

# AI-READY DATASETS

*Reuben Sarwal, Bakar Computational  
Health Sciences Institute, University of  
California, San Francisco*



# Agenda



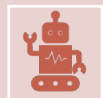
Overview of ImmPort



What makes an AI-ready dataset

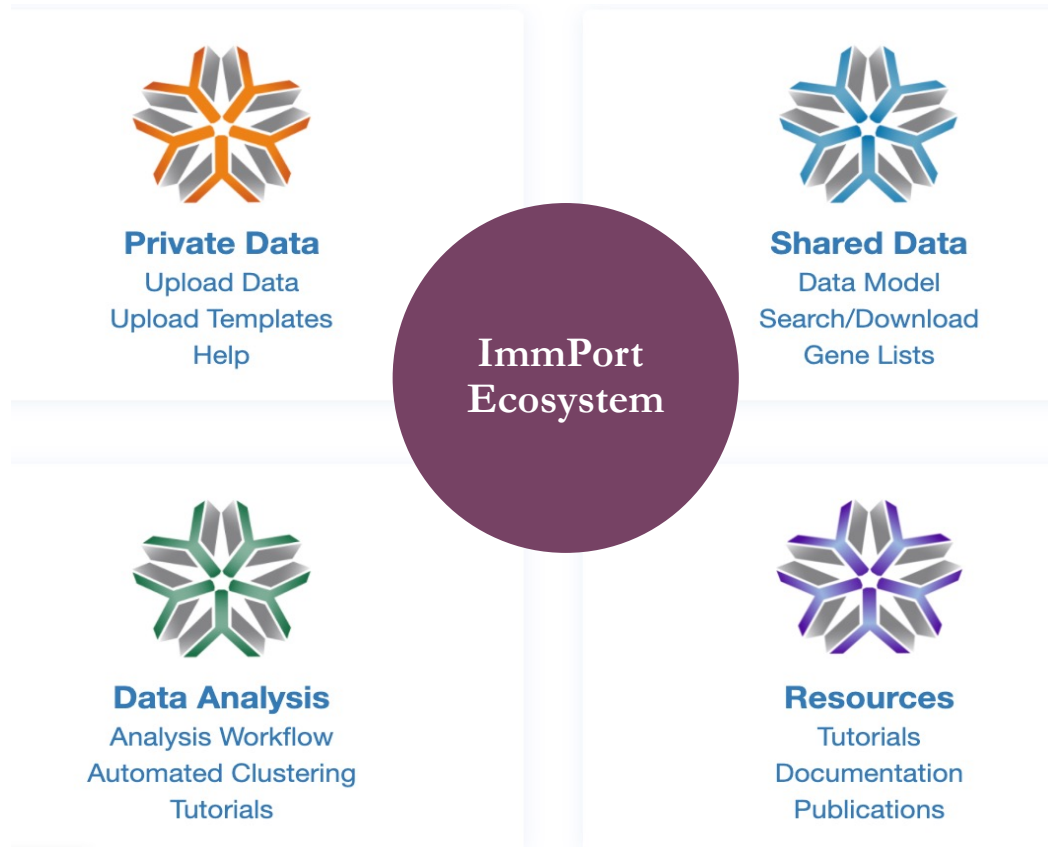


ImmPort and AI-ready data



Examples of data made AI-ready

# ImmPort data portal was developed to collect and share research and clinical trials data from NIAID/DAIT funded researchers



## FAIR Principles

Make your data:

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

### Findable

- Descriptive metadata
- Persistent

### Accessible

- Determining what to share
- Participant consent and risk management
- Access status

### Interoperable

- XML standards
- Data Documentation Initiative
- CDISC

### Reusable

- Rights and licence models
- Permitted and non-permitted use

<http://datafairport.org/>

# ImmPort Shares Data from Major NIAID-funded Programs and External Organizations



## Human Immunology Project Consortium



### COVID-19 - ImmPort

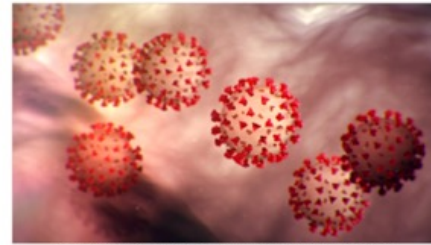


Image CDC Coronavirus Disease 2019 – COVID 19, Alissa Eckert

### Coronavirus Disease 19

Powered by ImmPort



### BILL & MELINDA GATES foundation

Powered by ImmPort

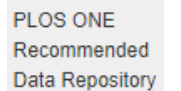
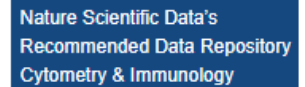


- Immunophenotyping Assessment in a COVID-19 Cohort (IMPACC)
- Serological Sciences Network (SeroNet)
- Multisystem Inflammatory Syndrome in Children (MIS-C)
- Impact of Initial Influenza Exposure on Immunity in Infants (U01)
- Atopic Dermatitis Research Network (ADRN)
- Population Genetics Analysis Program
- Protective Immunity for Special Populations
- HLA Region Genomics in Immune-mediated Diseases
- Modeling Immunity for Biodefense
- Reagent Development for Innate Immune Receptors
- Adjuvant Development Program
- Immunity in Neonates and Infants
- Asthma and Allergic Diseases Cooperative Research Centers
- HLA and KIR Region Genomics in Immune-Mediated Diseases
- Cooperative Study Group for Autoimmune Disease Prevention
- Immunobiology of Xenotransplantation
- Centers for Medical Countermeasures against Radiation Consortium
- Inner City Asthma Consortium
- Systems Approach to Immunity and Inflammation
- Innate Immune Receptors and Adjuvant Discovery Program
- Maintenance of Macaque Specific Pathogen-Free Breeding Colonies
- Non-human Primate Transplantation Tolerance Cooperative Study Group
- Consortium for Food Allergy Research
- Development of Sample Sparing Assays for Monitoring Immune Responses (U24)
- Asthma and Allergic Diseases Clinical Research Consortium (AACRC)
- The Clinical Islet Transplantation (CIT) Consortium
- Autoimmunity Centers of Excellence (ACE)
- Clinical Trials in Organ Transplantation (CTOC)
- Human Immunology Project Consortium (HIPC)
- Collaborative Influenza Vaccine Innovation Centers (CIVICs)
- Centers for Research in Emerging and Infectious Diseases (CREID)
- Cooperative Centers on Human Immunology
- Impact of Initial Influenza Exposure on Immunity in Infants (U01)
- A Multidisciplinary Approach to Study Vaccine-elicited Immunity and Efficacy Against Malaria (MVIE)

## 20 Years of FAIR Data Sharing

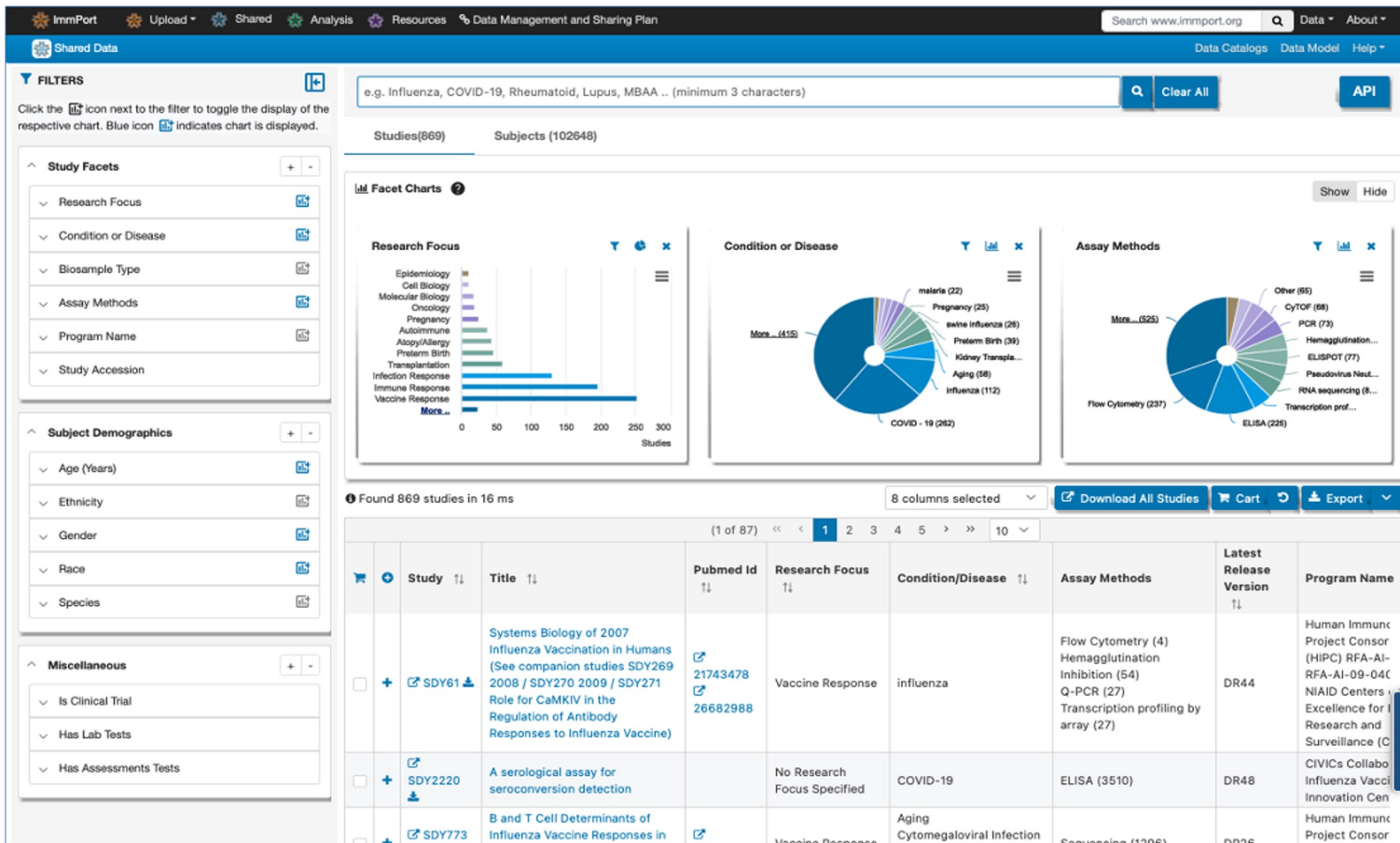


<http://doi.org/10.17616/R30J3F>









# ImmPort Shared Data Browser (Cohort Discovery Tool)

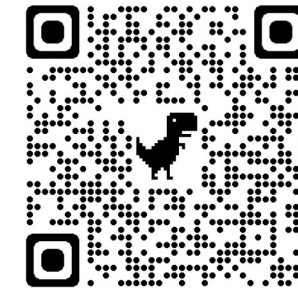
ImmPort currently shares **1047** studies encompassing a range of research areas, species & assay types including 181 Clinical Trials data



