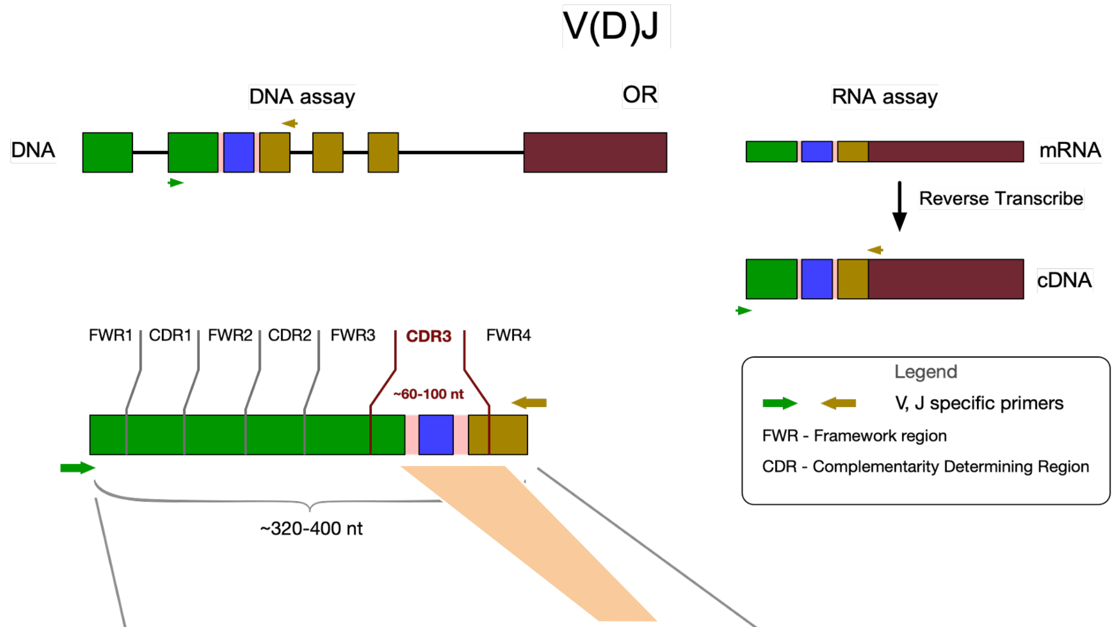
ML Education Challenges

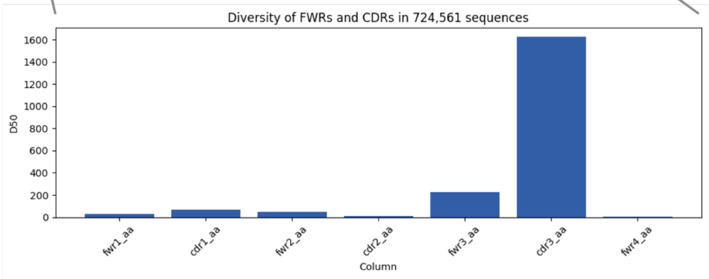
- Vocabulary, methods, appropriateness
- Infrastructure: data, tools, models
- Reproducing papers is hard
- Examples lack context
- Teaching: sysadmin >> coding

Can we?

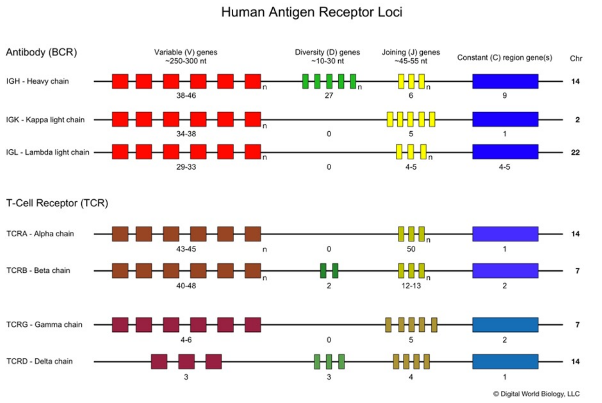
- Focus on a few concepts (regression, neural net, language models)
- Identify illustrative data sets
- Create infrastructure, libraries, install commands/scripts
- Document steps and concepts
- Accommodate a range of experience

