Supplemental Information: Multiplexed single-tier immunoassay for improved Lyme disease diagnosis across all disease stages

Supplemental Figure 1: Signal differences between healthy and control for IgG and IgM.

Supplemental Table 1: Binary classification cross validation results for 15 different models for Phase 1, 2, 3 datasets.

Supplemental Table 2. Characterization of the training cohort, with each antigen's predictive ability.

Supplemental Figure 2: Distribution of probability scores between healthy and disease for all tested models using the Phase 2b dataset.

Supplemental Figure 3: Youden's index was used to evaluate decision thresholds to maximize performance of the model. Figure shows the intersection of sensitivity and specificity.

Supplemental Figure 4: A feature reduction/feature ranking method was applied to the training data. Cross-validation results for a reduced set of features shown from 3 to 20 features.

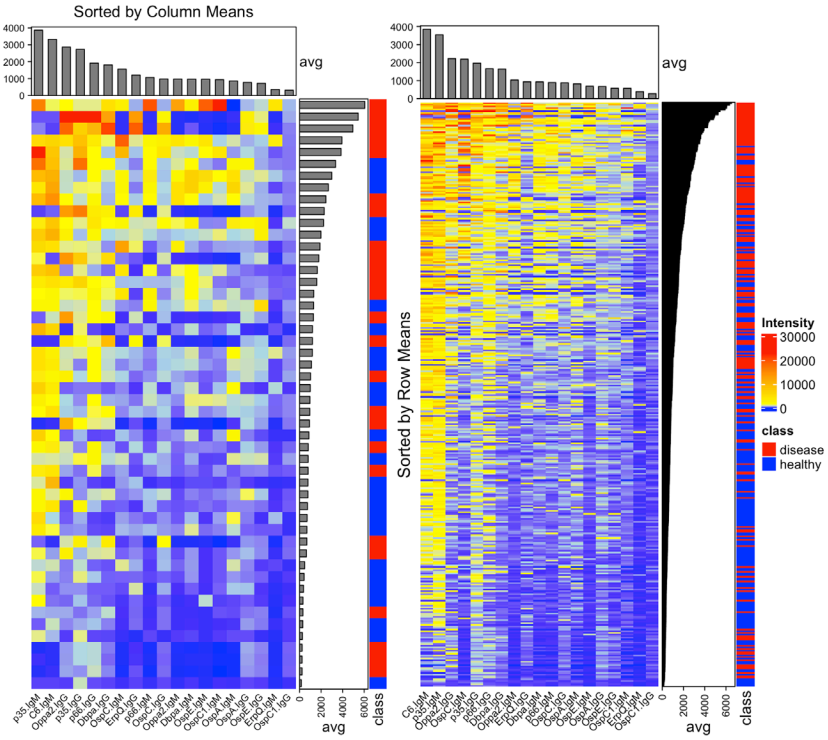
Supplemental Figure 5: Display of the probability scores for TP, TN, FP and FN cases for the Phase 1, 2 and 3 datasets.

Supplemental Figure 6: Calibration curves, AUC and boxplots highlighting the quality of the predictions, the specificity/sensitivity in a receiver-operator curve, a confusion matrix showing FP, FN, TP, TN counts, and boxplots showing the distribution of probabilities for TP, TN, FP and FN.

Supplemental Figure 7: SHAP feature importance bar-graph for the Phase 1, 2 and 3 datasets.

Binomial classifier selection method and description

While Luminex serological assays are straightforward, their multiplex nature and machine learning-based diagnosis require detailed explanation. The Supplemental Materials include raw data, classification results, decision thresholds, feature reduction methods, and ROC plots. To address potential overtraining and over-interpretation, we present multiple classifiers, cross-validation approaches, threshold determinations, and confidence intervals for all results.



Supplemental Figure 1: Heatmaps show signal intensity for patients (Y-axis) versus antigens (X-axis) in Phase 1 (left, N=100 discovery cohort) and Phase 2b (right, N=364 training cohort). Color indicates disease (red) or healthy (blue) status. Reactivity patterns are heterogeneous; not all disease patients show high reactivity across antigens, nor do all healthy controls show uniformly low reactivity.