Data Summary: Release 52.1, July 2024

Click on the counts with links to visualize the count breakdown

 Studies 1047	 Subjects 162392	 Diseases 165
 Experiments 4041	 Total Results 7297114	 Lab Tests 1305129



<https://import.org/shared/search>

# ImmPort Shared Data Browser (Cohort Discovery Tool): Easy to Investigate Multiple or Individual Studies

Summary Design Adverse Event Assessment Interventions Medications Substance Demographics Lab Tests Mechanistic Assays Study Files

+ -

Summary

Accession	SDY2564
Title	SARS-CoV-2 Spike-Binding Antibody Longevity and Protection from Reinfection with Antigenically Similar SARS-CoV-2 Variants.
DOI	<a href="https://doi.org/10.21430/m3o55f20tz">10.21430/m3o55f20tz</a>
Brief Description	The PARIS (Protection Associated with Rapid Immunity to SARS-CoV-2) cohort follows health care workers with and without documented coronavirus disease 2019 (COVID-19) since April 2020. We report our findings regarding severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike-binding antibody stability and protection from infection in the pre-variant era. We analyzed data from 400 health care workers (150 seropositive and 250 seronegative at enrollment) for a median of 84 days.
Research Focus	Immune Response
Condition Studied	COVID-19
Start Date	2020-03-01

Summary Design Adverse Event Assessment Interventions Medications Substance Demographics Lab Tests Mechanistic Assays Study Files

+ -

Mechanistic Assays

Search Mechanistic Assay ↻ 📄 Export

Accession	Title	Measurement Technique	# of Samples
EXP38975	ELISA_1B	ELISA	813
EXP38976	ELISA_1B-C	ELISA	125

<< < 1 > >> 10 ▾ Showing 1 to 2 of 2

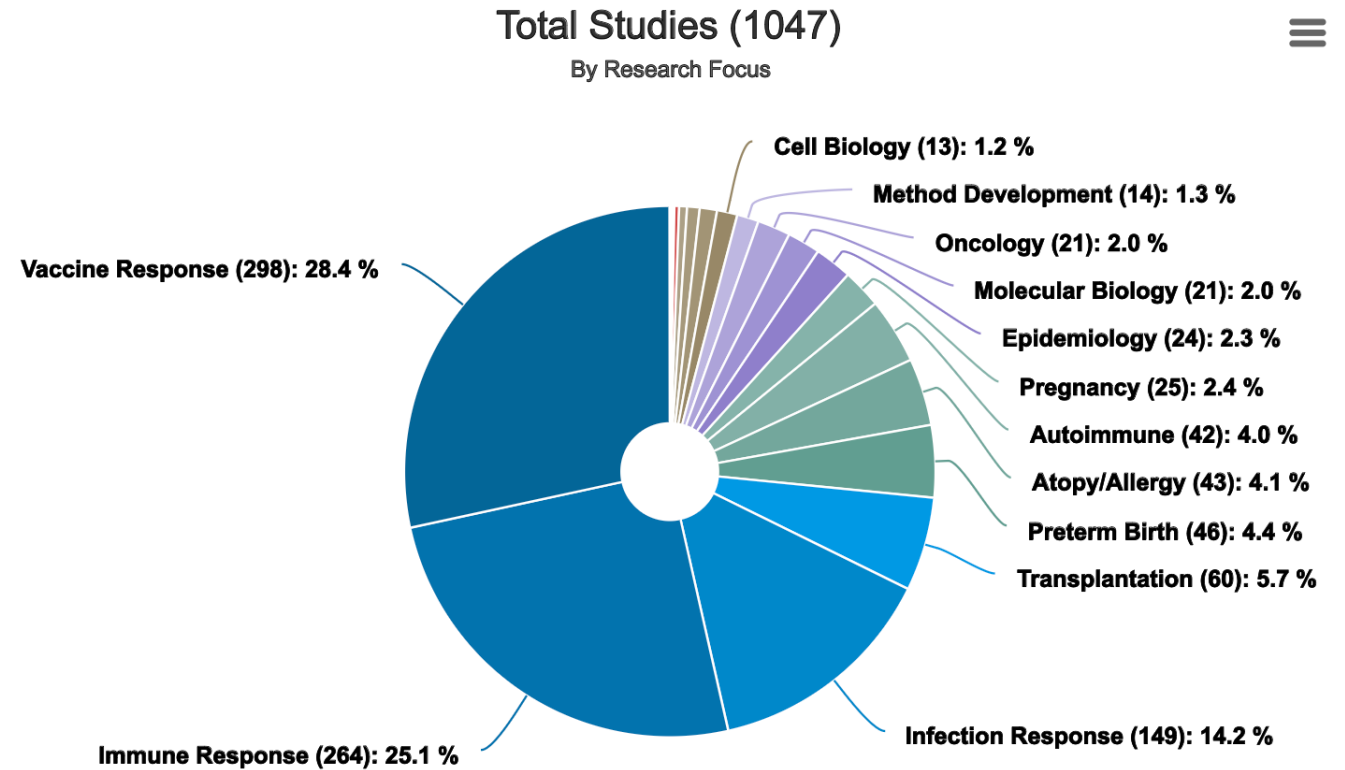
Experiment Detail for EXP38975 Protocols Reagents Treatments

Protocol	Name	Filename	Type	Description
PTL11721	All methods and details pertaining to study-121 as shared in the publication.	methods-121.PTL11721.docx	Experiment	All methods and details pertaining to study-121 as shared in the publication.

# Data Summary: Release 52.1, July 2024

Click on the counts with icon to visualize the count breakdown

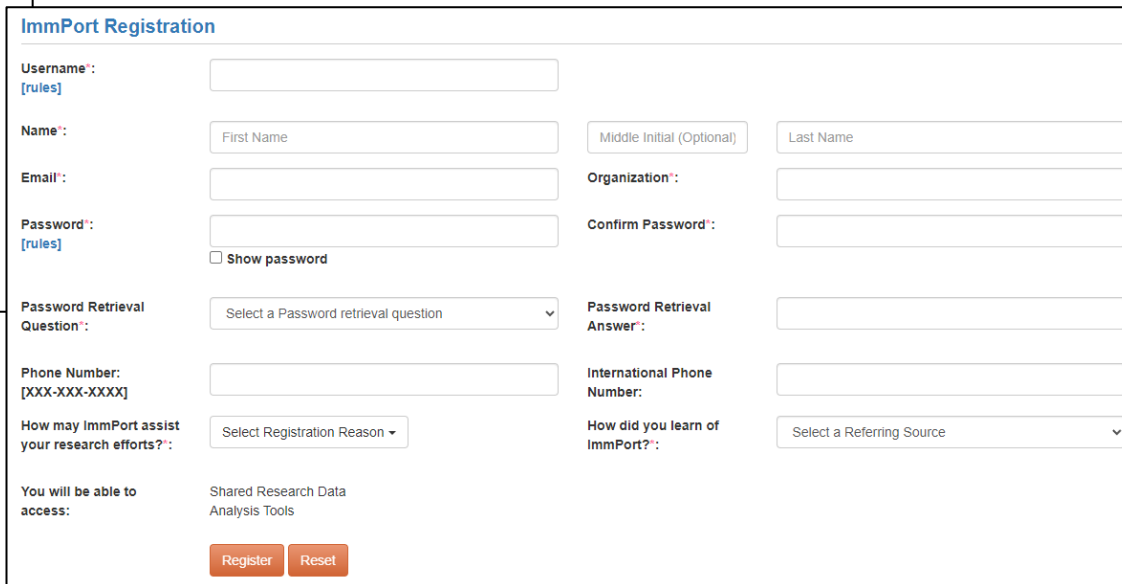
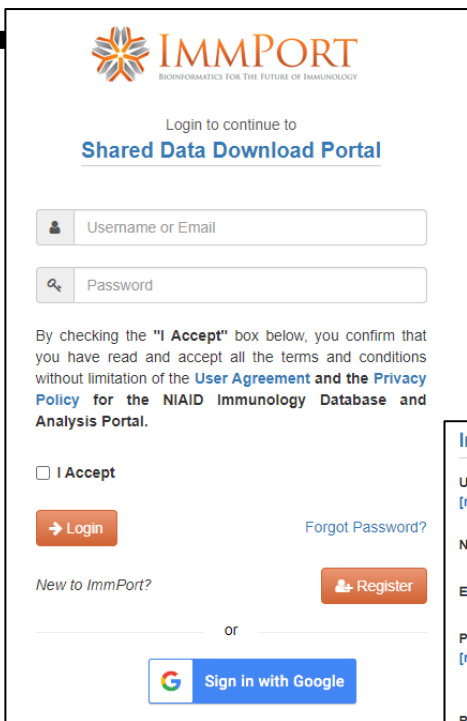
Studies	1047
Subjects	162392
Diseases	165
Protocols	2407
Experiments	4041
Total Results	7297114
Elisa Results	513602
Elispot Results	73829
Flow Cytometry Results	5212280
PCR Results	129657
HAI Results	32470
HLA Typing Results	149342
Luminex/MBAA Results	1095548
Neutralizing Antibody Results	37053
GEO/SRA/Other	53333
Lab Tests	1305129
Total Test Panels	435