Antibody diversity results from genetic recombination

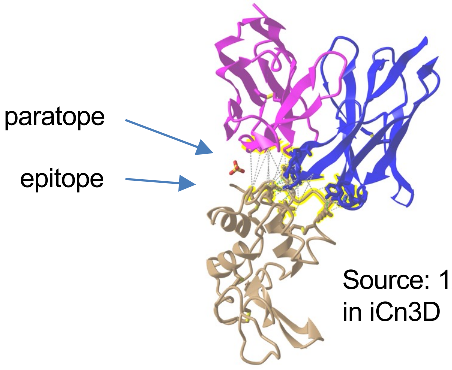




Diversity of FWRs and CDRs in 700,000+ sequences

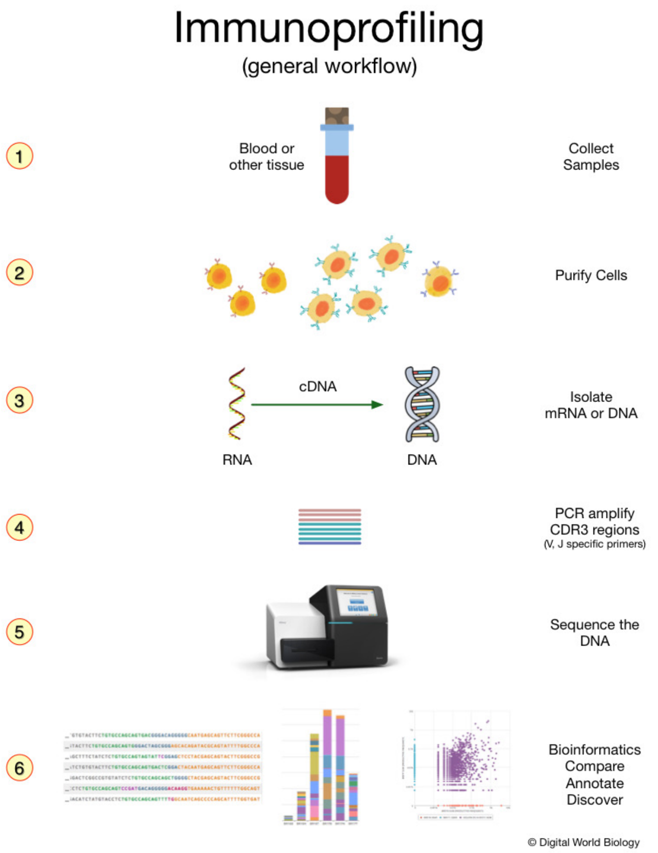


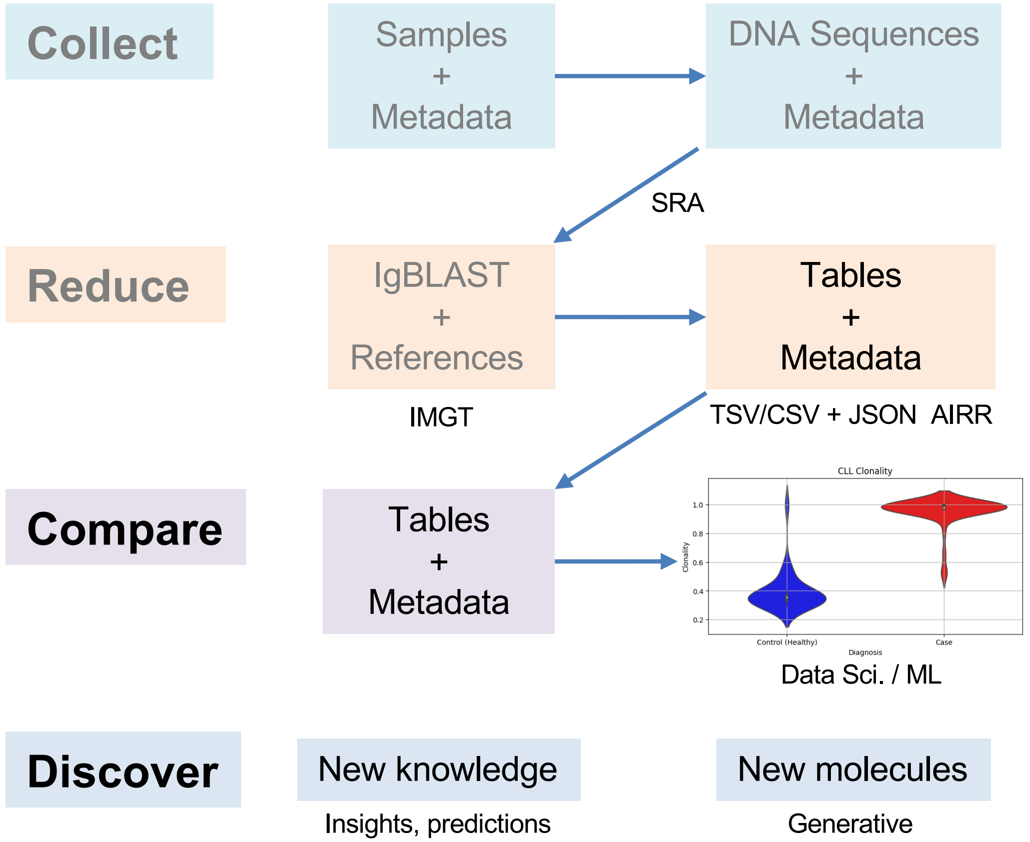
Antibody diversity results from imprecise gene rearrangements + mutation



