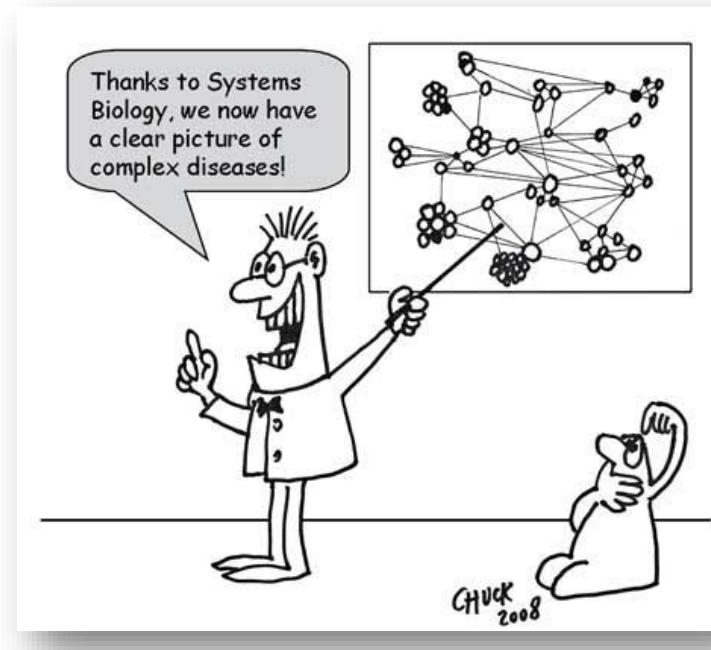


# Systems immunology: an intro to multi-omics data integration and machine learning



**Adriana Tomic**  
Systems Immunology | Oxford Vaccine Group



@TomicAdriana



adriana.tomic@paediatrics.ox.ac.uk

# Training course - overview

## **Part I – SIMON, pattern recognition and knowledge extraction platform (March 28<sup>th</sup> 2022)**

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

**Theoretical part (10-11am) ~1h**

- • Case study – example 1 (dealing with missing values, overfitting, model performance) **Case study (11-11:30am) ~0.5h**
- SIMON installation
  - Perform SIMON analysis using provided dataset
  - Performance metrics, evaluation and selection of high-quality models **Hands-on (1-3pm) ~2h**

## **Part II – Exploratory analysis (March 29<sup>th</sup> 2022)**

- Feature selection: scoring and elimination
- Correlation and clustering analysis

**Hands-on (9:30-11:30am) ~2h**

- Feature processing methods to avoid ‘curse of dimensionality’

**Theoretical part (1-1:30pm) ~0.5h**

- • Case study – example 2 (Multi-omics data integration) **Case study (1:30-2pm) ~0.5h**
- Discussion about project-specific problems **Discussion with practical examples ~1h**

# The art of feature engineering and selection



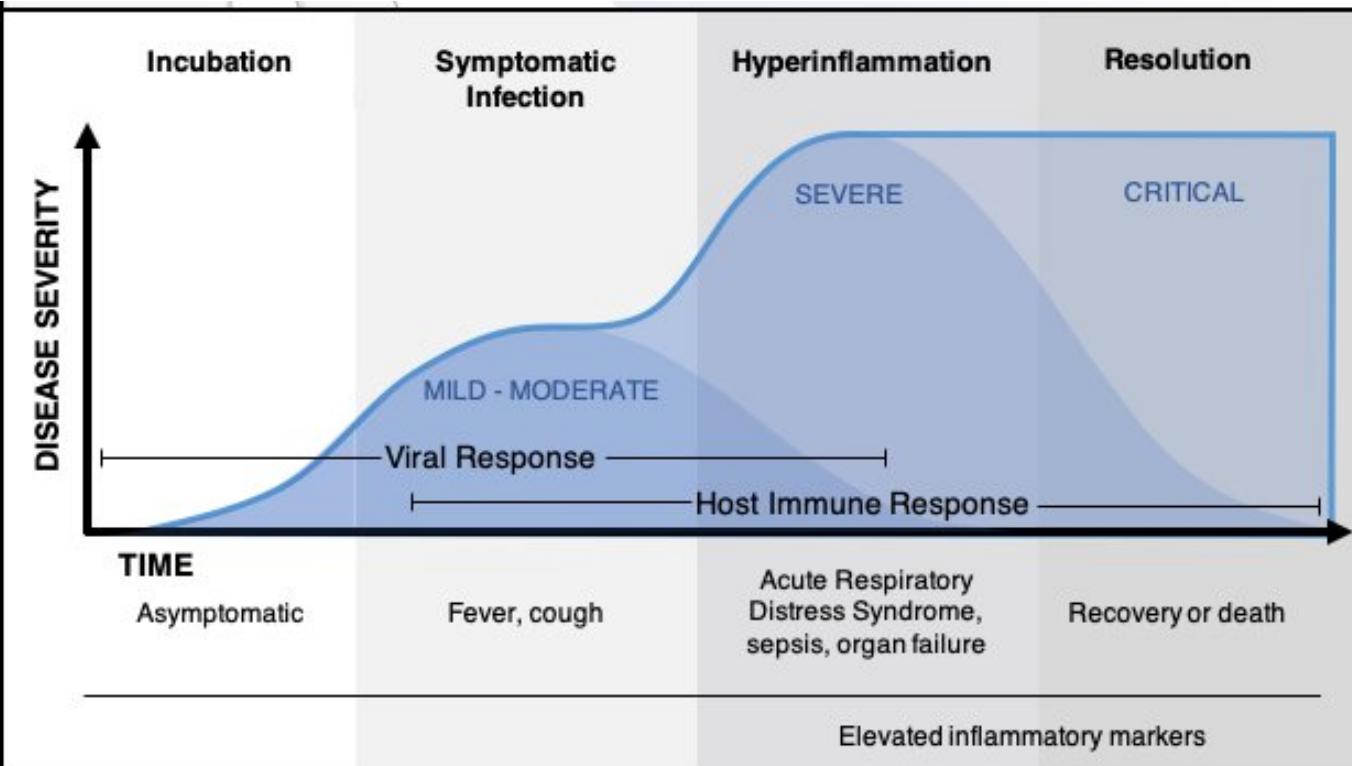
Free book online:  
<https://bookdown.org/max/FES/>

