Employing Feature Selection Algorithms to Determine the Immune State of Mice with Rheumatoid Arthritis

Brendon K. Colbert, Joslyn L. Mangal, Aleksandr Talitckii, Abhinav P. Acharya and Matthew M. Peet

Abstract

Background: The immune response is a dynamic process by which the body determines whether an antigen is self or nonself. The state of this dynamic process is defined by the relative balance and population of inflammatory and regulatory actors which comprise this decision making process. The goal of immunotherapy as applied to, e.g. Rheumatoid Arthritis (RA), then, is to bias the immune state in favor of the regulatory actors - thereby shutting down autoimmune pathways in the response. While there are several known approaches to immunotherapy, the effectiveness of the therapy will depend on how this intervention alters the evolution of this state. Unfortunately, this process is determined not only by the dynamics of the process, but the state of the system at the time of intervention - a state which is difficult if not impossible to determine prior to application of the therapy.

Results: We consider a mouse model of RA immunotherapy; collect high dimensional data on T cell markers and populations of mice after treatment with a recently developed immunotherapy for RA; and use feature selection algorithms to select a lower dimensional subset of this data which can be used to predict both the full set of T cell markers and populations, along with the progression of the immunotherapy treatment.

Conclusion: In this paper, we describe a process to identify a subset of immune cell populations that effectively represent the state of the immune response - which can measure the effectiveness of immunotherapy approaches to treating RA from measured T cell markers. We use a set of mouse-model experiments to obtain a robust dataset of T cell markers and populations and apply machine learning and feature selection techniques to determine specific T cells populations and models that predict the progression of RA, and can reconstruct any discarded features.

Keywords: Immune State; Immunotherapy; Feature Selection; Rheumatoid Arthritis; Flow Cytometry

Background

While a properly functioning immune system prevents illness by recognizing nonself antigens as foreign, a malfunctioning immune system can recognize self antigens as foreign, causing autoimmune diseases such as Rheumatoid Arthritis (RA). In recent years immune therapies have been proposed that attempt to treat autoimmune diseases such as RA by shifting the relative balance between inflammatory and regulatory immune responses in favor of the regulatory populations. For ex-
ample, sustained delivery of chemokines \cite{1, 2}, cytokines \cite{3, 4} and small molecule inhibitors \cite{5, 4} can modulate immune cell function (e.g. dendritic cells, T cells) in inflamed tissues to resolve RA and other autoimmune disease outcomes in pre-clinical animal models. However, the effect of the immunotherapy regimen is influenced by factors such as timing, dosage, and the current balance of inflammatory/regulatory response in the patient - thus making identification of effective treatment standards a challenging problem.

For this reason, there is a growing need for an observable measure of immune system health which can be used for the prediction and prevention of RA and other autoimmune diseases \cite{6, 7, 8}. However, the question of identifying observables is complicated by our relative lack of understanding of how the immune system determines self vs nonself and the number of potential observables which have been identified as contributing to the function of the immune system. To clarify the problem at hand, we therefore propose two relatively uncontroversial theses.

First, we presume that the question of identification of observables for prediction of autoimmune disease progression cannot be decoupled from the question of modeling, since in the absence of a predictive model, there is no way to verify that a certain set of observables can be used for prediction. That is, for any proposed set of observables, there must exist an associated predictive model with some associated accuracy in predicting autoimmune disease progression. Second, we presume that the immune system is deterministic in that the self-nonself decision (and hence autoimmune disease progression) is governed by a dynamical process wherein the relative populations of immunogenic and regulatory cells and molecules evolve over time and that the relative balance of these populations directly influences the establishment or elimination of autogenic response in autoimmune disease. That is, we presume that, given a method for modelling the immune system, there exists a set of observables capable of effectively predicting the process of self-nonself determination.

Given these assertions, we can propose three necessary components of any process for identification of observables with clinical predictive power. First, we require a method for modeling based on a given set of observables. While such a model may be based on physical principles, such a model may also be derived from data-based methods such as machine learning. Second, we require a way to test suitability of the predictive model associated with any given set of observables. Specifically, this test of suitability may include predictive accuracy of the associated model, along with other metrics such as clinical feasibility and robustness to patient variation. Finally, we require a methodology for selection and rejection of observables in order to obtain a set of observables with maximal suitability as defined previously. In this paper, we consider each of these requirements: using experimental data and a variety of machine learning algorithms to generate models; defining an appropriate metric for suitability; and using feature selection algorithms to find a set of observables with maximal suitability. Once we have addressed these requirements, we apply the proposed methodology - arriving at a set of maximally suitable observables, which we define as the “immune state”. An outline of our approach to addressing these required subproblems is listed below.

For the first problem we initially define our immunological dataset obtained from ongoing trials of RA immunotherapy. Then we define a set of machine learning
algorithms which use a given subset of data observables to identify outputs of interest such as other observables or the RA outcome - as measured by severity of inflammation.

For the second problem we propose a dual metric for suitability of a given set of observables based partially on predictive power of the associated model. The first part of this metric is based on minimality (not prediction), wherein we impose a penalty based on the number of observables in the set (cardinality) in order to reduce experimental and clinical complexity. Second, in order to ensure that relevant immunological data is not lost, we also add a penalty based on the error of the associated model to predict observables from the data not included in the given set. Third, to measure efficacy of the prediction, we impose a penalty based on the error in prediction of RA severity - a quantity we refer to as the “disease state”.

For the third problem we propose a variety of feature selection algorithms to determine the set of observables which are optimally suited using the suitability metric described above. We then report the results of applying the resulting algorithms to our dataset where we apply different weights to the three parts of the suitability metric and propose sets of maximally suitable observables for each case. We define the optimal sets as “immune states” and analyze the immune cells that were selected by the feature selection algorithms in each case.

