# Modelling Pathogen Response of the Human Immune System in a Reduced State Space

Pouria Tajvar[1], Rikard Forlin[2], Petter Brodin[2], and Dimos V. Dimarogonas[1]

***Abstract*— The immune system response to pathogens is organized by a network of cells communicating through expression of a variety of proteins and signaling molecules. A high number of genes are involved in encoding these communicating agents, but the relatively low number of data points is a major challenge in modelling the gene expression response. In this work we propose a feature-selection approach based on gene expression distributions at the single-cell level that improves dynamics identification at the population level. We investigate common approaches to differential expression analysis and show that Earth Mover's Distance (EMD) is a relatively robust measure for gene selection as reflected by the coefficient of variation as well as accuracy of a naive Bayes classifier based on the selected genes. We ultimately propose the bootstrap standard deviation metric as an estimate of state uncertainty and show that statistically significant signals in pathogen response can be recovered in the reduced state space constructed with the selected genes.**

## I. Introduction

Despite the general robustness of the immune response to a particular pathogen, there can be significant variability in its manifestation between individuals [1]. Understanding the dynamics of the coordinated immune response that explains this variability can be instrumental to identify susceptible individuals to certain pathogens or immune-mediated diseases and to guide vaccination strategies [2].

The study of the immune system dynamics is carried out at two primary levels: (i) intracellular gene networks and switching in cell states (such as [3]–[5]), and (ii) cell-cell interactions leading to adaptations analyzed at cell population levels (such as [6]–[9]). The timescale is another differentiating aspect among the studies, ranging from minutes (e.g. [3]) to hours and days (e.g. [10] and [11] respectively). Depending on the data-availability and time-steps, the dynamics are typically modelled as Boolean networks [12]–[14], Bayesian Networks [15], [16] or differential equations [3], [17], [18]. Tools from control theory and formal methods such as reachable-set and attractor computation can be used to integrate insights obtained from studying dynamical responses at different timescales. For example in [13], the steady state of the gene network is obtained as the attractor of the identified boolean network; this can be used in studying related signalling pathways with slower dynamics [19], [20].

A common challenge that is present both in single cell and population dynamics analysis across different timescales is the high number of measured features relative to the number of available sample points. To address this issue, various methods have been proposed for feature selection including a-priori filtering of features by differential expression assessment [21], [22], and simultaneous sparse optimization methods [23], [24]. The high number of genes measured may necessitate using both a-priori filtering and sparse optimization approach to minimize the rate of false discoveries (random artifacts that appear as statistically significant signals) [25], [26]. This problem is exacerbated in sparse nonlinear dynamics identification approaches as they rely on construction of nonlinear kernel libraries that further increase the state-space dimension [27], [28]. As a result, reliable a-priori filtering through differential expression assessment becomes even more critical.

Differential expression is most commonly measured as the difference between the logarithm of average gene expression between two sets of cells (Log); one reason for the popularity of this metric is that it can be applied using only average expression values which is available in *bulk measurement* methods. However, with the single-cell data becoming increasingly more accessible, other differential expression metrics such as earth mover's distance (EMD) and mutual information (MI) are also being investigated as alternatives. EMD specifically is shown to perform well in detecting changes in sub-populations and detection of sub-populations is essential in understanding pathogen responses as these responses are typically initiated by a small set of *early-responders* [29]. [30], [31] are two of the early works that showed that informative genes are reliably extracted using EMD between heterogenous cell classes.

In this work we formulate the dynamics identification problem as finding a set of discrete-time models where each model predicts one attribute such as expression level of a gene in Section II. The goal is to find one or more sets of attributes that are self-sufficient for prediction. Such attributes can be considered to belong to one signalling pathway. Following the discussion on the dimensinality challenge, in Section III we compare use of Log, EMD, and MI metrics for identifying significant gene expressions in pathogen response. We finally analyze the predictive power of sparse linear models in the reduced state space that is limited to the identified genes in Section IV. Throughout this paper, we are using the data-set published from [10]. The data-set comprises of samples collected from 120 individuals and exposed in-vitro to three pathogens, namely, C. Albicans (CA), M. Tuberculosis (MTB), and P. Aeruginosa (PA). The data includes sc-RNA assays of the untreated samples as well

[1]Division of Decision and Control Systems (DCS), KTH Royal Institute of Technology, tajvar@kth.se dimos@kth.se, [2]Department of Women's and Children's Health, Karolinska Institutet, rikard.forlin@ki.se petter.brodin@ki.se, * This work was supported by a WASP-DDLS grant from Knut and Alice Wallenberg foundation (KAW).

as samples after 3 hour and 24 hour of exposure to each pathogen. In total, expression of 23000 genes are measured in 1.1M cells.

