

## CASE STUDY

# Predicting class switch recombination in B-cells from antibody repertoire data

Lutecia Servius<sup>1</sup>  | Davide Pigoli<sup>1</sup> | Joseph Ng<sup>2</sup> | Franca Fraternali<sup>2</sup>

<sup>1</sup>Department of Mathematics, King's College London, London, UK

<sup>2</sup>Institute of Structural and Molecular Biology, University College London, London, UK

**Correspondence**

Lutecia Servius, Department of Mathematics, King's College London, UK.  
Email: [lutecia.servius@kcl.ac.uk](mailto:lutecia.servius@kcl.ac.uk)

**Funding information**

Biological Physics Across Scales CDT; Biotechnology and Biological Sciences Research Council, Grant/Award Number: BB/T002212/1



This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

**Abstract**

Statistical and machine learning methods have proved useful in many areas of immunology. In this paper, we address for the first time the problem of predicting the occurrence of class switch recombination (CSR) in B-cells, a problem of interest in understanding antibody response under immunological challenges. We propose a framework to analyze antibody repertoire data, based on clonal (CG) group representation in a way that allows us to predict CSR events using CG level features as input. We assess and compare the performance of several predicting models (logistic regression, LASSO logistic regression, random forest, and support vector machine) in carrying out this task. The proposed approach can obtain an unweighted average recall of 71% with models based on variable region descriptors and measures of CG diversity during an immune challenge and, most notably, before an immune challenge.

**KEYWORDS**

balanced accuracy, clonal groups, immune responses, predictive models

## 1 | INTRODUCTION

High-throughput measurements of cellular molecular profiles provide opportunities to investigate complex biological phenomena, taking advantage of a vast quantity of large-scale data (Goodwin et al., 2016). High-throughput sequencing methods are applied in a variety of fields, including immunology, to gain insights into immune function, dynamics, and the analysis of the factors that affect them. These data carry information about past and present immune challenges, enabling the prediction of immune health leading to the formulation of novel vaccines and therapeutics (Widrich et al., 2020).

The resultant generation of large data sets requires the use of appropriate statistical techniques to analyze and explore relevant physiological phenomena. Statistical and machine learning methods have been used in a variety of biological fields, including disease diagnosis (Hajipour et al., 2020; Pino-Mejías et al., 2008), and specifically in detecting immune responses (Horst et al., 2021; Wei et al., 2021). These models can be used to extract relationships in the data to generate a greater understanding of the variable of interest.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

For this comparison, we selected some widely used methods which allow us to balance performance and interpretability. For a classification problem where the number of features is smaller than the number of observations, logistic regression (LR; Hastie et al., 2009, Chap. 3) is a natural choice for a linear classifier. We also consider a regularized LR, the LASSO (Hastie et al., 2009, Chap. 3), where both coefficient shrinkage and variable selection occur in tandem. In contrast to the LR, both the random forest (RF; Breiman et al., 2017), a tree-based classifier, and kernel support vector machine (SVM) predictors (Hastie et al., 2009, Chap. 12) make no linearity assumption. There exist more complex models which prioritize predictive performance, like deep learning methods, however this comes at the cost of interpretability and requires, in our case, a substantial amount of data which can be prohibitive, thus they will not be included in our comparison.

In this paper, the focus is on the mechanism of class switch recombination (CSR) in B-cells and the use of statistical and machine learning methods to predict its occurrence solely using information from the variable region and donor characteristics. We propose a novel framework to analyze the antibody repertoire data, by leveraging the clonal group (CG) level characteristics—which can be repurposed for any cellular data—so that they can be fed as an input for predictive models and showcase how CG level representation successfully appropriates the biology of the cells and is relevant for our prediction problem. We then compare the performance of several statistical and machine learning techniques when applied to these novel data objects to predict CSR events under different experimental conditions. We show that CSR occurrence can be predicted on the basis of CG diversity in the variable region during an immune challenge, but most notably, before an immune challenge.

In Section 2, we will introduce the biological background of the data used in this paper. The data sets used to train, validate, and test the predictive methods will be presented in Section 3; Section 4 will propose a novel CG representation for the antibody repertoire data, informed by the biological background, while Section 5 details the data processing. In Section 6, we detail the methodology used to predict CSR using the five models of interest: LR, LASSO logistic regression (LASSO LR), RF, random forest with variable node size (RF-NSx), and SVM. Section 7 reports the performance of the predictive models when applied to the immunological data sets and discusses their interpretability. Finally, Section 8 is an overall discussion on what we can learn from this case study about the application of statistical and machine learning models to antibody repertoire and what are possible future directions in this regard.

## 2 | ANTIBODY REPERTOIRE AND CSR

B-cells are the main players in the humoral response and produce antibodies, a class of proteins specific to antigens found in invading pathogens and malignant cells (Feldkamp & Carey, 1996). Antigen binding occurs in the variable region; this is the element of the antibody which has direct contact with antigens and thus requires antigen specificity, acquired by accumulating mutations (Feldkamp & Carey, 1996) to fine-tune its capability to recognize millions of possible antigens that one may encounter in life. The huge diversity in antigen specificity, and thus in the variable region, is achieved through the recombination of three gene segments: Variable (V), Diversity (D), and Joining (J) (Lescale & Deriano, 2022). Comparatively, the constant region bears much less variability and determines the class (also known as “isotype”) and function of the antibody (Janeway et al., 2001). There are five major classes of antibodies in humans: IgM, IgD, IgG, IgA, and IgE. Both IgG and IgA can be subdivided into IgG1-4 and IgA1-2 subclasses. These antibody classes determine the downstream response a B-cell would initiate upon antigen binding; for example, IgE is the major type of antibody involved in allergic response (Janeway et al., 2001); IgG is typically the most abundant isotype in blood, and is involved in neutralizing bacterial and viral infections (Janeway et al., 2001).

In any antibody-mediated immune response, the first antibodies produced are IgM, yet multiple classes exist in the human body. The class of the antibody changes via the mechanism of CSR, thereby altering the function of the antibody and improving its ability to remove antigens that have induced the response (Stavnezer & Schrader, 2014). CSR helps to diversify B-cell responses and match antibody function to the immune challenge.

CSR is part of a complex process that generates effective antibodies against antigens. Upon confrontation with an antigen, an antibody response aims to perform two tasks simultaneously: (i) improve the antibody’s ability to recognize and target antigens (the specificity) by introducing mutations in the variable region; and (ii) undergo CSR to change the constant region such that the appropriate antibody classes are available to modulate downstream responses. This requires a balancing act between diversification, allowing highly-specific antibodies to arise; and focus, the selection of these specific antibodies and the expansion of their relative quantities in the “repertoire” of antibodies. CSR contributes to this balancing act via targeted changes of the constant regions. Since separate processes govern changes in the variable and

constant regions, they have traditionally been considered as separate entities—in fact, engineering of therapeutic antibodies typically involves “grafting” a variable region of interest (which binds to a desired antigen) onto a constant region with the desired biological functions (Chan & Carter, 2010). Albeit there are individual reports of interdependence of the constant and variable regions (Su et al., 2018; Tudor et al., 2012; Zhao et al., 2019), it is not clear whether this dependence is a general phenomenon or if instead there are features and/or rules indicating that certain variable regions are more susceptible to CSR than others.

