Profiling the baseline performance and limits of machine learning models for adaptive immune receptor repertoire classification

Chakravarthi Kanduri [1,*], Milena Pavlović [1], Lonneke Scheffer [1], Keshav Motwani [2], Maria Chernigovskaya [3], Victor Greiff [3] and Geir K. Sandve [1,*]

[1]Centre for Bioinformatics, Department of Informatics, University of Oslo, Oslo 0373, Norway
[2]Department of Pathology, Immunology and Laboratory Medicine, University of Florida, FL 32610, USA
[3]Department of Immunology and Oslo University Hospital, University of Oslo, Oslo, 0372, Norway
*Correspondence address. Chakravarthi Kanduri, Centre for Bioinformatics, Department of Informatics, University of Oslo, Oslo, 0373, Norway. E-mail: skanduri@uio.no; Geir K. Sandve, Centre for Bioinformatics, Department of Informatics, University of Oslo, Oslo, 0373, Norway. E-mail: geirksa@ifi.uio.no.

Abstract

Background: Machine learning (ML) methodology development for the classification of immune states in adaptive immune receptor repertoires (AIRRs) has seen a recent surge of interest. However, so far, there does not exist a systematic evaluation of scenarios where classical ML methods (such as penalized logistic regression) already perform adequately for AIRR classification. This hinders investigative reorientation to those scenarios where method development of more sophisticated ML approaches may be required.

Results: To identify those scenarios where a baseline ML method is able to perform well for AIRR classification, we generated a collection of synthetic AIRR benchmark data sets encompassing a wide range of data set architecture-associated and immune state–associated sequence patterns (signal) complexity. We trained ≈1,700 ML models with varying assumptions regarding immune signal on ≈1,000 data sets with a total of ≈250,000 AIRRs containing ≈46 billion TCRβ CDR3 amino acid sequences, thereby surpassing the sample sizes of current state-of-the-art AIRR-ML setups by two orders of magnitude. We found that L1-penalized logistic regression achieved high prediction accuracy even when the immune signal occurs only in 1 out of 50,000 AIR sequences.

Conclusions: We provide a reference benchmark to guide new AIRR-ML classification methodology by (i) identifying those scenarios characterized by immune signal and data set complexity, where baseline methods already achieve high prediction accuracy, and (ii) facilitating realistic expectations of the performance of AIRR-ML models given training data set properties and assumptions. Our study serves as a template for defining specialized AIRR benchmark data sets for comprehensive benchmarking of AIRR-ML methods.

Keywords: machine learning, benchmarking, baseline performance, adaptive immune receptor repertoires, AIRR, ML

Background

The adaptive immune system is responsible for mounting a tailored immune response against antigens (viruses, bacteria, cancer, self-antigens). The adaptive immune receptors (AIRs) expressed on the cell surface of T cells and B cells recognize and bind antigens [1]. To cover a broad space of potential antigens, AIRs maintain high diversity throughout an individual's lifetime by a stochastic process called V(D)J recombination [2–4]. For instance, in humans, the expected number of unique T-cell receptors (TCRs) is between 10^7 and 10^8, sampled from a set of $>10^{14}$ potential TCRs [5]. Upon antigen encounter, adaptive immune cells are activated and proliferate, with all daughter cells inheriting the same antigen-specific AIR sequence (clonal expansion). After clearance of the antigen, a fraction of the activated adaptive immune cells matures to a memory stage constituting long-term protection against antigen reexposure [6]. Therefore, a snapshot of the adaptive immune receptor repertoire (AIRR) by immune repertoire sequencing captures information on the current and past immune state of an individual, where patterns corresponding to a specific antigenic response may be traced [7–11].

Previous studies have shown that identical or similar B-cell receptors (BCRs) or TCRs (where similarity may be defined by edit distance or sharing of subsequence motifs) can be observed in multiple individuals who share a similar disease or phenotype [12–17]. Such evidence has become the basis for many further studies that developed machine learning (ML) methods to predict the immune states of individuals based on antigen-specific signatures recorded in AIRR data [18, 19]. Methods that use both classical ML and deep learning continue to emerge [20], with the aim of establishing AIRR-ML models for clinical diagnostics [21–34]. Most published methods use either the nucleotide or the amino acid sequence of the complementarity determining region 3 (CDR3) to search for and learn the antigen-specific patterns since the CDR3 loops of AIRs are known to be key determinants of the antigen specificity of AIRs [11, 35, 36]. Some of the published ML methods [23, 28, 32] aptly considered AIRR classification as a multiple instance learning problem (MIL) [37] consisting of repertoires as bags, receptors as instances, and immune state–associated receptors as witnesses (see Supplementary Fig. 1 for an illustration of AIRR-ML as MIL). One of the main challenges for AIRR-ML methodology is that the patterns associated with an immune

state (positive instances in MIL terminology) can be as rare as one antigen-binding AIR per million lymphocytes [19, 38]. A similarly low incidence of positive instances was also reported in previous studies that used ML for the classification of immune states based on AIRR data (e.g., [23, 26]). However, notably, the reported degree of the rareness of identified positive instances can vary between immune states as well as based on the sampling depth and statistical power of the studies.

Given the continued rise in the development and application of ML methods for immune state prediction, it is imperative to understand the capabilities and limits of ML methods applied to AIRR data sets. However, profiling the baseline performance of AIRR-ML methods requires a large suite of benchmark data sets representing a wide range of variable properties of AIRR data sets and immune signals. Although experimental AIRR data sets with large sample sizes are being generated occasionally (e.g., the recent ImmuneCODE database [39] containing SARS-CoV-2–specific TCR data sets), very few experimental studies have generated repertoire-labeled AIRR data sets at a high resolution and with a repertoire size >500 [26, 39, 40]. While the limited availability of experimental data sets is a challenge in itself in establishing a baseline performance of AIRR-ML methods, the biggest challenge is the lack of ground truth related to immune state–associated immune signals. Here, the immune state–associated immune signal refers to the pattern or information encoded in the AIRR data set that allows differentiating between two or more immune states (hereafter referred to as *immune signal* or *signal* for brevity). In the case of AIRR classification using CDR3 sequences, there exists little concrete, consensus knowledge on the size, shape, incidence levels, and diversity of immune state–associated patterns [11, 23]. In other words, it is not known whether the immune state–associated pattern will be in the form of full CDR3 sequences [26, 40, 41] or short subsequences (k-mers) [11, 23, 42–44], how large and diverse the pool of immune state–associated patterns is, and how frequently the immune state–associated patterns occur in repertoires characterized by a particular immune state. In the absence of knowledge on ground truth immune signals in experimental data sets, an alternative approach to overcoming these challenges is to use simulated data sets, where the data set and signal properties can be controlled while preserving and reflecting the complexities of experimental data sets to a large extent [45, 46].

In this study, we aimed to profile the limits and capabilities of baseline AIRR-ML models in predicting the immune state labels of repertoires (Fig. 1a). The generation of simulated benchmark data sets for such an endeavor should cover a diverse set of challenges representing characteristics of AIRR-ML study designs, immune state–associated signal assumptions, and ML model assumptions (hereafter collectively referred to as properties of AIRR-ML model training setup). To this end, we simulated a large suite of distinct benchmark data sets ($n \approx 1,000$ data sets with a total of $\approx 250,000$ repertoires containing approximately 46 billion TCRβ CDR3 amino acid sequences), in which the data set and signal properties were varied (Fig. 1c). To ensure nativeness of the simulated AIR sequences in terms of positional biases, amino acid usage, and sequence length distributions, we generated repertoires according to a human VDJ recombination model provided by AIRR simulation suite OLGA [47]. We then varied (i) immune signal properties (described below), (ii) sample size (number of *examples* available for training), (iii) repertoire size (number of sequences in each repertoire), (iv) class balance (balance between positive and negative class *examples* in training data set), and (v) noise in the negative class (signal incidence in negative class). Note that the italicized term *examples* commonly used in ML literature refers to repertoires throughout this article. Also, note that the term *positive class* refers to those repertoires that contain immune–state associated sequence patterns, whereas the term *negative class* refers to those repertoires where the signal can occur by chance. The immune signal properties that were varied in the benchmark data set simulations include (i) witness rate (the rate at which signal occurs in the positive class *examples*) (Fig. 1b), (ii) number of k-mers (also referred to as motifs) constituting the signal, (iii) the size of signal motifs, (iv) whether the signal motifs are continuous, and (v) distributional shift (difference in the witness rates of training data sets and future data sets met by the trained model). We then trained $\approx 1,700$ ML models with varying assumptions and complexity to profile the limits and scalability of AIRR-ML models. We found that even a baseline method such as logistic regression performed surprisingly well at a witness rate as low as 0.002%, comparable to a level of difficulty observed in AIR-based disease studies [38]. We also characterized several scenarios with increasing levels of signal complexities, in which logistic regression failed to learn the true signal and thus exhibited poor prediction performance. Overall, our findings shed light on the immune signal complexities and the basic data set properties that can pose a challenge to baseline ML models, thereby identifying the frontier where the development of novel methodologies by the AIRR-ML community is needed.

Analyses

The main goal of this study is not to comprehensively assess and benchmark the state-of-the-art machine learning methods for AIRR data set classification but rather provide empirical evidence on the performance levels of baseline ML models across a diverse set of challenges for AIRR classification. To this end, we profiled both the capabilities and limits of the models with varying data set and signal properties (Fig. 1). We replicated each investigation three times on separate data sets, and the average performance metric of the three replications together with variation is reported. The performance metric of each replication is the balanced accuracy ($\frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$) obtained through nested cross-validation in which 80% of the data is used for model training and hyperparameter tuning, while 20% is used for assessing prediction performance on independent test data. An overview of the variable properties of the AIRR-ML training setup assessed in this study and their corresponding benchmark data sets is shown in Table 1, Fig. 1c, and Supplementary Table 1.

Impact of k-mer implantation on the background k-mer frequency distributions

Previous studies have shown that most of the possible contacts between TCR and peptide antigens were made through only short and typically contiguous stretches of amino acid residues of CDR3s (IMGT positions 107–116) [26, 36]. Previously developed ML methods for receptor specificity (or receptor publicity) prediction or repertoire classification have used such evidence as a premise and often assumed that immune state–associated sequence patterns are short motifs (e.g., 2-mers, 3-mers, 4-mers, and 5-mers [21, 23, 27, 49]) while other studies considered the entire AIR sequence [26]. Here, we align with previous observations and profile the baseline performance of AIRR-ML models under the assumption that the immune state–associated sequence patterns are short motifs (k-mers). Therefore, to simulate immune state–associated signals in the construction of benchmark data sets,
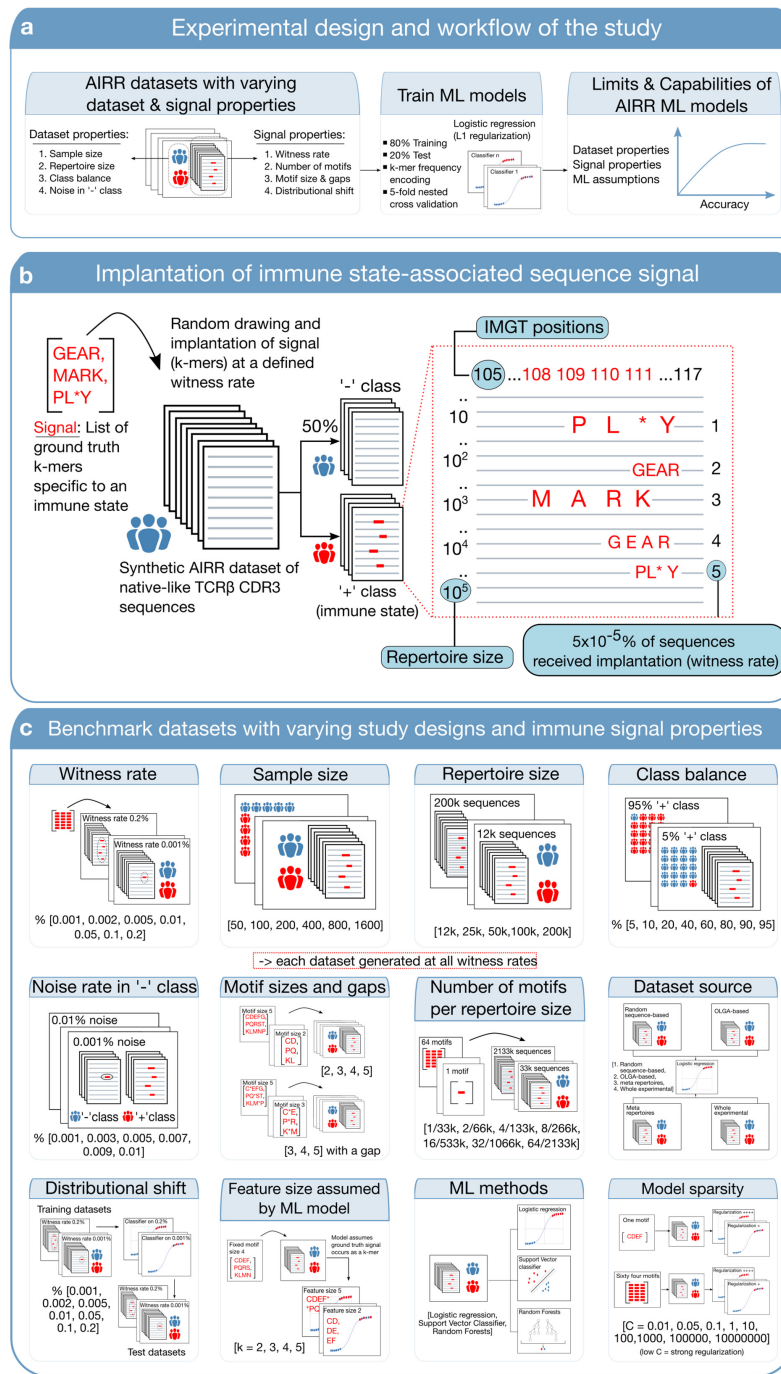
Figure 1: (a) Experimental design and workflow of the study. The overarching objective of this study is to profile and establish the limits and capabilities of baseline machine learning (ML) models for adaptive immune receptor repertoire (AIRR) classification. To meet this aim, we evaluate the performance of a baseline logistic regression model across multiple variations of AIRR data set properties (sample size, repertoire size, class balance, noise in negative class) and immune signal properties (witness rate, number of motifs, motif sizes and gaps, distributional shift). (b) An example of how immune state–associated signal is implanted into AIRR data sets: to mimic the realistic nature of AIRR data sets, we generated synthetic AIRR reference data sets (specifically, TCRβ CDR3 amino acid sequences) according to the VDJ recombination model provided by the AIRR simulation suite OLGA [47]. From a list of motifs, we randomly drew to implant motifs at a defined rate in a defined percentage of repertoires. Note that the wildcard character $*$ in the list of motifs refers to a gap, where any amino acid could occur with equal probability. The repertoires in which motifs are implanted are referred to as positive class and thus contain the patterns associated with a hypothetical immune state. The standard coordinates for CDR3 sequences according to the IMGT numbering system are positions 105–117. Here, IMGT positions [48] refer to the unique numbering system of the ImMunoGeneTics database that positions amino acids in a protein sequence in such a way that facilitates easy comparison of sequences irrespective of the antigen type, chain type, and species. We implant motifs only in the IMGT positions 108–111 with equal probability to not disrupt the positional biases inherent to the germline-encoded start and end portions of CDR3 sequences. In this illustration, a total of 5 sequences out of a repertoire size of 10^5 sequences received motif implantation. Each positive class repertoire receives a signal at this rate (5×10^{-5}% of sequence; referred to as witness rate throughout this article). (c) Construction of benchmark data sets with varying data set and signal properties. Each signal and data set property are varied at different rates (shown in square brackets for each property), resulting in multiple separate data sets. For the properties shown within a red-dotted boundary, each separate data set was further generated at all the witness rates explored (top left). Table 1 provides a granular overview of the benchmark data set suite.

