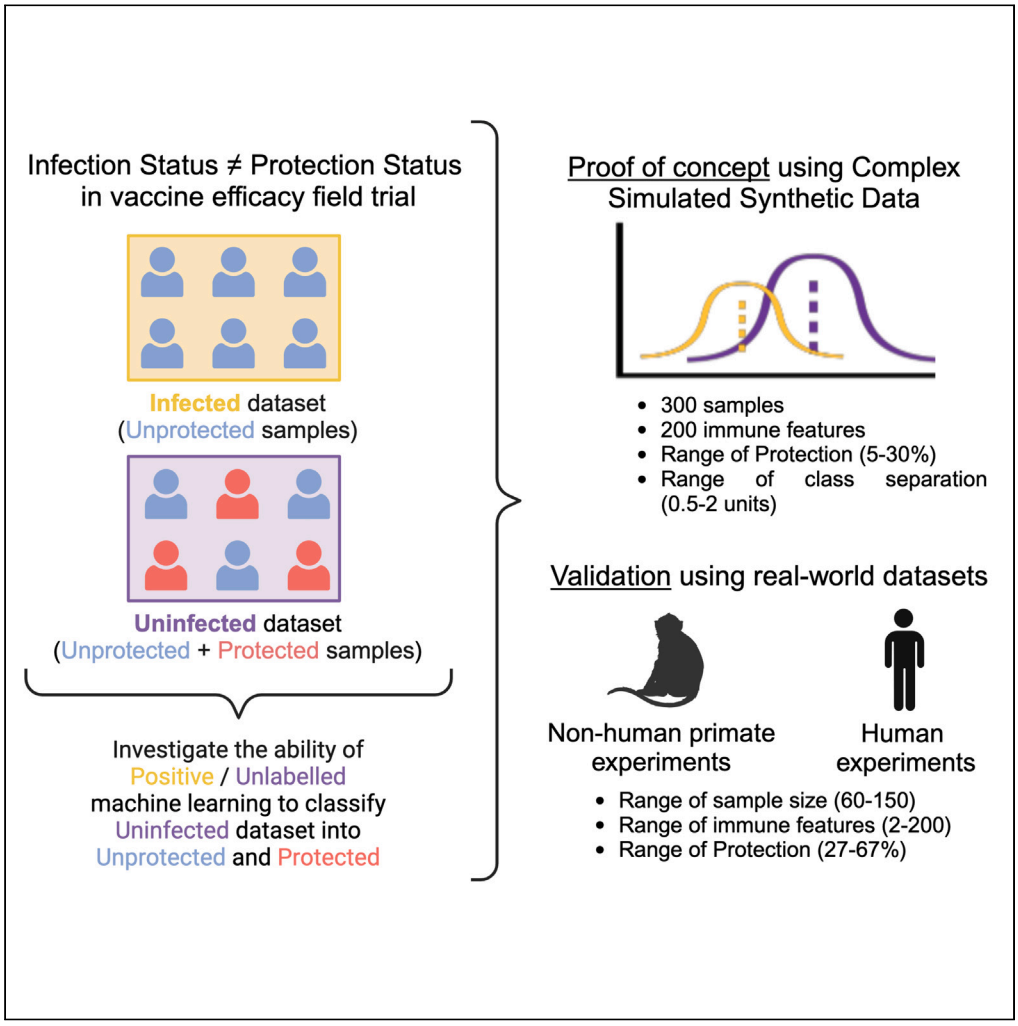
iScience

CellPress
OPEN ACCESS

Article

Positive-unlabeled learning to infer protection status and identify correlates in vaccine efficacy field trials



Shiwei Xu, Natasha S. Kelkar, Margaret E. Ackerman

margaret.e.ackerman@dartmouth.edu

Highlights

Infection status and protection status are not equivalent in vaccine efficacy trials

Protection status labels can be inferred by positive-unlabeled learning approaches

Inferred protection status labels can identify hidden immune correlates

Xu et al., iScience 27, 109086
March 15, 2024
https://doi.org/10.1016/j.isci.2024.109086

iScience

CellPress
OPEN ACCESS

Article

Positive-unlabeled learning to infer protection status and identify correlates in vaccine efficacy field trials

Shiwei Xu,[1] Natasha S. Kelkar,[2] and Margaret E. Ackerman[1,2,3,4,*]

SUMMARY

Correlates of protection (CoPs) are key guideposts that both support vaccine development and licensure as well as improve our understanding of the attributes of immune responses that may directly provide protection. Unfortunately, factors such as low rate of exposure and low efficacy can result in low power to discover correlates in field trials—making it difficult to identify these guideposts for the pathogens against which there is greatest need for further insights. To address this gap, we examine the ability of positive-unlabeled (PU) learning approaches to use immunogenicity data and infection status outcomes to accurately predict protection status. We report a combination of PU bagging and two-step reliable negative techniques that accurately classify the protection status of unlabeled (uninfected) samples from synthetic and real-world humoral immune response profiles in human trials and animal models and lead to the discovery of CoPs that are "missed" using conventional infection status case-control analysis.

INTRODUCTION

Though definitions can vary, in vaccine efficacy studies, correlates of protection (CoPs) are defined as immunization-induced markers that associate with vaccine efficacy against one or more specific clinical endpoints, such as infection, hospitalization, or death following exposure to a given pathogen.[1] Often treated as semantic equivalents, correlates of infection risk (CoRs) are markers that are associated with the clinical endpoint of infection among participants who may or may not have been exposed to the pathogen. This situation is typical of large field efficacy trials,[2,3] in which exposure status cannot be determined. Beyond providing insights into potential mechanisms of immunity, well-established correlates can be used to support approval of a new vaccine based on immunogenicity studies and without requiring larger-scale and longer-term efficacy trials.[4,5]

However, power to discover CoR via infection case-control analysis of immunogenicity data collected in field efficacy trials of vaccines can be limited by the combination of a low rate of exposure and low vaccine efficacy. Such circumstances have repeatedly presented in the setting of HIV-1 vaccines, among which only the RV144 Phase III multicenter, community-based efficacy trial met endpoint efficacy criteria, demonstrating a modest efficacy level of 31.2% at three years post-immunization.[6–8] More recent HVTN705 and HVTN706 trials, as well as HVTN702, designed to follow up on the RV144 trial, have failed to meet overall efficacy criteria.[9,10] Beyond the value of these outcomes in shaping the avenues by which continued clinical research proceeds, these studies nonetheless present the potential opportunity to define correlates,[11,12] which can be of exceptionally high utility in driving further vaccine research and development for such challenging pathogens.

Consistent with the difficulty of achieving vaccine-mediated protection from HIV-1, the first major passive immunization trials of the broadly neutralizing antibody VRC01 (antibody-mediated prevention) failed to meet their pre-defined endpoints for overall efficacy.[6] Nonetheless, correlates of reduced risk of infection, including antibody levels and sensitivity of the infecting virus to neutralization, have been observed.[10,13] These correlates demonstrate that even in the absence of overall efficacy, some individuals were protected from infection by some viruses. Correlates have also been observed for active vaccination studies in which overall efficacy criteria were not met.[14] While such correlates may not provide the same degree of value to regulatory authorities as those observed for interventions that meet efficacy endpoints, their discovery nonetheless suggests the possibility that further insights could be developed from optimized analysis of immune responses among infection cases and controls.

Such additional understanding could be supported by improved experimental and analytical techniques to identify candidate immune correlates of protection. The development of technologies that generate high-dimensional datasets has resulted in rich means to assess genetic, innate, and adaptive markers of protection. With appropriate modeling choices and effective approaches to address redundant and uninformative measures, such as regularization methods and evaluation in the context of cross validation, machine learning approaches have

[1]Quantitative Biological Sciences Program, Dartmouth College, Hanover, NH 03755, USA
[2]Department of Microbiology and Immunology, Geisel School of Medicine at Dartmouth, Dartmouth College, Hanover, NH 03755, USA
[3]Thayer School of Engineering, Dartmouth College, Hanover, NH 03755, USA
[4]Lead contact
*Correspondence: margaret.e.ackerman@dartmouth.edu
https://doi.org/10.1016/j.isci.2024.109086

proved their value in modeling complex data while evading the hazard of overfitting in the context of "wide" data in which features outnumber subjects. Such machine learning models applied to the analysis of high-throughput immunogenicity response profiles have robustly predicted different clinical endpoints and vaccination groups in multiple studies across diverse types of vaccines in both animal models and human trials.[14–18]

Beyond supporting the discovery of relationships with infection status outcomes, a thorough evaluation of immunogenicity data offers additional modeling opportunities. Here, we explore the ability of high-dimensional datasets to support inferences of protection status and uncover novel correlates of protection in vaccine efficacy field studies. For a pathogen with a low exposure rate, uninfected vaccine recipients include subjects that were exposed but protected from infection, and whose response profile offers insights into vaccine efficacy, as well as individuals that were not exposed during follow-up. For a vaccine with low efficacy, the majority of the uninfected vaccine recipients are expected to be uninfected only because they were not exposed.[19] Typical case-control study designs perform analysis to identify response attributes associated with infection status without considering the unprotected individuals among the uninfected cohort, which has the effect of "diluting" power to identify correlates.

