UNIVERSITY OF CATANIA

PhD IN COMPLEX SYSTEMS FOR PHYSICAL SOCIO-ECONOMIC AND LIFE SCIENCES
XXXIV cycle

DEPARTMENT OF PHYSICS AND ASTRONOMY *ETTORE MAJORANA*
DEPARTMENT OF CLINICAL AND EXPERIMENTAL MEDICINE - BIOINFORMATICS SECTION

Dr. Rosaria Valentina Rapicavoli

A System Biology approach for diseases modelling and drug repurposing

---

PhD Thesis

Tutor:

Dr. Salvatore Alaimo, PhD

Co-Tutors:

Prof. Alfredo Ferro

Prof. Stefania Stefani

Prof. Alfredo Pulvirenti

Tutor: Dr. Salvatore Alaimo Dr. Rosaria Valentina Rapicavoli
Co-Tutors: Prof. Alfredo Ferro
Prof. Stefania Stefani
Prof. Alfredo Pulvirenti

A System Biology approach for diseases modelling and drug repurposing

Abstract

Providing easy-to-use tools to support clinical and biological research in the analysis of complex biological phenomena is becoming a critical issue and a great challenge.
Systems biology is playing a central role in this direction. This type of approach can be strategic, for example, in helping clinicians to make better clinical decisions such as prescribing drugs or their combinations by tracking phenotypic deregulations consequences of a given disease, or specific to one or more patients. In fact, there is a growing interest in conducting patient-centered analyses and predictions to assess both patient-specific conditions and drug efficacy.
The difficulties, costs, and timelines required in the discovery and development of new pharmaceutical molecules is a main problem. On the other hand, the availability of a huge number of drugs, that have passed all critical phases of clinical trials and for which most side effects are already known, increase the interest in repositioning such drugs for new therapeutic purposes. This strategy constitutes one main issue in personalized medicine.
The aim of this thesis is to develop a new methodology, using a systems biology approach, for drug repositioning. This approach produces a drug prioritization based on a dynamic pathway-mechanistic prediction.
In silico simulations are powerful techniques able to integrate experimental data and bibliographic information in order to predict, at different scales of detail (pathways, genes, proteins, metabolites), the effect of diseases and drugs. Computational experiments produce data allowing a multi-level understanding of a specific disease. This will guide identification of new druggable targets highlighting mechanisms of action of repositioned drugs and their cascading effects.
Nowadays, infectious diseases represent a global emergency. Drug repurposing may represent a first-aid strategy waiting for vaccines or new specific molecules. The availability of cheap and easy-to-use computational tools for such task, appears to be of great importance for urgent clinical decision. The pandemic that has been developing since 2019 has raised this major challenge. For a virus as novel as SARS-CoV-2, knowledge of the host immune response to infection in order to design appropriate emergency therapies is crucial. Here we introduce a systems biology tool, the PHENotype SIMulator which, by leveraging available transcriptomic and proteomic data, allows *in silico* modeling of SARS-CoV-2 infection in host cells to i) determine with high sensitivity and specificity (both >96%) the viral effects on the host-immune cellular response, resulting in a specific cellular signature of SARS-CoV-2 and ii) use this specific signature to narrow promising repurposable therapeutic strategies. Using this tool, coupled with expertise in the field, we have identified several potential drugs for COVID-19, including methylprednisolone and metformin, and further discern key cellular pathways influenced by SARS-CoV-2 as potential new pharmacological targets in the pathogenesis of COVID-19.

Contents

Appendix

i) Figure S1 Comparison between PHENSIM with and without REACTOME for datasets where the altered gene belongs to the meta-pathway
   Figure S2 Comparison between PHENSIM with and without REACTOME for datasets where the altered gene was not in the meta-pathway
ii) Figure S3 From in vitro to in silico. Comparison of drug treatments between Stukalov et al. 2021 and PHENSIM predictions
iii) Figure S4. MITHrIL vs Reactome pathway analysis
iv) From Figure S5 to S11. Representation of resulted top pathways significantly affected by repurposed drugs.

List of Figures

List of Tables

Acknowledgments

I wish to thank all those who made this work possible and who played an important role for my PhD thesis and beyond.
First of all I wish to thank Prof. Alfredo Ferro who welcomed me in his wonderful research group with great enthusiasm and affection. I am, and I will be, infinitely grateful to him for letting me to feel immediately part of the group, but especially for giving me great respect and trust from the very beginning. I also thank him for giving me the opportunity to meet and cooperate with "the biggest" scientists, allowing me to grow from all points of view.
A special thank goes to Dr. Salvatore Alaimo who guided me during the whole project, supporting me every time I needed. He gave me constant stimuli and during the most crucial moments of my work he never made me feel alone, giving me his help but especially teaching me how to be independent.
I also wish to thank Prof. Alfredo Pulvirenti who welcomed me into his wonderful group with great positivity. I thank him for the stimuli, the opportunities, the encouragement and the serenity he always transmitted to the whole group, allowing us to work in the best way together with enthusiasm and spirit of collaboration.
I wish to thank Prof. Stefania Stefani, who gave me the courage to make fundamental choices for my professional and personal growth. She taught me that you can decide what to be in your life consciously, that obstacles can be overcome with courage, determination and sacrifice and that our dreams must be our engine, our energy.
I also wish to thank Prof. Andrea Rapisarda for the support and understanding he gave me in the choices I made throughout my PhD, giving me great trust. I thank him for being always present for each of us PhD students, overcoming our problems and taking care of us.
I also wish to thank the whole RxCOVEA group, especially Prof. Ashley Duits, Prof. Bud Mishra, Dr. Naomi I. Maria and Prof. Evelyne Bischof: my first international collaboration group. I thank all of them for the scientific stimuli, the cooperation and the peaceful atmosphere with which we worked during these last two very difficult years due to the COVID-19 pandemic. We have worked intensely and lovingly, breaking down any distance and becoming a big "family".
I thank Prof. Marino Zerial who introduced me to the MPI-CBG and in his wonderful research group since my first day in the Institute. I thank him for his friendliness and for making me feel always appreciated and welcomed. I thank all the Zerial-Lab, because each one of them contributed to make me feel at “home” during my time in Dresden.
I thank my family, my lovely parents, who allowed me to study until now. They were always caring and present, ever ready to provide me with great encouragement, reassurance, love and good advices.
Finally, I would like to thank my dear husband Mauro, for his immense love and comprehension. With him I shared my efforts and my achievements. He gave me the strongest and essential support staying near to me every single moment. I thank Mauro because he believed in me and in my dream infinitely, sometimes even more than myself. For this reason I dedicate my PhD thesis to him.
I thank each of the people mentioned with infinite affection and gratitude. Each of them has been essential to the success of this work.

1

Introduction

Our entire existence is based on continuous interactions. We are ourselves the result of interactions between thousands of genes, proteins and metabolites within our cells, and our state of health or disease is closely related to their collective behavior: we are complex systems.
Every cell is a complex and dynamic unit in which thousands of proteins regulate and target specific tasks with exceptional efficiency and precision. Cells in living organisms are constantly exposed to signals from both extracellular and intracellular microenvironments, and molecular interactions are modified with high sensitivity by evolving situations, thus regulating qualitative and quantitative biomolecules production. These signals regulate multiple cellular functions, including gene expression, chromatin remodeling, DNA replication and repair, protein synthesis and metabolism. The appropriate response to signals depends on the expression, activation or inhibition of interconnected sets of genes/proteins, acting in a well-defined order within vector-based biological processes, aimed to reach specific endpoints. In this context, the study of the genome and transcriptome, the definition of protein-protein interaction networks and the investigation of the association between multiple gene sets and molecular mechanisms in humans have provided valuable biological information.
In this perspective, new systems biology approaches are playing a central role in modelling and understanding the interactions between molecular entities behind biological phenomena, significantly improving manual analysis.
In addition, computational pathway analysis and/or *in silico* simulation techniques can be extensively applied on a massive scale, allowing thousands of hypotheses to be evaluated under various conditions.
In a world that is moving towards personalized medicine, tools allowing efficient and deep analyses of complex diseases are needed. These approaches should go beyond the simple identification of biomarkers supporting physicians in discriminating the most promising treatments. In particular this is very will be very helpful when rare or poorly known diseases are considered.

The research activity carried out during my PhD course and described in this thesis, is focused on the application of computational techniques to study molecular mechanisms underlying disease states and the development of a new Systems Biology approach for Drug Repositioning. Finally in silico-based screening may significantly lower preclinical study costs by filtering out less promising experiments.

Research on host-pathogen interactions is an ever-evolving field. To understand its social, health and economic impact, it is enough to consider that about a quarter of deaths in the world are caused by infectious diseases. Therefore, every two days a new pathogen is discovered implying new challenges for its prevention and treatment.
Nowadays, infectious diseases represent a global emergency. For this reason, the first application of the novel methodology proposed in the present thesis involved viral infections.

Understanding the mechanisms underlying host-pathogen interaction that lead to successful pathogen invasion and obtaining a comprehensive understanding of the molecular events and perturbations induced by infection is extremely important. Moreover, this can suggest druggable genes for new therapeutic strategies. Therefore, in order to correctly rank candidate repurposable drugs, it is necessary to design valid tools modeling both pathogen and drug molecular mechanism of action, through genes and pathways analyses.

Here we propose PHENSIM (Phenotype Simulator)[1] as such computational tool using a systems biology approach to simulate cell phenotypes such as drug, disease, or pathogen effects.

PHENSIM uses a probabilistic algorithm to calculate the effect of genes, proteins, microRNAs (miRNAs) and metabolites dysregulation on the KEGG[2,3] and REACTOME [4–6] pathways, allowing phenotype predictions to be made on selected cell lines or tissues in 25 different organisms.

To assess the reliability of the simulator, we built a benchmark from transcriptomics data collected from NCBI GEO and performed four case studies on known biological experiments. Our results showed high prediction accuracy, thus highlighting the capabilities of this methodology.

The new repositioning pipeline proposed in this thesis, involves the use of high-resolution but easily producible experimental data, i.e. transcriptomics or proteomics data, more specifically the differentially expressed genes/proteins (DEGs/DEPs) in a given cell line. Starting from these data, following a mechanistic approach, molecular consequences of viral infection are computationally predicted. Subsequently the results of such computation are analyzed highlighting both potentially druggable genes and pathways. On the other side, similar gene/pathways signatures of existing drugs will be produced. More precisely for each drug, specific targets known from the literature and databases such as Drug Bank or PubChem are collected and used as input for PHENSIM *in silico* simulations. Viral and drug simulations should be performed in the same biological context (cell lines, tissues etc.).

Finally, a ranking of repurposable drugs is constructed by Pearson anti-correlation between viral and drug signatures (see Fig.24). Indeed, this expresses how much drugs contrast viral infection. Those drugs which are positively correlated with the investigated virus will be disgorged from the list of candidates. On the other hand, negatively correlated drugs may be considered potential therapeutical suggestions according to their ranks.

The great advantage of this methodology is that it is an explainable not black-box approach. PHENSIM was chosen precisely because it allows the inspection of pathways deregulations and their associated genes, thus enabling the acquisition of new knowledge on the cellular and molecular mechanisms underlying the pathological condition being studied. Thus, for each drug it is possible to see not only whether it is positively or negatively correlated, but also why, i.e. which specific pathways and genes are involved.

As mentioned above, our original plan was to apply our methodology to study general viral infections. However, the emergency of COVID-19 pandemic addressed our investigation towards SARS-CoV-2 virus.

To validate our pipeline, drugs assayed at multiple concentrations *in vitro* against SARS-CoV-2 infection by *Stukalov et al. 2021*[7] were selected and repositioned *in silico*. Drugs signatures were constructed using data available in the L1000 database[8] and giving as input to the PHENSIM simulations the respective Differentially Expressed Genes (DEGs). Since *Stukalov et al. 2021*[7] tested several *in vitro* concentrations for which there was not an exact match in L1000, we decided to select the closest ones. Finally, our in silico results were compared to the *in vitro* assays described by *Stukalov et al. 2021*[7].

This study is part of a collaboration within an international and multidisciplinary group of scientists adhering to the challenging project RxCoVea[9], conceived and realized by prof. Bud Mishra (see

Chapter 4). This community involves about 100 volunteers, including senior scientists/professors and young researchers, from more than 15 countries, collaborating together with the common goal of being helpful in the COVID-19 pandemic.
Within this group research on the host-pathogen interactions and the validation of our method for drug repositioning was carried out through close collaboration with the *Department of Immunology of the Curacao Biomedical & Health Research Institute* (CBHRI), where I carried out part of my activities in smart working modalities due to the global pandemic, supervised by Prof. Ashley John Duits. The research was also carried out in collaboration with the Northwell Health Hospital represented by Dr. Naomi I. Maria, Shanghai University of Medicine represented by Prof. Evelyne Yehudit Bischof and the Courant Institute of Mathematical Sciences of New York University represented by Prof. Bud Mishra.
Given that currently available tools for pathway analysis are based on incomplete knowledge networks (for example KEGG contains only about one third of the known genes), it may happen that in some context missing genes may cause discrepancies with biological and clinical data. To overcome this problem, it is possible to extend PHENSIM model by introducing important missing genes and their relations with the other nodes of the network. In particular to correctly model SARS-CoV-2 action we extended the network by adding the crucial gene CD147. This integration was done manually by a detailed literature investigation for the reconstruction of the network to be integrated in the PHENSIM meta-pathway[1].
Although the manual extension of such networks by a careful and time-consuming literature review remains, in terms of accuracy and up-to-date information, the most reliable approach, a more efficient computational approach was clearly needed.
This led to the application of our text-processing system NETME[10]. This tool, starting from either a set of full texts obtained from PubMed, or pdfs directly provided by the user, interactively extracts biological elements from ontological databases and then synthesizes a network by inferring relationships between those elements.

This pandemic has brought the scientific community together, as never before. We have all embraced the idea of an open and shared science and we have experienced the importance of sharing data, codes and protocols, reporting and disseminating results.

The urgency of the pandemic has created an imperative to accelerate the adoption of open science [7, 11]. Multidisciplinary open science has emerged as a powerful mechanism to accelerate science and combat the rapidly evolving global pandemic of COVID-19.
Indeed, disseminating results through a collaborative environment allows for hypothesis testing, detecting contradictions, validating sources, and filtering out false data.
This prompted us to create SciKi (*Scientific wiKi)*, a toolbox developed primarily to integrate and disseminate results obtained using PHENSIM-based drug-discovery framework.
This new platform, which is not yet in its final form, is designed to interact with open scientific communities in an innovative way and to publicly disseminate reproducible and explainable scientific results in a simple and effective way.

The conceptual simplicity underlying the proposed pipeline makes it versatile and applicable in the most diverse contexts.
As a demonstration, in the context of host-pathogen interactions, another project is going on and it is almost in its final stage. The idea is comparing several viral infections by applying the above repositioning strategy.

In this context, several viral signatures will be constructed, analyzed and compared. The case of possible co-infection will be assessed. In particular, since we have not yet emerged from the COVID-19 pandemic, we are especially interested in studying the case of co-infections with SARS-CoV-2. In particular, HRV, IAV, RSV and HPIV3 with SARS-CoV-2 will be analyzed.

Given the potential of the Systems Biology methods described in this thesis, new ideas, projects and collaborations have come up.
A new project, started recently, involves bioinformatics along with systems biology approaches to helping in understanding complex diseases such as chronic liver disease leading to tissue degeneration. The new project will take place in close collaboration with the Max Planck Institute of Molecular Cell Biology and Genetics in Dresden (MPI-CBG), where I spent the last period of my PhD, working in the lab of Prof. Marino Zerial.

Some results illustrated in this thesis have been published:

· Alaimo S, Rapicavoli RV, Marceca GP, La Ferlita A, Serebrennikova OB, Tsichlis PN, et al. (2021) *PHENSIM: Phenotype Simulator*. PLoS Comput Biol 17(6): e1009069. https://doi.org/10.1371/journal.pcbi.1009069

· Alessandro Muscolino, Antonio Di Maria, **Rosaria Valentina Rapicavoli**, Salvatore Alaimo, Lorenzo Bellomo, Fabrizio Billeci, Stefano Borzì, Paolo Ferragina, Alfredo Ferro and Alfredo Pulvirenti. *NETME: On-the-fly knowledge network construction from biomedical literature*. Applied Network Science.

Results regarding the proposed new approach for drug repositioning is currently published in pre-print form:

· Naomi MARIA, **Rosaria Valentina Rapicavoli**, Salvatore Alaimo, Evelyne Bischof, Alessia Stasuzzo, Jantine Broek, Alfredo Pulvirenti, Bud Mishra, Ashley Duits, Alfredo Ferro. *Rapid Identification of Druggable Targets and the Power of the PHENotype SIMulator for Effective Drug Repurposing in COVID-19. Preprint, Research Squa*re. DOI:10.21203/rs.3.rs-287183/v1

SciKi has been presented at BITS 2021 conference as poster:

· Alaimo S., **Rapicavoli R. V**., Maria N., Bischof E., Broek J., Mishra B., Duits A., Ferro A., Pulvirenti A. *SciKi: Science Wiki for In Silico Target Discovery and Drug Repurposing to Combat Covid-19*.

2

Prerequisites

2.1 Biological Prerequisites

This section is a brief introduction to the necessary background in cell biology, genetics, virology and immunology.

2.1.1 The Cell and the Gene Information Flow

"*The key to any biological problem must be found in the cell, since every living thing is, or at some time in its history has been, a cell*", said by E. B. Wilson, pioneer of cell biology. In fact, cells are the fundamental units of life: all living organisms are made up of cells. They are small functional units surrounded by a membrane that possess many membrane-bound structures inside them, called organelles immersed in a solution called cytoplasm. Each cell possesses the machinery required to carry out all its vital functions and to create a new identical cell to itself, containing a new copy of the hereditary information.

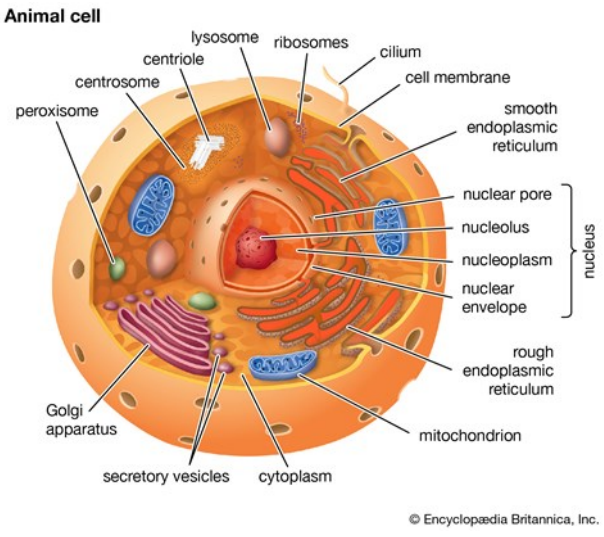


Figure 1. Typical Animal cell. Eukaryotic cells have a multitude of membrane-bound structures called organelles (e.g., nucleus, mitochondria, smooth and rough endoplasmic reticulum, ribosomes) and a cytoskeleton of microtubules, microfilaments, and intermediate filaments that play an important role in the structure and shape of the cell itself. The

nucleus, which is surrounded by a double membrane equipped with pores, houses the cell's DNA and directs the synthesis of proteins and ribosomes. Image courtesy of [12]

Among the organelles within the cell, the nucleus houses the genetic information in the form of DNA, a double-stranded molecule consisting of two long, paired polymeric chains made up of four types of monomers. Each monomer in a single strand of DNA is called *nucleotide* and consists of two parts: a sugar, deoxyribose, with a phosphate group attached, and a nitrogenous base, which can be Adenine, Guanine, Cytosine or Thymine. Nucleotides are referred to as A, G, C and T on the basis of the nitrogenous base they carry. In 1953, James Watson and Francis Crick put forward their double-helix model of DNA[13], based on crystallized X-ray structures being studied by Rosalind Franklin.

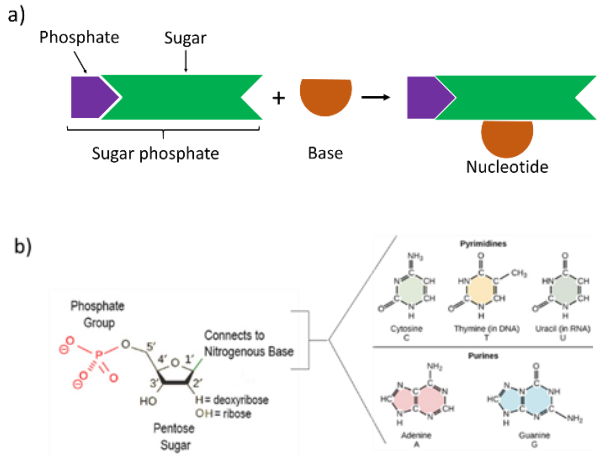


Figure 2. Nucleotide structure. a) Each nucleotide is made up of a sugar, a phosphate group, and a nitrogenous base. The sugar is deoxyribose in DNA and ribose in RNA. b) The nitrogenous base can be a purine, such as adenine (A) and guanine (G), or a pyrimidine, such as cytosine (C) and thymine (T). In the case of RNA molecules, thymine is replaced by another pyrimidine, uracil (U) [15].

In the living cell, DNA is not synthesized as an isolated free strand, but on a template consisting of a pre-existing DNA strand. Bases protruding from the existing strand bind to bases in the newly synthesized strand, according to a strict rule defined by complementary structures of the bases: A binds to T and C binds to G. This creates a double-stranded structure, consisting of two exactly complementary sequences of A-T and C-G. The two strands wrap around each other, forming a double helix. DNA has a predefined directionality and information is always interpreted and copied into the cells in a defined order.

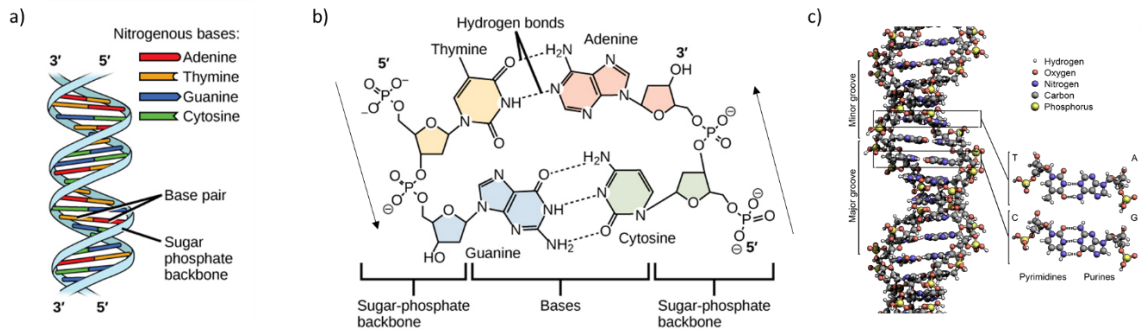


Figure 3. DNA double-stranded structure. a) Base pairing occurs between adenine and thymine, cytosine and guanine (purine-pyrimidine). These base couples are defined as complementary. b) The base pairs are stabilized by hydrogen bonds; adenine and thymine form two hydrogen bonds and cytosine and guanine form three hydrogen bonds. The two DNA strands are antiparallel in nature: the 3′ end of one strand faces the 5′ end of the other strand. c) The sugar and phosphate nucleotides form the backbone of the structure, while the nitrogenous bases are stacked on the inside. Each base pair is separated from the other base pair by a distance of 0.34 nm, and each turn of the helix measures 3.4 nm. The diameter of the DNA double helix is 2 nm. Only the pairing of a purine and a pyrimidine can explain the uniform diameter. Twisting the two strands around each other causes the formation of uniformly spaced major and minor grooves. Images courtesy [16, 17].

To perform its function, DNA needs to express genetic information to guide the synthesis of other molecules in the cell. This process, common to all living organisms, leads mainly to the production of two other key classes of polymers: RNA and proteins. Indeed, DNA does not directly code for protein synthesis, but instead employs RNA as an intermediary.

When a particular protein is needed by the cell, the nucleotide sequence of the appropriate portion of the long DNA molecule is first copied into RNA in a process called *transcription*. These DNA sequence copies on the RNA are used as templates to direct protein synthesis in a subsequent process called *translation*.
Genetic information always flows through *DNA à RNA à proteins*. All cells, from bacteria to humans, express their genetic information following this flow of information, a principle so fundamental called the "*central dogma of molecular biology*" [14].

GENES - The portions of the DNA that code for proteins are called genes, and genes are the molecular units of the hereditary [18]. Therefore, genes are portions of the genome located at precise *loci* in the DNA sequence (or, in for some viruses, in the RNA sequence) and contain the information needed to code for molecules such as RNA and proteins. During cell division and reproduction (mitosis and meiosis), the DNA and its genes are found in an extremely compacted form in the cell, the chromosomes. Each species has a certain number of chromosomes of a constant shape and size, humans have 46 of them.
A gene that is active in a given cellular context is evaluated according to its expression level, on which the downstream RNA and protein synthesis are dependent.

Upstream and downstream of the *protein coding region* are two transcribed but untranslated portions called the 5' untranslated region (5'UTR), with the important function of regulating the protein production process, and the 3' untranslated region (3'UTR), which regulates translation efficiency, the stability of the messenger and its location. Each gene is enclosed between a promoter region, in which the enzymes involved in transcription bind to begin the process, and a terminator, which is the portion where the RNA synthesis process ends. Finally, the region coding for the protein is composed of portions

that contain instructions, called *exons*, and unused portions called *introns*. Downstream and upstream of a gene, special sequences called enhancer, or silencer, may be present and they are used to facilitate or prevent the transcription process when bound to special proteins called *Transcription Factors*.

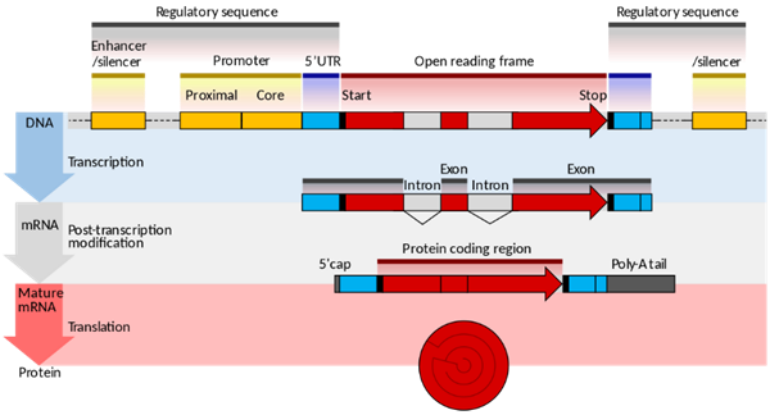


Figure 4. Structure of an eukaryotic protein coding gene. In yellow are depicted the regulatory features which determine where and when the protein coding sequence (shown in red) will be expressed. The gene is transcribed into a pre-mRNA which is spliced to remove introns and generate mature mRNA. 5′ and 3′ untranslated regions (UTRs) of the mRNA (shown in blue) direct which portions should be translated into the final protein product. Image courtesy [19].

Transcription and Translation - Through transcription and translation, cells express the genetic instructions in their genes. Each gene can be transcribed and translated with different efficiency, allowing the cell to produce large amounts of some proteins and small amounts of others. Depending on specialization, role and context, each cell can modify or regulate the expression of each gene, by regulating RNA production.

The first step is to copy a given portion of the DNA nucleotide sequence (gene) into a RNA nucleotide sequence. Despite the chemical form of RNA differing from DNA, the language in which the information is written is still essentially similar to the language of DNA (it is a nucleotide sequence). This explains why this process is called *transcription*.

The same segment of DNA can be used several times to drive the synthesis of many identical RNA transcripts.

RNA - *RiboNucleic Acid* is a polymeric molecule involved in different biological functions of coding, decoding, regulation and gene expression. It is composed of a slightly different sugar from DNA, the Ribose. RNA differs from DNA by one single base, as it has Uracyl (U) instead of Thymine. Therefore, those four bases pair up with their complementary counterparts in DNA: A-U, C-G. Furthermore, while DNA is found in cells as a double-stranded helix, RNA is single-stranded.

Transcription begins by unfolding a small portion of the double strand DNA to expose its bases on the two strands and thus to serve as a template for the synthesis of an RNA molecule. Any RNA which provides instructions to build a protein, is called messenger RNA (**mRNA**). The transcription process takes place in the nucleus by specific enzymes called RNA polymerases, which bind to a promoter and initiate the RNA production process. It is at this stage that Thymine is changed into Uracil. The RNA

chain produced by the transcription process is elongated one nucleotide at a time and its nucleotide sequence is exactly complementary to the strand of template DNA. The RNA strand does not remain hydrogen-bonded to the DNA template strand but, just behind the region where ribonucleotides are added, the RNA chain is displaced and the DNA helix is reformed. Thus, the RNA molecules produced by transcription are released from the DNA template as single strands.

The almost immediate release of the RNA strand means that many RNA copies can be produced from the same gene in a relatively short time, as the synthesis of other RNA molecules begins before the first RNA is completed. RNA polymerase operates at around 50 nucleotides per second, synthesizing more than a thousand transcripts from a single gene in an hour. The resulting RNA is processed and matures. The mature mRNA is transported into the cytoplasm where the translation process takes place.

Translation is carried out by ribosomes, cytoplasmic organelles, which use mRNA as a mold to produce a polypeptide sequence. Translation is accomplished by particular RNA molecules called transfer RNAs (tRNAs). Each tRNA associates a specific triplet of amino acids with a specific sequence of three nucleotides (codon). Once the translation of a given protein is finished, a stop codon is reached.

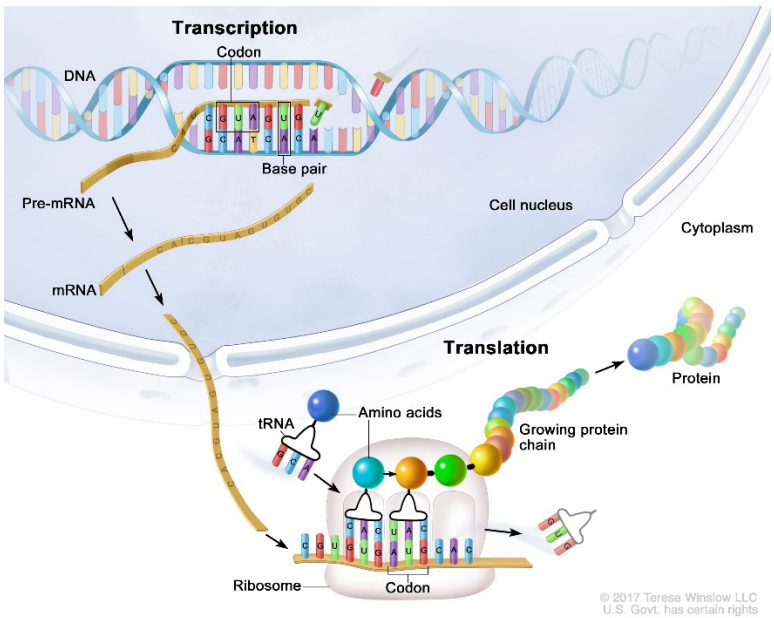


Figure 5. Transcription and translation. Transcription and translation are processes used by a cells to make all the proteins it requires to function. Transcription - the coding portion of DNA is copied into messenger RNA (mRNA) in the nucleus. The mRNA then carries the genetic information to the cytoplasm, where translation takes place. During translation, the mRNA attaches to the ribosome, which can read the genetic information to be translated into proteins. The transfer RNA (tRNA) carries an amino acid to the ribosome on the corresponding sequence in the mRNA. As each tRNA binds to the mRNA strand, that amino acid joins the other amino acids to form an amino acid chain. Once all the amino acids encoded in the mRNA piece have been linked, the completed protein is released from the ribosome. Image courtesy [18, 20].

Proteins - Proteins are large molecules made up of long chains of amino acids. To date, 20 amino acids are known, and they are synthesized directly from DNA. Proteins are therefore sequences of amino acids that differ from each other according to the order in which the amino acids are arranged. The sequence that makes up a protein is defined by the sequence of particular genes. Proteins are then organized at various levels of complexity and are typically folded to form a three-dimensional shape that will influence their activity. Proteins make up most of the dry mass of a cell and perform almost all cellular functions. They often work together to form what are called protein complexes. Once formed, proteins

exist for a limited period of time and they are then degraded and recycled through a cellular mechanism that determines their turnover.
The coding portions of DNA (genes) account for only 2% of the entire genome of eukaryotes. In fact, mRNA represents only 3-5% of the total RNA in a typical mammalian cell.

DNA has many non-coding protein regions that appear to have no function (junk DNA). However, many of these regions have been associated with important regulatory processes of gene activity. The product of non-coding genes is also RNA, and it is known as *non-coding RNA* precisely because it does not code for proteins. Therefore, non-coding RNA is a gene transcript that does not undergo translation.
non coding RNA (ncRNAs) - These include ribosomal RNA (rRNA), which is the most abundant type of RNA in the cell. It does not directly encode proteins, but is the essential component (about two-thirds) of ribosomes; nuclear RNAs (snRNAs), which direct the splicing of pre-mRNA (modification of the pre-mRNA, occurring along with or after transcription, in which introns are removed and exons are joined together) to give rise to mRNA; transfer RNAs (tRNAs) form the adaptors that choose amino acids and hold them in place on a ribosome to incorporate them into proteins; microRNAs (miRNAs) and small interfering RNAs (siRNAs) that serve as key regulators of the expression of eukaryotic genes; piwi-interacting RNAs (piRNAs) that protect the animal germline from transposons, genetic elements in the prokaryotic and eukaryotic genomes that are capable of moving 'transposons' from one position to another in the genome.
Among the ncRNAs, a brief description is devoted to miRNAs, since they are the ones that will be mentioned in this thesis. Indeed, they are included in the large network (meta pathway) that constitutes the knowledge base on which our simulator works.

Micro RNA (miRNAs) - microRNAs are small non-coding RNAs made up of approximately 22 nucleotides. They have been found not only in humans but also in animals, plants and some viruses. The main function attributed to miRNAs is regulating gene expression through post-transcriptional silencing [21-23]. They act through base pairing with the complementary sequence of the mRNA molecule. Once the pairing between miRNA and mRNA molecule has occurred, silencing can be achieved in several ways including cutting the mRNA into pieces, shortening the poly-A tail and destabilizing the mRNA molecule or reducing the efficiency of the translation process [24, 25]. miRNAs are produced from specific genes or introns of other genes and the DNA regions coding for miRNAs have a characteristic hairpin shape. The primary miRNA (pri-miRNA), once transcribed by RNA polymerase, is processed by the enzyme Drosha to free the hairpin from a pri-miRNA, which can contain more than one[24-26]. The transcript obtained is called precursor-miRNA (pre-miRNA) and this is then exported from the nucleus by a nucleocytoplasmic protein called Esportin-5 [27].
In the cytoplasm, pre-miRNAs are cut by the type III RNAase Dicer. The cut generates double-stranded RNA molecules that are approximately 22 bases long. Subsequently, miRNAs interact specifically with Argonaute proteins of the Ago subfamily and are incorporated into large ribonucleoprotein effector complexes called RNA-Induced Silencing Complexes (RISCs) where the interaction between miRNA and target takes place [28].

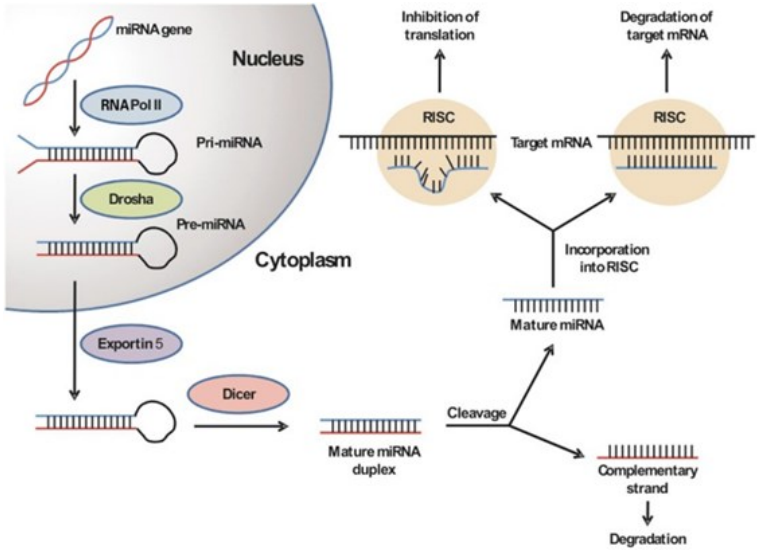


Figure 6. Overview of the endogenous miRNA pathway. In the nucleus The miRNA gene is initially transcribed into a primary miRNA (pri-miRNA) by RNA polymerase II. Subsequently Drosha enzyme processes the pri-miRNA into the 70 to 100 nt hairpin precursor miRNA (pre-miRNA), which is then translocated into the cytoplasm by Exportin-5. Here it is again cleaved by the ribonuclease Dicer enzyme into a mature miRNA duplex that through its guide strand binds to the RNA-induced silencing complex (RISC) to regulate gene expression by inducing either target mRNA degradation or translation repression, depending on the level of binding complementarity. Its complementary miRNA strand is released and degraded. Image courtesy [28, 29].

Genotype and Phenotype - All the genes possessed by each individual are called genotype, whereas the genetic component that confers all its observable characteristics, influenced both by its genotype and by the "environment", is called phenotype.

2.1.2 DNA Sequencing

DNA sequencing is the process for determining the ordered sequence of nucleotides in DNA. The development of advanced methods for DNA sequencing has led to a revolution in bio-medical research and discovery. The knowledge of DNA sequences is now fundamental for basic biological research, and in many applied fields such as medical diagnosis, virology and biological systematics.
The fast sequencing rate achieved with modern technology has been crucial in the complete sequencing of DNA, or genomes, of numerous organisms, including the human genome.
Sequencing data allows researchers to gain insight about changes in genes, associations with diseases and phenotypes, and identify potential drug targets. It was Frederick Sanger who in 1955 laid the foundation for protein sequencing by completing the sequence of all the amino acids in insulin, a small protein secreted by the pancreas. This served to demonstrate that proteins were chemical entities with a specific molecular pattern.
Sanger's discovery led *Crick* to develop the theory, published in 1958, which states that it is a specific arrangement of nucleotides in DNA that determines the sequence of amino acids in proteins, which in turn helps determine the function of the proteins themselves.
In 1970 at Cornell University the first method for determining DNA sequences was developed [30, 31].
In 1977 Frederick Sanger developed a more rapid DNA sequencing method at the MRC Centre, Cambridge, UK publishing a method for "*DNA sequencing with chain-terminating inhibitors*" in 1977 [32]. This approach, also known as *Sanger method*, after being developed, became the most widely used sequencing method for approximately 40 years and it is still considered the gold standard for sequencing.
Sanger sequencing is based on the selective incorporation of chain-terminating dideoxynucleotides (ddNTPs) by DNA polymerase during in vitro DNA replication [32, 33].
To perform sequencing, a single-stranded DNA template, a DNA primer, a DNA polymerase, normal triphosphate deoxynucleotides (dNTPs) and modified triphosphate dioxynucleotides (ddNTPs) are required, the latter is engineered to terminate DNA strand elongation. Indeed it lacks the 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, causing termination of extension of DNA when it is incorporated. The ddNTPs may be radioactively or fluorescently labelled for detection in automated sequencing machines.
Initially, Sanger's method required that all four DNA samples should be divided into separate tubes, each containing a specific ddNTP (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. The ratio of dNTPs to ddNTPs is approximately 100:1.
Next, the resulting DNA fragments were heat-denatured and separated by size using gel electrophoresis; the DNA bands visualized by autoradiography or UV light and the NDA sequence read directly on the X-ray film or gel image.
Sanger sequencing was the dominant method from the 1980s until the mid-2000s. Over that period, great advances in technology were made, such as fluorescent labelling, capillary electrophoresis, and general automation. It was through the Sanger method, in mass production form, that the first human genome was sequenced in 2001, ushering in the era of genomics.

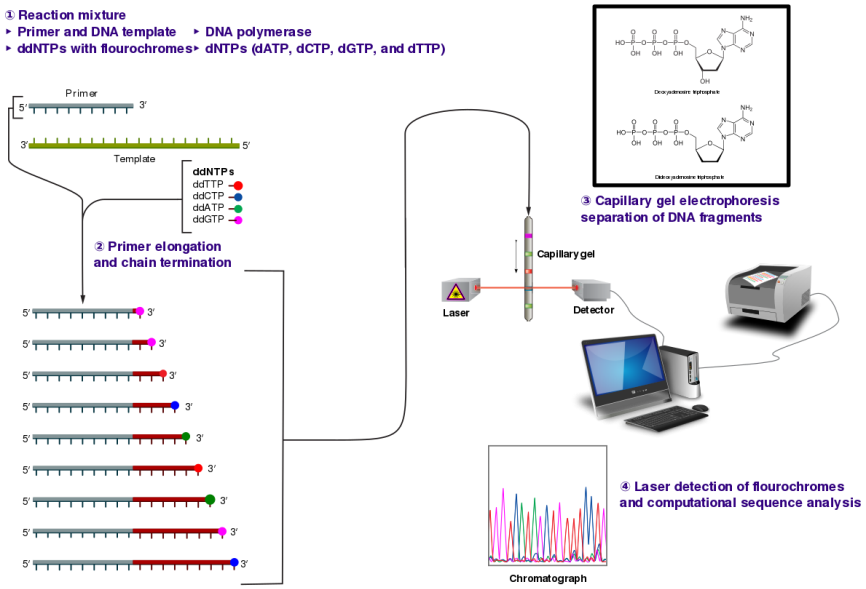


Figure 7. The Sanger method for DNA sequencing. (1) A primer is annealed to a sequence, (2) Reagents are added to the primer and template, including: DNA polymerase, dNTPs, and a small amount of all four dideoxynucleotides (ddNTPs) labeled with fluorophores. During primer elongation, the random insertion of a ddNTP instead of a dNTP terminates synthesis of the chain because DNA polymerase cannot react with the missing hydroxyl. This produces all possible lengths of chains. (3) The products are separated on a single lane capillary gel, where the resulting bands are read by an imaging system. (4) This produces several hundred thousand nucleotides a day, data which require storage and subsequent computational analysis. Image courtesy [34]

High costs and the need to reduce the time for acquiring more and more genomic data has been leading to the development of new, more efficient and faster methods.

With time, Sanger sequencing has given way to new high throughput technologies called "*Next-Generation Sequencing*"(**NGS**) or "*second-generation*" sequencing methods to distinguish them from earlier methods, including Sanger sequencing. These technologies parallel the sequencing process in a massive way, producing thousands or millions of sequences simultaneously (reads).

NGS technology is typically characterized by being highly scalable, allowing the entire genome to be sequenced at one time. For this reason, these sequencing techniques are also known as "*massively parallel*".

The first commercially available "next-generation" sequencing method was published and marketed by Lynx Therapeutics in 2000. This method, named *Massively Parallel Signature Sequencing* (MPSS), used a very complex approach to sequence four nucleotides at a time. However, this technology was simplified and made less expensive in subsequent years.

Basically, NGS is also defined as an "extension" method because the bases are identified during their addition to the parent chain. Very briefly the process could be summarized by starting with single-stranded DNA, a primer, DNA polymerase and labeled single nucleotides. Once the double-stranded DNA synthesis reaction is started, every time the DNA polymerase inserts a nucleotide on the elongating chain, it is immediately detected as a fluorescence signal specific for each of the nucleotides is released.

To date, there are several techniques for NGS that are increasingly accurate and cost-effective. Some of them are:

- Illumina: fluorescent sequencing, short reads;
- Life Technologies: pH sequencing, relies on the change in pH that occurs when a nucleotide is incorporated;

- Genereader NGS system (QIAGEN): fluorescence-based sequencing;
- Oxford Nanopore: single molecule sequencing based on the use of pores;
- 10X Genomics: short reads (usually sequenced on Illumina platform) that physically belong to the same DNA molecule (linked reads).

Through NGS techniques, it is possible to sequence:

Genomic DNA:
- Whole genome (the complete sequence - small genomes);
- Exome (only the DNA portion transcribed into RNA, exons);
- Targeted genes;
- Amplicons (only PCR products)

Transcriptome:
- Total RNA;
- mRNA;
- small RNA (<30 nt).

Epigenome:
- ChIP-Seq (Chromatin immunoprecipitation sequencing: DNA or RNA to which specific proteins are bound);
- Methyl-Seq (DNA methylation pattern study, epigenetics);
- Whole genome bisulfite sequencing (WGBS).

Extracting RNA expression

Quantification of RNA expression is a key factor in the analysis of biological and/or pathological processes. Among the techniques available today to sample gene expression, there are the **microarrays**. Microarrays are small solid supports, called *chips*, on which are immobilized, in fixed and known positions, thousands of DNA strand sequences (called *probes*) derived from different genes. Indeed, microarrays leverage on a technique of reverse hybridization consisting in fixing the probes on the support and labelling the nucleic acid we want to identify (target) to naturally form hydrogen bonds between pairs of complementary bases. A greater number of complementary base pairs implies a greater strength on the bond. Alla fine della reazione, la superficie dell'array viene lavata e rimarranno attaccati alle sonde solo i legami più forti. Through using color dyes, it is possible to identify where sequences have hybridized, and by comparing the color intensity between two different conditions, an estimate of expression can be determined.

In order to use microarrays with RNA molecules, the latter must be converted to complementary DNA (cDNA) through a process called reverse transcriptase.

Although microarrays are widely used and also relatively low-cost, they have some major limitations. First of all, the process of synthesis, purification, and storage of solutions necessary for manufacturing microarrays are extremely complex and expensive. Furthermore, when very similar RNA families are present in the sample, this technique becomes inaccurate because these molecules may hybridize to spots designed for other RNAs of the same family. This phenomenon is referred to as cross-hybridization. In addition, measuring the color intensity of microarrays can introduce biases in presence of overlapping spots or poorly expressed RNAs, where the color intensity is not sufficient and can cause failure in expression detection.

Cost reduction of NGS techniques has resulted in making these techniques suitable for detection of gene expression and they are defined as *RNA-sequencing* (**RNA-seq**). Therefore, these NGS techniques reveal

the presence and quantity of RNA in biological samples at a given time by analyzing the cellular transcriptome.
The use of these techniques has many advantages compared to microarrays. In particular, RNA-Seq facilitates the ability to observe alternative gene splicing transcripts, post-transcriptional modifications, gene fusion, mutations/SNPs, and changes in gene expression over time, or differences in gene expression in different groups or treatments [35]. In addition to mRNA transcripts, RNA-Seq can examine different RNA populations to include total RNA, small RNAs, such as miRNA, tRNA, and ribosomal profiles[36].
Recent advances in RNA-Seq include single-cell sequencing, in situ sequencing of fixed tissues, and sequencing of native RNA molecules with real-time single-molecule sequencing [37].
In order to analyze the data obtained from RNA-seq, it is necessary to use appropriate bioinformatics tools and subsequent analysis to validate the results obtained.

2.1.3 Omics Science

Omics sciences consist of those disciplines that, thanks to the use of advanced technologies of analysis, allow the production of a large amount of data, useful for the description and interpretation of the biological system studied (Fig.8).

Omics such as genomics, transcriptomics, proteomics, and metabolomics represent the data sources supporting systems biology, which aims to integrate data and provide predictive models for assessing the complex functioning of living systems. The first omics was genomics, as a discipline that studies and measures the set of genes in such an organism.

Whole DNA sequencing was rapidly followed by the development of innovative investigative technologies and there was a rapidly growing insight that knowledge of an organism's DNA gene sequence alone is not sufficient to understand its proper functioning. Thus, knowledge of the gene sequence does not take into account the effects of interaction with the overall environment.
This has led to change the previous hierarchical view of the functioning of a living system, described as a unidirectional flow between genes, transcripts, proteins and metabolites, into a view based on an interactive flow between the different levels of the system (gene system, transcripts, proteins and metabolites) and the external environment.

Similarly to the genome, systems were defined in terms of transcriptome, consisting of all messenger RNAs, proteome, consisting of all proteins, and metabolome, consisting of all metabolites present in a cell, tissue, organ, organism. The necessity to obtain quantitative data concerning the different component components (transcripts, proteins, metabolites), has led to the development of advanced technologies and new analytical disciplines (data-mining) that allow to interpret and summarize huge amounts of data.

Genome - The term genome, refers to the set of genes within a given organism and genomics is the discipline that investigates and measures that gene system.
This is the most advanced omics science and is focused on studying whole genome sequences and the information contained within them. Since the 1990s, several hundred genome sequencing projects have been completed on species representative of the three kingdoms of life. Of importance, the study of the human genome allowed the identification of groups of genes related to the development of diseases. As a consequence, groups of genes have been identified as related to the development of various human

diseases. This has directed the biomedical sciences to look for potential genetic markers of disorders, opening new perspectives in this field.

Transcriptome - The term transcriptome is derived from the two words *transcript* and *genome* to indicate the process of transcript production during the biological process of transcription. In general, the transcriptome is therefore the set of all transcribed RNAs, including coding and non-coding in an individual or a population of cells, although in some experiments it is used to refer to all mRNAs.

The advance of high-throughput technology has led to faster and more efficient ways of obtaining data about the transcriptome. To date, the most widely used techniques for studying the transcriptome are the DNA microarray, a hybridization-based technique, and the RNA-seq, a sequence-based approach, which is now the preferred method and the dominant transcriptomics technique.

The main experimental technologies used for transcriptomics are based on microarray techniques and Serial Analysis of Gene Expression (SAGE). The latter technique is based on the analysis of sequences of cDNA fragments derived from the reverse transcription of cellular or tissue RNA. The analysis allows us to evaluate the level of gene expression.

When physiological and pathological environmental conditions are altered and therefore also the exigencies of the cell change, it modifies its functions by modulating gene expression, thus affecting transcripts and, downstream, the translated proteins.

An in-depth analysis of the transcriptome allows researchers to obtain information on how the cell modulates the expression of certain genes rather than others, and thus on the biological processes associated with them, depending on specific contexts. Thus, for example, the transcriptome can be assessed during carcinogenesis, infections, metabolic dysfunction, for the detection of biomarkers, etc.

Transcriptomics has the potential to contribute to the development of new biomarkers useful for predicting disease progression and its potential response to treatments. Nonetheless, in order to obtain valid biomarkers, it is necessary to take into account that the analysis of transcripts may not be sufficient alone, since gene expression may be regulated at the post-transcriptional level.

Proteome - While the transcriptome is the set of all transcribed RNAs, the proteome allows the study of the proteins expressed in a cell (cells, organism, etc.), including all isoforms and post-translational modifications, according to the genome instructions. Effectively the term, coined by Wilkins, refers to *proteins* expressed by *genome* [38]. Proteins represent an increase in the level of biological complexity.

The proteome is dynamic over time, as it changes in response to external factors and differs substantially between different cell types of the same organism. Proteomics involves the large-scale study of proteins, particularly their structures and functions.

Nowadays, advanced techniques make it possible to obtain precise measurements of the molecular mass of polypeptides or of their proteolytic fragmentation products, and also to describe the primary structure of the polypeptides and thus obtain their complete identification.

These techniques are highly reliable due to the very high precision with which molecular masses can be measured. (errors no greater than 0,001%).

Metabolomics - This omics discipline aims to provide a quantitative measure of low molecular weight metabolites within cells, biofluids, tissues, or organisms. Collectively, these small molecules and their interactions within a biological system are known as *metabolome.*

To date, there is a large amount of published omics data available on databases such as Gene Expression Omnibus (GEO).

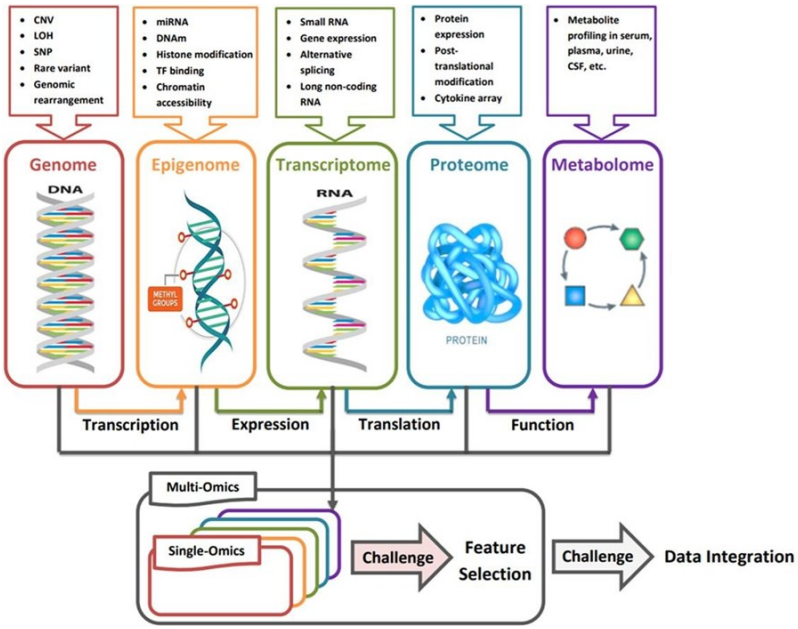


Figure 8. The overall diagram of relationship between single and multi-omics data analysis challenges.
Image courtesy (modified) [39]

2.1.4 Biological Pathways

A pathway is defined as a series of interactions among molecules within the cell leading to a certain product or to a change of the cell phenotype, displayed in a graph form. There are many types of biological pathways and each can describe different processes. Among the best known there are pathways involved in metabolism, gene regulation and signal transduction, those leading to the assembly of new molecules, such as a fat or protein, the cell cycle, etc. Pathway analysis allows the study and description of the change of a given biological process in response to external events or diseases, enabling a better understanding of its molecular details. In fact, for each pathway it is possible to inspect the contribution that each gene makes to determine its overall behavior.

A fundamental role in a pathway is played by its endpoints, that are those molecules which directly affect the phenotype, in relation to the phenomenon that is taking place at that moment.

Finally, it is very important to point out that the biological pathways that determine the functioning of our cells interact to form a larger network of interactions called the cellular interaction network. These interactions determine the overall behavior of each cell.

2.1.5 Host-Pathogen Interaction

Host is an entity that houses an associated microbiome/microbiota and interacts with microbes in such a way that the result is harm, benefit or indifference, resulting in the states of symbiosis, colonisation, commensalism, latency and disease [40, 41]. In the context of a host-pathogen interaction, the host is an entity that houses its own microbiota and interacts with pathogenic microorganisms. The result of this interaction is a trade-off between host, microbiota and pathogen [42, 43]**.**

Pathogen is any organism that can produce disease (bacteria, virus or fungi). Therefore, a pathogen refers to a microorganism as an infectious agent.

Host-pathogen interaction refers to how a pathogen sustains itself within host organisms on a molecular, cellular, organismal or population level and how the host reacts to the attack of a pathogen. Their interaction does not always result in disease. The host response to a microbial attack involves the activation of defenses at multiple levels, in other words it involves the activation of the immune system. This is why talking about host-pathogen interaction means describing all those biological processes in which the immune system plays a key role.

2.1.6 Immune System

The immune system is a network of biological processes designed to protect an organism from potential diseases. The immune system protects the host from infections and diseases with defenses of increasing specificity. To do this, the immune system is organized in a highly intricate and sophisticated manner. In many species, including humans, the immune system is categorized into an innate and an adaptive immune system.

Prior to the activation of the immune responses itself, every organism possesses physical and chemical barriers to the entry of the pathogen.
Effectively, the first encounter between pathogen and host takes place at the epithelial surface level comprising the skin, which lines the respiratory, digestive, urinary and reproductive tracts. The keratinized epithelial cells of the skin form a thick physical barrier; the sebaceous glands secrete fatty acids and lactic acid that prevent bacterial growth; epithelial cells lining internal organs, such as the respiratory and digestive tracts, secrete mucus that hinders the adhesion of agents and also contains substances that can kill or inhibit the proliferation of pathogens; cilia movements on epithelial cells lining the respiratory tract and represent an obstacle for pathogen adhesion [14].
However, physical and chemical barriers are not always sufficient to protect us from the invasion of pathogens.
When a pathogen successfully enters the body, the first form of defense comes from the *innate immune system*, which provides an immediate but non-specific response. It employs "*molecular sensors*" that recognize particular types or patterns of molecules that are common in pathogens. Innate immunity most of the time is efficient in destroying and clearing invading pathogens and in directing the development of an appropriate pathogen-specific *adaptive immune response*.

As this name suggests, the immune system adapts its response during an infection to improve its ability to detect the pathogen and improve its response. The adaptive immune system uses a class of white blood cells (leukocytes) called lymphocytes: B-lymphocytes (B-cells), which secrete antibodies that bind specifically to the pathogen, and T-lymphocytes (T-cells), which can either directly kill infected cells or produce signaling proteins, that are exposed on the cell surface or secreted, stimulating other host cells into contributing to the clearance of the pathogen.

The adaptive immune system then creates an immunological memory leading to an enhanced response to subsequent encounters with those pathogens. This process of acquired immunity is the basis of vaccination.
The innate response is active for a short time, whereas the adaptive response provides long-lasting protection.

Both innate and adaptive immunity have the ability to distinguish between the host's own molecules, called *self,* and “foreign” molecules, called *non-self.* Non-self molecules are called *antigens*, a term that means “*antibody generators*” [14]. They are defined as substances that bind to specific immune receptors leading to an immune response.

Innate Immune System

When pathogens successfully break through physical and chemical barriers and enter cells, the innate immune system, which corresponds to the non-specific response of the organism to pathogens, comes in action with the rapid intervention of a number of sentinel molecules that act as a proper alarm system warning about a potential infection in our cells. These molecules are protein receptors that recognize non-self molecules and initiate a sequence of events that are characteristic for innate immunity.

PRRs - Pattern recognition receptors (PRRs), as their name implies, recognize microbial molecules by the presence of characteristic repeated conserved patterns called *Pathogen-Associated Molecular Patterns* (PAMP).

