

Systems Immunology: an intro to multi-omics data integration and machine learning



Adriana Tomic

atomic-lab.org

aToMIClab



adrianatomic



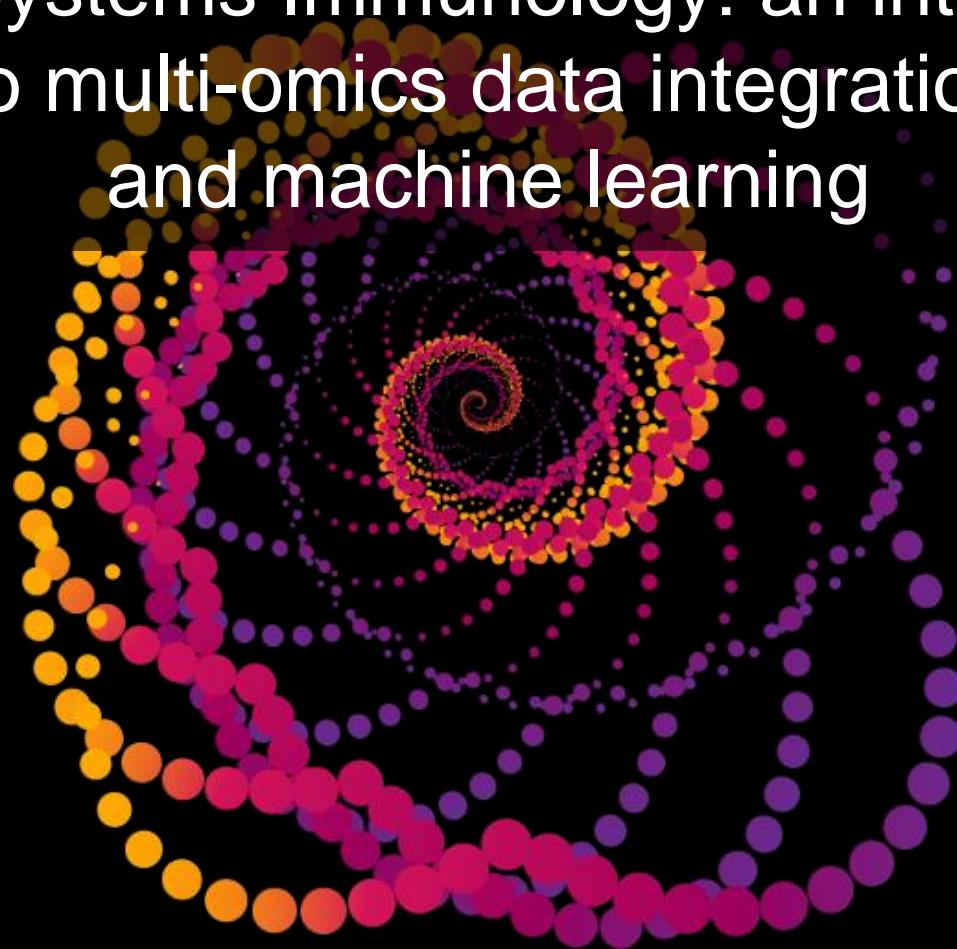
@TomicAdriana



atomic@bu.edu



atomiclaboratory



In this interactive course, we
will learn how to use
machine learning for
biological and biomedical
data integration and
knowledge discovery.

March 6th and 7th 2023, Oxford

Training course - overview

Part I – SIMON, pattern recognition and knowledge extraction platform (March 6th 2023)

- Machine learning and AI – what is all the fuss about?
- What is SIMON?

Theoretical part (9:30-10:30am) ~1h

- • Case study – example 1 (dealing with missing values, overfitting, model performance) **Case study (10:30-11:30am) ~1h**
- Perform SIMON analysis using provided dataset
 - Performance metrics, evaluation and selection of high-quality models **Hands-on (11:30-1:30pm) ~2h**

Part II – Exploratory analysis (March 7th 2023)

- Feature selection: scoring and elimination

Hands-on (9:30-10:30am) ~1h

- Feature processing methods to avoid ‘curse of dimensionality’

Theoretical part (10:30-11am) ~0.5h

- • Case study – example 2 (multi-omics data integration)

Case study (11-11:30am) ~0.5h

- • Case study – example 3 (unsupervised ML + practical demonstration)

Hands-on (11:30am-1:30pm) ~2h

- Discussion about project-specific problems

The art of feature engineering and selection

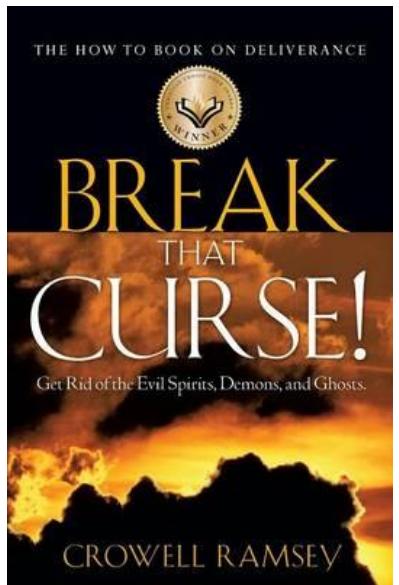


Free book online:
<https://bookdown.org/max/FES/>

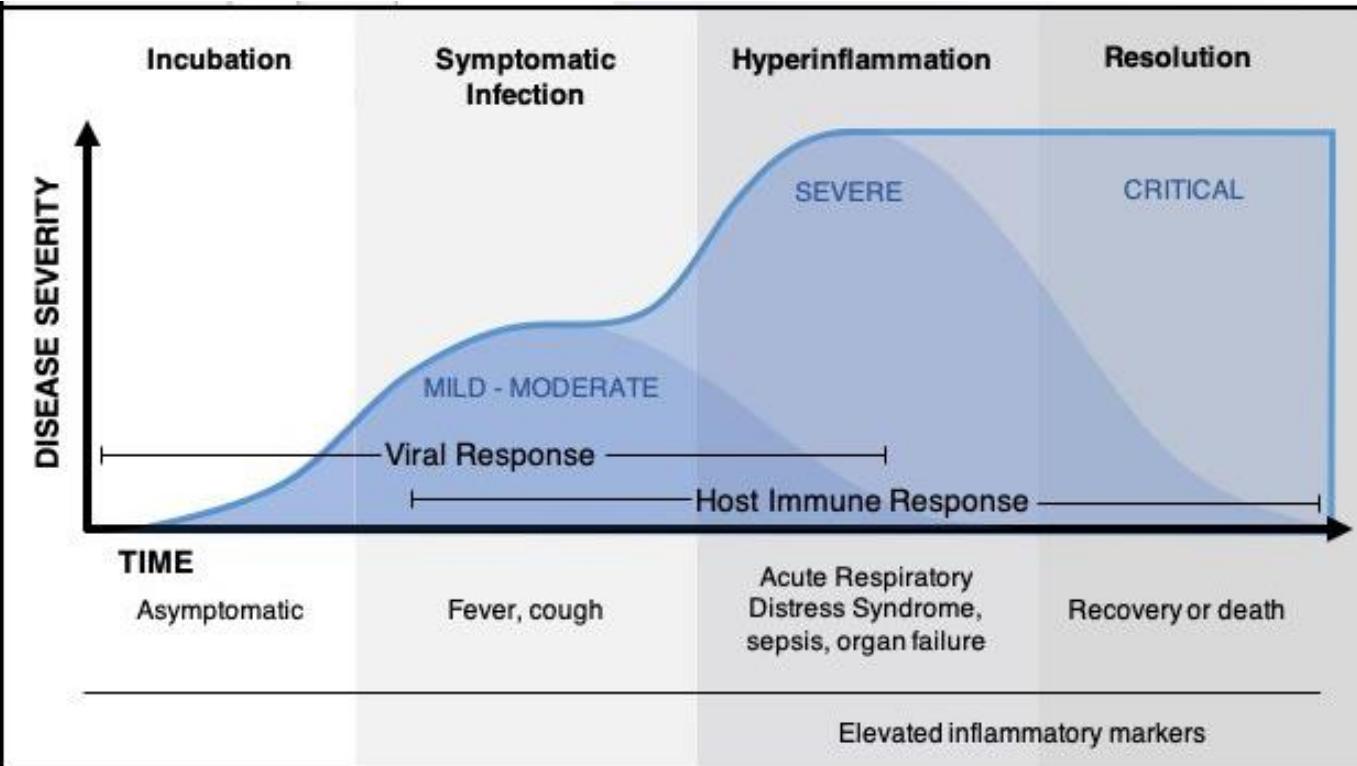
Max Kuhn and Kjell Johnson;
Chapman & Hall/CRC Data
Science Series: 2019

Part I. How to avoid the curse of dimensionality: Multi-omics data integration

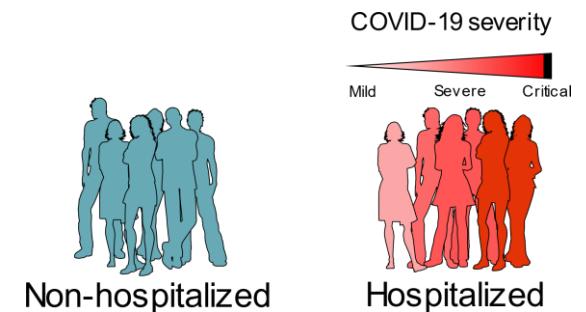
SIMON to the rescue: COMBATing COVID-19



Understanding COVID-19

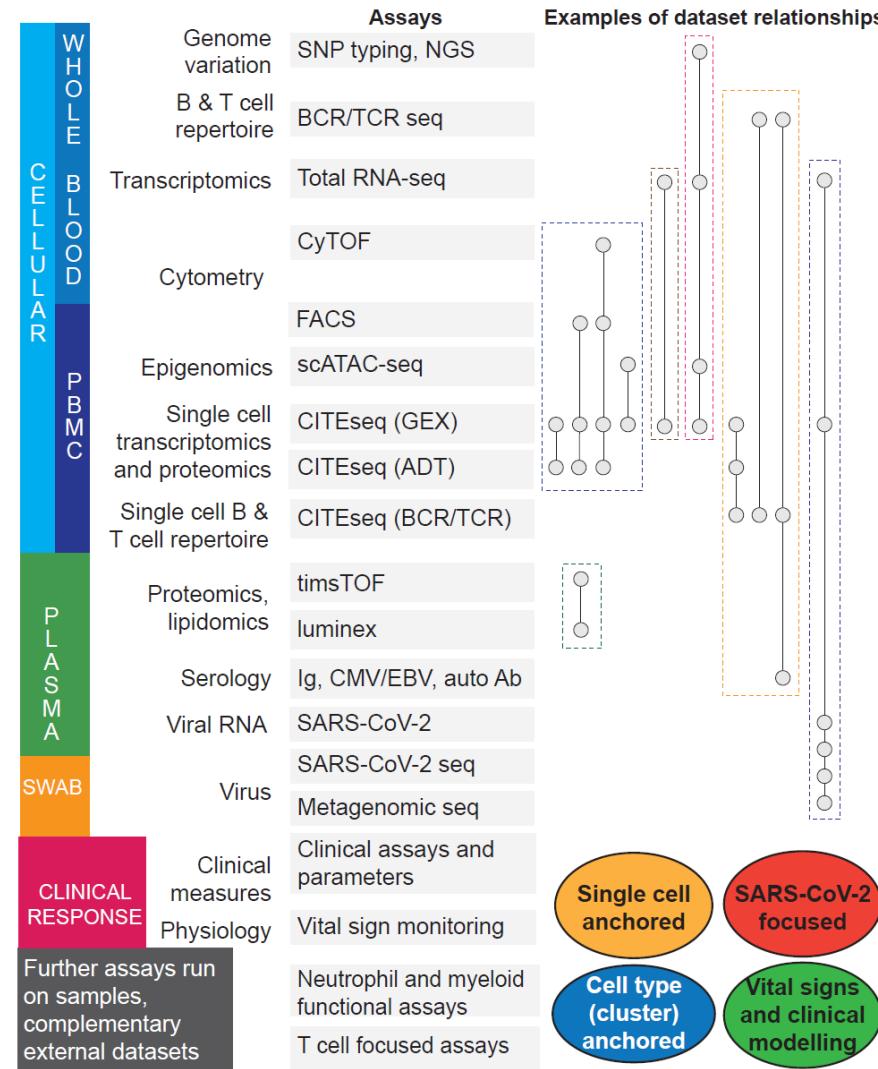


5-10% of patients with COVID-19 progress to severe disease

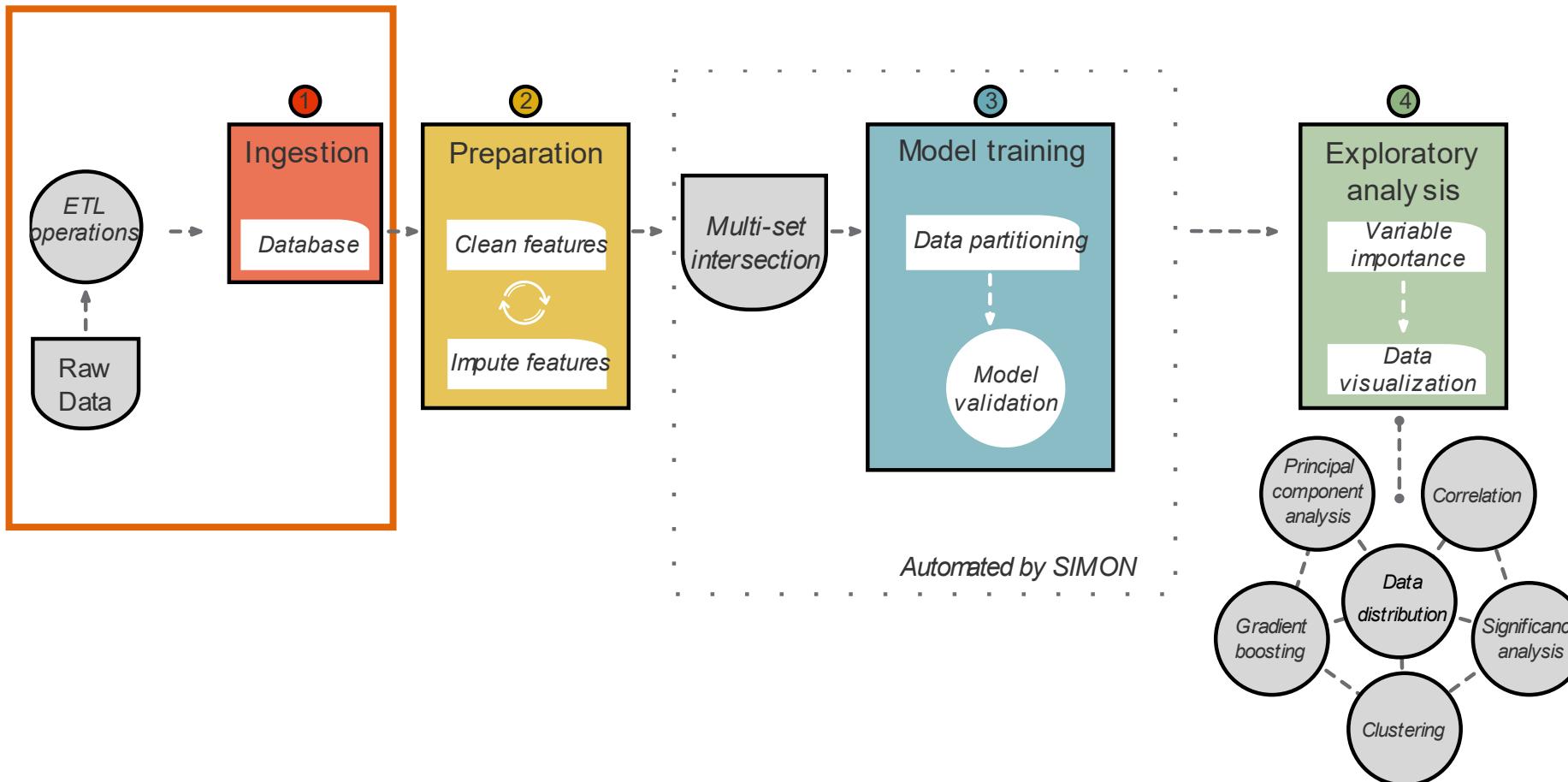


Oberfeld B. et al, SnapShot: COVID-19, Cell, 2020

Immunology research at Oxford aimed to COMBAT COVID-19



SIMON - Knowledge data discovery process



Tomic et al, JI, 2019 & Tomic et al, Patterns, 2021

'Everything in its place'



'Everything in its place'

CyTOF

Cell frequencies and numbers in the granulocyte depleted samples (8 files)

Iron measurement

Two measurements (1 file)

GSA

Chr3p21 COVID19-associated GWAS region and ABO type (2 files)

Mass Spec

Processed intensity matrix (1 file)

FACS

Cell frequencies, numbers and clusters (3 files)

CLINICAL dataset

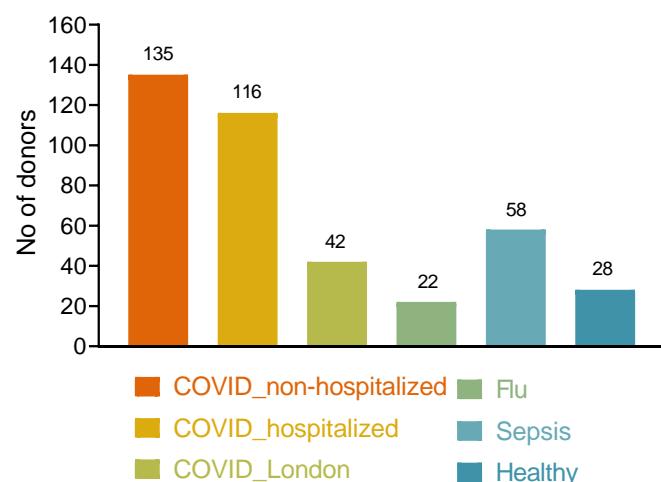
Total number of donors: 401
COVID_non-hospitalized: 135
COVID_hospitalized: 116
COVID_London: 42
Flu: 22
Sepsis: 58
Healthy: 28

RNAseq

Processed gene expression data (log cpm) and PCs
COVID samples (2 files)

Luminex

Concentrations and fluorescence intensities (2 files)