Positive-unlabeled (PU) learning, a subclass of semi-supervised learning, offers a means to revisit correlates analysis based on its potential to classify uninfected vaccine recipients according to their inferred protection status. With the assumption that the "unlabeled", or uninfected group consists of both positive (unprotected) and negative (protected) vaccine recipients, the goal of PU learning applied to vaccine efficacy field trials harmonizes with numerous real-world scenarios in other settings in which many samples are "unlabeled" due to the cost and feasibility of labeling.[20,21] In the past two decades, multiple PU learning strategies have been applied to a large variety of tasks in bioinformatic studies, such as disease gene identification, drug-drug interaction, and cancer metastasis prediction.[22–27] Many PU learning strategies fall into the category of "two-step reliable negative" approaches, where "Reliable Negative (RN)" examples are initially identified, and then used together with the Known Positive (KP) examples to label the remaining unlabeled samples using semi-supervised learning techniques. A more recent method developed by Mordelet and Vert termed "positive-unlabeled bagging" adapted bootstrapped aggregation to PU learning tasks, and outperformed alternative state-of-the-art PU learning approaches in scenarios in which positive examples are a minor class among all samples.[21,28]

Given alignment with the goal of correlate identification in human vaccine efficacy trials with low efficacy, we sought to investigate the empirical behavior of PU bagging algorithms in accurately inferring protection status and supporting discovery of correlates, and to compare these outcomes to traditional analysis of infection status labels. We modeled positive-unlabeled immunogenicity datasets with both high-dimensional synthetic datasets and real-world humoral response profiles for which actual protection status was known but could be blinded for analysis. We evaluate a combination of PU bagging and two-step reliable negative strategies to classify unlabeled (uninfected) samples with multiple labeling strategies and compare their protection status prediction and correlate discovery accuracy with infection status labels as well as with fully supervised models trained to predict infection status. This work demonstrates the potential of PU learning-based inferences of protection status to support the discovery of correlates of protection that are missed by traditional analysis of infection status in case-control study designs.

RESULTS

Inference of protection status from synthetic immunogenicity and infection outcome data

To evaluate the ability of PU bagging algorithms to accurately infer protection status class labels (Figure 1A) in the context of high-dimension immunogenicity data, synthetic datasets consisting of 300 samples with 200 features and ground truth protection status labels expected to present varying classification difficulty were generated. The unprotected class was assigned the True Positive label (P) and the protected class the True Negative label (N). To simulate the positive (P)-unlabeled (U) data presented by infection outcome status in vaccine efficacy field trials, the P/U label was generated from ground truth P/N label by randomly sampling a subset of subjects from True Positive (TP) group (Unprotected) as Known Positive (Infected) and pooling the rest of the True Positive group and the True Negative (TN) group (Protected) together as unlabeled (U), to simulate uninfected subjects. Varying baseline P/N classification difficulty was modeled by simulating protected and unprotected class immunogenicity features with different levels of class separation and class imbalance (Figure 1B). With P/U labels assigned to the training set, we performed an ensemble-based PU learning method, termed "PU bagging", on all the samples in training data and scored their probability of assignment to the unprotected class for the subjects in unlabeled and testing sets with aggregated votes from the bootstrapped classifiers (Figure 1C). This score can be considered to be a "pseudo label" that captures different degrees of protection suggested by immunological responses profiles.

We evaluated the performance of PU bagging algorithms with various classifiers, including support vector machines with radial (SVM-RBF) and linear (SVM-linear) kernels, Random Forest (RF), Multilayer Perceptron, Decision tree, Naive Bayes, and K nearest Neighbors (KNN, K = 5) to correctly identify P and N samples. Classification performance on unlabeled and test samples was scored using area under the receiver operator characteristic curve (ROC AUC). This metric provided a means to gauge performance without setting an explicit decision boundary. In many real-world P/U classification tasks, imbalance between Known Positive and unlabeled (U) samples is common, and for vaccines with low efficacy, the positive set is usually a minority among the unlabeled. To understand the influence of the number or proportion of KP samples on prediction performance, we varied the percentage of the randomly selected KP from 5% to 30% of all samples in the synthetic datasets. Under almost all conditions, classification accuracy was better than random, and when there were sufficient True Negatives among the unlabeled set, ROC AUC never fell below 0.8 (Figure 2A). As expected, for the same number of training set samples, higher AUC values were observed with increasing numbers of KP, as well as with increasing class separation. For P/N sample distributions with larger class separation

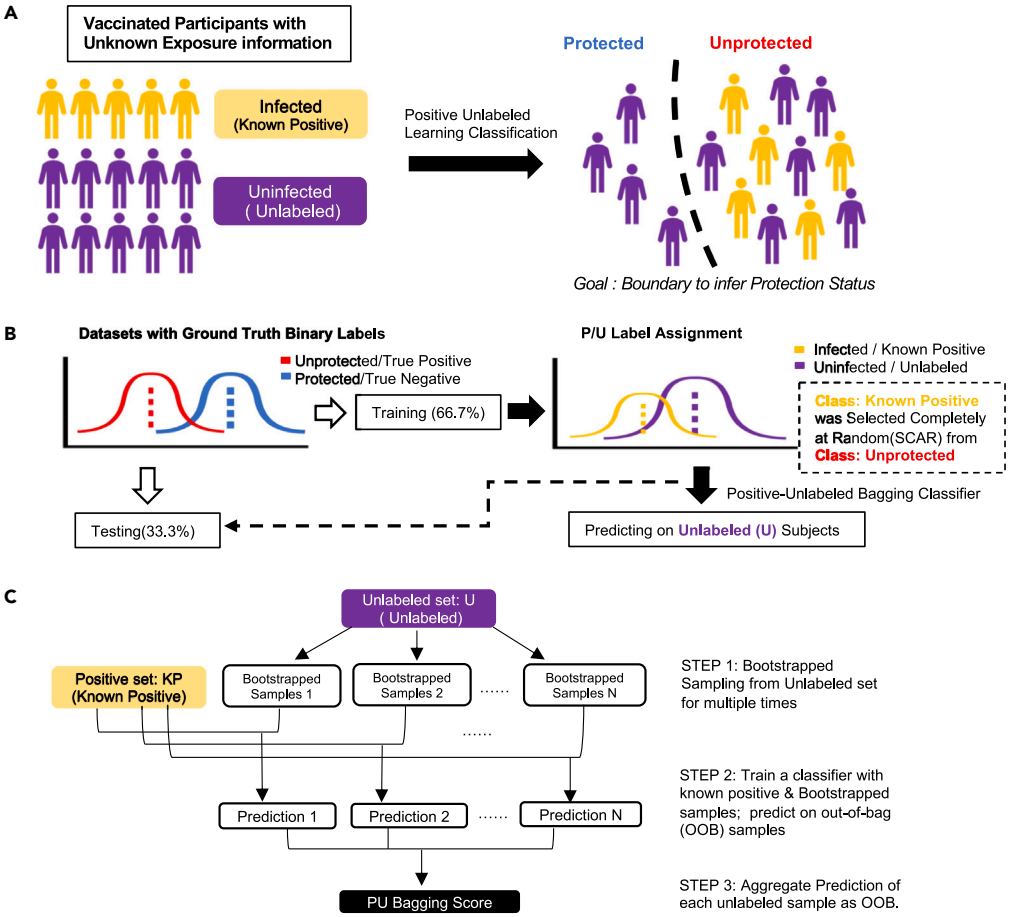


Figure 1. Application of positive-unlabeled learning to infer protection status in vaccine efficacy field trials

(A) Schematic representation of the protection status classification task in vaccine efficacy field trials. The exposure status of infected subjects is definitively known, but uninfected subjects may or may not have been exposed and protected. Positive-unlabeled (PU) learning methods may offer the ability to accurately infer protection status.

(B) Schematic representation of training and testing split in synthetic datasets, PU scenario simulation, and output from PU bagging classifier. Simulated immunogenicity data for protected and unprotected classes was generated. Samples from these "ground truth" classes were drawn at random to represent infected (sampled from the positive/unprotected class) and uninfected (sampled from both the positive/unprotected and negative/protected class) sets. A portion (2/3) was relabeled as positive or unlabeled, used to train a PU-learning based classifier and evaluated on the both the unlabeled samples in train set and the held-out test set (1/3) for validation. This labeling process and assignment of training and test sets was repeated 30 times.

(C) Illustration depicting the process for scoring unlabeled samples with positive-unlabeled bootstrapped aggregation algorithms (PU bagging).[28] By repeatedly (100 times) bootstrapping samples from U set, classifiers built with Known Positive (KP) and bootstrapped U set samples were used to predict the rest of the out-of-bag (OOB) samples. The predicted PU bagging score of U set samples were defined as the frequency of a sample to be predicted as "Positive" when OOB.

and balanced ground truth labels, PU bagging algorithms tended to perform well regardless of the percentage of KP. While substantial differences in prediction performance between the selected classifiers were not observed, and the PU bagging algorithm was developed with linear SVM, the radial kernel SVM classifier generally performed well and was used in further analysis.

To generalize these observations, we simulated P/U classes 30 times for each training dataset and scored subjects in the U and test sets to evaluate bias and variance. For consistency with prior correlates studies, the proportion of KP was set to 20% (n = 40) while class separation and the proportion of unlabeled samples from the protected class were varied. Again, performance was best with larger numbers of True Negatives and with greater distinctions between P/N distributions (Figure 2B). Despite high dimensionality of the underlying data, in which the number of features exceeded the sample size in each bootstrapped dataset replicate, overfitting was not observed, (Figures 2A and 2B). Furthermore, a permutation test, in which ground truth labels were randomly shuffled prior to modeling, was leveraged as a negative control, and exhibited random performance (Figure 2C). As another form of analytical control, a biased SVM classifier was trained to differentiate between infected and uninfected classes (i.e., P versus U). As expected, both overall AUC values across replicates (Figure 2C), and across different class compositions (Figure 2D) were higher for PU bagging scores than for SVM-based classification of infection status, demonstrating the ability of PU learning to capture actual differences in immunogenicity profiles among the unlabeled samples.