**Methods**

**A mouse model of RA and associated observables**

We propose a methodology for identifying observable measures for immune system health from measured features of the immune system. To better illustrate this methodology, we consider the approach as applied to a particularly rich dataset obtained from an ongoing series of experiments involving the use of biomaterials-based particles [9] containing metabolites that promote self tolerance in intermediate/late stage RA in a DBA/1j mouse model which develops severe arthritis when immunized with bovine collagen type 2 (bc2) autoantigen.

Six to eight week old male DBA/1j mice were obtained from Jackson Laboratories and, after one to two weeks of mice acclimating to the experimental location, mice were induced with RA. In this experimental series, the particles were synthesized either with or without autoantigen bc2 - a strategy designed to determine if the particles can generate antigen-specific anti-inflammatory response. The number of mice per group were determined using a statistical power of 80 percent and a significance level of alpha of 0.05. The arthritic scores were utilized to randomize mice into the control and treatment groups to assure that the overall average arthritic scores were comparable between each group. Researchers were aware of the group allocation throughout the study. An overview of the experimental procedure is provided in Fig. 1 and is further described in [10]. The chronology of the experiment is listed here in detail. The data collection used for model generation occurs exclusively on either day 62 or 70.

**Day 0 and 21:** RA was induced in mice to generate an autoimmune response for the development of severe polyarthritis. On day 35, the mice were divided into 3 groups, each receiving a distinct therapeutic regimen.
Group 0 - Days 35/42: The control group consists of 5 control mice, each receiving two subcutaneous injections of phosphate buffered saline (PBS) near the hind legs on days 35 and 42.

Group 1 - Days 35/42: Treatment group 1 consists of 5 mice. Each mouse receives two injections of 0.5 mg of biomaterials-based particles without embedded autoantigen bc2 near the hind legs on days 35 and 42.

Group 2 - Days 35/42: Treatment group 2 consists of 8 mice. Each mouse receives two injections of 0.5 mg of biomaterials-based particles with embedded autoantigen bc2 near the hind legs on days 35 and 42.

Measurements Taken on Days 62/70: Paw thickness measurements are used to determine arthritic scores for all mice and the end of study paw measurements were obtained either on day 62 or 70, and are defined on the interval [0,5]. Mice were euthanized by carbon dioxide asphyxiation according to the American Veterinary Medical Association (AVMA) guidelines and flow cytometry was performed on cells collected from the popliteal lymph node, cervical lymph node and spleen of each mouse on day 62 or 70. The flow cytometry procedure stained for CD4 (T helper (Th) cell marker), CD8 (cytotoxic T (Tc) cell marker), Ki67 (proliferation), CD25 (activation), Foxp3 (regulatory T (Treg) cell transcription factor (TF)), Tbet (Th1/Tc1 TF), GATA3 (Th2/Tc2 TF), RORyT (Th17/Tc17 TF), CD44 (effector memory marker), CD62L (central memory marker), and a tetramer that is specific to the autoantigen. Based on this staining, we identified 41 different combinations of markers which might be used to classify the phenotype of a T cell and determined the percentage of either CD4 or CD8 T cells presenting the associated combination of markers.

Summary of Associated Dataset: The data consists of 84 samples, based on 18 mice, each sample is associated with a mouse and sample location. There were no exclusions of mice, experimental units or data points. All samples are taken on day 62/70, and each sample consists of 43 features and one label. The first two features of each sample indicate group number (0-2) and sample location (1-3). The remaining 41 features defining the percentage (0-100) of the CD4/CD8 population exhibiting the associated combination of markers. The label for each sample is the arthritic score (0-5).

Based on this data we next propose several methods of machine learning to construct predictive models which use subsets of the features to predict both the label (disease progression) and remaining features. For generating these models, all features are scaled to the interval [0,1].

Predictive Model Generation via Machine Learning Algorithms
To identify clinically significant observables, we will use a metric of suitability combined with a feature selection algorithm to determine which observables have the most predictive power. However, the use of such feature selection algorithms requires a procedure for using a subset of the features to predict both the remaining features and the label.
Figure 1 A graphical description of the experimental procedure of inducing and treating RA in mice. The first two steps induce RA, the next two steps is the application of the treatment and the final step is the data generation using flow cytometry. CFA = complete Freund’s adjuvant, IFA = incomplete Freund’s adjuvant.

Suppose we are given a dataset of \( m \) samples, wherein each sample \( \{x_i, y_i\} \) defines a set of features \( \{x_i \in \mathbb{R}^n\}_{i=1}^m \) and an associated label \( \{y_i \in \mathbb{R}\}_{i=1}^m \). The regression problem, then, is to find a predictive model, \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) which minimizes the predictive errors \( f(x_i) - y_i \) in an appropriately defined metric. However, this metric and the resulting optimization problems vary significantly between algorithms. We next define several state-of-the-art machine learning algorithms which will be combined with feature selection algorithms to determine features with the most predictive power. Finally, we note that in the context of feature selection algorithms, when only a subset of the available features are used, the remaining “discarded” features become labels.

Before beginning, we note that the choice and tuning of ML algorithms is something more of an art than a science. Specifically, we want to avoid overfitting the training data - thus allowing our predictive models to perform well on unlabelled data. To this end, each of the ML algorithms we define has an associated set of “regularization parameters” which should be selected through some ad hoc process. These tuning parameters will then affect how well the resulting predictive model will generalize to unlabelled data. In each case, therefore, we specify these parameters but do not yet define how they are selected.