# Accessible

## ImmPort Registration & Login

- ImmPort study metadata (CDT Search) is browsable without login
- Registration and acceptance of Data Use Agreement is required to upload or download data
- Registration is free, simple, and immediate



## ImmPort Application Programming Interfaces (APIs)

API Documentation	Overview
Authentication	The ImmPort API includes endpoints for:
Sample API call using Shell Commands	
Tools for communicating with the ImmPort Data Query API	<ul style="list-style-type: none"><li>• Authentication/Authorization</li><li>• Shared Data<ul style="list-style-type: none"><li>• Study Data</li><li>• Assay Result Data</li><li>• Controlled Vocabulary or Lookup Tables</li><li>• Study Metadata (Used by UI)</li><li>• Download Files</li></ul></li><li>• Data Upload</li><li>• Batch Updater</li></ul>
ImmPort Auth Service >	
Shared Data API >	
Data Upload API >	
Batch Updater API >	

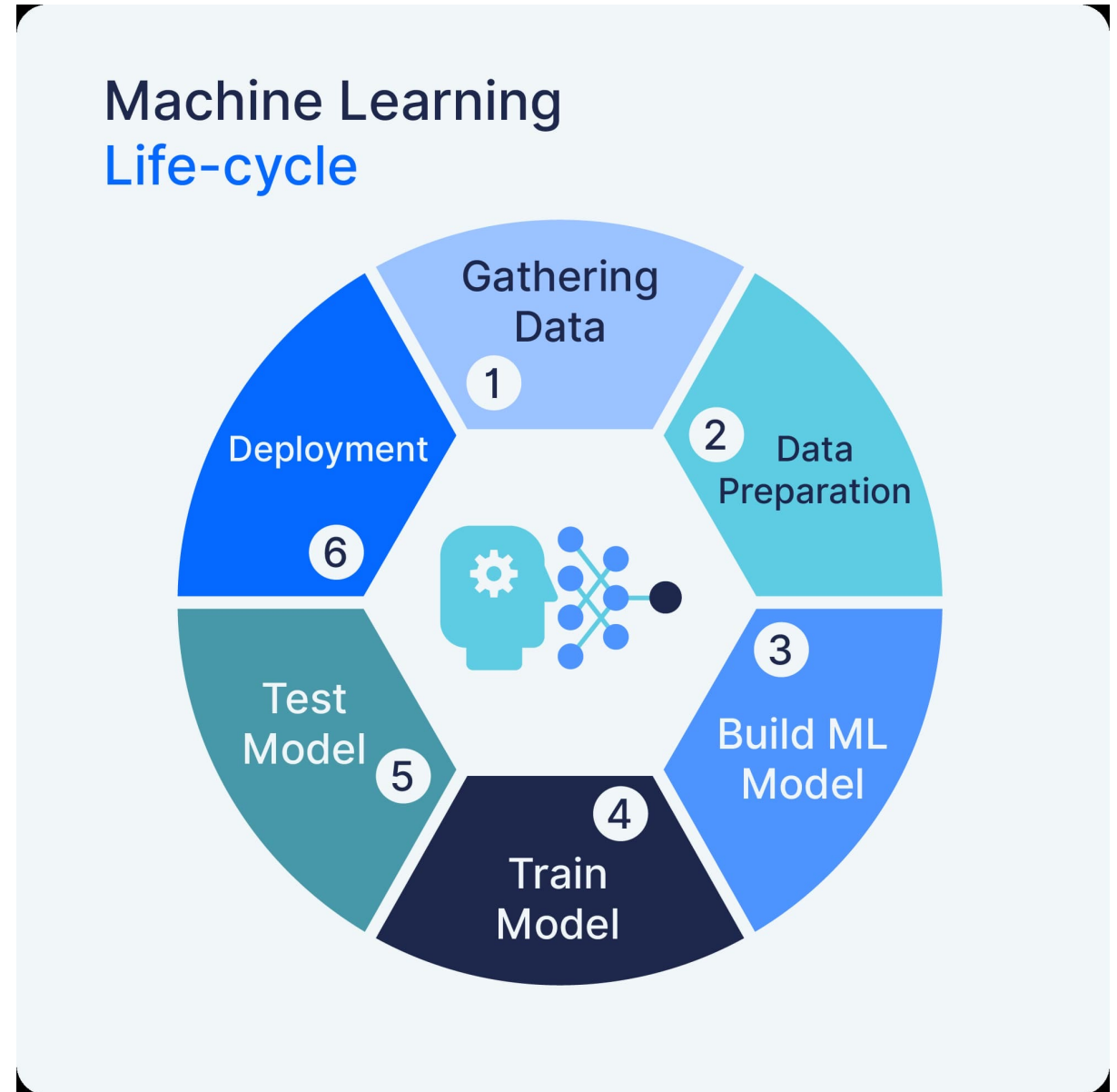
<https://docs.immport.org/apidocumentation/>

- ImmPort offers several APIs with detailed documentaiton for use

<https://www.immport.org/auth/login>



**HOW CAN WE  
UTILIZE ALL  
THIS DATA TO  
LEVERAGE  
ML/AI  
MODELS ?**



# The National Artificial Intelligence Research Resource (NAIRR) Pilot

The NAIRR Pilot aims to connect U.S. researchers and educators to computational, data, and training resources needed to advance AI research and research that employs AI. Federal agencies are collaborating with government-supported and non-governmental partners to implement the Pilot as a

## Why Create AI Ready Datasets?

- Spur innovation
- Increase diversity of talent
- Improve capacity
- Advance trustworthy AI

# ImmPort Featured As NAIRR AI-Ready Resource

**NAIRR Pilot** National Artificial Intelligence Research Resource Pilot

Current Opportunities ▾ NAIRR Secure Awarded Projects About ▾ Help ▾

## Open Data, Models, and More

This list does not include allocatable resources for research or education/teaching; please see the [Researcher Call](#) and [Classroom/Educator Call](#) for those resources.

All Courses **Datasets** Documentation Models Secure Other

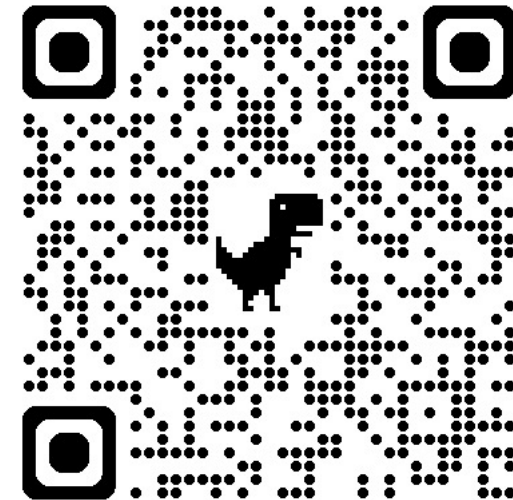
20 results in Datasets [Show all categories](#) [EXPAND ALL](#) [COLLAPSE ALL](#)

- NASA Earth Science AI Training Datasets ▾
- NASA Harmonized Landsat Sentinel-2 (HLS) Foundation Model ▾
- NASA HLS Burn Scars training dataset ▾
- NASA Multi-temporal crop classification training dataset ▾
- NASA Science Mission Directorate Curated AI training sets ▾
- NIH ImmPort ▴

Immport [Go to resource](#)

ImmPort from NIAID is a publicly accessible data sharing platform supporting immunology research and clinical studies. ImmPort offers curated datasets and reference datasets that adhere to the FAIR Principles and is certified by CoreTrustSeal.

NIH Medical Imaging and Data Resource Center (MIDRC) ▾



---

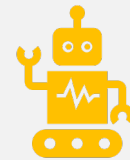
# Frequently Asked Questions



**What does AI-ready dataset mean?**



**How to make the datasets AI-ready?**



**Where can I get data to train my AI/ML model?**

# What is an AI-ready Dataset?



## DATA PREPARATION

SINGLE/AGGREGATED,  
NULL  
VALUES/OUTLIERS,  
ANNOTATION



## DATA QUALITY

COMPLETENESS,  
CONSISTENCY,  
INTEGRITY,  
PROVENANCE



## DATA DOCUMENTATION


METADATA, CODE  
BOOK/DATA  
DICTIONARY,  
ONTOLOGY, DOI



## DATA ACCESS

FORMATS, DELIVERY,  
USAGE RIGHTS,  
SECURITY/PRIVACY

**Clean and process the data into a usable format that can power an AI application**

Dataset	Description	AI-Readiness Assessment
<p><a href="#">10k Immunomes</a> </p> <p><a href="#">Download (138.2 MB)</a></p> <p><b>MD5 Checksum</b> </p> <p>303639c62f72196f65880f31f1819fa2</p> 	<p>The 10,000 Immunomes Project is a reference dataset for human immunology, derived from over 10,000 control subjects in the NIAID ImmPort Database. The dataset provides comprehensive profiles of the human immune system within healthy cohorts, representing various races.</p> <p><a href="#">Read the publication in Cell Reports</a></p> <p><a href="#">Browse ImmPort Studies in Shared Data</a></p> <p>Review the source code on <a href="#">GitHub</a> or <a href="#">Docker</a></p>	<p><a href="#">10k Immunomes Assessment</a> </p>