## II. Problem Formulation

Let us denote the set of real numbers as $\mathbb{R}$, non-zero real numbers by $\mathbb{R}^+$, and the inclusive set of integers between $a$ and $b$ by $[\![a,b]\!]$. We use lower-case letters for scalar values, tuples, and functions, lower-case bold letters for vectors, and upper-case letters for sets. We use $\mathbf{v}_i$ to refer to the $i$-th element of a vector $\mathbf{v} \in \mathbb{R}^n$. Let $I$ be a set of $m(\leq n)$ unique integers from 1 to $n$, vector $v' = v_I$ is an $m$-dimensional sub-vector of $v \in \mathbb{R}^n$ with elements corresponding to the indices in $I$. Superscripts in parenthesis such as $c^{(i)}$ are used throughout this paper to refer to different instances of a variable. We use $\mathcal{N}(\mu,\sigma)$ to indicate a normal distribution with $\mu$ and $\sigma$ respectively being the mean and standard deviation. We use $\mathbb{P}$ to denote the probability distribution and $\mathbb{E}$ for the expected value of random variables.

### A. Single-cell State

The state in a cell is expressed as a tuple $c =< \mathbf{x}, ct >$ where $\mathbf{x} \in \mathbb{R}^{+n_g}$ is the gene expression vector of $n_g$ genes and $ct \in CT$ is the cell-type, e.g. Monocyte. In many scRNA sequencing protocols, the cells are destroyed during the measurement, so typically only one measurement of each cell is available and the dynamics can only be analyzed at the population level.

### B. Sample data

We consider samples that are collected from healthy individuals and exposed to pathogens *in-vitro*. One instance of a sample data is a tuple $s =< C, p, t >$ that includes states of $n_c$ cells within the sample $C = \{c^{(i)} | 1 \leq i \leq n_c, c^{(i)} =< \mathbf{x}^{(i)}, ct^{(i)} >\}$, the pathogen type $p \in P$, and the time $t$ since the sample is exposed to the pathogen.

*Remark 1:* The single-cell state can be expressed as a vector by assigning a numeric value to the cell-type; however, at the population level, the state cannot be similarly expressed due to the permutation invariance of the cells in the sample. We can however select a vector of permutation invariant attributes $\mathbf{s}$ to represent the state of the sample $s$. Such attributes include average expression of genes in the population, or the relative number of cells that express that belong to a certain cell-type.

### C. Modelling in reduced space

Let us define a $\omega$-sparse pathogen response model $m^{(i)} =< f^{(i)}, I^i, e^{(i)} >$ as follows:

$$\mathbf{s}_i(t+\Delta t) = f^{(i)}(\mathbf{s}_{I^{(i)}}(t)) + \mathcal{N}(0, e^{(i)}) \tag{1}$$

where: $dim(I^i) \leq \omega$. Namely, $f^{(i)}$ predicts one attribute $\mathbf{s}_i$ at time $(t+\Delta t)$ from a set of attributes $\mathbf{s}_{I^{(i)}}$ at time $t$ as inputs and the prediction error is expected to belong to the Gaussian distribution $\mathcal{N}(0, e_i)$. Enforcing sparsity helps to avoid over-fitting given the limited number of data points.

*Problem 1:* Given an error threshold $e^{(th)}$, find a set of models $M$ such that

1) for all $m^{(i)} \in M$, $e^{(i)} < e^{(th)}$ and,
2) The set $M$ is *closed* , i.e. any attribute that is used as an input for one of the models in $M$ can itself be predicted with a model in $M$: $\forall m^{(i)} \in M : j \in I^i \rightarrow m^{(j)} \in M$.

Intuitively a closed set of models $M$ enables predicting attributes for multiple time-steps and can be considered as hypothesis for a signalling pathway.

## III. Differential expression for gene selection

The differential expression of a gene is measured between two sets of cells. Each set can be defined as a subset of cells in a dataset of $n_s$ samples $D = \{s^{(i)} | 0 \leq i \leq n_s, s^{(i)} =< C^{(i)}, p^{(i)}, t^{(i)} >\}$ that belong to a certain class; For instance the class of untreated cells $L^{(UT)}$, i.e. cells that are not exposed to pathogens, can be defined as follows: $L^{(UT)} = \{c^{(i,j)} | c^{(i,j)} \in C^{(i)}, p^{(i)} = \emptyset, < C^{(i)}, p^{(i)}, t^{(i)} > \in D\}$

Where $c^{(i,j)}$ is the $j$-th cell in the $i$-th sample in the dataset.

There are three common metrics that are used for measuring differential expression between cells belonging to two classes $L^{(1)}$, and $L^{(2)}$; in the following sections we compare the use of each metric for identifying the significant genes in pathogen response.