In each case below, we assume the data set contains \( m \) samples, \( \{x_i, y_i\}_{i=1}^m \), each with \( n \) features, \( x_i \in \mathbb{R}^n \) and a label \( y_i \in \mathbb{R} \).

**Regularized Linear Regression (LR)** The regularized linear regression algorithm returns a predictive model \( y = f(x) = w^T x + b \), where \( w \) solves the following optimization problem.

\[
\min_{w \in \mathbb{R}^n} \sum_{i=1}^m (y_i - w^T x_i - b)^2 + \alpha_1\|w\|^2 + \alpha_2\|w\|.
\]

In this case, \( \alpha_1 \geq 0 \) and \( \alpha_2 \geq 0 \) are the regularization parameters. Linear regression has the advantage of low computational complexity. However, the resulting predictor is linear and if the underlying physical process is nonlinear, accuracy of the predictive model will be poor.

**\( \epsilon \)-loss Support Vector Regression (SVR)** The support vector regression problem uses a predictive model of the form \( f(x) = \sum_{i=1}^m \alpha_i k(x, x_i) \) where \( \alpha \in \mathbb{R}^m \) is
the decision variable and k is a user selected positive kernel function. The objective function being minimized includes \( \sum_i |f(x_i) - y_i| \) for any i such that \(|f(x_i) - y_i| \geq \epsilon\), where \( \epsilon \) is a tuning parameter. In addition, there is a regularization parameter, C where regularization increases as C decreases. SVR can generate accurate nonlinear predictive models for appropriate choice of k. However, the selection of the kernel heavily influences the resulting accuracy and this process of selection is difficult to automate.

*Kernel Learning (PMKL)* Kernel learning algorithms improve on the SVR problem by automating the search for a kernel function. Note we consider the class of kernel learning algorithms to include Deep Learning (although the search problem in this case is non-convex). These approaches are limited, however, by the class of kernels over which they are able to search. The class of Tessellated Kernel functions have been shown in [11] to have the properties of universality, density, and tractability - meaning the resulting algorithms are rather accurate and generalize well to new data. Specifically, the PMKL algorithm for optimizing TK kernel functions was shown in [12] to be more robust than other tested ML algorithms (including multilayer neural networks) - at the cost of some additional computational complexity. The regularization parameters in this case are the \( \epsilon \) and C as defined above for SVR.

*Decision Tree Algorithms* Decision trees are composed of a series of conditional statements that branch in a “tree” like manner. We say the “depth” of a decision tree is how many conditional statements appear in a branch before leading to a label denoted the “leaf”. Both the depth of the decision trees and the maximum number of leaves are regularization parameters that can be modified by the user. Decision trees are often weak predictors alone and in this paper we use ensemble (random forest) or boosting (boosted trees) methods to increase predictive performance. These algorithms are defined as follows.

- **Random Forest**: The random forest algorithm is an ensemble machine learning method based on a combination of decision trees. Ensemble methods use a combination of predictive models (trees) that individually have poor generalization but when used in combination can have significantly improved predictions. The number of decision trees combined in the random forest algorithm can be used as a regularization parameter.

- **Boosted Trees**: Gradient boosting is another machine learning method also based on a combination of decision trees. In the boosted algorithm trees are added to the predictive model sequentially, and each additional tree is fit to the current residuals of the model. A “learning rate” is a weight applied to the addition of each decision tree, and is often used as a regularization parameter. Small learning rates tend to improve the generalization of the predictive models.

Next we will focus on a metric we may use to identify the observables which are most suitable to the task of predicting self vs nonself determination in autoimmune disease.
Quantifying Suitability of a Given Set of Observables

To identify a set of observables for predicting self vs nonself determination we rigorously define a metric for suitability in order to select the observables which lead to superior predictive models.

First, for the sake of generality, we define the algorithm, $OPT$, which we use as a placeholder for the machine learning algorithms described previously.

**Definition of $OPT$**: Given a dataset $\{(x_i, y_i)\}_{i=1}^m \subset \mathbb{R}^w \times \mathbb{R}^q$, $OPT(\{(x_i, y_i)\}_{i=1}^m)$, returns a predictive function, $f = argOPT(\{(x_i, y_i)\}_{i=1}^m)$, where $f : \mathbb{R}^w \rightarrow \mathbb{R}^q$.

Next, given a possible set of feature indices $F := \{1, \cdots, n\}$, we define the set of partitions of $F$ as $\mathcal{P}(F)$, and the set of all possible partitions of $F$ of length $w \leq n$ as follows.

$$B_w := \{v \in \mathbb{N}^w \mid v \in \mathcal{P}(F)\}$$

For a given selection of features, $b \in B_w$, we denote the associated projection $P_b : \mathbb{R}^n \rightarrow \mathbb{R}^w$ so that $(P_b(x))_i = x_b$, for $x \in \mathbb{R}^n$ and $i = 1, \cdots, w$.

To define a metric of suitability we consider three cost/penalty functions, $M_1, M_2$, and, $L$. The function $L$ is a function of the cardinality of the number of features selected, $L(|b|_C)$. The costs $M_1$ and $M_2$, however, measure how well the selection of features can be used to predict the disease state and the remaining features respectively. To accurately evaluate the performance of the predictor a partition of the data must be withheld from the training algorithm, $OPT$, and used solely for the purpose of testing the performance. For a given set of data, these metrics will vary depending on which data points are used for training $OPT$ and which are used to evaluate its performance. To explicitly account for the effect of choice in partitioning of data samples, we now define the set of samples $S := \{1, \cdots, m\}$, and the set of partitions of $S$ as $\mathcal{P}(S)$. As for features, we denote the set of sample partitions of length $r$ as

$$S_r := \{v \in \mathbb{N}^r \mid v \in \mathcal{P}(S)\}$$

and for a given selection of samples, $g \in S_r$, we denote the associated projected data set as $\mathcal{P}_g(X) := \{x_i \in X, \ i \in g\}$.

