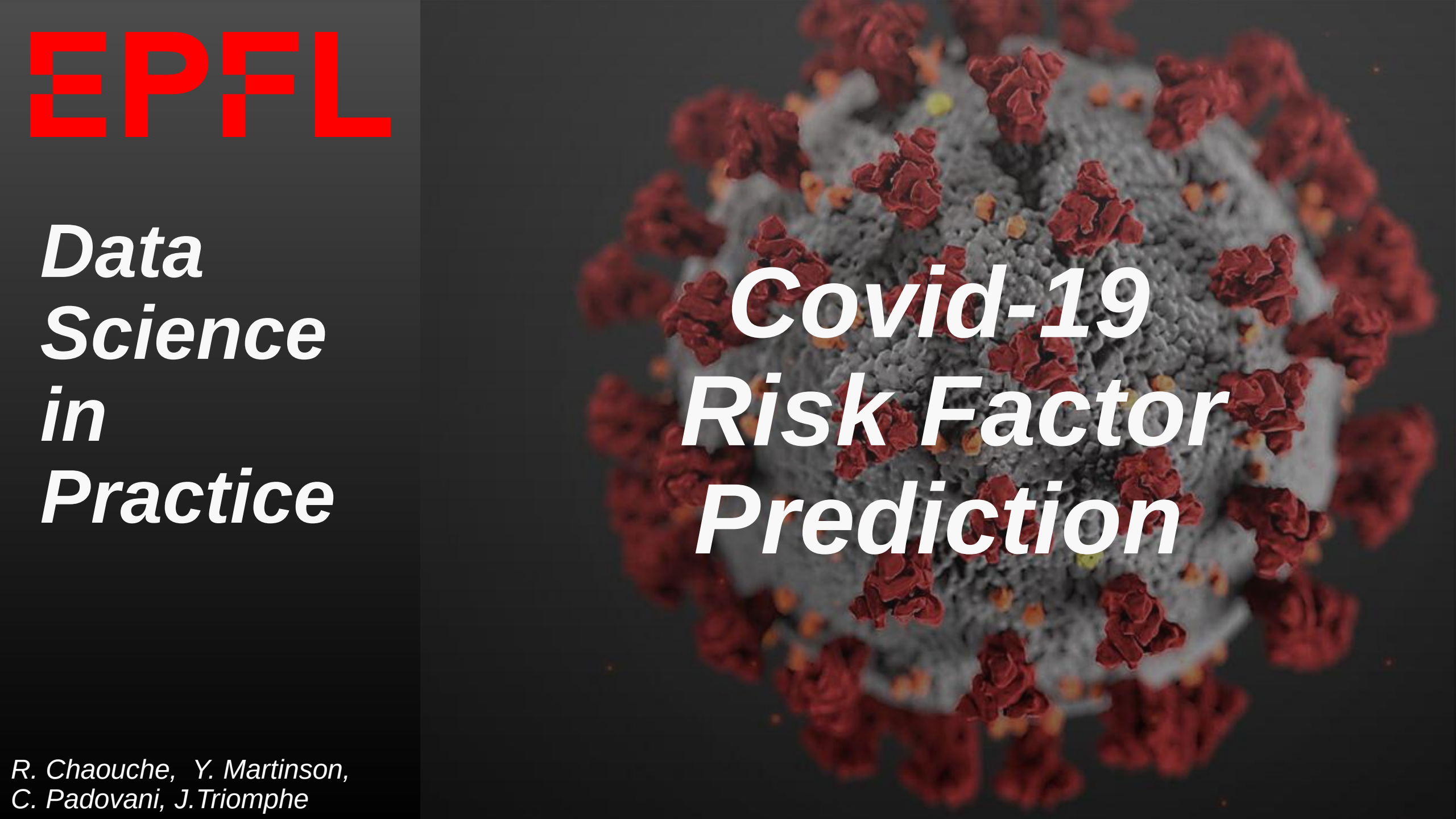
EPFL

Data Science in Practice



Covid-19 Risk Factor Prediction

R. Chaouche, Y. Martinson, C. Padovani, J.Triomphe

Context

Introduction

I. Problem description

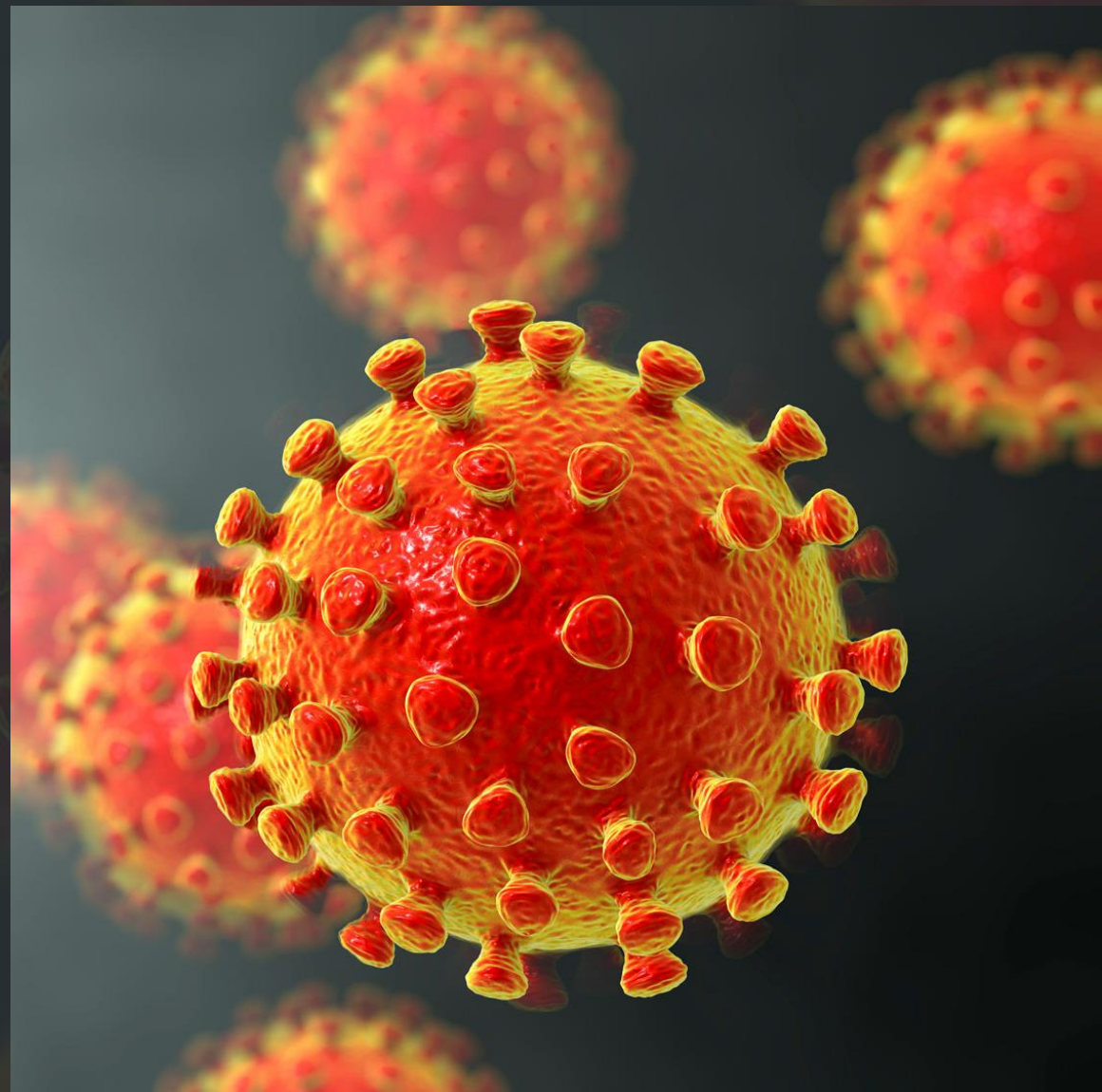
II. Data Management

III. Network analysis

IV. Random forest regression

V. Interpretations

Conclusion



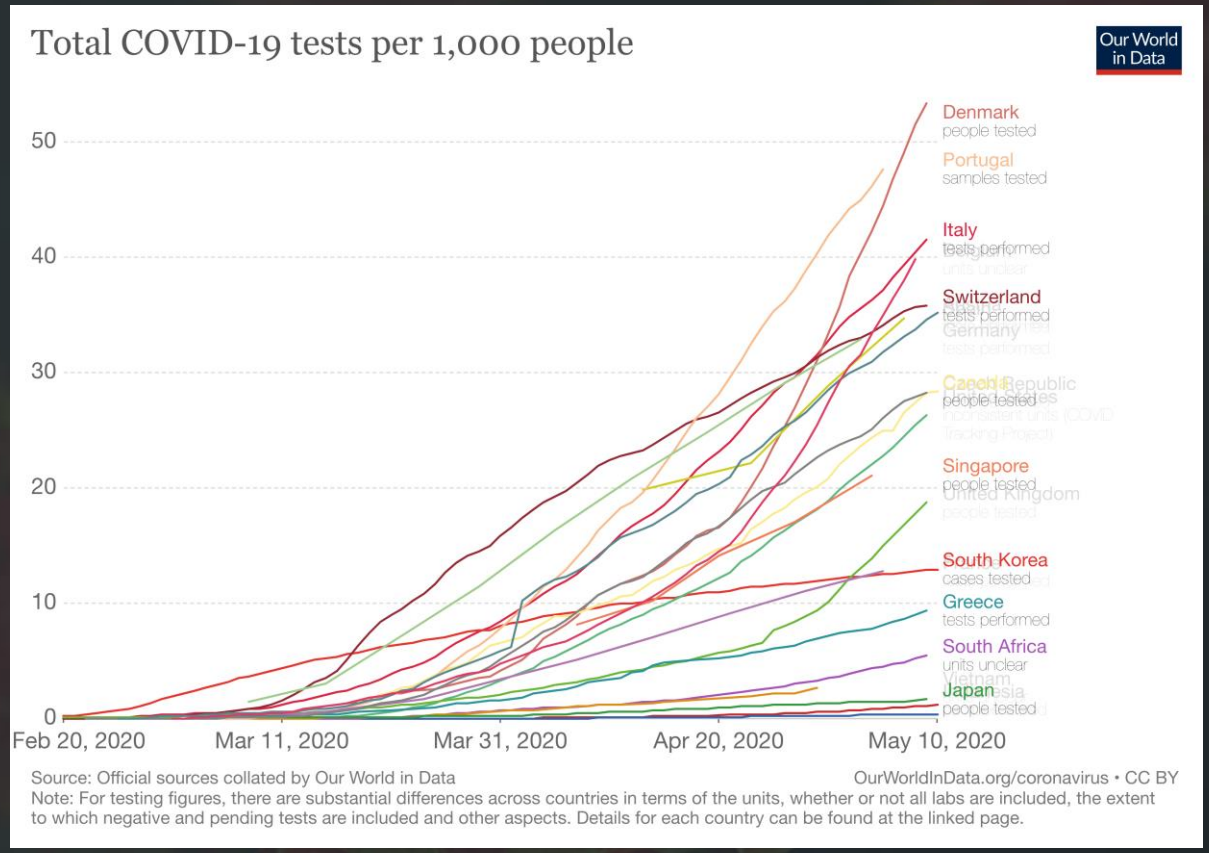

Introduction

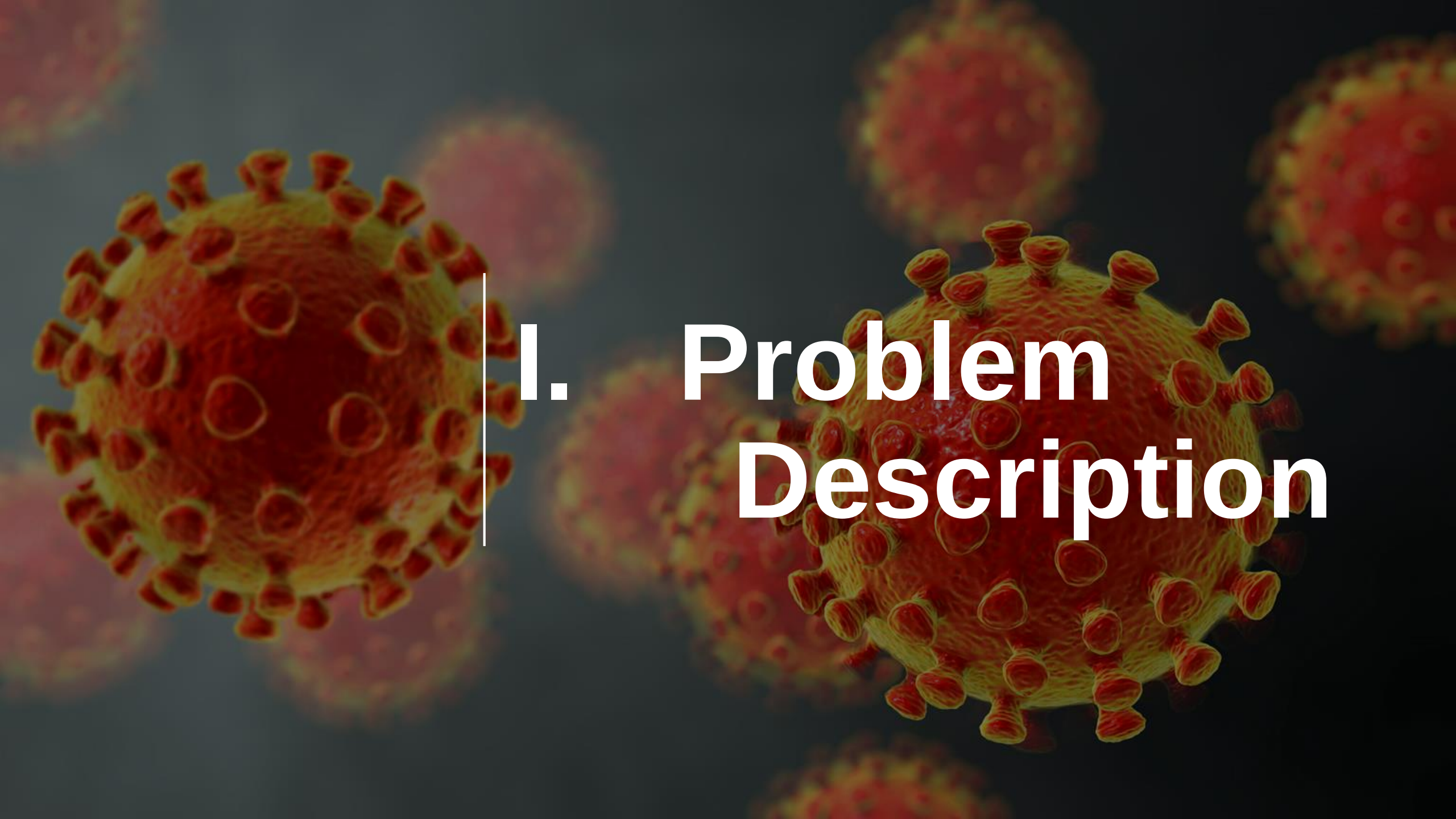
Introduction

→ **Virus has been spreading worldwide**
- 4 million confirmed cases
- Over 265,000 deaths

→ **Global Lock – down**
- Shut-down of large pans of the economy
- Long-term implications

→ **Tests to target SARS-CoV-2**





I. Problem Description

Understanding the need

→ **Health issues**

- High transmissibility and more than 265,000 deaths (06 May)
- Risk factors diseases known but poor literature about drug targets*

→ **Expensive testing campaigns**

- $100 /genome tests
- $30 million for testing Switzerland (35 test/1000 habitants)

→ **Tests do not have an optimal efficiency yet**

* Specific molecule , usually a protein, to which the drug binds to produce its effect

Challenges

→ **Mapping** a network of the major risk factor diseases for COVID-19