Supplemental Table 1: Complete classifier performance across validation phases. All 15 classifiers shown; XGBoost (main text) contextualized by alternative methods. Blue = cross-validation; green = test results. All analyses use locked 0.377 threshold. Table A (Phase 1, N=100): Cross-validation with 5-fold, 10-fold, and LOOCV. Table B (Phase 2a, N=264): Cross-validation plus Phase 1 holdout test (N=100). Table C (Phase 3): Blinded CDMOtest (Phase 3a, green) and cross-validation (Phase 3b, blue, N=259). Phase 3b LOOCV AUC (0.98) matches Phase 2b (N=364). Confidence intervals: DeLong's method for AUC; actual 95% CI from folds (5-fold/10-fold); Wilson's CI for LOOCV [Wilson 1927; Agresti 1998].

A

		Phase 1 - Discovery cohort model testing (N=50 control, 50 case)														
		C5.0	gbm	glmnet	knn	lda	mars	nb	nn	polyda	qda	rf	ripper	svm	svmLin	XGB
5 fold	AUC	0.94 (0.89-0.98)	0.97 (0.94-1.00)	0.92 (0.86-0.98)	0.84 (0.76-0.92)	0.86 (0.78-0.94)	0.95 (0.92-0.99)	0.88 (0.81-0.94)	0.94 (0.89-0.99)	0.85 (0.78-0.93)	0.87 (0.81-0.94)	0.96 (0.93-0.99)	0.89 (0.83-0.95)	0.89 (0.83-0.96)	0.90 (0.83-0.96)	**0.95 (0.91-0.99)**
	sensitivity	0.98 (0.90-1.00)	0.90 (0.79-0.96)	0.86 (0.77-0.93)	0.72 (0.58-0.83)	0.78 (0.65-0.87)	0.94 (0.87-0.98)	0.76 (0.63-0.86)	0.90 (0.79-0.96)	0.72 (0.58-0.83)	0.90 (0.79-0.96)	0.96 (0.90-1.00)	0.84 (0.72-0.92)	0.90 (0.79-0.96)	0.86 (0.74-0.93)	**0.90 (0.79-0.96)**
	specificity	0.56 (0.42-0.69)	0.86 (0.74-0.93)	0.86 (0.74-0.93)	0.88 (0.76-0.94)	0.88 (0.76-0.94)	0.88 (0.76-0.94)	0.78 (0.65-0.87)	0.84 (0.72-0.92)	0.84 (0.72-0.92)	0.72 (0.58-0.83)	0.70 (0.56-0.81)	0.70 (0.56-0.81)	0.64 (0.50-0.76)	0.70 (0.56-0.81)	**0.76 (0.63-0.86)**
10-fold	AUC	0.93 (0.89-0.98)	0.97 (0.94-1.00)	0.92 (0.86-0.98)	0.84 (0.76-0.92)	0.85 (0.77-0.93)	0.97 (0.95-1.00)	0.87 (0.81-0.94)	0.95 (0.90-0.99)	0.87 (0.80-0.94)	0.88 (0.82-0.95)	0.97 (0.94-1.00)	0.90 (0.83-0.96)	0.92 (0.87-0.97)	0.90 (0.84-0.97)	**0.97 (0.94-0.99)**
	sensitivity	0.92 (0.81-0.97)	0.92 (0.81-0.97)	0.88 (0.76-0.94)	0.78 (0.66-0.87)	0.68 (0.54-0.79)	0.90 (0.79-0.96)	0.76 (0.63-0.85)	0.90 (0.79-0.96)	0.74 (0.60-0.84)	0.88 (0.76-0.94)	0.98 (0.90-1.00)	0.88 (0.76-0.94)	0.96 (0.87-1.00)	0.88 (0.76-0.94)	**0.94 (0.84-0.98)**
	specificity	0.78 (0.65-0.87)	0.88 (0.76-0.94)	0.86 (0.74-0.93)	0.90 (0.79-0.96)	0.92 (0.81-0.97)	0.80 (0.67-0.89)	0.90 (0.79-0.96)	0.88 (0.76-0.94)	0.84 (0.72-0.92)	0.72 (0.58-0.83)	0.70 (0.56-0.81)	0.64 (0.50-0.76)	0.84 (0.72-0.92)	**0.84 (0.72-0.92)**	
LOOCV	AUC	0.93 (0.89-0.98)	0.97 (0.94-1.0)	0.92 (0.86-0.98)	0.83 (0.75-0.92)	0.86 (0.77-0.94)	0.95 (0.90-0.99)	0.87 (0.81-0.94)	0.94 (0.89-0.99)	0.87 (0.79-0.94)	0.87 (0.80-0.94)	0.96 (0.93-0.99)	0.83 (0.80-0.94)	0.92 (0.87-0.97)	0.91 (0.85-0.94)	**0.96 (0.91-0.97)**
	sensitivity	0.94 (0.84-0.98)	0.90 (0.79-0.96)	0.88 (0.76-0.92)	0.70 (0.56-0.81)	0.72 (0.61-0.97)	0.86 (0.74-0.93)	0.76 (0.63-0.86)	0.90 (0.79-0.96)	0.78 (0.65-0.87)	0.82 (0.69-0.90)	0.94 (0.84-0.98)	0.88 (0.76-0.94)	0.94 (0.84-0.98)	0.86 (0.74-0.93)	**0.92 (0.81-0.97)**
	specificity	0.76 (0.63-0.86)	0.84 (0.72-0.92)	0.84 (0.72-0.92)	0.88 (0.76-0.94)	0.92 (0.81-0.97)	0.86 (0.74-0.93)	0.78 (0.64-0.87)	0.88 (0.76-0.94)	0.86 (0.74-0.93)	0.74 (0.60-0.84)	0.72 (0.58-0.83)	0.73 (0.58-0.83)	0.68 (0.54-0.79)	0.84 (0.72-0.92)	**0.86 (0.74-0.93)**