Table 1: Overview of the variable properties of the AIRR-ML training setup assessed and their corresponding benchmark data sets

Assessed property pertaining to data set or signal or ML model	Range of values or categories at which the assessed property is varied	Total number of simulated data sets	Total number of repertoires across data sets	Average repertoire size across all data sets	Total number of sequences across all data sets (in millions)	Total number of ML models trained
Sample size (number of examples in training data) [Fig. 3a]	50, 100, 200, 400, 800, 1,600	126	66,150	100,000	6,615	126
Repertoire size (number of sequences in each repertoire) [Fig. 3b]	12k, 25k, 50k, 100k, 200k	105	21,000	77,400	1,625	105
Number of implanted k-mers constituting the signal relative to increasing repertoire size [Fig. 3c]	1/33k, 2/66k, 4/133k, 8/266k, 16/533k, 32/1,066k, 64/2,133k	147	29,400	846,658	24,892	147
The size of implanted k-mers excluding a gap position [Fig. 5a]	2, 3, 4, 5	84	16,800	100,000	1,680	84
The size of implanted k-mers including a gap position [Fig. 5b]	3, 4, 5	63	12,600	100,000	1,260	63
The feature size assumed by the model [Fig. 5c]	2, 3, 4, 5	21	4,200	100,000	420	84
Noise in the negative class [Fig. 5d]	0.001%, 0.003%, 0.005%, 0.007%, 0.009%, 0.01%	18	3,600	100,000	360	18
Regularization strength of the ML model (constant C indicating penalty level in logistic regression of scikit-learn) trained on data sets with two different signal definitions: 1 motif or 64 motifs [Fig. 2a, b]	0.01, 0.05, 0.1, 1, 10, 100, 1000, 100 000, 10 000 000	42	8,400	100,000	840	378
The witness rate in an unseen test data set [Fig. 4a]	0.001%, 0.002%, 0.005%, 0.01%, 0.05%, 0.1%, 0.2%	0	0	0	0	0
The percentage of positive examples in a repertoire data set (class balance) [Fig. 4b]	5%, 10%, 20%, 40%, 60%, 80%, 90%, 95%	168	33,600	100,000	3,360	168
Source of data set construction [Supplementary Fig. 9]	Random amino acid sequences, OLGA, meta repertoires, experimental repertoires	63	12,600	100,000	1,260	84
Machine learning model [Supplementary Fig. 10a]	Logistic regression, support vector classifier, random forest	0	0	0	0	42
L1-regularization strength of the SVC model (constant C indicating penalty level in scikit-learn implementation) [Supplementary Fig. 10b]	0.01, 0.05, 0.1, 1, 5, 10, 50, 100, 1,000, 100,000, 10,000,000	0	0	0	0	77
Number of estimators (hyperparameter) of random forest model [Supplementary Fig. 10c]	1, 2, 5, 10, 25, 50, 100, 200, 500, 1,000, 2,000, 5,000	0	0	0	0	84
Number of implanted subsequences (for comparison of feature selection-aided classifier with L1-penalized logistic regression) [Supplementary Fig. 11a, b]	4, 16, 64	63	12,600	100,000	1,260	126
Relation between sample size and the expected number of nonzero coefficients [Supplementary Fig. 12a]	Number of examples in minority class (20, 40, 80, 160, 320);the number of expected nonzero coefficients (2, 4, 8, 16)	60	18,600	100,000	1,860	60
Relation between sample size and the total number of predictors [Supplementary Fig. 13a]	Number of examples in minority class (20, 40, 80, 160); the number of predictors in log scale (6, 8, 10, 12)	48	9,000	100,000	900	48
Total		1,008	248,550	2,124,058	46,332	1,694

we implanted k-mers into the synthetic AIRR reference data sets (Fig. 1b).

We first quantitated the degree to which the implantation of 4-mers affects the background 4-mer frequency distributions of synthetic native-like AIRR data sets (Supplementary Fig. 2; see Methods for details). We focused on 4-mers for this investigation because we used 4-mers as the signal definition in a large majority (≈90%) of the benchmark data sets in this study. We observed that when 4-mers were implanted at lower witness rates (up to 10 of 100,000 sequences receiving implantation), only the implanted 4-mers exhibited significant differences in background frequency distributions. As more sequences received implantation (with an increase of the witness rate), the number of 4-mers that were significantly affected by overlapping partially with the implanted motifs increased. On average, per each implanted 4-mer, a total of approximately thirty-five 4-mers that overlapped three residues with implanted motifs and between ten and thirty 4-mers that overlapped two amino acid residues were significantly disturbed. Very few 4-mers (<4) that had zero amino acids overlapping with the implanted 4-mers exhibited differences in background frequency distributions across three independent replications.

The impact of model sparsity on prediction performance is dependent on the witness rate and the immune signal definition

The k-mer frequency encoding of AIRR data sets results in high dimensionality. For instance, decomposing a repertoire with 100,000 unique CDR3 sequences into 4-mers would result in approximately 160,000 unique 4-mers (with 20 amino acid residues that can occur at any of the four positions of a 4-mer, the total number of possible k-mers is $20^4 = 160,000$). Regularization or shrinkage is useful in high-dimensional problems to avoid overfitting and to improve the generalizability of models. We used scikit-learn's [50] implementations of L1-penalized logistic regression models throughout this study, where the hyperparameter controlling regularization strength is indicated by a variable C. Smaller values of C represent stronger regularization (see Methods section for details). In order to narrow down an appropriate parameter space for the regularization constant for all the benchmarks in this study, we first evaluated how the prediction performance of models scales with increasing regularization strength. To this end, we explored the impact of regularization strength in two different scenarios with distinct signal definitions, where the defined witness rates are composed of (i) only a single motif and (ii) multiple motifs ($n = 64$). At a high witness rate of 0.2%, the signal was so strong that a strong regularization (low C) was not particularly needed to attain higher accuracies (Fig. 2). This was irrespective of whether the witness rate was composed of a single motif or multiple motifs. As the witness rate decreased, an increase in accuracy was observed with increased regularization strength. When the immune signal was composed of a single motif, a strongly regularized model (C = 0.05) was able to classify almost perfectly (99% accuracy) at a witness rate of 0.005% (5 of 100,000 sequences containing a motif) and performed decently well (82% accuracy) even at a witness rate of 0.002% (2 of 100,000 sequences containing a motif). Even at a low witness rate of 0.001%, the strongly regularized model performed better than a random prediction by 10 percentage points (Fig. 2a). However, when the witness rate was composed of 64 motifs, the models did not exhibit good performance at the lower witness rates (1, 2, 5, 10 sequences of 100,000 containing a motif), although performed well with a strongly regular-

ized model on witness rates from 0.05% and upward. For both scenarios (Fig. 2a and b), the implanted motifs were ascribed higher weights when the models obtained decent performance (Supplementary Fig. 3 and Supplementary Fig. 4). In the case of poor performance, the weights for features corresponding to the implanted motifs were indistinguishable from those of other features (Supplementary Fig. 3 and Supplementary Fig. 4). This confirms that the implanted motifs were indeed the basis for the prediction performance of the models. We further found that the motifs that were ascribed higher weights by the model, in this case, were the very same motifs that exhibited significant differences in motif frequency distributions as observed through an independent statistical analysis (performed similarly to the previous section) (Supplementary Fig. 5 and Supplementary Fig. 6). This suggests that the users of AIRR classification methods who use a k-mer frequency encoding can use a univariate statistical test either as a feature selection tool or as an additional diagnostic to confirm the basis of classification.

Comparing the frequency distributions of both positive and negative class examples of both scenarios provided additional hints on why a signal definition composed of 64 motifs was a challenging problem to learn for the model (Supplementary Fig. 7 and Supplementary Fig. 8). When the signal was composed of 64 motifs, each of the motifs had an equal probability to occur in the sequences of the positive class. At a witness rate of 0.002%, in the single motif scenario, the same motif will occur in 2 of 100,000 sequences, whereas in the 64-motif scenario, any 2 of the 64 motifs will be implanted in each example. This would not only result in less overlap in the signal definition across positive class examples but also would restrict the implanted motifs from being over-represented relative to baseline frequencies, thus making the classes indistinguishable (Supplementary Fig. 8). Such a scenario can arise in experimental studies because of differences in sequencing depth, where the likelihood of detecting more phenotype-associated sequences increases with sequencing depth. These findings indicated that the prediction performance generally improved with model sparsity, although the model struggled at any regularization level as the number of receptors with implanted k-mers went below 2 of the 100,000 receptors for each repertoire.

A moderate increase in sample size is not sufficient for substantial performance gains

After choosing an appropriate hyperparameter interval for regularization strength (0.05, 0.1, 1, 5), we set out to understand if and how the prediction performance scales with increased sample size (number of repertoires available for training). In the previous section, we observed a decent prediction performance when a single motif was implanted at 0.002% (Fig. 2). To make the prediction problem slightly more complex, we used a signal definition composed of three motifs in the following experiments unless otherwise stated. This would mean that at 0.002% witness rate, any two of the three motifs will be implanted independently in each of the positive class examples. To obtain empirical sample complexity estimates, we evaluated the performance with different sample sizes at different witnesses rates (Fig. 3a). We observed that at higher witness rates, starting from 0.01% (10 of 100,000 sequences containing a motif) and above, even a smaller sample size of 50 repertoires was sufficient to reach a decent performance level (78% accuracy). At the lowest witness rate of 0.001%, the explored sample sizes did not provide evidence of a gain in performance even at a sample size of 1,600 repertoires. At lower wit-
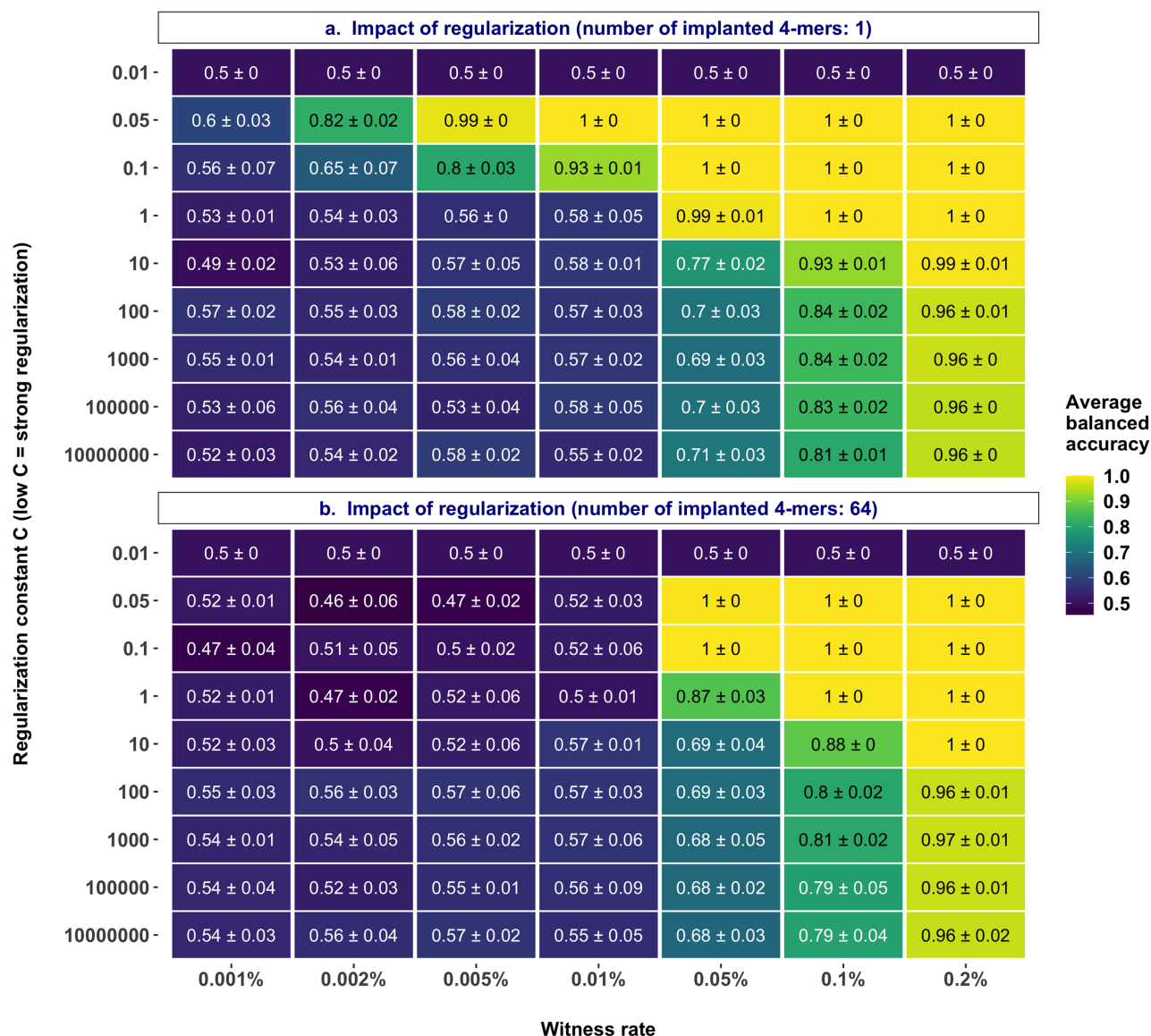
Figure 2: Model sparsity affects prediction performance: performance estimates of a logistic regression model regularized with a fixed regularization constant C (y-axis) in a binary classification of a balanced, labeled adaptive immune receptor repertoire data set of 200 repertoires with 100,000 amino acid sequences each where the signal in positive class examples composed of 4-mers is known at the explored witness rates (x-axis). The smaller the regularization constant C, the stronger the regularization. (a) Impact of regularization strength when the signal definition is composed of a single 4-mer. (b) Impact of regularization strength when the signal definition is composed of sixty-four 4-mers. The mean balanced accuracy of a fivefold cross-validation was computed in three independent replications. The color coding shows the mean and standard deviation of the performance estimate obtained by three independent replications.