Source: 1A2Y in iCn3D

Deep Ab DNA sequencing: workflow & data





Collect

Samples + Metadata → DNA Sequences + Metadata

SRA

Reduce

IgBLAST + References → Tables + Metadata

IMGT

TSV/CSV + JSON AIRR

Compare

Tables + Metadata →

Data Sci. / ML

Discover

New knowledge

Insights, predictions

New molecules

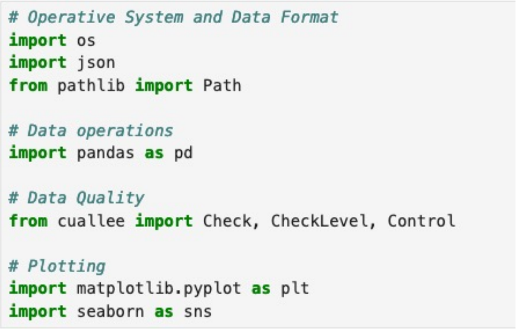
Generative

SRA = NCBI Short Read Archive; **IMGT** = ImMunoGenTics; **AIRR** = Adaptive Immune Receptor Repertoire

Tools and data

Tools

- Data Science



```
# Operative System and Data Format
import os
import json
from pathlib import Path

# Data operations
import pandas as pd

# Data Quality
from cuallee import Check, CheckLevel, Control

# Plotting
import matplotlib.pyplot as plt
import seaborn as sns
```

- Machine Learning

  AbLang1/2
  https://github.com/oxpig/AbLang,
  https://github.com/TobiasHeOl/AbLang2