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

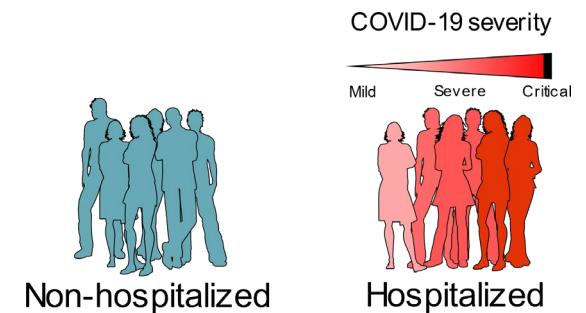
SIMON to the rescue: COMBATing COVID-19



# Understanding COVID-19

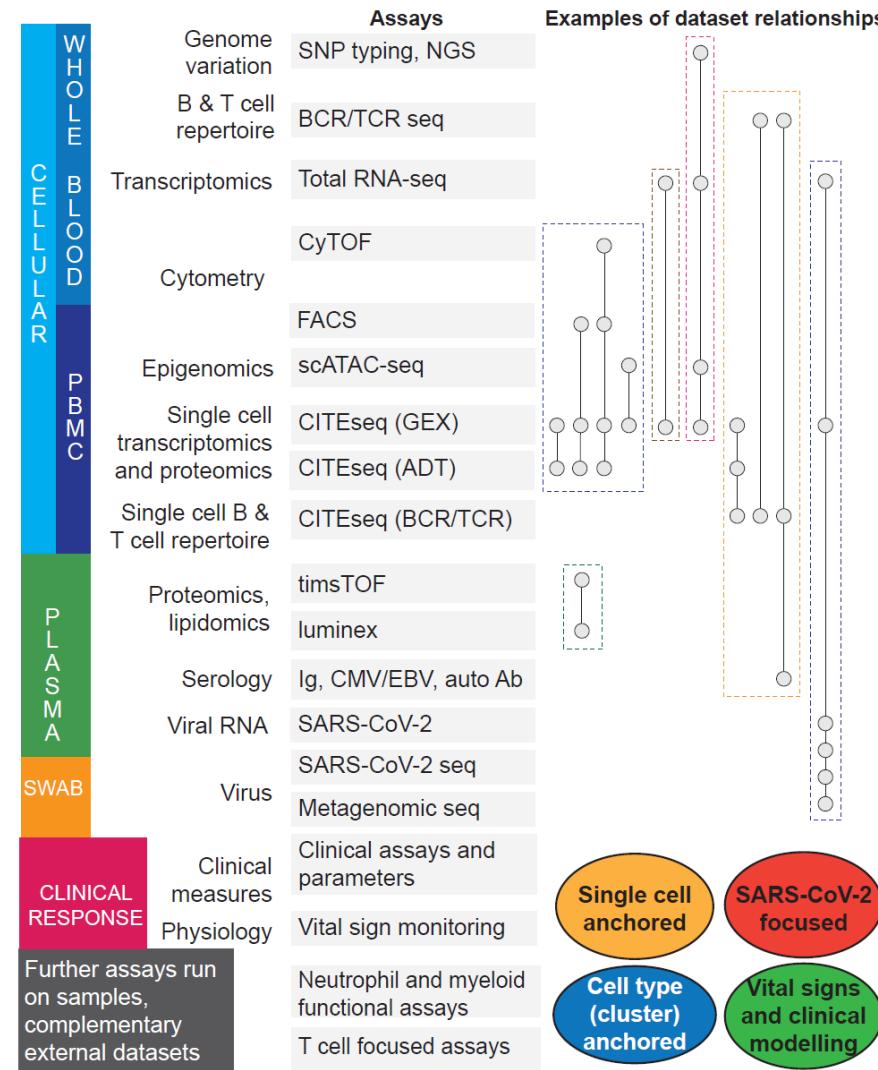


5-10% of patients with COVID-19 progress to severe disease

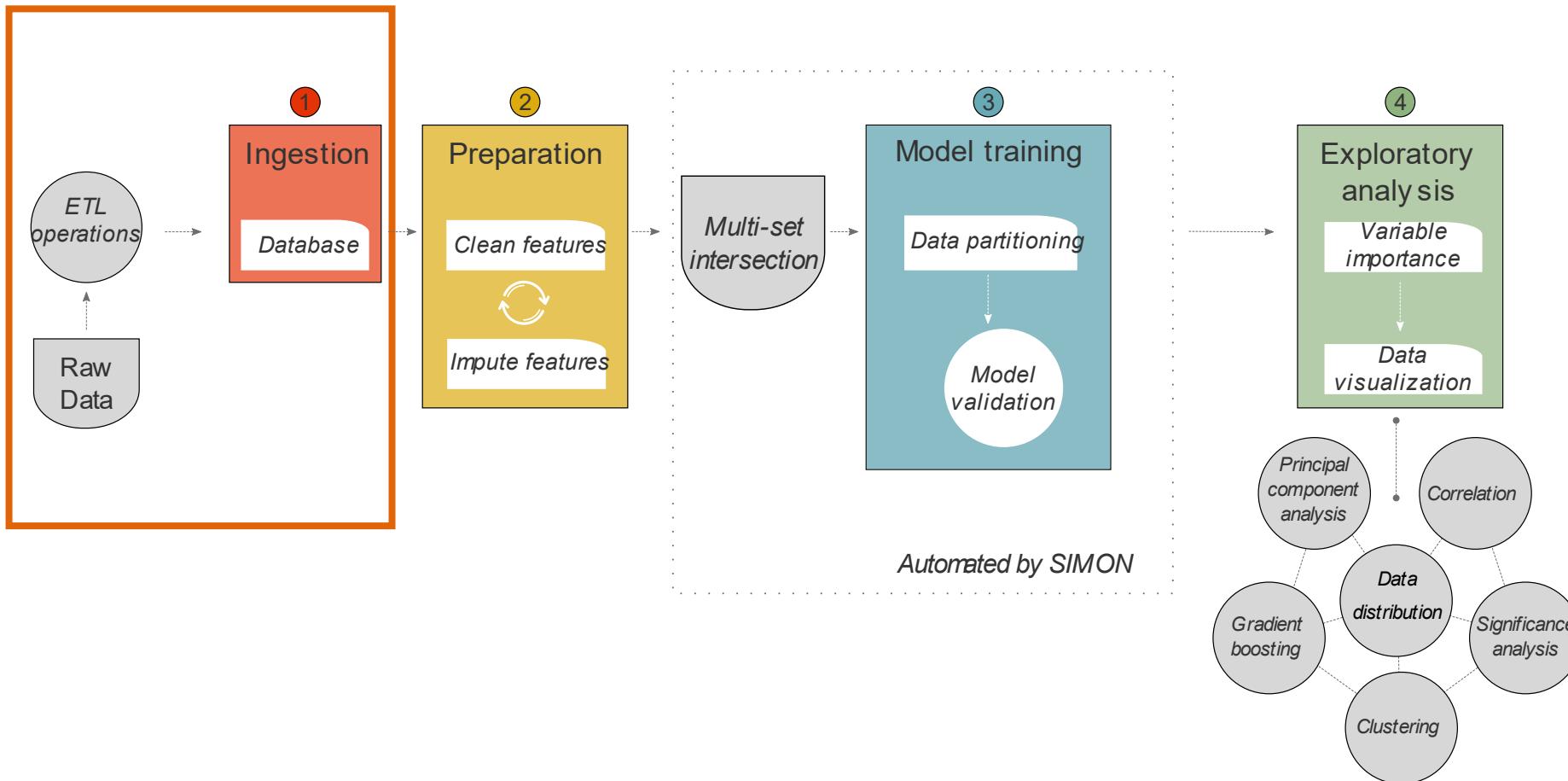


Oberfeld B. et al, SnapShot: COVID-19, Cell, 2020

# 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**

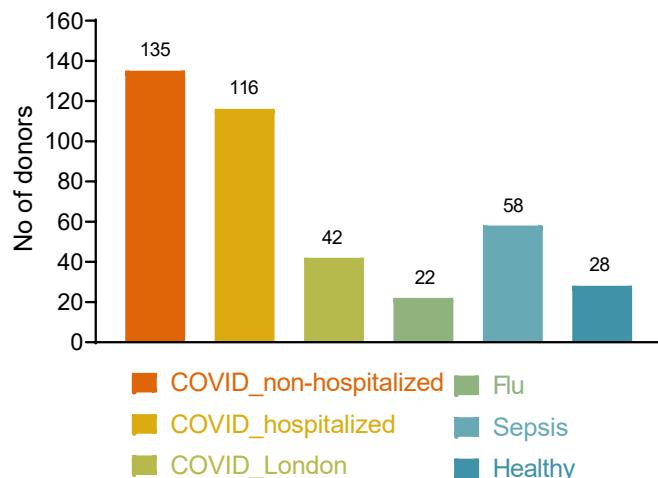
Total number of donors: 401  
COVID\_non-hospitalized: 135  
COVID\_hospitalized: 116  
COVID\_London: 42  
Flu: 22  
Sepsis: 58  
Healthy: 28

**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)

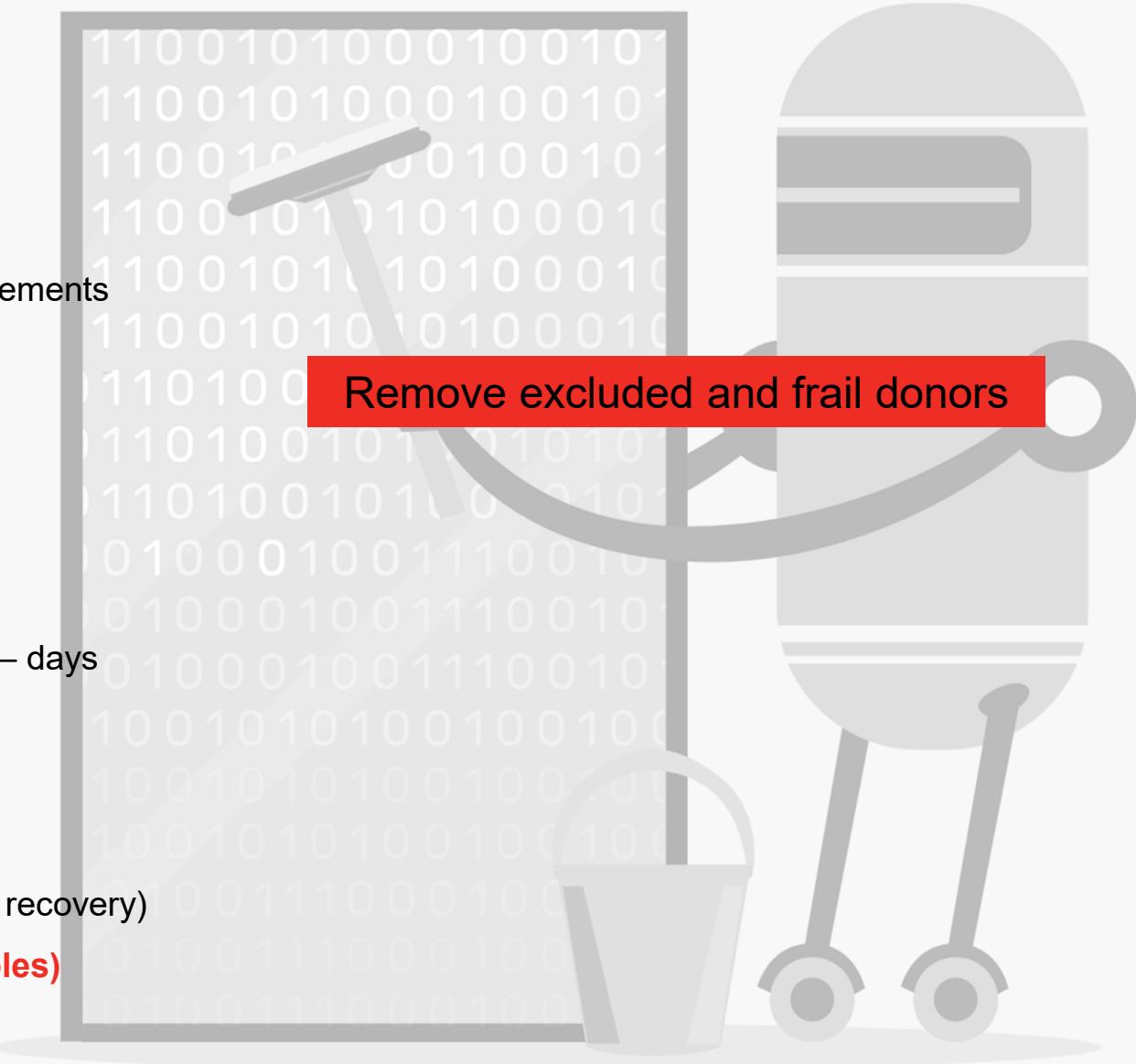
- Disease progress – for longitudinal samples (deterioration or recovery)

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

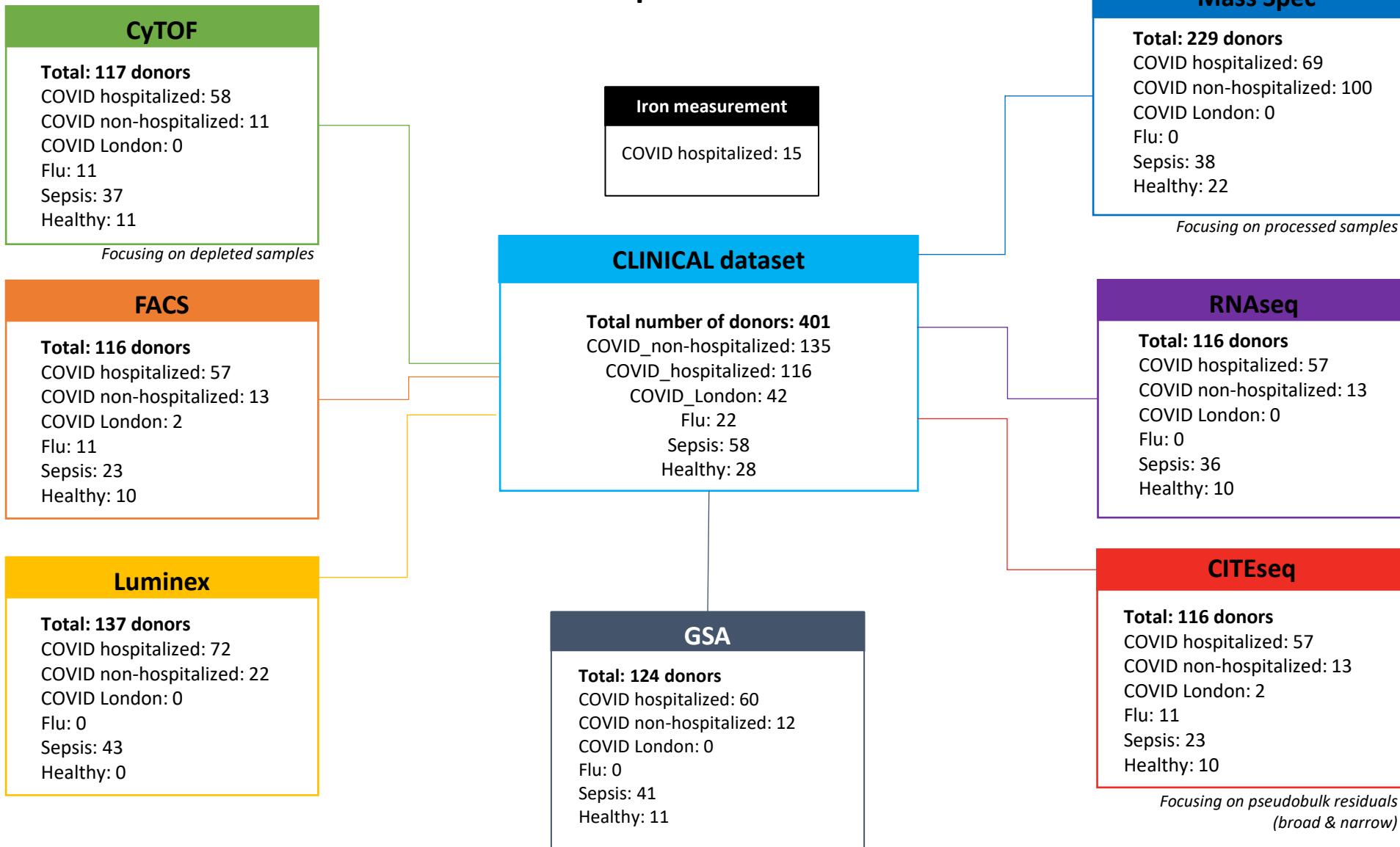
- sampling\_from\_max\_disease

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

*Written as factors*



**428 samples → 268 donors**



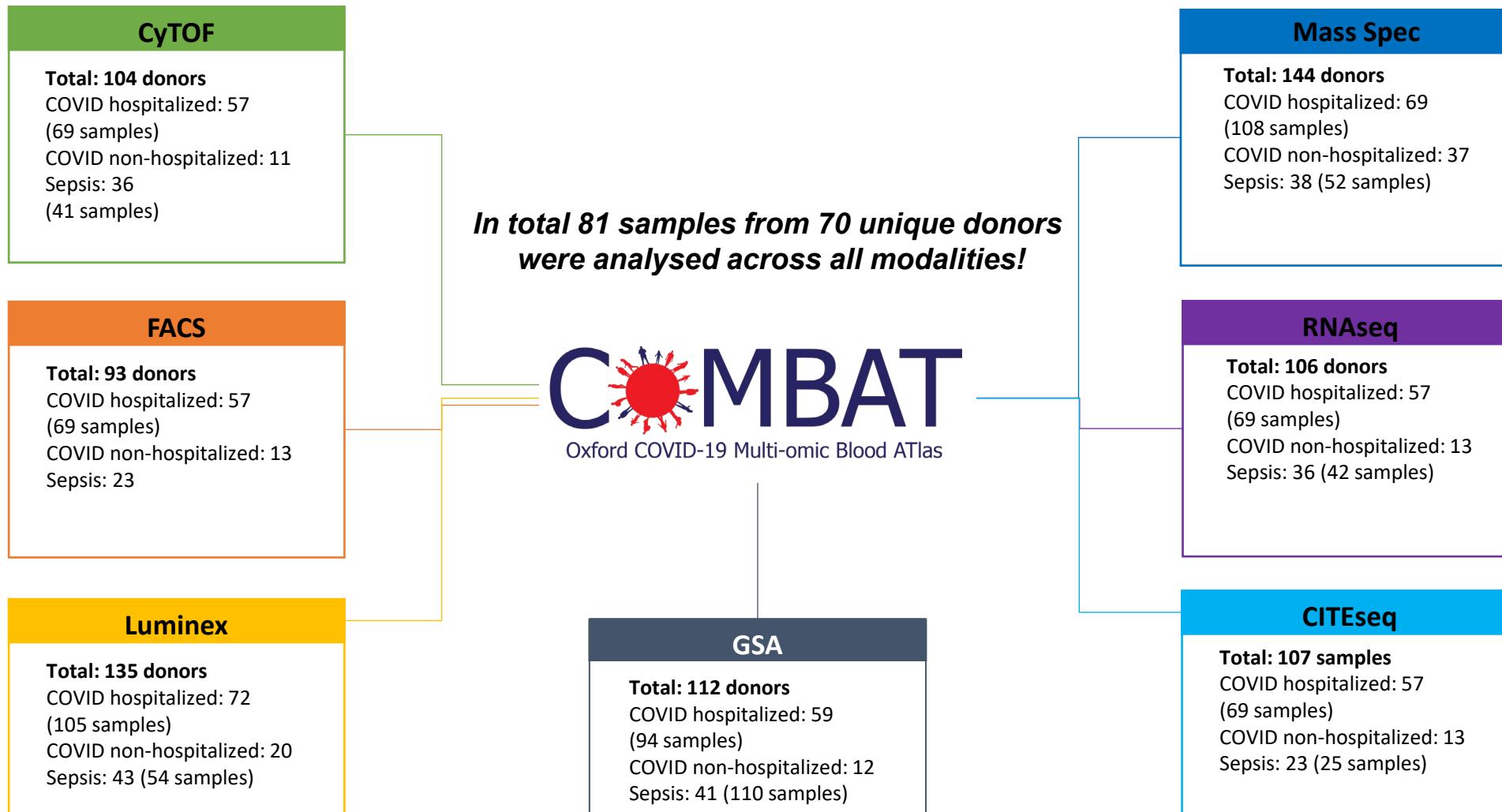
**COMBAT dataset: 428 samples → 268 donors**