→ **Create a model** to identify which targets of those risk factor diseases are more likely to be associated to COVID-19

Strategy

→ **Existing project from Kaggle**

Predicted relation between diseases and COVID-19 with a score system

→ **Our project**

Predicts relation between the targets of those diseases and the COVID-19

Expected outputs and outcomes

→ **Accurate network**

- Will help scientist figure out gene families.
- Expected communities could focus on the most represented genes.

→ **Accurate prediction of targets to pay attention on**

- Will help doctors to avoid additionnal infection due to COVID-19
- Will help to create more accurate tests

II. Data Set

Initial Data Set: Kaggle project (1)

→ CORD-19 Dataset challenge

59,000 scholarly articles, including over 47,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses created by The White House and a coalition of leading research groups

→ Mondo Disease DataBase

Aggregates and merges genetic associations curated from both literature and newly-derived loci1 from UK Biobank

→ Open Targets DataBase

Semi-automatically constructed ontology that merges in multiple disease resources to yield a coherent merged ontology

Initial Data Set: Kaggle project (2)

→ **Connects diseases**

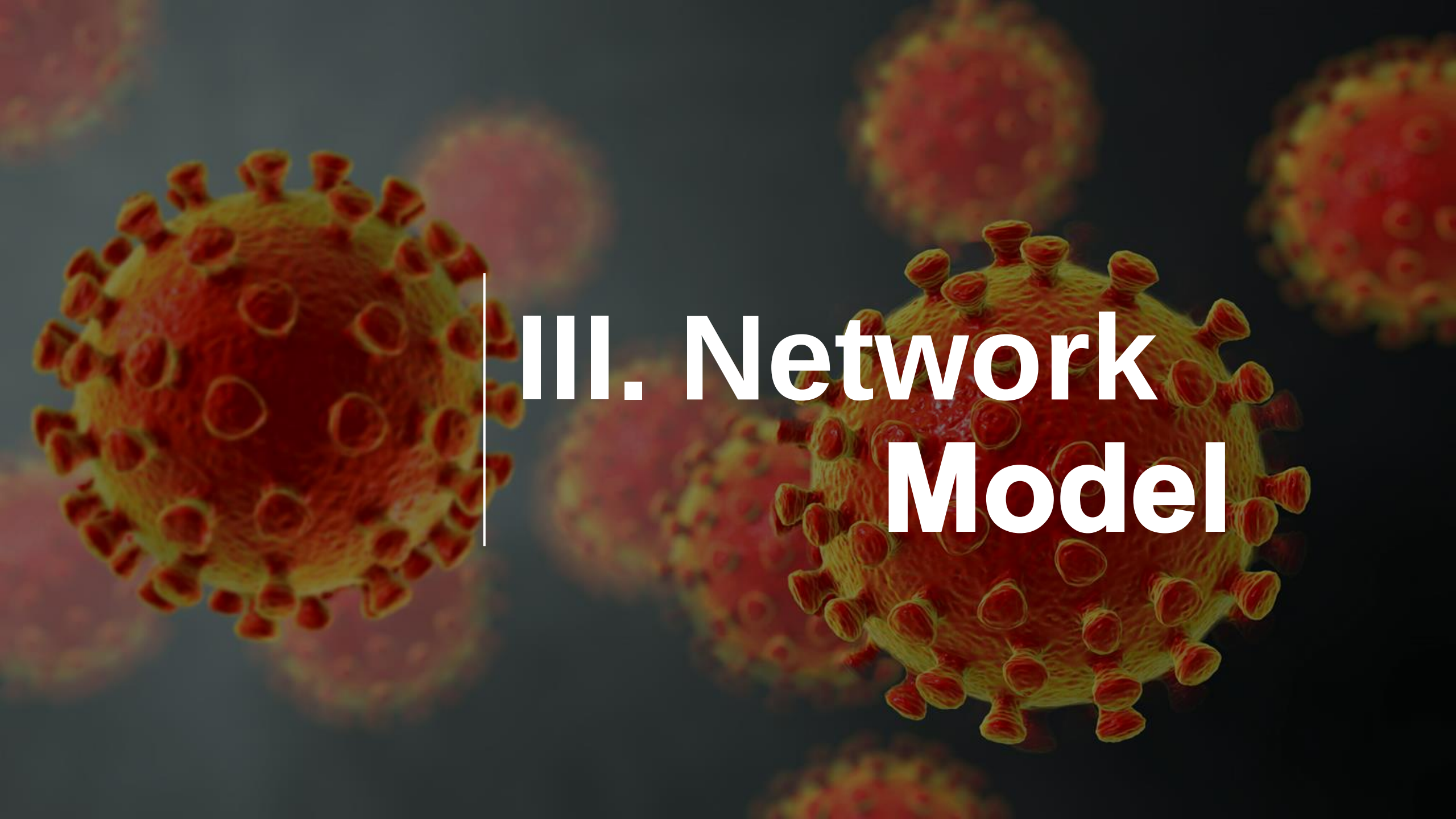
- With other diseases
- With drug targets