Considering CSR is in essence a state transition, it is not straightforward to directly detect using data collected from snapshots of this process. One possible method is by monitoring single cells; while emerging single-cell molecular profiling methods might hold the promise of tracing the occurrence and dynamics of CSR events, they typically involve setting up costly experimental methods to generate these data (Horton et al., 2022). Assessing the antibody response is most easily done by estimating the distribution of antibody classes and variable region using blood samples and/or samples from various immune organs (lymph nodes, spleen, etc.) (Becker et al., 2021); these “snapshots” of the response at multiple time points do not give direct indications as to *how* CSR occurs. However, analyzes of these snapshots obtained along a time course allow us to estimate the timing at which CSR occurs, as well as the start and end points of CSR (Ng et al., 2023; Stewart et al., 2022).

Since CSR operates at the level of individual B-cells, it is important to consider the heterogeneity of CSR status across the pool of B-cells. Upon interaction with an antigen, B-cells which are capable of recognizing this antigen will start to divide and produce a clone of identical progeny with receptors that bind to the same antigen, maintaining antigen specificity (Janeway et al., 2001). It is within these “clones” of B-cells where mutations and CSR operate to optimize antigen recognition. Methods to obtain sequences of the antibody repertoire provide means to define B-cell clones which allow us to define the occurrence of CSR and its diversity across different clones in the overall antibody response: given the variable region nucleotide sequence is the same, one can assemble them into CG—where a CG is a set of antibodies with variable regions which harbor an identical antigen-binding site, most commonly taken as the third complementarity determining region (CDR3) on the antibody heavy chain. CSR is said to have occurred if a CG contains more than one class. For any observation within this cross-class CG, they can be labeled as being amenable to CSR. Accordingly, it is necessary to monitor the progression of these CGs over time as repeated sampling over the course of an immune response (e.g., in a vaccination trial) would allow for observing increasing numbers of cross-class CGs.

### 3 | DATA SETS

To investigate the relationship between CSR and the variable region, it is necessary to use data sets with defined CGs and where each sequence has resolved isotype information to allow us to identify CSR as defined in Section 1. The hospitalized COVID-19 donors (COVID) and respiratory syncytial virus (RSV) data sets published in Stewart et al. (2022) suit this purpose.

Both these immunological data sets are B-cell repertoires with donor metadata as illustrated in Figure 1a. The repertoires contain sequenced heavy chain observations from donors with the variable and constant region descriptors mentioned in Section 4 which are both continuous and categorical in nature, for example, the V gene, J family, D gene, and Kidera factors. The observations are collected over a time course. A summary of the CGs in the two data sets at the first time point can be seen in Table S1.

It is important to note a limitation of this approach to define the presence or absence of CSR within a B-cell CG, namely, that it is dependent on the sampling depth in sequencing the repertoire. Specifically, the larger the CG, the higher the likelihood of observing sequences of different isotypes. Therefore, the binary variable of interest  $I_{\text{CrossClass}}$  is confounded by the sample size of the cells selected for the experiment. Without clarification on the sampling distribution and the dependence of the clonal size on the prevalence of switching, one can only conclude that  $I_{\text{CrossClass}} = 1$  (TRUE) refers to a successful *observation* of switching and  $I_{\text{CrossClass}} = 0$  (FALSE) is an unsuccessful *observation* of switching. This issue is partly mitigated by including the CG size at day 0 as a considered variable in the models. Furthermore, an intermediary test was conducted to investigate this effect by clustering the observations by clonal size and creating bins for the CG observed sizes to train separate predictive models. The results fell in line with those shown in Section 7.

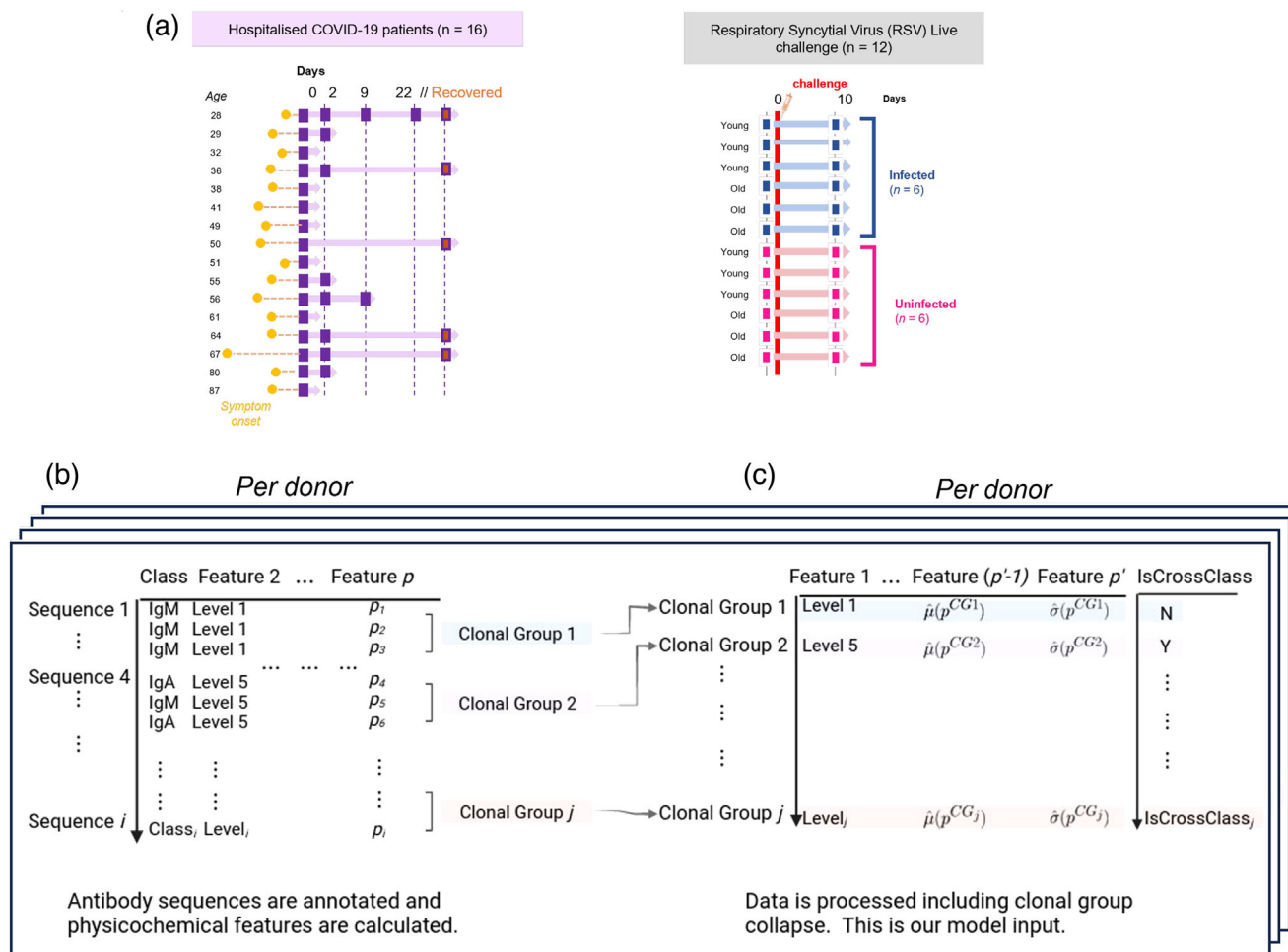


FIGURE 1 A schematic illustrating the data used including (a) the sampling regime for the COVID and respiratory syncytial virus (RSV) data sets outlined in Section 3 from Stewart et al. (2022); parts (b)–(c) show the workflow for the creation of the antibody sequence data sets and the effect of the data processing in Section 5 which includes the CG collapse. The  $\hat{\mu}(p^{CG_j})$  refers to the estimated mean of the variable  $p$  for CG  $j$  and  $\hat{\sigma}(\cdot)$  is the estimated standard deviation.