*1) Log difference:* The most common method in differential expression is looking at the difference between the logarithm of the average expression of the $g$-th gene:

$$DE_g^{Log}(L^{(1)}, L^{(2)}) = \frac{|log(\mathbb{E}(\mathbf{x}_g \mid L^{(1)}) + \varepsilon) - log(\mathbb{E}(\mathbf{x}_g \mid L^{(2)}) + \varepsilon)|}{|log(\varepsilon)|} \tag{2}$$

where $\mathbb{E}(\mathbf{x}_g \mid L^{(1)})$ and $\mathbb{E}(\mathbf{x}_g \mid L^{(1)})$ correspond to the expected values of gene expression $\mathbf{x}_g$ in a cells that respectively belong to classes $L^{(1)}$ and $L^{(2)}$. The regulating parameter $\varepsilon << 1$ is to avoid unbounded value for genes that are expressed close to zero. Higher values of $\varepsilon$ put a higher weight on linear difference (rather than ratio) between expression levels. Log difference is a reliable method for detecting significant changes at the population level but is likely to miss changes in sub-populations of the cells [32].

### A. Earth Mover's distance

The earth mover's distance (EMD) evaluates the effort required to transform one probability distribution to another one, assuming that they can be represented as two piles of earth and that the effort is proportional to the amount of earth moved. For 1-dimensional probability distribution functions, the earth mover's distance can be calculated as follows:

$$DE_g^{EMD}(L^{(1)}, L^{(2)}) = \int_{-\infty}^{\infty} \left| \int_{-\infty}^{v} (\mathbb{P}(\mathbf{x}_g = v' \mid L^{(1)}) - \mathbb{P}(\mathbf{x}_g = v' \mid L^{(2)})) dv' \right| dv \tag{3}$$

where $\mathbb{P}(\mathbf{x}_g \mid L^{(1)})$ and $\mathbb{P}(\mathbf{x}_g \mid L^{(2)})$ are respectively the probability distributions of the $g$-th gene in cells from the classes $L^{(1)}$ and $L^{(2)}$. We adopt a binning approach to obtain a discrete approximation of these probability distributions.

### B. Mutual Information

The mutual information (MI) between two variables is defined as the Kullback–Leibler (KL) divergence between their joint probability and product probability (i.e. assuming they are independent). The differential expression can be measured as the MI between the class and the gene expression assuming that the classes are equally likely:

$$DE_g^{MI}(L^{(1)},L^{(2)}) = \frac{\mathbb{P}(L^{(1)})}{\mathbb{P}(L^{(1)})+\mathbb{P}(L^{(2)})} D_{KL}\Big(\mathbb{P}(\mathbf{x}_g \mid L^{(1)}), \mathbb{P}(\mathbf{x}_g)\Big) + \frac{\mathbb{P}(L^{(2)})}{\mathbb{P}(L^{(1)})+\mathbb{P}(L^{(2)})} D_{KL}\Big(\mathbb{P}(\mathbf{x}_g \mid L^{(2)}), \mathbb{P}(\mathbf{x}_g)\Big). \quad (4)$$

where $D_{KL}$ refers to the KL divergence. $\mathbb{P}(L^{(1)})$ and $\mathbb{P}(L^{(2)})$ are the probabilities that a cell belongs to each of the classes $L^{(1)}$ and $L^{(2)}$. Similarly, binning can be used to approximate $\mathbb{P}(\mathbf{x}_g \mid L^{(1)})$ and $\mathbb{P}(\mathbf{x}_g \mid L^{(2)})$, however authors in [33] have shown that MI can be calculated with higher accuracy based on nearest neighbour calculation; we have therefore adopted this method (implemented in [34]) instead of binning.

### C. Metric comparison

As seen in Fig. 1, both Log difference and Mutual Information are highly correlated with the EMD as expected, however the top genes vary a lot between the metrics. With infinite samples all of the intra-class expression differences are expected to be zero, therefore the intra-class points in Fig. 1 show the error introduced by sample limits and should be accounted for when considering differential expressions. For example, errors of up to 0.25 are expected when using the Log metric while the gene with the highest inter-class differential expression is 0.5 and a change of 0.25 could put it outside the top 100 differentially expressed genes. This issue appears to be less problematic with MI and EMD metrics as the intra-class differences remain below 0.1 and the inter-class differences reach 0.3 and 0.5 respectively.

*1) Bootstrapping:* To compare the robustness of metrics we bootstrap (resample cells with replacement) for a number of times to see which metric is more consistent in selecting top genes based on differential expression. As seen in Fig. 2, Earth mover's distance provides the lowest coefficient of variation in it's top 100 genes when the cells are selected from different distributions. This confirms EMD as the most robust DE metric for identifying significant genes in a pathogen response.