→ **Association scores between diseases and COVID-19**

Data Reconstruction

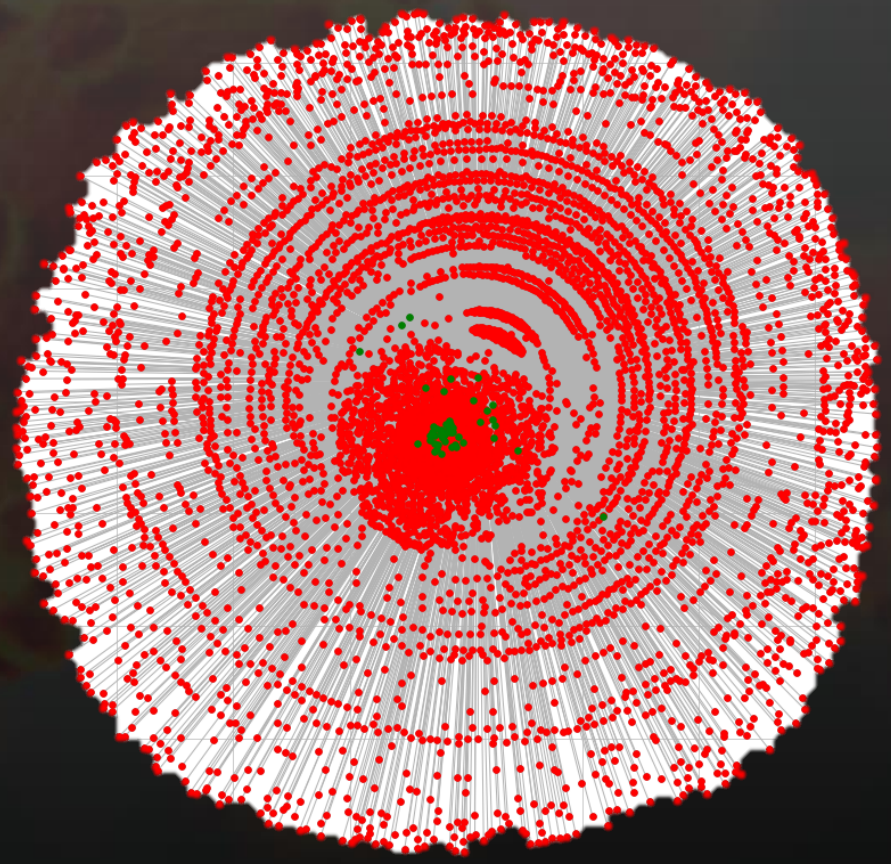
→ **From the output of the previous study**

- Combine disease-target association score with disease-COVID-19 association score
- Leverage the Open Targets Platform



III. Network Model

Model



→ **Dataset I**

- **Nodes: Diseases, Genes**
- **Edges: weighted score of gene on disease**
- **Goal: Centrality ranking**
- **Output: Overall Ranking**

Model

→ **Centralities: Degree, Betweenness, Closeness, PageRank**

→ **Scoring: Point ranking system for each node**

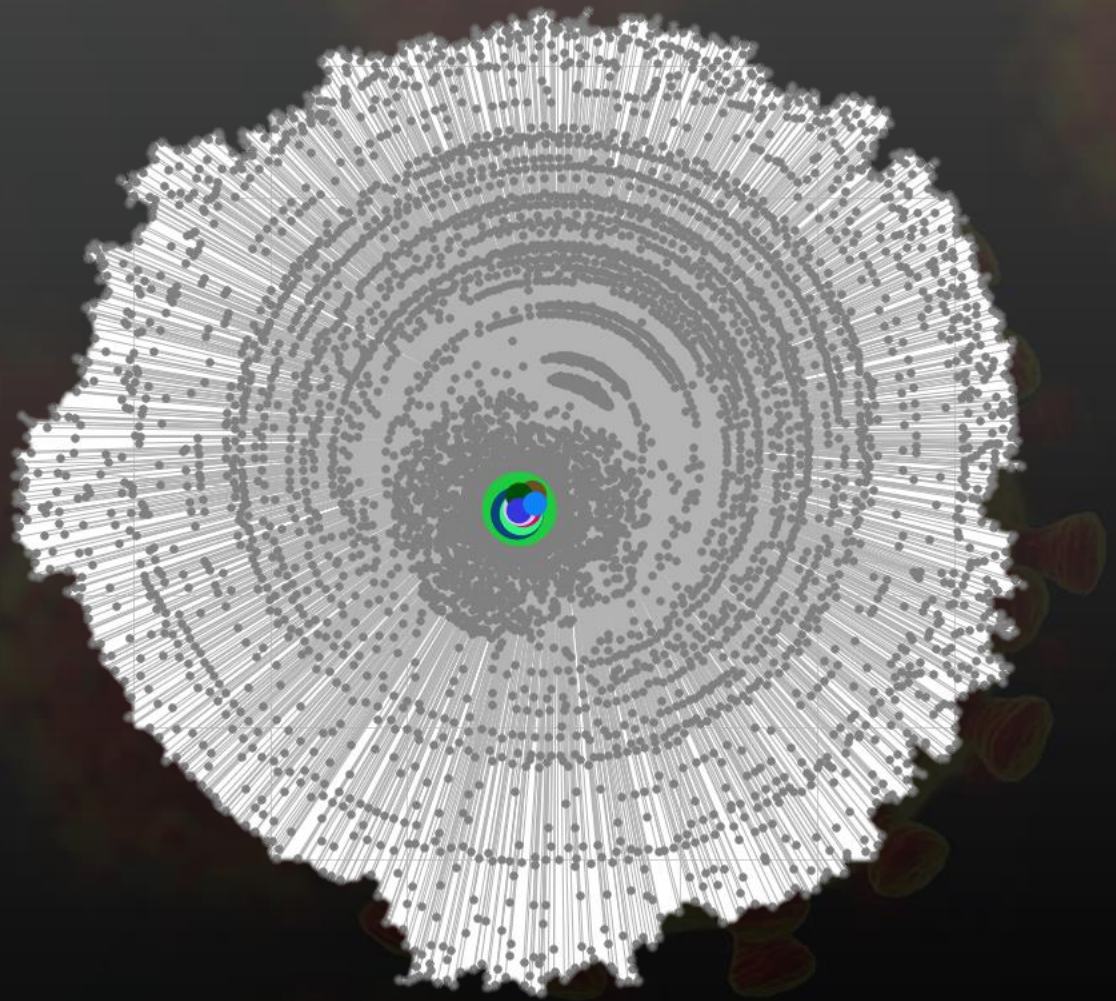
→ **Positive: Finds key players in the network, robust**

→ **Drawbacks: No Covid infos, unexpected outputs**

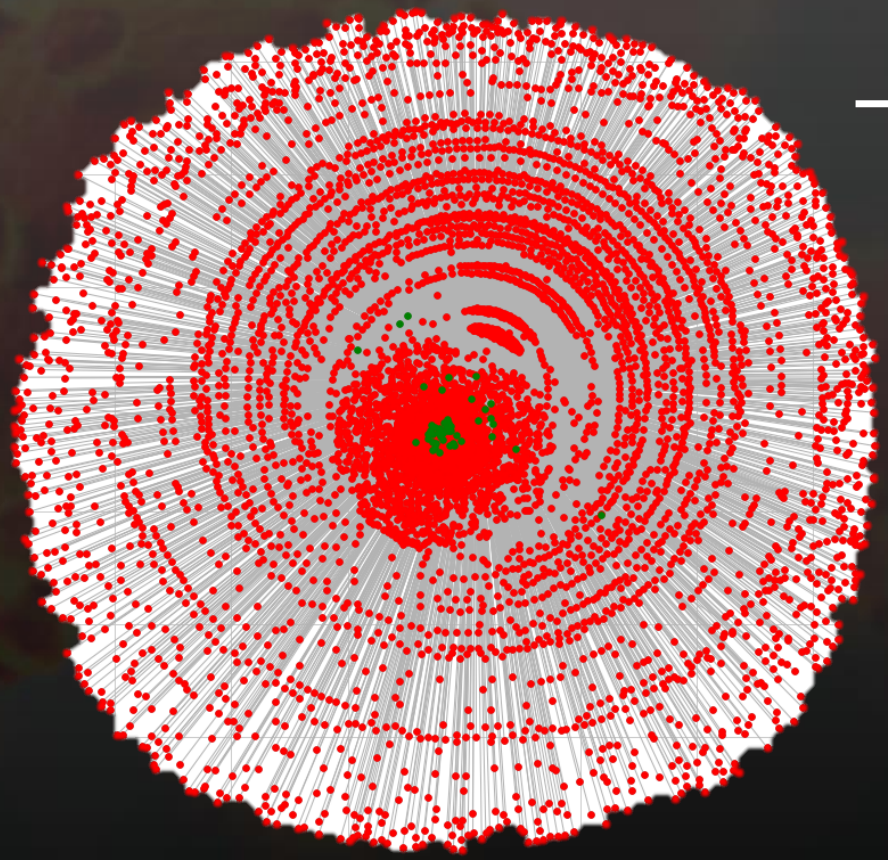
Results

→ **Very centered, close to gene locations**

1. **Bone Disease**
2. **Psychiatric Disorder**
3. **Biological process**
4. **Protein measurement**
5. **Diabetes Mellitus**