ImmPort AI-Readiness Assessment	
Dataset Name	<input type="text"/>
Dataset Version	<input type="text"/>
Dataset Location/Link	<input type="text"/>
Assessor Name	<input type="text"/>
Assessor Email	<input type="text"/>
	<b>Responses</b>
Is this raw data or a derived/processed data product?	(please select option) <input type="text"/>
Is this observational data, simulation/model output, or synthetic data?	(please select option) <input type="text"/>
Is the data single-source or aggregated from several sources?	(please select option) <input type="text"/>
<b>Data Quality</b>	
Will the dataset be updated?	(please select option) <input type="text"/>
If the data will be updated, how often will it be updated?	(please select option) <input type="text"/>
If the data will be updated, choose the update frequency that best describes the dataset.	(please select option) <input type="text"/>
Will there be different stages of the update (e.g., updated with preliminary data first and replaced by a later update of the full record)?	(please select option) <input type="text"/>
If yes, what is the delay between different stages?	(short answer) <input type="text"/>
Should the new version of the dataset supersede the current version?	(please select option) <input type="text"/>
Please provide an explanation for "Other" response.	(short answer) <input type="text"/>
Is there any documentation about the completeness of the dataset?	(please select option) <input type="text"/>
If there is documentation regarding the completeness of the dataset, please provide a link to report/document.	(link to report/document) <input type="text"/>

# We Have Created An Assessment Geared Towards Immunology Research

# Assessment Contains Guidelines and Instructions

## Instructions and Scoring

- ~ Do your best to answer each of the questions in the AI-Readiness Assessment tab.
- ~ Some questions may be easiest to answer for those with expertise on a given dataset. Others may be readily available through metadata files.
- ~ After answering all the questions for a given section, use the drop-down menu to choose a score from 0 to 4 that you feel best summarizes your answers to the questions for that section. See below for guidelines on how to assign a score.
- ~ After you have assigned a score for each of the four sections, the overall score for the dataset (the average of each of the four section scores) will automatically be calculated in a formula cell at the bottom of the assessment form.
- ~ Use this assessment as a tool to decide when a dataset is ready for release, to prioritize potential improvements to the dataset, or simply as additional documentation containing valuable information about a dataset that may be useful for others desiring to use the dataset.

## Section Scoring Guidelines

Scores are intended to be a holistic evaluation of all elements in a particular section. Not all questions will be equally relevant or applicable to every dataset. Depending on the dataset, a similar answer to a question may be "positive" in one context and "negative" in another. Use your best judgement. We suggest the below scale:

- 0 - None** of the elements of AI-readiness addressed in this section have been considered and/or implemented (where appropriate).
- 1 - A few** of the elements of AI-readiness addressed in this section have been considered and/or implemented (where appropriate).
- 2 - Some** of the elements of AI-readiness addressed in this section have been considered and/or implemented (where appropriate).
- 3 - Most** of the elements of AI-readiness addressed in this section have been considered and/or implemented (where appropriate).
- 4 - All** of the elements of AI-readiness addressed in this section have been considered and/or implemented (where appropriate).

# Why Have an Assessment

- Clarity for data generators and downstream users!

## AI-Readiness Assessment Overview and Instructions

Artificial intelligence (AI) provides a powerful suite of tools that can improve our understanding of immunology. The burgeoning field of AI is often collaborative bringing together diverse data producers, stewards, and users across agencies, academia, and industry. Quality AI-ready open data provides the link that enables these cross-sectoral teams to succeed. AI-ready data not only can enable practical AI applications but also provides an opportunity to modernize data management practices for all use cases.

### Overview

The AI-Readiness Assessment is meant to provide an assessment tool applicable to immunology datasets that may be intended for AI applications, or considered valuable to potential AI applications. Thus, it is likely that the utility of this tool will grow in the future as AI application becomes more common and integrated with agency science pipelines. It is not intended to create an extra requirement or to suggest that all previously curated datasets must conform to certain levels of AI-readiness.



This tool can be used to:

- ~ Provide AI practitioners with information to understand the work needed to use a particular dataset,
- ~ Identify areas to improve target datasets to streamline AI applications,
- ~ Provide agency personnel with a line of evidence to justify / prioritize resource investment to improve key datasets.

ImmPort utilized the [checklist](#) published by ESIP as a starting point and modified it to be more relevant for immunology data, as well as adding a suggested scoring system. (For details on modifications made, see the Changelog tab.) The questions on the assessment encompass four broad areas of AI-readiness: data quality, data documentation, data access and data preparation.

### Instructions and Scoring

- ~ Do your best to answer each of the questions in the AI-Readiness Assessment tab.
- ~ Some questions may be easiest to answer for those with expertise on a given dataset. Others may be readily available through metadata files.
- ~ After answering all the questions for a given section, use the drop-down menu to choose a score from 0 to 4 that you feel best summarizes your answers to the questions for that section. See below for guidelines on how to assign a score.
- ~ After you have assigned a score for each of the four sections, the overall score for the dataset (the average of each of the four section scores) will automatically be calculated in a formula cell at the bottom of the assessment form.
- ~ Use this assessment as a tool to decide when a dataset is ready for release, to prioritize potential improvements to the dataset, or simply as additional documentation containing valuable information about a dataset that may be useful for others desiring to use the dataset.

Dataset	Description	AI-Readiness Assessment
<p>10k Immunomes <a href="#">↗</a></p> <p><a href="#">Download (138.2 MB)</a></p> <p>MD5 Checksum <a href="#">📄</a></p> <p>303639c62f72196f65880f31f1819fa2</p> 	<p>The 10,000 Immunomes Project is a reference dataset for human immunology, derived from over 10,000 control subjects in the NIAID ImmPort Database. The dataset provides comprehensive profiles of the human immune system within healthy cohorts, representing various races.</p> <p><a href="#">Read the publication in Cell Reports</a></p> <p><a href="#">Browse ImmPort Studies in Shared Data</a></p> <p>Review the source code on <a href="#">GitHub</a> or <a href="#">Docker</a></p>	<p>10k Immunomes Assessment <a href="#">📄</a></p> 

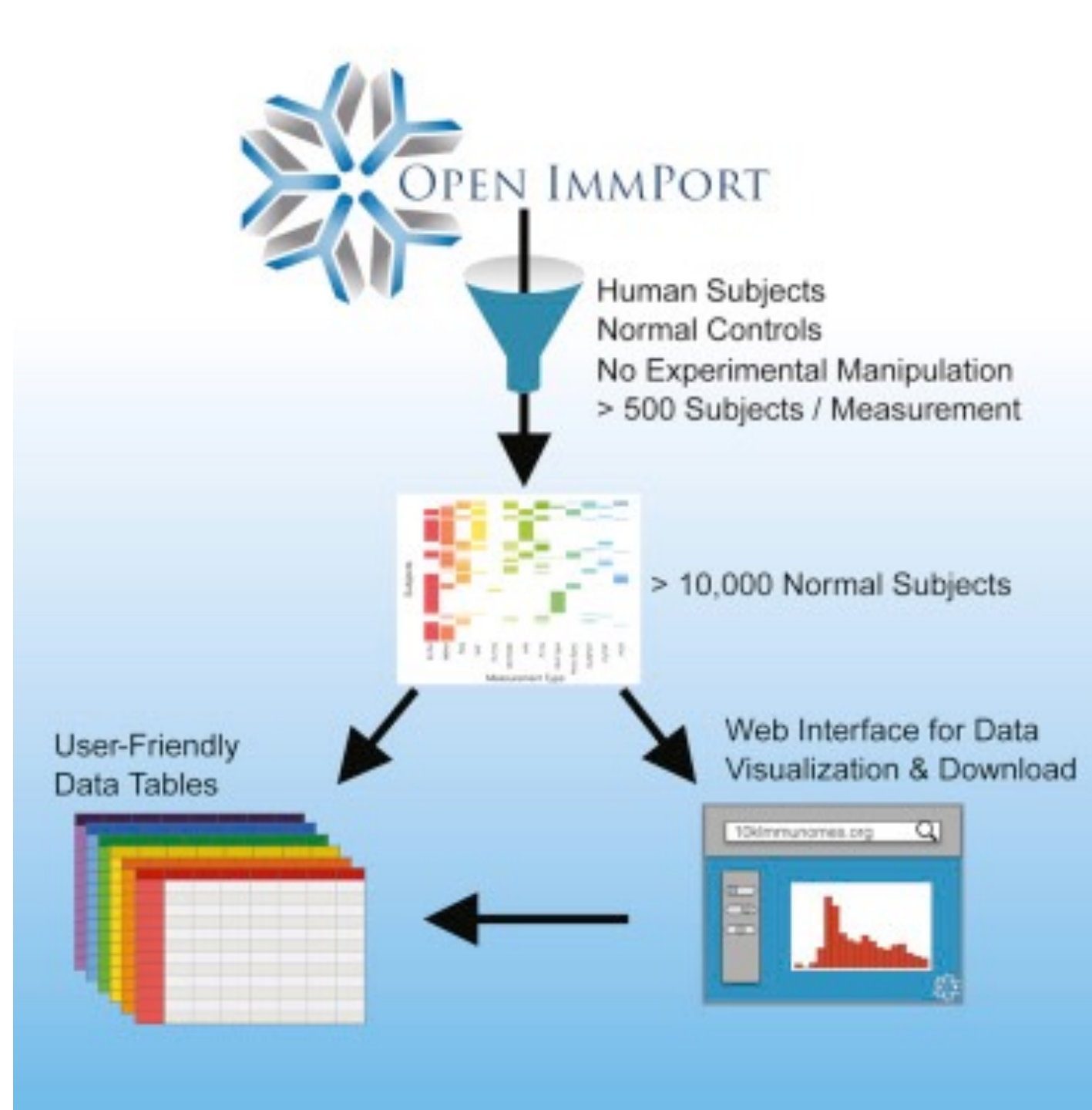
Is this raw data or a derived/processed data product?	<input type="radio"/> Derived
Is this observational data, simulation/model output, or synthetic data?	<input type="radio"/> Observed
Is the data single-source or aggregated from several sources?	<input type="radio"/> Aggregated
<b>Data Quality</b>	
Will the dataset be updated?	<input type="radio"/> No, it will not be updated
If the data will be updated, how often will it be updated?	<input type="radio"/> Not applicable
If the data will be updated, choose the update frequency that best describes the dataset.	<input type="radio"/> Not applicable
Will there be different stages of the update (e.g., updated with preliminary data first and replaced by a later update of the full record)?	<input type="radio"/> Not applicable
If yes, what is the delay between different stages?	(short answer)
Should the new version of the dataset supersede the current version?	<input type="radio"/> Not applicable
Please provide an explanation for "Other" response.	(short answer)
Is there any documentation about the completeness of the dataset?	<input type="radio"/> Yes
If there is documentation regarding the completeness of the dataset, please provide a link to report/document.	<input type="checkbox"/> 10K Data Completeness



# AI-powered opportunities from ImmPort studies and resources

Does ImmPort have AI-ready datasets?