### D. Cell classification based on the identified genes

After selecting a set of genes $G^{(i,j)}$ that are differentially expressed between two classes $L^{(i)}$ and $L^{(j)}$, we can assess how accurately the class of a cell can be inferred only from the expression of genes in $G^{(i,j)}$. For multi-class classification between classes $L^{(1)}$ to $L^{(k)}$, we use the union of differentially expressed genes for pairwise classification: $G = \bigcup_{i\in[\![1,k]\!], j\in[\![1,k]\!], i<j} G^{(i,j)}$

For a cell $c =< \mathbf{x}, ct >$ we can compute the proportional probability of it belonging to each class $L^{(i)}$ $(i \in [\![1,k]\!])$ as follows: $\mathbb{P}(c \in L^{(i)} \mid \mathbf{x}) \propto \mathbb{P}(c \in L^{(i)}) \cdot \prod_{g\in G} \mathbb{P}(\mathbf{x}_g \mid c \in L^{(i)})$ with $\propto$ indicating proportional relation. We can now construct a naive Bayes classifier for $c$ as follows:

$$\hat{L} = \arg\max_{L^{(i)}} \mathbb{P}(c \in L^{(i)} \mid \mathbf{x}), i \in [\![1,k]\!] \quad (5)$$

In Fig. 3 we compare the classification accuracy of cells, i.e. ratio of cells that are assigned to the correct class. In all classification problems, same number of cells from each class is selected. Three classification problems are considered: The first problem is to identify whether a cell has been exposed to a pathogen, the second problem is to identify whether 3 hours or 24 hours has passed since a cell's exposure to pathogen, and the last problem is to identify the pathogen type (CA, PA, or MTB) that a cell is exposed to.

It can be seen that the same naive Bayesian classifier performs significantly more accurately for the same number of genes when they are selected using the EMD measure rather than Log difference. Classifying the cells based on the pathogen type is applied only to the cells after 24 hours of exposure and appears to be the most challenging classification. Pathogen types are indistinguishable to the naive Bayes classifier after 3 hours of exposure.

## IV. Modelling dynamics

Having identified the highly differentially expressed genes in the response we can now investigate modelling the response dynamics in a reduced state space that consists of the relative population of each cell type together with the expression levels of the identified genes. To construct models, we introduce the *bootstrap standard deviation* for a gene denoted as $e_i^{(bts)}$ to estimate measurement uncertainty and is computed for each gene by calculating the standard deviation of its average expression level in different cell bootstraps of the samples. Intuitively, genes that are only expressed in a small number of cells have higher $e_i^{(bts)}$ since a small change in the number cells that express the gene affects it's average expression significantly. We only consider modelling genes in which the inter-subject standard deviation $e_i^{(0)}$ is five times higher than $e_i^{(bts)}$. We should note in spite of pre-filtering the genes, the number of remaining features, i.e. identified genes, is still high as compared to the number of samples. We adopt the standard LASSO formulation to identify linear models $\mathbf{s}_i(t+\Delta t) = A^{(i)}\mathbf{s}_{I^{(i)}}(t) + \mathcal{N}(0, e^{(i)})$: $A^{(i)} = arg\min(||\mathbf{s}_i(t+\Delta t) - A^{(i)}\mathbf{s}_{I^{(i)}}(t)||_2 + \alpha||A||_1)$. We recall that this linear model is a special case of formulation (1).

It can be seen in Fig. 4 that enforcing sparsity results in models with higher testing accuracy and avoids over-fitting.

Let us now construct the set of models $M$ by defining the error threshold $e^{(th)}$ from Problem 1 as improving the prediction accuracy by at least 2 times the bootstrap standard deviation, namely $e^{(th)} = e^{(0)} - 2e^{(bts)}$. In Fig. 5 we show the set of models that satisfy $e^{(i)} \leq e^{(th)}$ in the 20 highest expressed genes in monocytes in the $\Delta t = 3h$ response to the CA pathogen. Each row corresponds to one model, i.e.