  Generative
  Work in progress,

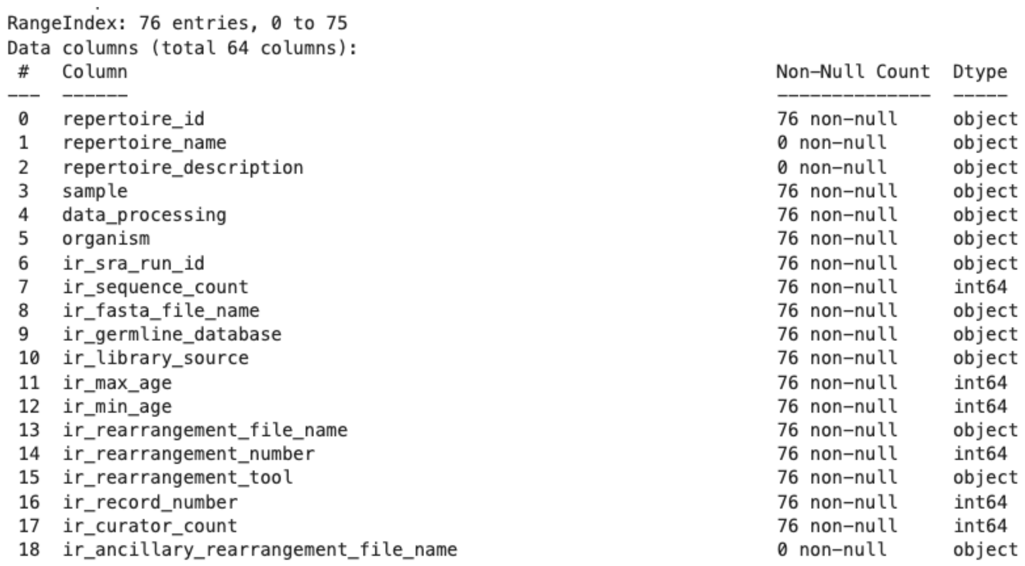
Data

- **Oxford Protein Informatics Group**
  https://opig.stats.ox.ac.uk/resources
  - >1 billion sequences from 80 studies
  - COV-AbDaB - 12,916 sequence CSV
    (all published/patented antibodies and nanobodies able to bind to coronaviruses, including SARS-CoV2, SARS-CoV1, and MERS-CoV)
- **iReceptor**
  http://ireceptor.irmacs.sfu.ca
  - 5.2 Billion annotated sequences from 10,019 repositories
  - Cancer case/control (1M+ sequences)
  - Somatic Mutation (1M+ sequences)
- NCBI – SRA
- IEDB

Exploring Data: TSV/CSV file (cancer case/control)

- AIRR => ~152 columns
- df.shape => (1063925, 152)
- +/- Does not include metadata (JSON file)

- Learn immune receptor biology from the data
- Many caveats: biology, lab, informatics