CITEseq

Pseudobulk residuals broad and narrow (2 files)

Data cleaning and generation of new variables

Pre-processing steps for each assay:

- **Data cleaning**

- replace special with alpha-numeric characters (+ → pos)

- data should only be numeric (replace ‘no data’ or ‘nd’ → NA)

- adding prefix/sufix to same parameters with different measurements

- (e.g. freq_cell subsets and Luminex parameter_intens)

- **Generating new features**

- hospitalization (yes or no)

- ventilation status (none or ventilated)

- oxygenation status (normal or abnormal)

- days_sample_taken_from_max_disease (days max disease – days

- sampling)

- Sampling (before or after max disease)

- Disease (recovered – convalescent samples vs ongoing)

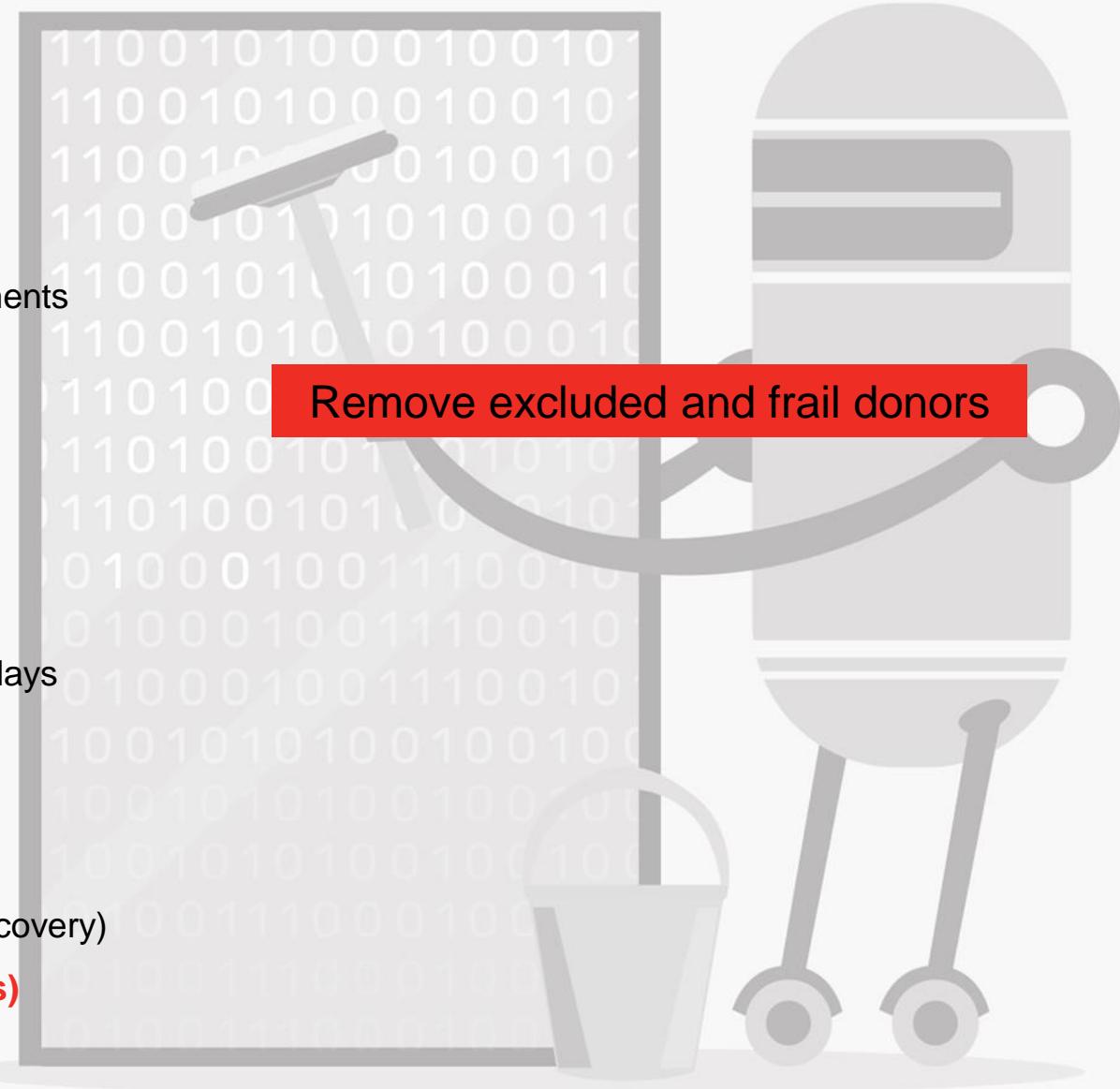
- Disease progress – for longitudinal samples (deterioration or recovery)

Sampling order (first sample taken or second/third samples)

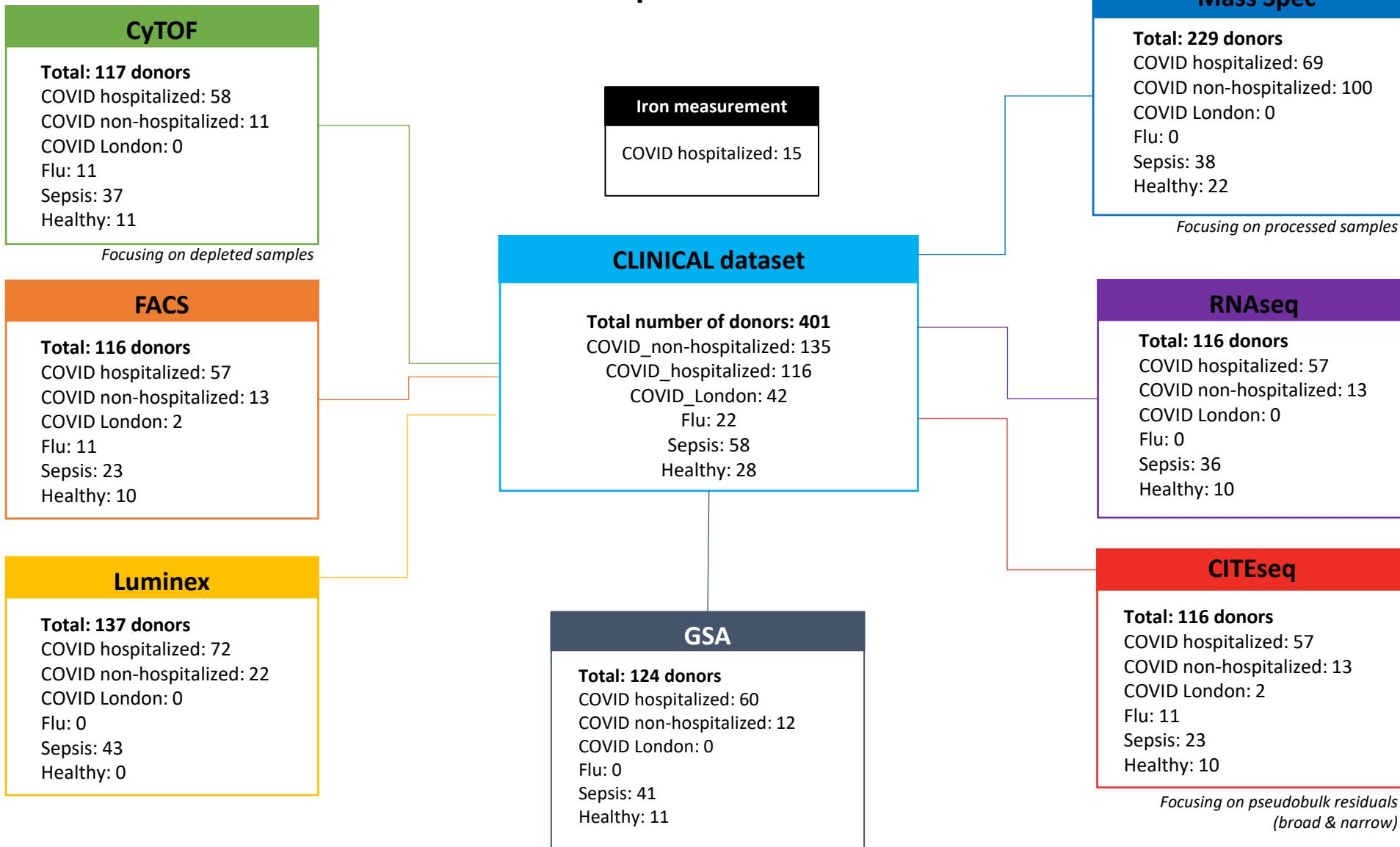
- sampling_from_max_disease

- Sampling after symptom onset (<6d - early and >6d – late)

Written as factors



428 samples → 268 donors



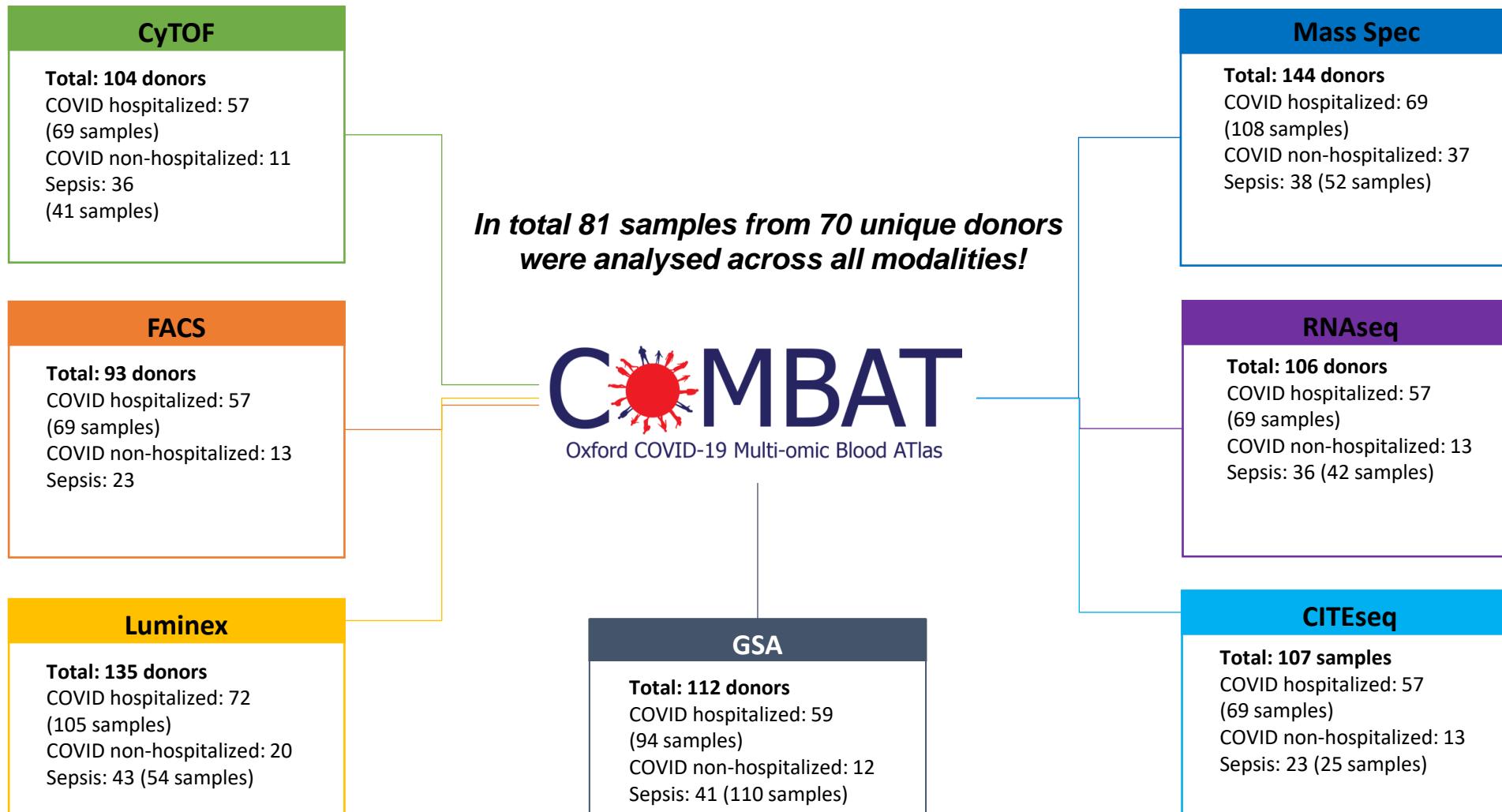
COMBAT dataset: 428 samples → 268 donors



Remove convalescent (covid, sepsis, hcw), Covid_London,
flu and healthy volunteers
281 samples → 167 donors

Which samples were analysed across modalities?

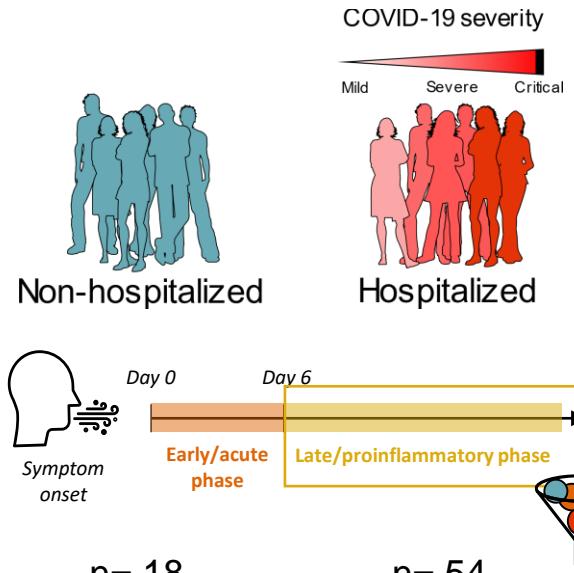
Samples and unique donors analysed across modalities



Knowledge discovery using COMBAT dataset

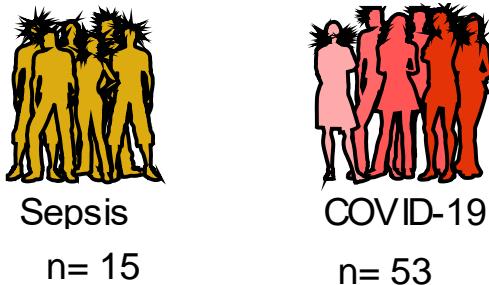
Outcome 1

Difference between Sars-CoV2-infected individuals?



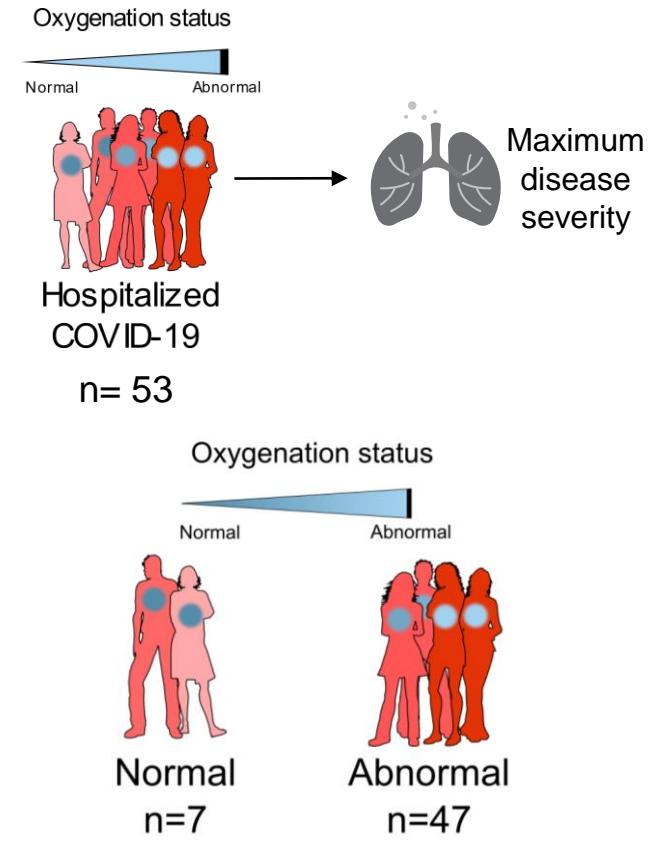
Outcome 2

Difference between hospitalized Sars-CoV2-infected and sepsis patients?

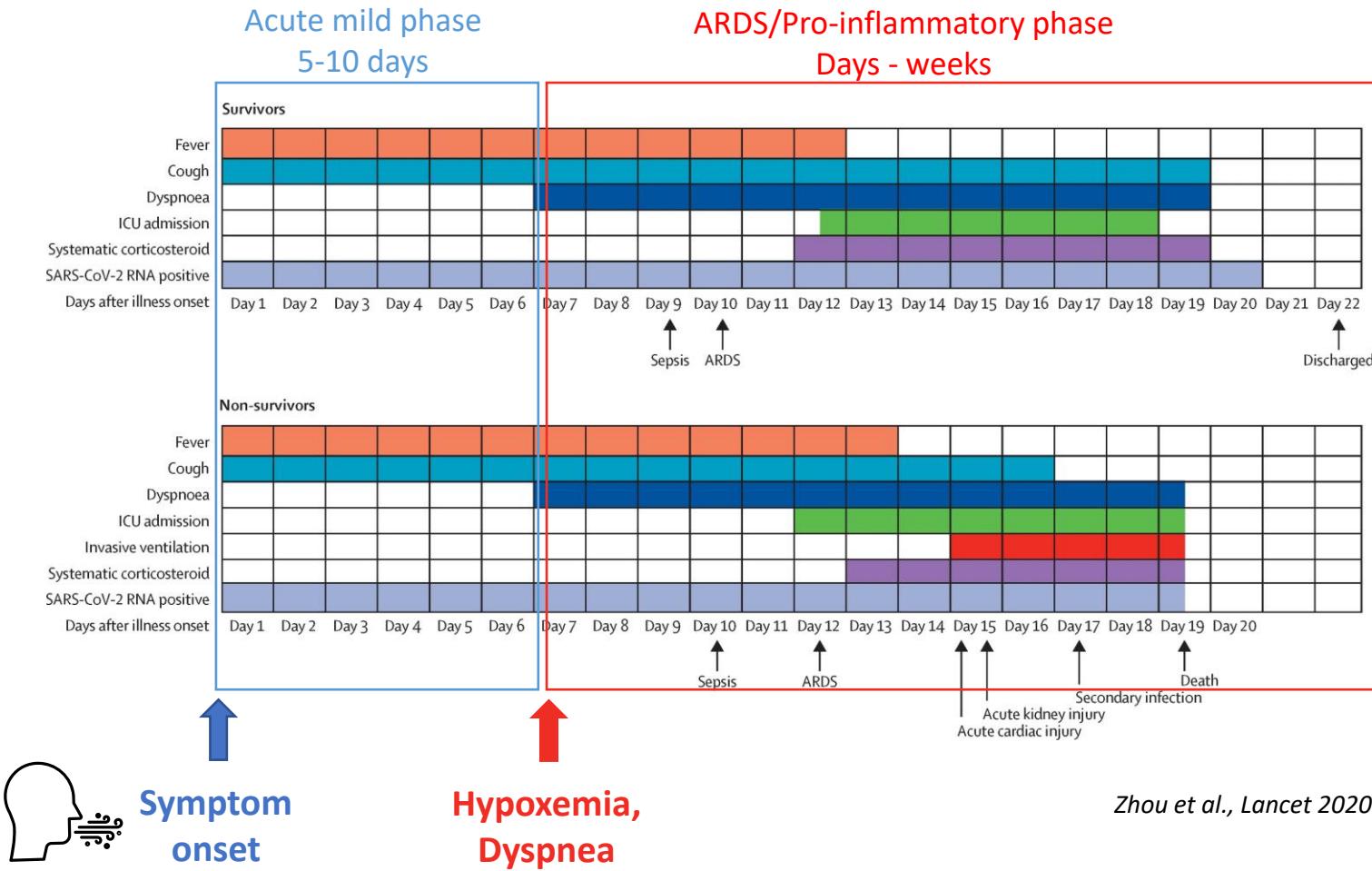


Outcome 3

Difference between hospitalized Sars-CoV2-infected patients?