### 3.1 | Hospitalized COVID-19 patients

Severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2) became known in late 2019 and is a pathogenic coronavirus that has resulted in a pandemic of acute respiratory disease named COVID-19 (Hu et al., 2021). The COVID data set was collected from SARS-Cov-2 positive patients who were hospitalized in Frimley and Wexham Park hospitals during the first wave of SARS-Cov-2 infections in the United Kingdom, 2020. The start of the symptoms for each donor varies as does the infection time. The data set contains observations from 16 donors taken across five time points: days 0 (admitted to hospital), 2, 9, 22, and follow-up 2–3 weeks later, although time-courses sampling were incomplete for some donors as can be seen in Figure 1a. Since the first time point (Day 0) is the only sampling point where all donors are represented, we used only the features summarized from data collected at the first time point for predicting the response variable defined across all available time points for each donor, using the models described in Section 6. The data metadata consist of the age and the sex of the donors.

### 3.2 | RSV live challenge

RSV is a major cause of lower respiratory tract disease in both the elderly and young children and has been shown to be an ever-present pathogen that infects nearly all children by the age of 2 (Battles & McLellan, 2019). Samples were

**TABLE 1** A nonexhaustive selection of the CG features used to describe the data introduced in Section 3 and used as the input for the models in Section 6.

Feature	Data type	Description
IsCrossClass	Binary	Response variable where 1 denotes a cross-class CG.
Vgene, Jfamily, etc.	Categorical	Gene segments that are recombined to produce the variable region.
Age	Categorical/ Continuous	Old/young label for RSV donors with the cutoff at 55 years; continuous variable of donor age for COVID.
V.REGION.id, J.REGION.id	Continuous	Measure of the mutation of a gene segment, it takes values between 0 and 100 where 100 refers to a complete match to the reference, thus no mutation. Can be taken as a proxy for B-cell maturation.
TotalN, TotalP, etc.	Continuous	Number of nucleotide additions/deletions in the CDR3 region during the recombination of V, D, and J gene segments.
Num_AAs	Continuous	Number of amino acids in the CDR3 sequence—measure of length.
Tiny, Small, etc.	Continuous	Physicochemical properties describing frequencies of specific amino acids in the CDR3 sequence.
Boman, pI_EMBOSS, etc.	Continuous	Physicochemical properties describing particular attributes of the CDR3 such as the potential protein interaction (Boman) and the isoelectric point according to EMBOSS (pI_EMBOSS).
kideral-10	Continuous	Set of 10 factors that aid in the description of the protein structure by describing orthogonal physicochemical protein properties (Nakai et al., 1988).
nobs_d0	Continuous	Number of sequence observations for a CG at day 0.

Abbreviations: CDR3, third complementarity determining region; RSV, respiratory syncytial virus.

taken from participants in an RSV challenge study where they were inoculated with the live virus and monitored for infections. The data set contains observations from 12 donors taken across two time points: days 0 (before inoculation) and 10. The data contain the immune receptor repertoire, the donor age group (i.e., young and old). For consistency with the COVID data set, we also collated features based on the first time point (Day 0) for predicting the response variable defined using the entire time course. Since for the antibody repertoire in the RSV data set the first sampled time point is before inoculation we can also investigate whether it is possible to predict CSR based on the variable region characteristics before an immune challenge.

## 4 | CLONAL GROUPS AS DATA OBJECTS

CSR is said to have occurred if a CG contains more than one class. Hence, the response variable of interest (i.e., the switching event) is observed at the level of the CG. As a consequence, the CG is the natural statistical unit to be considered when exploring and modeling CSR.

However, a CG is comprised of a set of sequences, so to develop a statistical or machine learning model to predict CSR (or lack thereof), we will need to summarize the distribution of the sequences in the CG into a vector of features that can become the input variables for the model. Each sequence in the CG is associated to both numeric and categorical variables—a selection can be seen in Table 1. The CGs are defined using the CDR3 which contains part of the V and J gene segment and all of the D gene segments (Lescale & Deriano, 2022). For the data sets used the CGs were already defined for the data in Stewart et al. (2022), by deriving pairwise comparisons of CDR3 nucleotide sequences using the Levenshtein distance. The resulting distance matrix was clustered using hierarchical clustering and cut at tree height of 0.05 to define CGs. It was further required that all observations within a CG have the same assigned germline V, D, and J genes, to minimize misattribution of sequences into CGs owing to mere CDR3 similarities. Thus within CGs sequence observations are highly similar. The categorical variables in Table 1 are identical across the clonal group.

For the numerical variables, one reasonable approach is to summarize their distribution within the CG with summary statistics of the distribution. It is expected that they too will be highly identical and display little dispersion across the CG. This was not the case. During the initial exploratory data analysis, we found a dispersion in the values within the CG



which was not expected given CGs are expected to have the same variable region as discussed in Section 2; skewness was close to zero and was not considered.

Given the variable region is much the same throughout the CG, the physicochemical properties of the sequence and the variable region mutation are not expected to display much, if any, variation. However, the data sets presented in Section 3 do contain some variation in the values across the CG. This is small and may either be attributed to measurement error by the CG classification procedure or this range could signify an inherent property of the CGs related to genuine diversification of sequence properties. We therefore decided to include the standard deviations of the variables in our features vector, to explore the possibility that the variable dispersion within the CG is predictive of CSR observation. The mean and the standard deviation were selected to represent location and dispersion of each numerical variable.

The categorical variables included in the feature vector are the variable region gene labels which define the CG, that is, V gene, J family, and D gene. The donor metadata have also been included in the feature vector. Care is then taken to remove the variables containing constant region information such as the proportion of class or subclass in each CG. As the CSR occurrence is defined in this instance by the presence of more than one class, these proportions would highlight a cross-class clonal group.

Typically, repertoire studies of immune response would sample antibody sequences at different time points with the expectation that CSR occurs postimmune challenge. As such it is of interest to use the features of the CG before the immune challenge to predict the outcome and investigate whether there are challenge-agnostic features that contribute to the probability of CSR.

## 5 | DATA PROCESSING

In this section, we describe the data processing steps that were carried out to get from the publicly available data described in Section 3 to the clonal-group level data sets used to train, validate, and test predictive methods in Section 7.

The data sets, as in Figure 1b, were processed on a per donor basis. They were filtered to only include heavy chain productive sequences (i.e., those encoding a full-length antibody). The data sets were then split by unique sequences; for antibody sequence data we expect to see many replicates of each unique sequence depending on the antibody expression level of the cell, these replicates were removed. Finally, singleton CGs (that is, those with only one sequence within the CG) were removed, since by definition singletons would not contain sequences of different isotypes. For singletons, it is not possible to define whether these clones are genuine negative samples in terms of CSR, or a false negative incidence owing to undersampling of sequences during data collection.