Remove convalescent (covid, sepsis, hcw), Covid\_London,  
flu and healthy volunteers  
**281 samples → 167 donors**

*Which samples were analysed across modalities?*

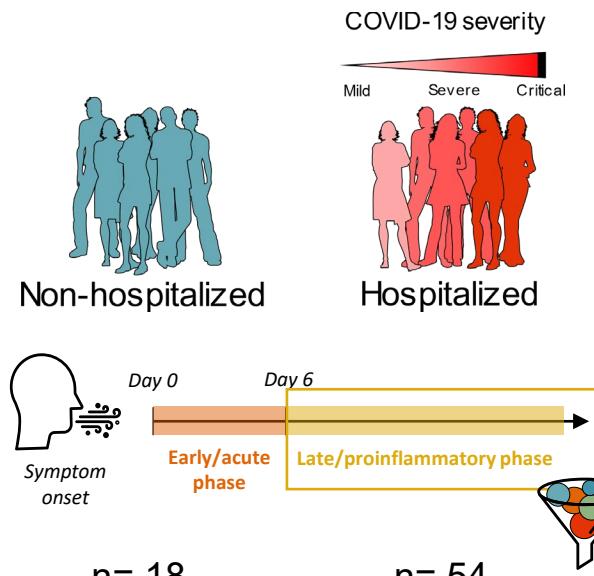
# Samples and unique donors analysed across modalities



# Knowledge discovery using COMBAT dataset

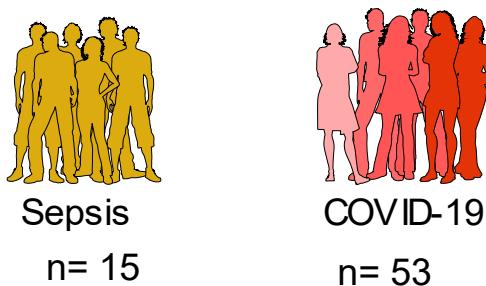
## Outcome 1

Difference between Sars-CoV2-infected individuals?



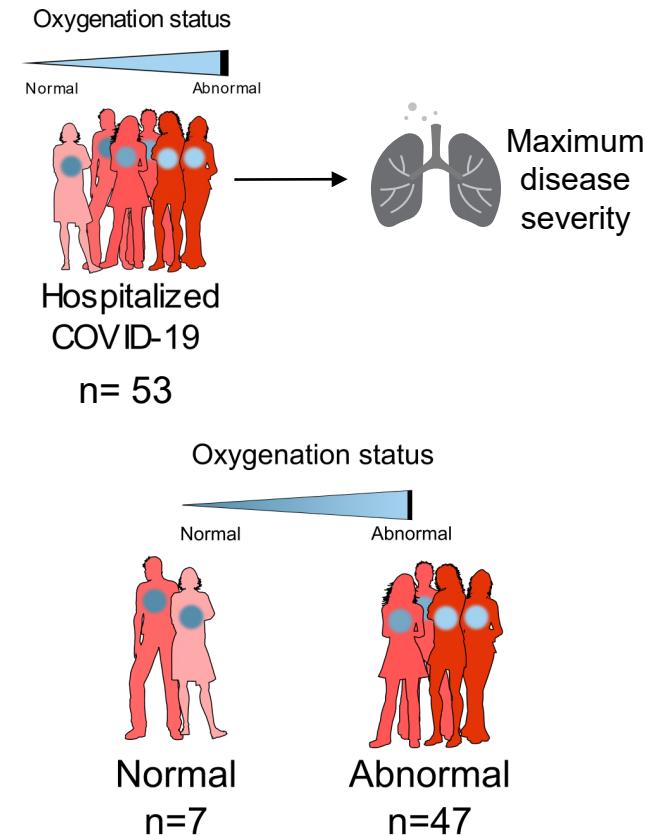
## Outcome 2

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

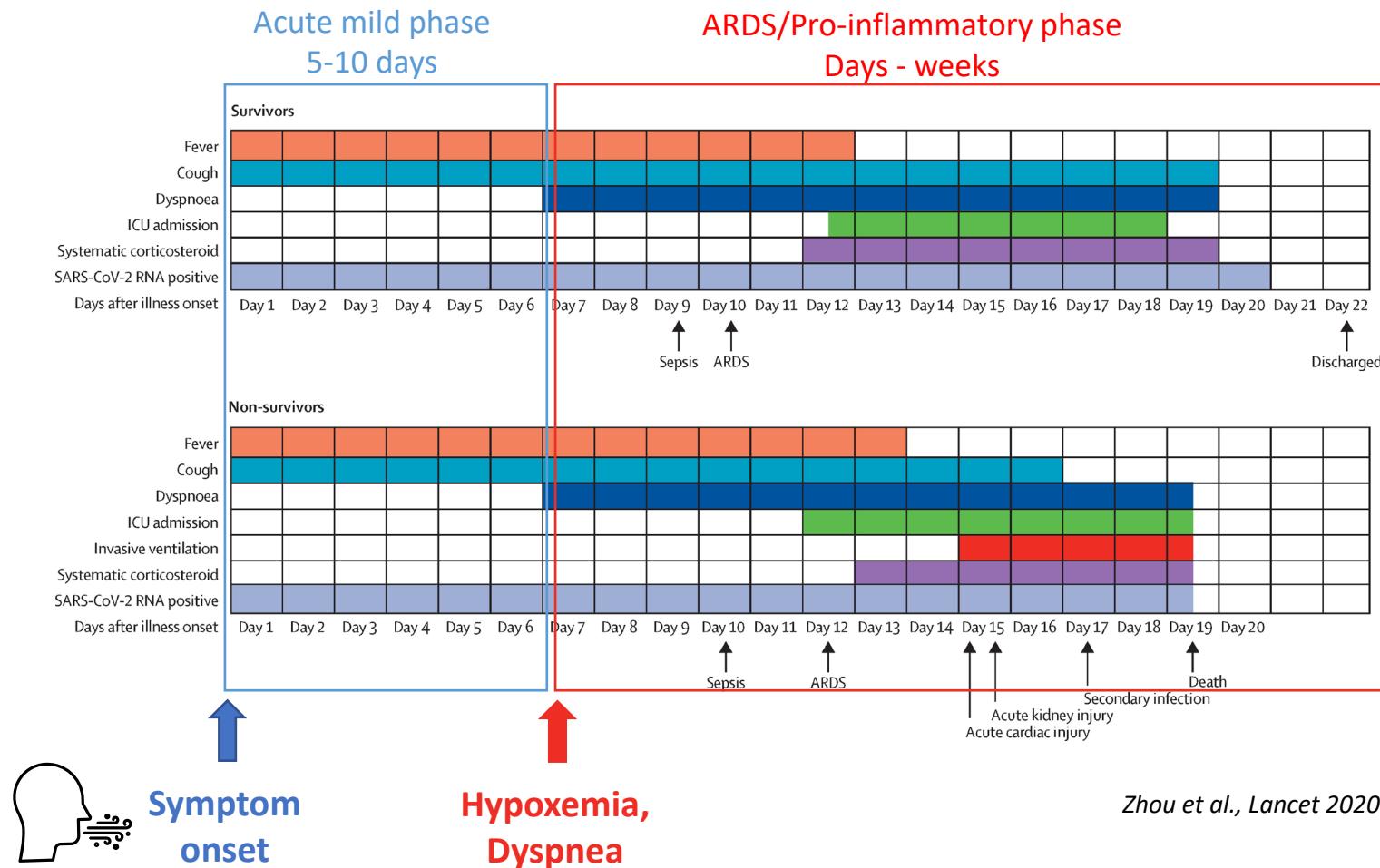


## Outcome 3

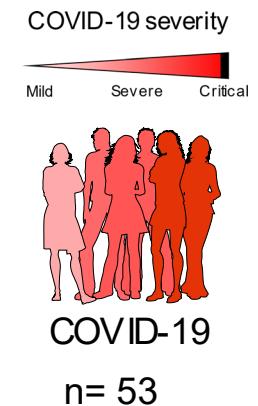
Difference between hospitalized Sars-CoV2-infected patients?



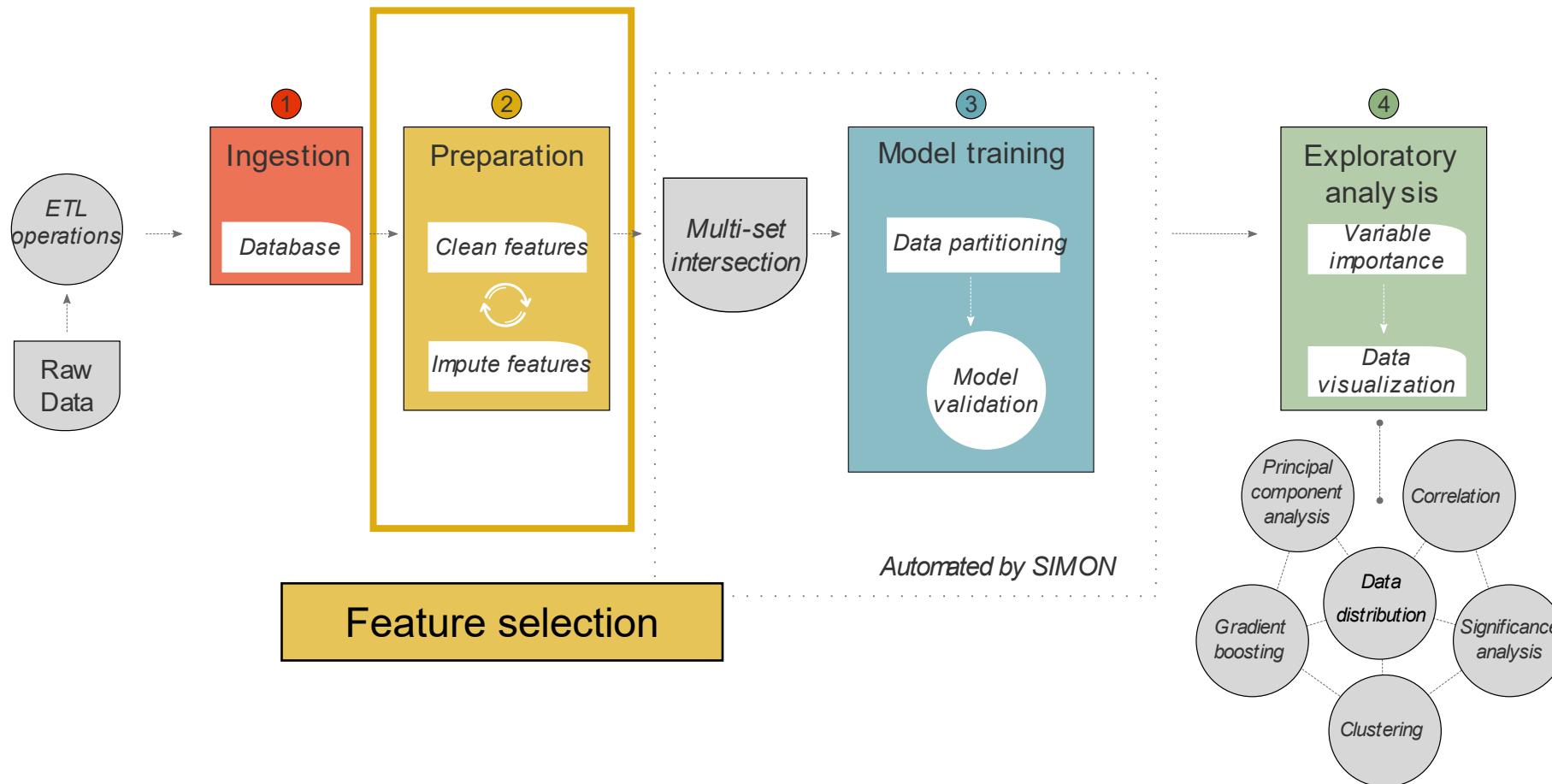
# Knowledge discovery using COMBAT dataset