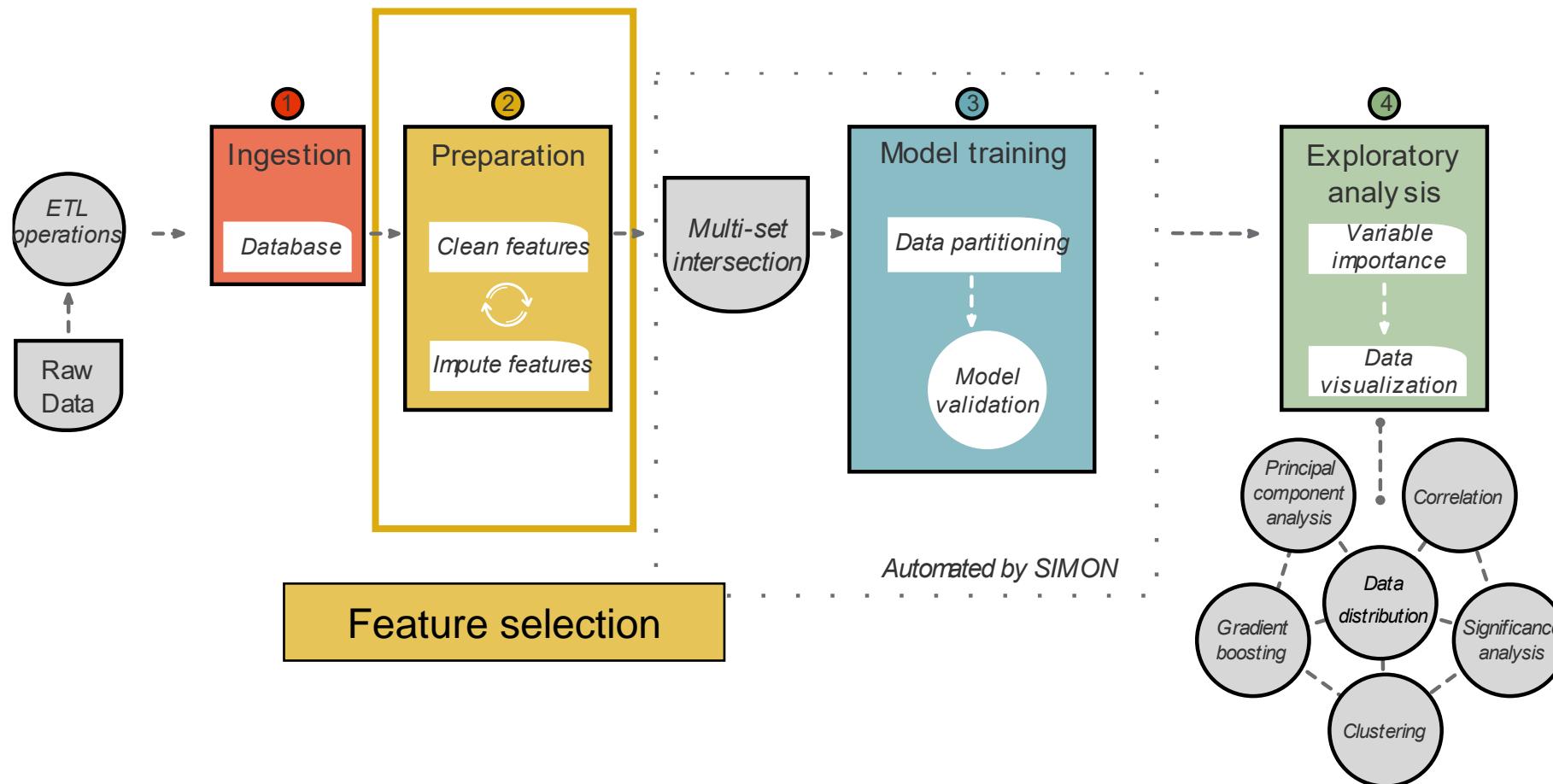
Knowledge discovery using COMBAT dataset



Difference between hospitalized SARS-CoV-2-infected and sepsis patients?

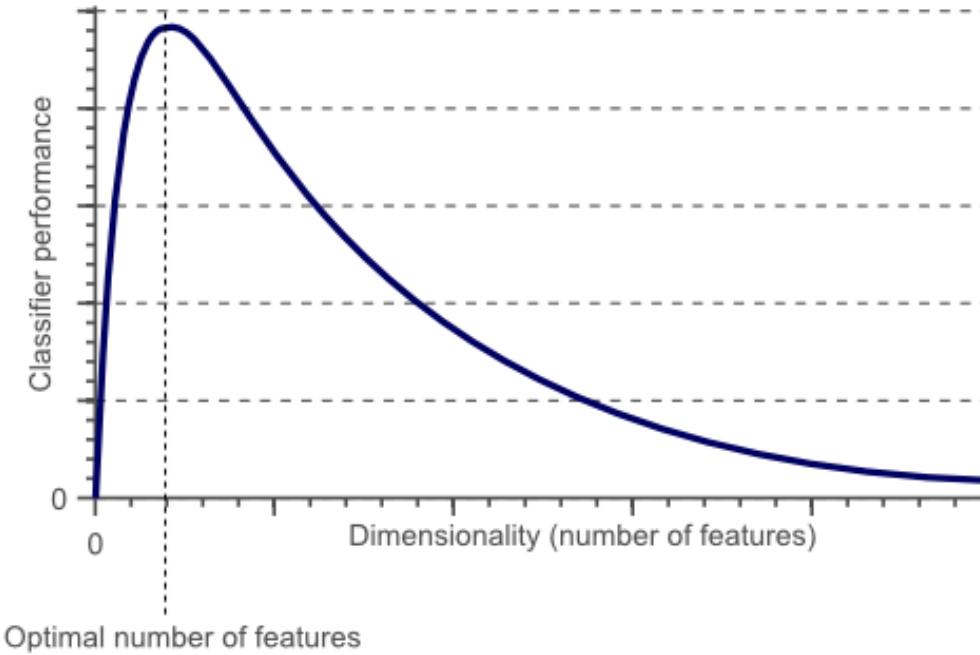
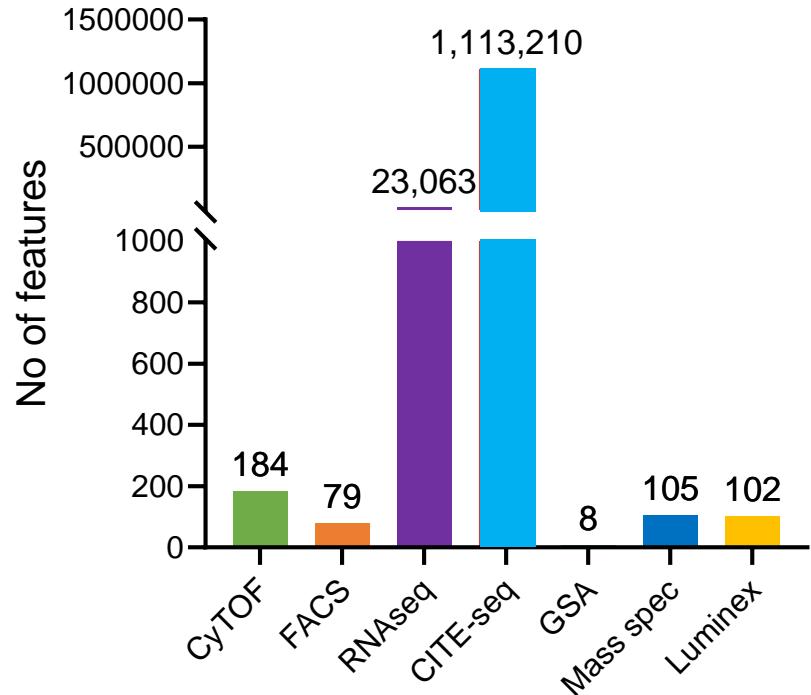


SIMON - Knowledge data discovery process



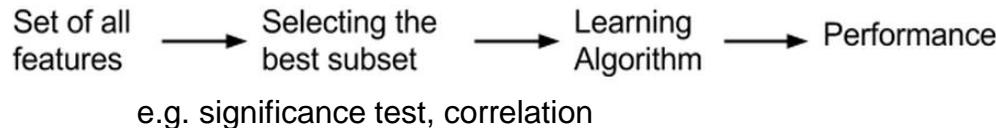
Tomic et al, JI, 2019 & Tomic et al, Patterns, 2021

'Curse of dimensionality'

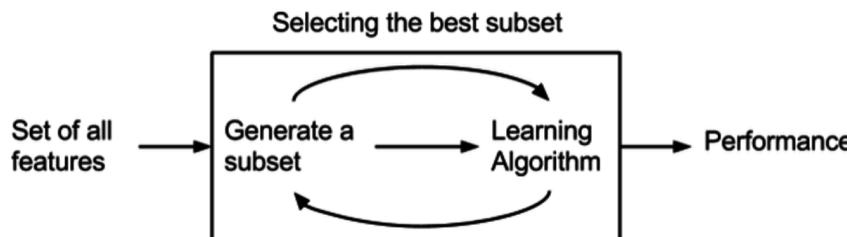


Feature selection

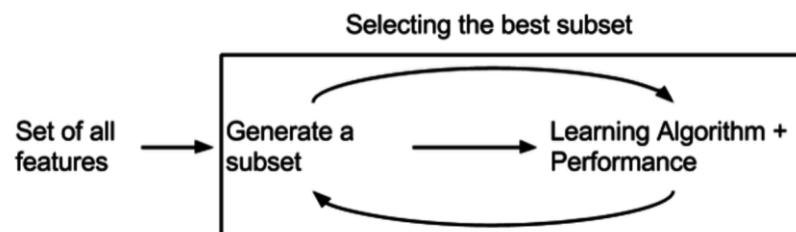
1. *Filter methods* - use a proxy measure to score a feature subset



2. *Wrapper methods* - use a predictive model to score feature subsets



3. *Embedded Methods* - catch-all group of techniques which perform feature selection as part of the model construction process (e.g. LASSO)



SIMON - Feature selection process

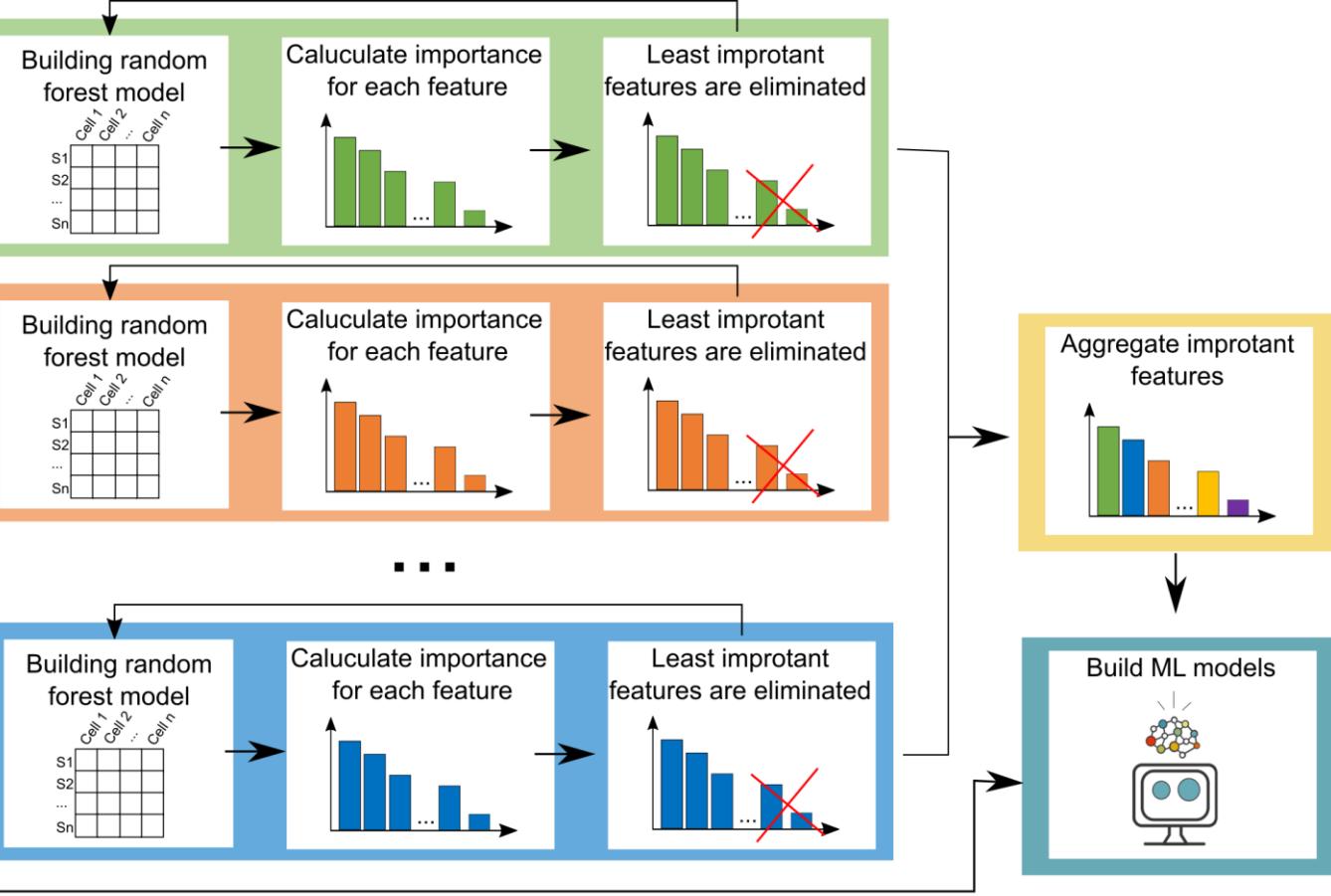
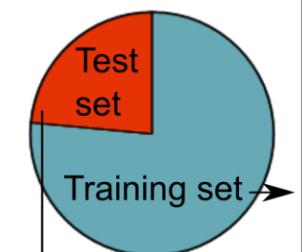
1. Dataset partitioning

2. Feature selection on training set within each assay (recursive feature elimination)

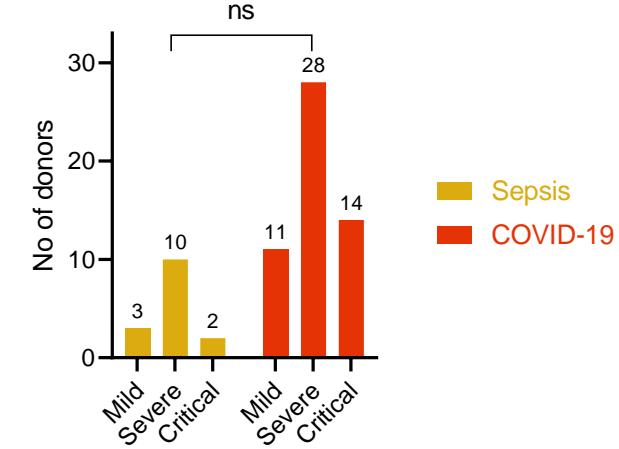
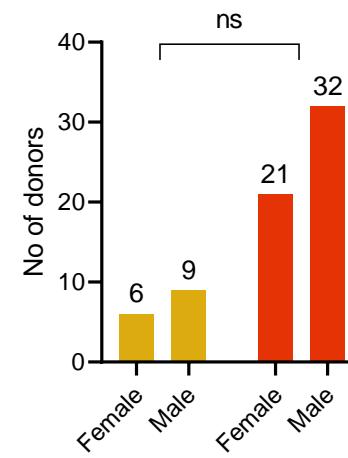
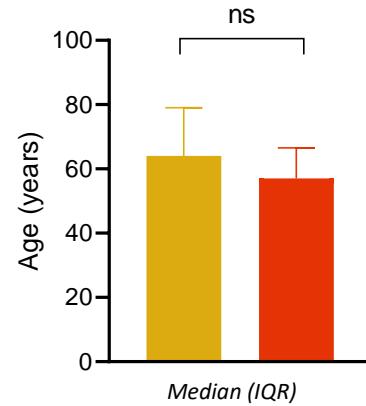
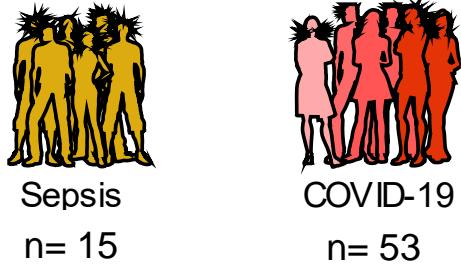
3. Merging selected features

4. Machine learning using SIMON

5. Model performance on test set



Difference between hospitalized SARS-CoV-2-infected and sepsis patients?