Model

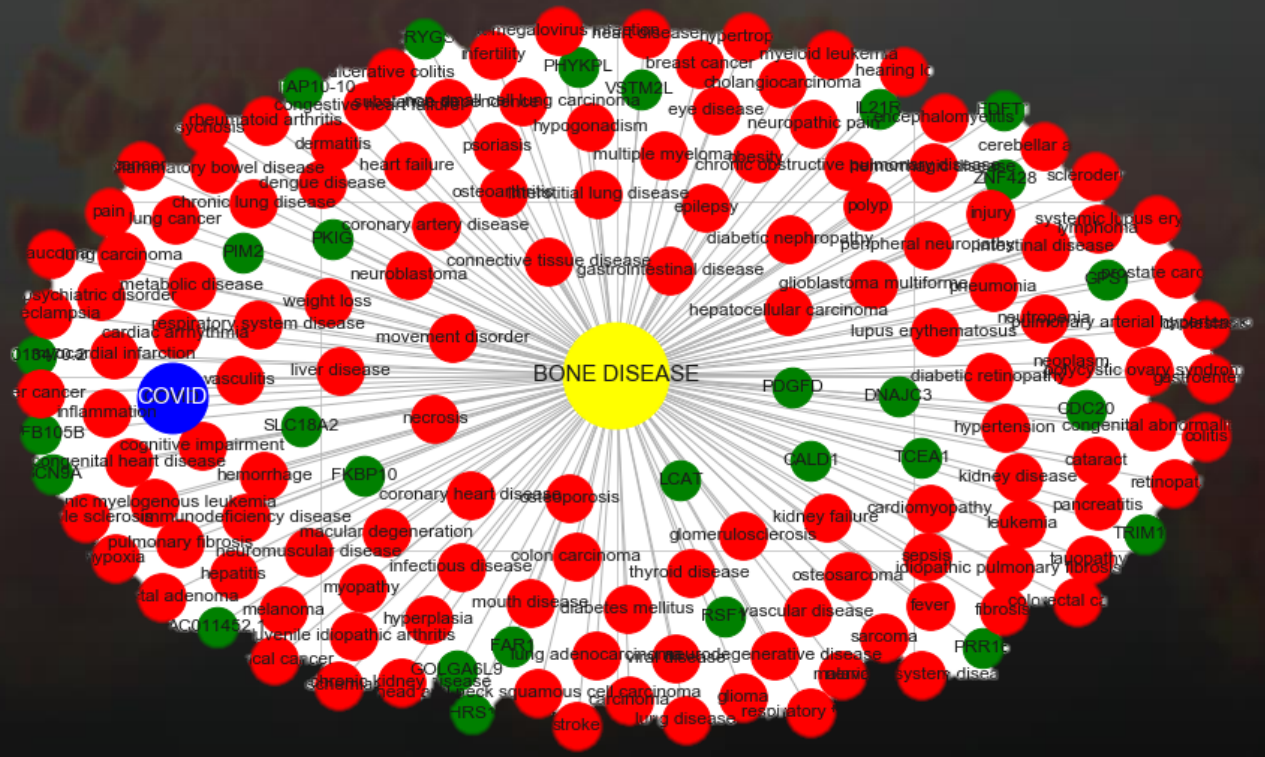


→ **Dataset II**

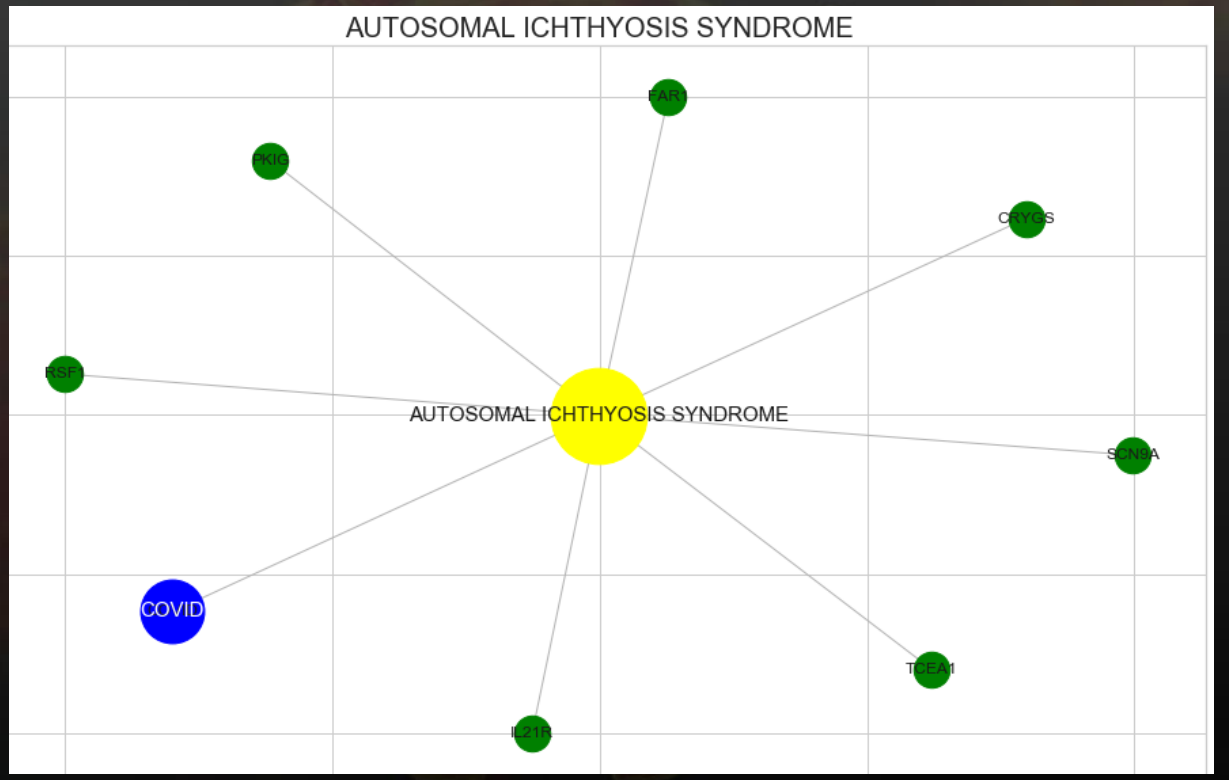
- **Nodes: Diseases, Genes**
- **Edges: weighted score of gene on disease + disease to disease**
- **Goal: Central nodes own network understanding + COVID-19 scoring comparison**