B

		Phase 2a - Holdout train-test (N=264 [147 control, 117 case])														
		C5.0	gbm	glmnet	knn	lda	mars	nb	nn	polyda	qda	rf	ripper	svm	svmLin	XGB
5-fold	AUC	0.95 (0.93-0.98)	0.97 (0.95-0.99)	0.93 (0.90-0.96)	0.90 (0.86-0.93)	0.89 (0.84-0.92)	0.96 (0.93-0.98)	0.87 (0.82-0.91)	0.95 (0.92-0.98)	0.91 (0.87-0.94)	0.82 (0.76-0.87)	0.97 (0.95-0.99)	0.96 (0.93-0.98)	0.91 (0.87-0.94)	0.92 (0.89-0.96)	**0.97 (0.95-0.98)**
	sensitivity	0.93 (0.87-0.97)	0.92 (0.85-0.95)	0.92 (0.85-0.95)	0.85 (0.77-0.90)	0.78 (0.69-0.84)	0.93 (0.87-0.97)	0.78 (0.64-0.87)	0.89 (0.82-0.93)	0.86 (0.79-0.91)	0.64 (0.55-0.72)	0.97 (0.92-0.99)	0.88 (0.81-0.93)	0.92 (0.86-0.96)	0.92 (0.85-0.95)	**0.92 (0.85-0.95)**
	specificity	0.80 (0.72-0.85)	0.90 (0.84-0.94)	0.83 (0.76-0.88)	0.71 (0.64-0.78)	0.83 (0.76-0.88)	0.90 (0.84-0.94)	0.86 (0.57-0.74)	0.88 (0.82-0.92)	0.82 (0.75-0.87)	0.88 (0.82-0.92)	0.81 (0.74-0.87)	0.91 (0.85-0.94)	0.71 (0.65-0.78)	0.78 (0.71-0.84)	**0.88 (0.82-0.92)**
10 fold	AUC	0.96 (0.94-0.98)	0.97 (0.96-0.99)	0.94 (0.91-0.97)	0.90 (0.87-0.94)	0.90 (0.86-0.94)	0.96 (0.94-0.98)	0.87 (0.83-0.91)	0.95 (0.93-0.98)	0.91 (0.87-0.94)	0.82 (0.77-0.87)	0.97 (0.95-0.99)	0.94 (0.91-0.97)	0.92 (0.89-0.95)	0.93 (0.89-0.96)	**0.97 (0.95-0.99)**
	sensitivity	0.93 (0.89-0.98)	0.92 (0.85-0.95)	0.97 (0.93-0.97)	0.86 (0.79-0.91)	0.79 (0.70-0.85)	0.90 (0.83-0.94)	0.67 (0.58-0.75)	0.91 (0.84-0.95)	0.84 (0.76-0.89)	0.62 (0.53-0.71)	0.96 (0.90-0.98)	0.89 (0.82-0.93)	0.94 (0.88-0.97)	0.92 (0.86-0.96)	**0.92 (0.86-0.96)**
	specificity	0.82 (0.73-0.88)	0.91 (0.86-0.95)	0.82 (0.75-0.88)	0.71 (0.63-0.78)	0.88 (0.82-0.93)	0.96 (0.73-0.91)	0.84 (0.77-0.89)	0.88 (0.82-0.91)	0.80 (0.73-0.85)	0.87 (0.81-0.92)	0.80 (0.73-0.86)	0.86 (0.79-0.91)	0.71 (0.64-0.78)	0.84 (0.77-0.89)	**0.85 (0.78-0.90)**