Demographic data of hospitalized COVID-19 and sepsis patients

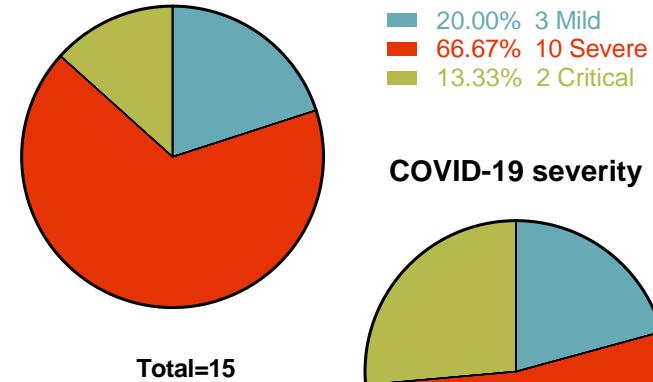
Characteristics	Sepsis (n=15)	COVID-19 (n=53)	P value
Age (y)			0.0969 ^a
Mean ± SD	63.6 ± 17.2	56.6 ± 14.2	
Median (IQR)	64 (52-79)	57 (47.5-66.5)	
Gender			0.3305 ^b
Female (%)	6 (40%)	21 (40%)	
Male (%)	9 (60%)	32 (60%)	
Disease severity			0.1250 ^b
Mild (%)	3 (20%)	11 (21%)	
Severe (%)	10 (67%)	28 (53%)	
Critical (%)	2 (13%)	14 (26%)	

Abbreviation: IQR, interquartile range

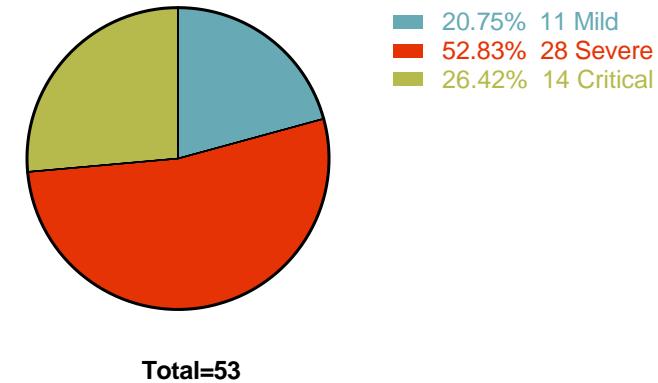
^aTwo-tailed Mann-Whitney test

^bTwo-way ANOVA

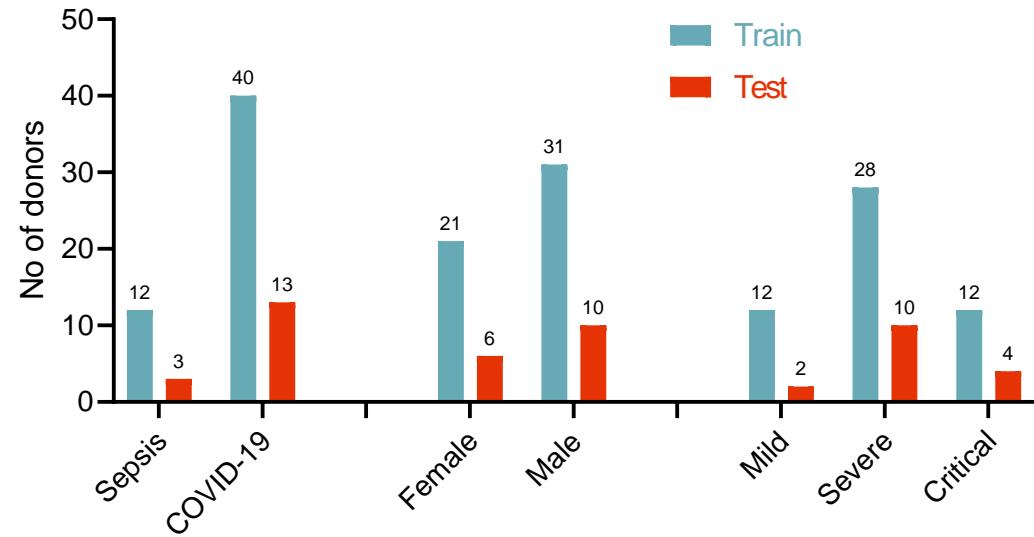
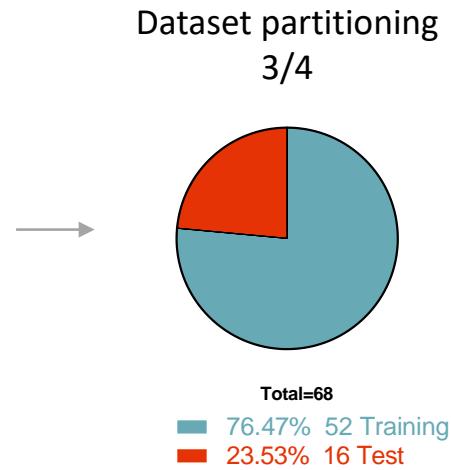
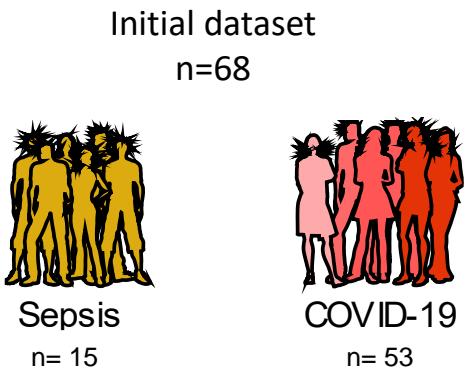
Sepsis severity



COVID-19 severity



Feature selection process



1. Dataset partitioning

2. Feature selection on training set
within each assay

3. Merging selected features

4. Machine learning using SIMON

5. Model performance on test set

Feature selection process

CyTOF	FACS	Luminex	Mass Spec	RNAseq	CITEseq	GSA
184 features (subsets frequencies and absolute numbers)	79 features (subsets frequencies, absolute numbers and clusters)	102 features (expression and fluorescence intensity)	105 features (intensity matrix)	23,063 features (counts per million)	1,112,210 features (20,615 genes in 54 cell populations)	8 features (dosage, ABO typing)

Wrapper method using recursive feature elimination

Step 1. Filter DEGs
Step 2. Wrapper method only on filtered genes

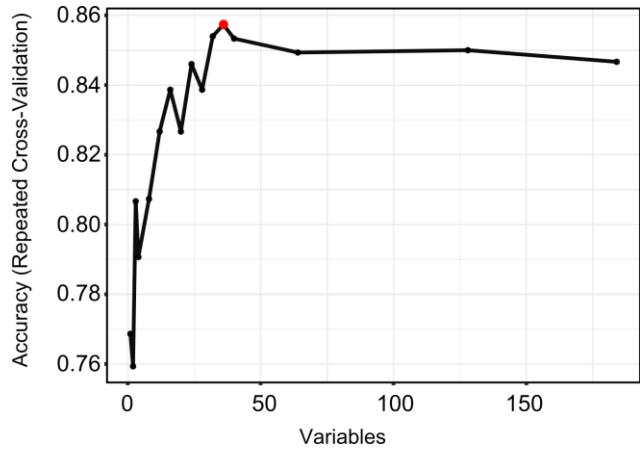
1. Dataset partitioning
2. Feature selection on training set within each assay
3. Merging selected features
4. Machine learning using SIMON
5. Model performance on test set

CyTOF

184 features
(subsets frequencies and absolute numbers)

36

36 features from CyTOF dataset selected by the top performing model

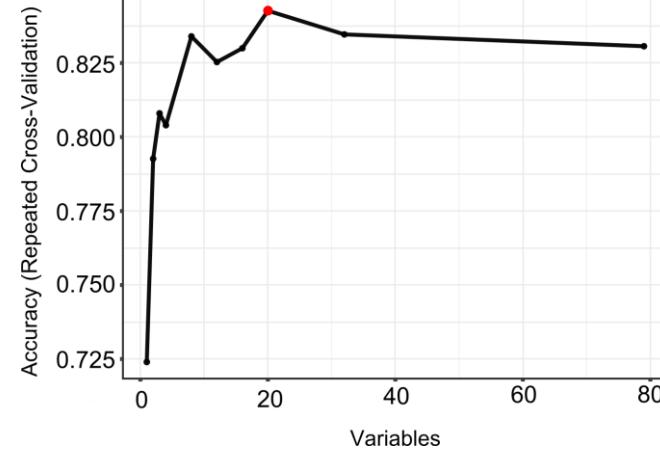


FACS

79 features
(subsets frequencies, absolute numbers and clusters)

20

20 features from FACS dataset selected by the top performing model

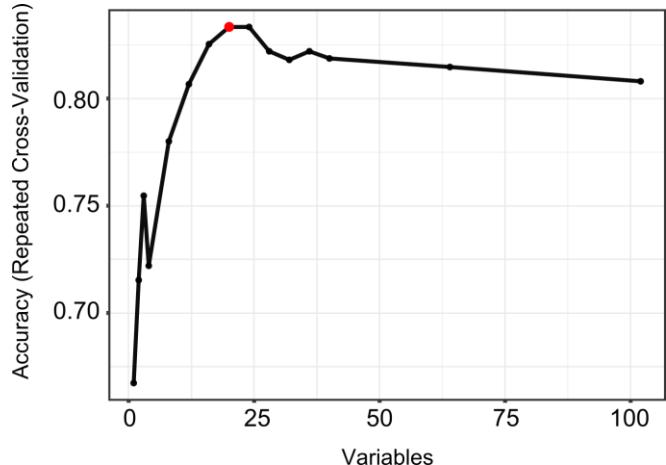


Luminex

102 features
(expression and fluorescence intensity)

20

20 features from Luminex dataset selected by the top performing model

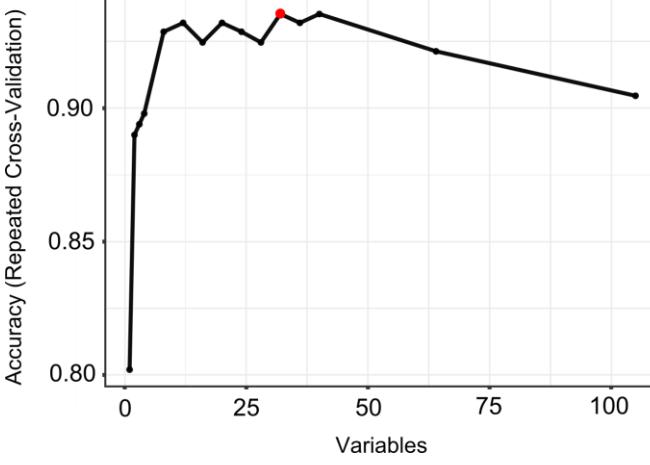


Mass Spec

105 features
(intensity matrix)

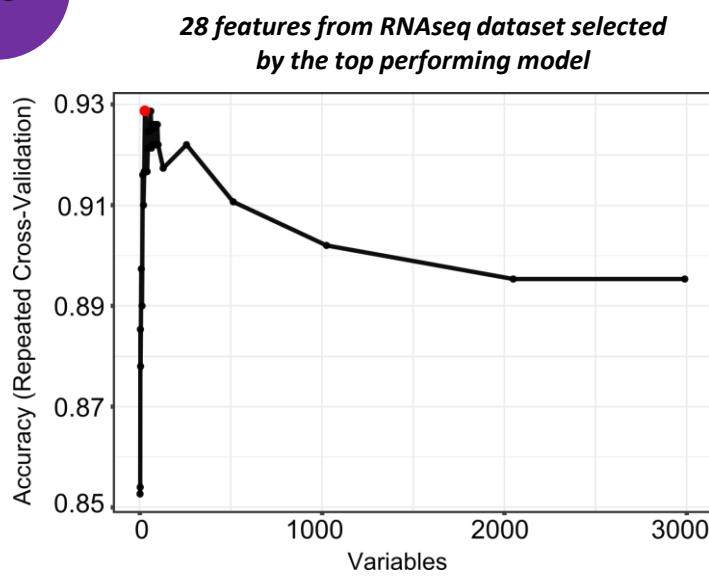
32

32 features from Mass Spec dataset selected by the top performing model



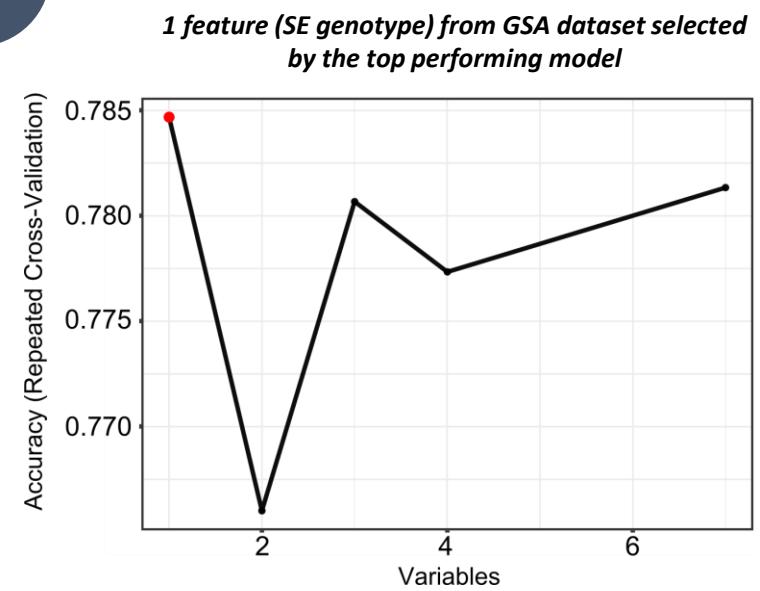
RNAseq
23,063 features (counts per million)
2,989 features (DEGs, FDR<0.05 and FC>1.5)

28

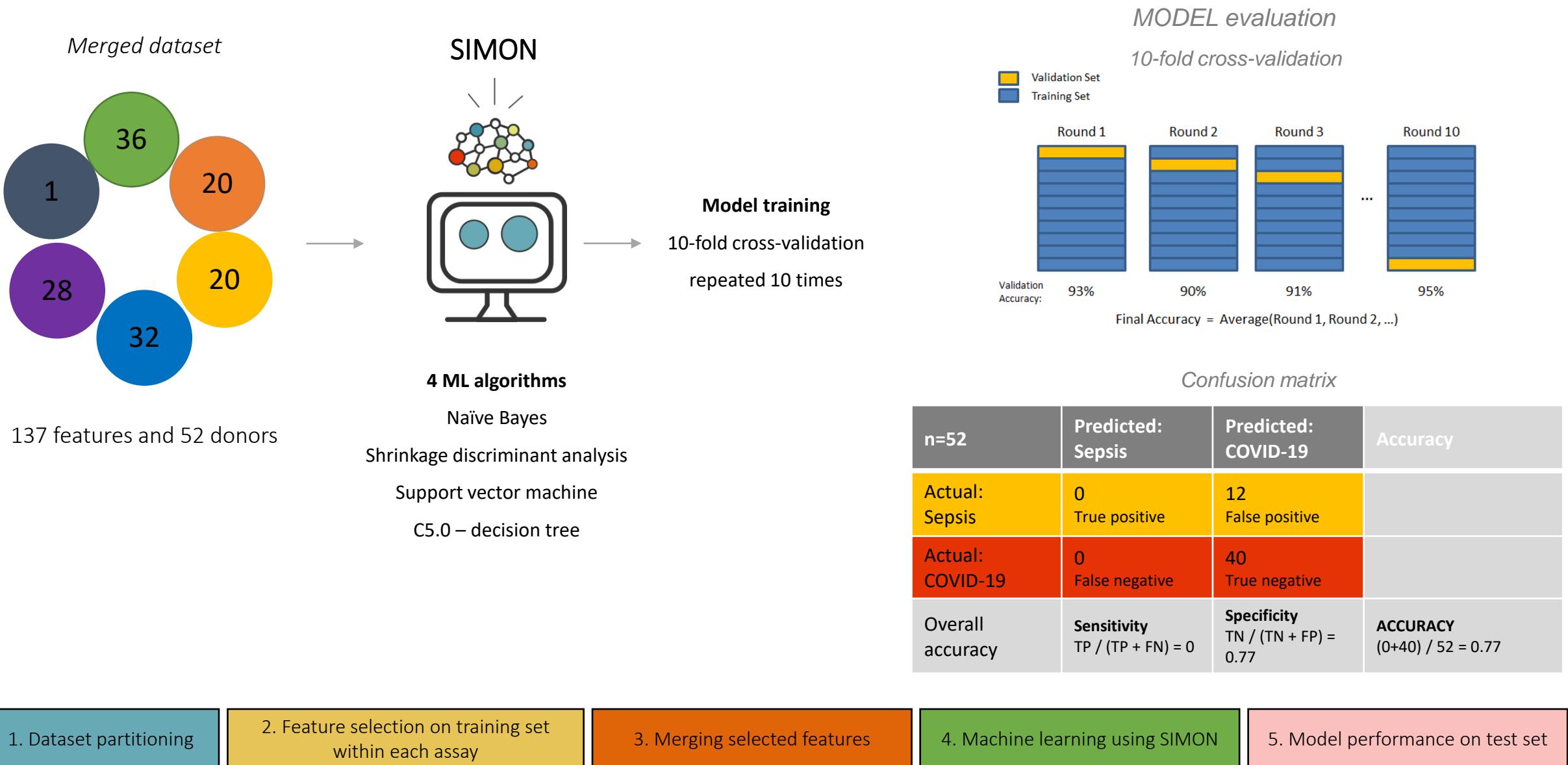


GSA
8 features (dosage, ABO typing)