Results

1. Overall Ranking

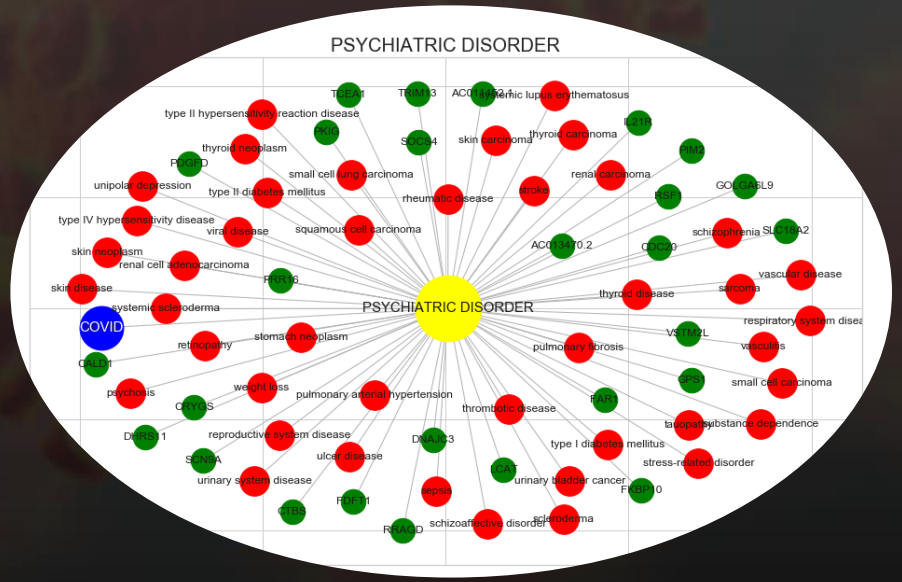


1. Covid Ranking

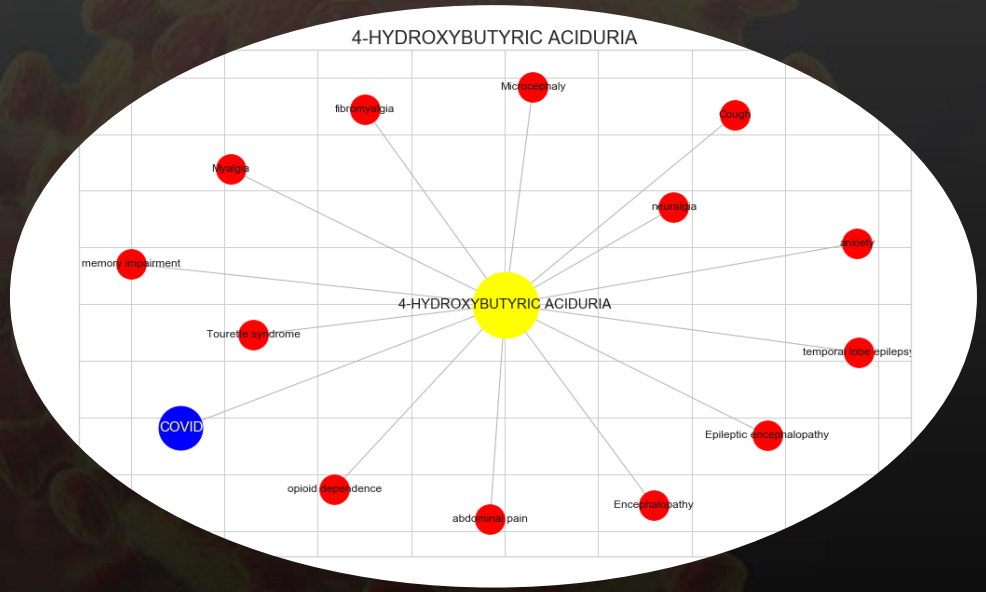


Results

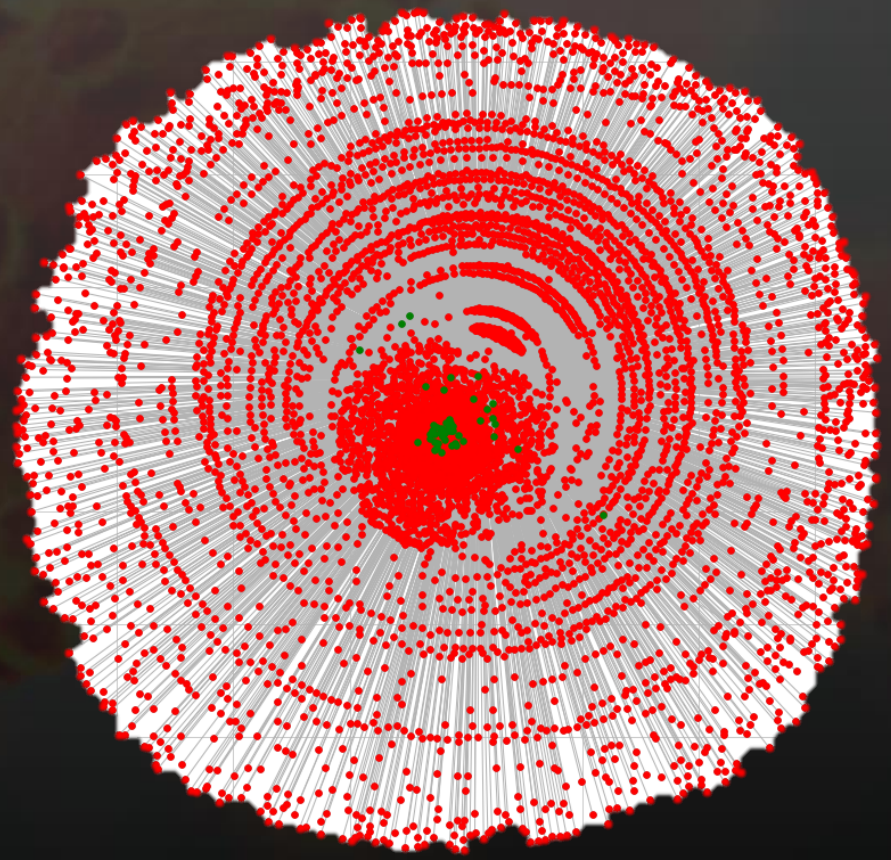
2. Overall Ranking



2. Covid Ranking



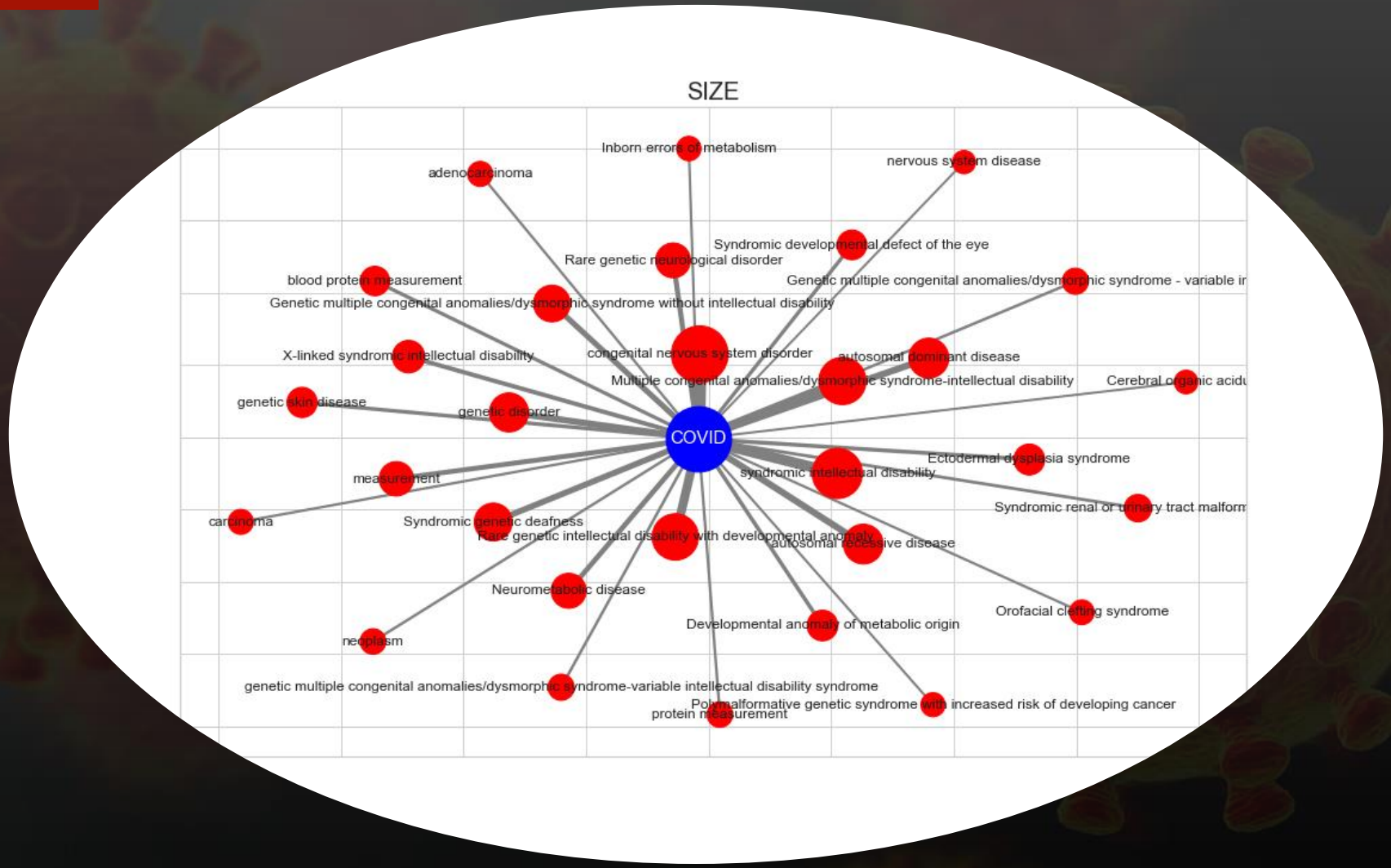
Model



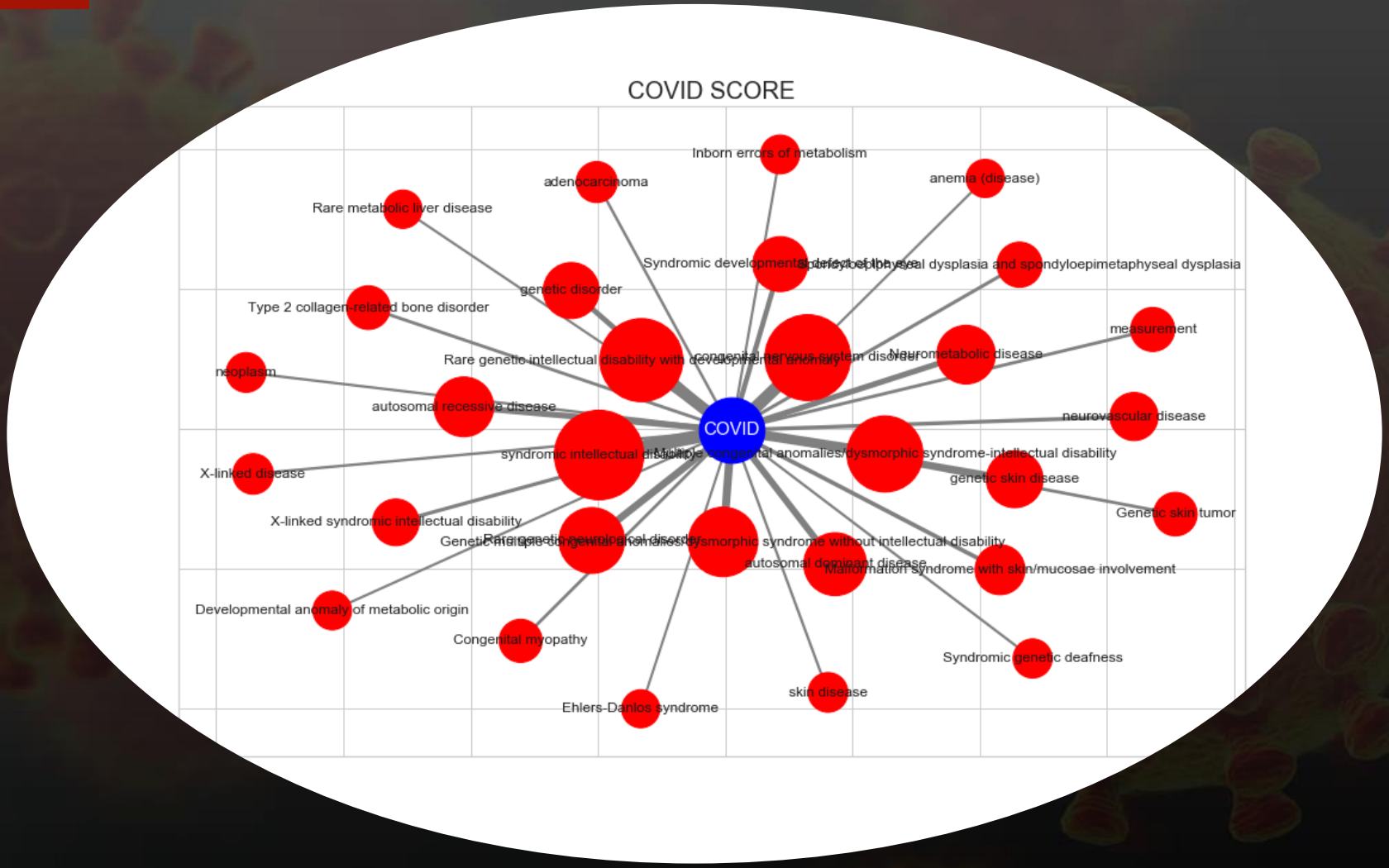
→ **Dataset II**

- **Nodes: Communites**
- **Edges: Covid score**
- **Goal: Community influence**

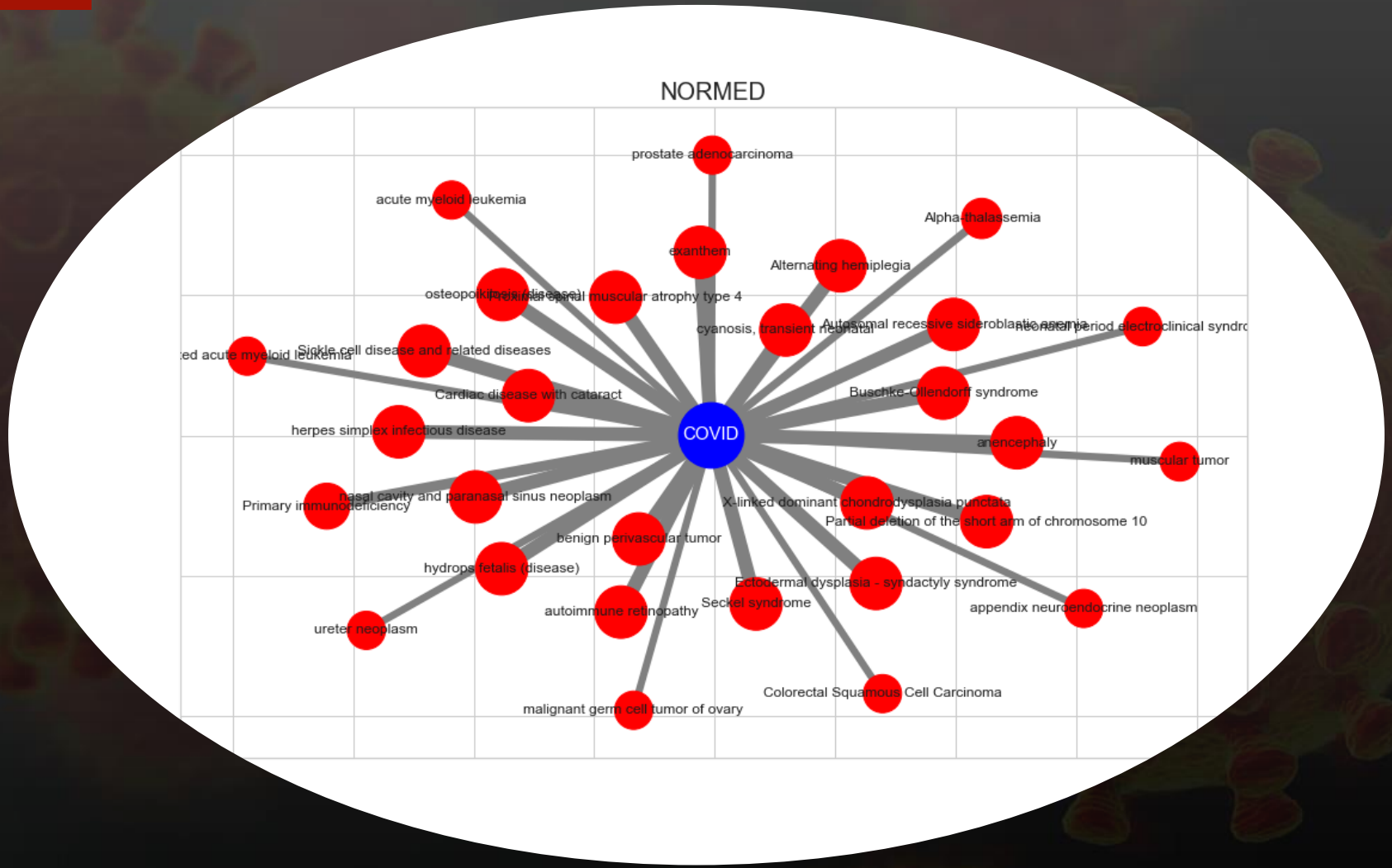
Results


SIZE
Inborn errors of metabolism
adenocarcinoma
nervous system disease
Syndromic developmental defect of the eye
Rare genetic neurological disorder
blood protein measurement
Genetic multiple congenital anomalies/dysmorphic syndrome - variable ir
Genetic multiple congenital anomalies/dysmorphic syndrome without intellectual disability
X-linked syndromic intellectual disability
congenital nervous system disorder
autosomal dominant disease
Multiple congenital anomalies/dysmorphic syndrome-intellectual disability
Cerebral organic acidu
genetic skin disease
genetic disorder
COVID
Ectodermal dysplasia syndrome
syndromic intellectual disability
measurement
Syndromic renal or urinary tract malform
carcinoma
Syndromic genetic deafness
Rare genetic intellectual disability with developmental anomaly
autosomal recessive disease
Neurometabolic disease
Orofacial clefting syndrome
Developmental anomaly of metabolic origin
neoplasm
genetic multiple congenital anomalies/dysmorphic syndrome-variable intellectual disability syndrome
Polymalformative genetic syndrome with increased risk of developing cancer
protein measurement

Results


COVID SCORE
Inborn errors of metabolism
adenocarcinoma
anemia (disease)
Rare metabolic liver disease
Syndromic developmental defect of the eye
spondyloepiphyseal dysplasia and spondyloepimetaphyseal dysplasia
genetic disorder
Type 2 collagen-related bone disorder
measurement