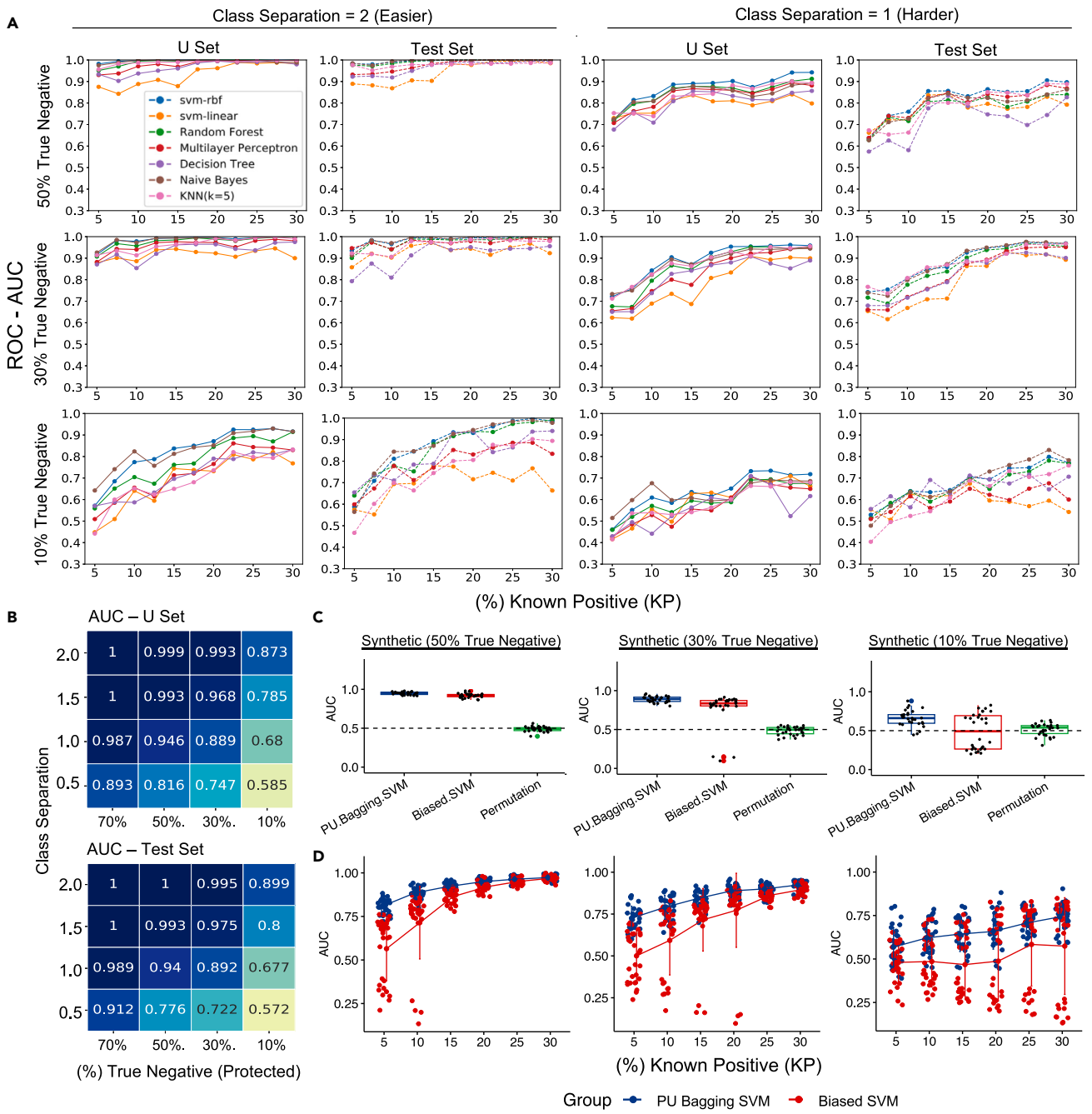


Figure 2. Accuracy of protection status inferences observed across synthetic datasets with varying classification difficulty

(A) Area under the receiver operator characteristic (ROC AUC) curve observed for protection status inferences. Synthetic datasets with varying proportions of (infected) Known Positives (x axis) and (protected) True Negatives (rows) were evaluated across a set of state-of-the-art classification approaches (inset) for unlabeled (U) training data and test data with varying class separation (columns).

(B) Heatmap visualization of mean AUC of PU bagging SVM with a fixed number of Known Positive samples (KP = 40(20%)) across varied class label ratios (columns) and class separations (rows) for unlabeled training data (top) and test data (bottom).

(C) AUC of PU bagging SVM for each replicate compared to a weighted SVM classifier and predictions of permuted class labels with a fixed number of Known Positive samples (KP = 40(20%)) across varied class label ratios (columns) for a class separation of 1. Boxplots illustrate medians and interquartile range observed across 30 PU label simulation replicates.

(D) AUC values observed for unlabeled (U) set samples using PU bagging SVM and supervised SVM classifier across varying percentages of Known Positives. Error bars represent standard deviation.

A

	Datasets	#Instance	# Feature	"Protected" (True Negative)	"Unprotected" (True Positive)	(%) True Negative	# Known Positive	(%) Known Positive
a	Lakhashe et al.	108	195	Group L	Group K + M	33.3	20	18.5
b	Ackerman et al.	60	165	AE239 + IM239	IM Mosaic	66.7	10	16.7
c	Brickley et al.	152	2	No Shedding	Shedding	27.0	30	19.7
d	Bandyopadhyay et al.	98	2	Challenge	Baseline	53.1	20	20.4

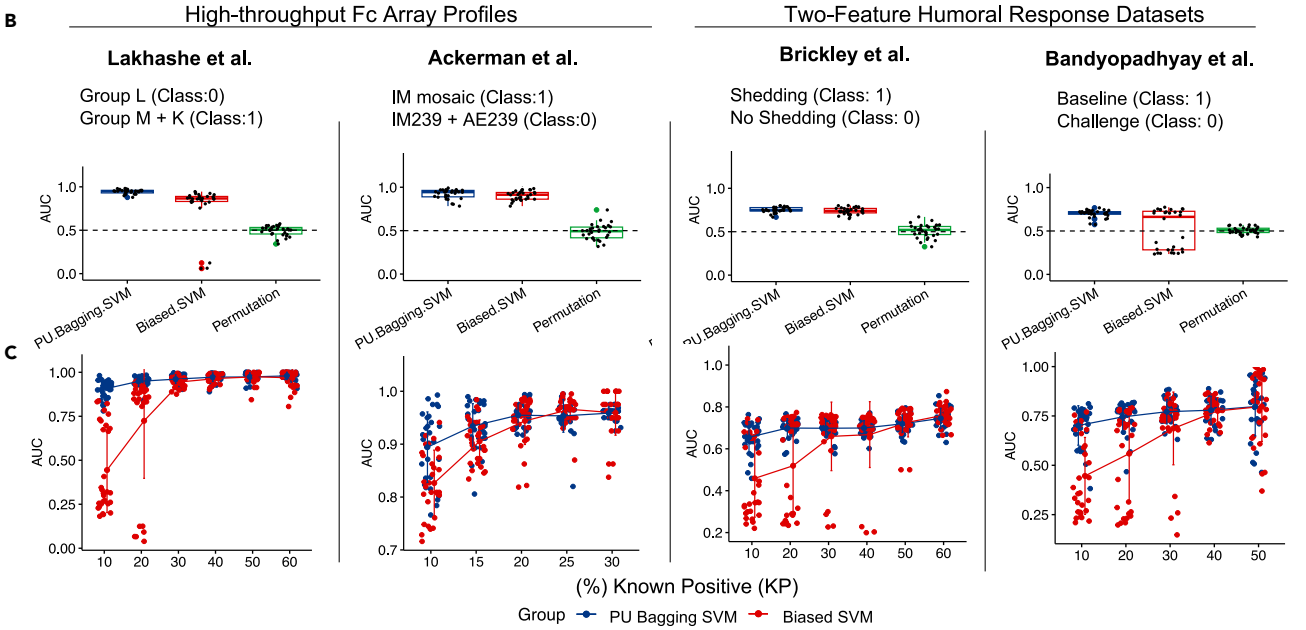


Figure 3. Accuracy of protection status inferences in diverse real-world vaccine efficacy studies

(A) Description of underlying study data and PU simulation conditions in inferences of protection status based on actual immunological profiles observed in each of four preclinical or clinical vaccine efficacy studies. The instance column indicates the number of samples or subjects whose profiles were available in each study, and the feature column indicates the number of immunological measurements evaluated.

(B) AUC of PU bagging SVM using log-transformed immunogenicity data for each replicate compared to weighted SVM classifier and predictions of permuted class labels with the indicated number of Known Positive (KP) samples (inset, n, %). Boxplots illustrate medians, and interquartile range observed across 30 PU label simulation replicates.

(C) AUC values observed for unlabeled (U) set samples using PU bagging and supervised SVM classifiers across varying proportions of Known Positives. Error bars represent standard deviation.

Inferences of protection status in preclinical and clinical efficacy studies

Biological research datasets are generally known to present challenges in modeling because of limited sample size, correlation between features, non-Gaussian distributions, and experimental noise, among other factors. To understand whether the state-of-the-art PU bagging techniques could overcome the practical difficulties in modeling humoral response profiles observed in preclinical and clinical trials, and to investigate the applicability of protection status inference strategies, the proposed algorithms were used to evaluate data from published vaccine efficacy studies (Figure 3A; Table S1). These datasets, which have different sample sizes and immunogenicity feature numbers, included immunization and challenge studies in nonhuman primates[15,29] and controlled human infection studies that employed replicating viral vaccines as a model of pathogen challenge[30,31] (Figure 3A). With the exception of Brickley et al., in which lack of detection of virus shedding was used to indicate protection status, samples from these studies were labeled as True Positive or negative according to protection status outcomes observed to be associated with immunization group or study time point (Table S1). For each study, 20% of the True Positives were randomly assigned as the infected (unprotected, or known P) class. The remainder were assigned the label of the uninfected (unlabeled, or U) class, as were samples in the protected group (True Negatives). As with synthetic data, to define reproducibility of prediction status inferences, this P and U set labeling process was repeated 30 times.