LOOCV	AUC	0.96 (0.94-0.98)	0.97 (0.96-0.99)	0.94 (0.90-0.97)	0.90 (0.86-0.93)	0.90 (0.86-0.94)	0.92 (0.86-0.96)	0.87 (0.83-0.91)	0.95 (0.92-0.98)	0.90 (0.86-0.94)	0.82 (0.77-0.87)	0.97 (0.95-0.99)	0.94 (0.92-0.97)	0.92 (0.90-0.96)	0.93 (0.90-0.97)	**0.97 (0.95-0.99)**
	sensitivity	0.93 (0.87-0.97)	0.91 (0.84-0.95)	0.92 (0.85-0.95)	0.87 (0.80-0.92)	0.81 (0.73-0.87)	0.92 (0.86-0.96)	0.66 (0.57-0.74)	0.90 (0.83-0.94)	0.83 (0.75-0.89)	0.67 (0.58-0.76)	0.96 (0.90-0.98)	0.87 (0.80-0.92)	0.93 (0.87-0.97)	0.92 (0.86-0.96)	**0.92 (0.85-0.95)**
	specificity	0.85 (0.78-0.90)	0.91 (0.87-0.94)	0.82 (0.75-0.88)	0.71 (0.63-0.78)	0.84 (0.78-0.89)	0.88 (0.80-0.91)	0.83 (0.76-0.88)	0.87 (0.81-0.92)	0.82 (0.75-0.87)	0.88 (0.82-0.92)	0.82 (0.75-0.88)	0.84 (0.77-0.89)	0.74 (0.67-0.81)	0.81 (0.74-0.87)	**0.88 (0.82-0.92)**
holdout test	AUC	0.99 (0.98-1.00)	0.99 (0.99-1.00)	0.93 (0.81-0.97)	0.91 (0.85-0.97)	0.84 (0.75-0.93)	0.99 (0.97-1.00)	0.88 (0.81-0.95)	0.95 (0.90-1.00)	0.87 (0.79-0.95)	0.92 (0.87-0.97)	0.99 (0.99-1.00)	0.96 (~.83-0.96)	0.99 (0.98-1.00)	0.96 (0.91-1.00)	**1.00 (0.99-1.00)**
	sensitivity	0.98 (0.90-1.00)	0.96 (0.87-1.00)	0.92 (0.81-0.97)	0.80 (0.67-0.89)	0.72 (0.58-0.83)	0.96 (0.87-0.99)	0.82 (0.69-0.90)	0.92 (0.81-0.97)	0.82 (0.69-0.90)	0.78 (0.65-0.87)	1.00 (0.93-1.00)	0.88 (0.76-0.94)	0.98 (0.90-1.00)	0.94 (0.84-0.98)	**1.00 (0.93-1.00)**
	specificity	0.88 (0.76-0.94)	0.94 (0.84-0.98)	0.88 (0.76-0.94)	0.84 (0.72-0.92)	0.80 (0.67-0.89)	0.88 (0.76-0.94)	0.78 (0.65-0.87)	0.90 (0.79-0.96)	0.84 (0.72-0.92)	0.96 (0.87-1.00)	0.76 (0.63-0.86)	0.88 (0.76-0.94)	0.86 (0.74-0.93)	0.88 (0.76-0.94)	**0.94 (0.84-0.98)**