Rare genetic intellectual disability with developmental anomaly
congenital nervous system disorder
Neurometabolic disease
neoplasm
autosomal recessive disease
COVID
neurovascular disease
syndromic intellectual disability
Multiple congenital anomalies/dysmorphic syndrome-intellectual disability
X-linked disease
genetic skin disease
Genetic skin tumor
X-linked syndromic intellectual disability
Genetic multiple congenital anomalies
Dysmorphic syndrome without intellectual disability
autosomal dominant disease
Malformation syndrome with skin/mucosae involvement
Developmental anomaly of metabolic origin
Congenital myopathy
Syndromic genetic deafness
skin disease
Ehlers-Danlos syndrome

Results


NORMED
prostate adenocarcinoma
acute myeloid leukemia
Alpha-thalassemia
exanthem
Alternating hemiplegia
osteopoikilosis (disease)
Proximal spinal muscular atrophy type 4
cyanosis, transient neonatal
Autosomal recessive sideroblastic anemia
neonatal period electroclinical syndro
ed acute myeloid leukemia
Sickle cell disease and related diseases
Cardiac disease with cataract
Buschke-Ollendorff syndrome
herpes simplex infectious disease
COVID
anencephaly
muscular tumor
Primary immunodeficiency
nasal cavity and paranasal sinus neoplasm
X-linked dominant chondrodysplasia punctata
Partial deletion of the short arm of chromosome 10
benign perivascular tumor
hydrops fetalis (disease)
Ectodermal dysplasia - syndactyly syndrome
Seckel syndrome
appendix neuroendocrine neoplasm
autoimmune retinopathy
ureter neoplasm
Colorectal Squamous Cell Carcinoma
malignant germ cell tumor of ovary

IV. Random Forest Model

Pre-processing

→ **Data set**

- Self generated
- No pre-processing needed
- 70/30 split for train/test data

Model

→ **Grid search model**

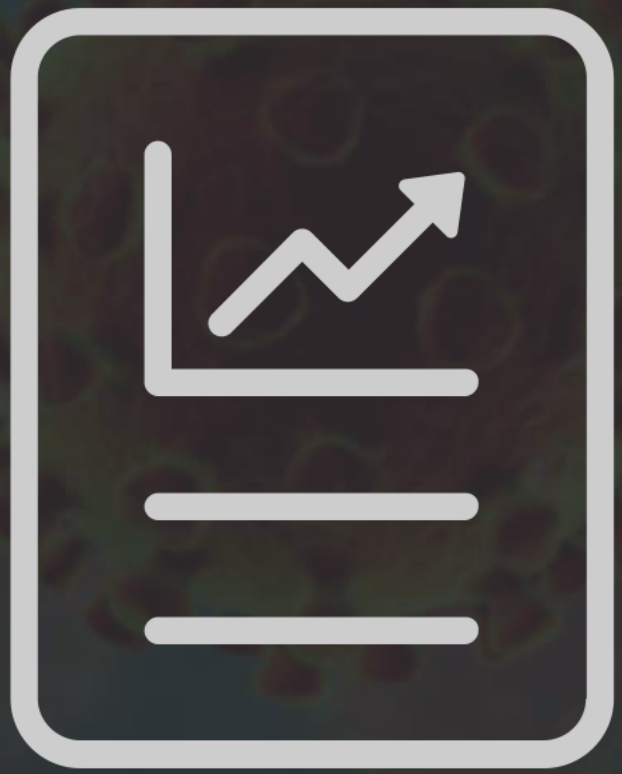
- **Parameters**
  - Maximum depth
  - Number of estimators
  - Minimum samples per leaf
- **385 fits**