PRRs can have different localizations thus differing also in the way they act. Some PRRs are transmembrane proteins present on the surface of many types of host cells, where they recognize extracellular pathogens. PRRs on cells such as macrophages and neutrophils may mediate the capture of pathogens in phagosomes, which then fuse with lysosomes, where the pathogens are destroyed. Other PRRs may be found within the cell either free in the cytosol or associated with the membranes of the endosomal system. Some PRRs, on the other hand, bind to the surface of extracellular pathogens, marking them for destruction via both phagocytes and proteins that are present in the blood that are part of the complement system.

There are several classes of PRRs:

TLRs - The first PRR to be described was the Drosophila Toll receptor, known for its ability to produce antimicrobial peptides that protect the midge from fungal infection [44]. Similar receptors were discovered shortly after in animals and plants and were called ***Toll-like receptors*** (**TLRs**). Mammals have about 10 TLRs, each recognizing distinct ligands such as: TLR3 recognizes viral double strand RNA in the lumen of endosomes, TLR4 recognizes the lipopolysaccharide (LPS) of the outer membrane of Gram-negative bacteria, TLR5 recognizes the protein that forms the bacterial flagellum, TLR9 recognizes short unmethylated sequences of viral, bacterial and protozoan DNA, called CpG motifs, which are not common in the vertebrates DNA[14].

NLR *NOD-Like Receptors* are a family of exclusively cytoplasmic PRRs that recognize bacterial molecules.

RLR *RIG- Like Receptors* are another class of exclusively cytoplasmic PRRs. They identify viral pathogens.

CLR *C-type Lectin Receptors* are a class of PRRs consisting of C-type lectin receptors. These are transmembrane proteins, present on the cell surface, which recognize carbohydrates on various microorganisms.

When activated by PAMPs, the numerous surface and intracellular PRRs stimulate the production of a wide variety of extracellular signal molecules that mediate the inflammatory response at the site of infection by activating intracellular signaling pathways that act on transcriptional regulators, including NFkB, to induce transcription of genes encoding the appropriate cytokines.

Some of the most important pro-inflammatory cytokines are *Tumor Necrosis Factor* α (**TNFα**), *Interferon γ* (**IFNγ**), various chemokines (which attract leukocytes) and several *Interleukins* (IL) such as **IL-1**, **IL-6**, **IL-12** and **IL-17**.

For instance, several cytoplasmic NLRs, once activated, assemble with adaptor proteins and specific protease precursors belonging to the caspase family to form *inflammasomes*, in which pro-inflammatory cytokines, such as IL-1, are activated by caspases. These cytokines are subsequently released from the cell by a secretion pathway that is not yet well understood. Apart from infections, NLR can also trigger the assembly of inflammasomes if cells are damaged or under stress for other reasons.

Complement System - The complement system consists of about thirty interacting soluble proteins produced continuously mainly by the liver. These proteins are known for their ability to amplify and "complement" the action of antibodies produced by B lymphocytes. Some complement components are also PRRs, which directly recognize PAMP on microbes. If there are no infections or special conditions requiring its activation, the complement system remains inactive. The destructive, inflammatory and self-amplifying properties of the complement cascade make it essential that key activated components are rapidly inactivated after being generated, to prevent the attack from spreading to neighbouring host cells.

Macrophages - In all animals, when the microbial invasor is detected, it is usually quickly encapsulated in a phagocytic cell. Macrophages are long-lived phagocytes found in most vertebrate tissues and are among the first cells to encounter pathogens, and whose PAMPs activate the secretion of pro-inflammatory molecules.

Neutrophiles - Neutrophils are short-living cells, abundant in blood but absent in normal healthy tissues, which are rapidly recruited to the site of infection by various types of molecules including chemokines secreted by activated macrophages. In the site of infection, neutrophils contribute with their pro-inflammatory cytokines. Neutrophils have a short half-life in the human blood system (several hours).

In addition to PRRs, macrophages and neutrophils possess various receptors on their cell surface that recognize fragments of complement proteins or antibodies bound to the pathogen surface. The binding of the pathogen to these receptors leads to its phagocytosis. Inside, phagocytes possess an impressive armamentarium to kill the invader, which includes enzymes, such as lysozyme and hydrolyse acids, that can degrade the pathogen's cell wall.
If a pathogen is too large to be successfully phagocytosed, a group of macrophages, neutrophils or eosinophils (another type of leukocyte) clusters around the invasor to attack it.

Blood and other extracellular fluids contain many proteins with antimicrobial activity, and some of these are produced in response to infection, while others are produced constitutively. The most important of these are the components of the complement system.

Dendritic Cells - Dendritic cells (DCs) are an important and heterogeneous class of cells belonging to the innate immune system. Immature DCs are located in most tissues of the body, for instance under the skin and intestinal epithelial layers, where they continuously select and process proteins present in their environment.
DCs express a wide variety of PRRs, which enable them to recognize and phagocyte pathogens and their products [14]. DCs become activated and mature when their PRRs encounter molecular patterns associated with pathogens (PAMP) or such products. When activated, they degrade pathogen proteins into peptide fragments, which bind to newly synthesized MHC proteins.

MHCs carry the fragments to the surface of DCs. The latter then migrate to a nearby lymphoid organ to present the MHC-peptide complexes to the T lymphocytes of the adaptive immune system to participate in fighting that specific pathogen.
In addition to MHC-peptide complexes, activated DCs also expose co-stimulatory proteins on their surface that help in activating T-cells to become effector or memory cell, and cell-cell adhesion molecules, which allow the T lymphocyte to bind to the dendritic cell for a sufficient time to be activated, generally several hours[14]. In addition, activated dendritic cells secrete various cytokines influencing the type of T-cell response, ensuring that it is appropriately tailored to fight the specific pathogen.

Dendritic cells provide the crucial link between the innate immune system, which provides a rapid first line of defense against invading pathogens, and the adaptive immune system, which organizes slower but much more powerful and specific responses.

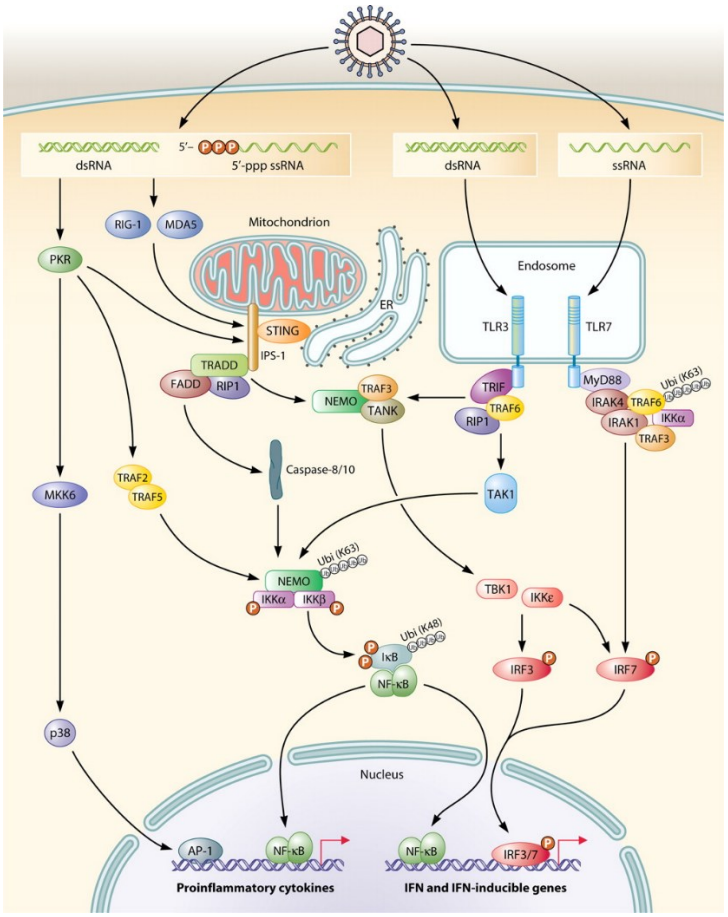


Figure 9. Overview Innate immunity pathway and intracellular RNA recognition and signaling. Innate immunity relies on the recognition of pathogen-associated molecular patterns (PAMPs) or endogenous danger signals through the detection of danger-associated molecular patterns (DAMPs) by pattern 1 recognition receptors (PRRs). Activation of PRRs triggers cell signaling leading to the production of proinflammatory cytokines, chemokines, and type 1 interferons, and the recruitment of phagocytic cells. The innate immune system comprises several classes of PRRs that allow for early detection of pathogens at the site of infection. Membrane-bound Toll-like receptors (TLRs) and C2-type lectin receptors (CLRs) detect PAMPs in extracellular and endosomal compartments. TLRs and CLRs cooperate with PRRs that detect the presence of cytosolic nucleic acids such as RIG-I like helicases/receptors (RLH/RLR). Another set of intracellular sensing PRRs are NOD-like receptors (NLRs) that can recognize PAMPs and DAMPs. Under stress (including infection and metabolic deregulation), some NLRs form high molecular weight complexes called inflammasomes. These autophagy-associated complexes play a central role in the control of innate and adaptive immunity. Figure shows the intracellular signaling involved in response to the foreign RNA recognition. Cytosolic dsRNA or 5′-triphosphate ssRNA is recognized primarily by the cytoplasmic RNA helicases RIG-I and MDA5, which mediate interaction with the mitochondria-localized adaptor IPS-1 and activate signaling to NF-κB and IRF3 through IKK and TBK/IKKε, respectively. The dsRNA can also be recognized by TLR3 located in the endosomal compartment

or by cytosolic PKR, but whereas TLR3 triggers signaling to NF-κB and IRF3, PKR instead activates NF-κB and MAPKs. Finally, ssRNA is recognized by TLR7/8 in endosomes and induces signaling to IRF7 as well as to NF-κB and MAPKs. Image courtesy [45, 46]

Innate immunity and viral infections

In case of viral infections, cells must adopt special strategies to try to prevent their replication. Since viruses use the host's molecular machinery to self-replicate, and host ribosomes will produce proteins and lipids forming the membranes of enveloped viruses, PAMPs are not present on the surface of those viruses. This problem is overcome by host cells because PRRs can recognize the presence of a virus by detecting unusual elements of the viral genome, such as double-stranded RNA (dsRNA), which is an intermediate in the life cycle of many viruses. dsRNA is recognized by several PRRs including TLR3.
Mammalian cells are highly efficient in recognizing the presence of dsRNA, which activates intracellular PRRs leading the host cell to produce and secrete two antiviral cytokines as a primary response: interferon α (IFN α) and interferon β (IFNβ)[14]. IFN α and β are described as type I interferons to distinguish them from IFN γ, which is a type II interferon and has different functions.

The production of type I IFNs is a primary cellular response to antiviral infection. They help in blocking viral replication in several ways and can act both in autocrine (on the infected cell that produced them) and paracrine (on adjacent uninfected cells) ways. Type I IFNs bind a surface receptor (IFNAR), which activates the intracellular JAK-STAT signaling pathway to stimulate transcription and the production of more than 300 proteins, including numerous cytokines. Thus, interferons are able to activate a latent ribonuclease, which nonspecifically degrades single-stranded RNA, and they also indirectly activate a protein kinase that inactivates the protein synthesis initiation factor eIF2, thus blocking the synthesis of most proteins in the cell. By adopting these strategies (destroying most of its RNA and blocking most protein synthesis), the host cell inhibits viral replication without killing itself. If those efforts fail, the host cell makes an even more extreme step to prevent virus replication: it commits self-suicide by apoptosis, often with the help of killer cells from the immune system.
Type I interferons can also have indirect ways to prevent viral replication. One of these is to increase the activity of natural killer cells (NK cells), which are leukocytes related to T- and B-lymphocytes but are part of the innate immune system and are recruited to inflammation sites very early. Similar to the cytotoxic T cells of the adaptive immune system, NK cells destroy virus-infected cells by inducing them to kill themselves by apoptosis.
IFN γ, which is a type II interferon, has the main functions of activating macrophages and inducing the expression of the Class II Major Histocompatibility Complex (MHC).

Adaptive Immune System

Differently from the innate immune system, which is programmed to react to broad categories of pathogens, the adaptive immune system is highly specific to each particular pathogen the body has encountered.
Adaptive immunity creates an immunological memory after a first response to a specific pathogen, leading to an enhanced response to future encounters with that pathogen.
This complex defense system is highly dependent on T and B lymphocytes (T and B cells). During their development, these cells can produce almost unlimited numbers of T (TCR) and B (BCR) receptors, rearranging particular DNA sequences in various combinations[14]. Collectively, these proteins can bind essentially to any molecule, including small chemical compounds, carbohydrates, lipids and

proteins; individually, they can distinguish very similar molecules, such as two proteins that differ by a single amino acid. Using this strategy, the adaptive immune system can recognize and respond specifically to any pathogen, including new mutant strains. Adaptive immunity can provide long-lasting protection, possibly extending for a person's entire life.
However, since the process of genetic rearrangement produces both receptors that can bind to self-molecules and receptors that can bind foreign molecules, vertebrates evolved specific mechanisms to ensure that B and T lymphocytes do not react against the host's own molecules and cells, a process called immunological self-tolerance[14]. In addition, many innocuous foreign substances enter our bodies and it would not make sense or be potentially dangerous to initiate adaptive immune responses against them. It is possible to trigger the adaptive immune system to respond to innocuous non-self molecules by co-injecting a molecule (often of microbial origin) called an adjuvant, which activates the PRR. This strategy is called immunization and is the basis of vaccination.
Any substance capable of stimulating B or T lymphocytes to initiate a specific adaptive immune response against it is called antigen or antibody generator.
There are two broad classes of adaptive immune responses: antibody responses and T-lymphocyte-mediated immune responses. Most pathogens induce both classes of response.

In **antibody responses**, B lymphocytes are activated to secrete antibodies, which are proteins transported in the bloodstream and capable of permeating other body fluids, where they can bind specifically to the antigen that stimulated their production. The antibody binding neutralizes extracellular viruses and microbial toxins by blocking their ability to bind the receptors on the host cell surface. The binding of antibodies also targets the pathogens for destruction, by either facilitating their phagocytosis and destruction by the phagocytes of the innate immune system, or by activating the complement system.

In **T-lymphocyte-mediated immune responses**, T-lymphocytes recognize foreign antigens that are bound to MHC proteins on the host cell surface, such as dendritic cells, which are specialized in antigen presentation to T-lymphocytes and are therefore defined as professional antigen-presenting cells (APC). Since MHC proteins transport protein fragments encoded by pathogens from the inside of a host cell to its cell surface, T lymphocytes can detect pathogens hiding in a host cell and can either kill the infected cell or trigger phagocytes or B cells to contribute to pathogen clearance.

B and T lymphocytes There are approximately 2×10^{12} lymphocytes in the human body, an amount that makes the immune system comparable in cell mass to the liver or the brain. Lymphocytes are abundant in the blood and lymph (the colorless fluid flowing in the lymphatic vessels, which connect lymph nodes to each other throughout the body and to the bloodstream) as well as in lymphoid organs, such as the thymus, lymph nodes and spleen; many are also located in other organs, such as the skin, lungs and gut.
T-cells and B-cells take their names from the organs in which they originate. T-cells are derived from thymus and B-cells, in adult mammals, are generated in the bone marrow[14]. Both types of lymphocytes develop from lymphoid progenitor cells, which are derived from multipotent hematopoietic stem cells, located mainly in the bone marrow.
The majority of B and T lymphocytes die in the primary lymphoid organs shortly after development, never becoming functional. Other lymphocytes, however, mature and migrate through the blood into peripheral lymphoid organs, mainly the lymph nodes, spleen and lymphoid tissues associated with the epithelium of the gastrointestinal tract, respiratory tract and skin. It is in these secondary lymphoid organs that foreign antigens activate T- and B-cells.

When B and T lymphocytes are not active, they are very similar, becoming morphologically distinguishable only once they have been activated by antigen. After activation by an antigen, both cell types proliferate and mature into effector cells. Effector B cells secrete antibodies; in their more mature form, called *plasma cells*, they contain an extended rugose endoplasmic reticulum responsible for antibody production[14].
Conversely, effector T cells contain a minimal percentage of endoplasmic reticulum and secrete various cytokines rather than antibodies. While antibodies produced by B lymphocytes are widely distributed by the blood circulation, cytokines produced by T lymphocytes act mainly locally on neighbouring cells, although some are carried by the circulatory stream and act over distant cells. The majority of antibody responses need T helper lymphocytes to begin.
The T-cell responses differ from those of B-cells. First of all, T lymphocytes are only activated to proliferate and differentiate into effector cells when antigen is shown on the surface of *Antigen-Presenting Cells* (APCs), usually dendritic cells in secondary lymphoid organs. T lymphocytes require APCs for activation, as the form of the antigen they recognize is different from that recognized by Ig produced by B lymphocytes.
Newly synthesized MHC proteins capture these peptide fragments and transport them to the host cell surface, where T lymphocytes can recognize them.
Effector T cells act only over short distances, within a secondary lymphoid organ or after migrating to the site of infection. Effector T cells interact directly with the targeted cells: they kill or mark it in some way. As for APCs, targeted cells must display an antigen bound to a MHC protein on their surface in order to be recognized by a T lymphocyte. There are three main classes of T lymphocytes: cytotoxic T lymphocytes, T helper lymphocytes and regulatory T lymphocytes.
T Cytotoxic effectors directly kill cells that are infected by a virus or some other intracellular pathogen.
T helper effector lymphocytes contribute stimulating responses from other immune cells, mainly macrophages, dendritic cells, B lymphocytes and cytotoxic T lymphocytes. There are several functionally distinct subtypes of T helper lymphocytes.
Regulatory effector T lymphocytes suppress the activity of other cells of the immune system.

MHC proteins capture and display peptide fragments of foreign proteins in order to present them to T lymphocytes. MHC proteins are divided into two main classes, which differ both structurally and functionally.

MHC class I - present peptides that are foreign to cytotoxic T lymphocytes. They are expressed by almost all nucleated cells in our body. Humans have three main groups of class I MHC proteins: HLA-A, HLA-B and HLA-C.
MHC class II - have peptides that are foreign to helper and regulatory T lymphocytes. This class of proteins is generally expressed only in APCs. All APCs charge their MHC class II proteins with peptides derived primarily from extracellular proteins that have been endocytosed and targeted to endosomes. Humans have three groups of MHC class II proteins: HLA-DR, HLA-DP and HLA-DQ.
The acronym HLA stands for human-leukocyte-associated, in fact these proteins were first identified in human leukocytes).

Immunological Memory - The most remarkable property of the adaptive immune system is its ability to respond to millions of different foreign antigens in a highly specific manner. Human B lymphocytes, for example, can collectively produce more than 10^{12} different antibodies which react specifically to the antigen that triggered their production[14].

When every lymphocyte develops in a primary lymphoid organ, it is designed to react to a particular antigen even before being exposed to it. The lymphocyte expresses this destiny in the form of receptors on its surface, which can specifically bind the antigen [14, 41, 42]. When a lymphocyte encounters its antigen in a secondary lymphoid organ, the binding of the antigen to the receptors activates the lymphocyte, which begins to proliferate, thereby producing many more cells with the same receptor, a process called clonal expansion[14]. Thus, an antigen activates only those lymphocytes that express complementary antigen-specific receptors, in other words, that have the appropriate characteristics to respond to it. This process, called clonal selection, provides an explanation for immunological memory, through which we develop lifelong immunity to many common infectious diseases after our initial exposure to the pathogen, either through natural infection or a vaccine.

Immunological memory depends not only on proliferation but also on the level of lymphocyte differentiation. In adults, peripheral lymphoid organs contain lymphocytes in at least three stages of maturation: naïve, effector, and memory cells. When naïve cells first encounter their specific antigen, they are stimulated to proliferate and differentiate into effectors that carry out the immune response (B effector cells secrete antibodies; T effector cells kill infected cells or influence the response of other cells, for example by secreting cytokines). Some of the naïve cells, triggered by the antigen, proliferate and differentiate into memory cells. These will be the first to become effectors if they ever encounter the same antigen again. Differently from most effector cells, which die within a few days or weeks, memory cells can persist for a lifetime. Most B and T effector cells die when the immune response is finished. A small portion will remain as effector cells helping to provide long-term protection against the pathogen.

Antibodies - Antibodies are proteins secreted exclusively by B lymphocytes as a defense against pathogens. They belong to the class of proteins called immunoglobulins (Ig) and are one of the most abundant protein components of the blood. The first immunoglobulins produced by a new B lymphocyte are not secreted, but incorporated into the plasma membrane as antigen receptors: B cell receptors (BCRs). Each B cell contains approximately 10^5 BCRs in its plasma membrane, which are all associated with an invariant transmembrane protein complex that activates intracellular signaling pathways once the antigen binds to the BCR.
When an antigen and a T helper cell activate a B lymphocyte (naive or memory), it proliferates and differentiates into an effector cell, which produces and secretes large amounts of soluble immunoglobulins called antibodies.
A typical Ig molecule possesses two identical antigen-binding sites and consists of four polypeptide chains, two identical light (L) chains and two identical heavy (H) chains. The N-terminal parts of both heavy and light chains generally cooperate to form the antigen-binding surface, while the more C-terminal parts of the heavy chains form the Y-shaped tail of the protein.

In mammals, there are five major classes of Ig: IgA, IgD, IgE, IgG and IgM, each with its own class of heavy chains, a, d, ε, g and m, respectively. In addition, there are four subclasses of IgG immunoglobulins (IgG1, IgG2, IgG3 and IgG4) with the heavy chains g1, g2, g3 and g4, respectively. In humans, there are also two subclasses of IgA. Each class (and subclass) has distinctive properties. IgM is always the first class of Ig produced by a B lymphocyte and develops in the bone marrow by forming the BCR on the surface of naïve immature B lymphocytes. IgM are secreted into the blood in the early stages of a primary antibody response, after initial exposure to the antigen.

T lymphocyte coreceptors: CD4 e CD8 The affinity of TCRs for MHC-peptide complexes on an APC is generally too low to mediate a functional interaction between the two cells. Thus, T lymphocytes need

accessory receptors to stabilize the interaction and increase the total strength of cell-cell adhesion. This increased adhesion allows the T lymphocyte to remain bound long enough to be activated.

An ancillary receptor that contributes directly in the activation of the T lymphocyte by generating its own intracellular signals is called a coreceptor. Among the most important are the CD4 and CD8 proteins, both single-pass transmembrane proteins with Ig-like extracellular domains.

CD4 is expressed on both helper and regulatory T cells and binds to MHC class II proteins;

CD8 is expressed on cytotoxic T cells and binds to MHC class I proteins.

CD4 and CD8 assist T-cell recognition by helping them to focus on particular MHC proteins and thus on particular target cell types.

A naive T helper cell, activated by the binding with a foreign peptide associated with an MHC class II protein on the surface of an activated dendritic cell, can differentiate into different types of effector T lymphocytes, depending on the nature of the pathogen and the cytokines it encounters: TH1, TH2, TFH, TH17, and regulatory (suppressor) T lymphocytes. These effector lymphocytes produce interferon-γ (IFNγ), which is crucial for triggering macrophages to destroy pathogens. IFNγ can also cause B lymphocytes to change the class of Ig they are producing.

Naïve TH lymphocytes activated in the presence of IL4 differentiate into TH2 lymphocytes, which are important for the control of extracellular pathogens, including parasites.

Naïve TH lymphocytes activated in the presence of IL6 and IL21 differentiate into follicular T helper lymphocytes (TFH), which are found in lymphoid follicles and secrete a variety of cytokines, including IL4 and IL21.

Naïve TH lymphocytes activated in the presence of IL6 and TGFβ differentiate into TH17 lymphocytes. These secrete IL17, which recruits neutrophils and stimulates skin and intestinal epithelial cells and fibroblasts to produce pro-inflammatory cytokines. TH17 lymphocytes are important in the control of extracellular bacterial and fungal infections and in wound healing, but they may also play a key role in autoimmune diseases and allergies.

Treg lymphocytes suppress the development, activation, or function of most other types of immune system cells, either through the secretion of suppressor cytokines such as IL10 and TGFb, or through the exposure of inhibitory proteins on the surface of Treg cells.

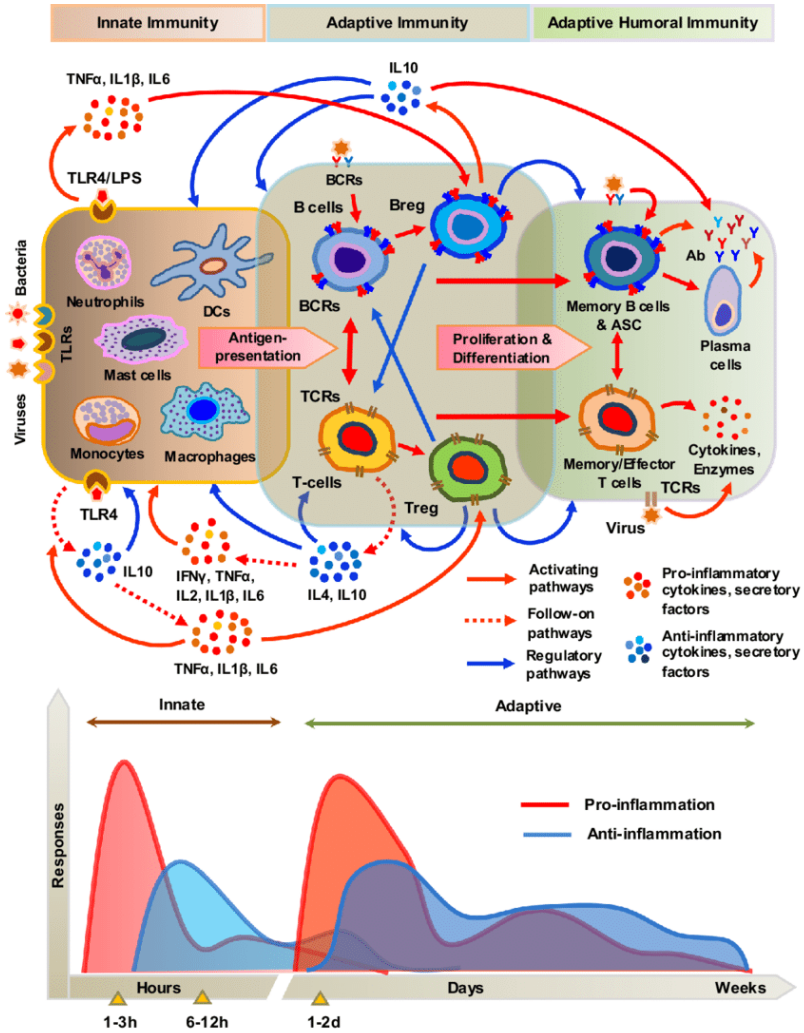


Figure 10. A simplified schematic diagram of the innate and adaptive immune response activating and regulatory pathways under normal physiological conditions. Image courtesy [38, 45]

2.1.7 Virus

Viruses are defined as obligate intracellular parasites that can replicate exclusively within the cells of host organisms. Viruses can infect all forms of life: animals, plants, microorganisms (including other infectious agents such as bacteria), and even other viruses.
When they are not in the active phase of infection or within an infected cell, viruses exist as independent, inactive particles also known as virions. Each virion consists of two or three parts: (I) the genetic material, which depending on the virus can be DNA or RNA; (II) a capsid, which is a protein coating that surrounds and protects the genetic material; and in some cases (III) a pocket of lipids that surrounds the protein coating when they are outside the cell. Virions can have simple, helical and icosahedral shapes as well as more complex architectures. Virions do not possess metabolism: they are therefore passively transported until they find a cell to infect. Infection of a host cell requires binding to specific membrane proteins.
In infected cells, viruses lose their structural individuality: they consist of the nucleic acids and their products that take over part of the cellular biosynthetic activity in order to produce new virions.
Alternatively, some viruses can physically insert their genome into that of the host so that it is replicated together with it. The viral genome inserted into that of the host, called a provirus, regains its individuality and produces new virions if the host cell is damaged.

Coronaviruses: main characteristics

Coronaviruses (CoVs) are a group of related RNA viruses that can infect the respiratory, gastrointestinal, hepatic, and central nervous systems of humans, livestock, birds, bats, mice, and many other wild animals[48-50].

Generally, human coronaviruses are members of the order Nidovirales which includes the families Coronaviridae, Arteriviridae, Roniviridae. Coronaviridae has two subfamilies: Coronavirinae and Torovirinae.
The Coronavirinae subfamily is further classified into four groups Alpha, Beta, Gamma and Delta. Among them, the first two (CoV α- and β-) infect mammals, γ-coronaviruses infect avian species, and δ-coronaviruses infect both mammals and avians.
Mild diseases in humans include some cases of the common cold (caused also by different viruses such as rhinoviruses), while potentially lethal strains can cause SARS, MERS, and COVID-19.

The first human coronaviruses to be identified were OC43 and 229E in the 1960s, followed by the identification of SARS-CoV in 2003, HCoV-NL63 in 2004, HKU1 in 2005, MERS-CoV in 2012, and finally SARS-CoV-2 in 2019[51].
For SARS-, MERS-, and SARS-CoV-2 zoonotic transmission is reported and they spread between humans through close contact. SARS-CoV-2 is relatively more infectious than SARSCoV and MERS-CoV probably due to different epidemiological dynamics.
CoVs are enveloped viruses with a single-stranded positive RNA genome and a nucleocapsid with helical symmetry[51, 52]. Their name is derived from their crown-shaped surface. CVs are spherical, polyhedral viruses, ranging from 80 to 160 nm in diameter. They have a positive-sense, single-stranded RNA genome (+ssRNA) that is one of the largest among RNA viruses, ranging from 26 to 32 kb in length [53, 54].
They are enclosed in an envelope embedded with a number of protein molecules [55]. The viral envelope is made up of a lipid bilayer in which membrane Glycoprotein (M), is the most abundant structural

protein of the virus. It reinforces the curvature of the membrane and attaches to the nucleocapsid. The envelope contains a small amount of a transient membrane protein known as the Envelope protein (E), which plays a role in virus assembly, release and pathogenesis [50, 56].
Nucleocapsid (N) is another viral protein that binds to the RNA genome, creating a symmetrical helical nucleocapsid. It has two domains, which can adhere to the RNA genome through various mechanisms [56, 57]. Characteristic are the club-like surface projections or peplomers, composed of trimers of a spike (S) protein, that are evident in electron microscopy images of coronavirus and which give them the classic crown from which they derive their name[57].
The S protein consists of two subunits S1 and S2. The homotrimeric S protein is a class I binding protein that mediates receptor binding and membrane fusion between the virus and the host cell. The S1 subunit has the receptor binding domain (RBD) instead the S2 subunit forms the stalk that anchors the tip into the viral envelope and, upon activation of the protease, allows fusion. The two subunits remain non-covalently bound as they are exposed on the viral surface until they attach to the membrane of the host cell[55].
The S1-S2 subunit complex is cleaved by host proteases such as those belonging to the cathepsin family and transmembrane serine protease 2 (TMPRSS2), when binding between the virus and the host cell occurs[58].
S1 proteins are critical components in terms of infection also because they are highly variable being responsible for host cell specificity.
The CVs +ssRNA genome has a 5′-terminal cap, a 3′-terminal poly (A) tail, and several open reading frames (ORFs). Their genome organization involves: 5′-leader-UTR-replicase (ORF1ab)-spike (S)-envelope (E)-membrane (M)-nucleocapsid (N)-3′UTR-poly (A) tail (Fig.11). ORFs 1a and 1b occupy the first two-thirds of the genome and encode for the replicase polyprotein (pp1ab) which cleaves to form 16 non-structural proteins (nsp1-nsp16) [58, 59].

Subsequent reading frames encode the four major structural proteins: spike, envelope, membrane, and nucleocapsid[60].

Spaced among the ORFs are accessory proteins that vary in number and function depending on the specific coronavirus[59, 60].

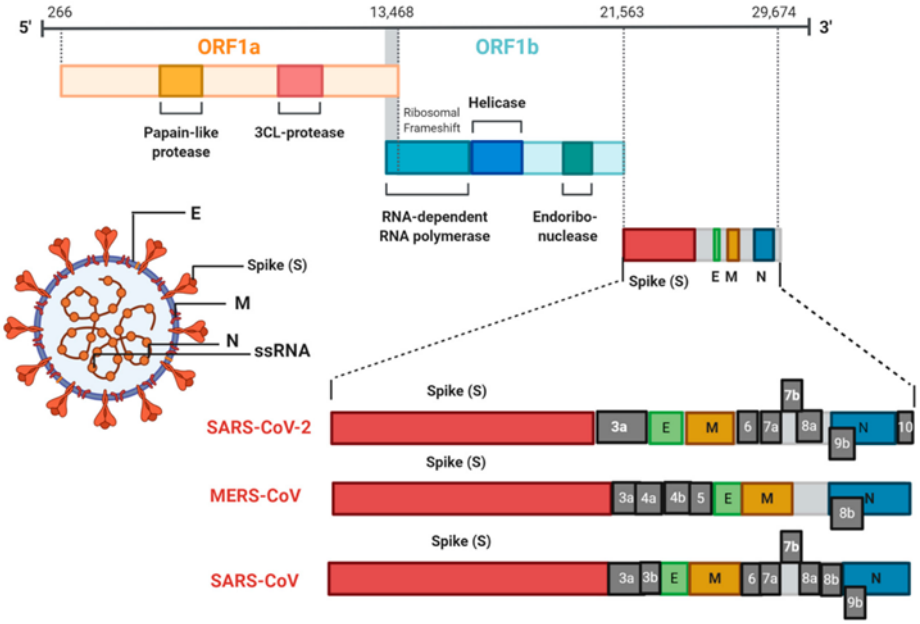


Figure 11. Coronavirus genomic organization. Coronaviruses are enveloped viruses with a positive-sense, single-stranded RNA of approximately 26-32 kb. They are spherical in shape and approximately 80-160 in diameter. The figure shows an overview of the genome organization of coronaviruses including: 5′-leader-UTR-replicase (ORF1ab)-spike (S)-envelope (E)-membrane (M)-nucleocapsid (N)-3′UTR-poly (A) tail. Notably represented are SARS-CoV, MERS-CoV, and SARS-CoV-2 that share ORF1 a/b encoding polyprotein pp1ab. The other ORFs are responsible for encoding the four major structural proteins: spike (S), envelope (E), membrane (M) and nucleocapsid (N) plus several accessory proteins. Image courtesy [61]

2.2 Mathematical and Algorithmic Prerequisites

2.2.1 Analysis of Biological data: fundamental of pathway enrichment analysis

Pathway enrichment analysis is a technique used to provide a reasoning for sets of biological elements generated by genome-scale experiments, particularly genes, in a specific context.
The accelerating availability of molecular sequences, particularly the sequences of entire genomes, has transformed both the theory and practice of experimental biology[62].

The richness and diversity of information progressively acquired, such as genes and different phenotypes, mutations with their pathological implications, proteins and their chemical characterization, have gradually emerged no more as different fields of research but as multiple aspects of a unified biology, highlighting the need to organize, describe, query and visualize biological knowledge at vastly different stages of completeness. So, the acquired knowledge has led to the construction of databases that categorize and describe the genes and the gene products for each living organism, in a very detailed way and, more importantly, with a standard format as for instance GeneOntology[62]. This has enabled the development of new methods that, using statistical approaches, can automatically reveal biological functions based on annotated information.

Knowledge Graph - Knowledge Graph (also known as a semantic network) is a systematic way to connect information and data to knowledge. It represents a collection of interlinked descriptions of entities, real-world objects, and events, or abstract concepts, obtained from knowledge-bases such as ontologies.

Ontology – It is a formal description of knowledge as a set of domain-based concepts in relationships among them. As a result, the ontology does not only introduce a shareable and reusable knowledge representation, but it can also provide new knowledge about the considered domain[63].

The Gene Ontology

The Gene Ontology (GO) is part of the wider Open Biological and Biomedical Ontology (OBO) project, created with the aim of unifying terminology in the biomedical field. It is considered one of the most important bioinformatics initiatives created with the aim to create a unified and standardized database of terminologies related to genes and biological functions of a wide range of categories, in order to facilitate the process of communication and data sharing. It defines concepts used to describe gene function, and relationships between these concepts. The Gene Ontology (GO) knowledge base has become and still remains one of the largest sources of information on gene function.
The project aims to develop and maintain a controlled vocabulary, to annotate genes and their products and also provide an easy access tool.
Within the GO database the terms are organized hierarchically. The most generic terms, at the highest level, are connected to their descendants by their relationship type, typically "is a" or "part of".
GO terms are organized in a Directed Acyclic Graph – DAG (Fig.12), where edges between the terms represent parent-child relationship. The level of specificity can be selected by the user.
GO classifies functions along three domains:
MF: Molecular Function - molecular activities of gene products;
CC: Cellular Component - parts of a cell or its external environment where gene products are active;

BP: Biological Process - pathways and larger processes made up of the activities of multiple gene products.

Every term within an ontology can be a word, a phrase, a unique identifier, a citation, and a domain to which it belongs. Synonyms are also often included.

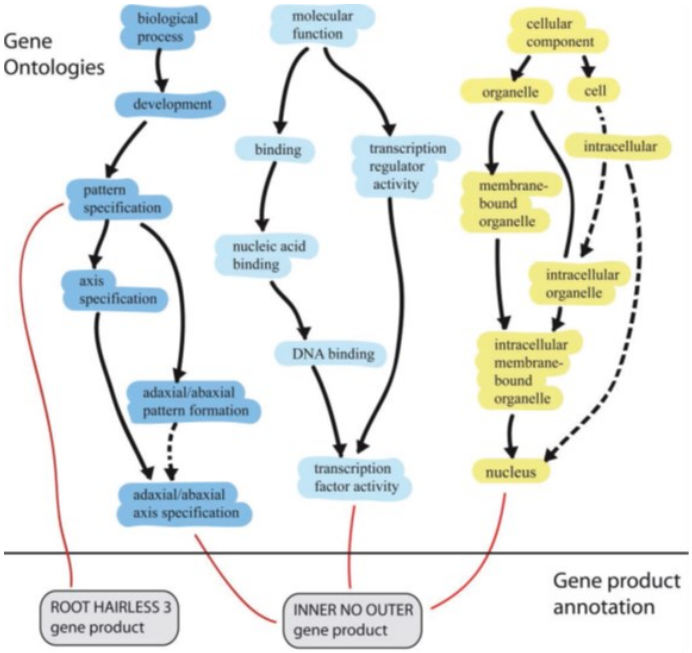


Figure 12. Ontologies - Directed Acyclic Graph This figure depicts the diagrams of the different sections of three ontologies, schematically: The biological process on the left side (dark blue), the molecular function in the middle (light blue), and the cellular component on the right side (yellow). The general concepts are at the top and the more specific concepts are at the bottom. The *is_a* relations (solid black lines) indicate that a child concept is a type of the parent concept, and the *part_of* relations (dashed black lines) indicate that the child concept is a part of the parent concept. Separately and concurrently with GO development, gene products are annotated (red lines) to the terms. The annotation indicates that the gene product (gray background, black outline) is involved in the process described, or has the function, or acts in the position described. Image from [64].

ReactomePA for enrichment analysis

ReactomePA is an R package designed for Reactome pathway based analysis[65]. It evaluates pathway associations with gene lists or genomic coordination obtained from high-throughput genomic and proteomic studies[65].

Reactome[4-6, 65] is a manually curated resource that describes chemical reactions, biological processes and pathways.

ReactomePA extends from the DOSE package[66] and supports hypergeometric testing and *gene set enrichment analysis* (GSEA)[66, 67] to provide Reactome and functional pathway analysis using variable NGS data.

Actually ReactomePA supports several model organisms, including *c.elegans*, fly, human, mouse, rat, yeast and zebrafish. It takes as input gene Entrez IDs.

The enrichPathway function allows users to select an appropriate background of genes as the baseline.

The gsePathway function supports GSEA to evaluate the enriched Reactome pathways of high-throughput data. In addition, ReactomePA provides several high-quality visualization features to facilitate the interpretation of the analysis. It is possible to graph the results including bar plot and dot

plot to summarize enrichment, cnetplot to visualize the gene-pathway association network, the enrichMap function to visualize the enriched pathway network, and gsea plot that displays the current sum of enrichment scores and its association with the phenotype.

2.2.2 Pearson Correlation

Among the different types of correlation coefficients, the Pearson correlation, also known as the linear correlation coefficient, is the most common correlation measure. The full name is *Pearson Product Moment Correlation* (PPMC) and it shows any linearity relationship between two statistical variables. A linear regression exists when the relationships between your variables can be described with a straight line.

Given two statistical variables X and Y, the Pearson correlation is defined as their covariance divided by the product of the standard deviations of the two variables:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma(X) \cdot \sigma(VY)}$$

The formulas return a value between -1 and 1, where:
A correlation coefficient of 1 indicates a strong positive relationship. It means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other.
A correlation coefficient of -1 indicates a strong negative relationship. It means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other.
Zero indicates no relationship at all. It means that for every increase, there isn't a positive or negative increase.
The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship.
Suppose we plot our linear relationship on a graph with one X axis and one Y axis, the X variable is sometimes called the independent variable and the Y variable is called the dependent variable.

DT-Web

DT-Web[104, 105] is a web-based interface to the Domain Tuned-Hybrid (DT-Hybrid)[104–106], which extends a well-established recommendation technique from domain-based knowledge that includes drug and target similarity.

This method, together with domain-specific knowledge expressing drug-target similarity, is used to calculate recommendations for each drug.

DT-Web can consider different matrices as input: known drug-target matrix, drug-drug similarity matrix, and target-target similarity matrix.

The drug-target interactions are taken from DrugBank, and from this data, an adjacency matrix is constructed. The drug-drug similarity is assessed using SIMCOMP[107] and then a similarity matrix is constructed. The target similarity matrix can be obtained by performing BLAST or using the Smith-Waterman local alignment technique.

Then, using these three matrices, a drug-target interaction network is constructed. Each target is mapped to its Entrez Identifier and annotated with Gene Ontology (GO) terms in this interaction network. For each pair of GO terms, the similarity score is calculated. Therefore, a p-value is calculated to evaluate the association between the predicted and validated targets.

DT-Web, given a set of candidate disease genes as input, can predict drug combinations whose targets are at an optimal distance from those genes.

Binding site parametrization

Binding sites are structural regions of macromolecules that bind ligands through interactions that are almost always reversible and can often be accompanied by conformational changes in the molecules. These are often conserved regions that can be used to search for other ligand-binding proteins that generally bind to other molecules by exploiting the structural similarity of these binding regions. Below, some of the methods designed to predict targets based on the binding sites of query molecules are explored.

ProBis

The ProBiS-ligands Web server predicts the binding of ligands to a protein structure. Starting with a protein structure or binding site, ProBiS-ligands identify model proteins in the Protein Data Bank (PDB) that share similar binding sites to the query [107, 108].

The algorithm uses the structure and physicochemical properties of the constituent amino acids and their backbones to compare two protein binding sites[108]. Then, it detects structures sharing similar 3D amino acid motifs to the searched protein within the PDB[108]. ProBiS-Database is a repository of non-redundant binding sites and associated PDB structures, which is updated weekly.

Pocket Similarity Search using Multiple-sketches - PoSSuM

PoSSuM searches the entire PDB database for binding similarity of all coupling molecules. Given a protein query, PoSSuM will search for all known ligand binding sites with a structure similar to the input. To obtain results, users can provide three types of input: protein structure; ligand binding site; and a ligand[108-110]. It uses a neighbor-searching algorithm called SketchSort. The similarity measure is determined based on cosine similarity and a p-value indicating significance [109, 110]. Dissimilarity values are given by the mean square deviation[109,110].

Disease-based approaches

This section is dedicated to tools that use disease association-dependent annotations. Disease-based approaches are used when drug pharmacology is not present or not considered.

MeSHDD

MeSHDD is a literature-based repositioning methodology that leverages drug-drug similarity based on the MeSH term co-occurrence[111]. MeSHDD, clusters drugs based on disease-centered Medical Subject Heading (MeSH) terms found in the MEDLINE Baseline Repository, which contains manually annotated MeSH terms for over 20 million biomedical articles, to predict shared indications[111].
MeSHDD uses drugs from DrugBank, including manually curated information on approved, investigational, and illicit drugs and their targets, mechanisms of action, and indications. Co-occurrence of drug-MeSH terms is calculated using a hypergeometric P-value, followed by a Bonferroni correction[111]. The drug-drug similarity is measured by calculating the bitwise distance from converting p-values to a binary representation. Drugs are clustered based on pairwise distances and bootstrap-means clustering techniques (implemented in R), and the Jaccard index was used to compare the clustering of various k-values[111].

RE:fine drugs

RE:fine drugs is a freely available interactive dashboard for integrated search and discovery of drug repurposing candidates from GWAS and PheWAS repurposing datasets constructed using previously reported methods in Nature Biotechnology[112].

Starting from a disease that users give as input, the tool returns a list of drugs that can be potentially useful for that case. Prediction results are classified as known/discovered if present in DrugBank, strongly supported if present in the NIH clinical trial registry and biomedical literature, Probable if the evidence is in the NIH clinical trial registry or biomedical literature, and Novel if not present in either[112].

Drug-induced gene expression to predict new connections

Drug-induced gene expression is the differential mRNA expression profiles in a cell line before and after drug treatment. This repurposing approach is accomplished by comparing disease-associated expression signatures with these drug-induced expression signatures, looking for drugs that have opposite effects on the disease and may be effective.

Connectivity map - CMap

CMap relies on a database of pre- and post-gene expression profiles from cellular samples in response to various types of perturbation, e.g., genetic perturbations in response to drug administration. CMap provides mRNA expression data from DNA microarrays for researchers who want to monitor differential expression to identify drugs that produce reverse signatures to query expression signatures. Connectivity are measured using the Kolmogorov-Smirnov statistical test. To date, CMap has generated a library containing over 1.5M gene expression profiles from ~5,000 small molecule compounds and ~3,000 gene reagents, tested in multiple cell types[8, 112, 113]. CMap has profoundly impacted therapeutic research and has opened new challenges in scientific investigations in drug repurposing, MoA elucidation, biological understanding, and systems biology[8][113].

Differentially expressed gene signatures - inhibitors DeSigN
DeSigN associates disease signatures with drug response signatures based on IC50 (quantitative measure of drug efficacy often used to prioritize compounds in vitro) data. Unlike CMap, which uses pre- and post-gene expression profiles, DeSigN uses only baseline gene expression profiles[114].

GoPredict
GoPredict uses gene expression data integrated with heterogeneous public information, such as signaling pathways and drug target information. It takes gene expression data as input and returns drug predictions as output. The reference databases used in GoPredict are TCGA, KEGGDrug, DrugBank, and Gene Ontology[115].

MANTRA 2.0
MANTRA 2.0 predicts molecular drug targets from gene expression profiles before and after drug perturbation in a collaborative and additive learning environment [115, 116].
An automated pipeline of MANTRA 2.0 transforms the gene expression profiles into a single drug "node" in the network and allows users to explore their neighbors to find new indications and interactions. It enables users to calculate a prototype ranked list (PRL) for each drug and then compare PRLs using a Gene Set Ensemble Approach (GSEA) based method[116].

NFFinder
NFFinder uses the MARQ method to compare molecular signatures. Performing this analysis requires two sets of expression data, up- and down-regulated genes compared to GEO, CMap, and DrugMatrix data[117].

Prediction of Drugs with Opposing Effects on Disease Genes - PDOD
The online server PDOD uses gene expression data and associates information regarding "effect-type" and "effect-direction" using KEGG pathway and drug target informations from DrugBank[117, 118]. It uses case/control expression datasets published in GEO to determine which gene expression changes happen due to a specific disease and looks for a drug that can counteract them[118].
To extract the gene signature, PDOD draws differentially expressed genes from the expression data by applying Limma and a function that evaluates the drug-disease score based on the parameterization of relationships[118].

2.2.3.3 Others Drug repositioning tools

Reverse Gene Expression Score - RGES
RGES is a system providing a predictive measure on how a given drug could reverse the gene expression profile for a given disease. The principle is to reverse overexpressed genes by increasing weakly expressed ones, restoring gene expression to levels closer to normal tissue[119].
First, the pipeline consists in calculating the gene expression signatures of the disease and the one generated by the drug-induced effect. From the two molecular signatures, the system then calculates the Reverse gene expression score (RGES) between the disease and the drug. This score ranges from -1 to 1, and it represents a measure of how much the drug under consideration can counteract the changes in expression due to disease. A low RGES value indicates higher potency to reverse disease gene expression and vice versa[119].

The data needed to perform the analysis can be taken from various publicly available databases such as TCGA, which includes gene expression profiles of tissue samples, LINCS, which includes perturber-mediated gene expression profiles, ChEMBL, which includes drug activity in tumor cells, and CCLE, which includes gene expression profiles of tumor cells[119]. Due to the progressively decreasing cost of many profiling technologies, large volumes of gene expression profiles of drugs under different biological conditions can be produced and made available to apply various drug repositioning and compound screening techniques such as RGES[119].

Searching off-lAbel dRUg aNd NEtwoRk - SAveRUNNER

SAveRUNNER is a freely available network-based algorithm for drug repurposing. Starting from a list of drug-target interactions and disease-gene associations, this tool predicts drug-disease associations by computing a new network-based similarity measure that prioritizes associations between drugs and diseases located in the same neighborhoods[120]. The pipeline consists first (i) in the construction of the proximity-based drug-disease network and then (ii) in the construction of a similarity-based bipartite drug-disease network.

The construction of the proximity-based drug-disease network comprises three phases:
Computation of network proximity (p) to measure how close the disease and drug modules are in the human interactome. Given two modules T and S that respectively represent the drug module, containing all t targets of the drug, and the disease module, comprising all s genes of the disease, p is described as the average length of the shortest path between the elements of T and S [120].
Computation of z-score proximity and p-values. SAveRUNNER calculates z-scores and their p-values by building a reference distance distribution corresponding to the expected distance between two randomly selected sets of proteins with the same size and degree distribution as the original sets of disease proteins and drug targets in the human interactome. The procedure is repeated 1000 times, and the z and its p-value are calculated through the mean and standard deviation of the reference distance distribution[120].
Statistically significant drug-disease associations (generally, p-value ≤ 0.05) are selected.
Next, the pipeline involves the construction of a similarity-based bipartite drug-disease network that comprises the following steps:

Computation of network similarity
The similarity measure is calculated from the network proximity measure p through the equation

$$similarity = max(p) - pmax(p) \qquad p = network\ proximity.$$

This measure assumes a value between 0 and 1[120].

Cluster detection
SAveRUNNER uses a clustering algorithm based on greedy optimization of the modularity network to define drug and disease groups. Each identified cluster is evaluated by the cluster quality score (QC)[120].

Adjustment of network similarity
The similarity of a drug-disease pair belonging to the same cluster increases proportionally to the QC score of the cluster. If the drug belongs to the same cluster as the disease then it might be considered eligible for repurposing. SAveRUNNER produces a list of predicted and prioritized drug-disease

associations in a weighted bipartite network format, where nodes represent drugs and diseases. A link between a drug and a disease occurs if the corresponding drug targets and disease genes are close in the interactome with a significant p-value ($p \leq 0.05$). Their interactions are represented by weighted edges in which the weight corresponds to the adjusted and normalized similarity value[120].

Bayesian ANalysis to determine Drug Interaction Targets – BANDIT

BANDIT is a machine learning algorithm that uses a Bayesian approach to integrate multiple data types and predict possible interactions with therapeutic effects. The rationale for this approach is integrating multiple data types to significantly improve the accuracy of target prediction[120, 121]. Indeed, BANDIT integrates data on drug efficacy, post-treatment transcriptional responses, drug structures, reported adverse effects, bioassay results and known targets[121]. The tool is based on a database containing approximately 2000 different drugs with 1670 different known targets and over 100,000 compounds without known targets (orphans)[121].

For each data type, a similarity score is calculated for all drug pairs with known targets. For each pair, BANDIT converts the similarity score into a likelihood ratio. These ratios are then combined to obtain a total likelihood ratio (TLR) proportional to the probability that two drugs share a target, given all available evidence[121].

The integrative approach of BANDIT can identify drugs that share targets, discern the mechanisms of approved drugs, explain existing but not fully known clinical phenotypes, and repurpose drugs for new therapeutic indications[121].

2.2.3.4 Data sources for drug repurposing

In the last decades, the gathering of genomic data has led to the acquisition of new knowledge on the genetic basis of diseases. It is enough to mention the numerous studies through which the association of gene loci with the risk of developing certain diseases has been discovered or the sequencing of human tumors, thanks to which somatic mutations underlying many types of cancer have been identified.

Thus, the acquisition of new knowledge about some disease phenotypes and drug-induced perturbations has increased the interest in new computational methods that can analyze and integrate large amounts of data to discover new disease targets.

These approaches have increased our understanding of the connection between genes, drugs, and disease leading to the generation of new hypotheses. Machine learning techniques and biomedical text mining approaches have been crucial in discovering hidden relationships between drugs and potential new therapeutic indications.

Systematic collection and analysis of gene expression data from human cell lines before and after drug treatment can be used to identify new opportunities for drug repurposing, discover new mechanisms of action for compounds, make small-molecule mimics of endogenous ligands, and predict side effects of such compounds[122].

In this direction, Connectivity Map was among the first databases to collect data about transcriptional responses of human cancer cell lines to various drugs/compounds and other small molecules. The first

version of this database had limitations due to its small scale, leading to the extension of the Connectivity Map project through the NIH Library of Integrated Network-based Cellular Signatures (LINCS) program. A new approach was introduced to increase the available experimental data. A cheaper technology than the classic RNA-seq, called L1000, was employed. The LINCS-L1000 provides the signatures of ~50 human cell lines in response to ~20,000 drugs (at various concentrations) for a total of over a million experiments[122].

In this section, I will provide an overview of CMap and its evolution LINCS L1000. These "big data" resources provide essential but straightforward platforms for characterizing small molecule-induced changes in gene expression and determining connections, similarities, or dissimilarities among diseases, drugs, genes, and pathways.

Connectivity Map - CMap
CMap, introduced in 2006 by Lamb et al., is a database collecting gene-expression profiles of drug-treated human cell cultures, which has been used for investigation of polypharmacology and drug repurposing.

Gene expression profiles are a series of experiments conducted using a microarray platform (Affymetrix HT_HG_U133 and HG_U133A) and standardized preprocessing (MAS 5.0). Experiments were done on different cell lines at different vehicle concentrations and time points compared to controls[123].

In the original CMap study, the initial reference database (Build 1) included 455 treatment-control pairs, where treatment constitutes a selection of 165 drugs, 42 different concentrations, 2-time points, and four human cell lines (MCF7, PC3, SKMEL5, and HL60). Subsequently, the database was significantly extended (Build 2), adding 1309 drugs with 156 different concentrations for a total of about 7000 gene expression profiles[123]. An “instance identifier” uniquely identifies each instance within the database. Thus, there is an instance representation in the reference database for each drug corresponding to treatment and control conditions[123].

The connectivity mapping methods
CMap's rationale is to use a reference database containing disease-specific gene expression profiles and compare it to the gene signature of a given drug. This approach is aimed to predict potential therapeutic candidate drugs. It also allows the identification of connections between drugs, genes, and diseases.

The CMap workflow comprises an initial query consisting of a set of gene signatures highly representative of a given biological state (e.g., disease). Although there is no definite way to generate the optimal gene signatures, the conventional approach identifies and uses a statistically significant list of differentially expressed genes (DEGs) calculated from disease and control samples. This list of genes will delineate the characteristic phenotype for a particular disease[123].

This kind of approach is platform-independent, allowing users to create query signatures from any gene expression platform[8, 123]. Then, the query is used to interrogate the CMap catalog.

Within the database, each of the signatures consists of a weighted average of the three biological replicate perturbations to mitigate the effects of unrelated replicates or outliers[8].

At this point, a connectivity score with a p-value is estimated using a non-parametric rank-ordered Kolmogorov-Smirnov (KS) test. The “connectivity score” is normalized through the random

permutation described by Lamb et al., assuming values from 1 to -1 to reflect the closeness between expression profiles[8, 123].

A positive correlation indicates the degree of similarity between a query signature and a perturbation-derived profile after specific treatment, whereas a negative correlation denotes an inverse similarity. These correlations are used to determine how exposure to a particular chemical may mimic or reverse the signature of the biological sample of interest.

A false discovery rate (FDR), which adjusts the p-value considering multiple hypothesis testing, and a t-parameter, which compares an observed enrichment score to all others in the database, are also calculated[8]. These metrics allow a comprehensive assessment of the relationship between a query and a perturbation, rather than just sorting by similarity.
Since the CMap method involves usage of expression profiles to define molecular signatures, it does not require prior knowledge of the detailed mechanism of action (MoA) or drug targets[8, 123]. This advantage makes it a widely used method in drug discovery and repositioning. The original CMap database had limited chemical and genetic perturbation data due to the high cost of commercial gene expression microarrays and RNA sequencing (RNA-seq). In addition, the expression profiles looked only at a few cell lines leaving the uncertainty of applicability to other cell lines, animal models, or human systems.
To improve the system and overcome these significant limitations, the same team of researchers developed a new simplified platform called L1000 to facilitate rapid and high-throughput gene expression profiles at a lower cost.

L1000
The L1000 platform, developed at the Broad Institute by the CMap team, is a method to facilitate high-throughput, low-cost gene expression profiling and is suitable for extending CMap at a large scale[8, 123].
The development of this method was part of the NIH LINCS (Library of Integrated Cellular Signatures) consortium, which funds the generation of expression profiles across multiple cell types and perturbations. To date, through L1000 technology, over one million gene expressions have been profiled and collected.