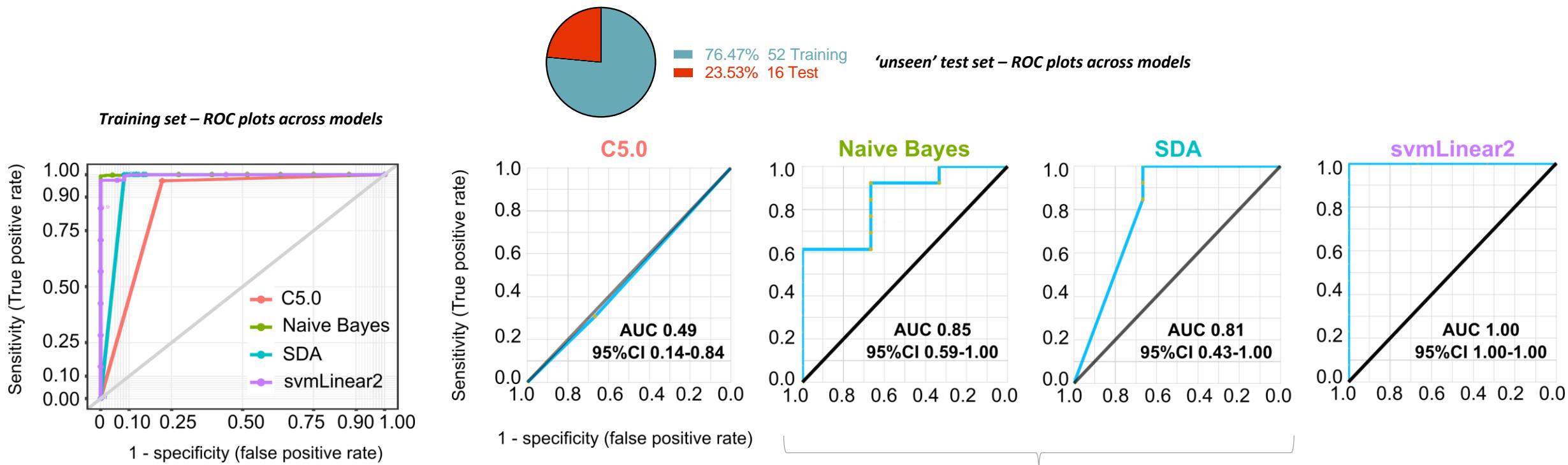
1



Feature selection process – merging selected features



Feature selection process – final model



RANDOM

GOOD DISCRIMINATIVE ABILITY

OVERFITTED

Specificity = 0
Sensitivity = 1

*model can only discriminate COVID-19 patients and always makes mistake on sepsis patients

1. Dataset partitioning

2. Feature selection on training set within each assay

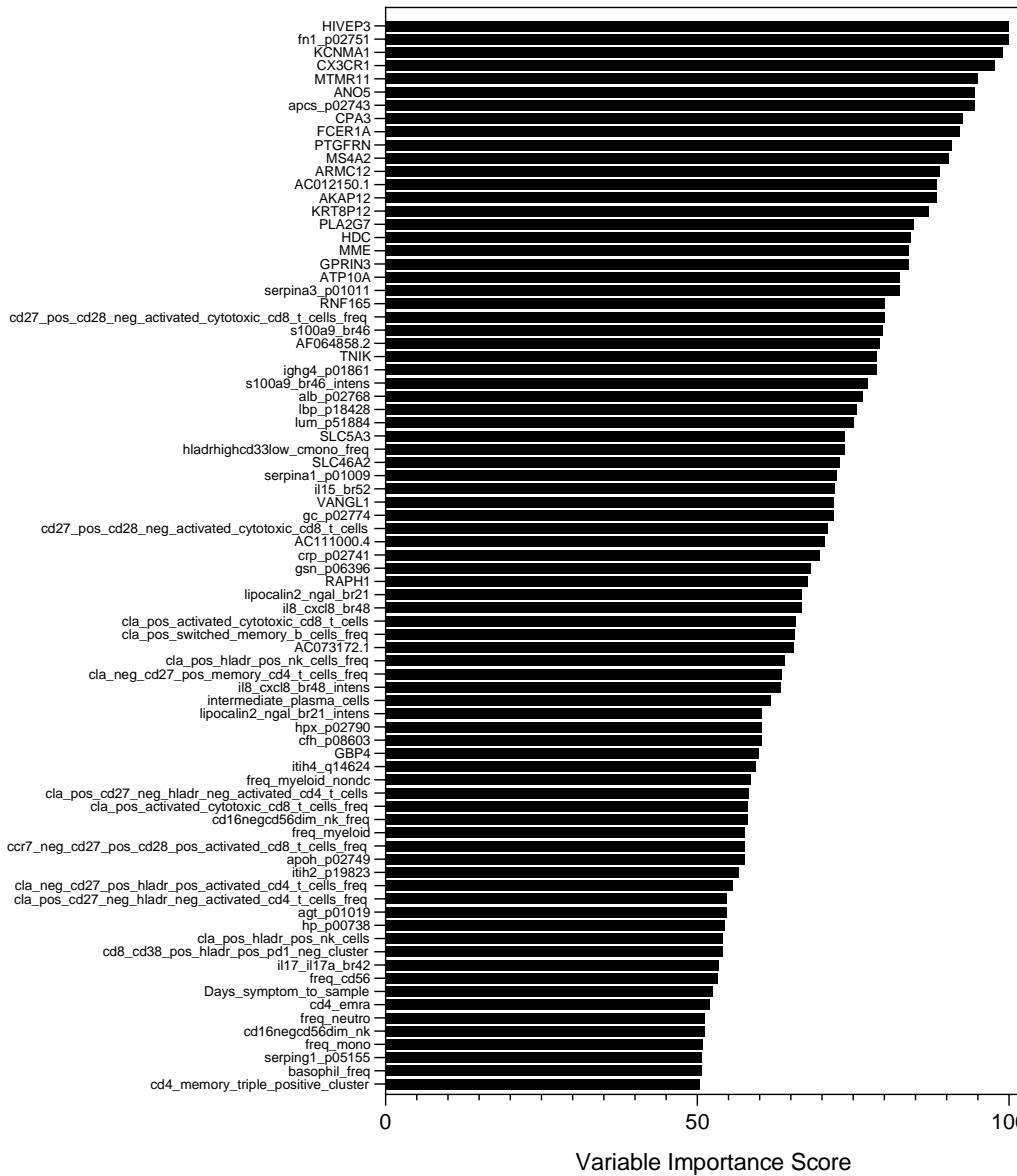
3. Merging selected features

4. Machine learning using SIMON

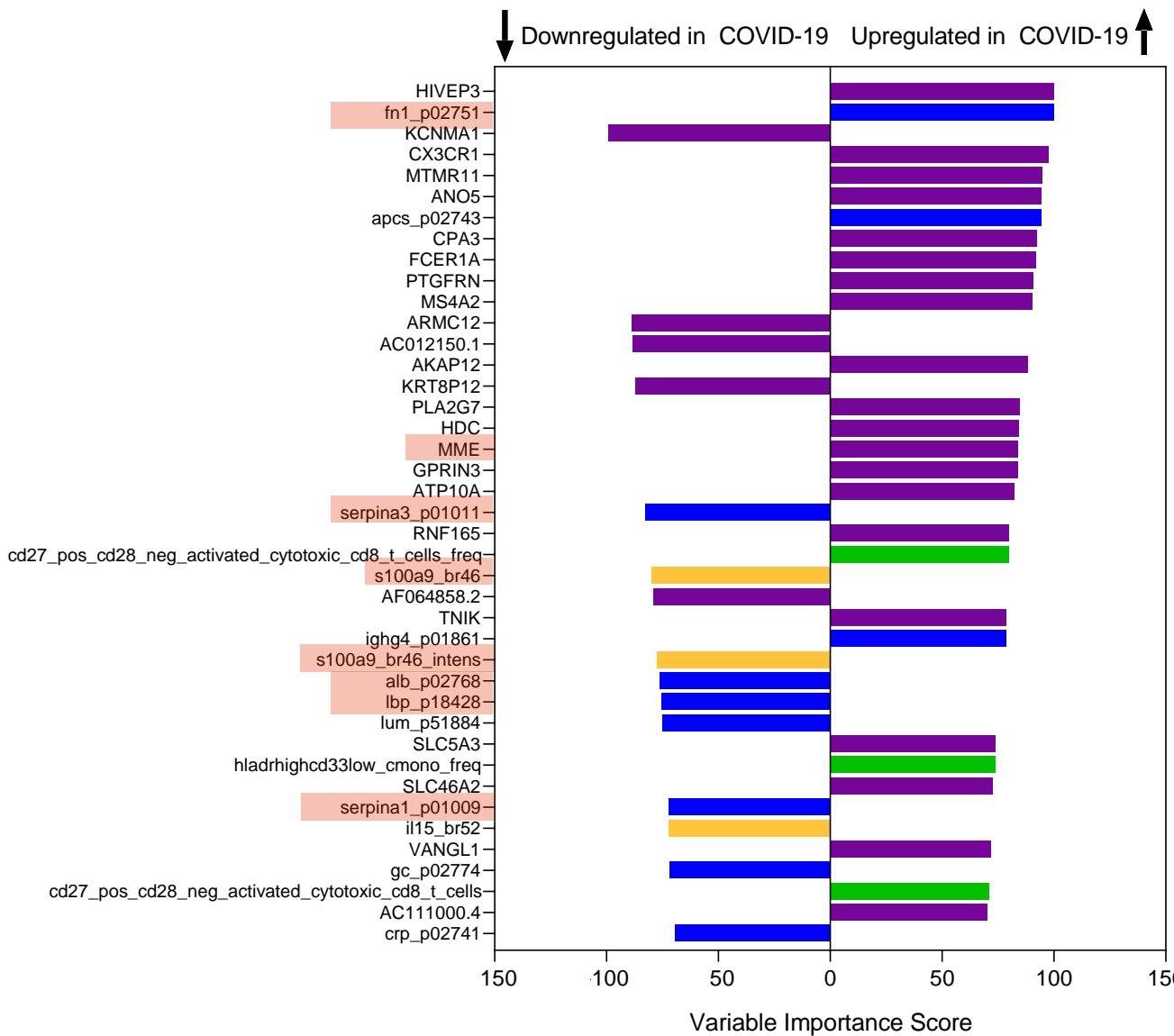
5. Model performance on test set

Pattern recognition in COMBAT using SIMON

**81 features strongly contribute to the final ML model
(variable importance score > 50)**



COVID-19: Knowns and unknowns



FN1, fibronectin → lung fibrosis in COVID-19 patients
Xu et al, 2020. doi: [10.1186/s12931-020-01445-6](https://doi.org/10.1186/s12931-020-01445-6).

MME - nephrilysin → part of the ACE2 complex
Emameh et al, 2020. doi: [10.1186/s12575-020-00124-6](https://doi.org/10.1186/s12575-020-00124-6).

Alpha-1-antichymotrypsin (Serpina 3) → up-regulated in 2006 SARS
Wan et al, 2006. doi: [10.1002/pmic.200500638](https://doi.org/10.1002/pmic.200500638).

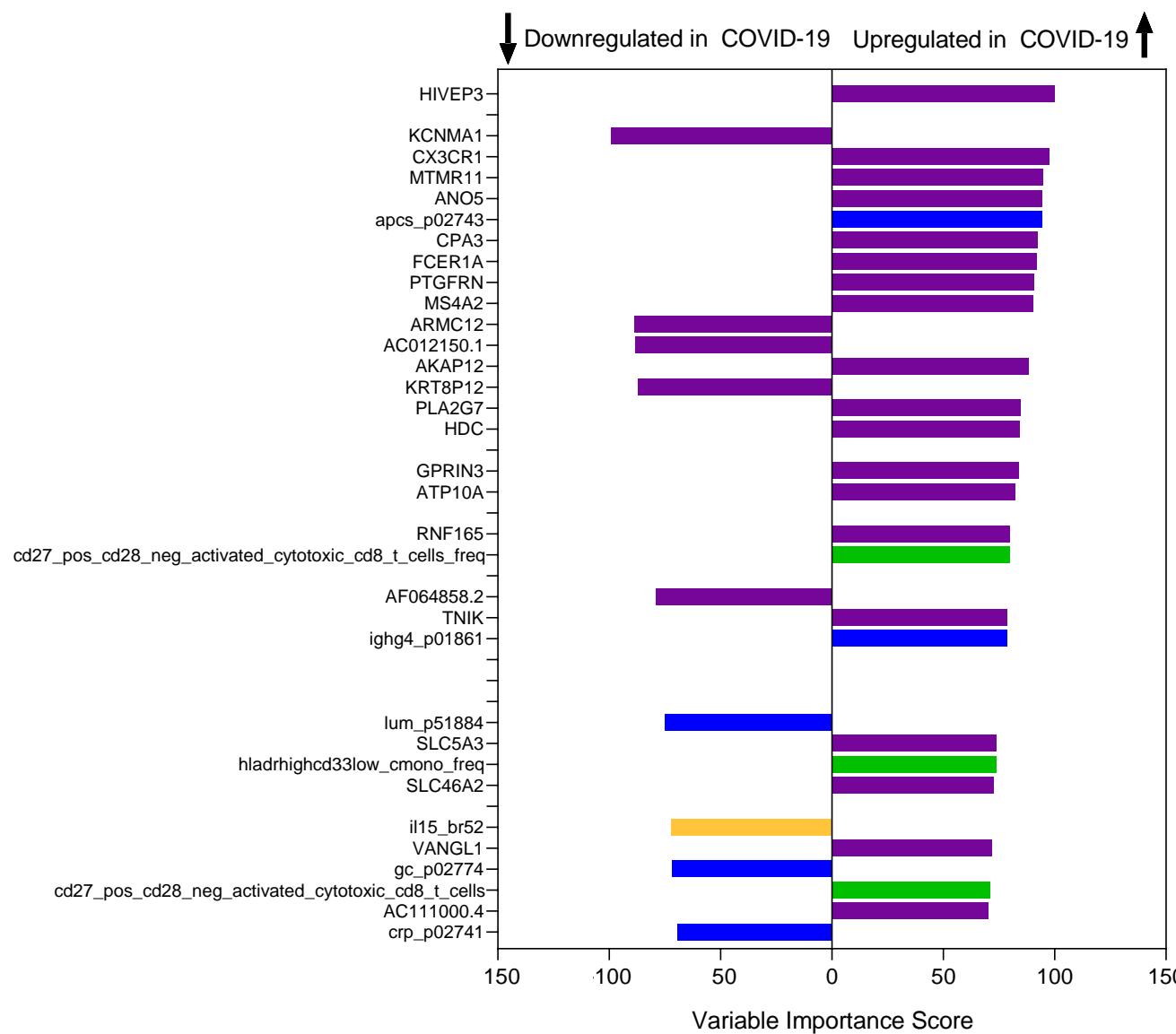
Calprotectin (S100A9) → increased in severe COVID-19 patients
Silvin et al, Cell 2020. doi: [10.1016/j.cell.2020.08.002](https://doi.org/10.1016/j.cell.2020.08.002).

Albumin (ALB) → hypoalbuminemia in COVID-19 patients

LBP, LPS binding protein → increased in COVID-19 patients
Hoel et al, J Intern Med 2020. doi: [10.1111/joim.13178](https://doi.org/10.1111/joim.13178).

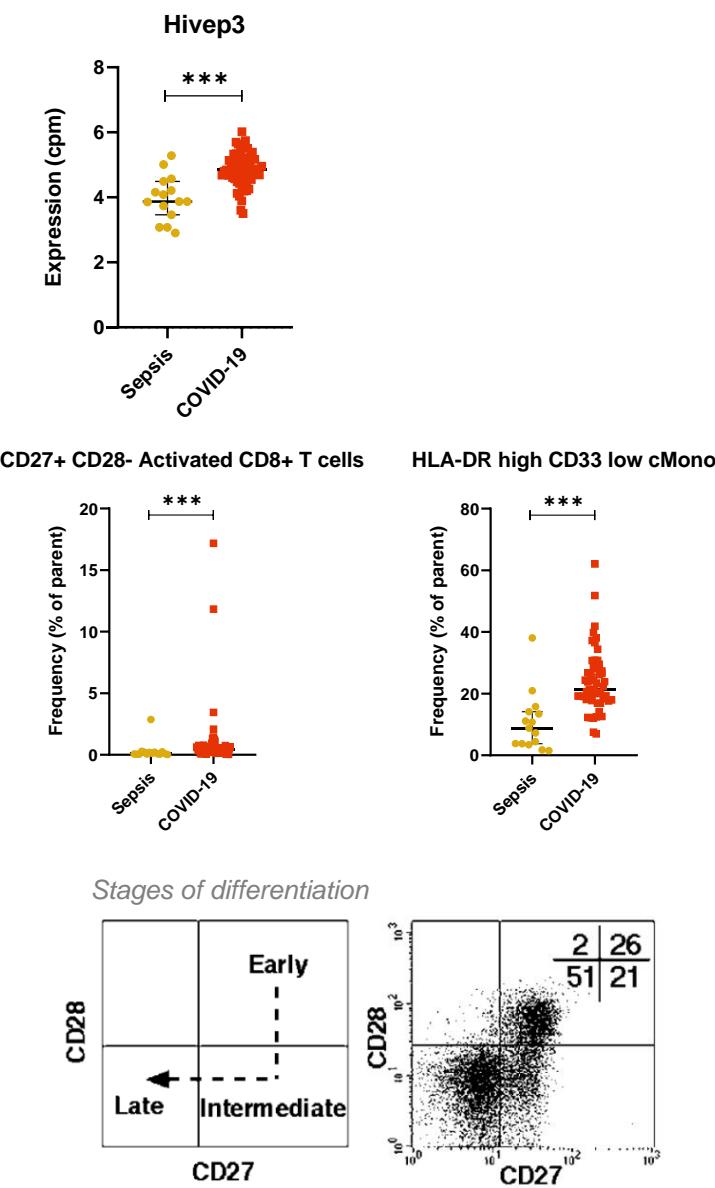
Alpha-1-antitrypsin (Serpina 1) → inhibits SARS CoV2 infection
de Loyola et al, 2020. doi: [10.1002/rmv.2157](https://doi.org/10.1002/rmv.2157)