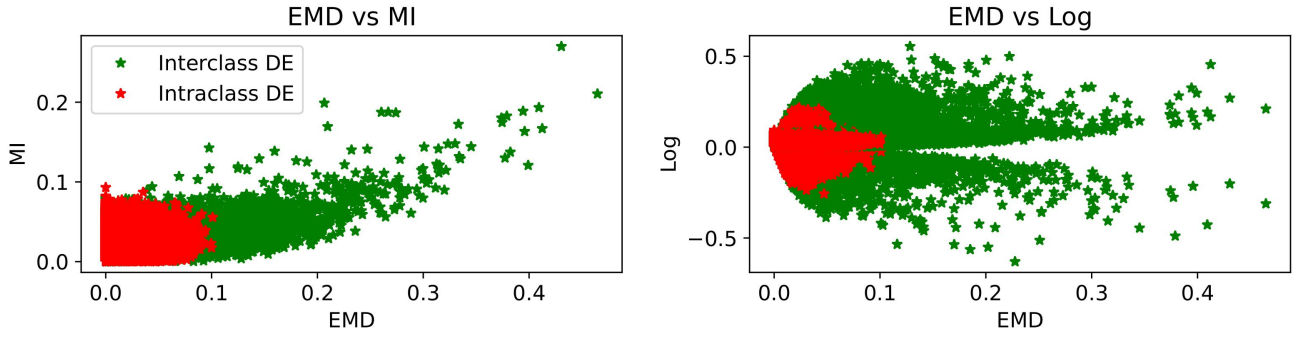


Fig. 1. Differential expression (DE) of genes based on three EMD, MI and Log metrics. The inter-class points show the difference in expression of each gene between untreated cells $L^{(UT)}$ (i.e. no pathogen) and cells that have been exposed to CA for 3 hours $L^{(3hCA)}$. The intra-class samples are where expression difference is measured between populations of cells that are selected with uniform probability from all cells. As the number of data-points approaches infinity, intraclass DE is expected to go zero for all metrics (i.e. all red points approach (0,0) in both figures.

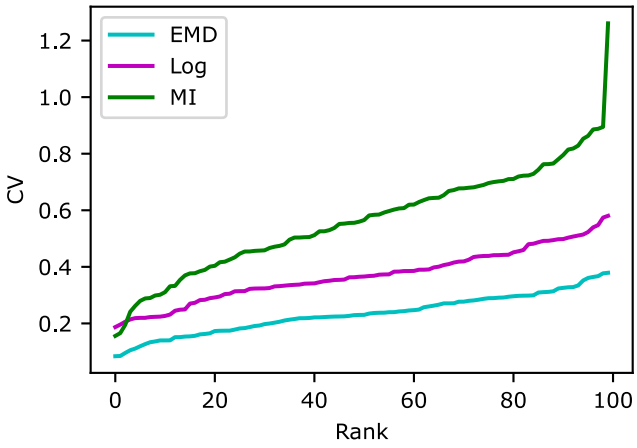


Fig. 2. Top 100 differentially expressed genes by each metric, sorted by their coefficient of variation in 10 bootstraps.

predicting one gene expression after 3 hours, and the non-zero columns in that row correspond to the genes that are required for this prediction.

We note that the requirement of model $M$ being closed in Problem 1 (i.e. condition 2) can also be expressed as: genes that correspond to columns of the matrix which contain non-zero values should be a subset of the genes corresponding to the rows of the matrix. We observe that this is not satisfied in the model shown in Fig. 5. This implies that even though we can predict the expression levels of a set of genes at 3 hours after exposure (i.e. one time step), we cannot predict their expression levels at 6 hours (i.e. two time steps) since the models rely on some genes that cannot be predicted at 3 hours with the required accuracy. Nevertheless, these models can provide insights to immune response pathways; for instance, DOCK4 and PTAFR are both associated with gamma interferon production [35], a cytokine known to be produced as a response to CA [36]. Fig. 5 suggests that PTAFR may be upstream of DOCK4 in this pathway. Such hypotheses can be used to inform gene knockout strategies that provide further insights on the underlying dynamics.

### A. Fast and slow responding genes

The over 90% classification accuracy between cells after 3h and cells after 24h hour exposure to pathogens seen in Fig. 3 suggests that there are genes that significantly change in expression levels within this time frame. Interestingly we observe that a number of these genes can be divided to (relatively) fast and slow reacting genes, suggesting that slower genes are further downstream in the signalling pathways. Fig. 6 shows an example of a fast reacting gene NCF1 and a slow reacting gene SLAMF8. At 3 hours after exposure NCF1 is highly down-regulated while SLAMF8 expression level is almost identical to the untreated cells; In contrast, after 24 hours, SLAMF8 is highly up-regulated while NCF1 has returned to the untreated expression level.

## V. Conclusion

In this paper we show that when single-cell data is available, using EMD is a more robust measure of differential expression in identifying pathogen response genes as compared to the more widely adopted log difference. We then propose bootstrapping standard deviation as a an approximation of sc-RNA measurement uncertainty. This measure can provide insight on the predictive power of dynamical models to reduce the rate of false hypotheses. We have shown that for certain highly differentially expressed genes, models with strong predictive power can be obtained; however, no subset of the obtained models satisfy the closed requirement expressed in Problem 1. This along with the observed fast responding genes suggests that gene expressions should be measured in shorter intervals to enable retrieving closed model sets corresponding to complete signalling pathways. Furthermore, we note that the obtained models serve as hypotheses regarding the signalling pathway dynamics that should be confirmed through gene knockouts or cell type depletion; nevertheless, we show that construction of the models in the reduced state space is instrumental in pruning weak hypotheses. In summary, we have shown that access to distributions and bootstrapping measures in the sc-RNA,

