Drug repurposing for SARS-CoV-2 infection using machine learning and mechanistic models of the COVID-19 Disease Maps

Carlos Loucera, Marina Esteban Medina, Kinza Rian, Matias M. Falco, Joaquin Dopazo and Maria Peña-Chilet

Clinical Bioinformatics Area, Fundación Progreso y Salud
Computational Systems Medicine, Institute of Biomedicine of Seville (IBIS)

DMCM2020: 5th Disease Maps Community Meeting (Online)
Paris/Luxembourg, France, November 12-14, 2020

Fundación Progreso y Salud
CONSEJERÍA DE SALUD Y FAMILIAS

The problem

Learn potential **relations** between proteins (**targets of drugs**) and the **COVID-19 Disease Map**

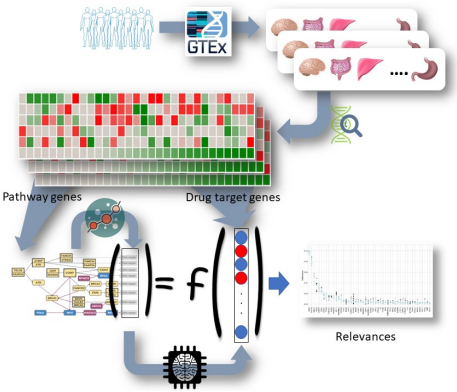
Mechanistic modeling of the MAP

Connect the **Drug** (Protein) Space with the MAP **Signalization** Space

Infer what is important to **regulate** the MAP

Suggest potential *repurposable* **drugs**.

Multi-output Supervised Learning



Build the Signalization Space

>**10k samples** across **53 tissue sites**
The **COVID-19 Disease Map**
277 circuits in **47** sig. **pathways**.

Hipathia Signal Transduction

$$S_n = v_n \left(1 - \prod_{s_a \in A} (1 - s_a)\right) \prod_{s_i \in I} (1 - s_i)$$

Connect the Dots

The **Map** results in **10k x 277** activities.
Over **2k KDT** targeted by **1.7k drugs**.
Most **KDT** lie outside the **Disease Map**.

Data-driven Validation

Performance

Model: Multi-Output Random Forest
Hyper-parameter optimization with TPE
Repeated Nested 10-fold CV
Use SHAP values for explanations
Summarize by `MeanAbs` and take 1st decil

Explanation stability

Robustness Vs Stochastic, Noise, Sampling
100 Holdout splits of half the sample size
Split *training* into *learning* and *validation*
Check unbiased performance over *test*
Compute Nogueiras statistic test and CI

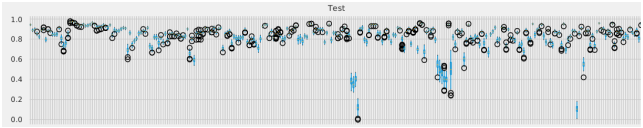
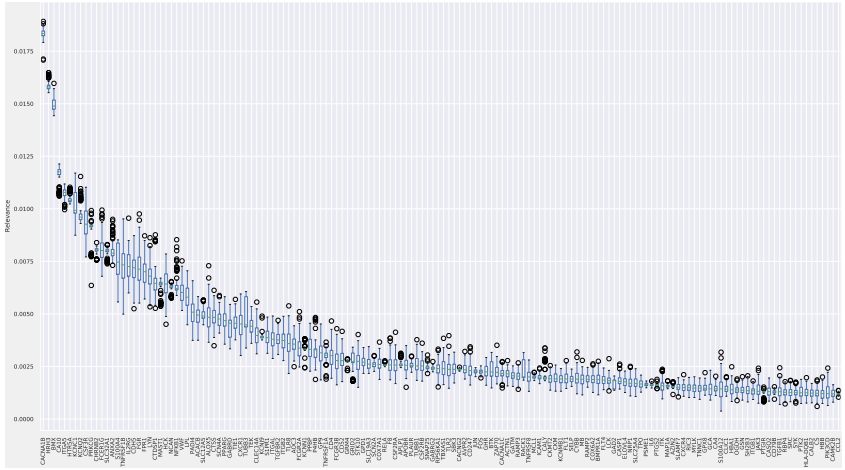


Figure: R^2 score distribution over the test (R10-foldCV).

`Mean` R^2: 0.82 ± 0.02
`N-Stat CI`: $(0.729, 0.735)$

Classical Model Explanations

Useful but too broad. Only speak about the whole map. Misleading?