As described previously, PU bagging scores were generated and used to calculate AUC. For all four real-world vaccine efficacy trial immunogenicity datasets, AUC values exceeded both the performance expected at random based on class composition or that observed across replicates following ground truth label permutation (Figure 3B). As might be expected from the skewness in raw data values, AUC values were

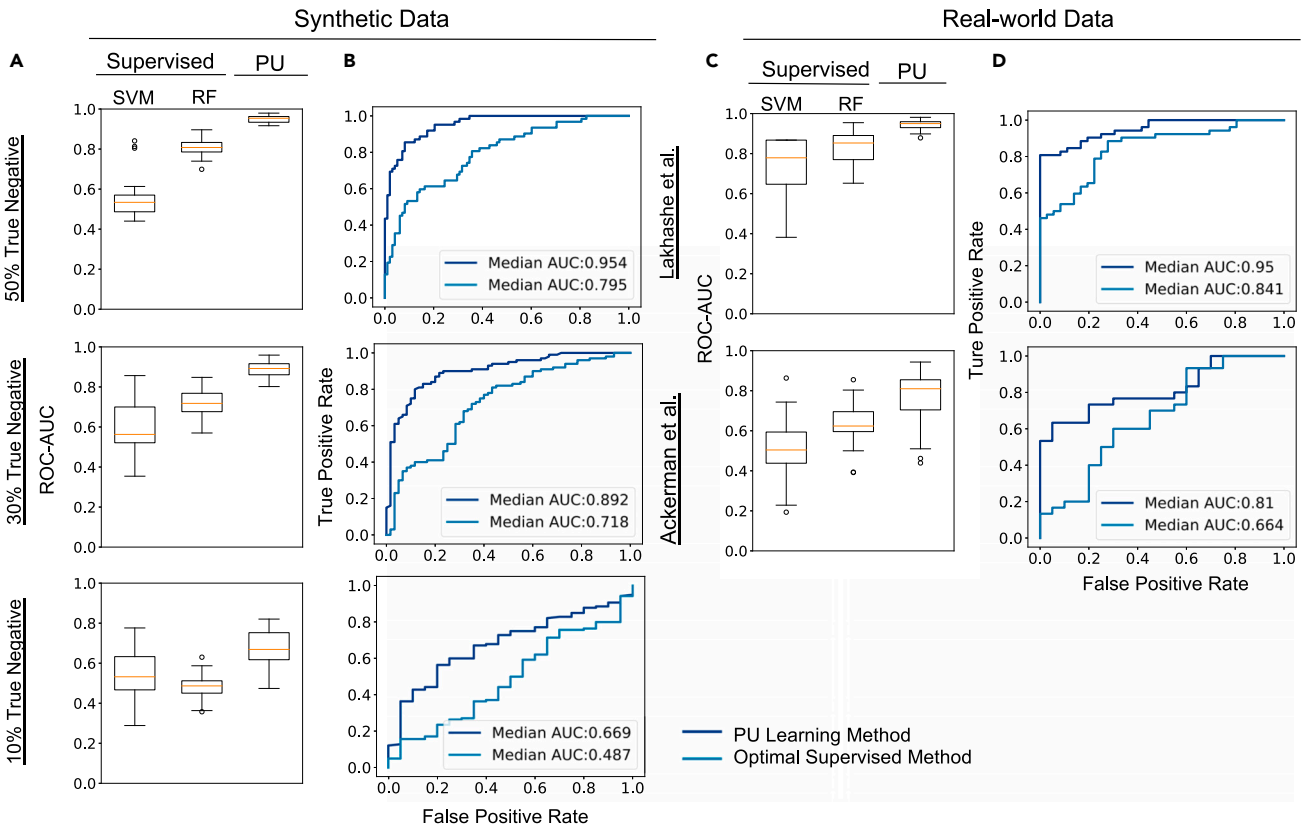


Figure 4. Comparison of PU bagging scores with infection status class probability predicted by supervised methods in high dimensional datasets for U set samples

Comparative performance of fully and semi-supervised models applied to synthetic (A and B) and real-world (C and D) data. Results for synthetic data are reported for fully supervised (SVM and RF) and PU models for varying proportions of True Negatives (rows).

(A and C) Area under the receiver operator characteristic (ROC AUC) curve observed for U set class inferences using probability score rank against ground truth label under a fully supervised SVM classifier, random forest (RF) classifier and PU bagging SVM across replicates. Boxplots illustrate medians and interquartile range observed across 30 PU label simulation replicates.

(B and D) Median ROC curve of the supervised learning method that obtained optimal average classification performance and PU learning method. AUCs for each ROC curve are annotated in bottom right.

improved when log-transformed rather than raw immunogenicity data were used (data not shown). Lastly, we compared the performance of the bagging approach to a biased SVM classifier. For both fixed (Figure 3B) and varied (Figure 3C) proportions of known positive samples, the PU learning method exhibited superior performance, particularly under conditions in which fewer known positives were available. These results demonstrate that PU bagging can reliably infer protection status from real-world immunogenicity and vaccine efficacy data.

Comparison between PU learning and conventional supervised learning

Other data-driven approaches to classify or score patterns of responses among uninfected may hold similar value. Whereas the emphasis in PU learning in this setting is to identify unprotected subjects from among uninfected individuals, supervised learning can be applied to classify subjects by infection status. In principle, both methods can be used to provide a score or rank for U set samples, and robust relationships between infection status and protection status could be discovered using this approach despite the mixed composition of the uninfected class. To determine whether conventional supervised learning methods provide similar value, two frequently used supervised machine learning classifiers, SVM and RF, were trained to differentiate infected from uninfected subjects for both synthetic and real-world datasets.

Here, comparisons were performed under a challenging scenario in which the positive class was the minority group compared to the unlabeled class, yielding relatively poor generalizability to predict infection status in the context of 3-fold cross validation (Table S2). Nonetheless, under many conditions, when infection status class probabilities were used as a score to predict the protection status of the unlabeled, uninfected samples, these approaches generally performed better than expected at random, providing a basis for their potential utility. The RF classifier generally outperformed SVM classifier in classifying the U set samples according to protection status (Figures 4A–4C). However, PU learning generally outperformed both of these fully supervised, cross-validated approaches (Figure 4). All three methods performed less well when there were few protected individuals represented in the unlabeled class (Figures 4A and 4B). Additionally, even though the

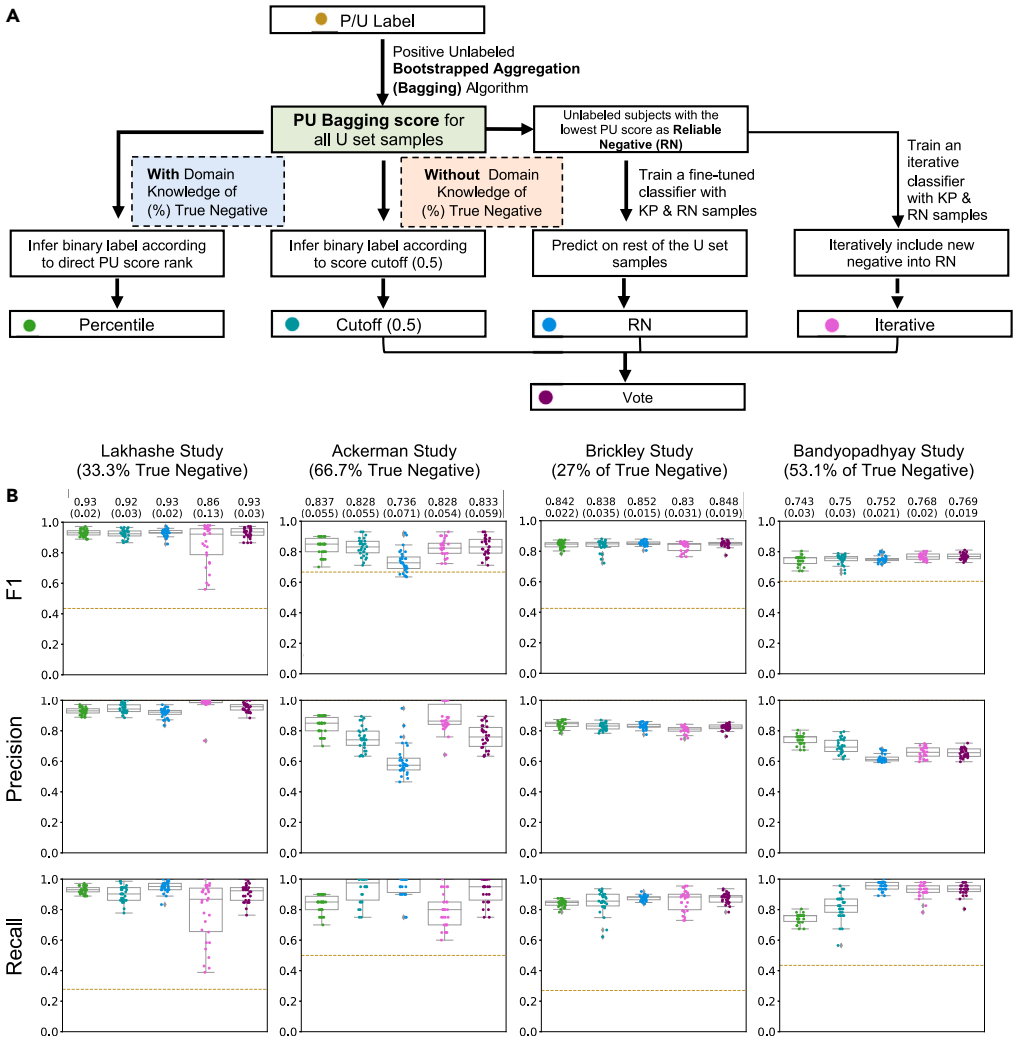


Figure 5. Prediction inference accuracy across distinct threshold setting strategies

(A) Schematic diagram of candidate labeling strategies to infer protection status labels based on scores obtained from PU bagging SVM classifier.

(B) Performance metrics including F1 (top), precision (center), and recall (bottom) scores of the binary labels inferred by each candidate labeling strategy from replicates across vaccine efficacy studies (columns). Mean and (standard deviation) of F1 scores from 30-time PU label simulation replicates are presented at top. Dotted lines depict classification performance of infection status labels.

supervised RF approach also employed bootstrapped aggregation (bagging), the PU bagging algorithm exhibited better performance in relabeling the U set samples (Figures 4B and 4D). These results further demonstrate the potential value of PU learning methods to inferring the protection status of uninfected subjects, especially in high-dimensional datasets where strong relationships with infection status are more likely to exist by chance as irrelevant covariates associated with a small KP class.