ness rates like 0.002% and 0.005%, there was no substantial gain in performance beyond a sample size of 200—at a witness rate of 0.002%, the sample size had to increase by eightfold in order to achieve a performance gain of 5 percentage points. Overall, these findings indicated that a moderate increase in sample size was not sufficient to obtain a substantial gain in the prediction performance. Although there was an improvement in prediction performance at lower witness rates, the improvement was moderate and gradual rather than a steep increase.

Classification performance improves with increasing repertoire size

We then set out to understand if prediction performance would improve with an increase in repertoire size (number of TCR sequences per repertoire). For this, we evaluated the prediction per-

formance with different repertoire sizes at different witness rates (Fig. 3b). We observed that at higher witness rates, including and beyond 0.05%, the models obtained good performance even at the lowest repertoire size that we explored. At moderate to low witness rates (0.002%, 0.005%, and 0.01%), increasing the repertoire size resulted in improved performance by more than 10 percentage points. This was because the increased repertoire size led to a larger absolute count of motifs or sequence patterns associated with the positive class in each repertoire. For instance, at a witness rate of 0.002% in our experiments, a repertoire size of 100,000 would on average carry two sequences that contain a phenotype-associated motif, whereas a repertoire size of 200,000 would carry four sequences that contain the motif. Since the signal definition in this experiment was composed of three motifs, it is more likely that a larger portion of signal definition will be included in
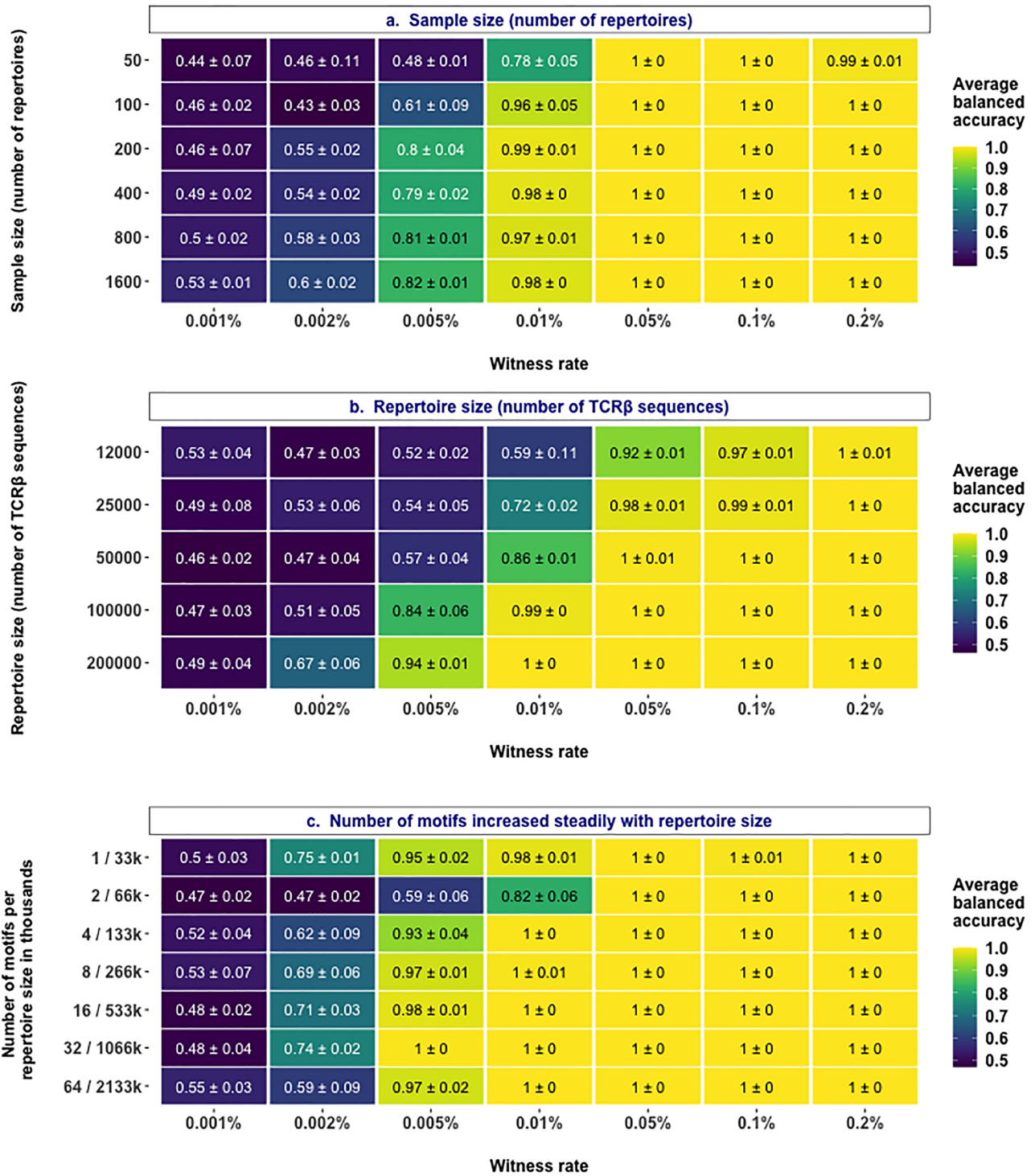
Figure 3: Impact of sample size, repertoire size, and the number of motifs constituting a signal. (a) Performance estimates of a regularized logistic regression model in a binary classification of balanced, labeled adaptive immune receptor repertoire (AIRR) data sets of varying sample sizes (y-axis) where the signal in positive class examples composed of 4-mers is known at the explored witness rates (x-axis). (b) Performance estimates of a regularized logistic regression model in a binary classification of balanced, labeled AIRR data sets with varying repertoire sizes (y-axis) where the signal in positive class examples composed of 4-mers is known at the explored witness rates (x-axis). (c) Performance estimates of a regularized logistic regression model in a binary classification of balanced, labeled AIRR data sets with a combination of varying repertoire sizes and signal definition (y-axis) where the signal in positive class examples composed of 4-mers is known at the explored witness rates (x-axis). The mean balanced accuracy of a fivefold cross-validation was computed in three independent replications. The color coding shows the mean and standard deviation of the performance estimate obtained by three independent replications.

a repertoire size of 200,000, leading to better performance. Notably, the average number of unique sequences generated in one of the recent experimental studies [26] was around 100,000. Overall, the findings show that prediction performance monotonically increased with increasing repertoire size.

The impact of repertoire size on performance gain is dependent on the true signal composition

We observed that when the signal definition was composed of relatively many motifs ($n = 64$), all the tested models failed to learn the implanted signal at moderate to low witness rates (Fig. 2). We also observed that performance improved with an increase in repertoire size when the number of motifs was kept constant ($n = 3$) (Fig. 3b). Combining the knowledge from both these experiments, we investigated whether and how the performance scales with increased repertoire size if the number of motifs also increases proportionally to the repertoire size. To this end, we trained models at different witness rates on data sets with increasing repertoire size, where the proportion of the number of motifs per repertoire size was kept constant. First, we observed that the performance of models gained at moderate to low witness rates (0.002%, 0.005%, 0.01%), even when the number of implanted motifs was 64 (Fig. 3c). This is in contrast to the observations that the models performed poorly at moderate to low witness rates when the number of implanted motifs was 64 in repertoires of a fixed size of 100,000 (Fig. 2). These findings validate the observations that an increase in repertoire size contributes to performance gains (Fig. 3b). Irrespective of the increase in the number of implanted motifs, there was a general trend of decent and comparable prediction accuracy at a witness rate of 0.002%, although not necessarily a linear improvement in performance. Overall, the findings validated the performance gains with increased repertoire size but indicated that the impact was dependent on the signal composition (number of implanted motifs) and the relation was nonlinear.

Classification performance strongly depends on the class balance in data sets

In the benchmark experiments described above, all the data sets were balanced in labels with 50% each of positive and negative class examples. However, for experimental data sets in biomedical research, it remains to sample as many negative examples as positive examples (or vice versa). To understand if and how classification performance scales with increased class imbalance, we evaluated the performance at different witness rates with multiple data sets that have varying degrees of class balance. We observed that the class balance did not have a large impact on the performance at moderate to high witness rates (Fig. 4a). However, at moderate to low witness rates such as 0.01% and 0.005%, having a high degree of class imbalance (e.g., containing only 20% of examples from either positive or negative class) resulted in a substantial decrease in performance. Overall, these findings demonstrated the dependence of classification performance on class balance.

Distributional shift impacts the classification performance

A distributional shift is a common problem encountered in real-world applications of machine learning [51]. It can take many forms but can simplistically be stated as a change in data distributions between the training data set and the examples the model meets in the future. To understand how the AIRR-ML models that we trained adapt to the distributional shift, we trained the models on data sets with varying witness rates and evaluated their prediction performance on test data sets that have different witness rates (Fig. 4b). We observed that the prediction accuracy decreased when the witness rate of either the training or the test data set decreased. We particularly noticed that the effect of decreased test witness rate was more substantial than a decrease in training witness rate (e.g., a model trained on a data set with a 0.005% witness rate performed better on a test data set with a 0.01% witness rate compared to the performance of a model trained on a 0.01% witness rate applied on a test data set with a 0.005% witness rate). Another striking observation was that given a particular training witness rate, a distributional shift that increased the test witness rate improved the performance. Even at very low training witness rates (0.001%), where prediction appears to be essentially random when tested on data from the same distribution, the accuracy increased up to 66% with a distributional shift that increased the test witness rate. This suggests that even though the model trained at 0.001% appears to not have learned any signal, it has indeed captured the signal, but the signal was not strong enough to allow accurate prediction at its native witness rate; rather if the same signal becomes stronger in an application setting, the same model indeed is able to exploit the signals it has learned to predict with better accuracy.

Classification performance depends on the shape of the ground truth signal and matched assumptions of the ML model

Previous studies suggested that short and linear subsequences (k-mers) of amino acid sequences make contact with the antigenic peptide residues [26, 36] and thus many of the previous studies considered the size of ground truth subsequences to be between 2 and 5 amino acid residues that are either continuous or contain gaps [21, 23, 27, 49]. We considered both gapped and ungapped k-mers, which we refer to together as ground truth signal shape. We evaluated if and how the performance of models scales with the shape of the ground truth signal. To this end, we first assessed the performance of models on data sets where the ground truth had implanted motifs of varied sizes (2-mers, 3-mers, 4-mers, and 5-mers) at different witness rates (Fig. 5a). We observed that when the signal was longer, a feature size-matched model was able to attain around 65% accuracy even at a witness rate of 0.001% and 80% accuracy at a 0.002% witness rate. As the size of the signal decreased, moderate to good performances were observed only at markedly higher witness rates (Fig. 5a). Next, we assessed the performance of models on data sets where the ground truth had gapped motifs with varied sizes (3-mers, 4-mers, and 5-mers with one gap in random position) implanted at different witness rates. When the true signal contained a gap (e.g., A∗C where ∗ could be any single amino acid residue), feature size-matched models did not perform as well as they did when the true signal did not contain any gaps. This was particularly evident at witness rates such as 0.005% and 0.01%, where the performance decreased substantially when compared to a scenario where the true signal did not contain any gaps (Fig. 5b). One of the contributory factors for such a decrease in performance might be the fact that the model had to retain multiple coefficients because of the wild-card amino acid residue in the gap position, thus capturing more noise along with the signal.