Model

→ **Random search model**

- **Parameters**
  - Number of estimators
  - Maximum depth
  - Minimum sample split
  - Minimum samples per leaf
  - Number of features considered when looking for the best split
- **25 fits due to simulation length**

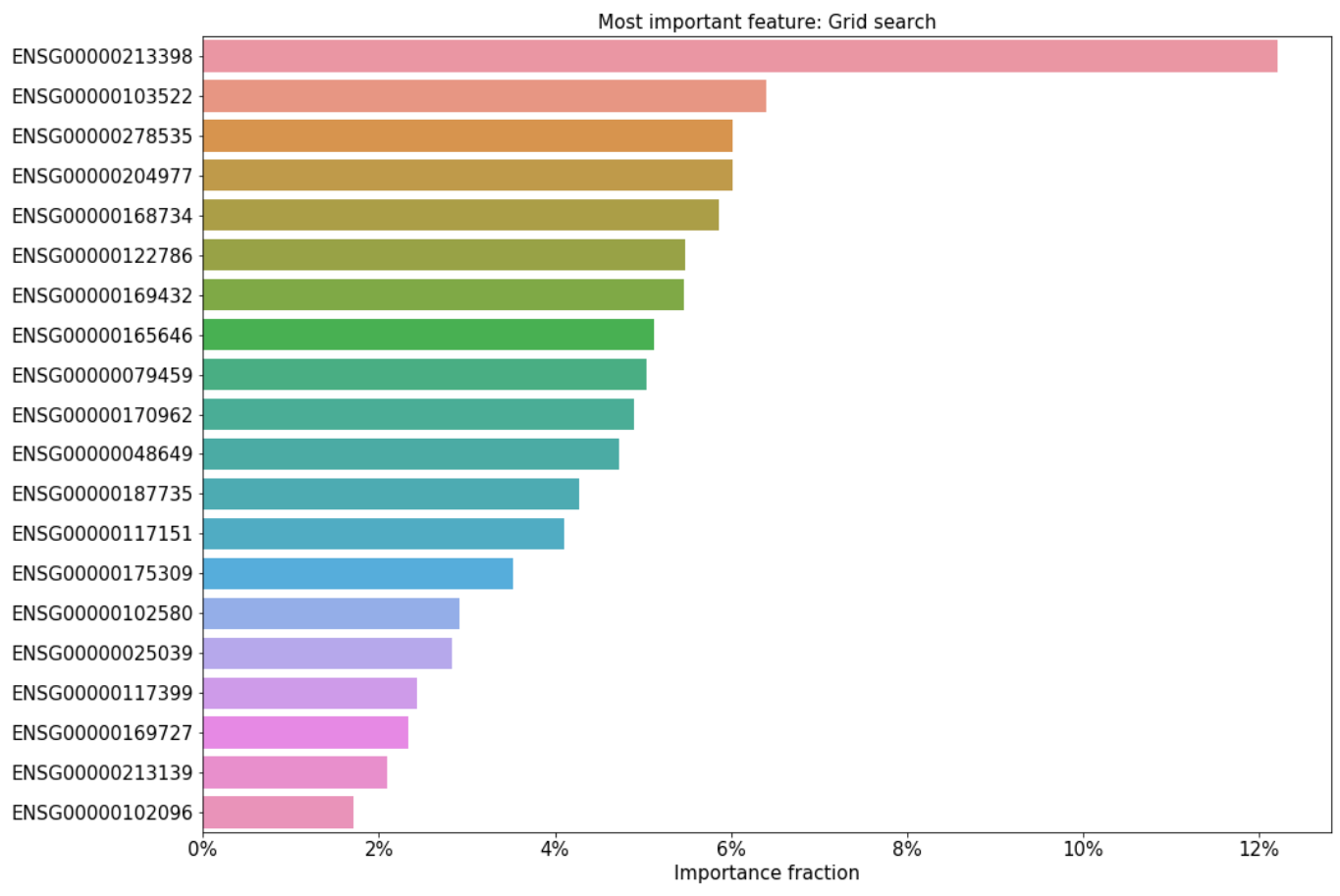
Results



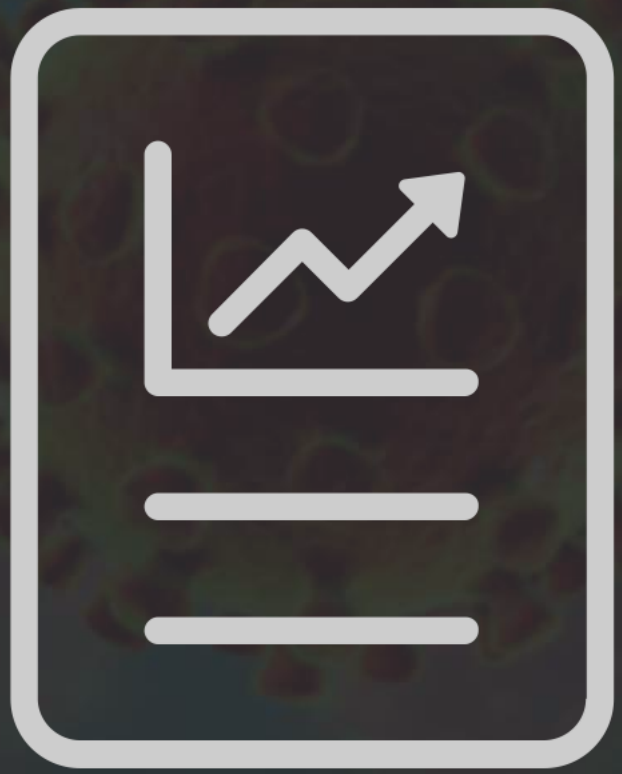
→ **Grid search**

- 22 targets out of 100 compose 95% of the importance
- Lecithin-cholesterol acyltransferase at 12%
- Next 18 between 2 and 8%
- $r^2 = 0.0013698372767038 11$
  - 35% less than baseline

Results


Most important feature: Grid search
ENSG00000213398
ENSG00000103522
ENSG00000278535
ENSG00000204977
ENSG00000168734
ENSG00000122786
ENSG00000169432
ENSG00000165646
ENSG00000079459
ENSG00000170962
ENSG00000048649
ENSG00000187735
ENSG00000117151
ENSG00000175309
ENSG00000102580
ENSG00000025039
ENSG00000117399
ENSG00000169727
ENSG00000213139
ENSG00000102096
0%
2%
4%
6%
8%
10%
12%
Importance fraction