As explained in Section 4, the statistical unit of interest is the CG rather than individual sequences. Once the CGs have been defined for each donor, summary statistics from all the sequences in the CG were calculated for the continuous variables. The mean and the standard deviation were used as predictors. The categorical variables remain unchanged as they are either CG level descriptors, that is, variable region genes, or they are donor metadata. The resultant data tables in Figure 1c have as many rows as CG and as many columns as the predictors that describe each CG, as detailed in Section 4 and with the features in Table 1. The predictor matrices for the data sets in Section 3 have dimensions  $28,798 \times 65$  and  $55,232 \times 65$  for RSV and COVID, respectively.

## 6 | PREDICTIVE METHODOLOGY FOR CSR EVENTS

In this section, we detail the application of five existing statistical and machine learning methods to the processed data sets, for the prediction of CSR events and the procedures we are going to use to compare their performance. The response variable is binary and indicates whether or not a CG contains more than one antibody class and can be used as an indicator that a CSR event has occurred. This binary variable will be denoted as `IsCrossClass` as in Table 1. The statistical and machine learning models below will take variable region descriptors as predictors to model the binary outcome `IsCrossClass`. All the models were fitted using R statistical software version 4.1.1 (R Core Team, 2021), relying on `doParallel` (Microsoft Corporation & Weston, 2022a), and `foreach` (Microsoft Corporation & Weston, 2022b) to parallelize the computations. The individual packages used for each model will be detailed below.

**TABLE 2** The five models considered with the associated hyperparameters that need to be selected, the search scope for the hyperparameters and the model selection method used. Note that the g-mean here refers to the geometric mean.

Model	Hyperparameters	Search space	Selection method
LR	Number of features	Scope set to include all variables in Table 1.	AIC minimizing step function.
LASSO LR	$\lambda$	Set of values generated using <code>glmnet</code> package. $\{2^0 \dots 2^6\}$ , $\{250, 500 \dots 1000\}$ $\{10^0 \dots 10^5\}$ , $\{2^0 \dots 2^6\}$ , $\{250, 500 \dots 1000\}$ $[2^{-15}, 2^3]$ , $[2^{-4}, 2^{10}]$	The g-mean on the validation set prediction was used to select the optimal hyperparameter combination.
RF	<code>mtry</code> , <code>ntree</code>		
RF-NSx	<code>nodesize</code> , <code>mtry</code> , <code>ntree</code>		
SVM	$\gamma$ , $c$		

Abbreviations: AIC, Akaike information criterion; LASSO LR, LASSO logistic regression; LR, logistic regression; RF, random forest; RF-NSx, random forest with variable node size; SVM, support vector machine.

## 6.1 | Performance test detail

For models to be fitted and evaluated, we split the data sets into distinct train and test sets, respectively. After the training set is split from the test set, two different strategies, namely, all donor (AD) and leave-one-donor-out (LIDO), were used on the training set for the cross-validation (CV) hyperparameter tuning for the models (Table 2). For the AD strategy, the data for all donors are sampled across the folds used for CV. Conversely, the LIDO strategy results in a  $d$ -fold CV—where  $d$  is the number of donors in the training set—as one donor is separated out to be sampled solely for the validation set for each fold. The LIDO strategy was employed to explore the generalizability of the models across donors while combating the weakness of predictive models (Efron, 2020; Shmueli, 2010).

In the pursuit of reducing the effect of the data imbalance of the model training, the cross-class observations were randomly oversampled with replacement in the training sets to achieve a 50:50 data balance. The model performance has been judged using the unweighted average recall (UAR).

## 6.2 | Model selection and hyperparameter tuning

As introduced in Section 3, the models use the antibody repertoire features and the donor metadata, where available, to predict the `IsCrossClass` response variable. In this work, all models were trained using day 0 data. When necessary, hyperparameters were selected using grid search and 10-fold CV or  $d$ -fold CV on the training set for the AD and the LIDO methods, respectively, by maximizing the geometric mean  $g\text{-mean} = \sqrt{\text{sensitivity} \times \text{specificity}}$  (Kubat et al., 1998). The sensitivity refers to the correct identification of cross-class clones and the specificity indicates the correct identification of non-cross-class clones. The g-mean allows us to give equal weight to the classification of both classes as opposed to the precision or F1 score; we wish to understand the variables contributing to both a cross-class and non-cross-clonal group.

The variables for the LR model were selected by minimizing the Akaike information criterion (AIC, Akaike, 1974) on the training set and the categorical variables were one-hot encoded. We carried out “both” direction selection using the `step()` function in R to minimize AIC, starting with an intercept-only model  $M_0$  with no predictors. Variable selection was repeated for each iteration.

Generalized linear mixed effects model was considered to improve the performance of LR by accounting for donors heterogeneity. The same model selection described for the LR was used to select the fixed effect variables with random effect on the intercept dependent on the donor. Given the number of variables considered and the dearth of existing model selection methods for generalized linear mixed effect models, it was not feasible to consider further random effects. The results for this model are included in Tables S2 and S3 as well as the time taken to run the strategies in Table S6. The random effect on the intercept improves the AD strategy performance. However, it should be noted that this is not a totally fair comparison, since due to the donor-dependent random effect the model is given access to information not available to other methods, that is, to which donor each CG belongs. Indeed, the performance of the method plummeted when tested with the LIDO strategy, and it becomes comparable with LR. Also, the computational cost of fitting the model was prohibitive. For all these reasons, we decided to focus the analysis on the LR and LASSO LR.

The LASSO LR was used for the variable selection and parameter estimation of an LR model. The model was fitted using the `glmnet` R package (Friedman et al., 2010). For the choice of the penalty  $\lambda$ , we use the `lambda.1se` in order to

achieve the most parsimonious model whose average CV g-mean is within 1 standard error from the average CV g-mean of the best model found in the grid search as discussed in Breiman et al. (2017).

The `randomForest` R package (Liaw & Wiener, 2002) was used to build the RF models in Section 7. In `randomForest`, the `mtry` is the number of randomly drawn candidate variables out of which each split is selected when growing a tree; `nodesize` is the minimum number of observations in a terminal node; and `ntree` is the number of trees that are grown.

To optimize the performance of an RF, it is necessary to fine-tune the learning process of the model. Some hyperparameters control this learning process. Existing literature (Scornet, 2017) suggests practitioners to allow trees to grow to their maximum depth, as this is shown to result in a better performance. Also, the default values for the hyperparameters are believed to yield good predictive performance (Scornet, 2017) even in the absence of theoretical justification.

In this work, we optimize the `mtry` and `ntree` hyperparameters using grid search and 10-fold CV (AD) or  $d$ -fold CV. The default settings in the `randomForest` package (Liaw & Wiener, 2002) allow the trees to grow to maximum depth and it is often stated that for RF classifiers the effect of overfitting is rarely seen (Hastie et al., 2009). However, this was not our experience with antibody repertoire data, so `nodesize` was additionally tuned to investigate this—we will denote the resulting predictor as RF-NSx. In Section 7, we will present the results of the RF model with the default `nodesize` = 1 value as well as RF-NSx. The grid was expanded to include default values of the hyperparameters, thus `mtry` =  $\{2^0..2^6\}$ , `nodesize` =  $\{10^0..10^5\}$ , `ntree` =  $\{250, 500..1000\}$  where the defaults are  $\sqrt{P} \approx 8, 1,$  and 500, respectively— $P$  is the number of predictors.