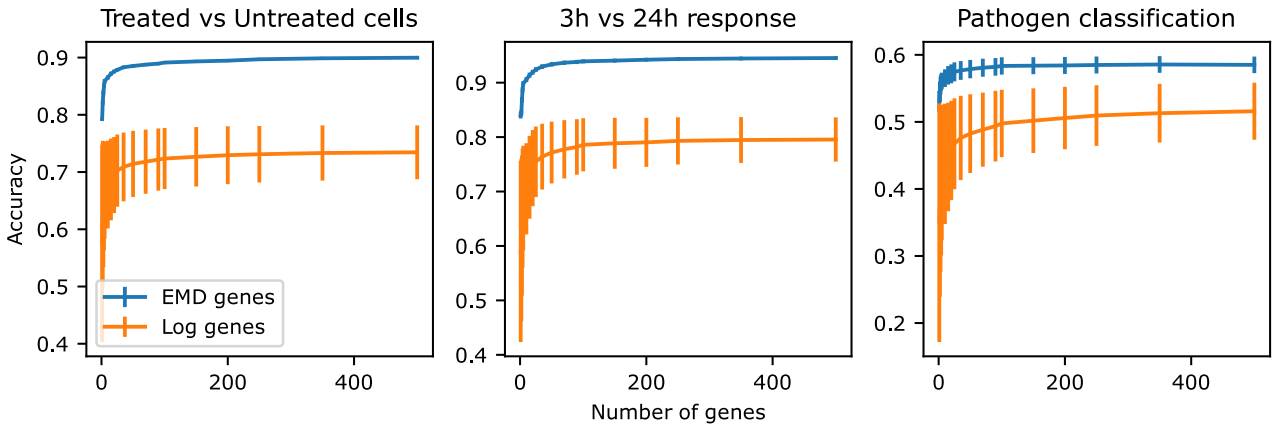


Fig. 3. Classification accuracy of naive Bayes classifiers ((5)). Selecting genes using the EMD measure results in higher classification accuracy.

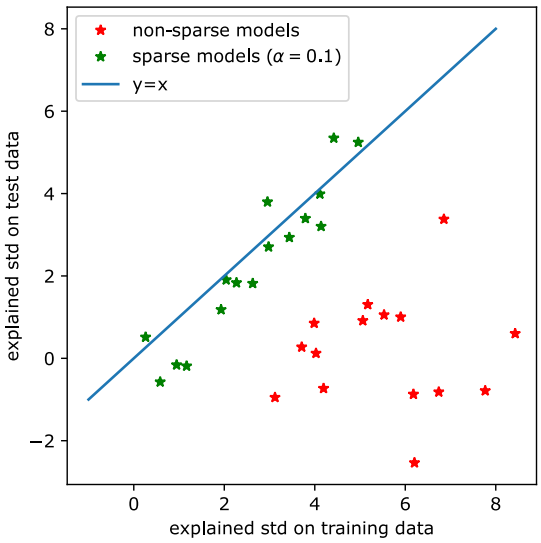


Fig. 4. Linear model identification of 3h response to the CA pathogen. Each point represents the model used to predict the expression of one gene. The explained std is measured in terms of bootstrap std.

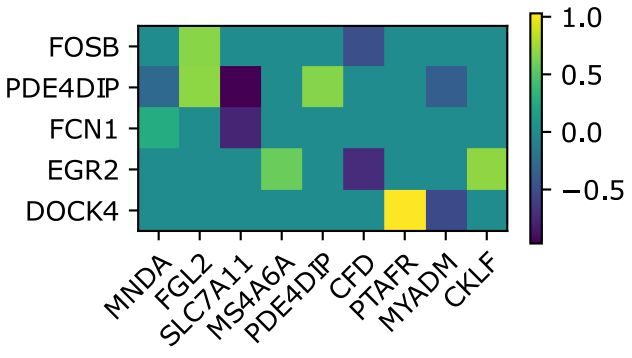


Fig. 5. Gene response models ($\Delta t = 3h$) to CA pathogen in monocytes.

enables identification of informative models regarding signalling pathway dynamics in the human immune system. We also note that restriction to linear models is a current limitation of our approach. We plan to investigate potential nonlinear formulations that still allow maintaining low false discovery rates, potentially by enforcing relations only between pairs of genes that are known to interact at the protein level. The methods presented in this paper are implemented in Python and can be accessed on https://github.com/KTH-DHSG/immune-system-pathogen-response.git.

## References


[1] P. Brodin and M. M. Davis, "Human immune system variation," *Nature reviews immunology*, vol. 17, no. 1, pp. 21–29, 2017.

[2] K. J. Kaczorowski, K. Shekhar, D. Nkulikiyimfura, C. L. Dekker, H. Maecker, M. M. Davis, A. K. Chakraborty, and P. Brodin, "Continuous immunotypes describe human immune variation and predict diverse responses," *Proceedings of the National Academy of Sciences*, vol. 114, no. 30, pp. E6097–E6106, 2017.