Furthermore, we set out to understand how the performance of models would vary when the feature size assumed by the model does not match the true signal size. For this, we evaluated the per-
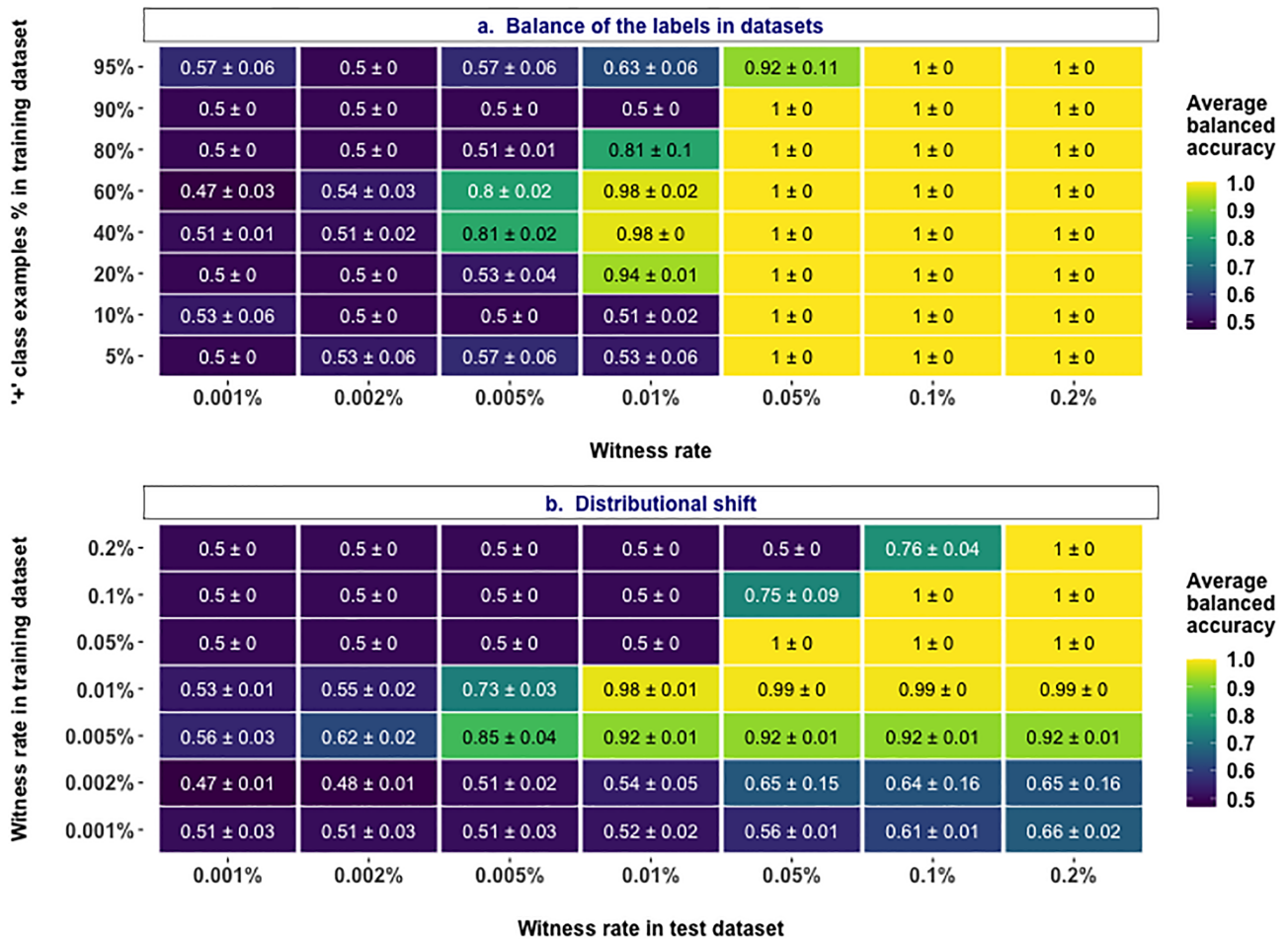
Figure 4: Impact of class balance and distributional shift. (a) Performance estimates of a regularized logistic regression model in a binary classification of unbalanced, labeled adaptive immune receptor repertoire (AIRR) data sets with varying degrees of class balance in the data sets (y-axis) where the signal in positive class examples composed of 4-mers is known at the explored witness rates (x-axis). (b) Performance of a regularized logistic regression model trained on balanced, labeled AIRR data sets with varying witness rates (y-axis) in the classification of a new unseen balanced, labeled AIRR test data sets where the signal in positive class examples composed of 4-mers is known at the explored witness rates (x-axis). The mean balanced accuracy of a fivefold cross-validation was computed in three independent replications. The color coding shows the mean and standard deviation of the performance estimate obtained by three independent replicates.

formance of models that made varied assumptions about the size of true signal (assuming the signal occurs as 2-mers, 3-mers, 4-mers, or 5-mers) on data sets where the true signal implanted at different witness rates was always held constant as 4-mers (Fig. 5c). We observed that when the feature size assumed by the model matched the size of the true signal (implanted motif size), the model performed almost perfectly at a witness rate of 0.01%, whereas the performance decreased when the feature size assumed by the model was larger or smaller than the true signal size. A similar trend was also evident at a witness rate of 0.005%. At higher witness rates (>0.01%), a model that assumed a feature size of 2, performed relatively poorly compared to the other models that achieved almost perfect classification accuracy at higher witness rates.

Overall, the findings indicated that learning smaller patterns at lower witness rates was challenging for the tested models compared to learning longer k-mers. The performance at lower witness rates was exacerbated when the ground truth signal contained gaps within the k-mers. ML models with mismatching assumptions on the ground truth signal fared markedly poorly compared with the models with matching assumptions.

Impact of noise on classification performance

In all experiments above, we evaluated the performance of models on synthetic benchmark data sets where we introduced signals into the positive class examples at defined rates. In real-world experiments, one cannot exclude the possibility of various sources of noise in the negative class examples. In this particular ML application, the signal that differentiates positive and negative class examples could, for instance, occur in both positive and negative class examples but be enriched above a baseline in positive class examples. In such a scenario, the background frequencies of the signal in the negative class represent noise. To understand the relation between classification performance and the noise levels in the negative class, we evaluated the performance of models that are trained on data sets with varying levels of noise (x-axis in Fig. 5d) while holding the witness rate in the positive class constant at 0.01%. We observed that the level of noise in negative class examples clearly had an impact on the performance of the model. The performance decreased as the noise in the negative class increased. However, the model performed well even when the noise in the negative class was half the witness rate in the positive class.
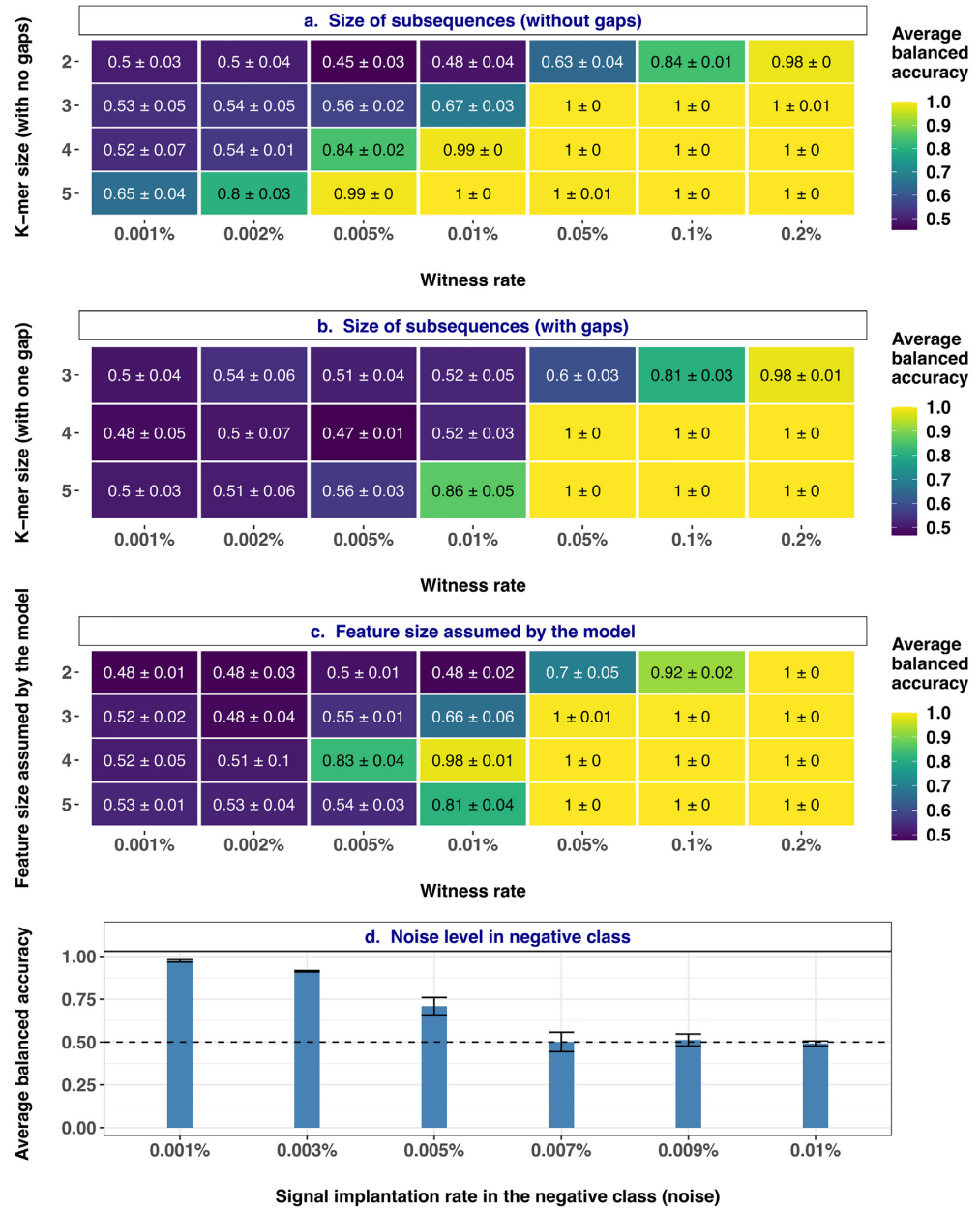
Figure 5: Impact of the size of k-mers (with/without gaps), feature size assumed by the model, and noise level in the negative class on the prediction performance. (a) Performance estimates of a regularized logistic regression model in a binary classification of balanced, labeled adaptive immune receptor repertoire (AIRR) data sets with the ground truth signal composed of varying k-mer sizes (y-axis) where the signal in positive class examples is known at the explored witness rates (x-axis). (b) Performance estimates of a regularized logistic regression model in a binary classification of balanced, labeled AIRR data sets with the ground truth signal composed of varying k-mer sizes that contained a gap in random position (y-axis), where the signal in positive class examples is known at the explored witness rates (x-axis). (c) Performance estimates of a regularized logistic regression model in a binary classification of balanced, labeled AIRR data sets that assumed varying feature sizes of the ground truth signal (y-axis) where the signal in positive class examples composed of 4-mers is known at the explored witness rates (x-axis). (d) Performance estimates of a regularized logistic regression model (y-axis) in a binary classification of balanced, labeled AIRR data sets where the witness rate in positive class examples composed of 4-mers is fixed at 0.01% and the noise in negative class examples is known at the explored rates (x-axis). The mean balanced accuracy of a fivefold cross-validation was computed in three independent replications. The color coding shows the mean and standard deviation of the performance estimate obtained by three independent replications. The illustrations in the left margin show examples of ground truth k-mer sizes and the feature sizes assumed by the model for each row of the charts in a, b, and c. Note that k-mers shown are only for the ease of illustration; in the simulations, a different set of k-mers ($n = 3$) was used.

Immune receptor sequence-specific biases make the classification problem unique

Adaptive immune receptor sequences are known to have positional biases, where specific amino acid residues occur with high probability in particular positions of the amino acid sequences. In addition, some amino acid residues might be overrepresented in the repertoires and there might be interdependencies between the positional frequencies of amino acids. These biases are specific to AIRR data sets. To understand if and how the performance of the tested models differs based on the presence/absence of AIRR data set–specific biases, we trained and evaluated models on data sets originating from three sources at different witness rates. For this, we generated four different types of data sets as described in the Methods section: random sequence repertoires, OLGA-based repertoires, meta repertoires, and experimental repertoires. The random sequence repertoires are free of any immune receptor–specific biases and can be a proxy for a special case of the general sequence pattern classification problem, while the OLGA-based repertoires contain immune receptor sequences according to the VDJ recombination model of OLGA and are thought to serve as a decent proxy for experimental AIRR sequence data sets. The meta repertoire data set contains TCRβ CDR3 sequences drawn randomly from repertoires of an experimental data set irrespective of their immune states or metadata, resulting in repertoires composed of randomly pooled experimentally determined sequences. The experimental repertoires are randomly selected whole experimental repertoires irrespective of their immune states or metadata subsampled to comparable repertoire size. The findings showed that the performance on random repertoires is poorer at witness rates such as 0.005% and 0.01% when compared to the other two data set sources (Supplementary Fig. 9). Although the performance estimates on OLGA-based repertoires looked similar to those on meta repertoires and subsampled experimental repertoires at a 0.005% witness rate, the models on meta repertoires and experimental repertoires exhibited better performance at 0.002%. This points toward the positional biases and dependency structures specific to AIRR data sets.

Comparison of the performance of selected machine learning methods on AIRR classification

Although a comprehensive assessment and benchmarking of ML methods for AIRR classification is not the goal of this study, we briefly explored how straightforward application of other classical ML methods like random forests (RFs) or a linear kernel-based support vector classifier (SVC) compared to the logistic regression model tested throughout this article. To this end, similarly to the exploration of hyperparameter space of logistic regression models, we first explored the hyperparameter spaces of both RF and SVC to narrow down the hyperparameter search space (Supplementary Fig. 10b and Supplementary Fig. 10c). We then assessed the performance estimates of $L1$-penalized logistic regression, RF, and SVC (both $L1$ and the default $L2$ regularizations) in classifying the AIRR data sets with different witness rates. We observed that $L1$-penalized logistic regression and SVC exhibited similar performance on this classification problem, whereas the $L2$-penalized SVC and RF performed poorly (Supplementary Fig. 10a). We believe this is due to the stronger degree of regularization of the $L1$-penalization used by both logistic regression and SVC that suited the problem domain characterized by sparse signals. This highlights that a stronger degree of regularization ($L1$ vs. $L2$) rather than the loss function (logistic regression vs. SVC) is important for

obtaining a better prediction performance for AIRR classification, given the assumption that immune state–associated patterns are rare.

Following the observation that sparse models based on $L1$-penalization suited this particular problem setup, we asked whether a feature selection step combined with a nonpenalized model would exhibit similar performance levels as penalized counterparts. To this end, we built a feature selection–aided classifier and compared its performance to $L1$-penalized logistic regression. Expectedly, as also observed in the analyses of Supplementary Fig. 5 and Supplementary Fig. 6 described above, both models exhibited similar performance across all the scenarios tested (Supplementary Fig. 11).