The SVM models were fitted using the `e1071` R package (Meyer et al., 2023). There are two main hyperparameters that require tuning to optimize the performance of the SVM, the cost  $c$ , and  $\gamma$ . For the models in Section 7, both  $c$  and  $\gamma$  were optimized. As discussed in Hsu et al. (2008), exponentially growing sequences of  $c$  and  $\gamma$  were used to identify optimal hyperparameters and given the inherent independence of the grid search method it was possible to parallelize the computation. The grid search was conducted with  $c \in [2^{-4}, 2^{10}]$ ,  $\gamma \in [2^{-15}, 2^3]$ .

## 7 | EMPIRICAL ANALYSIS AND RESULTS

Two experiments were conducted on the data to assess the performance of LASSO LR, LR, RF, RF-NSx, and SVM, namely, the AD and the LIDO experiments. For the AD experiment, the data set was randomly split into  $k = 5$  folds, four folds (80%) were used as the training set and one fold (20%) was left as the test set; the fold selection was repeated until each fold was used as a test set. The AD strategy was then employed on the  $k$  training sets for tuning and fitting the model. The LIDO experiment separated out data from a donor for the test set and then proceeds with the implementation of the LIDO strategy on the training set comprised by the other donors. This is iterated over all donors until each is used for the test set. The UAR, sometimes referred to as the balanced accuracy (Brodersen et al., 2010), is used to assess the model performance. It is the average of the sensitivity and the specificity and it was chosen because it does not directly depend on the prevalence of either class.

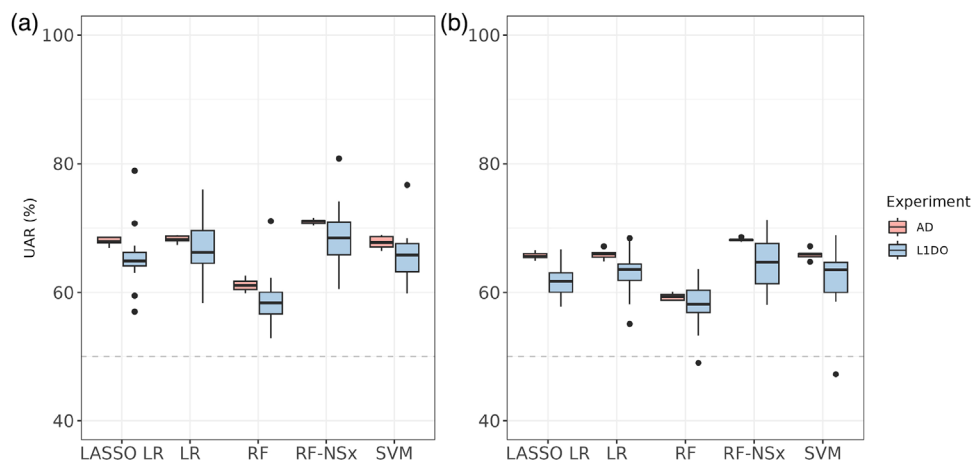
The purpose of the analysis is twofold. On the one hand, we wish to investigate if there are features in the variable region that contain information useful to predict the observation of a CSR event in the CG. On the other hand, we wish to compare the performance of the considered statistical and machine learning models in terms of predictive accuracy, generalizability to new individuals, and interpretability.

We evaluated the performance of the five models of interest on each of the data sets. For both the AD and LIDO test, the distribution of the UAR is presented in Figure 2 across both data sets and all iterations. The mean UAR of the LASSO LR, LR, RF, RF-NSx, and SVM performance in the AD test shown in Figure 2 across data sets is 67%, 67%, 60%, 70%, and 67%, respectively. For the LIDO test, it is 64%, 65%, 58%, 67%, and 64%, respectively.

For both data sets, the RF-NSx outperforms the RF in the AD experiment. The difference in performance is also visible in the LIDO experiment, but is more accentuated for the COVID data set. Looking at the value selected for `nodesize` in the RF-NSx model,  $10^3$  was the most common selection across experiments and data sets, chosen in 97% of cases. Moreover, Tables S4 and S5 (found in the Supporting Information) elucidate the overfitting occurring in the RF models compared to the RF-NSx. The difference between model fit to training data between the RF and RF-NSx is largest for COVID, where we expect to see more patient specificity as the data are taken during an immune response to a severe infection. The LIDO strategy allows for a model that is more robust against overfitting and is more generalizable which is vital for more complex data.

The highest mean UAR across the two data sets in the AD experiment is achieved by the RF-NSx model but there is negligible difference between LASSO LR, LR, SVM, and RF-NSx, as can be seen in Figure 2. However, the RF performance





**FIGURE 2** Boxplot of the unweighted average recall (UAR) performance of the LASSO logistic regression (LASSO LR), logistic regression (LR), random forest (RF), random forest with variable node size (RF-NSx), and support vector machine (SVM) in the all donor (AD) (red) and leave-one-donor-out (L1DO) (blue) tests across iterations for both the (a) COVID and (b) respiratory syncytial virus (RSV) data sets.

**TABLE 3** The performance of the LASSO logistic regression (LASSO LR) and random forest with variable node size (RF-NSx) based on a singular iteration of those shown in Figure 2.

Data	Model	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)	UAR (%)	G-mean (%)	AUC of ROC (%)
COVID	LASSO LR	73	85	23	35	79	79	84
	RF-NSx	91	70	16	27	81	80	87
RSV	LASSO LR	87	38	8	15	62	58	76
	RF-NSx	80	59	11	20	70	69	78

Abbreviations: AUC, area under the ROC curve; ROC, receiver operating characteristic; RSV, respiratory syncytial virus.

is notably worse than the other models. For the L1DO test, the RF-NSx performs best once more, yet there is strong overlap in the interquartile range with the LASSO LR, LR, RF-NSx, and SVM. The RF once more underperforms as compared to the other models and it, together with the SVM, are the only models with an experiment whose UAR is below 50% which is the baseline given the model solely predicts the majority class in an unbalanced data set. In addition, Table S6 elucidates the chasm in time spent for a single iteration of the model hyperparameter search and fitting for the SVM in comparison to the other models with as much as a magnitude difference compared to the closest model.