Comparison of methods to set explicit class boundaries

While PU bagging scores clearly captured differences among protected and unprotected subjects, these differences were oriented on a continuous scale. While it is likely that protection is not best represented as a categorical outcome, protection status class labels are of utility both to benchmarking model performance as well as to conducting analysis to identify correlates of protection. Furthermore, in simulations in which imbalanced numbers of TP and TN were present, the binary class prediction from PU bagging score with a cutoff at 0.5 performed less well than setting the cutoff according to the percentage of TN (Figure S1A). Hence, we decided to compare and combine the potential of different strategies to develop a binary classifier to distinguish between uninfected protected and infected protected groups (Figure 5A). First, we considered setting the boundary between classes at the midpoint value in PU scores (0.5). Additionally, we considered both standard and iterative reliable negative approaches. In the first, a biased SVM classifier was trained in the context of 3-fold cross validation to discriminate between Known Positives and RN, or the unlabeled samples with the lowest PU scores. Moreover, an iterative process was adapted to

expand newly predicted negative samples to the RN set.[32] In this approach, we employed a weighted SVM classifier in which the group of RN samples was expanded until either no additional RN was identified, or 20 iterations were reached. A combination of these three strategies, in which each was given a vote was used to aggregate predictions, based on the hypothesis that voting had the potential to decrease the risk of outlier prediction performance as compared to a single labeling strategy. The last approach considered a scenario in which the proportion of protected subjects among the uninfected group was known, for example, on the basis of overall efficacy data, and the boundary between classes was set based on the PU score rank percentile.

The performance of each method to set the protection status class boundary was defined for a set of evaluation metrics across 30 replicated positive-unlabeled simulations for both synthetic and actual immunogenicity datasets. Each boundary-setting approach yielded performance better than expected at random (Figures 5B and S1). In general, studies or conditions under which a higher ROC AUC was obtained from bagging SVM corresponded to those with better binary classification prediction metric outcomes for unlabeled samples. Between iterative and single-classifier RN methods, recall scores were generally higher for single-classifier RN while precision was higher for the iterative approach. The iterative approach tended to yield a superior F1 score, which represents the harmonic means of recall and precision metrics when the ground truth label class ratio was biased to include more True Negative samples; in contrast, the single-classifier RN method outperformed the iterative approach when there were fewer True Negative samples (Figures 5B and S1). Given these apparent tradeoffs, the vote-based approach demonstrated its hypothetical advantage. Performance was also good but not generally superior when domain knowledge regarding overall efficacy was modeled to set the boundary between protected and unprotected subjects based on the expected proportion of True Negatives among unlabeled samples. This result suggests that such knowledge can contribute but is not required to support accurate protection status predictions.

Identification of candidate correlates of protection

Lastly, we sought to assess the ability of the predicted labels from PU based methods to reliably identify CoP for the two studies with high-dimensional immunological profiles. Immune features were ranked according to statistical significance associated with ground truth protection status class labels (Figure 6A), and compared to that observed for Positive/Unlabeled class labels (Figure 6B), and for predicted protection status labels (Figure 6C). This analysis was repeated for different Positive/Unlabeled class compositions, and for each dataset, results are presented from three different simulation scenarios, in which the predicted labels achieved maximum, median, and minimum prediction performance based on F1 score (Figures 6B and 6C).

Whereas a number of features were associated with protection status group, these relationships were frequently obscured in the context of infection status (Positive/Unlabeled) labels. For example, no correlates were observed in many infection status label simulations. In cases where some features did reach nominal significance under infection status labels, confidence was considerably lower than that observed with protection status labels. Indeed, because these observations show how poor a proxy infection status can be for protection status in identifying correlates, they provide strong motivation for this work. In contrast, inferences of protection status based on immunogenicity data frequently supported discovery of these correlates, often with similar confidence as observed to ground truth protection class labels. As expected, replicates with a higher degree of agreement between predicted and ground truth labels resulted in greater and more confident recovery of correlates of protection, as well as a lower false discovery rate. However, some correlates were still missed, and false discoveries were made even in the context of even the most accurate protection status inferences (Figure 5B). While confidence tended to be lower for both false discoveries and for true correlates that failed to be rediscovered, caution is clearly warranted in interpretation.

DISCUSSION

Vaccine efficacy field trials help to establish the ability of a vaccine to protect at-risk populations. Beyond providing evidence for or against the effectiveness of a vaccine, trials present opportunities to define the attributes of protective or ineffective responses to immunization. When observed, correlates act as key signposts in directing and prioritizing subsequent vaccine clinical trials or alternative interventions. With the goal of supporting identification of candidate correlates from case-control study designs evaluating responses for vaccines with low efficacy, we applied PU learning approaches to compare immunogenicity profiles among vaccine recipients and to re-classify uninfected subjects as "was/would have been protected" and "was/would have been unprotected."

To this end, we evaluated the potential of PU learning methods that use bagging, or bootstrap aggregation, as a means to compensate for high variance in the context of biological datasets that are inherently noisy. For variably protective and immunogenic vaccines, synthetic immunogenicity datasets for protected and unprotected vaccine recipient response profiles were modeled. These class labels were partially blinded to recapitulate the infection status outcomes typical of field trials in which not all participants are exposed to the pathogen and means to differentiate subjects based on exposure do not exist. Whereas a prior study modeling highly simplified response profiles provided proof of concept for this approach,[19] immunogenicity profiles here better reflect a complex combination of responses, in which some but not all features associated with protection, and many features were assessed. Across diverse classification algorithms, degrees of class separation, and extents of vaccine efficacy, protection status could be accurately inferred. Correlates that did not relate to infection status could be discovered from the analysis of responses among subjects on the basis of inferred protection status labels.

While suggestive of the potential value of PU learning to contribute to correlate identification, simulated data may or may not reflect typical response profiles, class compositions, and other features of real-world studies. We focused on high-dimensional synthetic data as results with sparser feature sets have been previously reported,[19] and we expect that the need for insights into how to protect

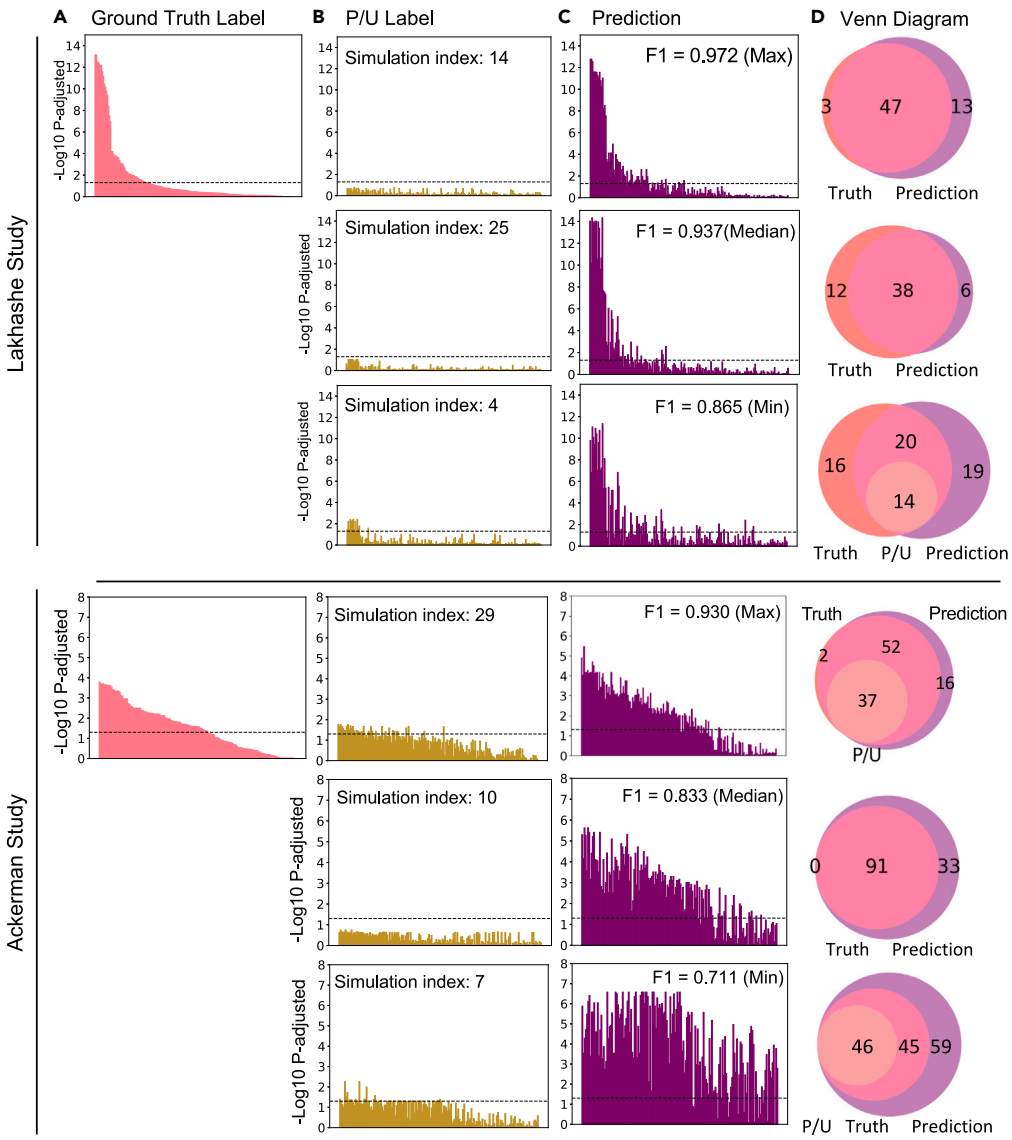


Figure 6. Discovery of correlates of protection based on protection status inferences

(A–C) Results of Mann-Whitney U tests for each immunogenicity feature for ground truth (A), Positive/Unlabeled (B), or predicted protection status (C) labels, ranked by confidence according to ground truth class labels. Dashed lines indicate an adjusted (Benjamini-Hochberg) p value of 0.05.

(D) Venn diagram demonstrating the overlap between statistically significant immunogenicity features identified using ground truth, P/U labels and predicted protection status labels. Results for PU simulations with maximum, median, and minimum F1 scores (inset) are shown.

against the most challenging pathogens is likely to motivate collection of high-dimensional data. Overall, while we aimed to cover parameter space sufficient to support generalizations, our analysis is not exhaustive, and cases poorly suited for this approach are certain to exist.