Discussion

In this study, we observed that a baseline method in the form of a strongly regularized logistic regression model was able to perform well in predicting the immune states of AIRR data sets even at a witness rate as low as 0.002% (2 of 100,000 sequences containing a motif), comparable to a level of difficulty observed in AIR-based disease studies [38]. We also identified several scenarios where the performance of the baseline method decreased with an increase in signal complexities (Fig. 6). These observations spark curiosity in how well the existing AIRR-ML methods would perform on the large and diverse benchmark data set suite generated in this study. The observation that a baseline method like logistic regression performs exceedingly well in a repertoire classification problem (albeit with a simple definition of potential immune signal) points toward the necessity of carrying out a comprehensive benchmarking of the existing AIRR-ML methods for repertoire classification (e.g., see references [23, 26, 28, 29, 32, 40, 41]) to understand if and where tailored ML methods with/without sophisticated architectures are particularly needed to improve the baseline performance. Although benchmarking of existing AIRR-ML methods was out of the scope of this study, an observation that one of the popular off-the-shelf ML methods, RF, performed poorly compared to suited sparse models ($L1$-penalized logistic regression, $L1$-penalized SVC, and a feature selection–aided classifier) (Supplementary Fig. 10, Supplementary Fig. 11) asserts the necessity of benchmarking state-of-the-art AIRR-ML methods. To avoid the biases of self-assessment trap [52] (e.g., less transparency in crucial aspects of ML models such as overfitting and poor generalizability) and ensure rigorous assessment of methods in biomedical research, a community-driven concerted effort known as "crowdsourced challenges" is increasingly becoming a popular way to benchmark methods [53–55]. The large benchmark data set suite generated in this study covering a wide variety of properties related to AIRR data sets and immune signals can already act as a suitable subset of scenarios to be assessed in such community-driven benchmarking efforts together with even more specialized benchmark data sets.

The benchmark data set suite generated in this study assumes that the signal that distinguishes immune states is embedded within the beta chain of TCRs in the form of linear subsequences (k-mers), which aligns with the observations that most of the possible contacts between TCR and peptide antigens were made through only short and typically linear stretches of amino acid residues of CDR3s (IMGT positions 107–116) [26, 36]. Nevertheless, existing AIRR-ML methods made varied assumptions regarding the immune signal [19] and reported decent performance metrics on their respective study data sets. For instance, existing methods assumed that immune signal occurs as (i) short linear sub-
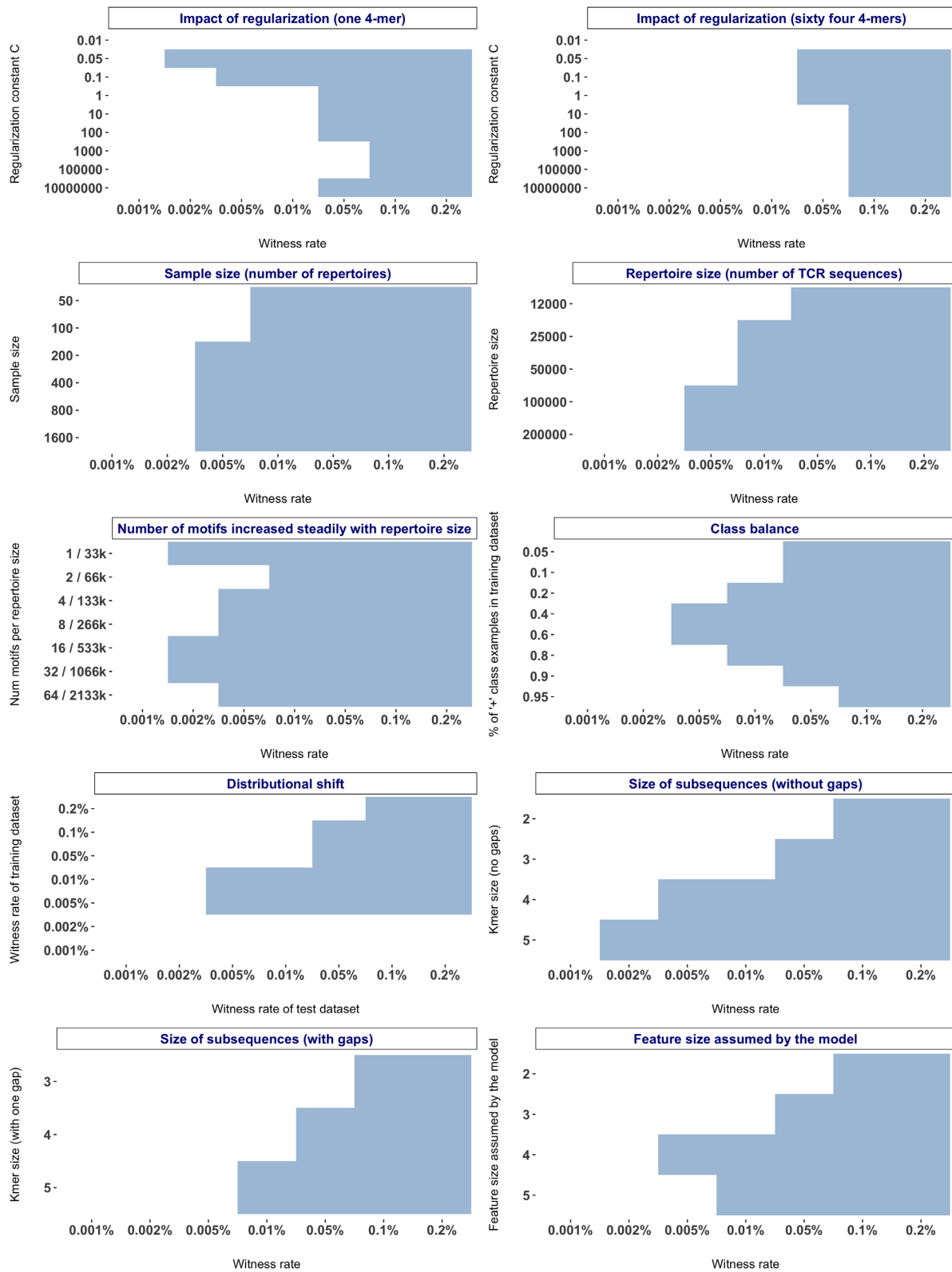
Figure 6: Boundaries and limits of a baseline adaptive immune receptor repertoire (AIRR) machine learning (ML) model across variations of the data set and signal properties explored: to demarcate the boundaries where baseline ML methods performed well for AIRR classification and where the baseline methods failed to learn adequately, we discretized the performance profiles of all the investigations (Figs. 2–5) by a defined threshold of average balanced accuracy ≥0.7 and standard deviation ≤0.1. The regions in blue show where the baseline methods performed adequately well and the white portions represent scenarios that need particular attention in future benchmarking endeavors of existing AIRR-ML methods as well as future methodological developments of sophisticated ML approaches if those scenarios remain intractable for existing AIRR-ML methods.

sequences (k-mers) [23, 42–44], (ii) full sequences with/without a particular combination of V gene usage [26, 40, 41], (iii) alternative representations of the sequences such as physicochemical properties of the amino acids [23], (iv) relevant features of the full sequences that can be captured as latent variables [28, 32], and (v) sequence-independent clonal abundance distributions captured as diversity profiles [56]. A comprehensive benchmarking of the state-of-the-art AIRR-ML methods thus requires an even more specialized benchmark data set suite than those generated in this study to assess the robustness of methods across varying assumptions of the ground truth immune signal.

We obtained several useful insights through profiling the baseline performance of AIRR-ML models in this study that will aid both the users and methodology developers of AIRR-ML models. First, we noticed that a traditional ML method like logistic regression even with stronger regularization failed to distinguish the classes by immune states at moderate to low witness rates (≤0.01%) when the total number of phenotype-associated sequence patterns is large (of which a fraction occurs in positive class examples) (Fig. 2b). Some of the existing AIRR-ML methods [23, 28, 32] aptly treated the repertoire classification problem as a two-staged approach known as multiple instance learning (MIL), where the phenotype-associated patterns are first determined followed by predicting a repertoire-level label based on some form of pooling function of the phenotype-associated patterns. Future studies should investigate if existing MIL-based methods or novel MIL approaches would perform well at moderate to low witness rates (≤0.01%) when the total number of phenotype-associated sequence patterns is large.

Second, we found that in our experimental setup, small sample sizes (≤100) were underpowered to learn rare immune signals in this study, and a marked improvement in performance was observed by increasing the sample size to 200. On the other hand, despite an eightfold increase in sample size from 200 repertoires to 1,600 repertoires, the improvement in the performance of the tested models was not substantial at lower witness rates (≤0.005%) (Fig. 3a). These observations have to be interpreted in the context of the constraints of contemporary AIRR-ML studies in generating data sets with large sample sizes. Several of the AIRR-ML studies [23, 25, 27, 41–44] that particularly generated AIRR data sets on clinical samples are constrained by sample sizes and thus trained ML models on sample sizes ≈50 repertoires, which is an underpowered sample size to learn signals at low witness rates in this study. A very few contemporary studies have generated data sets with as large a sample size as explored in this study. For instance, the recent ImmuneCODE database [39] containing SARS-CoV-2–specific TCR data sets has a curated sample size of ≈1,500 repertoires. Overall, these observations indicate that if the assumptions of the immune signals used in the benchmark data sets of this study match the ground truth, then the sample sizes of AIRR-ML studies may have to be substantially increased (beyond what is explored in this study) to be able to observe a considerable improvement in the performance of ML models at lower witness rates.

Third, we observed that the prediction performance of an *L1*-penalized logistic regression, at a constant witness rate of 0.01%, improved with increasing *n* proportional to $s*\log(p)$, where *n* refers to the number of examples of the minority class (the class that has few examples in an imbalanced binary classification problem), *p* refers to the total number of predictors (4-mers), and *s* refers to the number of nonzero coefficients of the ML model (Supplementary Text, Supplementary Fig. 12, and Supplementary Fig. 13). This observation is in agreement with the theoretical results of sparse

model theory [57], which may serve in the study design of AIRR studies that are characterized by high dimensionality and sparse signals. Notably, however, the effect sizes of coefficients and the sum of nonzero coefficients, which can vary with witness rate, sequencing depth, and between different immune states, may also have an impact on the success of suited sparse models at different sample sizes and warrant future investigations.

Our findings demonstrate that increasing the sequencing depth is more beneficial for training a generalizable model (than increasing sample size) as it increases the likelihood to include and learn more phenotype-associated sequence patterns, especially at moderate to low witness rates (≤0.01%) (Fig. 3). An increase in repertoire size not only improved the performance of the model at moderate to low witness rates (Fig. 3b) but also improved the performance when the total number of phenotype-associated sequence patterns is large (Fig. 3c). However, it might be a challenge for current experimental data sets to achieve as high sequencing depth as 10^6 unique sequences on average for all the repertoires and is currently limited to sizes around 10^5 unique sequences on average with a large dispersion [26, 39]. Future studies should investigate the trade-off between sample size and sequencing depth to understand if an increase in sample size would compensate for the shallow sequencing depths.

Although class imbalance [58] and distributional shifts [51] are known phenomena impacting the generalizability of ML models, we here charted out precisely how they impact the generalizability of AIRR-ML models (Fig. 4). We noticed a decrease in the performance of ML models with increased class imbalance. A straightforward approach used often to mitigate the class imbalance effect is to use some form of subsampling to bring balance in the classes. However, such techniques would further reduce the sample size and require a decently large number of examples of the least prevalent class to mitigate the impact of the sample size described above. Furthermore, we observed that a distributional shift involving a decrease in witness rate of either the training data set or test data set resulted in a decrease of prediction accuracy, although the impact of the latter was substantial. While the users of AIRR-ML methods should be mindful of the class imbalances and potential distributional shifts, the AIRR-ML community should explore, understand, and develop tailored approaches that remain robust to class imbalances and distributional shifts in AIRR data sets.

Previous studies reported motifs of different sizes contributing to the epitope specificity ranging from a single amino acid residue [59] to full sequences [26]. In light of such observations, it is important for the ML models to remain robust irrespective of the size of the ground truth k-mers. However, in this study, we noticed a marked deterioration of prediction performance when the assumed feature size did not match the ground truth, where a much higher witness rate was needed to compensate for a mismatched modeling assumption. The users of AIRR-ML methods should thus notice that assuming the ground truth signal as a single fixed feature size can turn out to be a major limitation of the models. Methods that learn relevant features of the sequences as latent variables [28, 29, 32] may overcome such limitations and thus should be investigated in future benchmarking endeavors. Further, we also observed that shorter k-mers were much harder to learn than longer k-mers and that complex k-mers (i.e., containing a gap) were much harder to learn than contiguous k-mers. Future studies should investigate if the state-of-the-art AIRR-ML methods are able to perform better in learning shorter and complex k-mers and develop methodologies to overcome these challenges faced by traditional encodings and ML methods.

Our exploration of the noise in the negative class can be viewed as a case of class noise [60] that sheds light on the signal frequency distributions, where the classes become inseparable. We observed that a model was able to perform well even when the negative class examples carried noise at a rate equivalent to half the witness rate in the positive class examples. In other words, when positive class repertoires carried a signal on average in 10 of 100,000 sequences, the model was able to separate them from negative class repertoires with decent accuracy as long as the negative class repertoires did not carry a signal in 5 of 100,000 sequences. In this simulation study, we were able to assess at what frequency level the ground truth signals are occurring in the negative class as we know the ground truth signals from the outset. When the ground truth signals are not known in experimental data sets, comparing the feature frequency distributions of positive and negative class examples for top features (selected based on model coefficients or some form of feature importance) can act as model diagnostic.