- 10K Immunomes

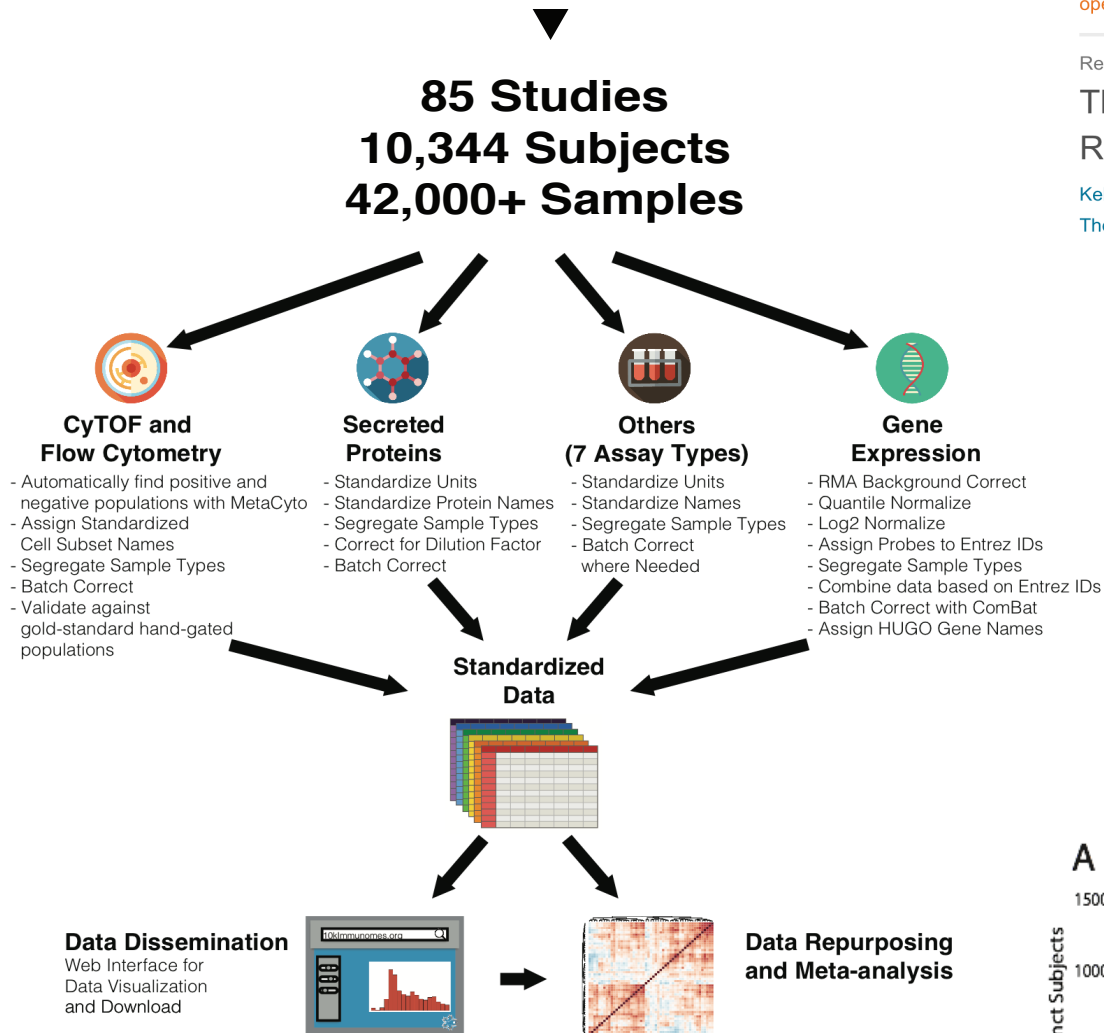
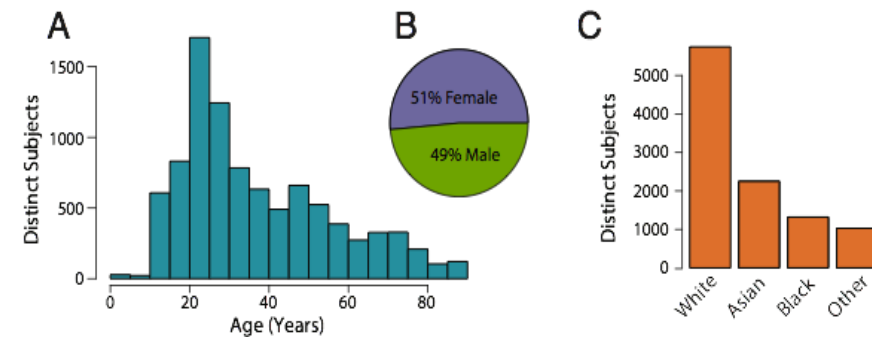


# 10KImmunes.org

## The 10,000 Immunomes Project: Building a Resource for Human Immunology

Kelly A. Zalocusky <sup>1, 2</sup>, Matthew J. Kan <sup>1, 2</sup>, Zicheng Hu <sup>1, 2</sup>, Patrick Dunn <sup>3</sup>, Elizabeth Thomson <sup>3</sup>, Jeffrey Wisner <sup>3</sup>, Sanchita Bhattacharya <sup>1, 2, 4</sup>, Atul J. Butte <sup>1, 2, 4, 5</sup>

- Large, diverse, cleaned reference dataset derived from ImmPort studies
- Interactive data visualization
- Custom control cohorts and standardized data download



### Data available in the 10,000 Immunomes Project

Total Samples	42117
Total Distinct Subjects	10344

### MEASUREMENT SUBJECTS

<u>Secreted Proteins</u>	4835
<i>ELISA</i>	4035
<i>Multiplex ELISA</i>	1286

<u>Virus Titer</u>	3609
<i>Virus Neutralization Titer</i>	2265
<i>HAI Titer</i>	1344

<u>Clinical Lab Tests</u>	2639
<i>Complete Blood Count</i>	1684
<i>Comprehensive Metabolic Panel</i>	664
<i>Fasting Lipid Profile</i>	664

<u>Questionnaire</u>	1422
----------------------	------

<u>Cytometry</u>	1415
<i>Flow Cytometry (PBMC)</i>	907
<i>CytoF (PBMC)</i>	583
<i>Flow Cytometry (Whole Blood)</i>	164

<u>HLA Type</u>	1093
-----------------	------

<u>Gene Expression Array</u>	476
<i>Whole Blood</i>	311
<i>PBMC</i>	165

# 10K Immunomes.org

## Immunoassays

Immunoassays are a useful tool for obtaining precise measurements on the concentration of molecules, bacteria, and viruses inside of a sample. This technique relies on the strong bonding affinity between an antibody and its corresponding antigen. Specific antibodies are chosen to 'grab' the particle of interest. Immunoassays can also be used to measure the relative effectiveness of an antibody on a specific virus of interest. We provide immunoassays from ELISA, Multiplex ELISA, HAI Titer, and Virus Neutralization studies.