To address performance in the context of actual immunogenicity data and protection outcomes, we repeatedly repeated the partial blinding, class inference, and correlate analysis process for a set of four distinct preclinical or clinical vaccine trials in which vaccine recipients were challenged and protection status was definitively known. These studies included exploratory nonhuman primate HIV vaccines, as well as human polio vaccines, and immunogenicity profiles comprised of up to almost 200 features for 60 to 152 samples. Samples with protective responses varied from 27 to 67% of unlabeled, uninfected groups, while the number of definitively labeled infected (unprotected) subjects varied from 10 to 30, similar to the 25–41 infection cases in the HIV vaccine efficacy field trials that motivated this work.

These real-world datasets presented greater challenges to modeling because of limited sample size, the correlation between features, non-Gaussian distribution of variables, and unavoidable experimental noise. Additionally, a number of these studies were powered to detect protection at the group level as compared to a placebo or other control, rather than to evaluate protection status among individuals within

immunization groups or time points. As a result, immunogenicity profiles were confounded by differences in regimen such as route of immunization or immunogen sequence. For these studies, we defined the protection status of individual samples based on the degree of protection observed at their group (Lakhashe et al., Ackerman et al.) or immunization time point (Bandyopadhyay et al.) level rather than based on challenge outcome results at the level of individual subjects. Nonetheless, while each individual subject in a protected group may not have exhibited a high degree of protection, we only evaluated studies in which each group modeled as comprised of protected subjects exhibited a statistically significant degree of protection. In one study (Brickley et al.), such confounding factors were not present, and we were able to assign protection status at the individual level—best reflecting typical efficacy trials in which only a single type of intervention is tested to a control. Further exploration of protection status outcome prediction in the context of immunogenicity differences that are not associated with different interventions would further support our observations. Nonetheless, evaluation of classification strategies on these diverse real-world datasets demonstrated that PU bagging classification possessed high potential to recover the ground truth protection status labels. Consistent with previous findings, PU bagging classifiers in high dimensional spaces outperformed a fully supervised machine learning classifier trained to predict infection status. Poor performance of supervised methods was more frequently observed when the proportion of Known Positive samples was low and for high dimensional datasets. These results suggest that when there are few infected subjects but many features, supervised results modeling infection status may be less reliable than PU learning. While fully and semi-supervised approaches were not extensively compared, these results suggest that considerable biological noise is introduced in the modeling process by treating all uninfected subjects as a single rather than mixed class. Conceptually, this outcome makes sense as features that relate to protection status are expected to classify infection status imperfectly, whereas features that relate to infection status only are best fit to the explicit modeling objective but less well fit to the biological objective.

A prior report demonstrated the potential of the reliable negative approach, a PU learning method to solve the "labeling problem" inherent in the analysis of case-control study designs in which identification of CoP is desired.[19] Comparison of bagging and reliable negative methods to infer protection status labels demonstrated comparable performance overall. Hence, we decided to build a pipeline that combines PU bagging and Reliable Negative approaches. Additionally, the consistency of this finding across repeated simulations from all the datasets reinforced confidence in applying the PU bagging classifier as the initial classifier used to define sample score and rank, as KP samples were frequently the minority group in real-world datasets. While definitive means to set a boundary between classes for binary class predictions are an unsolved challenge in PU learning, we employed a second step in which prior knowledge of overall efficacy rates, a mid-scale cutoff, RN, or iterative RN approaches were used or aggregated. Among these class labeling strategies, we identified a pattern between the P/N ground truth label ratio and the preferable choice: iteratively including negative samples with an SVM classifier outperformed a fine-tuned SVM classifier with more TN samples among the unlabeled. While overall efficacy rates are available for some vaccines, confidence intervals on these values can be wide, motivating exploration of strategies that don't rely on such prior information. To this end, aggregation across multiple predictions resulted in inferred protection status class labels that approximated ground truth despite the presence of outliers among individual prediction methods.

Notwithstanding the potential of this or other PU learning-based statistical pipelines to classify unlabeled samples with high dimensional biological datasets, we pinpointed several limitations of this approach. Firstly, with a fixed percentage of KP among all samples in the PU simulation, we observed a lower AUC when fewer samples were TN; meanwhile, a rapid decrease in prediction performance was observed when fewer KP were available, especially when baseline classification difficulty increased. While both results could be explained by the decreased percentage of KP among True Positive samples, extreme imbalance of class labels may be ubiquitous in many real-world classification tasks and particularly tricky when most samples are unlabeled. Several previous studies specifically discuss PU learning under imbalanced labels. Among them, Jiang et al. developed PU learning-based methods against "extreme imbalance" between KP and U sets, which was designed for the datasets with less than 5% of KP among overall samples.[33]

Secondly, we did not assess the labeling capability of this pipeline when selection bias was present among positive samples. In this work, the positive unlabeled scenarios were simulated under the "*select completely at random" (SCAR)* assumption, where the distribution of the KP examples was hypothesized to be equivalent to that of the underlying positive samples in the unlabeled set.[34] However, Bekker et al. have suggested that the positive examples might be subject to selection bias due to other covariates in some real-world applications such as in clinical diagnosis.[35] While we did not consider this additional confounding factor, a recent study that reweighted the PU samples with a propensity score under the framework of a bagging classifier suggests the potential to address the effects of selection bias in model building.[36] Prior studies have investigated the value of modeling exposure risk or other factors that are independent of immunological response profiles in more traditional analysis frameworks,[37] and could further inform application of PU learning techniques.

Thirdly, even in the context of accurate protection status class inferences, true CoP were not always uncovered, and false discoveries were made. While the strongest CoP were those most likely to be "rediscovered" and the weakest were the most likely to be lost, these results indicate that correlates discovered using this inference approach are certainly best considered to be candidate or inferred CoP rather than definitive observations. We envision those adaptations to this approach, such as subsampling of immunogenicity features or samples into distinct prediction and correlate sets could increase the robustness of this approach and lead to fewer false discoveries. However, we do not expect that candidate correlates discovered in this way are likely to provide the same value to regulators in the vaccine evaluation and approval process as correlates derived from infection status. Nonetheless, given the challenges to achieving vaccine-mediated protection against a number of infectious diseases, and the limited insights into protective mechanisms in humans to date for some pathogens, when candidate CoP are identified, the ability to follow up on such leads may offer alternatives to strictly empirical vaccine development approaches, as well as avenues for new interventions.

Lastly and most critically, while we have modeled the application of PU learning to discover candidate correlates in the context of ground truth knowledge of protection status, deployment of this approach in field trials of vaccines will not present this means to explicitly validate prediction accuracy. While combining permutation approaches and the spy positive method may offer a means to vet results by setting baseline expectations for outcomes expected at random,[38,39] poor prediction accuracy results using the pipeline described here will not be obvious. Beyond improving the ability of state-of-art methods to accurately predicting the ground truth label, future studies should also consider the validation of the predicted outcome with innovative methodologies and study designs. For example, in contexts where markers of exposure exist, they could offer means to validate protection status predictions. If reliable CoRs were discovered from the analysis of infection status, these observations would be expected to be strengthened when unprotected subjects are removed from the uninfected group. Indirect evidence suggestive of model reliability could come from differences in time to infection among True Positive that relates to classification confidence, or in outcomes available in further follow-up of study participants, or in the context of other vaccine or study groups. In other cases, inferred candidate CoP may be best validated in experiments designed to directly address their potential mechanistic contributions.

While we do not envision that the use of PU learning and protection status inferences will become a primary means of correlate discovery in field trials of vaccines and case-control study analysis, this work establishes its potential in this context. There is certainly broader utility to applying artificial intelligence methods in clinical research in which positive and unlabeled sample points are available. The objectives of PU learning also conform to tasks in biomarker identification for cancer prognosis and prediction in survival and estimating disease risks for child health surveillance.[25,40] Principally, the major barrier to direct application of the PU classification methods explored herein to real-world data are the lack of validation for the predicted outcome in the absence of ground truth labels. By virtue of this fact, traditional feature selection, parameter fine-tuning approaches, and evaluation metrics used in the setting of fully supervised learning tasks were difficult to leverage. Further P/U simulation experiments to compare multiple candidate methods on datasets that approximate real-world datasets in data type and characteristics could aid the researchers in optimization of choices of PU learning strategies. In the meantime, this work provides a strong rationale for the potential value of this semi-supervised approach to complement traditional case-control analysis designs in contributing to identification of key footholds in the development of vaccines against the most challenging pathogens.

Limitations of the study

Limitations of this study include reliance on protection outcomes at the immunization group level for some of the real-world datasets. Additionally, we did not assess the labeling capability of this pipeline when selection bias was present among positive samples, which may be expected in some real-world applications. Lastly, while we sought to analyze a diversity of datasets, how well the results reported here will extend to new data will certainly be context-dependent and determined by its specific characteristics.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Datasets
  - P/U Label assignment
  - Scoring and ranking unlabeled samples with PU learning
  - Permutation comparison
  - Supervised learning comparison
  - Binary classification of U set
  - Evaluation metrics
  - Analysis and visualization

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109086.

ACKNOWLEDGMENTS

This study was supported in part by NIAID R56AI165448 and P01AI162242. The graphical abstract was created using BioRender.