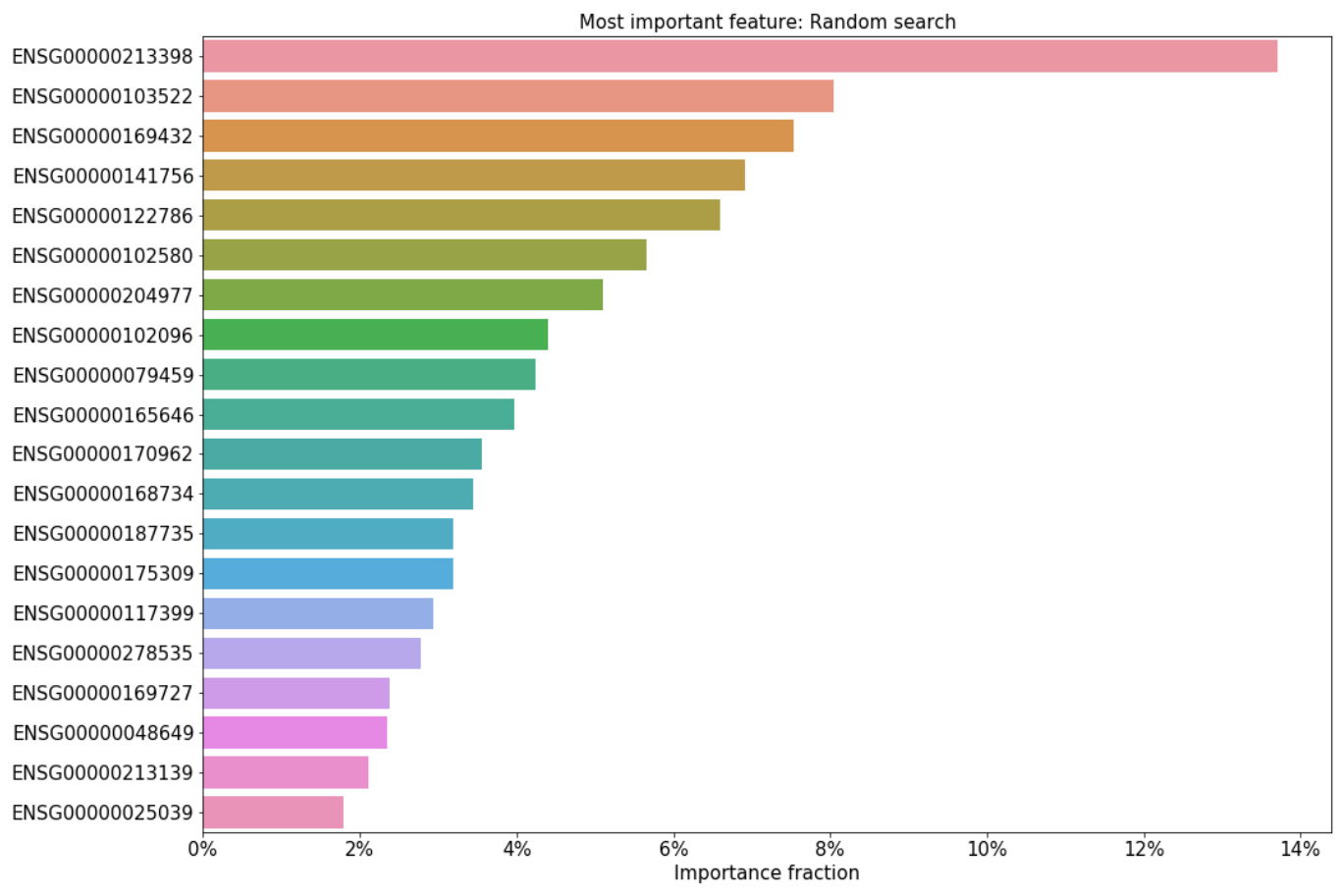
Results

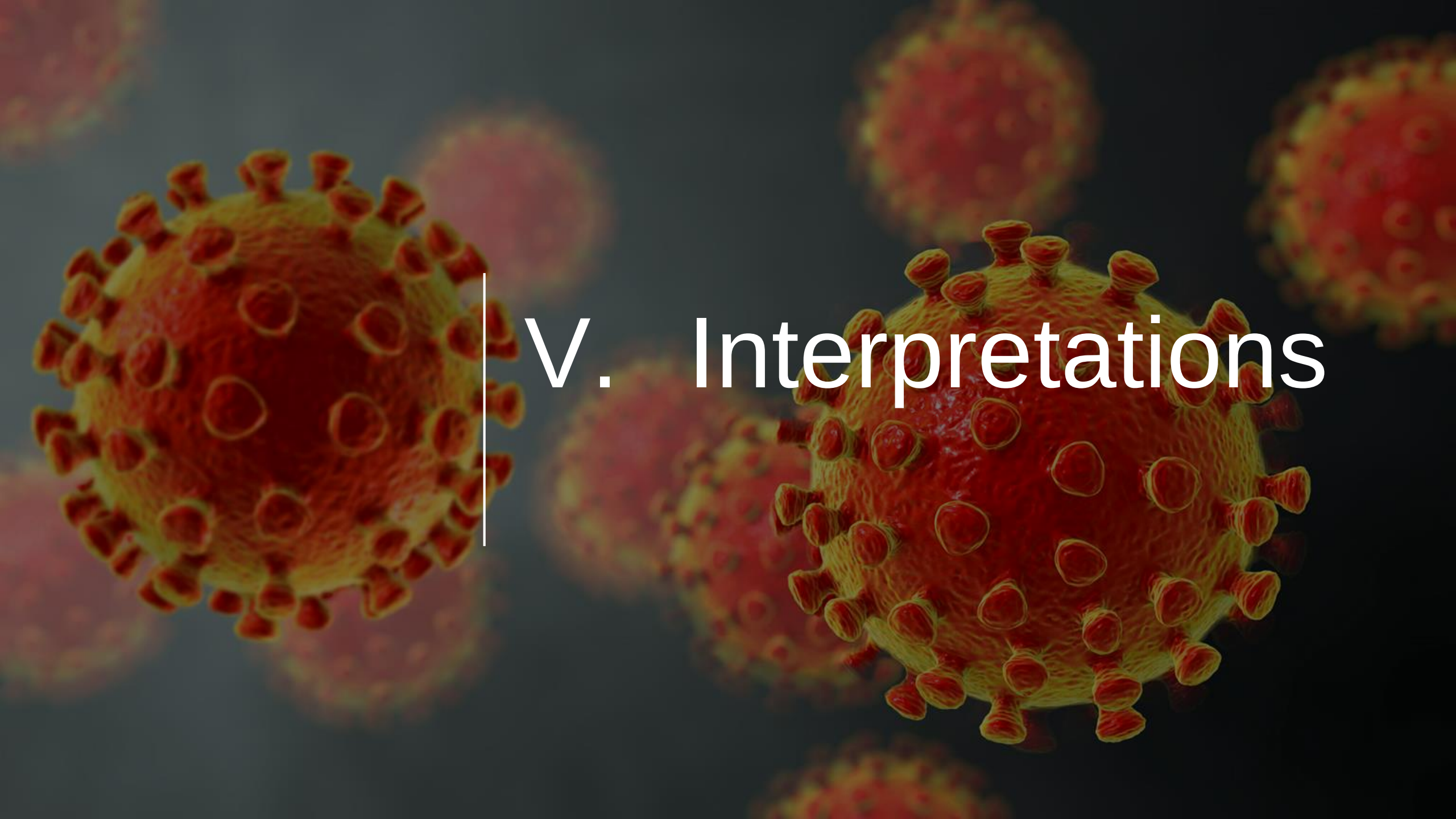


→ **Random search**

- 21 targets out of 100 compose 95% of the importance
- Lecithin-cholesterol acyltransferase at 14%
- Next 18 between 2 and 8%
- $r^2 = 0.00048797298004477074$
  - 47% less than baseline

Results


Most important feature: Random search
ENSG00000213398
ENSG00000103522
ENSG00000169432
ENSG00000141756
ENSG00000122786
ENSG00000102580
ENSG00000204977
ENSG00000102096
ENSG00000079459
ENSG00000165646
ENSG00000170962
ENSG00000168734
ENSG00000187735
ENSG00000175309
ENSG00000117399
ENSG00000278535
ENSG00000169727
ENSG00000048649
ENSG00000213139
ENSG00000025039
0%
2%
4%
6%
8%
10%
12%
14%
Importance fraction



V. Interpretations

Benefits for the research on the virus

→ **Network analysis**

- **Disease ranking:**
  - Bone disease, psychiatric disorder, biological process, protein measurement and diabetes mellitus
- **Target ranking:**
  - LCAT, SCN9A, SLC18A2, PDGFD and CALD1

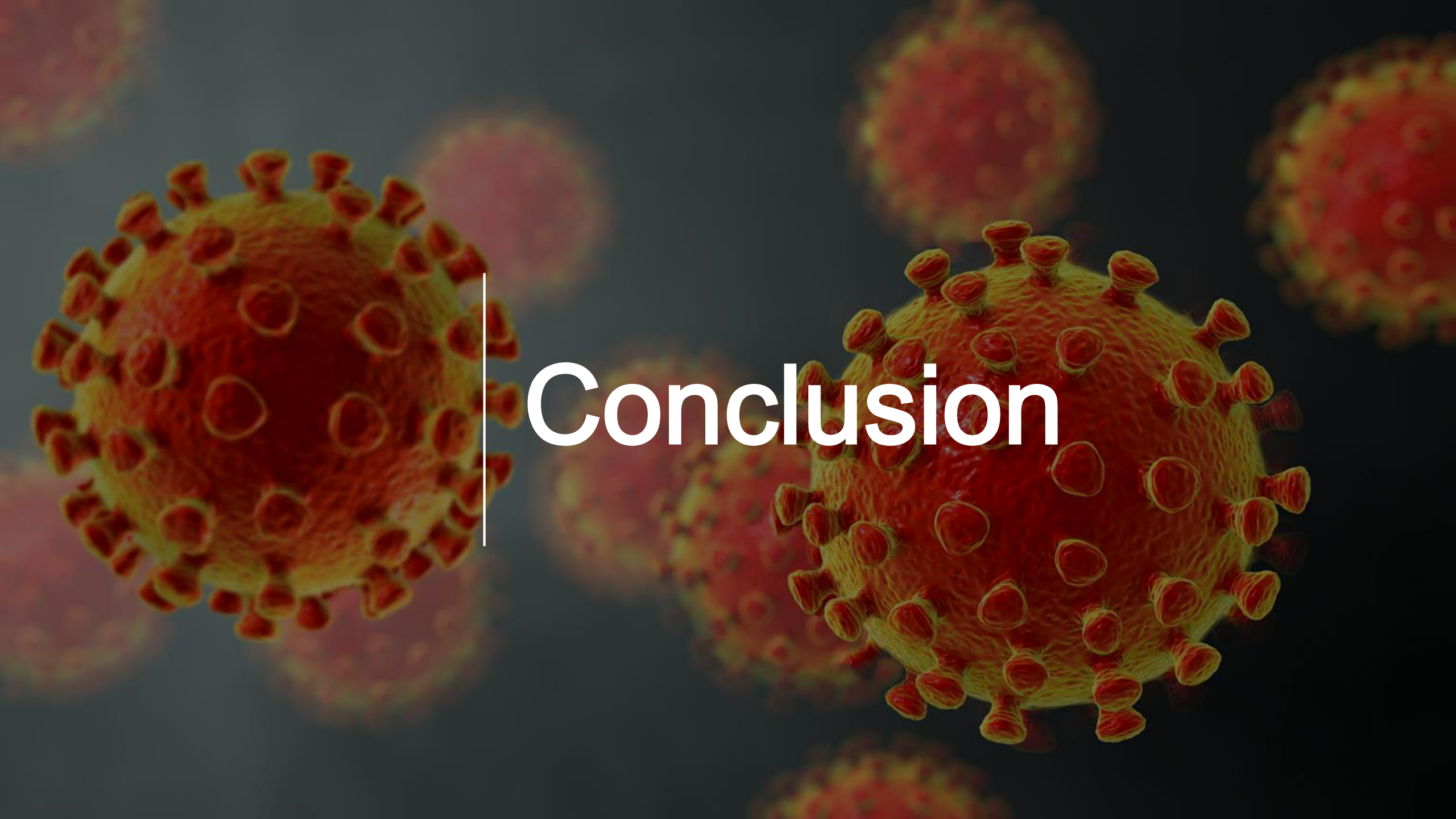
→ **Random forest regression**

- **Target identification helps reduce research time and efforts**
- **Reduce costs**
- **Speed up the vaccine development**

Improving and reusing the model

→ **Improvements and limitations**

- Based on the hypothesis that genetics and genomics play a role in the impact on COVID-19
- Association score between diseases and COVID-19 built on a partial data-set
- Main limitation relying in the amount of data used: targets and diseases
- Can be run with a larger amount of data to expect better results


Conclusion

Conclusions

- **Closest diseases to COVID-19 identified**
- **Most important target risk-factors identified**
- **Reduction of the time spent on vaccine development**
- **Help governments make proportionate decisions**

Thank you for your attention

R. Chaouche, Y. Martinson, C. Padovani, J.Triomphe