Therefore, the costs $M_1$ and $M_2$ are a function of the feature partition, $b$, the training partition, $g \in S_r \in \mathcal{P}(S)$ and the associated test partition, $h := S/g \in S_{m-r}$, so that $M_1(b, g)$ and $M_2(b, g)$ are the Root Mean Square Error (MSE) of predicting the test partition. Specifically, let $R(f, x, y) = \sqrt{\frac{1}{m-r}\sum_{i \in S/g} |f(x_i) - y_i|^2}$ and we have

$$M_1(b, g) = R(f_{b,g}, P_b(x), y) \quad \quad M_2(b, g) = \sum_{j \in F/b} R(d_{b,g}^{(j)}, P_b(x), P_j(x))$$

$$f_{b,g} = argOPT(\{P_b(x_g), y_g\}_{i=1}^r) \quad \quad d_{b,g}^{(j)} = argOPT(\{P_b(x_g), P_j(x_g)\}_{i=1}^r)$$

In the ideal case, we would average these costs over all possible partitions of the data set to give an estimate of the predictive power of $b \in B_w$. However, such
an approach would result in very large computational overhead. Therefore, we use
the \( k \)-fold cross validation approach, wherein we divide the samples into \( k \) training
partitions of size \( m(k-1) \), which we label as \( g(i) \in S_{m(k-1)} \) for \( i = 1, \cdots, k \). Then
the average cost of the feature partition \( b \) over the \( k \) sample partitions is

\[
J(b) = \frac{1}{k} \sum_{i=1}^{k} J'(b, g(i)).
\]

where

\[
J'(b, g) := \beta_1 M_1(b, g) + \beta_2 M_2(b, g) + L(|b|_C)
\]

and where \( \beta_1, \beta_2 \geq 0 \) are given weights.

First, we let \( \beta_1 = 1 \) and \( \beta_2 = L(w) = 0 - a \) case we denoted as the Minimal Disease
State (MDS). In this case, we are only concerned with predicting the progression
of the disease and are not concerned with predicting non-selected features or with
the number of features selected. Second, we let \( \beta_1 = 0 \) and \( \beta_2 = 1 \) and

\[
L(w) = \begin{cases} 
0 & \text{for } w \leq 10 \\
\infty & \text{for } w > 10.
\end{cases}
\]

In this case, we ignore the disease state and are only concerned with reducing the
number of features while retaining the ability to reconstruct discarded features - a
case we denote as the Minimal Immune State (MIS). Finally, we let \( \beta_1 = \beta_2 = 1 \) and
\( L(w) \) as defined for the MIS. We denote this final case as the Minimal Immune
and Disease State (MIDS).

Other Performance Metrics To show that the results of Optimization Problem (4)
as applied to MDS, MIS and MIDS are consistent with other learning metrics [13],
we also include data on these metrics for the chosen selection of features and associated predictor. These metrics are defined as follows. Let \( y \) be the vector of
labels (discarded features and the disease state) associated with features \( x \). Let \( \hat{y} \)
be the predicted labels as generated by the predictor when applied to features \( x \).
Let \( \tilde{y} \) and \( \tilde{\hat{y}} \) be the average values of \( y \) and \( \hat{y} \). Then we have the following.

**The Correlation Coefficient and relative Root Mean Squared Error** (CC and rRMSE):

\[
CC = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(\hat{y}_i - \tilde{\hat{y}})}{\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2 \sum_{i=1}^{N} (\hat{y}_i - \tilde{\hat{y}})^2}}; \quad \text{rRMSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}}
\]

**Mean Absolute Error and relative Mean Absolute Error** (MAE and rMAE):

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|; \quad \text{rMAE} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{\sum_{i=1}^{N} |y_i - \bar{y}|}
\]

Next we define the specific feature selection algorithms used to analyze the RA
data set.
Feature Selection Algorithms

We have now defined the metric of suitability as a function of the partition, \( b \in B_w \). Using this metric, the feature selection problem is defined as the following combinatoric optimization problem.

\[
\min_{b \in B_w, w \in \mathbb{N}} J(b) \quad (4)
\]

Optimization problems of this form are a special case of feature selection (typically solved using wrapper methods) and, being combinatorial optimization, Problem (4) is NP-hard [14]. As a consequence, most Feature Selection (FS) algorithms as applied to this problem are either heuristic, in that they are not guaranteed to converge to a globally optimal solution, or solve unrelated problems which may or may not yield reasonable values for Problem (4).

Nonetheless, several techniques have been proposed that enjoy relative accuracy and computational efficiency. We focus first on FS methods designed specifically for problems of the same form as Optimization Problem (4), then consider two other FS approaches that do not directly try to solve the optimization problem of interest but provide a comparison to the direct method.

Proposed Wrapper Method and Implementations

We first define the algorithm (a wrapper method) which will be used and then provide additional details on the various ML algorithms which are combined with this wrapper to solve Problem (4).

The most common wrapper methods are Sequential Feature Selection (SFS) algorithms [14]. SFS algorithms begin with an empty (or full) set of features and sequentially add (or remove) the highest value (or cost) feature until the set of features is a certain size or meets a performance metric.

The SFS algorithm used in this paper is as described in [15]. This SFS algorithm begins with \( b := \emptyset \), and iteratively selects a locally optimal feature (with respect to the objective function of Optimization Problem (4)) at each step.

Clearly, the effectiveness of Feature Selection depends on the ML algorithm (\( OPT \)) used to generate the predictive model. Therefore, in the Results Section, we test all the machine learning algorithms proposed herein. Unfortunately, the accuracy of the predictive model is influenced by user-selected parameters within the algorithm. For reproducibility, we list here the selections for these parameter values.

