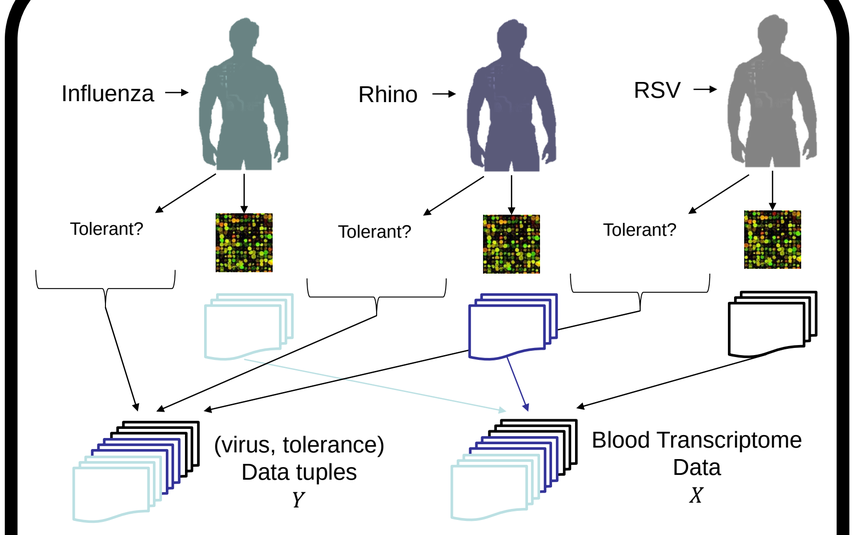
# Learning from the top: a hierarchical classification methodology for identifying tolerance biomarkers

Sahil Loomba[1], Diogo M. Camacho[1], James J. Collins[1,2]
[1]Wyss Institute for Biologically Inspired Engineering at Harvard University, [2]Massachusetts Institute of Technology

## Abstract

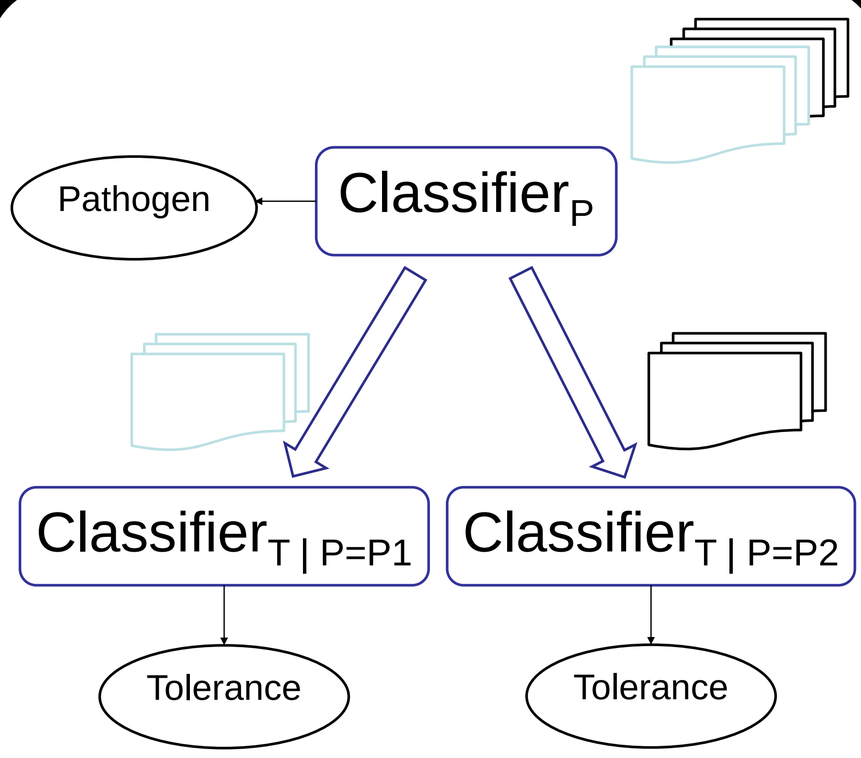
Pathogen infections, like the flu, can instigate different host immune responses and symptoms. The observation that some hosts show resilience or tolerance to pathogens leads us to identify alternative strategies for maintaining an individual's health, by comprehensively mapping and understanding what drives these mechanisms. Ahead of identifying biomarkers of tolerance from clinical studies, we make use of computational approaches to narrow down the search space of biology that is key to a tolerant host in response to a pathogen. We built a hierarchical meta-classifier on top of random forests, which are themselves an ensemble machine learning algorithm based on the decision tree classifier. We applied the method to transcriptomics data from humans infected with influenza H1N1, H3N2, rhinovirus, or respiratory syncytial virus (RSV) viruses in order to differentiate between states of tolerance and/or the infectious pathogen. Our results show that predicting tolerance conditioned on a prior pathogen classification improved overall accuracy compared to a standard classifier run on tolerance alone. Through this simple yet novel strategy, we uncover important transcripts in host-pathogen interactions that can drive a tolerant response, while allowing us to extract the similarities and differences between host responses across different pathogens. Because host response is a multifaceted process, involving a multitude of biomolecular networks, our future work will extend this approach to integrated multi-omics studies, while exploring different aspects of the classification problem.