In this study, we generated the synthetic reference repertoires (prior to simulating immune state–associated signal) according to the VDJ recombination model provided by OLGA [47], which acted as a reasonable proxy for real-world experimental repertoires (Supplementary Fig. 9). Although the synthetic repertoires used in this study retain the nativeness of AIRR data sets in terms of positional biases, amino acid usage, sequence length distributions, and typical repertoire sizes, future studies should also explore and understand how to mimic the other noises and biases that are inherent to experimental data sets (e.g., sequencing artifacts, library preparation issues, batch effects and impact of other covariates, coexistence of other signals that can further increase the complexity). A natural extension would then be to devise analytical strategies and ML methods that not only handle the idiosyncrasies of experimental data sets but also account for the effect of covariates to establish true causal relations between sequence patterns and immune states. Furthermore, the strategy of implanting k-mers as a means to introduce immune state–associated signals into the repertoires was found to be adequate as the implanted k-mers did not induce significant changes in the underlying baseline frequency distributions of other k-mers, thus not disrupting the positional biases of immune receptor sequences substantially (Supplementary Fig. 2). In future work, the impact of k-mer implantation on TCR generation probabilities should be investigated in order to investigate to what extent motif implantation changes the a priori likelihood of a sequence [4, 61]. In this study, we performed all the experiments using the TCRβ CDR3 amino acid sequences. However, the findings are applicable to adaptive immune receptor repertoires in general given that BCR and TCR sequences have very similar immunogenic architecture. That said, it will be of interest to measure the impact of BCR somatic hypermutation in the (BCR) AIRR-ML applications [45, 62].

The importance of the technical implementation setup of a large-scale computational study such as the one carried out here is noteworthy. To carry out a similar study involving AIRR-ML models, one would need streamlined operations to (i) read in AIRR-seq data sets in different file formats (both experimental and synthetic), (ii) represent the information of data sets in multiple ways (encoding schemes), (iii) simulate immune state–associated signals, (iv) have access to a wide variety of ML model implementations, (v) perform model interpretation easily through a range of exploratory diagnostic analyses, and (vi) ensure reproducibility, shareability, and transparency of the analyses. Handling all the aforementioned operations even when building on top of existing ML frameworks such as scikit-learn [50] with *ad*

hoc scripts can be challenging, time-consuming (for all the boilerplate code implementations), and less efficient if the implementations are not well tested or optimized. To overcome all these challenges, we used immuneML [63], which is a domain-adapted software ecosystem for ML analysis of AIRR data sets.

Conclusions: To help the scientific community in reorienting efforts of developing novel ML methodology for those scenarios of AIRR classification where baseline methods fail, we profiled the baseline performance of AIRR-ML models across a diverse set of challenges with increasing levels of complexities for the classification problem. The summarized findings (Fig. 6) showed the boundaries in terms of AIRR data set and immune signal characteristics, where baseline ML methods are able to classify repertoires according to immune state. Notably, the findings demonstrated the suitability of sparse models for AIRR classification, given the assumption that the patterns associated with an immune state occur at low frequency and can be as rare as one antigen-binding AIR per million lymphocytes. Future benchmarking studies should investigate if the state-of-the-art AIRR-ML methods are able to perform well on the intractable scenarios identified in this study. Such an endeavor can help narrow down the scenarios of AIRR classification where novel methodology development is needed. The benchmark data set suite and the knowledge of baseline performance levels of AIRR-ML models generated in this study serve multiple purposes by (i) providing a reference benchmark for novel AIRR classification methodology and enumerating scenarios where novel methodology development is not needed, (ii) providing realistic expectations of the performance of AIRR-ML models given the training data set properties and assumptions, and (iii) serving as a template for defining specialized AIRR benchmark data sets for comprehensive benchmarking of the AIRR-ML methods.

Methods
Generation of synthetic AIRR reference data sets

In this study, our goal was to profile the performance of AIRR-ML models across a wide range of distinct challenges listed in Table 1. Ideally, such an endeavor is best carried out with a combination of experimental and simulated data sets. However, often the ground truth in experimental data sets is not known. The impact of different variations of AIRR data set properties and signal properties on the performance of ML models cannot be sufficiently teased apart if there exists no knowledge on how hard the learning problem is at the outset. Therefore, we chose to generate synthetic AIRR reference data sets that retain the realistic nature of AIRR data sets in terms of positional biases, amino acid usage in the sequences of repertoires, sequence length distributions, and typical repertoire sizes. We generated the desired number of repertoires (default 200) for each benchmark data set by generating the desired number of TCRβ CDR3 amino acid sequences (default 100,000) according to the VDJ recombination model provided by OLGA [47]. In one of the benchmarks, we sought to evaluate whether the performance of ML models would vary if the data sets used for training the models deviate from AIRR-specific characteristics. For that, in addition to the repertoires generated according to the OLGA-provided VDJ recombination model, we generated repertoire data sets in three other ways: (i) the desired number of repertoires composed of the desired number of random amino acid sequences matching the length distribution of typical CDR3 amino acid sequences, (ii) desired number of repertoires composed of the desired number of amino acid sequences

randomly sampled from a large pool of experimental TCRβ CDR3 amino acid sequences from Emerson et al. [26], and (iii) desired number of whole repertoires from Emerson et al. [26] that have comparable repertoire sizes achieved through subsampling. Hereafter, we refer to the four synthetic reference data sets generated here as OLGA-based repertoires, random repertoires, meta repertoires (the one with pooled sequences from the experimental data set), and experimental repertoires (subsampled to have comparable repertoire sizes).

Construction of benchmark data sets through the simulation of immune state–associated signals

In all except one benchmark data set mentioned above (Table 1), we used OLGA-based repertoires to construct the benchmark data sets. Existing knowledge suggests that the majority of the possible contacts between TCR and peptide antigens were made through only short and typically linear stretches of amino acid residues of CDR3s (IMGT positions 107–116) [26, 36]. To account for the often unknown and varying lengths of the subsequences that make contact with peptide antigens, some of the previous studies either used a fixed assumption regarding the subsequence size (e.g., 4-mers [23]) or used varied definitions of the subsequence size (e.g., 2-mers, 3-mers, 4-mers, and 5-mers [21, 27]). In this study, while we use varying definitions of k-mer size to assess the impact of signal definition on classification performance, we use a fixed definition (4-mers) when the aims were to assess the impact of other properties of the AIRR-ML training setup. As a default definition, we used a signal composed of three k-mers of size 4 (4-mers) in the vast majority of the benchmarks. In each of the reference data set, signal is implanted at multiple different witness rates (0.001%, 0.002%, 0.005%, 0.01%, 0.05%, 0.1%, 0.2%) in 50% of the data set. The signal was implanted with equal probability at IMGT sequence positions 108–111 to not disrupt the germline signal in the conserved positions (the start and end portions of the sequences). The examples in which the signal was implanted were labeled positive class (with a hypothetical immune state) and the remaining were labeled negative class. In addition, for each investigation, one of the properties pertaining to the data set, signal, or ML model was varied at a range of values. Table 1 shows a list of the different data set or signal properties that were varied for the construction of benchmark data sets and the particular variations (range of values) of each property that were explored. For each property that was varied, a benchmark data set was constructed at all the witness rates explored. When investigating the noise in the negative class, the signal was implanted in the positive class examples at only one fixed witness rate (0.1%). Figure 1b shows a schematic illustration of implanting an immune state–associated sequence signal into synthetic AIRR reference data sets, and Fig. 1c shows schematic illustrations of the construction of benchmark data sets.

Sequence encoding and preprocessing

A k-mer frequency encoding was used in all the ML model training setups of this study. In other words, all the ML models of this study assume that the immune signal that differentiates positive and negative class occurs in the form of a contiguous subsequence of a defined size k (k-mers). The sequences in each repertoire were split into overlapping k-mers, and their frequencies were computed followed by L2 normalization of k-mer frequencies. Furthermore, each feature vector was standardized by scaling variance to 1 and mean to zero across examples. The default encoding scheme was subsequences of size 4 (4-mers). In a subset of the benchmarking

experiments, the k in k-mer frequency encoding varied from 2 to 5.

Machine learning models

We used scikit-learn's [50] implementation of an L1-regularized logistic regression (Lasso [64]) for most of the benchmarking experiments in order to establish a baseline for the predictive performance of ML models across diverse levels of challenges of AIRR classification. Consider a response variable Y with two classes "+" and "−" and predictor variables X with X_i denoting the ith example and X_{ij} denoting the jth feature of the ith example. In logistic regression, the conditional probability of $P(Y = +|X)$, shortly $p(X)$, is modeled using the logit function $log(\frac{p(X)}{1-p(X)}) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$, and the values of β_0 and β_j are found through minimizing the negative log-likelihood given by the following equation: $L = -log(\prod_{i:Y_i = +} p(X_i) \prod_{j:Y_j = -} (1 - p(X_j)))$

where $\prod$ represents product over i and j that run over positive (+) and negative (−) classes, respectively. In L1-penalized logistic regression, the cost function that is minimized is $L + \lambda \sum_{j=1}^{p} |\beta_j|$, where p is the number of predictors. Note that the hyperparameter controlling regularization strength is indicated by a variable C in scikit-learn's [50] implementations and is the inverse of regularization strength ($C = \lambda^{-1}$). Smaller values of C represent stronger regularization. The hyperparameter space used for controlling the regularization strength constant C was by default 0.05, 0.1, 1, or 5 (except for the benchmarking experiment that specifically explored the effect of regularization, where C varied across a wider range of values). A maximum of 500 iterations were allowed for the model's coefficients to converge. Although a comprehensive assessment and benchmarking of ML methods for AIRR classification falls outside the scope of this study, we briefly explored how straightforward application of other classical ML methods like RFs or a linear kernel-based SVC (using both L1 and L2 penalizations) compared to the L1-penalized logistic regression model tested throughout this article. The hyperparameter space for RF and SVC was narrowed down in a similar fashion as logistic regression. For L1 regularization, the regularization constant space and the maximum number of iterations used for SVC were the same as logistic regression. For RF, the number of trees explored in hyperparameter optimization was 5, 10, 50, and 100.

In addition to the off-the-shelf ML methods, we built a feature selection–aided classifier as another baseline to compare with the traditional sparse models during the revision process. Briefly, we used a Student's t -test to compare the two-class frequency distributions for each k-mer (as many tests as k-mers) and used the P value of the resulting tests as a hyperparameter that was optimized for in the inner loop of nested cross-validation. Only the features selected based on the optimal hyperparameter were included in the design matrix on which a traditional logistic regression with no penalization was fit. The method was implemented inside immuneML [63], and a separate docker image running the particular feature branch is provided under the code availability section.

Model training, selection, and evaluation

To estimate the prediction performance of the trained ML models on unseen AIRR data, we used a nested cross-validation (CV) setup. Briefly, we used a fivefold nested CV, where 80% of the data was used for training the model in the outer loop and the remaining 20% as a test set for estimating the model performance. In the

inner loop of the nested CV, we used a fivefold CV where the training data were again further split into 80% training and 20% validation set for tuning the hyperparameters and aiding in model selection. An exhaustive grid search was used for tuning the hyperparameters. The performance metric optimized in model training was accuracy. Note that we used balanced data sets for all the benchmarking experiments, except one scenario where the impact of class balance was explored. We replicated each investigation three times on separate data sets, and the average performance metric of the three replications together with variation is reported. The reported performance metric of each replication is the balanced accuracy [65] obtained through the nested CV.

Assessment of the impact of k-mer implantation on background k-mer frequency distribution

We assessed the impact of k-mer implantation on background k-mer frequency distributions in two different scenarios: (i) one where three 4-mers are implanted and (ii) another where sixty-four 4-mers are implanted. We used 200 repertoires generated by OLGA with a repertoire size of 100,000 sequences. We made an assumption that the 4-mer frequencies of the 200 repertoires will not differ in the absence of any immune state–associated sequence pattern (before 4-mer implantation). Thus, if we would compare the k-mer frequency distributions of two groups (the repertoires that would later become positive and negative class labels after implantation), there should be no significant differences between the groups. For this, we used a Student's t test to compare the two group frequency distributions for each k-mer (as many tests as the number of k-mers) followed by a multiple testing correction using the Benjamini–Hochberg method. We repeated the same process of statistical testing on data sets that received implantation of 4-mers at different witness rates in each scenario (three 4-mers implanted or sixty-four 4-mers implanted). We thresholded on the multiple testing corrected P values ($q < 0.05$) to identify 4-mers that exhibited differential incidence. The experiments were replicated on three separate data sets, and the average and standard deviation of the number of differentially incident motifs at different witness rates were reported.

Implementation details

Most steps of this benchmarking study were carried out using immuneML [63] (version 1.0.2), an integrated ecosystem for ML analysis of adaptive immune receptor repertoires. The ML methods used in this study were based on scikit-learn's [50] implementations. Specifically, the simulation of disease signals into the AIRR data sets for the construction of benchmark data sets, the sequence encoding and preprocessing steps, and the model training, selection, and evaluation were performed through immuneML [63] (version 1.0.2).

Docker container to improve reproducibility

To enable other researchers to reuse and reproduce the findings of this article, we created a Docker [66] image with a predefined computing environment maintaining all the dependencies required for the execution of the code with minimal overhead. Although newer versions of immuneML [63] and its dependencies have emerged during the preparation of this article, the Docker image froze the exact versions of all the software and thereby eliminated any barriers of reproducing the findings of this study. The Docker image can be accessed from the publicly hosted central repository of Docker (dockerhub): *kanduric/immuneml-v1:latest*.

Additionally, we demonstrated the ease of reusability and deployability of the containerized computational workflow by rerunning a subset of analyses from each experiment of this study on a toy data set (100 sequences in each repertoire compared to 100,000 in the original analyses) with fewer iterations in ML training [67]. Although the findings of such analyses are expected to be illogical, the main idea behind choosing a toy data set and fewer iterations in ML training for this demonstration purpose is to ensure rapid testing by other researchers on different computational environments. Notably, the original computational analysis of this article involves reading in ∼2 TB of data from the disk and writing ∼10 TB of data to the disk. To replicate the findings of this article, the original input data [68], analysis scripts [69], and the docker image that are made publicly available have to be used in a similar fashion as shown in the demo analyses [67].

Graphics

ggplot2 version 3.3.5 [70] was used for graphs and Inkscape version 1.0.1 [71] was used for illustrations.