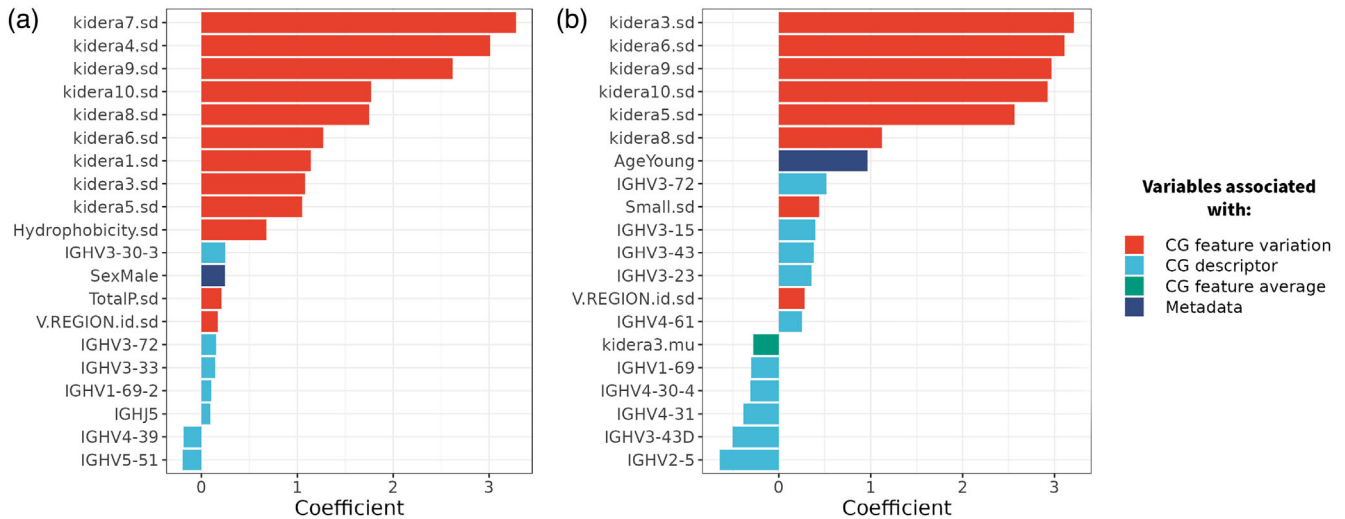
Overall, it is clear that regardless of the experiment type or model used, the features in the antibody repertoire contain information useful to distinguish between class homogeneous and heterogeneous CGs. We do also see that for all models the performance dips for the L1DO experiment with respect to the AD experiment.

## 7.1 | Model interpretation

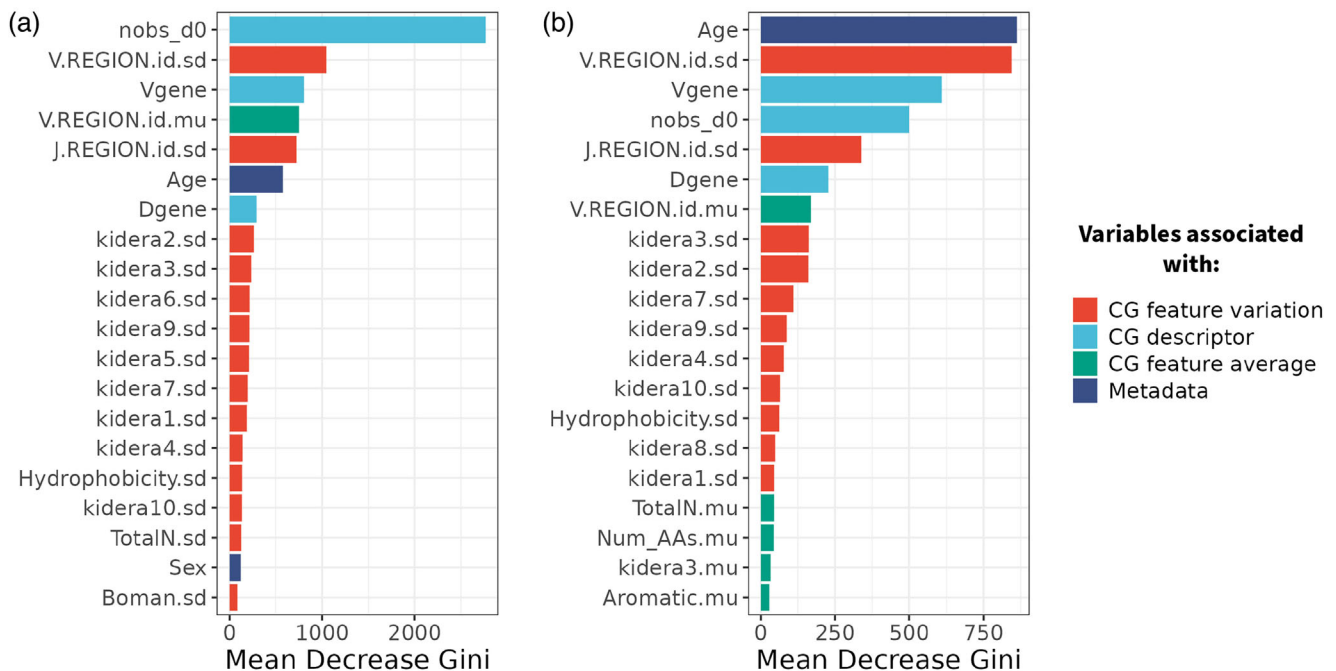
Since we have some evidence that the models are able to find predictive signal in the CG features, it would be useful to better understand where exactly this information comes from. To answer this question, we first investigate which features are selected for LASSO LR or deemed to be important for RF-NSx. Then, we wish to see if these features have a positive or negative association to observing a CSR event. We have selected both LASSO LR and RF-NSx as they display the best performance in Figure 2 and Table S3 as well as for their interpretability.

In order to compare these two models, we have used a single iteration of the L1DO strategy displayed in Figure 2. This means a single-fold combination was randomly selected for the train-test split to train, validate, and test. For both data sets, both LASSO LR and RF-NSx are trained and tested on the same data combination to generate the results in Table 3 and Figures 3–4.

In Figure 3, the estimates of the 20 features with the highest absolute value out of 34 nonzero coefficients for COVID and 57 for RSV are shown. We see that the variables associated with the variation of features within a CG are most impactful. This is more evident in the COVID data set. This implies that greater feature variation in the V region within a CG increases



**FIGURE 3** The LASSO logistic regression (LASSO LR) estimates of the 20 variables with the highest absolute values for (a) COVID (out of 34 nonzero coefficients) and (b) respiratory syncytial virus (RSV) (out of 57 nonzero coefficients). The  $\mu$  refers to the mean of the feature for a CG, that is,  $kidera6.\mu$  is the mean of the Kidera factor 6 for a CG. The  $sd$  refers to the standard deviation of a feature for a CG, that is,  $kidera1.sd$  is the standard deviation of the Kidera factor 1 for a CG. Features characterizing the CG itself (light blue), CG internal variation (red), CG average values (green), and the donor metadata (dark blue) have been highlighted.



**FIGURE 4** Gini index based random forest with variable node size (RF-NSx) variable importance rankings using `importance()` function in the `randomForest` R package (Liaw & Wiener, 2002). The set of variables shown is the top 20 variables with the highest mean decrease in node purity out of a total of 65 variables for (a) COVID and (b) respiratory syncytial virus (RSV). Features associated with characterization of the CG itself (light blue), CG internal variation (red), CG average values (green), and the donor metadata (dark blue) have been highlighted.

the odds of observing CSR. Particularly, for the RSV data set, we see that existing variation in physicochemical properties of the V region before a challenge distinguishes a resultant cross-class from a non-cross-class CG. As detailed in Table 1, the Kidera factors summarize different features of the protein structure, for our data set this specifically relates to the CDR3 loop, a vital component of the antibody for binding to antigens.