## Problem Overview

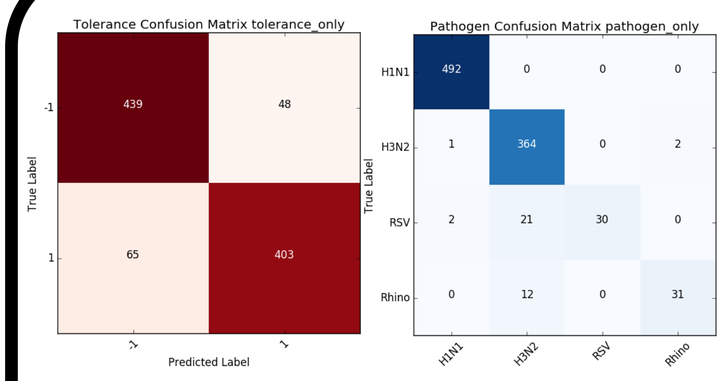



We aim to predict tolerance state from transcriptomics data, that is, find $f: X \rightarrow Y$
We define a meta-classifier to learn this function $f$

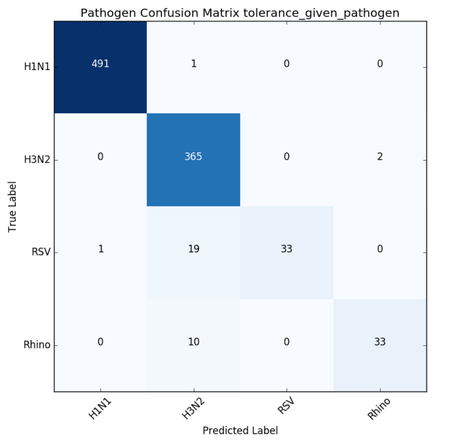
## Model Schematic




## Results




Confusion Matrix for Separate Tolerance and Pathogen Classifications




Confusion Matrix for Hierarchical Classification of Tolerance given Pathogen

We use one of the simplest classifiers, a single-output Random Forest, as the local classifier. Different strategies of constructing a hierarchical meta-classifier were tested, including soft-weighting of data samples for training, and conditioning pathogen on tolerance. We enclose accuracy results for tolerance conditioned on pathogen, which gave the best performance on a 5-cross-validated negative log loss score. This included comparisons to a multi-output Random Forest classifier, which learns a single model on the composite "tolerance-pathogen" classes.

Once we have a good classifier for tolerance, we can extract the top ranked features of this classifier, which serve as the key drivers of a tolerant response to pathogen infection. We enlist below the intersection of top-1000 features of all four level-2 classifiers; these are transcripts common to host response across the pathogens. Following which we use PANTHER for pathway enrichment, and look at key pathways that drive the tolerance mechanism.