Its name, L1000, is because it contains a number of reference transcripts equal to 1000, used to estimate the signature of the whole genome gene expression generated by microarrays. Effectively, the basic idea is that it is possible to capture any cellular state by starting from a certain number of representative transcripts at a low cost. The authors used a set (12,031) of Affymetrix HGU133A expression profiles available in the Gene Expression Omnibus (GEO) to define the threshold for the number of transcripts. From this analysis, it was estimated that 1,000 landmarks were sufficient to recover 82% of the information in the entire transcriptome[8].
CMap and its updated versions provide a hypothesis-generating tool to identify new therapeutic targets (drug repositioning), signaling pathways affected by a compound, and search for new Mechanisms of Action (MoA), including potential side effects. It allows identifying new or known disease-gene-drug connections, depending on the observed level of changes.
To facilitate the fruition and use of this system, a platform called CLUE - CMap Linked User Environment has been developed. It can provide several analyses and allow access to all data at multiple levels of pre-processing via Gene Expression Omnibus (GEO: GSE92742)[8].

The L1000 LINCS currently includes over one million gene expression profiles of chemically disrupted human cell lines. Several resources and databases derived from L1000 LINCS data are available, for example, the L1000 Characteristic Direction Signature (L1000CDS2) search engine described below.

L1000CDS2

L1000CDS2 is a web-based search engine software designed to query gene expression signatures versus LINCS data to discover and prioritize small molecules that reverse or mimic the entered gene expression profile[122].

To compute the signatures, L1000CDS2 uses a multivariate method called the Characteristic Direction (CD).

The L1000CDS2 search engine prioritizes thousands of small molecule signatures and their pairwise combinations predicted to mimic or reverse an input gene expression signature. The L1000CDS2 search engine also predicts drug targets for all small molecules profiled by the L1000 assay[122].

Rather than giving relevance to fold-change and assigning greater weight to single genes that show a big fold-change, the CD method assigns a higher weight to genes that move together in the same direction. Thus, a gene that changes less but "moves" along with a large group of other genes may have more weight than a single gene that has changed more in magnitude[122].

The method first identifies the linear hyperplane that best separates control samples from treatment samples using linear discriminant analysis and then uses the normal to this hyperplane to define the direction of change in expression space for each gene[122]. Signatures can be accessed through an advanced web-based application called L1000CDS2[122].

The platform allows inserting the initial queries in dedicated sections (e.g. some up- or down- regulated genes or a complete signature), to customize the search by selecting optional parameters. The system also supports searching for paired combinations of small molecules[122]. After starting the search by clicking the Search button, the first 50 signatures are shown in a table on the results page[122]. Each entry provides seven columns of signature information: rank, score, perturbation, cell-line, dose, time point, and overlap with input[122]. It is possible to download all the information about a signature as a JavaScript Object Notation file (JSON)[122]. Results can be downloaded in table format to a .csv file.

L1000CDS2 also allows users to perform enrichment analysis on the substructures of the best classified small molecules. The enrichment analysis results are displayed as a table where each row provides three pieces of information: the substructure, the p-value (calculated using Fisher's exact test), and the perturbation count. Enrichment analysis results can be shared through email, publication, or other documentation using a permanent URL provided on the page. Interestingly, there is a function that allows users to share their input signatures and metadata so that others can query those signatures[122].

The user may also decide to search for combinations of small molecules. In this case, L1000CDS2 compares each possible pair among the first 50 matching signatures and calculates the potential synergy between each pair by examining the level of orthogonality. The synergy score is calculated as the combined overlap of the differentially expressed genes of the two drug signatures with the input gene lists[122].

3

Related Works

The Systems Biology approach for in-depth analysis of cellular and molecular mechanisms underlying diseases and for the realization of a new methodology for drug repurposing, requires the use of algorithmic techniques of different nature.

Despite the unprecedented growth in our understanding of cell biology, it still remains challenging to link it to experimental data obtained with the pathophysiological state of cells and tissues under specific circumstances. Recently, computational approaches in systems biology have emerged as efficient means to bridge the gap between systems-level experimental biology and quantitative sciences[124].

Here, network analysis is playing a central role in modeling and understanding biological phenomena and, in this direction, algorithms that enable in-depth pathway analysis become key instruments. These algorithms, based on current biological knowledge, allow us to learn more about the characteristic disease-related phenotypes, classify them and make new hypotheses. In this perspective, *in silico* simulation methodologies can also assist in understanding the intricate patterns of interaction between molecular entities, significantly improving manual analysis. Moreover, *in silico* simulations can be extensively applied at massive scales, testing thousands of hypotheses under various conditions, which is usually experimentally impossible.

Computational analysis also allows us to filter results on the basis of the most promising hypothesis, becoming a valuable system to support experimental choices, helping to make well focused lab schedules, reducing time and costs.

In the following paragraphs will be described the algorithms used in my research work for pathway analysis and *in silico* phenotype simulations.

Pathway analysis is typically used in Omics data analysis to gain biological insights into the functional roles of predefined subsets of genes, proteins, and metabolites. Nowadays there are numerous methods proposed in the literature for this purpose. The method for pathway analysis used in this thesis, MITHrIL is a latest generation method that exploits not only information about the individual perturbed entities (genes, proteins, metabolites) and their relative level of deregulation (measured by LFCs) but also information about the topology of the underlying pathways, which, as the evidence from their evaluation reveals, results in improved sensitivity and specificity.

In silico simulations, on the other hand, will be performed using PHENSIM[1, 124], a computational tool using a systems biology approach to simulate how cell phenotypes are affected by the

activation/inhibition of one or multiple biomolecules, and it does so by exploiting signaling pathways. PHENSIM requires a set of nodes (at least one) together with their "deregulation type" (up-/down-regulation) as input to compute synthetic Log-Fold-Changes (LogFC) values that are then propagated within biological pathways using the MITHrIL algorithm Alaimo et al. 2016[125] to establish how these local perturbations can affect the cellular environment.

Since these models are based on knowledge networks such as KEGG or Reactome, it is necessary to address the problem of model incompleteness. Indeed, these networks contain partial information that could affect the success of *in silico* predictions. In this regard, a new system called NETME[10, 125] will be presented. This system, starting from a set of full texts obtained from PubMed, through an easy-to-use web interface, interactively extracts biological elements from ontological databases and then synthesizes a network by inferring relationships between these elements.

NETME allows to integrate large biological networks used for pathways analysis with missing information about genes, proteins, metabolites, drugs, etc., helping to develop more accurate and precise *in silico* models.

3.1 MiTHrIL - Mirna enrIched paTHway Impact anaLysis

The prediction of phenotypes, such as that related to diseases or to responses to therapies, starting from the large amount of genotypic high-dimensional data obtained through Next-Generation Sequencing techniques, is an extremely important task in translational biology and precision medicine [125].
These technologies enable the generation of a list of differentially regulated elements (genes or microRNAs) whose behavior varies significantly across phenotypes and in relation to different pathophysiological conditions.
Generally, to extrapolate from NGS data new insights about the biological processes in which differentially expressed genes are involved in a given phenotype, genes are grouped into smaller subsets according to some relationships that leverage on existing knowledge-bases such as ontologies or pathways. The analysis of this type of data at the functional level is crucial since it allows a strong reduction of dimensionality, thus providing greater insights on the biology of the phenomenon under study[125, 126].

An extensive class of techniques known as Pathway Analysis[127] goes in this direction. More recently, great interest has shifted toward a class of methods called Knowledge base-driven pathway analysis[128]. Those methods rely on existing databases, such as the Kyoto Encyclopedia of Gene and Genomes (KEGG)[2, 3,128] or Pathway Commons [129], to identify pathways that may be affected by expression changes in the observed phenotype.

There are three generations of approaches into which Knowledge base-driven pathway analysis techniques can be classified: i) Over-Representation Analysis (ORA); ii) Functional Class Scoring (FCS); iii) Pathway Topology-based (PT).
ORA methods statistically evaluate the number of deregulated genes in a pathway with respect to the set of all analyzed genes. These methods may be limiting because, by considering only the number of differentially expressed genes, while omitting their expression level, implies that their magnitude of change is not considered as important to the activity of the pathway. Furthermore, taking into account only statistically differentially expressed genes may lead to the exclusion of those genes whose coordinated alteration may lead to substantial effects, even though their differential expression may not be statistically significant.
Finally, they consider individual genes and pathways, respectively, in a manner independent of the surrounding biological context [125].

FCS methods compute a gene-level statistic from the expression levels, by means of a statistical approach (i.e. ANOVA, Q-statistic, signal-to-noise ratio, t-test, or Z-score). Such a statistic is calculated considering all genes in a pathway [130, 131] and its statistical significance is estimated through an appropriate null hypothesis [132-134].

This method identifies Functional Gene Sets by taking into account their relative positions in the complete list of genes studied and their expression level. One of the first and most popular methods using the FCS approach was Gene Set Enrichment Analysis (GSEA).

However, by using only expression values to compute the gene-level statistic, they do not take into account the magnitude of their deregulation when estimating pathway activity[125].

Finally, third-generation pathway analysis methods, PT, use specific topological information about the role, location, and interaction directions of elements (genes or other biomolecules) in the pathway to

compute scores. Effectively, pathways are modeled as graphs, where nodes represent genes and edges represent interactions between them.

MITHrIL (miRNA enriched pathway impact analysis) algorithm is a third generation method that extends the Draghici et al. [134, 135] and Tarca et al. [136] techniques (both third generation methods).

An important feature introduced with MITHrIL, that distinguishes it from the other pathway analysis techniques, is the extension of KEGG[2, 3 136]. pathways with information regarding microRNAs (miRNAs) and their interaction genes resulting in an improvement of a knowledge base with 10,537 experimentally validated interactions between 385 miRNAs and 3,080 genes.

These interactions are taken from validated databases as miRTarBase[3, 137] and miRecords [138]. The algorithm integrates also interactions between transcription factors (TFs) and miRNAs from TransmiR[139], increasing the knowledge stored within each pathway.

Mithril, starting from expression values of genes and/or microRNAs, returns a list of pathways sorted according to the degree of their deregulation, together with the corresponding statistical significance (p-values)[125, 139].

Alaimo et al. 2016 have proven that MITHrIL gives the best performances compared with PARADIGM[140], SPIA[136] and Micrographite [141] by employing the technique defined in Vaske et al. 2010[142] on a set of cancer types. The authors showed that taking into account the network topology and essential regulatory elements such as microRNAs, increases the results reliability. Indeed, miRNAs have been revealed to be crucial in the modulation of numerous cellular pathways via the exertion of their important regulatory function when targeting key genes[143, 144].

In contrast to SPIA, MITHrIL also returns the estimated perturbation for pathway endpoints. Importantly, the proper assessment of pathway endpoints can contribute to a much more accurate phenotype evaluation, as more detailed diversification among pathway-level disease phenotypes is reflected more in the endpoints than at any other node in the pathway network. This allows pathologies that also share a very similar set of deregulated genes to be distinguished more effectively.

Finally, in order to acquire information on which endpoints are contained in each pathway, MiTRiL employed a depth-first search algorithm (DFS) [145] to automatically mark which genes are located at the end of the chains of reactions in each pathway. The search for endpoints in a pathway starts from a random node. The DFS algorithm follows the interactions down to the nodes from which no other one can be reached (putative endpoints)[125, 145].

To start the MITHrIL analysis it is necessary to have a case/control expression data set from which statistically differentially expressed features have been extracted (genes, miRNAs, or both). The system requires as input the list of differentially expressed elements with the relative Log-Fold-Change.

Starting from such information, MITHrIL computes, for each gene in a pathway, a Perturbation Factor (PF), which is an estimate of how much its activity is altered considering its expression and 1-neighborhood[125]. Positive (negative) values of PF indicate that the gene is likely activated (inhibited)[125]. By combining each PF of a pathway, MITHrIL computes also an Impact Factor (IF) and an Accumulator (Acc). The IF of a pathway is a metric expressing how important are the changes detected in the pathway, the greater the value, the more significant are the changes[125]. The Acc indicates the total level of perturbation in the pathway and the general tendency of its genes: positive

Acc values indicate a majority of activated genes (or inhibited miRNAs), while negative ones corresponds to an abundance of inhibited genes (or activated miRNAs)[125]. Mithril calculates a p-value for the Acc, which estimates the probability of getting such accumulator by chance. Finally the Benjamini and Yekutieli [146] method is applied to estimate the false discovery rate and p-values are adjusted on multiple hypotheses [125, 146].
Therefore, the final result of the Mithril algorithm consists of a list of pathways along with their impact factor, accumulator and adjusted p-values.
Starting from such output it is possible to perform data analysis to extract new knowledge and informations.

3.2 PHENSIM - Phenotype Simulator

Nowadays, many simulation models have become available. They can be grouped into two broad categories: (i) discrete/logic or (ii) continuous models[147]. Discrete models represent each element's state in a biological network as discrete levels, and the temporal dynamic is also discretized. At each time step, the state is updated according to a function, determining how an entity's state depends on the state of other (usually connected) entities. Boolean networks [148, 149] and Petri nets [149, 150] represent two types of discrete models.

BioNSi (Biological Network Simulator)[151] is a tool for modeling biological networks and simulating their discrete-time dynamics, implemented as a Cytoscape 3 plugin [152], that uses KEGG pathways[153] as a network model. At each simulation time point, the state of a node is updated using an effect function. The simulation ends as soon as it reaches a steady state. The model is easy to use.

However, a more complex biological network might pose challenges to its performance.

Continuous models usually produce real continuous measurements instead of discretized values, simulating network dynamics over a continuous timescale. Although they could provide a greater degree of accuracy, these methods are limited by our current description of the biological systems and our measurement techniques' capabilities. Continuous linear models [154, 155] and flux balance analysis [156] are the most representative continuous models.

Pathway modeling is an essential step for building networks that simulation methodologies can use.

SBML is an open and interchange format for computer models of biological processes. However, converting pathways in annotated SBML files suitable for simulation models is not easy. Several tools such as KEGGconverter [156, 157] or KENeV [158] have been specifically developed for this objective. These tools can also consider crosstalk with neighboring pathways, providing improved simulation accuracy. However, KEGGconverter has not been updated recently, and KENeV does not integrate post-transcriptional regulatory interactions or Reactome pathways.

PHENSIM (PHENotype SIMulator), is a web-based, flexible, user-friendly pathway-based simulation technique, and an in silico tool based on it, allowing phenotype predictions on selected cell lines or tissues in 25 organisms, including models such as *Homo sapiens, Mus musculus, Rattus norvegicus, and Caenorhabditis elegans.*

PHENSIM has been mainly developed to predict the effects of one or multiple molecular deregulations on cell/tissue phenotype. Thus, we view PHENSIM as an easy-to-use, supportive pathway-based method that can make predictions of in vitro experiments targeting the expression of signaling processes' activity.

PHENSIM uses a probabilistic algorithm to predict the effect of deregulated (up/down) genes, metabolites, or microRNAs on the KEGG **meta-pathway**[159].

The **meta-pathway** is a network obtained by merging all KEGG pathways through their common nodes. This approach allows us to consider pathway crosstalk and, ideally, gives a more comprehensive representation of the human cell environment. Furthermore, the KEGG meta-pathway is annotated with experimentally validated miRNA-target and Transcription Factor-miRNA interactions to consider post-transcriptional expression modulation.

Currently, PHENSIM uses all KEGG pathways (downloaded on April 2020) with details on validated miRNA-targets inhibitory interactions downloaded from miRTarBase (release 8.0) [159, 160] and miRecords (updated to April 2013) [138], and TF-miRNAs interactions obtained from TransmiR (release 2.0) [138, 161]. Furthermore, since the method's architecture is easily extensible, we include the possibility of integrating Reactome pathways to the meta-pathway environment, yielding a richer and more comprehensive model.
Reactome is a free online database of biological pathways [4-6]. It includes databases of reactions, pathways and biological processes of several organisms. However, the largest one is dedicated to human biology. Reactome offers visual representations of biological pathways in full mechanistic detail, making source data available in a computationally accessible format. The nodes of the Reactome network therefore are entities (nucleic acids, proteins, complexes and small molecules) that participate in reactions, the latter being the arcs of the network, forming a network of biological interactions grouped into pathways (e.g. signaling, innate and acquired immunity, apoptosis, metabolism, etc.).
The pathways in Reactome are species-specific and experimentally validated. When there is no experimental validation that supports certain interactions, pathways may contain manually inferred steps from non-species-specific experimental data. This occurs only if an expert biologist, designated as the pathway author, and a second biologist, designated as the reviewer, concur to make such deductions as a valid one.
To date, Reactome contains a more extensive network than that of KEGG, which is why such extension becomes very important.

To start a simulation, PHENSIM requires a set of nodes (at least one) together with their "deregulation type" (up-/down-regulation) as input values. We can also provide: (i) a list of non-expressed genes, (ii) a set of new nodes or edges that will be added to the meta-pathway, and (iii) the organism.

The list of non-expressed genes is useful to specify the context (e.g. a specific cell line or tissue) in which we wish to perform the simulation. Generally, genes that report a value below a certain threshold in the expression data are considered to be unexpressed.

The addition of a set of new nodes or edges to the meta-pathway is very useful in case we want to simulate the action of "new entities" such as drugs, or in case we need to add to the meta-pathway missing nodes that are essential for the specific process to be analyzed.

Finally, since PHENSIM can work on different species, it is necessary to specify the organism in the appropriate section.

PHENSIM uses the input to compute synthetic Log-Fold-Changes (LogFC) values. These values are then propagated within biological pathways using the MITHrIL algorithm proposed in Alaimo et al. 2016[125] to establish how these local perturbations can affect the cellular environment.

This propagation result is called a "*Perturbation*," reflecting the change of expression for a gene in a pathway (negative/positive for down-/up-regulation). This value is computed for each gene in the meta-pathway. Finally, PHENSIM summarizes all results using two values for each gene: the "*Average Perturbation*" and the "*Activity Score*" (AS). The average perturbation is the mean for all perturbation values computed during the simulation process and reproduces the expected change of expression for the entire process. The function of the Activity Score is twofold. The sign gives the type of predicted effect: positive for activation, negative for inhibition. The value is the log-likelihood that this effect will occur. Together with the AS, PHENSIM also computes a *p-value* through a bootstrapping procedure. All p-values are then corrected for multiple hypotheses using the q-value approach PHENSIM *p-values*

are used to establish how biologically relevant the predicted alteration is for the simulated phenomena - i.e., the lower is a node p-value, the less likely it is that such alteration will occur by chance. An overview of the PHENSIM algorithm is depicted in Fig 13.

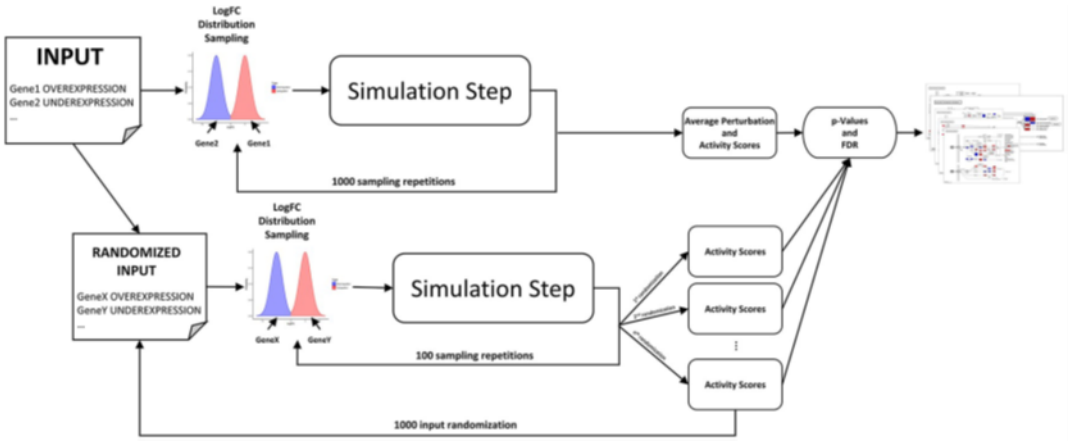


Figure13. Description of the PHENSIM algorithm. First, the user provides a set of genes and the type of alteration (over-/underexpression). Then, synthetic LogFCs are generated, and a simulation step is performed. This procedure is repeated 1000 times to compute the Activity Scores. Next, user input is randomized, and 100 synthetic LogFC are generated to estimate Activity Scores using the simulation step. This input randomization is repeated 1000 times for greater precision. Finally, p-values are computed, and the False Discovery Rate is estimated using the q-value methodology. The algorithm comprises 5 main steps. Given a user input, (i) synthetic LogFC are generated and a (ii) simulation step is performed. These steps are repeated 1000 times to (iii) compute the AS. Next, user input is (iv) randomized, and 100 synthetic LogFC are generated to estimate AS using the simulation step. The input is randomized 1000 times to obtain greater precision. Finally, (v) p-values are computed, and the False Discovery Rate is estimated using the q-value methodology.

The next section will describe the benchmarking procedure to which PHENSIM was subjected and the in silico experiments performed as case studies.

3.2.1 PHENSIM benchmarking procedure

To assess PHENSIM prediction reliability, we performed a comprehensive experimental analysis.
First, we built a benchmark based on data published in the GEO [162] database, specifically transcriptomics experiments performed on cell lines where a single gene was perturbed (knockdown, CRISPR, or transfection).
More in detail, we wanted to determine how much PHENSIM can correctly predict the biological outcomes of the up-/down-regulation of a gene in a cell line through comparisons with expression data collected before and after the alteration.
Therefore, we gathered 22 GEO series of cell lines with a perturbed gene. Since these series could contain multiple perturbation experiments of different genes or in several cell lines, we obtained a total of 50 case/control sample sets.
Their details are shown in [1] (supplementary Table S2) together with the name and code of the GEO series, the technology used to determine gene expression, the perturbed gene, the type of experiment (knockout, knockdown, transfection, CRISPR, etc.), whether the gene is present in KEGG pathways, and the GEO accessions of the case and control samples. Each sample set was then divided into two categories, which were analyzed differently: (i) samples whose altered gene is present in the meta-pathway (called DS1), and (ii) samples whose perturbed gene is not in the meta-pathway (called DS2).
For DS1, consisting of 30 sample sets, we directly simulated the alteration of the gene using PHENSIM.
For DS2, consisting of the remaining 20 sample sets, we simulated the alteration of the Differentially Expressed Genes (DEGs) computed between cases and controls. The rationale behind this choice is that DEGs somehow represent the effect of the source alteration.
For each dataset, non-expressed genes were identified according to the experiment type: Microarray or Sequencing. For sequencing, we chose all genes with an average count of less than 10. For microarrays, we selected all genes exhibiting an average expression less than the 10th percentile. DEGs were computed using Limma [163] with a p-value threshold of 0.05 and a LogFC threshold of 0.6. Each sample set was simulated as described above.
Then, we compared PHENSIM predictions (up/down-regulation) with LogFC computed on the expression data. All genes showing an absolute LogFC lower than 0.6 were considered as non-altered.
Finally, we assessed the results in terms of accuracy (the number of correctly predicted genes divided by the total number of genes). Furthermore, since accuracy can be influenced by class imbalance, we chose to compute *Positive Predictive Value* (PPV), *Sensitivity*, *Specificity*, and *False Negative Rate* (FNR) according to the type of alteration found in the expression data. More in detail, for altered genes (LogFC > 0.6), we want to identify upregulation and downregulation events correctly. Therefore, the True Positives (TPs) are genes predicted as upregulated with positive LogFC in the expression data. In contrast, genes predicted as downregulated with a negative LogFC are the True Negatives (TNs). Furthermore, genes predicted as upregulated with a negative LogFC are False Positives, and downregulated genes with a positive LogFC are False Negatives.

So that we determined the ability of PHENSIM to correctly identify upregulated genes by computing PPV and Sensitivity, while we assessed the performance regarding down regulated ones through Specificity:

$$PPV = \frac{TP}{TP+FP}, \quad Sensitivity = \frac{TP}{TP+FN}, \qquad Specificity = \frac{TN}{TN+FP},$$

Concerning non-altered genes, we were interested in determining whether PHENSIMis capable of correctly identifying them. In this case, a gene that is predicted as non-altered with a LogFC < 0.6 is considered as a True Positive, while a gene indicated as altered with a LogFC < 0.6 is a False Negative. Therefore, estimated the rate of correctly identified non-altered genes in terms of PPV, while the FNR shows us the percentage of non-altered genes that are wrongly identified as perturbed by PHENSIM:

$$FNR = \frac{FN}{FN + TP},$$

Then, among the competitors with which to compare the performances of our system, we have chosen BioNSi (Biological Network Simulator)[151, 163].

To compare performances with BioNSi, we ran the same simulations and computed the same metrics on the results. BioNSi requires an expression (in the range 0–9) for each gene and tracks how it changes until a steady state is reached. Therefore, a gene is up-/down-regulated if the simulated expression increases/decreases between the initial and the final state, respectively. If no change is observed, the gene is not perturbed. To run the simulation, we loaded the meta-pathway and set all genes' expression levels to 5. Next, we gave expression 9 for upregulated genes and 1 for down-regulated ones.

Moreover, since PHENSIM can extend KEGG pathways with REACTOME ones, we decided to run all tests on this extended network, comparing the results before and after the extension. However, we could not perform any comparison with BioNSi since it could not load the extended network due to its size.
To assess performance differences between the two systems for each dataset, we provide several graphs comparing each metric. In Fig 14, we summarized the DS1 datasets' results, and in Fig 15, we reported the results from the DS2 datasets. In each graph, we detail a single metric: *Positive Predictive Value* (PPV), *Sensitivity* and *Specificity* for genes showing altered expression, and *PPV* and *False Negative Rate* (FNR) for the non-altered ones. On the x-axis, we have PHENSIM performance, while on the y-axis, we have BioNSi. Each dot represents a dataset.
The black line marks the points where the two algorithms have the same performance. We summarize the comparisons before and after adding REACTOME pathways in Appendix i) Fig S1 for DS1 and Fig S2 for DS2. In these graphs, the x-axis represents the PHENSIM performance with REACTOME, while on the y-axis, we have PHENSIM without REACTOME.

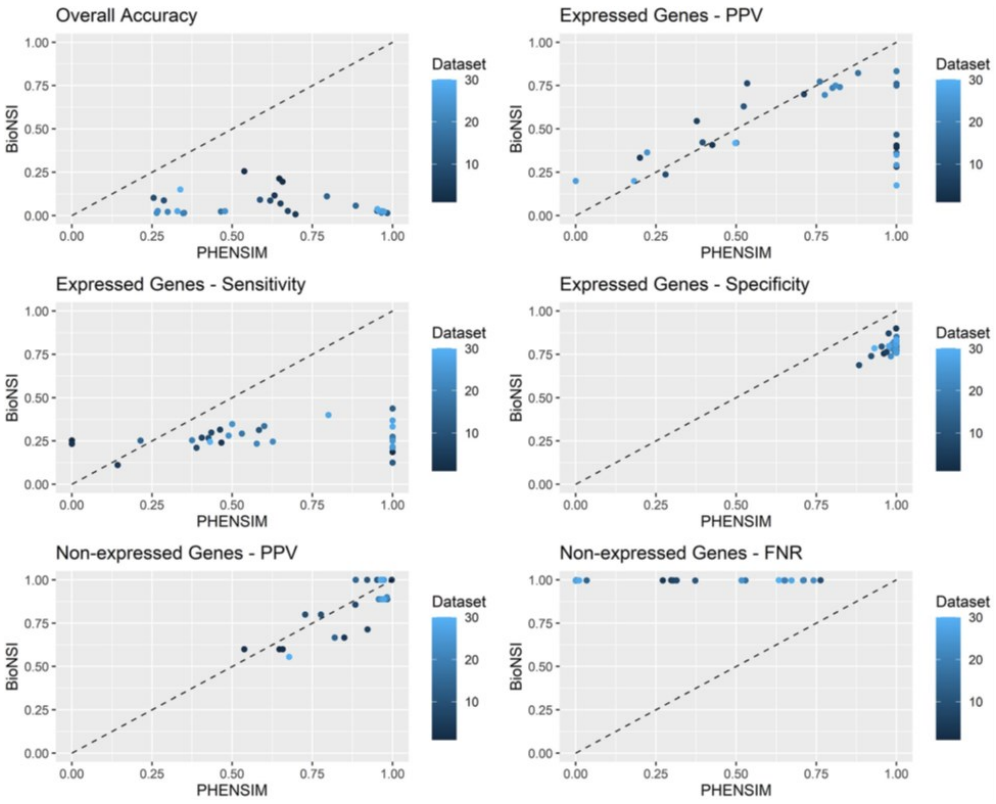


Figure 14. Comparison between PHENSIM and BioNSi for datasets where the altered gene was in the meta-pathway. Each graph reports one metric: Positive Predictive Value (PPV), Sensitivity and Specificity for genes showing altered expression, and PPV and False Negative Rate (FNR) for the non altered ones. On the x-axis, we report PHENSIMperformance, while on the y-axis, we present BioNSi. Each dot represents a dataset. The black line marks the points where the two algorithms have the same performance. On a dataset below the line, PHENSIMhas better performance than BioNSi; above the line, it is the opposite.

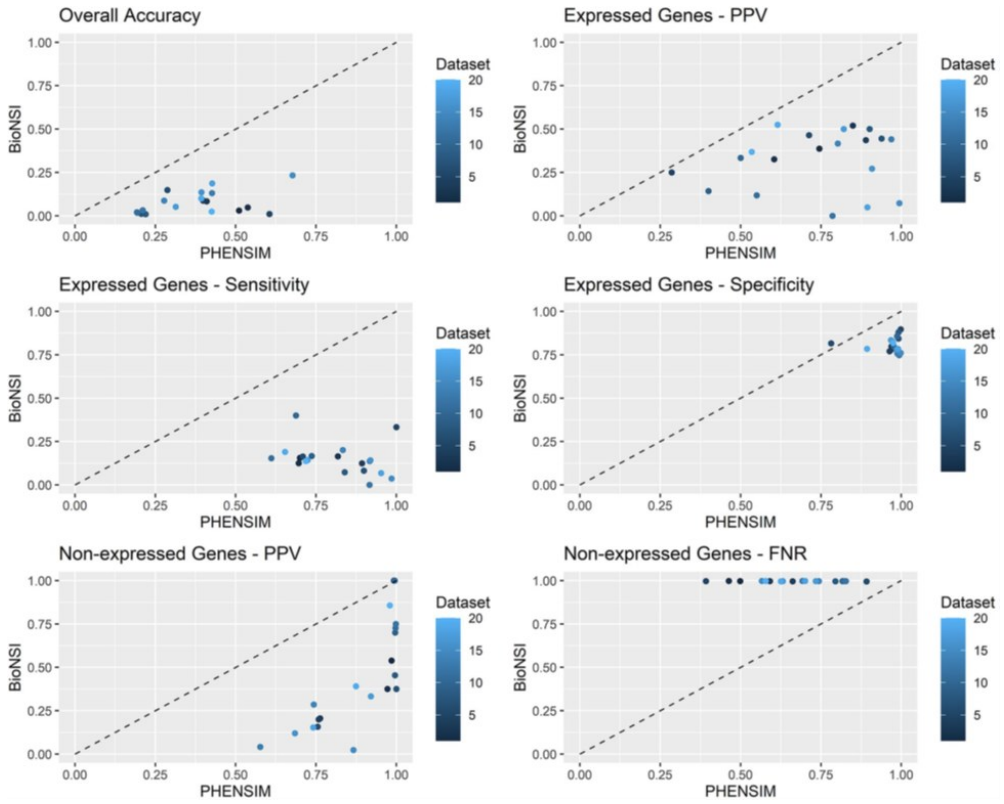


Figure 15. Comparison between PHENSIM and BioNSi for datasets where the altered gene was not in the meta-pathway. Each graph reports one metric: Positive Predictive Value (PPV), Sensitivity and Specificity for genes showing altered expression, and PPV and False Negative Rate (FNR) for the non-altered ones. On the x-axis, we report the PHENSIMperformance, while on the y-axis, we have BioNSi. Each dot represents a dataset. The black line marks the points where the two algorithms have the same performance. On a dataset below the line, PHENSIMhas better performance than BioNSi; above the line, it is the opposite.

Our results show that PHENSIM has an average accuracy of 0.6295 for the dataset in the first category and 0.3650 for the second category. BioNSi offers an average accuracy of 0.0640 and 0.0735 for the datasets in the first and second categories. Nevertheless, PHENSIM has higher PPV than BioNSi (0.6899 and 0.5075, respectively) in the first and second categories (PHENSIM = 0.7350, BioNSi = 0.3282). PHENSIM also shows a greater Sensitivity and Specificity to BioNSi. Furthermore, since PHENSIM can extend KEGG pathways with REACTOME, we performed the same tests on such an extended network, comparing the results before and after the integration. However, we could not evaluate BioNSi capabilities in this context since it could not load the extended network due to its size. In this setting, PHENSIM showed an average accuracy of 0.6437 with comparable PPV (0.6349) although lower Sensitivity (0.5416) and comparable Specificity (0.9854) for DS1. A slight decrease of performance can be observed for DS2 (Accuracy: 0.3291, PPV: 0.7571, Sensitivity: 0.7622, Specificity: 0.9716). Table 2 reports the detailed comparison in terms of average metrics.

	Algorithm	Accuracy	Altered Genes			Non-altered Genes	
			PPV	Sensitivity	Specificity	PPV	FNR
DS1							
	PHENSIM	0.6259	0.6899	0.6150	0.9829	0.8921	0.3078
	PHENSIM + Reactome	0.6437	0.6349	0.5416	0.9854	0.8972	0.3007
	BioNSi	0.0640	0.5075	0.2692	0.7925	0.8624	0.9970
DS2							
	PHENSIM	0.3650	0.7350	0.8105	0.9684	0.8797	0.6836
	PHENSIM + Reactome	0.3291	0.7571	0.7622	0.9716	0.8780	0.7201
	BioNSi	0.0735	0.3283	0.1500	0.8052	0.4345	0.9968

Table 2. Summary of the comparisons between PHENSIM and BioNSi. We computed for both software accuracy, Positive Predictive Value (PPV), Sensitivity and Specificity for genes showing altered expression, and PPV and False Negative Rate (FNR) for the non-altered ones. The sample sets were categorized based on the KEGG meta-pathway genes: DS1 contains all sample sets where the up- or down-regulated gene was in KEGG; DS2 all the remaining ones. https://doi.org/10.1371/journal.pcbi.1009069.s001

Moreover, to quantitatively evaluate network perturbation prediction, we chose an additional dataset containing experimental measurements of protein expression changes following drug treatment in a cell line[164]. The dataset comprises 124 protein levels in a time series from 10 minutes to 67 hours (8 timepoints). The authors followed the perturbation caused by the administration of 54 drug combinations, including several gene inhibitors (MEKi, AKTi, STAT3i, SRCi, mTORi, BETi, PKCi, RAFi, and JNKi). To perform the comparison, we first gathered all drug targets from Nyman et al. [164]. Then, we simulated the alteration of their targets for each drug combination and collected the results concerning the 124 proteins. Finally, we computed the Pearson Correlation Coefficient between our predictions and the actual measurement to indicate results consistency.

In Fig 16, we rep report the analysis results comparing PHENSIM steady-state predictions with each time point in terms of the Pearson Correlation Coefficient. Results show that PHENSIM predictions are coherent with the proteomics experiments, reaching the maximal correlation at 24h and 48h.

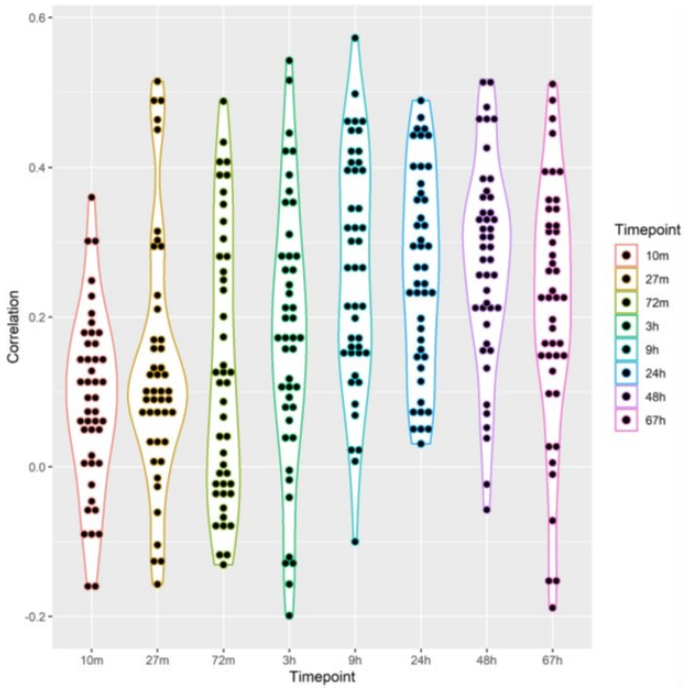


Figure 16. Comparison between PHENSIM predictions and the proteomics measurements of Nyman et al. [162]. We report the Pearson Correlation Coefficient computed between PHENSIM and the proteomics measurements for each timepoint and drug combination. Results are summarized through a violin plot detailing both the distribution and the values' density.

All raw data, input files, and other source codes are available for download at https://github.com/alaimos/phensim/tree/master/Benchmark.

Finally, to complete our assessment of PHENSIM capabilities, we run several simulations to perform 4 case studies on known biological experiments: (i) anti-cancer effects of metformin, (ii) Everolimus (RAD001) treatment in breast cancer, (iii) effects of exosomal vesicles on hematopoietic stem/progenitor cells (HSPCs) in the bone marrow (BM) and (iv) testing TNFα/siTPL2-dependent synthetic lethality on a subset of human cancer cell lines. We examined the ability of PHENSIM to correctly predict the activity status of both individual genes/proteins and signaling pathways by comparing PHENSIM predictions with experimental data. In the following sections, are reported the case studies and their results.

3.2.2 PHENSIM- Case Studies

Simulation 1: Anti-cancer effects of metformin

Metformin is a widely prescribed agent for the treatment of type 2 diabetes [164-166]. It inhibits glucose production in the liver and increases insulin sensitivity in the peripheral tissues. Furthermore, metformin treatment reduces insulin secretion by β-pancreatic cells. The key molecule that performs these functions is AMP-activated protein kinase (AMPK), a serine-threonine kinase regulating cellular energy metabolism.

Some evidence indicates that metformin possesses anti-cancer effects in various cancer types, especially in diabetic patients, directly and indirectly [165,167,168]. Indeed, metformin directly activates the LBK1-AMPK signaling pathway [168]. Metformin is known to uncouple the electron transport chain in the mitochondria by targeting Complex I [165, 167-169], leading to impaired mitochondrial function, decreased adenosine triphosphate (ATP) synthesis, and elevated cellular AMP/ATP ratio [165,167,168]. Increased AMP binding to AMPK activates AMPK by inducing phosphorylation of its catalytic subunit at residue Thr172 by liver kinase B1 (LKB1), a tumor suppressor and a regulator of cellular energy status[167,168]. The binding of AMP to AMPK also prevents the dephosphorylation of AMPK Thr172 by protein phosphatases. LKB1-activated AMPK phosphorylates and activates the tumor suppressor Tuberous Sclerosis Complex 1 and 2 (TSC1/2), which negatively regulates the activity of the mammalian target of rapamycin (mTOR), which is upregulated in most cancer cells and causes tumor proliferation and cell growth by inhibiting Ras homolog enriched in brain (Rheb) [165,168]. mTOR is a critical mediator of the phosphatidylinositol-3-kinase/protein kinase B/Akt (PI3K/PKB/Akt) signaling pathway, one of the most frequently deregulated molecular networks in human cancer[168].

Metformin-activated AMPK inhibits mTOR and reduces the phosphorylation of its downstream targets, the eukaryotic initiation factor 4E-binding proteins (4EBPs), and ribosomal S6 kinases (S6Ks), leading to an inhibition of global protein synthesis, cell cycle progression, cell proliferation, and angiogenesis[168]. Moreover, AMPK has been reported to suppress the mTOR signaling pathway independent of TSC2 via phosphorylation of mTOR binding protein Raptor.

Metformin has been shown to cause a G0/G1 cell cycle arrest by decreasing the expression of cyclin D1 [167].

Metformin-induced AMPK activation has been shown to phosphorylate insulin receptor substrate-1 (IRS-1) at Ser-794, which results in decreased recruitment of the p85 subunit of phosphoinositide-3-kinase (PI3K), thus, impairing the insulin-like growth factor (IGF)-stimulated PI3K/protein kinase B/ mammalian target of rapamycin complex 1 (PI3K/Akt/mTORC1) signaling pathway.

Metformin also inhibits the crosstalk between G-protein-coupled receptors (GPCR) and insulin/IGF1 receptors signaling, resulting in the inhibition of mTORC1 and reduction of cellular proliferation[165, 167].

Metformin induces nuclear degradation and decreased expression of Sp proteins, transcription factors for genes involved in cell proliferation (cyclin D1), metabolism (FAS), apoptosis (B-cell lymphoma 2, BCL-2, and survivin), and angiogenesis (vascular endothelial growth factor, VEGF, and its receptor VEGFR1) [167, 168].

The indirect mechanism of metformin in anti-cancer function is related to its ability to lower insulin and insulin-like growth factor 1 (IGF-1)[168].

Metformin disrupts insulin and IGF-1 signaling pathways by reducing insulin and IGF-1 levels, reducing total IGF-1 receptor and IR levels, and downregulating IGF-1 receptor and IR gene expression[170].

In parallel with this, metformin also downregulates the MAPK (mitogen-activated protein kinase) pathway, NF-κB (nuclear factor kappa B) signaling [169, 171], glycolysis, and the TCA (tricarboxylic acid) cycle [167, 168, 170]. An overview of the metformin-mediated effects is reported in Fig 18.

Based on these details, we run the PHENSIM simulation setting the simultaneous upregulation of LKB1 and the downregulation of insulin (Ins), IGF1, and GPD1[171, 172] as input.

As expected, PHENSIM returned significant downregulation of Insulin (pathway activity score = -8.7121, p-value 0.105) and mTOR signaling (pathway activity score = -8.7121, p-value 0.107).

Although mTOR's negative regulation should activate the repressor of translation initiation 4EBP, the simulation returns no activity score for this node. However, a low positive perturbation for 4EBP can be observed (perturbation = 0.00009). PHENSIM also predicted the inhibition of downstream nodes involved in protein synthesis, such as S6Ks (S6K-alpha3 activity score = -2.0019, p-value = 0.046).

MAPK signaling was predicted as downregulated (MAPK pathway activity score = -4.8203, p-value = 0.130). Several down regulated enzymes and metabolites were predicted for these two pathways, in full agreement with data from literature [170].

Finally, in accordance with literature, PHENSIM also predicted weak changes in cytokine gene expression as it can be seen from average nodes perturbations (IL6 perturbation = -0.0001; IL8 perturbation = -0.00002; IL17 perturbation = -0.00001; TNF-alpha perturbation = -0.00014)[170].

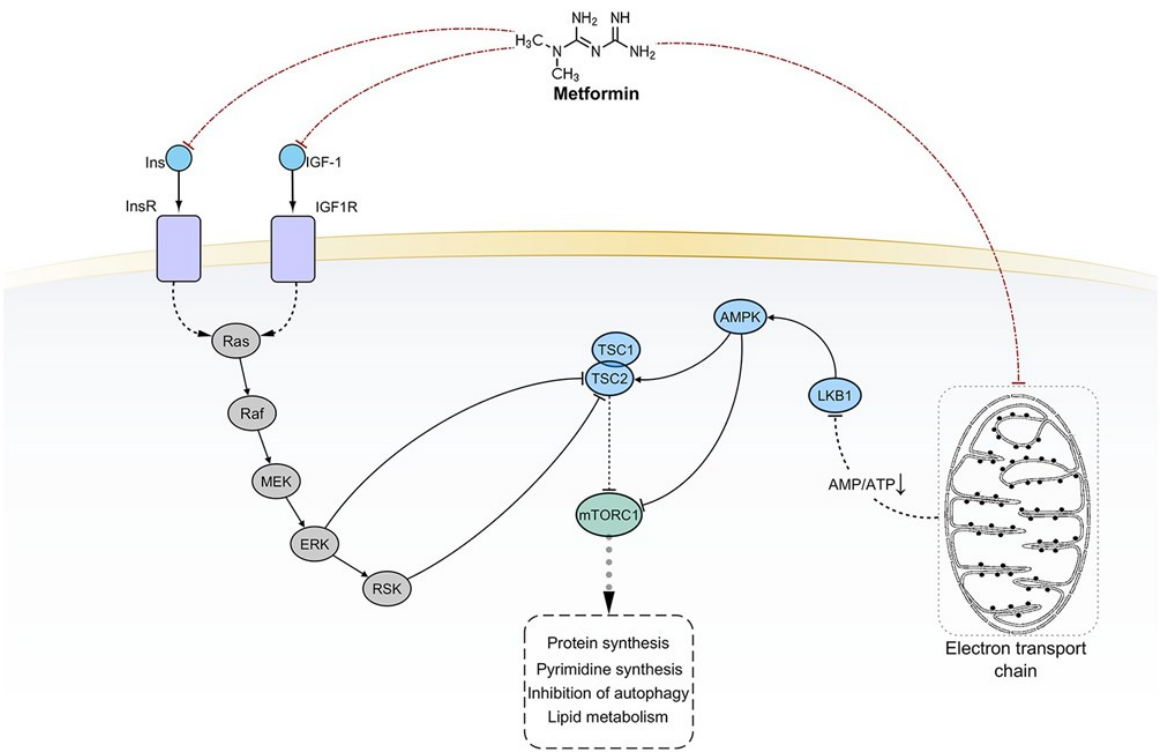


Figure 17. The current model of metformin-mediated pharmacological effects. Black solid edges represent direct interaction between first neighbor nodes. Dashed edges represent indirect interactions between nodes. Red dot-dashed edges evidence scientifically validated interactions considered for PHENSIM prediction.

Simulation 2: Everolimus (RAD001) and breast cancer

Everolimus (RAD001, Afinitor®), an analog of rapamycin, has shown immunosuppressive and anti-cancer activities [173-176]. It is currently approved to treat various cancer types, including metastatic breast cancer [177-179]. Everolimus has a growth inhibitory activity against tumor cells and can retard tumor growth through direct mechanisms against both the tumor cell and the solid tumor stromal components[175].

Everolimus inhibits the "mammalian target of rapamycin" (mTOR) to prevent the downstream signaling required for cell cycle progression, cell growth, and proliferation[173-175; 180-182].

In mammalian cells, mTOR exists in two complexes, mTORC1 and mTORC2[178, 179, 181, 183], which are differentially regulated and have distinct substrate specificities[178]. mTORC2 signaling is lower in breast tumors compared to normal breast tissue. This difference could suggest that mTORC1 signaling is more oncogenic than mTORC2[184].

mTORC1 promotes protein synthesis by (a) stimulating ribosome biogenesis via phosphorylation and inhibition of the RNA Polymerase III repressor MAF1[185]; (b) phosphorylating the p70S6K and 4EBP1 and modulating the activity of their downstream targets[186, 187]; (c) by regulating nucleocytoplasmic RNA transport[185, 186]. In addition, mTORC1 stimulates pyrimidine biosynthesis and lipid biosynthesis [186, 187] mTORC1 phosphorylates ULK1 (unc-51 like autophagy activating kinase 1) and DAP (death-associated protein) inhibiting autophagy[184, 188].

Finally, the upregulation of mTOR signaling can promote tumor growth and progression through several mechanisms, including the promotion of growth factor receptor signaling, angiogenesis, glycolytic metabolism, lipid metabolism, cancer cell migration, and suppression of autophagy [178, 179]. All these functions of mTORC1 are reversed by Everolimus and other mTORC1 inhibitors [174, 177] (Fig. 18).

Everolimus binds with high affinity to its intracellular receptor, the FKBP12, a protein belonging to the immunophilin family. The Everolimus–FKBP12 complex binds mTOR when associated with RAPTOR and mLST8 to form mTORC1 complex, resulting in decreased interaction between mTOR and RAPTOR, which could inhibit the phosphorylation and activation of the major mTORC1 downstream targets[176, 178-180, 183, 184].

Here we wanted to simulate the inhibition of mTORC1. Unfortunately, simulating mTORC1 inhibition was not feasible because KEGG does not distinguish the mTOR node in mTORC1 from the one included in mTORC2. To overcome such limitation, we have set the downregulation of p70S6K (p70S6Ka and p70S6Kb) and 4EBP and the upregulation of ULK1/2 because these are the well-known downstream targets of mTORC1. Then we uploaded a list of non-expressed genes in breast tissue to simulate the drug's effects on such tissue. Our simulation predicted that RNA transport factors would be downregulated, while factors involved in autophagy would be upregulated. The simulation showed that the RNA transport signaling pathway exhibits a negative activity scores (activity score = -4.4108; p-value = 0.13). Furthermore, we could predict several downregulated factors involved in RNA transport and protein synthesis, such as eukaryotic translation initiation factor 4A, 4B and ribosomal proteins S6Ks, p70-S6K and p70S6Kb (eIF4A1 activity score = -4.8203; eIF4A2 activity score = -4.8203; p70-S6K activity score = -4.8203; p70S6Kb activity score = -4.8203; p-value for all nodes < 0.01). PHENSIM also predicts the 4EBP1 inhibition (activity score = -4.8203; p-value = 0.013) and consequently the upregulation of eIF4E (activity score = 4.8203; p-value = 0.008).

PHENSIM predicts upregulation of the autophagy (activity score = 4.8203, p-value = 0.26) as a consequence of alterations in ULK1/2 phosphorylation levels and the downregulation of cyclin D. However, PHENSIM failed in predicting the deregulation of p21 (cyclin-dependent kinase inhibitor 1) and NF-kB [177]. This limitation is probably due to the presence of a single node for mTORC1 and mTORC2.

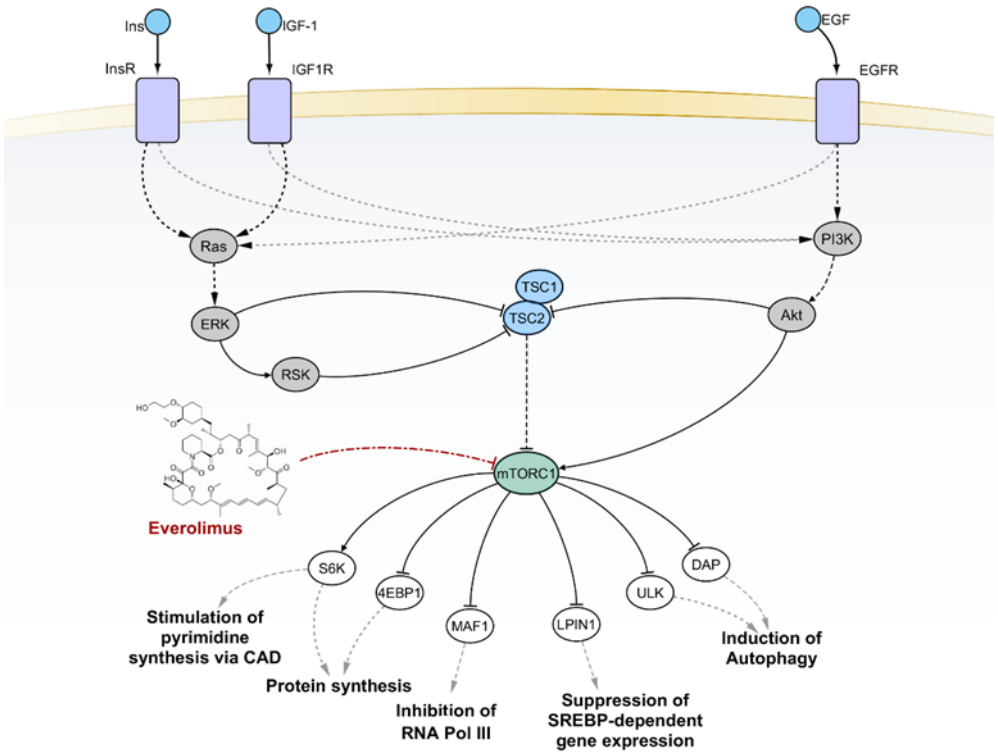


Figure 18. mTORC1 and its downstream signaling pathways. Black solid edges represent direct interaction between first neighbor nodes. Dashed edges represent indirect interactions between nodes. Red dot-dashed edges evidence scientifically validated interactions considered for PHENSIM prediction.

Simulation 3: effects of exosomal vesicles on hematopoietic stem/progenitor cells (HSPCs) in the bone marrow (BM)

Cancer-derived exosomes' functional relevance to tumor growth, metastasis, and treatment response has become increasingly evident[189,190]. Exosomes derived from AML blasts contain complex cargoes which function via paracrine mechanisms to modulate the properties of both the tumor cells themselves and the BM niche. Several microRNAs are selectively incorporated in these exosomes, including miR-150 and miR-155[191, 192]. One of these microRNAs targets is the transcription factor c-MYB, which is downregulated in tumor cells exposed to the exosomes[192]. Additional targets include c-KIT, DNMT1, Lymphoid Cell Helicase (HELLS), PAICS, an enzyme involved in purine biosynthesis, TAB2, and others. The downregulation of these molecules compromises hematopoiesis via stroma-independent mechanisms. However, the cargo of AML cell-derived exosomes also targets mesenchymal stromal progenitors, inhibiting/reducing the expression of hematopoietic stem cell supporting factors such as CXCL12 (C-X-C motif ligand 12), KITL (c-Kit ligand), IL-17, and IGF1 and interfering with both hematopoiesis and osteogenesis[189] (Fig. 19). Moreover, AML-derived exosomes increase gene expression supporting AML growth (DKK1, IL-6, CCL3).

To determine whether PHENSIM can make the correct predictions in this model, we simulated the uptake of the eight most representative miRNAs (miR-150, -155, -146a, -191, -221, -99b, -1246, and let-7a) included in AML-derived exosomes by hematopoietic stem cells [191].

The simulation predicts an inhibition of osteoclast differentiation (activity score = -8.7121, p-value = 0.132) and cytokine-cytokine receptor interaction pathways (activity score = -4.8203, p-value = 0.115). In agreement with the literature, some genes involved in modulation of normal hematopoiesis, like CXCL12 (activity score = -4.4108, p-value = 0.008) and the receptor IGF1R (activity score = -4.8203, p-value = 0.038), but not IGF1, were downregulated [189]. Similarly, c-MYB, which is involved in HSPC differentiation and proliferation, was also downregulated [192] (activity score = -4.5951; p-value = 0.012). However, PHENSIM failed to predict the upregulation of DKK, IL6 and CCL3 (DKK and CCL3 activity score = 0, IL6 activity score = -4.7015, p-value = 0.011), and the downregulation of KITL and IL17 (activity score = 0) [189].

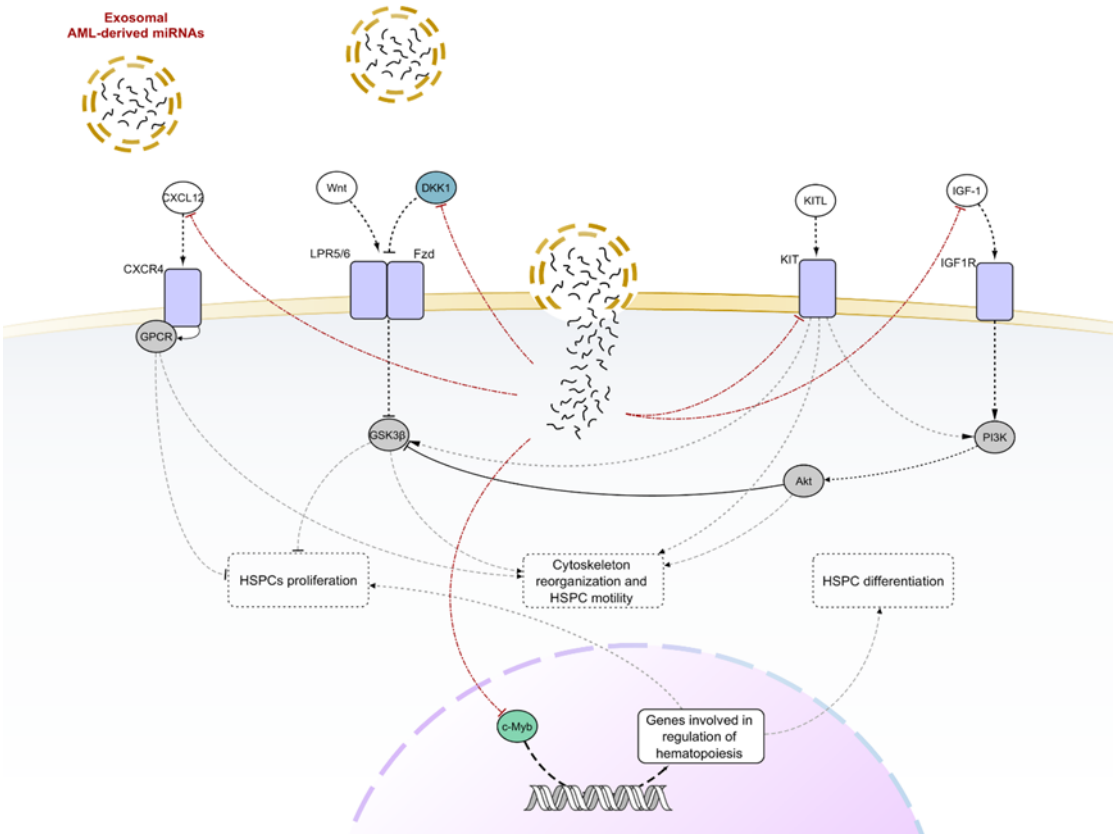


Figure 19. A reconstructed model showing cellular components involved in hematopoiesis and motility of HSPCs and their downregulation mediated by exosomal-miRNAs derived from AML cells. Black solid edges represent direct interaction between first neighbor nodes. Dashed edges represent indirect interactions between nodes. Red dot-dashed edges evidence scientifically validated interactions considered for PHENSIM prediction.