SHapley Additive exPlanations

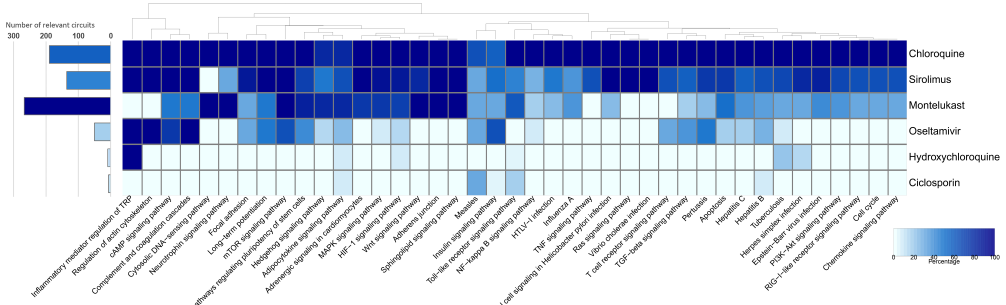
Fair feature responsibility attribution.
Dis-aggregated by circuit by construction.
Additive (use biologically-relevant groups).

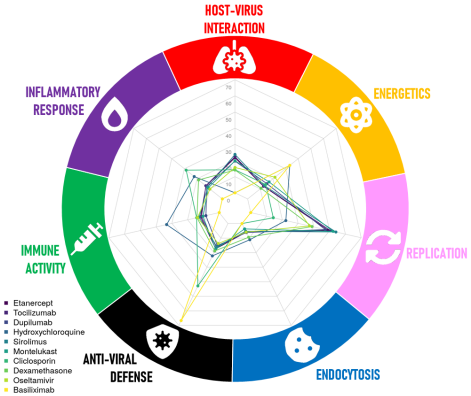
Template patterns

Conform a series of templates, e.g.:

Affect massively almost all MAP

Affect a few very specific circuits





380 KDTs (targeted by 679 drugs) have direct influence over the whole or partial parts of the map.

The GO biological processes enriched are mostly related to immune activity (T-cell, inflammatory response)

The COVID-19 Hallmarks are represented.

Future works

Speed the computation (GPU extensions).

Develop *sign* like SHAP aggregations.

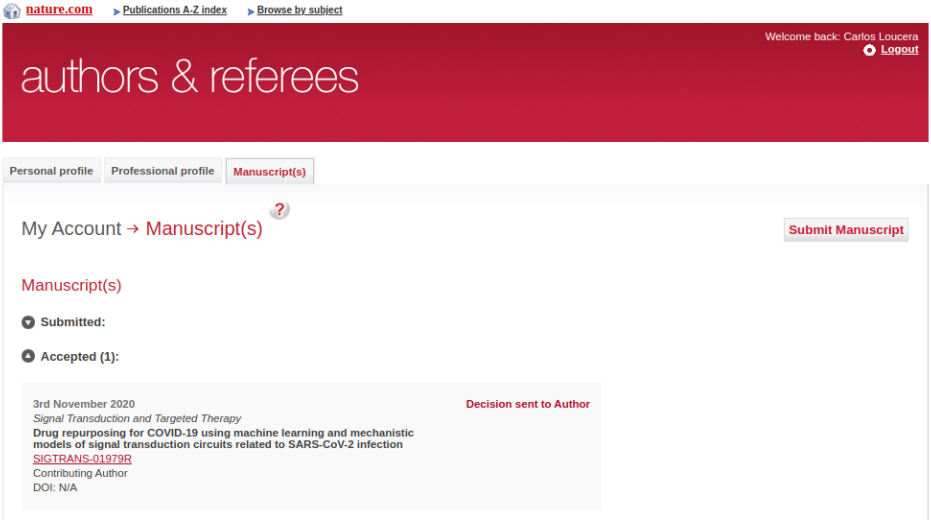
Aggregations in *task*-space.

Use directly the COVID-19 Disease Map (near done! see Kinza's talk, and our ML framework is DB-agnostic (**modular!**))

Machine learning software and other tools

python	ML coded in python 3.7 (and metadata parsing)
scikit-learn	ML pipelines, experimental design, etc
seaborn	ML plots
click	Command line interface
other	LaTeX, bash, slurm, jupyter lab, Steven's LaTeX styles, OBS studio ...

Accepted for publication in Signal Transduction & Targeted Therapy!



Fundación Progreso y Salud
CONSEJERÍA DE SALUD Y FAMILIAS

Clinical Bioinformatics Area
Fundación Progreso y Salud, Sevilla, Spain, and…

…the INB-ELIXIR-ES, National Institute of Bioinformatics and the BiER (CIBERER Network of Centers for Research in Rare Diseases)

https://www.slideshare.net/xdopazo/

Follow us on twitter
@xdopazo
@ClinicalBioinfo
@loucerac
@CSoriguer

INB

Fundación BBVA

NAGEN
PROYECTO GENOMA 1000 NAVARRA

Instituto de Salud Carlos III · ciberer · MLFPM · elixir SPAIN · MARIE CURIE ACTIONS