Pattern recognition in COMBAT using SIMON

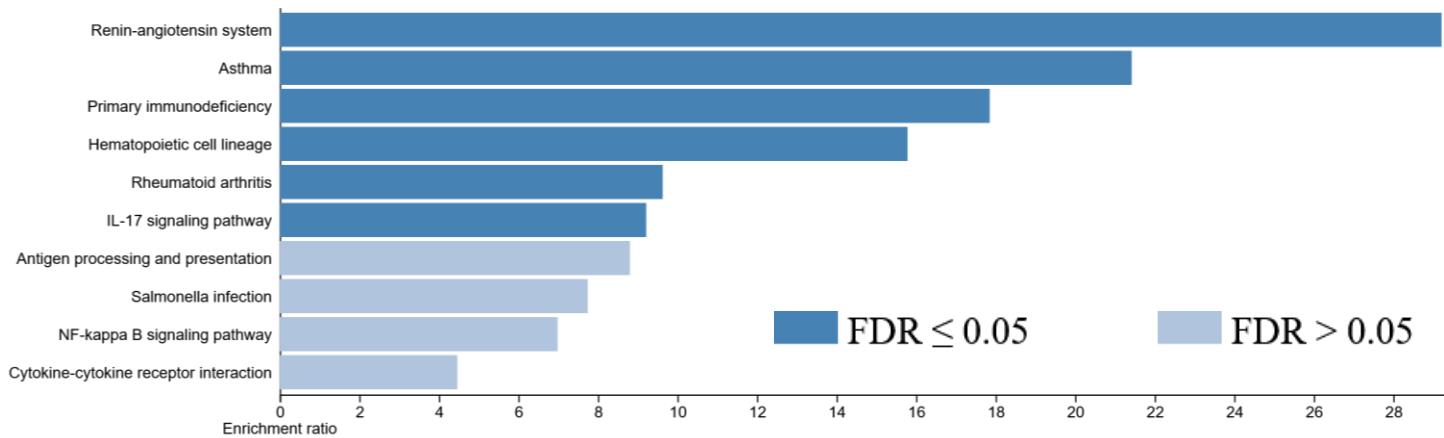


41 features with score above 69

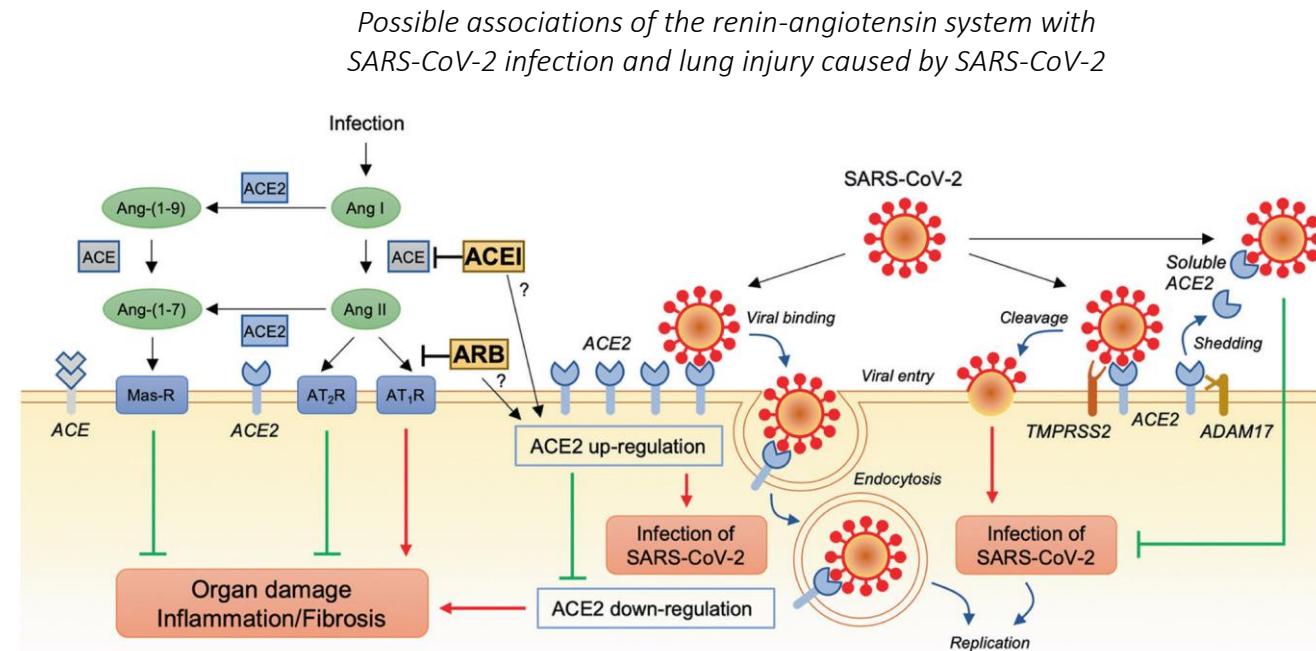
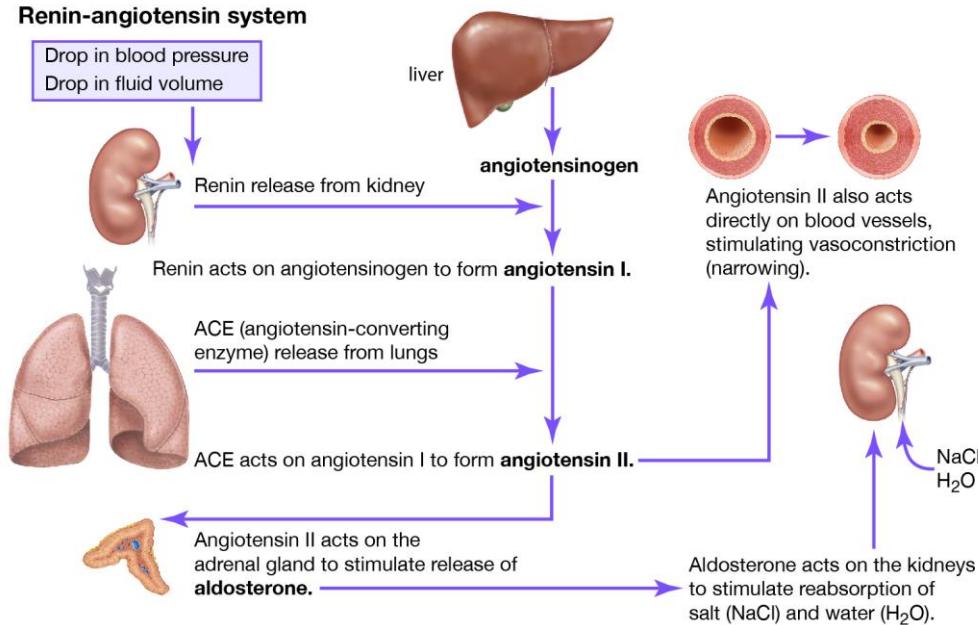
RNAseq – purple, Mass Spec – blue, CyTOF – green, Luminex - yellow



COVID-19 vs sepsis patients – KEGG pathway analysis

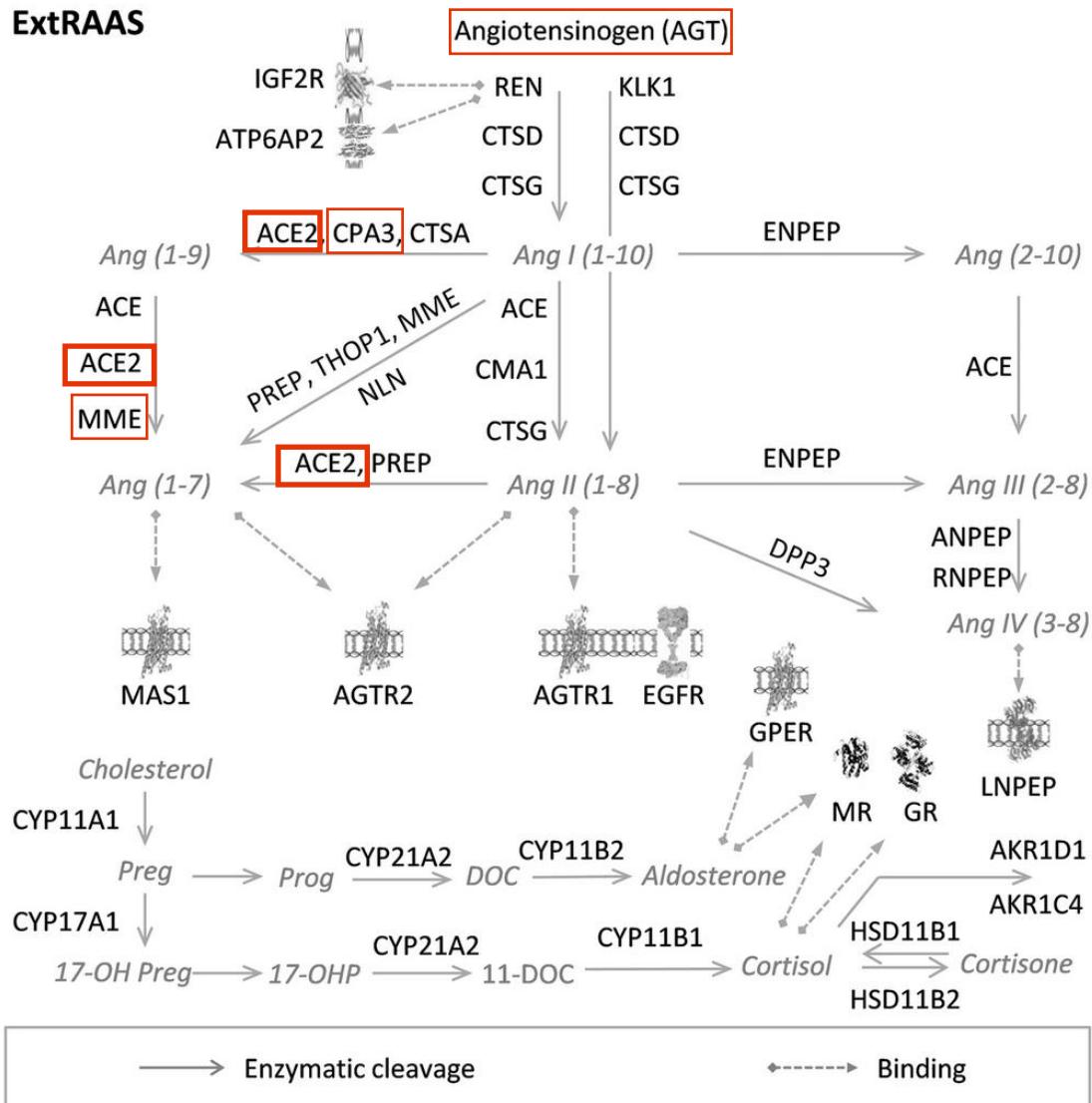


Renin-angiotensin system

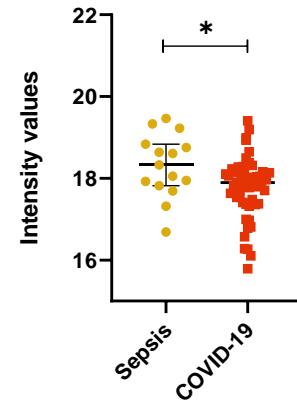


Renin-angiotensin system in COVID-19

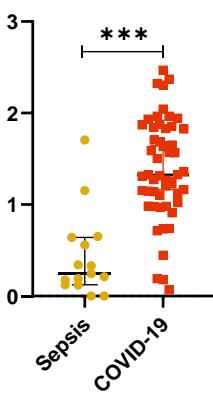
ExtRAAS



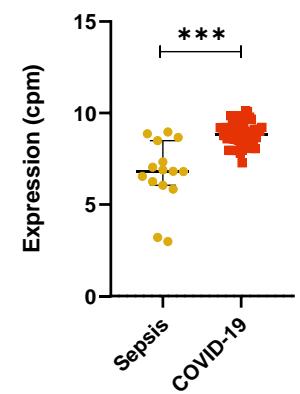
Angiotensinogen AGT



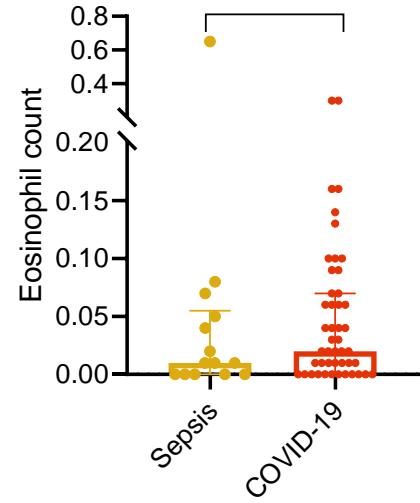
CPA3



MME



ns



carboxypeptidase 3 (CPA3)

- Correlated with poorly controlled asthma
- Correlating with increased numbers of eosinophils
- Associated with **eosinophilic esophagitis**

Fricker et al, J Allergy Clin Immunol . 2019. doi: 10.1016/j.jaci.2018.12.1020.



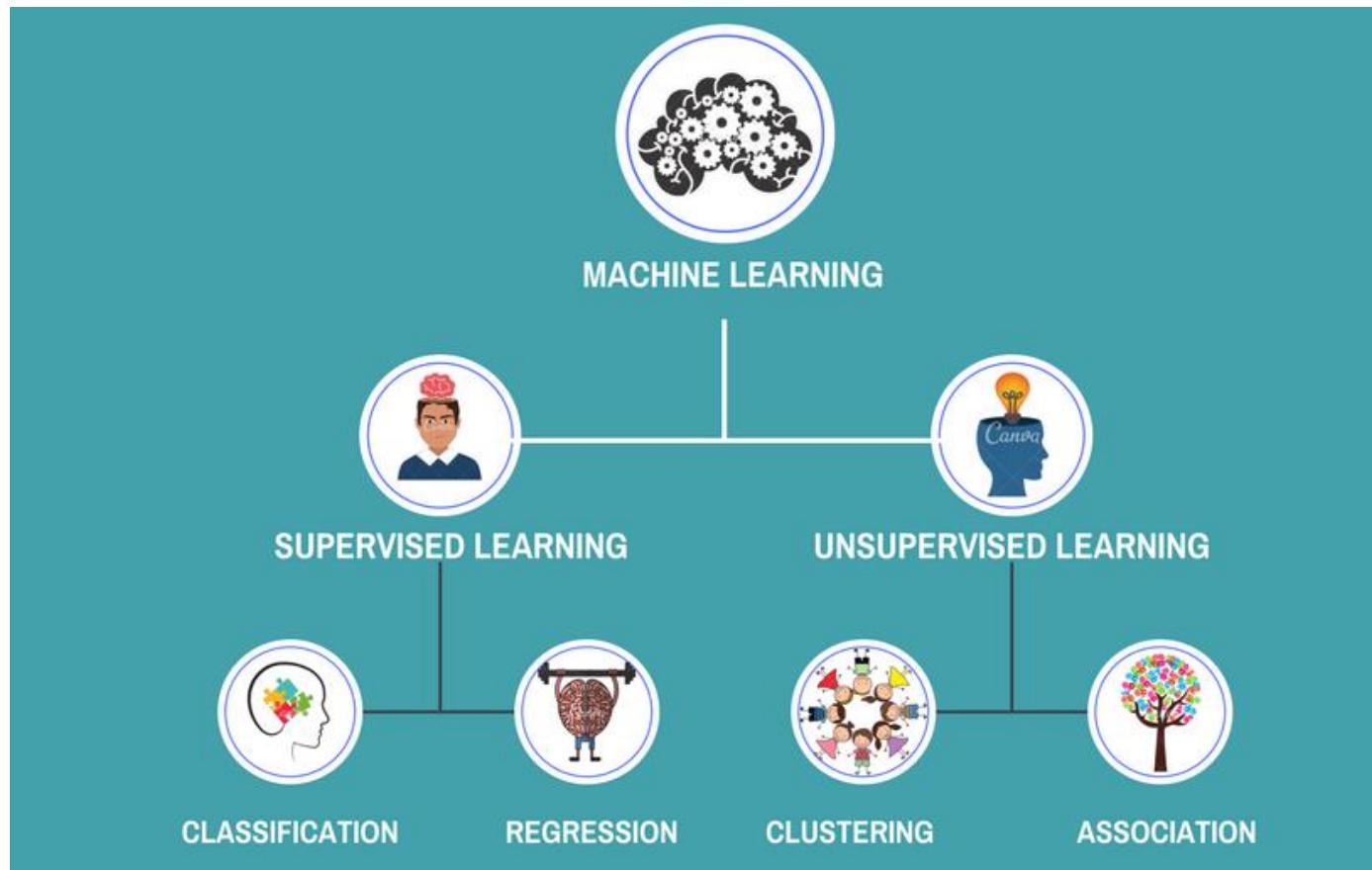
KUKA

KNEXT

KNEXT barista robot

Perfect coffee every time

Part II – Let the algorithms run loose: the power of unsupervised ML



PITCH Consortium

Protective Immunity from T cells to Covid-19 in Healthcare workers



Extension of the UK SIREN Study

Dept of Health & Social Care Funded

- Prospective longitudinal cohort study in 5 sites
 - Oxford (Eleanor Barnes, Philippa Matthews, Chris Conlon, Katie Jeffrey)
 - Liverpool (Lance Turtle)
 - Sheffield (Thushan de Silva, Sarah Rowland Jones)
 - Birmingham (Alex Richter)
 - Newcastle (Chris Duncan, Rebecca Payne)



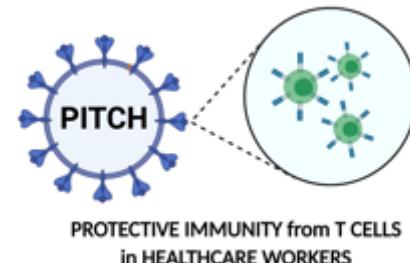
Susanna Dunachie
Chief Investigators, University of Oxford