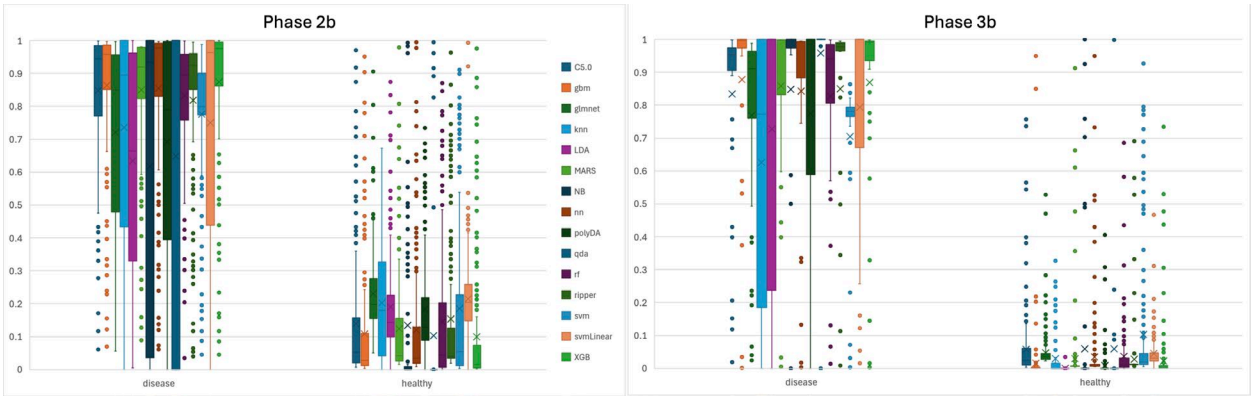
C

		Phase 3 - Tech transfer train-test (train: N=364, test: N=269)														
		C5.0	gbm	glmnet	knn	lda	mars	nb	nn	polyda	qda	rf	ripper	svm	svmLin	XGB
5 fold	AUC	0.99 (0.97-1.00)	0.98 (0.95-1.00)	0.97 (0.92-1.00)	0.92 (0.87-0.97)	0.93 (0.86-0.99)	0.98 (0.95-1.00)	0.98 (0.96-1.00)	0.95 (0.90-1.00)	0.95 (0.90-1.00)	0.96 (0.92-0.99)	0.99 (0.98-1.00)	0.97 (0.93-1.00)	0.92 (0.86-0.98)	0.95 (0.90-1.00)	**0.98 (0.96-1.00)**
	sensitivity	0.98 (0.89-1.00)	0.92 (0.81-0.97)	0.94 (0.84-0.98)	0.86 (0.73-0.93)	0.74 (0.60-0.84)	0.94 (0.84-0.98)	0.91 (0.84-0.98)	0.90 (0.79-0.96)	0.84 (0.71-0.92)	0.96 (0.86-1.00)	0.96 (0.86-0.99)	0.94 (0.84-0.98)	0.94 (0.84-0.98)	0.90 (0.78-0.96)	**0.94 (0.84-0.98)**
	specificity	0.81 (0.75-0.86)	0.99 (0.97-1.00)	0.94 (0.90-0.97)	0.86 (0.81-0.90)	1.00 (0.92-1.00)	0.97 (0.93-0.98)	0.89 (0.84-0.92)	0.91 (0.86-0.94)	0.99 (0.96-1.00)	0.93 (0.89-0.96)	0.91 (0.86-0.94)	0.96 (0.91-0.98)	0.82 (0.76-0.87)	0.94 (0.90-0.97)	**0.96 (0.92-0.98)**
10 fold	AUC	0.99 (0.97-1.00)	0.98 (0.95-1.00)	0.97 (0.93-1.00)	0.95 (0.91-0.99)	0.96 (0.92-1.00)	0.99 (0.99-1.00)	0.97 (0.95-1.00)	0.95 (0.90-1.00)	0.97 (0.93-1.00)	0.96 (0.93-0.99)	0.99 (0.97-1.00)	0.95 (0.91-1.00)	0.93 (0.87-0.98)	0.95 (0.90-1.00)	**0.98 (0.95-1.00)**
	sensitivity	0.98 (0.89-1.00)	0.92 (0.81-0.97)	0.94 (0.84-0.98)	0.88 (0.76-0.94)	0.80 (0.66-0.89)	0.96 (0.86-0.99)	0.92 (0.81-0.97)	0.92 (0.81-0.97)	0.86 (0.73-0.93)	0.96 (0.86-0.99)	0.96 (0.86-0.99)	0.92 (0.81-0.97)	0.94 (0.84-0.98)	0.92 (0.81-0.97)	**0.94 (0.84-0.98)**
	specificity	0.82 (0.77-0.87)	0.98 (0.95-0.99)	0.94 (0.90-0.97)	0.85 (0.79-0.89)	1.00 (0.98-1.00)	0.96 (0.92-0.98)	0.88 (0.83-0.91)	0.92 (0.88-0.95)	0.98 (0.95-0.99)	0.94 (0.90-0.96)	0.91 (0.86-0.94)	0.96 (0.93-0.98)	0.81 (0.76-0.86)	0.93 (0.89-0.96)	**0.97 (0.93-0.98)**