Simulation 4: testing TNFα/siTPL2-dependent synthetic lethality on a subset of human cancer cell lines

TNFα (tumor necrosis factor alpha), a type II transmembrane protein, is a member of the tumor necrosis factor cytokine superfamily and has an essential role in innate immunity and inflammation.
Although it can induce cell death, most cells are protected by a variety of mechanisms.
In a recent paper, *Serebrennikova et al*.[192, 193] showed that one of the checkpoints of TNFα-induced cell death is TPL2 (MAP3K8), a MAP3 kinase that is known to have an important role in immunity, inflammation, and oncogenesis. The knockdown of TPL2 resulted in the downregulation of miR-21 and the upregulation of its target CASP8 (caspase-8). This effect, combined with the downregulation of the caspase-8 inhibitor cFLIP (FADD-like IL-1β-converting enzyme inhibitory protein), resulted in the activation of caspase-8 by TNFα and the initiation of apoptosis (Fig. 20). The activation of caspase-8 promotes the activation of the mitochondrial pathway of apoptosis. However, some molecules such as BIML (Bcl-2-like protein 11, isoform L), which are also involved in the activation of the mitochondrial pathway, may be activated via caspase-8-independent mechanisms. A crucial upstream regulator of this pathway is NF-κB. The knockdown of TPL2 also inhibits the activation of ERK (MAPK1/2), JNK (c-Jun Nterminal kinase), and p38MAPK, the activation of AKT, and the phosphorylation of GSK3 (glycogen synthase kinase 3) at Ser9/21. However, their inhibition does not appear to have a role in the initiation of TNFα/siTPL2-induced apoptosis. It is worth noting that the activation of the apoptotic (caspase-8-dependent) pathway in TNFα/siTPL2 treated cells was observed in some but not all cancer cell lines, suggesting that correct prediction will depend on whether the data analyzed by PHENSIM are derived from sensitive or resistant cells.

To launch the simulation, we set TPL2 and miR-21-5p as downregulated and TNFα as upregulated. Since our goal was to simulate the outcome of such treatment in six cell lines, i.e., HeLa, HCT116, U2-OS, CaCo-2, RKO, and SW480, we launched six different simulations. Each simulation had a separate list of non-expressed genes, one for each cell line.
Among these tumor cell lines, only HeLa, HCT116, U2-OS were sensitive to treatment with TNFα/siTPL2. At the end of the computations, PHENSIM could not predict the upregulation of caspase-8 for any of the six cell lines nor the downregulation of cFLIP. This limitation could be the result of missing information in KEGG pathways. PHENSIM did not indicate any activity score for MLC1 (Mcl-1 apoptosis regulator) and XIAP (X-linked inhibitor of apoptosis) nodes.
PHENSIM could not predict the upregulation of the apoptosis inhibitors BCL2 and BCL-XL in all cell lines except for HCT116, where BCL2 results positively perturbed (perturbation = 0.001). PHENSIM showed a negative perturbation of the inducer of mitochondrial apoptosis BAX only in HCT116.
Although these results do not precisely reflect our expectations as there are discrepancies between the in vitro and in silico experiment done by PHENSIM, it was confirmed by results obtained by the previously mentioned experimental study [193], which suggested that the change in the expression of such molecules was due to the activation of feedback mechanisms.
Besides, phosphorylated ERK, MEK, JNK, and p38 activity were strongly downregulated for all of the six cell lines except for RKO where PHENSIM predict correctly just ERK and p38, and for Caco-2 cells, which result in a negative activity score for ERK and a weak perturbation for JNK and p38 genes. Finally, PHENSIM did not predict cIAP2 (baculoviral IAP repeat containing 2) negative perturbation, which has an activity score of 0 and a weak negative perturbation, in RKO cells as confirmed by the experimental data.

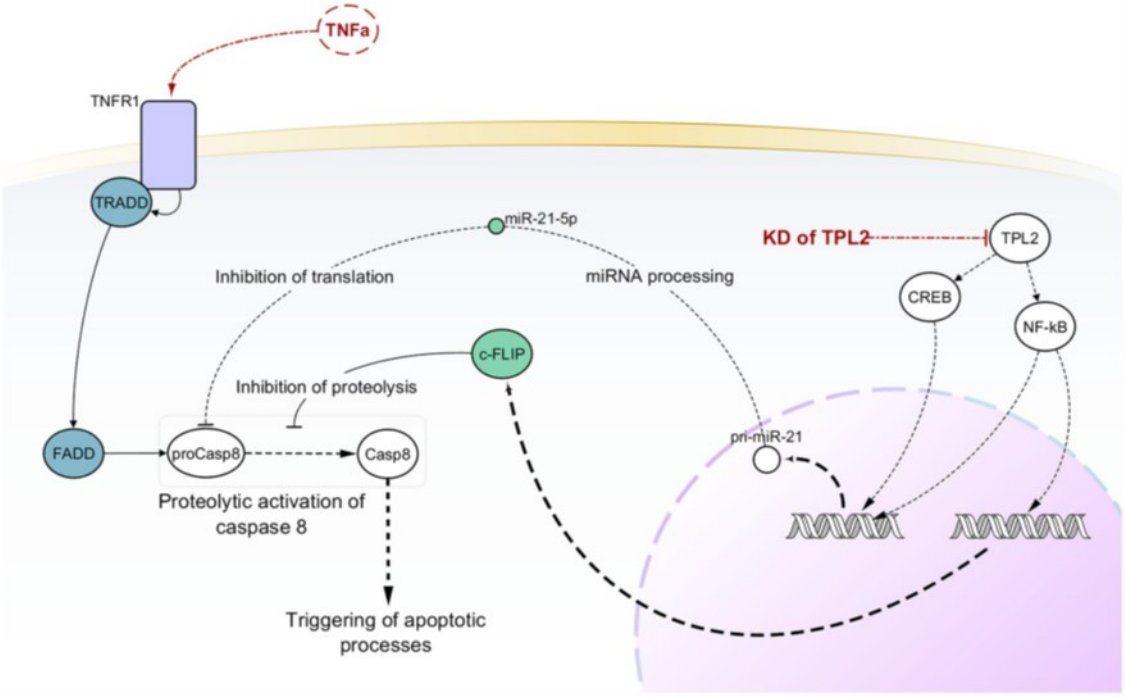


Figure 20. Generalized model showing molecular mechanisms underlying the TNFα/siTPL2-dependent synthetic lethality. Black solid edges represent direct interaction between first neighbor nodes. Dashed edges represent indirect interactions between nodes. Red dot-dashed edges evidence scientifically validated interactions considered for PHENSIM prediction.

The comparison between PHENSIM predictions and experimental data reported shows that simulation 1 results were in almost full agreement with literature and that there is a partial agreement with the three remaining simulations, showing a discrete degree of accuracy.

Discrepancies with baseline data suggest some limitations in the predictive potential of our method. However, since pathway analysis relies on prior knowledge about how genes, proteins, and metabolites interact, we hypothesize that such a negative outcome is at least partly due to the incompleteness of the existing knowledge employed in the study. Indeed, since the biological pathways on current databases are still largely fragmented, calculations based on them will inevitably produce less than ideal results[128]. One example of this limitation is provided by mTORC1 downstream signaling. It is known that mTORC1 promotes protein synthesis by phosphorylating p70SK and 4EBP. It also stimulates ribosome biogenesis via inhibitory phosphorylation of the RNA Polymerase III repressor MAF1[185]. mTORC1-induced pyrimidine biosynthesis is stimulated by p70S6K-mediated phosphorylation of the CAD enzyme (carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase). Furthermore, the upregulation of 5-phosphoribosyl-1 pyrophosphate (PRPP) is an allosteric CAD activator [187, 194].

KEGG Pathways do not consider such interactions. Therefore, our tool could not predict any perturbations for these biological processes. Similar observations can be made for the downregulation of cFLIP in the siTPL2/TNFα-resistant cell lines by our method. However, we were able to identify indirect evidence of such activity. On the other hand, the correct predictions obtained for autophagy, RNA transport, and mTOR signaling in simulation 2, and the mitochondrial apoptotic pathway activation in simulation 4, suggest that, provided with the right information, PHENSIM is likely to obtain significantly better results.

A further limitation for pathway analysis methods is the current knowledge-base inability to contextualize gene expression and pathway activation in a cell- and condition-specific manner [128]. Furthermore, pathways do not consider protein isoforms encoded by different genes or differently processed mRNAs derived from a single gene. This poses a significant limitation since such isoforms may have unique and sometimes opposite signaling properties. By developing a strategy that allows removing non-expressed genes from the computation, we offer the user the possibility to contextualize predictions in a cell- or tissue-dependent manner. In conjunction with this, integrating KEGG pathways with information from post-transcriptional regulators such as miRNAs increased the results' accuracy, leading to considerable improvements in predictions[195]. Moreover, using the meta-pathway approach, instead of single disjointed pathways, partially addresses pathway independence[128].

In conclusion, PHENSIM showed good accuracy in most applications and could predict the effects of several biological events starting from the analysis of their impact on KEGG. We believe that several discrepancies can be traced to the incompleteness of knowledge in KEGG or Reactome pathways or the lack of appropriate cell- and condition-specific information. Such incompleteness can be partially addressed through a manual annotation of the pathways with the missing elements and links, including miRNA-target and TF-miRNA interactions.

Despite these limitations, our approach shows appreciable utility in the experimental field as a tool for the reliable prioritization of experiments with greater success chances.

In the next paragraph a new system for knowledge network construction will be introduced. It works by applying text mining and machine learning algorithms. Indeed, although the meta-pathway is easily extensible manually, reconstructing the missing information (nodes and arcs) on the basis of extensive bibliographic research, it can be quite expensive in terms of time. For this reason, having a system that builds biological knowledge networks annotating automatically the information present in the literature becomes extremely interesting.

3.3 NetMe

In the previous paragraphs were discussed some of the existing tools [1] and computational models relying on existing network databases, such as KEGG[2, 195, 196] and Reactome [4-6, 153]. It was also seen that, despite the enormous amount of available data, these databases are still incomplete and therefore have partial information[197]. As an example, KEGG includes approximately one-third of the known genes[153].

The algorithms described in previous paragraphs are manually extensible, which means that it is possible to add missing nodes and edges as needed. Since especially in research areas like biology or bio-medicine, thanks to fast-track publication journals, the number of published papers increases significantly fast, it becomes extremely expensive in terms of energy and time to make a totally manual research, especially if it is necessary to integrate highly connected nodes, or multiple nodes, in the network. Moreover, even if our research can be carefully and detailed, it cannot be totally complete and updated with all the knowledge available in literature.

In the last few years, thanks to the availability of sizeable open-access article repositories such as PubMed Central[198, 199], arxiv [197, 200] biorxiv [201] as well as ontology databases which hold entities and their relations[202], the research community has focused on text mining tools and machine learning algorithms to digest these corpora and extract valuable semantic knowledge from them. Text mining[202, 203], and Natural Language Processing[204] tools employ information extraction methods to translate unstructured textual knowledge in a form that can be easily analyzed and used to build a functional knowledge network or graph [198, 205, 206].

This technology allows us to infer putative relations among molecules, such as understanding how proteins interact with each other or determining which gene mutations are involved in a disease.

NETME is a novel web-based app which is able to extract knowledge from a collection of full-text documents.

Moreover it is the first tool that allows to interactively synthesize biological knowledge-graphs on-the-fly starting from a set of n documents obtained through: i) a query to the PubMed database; ii) list of PMCID/PMID provided by the users; iii) a set of PDF documents. The inferred network contains biological elements (i.e., genes, diseases, drugs, enzymes) as nodes and edges as possible relationships.

To build a knowledge-network NETME operates in two main steps:

1. First, through OntoTAGME tool, it converts the full-text of the input documents into a list of entities using literature databases and ontologies (such as GeneOntology[207], Drug-Bank[208], DisGeNET [209], and Obofoundry [210] as corpus. These entities will be the knowledge graph nodes.

2. Next, an NLP model based on Python SpaCy[211], and NLTK[210–212] libraries, is executed to infer the relations among nodes entity-nodes belonging to the same sentence (S_i) or to the adjacent ones (S_i, S_{i+1}) of the same document.

These relationships can indicate disease treatment, gene regulations, molecular functions, gene-gene interactions, gene-disease interactions, gene-drug interactions, drug-disease interactions, disease-disease interactions and drug-drug interactions.

NETME allows users to build networks composed of several biological entities such as: genes, variants, diseases, drugs, compounds, molecular function, biological proves, pathways, enzymes, etc.
In building this network it handles the disambiguation among gene symbols and the acronyms of diseases or other biological elements, a very common problem since in many documents, authors assign acronyms for very long biological elements that are usually equal to genes symbols.
Moreover our system has an easy-to-use web interface in which users can specify various search parameters (number of articles from PubMed to work on, criteria used to sort articles, specific items on which one wants to focus the query, etc.).
The result of the network inference procedure is a direct graph (network) which shows all inference details in three main tables containing: the list of extracted papers, the list of annotations, and the list of edges together with their weight.
Users can then click on nodes in the network to view all incoming and outgoing connections, or click on particular edges to view the type and verbal relationship between the nodes they link.
Examples of applications include: i) analyzing disease networks for identifying disease-causing genes and pathways [213]; ii) discovering the functional interdependence among molecular mechanisms through textcolorrednetwork inference and construction[205]; iii) releasing Network-based inference models with application on drug repurposing [214].
The reliability of the knowledge graphs generated by NETME was tested through two case studies. The first one aims to provide a comprehensive analysis of the NETME performances by verifying its ability to predict known relationships between genes drawn from Kyoto Encyclopedia of Genes and Genomes - KEGG[2, 153, 196, 214] or Reactome[4-6] and, on the other hand, its ability to avoid the inference of false connections between proteins using the Negatome 2.0 database[215, 216]. The second case study is a more specific application and focuses on developing a network starting from a precise gene, that is CD147, this time using a selection on previously selected papers. In both cases, the performance of NETME was measured in terms of the precision/recall curve.

In the following section, will be described the first case study, whereas the second will be widely discussed later in the thesis, since it concerns a very important application related to the proposed novel drug repositioning pipeline.

3.3.1 Case study

The first case study focuses on assessing NETME performance through its capability to recover known gene interactions. For this purpose, we selected a subset of gene-gene interactions from KEGG/REACTOME by making use of STRING API.
More precisely, such interactions were obtained by selecting 100 random gene-gene interactions for each of the following STRING text-mining score intervals: 500-600, 600-700, 700-800, 800-900, >= 900. These interactions form the true-positive set.
Next, we selected 100 random pairs of non-interacting genes from the Negatome 2.0 database as a true-negative set (listed in Table 3). For each interacting gene-pairs, we queried NETME with the papers used by STRING to infer the interactions. On the other hand, to annotate non-interacting genes, we queried NETME with the pair of genes of interest, selecting the top 20 papers from PubMed.

Non-interacting genes from Negatome 2.0							
SOURCE	**TARGET**	**SOURCE**	**TARGET**	**SOURCE**	**TARGET**	**SOURCE**	**TARGET**
AKT1	TSC1	MAD2L2	MAD1L1	CTNND1	APC	TANK	RBCK1
ARAF	BCL2L1	NCK1	EGFR	CTNND1	CTNNA1	TBC1D7	TSC2
ARAF	BCL2	OSM	LIFR	CTNND1	CTNND1	TFDP1	CDK2
BCL10	BIRC3	PARD3	LIMK1	CTNND1	CTNNB1	TFDP1	CCNA1
BCL2L1	MAVS	PDGFC	FLT1	DKK1	WNT1	TICAM1	TLR4
BMPR1A	TGFB1	PFN4	ACTB	DKK1	SOST	TJAP1	F11R
BMPR1A	BMP5	PGF	KDR	DVL1	TSC1	TJAP1	CLDN1
BMPR1A	BMP6	PIAS3	STAT1	EIF3I	ACVR2A	TJAP1	TJP1
BMPR1B	TGFB1	PIK3CG	PIK3R2	EIF3I	ACVR1	TNF	EGFR
BMPR1B	BMP5	PKN1	RPS6KA1	EIF3I	TGFBR1	TRADD	TNFRSF10A
BMPR1B	BMP6	PKN1	RPS6KA3	EP300	CD44	TRADD	TNFRSF10B
BMPR2	BMP2	PKN1	MAP3K2	ERBB2	PIK3R2	TRAF6	IRF3
CCND1	MCM2	PKN2	RPS6KA1	ETS1	CREBBP	TSC1	CDKN1B
CCR3	CCL3	PKN2	RPS6KA3	FOXO1	TSC1	VAV1	SHC1
CCR3	CCL4L2	PKN2	MAP3K3	GRAP2	SOS1	VEGFB	KDR
CD274	CD28	RB1	SMAD3	GRAP2	CBL	VEGFB	FLT4
CD274	CTLA4	RBL2	SMAD3	HDAC2	RELA	VEGFC	FLT1
CD274	ICOS	RIPK1	TNFRSF10A	HIPK2	MDM2	VIPR2	RAMP1
CD3G	ZAP70	RIPK1	TNFRSF10B	HSPA4	BAX	VIPR2	RAMP2
CD74	NOTCH1	SFN	TSC1	IGF2	IGF1R	VIPR2	RAMP3
CDKN1B	TSC1	SH3KBP1	TNFRSF14	IL15	IL2RA	VWF	F8
CSF2	IL3RA	SMAD1	ANAPC10	IL1A	EGFR	YWHAB	TSC1
CTNNB1	HSP90AA1	SMAD4	ANAPC10	IL22	IL10RA	YWHAE	TSC1
CTNNB1	DDIT3	SOCS3	JAK2	IL4R	IL13	YWHAZ	TSC1
CTNND1	IL2	STIM1	TRPC6	KDR	FLT1	NFKBIA	CREB3L2

Table 3. List of the first 100 pairs of non-interacting genes from the Negatome 2.0 database. The column "SOURCE" indicates the starting gene, instead the column "TARGET" indicates the gene to which the action of the source gene is directed.

Accuracy, sensitivity, specificity and PPV values, detected by NETME, are listed in Table 4

Text-Mining Score Interval	Accuracy	Sensitivity	Specificity	PPV
500-600	58.5%	31%	86%	68.8%
600-700	66.5%	47%	86%	77.05%
700-800	72.5%	59%	86%	80.8%
800-900	73.5%	61%	86%	81.3%
≥ 900	84%	82%	86%	85.4%

Table 4. Metrics on NETME's ability to predict known interactions (from KEGG/Reactome) and non-interactions (from Negatome 2.0) between genes.

The results clearly show that NETME produces reliable results when the annotations are performed on top of relevant literature (STRING text-mining score higher than 700). On the other hand, when the STRING text-mining score is lower than 700, the NETME performances degrade in accordance with STRING predicted confidence as highlighted by their score . The reason behind such a behavior is due:

(i) not enough literature about these interactions; (ii) the interactions have been inferred by human curators as a combination of other interactions occurring in the text. Furthermore, when the text-mining score is small, STRING predictions could be wrong. Indeed, as reported in [205], a score of 500 would indicate that roughly every second term of an interaction might be erroneous (i.e., a false positive). Therefore, the computed value of accuracy, sensitivity, specificity and PPV could be incorrect.

4

From Diseases Mechanisms of Action to Drug Repurposing: A novel Systems Biology approach

The following chapter aims to illustrate the new pipeline built for drug repositioning. This project should be a response to the necessity for specialists in the various fields of bio-medicine, to have an easy to use tool that allows to use the impressive amount of data from high-throughput experimentation to produce new knowledge regarding disease and drug mechanisms of action and therefore may help in the identification of potential drug candidates for therapies, including personalized medicines.
The pipeline that will be presented is designed to be a simple methodology that is easily applicable in different contexts, in other words, it is transferable..
The methodology exploits the potential of *in silico* simulations to predict disease-induced effects on the one hand and drug effects on the other, and then perform drug repositioning and ranking on the basis of the acquired knowledge.
For this purpose, we chose to use PHENSIM-Phenotype Simulator [1], to perform the predictions in silico. As previously described this is a web-based, easy-to-use computational tool that adopts a Systems Biology approach, and allows us to simulate how cellular phenotypes are affected by perturbation (activation/inhibition) of one or more biomolecules.
To calculate the deregulation effect of genes, proteins, microRNAs (miRNAs) and metabolites, PHENSIM uses a probabilistic algorithm, on all pathways present in KEGG, combined into a single meta-pathway[159] and integrates miRNA-target and transcription factor (TF)-miRNA information extracted from online public knowledge bases [125, 159].
As seen previously, PHENSIM offers the possibility to extend the meta-pathway with REACTOME to integrate a wider source of information for cellular networks. In addition, given the partial completeness of these networks (KEGG or REACTOME), PHENSIM allows to integrate the meta-pathway with new nodes and edges to complete biological pathways that are not completely connected and lacking in information, which is often a problem especially when the processes are crucial to the phenomenon under study.
Moreover, it is worth mentioning that PHENSIM offers the possibility to perform contextualized simulations in as many as 25 different organisms.
As the first step, the strategy involves (**i**) the creation of a "*molecular signature*" of the disease, i.e. the prediction of the effects caused by the disease in a specific biological context (e.g. cell line). This first step becomes extremely important taking into account that, having a systematic and global view of the effects caused by a specific disease, can improve knowledge in the field and also help the identification of potential drug targets. Thus, despite the enormous progress in understanding the molecular

mechanisms underlying diseases, knowledge is often only partial and selecting the appropriate approach to make full use of the large amounts of available data is still a challenge. The conceptual simplicity of the methodology is also reflected in the required data to perform it. In fact, it is expected, in this first step, the use of easily available experimental data, such as transcriptomics and/or proteomics. In fact, the signature can be determined using as input to the simulator two types of data: Differentially Expressed Genes (DEGs) calculated from transcriptomic data (transcriptomic approach) or Differentially Expressed Proteins (DEPs) from proteomic data (proteomic approach).
Simultaneously, the method involves (**ii**) predicting the effects of given drugs in the same biological context, i.e., building a database of drug signatures. Drug signatures are calculated by inputting multiple targets reported on databases such as DrugBank and/or derived from a careful literature survey.

To contextualize the input simulation will have to specify the genes not expressed [see chapter Related Work - PHENSIM- Phenotype Simulator].

After this, the methodology involves (**iii**) the proper application of the repurposing strategy, through the calculation of Pearson's correlation between the viral signature and the drug one. This will give rise to a correlation scoring system to evaluate candidates for drug repositioning in a given sample (cells, tissue). Eventually, ranking of possible repositioning candidates will be possible. Negative correlation will indicate that the drug is counteracting the action of the disease.

The method is designed to be transparent and not a black box. Specifically, during all steps, it is possible to inspect the perturbations for each node within the meta-pathway and understand their role within each biological pathway. Indeed, anti-correlation will then show, for each node within the meta-pathway and for each biological pathway, where the drug does or does not counteract the disease.
As discussed in the introduction to the thesis, the first area of application of the methodology concerns the study of host-pathogen interactions, with particular interest in viral infections in humans.
However, exactly when this new methodological approach was coming to fruition, at the end of 2019, a new virus of only a few nanometers (50-200 nm) in size showed its power by kneeling the entire planet. A novel member of the coronaviruses was reported in Wuhan, Hubei Province, China. It was a new viral strain more closely related to a bat coronavirus (RaTG13) that is now formally known as *Severe Acute Respiratory Syndrome Coronavirus 2* (**SARS-CoV-2**). The disease caused by this virus has been named *Coronavirus 2019* (**COVID-19**).

Starting with a local spread from Wuhan, the infection quickly diffused worldwide. On March 11, 2020, after assessing the levels of spread and severity of the viral infection, the World Health Organization (WHO) declared it a global pandemic, as it continuously spreads and holds the world hostage.

From the very beginning of the pandemic, in the early 2020s, we became part of an international group called *RxCovea*, which to date, gathers about 100 members from more than 15 countries, representing a broad swath of disciplines relevant to the present crisis including, but not limited to, epidemiological modeling, artificial intelligence, immunology, game theory, drug development, diagnostic screening and testing, economics, and data management (Fig. 21). In RxCovea we are all volunteers, and our objectives are neither fame nor fortune; we simply hope to be helpful [9].
The group was founded with the assumption that existing methods for understanding and managing this pandemic were inadequate, and that significant innovations are needed to address the current challenge, as well as to prepare to respond to any future pandemic effectively and efficiently.

Innovations cannot be imposed, managed, or predicted, but the chances are greatly increased when disciplinary boundaries are crossed and young scientists and technologists have unlimited access to mentors as they strategize against unprecedented challenges.

RxCovea using the discipline, experience, and leadership of the older generation, combined with the technological expertise of the younger generation, aims to use the talents of each individual to come up with innovative strategies on how to solve the unique challenges of Coronavirus Disease 2019 (COVID-19).

The complexity of the COVID-19 problem led to unconventional interactions and a non-hierarchical, multidisciplinary organization within the RxCovea group to generate ideas and tools, and then submit them to rigorous scrutiny and testing by proven experts.

Particularly along with my research group, we are part of the RxCOVEA subgroup that is specifically focused on drug repositioning for COVID19. It is in collaboration with the Curacao Biomedical & Health Research Institute (CBHRI) - Department of Immunology, Northwell Health Hospital and Courant Institute of Mathematical Sciences at New York University that we have developed and applied our novel drug repositioning pipeline.

Figure 21. The "cure COVid for Ever and for All" (RxCOVEA) Framework: A Global Network. Pan-national members of RxCOVEA superimposed on the pandemic viral genomic tracing map of COVID-19 spread (Gisaid.org). The serious health and economic consequences of the interconnected world in spreading the infection are contrasted by the rigorous science and technology projects, self-assembling spontaneously and equally rapidly around the globe (representative images shown of RxCOVEA member locality).

4.1 SARS-CoV-2

At the end of 2019, a new member of the coronavirus was reported in Wuhan, Hubei Province, China. The virus was a new strain most closely related to a bat coronavirus, RaTG13, that is now formally known as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2).
Recent studies have shown that RaTG13 has the highest similarity to SARS-CoV-2 (92-96% similarity) and establishes a separate order from other coronaviruses[217]. The disease caused by SARS-CoV-2 has been named Coronavirus 2019 (COVID-19).
Starting with a local spread from Wuhan, the infection spread very rapidly around the world and on 11 March 2020, the World Health Organization (WHO) declared it a global pandemic.
COVID-19 is considered one of the largest fast expanding pandemics since the 1918 Spanish flu with serious consequences for global health and economy. As of November 2021, SARS-CoV-2 has infected more than 250 million people, and caused 5.148.221 deaths (WHO Coronavirus Disease Dashboard, https://covid19.who.int)(Fig.20)[217, 218].

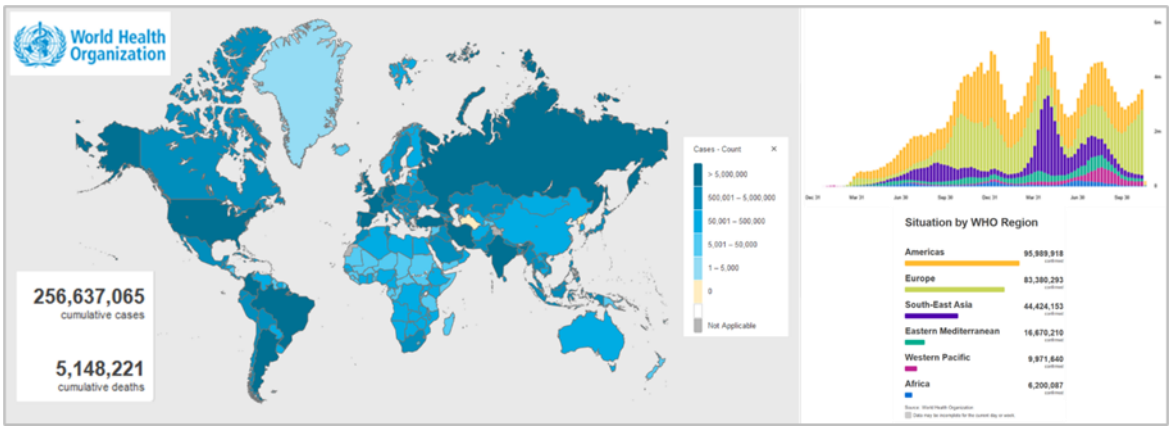


Figure 22. Global Report on COVID-19 pandemic by WHO. Globally, as of 1:11pm CET, 22 November 2021, there have been 256.637.065 confirmed cases of COVID-19, including 5.148.221 deaths, reported to WHO. As of 18 November 2021, a total of 7.370.902.499 vaccine doses have been administered. WHO Coronavirus Disease Dashboard, Image courtesy [217].

SARS-CoV-2 is a single-stranded, positive-sense RNA (+ssRNA) virus, which belongs to the β-coronavirus family and therefore it is an enveloped, single-stranded RNA virus. β-coronaviruses are able to infect wild animals, livestock and humans. While bats are the first suspected source of this virus, there may be an intermediate host in the bat-human transmission chain [220], for SARS-CoV-2 the pangolin has been hypothesized.
The single-stranded, positive-sense RNA genome has a 5′-terminal cap, a 3′-terminal poly (A) tail, and several open reading frames (ORFs). SARS-CoV-2 is composed of 13-15 ORFs (12 functional) containing about 30,000 nucleotides, encoding 27 proteins [220, 221].
At the 5′ terminus of the genome, there are ORF1a and ORF1b genes encoding 15-16 non-structural proteins (NSPs) from nsp1 to nsp10 and nsp12 to nsp16, respectively. Of those 15 make up the viral replication and transcription complex (RTC) that includes enzymes that process and modify RNA and an RNA repair function required to maintain genome integrity [222].
On the other hand, the 3′ term of the genome contains four structural proteins (S, E, M and N) and eight secondary proteins (3a, 3b, p6, 7a, 7b, 8b, 9b and orf14) [221].
Most of the viral nucleotide content (two thirds of the capped and polyadenylated genome) is held by the two non-structural proteins ORF1a and ORF1b followed by the structural proteins. Among the

proteins encoded by ORF 1a and 1b are the polyproteins pp1a and pp1ab, which are well conserved in all CoVs belonging to the same family [221].
The many NSPs play numerous roles in virus replication and assembly processes[222, 223].
These proteins participate in viral pathogenesis by modulating the regulation of early transcription, helicase activity, immunomodulation, gene transactivation and by counteracting the antiviral response [58][224]. Figure 23 shows a schematic representation of SARS-CoV-2 structure and genome.

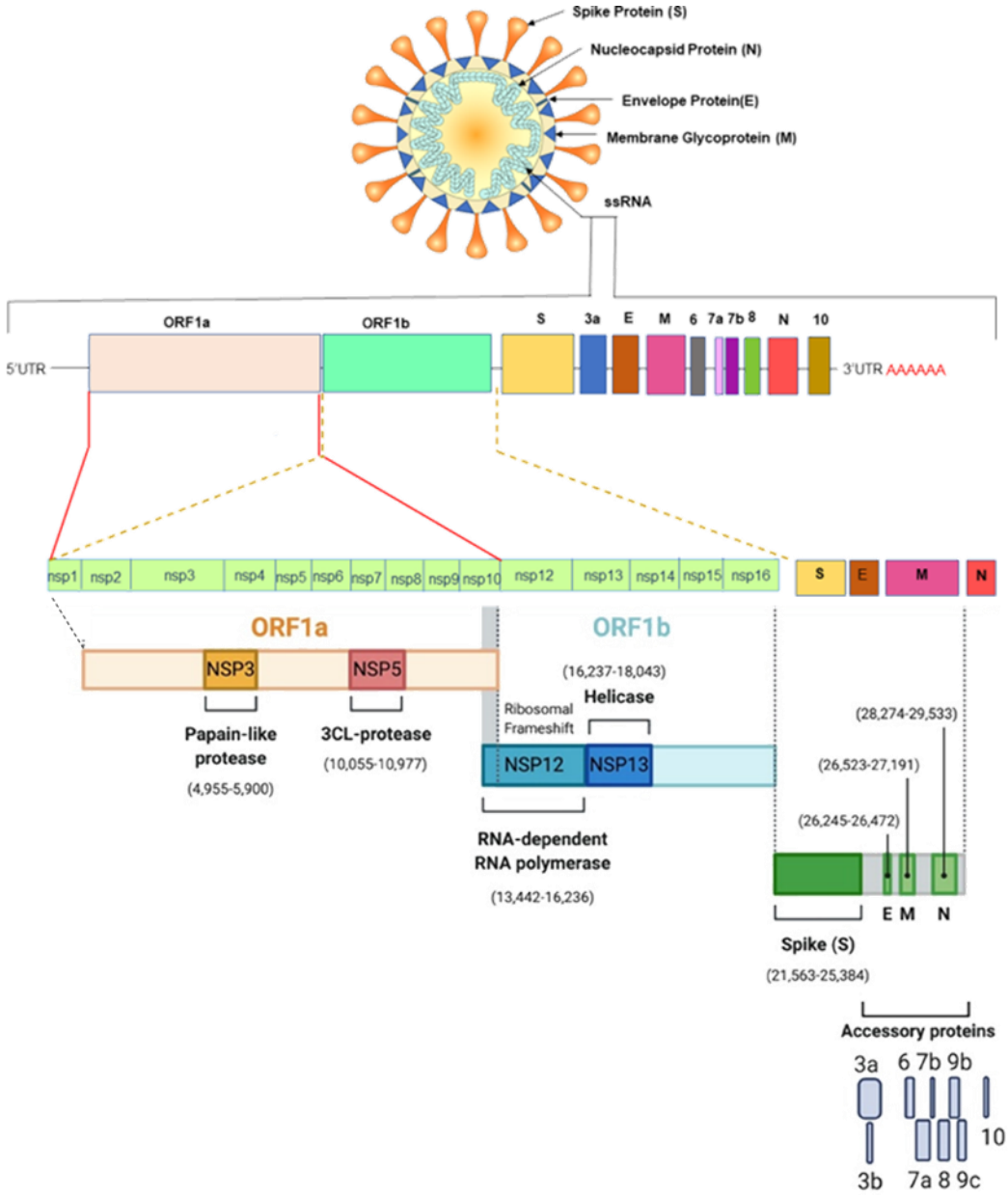


Figure 23. Representation of SARS-CoV-2: structure and genome. SARS-CoV-2 is a Betacoronavirus, with a spherical shape and surrounded by an external lipid envelope, covered by spike glycoprotein. Its complete RNA (single-stranded, positive-sense RNA (+ssRNA)) genome comprises approximately 29,903 nucleotides and has a replicase complex, composed of ORF1a and ORF1b, at the 5′UTR. ORF1a and ORF1b encode 16 non-structural proteins (nsp1-16).
ORF1a encodes from nsp1 to nsp10, whereas ORF1b encodes for nsp12-nsp16. There are then four genes that encode for structural proteins: Spike gene (S), Envelope gene (E), Membrane gene (M), Nucleocapsid gene (N) and a poly tail (A) at the 3′UTR. Accessory genes are distributed among the structural genes. Image adapted from [225]and [226]

4.1.1 Viral infection and molecular mechanism of action

SARS-CoV-2 spreads from one person to another through direct contact or over short distances in the air, either impacted in aerosol droplets or carried on fomites. The primary reproduction number (R0) of the person-to-person spread of SARS-CoV-2 is about 2.6, which means that infected cases grow at an exponential rate [227].

After inhalation of droplets containing viral particles, infection of mammalian lung epithelial cells begins with the virus binding to a specific receptor on the cell surface, via its Spike (S) protein. Like all the other viruses, coronaviruses require the host's cellular mechanisms to survive and replicate. Viral replication in these cells causes direct negative effects including the rapid and abundant secretion of cytokines and chemokines by local immune cells that cause the well-known cascade phenomenon called cytokine storm with harmful effects on the lungs and beyond.

In the following section, the mechanism of action of SARS-CoV-2 will be illustrated by organizing it into three distinct steps as follows, for simplicity:

1. Binding and Entry into the Host Cell;

2. Viral Transcription and Translation;

3. Virion Assembly and Release

Binding and Entry into the Host Cell

SARS-CoV-2 enters into the host cell by direct fusion of the viral envelope with the host cell membrane, or membrane fusion within the endosome after endocytosis. Virus entry into the host cell begins with the attachment of the virus to the cell surface through the binding of the Spike protein to a human host cell surface receptor. In SARS-CoV-2, as well as SARS-CoV, the main entry point to the host cell is the angiotensin-converting enzyme 2 (ACE2) receptor [228-230], which is widely expressed with the original structure conserved in a variety of animals, including fish, amphibians, reptiles, birds and mammals [230, 231]. In humans, ACE-2 is expressed on lung and gut epithelial cells and, in lower proportion, in kidney, heart, adipose, and both male and female reproductive tissues [232, 233].

The wide expression of ACE-2 in various tissues contributes to the multi-tissue infection by SARS-CoV-2 in humans. The entry of SARS-CoV-2 into the cells markedly down-regulates ACE-2 receptors, which favors the progression of inflammatory and thrombotic processes [233]

Coronavirus Spike (S) proteins are homotrimeric class I fusion glycoproteins that consists of two subunits, S1 and S2, with S1 at the N-terminus providing Receptor Binding Function (RBD) and S2 at the C-terminus providing fusion activity (Fig.24). After RBD-receptor interaction proteolytic cleavage of coronavirus S-proteins (S1/S2) by host cell-derived proteases such as furin, TMPRSS2, and cathepsin B/L (CatB and CatL), is essential for viral-host membrane fusion [234-237].

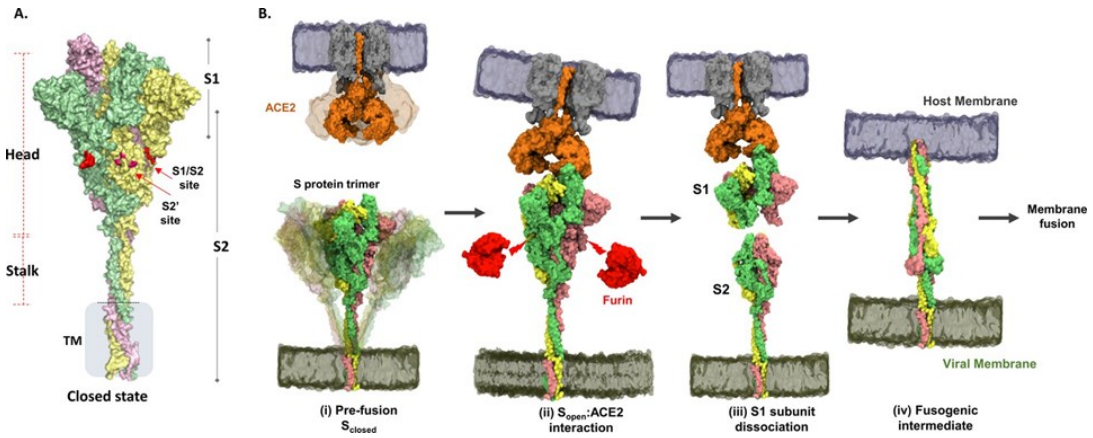


Figure 24. Structure of the trimeric spike (S) protein and virus-host entry initiated by S recognition and binding to the ACE2 receptor. (A) S protein construct shows a head, stalk, and transmembrane (TM) S1/S2 and S2′ cleavage sites are depicted in red. Proteolytic processing (furin) of the S protein generates the S1 and S2 subunits. (B) Schematic of viral entry into the host cell mediated by S-ACE2 interactions [238]. Binding to ACE2 induces conformational changes that promote proteolysis of Furin (red) at the cleavage site (red arrows), leading to dissociation of the S1 and S2 subunits, the mechanism of which is unknown. Furin here also denotes relevant related proteases. The residual ACE2-bound S1 subunit becomes stably bound to ACE2, and the S2 subunits dissociate. Image courtesy [239]

The way SARS-CoV-2 enters into a given cell type depends largely on the expression of proteases, in particular TMPRSS2. When TMPRSS2 (or other serine proteases such as TMPRSS4 or the human airway trypsin-like protease [HAT]) is expressed, the early entry pathway is preferred, whereas in the absence of this protease, the virus relies on the late pathway involving endocytosis and activation by cathepsin L (CTSL)[229, 230].

The understanding of the relationship between TMPRSS2 and SARS-CoV-2 infection since the early stages of the emergence was demonstrated starting from the pre-existing knowledge that expression of TMPRSS2, leads to activation of the SARS-CoV spike protein, allowing membrane fusion [240]. The TMPRSS2 gene encodes a type II transmembrane serine protease (TTSP) and is androgen-regulated and highly expressed in the prostate epithelium [241] this could support the increased prevalence of the virus in the male sex. Although ACE2 is considered the preferential entry point to the host cells other host receptors and/or co-receptors have been reported to promote the entry of SARS-CoV-2 into cells of the respiratory system.
Apart from ACE-2, one of the most important alternative entry points for the virus is Basigin receptor (BSG), also known as CD147 or EMMPRIN[242] (Fig. 25).

Furin, a proprotein convertase, is also known since a while ago to play a role in viral entry, and recent data support a role for this enzyme in particular in TMPRSS2-mediated cell surface entry (Fig. 25).

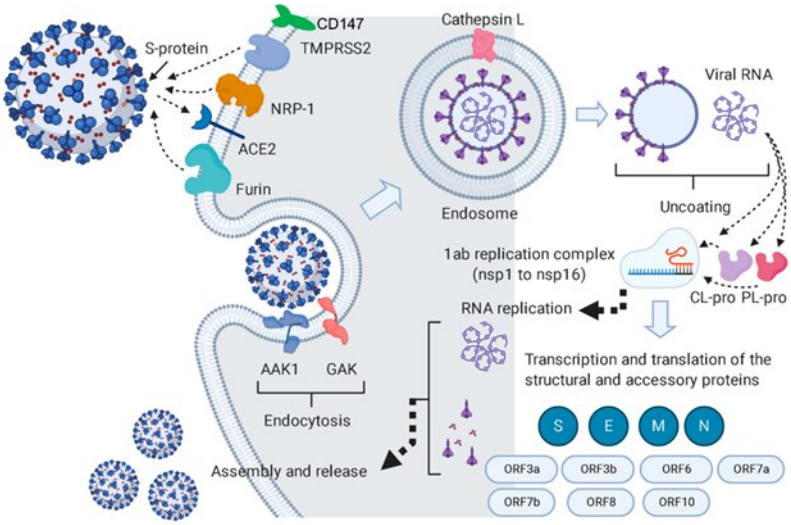


Figure 25. Schematic representation of the pathogenesis of SARS CoV-2 inside the host cell. It shows the key proteins involved in the viral life cycle. Image (modified) courtesy [242, 243].

Cleavage of the spike protein by furin at the S1/S2 cleavage site is thought to occur after viral replication in the intermediate compartment of the Golgi endoplasmic reticulum (ERGIC) (Fig.26)[244]. The S1/S2 Furin cleavage site is present in SARS-CoV-2 and MERS but not in SARS-CoV[245].
Recently, the VEGF-A receptor Neuropilin 1 (NRP1) has also been shown to be a host factor for SARS-CoV-2 spike protein flayed by Furin[245-247].

Unlike ACE2, which directly binds the RBD of SARS-CoV-2, neuropilin-1 interacts with RRAR residues (amino acids 682–685) only after the C-terminus of SARS-CoV-2 S1 protein is exposed by protease cleavage[248]. Therefore, neuropilin-1 serves as a "post-proteolysis receptor" for viral attachment on the surface of host cells [248].

Beyond such preferential entry routes for the virus, several in vitro studies have demonstrated an alternative endosomal-lysosomal pathway for SARS-CoV-2 entry. Indeed, coronaviruses in general, and SARS-CoV-2 in detail, are able to establish robust infection through endosomal entry within the in vitro cell culture systems commonly used. Even the understanding of the molecular events involved in the endosomal entry pathway is not fully understood, it is known that relevant functions in this pathway include CTSL, 1 of 11 cathepsins in humans [249, 250].

When the spike protein binds to the receptor via the S1 subunit, the receptor initiates severe conformational changes in the S2 subunits, which as a consequence lead to the fusion between the virus and the host cell membrane. This results in the release of the nucleocapsid into the cytoplasm where viral replication can take place[57].

Once the virus enters the cellular environment and the envelope is removed, in the cytoplasm the viruses express and replicate their genomic RNA using the host machinery to replicate its genetic material and assemble new viral particles.

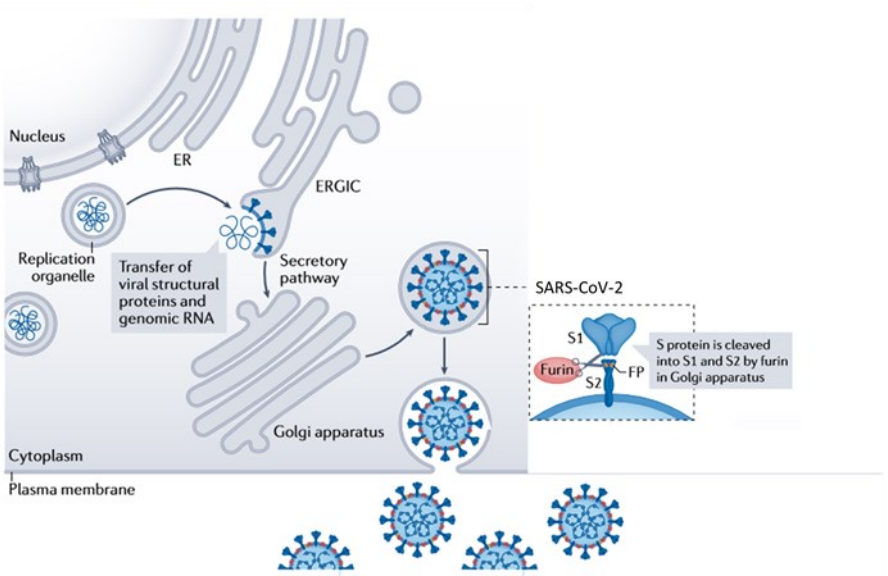


Fig.26 Coronavirus maturation. Infection by a coronavirus induces in the perinuclear area the formation of new membranous structures of various sizes and shapes, which as a whole are referred to as *replication organelles* [251, 252]. These structures originate from the endoplasmic reticulum (ER) and house viral replication complexes, sequestering them from cellular innate immune molecules. Viral structural proteins and genomic RNA synthesized at the replication site are then translocated through an unknown mechanism to the ER–Golgi intermediate compartment (ERGIC), where virus assembly and budding occur[253, 254]. Only four viral proteins, spike (S), envelope (E), membrane (M) and nucleocapsid (N) proteins, are incorporated into the virion. While the N protein bound to the viral genomic RNA is packed inside the virion, the structural proteins S, E and M are incorporated in the virion membrane. The S protein, assembled as a trimer, giving the appearance of a crown (corona), mediates major entry steps, including receptor binding and membrane fusion. During biosynthesis and maturation in the infected cell, the S protein is cleaved by furin or furin- like proprotein convertase in the Golgi apparatus into the S1 and S2 subunits, which remain associated [255, 256]. The S protein on the virus therefore consists of two non- covalently associated subunits with different functions: in the new target cell, the S1 subunit binds the receptor and the S2 subunit anchors the S protein to the virion membrane and mediates membrane fusion. The E and M proteins contribute to virus assembly and budding through the interactions with other viral proteins [257, 258]. Assembled viruses bud into the ERGIC lumen and reach the plasma membrane via the secretory pathway, where they are released into the extracellular space after virus- containing vesicles fuse with the plasma membrane. FP, fusion peptide. Image courtesy [259]

Viral Transcription and Translation

Viral genomic replication begins with the synthesis of full-length negative-sense genomic copies, which act as a template for the production of new positive-sense genomic RNA. The polymerase can alter the string pattern during discontinuous genome transcription and create a complex set of negative subgenomic RNA (sgRNA) strands used as a template for creating a complex set of positive sgRNA strands [57]. The S, M, and E proteins are translated and released in the endoplasmic reticulum, whereas N protein is released in the cytoplasm after translation. Most structural proteins of the virus undergo post-translational changes that are essential for their function[221].

Virion Assembly and Release

Before its release, the virus is assembled in the ER-Golgi Intermediate Compartment (ERGIC) in coordination with the M proteins [57]. The M protein plays a crucial role in the formation of the viral envelope; in fact, the interaction between M proteins generates the viral envelope scaffold, and M-S and M-N interactions facilitate the transport of other viral proteins to the assembly site[221, 260]. On the other hand, the interaction of the M-protein with the E-protein plays a role in component assembly and also induces membrane curvature [260].

Once assembled, viral particles bind to vesicles and are transported by exocytosis through the secretory pathway[221]

Like other RNA viruses, SARS-CoV-2, while adapting to their new human hosts, is prone to genetic evolution with the development of mutations over time, resulting in mutant variants that may have different characteristics than its ancestral strains. Several variants of SARS-CoV-2 have been described during the course of this pandemic, among which only a few are considered Variants Of Concern (VOCs) by the WHO, given their impact on global public health. Based on the recent epidemiological update by the WHO, as of June 22, 2021, four SARS-CoV-2 VOCs have been identified since the beginning of the pandemic:

Alpha (B.1.1.7 lineage): first variant of concern described in the United Kingdom (UK) - September 2020

Beta (B.1.351 lineage): first reported in South Africa - May 2020

Gamma(P.1 lineage): first reported in Brazil - November 2020

Delta (B.1.617.2 lineage): first reported in India - October 2020

Lambda (C.37 lineage): first reported in Peru - December 2020

Mu (B.1.621 lineage): first reported in Columbia - January 2020

Omicron (B.1.1.529 lineage): first reported in South Africa and Botswana - November 2021

Despite the unprecedented speed of vaccine development against the prevention of COVID-19 and robust global mass vaccination efforts, the emergence of these new SARS-CoV-2 variants threatens to overturn the significant progress made so far in limiting the spread of this viral illness[261].

4.1.2 Pathogenesis and Molecular mechanism of infection

SARS-CoV-2 infection activates the innate and adaptive immune response, supporting, for most cases, the resolution of COVID-19 disease.
Indeed, effective antiviral responses of host innate and adaptive immunity (production of proinflammatory cytokines, activation of T cells, CD4 and CD8+), are essential for clearance of infected cells[261-263]. However, the tissue damage caused by the virus might induce the exaggerated production of proinflammatory cytokines, recruitment of proinflammatory macrophages and granulocytes leading to macrophage activation syndrome - MAS (or secondary hemophagocytic lymphohistiocytosis - sHLH), and to tissue damage [264, 265].

Only a small percentage of patients, characterized by a huge production of cytokines (cytokine storm) progresses to severe pneumonia and eventually develops acute respiratory distress syndrome (ARDS), septic shock and/or multiple organ failure. The severity of COVID-19 is related to the level of proinflammatory cytokines and immune cell subsets [265-267].

From the molecular point of view, following viral Spike protein binding to host cells via the ACE2 receptor, viral RNAs, as pathogen-associated molecular patterns (PAMPs), are detected by recognition

receptors, which include the family of Toll like receptors (TLRs). Viral RNA or intermediates during viral replication, including dsRNA, are recognized by both endosomal RNA receptors, TLR3 and TLR7/8, and the cytosolic RNA sensor, retinoic acid-inducible gene (RIG-I)/MDA5[268].

After identifying the RNA of the virus, the signaling cascades pathway of NF-kB and IRF3 is activated. These transcription factors, once translocated into the nucleus induce expression of inflammatory cytokines and IFN-I, which results in the immune system's initial response to a viral attack [268, 269].

Studies of the molecular mechanisms underlying SARS-CoV-2 infection show that, as also other viruses, it can adopt strategies to evade and modulate the host's innate immune response, hence evading immune detection and dampening human defenses at least at the onset of infection.

For instance it was seen that it evades the antiviral effects of type I and III IFNs at multiple levels, including the induction of IFN expression and cellular responses to IFNs [270]. Furthermore, SARS-CoV-2 and other coronaviruses replicate within double-membrane vesicles, preventing recognition of dsRNA replication intermediates by cytosolic RLRs [253].
Furthermore, a modification of the cap structure of viral RNA by Nsp16, which has 2′-O-methyltransferase activity, prevents MDA-5-mediated detection of viral RNA [251, 269] e la successiva produzione di IFN-β [272]. Moreover, modification of the viral RNA by Nsp14 of SARS-CoV, which has guanine-N7-methyltransferase activity, mimics the 5′ cap structure of host mRNAs, allowing the efficient escape of viral RNA from detection by RIG-I [271, 273].
Failure to suppress the virus with an adequate primary response leads to viral replication and propagation, resulting in the production of large amounts of INF-I (at advanced infection), followed by chemotaxis of neutrophils and macrophages in the lungs, causing the release of inflammatory cytokines [269, 271, 273] as described below.

Patients with COVID-19 showed increased plasma levels of proinflammatory cytokines including IL-1β, IL-1RA, IL-2, IL-6, IL-7, IL-8, IL-9, IL-10, IL-17, FGF, granulocyte-colony-stimulating factor (G-CSF), GMCSF, IP-10, monocyte chemoattractant peptide(MCP)-1, macrophage inflammatory protein (MIP)-1α, MIP-1β, PDGF, VEGF, CCL3, IFN-γ, and tumor necrosis factor (TNF)α [135, 268, 269, 274].

Specifically, it was found that plasma levels of IL-1β, IL-1RA, IL-7, IL-8, IL-10, IFN-ɣ, MCP-1, MIP-1A, MIP-1B, G-CSF, and TNF-α were increased since initial stage. Then, further analysis has shown that plasma concentrations of IL-2, IL-7, IL-17, IL-10, MCP- 1, MIP-1A, and TNF-α in ICU patients are higher than in non-ICU patients[275]. In addition, plasma levels of IL-2, IL-6, IL-8, IL-10, and TNF-α observed in severe infections are significantly higher than those in non-severe infections (Fig.27) [275, 276].

Increased plasma cytokine and chemokine levels and neutrophil-to-lymphocyte ratio (NLR) in patients infected with SARS-CoV-2 were correlated with disease severity ranging from mild to severe phase.
Such hypercytokinemia, the so-called "cytokine storm", has been proposed as one of the key leading factors that trigger the pathological processes leading to plasma leakage, vascular permeability, and disseminated vascular coagulation, observed in COVID-19 patients, and accounting for life-threatening respiratory symptoms [220].
While a rapid and well-coordinated immune response represents the first line of defense against viral infection, an exaggerated host inflammatory response and a dysregulated host adaptive immune response can cause tissue damage both at the site of infection and systemically. The excessive proinflammatory host response has been hypothesized to induce an immune pathology resulting in the

rapid course of acute lung injury (ALI) and ARDS occurring in SARS-CoV-2 infected patients [220, 275, 277].
During the infection, both innate and adaptive immune cells synergistically participate in the antiviral response[278].
After entry and replication of SARS-CoV-2 into epithelial cells, the cells are damaged and lysed so that the viral content can spread. Viral antigens are presented to CD8+ T cells and natural killer (NK) cells. Subepithelial dendritic cells (DCs), as well as tissue macrophages (MΦ), can recognize SARS-CoV-2 antigens and present them to CD4+ T cells via MHC class II molecules after epithelial destruction, and trigger T cell differentiation into Th1, Th17, and follicular helper (TFH) memory cells[217, 278]. TFH cells can help B lymphocytes to differentiate into plasma cells (PCs) and promote the generation of different antibody isotypes that start with IgM production [279].
B cells are able to directly recognize SARS-CoV-2 and present viral antigens to Th lymphocytes via MHC class II molecules that then contribute to the induction of MΦ. Typically, after the primary IgM response, there is an isotypic switch to IgG that can lead to long-term immunity[279, 280]. However, SARS-CoV-2 appears to be able to impair the development of long-term antibody responses by blunting the germinal center (GC) response [281].

In patients with severe COVID-19, there was a marked decrease in absolute number levels of CD4+ cells, CD8+ cells, B cells[282] and circulating natural killer (NK) cells [220, 275, 282, 283] as well as a decrease in monocytes, eosinophils, and basophils [284-286]. It is supposed that one of the causes may be functional impairment and overexpression of activation or depletion markers, such as FAS, TRAIL and caspase 3, in the case of CD4+ and CD8+ T cells [287].

Novel SARS-CoV-2 has been shown to primarily affect lymphocyte count and balance.
Specifically, it was seen that deceased COVID-19 patients had a lower percentage of CD3+, CD4+, and CD8+ lymphocyte populations than survivors, which are strong predictors of mortality, organ injury, and severe pneumonia [288].
SARS-CoV-2 infection can lead to immune dysregulation through affecting the subsets of T cells. A retrospective study done in Wuhan, shows a significantly lower numbers of total T cells, both helper T cells and suppressor T cells, in patients with severe COVID-19 [276].

Particularly, naïve and memory T cells are key immune components, whose balance is crucial for maintaining a highly efficient defensive response [268]. Naïve T cells enable defenses against novel and previously unrecognized infections through a massive and tightly coordinated release of cytokines, whereas memory T cells mediate the antigen-specific immune response[268]. A dysregulation in their balance, favoring naïve T cell activity over regulatory T cells, could highly contribute to hyperinflammation [268]. A reduction in memory T cells on the other hand could be implicated in COVID-19 recurrence [268, 289-291].

Therefore, COVID-19 causes immune dysregulation by inducing an abnormal cytokine and chemokine response, modulation of total neutrophils, and lymphocytopenia, all of which could enhance the cytokine storm and cause further tissue damage, resulting in a severe disease course and worsening prognosis.