[3] V. A. Huynh-Thu and G. Sanguinetti, "Combining tree-based and dynamical systems for the inference of gene regulatory networks," *Bioinformatics*, vol. 31, no. 10, pp. 1614–1622, 2015.

[4] O. Zeyen and J. Pang, "Target control of boolean networks with permanent edgetic perturbations," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 4236–4243.

[5] J. Gebert, N. Radde, and G.-W. Weber, "Modeling gene regulatory networks with piecewise linear differential equations," *European Journal of Operational Research*, vol. 181, no. 3, pp. 1148–1165, 2007.

[6] L. You, R. S. Cox Iii, R. Weiss, and F. H. Arnold, "Programmed population control by cell–cell communication and regulated killing," *Nature*, vol. 428, no. 6985, pp. 868–871, 2004.

[7] A. Rivera, M. C. Siracusa, G. S. Yap, and W. C. Gause, "Innate cell communication kick-starts pathogen-specific immunity," *Nature immunology*, vol. 17, no. 4, pp. 356–363, 2016.

[8] E. Armingol, A. Officer, O. Harismendy, and N. E. Lewis, "Deciphering cell–cell interactions and communication from gene expression," *Nature Reviews Genetics*, vol. 22, no. 2, pp. 71–88, 2021.

[9] S. Naik and A. Mohammed, "Coexpression network analysis of human candida infection reveals key modules and hub genes responsible for host-pathogen interactions," *Frontiers in Genetics*, vol. 13, 2022.

[10] R. Oelen, D. H. de Vries, H. Brugge, M. G. Gordon, M. Vochteloo, single-cell eQTLGen consortium, B. Consortium, C. J. Ye, H.-J. Westra, L. Franke *et al.*, "Single-cell rna-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure," *Nature Communications*, vol. 13, no. 1, p. 3267, 2022.

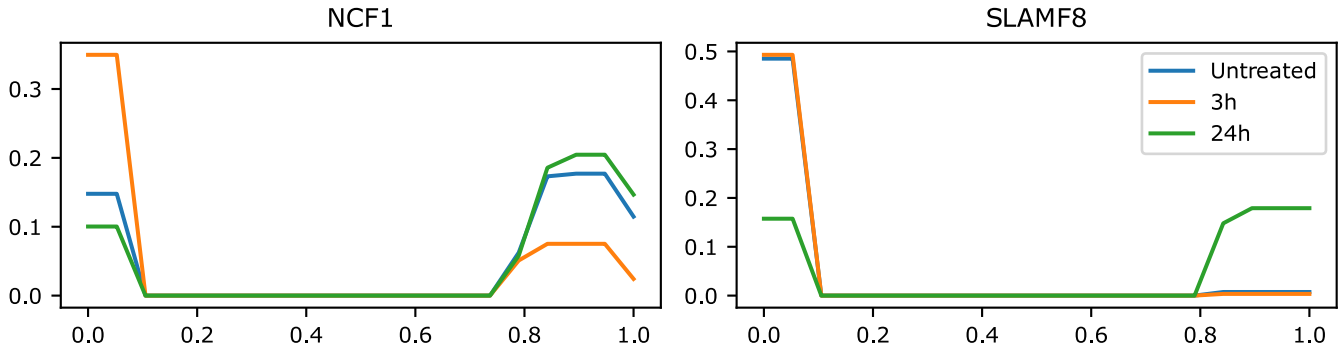


Fig. 6. Expression levels of two genes (NCF1 and SLAMF8) in untreated samples and after 3h and 24h of exposure to pathogens. The expression level (x-axis) is normalized using quantile transformation and the y-axis corresponds to the ratio of cells that are expressing the gene at a given expression level.


[11] X.-N. Zhao, Y. You, X.-M. Cui, H.-X. Gao, G.-L. Wang, S.-B. Zhang, L. Yao, L.-J. Duan, K.-L. Zhu, Y.-L. Wang *et al.*, "Single-cell immune profiling reveals distinct immune response in asymptomatic covid-19 patients," *Signal transduction and targeted therapy*, vol. 6, no. 1, p. 342, 2021.

[12] Z. Gao, "Stability analysis of breast cancer gene regulation network based on boolean network model," in *2022 International Conference on Computers, Information Processing and Advanced Education (CIPAE)*. IEEE, 2022, pp. 122–125.

[13] W. Abou-Jaoudé, P. Traynard, P. T. Monteiro, J. Saez-Rodriguez, T. Helikar, D. Thieffry, and C. Chaouiya, "Logical modeling and dynamical analysis of cellular networks," *Frontiers in genetics*, vol. 7, p. 94, 2016.