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

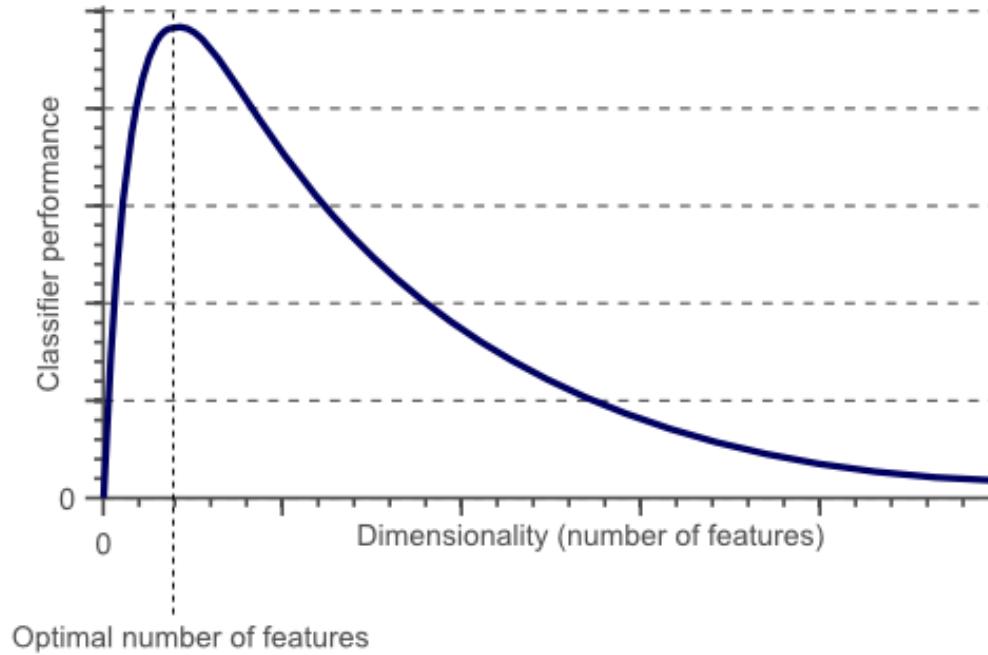
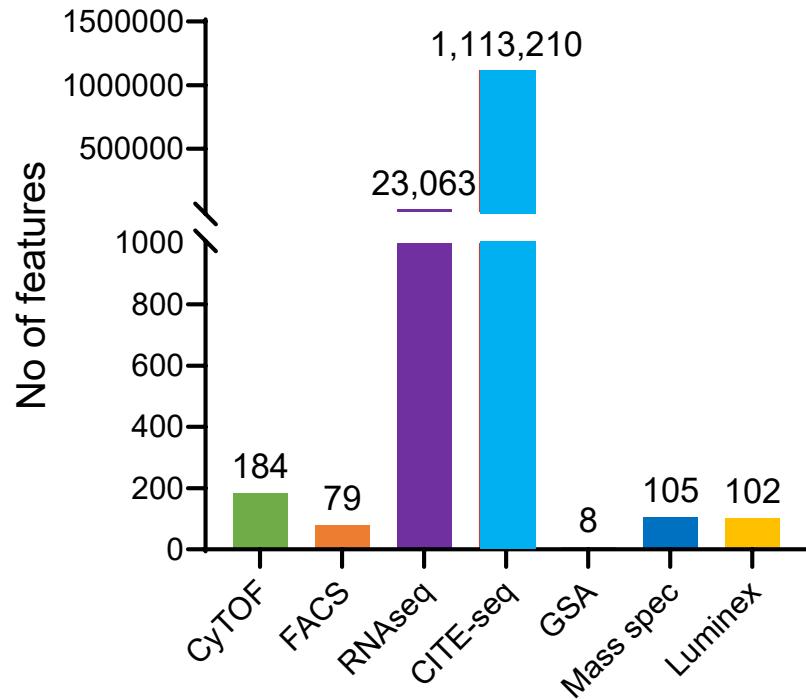


# SIMON - Knowledge data discovery process



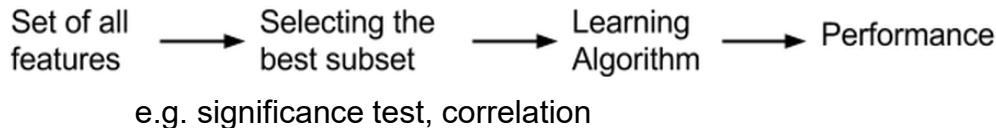
Tomic et al, JI, 2019 & Tomic et al, Patterns, 2021

# 'Curse of dimensionality'

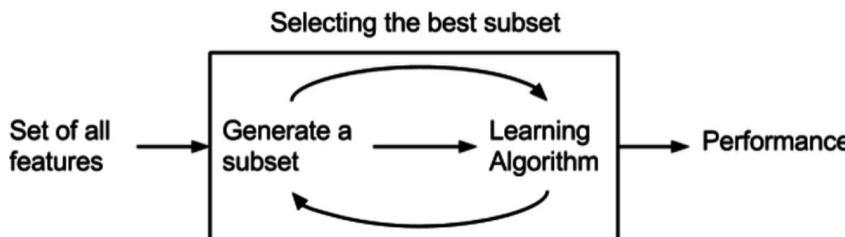


# Feature selection

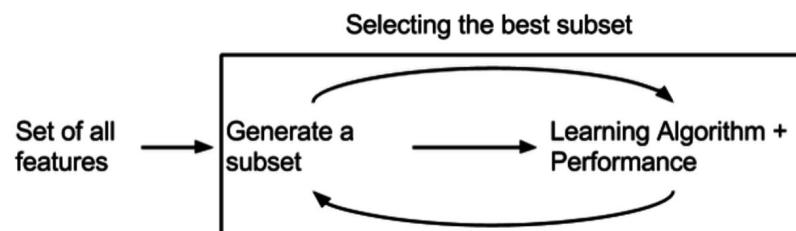
## 1. *Filter methods* - use a proxy measure to score a feature subset



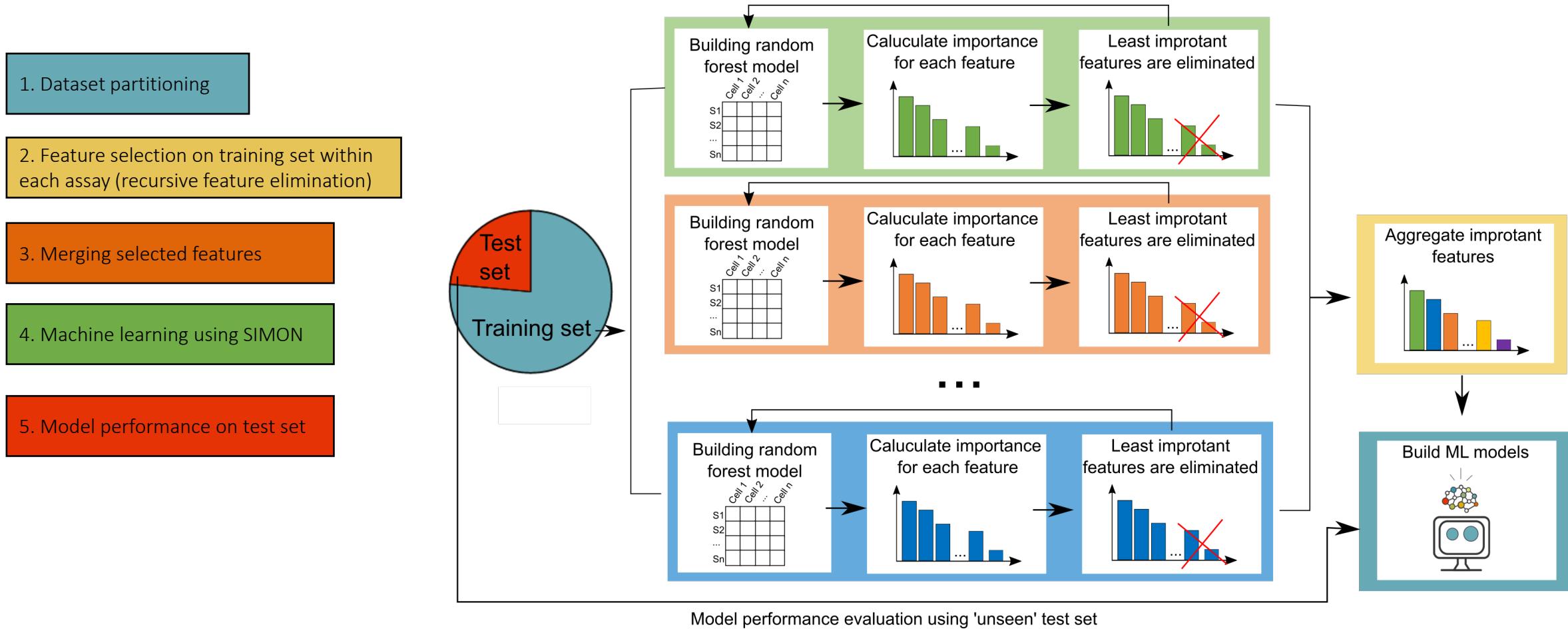
## 2. *Wrapper methods* - use a predictive model to score feature subsets



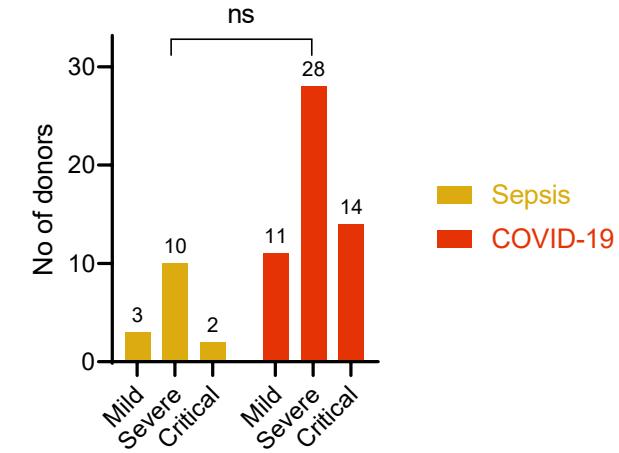
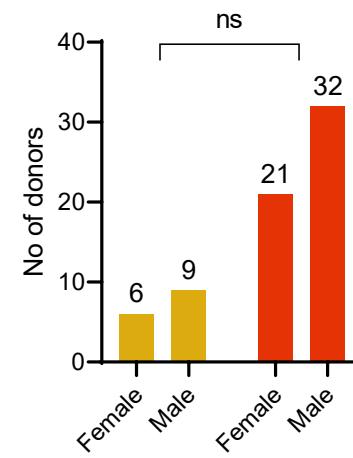
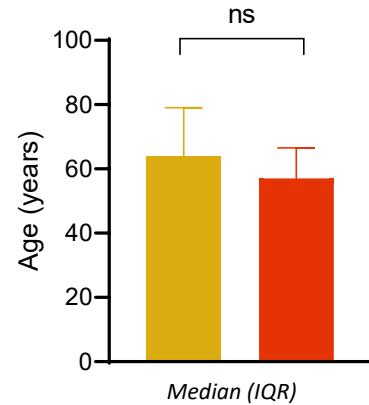
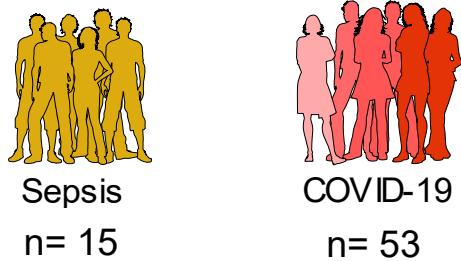
## 3. *Embedded Methods* - catch-all group of techniques which perform feature selection as part of the model construction process (e.g. LASSO)



# SIMON - Feature selection process



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



Demographic data of hospitalized COVID-19 and sepsis patients

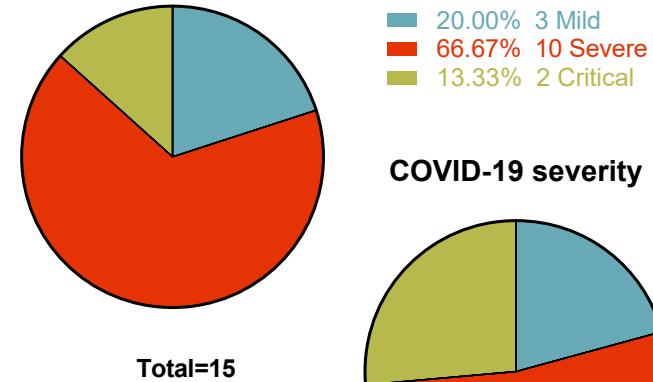
Characteristics	Sepsis (n=15)	COVID-19 (n=53)	P value
<b>Age (y)</b>			0.0969 <sup>a</sup>
Mean ± SD	63.6 ± 17.2	56.6 ± 14.2	
Median (IQR)	64 (52-79)	57 (47.5-66.5)	
<b>Gender</b>			0.3305 <sup>b</sup>
Female (%)	6 (40%)	21 (40%)	
Male (%)	9 (60%)	32 (60%)	
<b>Disease severity</b>			0.1250 <sup>b</sup>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