- Transcriptomics
- Proteomics
- Immunoassays
- Lab Tests

**Select Data Type:**

Normalized       Outliers

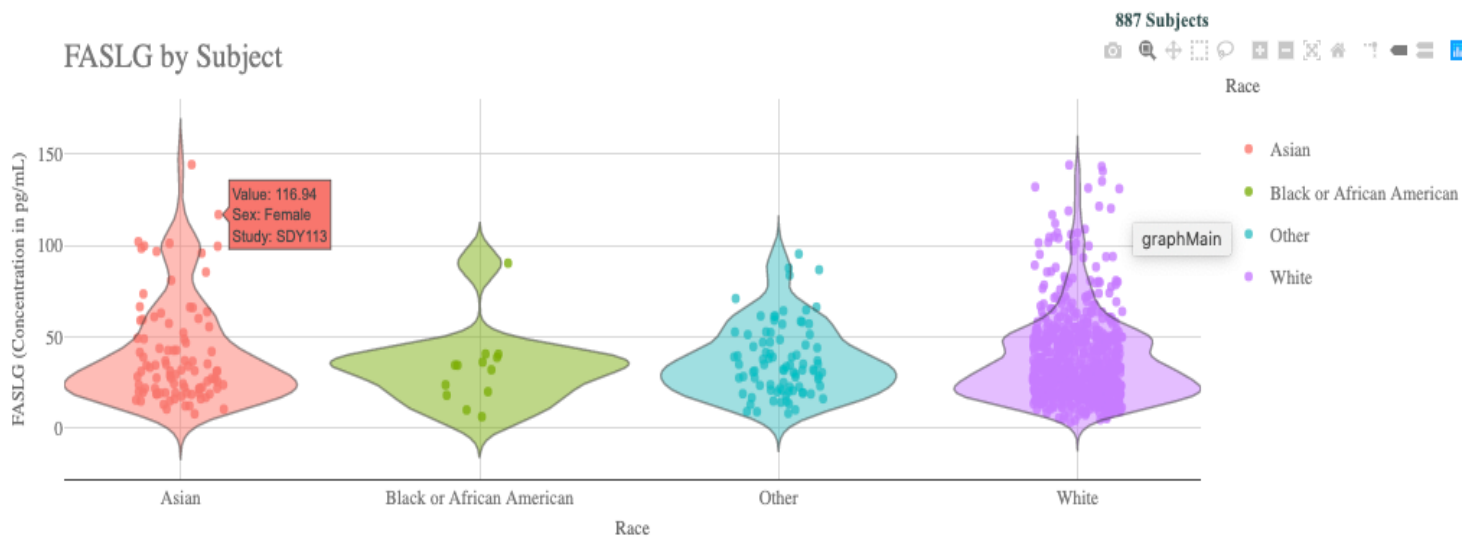
**Select Analyte:**

**Age Range:**

**Ethnicities:**  
 White  
 Black or African American  
 Asian  
 Other

**Plot By:**  
 Age & Sex  
 Ethnicity  
 Study

**Sex:**  
 Female     Male



### Download

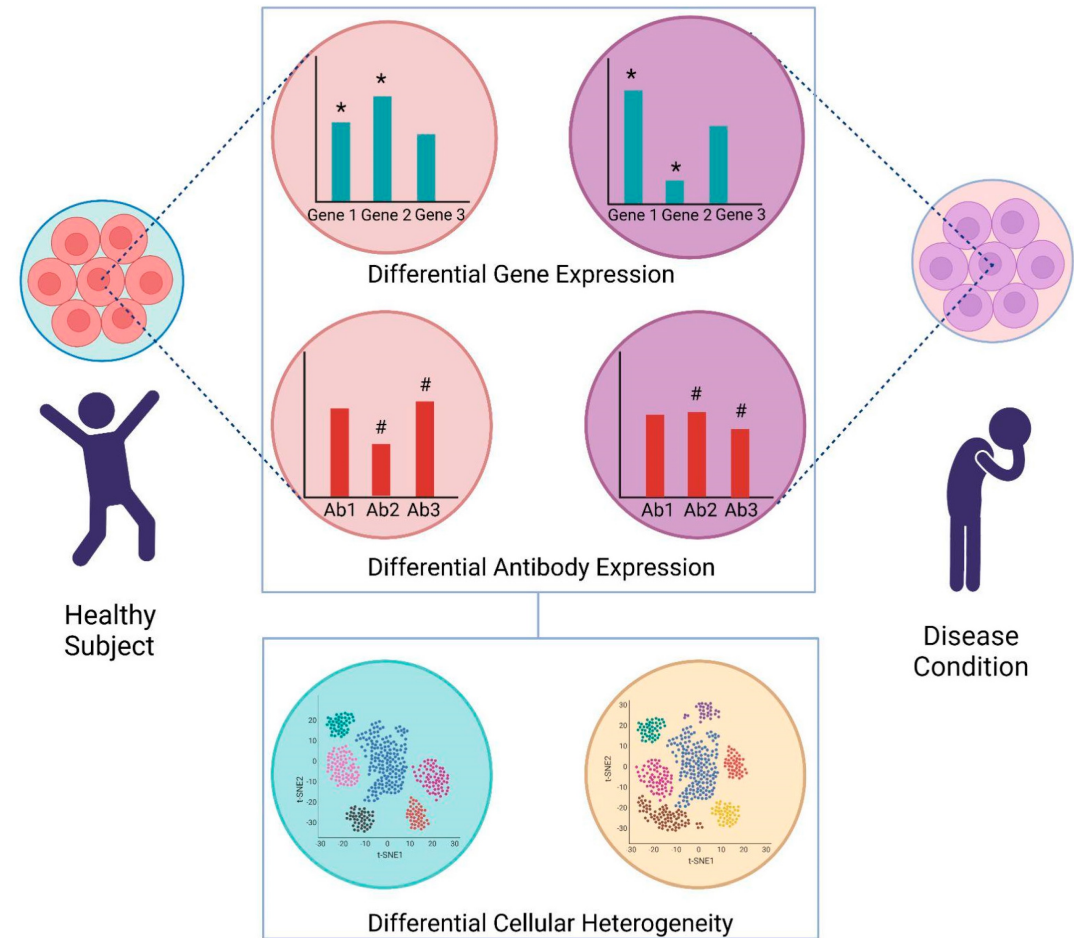
- Image
- Plot Data
- All Data

\* All Data is dataset's raw and formatted files

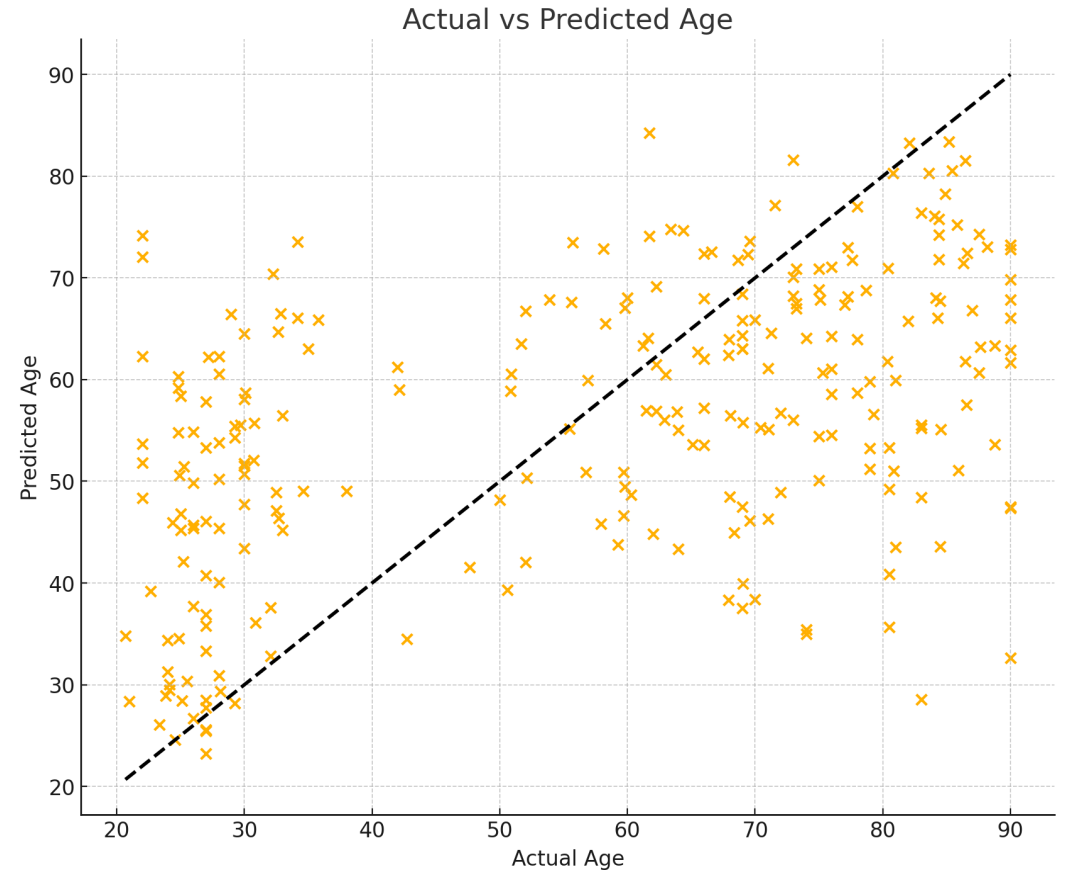
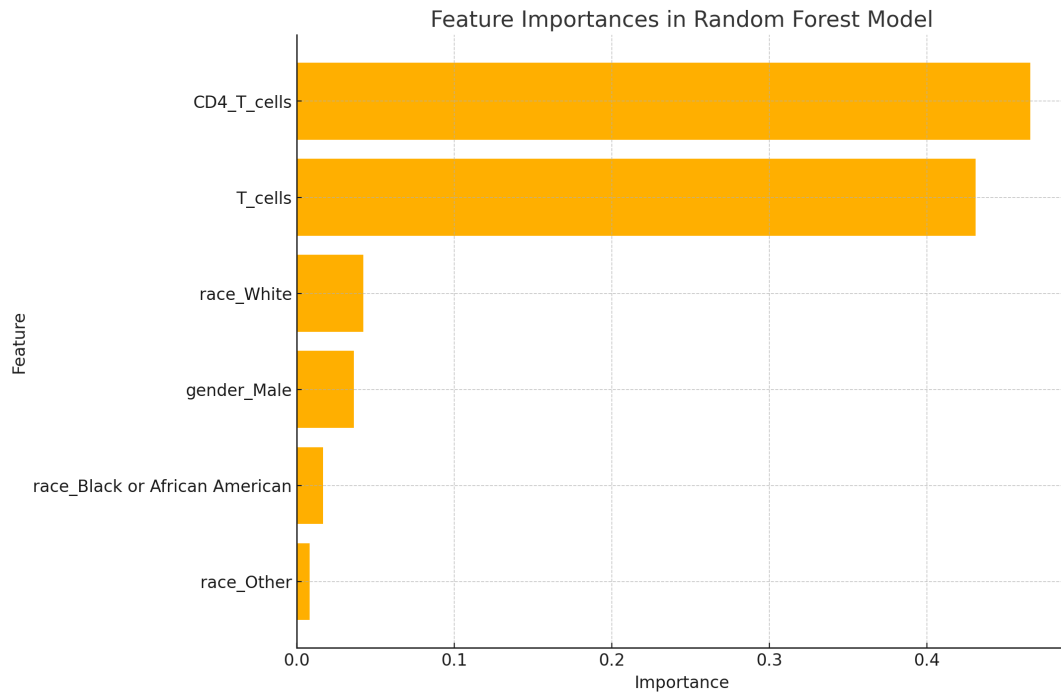
# Use Case: AI-Ready dataset/resource

## 10KImmunes

- Large training dataset to profile healthy subjects
- Features to build the classifiers
- Key variables -Age, Race, Sex



# Clean, structured data such as 10K Immunomes can easily be used in traditional ML models (ex: PBMC Flow Data)



# LLM's like GPT4 Can Explore Flow Cytometry Data...

I am giving you some normalized flow cytometry data from whole blood and from PBMCs downloaded from 10K Immunomes. Please run an initial data exploration: how many samples, how many subjects, distributions of different cell type data available etc. Please create helpful visualizations along the way when appropriate.