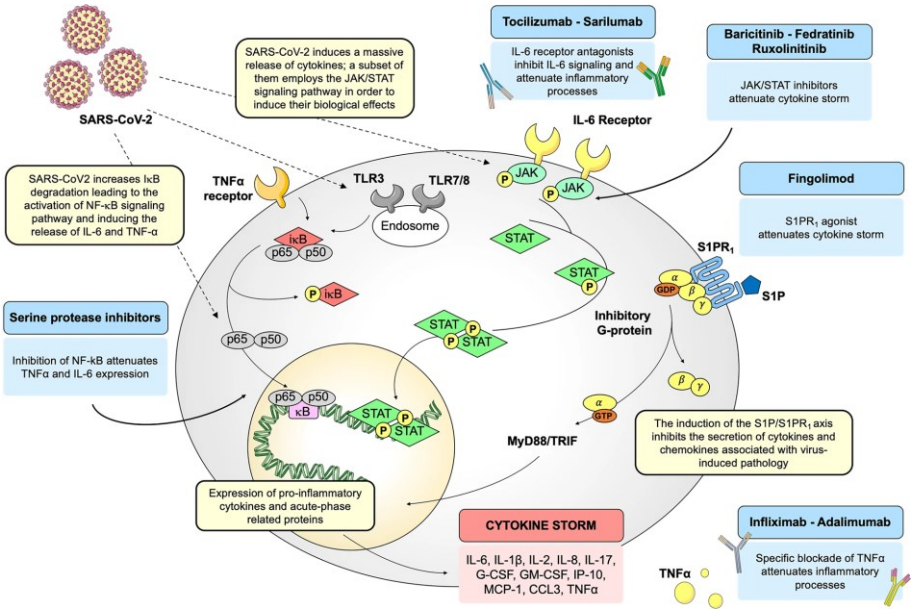


Figure 27. Schematic representation of SARS-CoV-2-driven signaling pathways and potential drug targets. Schematic representation of host intracellular signaling pathways induced by SARS-CoV-2 infection. Selected drugs, acting on these pathways, are repurposed to manage the cytokine storm induced by the viral infection. SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; IκB, inhibitor of nuclear factor κB; NF-κB, p65-p50, nuclear factor κB; IL-6, interleukin 6; IL-1β, interleukin 1β; IL-2, interleukin 2; IL-8, interleukin 8; IL-17, interleukin 17; G-CSF, granulocyte-colony stimulating factor; GM-CSF, granulocyte macrophage-colony stimulating factor; IP-10, IFN-γ-induced protein 10; MCP-1, monocyte chemoattractant protein 1; CCL3, chemokine (C-C motif) ligand 3; TNFα, Tumor necrosis factor α; JAK, Janus kinase; STAT, signal transducer and activator of transcription; S1P, sphingosine-1-phosphate; $S1PR_1$, sphingosine-1-phosphate receptor 1; MyD88, myeloid differentiation primary response gene 88; TRIF, TIR-domain-containing adapter-inducing IFN-β. Image courtesy [268]

4.1.3 COVID 19: more than a pulmonary disease

The first reports describing pneumonia due to infection with the novel coronavirus were published in NEJM on January 24, 2020 [274] and in Nature on February 3, 2020 [277]. In these reports, the initial identification of SARS-CoV-2 was done by sequencing and phylogenetic analysis on lower respiratory tract and bronchoalveolar lavage (BAL) fluids collected from patients in Wuhan from December 21, 2019 onward.

Approximately 80% of COVID-19 cases are asymptomatic or present mild to moderate symptoms, but about 15% progress to severe pneumonia and about 5% eventually develop acute respiratory distress syndrome (ARDS), septic shock and/or multiple organ failure [220, 275]. The most common symptoms of COVID-19 are fever, fatigue and respiratory symptoms, including cough, sore throat and shortness of breath, respiratory disorders, including neurological, cardiovascular, intestinal, and kidney dysfunction [268].

The primary site of infection of SARS-CoV-2 is the lower respiratory tract. In the respiratory tract, SARS-CoV-2 RNA and/or antigens were observed primarily in respiratory ciliated epithelial cells and type I and II pneumocytes, but also in alveolar macrophages [292, 293].

Underlying the respiratory disease there is an alveolar damage that includes alveolar edema, vascular decongestion, and inflammatory infiltration and appears to lead to pulmonary fibrosis [294]. Rarely hypercoagulation leading to death via pulmonary embolism [295]. In cases where there are severe symptomatic manifestations, it causes severe pneumonia, mainly due to activation of the inflammasome and pyroptosis[296]. It is known that canonical and non-canonical activation of pyroptosis can trigger the release of IL-1β, an interleukin that has been found increased in the serum of SARS-CoV-2 positive patients. Specifically, it appears that SARS-CoV-2 causes cellular pyroptosis in lymphocytes through activation of NLRP3[297].

Although the lungs are considered "*viral ground zero*", it is now known that SARS-CoV-2 affects many organs in the human body by both a direct viral infection or indirect effects of the immune response. Given the role of ACE2 as a major cellular entry point for SARS-CoV-2, some studies have attempted to map ACE2 expression to obtain information about the tissues or cell types that are theoretically susceptible to SARS-CoV-2 infection. Interestingly, the presence of SARS-CoV-2 components in different tissues does not always correlate with ACE2 expression levels.

At the neurological level, ACE2 has been shown to be expressed in both neuronal and nonneuronal cells. In the human central nervous system, it is especially expressed in the spinal cord, dorsal root ganglion, brainstem substantia nigra, choroid plexus, hypothalamus, hippocampus, middle temporal gyrus, and posterior cingulate cortex[298, 299]. Interestingly, non-neuronal cells rather than neuronal cells residing in the epithelium and olfactory bulb express ACE2 and TMPRSS2[298, 2300-302].

Given the presence of these receptors, it is believed that SARS-CoV-2 can directly invade the central nervous system. Furthermore, induced damage to olfactory receptors is correlated with the characteristic olfactory sensorineural loss that occurs early in the course of the disease. Neurological damage from COVID19 is of primary importance. It is enough to consider that approximately 78% of COVID-19 patients show neurological symptoms ranging from headache, loss of smell (anosmia) and taste (ageusia), imbalance, altered consciousness, delirium and paresthesia to paralysis of the extremities and seizures[301–309]. Severe neurological symptoms can mostly be attributed to abnormalities located in

the brain (trunk) and spine such as edema, hemorrhage, and thrombotic events with or without stroke, demyelination, and encephalomyelitis [309-312].

However, it remains unclear whether the severe neurological manifestations are triggered by direct virus-induced damage or by virus-induced endothelial damage and/or cytokine disturbances.

Ocular symptoms remained rare. COVID-19 hospitalized patients have manifested dry eyes, blurred vision, foreign body sensation, and conjunctivitis with conjunctival congestion. If we go to see how ACE2 is expressed in these organs, it has been found to be restricted to the retina and specifically in the retinal epithelium[312] whereas in corneal and conjunctival epithelial cells both ACE2 and TMPRSS2 are co-expressed[314, 315]. Therefore, infection could occur via droplets that enter the eye and travel through the nasolacrimal canal to the respiratory tract.

The cardiovascular system is also compromised in cases of severe disease. Approximately 20% of patients admitted to intensive care units have developed acute cardiac injury during the course of infection [275, 316]. Although the relatively high expression of ACE2 in cardiac tissue supports potential direct infection [317, 318], it is unknown whether SARS-CoV-2 facilitates cardiac injury through direct infection or by triggering inappropriate immune activation.

Approximately 10-15% of COVID-19 cases report gastro-intestinal (GI) tract symptoms including diarrhea, nausea, vomiting, or abdominal pain [319-321]. Expression of ACE2 is elevated in epithelial cells throughout the gastrointestinal tract, including the oral mucosa of the tongue and in enterocytes of the ileum and colon[321-325]. 35-56% of COVID-19 patients report abnormal liver tests (aspartate aminotransferase (AST), alanine aminotransferase (ALT), and bilirubin)[326, 327]. Virus was also detected in stools.

The human kidney may also be a target for SARS-CoV-2. Seventy-five percent of infected patients report abnormalities such as proteinuria, hematuria, and leukocyturia on urine tests [125, 126]. In addition, up to 27% of COVID-19 patients even develop acute renal failure, especially the elderly with comorbidities such as hypertension and heart failure [328, 329]. Virus (RNA, protein, and viral particles) has been found in the tubular epithelium of the kidney, where ACE2 is highly expressed[330-334], in podocytes, and to a lesser extent in the renal endothelium of deceased patients with COVID-19 [292, 329, 334-337].

Regarding the reproductive system, however, only a few cases of testicular pain in men are reported, nothing for women[338]. There is no evidence of viral transmission from mother to fetus and no evidence of SARS-CoV-2 has been found in the placenta, amniotic fluid or cord blood. In addition, the human placenta does not express high levels of ACE2[339-343].

Although obesity has been found to be the primary non-genetic risk factor for the onset of severe disease, it has not yet been detected in adipose tissue. *Poort et al. 2020* found that patients who have an increased body mass index (BMI) also have increased serum leptin levels, hypothesizing that this might correlate with the severity of infection[342-345]. In addition, a particularly interesting finding is that obese subjects report increased ACE2 receptor expression in lung epithelia, which correlates positively with infection and cardiovascular disease[344, 345].

Immune dysfunction has been extensively characterized in COVID-19 patients, such as dysregulation of T cells, B cells, and innate immune cells[345-347]. It is interesting that despite the determinant role of the immune response during SARS-CoV-2 infection, no viral replication is observed in patients'

immune cells. Postmortem studies of secondary lymphoid organs from COVID-19 patients confirmed the expression, albeit minimal, of the ACE2 receptor and the presence of the SARS-CoV-2 nucleoprotein in CD169+ macrophages[347, 348].

Recent studies also report expression, albeit minimal, of ACE2 and TMPRSS2 in innate immune cells such as monocytes and macrophages even in humans[349]. In addition, monocytes can express high levels of CD147[349, 350].

ACE2 expression levels are strongly influenced by prior conditions in patients and comorbidities. Cigarette smoking, diabetes, obesity, and hypertension induce elevated levels of ACE2 expression in the respiratory tract and are recognized as factors associated with COVID-19 disease severity[351-352].

The distribution of ACE2 expression levels may give indications about the organs targeted by the virus, although the existence of different entry points to the cellular environment implies that there is no absolute correspondence between the distribution of SARS-CoV-2 and ACE2.

Moreover, regarding the distribution and expression of ACE2, it is very important to distinguish between membrane-bound and soluble ACE2 molecules. Indeed, while the membrane-bound form acts as a host cell receptor for SARS-CoV-2, soluble ACE2 can neutralize free virions by shielding the viral spike (S) binding protein[354].Elevated levels of soluble ACE2 in the plasma of children may be one explanation why children often exhibit minor symptoms in SARS-CoV-2 infection, whereas the elderly are at increased risk for severe disease [355, 356].

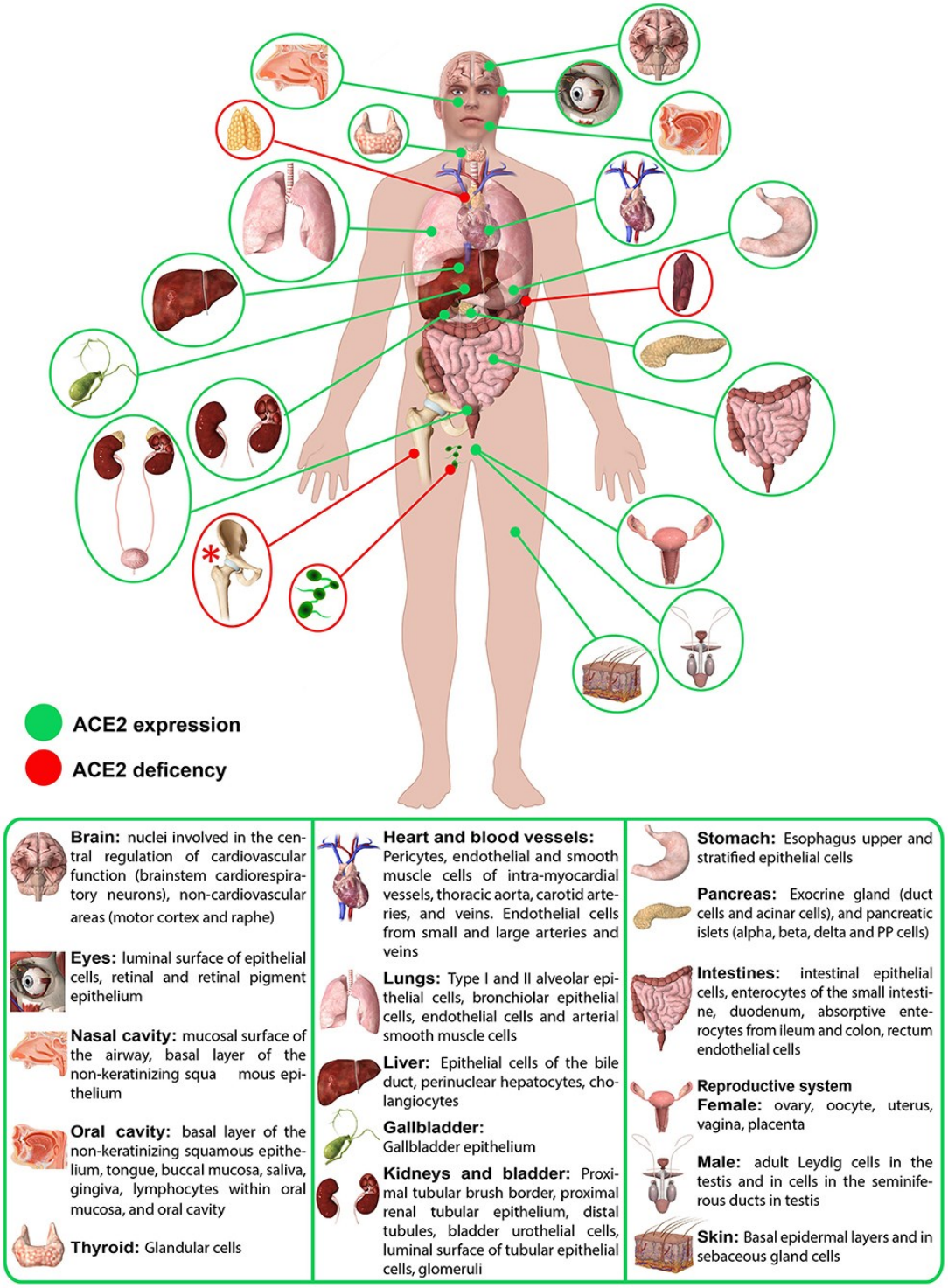


Figure 28. Schematic representation of ACE2 expression in human organs. ACE2 mRNA is present in all organs [268,357]. ACE2 protein expression is present in heart, kidney, testis, lung (type I and type II alveolar epithelial cells), nasal, and oral mucosa and nasopharynx (basal layer of the non-keratinizing squamous epithelium), smooth muscle cells and endothelium of vessels from stomach, small intestine and colon, in smooth muscle cells of the muscularis mucosae and the muscularis propria, in enterocytes of all parts of the small intestine including the duodenum, jejunum, and ileum (but colon), skin (basal cell layer of the epidermis extending to the basal cell layer of hair follicles smooth muscle cells surrounding the sebaceous glands, cells of the eccrine glands), endothelial, and smooth muscle cell of the brain [357]. Red asterisk (*): ACE2 deficiency only hypothesized. Image from [355]

4.1.4 Treatment

In the context of the current COVID-19 pandemic, significant progress has been made to develop prophylactic and therapeutic strategies that could successfully resolve the disease. The global scientific community has made immense efforts to shorten the timeframe for finding convalescent plasma, vaccines, neutralizing antibodies, and other antiviral drugs. Although several vaccines have been approved in various countries, through emergency authorization, a large proportion of the world's population currently remains unvaccinated due to disparities in vaccine distribution and limited production capacity and, unfortunately, also due to scientific misinformation circulating through the mass media (from TV to social networks).

Although vaccine development and deployment is underway, to date, only 42.4% of the world's population has undergone full vaccination. Indeed, widespread distribution remains a challenge, and at present only an antiviral (Remdesivir, given intravenously in patients with severe COVID-19 disease) and glucocorticoids (Dexamethasone/ Methylprednisolone) have been approved for treatment of severe COVID-19. The use of glucocorticoids has been proposed only in patients with markedly elevated C-reactive protein levels. In severe patients, early use of glucocorticoids appears to be associated with a significant reduction in mortality or mechanical ventilation. In contrast, glucocorticoid treatment in patients with lower levels of C-reactive protein is associated with worse outcomes [358].

Recently, the virus-neutralizing antibody cocktail (Casirivimab and Imdevimab, termed REGN-COV2) also received emergency use authorization for treatment of mild to moderate COVID-19 in high-risk patients. Otherwise, no established drug is available to prevent or adequately treat COVID-19 and in the absence of a clear etiological understanding, treatment has remained largely supportive and symptomatic[268, 359].

Due to the lack of treatment options (particularly in low- and middle-income countries), the slow progression in vaccination, and the emergence of SARS-CoV-2 variants eliciting reduced responses to vaccines, it becomes of utmost importance and urgency to implement new strategies for the development and identification of drugs that can help the specific treatment of the disease and reduce the morbidity and mortality of COVID-19.

An important contribution to speeding up experimental processes and choices has been made by advanced bioinformatics modelling tools. Therefore, alongside in vitro studies, in silico studies are of great importance for rapid and effective drug discovery. Indeed, computational structure-based drug design and immuno-informatics have recently resulted in identification of potential SARS-CoV-2 target proteins and drugs that are being selected for further testing[268, 360, 361]. Another promising avenue for obtaining effective and readily available therapeutic strategies is the repurposing of drugs already approved for other indications. Drug repurposing strategies provide an attractive and effective approach based on available drug characteristics – drug-related pharmacology and toxicology – for rapid therapeutic selection[360–362]. If we could, with higher probability, identify and pre-select the most promising hypothesis-based candidates using *in silico* systems biology tools, prior to costly and laborious *in vitro* and *in vivo* experiments and ensuing clinical trials, we could significantly improve disease-specific drug development[362, 363].

4.1.5 Rapid Identification of Druggable Targets and the Power of the PHENotype SIMulator for Effective Drug Repurposing in COVID-19

In the previous chapters it has been pointed out how drug repurposing strategies could provide an efficient alternative for a rapid treatment selection[362]. In this direction, a systems biology approach, which allows us to identify and pre-select the most promising candidates based on our specific requirements, prior to expensive and laborious in vitro and in vivo experiments and subsequent clinical trials, can significantly improve and accelerate the development of new drug schemes.

During a health emergency such as the one we are undergoing due to the outbreak of SARS-CoV-2, where urgent need for efficient care is manifested daily, having an easy-to-use instrument that allows fast and high-reliability drug screening becomes extremely powerful.

Several *in silico* techniques have been developed, mainly making use of molecular modeling of key viral proteins for virtual screening of drug candidates simulating receptor-drug molecular dynamics [268, 359, 360]. In order to increase the effectiveness of identifying candidate drugs for combating COVID-19, it is crucial to build on a more in-depth knowledge of the molecular basis of the immune signaling pathways regarding host-virus interaction and SARS-CoV-2-induced immunopathology. Only if we better understand how this particular virus affects host cells in detail, on a transcriptomic, proteomic level and beyond[268, 359, 360, 364], will we be able to effectively treat COVID-19 patients. It is becoming evident that treatment should not only focus on direct antiviral effects in mild cases but should also encompass potential (cytokine storm induced) aberrant host-response in severe cases[268, 365, 366].

Taken together, this points towards the importance of a more detailed and targeted approach for COVID-19, where antivirals or steroids alone might not suffice and specifically targeting the (aberrant) host-response is imperative [268, 360, 361, 366]. Recently in literature, tools and algorithms devised to perform simulation on biological networks have been described[151, 152].

Here we aim to utilize our systems biology tool, the PHENotype SIMulator (PHENSIM), to leverage the power of pathway analysis by simulating tissue-specific infection of host cells of SARS-CoV-2 and subsequently perform in silico drug selection for potential repurposing.

In the following paragraphs, the new approach for drug repositioning using PHENSIM will be described.

Next to that, will be presented the validation of the methodology by comparing our results with available data from recently published *in vitro* studies based on transcriptomics and proteomics in different model systems[359, 364]. Relevant and significantly affected pathways are further detailed on a protein interaction level. Finally, we show the potential of the PHENSIM in selecting promising hypothesis-driven COVID-19 drug candidates, which has applicability to other diseases and broader aspects of clinical practice, thereby outlining the potential power of PHENSIM in drug repurposing in COVID-19 and beyond.

4.1.5.1 PHENSIM approach

As described in chapter 3 PHENSIM is a systems biology approach, simulating the effect of the alteration of one or more biomolecules (genes, proteins, microRNAs, or metabolites) in a specific cellular context using KEGG (Kyoto Encyclopedia of Genes and Genomes) meta-pathways[1, 125, 364]. The meta-pathway concept, introduced by us previously[159], has been devised to account for pathway cross-talk in analysis. Essentially, all KEGG pathways are merged in a single graph through common nodes, where the meta-pathway is a graph in which the nodes represent molecular entities (genes, metabolites) and the edges are the known biological interactions present in the KEGG database. The meta-pathway is further completed by adding validated miRNA-targets downloaded from miRTarBase (release 8.0), miRecords and TF-miRNA-interactions obtained from TransmiR (release 2.0)[159-161]. The algorithm thus computes, under-specified biological contexts, by iteratively propagating the effects and alterations of one or more biomolecules (differentially expressed genes (DEGs), proteins, microRNAs, or metabolites), thus making use of published virus-human interaction data[161, 361].

To start the simulation, PHENSIM requires a set of biomolecules as input, their direction of deregulation (activation/up-regulation or inhibition/down-regulation), and a set of inactive biomolecules in the cellular context (cell lines, tissue, e.g). The algorithm uses these details to compute synthetic Log Fold Changes (LogFC). Synthetic LogFCs are computed by sampling the normal distribution fitted to the actual LogFCs of a particular gene, as described previously[1]. Such values are then propagated within the meta-pathway, using the MITHrIL (Mirna enrIched paTHway Impact anaLysis) pathway perturbation analysis[125]. MITHrIL determines how local change can affect the cellular environment by computing a “perturbation”. For each gene in the meta-pathway, the perturbation reflects its expected change of expression/activity (negative/positive for down-/up-regulation, respectively). Finally, these results are collected and synthesized using two values: the “Average Perturbation” and the "Activity Score" (AS). To recall what was described previously, given a node, the average perturbation is the mean for its perturbation values computed at each simulation step. It reproduces the expected change of expression for the entire process. The function of AS is twofold: 1) the sign gives the type of predicted effect (activation(+); inhibition(-)), 2) the value is the log-likelihood that this effect will occur. Together with AS, PHENSIM also calculates a p-value which determines how biologically relevant the predicted alteration is for the phenomena being simulated. All p-values computed by PHENSIM are corrected for multiple hypotheses using the q-value algorithm[125, 367]. To determine this probability, PHENSIM randomly selects genes in the meta-pathway and runs the simulation on this random set. By repeating this procedure (n=1000 for our simulations), it is possible to empirically estimate the probability that a node has a higher activity score than the observed one. For this reason, we can employ such a value to determine which alterations are most specific for a particular infection, gaining novel hypotheses on the molecular action of the pathogen.

PHENSIM pathogen alterations profile

Our approach defines a protocol for the *in silico* simulation of emerging pathogen infection, aimed at defining candidate drugs for repositioning. First, we find a representation of the pathogen in the KEGG meta-pathway, which allows us to perform simulations. For a novel pathogen, such as SARS-CoV-2, interactions with the host genes might be unknown. Therefore, we can approximate this by employing expression data of pre-/post-infection samples. The rationale is that differentially expressed genes (DEGs) represent the downstream effects of the viral infection on the host; *i)* we compute DEGs between pre- and post-infection samples, *ii)* we extend the meta-pathway by adding a new node representing the

virus, *iii)* the viral node is connected to each DEG with an activating (/inhibiting) edge if its LogFC computed between post- and pre-infection is positive (/negative), *iv)* we run a simulation by giving the upregulation of the viral node as input.
To build the *pathogen signature*, we use pathway endpoints; *An endpoint is a biological element in a pathway whose alteration, based on current knowledge, affects the phenotype in a specific way[125].* Given the output of this simulation, we collect all Endpoint Activity Scores in a single signature, the 'pathogen alterations profile'. This profile can be exploited to search for possible repositioning candidates, by building a *drug signature* database queried by means of a similarity measure (Fig. 29). When building the simulation profile, we do not use any p-value. Indeed, we need to consider not only the alterations, which are the most specific for a particular infection, but also the alterations caused by any cellular response to the infection. Since the p-value represents the biological relevance for the phenomena that is being simulated, we can ignore its value to build the signature.

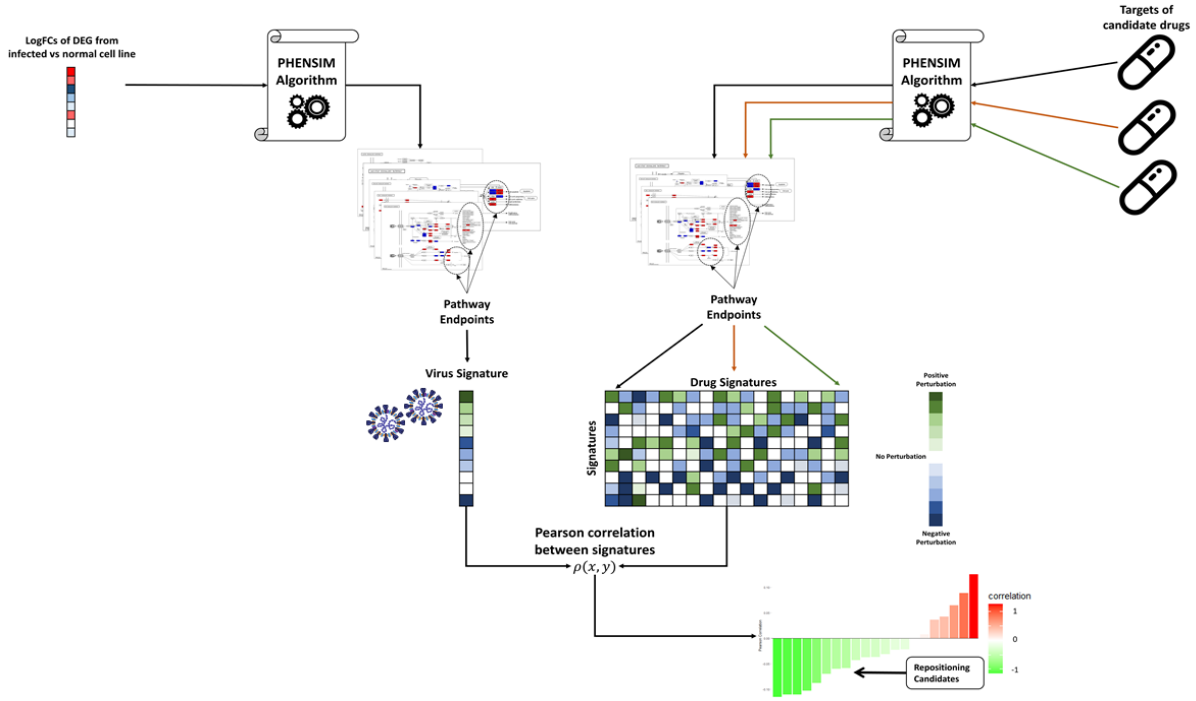


Figure 29. Schematic representation of the PHENSIM Drug repurposing Strategy. Outline for our approach to acquire a cell-specific viral signature in silico using a Transcriptomic strategy: logFold Changes (logFCs) of Differentially Expressed Genes (DEGs) arising from transcriptomic genome wide expression analysis of SARS-CoV-2 infected vs. baseline uninfected cells, cell-lines and tissues are the main input for the PHENotype SIMulator. Once a cell-specific viral signature is defined based on gene and signaling pathway endpoints using KEGG meta-pathway analysis, PHENSIMcan be exploited to search for possible repositioning candidates by building a drug signature database using the Drug repurposing strategy: multiple targets of drug candidates are used as input for PHENSIMto define drug signatures based on pathway endpoints. A Pearson correlation between the acquired virus and drug signatures ρ(x,y) gives rise to a correlation scoring system to evaluate drug repositioning candidates in a certain infected cell or tissue. Negative correlation (green) predicts promising targets that inhibit the viral signature and positive correlation (red) suggests exacerbation of the viral signature when introducing the drug.

PHENSIM drug signature database

Given a particular drug identified through databases (i.e. Drugbank or Pubchem) and literature (Pubmed) searches, we define all known targets and their alterations (up/down-regulations caused by the drug), and these alterations are provided as input to PHENSIM together with the same cellular context specified for the viral simulation. The results are used to define a drug signature using pathway endpoints as described above, which are collected in a database used for repositioning.

Furthermore, for each drug, we compute random models to empirically estimate repurposing p-values. Let *T* be the set of targets for a drug. First, we select a random set of targets of the same size as *T*. Therefore, we perform a PHENSIM simulation with the random set using the same alterations of the drug. Finally, we collect the signature as described above. This procedure is repeated to gather for each drug N random signatures (N = 1000 in our experiments) that are stored for the p-value computation phase.

PHENSIM drug repurposing approach

Our drug repurposing methodology is based on a similarity search performed on the drug signature database. Given a pathogen profile computed with PHENSIM, we use a correlation function to scan through each record in a drug profile database. This procedure yields a ranking on each drug in the range [-1;1], where negative values indicate that the virus alteration profile is opposite to the drug and positive values indicate the reverse; drugs with a negative correlation are considered possible candidates for repositioning. In our experiments, we employ a Pearson correlation function to run the similarity search. Since PHENSIM is based on MITHrIL pathway perturbation analysis, which computes results in a log-linear space[125], we can assume a Pearson correlation is sufficient to determine similarity between the viral and drug signature.

A key characteristic of this approach is the capability to simulate both single and drug combinations. Furthermore, PHENSIM also provides a framework for extending pathways by adding new nodes and edges coming from results in the literature as well as other reputable sources.

Finally, to assess whether each drug candidate targets relevant infection processes, we decomposed the Pearson correlation in terms of KEGG pathways and reviewed the results. More in detail, let D and V be drug and pathogen alteration profiles, respectively. That is, *D[e]* is the activity score of the endpoint "*e*" computed by PHENSIM for a drug simulation, and *V[e]* is the activity score for the same endpoint in the pathogen simulation. Pearson correlation $\rho_{D,V}$ can be written as *equation (1):*

$$\rho(D,V) = \frac{\sum_e (De - \overline{D})(V[e] - \overline{V})}{\sigma(D) \cdot \sigma(V)},$$

Where and are the means of $\overline{D}$ and $\overline{V}$, respectively, and $\sigma\cdot$ is the standard deviation. Therefore, given a pathway P, we can sum the Pearson correlation components belonging to its endpoints to estimate how much it contributes to the final correlation value. More in detail, the partial correlation $\hat{\rho}(D,V,P)$ can be computed as *equation (2):*

$$\hat{\rho}(D,V,P) = \frac{\sum_{e \in P} (D[e] - \overline{D}\,)\,(V[e] - \overline{V})}{\sigma(D) \cdot \sigma(V)},$$

A significant feature of this partial correlation approach is that we obtain the total correlation by summing up all values for each pathway P. Therefore, we can determine which biological processes are impacted by the drug administration.

Drug repurposing approach: p-value computation

To determine the significance of the results, we use an empirical approach based on a bootstrapping procedure. First for each drug *D*, we build *N* random signatures D'_i as described above. Next, for each

signature we compute the Pearson correlation coefficient against the pathogen profile, $\rho(D_i', V)$. Finally, we count the number of times we obtain a greater correlation to empirically estimate the p-value as:

$$p = \frac{|\{D_i' || \rho(D_i', V)| > |\rho(D, V)|\}|}{N}.$$

Finally, the p-values are corrected for multiple hypotheses using the Benjamini-Hochberg Procedure[368].

PHENSIM combined drug/pathogen simulation

To further evaluate whether the results of the correlation could be confirmed by PHENSIM, we devised a strategy to simultaneously simulate drug action and pathogen infection on a host cell line. First, we collected DEGs between pre- and post-infection samples as described in the previous section. Then, given a drug, we gather its known targets and their alterations (up/down-regulations caused by the drug) through databases (i.e. Drugbank or Pubchem) and literature (Pubmed) searches. Therefore, we extend the meta-pathway by adding two nodes, representing the virus and the drug, respectively. The virus node is connected to each DEG with an edge as described in the "*PHENSIM pathogen alterations profile*" section. Then, an activating (inhibiting) edge is added between the drug node and a target, for any up-regulated (/down-regulated) target. Finally, we can run a simulation by giving as input the simultaneous upregulation of both virus and drug nodes (results depicted in Fig. 29).

4.1.5.2 PHENSIM method validation

To determine the efficacy of our model we used several datasets obtained in the context of SARS-CoV-2 infection. For each dataset, we computed the genome-wide Log Fold Changes (FC). As PHENSIM does not require any quantitative information, DEGs were termed upregulated if LogFC > 0.6, and downregulated if LogFC < -0.6.

PHENSIM transcriptomic reliability assessment

To assess the reliability of the results, we focused on two fronts: ii) the ability of PHENSIMto predict genes altered in the expression data, and ii) the ability to predict the correct direction of the alteration. In detail, we define as altered all genes having an absolute LogFC > 0.6. The type of the alteration is given by the sign of the LogFC (+LogFC for upregulation, -LogFC for downregulation). Predictive power of PHENSIM was assessed by means of Positive Predictive Value (PPV), Sensitivity, and Specificity. The PPV is the proportion of true positive results with respect to all positive predictions, the sensitivity is the percentage of true positives with respect to the entire population, and the specificity is the percentage of true negatives with respect to all negative cases.

PHENSIM transcriptomic approach

For the evaluation of our strategy, we exploited transcriptomics data published in Blanco-Melo *et al*. 2020 [GSE147507][364, 368] and proteomics data coming from Bojkova *et al*. 2020[359].

The Blanco-Melo *et al*. dataset comprises RNA-seq data of infected vs. mock-treated cell-lines from human and ferret. The data were obtained by using the Illumina NextSeq 500 platform[364]. In our analysis, we focused solely on human cell data. In detail, 4 cell-lines were evaluated: primary human lung epithelium (NHBE), transformed lung alveolar (A549) cells, transformed lung alveolar (A549) transduced with a vector expressing human ACE2 and transformed lung-derived Calu-3 cells. For all cell-lines, sequencing data of biological replicates was obtained from mock treated or SARS-CoV-2 infected experiments. Furthermore, for both A549 cell lines different MOIs (multiplicity of infection) were used at low 0.2 and high MOI 2.0. Following the same procedure used by Blanco-Melo *et al*., raw counts were normalized and analyzed for differential expression using the DESeq2 pipeline[364, 369]. All genes with an FDR-adjusted p-value < 0.05 and absolute LogFC>0.6 were considered differentially expressed. Non-expressed genes for a specific cell-line were defined as genes that showed an average read count lower than 10.

PHENSIM proteomic approach

To determine if our methodology can also exploit proteomic data, we leveraged data from Bojkova *et al.* 2020, namely proteome measurement by LC–MS/MS of control vs. SARS-CoV-2-infected human Caco-2 cell lines[359]. All cell lines were analyzed in triplicates (n=3) at 2, 4, 10 and 24h. Log2-ratios between infected and normal differentially expressed proteins (DEPs) (p-value < 0.05) were used as input for the simulation algorithm. Non-expressed proteins for the Caco-2 cells were taken from the

Human Protein Atlas (using the query "celline_category_rna:CACO-2;Not detected"). Since PHENSIM uses Entrez Gene Identifiers, we mapped all proteins to their gene, yielding 5809 mapped proteins. Of these 5809 mapped proteins, we could find only 1914 in the KEGG meta-pathway. We therefore combined our PHENSIM analysis with an enrichment analysis to determine if a prediction made by our methodology was based on adequate data.

More in detail, for each time point and each pathway, we compared the number of altered proteins predicted by PHENSIM to the expected number of altered proteins using a hypergeometric distribution. This analysis yielded an enrichment p-value combined with PHENSIM one using the Fisher's Method [368, 369] due to their independence. P-values were corrected for multiple hypotheses using the Benjamini-Hochberg correction, and all pathways with a p-value < 0.05 were considered significant for further analyses.

As Bojkova *et al.* reported their results using Reactome Pathway analysis [359], we selected all pathways mentioned in their paper and also in the supplementary material published (translation, splicing, carbon metabolism and nucleic acid metabolism, for instance). Comparisons were performed using the Average Pathway Perturbations as reported by PHENSIM(Fig. 32D). Finally, since we found many similar metabolic pathways significantly affected *in silico* as described *in vitro* by Bojkova *et al.*, we aimed to determine if a core set of proteins was common between pathways; results are displayed in the VENN diagrams in Fig. 33.

PHENSIMCD147 gene and pathway extension

Prior to running all simulations, we verified if viral entry points were present in KEGG to better represent viral activity. SARS-CoV-2 is said to invade host cells via these two receptors: angiotensin-converting enzyme 2 (ACE2) and CD147 (also known as Basigin or EMMPRIN[372]. In KEGG the latter gene is missing and to extend our simulation model, a new node, representing CD147 was added and connected with its known interactions and downstream nodes according to literature [373-382]. The incoming edges to CD147 represent the possible activators/inhibitors (upstream genes), the outgoing edges represent the actions performed by CD147 towards its downstream genes. CD147 is a transmembrane protein of the immunoglobulin super family, expressed in many tissues and cells, acting as the main upstream the stimulator of matrix metalloproteinases (MMPs) and playing a crucial role in intercellular recognition[374]. Over the last decade, several groups have shown that CD147 acts as a key molecule in the pathogenesis of several human diseases including infectious diseases (HIV, HBV, HCV, KSHV)[372, 374], and it has now been posed to recognize and internalize/endocytose SARS-CoV-2 in certain cell types[372].

Drug Repurposing Validation

To evaluate our drug repurposing approach, we compared predictions made using our pipeline with the in vitro drug screening performed in *Stukalov et al. 2021*[7], using the previously described transcriptomic approach to build both pathogen and drug signatures. First, we gathered all drugs expression data from the LINCS L1000[8, 57] dataset for the A549 cell line. Next, we selected all drugs that were also tested in *Stukalov et al. 2021*[7], yielding 27 drugs. Therefore, for each drug we matched the concentration values from the L1000 datasets with the nearest one in *Stukalov at al*., building a dataset of 81 drug-concentration transcriptomic experiments. Then for each experiment, we gathered the differentially expressed genes and performed PHENSIM simulation as described above to produce drug signatures.

For the pathogen signatures we used the expression data from Blanco-Melo et al. for A549 and A549-ACE2 cell lines.

Finally, using the repurposing procedure described in the ”PHENSIM drug repurposing approach” section, we determined the correlation values for each drug experiment and computed the FDR-adjusted p-values. The results were, then, compared against the in vitro drug screening results from Stukalov et al. Such screening results were reported as treatment-induced changes in virus growth over time using the $\log_2$ fold change of the GFP signal normalized to the total cell confluence between the treated and control conditions (Figure 30).

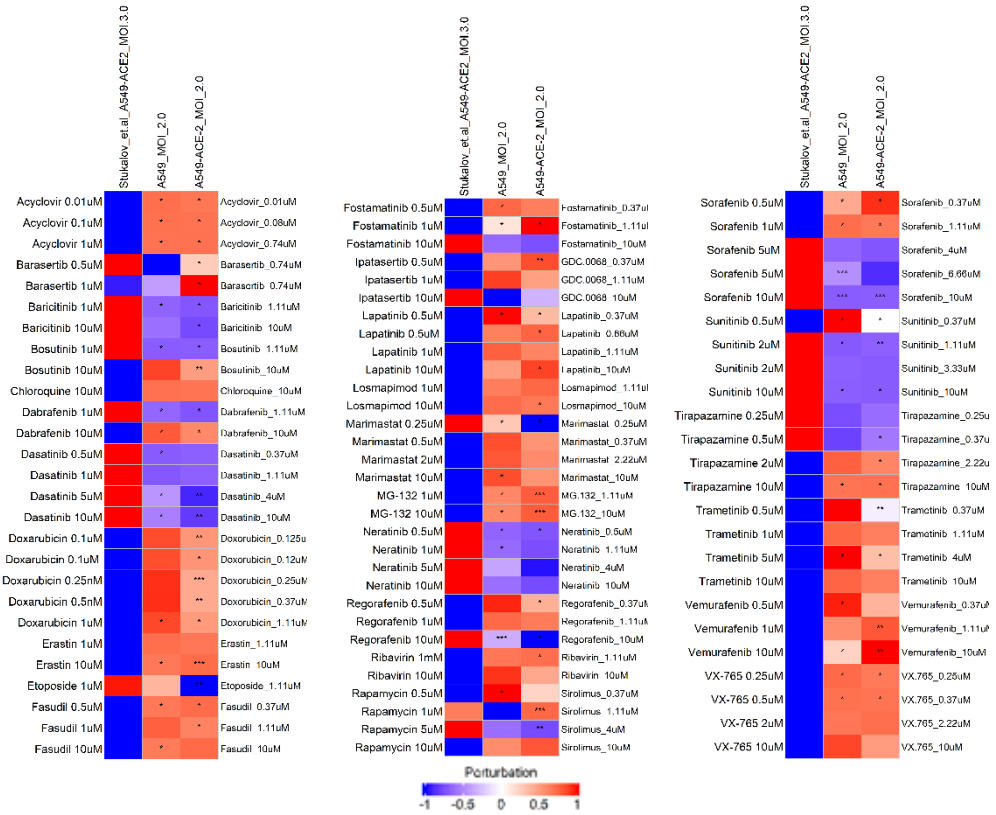


Figure 30 Validation outcomes for the PHENSIM repositioning approach. The test was done on all drugs tested by Stukalov et al. that are also present in the L1000 database at comparable concentrations. The L1000 drug signatures (DEGs caused by the drug effects) were gived as input to PHENSIM to built up our drug signatures. Then, Pearson correlations were performed between the PHENSIM drug- and viral- signatures obtained before.
Although the validation was satisfactory, it must be taken into account that the in vitro assays were performed on A549-ACE-2 cells whereas the data available on L1000 concern A549 cells. In addition, Pearson's correlation was performed using the viral model of A549 at both MOI 0.2 (see Appendix Figure S3) and MOI 2.0 instead Stukalov et al. infected cells with SARS-CoV-2 at MOI 3.0. In addition, a prediction was also made by performing the Pearson correlation against the in silico model of A549-ACE-2 at MOI 0.2 (see Appendix) and 2.0.
The drugs (and concentrations) tested by Stukalov et al. are listed at the top of the heatmap, the corresponding L1000 drugs (and the nearest concentrations to the Stukalovo et al. ones) are listed at the bottom side. The colours in the heatmap represent the correlation values. In red are showed the drugs that are positively correlated with SARS-CoV-2 and in blue the negative correlated ones. Stars denote the adjusted p-values, pv ≤ 0.05 * ; pv ≤ 0.01 **; pv ≤ 0.001 ***.

Although the validation was satisfactory, it is to be taken into account that there were important differences in terms of cell lines, multiplicity of infection and concentration of drugs used for in vitro versus in silico experiments.

Stukalov et al. perform the in vitro experiments on A549-ACE-2 cells while the data available on L1000 concern A549 cells. In addition, Pearson's correlation was performed using our viral model at both MOI 0.2 (see supplementary figure) and MOI 2.0 whereas Stukalov et al. infected the cells with SARS-CoV-2 at MOI 3.0 to perform the in vitro viral inhibitor assay.

Although the drug signatures available on L1000 were on A549 cells, we performed repositioning also on A549-ACE-2 at MOI 0.2 (see Appendix) and 2.0 by exploiting our in silico viral infection model. Thus, our prediction on A549-ACE-2 was made using our viral signature as obtained previously and the drug signature resulting from the L1000 data on A549.

Concerning the different concentrations, we have taken into account, for each drug, the ones closest to those tested by Stukalov et al. (see Fig. 30).

Figure 30 shows the correlation values between the different drugs' effects and the in silico infection models, compared with the in vitro results reported by *Stukalov et al.*

Our repositioning approach shows, according to *Stukalov et al*., that the B-RAF inhibitors Sorafenib, Regorafenib and Dabrafenib and the JAK1/2 inhibitor Baricitinib, which are commonly used to treat cancer and autoimmune diseases [7, 383, 384], led to a significant increase of virus infection.

Our results reveal a slight anticorrelation of Tirapazamine, an inducer of DNA damage at a concentration of 2.22 uM on A549-ACE-2 MOI 2.0 cells, reflecting the findings of *Stukalov et al.* however in their data the strongest effects appear to be at concentrations as high as 10 uM. In addition, our results also report potential effects for the mTOR inhibitor Ramapicins, and in particular we evaluated Sirolimus. In addition, according to *Stukalov et al.,* Marimastat, a potent inhibitor of matrix metalloproteinases (MMP) proteinases, seems to be a promising drug.

In addition, our results also reveal that Bosutinib, a small molecule BCR-ABL and src tyrosine kinase inhibitor used for the treatment of chronic myelogenous leukemia, and Doxorubicin, an antibiotic and antineoplastic that binds to cellular DNA and inhibits nucleic acid synthesis and mitosis (acting mainly in the S phase of the cell cycle) causing chromosome aberrations, should be also potentially good candidates.

4.1.5.3 Performance Evaluation: PHENSIM Genome-wide and Proteome network analysis

We compared our results with published in vitro experiments from Blanco-Melo *et al.,* Bojkova *et al*., and Draghici *et al*.[359, 364, 384, 385]. First, we compared the results from Blanco-Melo *et al*. with our *in silico* predictions for NHBE, Calu-3, A549 (MOI 0.2 and 2) and A549 transduced with ACE-2 (MOI 0.2 and 2) cells. Transcriptomics data for all cell lines were collected from the GEO *dataset* GSE147507 and Log2-LogFCs were computed. Next, LogFCs were compared with our predicted Activity Scores (AS) by accounting for their direction of perturbation. We compared our predictions with the genes that *Blanco-Melo et al*. reports as important in the antiviral host-response to SARS-CoV-2. Furthermore, using an unbiased approach and to verify the accuracy of PHENSIM, we assessed the top-10 upregulated and top-10 downregulated genes for each cell-line. Finally, we assessed our viral simulation with results from Draghici *et al.* 2020 and Catanzaro *et al.* 2020[268, 385].

Pathway Analysis

Pathway analysis was applied to the transcriptomics data to determine which biological processes were altered by the viral infection. We used 4 pathway analysis approaches to assess the most impacted pathways: *1)* MITHrIL, *2)* SPIA, *3)* Reactome Pathways and *4)* Gene Ontology Enrichment analysis. MITHrIL pathway analysis was performed as described in Alaimo *et al.,* 2016[268, 386]. We used the LogFC of DEGs for all cell lines to perform MITHrIL perturbation analysis on the KEGG meta-pathway. Therefore, all values were aggregated on a pathway basis to compute an Accumulator and a p-value. Finally, p-values were adjusted for multiple hypotheses using the Benjamini-Hochberg FDR correction. Results were filtered by an FDR-adjusted p-value of 0.05 and ranked using the Accumulator. The top-25 significant pathways were reported in Fig. 31D&E. SPIA analysis was performed as previously described by Tarca *et al*., 2009[136]; the LogFC of DEGs and ranked pathways were calculated using FDR-adjusted p-values as computed by SPIA. Pathways with a $p<0.05$ were considered significant.

Finally, to further expand our understanding of the biological processes affected by the infection, we performed enrichment analysis on both Reactome Pathways, using the ReactomePA package [65, 136], and Gene Ontology, using the GOfuncR package [387]. All results produced by the 4 pathway methodologies were collected and considered significant with an FDR-adjusted p-value<0.05 (Seey Fig. S3 and Supplementary material in [388]).

Statistical analysis

Statistical methods for transcriptomics and proteomics were applied as described by Blanco-Mello and Bojkova *et al.* respectively[359, 364] For transcriptomic data, raw counts were normalized and analyzed for differential expression using the DESeq2 pipeline as previously described [369] . All genes with an FDR-adjusted p-value<0.05 and absolute LogFC>0.6 were considered differentially expressed. In addition, we considered all genes showing an average read count<10 as non-expressed. All p-values computed by PHENSIM are corrected for multiple hypotheses testing, using the q-value algorithm [359]. For proteomic data, Normalized LC-MS/MS data were downloaded and significance was tested using unpaired two-sided Student's t-tests with equal variance assumed. All values were aggregated on a pathway basis to compute an Accumulator and a p-value, and p-values were adjusted for multiple hypotheses using the Benjamini-Hochberg FDR correction. Results were filtered by an FDR-adjusted p-value of 0.05 and ranked using the Accumulator.

4.1.5.4 PHENSIM model: from in vitro *to* in silico

Innovative approaches to rapidly elucidate a pathogens' mechanism of action have proven crucial for containing the global burden of communicable diseases. The PHENSIM approach, described here, is based on the definition of a newly introduced protocol for *in silico* simulation of novel emerging pathogens, such as SARS-CoV-2, and it aims at elucidating distinct host-responses and molecular mechanisms triggered by that particular pathogen, all while defining possible candidate drugs for indication repositioning.

For our strategy to be viable, even when only limited direct knowledge is available on the host-pathogen interaction, we need direct infection (*in vitro*) data that can be exploited to predict such interactions. To acquire this knowledge, we therefore employ transcriptomic and proteomic experiments of *in vitro* infected vs. normal, pathogen-free cell lines. When available, we leverage Differentially Expressed Genes (DEGs) as a means to simulate the direct and indirect effect of the virus on a host without a priori knowledge regarding the mechanism of infection. Using DEGs as input for our cell *PHEN*otype *SIM*ulator PHENSIM[1, 105] we define a signature of pathogen predicted effects on human pathways (*pathogen alterations profile*; here termed the "*viral signature*"; see Fig. 29). To build the viral signature, we use pathway endpoints; an endpoint is a biological element in a pathway whose alteration, based on current knowledge, affects the phenotype in a specific way[125].

By leveraging PHENSIM we aimed to determine the impact of such viral infection induced alterations on an array of human cell lines *in silico.* Simulation results are used to define a "viral signature" that can then be employed to identify candidate drugs. Once a cell-specific SARS-CoV-2 viral signature is defined, potential repositioning drugs can be identified by building a "*drug signature*" database queried by means of a similarity measure using pathway endpoints (Fig. 29). Given a candidate drug identified through a database (i.e. Drugbank or Pubchem) and literature (Pubmed) search, we define all known targets and alterations (up/down-regulations caused by the drug). Alterations are then provided as input to PHENSIM, together with the corresponding cell-specific *viral signature.* Next, distinct endpoint pathways[125] are identified and resulting *drug signatures* relating to a specific candidate can subsequently be compared with acquired *viral signatures* to evaluate the inhibitory potential of that candidate drug. Both viral and drug signatures are collected in a database, where a similarity search is performed using a Pearson correlation $\rho(x,y)$ since the propagation algorithm is linear in time complexity[125, 389]; see methods section *equation (1).* All drugs whose correlation with the virus is negative (green) are considered possible repositioning candidates, since they predict inhibition of the viral signature, whereas a positive correlation (red) suggests exacerbation of the viral signature when introducing the candidate drug.

4.1.5.4 Validation of PHENSIM transcriptomic strategy in SARS-CoV-2-infected host cells

To validate our PHENSIM model on a transcriptomic level in the context of SARS-CoV2, we sought to replicate the *in vitro* experiments using publicly available data presented by *Blanco-Melo et al*[125, 364, 389]. The in-depth transcriptomic analysis of SARS-CoV-2 elicited host-response by Blanco-Melo *et al.* recently revealed an inappropriate inflammatory response driven by reduced innate antiviral defenses, with low or delayed type I and type III interferon (IFN) and exaggerated inflammatory cytokine response, with elevated chemokines and IL-6[364].

As SARS-CoV-2 largely affects the lungs and respiratory tract, and because of its apparent affinity for lung tissue, the authors make use of several respiratory epithelial cell lines to assess the transcriptomic host-response. Here we use PHENSIM to reproduce transcriptomic effects in silico, as described *in vitro* for the following cell lines, namely undifferentiated normal human bronchial epithelial (NHBE) cells, cultured human airway epithelial cells (Calu-3) cells and A549 lung alveolar cells. The comparison of these results is depicted in Fig. 31. A549 cells are described to be relatively non-permissive to SARS-CoV-2 replication in comparison to Calu-3 cells, which is attributed to low expression of the viral entry receptor angiotensin-converting enzyme (ACE)2 [229, 364]. Thus, A549 cells were transduced with human ACE2 (A549-ACE2), which enabled SARS-CoV-2 replication at low-MOI (multiplicity of infection of 0.2). Furthermore, to induce significant IFN-I and -III expression, a high MOI of approximately 2-5 was necessary.

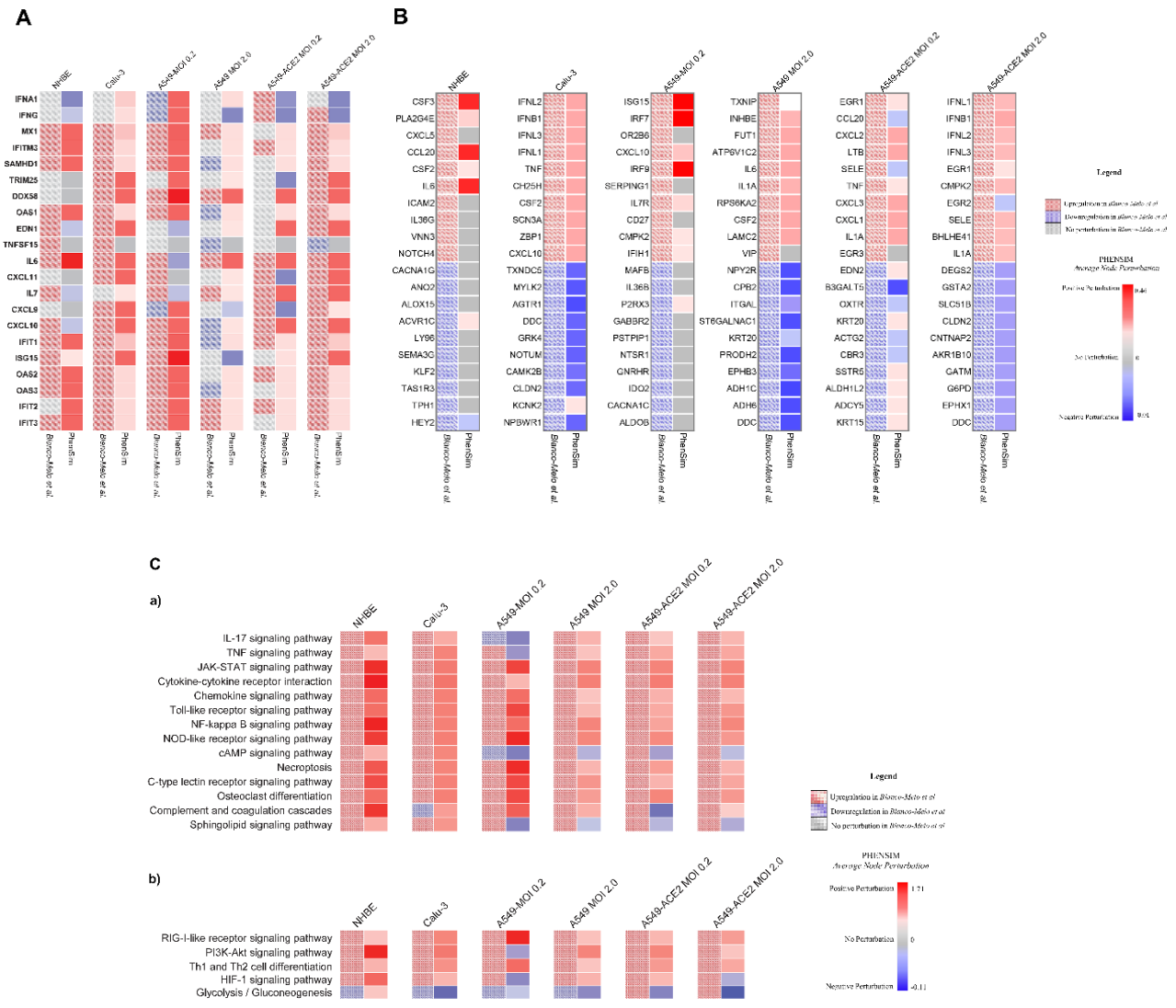
Here we leveraged the data published by *Blanco-Melo et al*. to run our PHENSIM simulation pipeline. In Fig. 31A we show representative genes, namely anti-viral, IFN stimulated genes (ISGs) and inflammatory cytokines and chemokines, considered important for the course of SARS-CoV-2 infection.

The heatmap shows perturbed expression, either up- or down-regulated, based on results obtained by *in vitro* (left column for each depicted cell-line) experiments for the different cells assessed in comparison to *in silico* PHENSIM predictions (right column; Fig. 31A). An unbiased approach of this predictive comparison is shown in Fig. 31B, displaying the top 10 up- and downregulated DEGs based on *in vitro* SARS-CoV-2 infection, as assessed in the different cells at low and high MOI (0.2 and 2) and with ACE2 addition in A549 lung alveolar cells. For each of the top *in vitro* acquired DEGs (left; checkered boxes), the PHENSIM predicted result is shown side-by-side (right). At first glance, PHENSIM reaches high predictive accuracy for Calu-3 human airway epithelial cells and A549-ACE2 and high MOI of 2, at least for the top DEGs (Fig. 31B). To quantify the overall predictive accuracy of PHENSIM, genome-wide transcriptomic data was assessed for all scenarios as described in Fig. 31. Overall accuracy of *in vitro* predicted transcriptomic results are shown in Table 1, ranging from 51.66-83.74% for A549-ACE2 MOI 0.2 - to NHBE cells. Sensitivity of perturbation prediction for nodes accurately predicted as perturbed, ranged from 95.83-100.00% sensitivity with 97.67-99.86% specificity for this in-depth SARS-CoV-2 transcriptomic analysis. Furthermore, the positive predictive value (PPV) and False negative rate (FNR) are shown for each tested scenario (see Table 4).