LOOCV	AUC	0.98 (0.96-1.00)	0.98 (0.95-1.00)	0.97 (0.93-1.00)	0.94 (0.90-0.98)	0.95 (0.90-1.00)	0.97 (0.93-1.00)	0.97 (0.95-1.00)	0.95 (0.89-1.00)	0.97 (0.92-1.00)	0.95 (0.92-0.99)	0.99 (0.97-1.00)	0.96 (0.92-1.00)	0.91 (0.85-0.97)	0.96 (0.91-1.00)	**0.98 (0.96-1.00)**
	sensitivity	0.98 (0.89-1.00)	0.92 (0.81-0.97)	0.94 (0.84-0.98)	0.89 (0.78-0.94)	0.80 (0.66-0.89)	0.94 (0.84-0.98)	0.92 (0.81-0.97)	0.92 (0.81-0.97)	0.88 (0.76-0.94)	0.96 (0.86-0.99)	0.96 (0.86-0.99)	0.92 (0.81-0.97)	0.92 (0.81-0.97)	0.92 (0.81-0.97)	**0.94 (0.84-0.98)**
	specificity	0.96 (0.79-0.99)	0.97 (0.94-0.99)	0.95 (0.91-0.97)	0.85 (0.79-0.89)	1.00 (0.98-1.00)	0.96 (0.93-0.98)	0.89 (0.84-0.93)	0.92 (0.88-0.95)	0.98 (0.95-0.99)	0.93 (0.89-0.96)	0.92 (0.87-0.95)	0.96 (0.93-0.98)	0.82 (0.76-0.87)	0.94 (0.90-0.96)	**0.96 (0.92-0.98)**
technology transfer test	AUC	0.99 (0.98-1.00)	0.99 (0.99-1.00)	0.93 (0.81-0.97)	0.91 (0.85-0.97)	0.84 (0.75-0.93)	0.99 (0.97-1.00)	0.88 (0.81-0.95)	0.95 (0.90-1.00)	0.87 (0.79-0.95)	0.92 (0.87-0.97)	0.99 (0.99-1.00)	0.96 (~.83-0.96)	0.99 (0.98-1.00)	0.96 (0.91-1.00)	**1.00 (0.99-1.00)**
	sensitivity	0.98 (0.90-1.00)	0.96 (0.87-1.00)	0.92 (0.81-0.97)	0.80 (0.67-0.89)	0.72 (0.58-0.83)	0.96 (0.87-0.99)	0.82 (0.69-0.90)	0.92 (0.81-0.97)	0.82 (0.69-0.90)	0.78 (0.65-0.87)	1.00 (0.93-1.00)	0.88 (0.76-0.94)	0.98 (0.90-1.00)	0.94 (0.84-0.98)	**1.00 (0.93-1.00)**
	specificity	0.88 (0.76-0.94)	0.94 (0.84-0.98)	0.88 (0.76-0.94)	0.84 (0.72-0.92)	0.80 (0.67-0.89)	0.88 (0.76-0.94)	0.78 (0.65-0.87)	0.90 (0.79-0.96)	0.84 (0.72-0.92)	0.96 (0.87-1.00)	0.76 (0.63-0.86)	0.88 (0.76-0.94)	0.86 (0.74-0.93)	0.88 (0.76-0.94)	**0.94 (0.84-0.98)**