Abbreviation: IQR, interquartile range

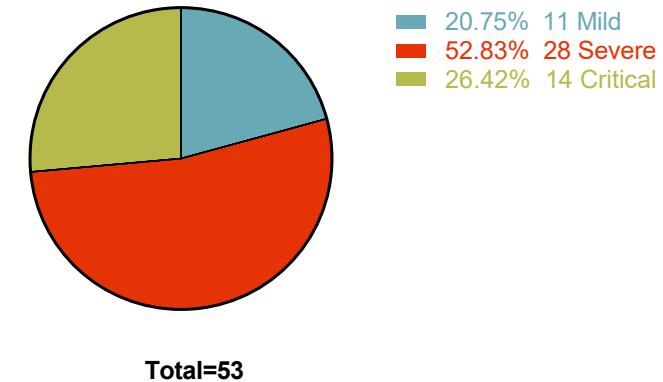
<sup>a</sup>Two-tailed Mann-Whitney test

<sup>b</sup>Two-way ANOVA

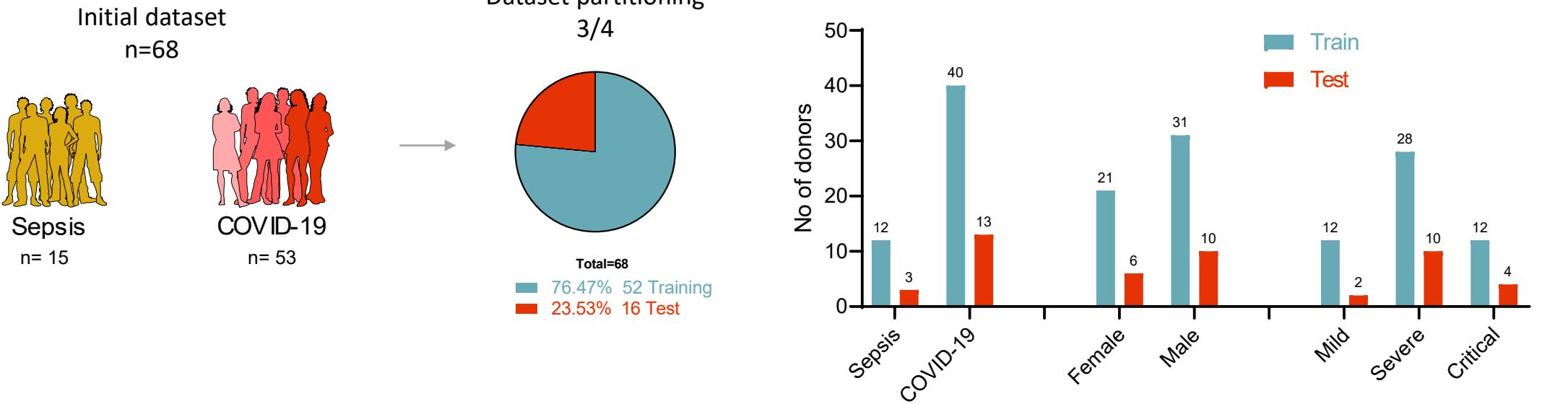
Sepsis severity



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
<b>184 features</b> (subsets frequencies and absolute numbers)	<b>79 features</b> (subsets frequencies, absolute numbers and clusters)	<b>102 features</b> (expression and fluorescence intensity)	<b>105 features</b> (intensity matrix)	<b>23,063 features</b> (counts per million)	<b>1,112,210 features</b> (20,615 genes in 54 cell populations)	<b>8 features</b> (dosage, ABO typing)

Wrapper method using recursive feature elimination

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

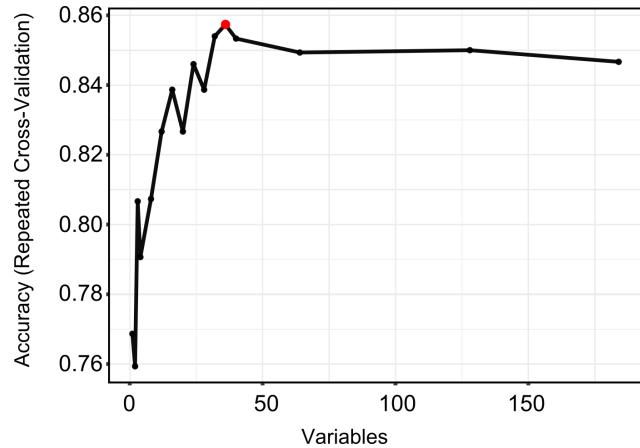
1. Dataset partitioning
2. Feature selection on training set within each assay
3. Merging selected features
4. Machine learning using SIMON
5. Model performance on test set

### CyTOF

**184 features**  
(subsets frequencies and absolute numbers)

36

*36 features from CyTOF dataset selected by the top performing model*

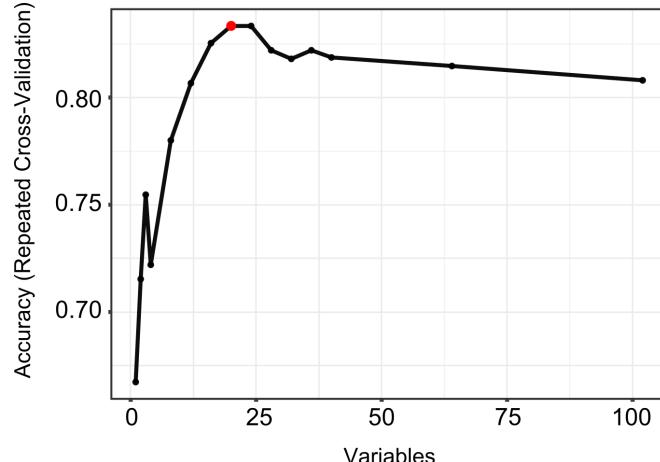


### Luminex

**102 features**  
(expression and fluorescence intensity)

20

*20 features from Luminex dataset selected by the top performing model*

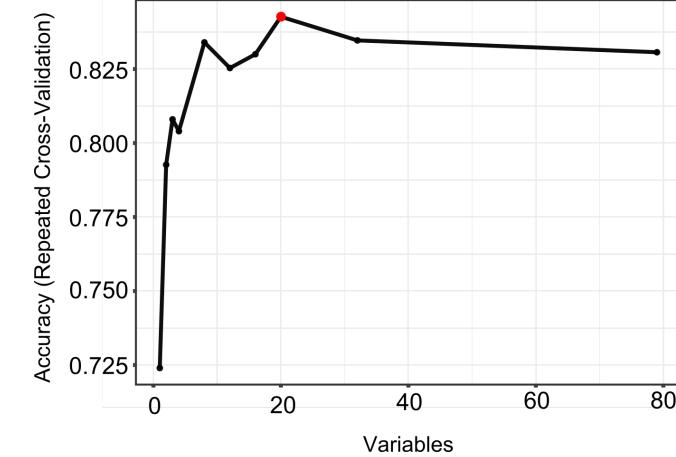


### FACS

**79 features**  
(subsets frequencies, absolute numbers and clusters)

20

*20 features from FACS dataset selected by the top performing model*

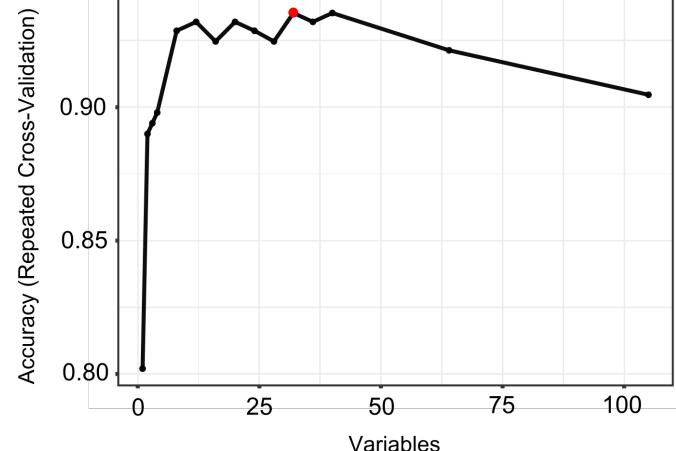


### Mass Spec

**105 features**  
(intensity matrix)

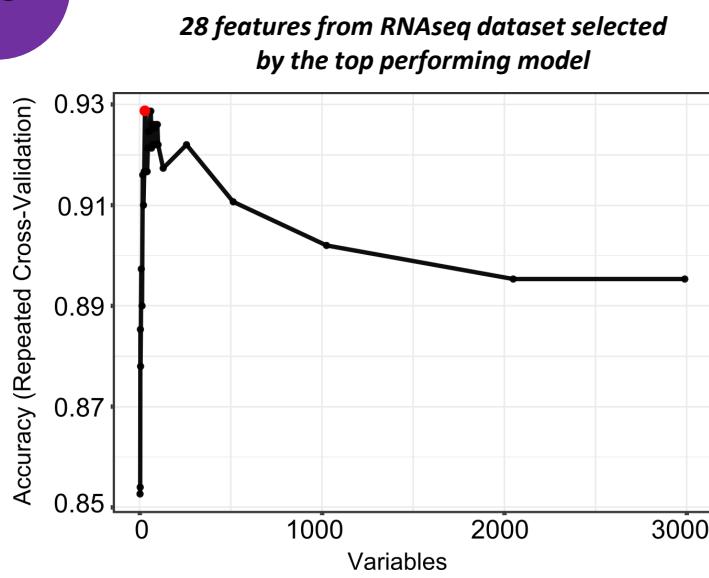
32

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



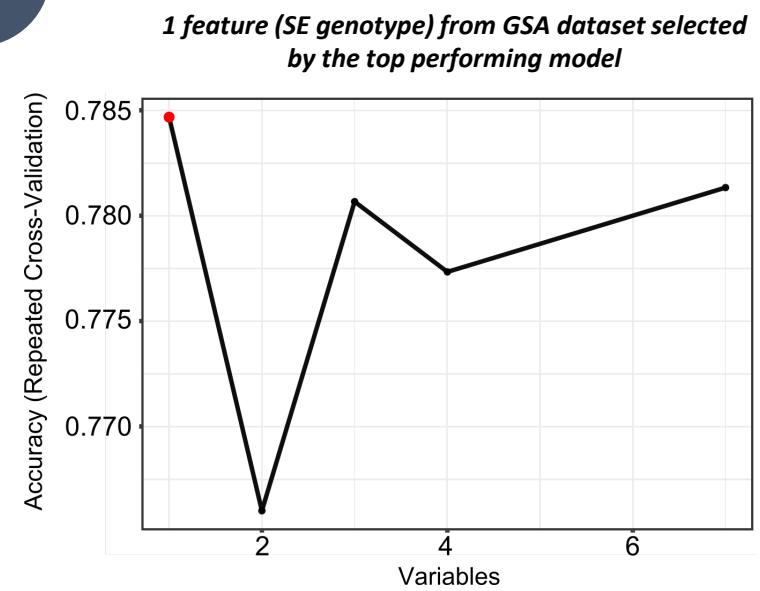
RNAseq
23,063 features (counts per million)
2,989 features (DEGs, FDR<0.05 and FC>1.5)