Availability of Source Code

The bash scripts used to generate all the synthetic repertoires are made publicly available [69]. All the synthetic repertoires (∼2.1 TB) used as input for the simulation of disease signals and ML model training are made publicly available at [68]. The analytical details for all investigations including simulation of disease signals and ML model training are made available in the form of immuneML YAML specification files that describe each step and choice of the analyses at [69]. These specifications are both human readable (analysis transparency) and executable using the publicly available immuneML platform [63] version 1.0.2. To ensure the ease of reusability and reproducibility, we containerized the whole computational environment that is necessary to remove any barriers of replicating the findings of this study. The docker image of the computational workflow of this article is available from a publicly hosted docker repository (dockerhub) at *kanduric/immuneml-v1:latest*. A demo of using the provided docker image to rerun each category of experiment in this article has been shown on toy data sets at [67]. Additional analyses were performed during the revision using an updated version of the immuneML platform [63], and the docker image of the updated version is available from the publicly hosted docker repository (dockerhub) at *kanduric/immuneml-v2:latest*. All the analytical details of the analyses performed during the revision process were made available in the form of immuneML YAML specification files at [72].

Data Availability

An archival copy of the github repository [67] is available via the *GigaScience* database GigaDB [73]. Synthetic repertoires used as input for the simulation of disease signals and ML model training are made publicly available at [68].

Additional files

Supplementary Fig. 1. An illustration of formulating classification of immune states based on AIRR data as a multiple instance learning (MIL) problem. Some of the published methods [23, 28, 32] aptly considered AIRR classification as an MIL [37] consisting of repertoires as bags, receptors as instances, and immune state–associated receptors as witnesses. Here, the witnesses (positive in-

stances) are shown in red text, whereas the blue text represents negative instances. Notably, the positive instances can be as rare as one in a million instances. The MIL methods use different pooling functions to learn the relationship between the labels of instances of a bag and the label of a bag (color of bags here).

Supplementary Fig. 2. (Relates to Fig. 2b and Fig. 3a). Impact of k-mer implantation on the background k-mer frequency distribution: the y-axis shows the average number of motifs with induced significant differential incidence scaled by the total number of implanted motifs in $\log_{10}$ scale when the signal (number of implanted 4-mers in upper and lower panels) was implanted at different witness rates (on the x-axis) in 200 repertoires of size 100,000 sequences. The color coding separates the motifs with induced differential incidence by how much they overlap with any implanted motif. For instance, yellow indicates that the differentially incident motif overlaps all its four amino acid residues with one of the implanted 4-mers. In other words, those are the true implanted motifs. The error bars are based on the standard deviation of values across three independent replications. The chart shows that at lower witness rates (up to 0.01%), the implantation of 4-mers predominantly disturbs the frequency distributions of only the implanted 4-mers between positive and negative class repertoires. As the witness rate increases, the signal gets implanted in more sequences, thus overlapping with many different 4-mers by one, two, or three amino acid residues. The total number of disturbed 4-mers increased proportionally to the number of implanted 4-mers (as evident through the scaling of the y-axis per the number of implanted 4-mers).

Supplementary Fig. 3. (Relates to Fig. 2a). Distribution of the coefficient values of the trained model categorized by the level of overlap with the implanted single 4-mer (as in Fig. 2a). The y-axis shows the coefficient values of different 4-mers and the x-axis shows the number of amino acid residues of the 4-mers that overlap the true implanted motifs. In this case, an overlap value of 4 indicates that the 4-mer is the true implanted motif. Different panels in columns show different witness rates, and the panels in different rows correspond to the different splits of the cross-validation. The charts confirm that at lower witness rates, it was only the truly implanted motif that was ascribed the maximum weight (coefficient) by the model. As the witness rate increases, the motif gets implanted in more sequences and thereby overlaps with many other motifs partially by one, two, or three amino acid residues. Thus, such partially overlapping motifs would also be ascribed higher weights by the model as the witness rate increases.

Supplementary Fig. 4. (Relates to Fig. 2b). Distribution of the coefficient values of the trained model categorized by the level of overlap with the implanted sixty-four 4-mers (as in Fig. 2b). The y-axis shows the coefficient values of different 4-mers and the x-axis shows the number of amino acid residues of the 4-mers that overlap the true implanted motifs. In this case, an overlap value of 4 indicates that the 4-mers are the true implanted motifs. Different panels in columns show different witness rates and the panels in different rows correspond to the different splits of the cross-validation. The charts show that at lower witness rates, none of the 4-mers was ascribed extreme weights by the model. As the witness rate increased, the true motifs were ascribed higher weights.

Supplementary Fig. 5. (Relates to Fig. 2a and Supplementary Fig. 3). Distribution of the coefficient values of the trained model categorized by the level of overlap with the implanted single 4-mer (as in Fig. 2a)—this chart, unlike Supplementary Fig. 3, shows the distribution only for the statistically significant 4-mers that were identified through univariate statistical testing. From all the 4-mers that were plotted in Supplementary Fig. 3, we filtered out all those 4-mers that were found with no significant differences in frequency distributions in a univariate statistical test (see Methods) and retained only the significant motifs. This chart confirms that statistically significant motifs that can be identified through a univariate statistical test comparing the frequency distributions are indeed the same motifs that were ascribed higher weights by the penalized logistic regression model. The y-axis shows the coefficient values of different 4-mers, and the x-axis shows the number of amino acid residues of the 4-mers that overlap the true implanted motifs. In this case, an overlap value of 4 indicates that the 4-mer is the true implanted motif. Different panels in columns show different witness rates, and the panels in different rows correspond to the different splits of the cross-validation. Our analysis confirms that at lower witness rates, it was only the truly implanted motif that was ascribed the maximum weight (coefficient) by the model. As the witness rate increases, the motif gets implanted in more sequences and thereby overlaps with many other motifs partially by one, two, or three amino acid residues. Thus, such partially overlapping motifs would also be ascribed higher weights by the model as the witness rate increases.

Supplementary Fig. 6. (Relates to Fig. 2b and Supplementary Fig. 4). Distribution of the coefficient values of the trained model categorized by the level of overlap with the implanted sixty-four 4-mers (as in Fig. 2b)—this chart, unlike Supplementary Fig. 4, shows the distribution only for the statistically significant 4-mers that were identified through univariate statistical testing. From all the 4-mers that were plotted in Supplementary Fig. 4, we filtered out all those 4-mers that were found with no significant differences in frequency distributions in a univariate statistical test (see Methods) and retained only the significant motifs. This chart confirms that statistically significant motifs that can be identified through a univariate statistical test comparing the frequency distributions are indeed the very same motifs that were ascribed higher weights by the penalized logistic regression model. *Note that, unlike* Supplementary Fig. 4*, this chart shows fewer column panels corresponding to witness rates. This was because there were no significant motifs that were identified at lower witness rates.* The y-axis shows the coefficient values of different 4-mers and the x-axis shows the number of amino acid residues of the 4-mers that overlap the true implanted motifs. In this case, an overlap value of 4 indicates that the 4-mers are the true implanted motifs. Different panels in columns show different witness rates and the panels in different rows correspond to the different splits of the cross-validation. The charts show that at lower witness rates, none of the 4-mers was ascribed extreme weights by the model. As the witness rate increased, the true motifs were ascribed higher weights.

Supplementary Fig. 7. (Relates to Fig. 2a). Frequency distribution of implanted 4-mer (as in Fig. 2a) in both positive and negative class examples. On the y-axis, the frequency of the implanted 4-mer is shown in the $\log_{10}$ scale at different witness rates on the x-axis. The color indicates positive and negative class examples. The chart shows that already at a witness rate of 0.002%, the frequency distributions of both positive and negative classes started to become distinguishable.

Supplementary Fig. 8. (Relates to Fig. 2b). Frequency distribution of implanted sixty-four 4-mers (as in Fig. 2b) in both positive and negative class examples. On the y-axis, the frequency of the implanted 4-mers is shown in the $\log_{10}$ scale at different witness rates (in different panels). The x-axis shows the index of each implanted 4-mer. The color indicates positive and negative class examples. The chart shows that at lower witness rates, the fre-

quency distributions of both positive and negative classes are indistinguishable.

Supplementary Fig. 9. (Relates to Figs. 2–5). Impact of the source of data set construction. Performance estimates of a regularized logistic regression model in a binary classification of balanced, labeled data sets of varying sources (on the y-axis), where the signal in positive class examples composed of 4-mers is known at the explored witness rates (on the x-axis).

Supplementary Fig. 10. (Relates to Figs. 2–5). Impact of the choice of ML method and the hyperparameter spaces of explored ML methods. (a) Performance estimates of different ML models (on the y-axis) in a binary classification of balanced, labeled AIRR data sets, where the signal in positive class examples composed of 4-mers is known at the explored witness rates (on the x-axis). The mean balanced accuracy of a fivefold cross-validation was computed in three independent replications. The color coding shows the mean and standard deviation of the performance estimate obtained by three independent replications. (b) Impact of $L1$ regularization on the performance estimates of SVC. Performance estimates of a SVC $L1$-regularized with a fixed regularization constant C (explored on the y-axis) in a binary classification of a balanced, labeled AIRR data set where the signal in positive class examples composed of 4-mers is known at the explored witness rates (on the x-axis). The smaller the value of regularization constant C, the stronger the regularization. The signal definition is composed of three motifs. The color coding shows the mean and standard deviation of the balanced accuracy estimated by a fivefold cross-validation. (c) Impact of the number of estimators on the performance estimates of the RF classifier. Performance estimates of a RF classifier parametrized with a fixed number of estimators (explored on the y-axis) in a binary classification of a balanced, labeled AIRR data set where the signal in positive class examples composed of 4-mers is known at the explored witness rates (on the x-axis). The signal definition is composed of three motifs. The color coding shows the mean and standard deviation of the balanced accuracy estimated through a fivefold cross-validation.

Supplementary Fig. 11. Comparison of the performance of a feature selection–aided classifier with that of an $L1$-penalized logistic regression. Performance estimates of two different models in a binary classification of a balanced, labeled AIRR data set where the signal in positive class examples was composed of varying numbers of 4-mers (on the y-axis) and is known at the explored witness rates (x-axis). (a) Performance estimates of a classifier that was aided by a feature selection step based on a univariate statistical test. (b) Performance estimates of an $L1$-penalized logistic regression model. The mean balanced accuracy of fivefold nested cross-validation was computed in three independent replications. The color coding shows the mean and standard deviation of the performance estimate obtained by three independent replications.

Supplementary Fig. 12. Performance as a function of the number of examples in the minority class and the expected number of nonzero coefficients. Performance estimates of an $L1$-penalized logistic regression in a binary classification of a balanced, labeled AIRR data set where the signal in positive class examples was composed of varying numbers of 4-mers (on the y-axis) and is known at a witness rate of 0.01%, and the number of examples of minority class available for training within each cross-validation loop is varied (on the x-axis). (a) Performance estimates of an $L1$-penalized logistic regression model. The mean balanced accuracy of fivefold nested cross-validation was computed in three independent replications. The color coding shows the mean and standard deviation of the performance estimate obtained by three independent replications. (b) Relation between the expected num-

ber of nonzero coefficients, s (y-axis), and the number of examples in minority class, n (x-axis), and number of predictors in log scale, $\log(p)$, which is held constant as 12. For each cell of the heatmap in panel (a), the corresponding cell in panel (b) shows the ratio $s*\log(p)/n$, and the color coding represents the range of this ratio. The figure shows that the low prediction performance on panel (a) correlates negatively with a high ratio of $s*\log(p)/n$.

Supplementary Fig. 13. Performance as a function of the number of examples in the minority class and the total number of predictors. Performance estimates of an $L1$-penalized logistic regression in a binary classification of a balanced, labeled AIRR data set where the signal in positive class examples was composed of sixteen 4-mers at a witness rate of 0.01% that was held constant, while the number of examples of the minority class available for training within each cross-validation loop is varied (on the x-axis) and the total number of predictors was varied on the y-axis. Notably, the varying total number of predictors on the y-axis always included the predictors that constitute the signal. (a) Performance estimates of an $L1$-penalized logistic regression model. The mean balanced accuracy of fivefold nested cross-validation was computed in three independent replications. The color coding shows the mean and standard deviation of the performance estimate obtained by three independent replications. (b) Relation between the expected number of nonzero coefficients, s, and the number of examples in minority class, n, and number of predictors in log scale, $\log(p)$. For each cell of the heatmap in panel (a), the corresponding cell in panel (b) shows the ratio $s*\log(p)/n$, and the color coding represents the range of this ratio.

Supplementary Table 1. Descriptive statistics of the benchmarking setup

Relating the empirical sample complexity with sparse model theory

The patterns associated with an immune state can be as rare as one antigen-binding AIR per million lymphocytes and were found to occur at a low incidence in previous studies [19, 23, 26, 38]. Sparse models tested in this study are thus particularly suited to the problem setup. In Fig. 3a, by considering the prediction performance as a function of sample size and witness rate in balanced data sets, we empirically assessed the sample complexity while holding the total number of predictors (4-mers) and the expected number of nonzero coefficients of the ML model (number of implanted 4-mers) constant. The theoretical results of sparse model theory [57] state that a sparse model will be successful for n proportional to $s*\log(p)$, where n refers to the number of examples of the minority class (here minority class refer to the class that has few examples in an imbalanced binary classification problem), p refers to the total number of predictors, and s refers to the number of nonzero coefficients. We set out to understand if the sample complexity of sparse models used in this study match with the theoretical results. Specifically, we assessed the prediction performance of the $L1$-penalized logistic regression as a function of the number of examples in minority class with (i) a varying number of expected nonzero coefficients while holding the number of predictors constant (Supplementary Fig. 12) and (ii) a varying number of predictors while holding the expected number of nonzero coefficients constant (Supplementary Fig. 13). In agreement with the theoretical results, we observed that the higher the expected number of nonzero coefficients relative to the number of examples in the minority class, the lower the prediction performance (Supplementary Fig. 12). Similarly, we observed a qualitative distinction in performance levels between high and low

$s*log(p)$ when holding s constant (Supplementary Fig. 13). However, we observed no differences in performance with a decreasing number of total predictors when the number of examples in the minority class was constant.