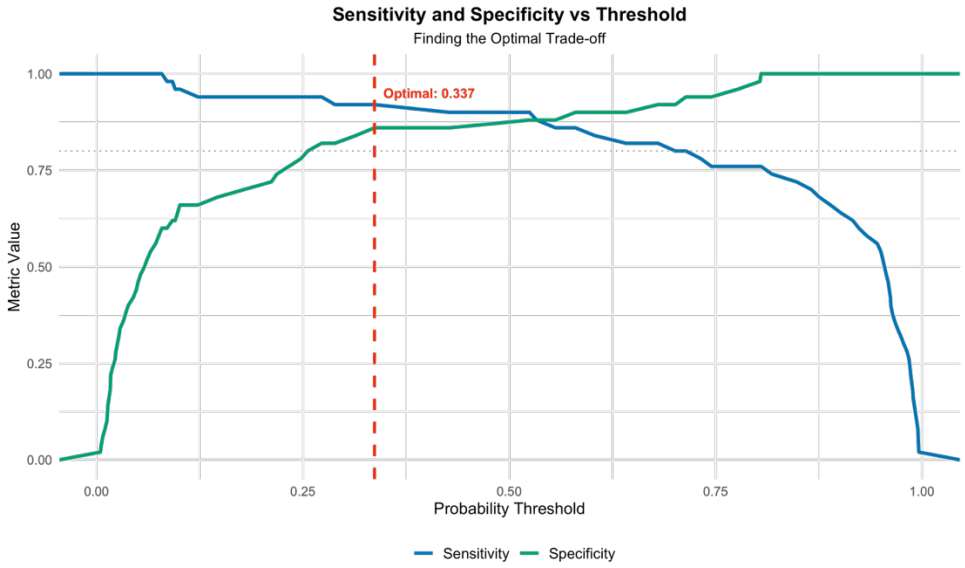
Supplemental Table 2: Characterization of the 364-person training cohort. Each antigen tested independently for predictive ability. Mean signals and standard deviations show relative differences between healthy and disease. Youden's threshold optimizes sensitivity/specificity; Youden's J measures stand-alone diagnostic performance. Raw and Bonferroni-corrected p-values test statistical significance; Cohen's d quantifies effect size. Ensemble rank averages informativeness across 5 classifiers (AUC-optimized). XGB rank uses XGBoost alone. SHAP (Shapley Additive exPlanation) quantifies each antigen's contribution to correct predictions. Rankings show significant overlap across XGB, ensemble, and SHAP methods.

Phase 2b LOOCV	AUC	healthy mean	healthy SD	disease mean	disease SD	Youden's threshold	Youden's J	sensitivity	specificity	p-value	Bonferroni	Cohen's d	ensembl e rank	XGB rank	SHAP rank
C6 IgG	0.81	897.27	2108.44	3562.31	4799.89	268.90	0.55	0.99	0.55	1.00E-11	2.01E-10	0.74	6	5	5
C6 IgM	0.48	3661.56	3960.17	4072.77	4594.49	3898.05	0.11	0.42	0.69	3.60E-01	1.00E+00	0.10	19	16	12
Dbpa IgG	0.82	466.52	448.99	3020.65	4622.59	299.40	0.47	0.98	0.49	1.19E-13	2.39E-12	0.81	7	9	8
Dbpa IgM	0.53	764.20	997.51	1130.51	1491.42	881.15	0.20	0.43	0.77	5.56E-03	1.11E-01	0.29	20	20	20
Erpq IgG	0.62	426.98	354.67	1540.58	3473.49	916.80	0.21	0.28	0.94	1.03E-05	2.06E-04	0.47	11	13	16
Erpq IgM	0.53	225.26	332.86	582.75	1309.31	214.40	0.17	0.43	0.74	2.54E-04	5.08E-03	0.39	16	10	9
Oppa2 IgG	0.87	591.70	751.60	4154.60	5303.28	764.50	0.56	0.78	0.78	1.15E-18	2.30E-17	0.98	1	2	3
Oppa2 IgM	0.56	461.62	574.38	1709.96	2742.45	910.75	0.27	0.39	0.88	1.29E-09	2.57E-08	0.66	12	14	15
OspA IgG	0.58	643.93	640.03	718.80	672.09	359.25	0.25	0.87	0.39	2.78E-01	1.00E+00	0.11	4	3	2
OspA IgM	0.62	934.43	1050.44	689.92	843.83	189.28	0.00	0.35	0.91	1.61E-02	3.22E-01	0.25	5	7	7
OspC IgG	0.76	453.37	343.05	1385.46	2161.24	530.75	0.37	0.66	0.71	5.77E-09	1.15E-07	0.63	10	6	4
OspC IgM	0.83	1342.65	1748.79	10864.23	10178.82	1281.25	0.33	0.82	0.72	1.57E-13	3.13E-12	0.81	2	4	6
OspC1 IgG	0.75	220.80	184.75	353.14	183.58	181.25	0.62	0.99	0.63	3.60E-11	7.20E-10	0.72	3	1	1
OspC1 IgM	0.50	415.66	351.31	752.47	2175.67	598.25	0.20	0.32	0.87	3.30E-02	6.59E-01	0.23	8	8	10
OspE IgG	0.73	442.93	596.44	728.97	979.82	259.25	0.51	0.97	0.54	7.02E-04	1.40E-02	0.36	9	12	11
OspE IgM	0.54	406.55	596.98	1027.32	2250.05	587.15	0.22	0.38	0.61	2.29E-04	4.59E-03	0.39	14	18	18
p35 IgG	0.65	1334.81	2133.15	2713.52	3603.19	1310.75	0.26	0.46	0.80	8.47E-06	1.69E-04	0.48	17	17	19