## Key Transcripts

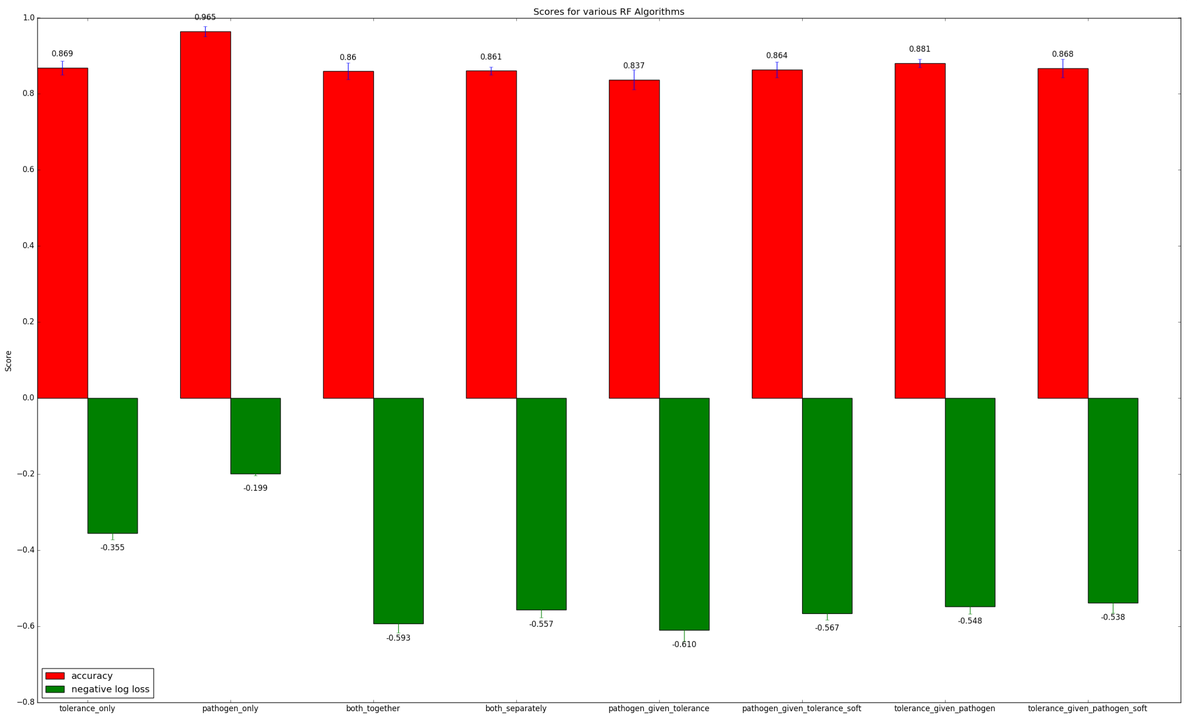
| Entrez ID | Function |
|---|---|
| 9693 | Rap guanine nucleotide exchange factor 2 |
| 2352 | Folate receptor |
| 9631 | Nucleoporin 155 |
| 5413 | Septin 5 |
| 2812 | Glycoprotein Ib platelet beta subunit |
| 3430 | Interferon induced protein 35 |
| 3431 | SP110 nuclear body protein |
| 6772 | Signal transducer and activator of transcription 1 |
| 53335 | B-cell CLL/lymphoma 11A |
| 115207 | Potassium channel tetramerization domain containing 12 |
| 6925 | Transcription factor 4 |
| 440026 | Transmembrane protein 41B |
| 3119 | Major histocompatibility complex, class II, DQ beta 1 |
| 23189 | KN motif and ankyrin repeat domains 1 |
| 26137 | Zinc finger and BTB domain containing 20 |
| 375346 | Transmembrane protein 110 |
| 9380 | Glyoxylate and hydroxypyruvate reductase |