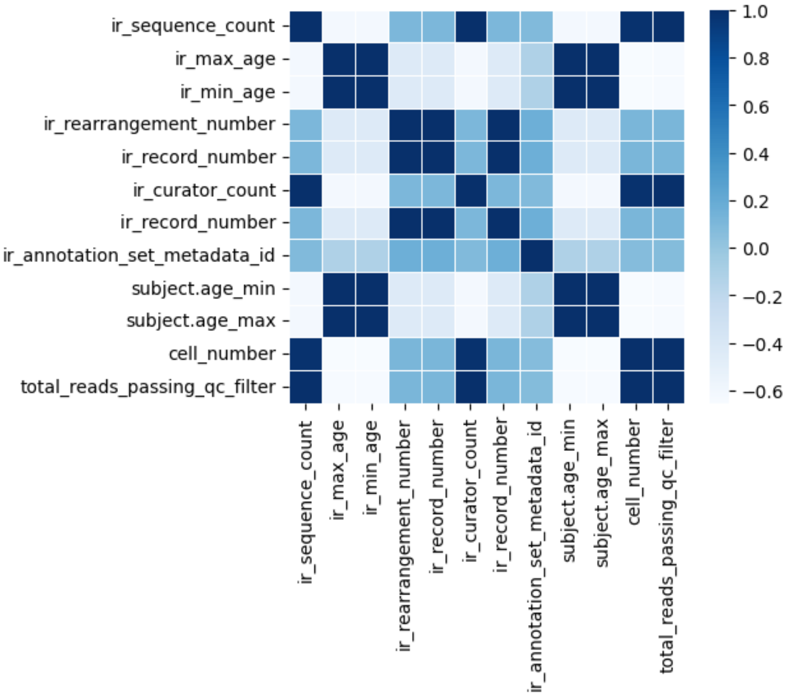


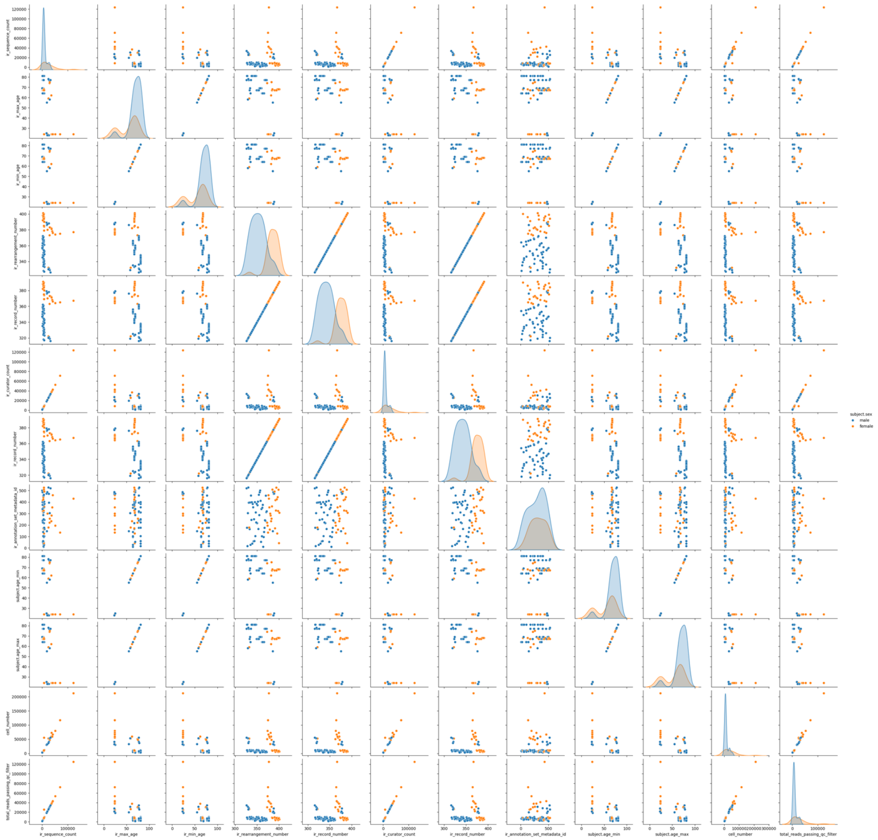
```
RangeIndex: 76 entries, 0 to 75
Data columns (total 64 columns):
 #   Column                                                           Non-Null Count  Dtype
---  ------                                                           --------------  -----
 0   repertoire_id                                                    76 non-null     object
 1   repertoire_name                                                  0 non-null      object
 2   repertoire_description                                           0 non-null      object
 3   sample                                                           76 non-null     object
 4   data_processing                                                  76 non-null     object
 5   organism                                                         76 non-null     object
 6   ir_sra_run_id                                                    76 non-null     object
 7   ir_sequence_count                                                76 non-null     int64
 8   ir_fasta_file_name                                               76 non-null     object
 9   ir_germline_database                                             76 non-null     object
 10  ir_library_source                                                76 non-null     object
 11  ir_max_age                                                       76 non-null     int64
 12  ir_min_age                                                       76 non-null     int64
 13  ir_rearrangement_file_name                                       76 non-null     object
 14  ir_rearrangement_number                                          76 non-null     int64
 15  ir_rearrangement_tool                                            76 non-null     object
 16  ir_record_number                                                 76 non-null     int64
 17  ir_curator_count                                                 76 non-null     int64
 18  ir_ancillary_rearrangement_file_name                             0 non-null      object
```