We do see a marked differences in the genes selected between the RSV and COVID data sets which may point to a certain level of challenge specificity in the antibody repertoire. Furthermore, in Figure S1A we can see that there are no genes that are consistently selected in both AD and LIDO strategy for the COVID data set which may be attributed to patient specificity during this immune challenge. In contrast, Figure S1B we see both IGHV1-69 and IGHV2-5 consistently selected for RSV in both AD and LIDO strategies and the direction of the impact is consistent. Together with Figure S1, we can see that some of the top features consistently selected are specific genes which may reflect gene preference relevant to the immune challenge, however the effect size is much smaller than the dispersion within a clonal group.

In Figures 3–4, both data sets show that the standard deviation of Kidera factor 3 is an indicator of an increase in the odds of a cross-class CG. This is also illustrated in Figures S1 and S2 where we see that *kidera3.sd* is shown to have a consistent positive association with CSR observation across iterations, data sets, and strategies. Kidera 3 refers to the extended structural preference, that is, a lower propensity to form compact, helical conformations. From Figure 3b, we see that the more extended (thus less helical) the average structure of the CDR3 is (*kidera3.mu*), the less likely it will be predicted as a cross-class CG. Interestingly, a more helical CDR3 loop is an indication of antigen specificity (Lowe et al., 2011). Thus, implying that the less antigen-specific a CG appears to be before a challenge, the lesser the odds of CSR postchallenge as the Kidera factor is calculated based on the constituents of the CG *before* the challenge in the RSV data set. This association holds for RSV across iterations and strategies.

Comparatively, the variable importance metric used in Figure 4 for the RF-NSx models is the mean decrease in the Gini index (Breiman et al., 2017) which relates to the node purity. This was calculated using the `importance()` function in the `randomForest` R package (Liaw & Wiener, 2002). A smaller value means a node contains more observations from a single class compared to a node with a larger value, hence the mean decrease looks to quantify the reduction in node impurity the variable's inclusion incurs. Unlike the estimated coefficients in LASSO LR, the mean decrease in Gini does not illuminate whether the feature has a positive or negative association with observing a cross-class clonal group.

We see in both Figures 3b and 4b that the age of the donor has an impact on the prediction of a cross-class CG. From the LASSO LR, we can see that being young increases the odds of predicting a cross-class CG. In addition, the V (VRI) and J region identity (JRI) can be used to probe the B-cell maturation effect within our models. We can see that the standard deviation of the VRI is positively associated with the odds of a cross-class CG for both the COVID and RSV data set in Figures 3–4. This may reflect that being amenable to CSR is associated with variable region diversification.

## 8 | DISCUSSION AND CONCLUSIONS

It appears clear from the results in Section 7 that there is indeed evidence that the CG features contain useful information to predict CSR events. Moreover, it can be seen that the performance of the five models differs between the AD and the LIDO experiments, confirming the expectation that generalizing to unseen individuals is a more challenging task, especially given the relatively small numbers of donors in the data sets.

The COVID data set overall resulted in models with a greater range of performance, particularly in the LIDO experiment. Unlike RSV where the data were collected as result of a controlled immune challenge experiment, COVID day 0 data have internal variation associated to the time of the individual immune response. The dissimilarity in the immune challenge progression is likely to have compounded the inherent donor specific variations and resulted in the higher variability in performance in the LIDO experiment as compared to the RSV data set. Nevertheless, the relative performance of the models is consistent across the data set, with the median performance dipping from AD to LIDO experiment and the RF scoring lowest among the methods.

It is also notable that the models trained on RSV resulted in greater overlap in the use of predictors between AD and LIDO, suggesting that the consistency in day 0 across donors may lead to models which are more suitable to generalization. Withal, both models highlight the importance of the number of sequences observed in the CG at day 0 and the V genes of the CG for the prediction performance. The number of sequences observed in the CG at day 0 is consistently shown to be associated with CSR but it has a greater effect for the COVID data set; as discussed in Section 2, CSR necessitates clonal expansion hence a larger clone in the midst of an immune response, as is the case for the donors in COVID, would be expected to have a higher likelihood of CSR.

Strikingly, the RF-NSx model agrees with LASSO LR about the importance of the standard deviation of the mutation level as well as the physicochemical properties suggesting an association between cross-class CGs and variation within a CG before a challenge. We are also able to identify the importance of Kidera factor 3 in the models trained on the RSV data set. However, solely the bidirectionality of the LASSO LR coefficients allows us to unearth the relationship between a less

helical CDR3 and the diminished odds of CSR. The prediction of CSR relies on the diversity found in the CG at day 0. This holds for both the models trained on data before the challenge (RSV) and within the challenge (COVID). We also see this in the tendency for the models to favor the standard deviations of the Kidera factors as opposed to their mean values.

Regarding the RF and RF-NSx, the literature tends to suggest to allow trees to grow to maximum depth, which does not seem to be optimal for this prediction problem. Indeed, the RF-NSx model outperforms RF for both tests and all iterations with a `nodesize`  $10^3$  greater than the RF. For data as complex and prone to patient specificity, it seems the shallower trees allow the RF-NSx model to better respond to unseen data and consequently be more generalizable. Furthermore, it seems clear that including `nodesize` in the hyperparameter search should be best practice given the impact this value has on the performance. For the SVM, we see similar performance to the LR model, but the additional computational cost makes it rather prohibitive for these complex data sets.

In conclusion, we have introduced a framework to predict CSR events from the antibody repertoire using CG collapse. The peculiarity of this problem is that CSR events can be observed only at the CG level so the set of antibody sequences in the CG needs to be summarized in a vector of features that can be used as input for statistical and machine learning models. We expect this kind of approach to be of interest to other tasks in Repertoire data analysis in immunology, including tackling the classification of single-cell RNA (scRNA) sequences that require CG level features for analysis. This method can be trained on bulk sequencing data, as done in this paper, and applied to the scRNA data as an imputation method to identify the likelihood that it belongs to a cross-class clonal group.

We have compared the performance of some popular predictive methods to address their strengths and weaknesses for this specific task and highlighted the importance of comprehensive hyperparameter tuning for RF models. We have identified a signal across models for the observation of CSR simply using the variable region as features suggesting an important association between the variable region and the constant region. From this point, it would be of great interest to see if we can extend this work to predicting the class proportions that would be seen post-CSR and which features contribute. We also discussed the differences between the observation of CSR and CSR, which we believe can be addressed through the construction of controlled experiments where CSR is induced using agents that are known to result in switching, thus we would have access to the entire CG rather than a sample.

## ACKNOWLEDGMENTS

This work was supported by the Biological Physics Across Scales Centre for Doctoral Training, funded by King's College London. The authors wish to acknowledge the King's Computational Research, Engineering, and Technology Environment (CREATE) (King's College London, 2022) platform for providing computational resources for this work. J. Ng and F. Fraternali are supported by the Biotechnology and Biological Sciences Research Council (BB/T002212/1). The funders had no role in the collection and analysis of the samples, in the interpretation of data, in writing the report, nor in the decision to submit the paper for publication.


## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Zenodo at <https://doi.org/10.5281/zenodo.5146019>.

## OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

## ORCID

Lutecia Servius  <https://orcid.org/0009-0004-4573-1912>

## REFERENCES

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.
- Battles, M. B., & McLellan, J. S. (2019). Respiratory syncytial virus entry and how to block it. *Nature Reviews Microbiology*, 17, 233–245.