## Initial Data Exploration

### Blood Data

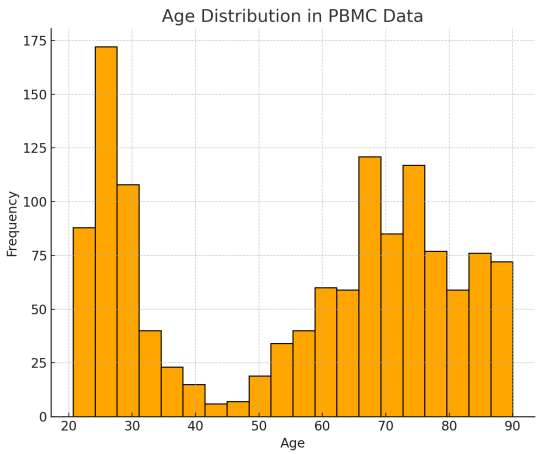
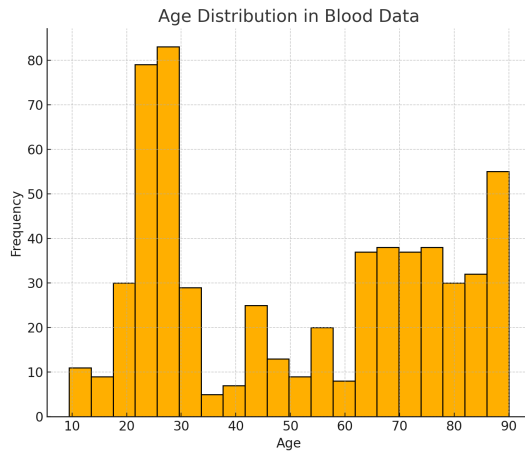
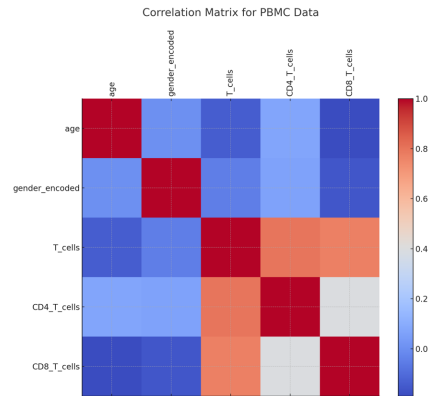
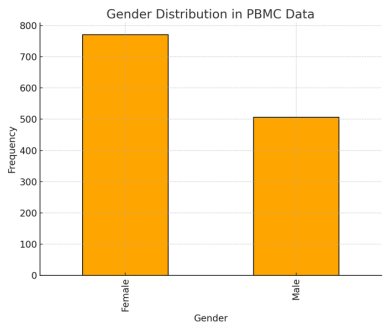
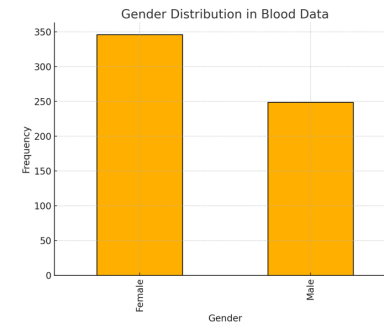
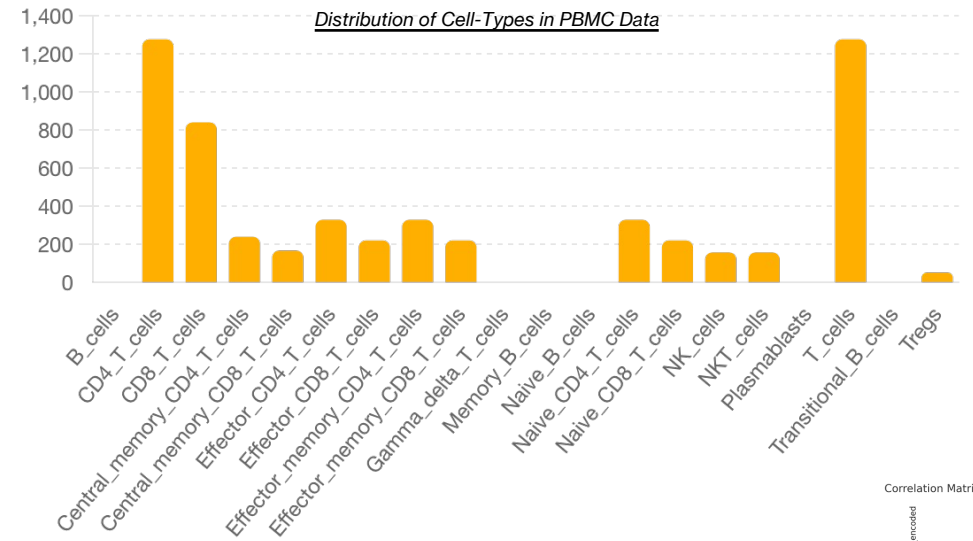
- Number of Samples: 595
- Columns: 7
  - ``subject_accession`, `study_accession`, `age`, `race`, `gender`, `T_cells``

### PBMC Data

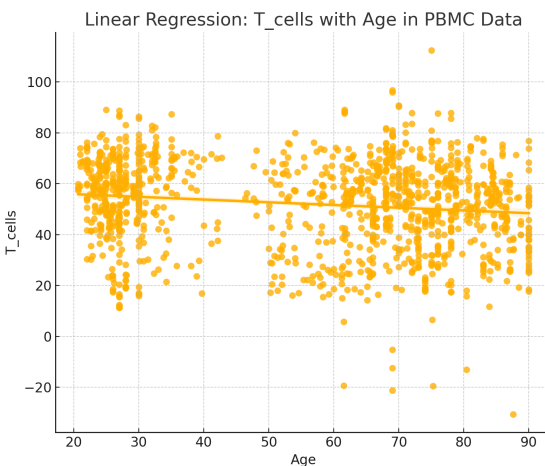
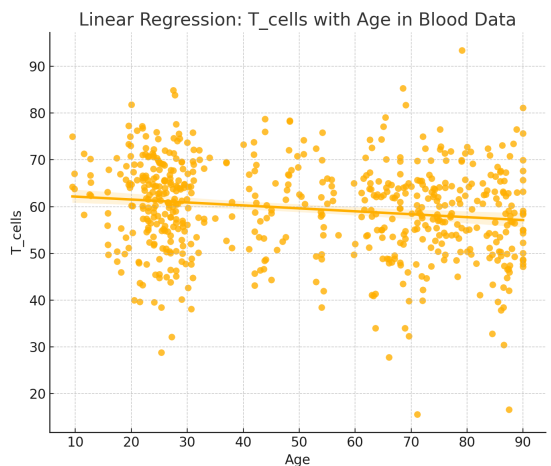
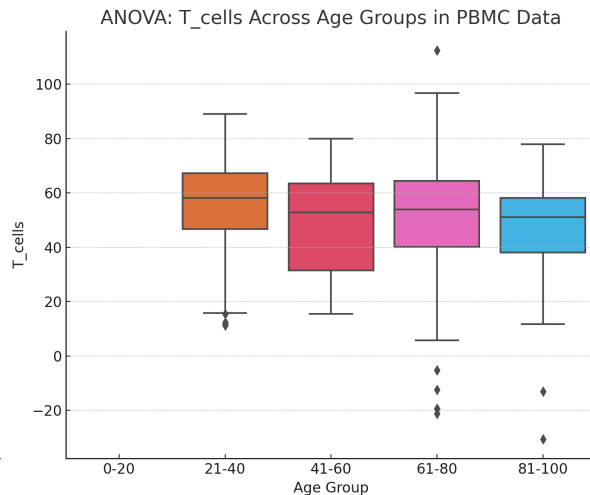
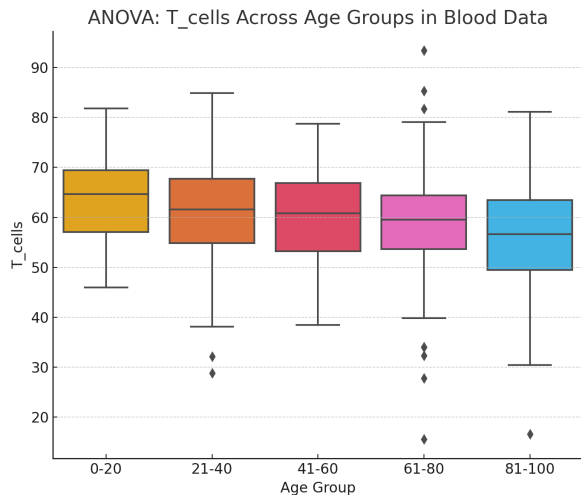
- Number of Samples: 1278
- Columns: 25
  - ``study_accession`, `subject_accession`, `age`, `race`, `gender`, `B_cells`, `CD4_T_cells`, `CD8_T_cells`, `Central_memory_CD4_T_cells`, `Central_memory_CD8_T_cells`, `Effector_CD4_T_cells`, `Effector_CD8_T_cells`, `Effector_memory_CD4_T_cells`, `Effector_memory_CD8_T_cells`, `Gamma_delta_T_cells`, `Memory_B_cells`, `Naive_B_cells`, `Naive_CD4_T_cells`, `Naive_CD8_T_cells`, `NK_cells`, `NKT_cells`, `Plasmablasts`, `T_cells`, `Transitional_B_cells`, `Tregs``

### Observations

1. Blood data has only one type of cell data: ``T_cells``.
2. PBMC data contains a wider range of cell types, but several columns have many missing values.