AIRR = Adaptive Immune Receptor Repertoire

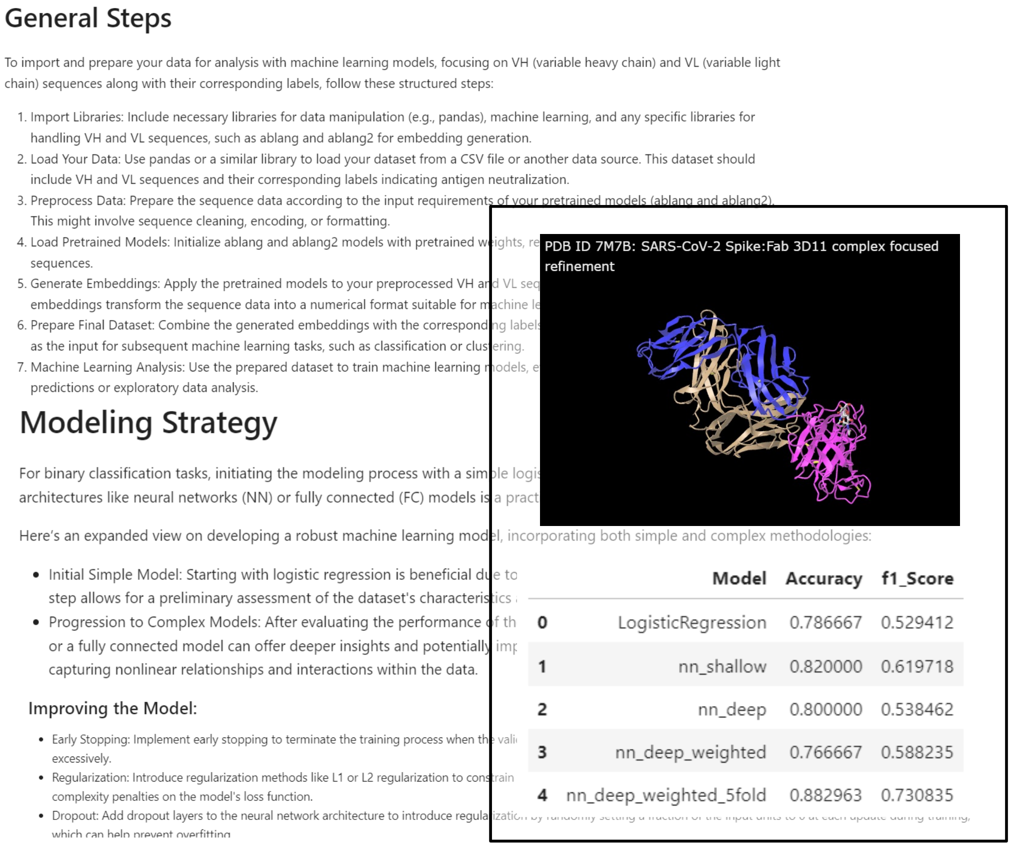
Exploring Data: Data correlations

Plots are used to visualize data correlations between columns





Machine learning: CoV-AbDab neutralizing Abs



General Steps

To import and prepare your data for analysis with machine learning models, focusing on VH (variable heavy chain) and VL (variable light chain) sequences along with their corresponding labels, follow these structured steps:

1. Import Libraries: Include necessary libraries for data manipulation (e.g., pandas), machine learning, and any specific libraries for handling VH and VL sequences, such as ablang and ablang2 for embedding generation.
2. Load Your Data: Use pandas or a similar library to load your dataset from a CSV file or another data source. This dataset should include VH and VL sequences and their corresponding labels indicating antigen neutralization.
3. Preprocess Data: Prepare the sequence data according to the input requirements of your pretrained models (ablang and ablang2). This might involve sequence cleaning, encoding, or formatting.
4. Load Pretrained Models: Initialize ablang and ablang2 models with pretrained weights, re sequences.
5. Generate Embeddings: Apply the pretrained models to your preprocessed VH and VL seq embeddings transform the sequence data into a numerical format suitable for machine le
6. Prepare Final Dataset: Combine the generated embeddings with the corresponding labels as the input for subsequent machine learning tasks, such as classification or clustering.
7. Machine Learning Analysis: Use the prepared dataset to train machine learning models, e predictions or exploratory data analysis.

Modeling Strategy

For binary classification tasks, initiating the modeling process with a simple logis architectures like neural networks (NN) or fully connected (FC) models is a pract

Here's an expanded view on developing a robust machine learning model, incorporating both simple and complex methodologies:

- Initial Simple Model: Starting with logistic regression is beneficial due to step allows for a preliminary assessment of the dataset's characteristics
- Progression to Complex Models: After evaluating the performance of th or a fully connected model can offer deeper insights and potentially imp capturing nonlinear relationships and interactions within the data.

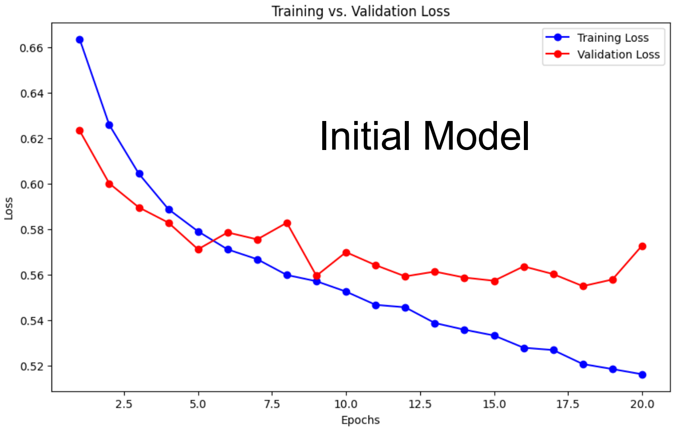
Improving the Model:

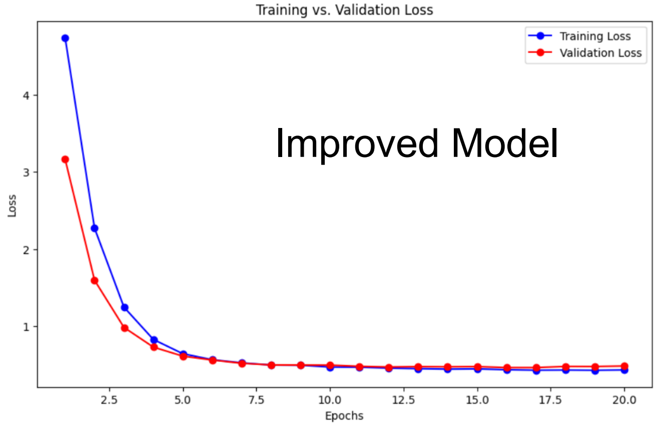
- Early Stopping: Implement early stopping to terminate the training process when the vali excessively.
- Regularization: Introduce regularization methods like L1 or L2 regularization to constrain complexity penalties on the model's loss function.
- Dropout: Add dropout layers to the neural network architecture to introduce regularization by randomly setting a fraction of the input units to 0 at each update during training, which can help prevent overfitting

PDB ID 7M7B: SARS-CoV-2 Spike:Fab 3D11 complex focused refinement

	Model	Accuracy	f1_Score
0	LogisticRegression	0.786667	0.529412
1	nn_shallow	0.820000	0.619718
2	nn_deep	0.800000	0.538462
3	nn_deep_weighted	0.766667	0.588235
4	nn_deep_weighted_5fold	0.882963	0.730835

. . .





© www.digitalworldbiology.com

Progress

Can we?	Progress
Focus on a few concepts (regression, neural net, language models)	• Working with very large TSV files + meta data • Evaluating data quality, correlations • Principle component analyses (PCA) • Machine learning for classification
Identify illustrative data sets	• **iReceptor cancer case/control:** Pandas, clonality concepts, data correlation, distribution, basis for how to proceed • **OPIG CoV-AbDab:** Pandas, data exploration, ML to predict neutralizing antibodies
Create infrastructure, libraries, install commands/scripts	• Many jupyter notebooks to build from • Include the needed packages
Documents steps and concepts	• Some of the jupyter notebooks are well annotated & explanatory • Markdown serves as documentation
Accommodate a range of experience	• Team was learning antibody concepts • Members with strong computer backgrounds taught • Use cases support novice and strong programming experience • Data analysis concepts and introduction to machine learning

Love hackathons?

Antibody Engineers

Next Hackathon: Mon Aug 5th - Thu Aug 8th, 2024
Required experience: students, faculty, new to programming, industry/academic experts

Projects:

- **NIST CHO cells:** cell line stability and developing the materials and an ELISA to measure the antibody
- **CEDAR:** IEDB's Cancer Epitope Database and Analysis Resource, explore neoantigens, antigen processing, and immunotherapies.
- **Antibodies & AI:** Continue the presented work
- **Affordable Antibody Engineering:** Purifying single-chain antibodies to green fluorescent protein and ELISAs
- **Project Sea Star:** Can we use homology modeling to find antibodies for non-model organisms?
- **Pathogens:** Use the iCn3D, the SabDab database, and viral sequence databases (nextstrain.org) to explore sequence variation and it's impact on antibody binding.
- **Immune Defense:** Help test an immunology-based video game
- **iCn3D datasets and collections:** Identify antibody-antigen structures that will be useful for teaching and developing protocols that faculty can use in creating their own molecular datasets.

Learn More: https://antibody-engineers.org/event/antibody-engineering-hackathon-august-2024
Questions: todd@digitalworldbiology.com