28

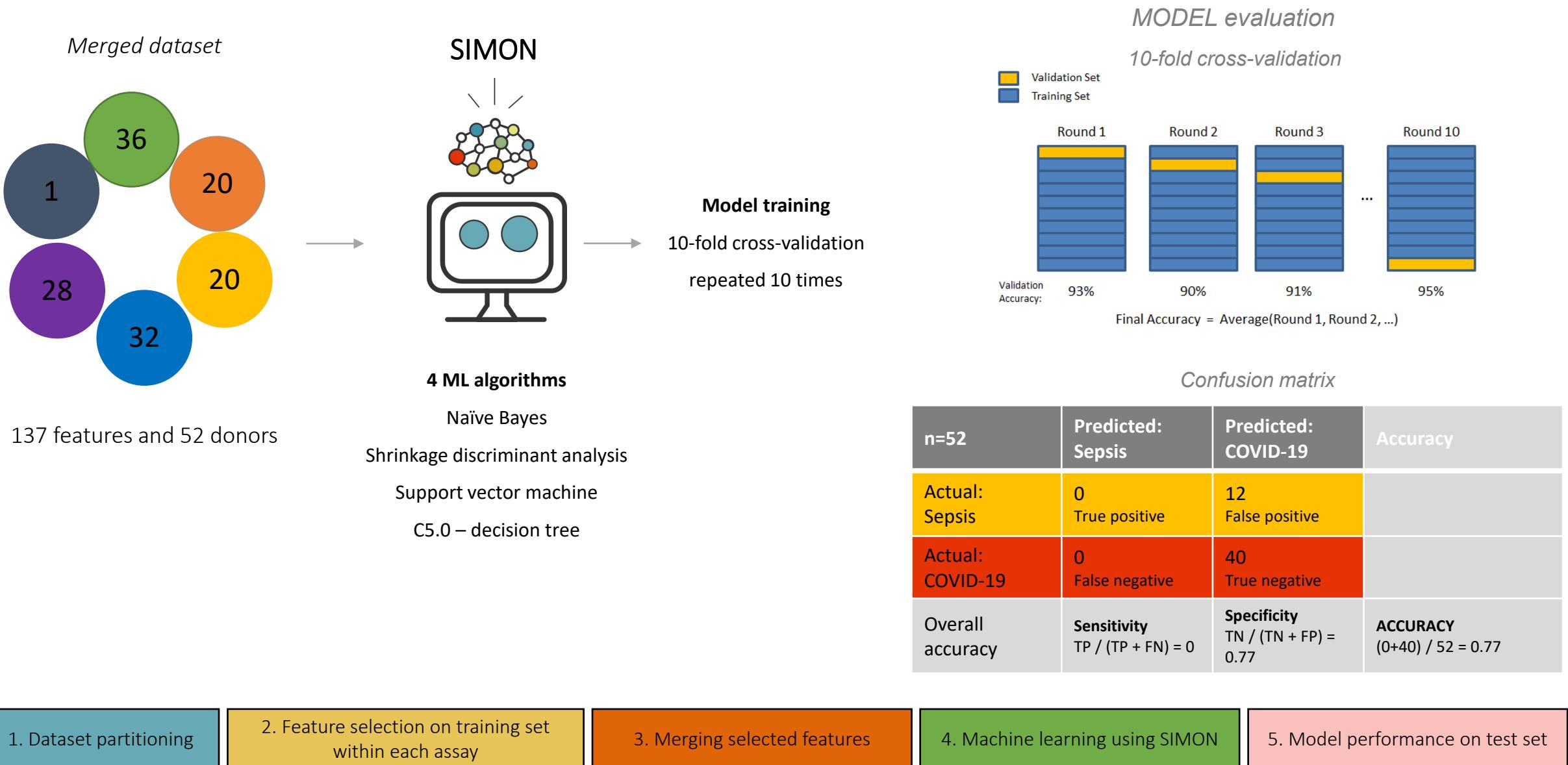


GSA
8 features (dosage, ABO typing)

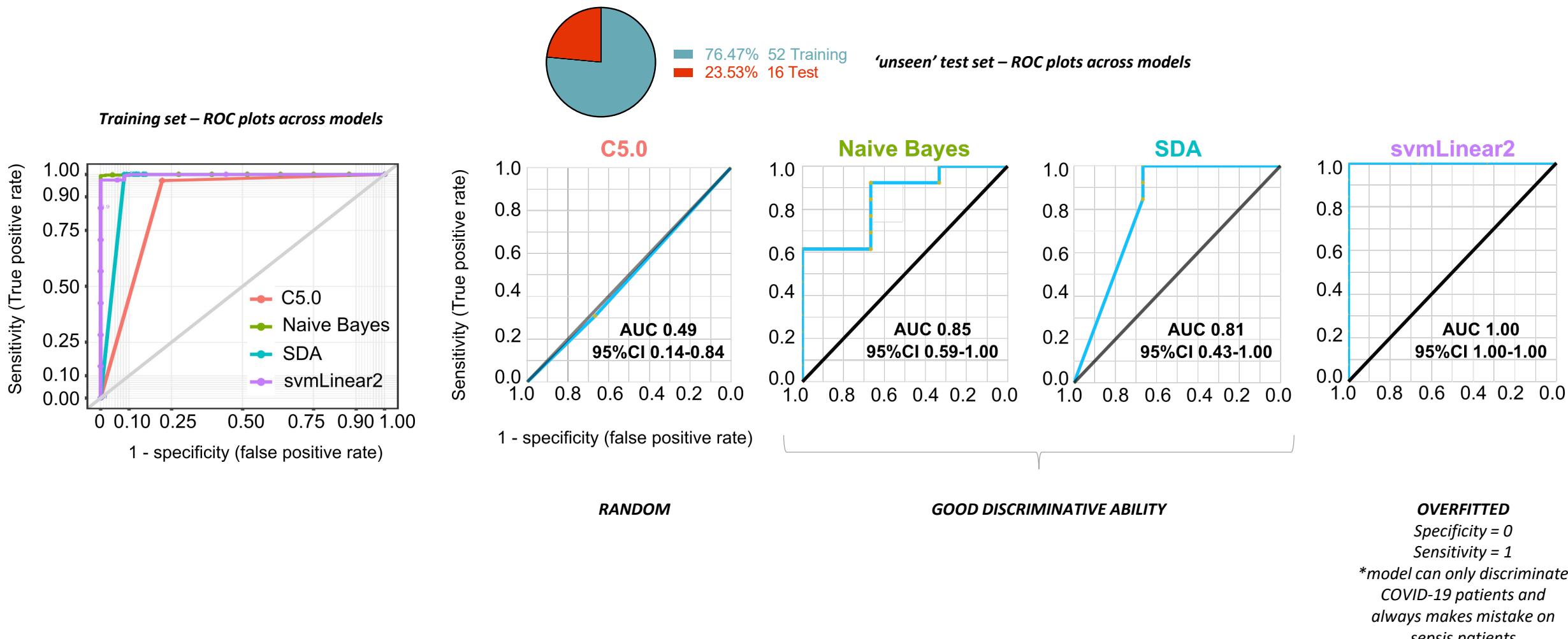
1



# Feature selection process – merging selected features



# Feature selection process – final model



1. Dataset partitioning

2. Feature selection on training set within each assay

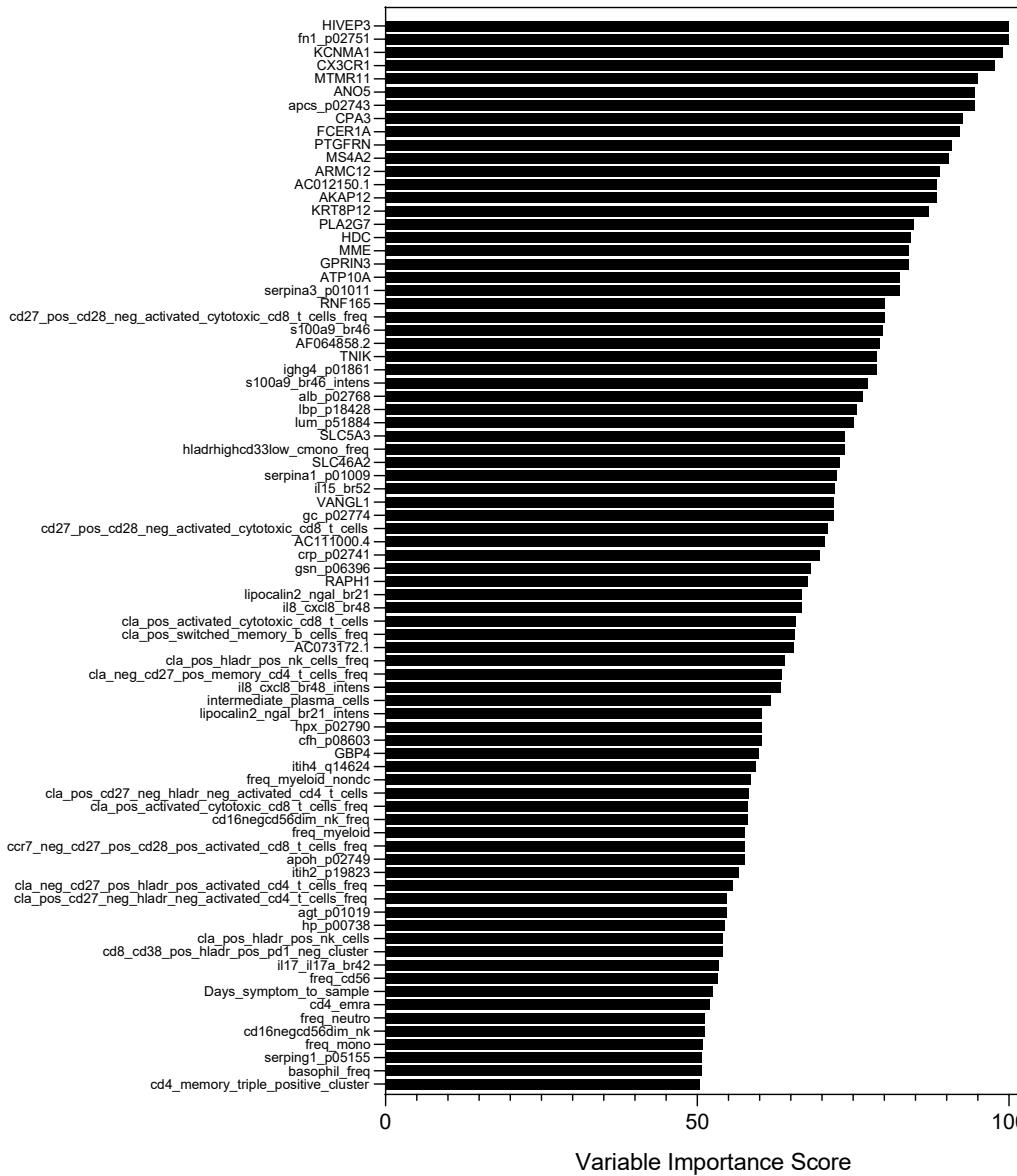
3. Merging selected features

4. Machine learning using SIMON

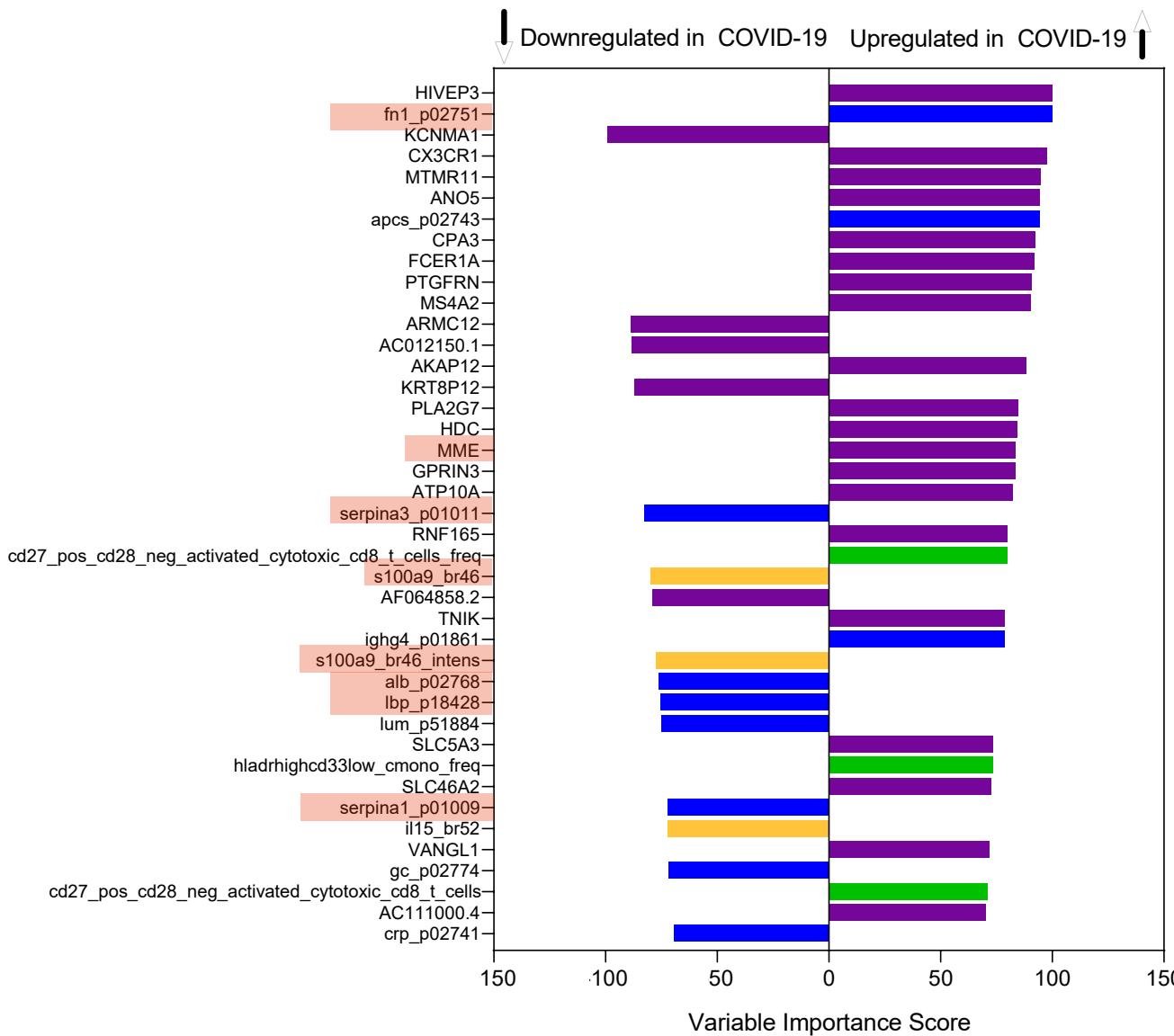
5. Model performance on test set

# 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.

**MME - neprilysin → part of the ACE2 complex**  
Emameh et al, 2020. doi: 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.