# And with Clean Structured Data, GPT4 Can Also Run Statistical Analyses



- **Blood Data:**
  - F-value (Gender): 3.890
  - p-value (Gender): 0.049
  - F-value (Age): 15.666
  - p-value (Age): <0.0001
  - **Interpretation:** Gender and age both have significant effects on `T\_cells` in the blood data, with gender being significant at the 5% level.
- **PBMC Data:**
  - F-value (Gender): 3.060
  - p-value (Gender): 0.080
  - F-value (Age): 25.771
  - p-value (Age): <0.0001
  - **Interpretation:** Age has a significant effect on `T\_cells` in the PBMC data, while gender is not significant at the 5% level (though it is close at 8%).

## Linear Regression

- **Blood Data:**
  - R-squared: 0.030
  - **Significant predictors:**
    - Age (p-value: 0.0001)
    - Gender (p-value: 0.049)
  - **Interpretation:** Both age and gender are significant predictors of `T\_cells` in the blood data, but they only explain a small portion of the variance (3%).

*With data clearly labelled, documented, and prepared for public use, novel AI models can easily ingest, clean, and analyze datasets--shown here with just one prompt in GPT4!*

# Investigate Code GPT4 is Running

Analysis

x

```
python Always show details  Copy code

# ANOVA for CD4_T_cells across age groups in PBMC Data
anova_cd4 = ols('CD4_T_cells ~ C(age_group)', data=pbmc_d
anova_table_cd4 = sm.stats.anova_lm(anova_cd4, typ=2)

# ANCOVA for CD4_T_cells between genders while controllin
ancova_cd4 = ols('CD4_T_cells ~ C(gender) + age', data=pb
ancova_table_cd4 = sm.stats.anova_lm(ancova_cd4, typ=2)

# Linear Regression for CD4_T_cells with age and gender i
linear_regression_cd4 = ols('CD4_T_cells ~ age + C(gender
linear_regression_summary_cd4 = linear_regression_cd4.sum

# Visualizations for CD4_T_cells

# ANOVA Visualization
plt.figure(figsize=(12, 6))
sns.boxplot(x='age_group', y='CD4_T_cells', data=pbmc_dat
plt.title('ANOVA: CD4_T_cells Across Age Groups in PBMC D
plt.xlabel('Age Group')
plt.ylabel('CD4_T_cells')
plt.show()

# ANCOVA Visualization
plt.figure(figsize=(12, 6))
sns.scatterplot(x='age', y='CD4 T cells', hue='gender', d
```



Sample Code for Hyperparameter Tuning and Model Training:

```
python Copy code

from sklearn.model_selection import RandomizedSearchCV
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import numpy as np

# Defining the parameter distribution for Randomized Search
param_distributions = {
    'n_estimators': [100, 200, 300],
    'max_features': ['auto', 'sqrt', 'log2'],
    'max_depth': [10, 20, 30, None],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'bootstrap': [True, False]
}

# Setting up the Randomized Search
rf_random_search = RandomizedSearchCV(estimator=RandomForestRegressor(random_sta
                                n_iter=50, cv=3, n_jobs=-1, verbose=2, sco


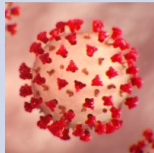
# Fitting the Randomized Search to the data
rf_random_search.fit(X_train_cleaned, y_train_cleaned)

# Best parameters and best score
best_params_random = rf_random_search.best_params_
best_score_random = rf_random_search.best_score_

# Fitting the model with the best hyperparameters
rf_model_best = RandomForestRegressor(**best_params_random, random_state=42)
```



## ImmPort Provided AI-Ready Datasets:

Dataset	Description	AI-Readiness Assessment
<p>10k Immunomes <a href="#">↗</a></p> <p><a href="#">Download (138.2 MB)</a></p> <p>MD5 Checksum <a href="#">📄</a></p> <p>303639c62f72196f65880f31f1819fa2</p>  <p>DOI: 10.21430/RV23-WR13</p>	<p>The 10,000 Immunomes Project is a reference dataset for human immunology, derived from over 10,000 control subjects in the NIAID ImmPort Database. The dataset provides comprehensive profiles of the human immune system within healthy cohorts, representing various races.</p> <p><a href="#">Read the publication in Cell Reports</a></p> <p><a href="#">Browse ImmPort Studies in Shared Data</a></p> <p>Review the source code on <a href="#">GitHub</a> or <a href="#">Docker</a></p>	<p>10k Immunomes Assessment <a href="#">📄</a></p>
<p>COVID-19 Compendium (Coming Soon)</p> <p><a href="#">Download</a></p> 	<p>The COVID-19 compendium is a harmonized collection of curated COVID-19 studies, with a large number being part of NCI's SeroNet program. The dataset is compiled to enable meta-analysis of COVID-19 data by SeroNet and other communities.</p>	<p>COVID-19 Compendium Assessment (Coming Soon)</p>

1. Data use agreement pop-up when user clicks download data button
2. Create and publish DOI for AI-ready datasets
  - DOI info will be consistent with other shared ImmPort studies

**User Agreement for the NIAID Immunology Database and Analysis Portal (ImmPort)<sup>1</sup>**

ImmPort is a data sharing and data analysis portal for immunology research community funded by the National Institute of Allergy and Infectious Diseases (NIAID), Division of Allergy, Immunology, and Transplantation (DAIT). You will be asked to accept the terms and conditions of this agreement without exception when you log in to ImmPort.

**1.1 You will not attempt to identify individuals from ImmPort data sets**

As a condition of obtaining access to the ImmPort database you agree to not use the ImmPort data, alone or in combination with other data, to identify any individual or entity or otherwise link information from these data with information in another dataset in a manner that includes the identity of an individual or entity. If you inadvertently discover the identity of any patient, then (a) You agree that you will make no use of this knowledge, (b) that you will notify the NIAID Program Officer (Quan Chen, [quan.chen@nih.gov](mailto:quan.chen@nih.gov)) of the incident, and (c) that you will inform no one else of the discovered identity.

**1.2 Do not share your username and password**

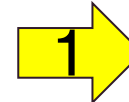
You will use reasonable efforts to maintain the secrecy of the user name issued to you by ImmPort and the password corresponding to the user name. Without limiting the foregoing, you will not share password or user name information with others or allow others to use your password and/or user name.

**1.3 Data Provider Obligations**

If errors with provided data are identified at a later date, the Data Provider agrees to update uploaded data for accuracy.

**Warranties and Liability**

[Accept & Continue](#)



# In Progress: COVID-19 AI-Ready Data



**200+ Studies  
Captured in  
Workbook**

Remove studies  
with low data  
readiness scores

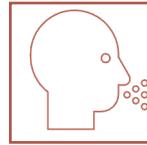
Investigate  
remaining studies



**115 Studies**

Infection and  
Immune response  
studies captured

Studies with  
similar assays and  
conditions



**25 Studies**

All specified as  
infection/immune  
response studies

- ELISA, flow, or  
neutralization  
assay data  
captured

**Total N:**

- 38,573 samples
- 2,377 subjects

# Conclusions: Benefits of ImmPort and Utilization of AI

- 1. Visualization and Interpretation:** AI-powered tools can enhance the visualization and interpretation of immunological data, helping researchers to identify trends and patterns that might be missed with traditional analysis methods
- 2. Predictive Modeling:** AI can analyze ImmPort data to predict disease outcomes, treatment responses, and patient stratification. For example, machine learning models can be trained on immunological assay data to forecast the progression of autoimmune diseases.
- 3. Automated Data Processing:** AI can automate the processing and analysis of large-scale immunological data, making it easier for researchers to handle and interpret complex datasets
- 4. Data Integration and Analysis:** ImmPort's multi-modal data, including ELISA, flow cytometry, and RNA sequencing, can be integrated and analyzed using AI to uncover new insights into immune system functioning and disease mechanisms.
- 5. Drug Discovery:** By leveraging AI to analyze the vast datasets available on ImmPort, researchers can identify potential drug targets and biomarkers. This can accelerate the development of new therapies for various diseases.
- 6. Personalized Medicine:** AI can help in tailoring treatments to individual patients by analyzing their immunological profiles. This can lead to more effective and personalized healthcare solutions.

# ImmPort Team



## UCSF

Atul Butte, PI

**Sanchita Bhattacharya, Co-I**

**Immune System Sciences**

Steven H. Kleinstein, Yale  
University

## NIAID

Anupama Gururaj

Quan Chen

Dawei Lin

## ICF

Srinivas Chepuri

Karen Ketchum

Matthew Strub

Olivier Toujas-Bernate

Alicia Williamson

## Funding Support

National Institute of Allergy and Infectious  
Diseases (NIAID)

National Institutes of Health (NIH)

Health and Human Services (HHS)

Contract #: HHSN316201200036W

\*This research was funded by the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health, with additional support from the NIH Office of Data Science Strategy (ODSS) and the Office of Data Science and Emerging Technologies (ODSET). We appreciate the collaboration with the National Artificial Intelligence Research Resource (NAIRR) to facilitate sharing open data to accelerate health discovery. This research is solely the responsibility of the authors, and does not necessarily represent the official views of the NIH.

## Peraton

**Sammi Smith**

Zhiping Gu

Morgan Crafts

Elizabeth Thomson

Emma Afferton

Sanjiv Desai

John Campbell

Kate Hypes

Jaya Kannan

Ruth Monteiro

Zullinel Trilla-Flores

Vilma Thomas

Bryan Walters

Shujia Zhou

National Institute of Allergy and Infectious Diseases (NIAID)  
National Institutes of Health (NIH)  
Health and Human Services (HHS)