Paul Klenerman
Chief Investigators, University of Oxford

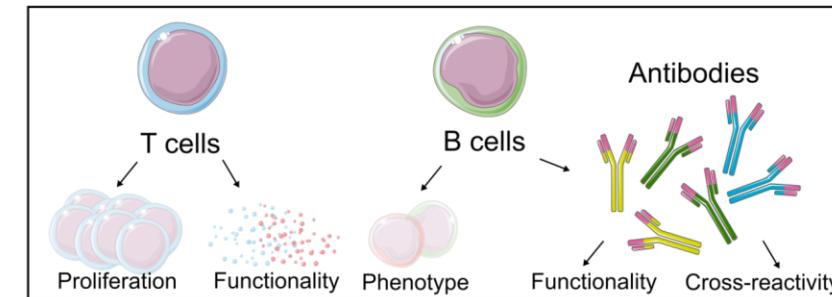
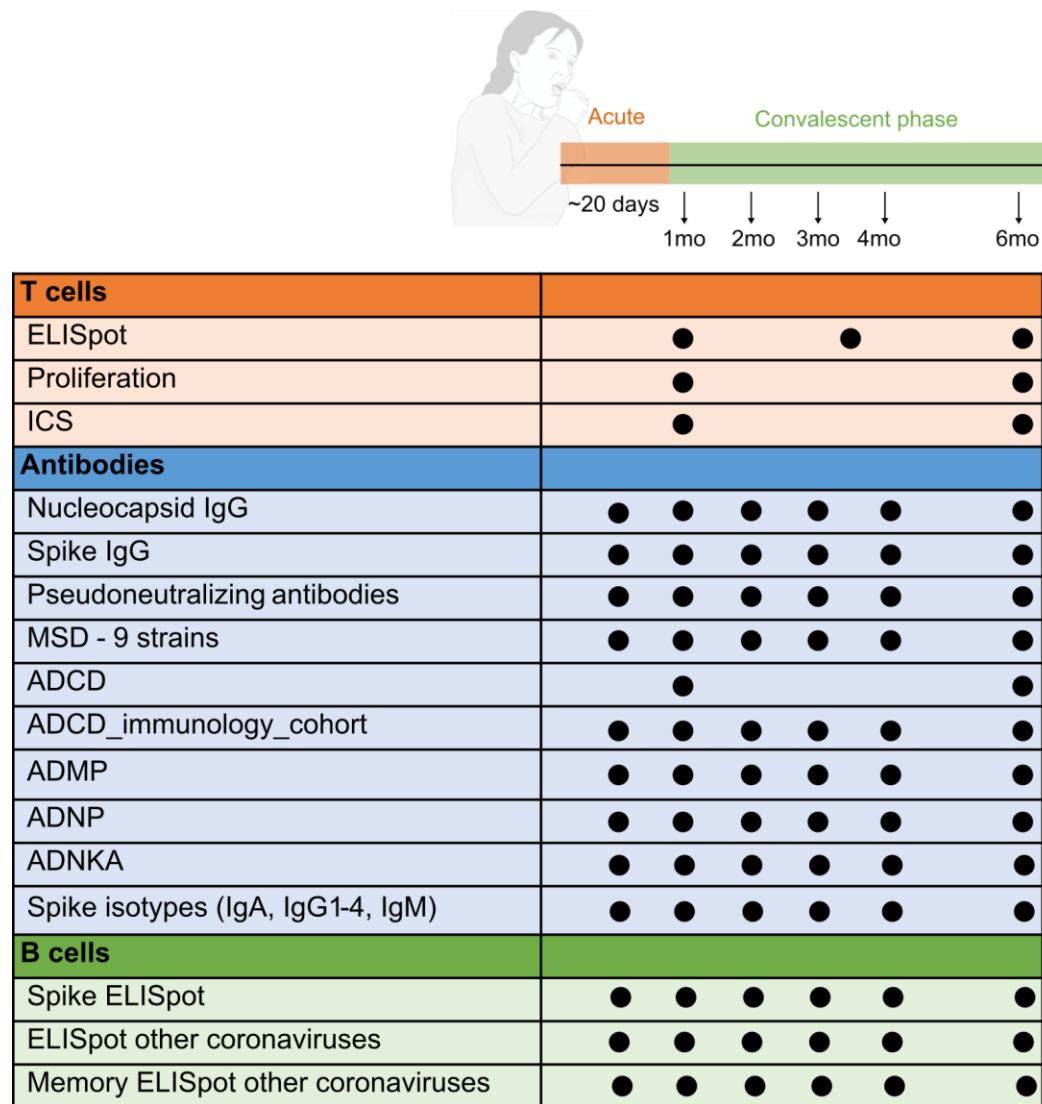
1750 Healthcare workers recruited to date



Department
of Health &
Social Care



Identifying immunological signature of long-term protective immunity to SARS-CoV-2



Total:

- **86 donors** (12 asymptomatic, 66 mild, 8 severe)
- **433 samples**

Research questions

Research question 1

What is driving the persistence of SARS-CoV-2-specific immune memory after infection?
Is there a difference between asymptomatic, mild and severe individuals?

Research question 2

Which immunological parameters at the baseline[◊] are important for **durable response that can provide protection*** against SARS-CoV2 infection?

***Durable response and protection**

→ Being seropositive 6 months pso (anti-N antibodies ≥ 1.4)

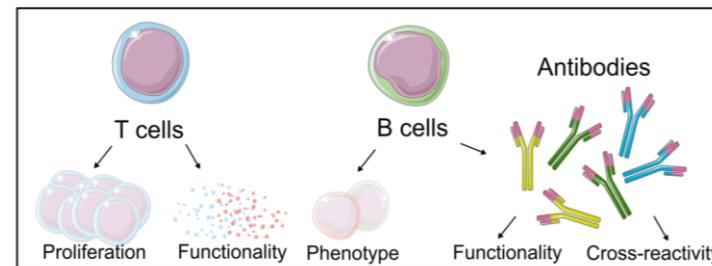
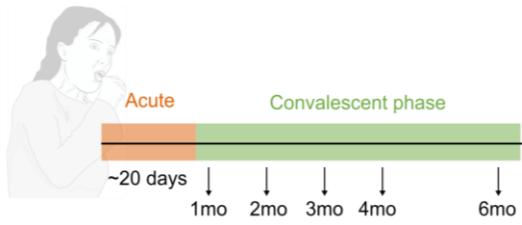
ref. Lumley et al, NEJM, 2021. Antibody Status and Incidence of SARS-CoV-2 Infection in Health Care Workers

[◊] Day 28 PSO/PCR positivity

Research question 1

What is driving the persistence of SARS-CoV-2-specific immune memory after infection?
Is there a difference between asymptomatic, mild and severe individuals?

Integrative analysis of clinical and longitudinal immunological data of SARS-CoV2 infected individuals



Clinical data: 20 features

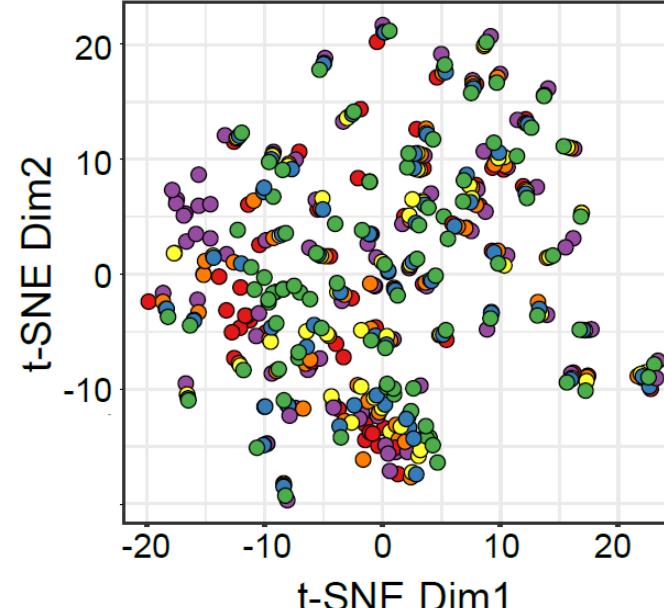
Immunological data: 51 parameters, 433 samples



Integrated clinical and immunological data

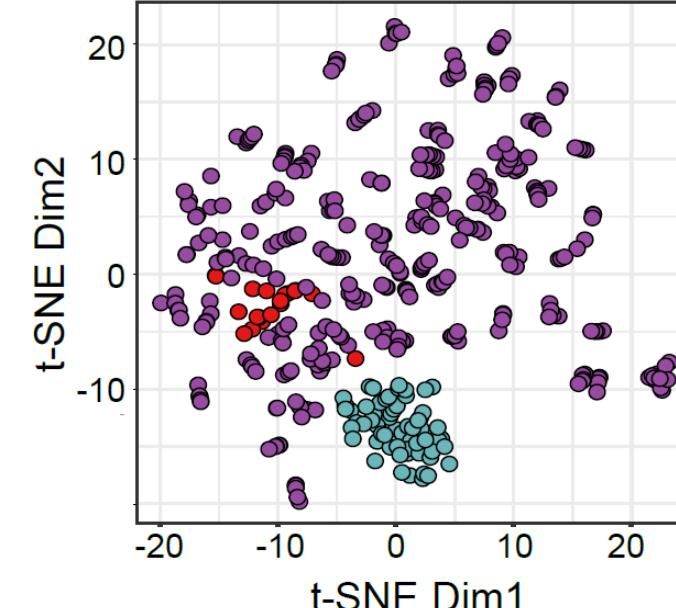
B

Timepoint: ● <1mo ● 1mo ● 2mo
● 3mo ● 4mo ● 6mo



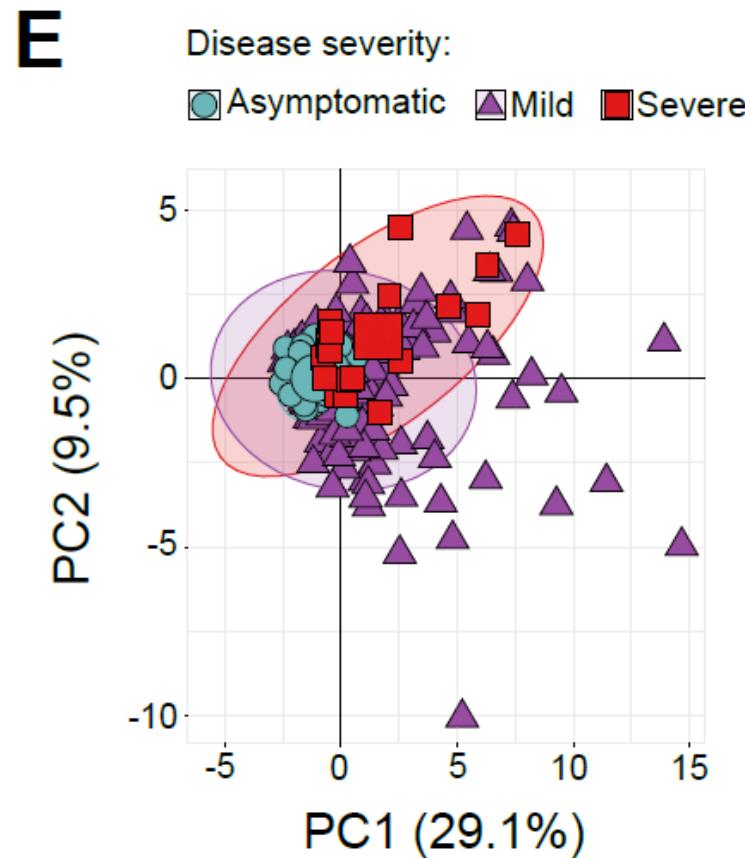
Disease severity:

● Asymptomatic ● Mild ● Severe

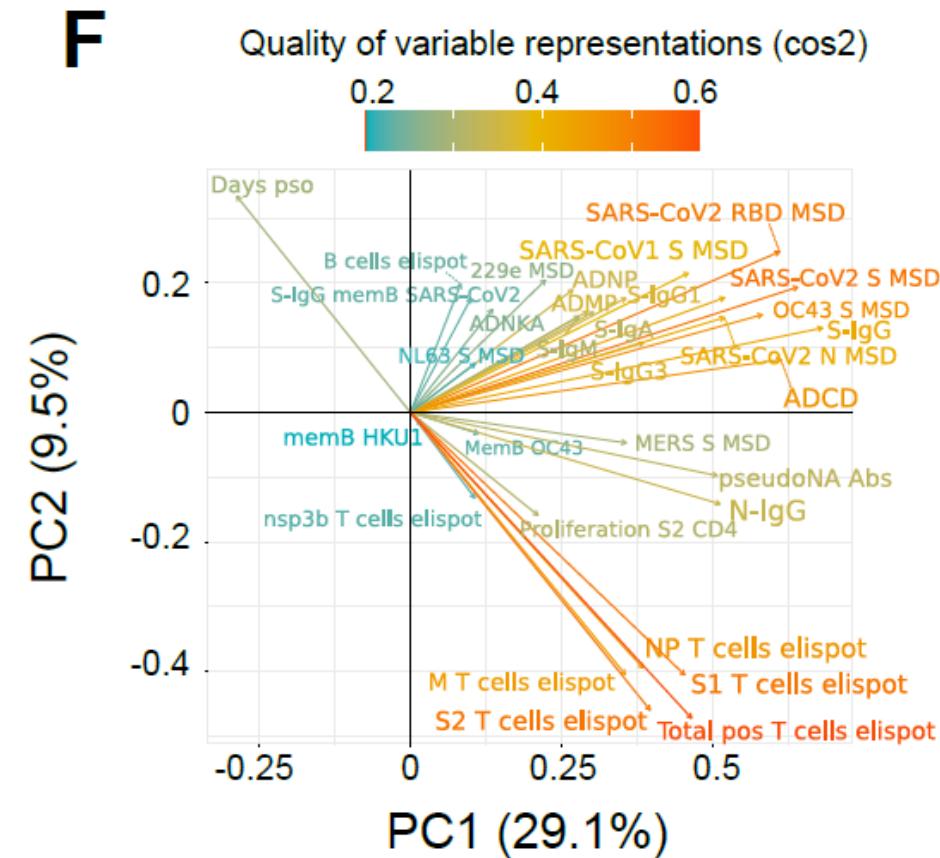


Integrated analysis of longitudinal immune response of SARS-CoV2-infected individuals identifies three immunophenotypic groups

Mild individuals have heterogenous immune responses



38.6% of the variance is explained by all measured immunological data



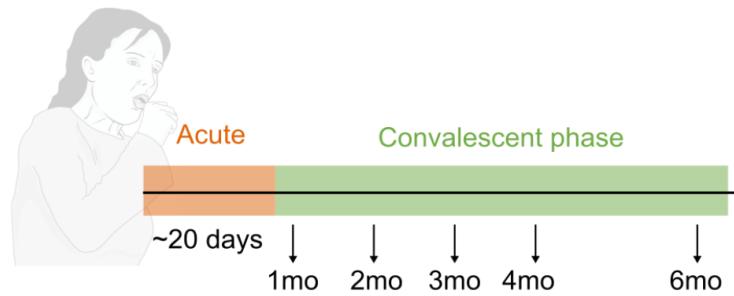
Conclusions I

SARS-CoV-2-infected individuals are separated in 3 distinct immunophenotypical groups (tSNE, PCA and hierarchical clustering):

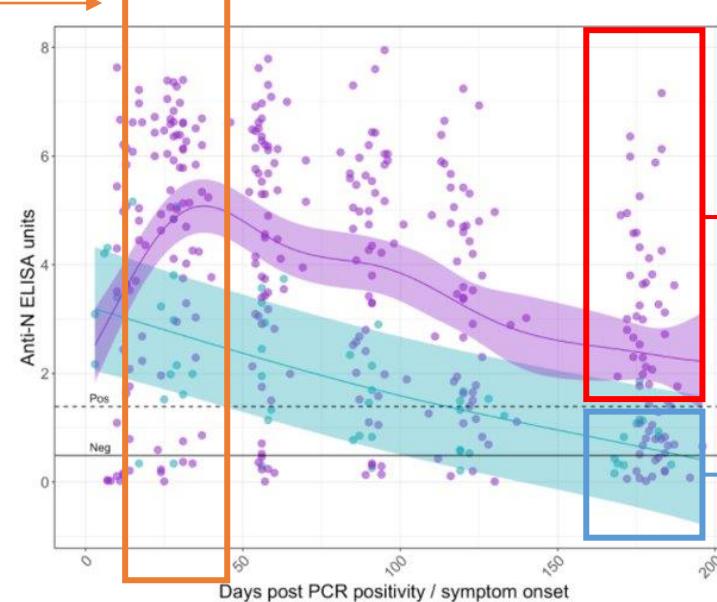
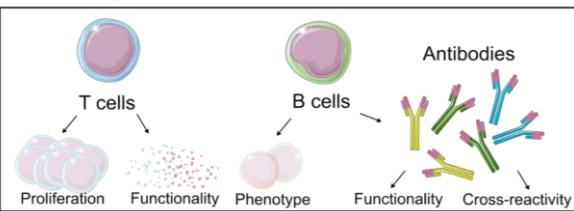
1. **Immunophenotype I:** Lower Ab and T cell responses (similar to asymptomatic)
2. **Immunophenotype II:** Higher Ab responses (similar to severe)
3. **Immunophenotype III:** Higher T cell responses

Research question 2

Which immunological parameters at the baseline are important for durable response that can provide protection against SARS-CoV2 infection?