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

**Albumin (ALB) → hypoalbuminemia in COVID-19 patients**

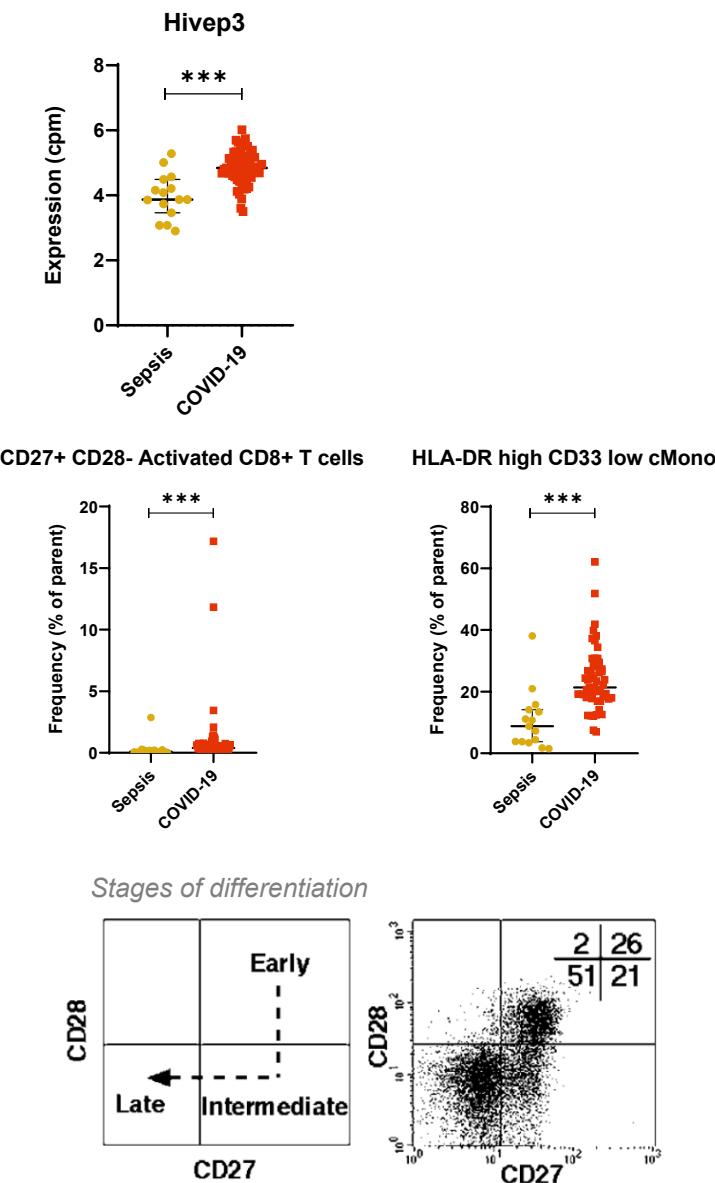
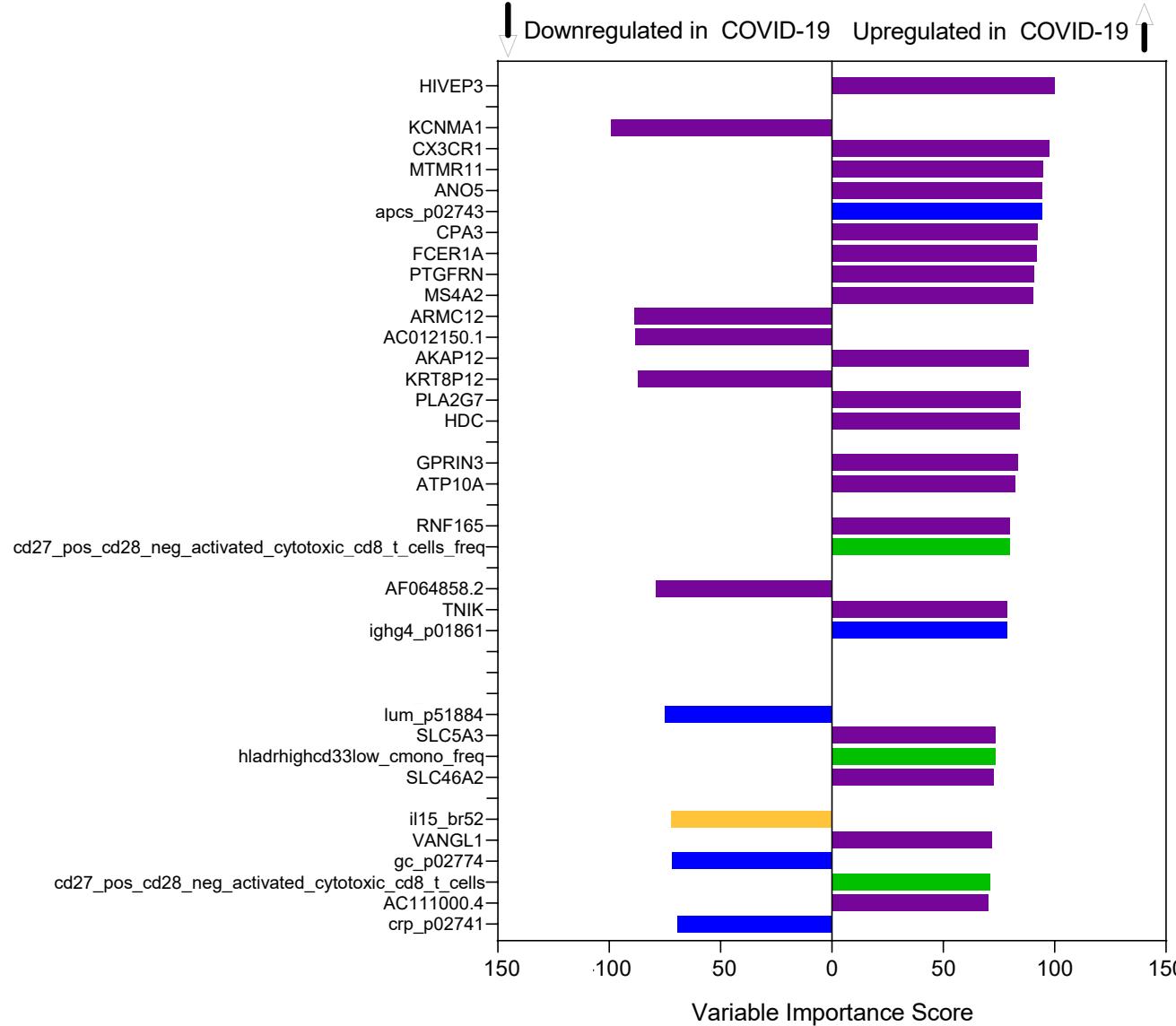
*LBP, LPS binding protein* → increased in COVID-19 patients  
Hoel et al, J Intern Med 2020. doi: 10.1111/joim.13178.

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

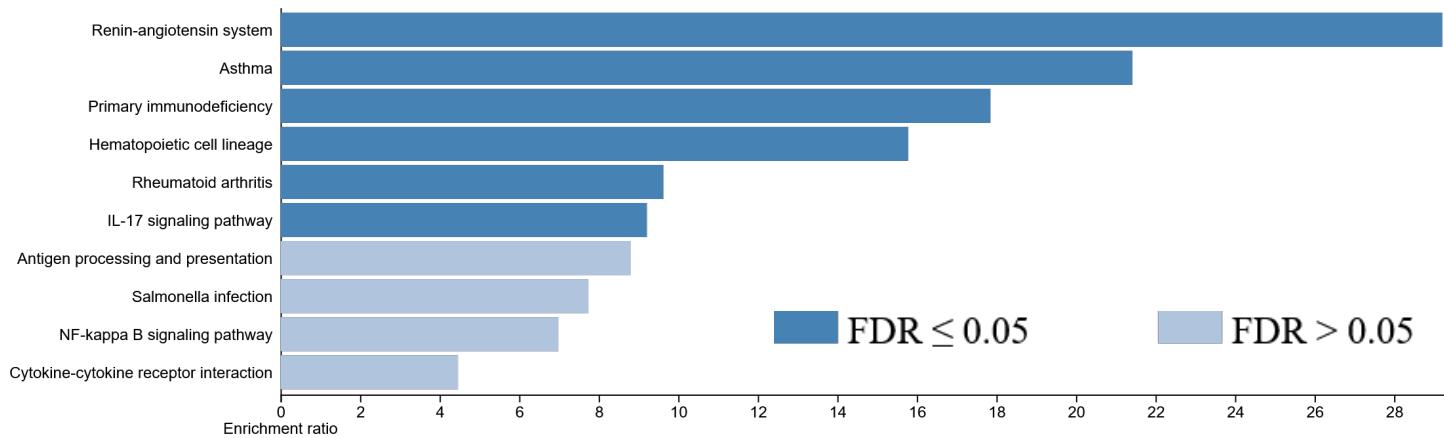
*41 features with score above 69*

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

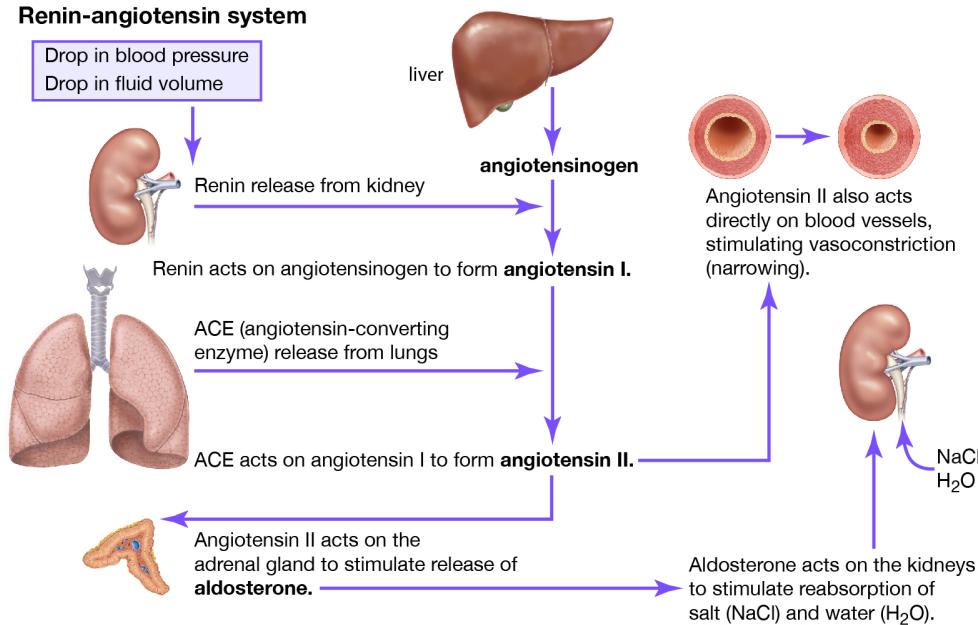
# Pattern recognition in COMBAT using SIMON



# COVID-19 vs sepsis patients – KEGG pathway analysis

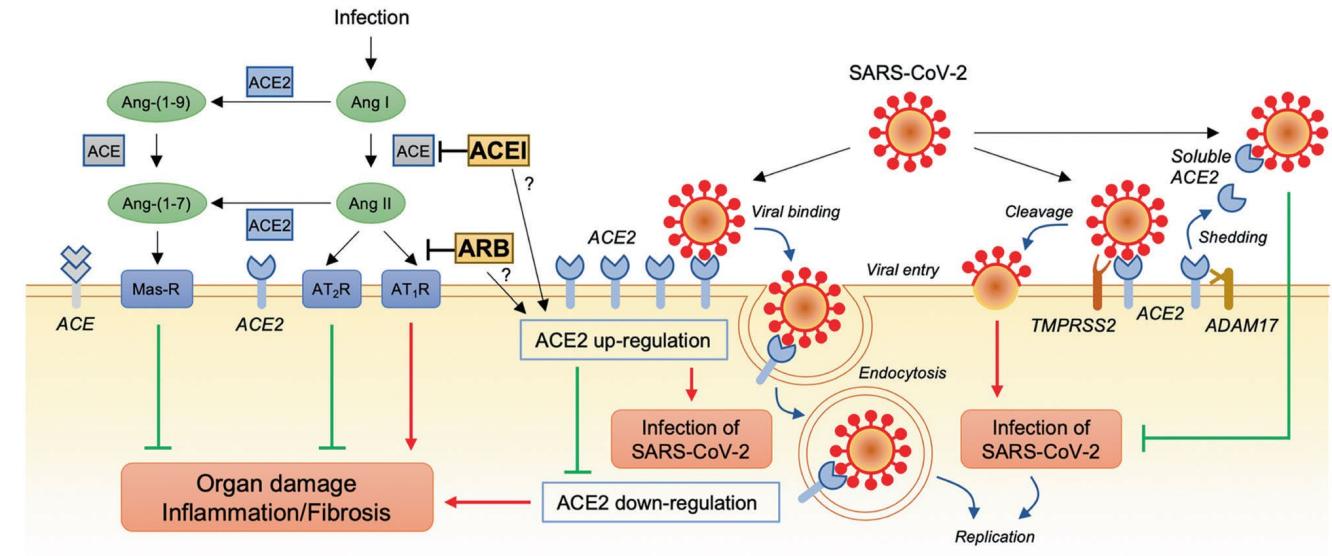


## Renin-angiotensin system



© Encyclopædia Britannica, Inc.