AUTHOR CONTRIBUTIONS

Investigation and coding: S.X. and N.K.; data visualization: S.X.; writing- original draft: S.X.; writing-reviewing and editing: all authors. supervision: M.E.A.; conceptualization and funding: M.E.A.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 29, 2023
Revised: November 29, 2023
Accepted: January 29, 2024
Published: February 2, 2024

REFERENCES

1. Gilbert, P.B., Donis, R.O., Koup, R.A., Fong, Y., Plotkin, S.A., and Follmann, D. (2022). A Covid-19 Milestone Attained - A Correlate of Protection for Vaccines. N. Engl. J. Med. *387*, 2203–2206. https://doi.org/10.1056/NEJMp2211314.
2. Plotkin, S.A., and Gilbert, P.B. (2012). Nomenclature for immune correlates of protection after vaccination. Clin. Infect. Dis. *54*, 1615–1617. https://doi.org/10.1093/cid/cis238.
3. World Health Organization (2013). Correlates of Vaccine- Induced Protection: Method and Implications. *WHO/IVB/13.01*. https://apps.who.int/iris/bitstream/handle/10665/84288/WHO_IVB_13.01_eng.pdf.
4. Koup, R.A., Donis, R.O., Gilbert, P.B., Li, A.W., Shah, N.A., and Houchens, C.R. (2021). A government-led effort to identify correlates of protection for COVID-19 vaccines. Nat. Med. *27*, 1493–1494. https://doi.org/10.1038/s41591-021-01484-6.
5. Jiang, H.D., Zhang, L., Li, J.X., and Zhu, F.C. (2021). Next Steps for Efficacy Evaluation in Clinical Trials of COVID-19 Vaccines. Eng. Plast. *7*, 903–907. https://doi.org/10.1016/j.eng.2021.04.013.
6. Rerks-Ngarm, S., Pitisuttithum, P., Nitayaphan, S., Kaewkungwal, J., Chiu, J., Paris, R., Premsri, N., Namwat, C., de Souza, M., Adams, E., et al. (2009). Vaccination with ALVAC and AIDSVAX to Prevent HIV-1 Infection in Thailand. N. Engl. J. Med. *361*, 2209–2220. https://doi.org/10.1056/NEJMoa0908492.
7. Kim, J., Vasan, S., Kim, J.H., and Ake, J.A. (2021). Current approaches to HIV vaccine development: a narrative review. J. Int. AIDS Soc. *24*, e25793.
8. Karasavvas, N., Billings, E., Rao, M., Williams, C., Zolla-Pazner, S., Bailer, R.T., Koup, R.A., Madnote, S., Arworn, D., Shen, X., et al. (2012). The Thai Phase III HIV Type 1 Vaccine Trial (RV144) Regimen Induces Antibodies That Target Conserved Regions Within the V2 Loop of gp120. AIDS Res. Hum. Retrovir. *28*, 1444–1457. https://doi.org/10.1089/aid.2012.0103.
9. Hammer, S.M., Sobieszczyk, M.E., Janes, H., Karuna, S.T., Mulligan, M.J., Grove, D., Koblin, B.A., Buchbinder, S.P., Keefer, M.C., Tomaras, G.D., et al. (2013). Efficacy trial of a DNA/rAd5 HIV-1 preventive vaccine. N. Engl. J. Med. *369*, 2083–2092. https://doi.org/10.1056/NEJMoa1310566.
10. Corey, L., Gilbert, P.B., Juraska, M., Montefiori, D.C., Morris, L., Karuna, S.T., Edupuganti, S., Mgodi, N.M., DeCamp, A.C., Rudnicki, E., et al. (2021). Two randomized trials of neutralizing antibodies to prevent HIV-1 acquisition. N. Engl. J. Med. *384*, 1003–1014.
11. Rolland, M., Tovanabutra, S., deCamp, A.C., Frahm, N., Gilbert, P.B., Sanders-Buell, E., Heath, L., Magaret, C.A., Bose, M., Bradfield, A., et al. (2011). Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. Nat. Med. *17*, 366–371. https://doi.org/10.1038/nm.2316.
12. Ng'uni, T., Chasara, C., and Ndhlovu, Z.M. (2020). Major Scientific Hurdles in HIV Vaccine Development: Historical Perspective and Future Directions. Front. Immunol. *11*, 590780. https://doi.org/10.3389/fimmu.2020.590780.
13. Seaton, K.E. (2023). Pharmacokinetic Serum Concentrations of VRC01 Correlate with Prevention of HIV-1 Acquisition.
14. Neidich, S.D., Fong, Y., Li, S.S., Geraghty, D.E., Williamson, B.D., Young, W.C., Goodman, D., Seaton, K.E., Shen, X., Sawant, S., et al. (2019). Antibody Fc effector functions and IgG3 associate with decreased HIV-1 risk. J. Clin. Invest. *129*, 4838–4849. https://doi.org/10.1172/JCI126391.
15. Ackerman, M.E., Das, J., Pittala, S., Broge, T., Linde, C., Suscovich, T.J., Brown, E.P., Bradley, T., Natarajan, H., Lin, S., et al. (2018). Route of immunization defines multiple mechanisms of vaccine-mediated protection against SIV. Nat. Med. *24*, 1590–1598. https://doi.org/10.1038/s41591-018-0161-0.
16. Pittala, S., Bagley, K., Schwartz, J.A., Brown, E.P., Weiner, J.A., Prado, I.J., Zhang, W., Xu, R., Ota-Setlik, A., Pal, R., et al. (2019). Antibody Fab-Fc properties outperform titer in predictive models of SIV vaccine-induced protection. Mol. Syst. Biol. *15*, e8747. https://doi.org/10.15252/msb.20188747.
17. Bradley, T., Pollara, J., Santra, S., Vandergrift, N., Pittala, S., Bailey-Kellogg, C., Shen, X., Parks, R., Goodman, D., Eaton, A., et al. (2017). Pentavalent HIV-1 vaccine protects against simian-human immunodeficiency virus challenge. Nat. Commun. *8*, 15711. https://doi.org/10.1038/ncomms15711.
18. Felber, B.K., Lu, Z., Hu, X., Valentin, A., Rosati, M., Remmel, C.A.L., Weiner, J.A., Carpenter, M.C., Faircloth, K., Stanfield-Oakley, S., et al. (2020). Co-immunization of DNA and Protein in the Same Anatomical Sites Induces Superior Protective Immune Responses against SHIV Challenge. Cell Rep. *31*, 107624. https://doi.org/10.1016/j.celrep.2020.107624.
19. Kelkar, N.S., Morrison, K.S., and Ackerman, M.E. (2023). Foundations for improved vaccine correlate of risk analysis using positive-unlabeled learning. Hum. Vaccines Immunother. *19*, 2204020. https://doi.org/10.1080/21645515.2023.2204020.
20. Youngs, N., Shasha, D., and Bonneau, R. (2015). Positive-unlabeled Learning in the Face of Labeling Bias (IEEE), pp. *639–645*.
21. Liu, B., Dai, Y., Li, X., Lee, W.S., and Yu, P.S. (2003). Building Text Classifiers Using Positive and Unlabeled Examples (IEEE), pp. 179–186.
22. Kolosov, N., Daly, M.J., and Artomov, M. (2021). Prioritization of disease genes from GWAS using ensemble-based positive-unlabeled learning. Eur. J. Hum. Genet. *29*, 1527–1535. https://doi.org/10.1038/s41431-021-00930-w.
23. Zeng, X., Zhong, Y., Lin, W., and Zou, Q. (2020). Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. Briefings Bioinf. *21*, 1425–1436. https://doi.org/10.1093/bib/bbz080.
24. Yang, P., Li, X.-L., Mei, J.-P., Kwoh, C.-K., and Ng, S.-K. (2012). Positive-unlabeled learning for disease gene identification. Bioinformatics *28*, 2640–2647. https://doi.org/10.1093/bioinformatics/bts504.
25. Zhou, J., Lu, X., Chang, W., Wan, C., Lu, X., Zhang, C., and Cao, S. (2022). PLUS: Predicting cancer metastasis potential based on positive and unlabeled learning. PLoS Comput. Biol. *18*, e1009956. https://doi.org/10.1371/journal.pcbi.1009956.
26. Zheng, Y., Peng, H., Zhang, X., Zhao, Z., Gao, X., and Li, J. (2019). DDI-PULearn: a positive-unlabeled learning method for large-scale prediction of drug-drug interactions. BMC Bioinf. *20*, 661. https://doi.org/10.1186/s12859-019-3214-6.
27. Ju, Z., and Wang, S.Y. (2020). Computational Identification of Lysine Glutarylation Sites Using Positive-Unlabeled Learning. Curr. Genom. *21*, 204–211. https://doi.org/10.2174/1389202921666200511072327.
28. Mordelet, F., and Vert, J.P. (2014). A bagging SVM to learn from positive and unlabeled examples. Pattern Recogn. Lett. *37*, 201–209. https://doi.org/10.1016/j.patrec.2013.06.010.
29. Lakhashe, S.K., Amacker, M., Hariraju, D., Vyas, H.K., Morrison, K.S., Weiner, J.A., Ackerman, M.E., Roy, V., Alter, G., Ferrari, G., et al. (2022). Cooperation Between Systemic and Mucosal Antibodies Induced by Virosomal Vaccines Targeting HIV-1 Env: Protection of Indian Rhesus Macaques Against Low-Dose Intravaginal SHIV Challenges. Front. Immunol. *13*, 788619. https://doi.org/10.3389/fimmu.2022.788619.
30. Brickley, E.B., Strauch, C.B., Wieland-Alter, W.F., Connor, R.I., Lin, S., Weiner, J.A.,

Ackerman, M.E., Arita, M., Oberste, M.S., Weldon, W.C., et al. (2018). Intestinal Immune Responses to Type 2 Oral Polio Vaccine (OPV) Challenge in Infants Previously Immunized With Bivalent OPV and Either High-Dose or Standard Inactivated Polio Vaccine. J. Infect. Dis. *217*, 371–380. https://doi.org/10.1093/infdis/jix556.