List of abbreviations

AIR: adaptive immune receptor; AIRR: adaptive immune receptor repertoire; CDR3: complementarity determining region 3; CV: cross-validation; IMGT: ImMunoGeneTics; MIL: multiple instance learning problem; ML: machine learning; RF: random forest; SVC: support vector classifier; TCR: T-cell receptor; TCRβ: T-cell receptor beta chain.

Ethics approval and consent

Not applicable

Consent for publication

Not applicable

Competing interests

VG declares advisory board positions in aiNET GmbH and Enpicom B.V. VG is a consultant for Adaptyv Biosystems, Specifica Inc, and Roche/Genentech.

Funding

Authors' contributions

CK and GKS conceived the overall study. MP, LS, KM, MC, and VG participated in brainstorming and provided critical conceptual feedback. CK performed all analyses and drafted the manuscript. VG and GKS provided critical edits to the manuscript. All authors read and approved the final manuscript.

This work was performed using the Immunohub eInfrastructure funded by the University of Oslo and operated by the authors in close collaboration with the University Center for Information Technology, University of Oslo.

References

1. Glanville, J, Zhai, W, Berka, J *et al.* Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc Natl Acad Sci* 2009;**106**(48):20216–21.
2. Davis, MM, Bjorkman, PJ. T-cell antigen receptor genes and T-cell recognition. *Nature* 1988;**334**(6181):395–402.
3. Tonegawa, S. Somatic generation of antibody diversity. *Nature* 1983;**302**(5909):575–81.
4. Slabodkin, A, Chernigovskaya, M, Mikocziova, I *et al.* Individualized VDJ recombination predisposes the available Ig sequence space. *Genome Res* 2021;**31**(12):2209–24.
5. Arstila, TP, Casrouge, A, Baron, V *et al.* A direct estimate of the human alphabeta T cell receptor diversity. *Science* 1999;**286**(5441):958–61.
6. Farber, DL, Yudanin, NA, Restifo, NP. Human memory T cells: generation, compartmentalization and homeostasis. *Nat Rev Immunol* 2014;**14**(1):24–35.
7. Calis, JJA, Rosenberg, BR. Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends Immunol* 2014;**35**(12):581–90.
8. Georgiou, G, Ippolito, GC, Beausang, J *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol* 2014;**32**(2):158–68.
9. Arnaout, RA, Prak, ETL, Schwab, N *et al.* Adaptive immune receptor repertoire community: the future of blood resting is the immunome. *Front Immunol* 2021;**12**. doi: 10.3389/fimmu.2021.626793.
10. Miho, E, Yermanos, A, Weber, CR *et al.* Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. *Front Immunol* 2018;**9**. doi: 10.3389/fimmu.2018.00224.
11. Akbar, R, Robert, PA, Pavlović, M *et al.* A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Rep* 2021;**34**(11):108856.
12. Venturi, V, DA, Price, Douek, DC *et al.* The molecular basis for public T-cell responses? *Nat Rev Immunol* 2008;**8**(3):231–8.
13. Li, H, Ye, C, Ji, G *et al.* Determinants of public T cell responses. *Cell Res* 2012;**22**(1)33–42.
14. Madi, A, Shifrut, E, Reich-Zeliger, S *et al.* T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res* 2014;**24**(10):1603–12.
15. Madi, A, Poran, A, Shifrut, E *et al.* T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. *eLife* 2017;**6**:e22057. doi: 10.7554/eLife.22057.
16. Pogorelyy, MV, Elhanati, Y, Marcou, Q *et al.* Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. *PLoS Comput Biol* 2017;**13**(7):e1005572.
17. Greiff, V, Menzel, U, Miho, E *et al.* Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B cell development. *Cell Rep* 2017;**19**(7):1467–78.
18. Pogorelyy, MV, Minervina, AA, Shugay, M *et al.* Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLoS Biol Public Library of Science*; 2019;**17**(6):e3000314.
19. Greiff, V, Yaari, G, Cowell, LG. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Curr Opin Syst Biol* 2020;**24**:109–19. https://doi.org/10.1016/j.coisb.2020.10.010.
20. Pertseva, M, Gao, B, Neumeier, D *et al.* Applications of machine and deep learning in adaptive immunity. *Annu Rev Chem Biomol Eng* 2021;**12**(1):39–62.
21. Glanville, J, Huang, H, Nau, A *et al.* Identifying specificity groups in the T cell receptor repertoire. *Nature* 2017;**547**(7661):94–98.
22. Dash, P, Fiore-Gartland, AJ, Hertz, T *et al.* Quantifiable predictive features define epitope specific T cell receptor repertoires. *Nature* 2017;**547**(7661):89–93.
23. Ostmeyer, J, Christley, S, Toby, IT *et al.* Biophysicochemical motifs in T-cell receptor sequences distinguish repertoires from tumor-

infiltrating lymphocyte and adjacent healthy tissue. *Cancer Res* 2019;**79**(7):1671–80.

24. Pogorelyy, MV, Minervina, AA, Chudakov, DM *et al*. Method for identification of condition-associated public antigen receptor sequences. *eLife* 2018;**7**:e33050. doi: 10.7554/eLife.33050.

25. Zhang, H, Liu, L, Zhang, J *et al*. Investigation of antigen-specific T-cell receptor clusters in human cancers. *Clin Cancer Res* American Association for Cancer Research; 2020;**26**(6):1359–71.

26. Emerson, RO, DeWitt, WS, Vignali, M *et al*. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet* 2017;**49**(5):659–65.

27. Huang, H, Wang, C, Rubelt, F *et al*. Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nat Biotechnol* 2020;**38**(10):1194–202.

28. Widrich, M, Schäfl, B, Pavlović, M *et al*. Modern Hopfield networks and attention for immune repertoire classification. *bioRxiv* 2020. doi: 10.1101/2020.04.12.038158.

29. Ostmeyer, J, Christley, S, Cowell, L. Dynamic kernel matching for non-conforming data: a case study of T-cell receptor Datasets. *arXiv* 2103. 10472v12021. https://doi.org/10.48550/arXiv.2103.10 472.

30. Beshnova, D, Ye, J, Onabolu, O *et al*. De novo prediction of cancer-associated T cell receptors for noninvasive cancer detection. *Sci Transl Med* 2020;**12**(557):eaaz3738.

31. Shemesh, O, Polak, P, Lundin, KEA *et al*. Machine learning analysis of naïve B-cell receptor repertoires stratifies celiac disease patients and controls. *Front Immunol* 2021;**12**:627813. doi: 10.3389/fimmu.2021.627813.

32. Sidhom, J-W, Larman, HB, Pardoll, DM *et al*. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nat Commun* 2021;**12**(1):1605.

33. Rognes, T, Scheffer, L, Greiff, V *et al*. CompAIRR: ultra-fast comparison of adaptive immune receptor repertoires by exact and approximate sequence matching. *bioRxiv* 2022: https://doi.org/ 10.1101/2021.10.30.466600.

34. Weber, CR, Rubio, T, Wang, L *et al*. Reference-based comparison of adaptive immune receptor repertoires. *bioRxiv* 2022: https:// doi.org/10.1101/2022.01.23.476436.

35. Xu, JL, Davis, MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* 2000;**13**(1):37–45.

36. Rudolph, MG, Stanfield, RL, Wilson, IA. How TCRs bind MHCs, peptides, and coreceptors. *Annu Rev Immunol* 2006;**24**(1):419–66.

37. Dietterich, TG, Lathrop, RH, Lozano-Pérez, T. Solving the multiple instance problem with axis-parallel rectangles. *Artif Intell* 1997;**89**(1–2):31–71.

38. Christophersen, A, Ráki, M, Bergseng, E *et al*. Tetramer-visualized gluten-specific CD4+ T cells in blood as a potential diagnostic marker for coeliac disease without oral gluten challenge. *United Eur Gastroenterol J* 2014;**2**(4):268–78.

39. Nolan, S, Vignali, M, Klinger, M *et al*. A large-scale database of T-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. *Res Sq* 2020. doi: 10.21203/rs.3.rs-51964/v1.

40. Liu, X, Zhang, W, Zhao, M *et al*. T cell receptor β repertoires as novel diagnostic markers for systemic lupus erythematosus and rheumatoid arthritis. *Ann Rheum Dis* 2019;**78**(8):1070–8.

41. De Neuter, N, Bartholomeus, E, Elias, G *et al*. Memory CD4+ T cell receptor repertoire data mining as a tool for identifying cytomegalovirus serostatus. *Genes Immun* 2019;**20**(3):255–60.

42. Cinelli, M, Sun, Y, Best, K *et al*. Feature selection using a one dimensional naïve Bayes' classifier increases the accuracy of support vector machine classification of CDR3 repertoires. *Bioinformatics* 2017;**33**(7):951–5.

43. Sun, Y, Best, K, Cinelli, M *et al*. Specificity, privacy, and degeneracy in the CD4 T cell receptor repertoire following immunization. *Front Immunol* 2017;**8**:430. doi: 10.3389/fimmu.2017.00430.

44. Thomas, N, Best, K, Cinelli, M *et al*. Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics* 2014;**30**(22):3181–8.

45. Weber, CR, Akbar, R, Yermanos, A *et al*. immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking. *Bioinformatics* 2020;**36**(11):3594–6.

46. Robert, PA, Akbar, R, Frank, R *et al*. One billion synthetic 3D-antibody-antigen complexes enable unconstrained machine-learning formalized investigation of antibody specificity prediction. *bioRxiv* 2022: https://doi.org/10.1101/2021.07.06.451258.

47. Sethna, Z, Elhanati, Y, Callan, CG *et al*. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics* 2019;**35**(17):2974–81.

48. Giudicelli, V, Lefranc, MP. Ontology for immunogenetics: the IMGT-ONTOLOGY. *Bioinformatics* 1999;**15**(12):1047–54.

49. Greiff, V, Weber, CR, Palme, J *et al*. Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. *J Immunol* 2017;**199**(8):2985–97.

50. Pedregosa, F, Varoquaux, G, Gramfort, A *et al*. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011 **12**: 2825–302011.

51. Quionero-Candela, J, Sugiyama, M, Schwaighofer, A *et al*. *Dataset Shift in Machine Learning*. Cambridge, MA: The MIT Press.

52. Norel, R, Rice, JJ, Stolovitzky, G. The self-assessment trap: can we all be better than average? *Mol Syst Biol* 2011;**7**(1):537.

53. Costello, JC, Stolovitzky, G. Seeking the wisdom of crowds through challenge-based competitions in biomedical research. *Clin Pharmacol Ther* 2013;**93**(5):396–8.

54. Saez-Rodriguez, J, Costello, JC, Friend, SH *et al*. Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat Rev Genet* 2016;**17**(8):470–86.

55. Ellrott, K, Buchanan, A, Creason, A *et al*. Reproducible biomedical benchmarking in the cloud: lessons from crowd-sourced data challenges. *Genome Biol* 2019;**20**(1):195. doi: 10.1186/s13059-019-1794-0.

56. Greiff, V, Bhat, P, Cook, SC *et al*. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med* 2015;**7**(1):49.

57. Hastie, T, Tibshirani, R, Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC. ISBN-13: 978-1498712163.

58. Japkowicz, N, Stephen, S. The class imbalance problem: a systematic study. *Intelligent Data Anal* 2002;**6**(5):429–49.

59. Dahal-Koirala, S, Risnes, LF, Neumann, RS *et al*. Comprehensive analysis of CDR3 sequences in gluten-specific T-cell receptors reveals a dominant R-motif and several new minor motifs. *Front Immunol* 2021;**12**:639672. doi: 10.3389/fimmu.2021.639672.

60. Wang, RY, Storey, VC, Firth, CP. A framework for analysis of data quality research. *IEEE Trans Knowl Data Eng* 1995;**7**(4): 623–40.

61. Marcou, Q, Mora, T, Walczak, AM. High-throughput immune repertoire analysis with IGoR. *Nat Commun* 2018;**9**(1):571.

62. Yermanos, A, Greiff, V, Krautler, NJ *et al*. Comparison of methods for phylogenetic B-cell lineage inference using time-resolved antibody repertoire simulations (AbSim). *Bioinformatics* 2017;**33**(24):3938–46.

63. Pavlović, M, Scheffer, L, Motwani, K *et al.* The immuneML ecosystem for machine learning analysis of adaptive immune receptor repertoires. *Nat Machine Intelligence* 2021;**3**(11):936–44.

64. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B* 1996;**58**(1):267–88.

65. Brodersen, KH, Ong, CS, Stephan, KE *et al.* The balanced accuracy and its posterior distribution. *2010 20th Int Conf Pattern Recognit*. International conference on Pattern Recognition. 2010. Istanbul, Turkey.

66. Docker, https://www.docker.com. Accessed: 2021 Jul 16.

67. Demonstration of, reproducibility and deployability of the computational workflow of the article. *Profiling the baseline performance and limits of machine learning models for adaptive immune receptor repertoire classification*. 2021. https://github.com/KanduriC/demo_reproducibility_kanduricetal2021. Accessed 2021 Jul 16.

68. Kanduri, C. Data used in the article "Profiling the baseline performance and limits of machine learning models for adaptive immune receptor repertoire classification" NIRD Research Data Archive 2021. https://doi.org/10.11582/2021.00064.

69. Kanduri, C. Code and software used in the article "Profiling the baseline performance and limits of machine learning models for adaptive immune receptor repertoire classification". *NIRD Research Data Archive* 2021. https://doi.org/10.11582/2021.00038.

70. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag. https://doi.org/10.1007/978-0-387-98141-3.

71. Inkscape Project. Inkscape. https://inkscape.org/. Accessed: 08, May, 2022.

72. Kanduri, C. Code used in the manuscript revision process of the article "Profiling the baseline performance and limits of machine learning models for adaptive immune receptor repertoire classification" NIRD Research Data Archive. 2021. https://doi.org/10.11582/2021.00146.

73. Kanduri, C, Pavlović, M, Scheffer, L *et al.* Supporting data for "Profiling the baseline performance and limits of machine learning models for adaptive immune receptor repertoire classification." *GigaScience Database*. 2022. http://dx.doi.org/10.5524/102212.