- Becker, M., Dulovic, A., Junker, D., Ruetalo, N., Kaiser, P. D., Pinilla, Y. T., Heinzel, C., Haering, J., Traenkle, B., Wagner, T. R., Layer, M., Mehrlaender, M., Mirakaj, V., Held, J., Planatscher, H., Schenke-Layland, K., Krause, G., Strengert, M., Bakchoul, T., ... Schneiderhan-Marra, N. (2021). Immune response to SARS-CoV-2 variants of concern in vaccinated individuals. *Nature Communications*, *12*, 3109.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge.
- Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *Proceedings of the International Conference on Pattern Recognition* (pp. 3121–3124).
- Chan, A. C., & Carter, P. J. (2010). Therapeutic antibodies for autoimmunity and inflammation. *Nature Reviews Immunology*, *10*, 301–316.
- Efron, B. (2020). Prediction, estimation, and attribution. *International Statistical Review*, *88*, S28–S59.
- Feldkamp, C. S., & Carey, J. L. (1996). Immune function and antibody structure. In *Immunoassay* (pp. 5–24). Elsevier.
- Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22.
- Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351.
- Hajipour, F., Jozani, M. J., & Moussavi, Z. (2020). A comparison of regularized logistic regression and random forest machine learning models for daytime diagnosis of obstructive sleep apnea. *Medical & Biological Engineering & Computing*, *58*, 2517–2529.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer Series in Statistics. Springer New York.
- Horst, A., Smakaj, E., Natali, E. N., Tosoni, D., Babrak, L. M., Meier, P., & Miho, E. (2021). Machine learning detects anti-DENV signatures in antibody repertoire sequences. *Frontiers in Artificial Intelligence*, *4*, 715462.
- Horton, M. B., Cheon, H., Duffy, K. R., Brown, D., Naik, S. H., Alvarado, C., Groom, J. R., Heinzel, S., & Hodgkin, P. D. (2022). Lineage tracing reveals B cell antibody class switching is stochastic, cell-autonomous, and tuneable. *Immunity*, *55*, 1843–1855.
- Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2008). A practical guide to support vector classification. *BJU International*, *101*, 1396–1400.
- Hu, B., Guo, H., Zhou, P., & Shi, Z.-L. (2021). Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology*, *19*, 141–154.
- Janeway, C. A., Travers, P., Walport, M., & Shlomchik, M. J. (2001). *Immunobiology: The immune system in health and disease* (5th ed.). Garland Science.
- King's College London. (2022). *King's Computational Research, Engineering and Technology Environment (CREATE)*. <https://doi.org/10.18742/rnvf-m076>
- Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, *30*, 195–215.
- Lescale, C., & Deriano, L. (2022). V(D)J recombination: Orchestrating diversity without damage. In *Reference module in life sciences*. Elsevier.
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*, 18–22.
- Lowe, D. C., Gerhardt, S., Ward, A., Hargreaves, D., Anderson, M., Ferraro, F., Pauptit, R. A., Pattison, D. V., Buchanan, C., Popovic, B., Finch, D. K., Wilkinson, T., Sleeman, M., Vaughan, T. J., & Mallinder, P. R. (2011). Engineering a high-affinity anti-IL-15 antibody: Crystal structure reveals an  $\alpha$ -helix in VH CDR3 as key component of paratope. *Journal of Molecular Biology*, *406*, 160–175.
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2023). e1071: Misc functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. <https://CRAN.R-project.org/package=e1071>
- Weston, S., & Microsoft Corporation. (2022a). *doParallel: Foreach parallel adaptor for the "parallel" package*.
- Weston, S., & Microsoft Corporation. (2022b). *foreach: Provides Foreach looping construct*.
- Nakai, K., Kidera, A., & Kanehisa, M. (1988). Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Engineering, Design and Selection*, *2*, 93–100.
- Ng, J. C. F., Montamat Garcia, G., Stewart, A. T., Blair, P., Mauri, C., Dunn-Walters, D. K., & Fraternali, F. (2023). sciCSR infers B cell state transition and predicts class-switch recombination dynamics using single-cell transcriptomic data. *Nature Methods*, <https://doi.org/10.1038/s41592-023-02060-1>
- Pino-Mejías, R., Carrasco-Mairena, M., Pascual-Acosta, A., Cubiles-De-La-Vega, M. D., & Muñoz-García, J. (2008). A comparison of classification models to identify the Fragile X Syndrome. *Journal of Applied Statistics*, *35*, 233–244.
- R Core Team. (2021). *R: A language and environment for statistical computing*.
- Scornet, E. (2017). Tuning parameters in random forests. *ESAIM: Proceedings and Surveys*, *60*, 144–162.
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*, 289–310.
- Stavnezer, J., & Schrader, C. E. (2014). IgH chain class switch recombination: Mechanism and regulation. *Journal of Immunology*, *193*, 5370–5378.
- Stewart, A., Sinclair, E., Ng, J. C. F., O'Hare, J. S., Page, A., Serangeli, I., Margreitter, C., Orsenigo, F., Longman, K., Frampas, C., Costa, C., Lewis, H.-M., Kasar, N., Wu, B., Kipling, D., Openshaw, P. J., Chiu, C., Baillie, J. K., Scott, J. T., ... Dunn-Walters, D. K. (2022). Pandemic, epidemic, endemic: B cell repertoire analysis reveals unique anti-viral responses to SARS-CoV-2, Ebola and Respiratory Syncytial Virus. *Frontiers in Immunology*, *13*, 807104.
- Su, C., Lua, W.-H., Ling, W.-L., & Gan, S. (2018). Allosteric effects between the antibody constant and variable regions: A study of IgA Fc mutations on antigen binding. *Antibodies*, *7*, 20.
- Tudor, D., Yu, H., Maupetit, J., Drillet, A.-S., Bouceba, T., Schwartz-Cornil, I., Lopalco, L., Tuffery, P., & Bomsel, M. (2012). Isotype modulates epitope specificity, affinity, and antiviral activities of anti-HIV-1 human broadly neutralizing 2F5 antibody. *Proceedings of the National Academy of Sciences*, *109*, 12680–12685.



- Wei, J., Matthews, P. C., Stoesser, N., Maddox, T., Lorenzi, L., Studley, R., Bell, J. I., Newton, J. N., Farrar, J., Diamond, I., Rourke, E., Howarth, A., Marsden, B. D., Hoosdally, S., Jones, E. Y., Stuart, D. I., Crook, D. W., Peto, T. E. A., Pouwels, K. B., . . . Kavanagh, J. (2021). Anti-spike antibody response to natural SARS-CoV-2 infection in the general population. *Nature Communications*, *12*, 6250.
- Widrich, M., Schäfl, B., Ramsauer, H., Pavlović, M., Gruber, L., Holzleitner, M., Brandstetter, J., Sandve, G. K., Greiff, V., Hochreiter, S., & Klambauer, G. (2020). Modern Hopfield networks and attention for immune repertoire classification. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems 33*. Neural Information Processing Systems Foundation, Inc. (NeurIPS).
- Zhao, J., Nussinov, R., & Ma, B. (2019). Antigen binding allosterically promotes Fc receptor recognition. *mAbs*, *11*, 58–74.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Servius, L., Pigoli, D., Ng, J., & Fraternali, F. (2024). Predicting class switch recombination in B-cells from antibody repertoire data. *Biometrical Journal*, *66*, 2300171.  
<https://doi.org/10.1002/bimj.202300171>