**Linear Regression:** We test all 16 combinations of \( \alpha_1 \in [0, 0.1, 1, 5] \) and \( \alpha_2 \in [0, 0.1, 1, 5] \) and the data from choice yielding highest suitability (\( J \)) is listed in Table 2.

**PMKL:** We use the default TK kernel parameters and test \( \epsilon = .005 \), and \( C \in [0.01, 1, 3, 5, 1] \) and the data from the choice yielding highest suitability (\( J \)) is listed in Table 2.

**SVR:** We test all combinations of \( \epsilon = .1 \), \( C \in [1, 5, 10] \) and 3 kernel functions (linear, RBF, or 3rd degree polynomial) and the data from choice yielding highest suitability (\( J \)) is listed in Table 2. For the RBF kernel the features are normalized by their variance and a bandwidth of \( \frac{1}{n} \) is selected.

**Random Forest** We test 9 combinations of number of trees (\( n_{trees} \in [50, 100, 150] \))
and the maximum tree depth of \((\text{max}\_\text{depth} \in [5, 10, 20])\) and the data from choice yielding highest suitability \((J)\) is listed in Table 2. **Boosted Trees** We test 15 combinations of number of trees \((n_{\text{trees}} \in [50, 100, 150, 250])\) and learning rate \((\text{LR} \in [0.01, 0.1, 0.5])\) and the data from choice yielding highest suitability \((J)\) is listed in Table 2.

Suitability of Filter and Embedded Methods

Alternative feature selection algorithms will be used as a baseline by which we may compare the wrapper method. We use three filter methods and one embedded method in the analysis.

**Filter Methods**

Given a set of data, filter methods use a rating function to rank each features relative “importance”. After the features have been ranked, the user may select \(w\) features to be kept and the remaining features will be discarded. The rating functions used to generate the data in Table 2 are as follows.

**Mutual Information (MI)** The Mutual Information criteria [16] is a statistical function of two random variables that describes the amount of information contained in one random variable relative to the other.

**Analysis of Variance (ANOVA)** The ANOVA method [17] is a commonly used method for analyzing variable dependencies. The F-test is used to estimate the features importance.

**Principle component analysis (PCA)** This method approximates the data with linear manifolds [18]. The main methods used to perform PCA are based on the singular value decomposition and diagonalization of the correlation matrix. We calculate the importance based on the first 3 eigenvectors.

In all cases, once a set of features has been selected, suitability \((J)\) is determined using each of the ML algorithms and we report the minimum of these values.

**Embedded Methods**

Embedded FS methods attempt to embed the process of feature selection directly into the model generation process - typically adding a cost for inclusion of a particular feature in the model. These methods have been used in the gene expression domains as in [19] and have been successfully applied to mass spectrometry analysis in [20, 21, 22]. For this analysis, only a single embedded method was considered.

**Mean Decrease in Impurity (RF)** The Gini Importance or Mean Decrease in Impurity [23] is an embedded method for the Random forest algorithm. It calculates the importance of features as the mean of the number of splits (over all trees) that include this feature, weighted by the probability of reaching this node.

We next apply these methods to the RA dataset and report the results.

**Results**

Here we define three immune states generated by varying the suitability metric. These three immune states are lower dimensional subsets of the data which can be used to either predict the progression of RA, reconstruct the full set of T cell markers and populations, or perform both tasks simultaneously.
Table 1  Metrics of fit for Feature Selection algorithms designed to select features for predicting the disease progression using the RA data described and feature selection algorithms described in the Methods Section. For each algorithm we report only the results of the regularization parameters that generated the best objective value.

| Algorithm       | $J$ | MAE | rRMSE | rMAE | cc  |
|-----------------|-----|-----|-------|------|-----|
| Linear Model    | 0.32| 0.27| 0.78  | 0.72 | 0.55|
| PMKL            | 0.34| 0.29| 0.84  | 0.84 | 0.51|
| Random Forest   | 0.35| 0.29| 0.83  | 0.87 | 0.53|
| Boosted Trees   | 0.36| 0.31| 0.89  | 0.88 | 0.49|
| SVR             | 0.37| 0.31| 0.90  | 0.93 | 0.40|
| MI              | 0.38| 0.34| 0.99  | 0.99 | 0.28|
| RF              | 0.38| 0.34| 0.98  | 1.02 | 0.30|
| ANOVA           | 0.38| 0.35| 1.02  | 1.01 | 0.28|
| PCA             | 0.40| 0.38| 1.11  | 1.10 | 0.28|

Case 1: Features for Predicting Disease Progression (MDS)

First we consider the problem of selecting features that are optimal for predicting the disease progression, a state we denote the Minimal Disease State (MDS).

**Performance of FS Algorithms** In Table 1 we rank a collection of feature selection methods via the proposed comparison metric of fit, $J$ in Optimization Problem (4), for the MDS case. We also report the other metrics of fit as defined in Eq. (3) and (2).

For MDS the Sequential Forward Selection (SFS) based algorithms had objective values, significantly less than the embedded and filter methods. This large difference is likely due to the fact that the embedded and filter methods are heuristics.

Looking specifically at the SFS based algorithms we see that for MDS, the SFS Linear, SFS TK, SFS Random Forest, SFS Boosted Trees and SFS SVR algorithms had the best performance respectively. Two of these algorithms selected significantly fewer features: The SFS Random Forest (4 features) and the SFS TK (5 features). Therefore, in terms of predictive accuracy per selected number of features the SFS TK and SFS Random Forest algorithms performed best, selecting less than half as many features as the 12 average features selected by the other methods. However the linear model which selected 12 features provided the best overall predictive accuracy for the disease state.