	Overall Accuracy	Nodes Predicted as perturbed			Nodes predicted as non-perturbed	
		PPV	Sensitivity	Specificity	PVV	FNR
A549-ACE2 MOI 0.2	51.66%	93.50%	96.72%	97.67%	58.90%	41.10%
A549-ACE2 MOI 2	71.72%	96.88%	99.24%	99.13%	60.04%	39.96%
A549 MOI 0.2	83.74%	68.75%	100.00%	99.86%	85.64%	14.36%
A549 MOI 2	78.20%	97.41%	97.20%	99.58%	77.47%	22.53%
Calu-3	77.17%	96.93%	99.34%	99.30%	76.55%	23.45%
NHBE	82.43%	67.65%	95.83%	99.69%	86.48%	13.52%

Table 4. PHENSIM transcriptomic predicted values from *Blanco-Melo et al.* 2020.

In order to further verify PHENSIM's robustness in whole genome pathway analysis, we next explored PHENSIM's ability to predict significantly affected signaling pathways in SARS-CoV-2 infection. In Fig. 31C we highlight PHENSIM's predicted perturbation of a select set of affected pathways during infection, as recently identified to be of importance by Catanzaro *et al.* 2020 and Draghici *et al.* 2020, such as IL-17, JAK-STAT and TNF signaling pathways, Toll-like Receptor (TLR), NOD-like receptor and RIG-I-like receptor signaling pathways as well as complement and coagulation cascades.



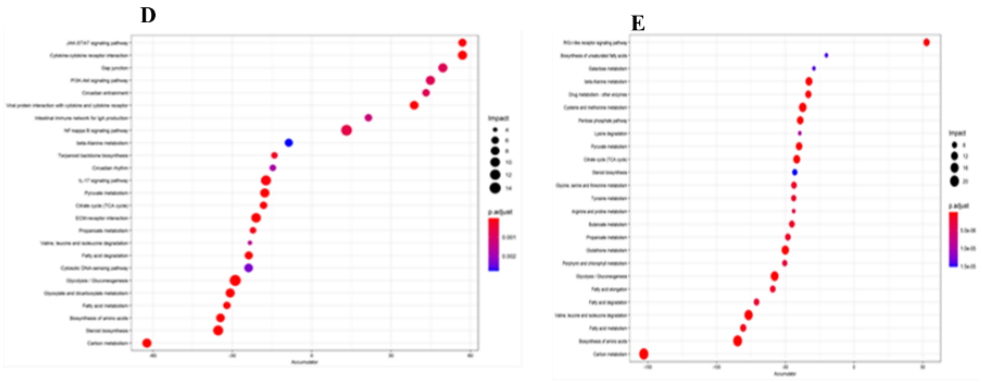


Figure 31. In silico PHENSIM prediction of host transcriptional response to SARS-Cov-2. In vitro results from Blanco-Melo et al. (left column; checkered boxes) are compared to in silico PHENSIM predictions (right; solid) for all evaluated respiratory related cells assessed; NHBE, Calu-3, A549 cells at low (0.2) and high (2.0) MOI, ± ACE2 transduction respectively. A) Heatmap depicting the perturbation of a select subset of anti-viral, ISGs and inflammatory genes. B) Heatmaps depicting unbiased analysis of the top-10 upregulated (red) and top-10 downregulated (blue) DEGs from Blanco-Melo et al. (left) with side-by-side PHENSIM predictions (right). For A&B, legend shows denoted perturbations for PHENSIM prediction and Blanco-Melo et al. See legend box for DEG annotation. C) Heatmap depicts whole genome pathway analysis as predicted by PHENSIM for a select set of signaling pathways of interest in all assessed cell types. Pathway selection was based on highlighted pathways affected by SARS-CoV-2 infection. Color gradient depicts the average pathway perturbation as predicted in our PHENSIM in silico experiments. D&E) MITHrIL pathway analysis was used to assess top meta-pathways for D) A549-ACE2 MOI 0.2 (low viral load) and E) A549-ACE2 MOI 2.0 (high viral load), according to impact (circle size) and significance (color-gradient for adjusted p-value) for the top 12 up- (+accumulator) and down-regulated pathways. The accumulator is the accumulation/sum of all perturbations computed for that particular pathway. NHBE; Normal Human Bronchial Epithelial cells, Calu-3; Cultured human airway epithelial cells, A549; Transformed lung alveolar cells, ACE2; angiotensin-converting enzyme, MOI; multiplicity of infection. DEGs; Differentially expressed genes, ISGs; IFN-stimulated genes.

For further verification of our PHENSIM pathway analysis prediction *in silico*, we compared our results with those obtained using our previously described MITHrIL (Mirna enrIched paTHway Impact anaLysis) tool[125, 229] to analyze the *Blanco-Melo et al*. acquired *in vitro* data (Fig. 31D-E). Given DEGs, MITHrIL first computes a perturbation for each gene in the meta-pathway (as described in Methods section). The perturbation can be considered as the predicted state that the node will have given the input DEGs. Next, we sum the perturbation of all nodes for each pathway to acquire the "accumulated perturbation," or the Accumulator. The accumulator is equivalent to a pathway expression and is a sum of all perturbations computed for that particular pathway. MITHrIL pathway analysis for A549-ACE2 at low viral load (MOI 0.2) revealed Chemokine, JAK-STAT, PI3K-Akt signaling and cytokine-cytokine interaction as a few of the top upregulated pathways, according to impact (circle size), significance (color-gradient for adjusted p-value) and accumulated perturbation computed for that particular pathway (accumulator).

For A549-ACE2 at high viral load (MOI 2.0; Fig. 31E), next to similar pathways at low viral MOI, Toll-like receptor (TLR) and NOD-like receptor signaling were among the top pathways observed, corresponding to the observation that high viral MOI was needed to induce significant type I IFN signaling[364]. Interestingly, both at low and high MOI various metabolic pathways were significantly affected with a negative accumulator. Overall, the MITHrIL analysis results show the most affected pathways to be similar to the PHENSIM *in silico* predicted results.

4.1.5.5 Modeling proteomics in SARS-CoV-2-infected host cells leveraging PHENSIM

Using combinatorial profiling of proteomics and translatomics to study host-infection on a cellular and molecular level gives opportunity to study relevant viral pathogenicity in the search of potential drug targets [359]. As SARS-CoV-2 has been detected in stool and can replicate in gastrointestinal cells [172, 193, 359], Bojkova *et al.* use the human colon epithelial carcinoma cell line *Caco-2* to study SARS-CoV-2 infection [359]. With their novel method, multiplexed enhanced protein dynamics (mePROD) proteomics, they determined SARS-CoV-2-specific translatome and proteome changes at high temporal resolution [128], and were able to quantify translational changes occurring during SARS-CoV-2 infection *in vitro* over the course of 24 hours at multiple timepoints (at 2, 4, 10 and 24h) [359].

PHENSIM proteomic validation

To validate PHENSIM on a proteomic level, we used our *in silico* approach to replicate the *in vitro* SARS-CoV-2 infection of human Caco-2 cells[125, 229, 359]. As viral genome copy number in cell culture supernatant and all viral protein levels assessed reached peak levels at 24h post infection, and the proteome underwent most extensive modulation[359], we focused on this particular time-point for more in-depth comparison of protein expression and functional pathway analysis (Fig. 32). The PHENSIM simulation results obtained by leveraging the proteomic data 24hrs post SARS-CoV-2 infection are shown in Fig. 32. We provide an unbiased assessment by comparing the PHENSIM obtained Average Node perturbation *in silico*, to the 30 most perturbed proteins according to Bojkova *et al*. In order to compare *in vitro* to *in silico* protein expression levels a representative selection of relevant proteins involved in infection is depicted in the heatmaps in Fig. 32B and C. In Fig. 32B, proteomic perturbation of the top differentially expressed proteins (DEPs; n=30) as predicted by PHENSIM(right, solid) is compared side-by-side to perturbation results from Bojkova *et al.* (left, checkered). Next, in Fig. 32C the top DEPs described by *Bojkova et al.* (right) are compared to PHENSIM predicted perturbation. Based on this selection of proteins we can denote a relatively high prediction rate for PHENSIM, although not all proteins are predicted to full accuracy. When quantifying the predictive power of PHENSIM on this protein-wide analysis, PHENSIM simulated results showed a predictive accuracy of 97.9% to the described *in vitro* proteomic data at 24hrs, where significant perturbation prediction was at 97.87% sensitivity and 97.96% specificity for this particular dataset (see Table 5).

Time (hours)	All Proteins	Proteins in Meta-pathway	Predicted Percentage	Accuracy	PPV	Sensitivity	Specificity
2	5809	1914	6.95%	93.98%	95.45%	92.65%	95.38%
6	5809	1914	11.70%	93.75%	94.35%	97.66%	81.13%
10	5809	1914	10.45%	94.50%	95.27%	98.17%	77.78%
24	5809	1914	34.95%	97.91%	98.39%	97.87%	97.96%

Table 5. PHENSIM proteomic predicted values from Bojkova et al. 2020.

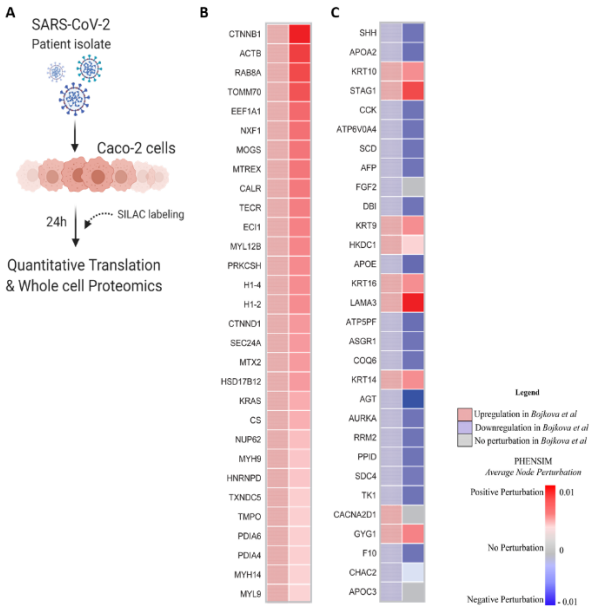
PHENSIM proteomics from in vitro to in silico

Next, to compare the Reactome-based *in vitro* functional pathway analysis to our PHENSIM *in silico* approach, a representative selection of significantly affected pathways correctly predicted by PHENSIM is depicted in Fig. 32D. Pathways were selected according to the cellular mechanisms highlighted by *Bojkova et al.*[359]. The centered heatmap shows an increasing activity score (top to bottom) as predicted by PHENSIM for each pathway. An in-depth analysis of proteomic pathways at 24hrs revealed distinct upregulation of various pathways involving cellular metabolism such as fatty acid degradation, glycolysis and gluconeogenesis, carbon metabolism, inflammatory and immune signaling pathways and also cellular senescence signaling pathways (Fig. 32D).

PHENSIM predicts a metabolic signature in SARS-CoV-2 infection in silico

As a metabolic signature was identified by PHENSIM's proteomic in silico simulation of SARS-CoV-2 infection (Fig. 32D), we next assessed the degree of intersection between the perturbed genes of these metabolic pathways in order to reject the hypothesis that a common set of altered proteins is driving the significant perturbation of these closely related metabolic pathways. All metabolic pathways considered essential for SARS-CoV-2 infection according to the acquired Bojkova et al. data (Fig. 32) were included in the analysis (FDR-adjusted p-value < 0.05) and a PHENSIM activity score was determined (see Table 5).

The affected general metabolic pathways showed very low degree of shared sub-pathway overlap. The Venn diagrams in Fig. 33 show all possible intersections for the following top metabolic pathways: (i) Fatty acid degradation, Amino sugar and nucleotide sugar metabolism, Glycolysis /Gluconeogenesis, Citrate cycle (TCA cycle), and Purine metabolism; (ii) Glycolysis /Gluconeogenesis, Citrate cycle (TCA cycle), Purine metabolism, Carbon metabolism, and Pyrimidine metabolism.



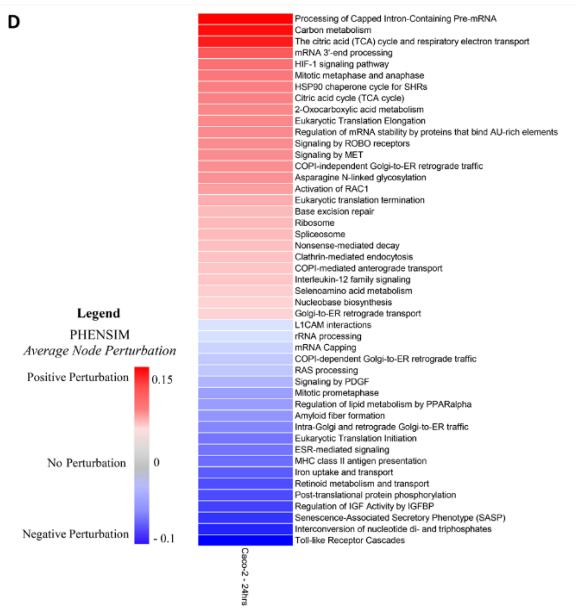


Figure 32. PHENSIM proteomic pathway analysis in SARS-CoV-2-infected human host cells. PHENSIM pathway analysis of the Caco-2 cell experiment was simulated in silico to reproduce in vitro results presented by Bojkova et al. at the 24hour time-point post SARS-CoV-2 infection A) Schematic representation depicting the experimental design as described by Bojkova et al. in vitro: the human colon epithelial carcinoma cell line, Caco-2 cells, were infected and monitored for 24hrs post SARS-CoV-2 infection. Naturally occurring heavy isotype SILAC labelling was used to quantify translational changes, as this method does not affect cellular behavior allowing for unbiased pathway analysis. Quantitative translation and whole cell proteomics by LC-MS/MS was performed 5. B&C) Heatmaps depicting a representative subset of the 30 top differentially expressed proteins (FDR<0.05) involved in viral infection after 24hr SARS-CoV-2 infection B) as predicted by PHENSIM in silico (right column, solid squares), compared to expression results as determined by Bojkova et al. (left column, checkered squares) and C) as described by Bojkova et al. (left column, checkered) with side-by-side PHENSIM expression prediction for that protein (right column, solid). D) Heatmap depicts PHENSIM simulated results in silico for the signaling pathways significantly affected at 24h post infection; Up- (red) and Down-regulated (blue). These signaling pathways were described as significant by Bojkova et al. in their analysis. Color gradient reflects PHENSIM activity; the value of the activity score attributed to each pathway from blue (downregulation) to red (maximum upregulation). Caco-2; the human colon epithelial carcinoma cell line, SILAC; Stable Isotype Labeling by Amino Acids in Cell culture, LC-MS/MS; Liquid chromatography mass spectrometry, DEPs; Differentially expressed proteins, Max; maximum.

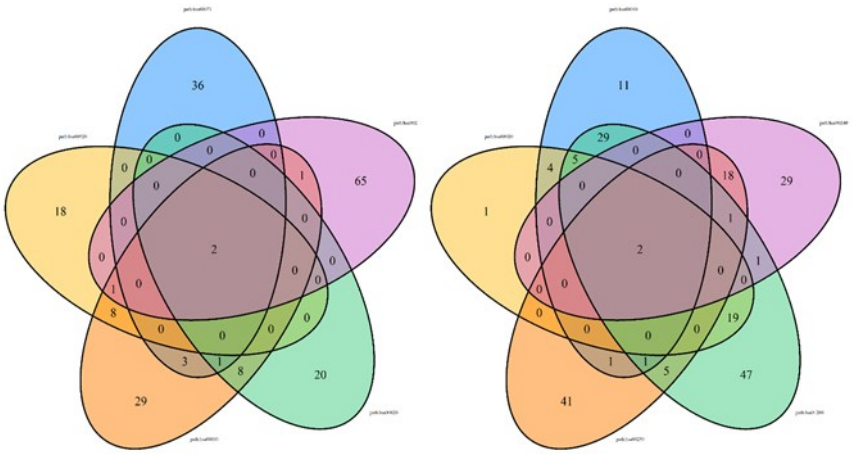


Figure 33. Venn diagrams of the perturbed genes of the significant metabolic pathways, related to Figure 32 and Table S2 in [386]. This figure illustrates the Venn diagrams obtained by calculating the intersections between the perturbed genes of the metabolic pathways identified in the PHENSIM simulation for the proteomics data illustrated in Figure 32. The analysis was conducted to exclude the hypothesis that a common core of altered enzymes was driving the significant perturbation of these closely related metabolic pathways. To construct the graphs, we took all the metabolic pathways considered essential for the infection in Bojkova et al. 2020 (Figure 32), which presented an FDR-adjusted p-value < 0.05 (Table S2). For each of these pathways, the perturbed genes were determined according to the PHENSIM activity score. Finally, all possible intersections were calculated. To better visualize the results, the eight pathways obtained with the previous criteria were divided into two groups of 5 pathways: (i) Fatty acid degradation, Amino sugar and nucleotide sugar metabolism, Glycolysis /Gluconeogenesis, Citrate cycle (TCA cycle), and Purine metabolism; (ii) Glycolysis /Gluconeogenesis, Citrate cycle (TCA cycle), Purine metabolism, Carbon metabolism, and Pyrimidine metabolism.

4.1.5.6 PHENSIM Drug repurposing strategy for COVID-19

The next step in our PHENSIM approach is the employment of our drug strategy in order to test candidate drugs for potential COVID-19 repurposing. This approach takes advantage of existing knowledge on drug-related pharmacology and toxicology for rapid therapeutic selection[362]. As schematically described in Fig. 29, once a cell-specific viral signature is defined, it can be exploited to search for possible repositioning candidates by leveraging our select drug signature database. We used a *Pearson correlation p*(x,y) to compare the viral and drug signatures, which gives rise to a correlation score specific to that candidate drug, computed for SARS-CoV-2 infection in a particular setting. Here we set out to test a selection of hypothesis- and data-driven candidate drugs as shown in Fig. 34. One such drug which regrettably failed to live up to its anticipated potential to effectively treat COVID-19 is the antimalarial drug hydroxychloroquine (HCQ), currently approved for rheumatologic implications, although associated with cardiac toxicity[362, 390, 391].

Although the efficacy of corticosteroids in viral acute respiratory distress syndrome (ARDS) remains controversial, recent evidence on drugs such as Dexamethasone and Methylprednisolone are showing promise in COVID-19[385, 392]. Furthermore, the potential beneficial effects of blocking the mTOR pathway with use of mTOR-inhibitors such as Metformin, Everolimus or Rapamycin (the later not evaluated here) in COVID-19 patients have been hypothesized, however its effects on gene expression and distinct signaling pathways remain to be satisfyingly established. In light of targeting cell immunometabolism, 2-Deoxy-Glucose (2DG) was recently proposed as a possible therapeutic in COVID-19[359, 385]. Lastly, therapeutic targeting of excessive host inflammation by inhibiting Bruton tyrosine kinase (BTK) – for example the BTK-inhibtor Acalabrutinib – in severe COVID-19 was recently described[393].

We evaluated a select set of candidate drugs for potential repurposing in SARS-CoV-2 infection as shown in Fig. 34 The drug candidates having a positive effect on ameliorating SARS-CoV-2 infection *in silico* have a negative correlation score (green) between viral and drug signature, whereas candidate drugs worsening the disease phenotype have a positive correlation (red). Indeed, for both low and high viral load (MOI), Methylprednisolone, Metformin, Dexamethasone and Acalabrutinib positively correlated with the viral signature (green) which points to an effective therapeutic to target SARS-CoV-2 infection in A549 cells in the presence of ACE2, however, the order of the candidate drugs differed somewhat between the two.

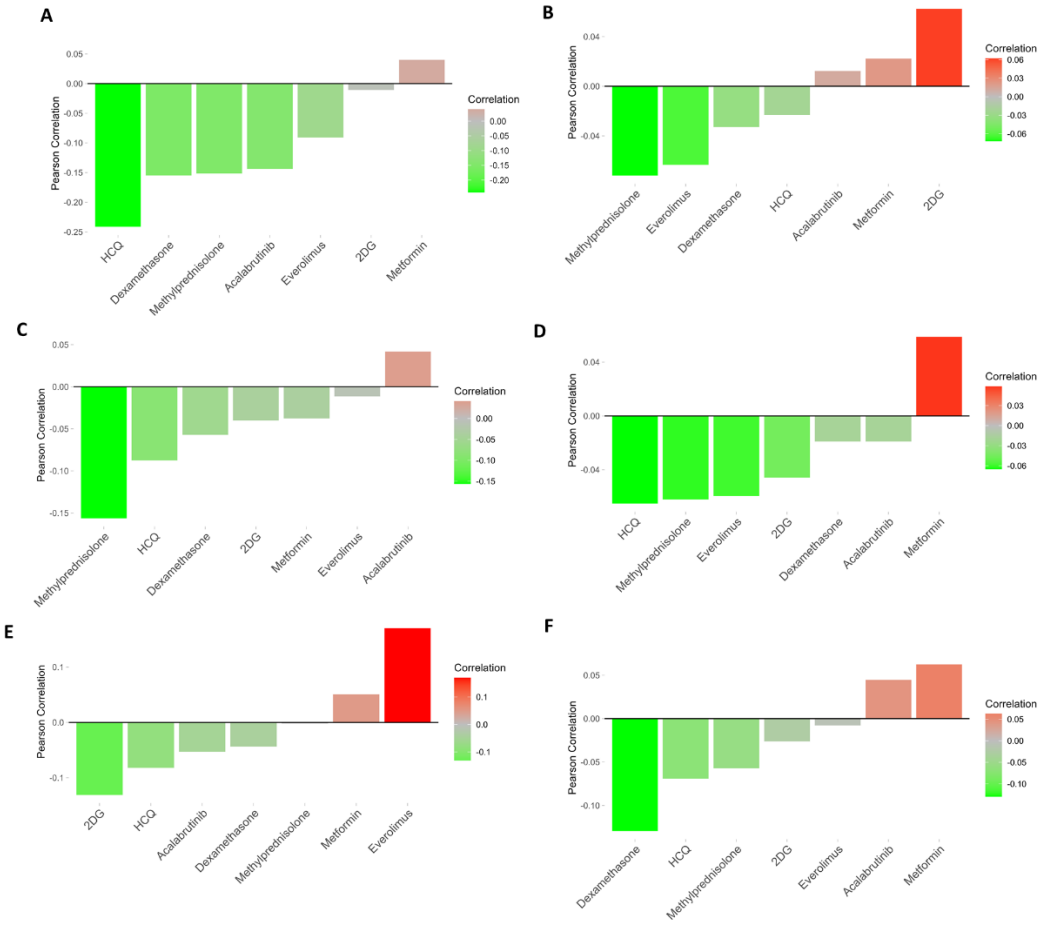


Figure 34. Drug repositioning candidates for COVID-19. We leverage our PHENSIM drug strategy approach to test candidate drugs for potential repurposing for COVID-19 treatment. Once a cell-specific viral signature is defined, it can be exploited to search for possible repositioning candidates by building a drug signature database. A Pearson correlation p(x,y) between the viral and drug signatures gives rise to a correlation score. Drug candidates having a positive effect on ameliorating SARS-CoV-2 infection have a negative correlation score (green) between viral and drug signature, whereas candidate drugs worsening disease correlate positively (red). Here we show distinct candidate drugs having a variable effect depending on cell types and on the multiplicity of infection (MOI) of virus infection. A) NHBE; B) Calu-3; C) A549 MOI 0.2; D) A549 MOI 2.0; E) A549-ACE-2 MOI 0.2; F) A549-ACE-2 MOI 2.0. This analysis shows the modeling viral load dynamics and discerning what candidate could work best in low vs higher viral load. Drug candidates represented here: Methylprednisolone, Metformin (mTOR-inhibitor), (Hydroxy)chloroquine (HCQ-CQ), Acalabrutinib (BTK-inhibitor), Dexamethasone, 2-Deoxy-Glucose (2DG) and Everolimus (mTOR-inhibitor). ACE2; angiotensin-converting enzyme, MOI; multiplicity of infection.

Using CoVariation analysis, we next looked at individual pathway contributions for each of the repositioning candidates evaluated here. The acquired Pearson correlation when comparing viral and drug-based signatures was dissected into components to show individual pathway contribution (see Fig. 35 and Appendix Fig. S6).

The overall effect of a candidate drug can be seen as the sum of the individually affected pathways, where anti-correlation is depicted in purple and positive correlation in orange. In Fig.35 we use Methylprednisolone as an example for A549 cells expressing ACE2 receptors at low (0.2, Fig. 35A) and high MOI (2.0, Fig. 35B). Only significantly affected pathways are depicted to illustrate the variation and effectiveness of the tested drug candidates (top; most pathways are anti-correlated shown in purple), to least likely candidate of interest (bottom; mostly positively correlated pathways in orange). Some top anti-correlated pathways for Methylprednisolone, highly contributing to the final result of this drug candidate based on our PHENSIM analysis include the JAK-STAT pathway, the Toll-like receptor

pathway, MAPK and PI3K-AKT signaling pathways. Next to similar pathways of importance affected for A549-ACE2 at both low and high viral MOI such as JAK-STAT, Toll-like receptor (TLR), NOD-like receptor, RIG-I-like receptor and MAP-kinase (MAPK) signaling, Focal adhesion and Neurotrophin signaling pathway were among the top pathways observed at high viral load (MOI 2.0; Fig. 35B). Pathway accumulation plots for the other drugs are shown in Appendix (iv) Fig. S5-S11.

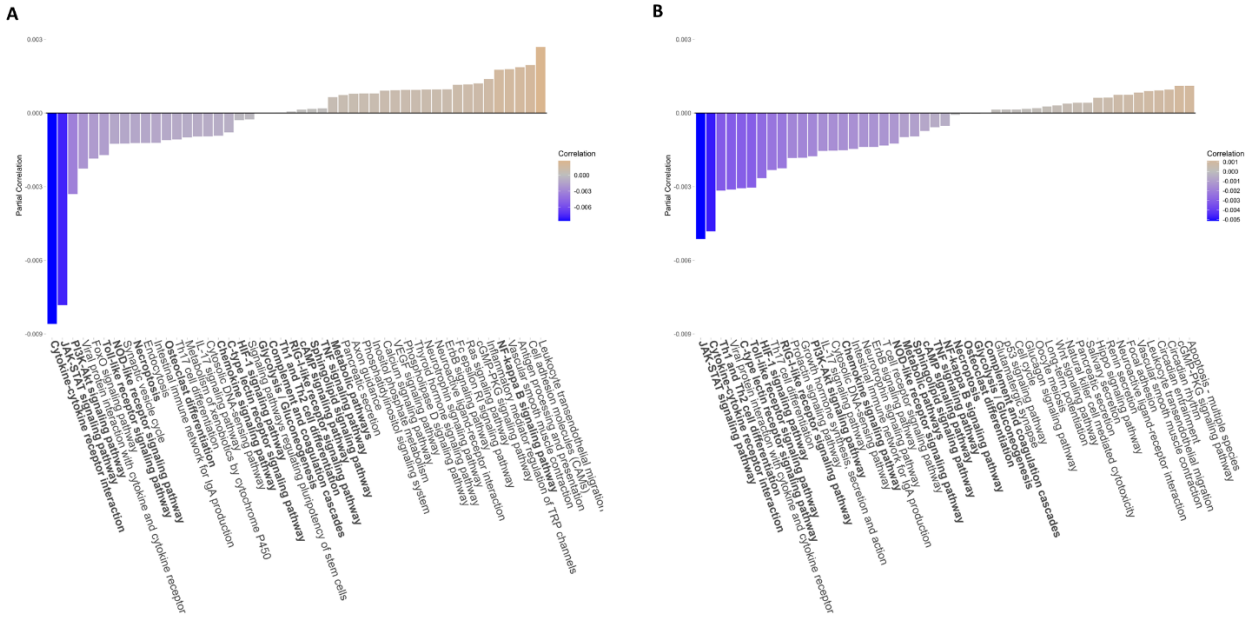


Figure 35. Resulted top pathways significantly affected by Methylprednisolone treatment in A549-ACE-2 cells.
A) A549-ACE-2 MOI 0.2 and B) MOI 2.0.

Methylprednisolone treatment of SARS-CoV-2 infected host cells in silico

As a next step in our drug repurposing efforts, we simulate the simultaneous host-cell infection of SARS-CoV-2 and *in silico* treatment with Methylprednisolone (MP), hereby combining the drug action and pathogen infection on a host-cell, in order to further assess MP as top candidate. We simulate SARS-CoV-2 viral infection and simultaneous MP treatment *in silico*, in order to more closely resemble the *in vivo* situation (Fig. 36). The heatmap in Fig. 5 depicts the results of transcriptomic pathways analysis of host-cell SARS-CoV-2 infection, based on Blanco melo *et al. in vitro* (Fig. 5; left column A) and PHENSIM simulation *in silico* (Fig. 36.; middle column B), compared to MP treatment of *in silico* SARS-CoV-2 infected host cells (Fig. 36; right column C). Here we visualize the pathways identified in Fig. 31C, and show the effects of MP treatment on these top affected pathways during SARS-CoV-2 infection in particular host-cells. All identified upregulated pathways during infection were significantly inhibited by MP treatment, showing it's known anti-inflammatory and immunosuppressive effects.

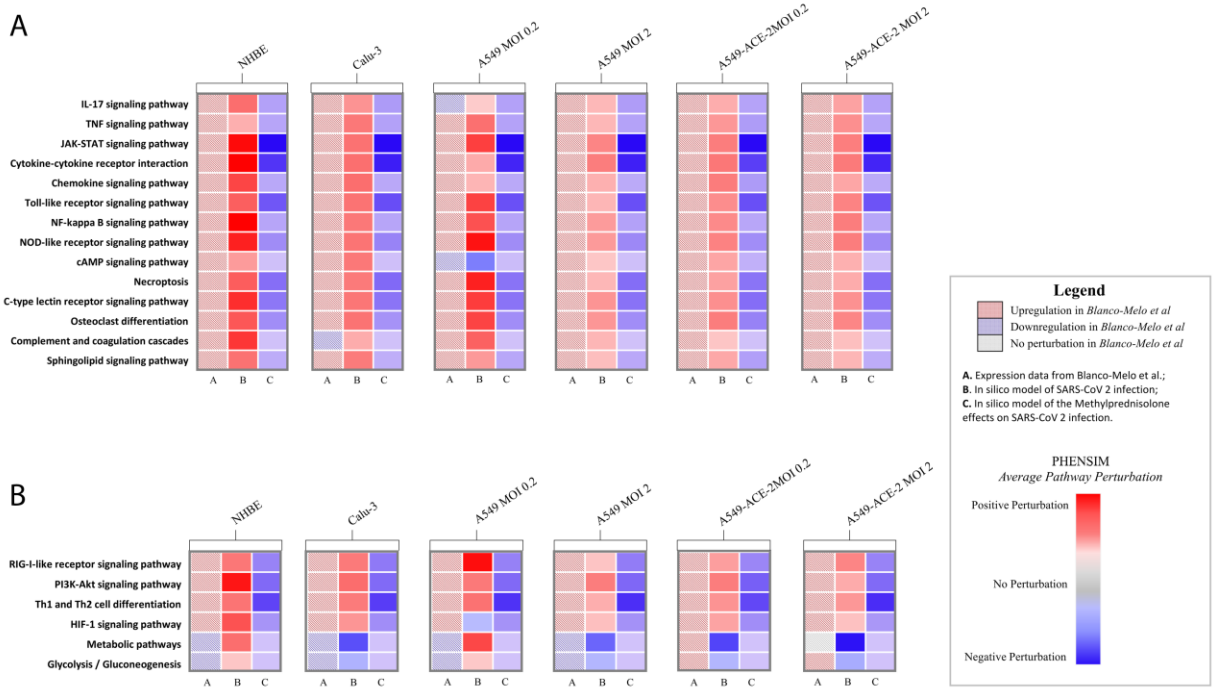


Figure 36. Methylprednisolone inhibits key inflammatory and viral signaling pathways in host lung and airway cells after SARS-CoV-2 infection. Heatmap depicts the effects of Methylprednisolone in silico in SARS-CoV-2 infection on select signaling pathways of interest (similar pathways to Fig. 2C). From left to right, **column A** shows pathway analysis results of SARS-CoV-2 infection in vitro as performed using the MITHrIL algorithm; **column B** shows PHENSIM results of SARS-CoV-2 infection in silico; **column C** shows PHENSIM simulation results of Methylprednisolone on SARS-CoV-2 infected cells in silico. Color gradient depicts the average pathway perturbation as predicted in our PHENSIM in silico experiments for column B&C. NHBE; Normal Human Bronchial Epithelial cells, Calu-3; Cultured human airway epithelial cells, A549; Transformed lung alveolar cells, ACE2; angiotensin-converting enzyme, MOI; multiplicity of infection.

4.2 NETME to extend biological networks: The case of CD147.

Previously, it was pointed out the problem of incompleteness of the main biological knowledge networks underlying many of the in silico models such as KEGG.

A concrete example as previously described, is CD147, also known as Basigin (BSG) or EMMPRIN, a transmembrane glycoprotein of the immunoglobulin superfamily, expressed in many tissues and cells, which is known to participate in several high biological and clinical relevance processes and is a crucial molecule in the pathogenesis of several human diseases[374, 394]. Moreover, as mentioned above, CD147 plays a fundamental role in SARS-CoV-2 infection since, together with Angiotensin-Converting Enzyme 2 (ACE2), it interacts with viral spike protein as secondary cellular entry point.

In this direction, CD147 is an example of how a missing crucial gene within a biological network can compromise scientists' efforts to understand certain molecular phenomena. In literature, there are many valuable tools [214, 395] to integrate the missing information into bio-databases, such as KEGG. However, the most reliable approach in terms of accuracy and updated information remains the manual curation of such networks through careful and time-consuming literature analysis. On the other hand, a manually constructed network provides partial information due to the limited number of articles that a scientist could read. Among the bibliography consulted to build the network manually, we have carefully selected 11 papers containing a significant amount of helpful information for our purpose. On the other hand, we also assessed the capabilities of NETME in inferring CD147-diseases relations. For this purpose we selected 100 random interactions from DisGenNET [209], as well as the same abstracts used by DisGenNET for inferring such interactions.

This NETME application aims to be a practical example of how this system can create valuable networks by analyzing quickly and automatically larger sets of publications.

The set of 11 selected papers, described in Figure 37a, was analyzed by a bio-expert to derive a CD147-genes interactions network manually. This process resulted in 50 genes and 64 interactions, as shown in Figure 37a. Next, by using the same set of papers, we run NETME with no upstream filter. The automatically generated network consisted of 86 genes and 139 relationships between them (see Figure 37a-b). As the manually curated network consists of genes and proteins, only elements from these two categories were selected for the evaluation. This was performed by considering edges with the lowest "bio" score for each node pair. Qualitatively, this network includes most of the interconnections mentioned in the papers, thus providing a reliable and comprehensive overview of the molecular function of Basigin.

Quantitatively, NETME achieved an accuracy of 98.99%, a sensitivity of 100%, a specificity of 98.98%, and a positive predicted value of 46.32%. Figure 38a-b-c depicts the precision/recall curve (AUC 0.997), the sensitivity/specificity curve and the True positive rate/False Positive Rate one. The construction of the curves considered all possible gene-pairs and their edges.

Finally, we queried NETME with the selected 100 random CD147-diseases interactions in DisGenNET, selecting the same PubMed abstract used by DisGenNET for inferring those interactions. NETME detected 63 True Positive values out of 100, revealing a sensitivity of 63%. It is essential to stress that NETME allows us to extract a satisfactory and valid amount of information in a few minutes, compared to a manual search that may take days or weeks. We also believe that this case study is significant because, in the evaluation, we considered not only the presence of a link between two nodes but even

more closely the type of edge, hence the adequacy and specificity of the annotated edge in its biological context.

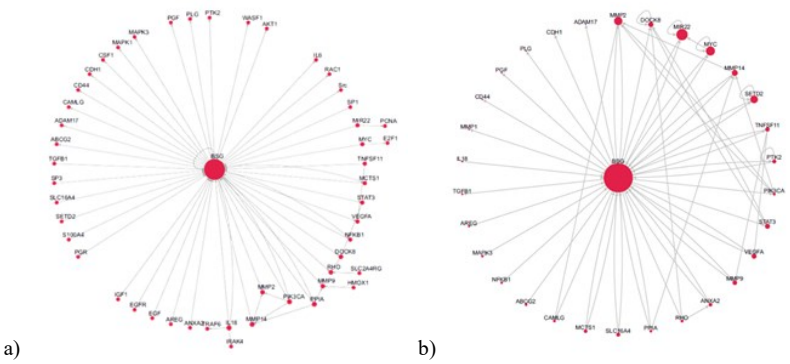


Figure 37 CD147 Network reconstruction using NETME. a) depicts the pathway constructed by hand from the selected papers [373-382], with CD147(BSG) as the central node. b) Shows the molecular mechanisms summarized in the knowledge network developed by NETME in accordance with the same papers used in a) NETME shows that CD147 is a potent inducer of metalloproteinases (MMPs) such as MMP2, MMP14 and MMP9 as reported in [374, 378, 379]. Furthermore, the overexpression of CD147, which results in increased phosphorylation of PI3K(PIK3CA), Akt(AKT1), leads to the secretion of vascular endothelial growth factor (VEGFA) in several biological contexts such as KSHV infection [374, 378]. In addition to its ability to induce MMPs, CD147 regulates spermatogenesis, lymphocyte reactivity and MCT system, in particular MCT1 and MCT4 (MCTS1 and SLC16A4) expression [374, 382]. Our results also show that CD147 can increase the expression of ATP-binding cassette transporter G2 (ABCG2) protein, regulating its function as a drug transporter, as mentioned by Xiong et al. for MCF-7 cells[374]. NETME identifies also BSG as an upstream activator of STAT3, highlighting its involvement in tumor development in agreement with the literature[381]. As summarized by our knowledge network, CD147 is regulated by various inflammatory mediators, such as RANKL (TNFSF11), denoting its involvement in inflammatory processes [377, 378]. Among the potential activators of BSG, NETME also find the transcription factor c-Myc (MYC) [375].

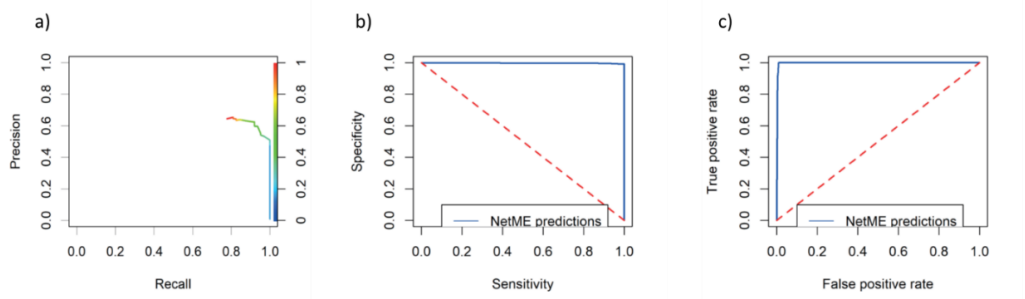


Figure 38. Metrics of BSG-network performed by NETME. The plots show a) Precision/Recall curve; b) Sensitivity/Specificity; c) True positive rate/False Positive Rate. The red dashed line in b) and c), indicates the expected result if the used method was random, that is any method which, given a pair of nodes, elects whether between them there is a link with a probability of 0.5.

4.3 The Value of Sharing: Scientific Wiki

The COVID-19 pandemic has reinforced to the entire community the value of collaboration. The scientific world has experienced the importance of sharing protocols, results, data, and codes succeeding in obtaining impressive scientific results in a very short time. This project is one of the concrete proofs that collaboration, at every level, always helps to build and build better.

Multi-disciplinary open science has emerged as a powerful mechanism to accelerate science and fight the rapidly evolving worldwide COVID-19 pandemic.

The current pandemic has raised the need for efficient and effective identification of potential drug candidates, creating an urgency for spreading knowledge and innovation.

This prompted us to create SciKi, short for *Scientific wiKi*, a toolbox developed primarily to integrate and disseminate results obtained using the PHENSIM(Phenotype Simulator) based drug-discovery framework. Indeed, disseminating the results through a collaborative environment allows the verification of hypotheses, by detecting contradictions, validating sources, and filtering fake data. Therefore such a tool is critically needed.

Sciki is a toolbox primarily developed to interpret and disseminate the results obtained by using our drug-discovery framework based on PHENSIM(Phenotype Simulator).

Designed to interact with open science communities innovatively by helping researchers search for candidate drugs based on publications, wikis, leaderboards, and comments and machine-generated "interpretations" for successful (e.g. thresholded by statistical significance) candidates.

This new platform, which is not yet in its final form, is designed to interact with open scientific communities in an innovative way and to publicly disseminate reproducible and explainable scientific results in a simple and effective way.

https://sciki.eu/

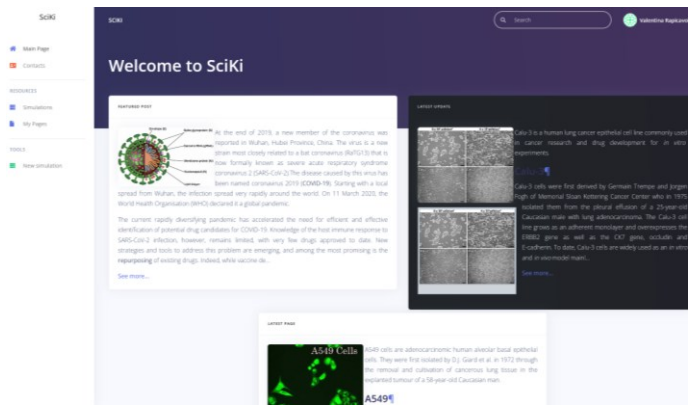


Figure 39 Screen shot of SciKi's main page.

5

Conclusions

Technological advancement in biomedicine allowed the production of a large amount of omics and clinical data, increasing the necessity to develop new techniques to analyze them.
Providing easy-to-use tools to support clinical and biological research in the analysis of complex diseases or, in general, complex biological phenomena is becoming a critical issue and a great challenge. Systems biology is playing a central role in this direction. Abstract biological modeling can be strategic for instance, in helping physicians to make better clinical decisions such as prescribing new therapeutic schemes based on specific disease-related gene and/or pathway deregulations. This framework allows patient-centered analyses and predictions to evaluate both the patient's disease status and the efficacy of treatments moving towards personalized medicine.
The difficulties, costs and time required in the discovery and development of new pharmaceutical molecules taken together with the availability of a huge number of drugs that passed all the critical phases of clinical trials, and for which most of the side effects are already known, increase the interest in repositioning existing drugs for new therapeutic purposes.

Various databases and toolboxes are available to systematically produce exciting predictions for drug repositioning and identify thousands of drug-disease pairs by computational studies [98,146,147]. Although advanced computational tools for drug repositioning exist, they are often difficult to understand or use, limiting their accessibility to scientists without a strong computational background [68].

The purpose of this thesis is to develop a new methodology for drug repositioning, using a typical systems biology approach, differentiating from most existing methods for being explainable.

In silico simulations, based on experimental data and literature information, allow us to explore the obtained results at multiple levels of detail (overall phenotypic profile, pathways, genes, metabolites) both from the disease-specific and drug-specific point of view.

This thesis addresses a multi-hypothesis-driven systems biology approach to improve drug repositioning across multiple contexts.

This dissertation was motivated by the current SARS-Cov-2 pandemic, which has hyper-accelerated the need for efficient and effective identification of potential drug candidates.

We find that PHENSIM performs at a high overall accuracy with high PPV, sensitivity and specificity for all airway and lung-related cell lines evaluated in our applications. PHENSIM predictive performance was further validated using our previously described MITHRIL transcriptomic pathway analysis[125, 393], showing similar results. Interestingly, key signaling pathways proposed to be crucial in SARS-CoV2 infection[268], were shown to be significantly perturbed in all cell lines studied *in silico* using PHENSIM, simulation offering promising potential molecular drug targets.

The PHENSIM strategy was also suitable for a proteomics/translatome-data based approach. PHENSIM simulation was compared to published SARS-CoV2 infection specific proteomic effects in host cell lines[268, 359]. Comparing simulation results to proteomic data 24 hours post infection showed, for the proteins available in KEGG, high accuracy with a PPV, sensitivity and specificity well above 97%. Inhibition of several of these protein-associated pathways was shown to prevent viral replication in human cells[268].

We next used the transcriptomic-based PHENSIM approach to compare the viral signatures, computed with respect to model cell lines, to *in-silico*-derived drug signatures. Our overall correlation results show several potential drug repurposing candidates negatively correlating with SARS-CoV-2, varying from corticosteroids such as MP (already approved for treatment of COVID-19 patients) to biologics such as BTK-inhibitors that are currently being studied in clinical trials[268, 393] to metformin[396]. Individual signaling pathway contribution to the observed correlation score could be further delineated for each individual drug, providing specific targets for in depth analysis and potential for pathway-specific therapeutic targeting. As expected, the individual pathways most targeted by the *in silico* drug interventions (Fig. 36A&B) were similar to pathways found most perturbed by PHENSIM during host transcriptomic response to SARS-CoV-2 viral infection (Fig. 31C&D), emphasizing their potential therapeutic effects. HCQ, although hypothesized to be a good potential candidate to treat COVID-19, has not proven effective *in vivo*[396, 397]. The exact reason why HCQ has failed in COVID-19 remains to be fully understood. Interestingly, COVID-19 is associated with a variety of hematologic complications[398], and increased HCQ use during the COVID-19 pandemic has induced the emergence of methemoglobinemia, including tissue hypoxia and reduced oxygenation[398-400]. Evidently, evaluating the risk-benefit ratios – drug safety and efficacy – is crucial when selecting drugs to be repurposed for COVID-19[362, 400], which particularly holds true for HCQ[391, 401, 402].

In Fig. 34, we depict our drug repurposing PHENSIM approach that functions as a selection tool for initial drug candidate screening, based on the anti-correlation of viral and drug signatures. In this ranking system a negative correlation constitutes higher potential for that particular drug. This broader correlation approach can be used to screen large sets of candidate drugs. Next, as depicted in Fig. 36, a more dynamic and extensive analysis can be performed, in order to compare simulations of SARS-CoV-2 host-cell infection (column B) and *in silico* treatment with a candidate drug such as Methylprednisolone (column C). Although Methylprednisolone is a known broad-spectrum corticosteroid, with clear anti-inflammatory and immunosuppressive effects (as shown in Fig.35), complete inhibition of these crucial immune signaling pathways might not be beneficial to COVID-19 patients at every stage of disease as described in clinical practice. Other, more targeted drug candidates might be more beneficial to the overall functioning of the patient's immune system during the fight and recovery from COVID-19. Indeed, our detailed approach can be implemented for all other top candidates, for further in-depth evaluation of their potential. However, we should bear in mind that the simulation is simultaneous (both virus and drug) and not completely reflective of a sequential treatment of a drug during infection. We are currently leveraging our simultaneous approach to evaluate the use of Metformin in COVID-19 in more detail.

Drug repurposing towards COVID-19 is challenging, but also opens many new opportunities. Several innovative approaches have been used varying from structure assisted computer designed mini inhibitors of receptor binding domain (RBD) [402–404], inhibitors of viral key enzymes like Mpro[360, 405, 406], machine learning models predicting compound protein inhibiting activity[406, 407] to infected cell-based assays drug screening[406–409]. Using computational tools, such as PHENSIM, allows for safe exploration of potential candidate drugs and uses previously acquired knowledge from biomedical databases to narrow the scope of possible viable biomarkers and druggable targets. One of the clear advantages of PHENSIM is a more effective selection of hypothesis driven drugs, before initiating extensive, time-consuming and costly *in vitro* experiments that should eventually provide the basis for clinical studies. PHENSIM requires on average (depending on data availability) about 3 hours of for each simulation. Another interesting possibility enabled by our approach is the potential capability to not only simulate the effect of a single drug, but also drug combinations. This expansion of PHENSIM is currently being developed (see Methods section; *data availability).* By making use of not just viral targets but also host proteins and structured pathways in the computation of the PHENSIM viral signature, we broaden the scope of potential drug targets with the added advantage that these are less prone to resistance development[410]. Here we simulated a select set of candidate drugs for repurposing in COVID-19, however, there are many candidates with high potential that can be added to this list, and further evaluated by our PHENSIM system *in silico* in the near future. We can also learn from and identify additional candidates based on the results obtained in this study. Starting from SARS-CoV-2 viral signature acquired by PHENSIM and recent data on IFN-involvement in COVID-19 [411], targeting the JAK/STAT pathway using Baracitinib – approved for moderate to severe arthritis[411, 412] – was recently shown to reduce time-to-recovery for hospitalized COVID-19 patients in combination with Remdesivir [411–413], however, caution is warranted[414].

An important featrure of PHENSIM is its extensibility with additional information on crucial genes, absent from KEGG, when specific knowledge becomes available. The in silico model presented here provides an interesting framework that could be further developed and expanded, achieving a more complete cell signature by new available data on processes. These may include cell-cell communication through ligand-receptor complexes[414, 415] or viral immune evasion e.g.[414–416]. For example, in the case of SARS-CoV2 infection, the absence of some important genes in the KEGG , was considered a severe limitation. More specifically Basigin (BSG), also known as CD147, was added to KEGG, in order to investigate the role of extracellular matrix metalloproteinase inducer (EMMPRIN) in COVID-19.

Although the extension of the model has been manually executed, we have shown that, using our new text-processing NETME system, it is possible to build knowledge networks in an easy, fast and automatic way, obtaining results comparable to manual bibliographic research. However in-silico literature mining allows to consult many more papers in a much shorter time. Since models based on biological networks such as KEGG or Reactome are incomplete, providing a way to solve the problem in a simple, fast and reliable way becomes extremely important.

As most *in vitro* studies are performed on cell lines, tissue tropism characteristics of viral infection seem key to better understanding viral activity[417, 418]. The same model could be adapted to study specific cells involved in viral infection like tissue-specific epithelial cells and immune cells (e.g. T cells and NK cells)[417, 419]. Moreover, many interesting avenues can potentially be explored using PHENSIM, such as modeling immune-related effects of this pathogen and others, in distinct tissue-specific non-immune epithelial cells, stem cells, and beyond[293, 365, 417, 419]. The system can be further adapted to include new data gathered on the viral translational landscape related to newly discovered open reading frames (ORFs) and potential novel polypeptides/proteins and infectivity potentiating cell surface

structures like neuropilin [246, 420]. Interestingly, integration of all the aforementioned schemes could potentially yield novel and effective drug targets[421].

As demonstrated by our results, we believe that the PHENSIM system provides a multitude of powerful systems biology functions and implements them easily and efficiently. PHENSIM is a simulation algorithm which follows the biological processes modeled by pathways. Therefore, PHENSIM is able to make a prediction of such processes and not only of the final effect, going beyond methods based on pathway enrichment. Furthermore, since pharmacological treatments may depend on the state of biological processes, PHENSIM may be of more appropriate use in this context. Comparison with other simulation algorithms such as BIONSI[151, 152] has shown excellent performance by PHENSIM[1]. PHENSIM creates and builds on interpretable and intervenable mechanistic bio-chemical models, rather than combinatorial and statistical "black-box" models for joint stationary distribution of biological data, as in, say protein-protein interaction (PPI) networks, Graphical or Deep-net models.

PHENSIM gives rise to feasible validation and comparison of *in vitro* and *in vivo* experimental data[268, 359], gives insight into drug efficacy[268, 359, 385, 410], tracks specific host signal transduction pathways[268], *in silico* testing of single drugs and drug combinations and further delineation of future targets (e.g. CD147) and identification of specific pathways of action of both pathogen and therapeutic compound in healthy and infected systems. For cost efficiency, validated predictive methods and assays for early elimination of potential drug candidates are of great value[370, 410]. The overall efficiency (time, costs, safety) prompts to suggest implementing PHENSIM not only in viral acute pandemic settings[394], but in additional curative and non-curative diseases, especially complex chronic disorders, where clinical trials are time-consuming or impossible to reduce to practice. Optimally leveraging the power of pathway analysis by simulating host cell and tissue-specific infection and performing *in silico* drug selection, has a tremendous potential beyond COVID-19, with applicability to high global burden communicable diseases, translatable to pathogens of viral, bacterial and fungal origin, and potentially chronic disease such as inflammaging and diabetes. In conclusion, our PHENSIM approach will enable more rapidly initiated clinical trials and accelerated regulatory review of already pre-selected drugs with a high repurposable potential. However, there are critical considerations for the clinical use of repurposed drugs related to drug combinations, alternative doses, and routes of administration that need to be systematically explored. It becomes critical in this perspective, to develop new algorithms, methods and tools that enable quantitative analysis of omics data using the framework of systems biology. In this perspective we propose to extend PHENSIM by integrating new features that allow to simulate the timing of drug administration, in fact at the moment we can only perform simulations on the effects of simultaneous administration of drugs, and also add new features that allow to perform quantitative investigations (e.g. dosages). The addition of the temporal factor also can be very useful *in silico* modeling of infections that occur over time.

6

Future and outlook

The flexibility and transposability of our method make it extremely interesting for several applications. Among these, we are currently carrying out a study on the comparison of the effects of different viral infections including HRV, HPIV3, HIAV and RSV. Through this study we would like to investigate the mechanisms of action of the various viruses and explore the potential effects of overlapping infections. Indeed, it is known that viruses, in their interaction with the host, implement very different strategies to make a successful invasion. Knowing the differences and similarities might open new avenues and perspectives in the research for new potential pharmacological targets together with the acquisition of new knowledge on some previously underestimated biological processes.

Systems biology and other bioinformatics techniques are cross disciplinary sciences which have shown to be very useful in enhancing research in cancer and other complex diseases.
In this direction new research plans and collaborations are being developed. In particular, a project concerning important liver diseases such as Nonalcoholic steatohepatitis (NASH) and Primary sclerosing cholangitis (PSC), is being carried out by biologists and bioinformaticians in collaboration with clinicians.
NASH is a progressive liver disease involving hepatocyte injury, accumulation of lipid droplets in the liver, inflammation and fibrosis.
PSC is a progressive cholangiopathy involving bile duct destruction and strictures, concentric periductal fibrosis and periportal inflammation. PSC patients are forced, in 50% of cases, to resort to transplantation within 10-15 years from diagnosis.
Both diseases lead to cirrhosis, malignancies and end-stage liver disease.
The aim of the project is to outlight related molecular mechanisms, supporting clinical and biological research in order to formulate new hypotheses and their possible applications.

7

Bibliography

1. Alaimo S, Rapicavoli RV, Marceca GP, La Ferlita A, Serebrennikova OB, Tsichlis PN, Mishra B, Pulvirenti A, Ferro A (2021) PHENSIM: Phenotype Simulator. PLoS Comput Biol 17:e1009069

2. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30

3. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42:D199–205

4. Croft D, O'Kelly G, Wu G, et al (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 39:D691–7

5. Joshi-Tope G, Gillespie M, Vastrik I, et al (2005) Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 33:D428–32

6. Croft D, Mundo AF, Haw R, et al (2014) The Reactome pathway knowledgebase. Nucleic Acids Res 42:D472–7

7. Stukalov A, Girault V, Grass V, et al (2021) Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV. Nature 594:246–252

8. Subramanian A, Narayan R, Corsello SM, et al (2017) A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell 171:1437–1452.e17

9. Bischof E, Broek JAC, Cantor CR, et al (2020) ANERGY TO SYNERGY-THE ENERGY FUELING THE RXCOVEA FRAMEWORK. Int J Multiscale Comput Eng 18:329–333

10. Muscolino A, Di Maria A, Alaimo S, Borzì S, Ferragina P, Ferro A, Pulvirenti A (2021) NETME: On-the-Fly Knowledge Network Construction from Biomedical Literature. Complex Networks & Their Applications IX 386–397

11. Kadakia KT, Beckman AL, Ross JS, Krumholz HM (2021) Leveraging Open Science to Accelerate Research. N Engl J Med 384:e61

12. cell. https://www.britannica.com/science/cell-biology. Accessed 24 Nov 2021

13. Watson JD (1953) Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid.

14. Alberts B (2017) Molecular Biology of the Cell. Garland Science

15. Lumen Learning Storing Genetic Information. https://courses.lumenlearning.com/wmopen-nmbiology1/chapter/storing-genetic-information/. Accessed 24 Nov 2021

16. 9.1 The Structure of DNA. https://openstax.org/books/concepts-biology/pages/9-1-the-structure-of-dna. Accessed 17 Dec 2021

17. File:DNA Structure+Key+Labelled.pn NoBB.png - Wikipedia. https://en.wikipedia.org/wiki/File:DNA_Structure%2BKey%2BLabelled.pn_NoBB.png. Accessed 17 Dec 2021

18. Blackstone NW (2001) Molecular Cell Biology. Harvey Lodish , Arnold Berk , S. Lawrence Zipursky , Paul Matsudaira , David Baltimore , James Darnell. The Quarterly Review of Biology 76:76–76

19. https://upload.wikimedia.org/wikipedia/commons/5/54/Gene_structure_eukaryote_2_annotated.svg. Accessed 17 Dec 2021

20. (2011) NCI Dictionary of Cancer Terms. https://www.cancer.gov/publications/dictionaries/cancer-terms. Accessed 24 Nov 2021

21. Website. Ambros, V. The functions of animal microRNAs. Nature 431, 350–355 (2004). https://doi.org/10.1038/nature02871.

22. Ambros V (2004) The functions of animal microRNAs. Nature 431:350–355

23. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116:281–297

24. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. Cell 136:215–233

25. Fabian MR, Sonenberg N, Filipowicz W (2010) Regulation of mRNA translation and stability by microRNAs. Annu Rev Biochem 79:351–379

26. Han J, Lee Y, Yeom K-H, Kim Y-K, Jin H, Kim VN (2004) The Drosha-DGCR8 complex in primary microRNA processing. Genes Dev 18:3016–3027

27. Murchison EP, Hannon GJ (2004) miRNAs on the move: miRNA biogenesis and the RNAi machinery. Curr Opin Cell Biol 16:223–229

28. Rana TM (2007) Illuminating the silence: understanding the structure and function of small RNAs. Nat Rev Mol Cell Biol 8:23–36

29. Li W, Lebrun DG, Li M (2011) The expression and functions of microRNAs in pancreatic adenocarcinoma and hepatocellular carcinoma. Chin J Cancer 30:540–550

30. Wu R (1972) Nucleotide Sequence Analysis of DNA. Nature New Biology 236:198–200

31. Padmanabhan R, Jay E, Wu R (1974) Chemical Synthesis of a Primer and Its Use in the Sequence Analysis of the Lysozyme Gene of Bacteriophage T4. Proceedings of the National Academy of Sciences 71:2510–2514

32. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74:5463–5467

33. Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology 94:441–448

34. Website. Wikimedia. The Sanger (chain-termination) method for dna sequencing, 2012. URL https://upload.wikimedia.org/wikipedia/commons/b/b2/Sanger-sequencing.svg.

35. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM (2009) Transcriptome sequencing to detect gene fusions in cancer. Nature 458:97–101

36. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nature Protocols 7:1534–1550

37. Lee JH, Daugharthy ER, Scheiman J, et al (2014) Highly multiplexed subcellular RNA sequencing in situ. Science 343:1360–1363

38. Godovac-Zimmermann J (2008) 8th Siena meeting. From genome to proteome: integration and proteome completion. Expert Rev Proteomics 5:769–773

39. Momeni Z, Hassanzadeh E, Saniee Abadeh M, Bellazzi R (2020) A survey on single and multi omics data mining methods in cancer data classification. J Biomed Inform 107:103466

40. Rhodes DR, Chinnaiyan AM (2005) Integrative analysis of the cancer transcriptome. Nat Genet 37 Suppl:S31–7

41. Casadevall A, Pirofski L-A (2015) What is a host? Incorporating the microbiota into the damage-response framework. Infect Immun 83:2–7

42. Alizon S, Hurford A, Mideo N, Van Baalen M (2009) Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future. J Evol Biol 22:245–259

43. Bahia D, Satoskar AR, Dussurget O (2018) Editorial: Cell Signaling in Host-Pathogen Interactions: The Host Point of View. Front Immunol 9:221

44. Lemaitre B, Nicolas E, Michaut L, Reichhart JM, Hoffmann JA (1996) The dorsoventral regulatory gene cassette spätzle/toll/cactus controls the potent antifungal response in Drosophila adults. Cell 86:973–983

45. Website. EMBL-EBI The home for big data in biology. https://www.ebi.ac.uk/training/online/courses/metabolomics-introduction/what-is/.

46. Trine H. Mogensen. Pathogen Recognition and Inflammatory Signaling in Innate Immune Defenses. American Society for Microbiology, Clinical Microbiology Reviews. Volume 22, Issue 2, April 2009, Pages 240-273. https://doi.org/10.1128/CMR.00046-08

47. Nguyen TG (2020) Harnessing Newton's third-law paradigm to treat autoimmune diseases and chronic inflammations. Inflammation Research 69:813–824

48. Wang L-F, Shi Z, Zhang S, Field H, Daszak P, Eaton BT (2006) Review of bats and SARS. Emerg Infect Dis 12:1834–1840

49. Ge X-Y, Li J-L, Yang X-L, et al (2013) Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature 503:535–538

50. Chen Y, Guo D (2016) Molecular mechanisms of coronavirus RNA capping and methylation. Virol Sin 31:3–11

51. Corman VM, Muth D, Niemeyer D, Drosten C (2018) Hosts and Sources of Endemic Human Coronaviruses. Adv Virus Res 100:163–188

52. Cherry J, Demmler-Harrison GJ, Kaplan SL, Steinbach WJ, Hotez PJ (2017) Feigin and Cherry's Textbook of Pediatric Infectious Diseases E-Book. Elsevier Health Sciences

53. Navas-Martín SR, Weiss S (2004) Coronavirus replication and pathogenesis: Implications for the recent outbreak of severe acute respiratory syndrome (SARS), and the challenge for vaccine development. J Neurovirol 10:75–85

54. Lu R, Zhao X, Li J, et al (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395:565–574

55. Lalchhandama K (2020) The chronicles of coronaviruses: the electron microscope, the doughnut, and the spike. Science Vision 20:78–92

56. Chen Y, Liu Q, Guo D (2020) Emerging coronaviruses: Genome structure, replication, and pathogenesis. Journal of Medical Virology 92:2249–2249

57. Lai MM, Cavanagh D (1997) The molecular biology of coronaviruses. Adv Virus Res 48:1–100

58. Alsaadi EAJ, Jones IM (2019) Membrane binding proteins of coronaviruses. Future Virology 14:275–286

59. Fehr AR, Perlman S (2015) Coronaviruses: An Overview of Their Replication and Pathogenesis. Coronaviruses 1–23

60. Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LLM, Guan Y, Rozanov M, Spaan WJM, Gorbalenya AE (2003) Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. J Mol Biol 331:991–1004

61. Zhand S, Jazi MS, Mohammadi S, Rasekhi RT, Rostamian G, Kalani MR, Rostamian A, George J, Douglas MW (2020) COVID-19: The Immune Responses and Clinical Therapy Candidates. International Journal of Molecular Sciences 21:5559

62. Ashburner M, Ball CA, Blake JA, et al (2000) Gene Ontology: tool for the unification of biology. Nature Genetics 25:25–29

63. Ma X, Gao L (2012) Biological network analysis: insights into structure and functions. Briefings in Functional Genomics 11:434–442

64. Clark JI, Brooksbank C, Lomax J (2005) It's All GO for Plant Scientists. Plant Physiology 138:1268–1279

65. Yu G, He Q-Y (2016) ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Molecular BioSystems 12:477–479

66. Yu G, Wang L-G, Yan G-R, He Q-Y (2015) DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics 31:608–609

67. Subramanian A, Tamayo P, Mootha VK, et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102:15545–15550

68. Sam E, Athri P (2019) Web-based drug repurposing tools: a survey. Brief Bioinform 20:299–316

69. Jin G, Wong STC (2014) Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. Drug Discov Today 19:637–644

70. Pfister DG (2012) Off-label use of oncology drugs: the need for more data and then some. J Clin Oncol 30:584–586

71. Burstein HJ (2013) Off-Label Use of Oncology Drugs: Too Much, Too Little, or Just Right? Journal of the National Comprehensive Cancer Network 11:505–506

72. Stephen R, Knopf K, Reynolds MW, Luo W, Fraeman K (2009) PCN112 OFF-LABEL USE OF ONCOLOGY DRUGS IN A COMMUNITY ONCOLOGY EMR DATABASE. Value in Health 12:A57

73. Swamidass SJ (2011) Mining small-molecule screens to repurpose drugs. Brief Bioinform 12:327–335

74. Ekins S, Mestres J, Testa B (2007) In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. Br J Pharmacol 152:9–20

75. Yildirim MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M (2007) Drug-target network. Nat Biotechnol 25:1119–1126

76. Kinnings SL, Liu N, Buchmeier N, Tonge PJ, Xie L, Bourne PE (2009) Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. PLoS Comput Biol 5:e1000423

77. Yang L, Agarwal P (2011) Systematic drug repositioning based on clinical side-effects. PLoS One 6:e28025

78. Campillos M, Kuhn M, Gavin A-C, Jensen LJ, Bork P (2008) Drug target identification using side-effect similarity. Science 321:263–266

79. Bisgin H, Liu Z, Kelly R, Fang H, Xu X, Tong W (2012) Investigating drug repositioning opportunities in FDA drug labels through topic modeling. BMC Bioinformatics 13 Suppl 15:S6

80. Gong J, Cai C, Liu X, Ku X, Jiang H, Gao D, Li H (2013) ChemMapper: a versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method. Bioinformatics 29:1827–1829

81. Kringelum J, Kjaerulff SK, Brunak S, Lund O, Oprea TI, Taboureau O (2016) ChemProt-3.0: a global chemical biology diseases mapping. Database 2016:bav123

82. Liu X, Vogt I, Haque T, Campillos M (2013) HitPick: a web server for hit identification and target prediction of chemical screenings. Bioinformatics 29:1910–1912

83. Xiao X, Min J-L, Lin W-Z, Liu Z, Cheng X, Chou K-C (2015) iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach. J Biomol Struct Dyn 33:2221–2233

84. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16:321–357

85. Awale M, Reymond J-L (2017) The polypharmacology browser: a web-based multi-fingerprint target prediction tool using ChEMBL bioactivity data. J Cheminform 9:11

86. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK (2007) Relating protein pharmacology by ligand chemistry. Nature Biotechnology 25:197–206

87. Nickel J, Gohlke B-O, Erehman J, Banerjee P, Rong WW, Goede A, Dunkel M, Preissner R (2014) SuperPred: update on drug classification and target prediction. Nucleic Acids Res 42:W26–31

88. Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, Zoete V (2014) SwissTargetPrediction: a web server for target prediction of bioactive small molecules. Nucleic Acids Res 42:W32–8

89. Daina A, Michielin O, Zoete V (2019) SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules. Nucleic Acids Res 47:W357–W364

90. Liu X, Gao Y, Peng J, Xu Y, Wang Y, Zhou N, Xing J, Luo X, Jiang H, Zheng M (2015) TarPred: a web application for predicting therapeutic and side effect targets of chemical compounds. Bioinformatics 31:2049–2051

91. Wang L, Ma C, Wipf P, Liu H, Su W, Xie X-Q (2013) TargetHunter: An In Silico Target Identification Tool for Predicting Therapeutic Potential of Small Organic Molecules Based on Chemogenomic Database. The AAPS Journal 15:395–406

92. Roy K (2019) In Silico Drug Design: Repurposing Techniques and Methodologies. Academic Press

93. Wang J-C, Chu P-Y, Chen C-M, Lin J-H (2012) idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. Nucleic Acids Res 40:W393–9

94. Wang C, Hu G, Wang K, Brylinski M, Xie L, Kurgan L (2016) PDID: database of molecular-level putative protein-drug interactions in the structural human proteome. Bioinformatics 32:579–586

95. Li H, Gao Z, Kang L, et al (2006) TarFisDock: a web server for identifying drug targets with docking approach. Nucleic Acids Res 34:W219–24

96. Cobanoglu MC, Oltvai ZN, Taylor DL, Bahar I (2015) BalestraWeb: efficient online evaluation of drug-target interactions. Bioinformatics 31:131–133

97. Lo Y-C, Senese S, Li C-M, Hu Q, Huang Y, Damoiseaux R, Torres JZ (2015) Large-scale chemical similarity networks for target profiling of compounds identified in cell-based chemical screens. PLoS Comput Biol 11:e1004153

98. Ba-Alawi W, Soufan O, Essack M, Kalnis P, Bajic VB (2016) DASPfind: new efficient method to predict drug-target interactions. J Cheminform 8:15

99. Martínez-Jiménez F, Marti-Renom MA (2015) Ligand-target prediction by structural network biology using nAnnoLyze. PLoS Comput Biol 11:e1004157

100. von Eichborn J, Murgueitio MS, Dunkel M, Koerner S, Bourne PE, Preissner R (2011) PROMISCUOUS: a database for network-based drug-repositioning. Nucleic Acids Res 39:D1060–6

101. Gallo K, Goede A, Eckert A, Moahamed B, Preissner R, Gohlke B-O (2021) PROMISCUOUS 2.0: a resource for drug-repositioning. Nucleic Acids Res 49:D1373–D1380

102. Chen B, Ding Y, Wild DJ (2012) Assessing drug target association using semantic linked data. PLoS Comput Biol 8:e1002574

103. Kuhn M, Szklarczyk D, Pletscher-Frankild S, Blicher TH, von Mering C, Jensen LJ, Bork P (2014) STITCH 4: integration of protein-chemical interactions with user data. Nucleic Acids Res 42:D401–7

104. Kuhn M, Szklarczyk D, Franceschini A, Campillos M, von Mering C, Jensen LJ, Beyer A, Bork P (2010) STITCH 2: an interaction network database for small molecules and proteins. Nucleic Acids Res 38:D552–6

105. Alaimo S, Bonnici V, Cancemi D, Ferro A, Giugno R, Pulvirenti A (2015) DT-Web: a web-based application for drug-target interaction and drug combination prediction through domain-tuned network-based inference. BMC Syst Biol 9 Suppl 3:S4

106. Alaimo S, Pulvirenti A, Giugno R, Ferro A (2013) Drug-target interaction prediction through domain-tuned network-based inference. Bioinformatics 29:2004–2008

107. Hattori M, Tanaka N, Kanehisa M, Goto S (2010) SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. Nucleic Acids Res 38:W652–6

108. Konc J, Janezic D (2012) ProBiS-2012: web server and web services for detection of structurally similar binding sites in proteins. Nucleic Acids Res 40:W214–21

109. Ito J-I, Tabei Y, Shimizu K, Tsuda K, Tomii K (2012) PoSSuM: a database of similar protein-ligand binding and putative pockets. Nucleic Acids Res 40:D541–8

110. Ito J-I, Ikeda K, Yamada K, Mizuguchi K, Tomii K (2015) PoSSuM v.2.0: data update and a new function for investigating ligand analogs and target proteins of small-molecule drugs. Nucleic Acids Research 43:D392–D398

111. Brown AS, Patel CJ (2017) MeSHDD: Literature-based drug-drug similarity for drug repositioning. J Am Med Inform Assoc 24:614–618

112. Moosavinasab S, Patterson J, Strouse R, Rastegar-Mojarad M, Regan K, Payne PRO, Huang Y, Lin SM (2016) “RE:fine drugs”: an interactive dashboard to access drug repurposing opportunities. Database . https://doi.org/10.1093/database/baw083

113. Lamb J, Crawford ED, Peck D, et al (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313:1929–1935

114. Lee BKB, Tiong KH, Chang JK, Liew CS, Abdul Rahman ZA, Tan AC, Khang TF, Cheong SC (2017) DeSigN: connecting gene expression with therapeutics for drug repurposing and development. BMC Genomics 18:934

115. Louhimo R, Laakso M, Belitskin D, Klefström J, Lehtonen R, Hautaniemi S (2016) Data integration to prioritize drugs using genomics and curated data. BioData Min 9:21

116. Carrella D, Napolitano F, Rispoli R, Miglietta M, Carissimo A, Cutillo L, Sirci F, Gregoretti F, Di Bernardo D (2014) Mantra 2.0: an online collaborative resource for drug mode of action and repurposing by network analysis. Bioinformatics 30:1787–1788

117. Setoain J, Franch M, Martínez M, Tabas-Madrid D, Sorzano COS, Bakker A, Gonzalez-Couto E, Elvira J, Pascual-Montano A (2015) NFFinder: an online bioinformatics tool for searching similar transcriptomics experiments in the context of drug repositioning. Nucleic Acids Res 43:W193–9

118. Yu H, Choo S, Park J, Jung J, Kang Y, Lee D (2016) Prediction of drugs having opposite effects on disease genes in a directed network. BMC Systems Biology. https://doi.org/10.1186/s12918-015-0243-2

119. Chen B, Ma L, Paik H, Sirota M, Wei W, Chua M-S, So S, Butte AJ (2017) Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. Nat Commun 8:16022

120. Fiscon G, Paci P (2021) SAveRUNNER: an R-based tool for drug repurposing. BMC Bioinformatics 22:150

121. Madhukar NS, Khade PK, Huang L, Gayvert K, Galletti G, Stogniew M, Allen JE, Giannakakou P, Elemento O (2019) A Bayesian machine learning approach for drug target identification using diverse data types. Nat Commun 10:5221

122. Duan Q, Reid SP, Clark NR, et al (2016) L1000CDS2: LINCS L1000 characteristic direction signatures search engine. npj Systems Biology and Applications. https://doi.org/10.1038/npjsba.2016.15

123. Musa A, Ghoraie LS, Zhang S-D, Glazko G, Yli-Harja O, Dehmer M, Haibe-Kains B, Emmert-Streib F (2017) A review of connectivity map and computational approaches in pharmacogenomics. Brief Bioinform 18:903

124. Wang R-S, Maron BA, Loscalzo J (2015) Systems medicine: evolution of systems biology from bench to bedside. Wiley Interdiscip Rev Syst Biol Med 7:141–161

125. Alaimo S, Giugno R, Acunzo M, Veneziano D, Ferro A, Pulvirenti A (2016) Post-transcriptional knowledge in pathway analysis increases the accuracy of phenotypes classification. Oncotarget 7:54572–54582

126. Glazko GV, Emmert-Streib F (2009) Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. Bioinformatics 25:2348–2354

127. Green ML, Karp PD (2006) The outcomes of pathway database computations depend on pathway ontology. Nucleic Acids Res 34:3687–3697

128. Khatri P, Sirota M, Butte AJ (2012) Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol 8:e1002375

129. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C (2011) Pathway Commons, a web resource for biological pathway data. Nucleic Acids Research 39:D685–D690

130. Kong SW, Pu WT, Park PJ (2006) A multivariate approach for integrating genome-wide expression data and biological knowledge. Bioinformatics 22:2373–2380

131. Klebanov L, Glazko G, Salzman P, Yakovlev A, Xiao Y (2007) A multivariate extension of the gene set enrichment analysis. J Bioinform Comput Biol 5:1139–1153

132. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ (2005) Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci U S A 102:13544–13549

133. Goeman JJ, Bühlmann P (2007) Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 23:980–987

134. Efron B, Tibshirani R (2007) On testing the significance of sets of genes. The Annals of Applied Statistics. https://doi.org/10.1214/07-aoas101

135. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R (2007) A systems biology approach for pathway level analysis. Genome Research 17:1537–1545

136. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim J-S, Kim CJ, Kusanovic JP, Romero R (2009) A novel signaling pathway impact analysis. Bioinformatics 25:75–82

137. Hsu S-D, Lin F-M, Wu W-Y, et al (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res 39:D163–9

138. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T (2009) miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res 37:D105–10

139. Wang J, Lu M, Qiu C, Cui Q (2010) TransmiR: a transcription factor-microRNA regulation database. Nucleic Acids Res 38:D119–22

140. Sedgewick AJ, Benz SC, Rabizadeh S, Soon-Shiong P, Vaske CJ (2013) Learning subgroup-specific regulatory interactions and regulator independence with PARADIGM. Bioinformatics 29:i62–70

141. Calura E, Martini P, Sales G, Beltrame L, Chiorino G, D'Incalci M, Marchini S, Romualdi C (2014) Wiring miRNAs to pathways: a topological approach to integrate miRNA and mRNA expression profiles. Nucleic Acids Res 42:e96

142. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26:i237–45

143. Calin GA, Dumitru CD, Shimizu M, et al (2002) Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci U S A 99:15524–15529

144. Acunzo M, Romano G, Wernicke D, Croce CM (2015) MicroRNA and cancer – A brief overview. Advances in Biological Regulation 57:1–9

145. Cormen TH, Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) Introduction To Algorithms. MIT Press

146. Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics. https://doi.org/10.1214/aos/1013699998

147. Website. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology. 2008; 9(10):770–80. https://doi.org/10.1038/nrm2503 PMID: 18797474.

148. Cohen DPA, Martignetti L, Robine S, Barillot E, Zinovyev A, Calzone L (2015) Mathematical Modelling of Molecular Pathways Enabling Tumour Cell Invasion and Migration. PLoS Comput Biol 11:e1004571

149. Sizek H, Hamel A, Deritei D, Campbell S, Ravasz Regan E (2019) Boolean model of growth signaling, cell cycle and apoptosis predicts the molecular mechanism of aberrant cell cycle progression driven by hyperactive PI3K. PLoS Comput Biol 15:e1006402

150. Barbuti R, Gori R, Milazzo P, Nasti L (2020) A survey of gene regulatory networks modelling methods: from differential equations, to Boolean and qualitative bioinspired models. Journal of Membrane Computing 2:207–226

151. Rubinstein A, Bracha N, Rudner L, Zucker N, Sloin HE, Chor B (2016) BioNSi: A Discrete Biological Network Simulator Tool. J Proteome Res 15:2871–2880

152. Yeheskel A, Reiter A, Pasmanik-Chor M, Rubinstein A (2017) Simulation and visualization of multiple KEGG pathways using BioNSi. F1000Res 6:2120

153. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45:D353–D361

154. Sauer U, Hatzimanikatis V, Hohmann HP, Manneberg M, van Loon AP, Bailey JE (1996) Physiology and metabolic fluxes of wild-type and riboflavin-producing Bacillus subtilis. Appl Environ Microbiol 62:3687–3696

155. Hellerstein MK (2003) In vivo measurement of fluxes through metabolic pathways: the missing link in functional genomics and pharmaceutical research. Annu Rev Nutr 23:379–402

156. Raman K, Chandra N (2009) Flux balance analysis of biological systems: applications and challenges. Brief Bioinform 10:435–449

157. Moutselos K, Kanaris I, Chatziioannou A, Maglogiannis I, Kolisis FN (2009) KEGGconverter: a tool for the in-silico modelling of metabolic networks of the KEGG Pathways database. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-10-324

158. Pilalis E, Koutsandreas T, Valavanis I, Athanasiadis E, Spyrou G, Chatziioannou A (2015) KENeV: A web-application for the automated reconstruction and visualization of the enriched metabolic and signaling super-pathways deriving from genomic experiments. Comput Struct Biotechnol J 13:248–255

159. Alaimo S, Marceca GP, Ferro A, Pulvirenti A (2017) Detecting Disease Specific Pathway Substructures through an Integrated Systems Biology Approach. Noncoding RNA. https://doi.org/10.3390/ncrna3020020

160. Huang H-Y, Lin Y-C-D, Li J, et al (2020) miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. Nucleic Acids Res 48:D148–D154

161. Tong Z, Cui Q, Wang J, Zhou Y (2019) TransmiR v2.0: an updated transcription factor-microRNA regulation database. Nucleic Acids Res 47:D253–D258

162. Barrett T, Wilhite SE, Ledoux P, et al (2013) NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res 41:D991–5

163. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43:e47

164. Nyman E, Stein RR, Jing X, Wang W, Marks B, Zervantonakis IK, Korkut A, Gauthier NP, Sander C Perturbation biology links temporal protein changes to drug responses in a melanoma cell line. https://doi.org/10.1101/568758

165. Bahrambeigi S, Shafiei-Irannejad V (2020) Immune-mediated anti-tumor effects of metformin; targeting metabolic reprogramming of T cells as a new possible mechanism for anti-cancer effects of metformin. Biochem Pharmacol 174:113787

166. Saraei P, Asadi I, Kakar MA, Moradi-Kor N (2019) The beneficial effects of metformin on cancer prevention and therapy: a comprehensive review of recent advances. Cancer Manag Res 11:3295–3313

167. Cantoria MJ, Patel H, Boros LG, Meuillet EJ (2014) Metformin and Pancreatic Cancer Metabolism. Pancreatic Cancer - Insights into Molecular Mechanisms and Novel Approaches to Early Detection and Treatment. https://doi.org/10.5772/57432

168. Yu X, Mao W, Zhai Y, Tong C, Liu M, Ma L, Yu X, Li S (2017) Anti-tumor activity of metformin: from metabolic and epigenetic perspectives. Oncotarget 8:5619–5628

169. Sekino N, Kano M, Matsumoto Y, et al (2018) Antitumor effects of metformin are a result of inhibiting nuclear factor kappa B nuclear translocation in esophageal squamous cell carcinoma. Cancer Sci 109:1066–1074

170. Gong J, Kelekar G, Shen J, Shen J, Kaur S, Mita M (2016) The expanding role of metformin in cancer: an update on antitumor mechanisms and clinical development. Target Oncol 11:447–467

171. Sultuybek GK, Soydas T, Yenmis G (2019) NF-κB as the mediator of metformin's effect on ageing and ageing-related diseases. Clinical and Experimental Pharmacology and Physiology 46:413–422

172. Schulten H-J (2018) Pleiotropic Effects of Metformin on Cancer. Int J Mol Sci. https://doi.org/10.3390/ijms19102850

173. Citi V, Del Re M, Martelli A, Calderone V, Breschi MC, Danesi R (2018) Phosphorylation of AKT and ERK1/2 and mutations of PIK3CA and PTEN are predictive of breast cancer cell sensitivity to everolimus in vitro. Cancer Chemother Pharmacol 81:745–754

174. Hurvitz SA, Kalous O, Conklin D, et al (2015) In vitro activity of the mTOR inhibitor everolimus, in a large panel of breast cancer cell lines and analysis for predictors of response. Breast Cancer Research and Treatment 149:669–680

175. O'Reilly T, McSheehy PM (2010) Biomarker Development for the Clinical Activity of the mTOR Inhibitor Everolimus (RAD001): Processes, Limitations, and Further Proposals. Transl Oncol 3:65–79

176. Wazir U, Wazir A, Khanzada ZS, Jiang WG, Sharma AK, Mokbel K (2014) Current state of mTOR targeting in human breast cancer. Cancer Genomics Proteomics 11:167–174

177. Lui A, New J, Ogony J, Thomas S, Lewis-Wambi J (2016) Everolimus downregulates estrogen receptor and induces autophagy in aromatase inhibitor-resistant breast cancer cells. BMC Cancer 16:487

178. Hua H, Kong Q, Zhang H, Wang J, Luo T, Jiang Y (2019) Targeting mTOR for cancer therapy. J Hematol Oncol 12:71

179. Formisano L, Napolitano F, Rosa R, D'Amato V, Servetto A, Marciano R, De Placido P, Bianco C, Bianco R (2020) Mechanisms of resistance to mTOR inhibitors. Critical Reviews in Oncology/Hematology 147:102886

180. Yamanaka K, Petrulionis M, Lin S, et al (2013) Therapeutic potential and adverse events of everolimus for treatment of hepatocellular carcinoma - systematic review and meta-analysis. Cancer Med 2:862–871

181. Royce ME, Osman D (2015) Everolimus in the Treatment of Metastatic Breast Cancer. Breast Cancer 9:73–79

182. Kuo C-T, Chen C-L, Li C-C, Huang G-S, Ma W-Y, Hsu W-F, Lin C-H, Lu Y-S, Wo AM (2020) Author Correction: Immunofluorescence can assess the efficacy of mTOR pathway therapeutic agent Everolimus in breast cancer models. Sci Rep 10:14139

183. Houghton PJ (2010) Everolimus. Clin Cancer Res 16:1368–1372

184. Hare SH, Harvey AJ (2017) mTOR function and therapeutic targeting in breast cancer. Am J Cancer Res 7:383–404

185. Michels AA (2011) MAF1: a new target of mTORC1. Biochem Soc Trans 39:487–491

186. Ben-Sahra I, Manning BD (2017) mTORC1 signaling and the metabolic control of cell growth. Curr Opin Cell Biol 45:72–82

187. Ben-Sahra I, Howell JJ, Asara JM, Manning BD (2013) Stimulation of de novo pyrimidine synthesis by growth signaling through mTOR and S6K1. Science 339:1323–1328

188. Eisenberg-Lerner A, Bialik S, Simon H-U, Kimchi A (2009) Life and death partners: apoptosis, autophagy and the cross-talk between them. Cell Death Differ 16:966–975

189. Kumar B, Garcia M, Weng L, et al (2018) Acute myeloid leukemia transforms the bone marrow niche into a leukemia-permissive microenvironment through exosome secretion. Leukemia 32:575–587

190. Wang X, Chen H, Bai J, He A (2017) MicroRNA: an important regulator in acute myeloid leukemia. Cell Biol Int 41:936–945

191. Hornick NI, Huan J, Doron B, Goloviznina NA, Lapidus J, Chang BH, Kurre P (2015) Serum Exosome MicroRNA as a Minimally-Invasive Early Biomarker of AML. Sci Rep 5:11295

192. Hornick NI, Doron B, Abdelhamed S, Huan J, Harrington CA, Shen R, Cambronne XA, Chakkaramakkil Verghese S, Kurre P (2016) AML suppresses hematopoiesis by releasing exosomes that contain microRNAs targeting c-MYB. Sci Signal 9:ra88

193. Serebrennikova OB, Paraskevopoulou MD, Aguado-Fraile E, Taraslia V, Ren W, Thapa G, Roper J, Du K, Croce CM, Tsichlis PN (2019) The combination of knockdown and TNFα causes synthetic lethality via caspase-8 activation in human carcinoma cell lines. Proc Natl Acad Sci U S A 116:14039–14048

194. Robitaille AM, Christen S, Shimobayashi M, Cornu M, Fava LL, Moes S, Prescianotto-Baschong C, Sauer U, Jenoe P, Hall MN (2013) Quantitative phosphoproteomics reveal mTORC1 activates de novo pyrimidine synthesis. Science 339:1320–1323

195. Alaimo S, Micale G, La Ferlita A, Ferro A, Pulvirenti A (2019) Computational Methods to Investigate the Impact of miRNAs on Pathways. Methods in Molecular Biology 183–209

196. Kanehisa M (2019) Toward understanding the origin and evolution of cellular organisms. Protein Sci 28:1947–1951

197. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási A-L (2015) Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science 347:1257601

198. Nicholson DN, Greene CS (2020) Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J 18:1414–1428

199. Beck J Report from the Field: PubMed Central, an XML-based Archive of Life Sciences Journal Articles. Proceedings of the International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML. https://doi.org/10.4242/balisagevol6.beck01

200. Website. P.: arXiv. Available at https://arxiv.org.

201. Website. bioRxiv https://www.biorxiv.org/.

202. Lambrix P, Tan H, Jakoniene V, Strömbäck L Biological Ontologies. Semantic Web 85–99

203. Cohen AM (2005) A survey of current work in biomedical text mining. Briefings in Bioinformatics 6:57–71

204. Krallinger M, Erhardt RA-A, Valencia A (2005) Text-mining approaches in molecular biology and biomedicine. Drug Discov Today 10:439–445

205. Szklarczyk D, Morris JH, Cook H, et al (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 45:D362–D368

206. Dörpinghaus J, Apke A, Lage-Rupprecht V, Stefan A (2019) Data Exploration and Validation on dense knowledge graphs for biomedical research.

207. Consortium GO, Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 32:258D–261

208. Wishart DS, Feunang YD, Guo AC, et al (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research 46:D1074–D1082

209. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 48:D845–D855

210. Smith B, Ashburner M, Rosse C, et al (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25:1251–1255

211. Honnibal M, Johnson M (2015) An Improved Non-monotonic Transition System for Dependency Parsing. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/d15-1162

212. Loper E, Bird S (2002) NLTK. Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics -. https://doi.org/10.3115/1118108.1118117

213. Barabási A-L, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet 12:56–68

214. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, Green A, Khankhanian P, Baranzini SE (2017) Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife. https://doi.org/10.7554/elife.26726

215. Blohm P, Frishman G, Smialowski P, Goebels F, Wachinger B, Ruepp A, Frishman D (2014) Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. Nucleic Acids Research 42:D396–D400

216. Smialowski P, Pagel P, Wong P, et al (2010) The Negatome database: a reference set of non-interacting protein pairs. Nucleic Acids Research 38:D540–D544

217. Parikhani AB, Bazaz M, Bamehr H, Fereshteh S, Amiri S, Salehi-Vaziri M, Arashkia A, Azadmanesh K (2021) The Inclusive Review on SARS-CoV-2 Biology, Epidemiology, Diagnosis, and Potential Management Options. Current Microbiology 78:1099–1114

218. WHO Coronavirus Disease (COVID-19) Dashboard (2020). Bangladesh Physiotherapy Journal. https://doi.org/10.46945/bpj.10.1.03.01

219. WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int. Accessed 4 Dec 2021

220. Xu Z, Shi L, Wang Y, et al (2020) Pathological findings of COVID-19 associated with acute respiratory distress syndrome. Lancet Respir Med 8:420–422

221. Naqvi AAT, Fatima K, Mohammad T, Fatima U, Singh IK, Singh A, Atif SM, Hariprasad G, Hasan GM, Hassan MI (2020) Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. Biochim Biophys Acta Mol Basis Dis 1866:165878

222. Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ (2006) Nidovirales: evolving the largest RNA virus genome. Virus Res 117:17–37

223. Xue B, Blocquel D, Habchi J, Uversky AV, Kurgan L, Uversky VN, Longhi S (2014) Structural Disorder in Viral Proteins. Chemical Reviews 114:6880–6911

224. Müller C, Schulte FW, Lange-Grünweller K, Obermann W, Madhugiri R, Pleschka S, Ziebuhr J, Hartmann RK, Grünweller A (2018) Broad-spectrum antiviral activity of the eIF4A inhibitor silvestrol against corona- and picornaviruses. Antiviral Res 150:123–129

225. Rastogi M, Pandey N, Shukla A, Singh SK (2020) SARS coronavirus 2: from genome to infectome. Respir Res 21:318

226. Lee J-E, Chung JK, Kim TS, et al Genomic and phylogenetic analyses of SARS-CoV-2 strains isolated in the city of Gwangju, South Korea. https://doi.org/10.1101/2020.12.16.423178

227. Hellewell J, Abbott S, Gimma A, et al (2020) Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. Lancet Glob Health 8:e488–e496

228. Yang N, Shen H-M (2020) Targeting the Endocytic Pathway and Autophagy Process as a Novel Therapeutic Strategy in COVID-19. Int J Biol Sci 16:1724–1731

229. Hoffmann M, Kleine-Weber H, Schroeder S, et al (2020) SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell 181:271–280.e8

230. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, Graham BS, McLellan JS (2020) Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation. bioRxiv. https://doi.org/10.1101/2020.02.11.944462

231. Damas J, Hughes GM, Keough KC, et al (2020) Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. Proc Natl Acad Sci U S A 117:22311–22322

232. Lukassen S, Chua RL, Trefzer T, et al (2020) SARS-CoV-2 receptor ACE2 and TMPRSS2 are primarily expressed in bronchial transient secretory cells. EMBO J 39:e105114

233. Subramanian A, Vernon KA, Slyper M, et al RAAS blockade, kidney disease, and expression of ACE2, the entry receptor for SARS-CoV-2, in kidney epithelial and endothelial cells. https://doi.org/10.1101/2020.06.23.167098

234. Gierer S, Bertram S, Kaup F, et al (2013) The spike protein of the emerging betacoronavirus EMC uses a novel coronavirus receptor for entry, can be activated by TMPRSS2, and is targeted by neutralizing antibodies. J Virol 87:5502–5511

235. Matsuyama S, Nagata N, Shirato K, Kawase M, Takeda M, Taguchi F (2010) Efficient activation of the severe acute respiratory syndrome coronavirus spike protein by the transmembrane protease TMPRSS2. J Virol 84:12658–12664

236. Hoffmann M, Kleine-Weber H, Krüger N, Müller M, Drosten C, Pöhlmann S The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. https://doi.org/10.1101/2020.01.31.929042

237. Ou X, Liu Y, Lei X, et al (2021) Author Correction: Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nat Commun 12:2144

238. Shang J, Wan Y, Liu C, Yount B, Gully K, Yang Y, Auerbach A, Peng G, Baric R, Li F (2020) Structure of mouse coronavirus spike protein complexed with receptor reveals mechanism for viral entry. PLoS Pathog 16:e1008392

239. Raghuvamsi PV, Tulsian NK, Samsudin F, et al (2021) SARS-CoV-2 S protein:ACE2 interaction reveals novel allosteric targets. Elife. https://doi.org/10.7554/eLife.63646

240. Glowacka I, Bertram S, Müller MA, et al (2011) Evidence that TMPRSS2 activates the severe acute respiratory syndrome coronavirus spike protein for membrane fusion and reduces viral control by the humoral immune response. J Virol 85:4122–4134

241. Lin B, White JT, Utleg AG, Wang S, Ferguson C, True LD, Vessella R, Hood L, Nelson PS (2003) Isolation and characterization of human and mouse WDR19,a novel WD-repeat protein exhibiting androgen-regulated expression in prostate epithelium☆. Genomics 82:331–342

242. Ganier C, Du-Harpur X, Harun N, Wan B, Arthurs C, Luscombe NM, Watt FM, Lynch MD CD147 (BSG) but not ACE2 expression is detectable in vascular endothelial cells within single cell RNA sequencing datasets derived from multiple tissues in healthy individuals. https://doi.org/10.1101/2020.05.29.123513

243. Alhadrami HA, Sayed AM, Hassan HM, et al (2021) Cnicin as an Anti-SARS-CoV-2: An Integrated In Silico and In Vitro Approach for the Rapid Identification of Potential COVID-19 Therapeutics. Antibiotics (Basel). https://doi.org/10.3390/antibiotics10050542

244. Lu G, Hu Y, Wang Q, et al (2013) Molecular basis of binding between novel human coronavirus MERS-CoV and its receptor CD26. Nature 500:227–231

245. Yeager CL, Ashmun RA, Williams RK, Cardellichio CB, Shapiro LH, Thomas Look A, Holmes KV (1992) Human aminopeptidase N is a receptor for human coronavirus 229E. Nature 357:420–422

246. Cantuti-Castelvetri L, Ojha R, Pedro LD, et al (2020) Neuropilin-1 facilitates SARS-CoV-2 cell entry and infectivity. Science 370:856–860

247. Moutal A, Martin LF, Boinon L, et al (2020) SARS-CoV-2 Spike protein co-opts VEGF-A/Neuropilin-1 receptor signaling to induce analgesia. bioRxiv. https://doi.org/10.1101/2020.07.17.209288

248. Zhang Q, Xiang R, Huo S, Zhou Y, Jiang S, Wang Q, Yu F (2021) Molecular mechanism of interaction between SARS-CoV-2 and host cells and interventional therapy. Signal Transduct Target Ther 6:233

249. Rossi A, Deveraux Q, Turk B, Sali A (2004) Comprehensive search for cysteine cathepsins in the human genome. Biol Chem 385:363–372

250. Simmons G, Gosalia DN, Rennekamp AJ, Reeves JD, Diamond SL, Bates P (2005) Inhibitors of cathepsin L prevent severe acute respiratory syndrome coronavirus entry. Proc Natl Acad Sci U S A 102:11876–11881

251. Gosert R, Kanjanahaluethai A, Egger D, Bienz K, Baker SC (2002) RNA Replication of Mouse Hepatitis Virus Takes Place at Double-Membrane Vesicles. Journal of Virology 76:3697–3708

252. Wolff G, Limpens RWAL, Zevenhoven-Dobbe JC, et al (2020) A molecular pore spans the double membrane of the coronavirus replication organelle. Science 369:1395–1398

253. Stertz S, Reichelt M, Spiegel M, Kuri T, Martínez-Sobrido L, García-Sastre A, Weber F, Kochs G (2007) The intracellular sites of early replication and budding of SARS-coronavirus. Virology 361:304–315

254. Goldsmith CS, Tatti KM, Ksiazek TG, Rollin PE, Comer JA, Lee WW, Rota PA, Bankamp B, Bellini WJ, Zaki SR (2004) Ultrastructural characterization of SARS coronavirus. Emerg Infect Dis 10:320–326

255. Hoffmann M, Kleine-Weber H, Pöhlmann S (2020) A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. Mol Cell 78:779–784.e5

256. Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, Li F (2020) Cell entry mechanisms of SARS-CoV-2. Proc Natl Acad Sci U S A 117:11727–11734

257. Neuman BW, Kiss G, Kunding AH, et al (2011) A structural analysis of M protein in coronavirus assembly and morphology. J Struct Biol 174:11–22

258. Schoeman D, Fielding BC (2019) Coronavirus envelope protein: current knowledge. Virol J 16:69

259. Jackson CB, Farzan M, Chen B, Choe H (2021) Mechanisms of SARS-CoV-2 entry into cells. Nat Rev Mol Cell Biol. https://doi.org/10.1038/s41580-021-00418-x

260. Hogue BG, Machamer CE (2014) Coronavirus Structural Proteins and Virus Assembly. Nidoviruses 179–200

261. Cascella M, Rajnik M, Aleem A, Dulebohn SC, Di Napoli R (2021) Features, Evaluation, and Treatment of Coronavirus (COVID-19). StatPearls

262. Ivashkiv LB, Donlin LT (2014) Regulation of type I interferon responses. Nat Rev Immunol 14:36–49

263. Li G, Fan Y, Lai Y, et al (2020) Coronavirus infections and immune responses. J Med Virol 92:424–432

264. George MR (2014) Hemophagocytic lymphohistiocytosis: review of etiologies and management. J Blood Med 5:69–86

265. McGonagle D, Sharif K, O'Regan A, Bridgewood C (2020) The Role of Cytokines including Interleukin-6 in COVID-19 induced Pneumonia and Macrophage Activation Syndrome-Like Disease. Autoimmunity Reviews 19:102537

266. Wan S, Yi Q, Fan S, et al Characteristics of lymphocyte subsets and cytokines in peripheral blood of 123 hospitalized patients with 2019 novel coronavirus pneumonia (NCP). https://doi.org/10.1101/2020.02.10.20021832

267. Wang F, Nie J, Wang H, Zhao Q, Xiong Y, Deng L, Song S, Ma Z, Mo P, Zhang Y (2020) Characteristics of Peripheral Lymphocyte Subset Alteration in COVID-19 Pneumonia. J Infect Dis 221:1762–1769

268. Catanzaro M, Fagiani F, Racchi M, Corsini E, Govoni S, Lanni C (2020) Immune response in COVID-19: addressing a pharmacological challenge by targeting pathways triggered by SARS-CoV-2. Signal Transduct Target Ther 5:84

269. Prompetchara E, Ketloy C, Palaga T (2020) Immune responses in COVID-19 and potential vaccines: Lessons learned from SARS and MERS epidemic. Asian Pac J Allergy Immunol 38:1–9

270. Kim Y-M, Shin E-C (2021) Type I and III interferon responses in SARS-CoV-2 infection. Exp Mol Med 53:750–760

271. Menachery VD, Debbink K, Baric RS (2014) Coronavirus non-structural protein 16: evasion, attenuation, and possible treatments. Virus Res 194:191–199

272. Hackbart M, Deng X, Baker SC (2020) Coronavirus endoribonuclease targets viral polyuridine sequences to evade activating host sensors. Proc Natl Acad Sci U S A 117:8094–8103

273. Chen Y, Cai H, Pan J 'an, Xiang N, Tien P, Ahola T, Guo D (2009) Functional screen reveals SARS coronavirus nonstructural protein nsp14 as a novel cap N7 methyltransferase. Proc Natl Acad Sci U S A 106:3484–3489

274. Zhu N, Zhang D, Wang W, et al (2020) A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine 382:727–733

275. Huang C, Wang Y, Li X, et al (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395:497–506

276. Qin C, Zhou L, Hu Z, et al (2020) Dysregulation of Immune Response in Patients With Coronavirus 2019 (COVID-19) in Wuhan, China. Clin Infect Dis 71:762–768

277. Wu F, Zhao S, Yu B, et al (2020) Author Correction: A new coronavirus associated with human respiratory disease in China. Nature 580:E7–E7

278. Zinkernagel RM (1996) Immunology taught by viruses. Science 271:173–178

279. Azkur AK, Akdis M, Azkur D, Sokolowska M, van de Veen W, Brüggen M-C, O'Mahony L, Gao Y, Nadeau K, Akdis CA (2020) Immune response to SARS-CoV-2 and mechanisms of immunopathological changes in COVID-19. Allergy 75:1564–1581

280. Long Q-X, Liu B-Z, Deng H-J, et al (2020) Antibody responses to SARS-CoV-2 in patients with COVID-19. Nat Med 26:845–848

281. Cañete PF, Vinuesa CG (2020) COVID-19 Makes B Cells Forget, but T Cells Remember. Cell 183:13–15

282. Melenotte C, Silvin A, Goubet A-G, et al (2020) Immune responses during COVID-19 infection. Oncoimmunology 9:1807836

283. Shi Y, Tan M, Chen X, Liu Y, Huang J, Ou J, Deng X Immunopathological characteristics of coronavirus disease 2019 cases in Guangzhou, China. https://doi.org/10.1101/2020.03.12.20034736

284. Shi P, Ren G, Yang J, et al (2020) Clinical characteristics of imported and second-generation coronavirus disease 2019 (COVID-19) cases in Shaanxi outside Wuhan, China: a multicentre retrospective study. Epidemiology and Infection. https://doi.org/10.1017/s0950268820002332

285. Zhang B, Zhou X, Zhu C, Feng F, Qiu Y, Feng J, Jia Q, Song Q, Zhu B, Wang J Immune phenotyping based on neutrophil-to-lymphocyte ratio and IgG predicts disease severity and outcome for patients with COVID-19. https://doi.org/10.1101/2020.03.12.20035048

286. Zhang B, Zhou X, Zhu C, et al (2020) Immune Phenotyping Based on the Neutrophil-to-Lymphocyte Ratio and IgG Level Predicts Disease Severity and Outcome for Patients With COVID-19. Front Mol Biosci 7:157

287. Chen Z, John Wherry E (2020) T cell responses in patients with COVID-19. Nature Reviews Immunology 20:529–536

288. Li D, Chen Y, Liu H, Jia Y, Li F, Wang W, Wu J, Wan Z, Cao Y, Zeng R (2020) Immune dysfunction leads to mortality and organ injury in patients with COVID-19 in China: insights from ERS-COVID-19 study. Signal Transduct Target Ther 5:62

289. Zhou L, Liu K, Liu HG (2020) [Cause analysis and treatment strategies of "recurrence" with novel coronavirus pneumonia (COVID-19) patients after discharge from hospital]. Zhonghua Jie He He Hu Xi Za Zhi 43:281–284

290. Chen D, Xu W, Lei Z, Huang Z, Liu J, Gao Z, Peng L (2020) Recurrence of positive SARS-CoV-2 RNA in COVID-19: A case report. International Journal of Infectious Diseases 93:297–299

291. Tan L, Wang Q, Zhang D, Ding J, Huang Q, Tang Y-Q, Wang Q, Miao H (2020) Correction: Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study. Signal Transduct Target Ther 5:61

292. Bradley BT, Maioli H, Johnston R, et al (2020) Histopathology and ultrastructural findings of fatal COVID-19 infections in Washington State: a case series. Lancet 396:320–332

293. Hou YJ, Okuda K, Edwards CE, et al (2020) SARS-CoV-2 Reverse Genetics Reveals a Variable Infection Gradient in the Respiratory Tract. Cell 182:429–446.e14

294. Carsana L, Sonzogni A, Nasr A, et al (2020) Pulmonary post-mortem findings in a series of COVID-19 cases from northern Italy: a two-centre descriptive study. The Lancet Infectious Diseases 20:1135–1140

295. Polat V, Bostancı Gİ (2020) Sudden death due to acute pulmonary embolism in a young woman with COVID-19. Journal of Thrombosis and Thrombolysis 50:239–241

296. Yap JKY, Moriyama M, Iwasaki A (2020) Inflammasomes and Pyroptosis as Therapeutic Targets for COVID-19. J Immunol 205:307–312

297. Yang M Cell Pyroptosis, a Potential Pathogenic Mechanism of 2019-nCoV Infection. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3527420

298. Chen R, Wang K, Yu J, Howard D, French L, Chen Z, Wen C, Xu Z (2020) The Spatial and Cell-Type Distribution of SARS-CoV-2 Receptor ACE2 in the Human and Mouse Brains. Front Neurol 11:573095

299. Shiers S, Ray PR, Wangzhou A, Tatsui CE, Rhines L, Li Y, Uhelski ML, Dougherty PM, Price TJ ACE2 expression in human dorsal root ganglion sensory neurons: implications for SARS-CoV-2 virus-induced neurological effects. https://doi.org/10.1101/2020.05.28.122374

300. Fodoulian L, Tuberosa J, Rossier D, et al SARS-CoV-2 receptor and entry genes are expressed by sustentacular cells in the human olfactory neuroepithelium. https://doi.org/10.1101/2020.03.31.013268

301. Butowt R, Bilinska K (2020) SARS-CoV-2: Olfaction, Brain Infection, and the Urgent Need for Clinical Samples Allowing Earlier Virus Detection. ACS Chem Neurosci 11:1200–1203

302. Brann DH, Tsukahara T, Weinreb C, et al (2020) Non-neuronal expression of SARS-CoV-2 entry genes in the olfactory system suggests mechanisms underlying COVID-19-associated anosmia. Sci Adv. https://doi.org/10.1126/sciadv.abc5801

303. Mao L, Jin H, Wang M, et al (2020) Neurologic Manifestations of Hospitalized Patients With Coronavirus Disease 2019 in Wuhan, China. JAMA Neurol 77:683–690

304. Dawson P, Rabold EM, Laws RL, et al (2021) Loss of Taste and Smell as Distinguishing Symptoms of Coronavirus Disease 2019. Clin Infect Dis 72:682–685

305. Vaira LA, Deiana G, Fois AG, et al (2020) Objective evaluation of anosmia and ageusia in COVID-19 patients: Single-center experience on 72 cases. Head Neck 42:1252–1258

306. Bernard-Valnet R, Pizzarotti B, Anichini A, Demars Y, Russo E, Schmidhauser M, Cerutti-Sola J, Rossetti AO, Du Pasquier R Two patients with acute meningo-encephalitis concomitant to SARS-CoV-2 infection. https://doi.org/10.1101/2020.04.17.20060251

307. Alberti P, Beretta S, Piatti M, et al (2020) Guillain-Barré syndrome related to COVID-19 infection. Neurol Neuroimmunol Neuroinflamm. https://doi.org/10.1212/NXI.0000000000000741

308. Benussi A, Pilotto A, Premi E, et al (2020) Clinical characteristics and outcomes of inpatients with neurologic disease and COVID-19 in Brescia, Lombardy, Italy. Neurology 95:e910–e920

309. Yin R, Yang Z, Wei Y, et al Clinical characteristics of 106 patients with neurological diseases and co-morbid coronavirus disease 2019: a retrospective study. https://doi.org/10.1101/2020.04.29.20085415

310. Coolen T, Lolli V, Sadeghi N, et al (2020) Early postmortem brain MRI findings in COVID-19 non-survivors. Neurology 95:e2016–e2027

311. Hess DC, Eldahshan W, Rutkowski E (2020) COVID-19-Related Stroke. Transl Stroke Res 11:322–325

312. Poyiadji N, Shahin G, Noujaim D, Stone M, Patel S, Griffith B (2020) COVID-19–associated Acute Hemorrhagic Necrotizing Encephalopathy: Imaging Features. Radiology 296:E119–E120

313. Bozkurt B, Eğrilmez S, Şengör T, Yıldırım Ö, İrkeç M (2020) The COVID-19 Pandemic: Clinical Information for Ophthalmologists. Turk J Ophthalmol 50:59–63

314. Hamashima K, Gautam P, Lau KA, Khiong CW, Blenkinsop TA, Li H, Loh Y-H Potential modes of COVID-19 transmission from human eye revealed by single-cell atlas. https://doi.org/10.1101/2020.05.09.085613

315. Zhou L, Xu Z, Castiglione GM, Soiberman US, Eberhart CG, Duh EJ (2020) ACE2 and TMPRSS2 are expressed on the human ocular surface, suggesting susceptibility to SARS-CoV-2 infection. Ocul Surf 18:537–544

316. Argenziano MG, Bruce SL, Slater CL, et al (2020) Characterization and clinical course of 1000 patients with coronavirus disease 2019 in New York: retrospective case series. BMJ 369:m1996

317. Venkatakrishnan AJ, Puranik A, Anand A, et al Knowledge synthesis from 100 million biomedical documents augments the deep expression profiling of coronavirus receptors. https://doi.org/10.1101/2020.03.24.005702

318. Xu D, Ma M, Xu Y, Su Y, Ong S-B, Hu X, Chai M, Zhao M, Li H, Xu X Single-cell Transcriptome Analysis Indicates New Potential Regulation Mechanism of ACE2 and NPs signaling among heart failure patients infected with SARS-CoV-2. https://doi.org/10.1101/2020.04.30.20081257

319. Lin L, Jiang X, Zhang Z, et al (2020) Gastrointestinal symptoms of 95 cases with SARS-CoV-2 infection. Gut 69:997–1001

320. Gupta S, Parker J, Smits S, Underwood J, Dolwani S (2020) Persistent viral shedding of SARS-CoV-2 in faeces - a rapid review. Colorectal Dis 22:611–620

321. Parasa S, Desai M, Thoguluva Chandrasekar V, et al (2020) Prevalence of Gastrointestinal Symptoms and Fecal Viral Shedding in Patients With Coronavirus Disease 2019: A Systematic Review and Meta-analysis. JAMA Netw Open 3:e2011335

322. Song J, Li Y, Huang X, Chen Z, Li Y, Liu C, Chen Z, Duan X (2020) Systematic analysis of ACE2 and TMPRSS2 expression in salivary glands reveals underlying transmission mechanism caused by SARS-CoV-2. J Med Virol 92:2556–2566

323. Wang J, Zhao S, Liu M, et al ACE2 expression by colonic epithelial cells is associated with viral infection, immunity and energy metabolism. https://doi.org/10.1101/2020.02.05.20020545

324. Xu H, Zhong L, Deng J, Peng J, Dan H, Zeng X, Li T, Chen Q (2020) High expression of ACE2 receptor of 2019-nCoV on the epithelial cells of oral mucosa. Int J Oral Sci 12:8

325. Zabetakis I, Matthys C, Tsoupras A (2021) Coronavirus Disease (COVID-19): Diet, Inflammation and Nutritional Status. Frontiers Media SA

326. Wang Q, Zhao H, Liu L-G, et al (2020) Pattern of liver injury in adult patients with COVID-19: a retrospective analysis of 105 patients. Mil Med Res 7:28

327. Fan Z, Chen L, Li J, Cheng X, Yang J, Tian C, Zhang Y, Huang S, Liu Z, Cheng J (2020) Clinical Features of COVID-19-Related Liver Functional Abnormality. Clin Gastroenterol Hepatol 18:1561–1566

328. Chu KH, Tsang WK, Tang CS, et al (2005) Acute renal impairment in coronavirus-associated severe acute respiratory syndrome. Kidney International 67:698–705

329. Diao B, Wang C, Wang R, et al (2021) Human kidney is a target for novel severe acute respiratory syndrome coronavirus 2 infection. Nat Commun 12:2506

330. Fan C, Li K, Ding Y, Lu W, Wang J ACE2 Expression in Kidney and Testis May Cause Kidney and Testis Damage After 2019-nCoV Infection. https://doi.org/10.1101/2020.02.12.20022418

331. Farooqui T, Farooqui AA (2021) Gut Microbiota in Neurologic and Visceral Diseases. Academic Press

332. Lin W, Fan J, Hu L-F, et al (2021) Single-cell analysis of angiotensin-converting enzyme II expression in human kidneys and bladders reveals a potential route of 2019 novel coronavirus infection. Chin Med J 134:935–943

333. Hikmet F, Méar L, Edvinsson Å, Micke P, Uhlén M, Lindskog C (2020) The protein expression profile of ACE2 in human tissues. Mol Syst Biol 16:e9610

334. Ren X, Wang S, Chen X, et al (2020) Multiple Expression Assessments of ACE2 and TMPRSS2 SARS-CoV-2 Entry Molecules in the Urinary Tract and Their Associations with Clinical Manifestations of COVID-19. Infect Drug Resist 13:3977–3990

335. Monteil V, Kwon H, Prado P, et al (2020) Inhibition of SARS-CoV-2 Infections in Engineered Human Tissues Using Clinical-Grade Soluble Human ACE2. Cell 181:905–913.e7

336. Su H, Yang M, Wan C, et al (2020) Renal histopathological analysis of 26 postmortem findings of patients with COVID-19 in China. Kidney International 98:219–227

337. Pesaresi M, Pirani F, Tagliabracci A, Valsecchi M, Procopio AD, Busardò FP, Graciotti L (2020) SARS-CoV-2 identification in lungs, heart and kidney specimens by transmission and scanning electron microscopy. Eur Rev Med Pharmacol Sci 24:5186–5188

338. Pan F, Xiao X, Guo J, et al (2020) No evidence of severe acute respiratory syndrome–coronavirus 2 in semen of males recovering from coronavirus disease 2019. Fertility and Sterility 113:1135–1139

339. Pique-Regi R, Romero R, Tarca AL, Luca F, Xu Y, Alazizi A, Leng Y, Hsu C-D, Gomez-Lopez N (2020) Does the human placenta express the canonical cell entry mediators for SARS-CoV-2? Elife. https://doi.org/10.7554/eLife.58716

340. Chen H, Guo J, Wang C, et al (2020) Clinical characteristics and intrauterine vertical transmission potential of COVID-19 infection in nine pregnant women: a retrospective review of medical records. Lancet 395:809–815

341. Kalafat E, Yaprak E, Cinar G, Varli B, Ozisik S, Uzun C, Azap A, Koc A (2020) Lung ultrasound and computed tomographic findings in pregnant woman with COVID-19. Ultrasound Obstet Gynecol 55:835–837

342. Li Y, Zhao R, Zheng S, et al (2020) Lack of Vertical Transmission of Severe Acute Respiratory Syndrome Coronavirus 2, China. Emerg Infect Dis 26:1335–1336

343. Goad J, Rudolph J, Rajkovic A (2020) Female reproductive tract has low concentration of SARS-CoV2 receptors. PLoS One 15:e0243959

344. Voort PHJ van der, van der Voort PHJ, Moser J, Zandstra DF, Muller Kobold AC, Knoester M, Calkhoven CF, Hamming I, van Meurs M A clinical and biological framework on the role of visceral fat tissue and leptin in SARS-CoV-2 infection related respiratory failure. https://doi.org/10.1101/2020.04.30.20086108

345. Heialy SA, Al Heialy S, Hachim M, Senok A, Tayoun AA, Hamoudi R, Alsheikh-Ali A, Hamid Q Regulation of angiotensin converting enzyme 2 (ACE2) in obesity: implications for COVID-19. https://doi.org/10.1101/2020.04.17.046938

346. Tay MZ, Poh CM, Rénia L, MacAry PA, Ng LFP (2020) The trinity of COVID-19: immunity, inflammation and intervention. Nat Rev Immunol 20:363–374

347. Kuri-Cervantes L, Pampena MB, Meng W, et al (2020) Comprehensive mapping of immune perturbations associated with severe COVID-19. Sci Immunol. https://doi.org/10.1126/sciimmunol.abd7114

348. Feng Z, Diao B, Wang R, et al The Novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Directly Decimates Human Spleens and Lymph Nodes. https://doi.org/10.1101/2020.03.27.20045427

349. Zhang H, Kang Z, Gong H, et al (2020) Digestive system is a potential route of COVID-19: an analysis of single-cell coexpression pattern of key proteins in viral entry process. Gut 69:1010–1018

350. Kim J-Y, Kim W-J, Kim H, Suk K, Lee W-H (2009) The Stimulation of CD147 Induces MMP-9 Expression through ERK and NF-kappaB in Macrophages: Implication for Atherosclerosis. Immune Netw 9:90–97

351. Smith JC, Sausville EL, Girish V, Yuan ML, Vasudevan A, John KM, Sheltzer JM (2020) Cigarette Smoke Exposure and Inflammatory Signaling Increase the Expression of the SARS-CoV-2 Receptor ACE2 in the Respiratory Tract. Dev Cell 53:514–529.e3

352. Rao S, Lau A, So H-C (2020) Exploring Diseases/Traits and Blood Proteins Causally Related to Expression of ACE2, the Putative Receptor of SARS-CoV-2: A Mendelian Randomization Analysis Highlights Tentative Relevance of Diabetes-Related Traits. Diabetes Care 43:1416–1426

353. Jacobs M, Van Eeckhoutte HP, Wijnant SRA, Janssens W, Joos GF, Brusselle GG, Bracke KR (2020) Increased expression of ACE2, the SARS-CoV-2 entry receptor, in alveolar and bronchial epithelium of smokers and COPD subjects. Eur Respir J. https://doi.org/10.1183/13993003.02378-2020

354. Ciaglia E, Vecchione C, Puca AA (2020) COVID-19 Infection and Circulating ACE2 Levels: Protective Role in Women and Children. Front Pediatr 8:206

355. Bénéteau-Burnat B, Baudin B, Morgant G, Baumann FC, Giboudeau J (1990) Serum angiotensin-converting enzyme in healthy and sarcoidotic children: comparison with the reference interval for adults. Clin Chem 36:344–346

356. Chen J, Jiang Q, Xia X, Liu K, Yu Z, Tao W, Gong W, Han J-DJ (2020) Individual variation of the SARS-CoV-2 receptor ACE2 gene expression and regulation. Aging Cell. https://doi.org/10.1111/acel.13168

357. Hamming I, Timens W, Bulthuis MLC, Lely AT, Navis GJ, van Goor H (2004) Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. J Pathol 203:631–637

358. Keller MJ, Kitsis EA, Arora S, Chen J-T, Agarwal S, Ross MJ, Tomer Y, Southern W (2020) Effect of Systemic Glucocorticoids on Mortality or Mechanical Ventilation in Patients With COVID-19. Journal of Hospital Medicine 15:489–493

359. Bojkova D, Klann K, Koch B, Widera M, Krause D, Ciesek S, Cinatl J, Münch C (2020) Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. Nature 583:469–472

360. Panda PK, Arul MN, Patel P, Verma SK, Luo W, Rubahn H-G, Mishra YK, Suar M, Ahuja R (2020) Structure-based drug designing and immunoinformatics approach for SARS-CoV-2. Sci Adv 6:eabb8097

361. Parks JM, Smith JC (2020) How to Discover Antiviral Drugs Quickly. N Engl J Med 382:2261–2264

362. Guy RK, Kiplin Guy R, DiPaola RS, Romanelli F, Dutch RE (2020) Rapid repurposing of drugs for COVID-19. Science 368:829–830

363. Levin JM, Oprea TI, Davidovich S, Clozel T, Overington JP, Vanhaelen Q, Cantor CR, Bischof E, Zhavoronkov A (2020) Artificial intelligence, drug repurposing and peer review. Nat Biotechnol 38:1127–1131

364. Blanco-Melo D, Nilsson-Payant BE, Liu W-C, et al (2020) Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. Cell 181:1036–1045.e9

365. Mangalmurti N, Hunter CA (2020) Cytokine Storms: Understanding COVID-19. Immunity 53:19–25

366. Zhong J, Tang J, Ye C, Dong L (2020) The immunology of COVID-19: is immune modulation an option for treatment? Lancet Rheumatol 2:e428–e436

367. Website. Storey J.D., B. A. J., Dabney A. and Robinson D. qvalue: Q-value estimation for false discovery rate control. R package version 2.20.0. http://github.com/jdstorey/qvalue. Github (2020).

368. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) 57:289–300

369. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550

370. Gupta R (2020) Advancing new tools for infectious diseases. Science 370:913–914

371. Mosteller F, Fisher RA (1948) Questions and Answers. The American Statistician 2:30

372. Wang K, Chen W, Zhou Y-S, et al SARS-CoV-2 invades host cells via a novel route: CD147-spike protein. https://doi.org/10.1101/2020.03.14.988345

373. Jiang Z, Hu S, Hua D, Ni J, Xu L, Ge Y, Zhou Y, Cheng Z, Wu S (2014) β3GnT8 plays an important role in CD147 signal transduction as an upstream modulator of MMP production in tumor cells. Oncol Rep 32:1156–1162

374. Xiong L, Edwards C, Zhou L (2014) The Biological Function and Clinical Utilization of CD147 in Human Diseases: A Review of the Current Scientific Literature. International Journal of Molecular Sciences 15:17411–17441

375. Kong L-M, Liao C-G, Zhang Y, Xu J, Li Y, Huang W, Zhang Y, Bian H, Chen Z-N (2014) A regulatory loop involving miR-22, Sp1, and c-Myc modulates CD147 expression in breast cancer invasion and metastasis. Cancer Res 74:3764–3778

376. Ke X, Fei F, Chen Y, Xu L, Zhang Z, Huang Q, Zhang H, Yang H, Chen Z, Xing J (2012) Hypoxia upregulates CD147 through a combined effect of HIF-1α and Sp1 to promote glycolysis and tumor progression in epithelial solid tumors. Carcinogenesis 33:1598–1607

377. Grass GD, Daniel Grass G, Toole BP (2016) How, with whom and when: an overview of CD147-mediated regulatory networks influencing matrix metalloproteinase activity. Bioscience Reports. https://doi.org/10.1042/bsr20150256

378. Rucci N, Millimaggi D, Mari M, Del Fattore A, Bologna M, Teti A, Angelucci A, Dolo V (2010) Receptor Activator of NF-κB Ligand Enhances Breast Cancer–Induced Osteolytic Lesions through Upregulation of Extracellular Matrix Metalloproteinase Inducer/CD147. Cancer Research 70:6150–6160

379. Ding P, Zhang X, Jin S, Duan B, Chu P, Zhang Y, Chen Z-N, Xia B, Song F (2017) CD147 functions as the signaling receptor for extracellular divalent copper in hepatocellular carcinoma cells. Oncotarget 8:51151–51163

380. Ulrich H, Pillat MM (2020) CD147 as a Target for COVID-19 Treatment: Suggested Effects of Azithromycin and Stem Cell Engagement. Stem Cell Reviews and Reports 16:434–440

381. Wang S-J, Cui H-Y, Liu Y-M, Zhao P, Zhang Y, Fu Z-G, Chen Z-N, Jiang J-L (2015) CD147 promotes Src-dependent activation of Rac1 signaling through STAT3/DOCK8 during the motility of hepatocellular carcinoma cells. Oncotarget 6:243–257

382. Kirk P, Wilson MC, Heddle C, Brown MH, Barclay AN, Halestrap AP (2000) CD147 is tightly associated with lactate transporters MCT1 and MCT4 and facilitates their cell surface expression. EMBO J 19:3896–3904

383. Seebacher NA, Stacy AE, Porter GM, Merlot AM (2019) Clinical development of targeted and immune based anti-cancer therapies. J Exp Clin Cancer Res 38:156

384. Yong H-Y, Koh M-S, Moon A (2009) The p38 MAPK inhibitors for the treatment of inflammatory diseases and cancer. Expert Opin Investig Drugs 18:1893–1905

385. Draghici S, Nguyen T-M, Sonna LA, et al (2021) COVID-19: disease pathways and gene expression changes predict methylprednisolone can improve outcome in severe cases. Bioinformatics. https://doi.org/10.1093/bioinformatics/btab163

386. Alaimo S, Giugno R, Pulvirenti A (2016) Recommendation Techniques for Drug-Target Interaction Prediction and Drug Repositioning. Methods Mol Biol 1415:441–462

387. Institute NC, National Cancer Institute (2020) Gene Ontology. Definitions. https://doi.org/10.32388/kp0fz4

388. Maria N, Rapicavoli RV, Alaimo S, Bischof E, Stasuzzo A, Broek J, Pulvirenti A, Mishra B, Duits A, Ferro A (2021) Rapid Identification of Druggable Targets and the Power of the PHENotype SIMulator for Effective Drug Repurposing in COVID-19. Res Sq. https://doi.org/10.21203/rs.3.rs-287183/v1

389. Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet-Cordero A, Sage J, Butte AJ (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Transl Med 3:96ra77

390. Schrezenmeier E, Dörner T (2020) Mechanisms of action of hydroxychloroquine and chloroquine: implications for rheumatology. Nat Rev Rheumatol 16:155–166

391. Lane JCE, Weaver J, Kostka K, et al (2020) Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study. Lancet Rheumatol 2:e698–e711

392. Group TRC, The RECOVERY Collaborative Group (2021) Dexamethasone in Hospitalized Patients with Covid-19. New England Journal of Medicine 384:693–704

393. Roschewski M, Lionakis MS, Sharman JP, et al (2020) Inhibition of Bruton tyrosine kinase in patients with severe COVID-19. Sci Immunol. https://doi.org/10.1126/sciimmunol.abd0110

394. Abrams RPM, Yasgar A, Teramoto T, et al (2020) Therapeutic candidates for the Zika virus identified by a high-throughput screen for Zika protease inhibitors. Proceedings of the National Academy of Sciences 117:31365–31375

395. Himmelstein DS, Baranzini SE Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. https://doi.org/10.1101/011569

396. Bramante CT, Ingraham NE, Murray TA, et al (2021) Metformin and risk of mortality in patients hospitalised with COVID-19: a retrospective cohort analysis. Lancet Healthy Longev 2:e34–e41

397. Peiffer-Smadja N, Rebeaud ME, Guihur A, Mahamat-Saleh Y, Fiolet T (2021) Hydroxychloroquine and COVID-19: a tale of populism and obscurantism. Lancet Infect Dis 21:e121

398. Terpos E, Ntanasis-Stathopoulos I, Elalamy I, Kastritis E, Sergentanis TN, Politou M, Psaltopoulou T, Gerotziafas G, Dimopoulos MA (2020) Hematological findings and complications of COVID-19. Am J Hematol 95:834–847

399. Naymagon L, Berwick S, Kessler A, Lancman G, Gidwani U, Troy K (2020) The emergence of methemoglobinemia amidst the COVID-19 pandemic. Am J Hematol 95:E196–E197

400. Faisal H, Bloom A, Gaber AO (2020) Unexplained Methemoglobinemia in Coronavirus Disease 2019: A Case Report. A A Pract 14:e01287

401. Pers Y-M, Padern G (2020) Revisiting the cardiovascular risk of hydroxychloroquine in RA. Nat Rev Rheumatol 16:671–672

402. Fiolet T, Guihur A, Rebeaud ME, Mulot M, Peiffer-Smadja N, Mahamat-Saleh Y (2021) Effect of hydroxychloroquine with or without azithromycin on the mortality of coronavirus disease 2019 (COVID-19) patients: a systematic review and meta-analysis. Clinical Microbiology and Infection 27:19–27

403. Cardin R (2020) Faculty Opinions recommendation of De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature. https://doi.org/10.3410/f.738636719.793578583

404. Cao L, Goreshnik I, Coventry B, et al (2020) De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. Science 370:426–431

405. Menéndez CA, Byléhn F, Perez-Lemus GR, Alvarado W, de Pablo JJ (2020) Molecular characterization of ebselen binding activity to SARS-CoV-2 main protease. Sci Adv. https://doi.org/10.1126/sciadv.abd0345

406. Jin Z, Du X, Xu Y, et al (2020) Structure of M from SARS-CoV-2 and discovery of its inhibitors. Nature 582:289–293

407. Kowalewski J, Ray A (2020) Predicting novel drugs for SARS-CoV-2 using machine learning from a >10 million chemical space. Heliyon 6:e04639

408. Riva L, Yuan S, Yin X, et al (2020) Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. Nature 586:113–119

409. Touret F, Gilles M, Barral K, Nougairède A, van Helden J, Decroly E, de Lamballerie X, Coutard B (2020) In vitro screening of a FDA approved chemical library reveals potential inhibitors of SARS-CoV-2 replication. Sci Rep 10:13093

410. Gordon DE, Jang GM, Bouhaddou M, et al (2020) A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 583:459–468

411. Calabrese LH, Winthrop K, Strand V, Yazdany J, Walter JE (2021) Type I interferon, anti-interferon antibodies, and COVID-19. Lancet Rheumatol 3:e246–e247

412. Bronte V, Ugel S, Tinazzi E, et al (2020) Baricitinib restrains the immune dysregulation in patients with severe COVID-19. J Clin Invest 130:6409–6416

413. Kalil AC, Patterson TF, Mehta AK, et al (2021) Baricitinib plus Remdesivir for Hospitalized Adults with Covid-19. N Engl J Med 384:795–807

414. Calabrese LH, Calabrese C (2021) Baricitinib and dexamethasone for hospitalized patients with COVID-19. Cleve Clin J Med. https://doi.org/10.3949/ccjm.88a.ccc073

415. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R (2020) CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat Protoc 15:1484–1506

416. Thoms M, Buschauer R, Ameismeier M, et al (2020) Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2. Science 369:1249–1255

417. Altmann DM, Boyton RJ (2020) SARS-CoV-2 T cell immunity: Specificity, function, durability, and role in protection. Sci Immunol. https://doi.org/10.1126/sciimmunol.abd6160

418. Yang L, Han Y, Nilsson-Payant BE, et al (2020) A Human Pluripotent Stem Cell-based Platform to Study SARS-CoV-2 Tropism and Model Virus Infection in Human Cells and Organoids. Cell Stem Cell 27:125–136.e7

419. Krausgruber T, Fortelny N, Fife-Gernedl V, et al (2020) Structural cells are key regulators of organ-specific immune responses. Nature 583:296–302

420. Daly JL, Simonetti B, Klein K, et al (2020) Neuropilin-1 is a host factor for SARS-CoV-2 infection. Science 370:861–865

421. Finkel Y, Mizrahi O, Nachshon A, et al (2021) The coding capacity of SARS-CoV-2. Nature 589:125–130

Appendix

i)

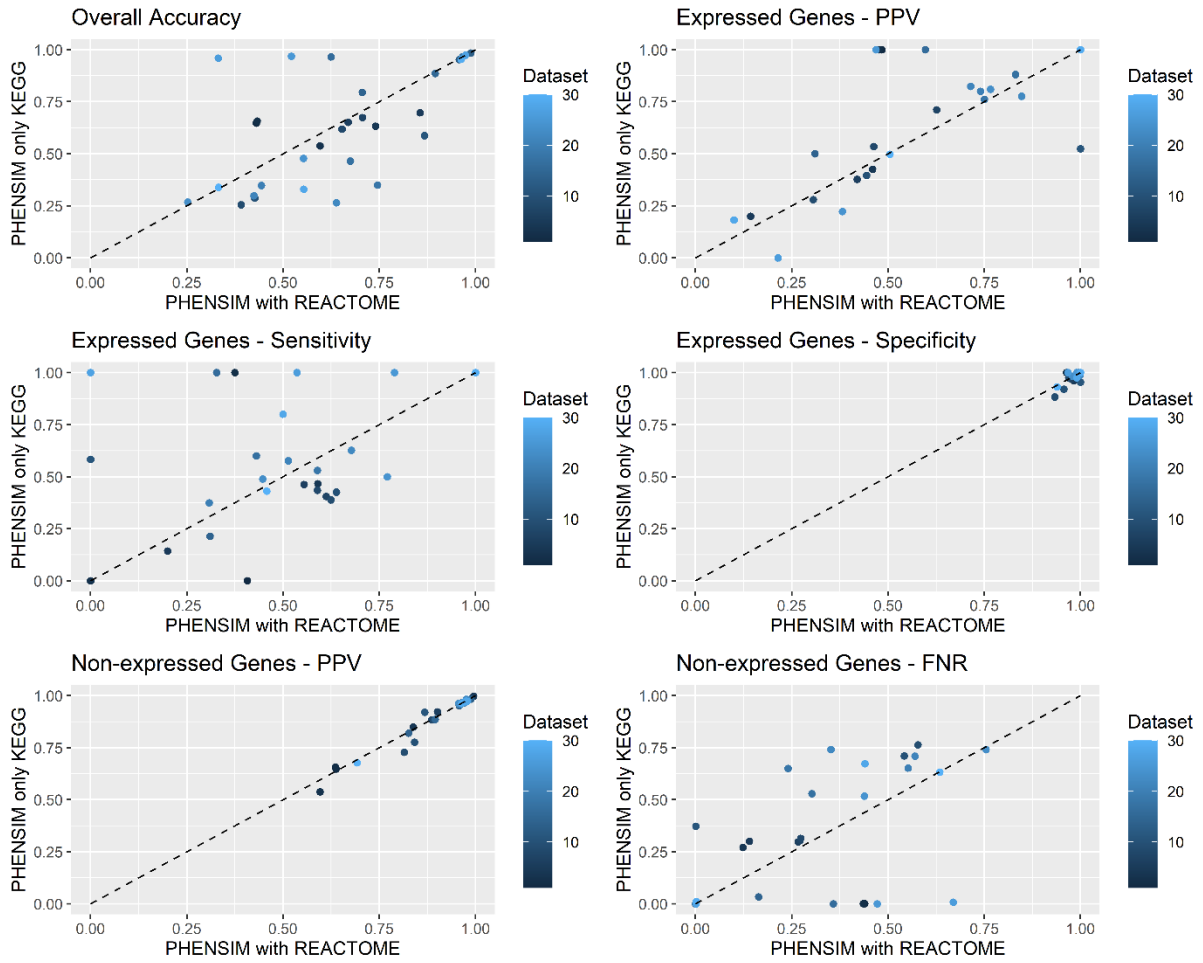


Figure S1 Comparison between PHENSIM with and without REACTOME for datasets where the altered gene belongs to the meta-pathway. Each graph reports one metric: Positive Predictive Value (PPV), Sensitivity and Specificity for genes showing altered expression, and PPV and False Negative Rate (FNR) for the non-altered ones. On the x-axis, we report PHENSIM performance with REACTOME, while on the y-axis, we present PHENSIM without REACTOME. Each dot is a dataset. The line marks the points where the two variants have the same performance.

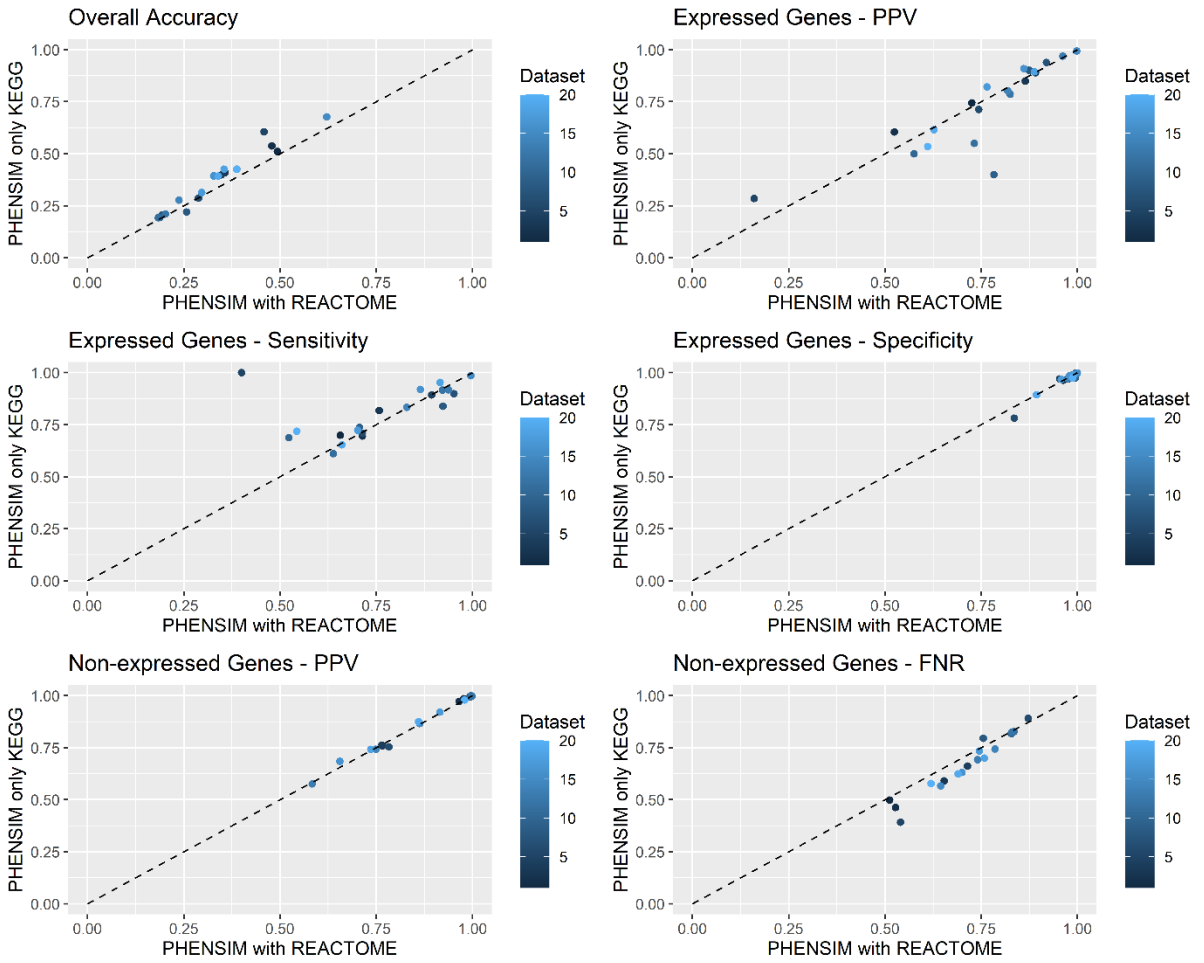


Figure S2 Comparison between PHENSIM with and without REACTOME for datasets where the altered gene was not in the meta-pathway. Each graph reports one metric: Positive Predictive Value (PPV), Sensitivity and Specificity for genes showing altered expression, and PPV and False Negative Rate (FNR) for the non-altered ones. On the x-axis, we report the PHENSIM performance with REACTOME, while on the y-axis, we have PHENSIM without REACTOME. Each dot is a dataset. The black line marks the points where the two algorithms have the same performance.

ii)

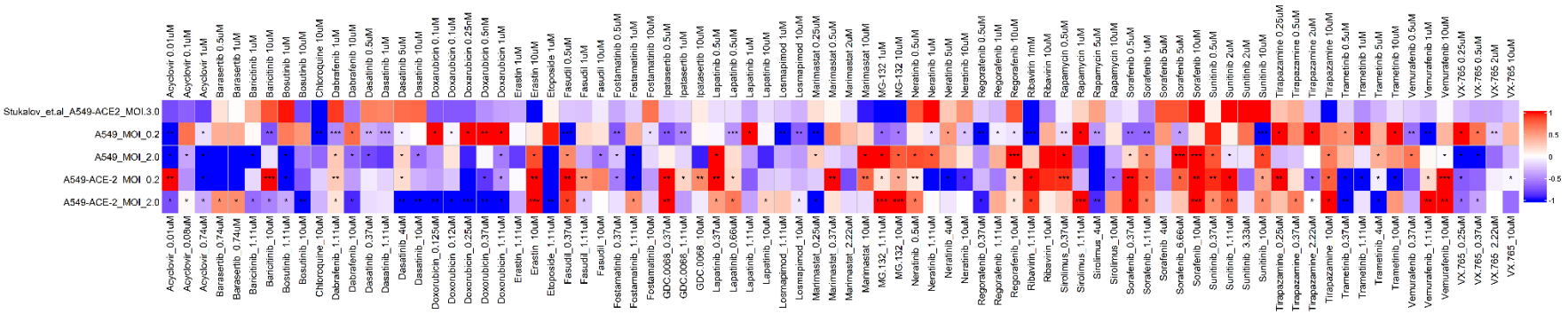


Figure S3 From in vitro to in silico. Comparison of drug treatments between Stukalov et al. 2021 and PHENSIM predictions. The heatmap depicts the validation outcomes for the PHENSIM repositioning approach. The test was done on all drugs tested by Stukalov et al. that are also present in the L1000 database at comparable concentrations. The L1000 drug signatures (DEGs caused by the drug effects) were gived as input to PHENSIM to built up our drug signatures. Then, Pearson correlations were performed between the PHENSIM drug- and viral- signatures obtained before.

Although the validation was satisfactory, it must be taken into account that the *in vitro* assays were performed on A549-ACE-2 cells whereas the data available on L1000 concern A549 cells. In addition, Pearson's correlation was performed using the viral model of A549 at both MOI 0.2 (see supplementary figure) and MOI 2.0 whereas Stukalov et al. infected cells with MOI 3.0.

Drugs (and concentrations) tested by Stukalov et al. are listed at the top of the heatmap, the corresponding L1000 drugs (and the nearest concentrations to the Stukalovo et al. ones) are listed at the bottom side. The colours in the heatmap represent the correlation values. In red are showed the drugs that are positively correlated with SARS-CoV-2 and in blue the negative correlated ones. Stars denote the adjusted p-values, $pv \leq 0.05$ * ; $pv \leq 0.01$ **; $pv \leq 0.001$ ***.

iii)

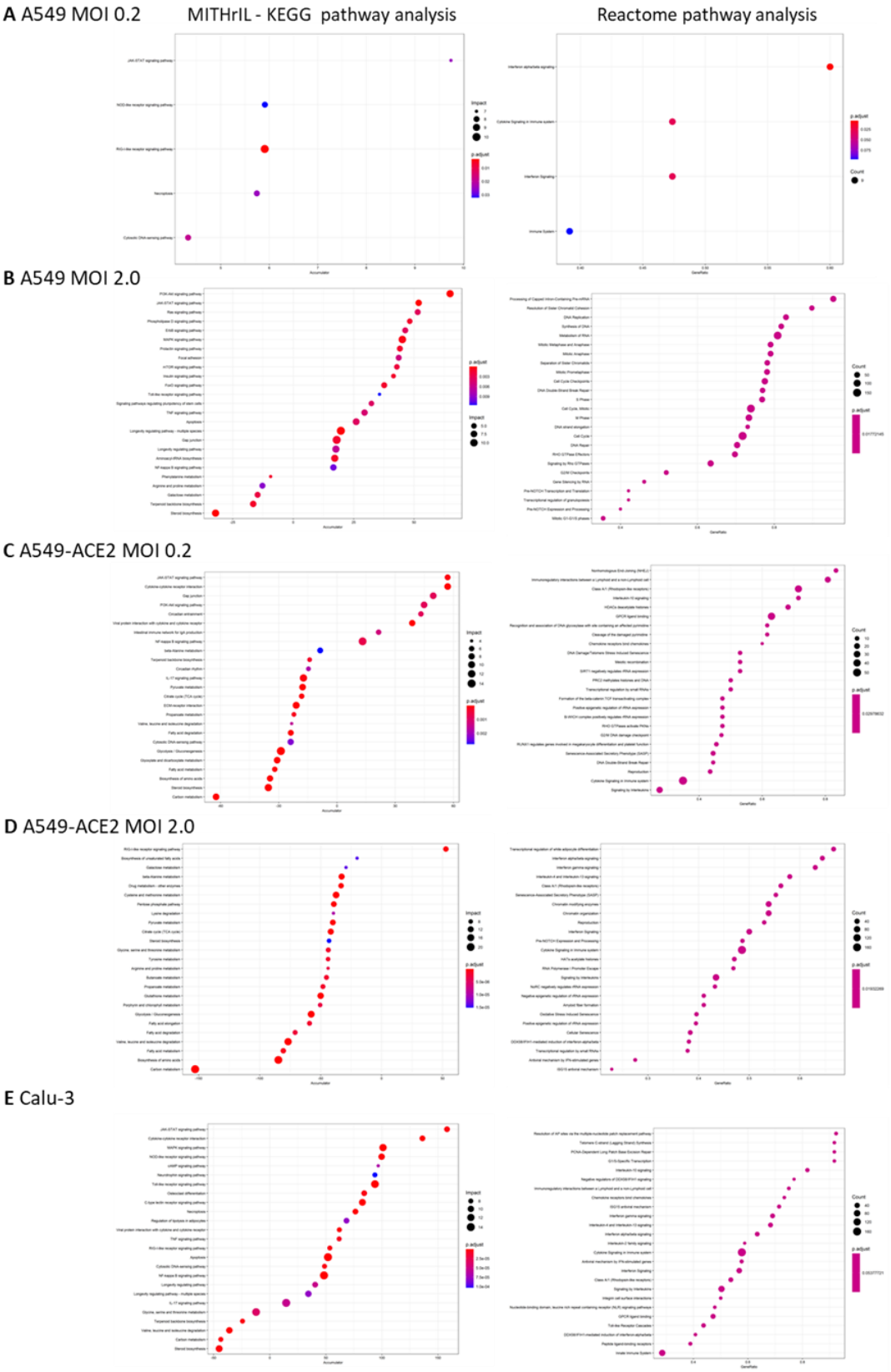


Figure S4. MITHrIL vs Reactome pathway analysis. Figure is related to Figure 31D&E. MITHrIL (using KEGG pathways; left) and Reactome pathway analysis (right) was used to assess top meta-pathways for A549 lung alveolar cells¬ +/- transduction with human ACE2 (A549-ACE2), at low and high multiplicity of infection (MOI), and in cultured human airway epithelial (Calu-3) cells. according to impact (circle size) and significance (color-gradient for adjusted p-value). The accumulator is the accumulation of all perturbations computed for that particular pathway.

iv)

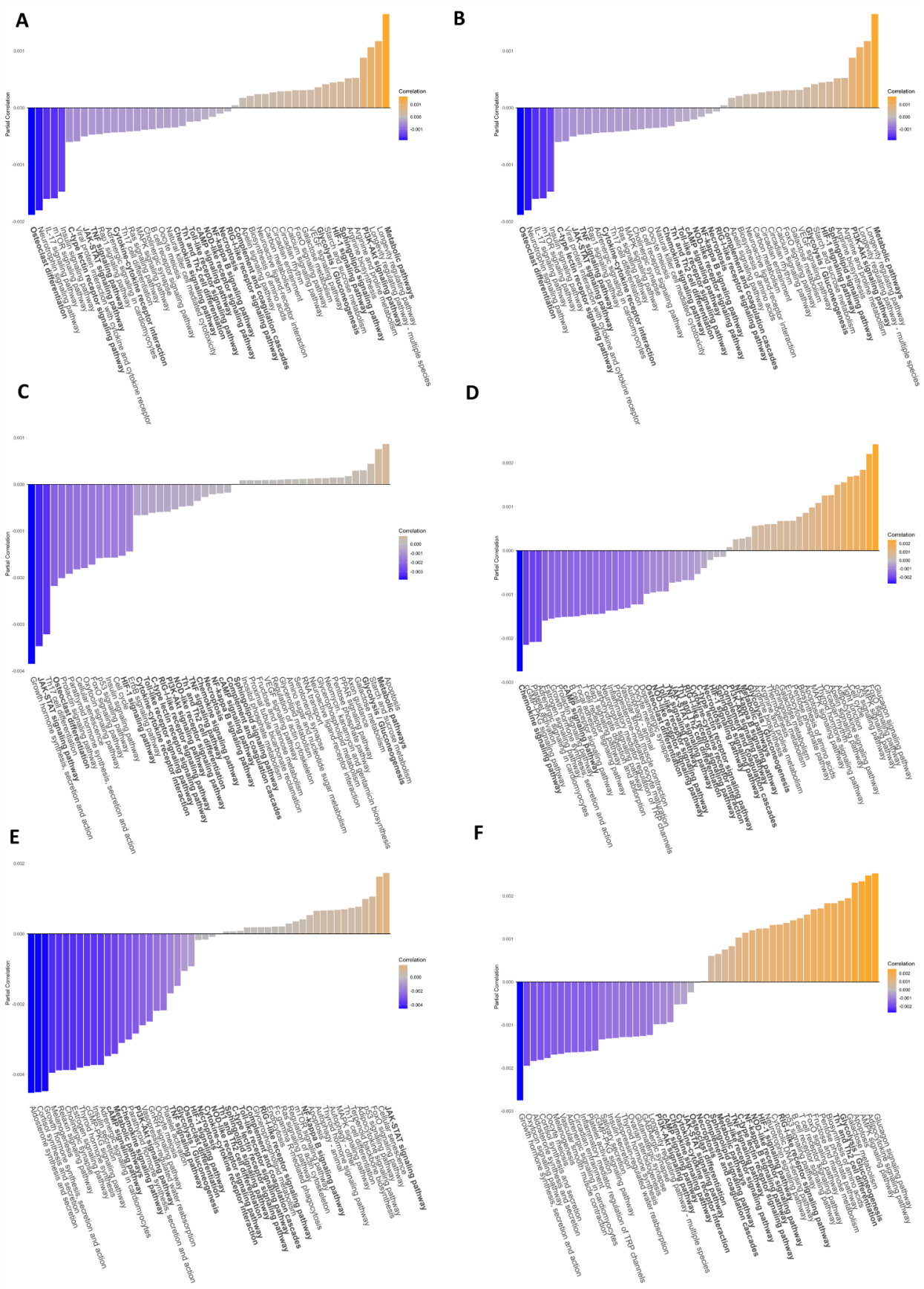


Figure S5 Resulted top pathways significantly affected by 2DG. A)NHBE; B) Calu-3; C)A549 MOI 0.2; D) A549 MOI 2.0 E) A549-ACE-2 MOI 0.2; F) A549-ACE-2 MOI 0.2.

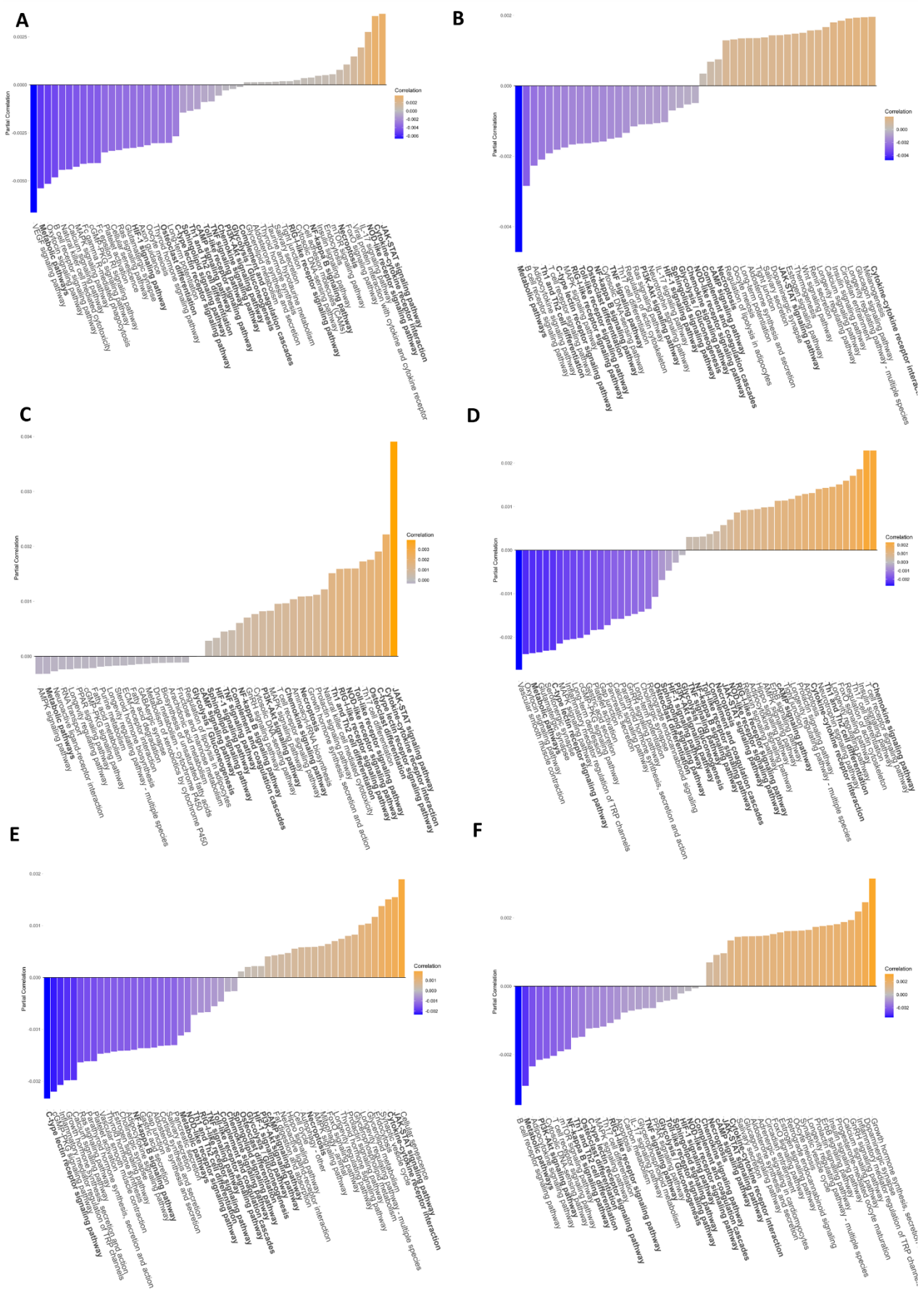


Figure S6 Resulted top pathways significantly affected by Acalabrutinib. A)NHBE; B) Calu-3; C)A549 MOI 0.2; D) A549 MOI 2.0 E) A549-ACE-2 MOI 0.2; F) A549-ACE-2 MOI 0.2.

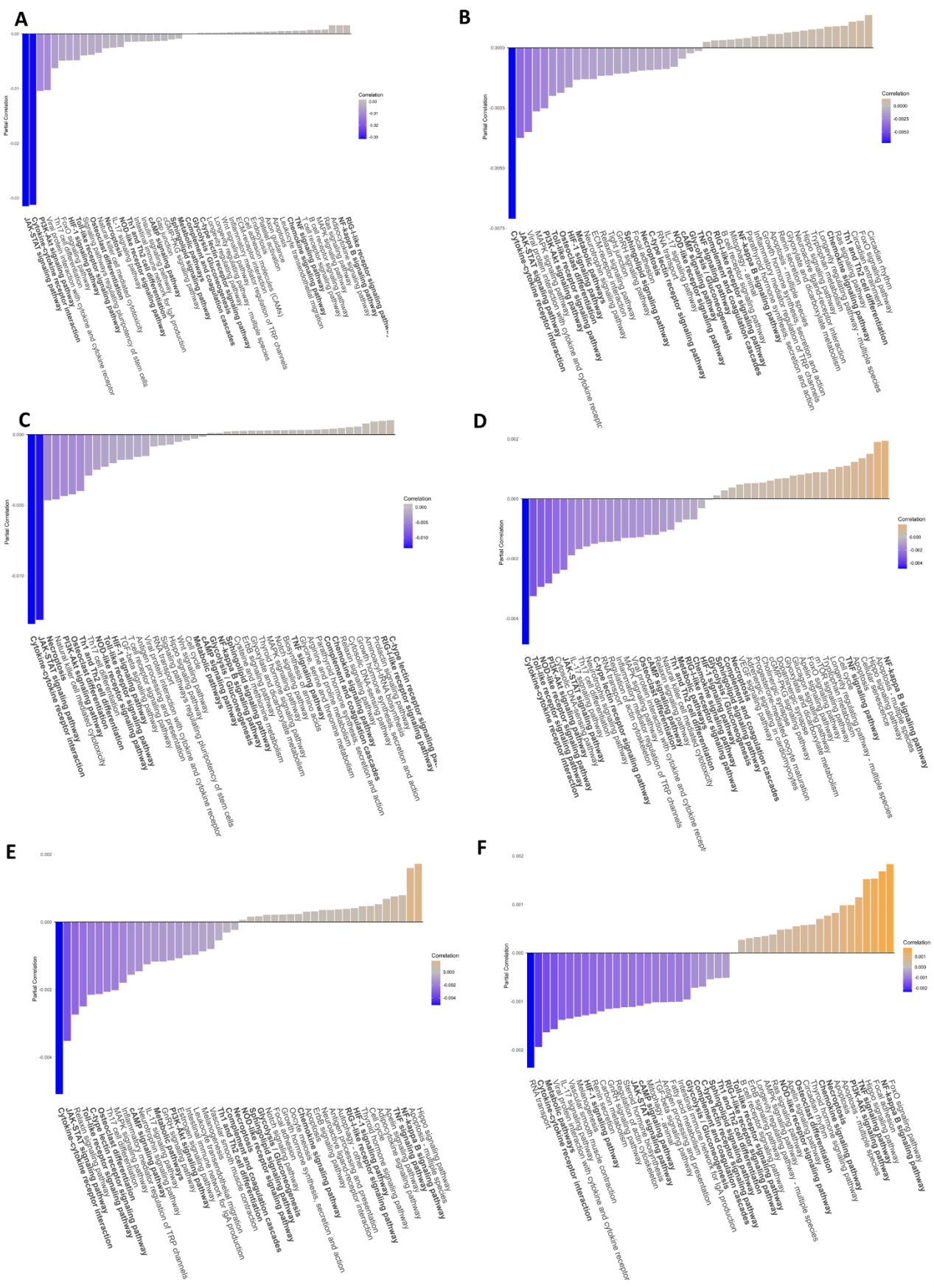


Figure S7 Resulted top pathways significantly affected by Dexamethasone. A)NHBE; B) Calu-3; C)A549 MOI 0.2; D) A549 MOI 2.0 E) A549-ACE-2 MOI 0.2; F) A549-ACE-2 MOI 0.2.

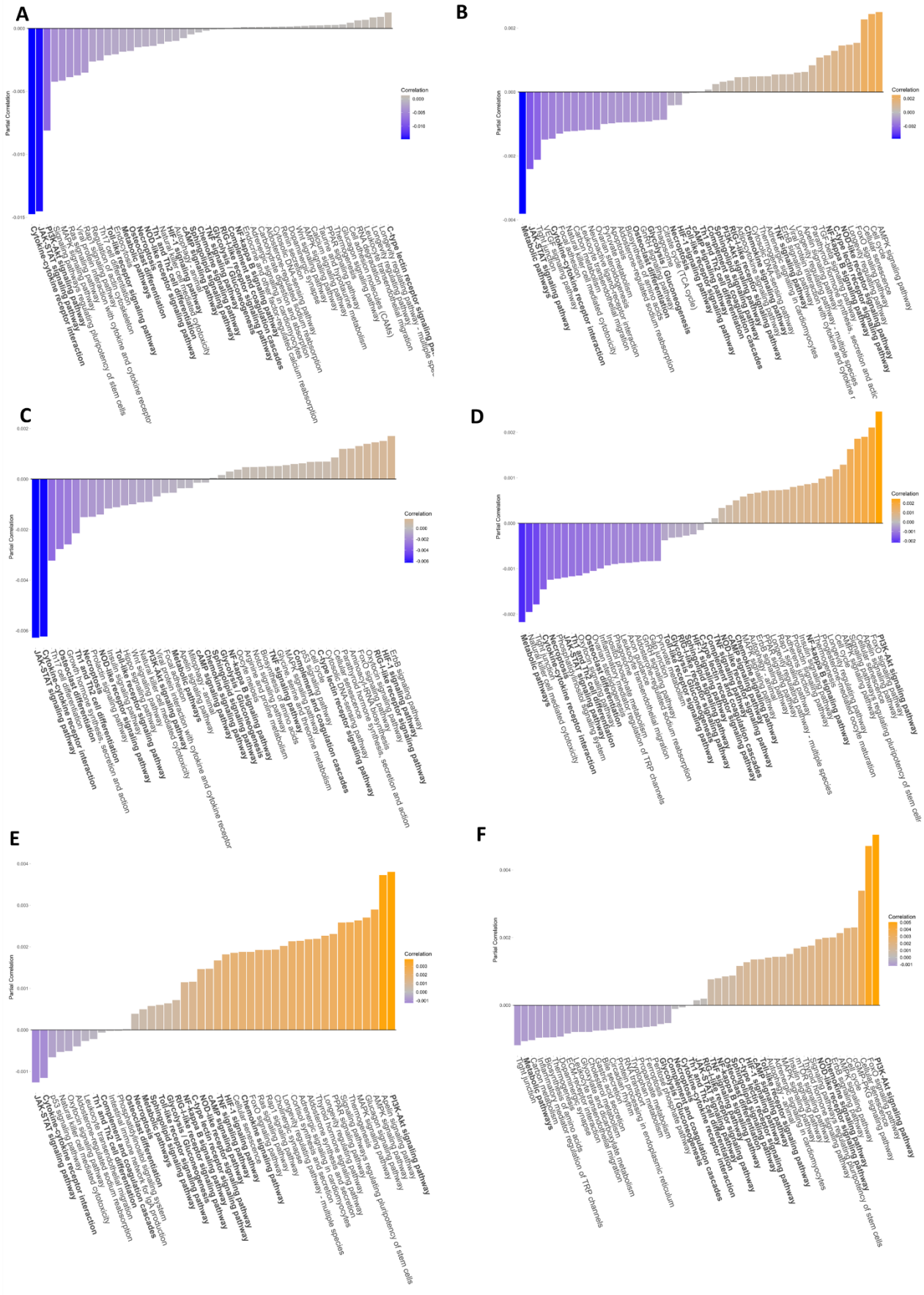


Figure S8 Resulted top pathways significantly affected by Everolimus. A)NHBE; B) Calu-3; C)A549 MOI 0.2; D) A549 MOI 2.0 E) A549-ACE-2 MOI 0.2; F) A549-ACE-2 MOI 0.2.

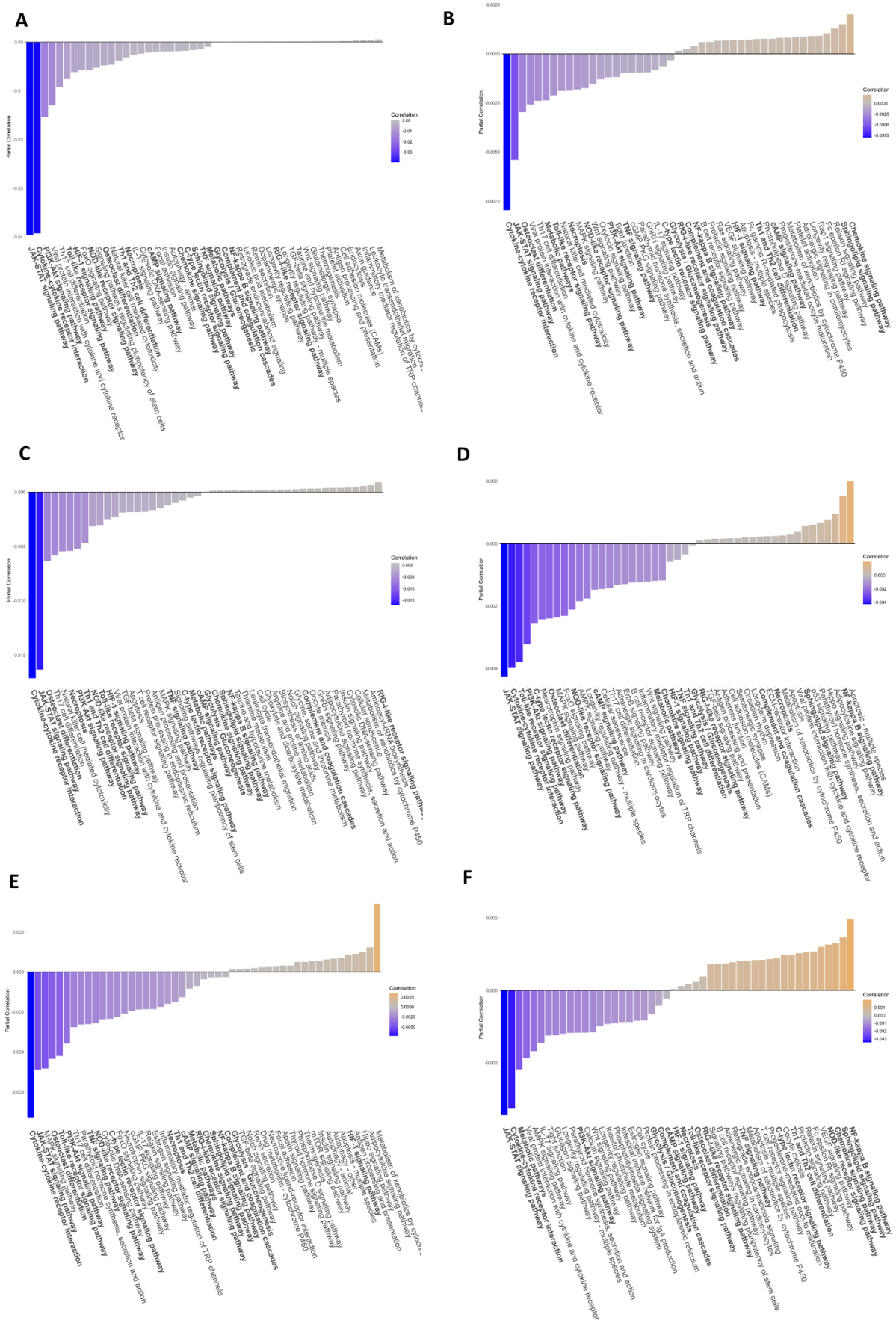


Figure S9 Resulted top pathways significantly affected by Hydroxychloroquine (HCQ). A)NHBE; B) Calu-3; C)A549 MOI 0.2; D) A549 MOI 2.0 E) A549-ACE-2 MOI 0.2; F) A549-ACE-2 MOI 0.2.

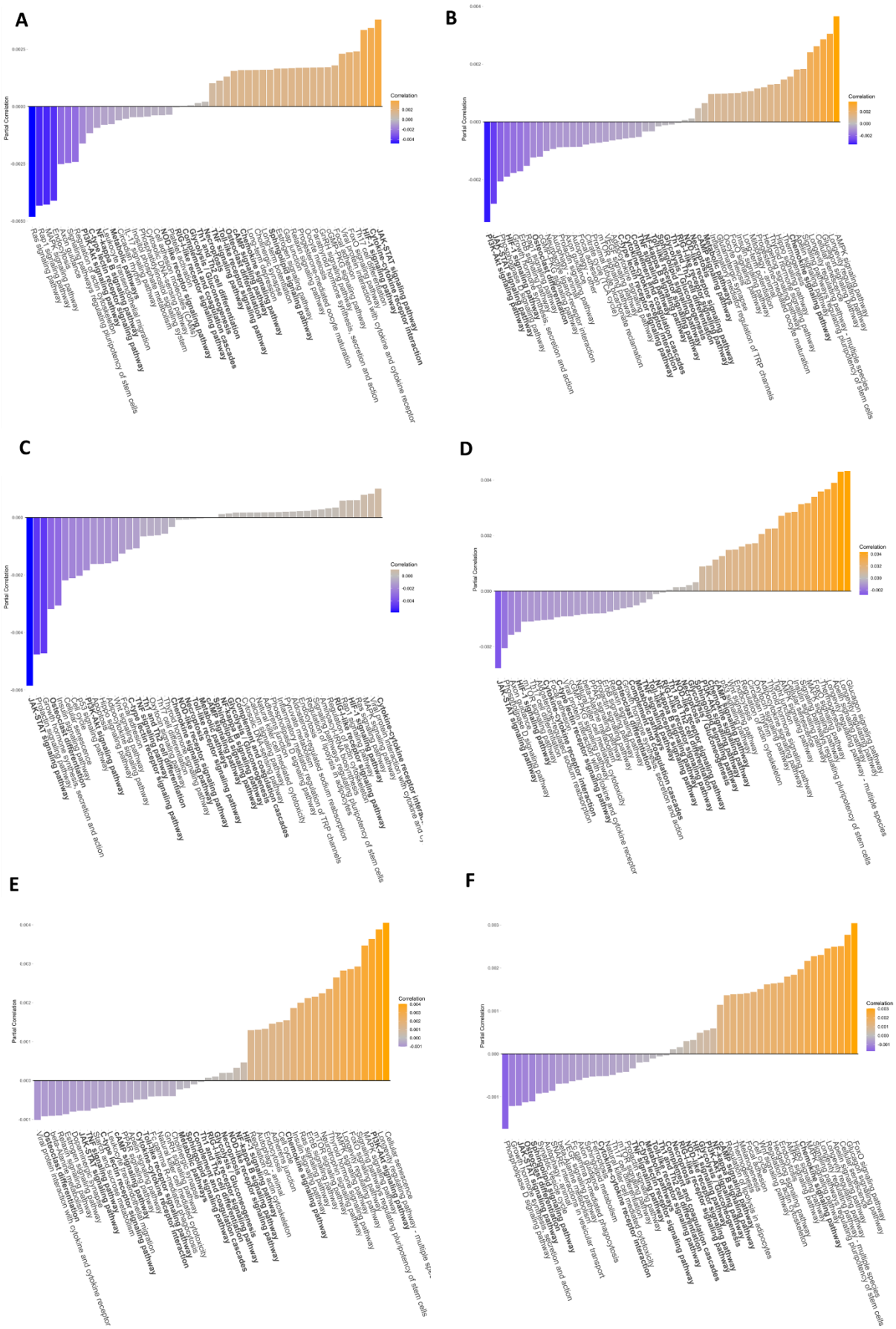


Figure S10 Resulted top pathways significantly affected by Metformin. A)NHBE; B) Calu-3; C)A549 MOI 0.2; D) A549 MOI 2.0 E) A549-ACE-2 MOI 0.2; F) A549-ACE-2 MOI 0.2.

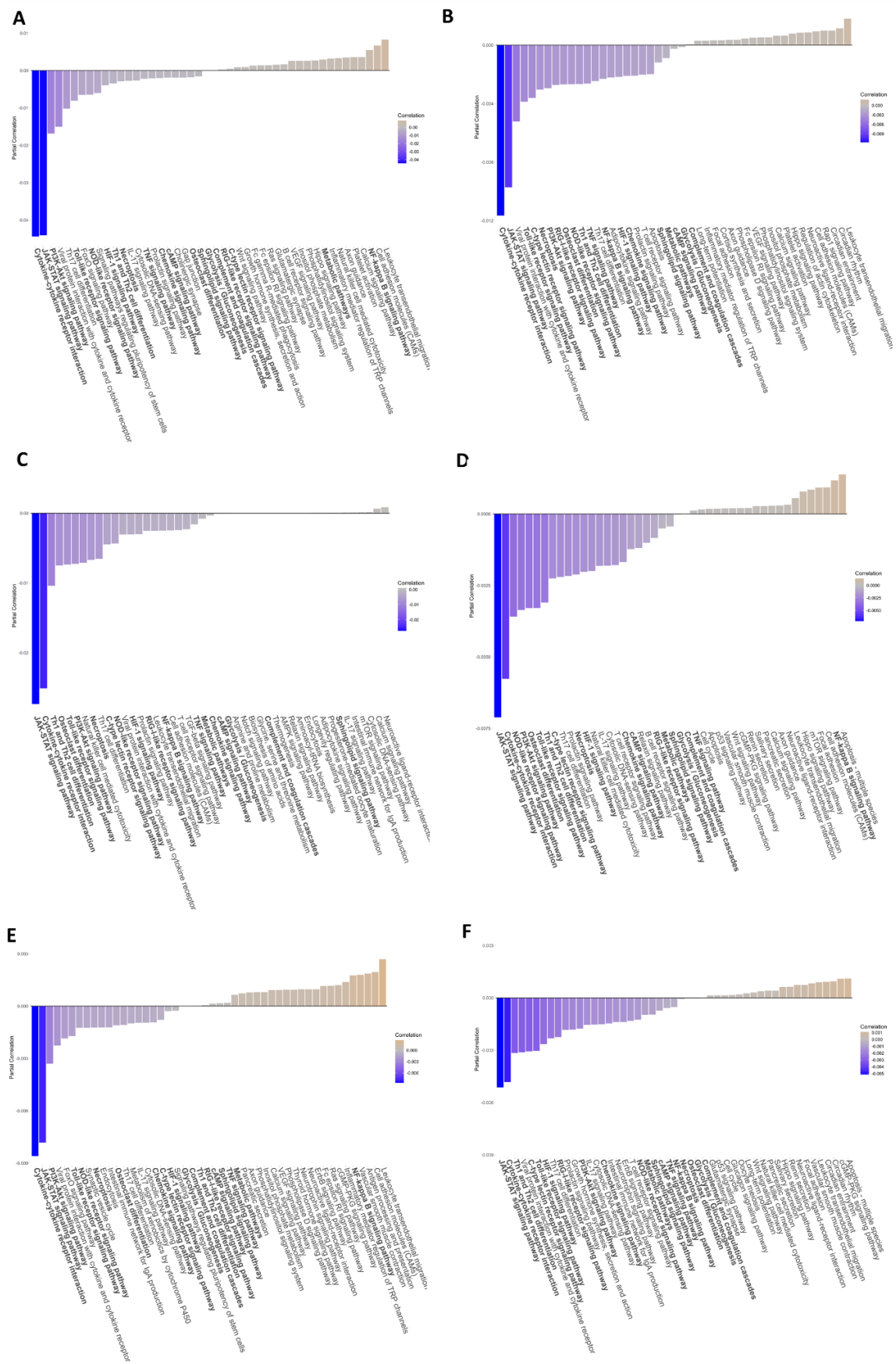


Figure S11 Resulted top pathways significantly affected by Methylprednisolon. A)NHBE; B) Calu-3; C)A549 MOI 0.2; D) A549 MOI 2.0 E) A549-ACE-2 MOI 0.2; F) A549-ACE-2 MOI 0.2.