Predictors (1mo pso)



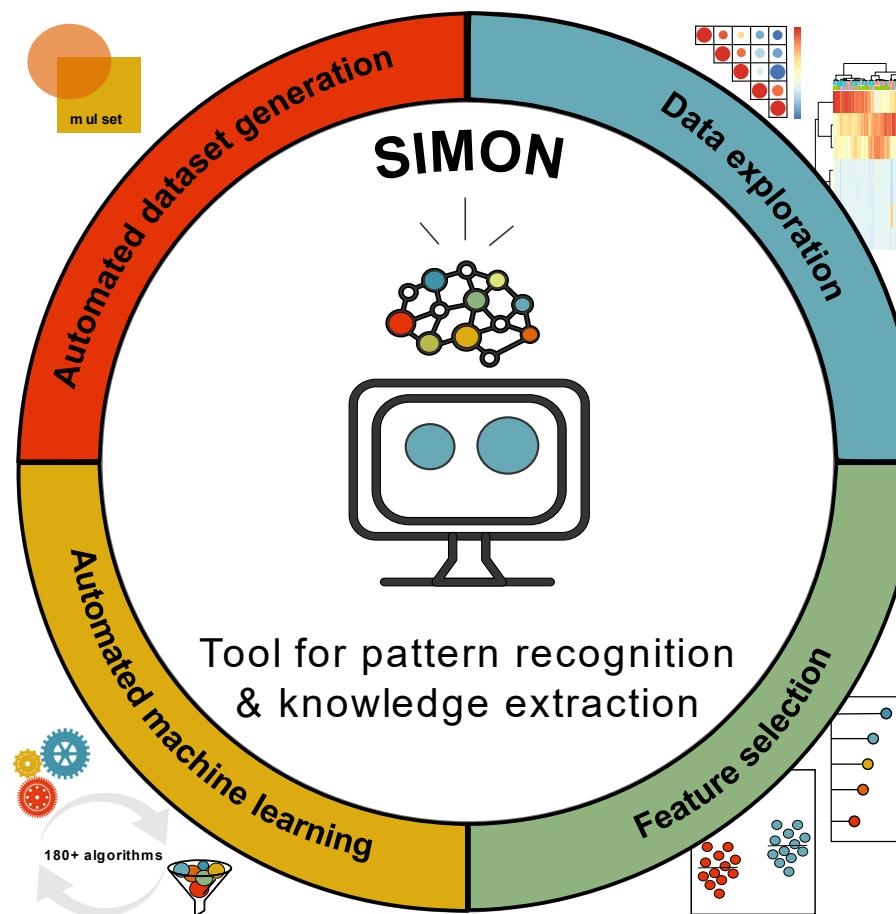
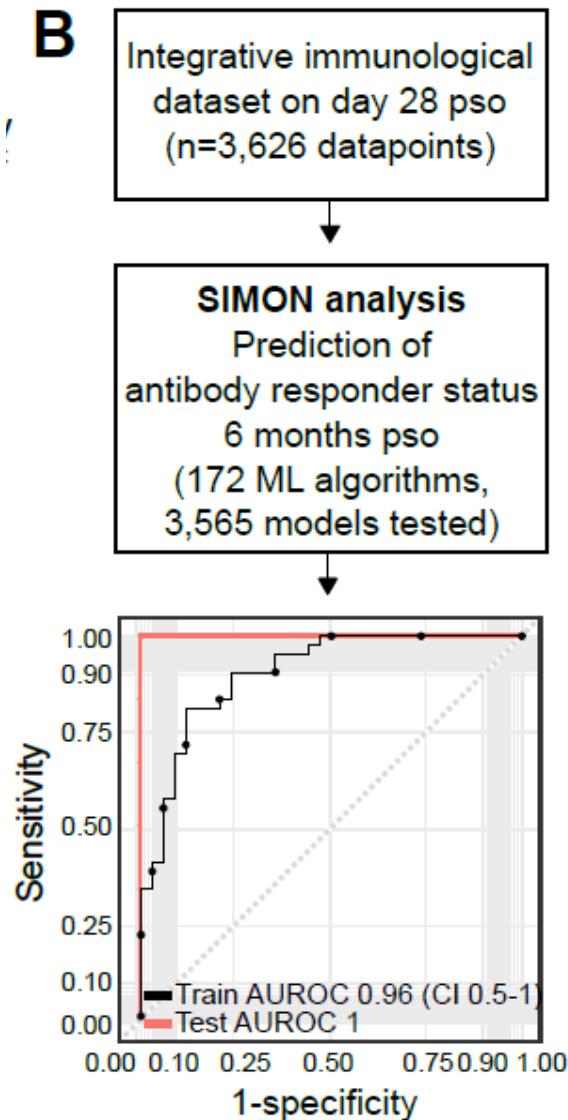
→ **High responders (protected)**

Lumley et al, NEJM, 2021. Antibody Status and Incidence of SARS-CoV-2 Infection in Health Care Workers

→ **Low responders (unprotected)**

Which immunological parameters at the baseline can predict durable and protective response against SARS-CoV-2 infection?

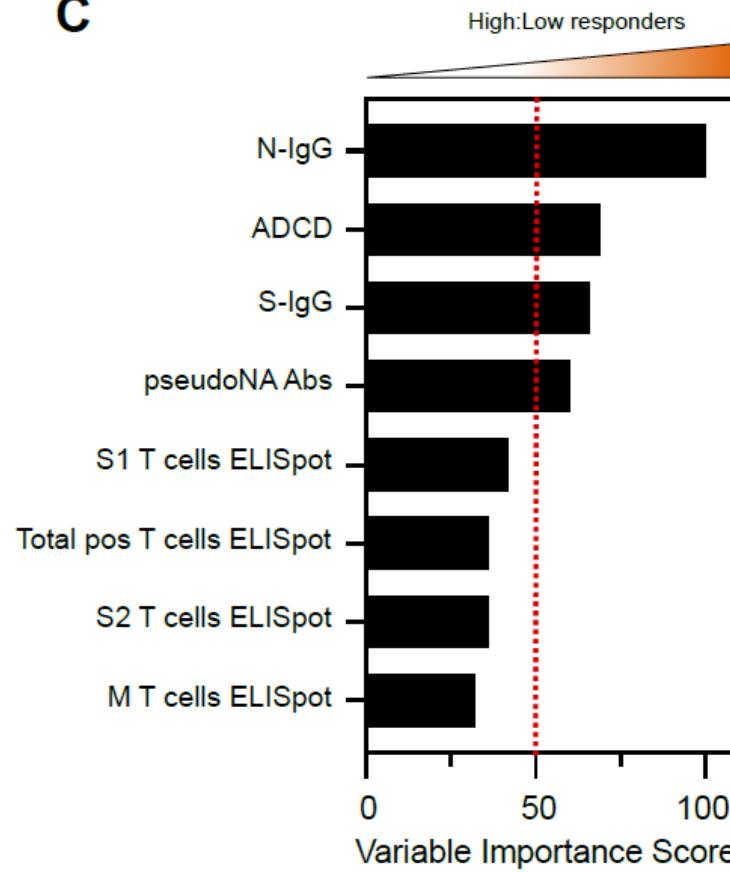
B



SIMON identifies early signature of SARS-CoV-2 protective immunity

*Baseline parameters that can predict individuals
on a trajectory for long-term immunity*

C



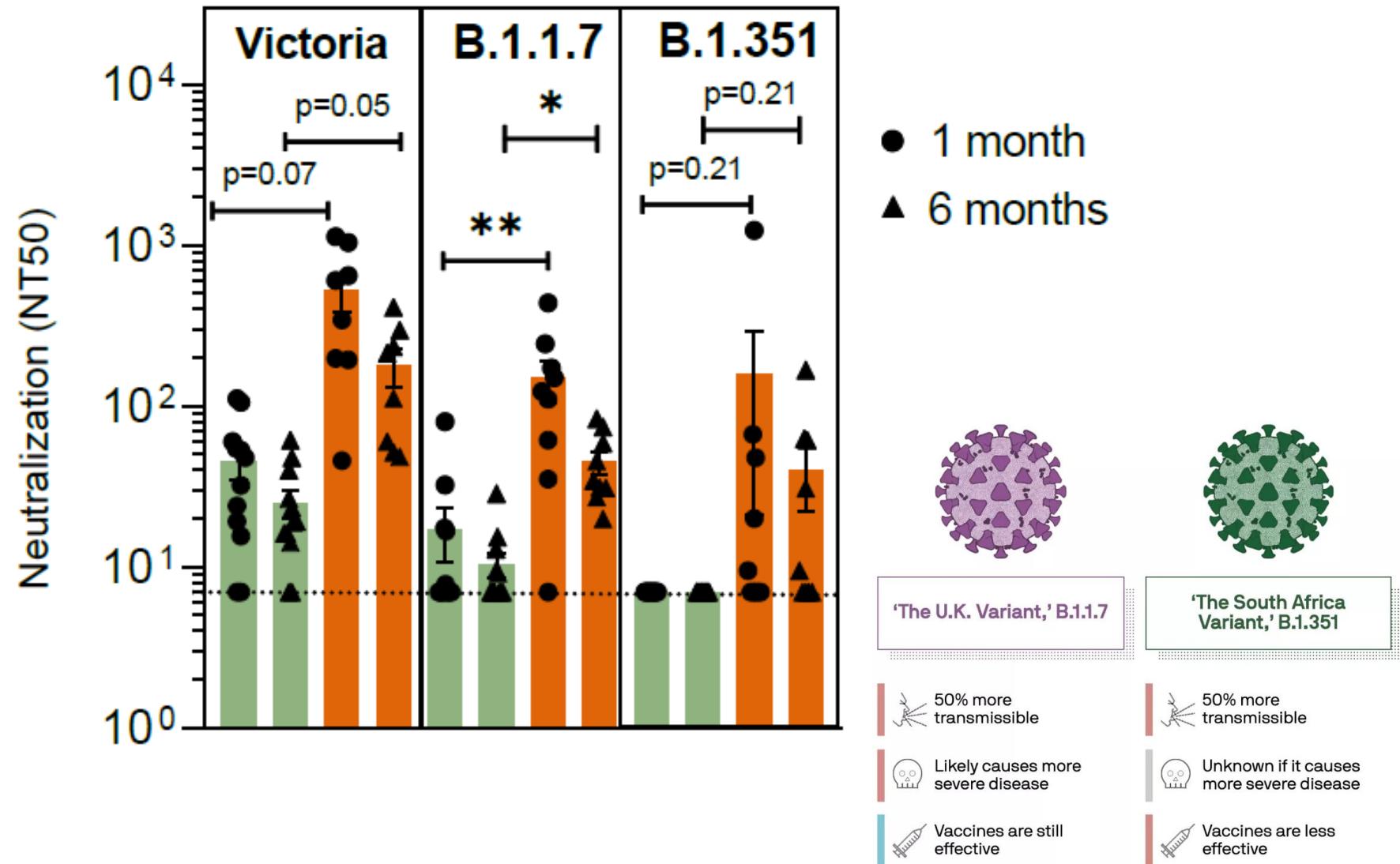
Conclusions II:

Immunological signature of long-term protective immunity to SARS-CoV2

1. Early generation of antibodies (S, N, pnAbs)
2. Early generation of T cell responses (S, M)

Could identified individuals with durable and protective SARS-CoV2 immunity be protected against novel variants?

SARS-CoV-2-infected individuals with long-term immunity have higher levels of neutralizing antibodies against variants of concern



Final conclusions

- persistence of SARS-CoV-2-specific immune memory depends on the magnitude of the immune response early after infection
- individuals with higher anti-S and anti-N antibodies and T cell responses generated early after infection mount protective SARS-CoV-2 immunity
- variable protection against novel variants



Paul Klenerman
Donal Skelly
Susanna Dunachie
Ane Ogbe
Carl-Philipp Hackstein
Hossain Delowar Akhter
Patpong Rongkard
Mohammed Ali
Barbara Kronsteiner-Dobramysl
Anthony Brown
Azim Ansari
John Frater
Matt Pace
Panagiota Zacharopoulou
Helen Brown
Philip Goulder
Vinicius Vieira
(and many more!)



Alex Mentzer
Julian Knight



Daniel O'Connor
Andrew Pollard
Jennifer Hill
Lisa Stockdale
Laura Silva-Reyes
Aline Linder
Luke Blackwell
Sagida Bibi
Elizabeth Clutterbuck
(and many more!)



NHS Trust

Lizzie Stafford
Bea Simmons
Síle Johnson
Tim James
James Grist
Chris Conlon
Katie Jeffreys
University of Oxford medical students



William James
Adam Harding



Teresa Lambe
Sandra-Belij Rammerstorfer
Amy Flaxman



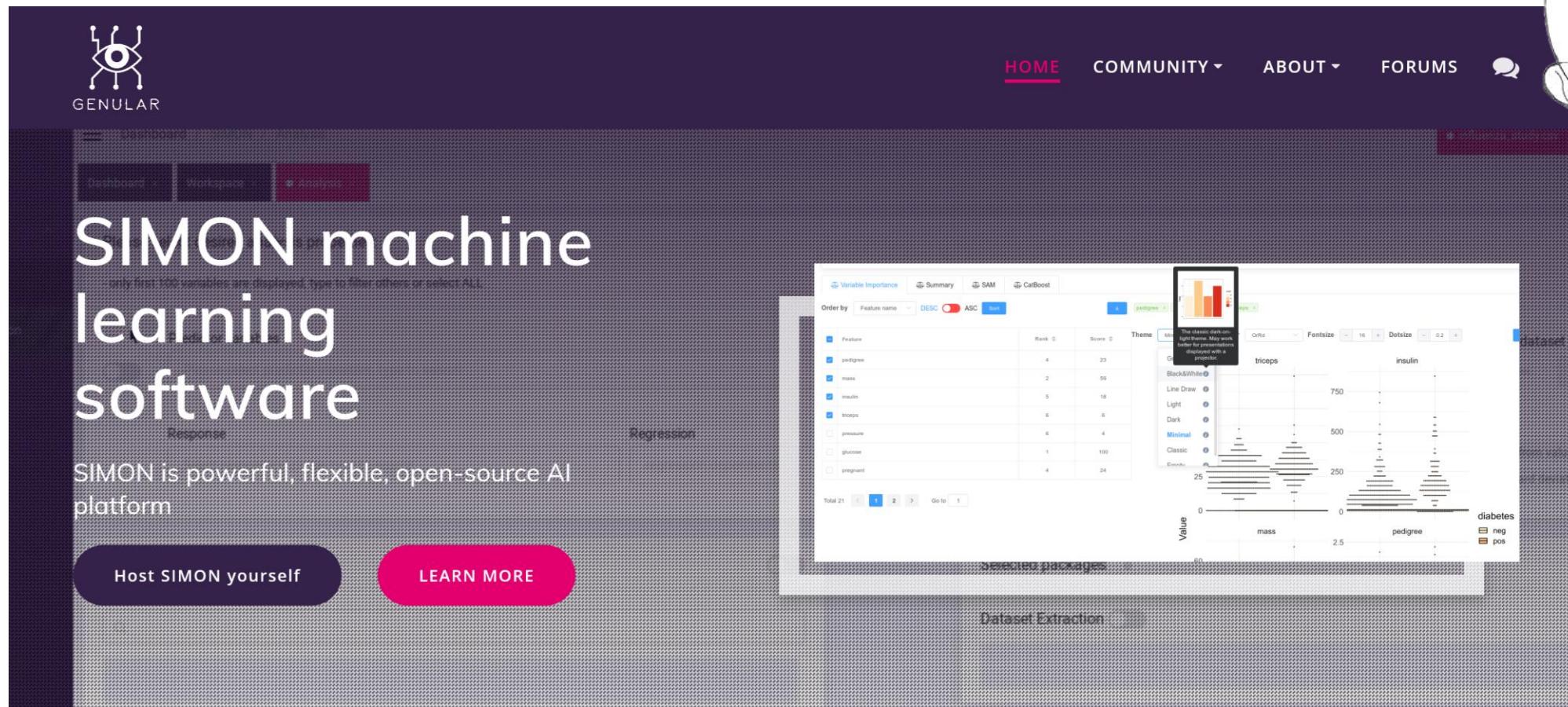
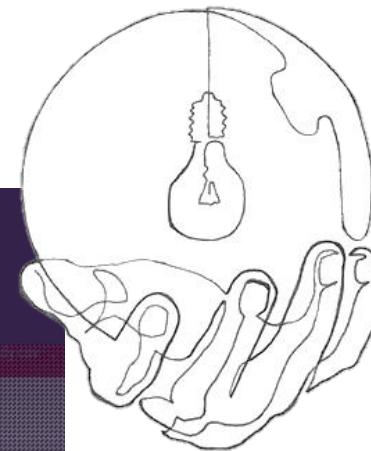
Public Health
England

Miles Carroll
Stephanie Longet
Steve Taylor
Breeze Cavell
& the Pathogen
Research Group

the
HUO FAMILY
FOUNDATION



Join open-source community supporting SIMON!



The screenshot shows the SIMON machine learning software homepage. At the top, there's a navigation bar with links for HOME, COMMUNITY, ABOUT, FORUMS, and a message icon. The main title "SIMON machine learning software" is prominently displayed in large white font. Below the title, a sub-section titled "Regression" shows a table of variable importance for a regression model. The table lists features like pedigree, mass, insulin, triceps, pressure, glucose, and pregnant, along with their ranks and scores. To the right of the table is a data visualization interface showing histograms for triceps, insulin, and diabetes, with sliders for theme, font size, and dot size.

GENULAR

SIMON machine learning software

SIMON is powerful, flexible, open-source AI platform

Host SIMON yourself

LEARN MORE

Variable Importance

Feature	Rank	Score
pedigree	4	23
mass	2	59
insulin	5	18
triceps	6	6
pressure	8	4
glucose	1	100
pregnant	4	24

Theme

The classic dark-on-light theme. May work better when displayed with a provider.

Black&White

Line Draw

Light

Dark

Minimal

Classic

Fontsize

Dotsize

Value

Dataset Extraction

diabetes

neg pos



Check out SIMON at genular.org



GENULAR

HOME

COMMUNITY ▾

ABOUT ▾

FORUMS

SIMON Knowledge Base

Have a Question?

Search the documentation...

Search

Installation

📄 Installation Quickstart

Machine Learning

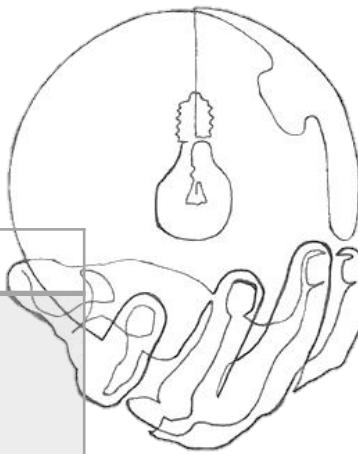
📄 How to perform SIMON analysis?

📄 Instruction videos



Star us on
GitHub!!!

Join open-source community supporting SIMON!



Project	How To Help	Next Step
Localization (English, German, French, Chinese, Arabic)	Help us translate SIMON into your language. If some translation is missing or incorrect you can easily help us by correcting it.	Join our Translation Community
Tutorials	Help others use and understand SIMON	Write a tutorial or record it, with usage examples
Organizing	Ask questions on recently opened GitHub issues to move the discussion forward	Go to GitHub Issues
Write article	Help other understand what is Machine Learning & how can they apply it, by publishing blog post	e-mail us



Check out SIMON at **genular.org**



Mark M. Davis Lab
Department of Microbiology and Immunology

Elsa Sola Verges
Allison Nau
Lisa Wagar



Stanford-LPCH
Vaccine Program

Cornelia L. Dekker



Institute for Immunity,
Transplantation and Infection

The Human Immune Monitoring Center

Mike Leipold
Yael Rosenberg-Hasson
Janine Bodea Sung
Holden Maecker



Institute for Immunity,
Transplantation and Infection

Biomedical Data Science

Purvesh Khatri



National Institutes of Health
Turning Discovery Into Health



Marie Skłodowska-Curie grant
(FluPRINT, Project No 796636)

Thank you