**Most Important Features Using the SFS Algorithms**

In Fig. 2 we show the observables that were selected by each of the proposed algorithms. If we consider only the top performing algorithms (the SFS based algorithms) and the markers specific to helper and regulatory cells, then counting the number of times a feature was selected by the SFS algorithms, the following features were chosen by at least three of the algorithms.

1. $CD4^+GATA3^+CD44^+CD62^{(Lo)}$ (3 times)
2. $CD4^+GATA3^+Ki67^+$ (3 times)
3. $CD4^+Foxp3^+CD25^+$ (3 times)
4. $CD4^+Foxp3^+CD25^+Ki67^+Autoantigen$ (3 times)
5. $CD4^+Tbet^+$ (3 times)

Among the cytotoxic cells, the algorithms were most consistent, with all five of the algorithms selecting one feature in common.

6. $CD8^+Ki67^+$ (4 times)
(7) CD8+GATA3+ (3 times)
(8) CD8+Tbet+ (3 times)

This group of cells consists of cytotoxic (6,7,8), Th memory (1), Th (2,5), and Treg (3,4) T cell sub-populations. The location feature (origin of the tested cells), was selected only once by an SFS based algorithm. In this case we do not include the treatment as a possible feature, since we are primarily interested in the prediction of the disease state using sub-populations of T cells as opposed to the already known correlation between treatment and disease state. In the next two cases treatment is considered a feature.

Case 2: Features for Reconstructing Discarded Features (MIS)
Next we consider the problem of selecting features that are optimal for reconstructing discarded features to determine a Minimal Immune State (MIS).

Performance of FS Algorithms
In Table 2 we rank a collection of feature selection methods via the proposed comparison metric of fit, $J$ in Optimization Problem (4), for the MIS case. We also report the other metrics of fit as defined in Eq. (3) and (2).

Looking specifically at the SFS based algorithms we see that for MIS, the SFS Random Forest and Boosted Trees algorithms were the best performing methods. All algorithms selected the maximum number of 10 features.

Most Important Features Using the SFS Algorithms
In Fig. 3 we show the features that were selected by each of the proposed algorithms.
Table 2: Metrics of fit for Feature Selection algorithms designed to select features best at reconstructing discarded features using the RA data described and feature selection algorithms described in the Methods Section. For each algorithm we report only the results of the regularization parameters that generated the best objective value.

| Algorithm       | J    | MAE | rRMSE | rMAE | cc  |
|-----------------|------|-----|-------|------|-----|
| Random Forest   | 0.11 | 0.08| 0.37  | 0.29 | 0.89|
| Boosted Trees   | 0.12 | 0.08| 0.38  | 0.31 | 0.88|
| PMKL            | 0.12 | 0.09| 0.37  | 0.22 | 0.89|
| Linear Model    | 0.13 | 0.09| 0.44  | 0.31 | 0.87|
| SVR             | 0.13 | 0.09| 0.46  | 0.32 | 0.85|
| PCA             | 0.15 | 0.12| 0.58  | 0.49 | 0.76|
| MI              | 0.17 | 0.13| 0.61  | 0.53 | 0.74|
| RF              | 0.17 | 0.13| 0.59  | 0.49 | 0.75|

Unlike in the previous subsection, there was less of an agreement among the high-performing SFS algorithms as to the most significant features. For MIS only 6 different features were selected by at least three algorithms. First, if we consider markers specific to helper and regulatory cells, and counting the number of times a feature was selected by the SFS methods (each method selected 10 features), the following features were each chosen by at least 3 algorithms.

1. $CD4^+GATA3^+CD44^+CD62L(Lo)$ (4 times)
2. $CD4^+Tbet^+Ki67^+ (4 times)$
3. $CD4^+GATA3^+Ki67^+Autoantigen$ (3 times)
4. $CD4^+Tbet^+Autoantigen$ (3 times)

We note that two of the selected features are autoantigen specific as opposed to the single autoantigen specific feature selected for cells in MDS.

Among the cytotoxic cells, the algorithms were less consistent, with only three of the algorithms selecting similar sub-populations.

5. $CD8^+Ki67$ (3 times)
6. $CD8^+GATA3^+CD44^+CD62(Lo)$ (3 times)

We note that the central memory T cells (CD62L) appear in both the helper/regulatory populations and the cytotoxic cell populations. In this case, data-rich biomarkers (those containing multiple markers), were selected slightly more often when compared to MDS. The average number of markers in the selected features is 3.33 in this case compared to 2.875 in the MDS case.

Of particular note is the fact that the location feature (origin of the tested cells) and the treatment feature (which treatment was applied) were both selected by almost every algorithm.

Case 3: Features for Disease Progression and Reconstruction (MIDS)

Next we consider the problem of selecting features that are optimal for predicting a combination of the MIS and MDS objectives, the Minimal Immune and Disease State (MIDS).

Overall Performance of FS Algorithms: In Table 3 we rank a collection of feature selection methods via the proposed comparison metric of fit, $J$ in Optimization Problem (4), for the MIDS case. We also report the other metrics of fit as defined in Eq. (3) and (2).

As in the MIS and MDS case, the SFS algorithms had the best performance, and consequently we consider only the SFS algorithms in the analysis. Looking
The green squares indicate that the feature selection method (left) selected the feature (top). We show the observables selected by the nine different FS algorithms from Section that compose the MIS. The methods are ordered from lowest objective function, $J(b)$, at the top to lowest objective at the bottom. The SFS methods and the features most commonly selected by those methods are bolded.