## Enriched Pathways

| PANTHER Pathway | Function |
|---|---|
| JAK/STAT signaling pathway (P00038) | Activated by both cytokines and interferons; allows for rapid and direct transduction of an extracellular signal into the nucleus |
| Enkephalin release (P05913) | Opiod peptides that are found at high levels in the brain and endocrine tissues; play an important role in behavior, pain, cardiac function, cellular growth, immunity, and ischemic tolerance |
| Metabotropic glutamate receptor group I pathway (P00041) | Found in forebrain and cerebellum; its antagonists have antidepressant-like activity in a variety of preclinical models; potential drug targets for ischemia, schizophrenia, and epilepsy |
| Histamine H2 receptor mediated signaling pathway (P04386) | Activation results in many physiological responses: secretion of gastric juices, smooth muscle relaxation, inhibition of antibody synthesis, T-cell proliferation and cytokine production |
| Interferon-gamma signaling pathway (P00035) | IFNs are pleiotropic cytokines that mediate anti-viral responses, inhibit proliferation and participate in immune surveillance and tumor; IFN-gamma, that is produced by activated T cells and natural killer cells, activates JAK-STAT pathway |
| Nicotine pharmacodynamics pathway (P06587) | Nicotine causes cell depolarization and an influx of calcium through voltage dependent calcium channels, triggering the release of epinephrine from the chromaffin vesicles to the bloodstream, which leads to increase of heart rate and blood pressure, and elevation of blood glucose level |
| GABA-B receptor II signaling (P05731) | Stimulate the opening of $K^+$ channels which hyperpolarizes the neuron; considered inhibitory receptors and decrease the cell's conductance to $Ca^{2+}$ |
| EGF receptor signaling pathway (P00018) | Mediate cellular signaling pathways involved in growth and proliferation in response to the binding of a variety of growth factor ligands |
| CCKR signaling map (P06959) | Binding of gastrin (responsible for stimulation of acid secretion from the parietal cell) or CCK to their common cognate receptor triggers the activation of multiple signal transduction pathways that relay the mitogenic signal to the nucleus and promote cell proliferation |
| Cadherin signaling pathway (P00012) | Involved in many biological processes, such as development, neurogenesis, cell adhesion, and inflammation; implicated to be involved in many disease, such as cancer; cadherin-catenin complexes are important sensors and transmitters of the extracellular cues inside the cell body and into the nucleus |
| Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway (P00026) | G-protein receptor activated pathways; a number of activated receptors can bind to and activate the associated heterotrimeric G-protein consisting of either Gi alpha or Gs alpha |
| Inflammation mediated by chemokine and cytokine signaling pathway (P00031) | Upon binding to a family of G-protein coupled seven-transmembrane receptors, chemokines (chemotactic cytokines) control and direct trafficking and migration of immune cells |

## Future Work

The advantages of this meta-classifier is that it is very flexible. Any classifier can be used locally to classify on the output of a particular hierarchy level. Also, nothing stops us from creating very deep models, by incorporating more than just two outputs. For instance, a "severity score" that signifies the viral load could be a potential intermediate layer between pathogen and tolerance. Restricting tolerance to be the output for the last layer, one can define a greedy algorithm for constructing the upper layers of the classifier on the remainder outputs, based on the highest-scoring-output-first heuristic. The meta-classifier allows lower outputs to be learnt better by the virtue of highly-discriminating outputs in upper layers (see adjoining figure). We plan to incorporate more such output features, and expand the feature space to multi-omics data, for a larger number of pathogen families.




Accuracy and Negative-Log-Loss scores for various Random Forest Models; notice highest NLL score of tolerance_given_pathogen_soft method while predicting both pathogen and tolerance

## References

DATA SOURCE PAPER

## Acknowledgements

We would like to acknowledge contributions of Ivan Stojkovic and Zoran Obradovic at Temple University

## Funding

This work is supported by Wyss Institute for Biologically Inspired Engineering at Harvard University, and DARPA THoR 15-21