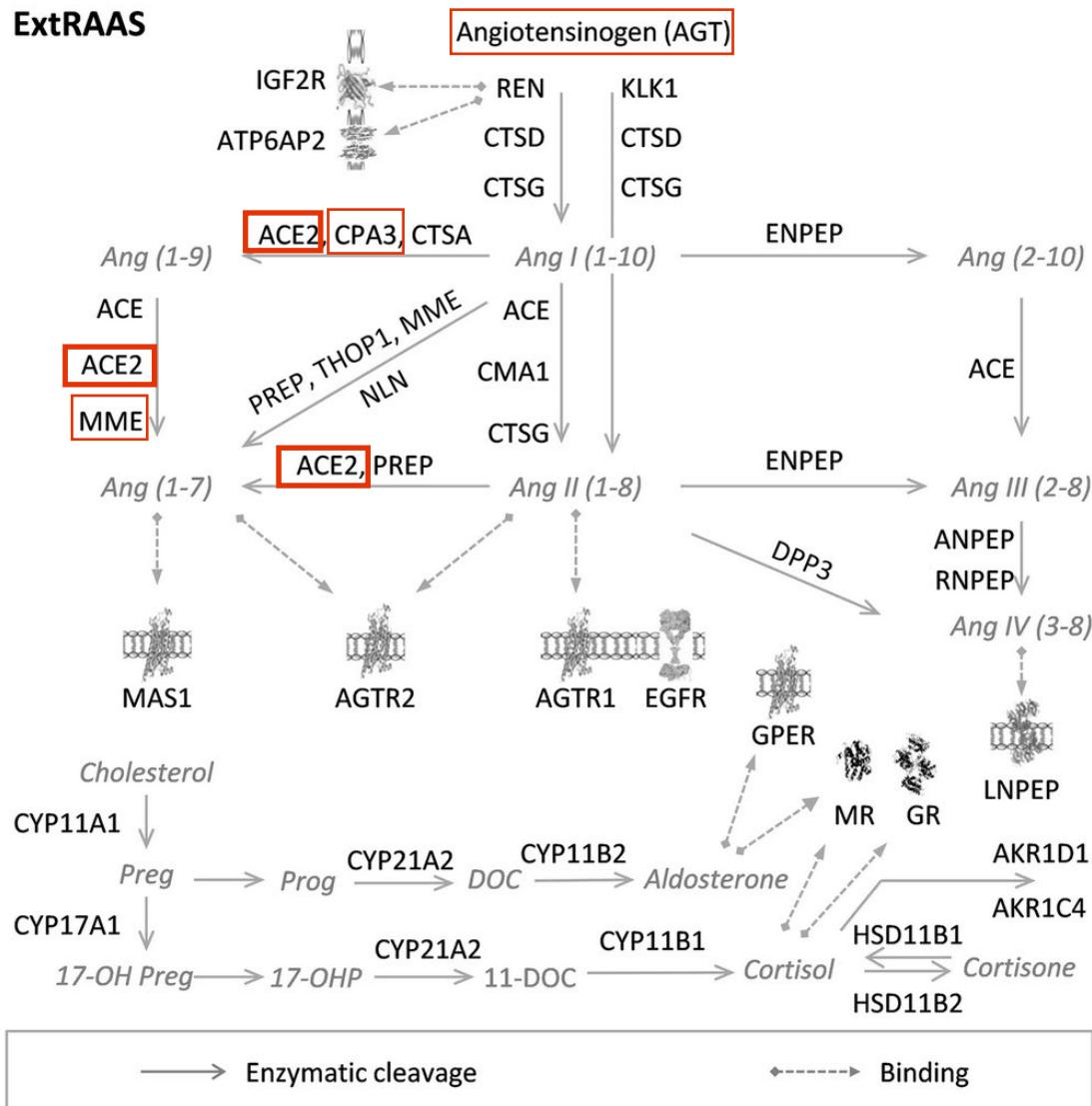
Possible associations of the renin-angiotensin system with SARS-CoV-2 infection and lung injury caused by SARS-CoV-2



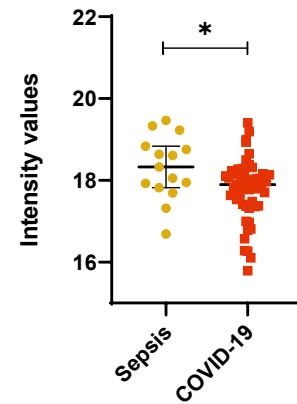
Furuhashi et al, Hypertension Research, 2020, doi: 10.1038/s41440-020-0478-1

# Renin-angiotensin system in COVID-19

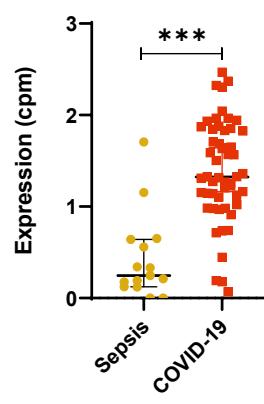
## ExtRAAS



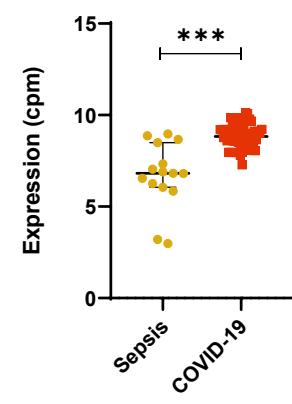
## Angiotensinogen AGT



## CPA3



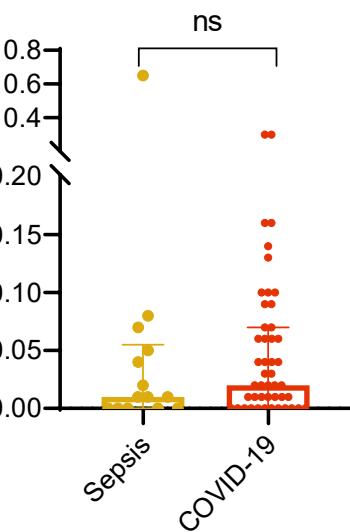
## MME



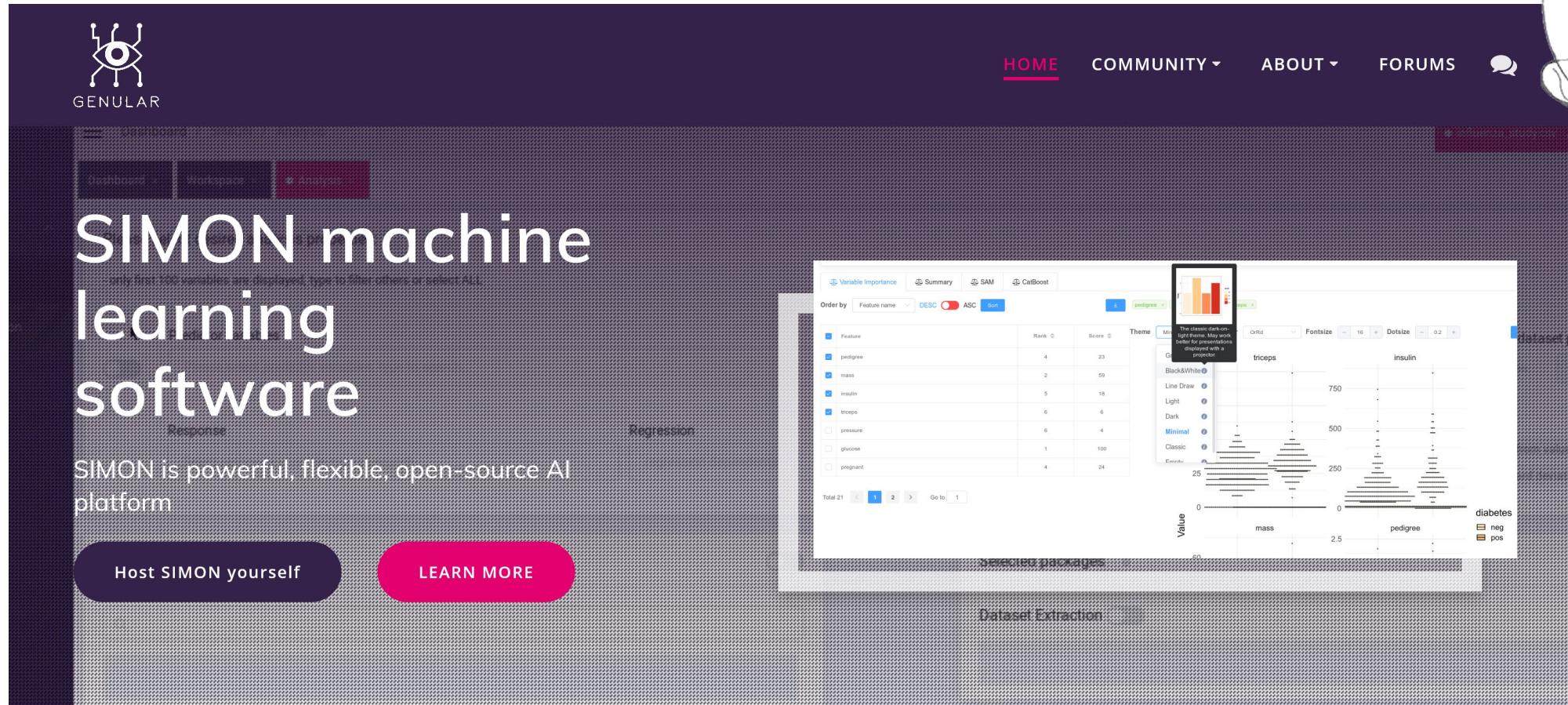
## carboxypeptidase 3 (CPA3)

- 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.



# Join open-source community supporting SIMON!



The image shows the homepage of the SIMON machine learning software. At the top left is the GENULAR logo, which features a stylized eye with three lines extending from it. The main title "SIMON machine learning software" is displayed prominently in large white font. Below the title, a subtitle reads "SIMON is powerful, flexible, open-source AI platform". Two buttons are present: "Host SIMON yourself" in a dark purple button and "LEARN MORE" in a pink button. The background has a subtle grid pattern. At the top right of the page is a navigation bar with links for "HOME", "COMMUNITY", "ABOUT", "FORUMS", and a speech bubble icon. A large screenshot of the SIMON interface is centered on the page, showing a "Variable Importance" table and a "Dataset Extraction" plot.

**SIMON machine learning software**

SIMON is powerful, flexible, open-source AI platform

Host SIMON yourself    LEARN MORE

Variable Importance    Summary    SAM    CatBoost

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

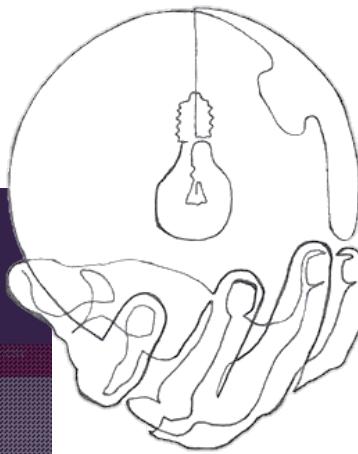
Dataset Extraction

The classic dark-on-light theme. May work better for some environments.  
The classic dark-on-light theme. May work better for some environments.

Value

diabetes

neg pos



Check out SIMON at [genular.org](http://genular.org)



GENULAR

HOME

COMMUNITY ▾

ABOUT ▾

FORUMS

# SIMON Knowledge Base

## Have a Question?

Search the documentation...

Search

### Installation

📄 Installation Quickstart

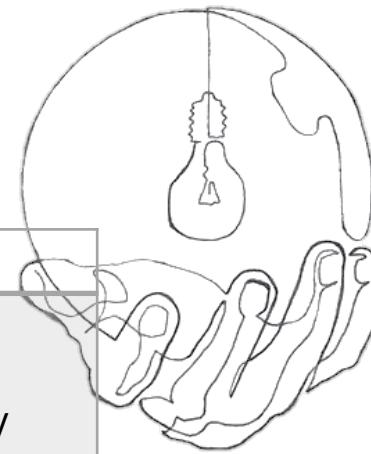
### Machine Learning

📄 How to perform SIMON analysis?

📄 Instruction videos



# 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](http://genular.org)