**Most Important Features Using the SFS Algorithms**

In Fig. 4 we show the features that were selected by each of the proposed algorithms. If we consider markers specific to helper and regulatory cells, the following features were each chosen by at least three of the five algorithms.

1. $\text{CD}_4^+\text{GATA}_3^+\text{CD}_{44}^+\text{CD}_{62L}^\text{Lo}$ (4 times)
2. $\text{CD}_4^+\text{Thet}^+\text{Ki67}^+$ (4 times)
3. $\text{CD}_4^+\text{Thet}^+\text{Autoantigen}$ (4 times)
4. $\text{CD}_4^+\text{GATA}_3^+\text{Ki67}^+\text{Autoantigen}$ (3 times)

As in the MIS case two of the selected features are autoantigen specific, however no cytotoxic cells were consistently selected by at least three of the SFS algorithms.

Similar to the MIS case, no regulatory cells were consistently selected by all five of the top performing algorithms. Just as in the case of the MIS the location feature (origin of the tested cells) were selected by almost every algorithm and the treatment attribute was likewise selected by every algorithm.

**Discussion**

Here we discuss the results of each each proposed case, MDS, MIS, and MIDS separately.
Table 3: Metrics of fit for Feature Selection algorithms designed to select features best at reconstructing discarded features and predict disease progression using the RA data described and feature selection algorithms described in the Methods Section. For each algorithm we report only the results of the regularization parameters that generated the best objective value.

| Algorithm       | J MAE | rRMSE | rMAE | cc  |
|-----------------|-------|-------|------|-----|
| Random Forest   | 0.11  | 0.09  | 0.38 | 0.29 | 0.90 |
| Boosted Trees   | 0.12  | 0.09  | 0.39 | 0.31 | 0.88 |
| PMKL            | 0.12  | 0.09  | 0.38 | 0.23 | 0.89 |
| Linear Model    | 0.13  | 0.10  | 0.45 | 0.31 | 0.86 |
| SVR             | 0.14  | 0.10  | 0.47 | 0.34 | 0.83 |
| RF              | 0.14  | 0.09  | 0.47 | 0.63 | 0.82 |
| MI              | 0.17  | 0.12  | 0.52 | 0.48 | 0.77 |
| ANOVA           | 0.18  | 0.12  | 0.56 | 0.55 | 0.73 |

Case 1: Features for Predicting Disease Progression (MDS)  High predictive accuracy is important for tracking the disease progression and predicting the effectiveness of treatments based on measurements of the T cell markers - which is of particular importance for autoimmune diseases where the disease state may be difficult to measure.

The location feature was not selected by the top performing feature selection algorithms suggesting that the location where the T cells were collected is inconsequential to predicting the disease state. This implies that there is significant uniformity in the disease state among the lymph nodes and spleen.

In addition, only one selected feature (4) was autoantigen specific - indicating that most of these T cell markers may be correlated to, not just the progression of RA, but the progression of other similar autoimmune diseases as well.

Finally, the MDS consisted of biomarkers that were not data-rich but contained general classes of T cells rather than the more specific sub-populations selected in the MIS and MIDS cases. This suggests that the disease state is caused by a large irregularity in general populations of T cells rather than irregularities in a few smaller specific populations.

Case 2: Features for Reconstructing Discarded Features (MIS)  This case studied the features most important for determining the overall state of the immune system (and not the progressive state of any particular disease).

When compared to the disease state the MIS consisted of significantly more data-rich markers. This may be due to the fact that estimating the entire immune state is significantly more difficult then simply estimating the disease state alone, and more data-rich markers may therefore be necessary. No regulatory cells were consistently selected by all five of the top performing algorithms, suggesting that no specific regulatory cell type was most important for reconstructing the immune state.

Nearly all algorithms selected both the location the T cells were collected and the treatment as important features for predicting the T cell populations. This implies that the immune state is not uniform across the lymph nodes and spleen, and that the specific treatment given to the mice has a large impact on the immune state.

Case 3: Features for Disease Progression and Reconstruction (MIDS)  This case studied the features most important for determining both the overall state of the immune system as well as the progressive state of RA in the mice.
The green squares indicate that the feature selection method (left) selected the feature (top). We show the observables selected by the nine different FS algorithms from Section 7 that compose the MIDS. The methods are ordered from lowest objective function, \( J(b) \), at the top to lowest objective at the bottom. The SFS methods and the features most commonly selected by those methods are bolded.

The combined state had many of the same features as the immune state, consisting of more data-rich markers than the disease state alone, inconsistent selection of regulatory cells, and the selection of both location and treatment as important features.

However, in this case the algorithms were least consistent on selecting cytotoxic cells, implying that no single cytotoxic cell type was consistently selected to both reconstruct the immune system and determine the disease progression. This would imply, therefore, that the Th cell populations are more essential to both predicting the disease progression and reconstructing the immune state.

We note that the memory T cell sub-population \( CD_{4}^{+}GATA3^{+}CD_{44}^{+}CD_{62L}^{(Lo)} \) was selected in all three cases. It is clear that this sub-population is significant to both the immune and disease states.

Conclusions

In this paper, we have considered the problem of identification of three different states of the immune system of increasing complexity. Specifically, we have used a set of mouse-model experiments to obtain a robust dataset of T cell markers and populations at the end stage of a proposed immunotherapy treatment. The first state is the disease state, which is important for tracking the disease progression and predicting the effectiveness of treatments. Next is the immune state which is the minimal number of sub-populations needed to reconstruct the remaining discarded features. Finally we find the combined overall immune state for predicting both the disease state and reconstructing the discarded features. From these experiments
we were able to determine that the $\text{CD}^4+\text{GATA3}+\text{CD}44+\text{CD}62\text{L(Lo)}$ memory T cell sub-population is significant to both the immune state and disease state of mice with RA - and have developed a set of T cell sub-populations important for tracking the disease progression and outcome of immunotherapy.