[14] R. Barbuti, R. Gori, P. Milazzo, and L. Nasti, "A survey of gene regulatory networks modelling methods: from differential equations, to boolean and qualitative bioinspired models," *Journal of Membrane Computing*, vol. 2, pp. 207–226, 2020.

[15] Y. Li, H. Chen, J. Zheng, and A. Ngom, "The max-min high-order dynamic bayesian network for learning gene regulatory networks with time-delayed regulations," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 13, no. 4, pp. 792–803, 2015.

[16] C. Wang, S. Xu, and Z.-P. Liu, "Evaluating gene regulatory network activity from dynamic expression data by regularized constraint programming," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 11, pp. 5738–5749, 2022.

[17] F. Parise, M. E. Valcher, and J. Lygeros, "Computing the projected reachable set of stochastic biochemical reaction networks modeled by switched affine systems," *IEEE Transactions on Automatic Control*, vol. 63, no. 11, pp. 3719–3734, 2018.

[18] M. Pasquini and D. Angeli, "On convergence for piecewise affine models of gene regulatory networks via a lyapunov approach," *IEEE Transactions on Automatic Control*, vol. 65, no. 8, pp. 3333–3348, 2019.

[19] B. Melykuti, E. August, A. Papachristodoulou, and H. El-Samad, "Discriminating between rival biochemical network models: three approaches to optimal experiment design," *BMC systems biology*, vol. 4, no. 1, pp. 1–16, 2010.

[20] J. Kuntz, D. Oyarzún, and G.-B. Stan, "Model reduction of genetic-metabolic networks via time scale separation," *A systems theoretic approach to systems and synthetic biology I: models and system characterizations*, pp. 181–210, 2014.

[21] F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci, and D. Betel, "Comprehensive evaluation of differential gene expression analysis methods for rna-seq data," *Genome biology*, vol. 14, no. 9, pp. 1–13, 2013.

[22] A. McDermaid, B. Monier, J. Zhao, B. Liu, and Q. Ma, "Interpretation of differential gene expression results of rna-seq data: review and integration," *Briefings in bioinformatics*, vol. 20, no. 6, pp. 2044–2054, 2019.

[23] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, pp. 389–422, 2002.

[24] P. García-Díaz, I. Sánchez-Berriel, J. A. Martínez-Rojas, and A. M. Diez-Pascual, "Unsupervised feature selection algorithm for multiclass cancer classification of gene expression rna-seq data," *Genomics*, vol. 112, no. 2, pp. 1916–1925, 2020.

[25] L. Liu and J. Liu, "Reconstructing gene regulatory networks via memetic algorithm and lasso based on recurrent neural networks," *Soft Computing*, vol. 24, pp. 4205–4221, 2020.

[26] M. Saint-Antoine, R. Grima, and A. Singh, "Fluctuation-based approaches to infer kinetics of cell-state switching," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 3878–3883.

[27] N. M. Mangan, S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Inferring biological networks by sparse identification of nonlinear dynamics," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 2, no. 1, pp. 52–63, 2016.

[28] A. B. Brummer, A. Xella, R. Woodall, V. Adhikarla, H. Cho, M. Gutova, C. E. Brown, and R. C. Rockne, "Data driven model discovery and interpretation for car t-cell killing using sparse identification and latent variables," *bioRxiv*, pp. 2022–09, 2022.

[29] L. C. Van Eyndhoven, V. P. Verberne, C. V. Bouten, A. Singh, and J. Tel, "Transiently heritable fates and quorum sensing drive early ifn-i response dynamics," *Elife*, vol. 12, p. e83055, 2023.

[30] S. Nabavi, D. Schmolze, M. Maitituoheti, S. Malladi, and A. H. Beck, "Emdomics: a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes," *Bioinformatics*, vol. 32, no. 4, pp. 533–541, 2016.

[31] S. Nabavi and A. H. Beck, "Earth mover's distance for differential analysis of heterogeneous genomics data," in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2015, pp. 963–966.

[32] C. A. Vallejos, S. Richardson, and J. C. Marioni, "Beyond comparisons of means: understanding changes in gene expression at the single-cell level," *Genome biology*, vol. 17, no. 1, pp. 1–14, 2016.

[33] B. C. Ross, "Mutual information between discrete and continuous data sets," *PloS one*, vol. 9, no. 2, p. e87357, 2014.

[34] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[35] G. Pètre, M. E. Atifi, A. Millet *et al.*, "Proteomic signature reveals modulation of human macrophage polarization and functions under differing environmental oxygen conditions," *Molecular & Cellular Proteomics*, vol. 16, no. 12, pp. 2153–2168, 2017.

[36] D. Gozalbo, V. Maneu, M. L. Gil *et al.*, "Role of ifn-gamma in immune responses to candida albicans infections," 2014.