31. Bandyopadhyay, A.S., Gast, C., Brickley, E.B., Rüttimann, R., Clemens, R., Oberste, M.S., Weldon, W.C., Ackerman, M.E., Connor, R.I., Wieland-Alter, W.F., et al. (2021). A Randomized Phase 4 Study of Immunogenicity and Safety After Monovalent Oral Type 2 Sabin Poliovirus Vaccine Challenge in Children Vaccinated with Inactivated Poliovirus Vaccine in Lithuania. J. Infect. Dis. *223*, 119–127. https://doi.org/10.1093/infdis/jiaa390.
32. Yu, H., Han, J., and Chang, K.-C. (2004). PEBL: Web page classification without negative examples. IEEE Trans. Knowl. Data Eng. *16*, 70–81.
33. Jiang, L., Li, D., Wang, Q., Wang, S., and Wang, S. (2020). Improving positive unlabeled learning: Practical aul estimation and new training method for extremely imbalanced data sets. Preprint at arXiv. https://doi.org/10.48550/arXiv.2004.09820.
34. Bekker, J., and Davis, J. (2018). Learning from Positive and Unlabeled Data under the Selected at Random Assumption (PMLR), pp. 8–22.
35. Bekker, J., Robberechts, P., and Davis, J. (2020). Beyond the Selected Completely at Random Assumption for Learning from Positive and Unlabeled Data (Springer), pp. 71–85.
36. De Block, S., and Bekker, J. (2022). Bagging Propensity Weighting: A Robust Method for Biased PU Learning (PMLR), pp. 23–37.
37. Dunning, A.J. (2006). A model for immunological correlates of protection. Stat. Med. *25*, 1485–1497. https://doi.org/10.1002/sim.2282.
38. Ojala, M., and Garriga, G.C. (2010). Permutation tests for studying classifier performance. J. Mach. Learn. Res. *11*.
39. Liu, B., Lee, W.S., Yu, P.S., and Li, X. (2002). Partially supervised classification of text documents, *485*, pp. 387–394.
40. Lotfnezhad Afshar, H., Jabbari, N., Khalkhali, H.R., and Esnaashari, O. (2021). Prediction of Breast Cancer Survival by Machine Learning Methods: An Application of Multiple Imputation. Iran. J. Public Health *50*, 598–605. https://doi.org/10.18502/ijph.v50i3.5606.
41. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
42. Kassambara, A., and Kassambara, M.A. (2020). Package 'ggpubr'. R Package version 0.1 6.
43. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.
44. Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Comput. Sci. Eng. *9*, 90–95.
45. Tretyakov, K. (2017). Matplotlib-Venn: Functions for Plotting Area-Proportional Two-And Three-Way Venn Diagrams in Matplotlib.
46. Waskom, M. (2021). Seaborn: statistical data visualization. J. Open Source Softw. *6*, 3021.
47. Lakhashe, S.K., Amacker, M., Hariraju, D., Vyas, H.K., Morrison, K.S., Weiner, J.A., Ackerman, M.E., Roy, V., Alter, G., Ferrari, G., et al. (2022). Cooperation Between Systemic and Mucosal Antibodies Induced by Virosomal Vaccines Targeting HIV-1 Env: Protection of Indian Rhesus Macaques Against Low-Dose Intravaginal SHIV Challenges. Front. Immunol. *13*, 788619.
48. Bekker, J., and Davis, J. (2020). Learning from positive and unlabeled data: a survey. Mach. Learn. *109*, 719–760. https://doi.org/10.1007/s10994-020-05877-5.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Python	https://www.python.org/downloads/	3.11.0
Jupyter Notebook	https://jupyter.org/install	6.5.2
Scikit-Learn	Pedragosa et al.[41]	1.1.3
Transudative Positive-Unlabeled Bootstrapped Aggregation (PU Bagging)	Mordelet et al.[28]	
ggpubr	Kassambara and Kassambara[42]	
Scipy	Virtanen et al.[43]	
matplotlib	Hunter[44]	
matplotlib-venn	Tretyakova[45]	
seaborn	Waskorn[46]	

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the corresponding author, Prof. Margaret E. Ackerman (margaret.e.ackerman@dartmouth.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Raw data used in this study has been previously published or is available by request from the corresponding authors of publications in which it was originally reported.[15,30,31,47]
- This paper does not report original code. The code adapted for this study is available upon request.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This study conducted analysis of previously reported pre-clinical and clinical vaccine studies.[15,30,31,47]

METHOD DETAILS

Datasets

Multiple high dimensional synthetic datasets (P>=N) with binary ground truth protection status labels were generated by "make_classification" class from "scikit-learn" library.[41] For each dataset, 300 samples with binary protection status class labels were generated with varying classification task difficulty by tuning the parameter "*class_sep*" to manipulate the spread of hypercube and the parameter "*weights*" to determine the proportion of samples assigned into each class. A set of 200 immunogenicity features were modeled. This dataset was composed of *30%* "informative" (parameter *"n_informative"*) features and *70%* "redundant" (parameter *"n_redundant"*) features, which were generated from random linear combination of the informative features to introduce covariance between features. Synthetic data were split into training (including Positive (P) set and Unlabeled (U) Set) and test sets at a ratio of 2:1 (Figure 1B). Immune response profiles from published preclinical and clinical vaccine efficacy trials were obtained (Figure 3A; Table S1).[15,30,31,47] For immunogenicity data, groups (Lakhashe, Ackerman, and Brickley et al.) or timepoints (Bandyopadhyay et al.) for which protection was demonstrated were assigned as the protected group for simulation purposes.

P/U Label assignment

The Selected Completely at Random (SCAR) strategy was adopted for both synthetic data and real-world immunological profiles with ground truth P/N Label to assign positive and unlabeled Labels.[48] Known Positive (KP) subjects were randomly sampled from Class:1 (TP) and

remaining TP and TN subjects were assigned as the unlabeled (U) class. To evaluate the reproducibility of modeling results, for each data set, P/U labels were assigned in this way 30 times.

Scoring and ranking unlabeled samples with PU learning

The Transudative Positive-Unlabeled Bootstrapped Aggregation (PU Bagging) Algorithm by Moderlet et al. was adapted to score unlabeled (U) subjects.[28] According to the number of KP in each dataset, a matched number of U subjects were randomly sampled with replacement (bootstrapped) from the U set and temporarily labeled as "negative". A classifier was built with all KP and bootstrapped "negative" (Bagged) samples. As underlying immunogenicity data was typically skewed, feature values were log transformed prior to modeling. To reduce the influence of features with high ranges on classification, the initial parameters for standardization were calculated from the "Bagged" samples; then the parameters were used to transform the "Bagged", "Out-of-Bag", and test set samples for the simulations with synthetic datasets. The classifier was then used to predict the out-of-bag samples (OOB) samples in U set as positive or negative. This process was repeated 100 times and the PU bagging score for each unlabeled sample was calculated by aggregating the prediction of it as an OOB from an ensemble of the 100 classifiers:

$$\text{PU Bagging Score (U set)} = \frac{\text{Sum up prediction as "OOB"}}{\text{Number of "OOB"}}$$

In the synthetic datasets, the classifiers of bagged samples were predicted on a held-out test set as an external validation to understand the generalizability of the PU Bagging classifier built with KP and targeted unlabeled (U) samples.[28] The PU Bagging Score for holdout/test set samples was calculated as following:

$$\text{PU Bagging Score (test set)} = \frac{\text{Sum up prediction from all bootstrapped classifier}}{\text{Number of Bootstrap}}$$

To compare the prediction performance under varied percentages of KP among all samples, a single biased RBF-SVM classifier was established with P/U labels and reweighting the samples by setting the parameter *"class weight"* as *"balanced"*. The "*predict_proba*" method was applied to score each unlabeled subject according to its class 1 probability. All machine learning classifiers were built with "Scikit-Learn" package (version 1.1.3).[41]

Permutation comparison

To investigate whether model classification performance exceeded that expected at random, ground truth label permutation was employed to define baseline performance.[38] In each permutation scenario, the ground truth label was first randomly shuffled, followed by P/U label assignment and U set sample scoring exactly as employed to classify U set samples under "actual" ground truth labels.

Supervised learning comparison

To investigate whether conventional supervised learning pipeline outperforms PU learning methods in classifying U set samples, an SVM classifier and a Random Forest classifier was trained for each simulated scenario by treating the U set as "negative". The best combination of hyperparameters was searched under 3-fold cross validation using *GridSearchCV* scoring by ROC-AUC. The ability of these models to classify protection status in held out U set samples in the context of three-fold cross-validation was reported. All machine learning classifiers were built with "Scikit-Learn" package (version 1.1.3).[41]

Binary classification of U set

Multiple strategies based on the score rank from PU bagging algorithm to determine decision boundary for binary classification among unlabeled samples were proposed and applied (Figure 4A).

- *Label: Percentiles* – With the knowledge of the proportion (*p*) of underlying TN among all samples, a threshold was employed to split ($p * N_U$) unlabeled samples as "predicted negative" according to the rank of PU Bagging Score.
- *Label: Cutoff (0.5)* – In the absence the knowledge in P/N proportion, an unlabeled sample was "negative" with a PU Bagging Score smaller than 0.5; otherwise, it was predicted as "positive"
- *Label: RN* – Reliable Negative (RN) group was initially defined by the samples of the lowest PU Bagging Score rank ($N_{RN} = N_{KP}$). The hyperparameter of the SVM model was trained with KP and RN using 3-fold stratified cross-validation. The model was used to predict the rest of the U into positive and negative.
- *Label: iterative* – An SVM classifier with default hyperparameters was initially trained with KP and RN. The RN was expanded by iteratively including newly predicted N from the rest of the unlabeled samples. The iteration was halted when no unlabeled sample was predicted as "negative", or the number of iterations reached 20.
- *Label: Vote* – The averaged prediction from *Label: Cutoff (0.5), Label: RN and Label: iterative.*

Evaluation metrics

Area under the receiver operating characteristic curve (ROC AUC) was computed between the continuous PU bagging score-based class labels and ground truth labels of the U set samples and holdout test set samples for simulations in the synthetic datasets. ROC-AUC was also used between the class:1 probability of U set samples from the supervised SVM classifier and ground truth labels for the evaluation of class assignments made for real-world immunogenicity data. The binary predictions were compared to the Ground Truth Label of all samples using F1 score, Recall, and Precision.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{F1} = 2 * \frac{\text{precision} + \text{recall}}{\text{precision} * \text{recall}}$$

Analysis and visualization

The R package "ggpubr" was employed for statistic output and visualization.[42] Mann Whitney U test was implemented to annotate significant features under different sets of labels. The Benjamini-Hochberg procedure was used to control the false discovery rate at a level of 0.05 for *p* value adjustment. Python packages "Scipy", "matplotlib-venn", "matplotlib", "seaborn" were utilized for statistical analysis and visualization.[43–46]