Acknowledgements
Not applicable.

Funding
This material is based upon work supported by the National Science Foundation under grant No. 1931270, No. 1935453 to MMP, the National Science Foundation Graduate Research Fellowship Program under Grant No. 026257-001 to BKC and by the following grants to APA - NIH R01AR078343 and NIH R01AI155907.

Availability of data and materials
The dataset supporting the conclusions of this article is included within the additional file in the supplementary material.

Ethics approval and consent to participate
The Institutional Animal Care and Use Committee (IACUC) of Arizona State University approved animal studies regarding the rheumatoid arthritis protocol number: 19-1712R. All animal experiments were conducted in accordance with the guidelines of Arizona State University and IACUC.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Authors’ contributions
BKC and AT were responsible for the data analysis, feature selection and machine learning. BKC also contributed to the writing of the paper. JLM helped design and perform animal experiments. APA contributed to the design of animal experiments and the development of the work. MMP contributed to the development of the results and writing of the paper. All authors read and approved the final manuscript.

References
1. Fisher, J.D., Zhang, W., Balmert, S.C., Aral, A.M., Acharya, A.P., Kulahci, Y., Li, J., Turnquist, H.R., Thomson, A.W., Solari, M.G., et al.: In situ recruitment of regulatory t cells promotes donor-specific tolerance in vasculatized composite allotransplantation. Science Advances 6(11) (2020)
2. Ratay, M.L., Glowacki, A.J., Balmert, S.C., Acharya, A.P., Polat, J., Andrews, L.P., Fedorchak, M.V., Schuman, J.S., Vignali, D.A.A., Little, S.R.: Treg-recruiting microspheres prevent inflammation in a murine model of dry eye disease. Journal of Controlled Release (2017)
3. Fisher, J.D., Balmert, S.C., Zhang, W., Schweizer, R., Schneider, J.T., Komatsu, C., Dong, L., Erbas, V.E., Unadkat, J.V., Aral, A.M., et al.: Treg-inducing microparticles promote donor-specific tolerance in experimental vasculatized composite allotransplantation. Proceedings of the National Academy of Sciences 116(51) (2019)
4. Ratay, M.L., Balmert, S.C., Acharya, A.P., Greene, A.C., Meyyappan, T., Little, S.R.: Tri microspheres prevent key signs of dry eye disease in a murine, inflammatory model. Scientific reports 7(1) (2017)
5. Acharya, A.P., Sinha, M., Ratay, M.L., Ding, X., Balmert, S.C., Workman, C.J., Wang, Y., Vignali, D.A.A., Little, S.R.: Localized multi-component delivery platform generates local and systemic anti-tumor immunity. Advanced Functional Materials 27(5) (2017)
6. Brodin, P., Davis, M.M.: Human immune system variation. Nature reviews immunology 17(1) (2017)
7. Davis, M.M.: A prescription for human immunology. Immunity 29(6) (2008)
8. Germain, R.N., Schwartzberg, P.L.: The human condition: an immunological perspective. Nature immunology 12(5) (2011)
9. Mangal, J.L., Inamdar, S., Yang, Y., Dutta, S., Wankhede, M., Shi, X., Gu, H., Green, M., Rege, K., Curtis, M., et al.: Metabolite releasing polymers control dendritic cell function by modulating their energy metabolism. Journal of Materials Chemistry B 8(24) (2020)
10. Mangal, J.L., Inamdar, S., Le, T., Shi, X., Curtis, M., Gu, H., Acharya, A.P.: Inhibition of glycolysis in the presence of antigen generates suppressive antigen-specific responses and restrains rheumatoid arthritis in mice. Biomaterials 277 (2021)
11. Colbert, B.K., Peet, M.M.: A convex parametrization of a new class of universal kernel functions. Journal of Machine Learning Research 21(45) (2020)
12. Colbert, B.K., Peet, M.M.: A New Algorithm for Tessellated Kernel Learning (2020). 2006.07693
13. Borchani, H., Varando, G., Bielza, C., Larrañaga, P.: A survey on multi-output regression. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 5(5) (2015)
14. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Computers and Electrical Engineering 40(1) (2014)
15. Dash, M., Liu, H.: Feature selection for classification. Intelligent data analysis 1 (1997)
16. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE transactions on neural networks 5(4) (1994)
17. Kumar, M., Rath, N.K., Swain, A., Rath, S.K.: Feature selection and classification of microarray data using mapreduce based anova and k-nearest neighbor. Procedia Computer Science 54 (2015)
18. Song, F., Guo, Z., Mei, D.: Feature selection using principal component analysis. 2010 International Conference on System Science, Engineering Design and Manufacturing Informatization I (2010)
19. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning 46 (2004)
20. Zhang, X., Lu, X., Shi, Q., Xu, X.-q., Leung, H.E., Harris, L., Iglehart, J.D., Miron, A., Liu, J., Wong, W.: Recursive svm feature selection and sample classification for mass-spectrometry and microarray data. BMC Bioinformatics 7 (2006)
21. Jong, K., Marchiori, E., Sebag, M., Vaart, A.W.: Feature selection in proteomic pattern data with support vector machines. 2004 Symposium on Computational Intelligence in Bioinformatics and Computational Biology (2004)
22. Prados, J., Kalousis, A., Sanchez, J., Allard, L., Carrette, O., Hilario, M.: Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. Proteomics 4(8) (2004)
23. Breiman, L., Friedman, J., Olshen, R., Stone, C.J.: Classification and regression trees. Biometrics 40(3) (1984)