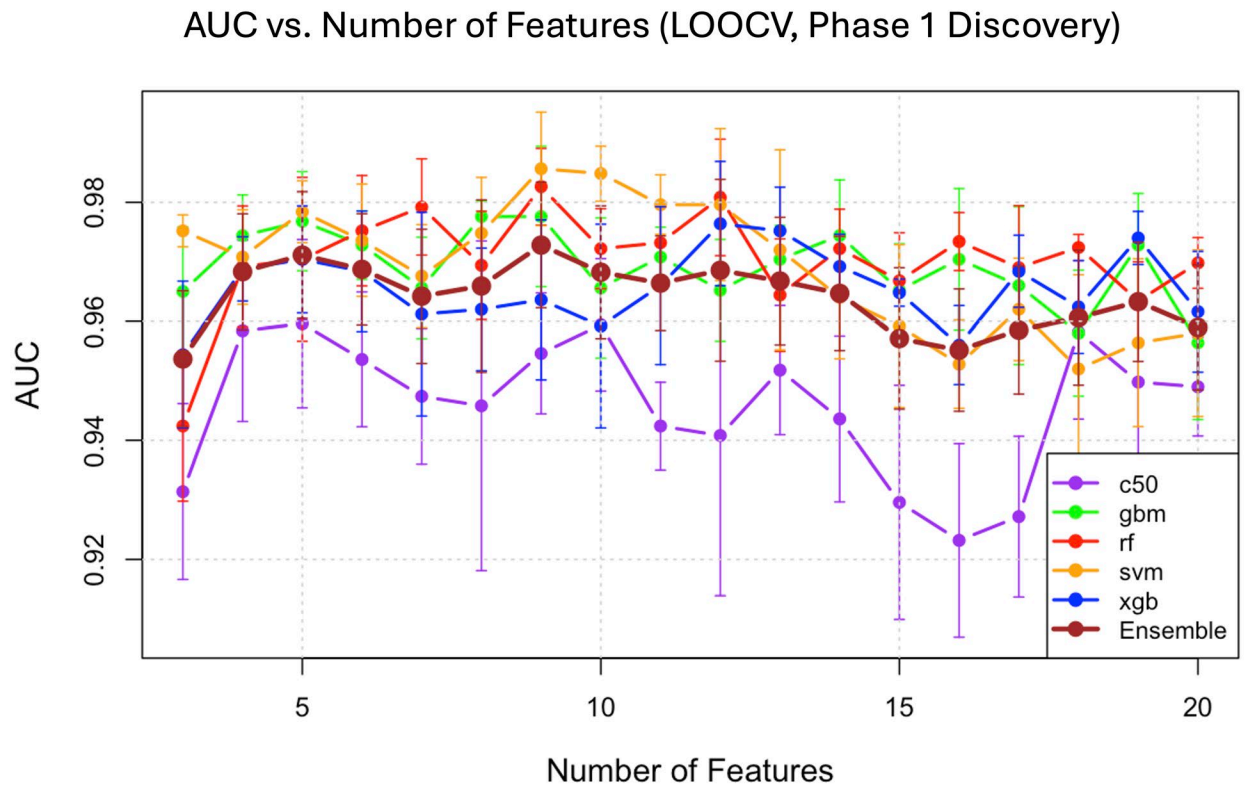
p35 IgM	0.53	3321.11	2990.83	3819.11	4529.86	715.75	0.08	0.29	0.93	2.11E-01	1.00E+00	0.13	18	15	14
p66 IgG	0.82	132.90	586.12	962.46	1547.99	134.75	0.48	0.84	0.86	6.87E-08	1.37E-06	0.58	15	19	17
p66 IgM	0.52	652.51	818.79	1180.41	2261.12	891.50	0.18	0.38	0.81	2.46E-03	4.91E-02	0.32	13	11	13



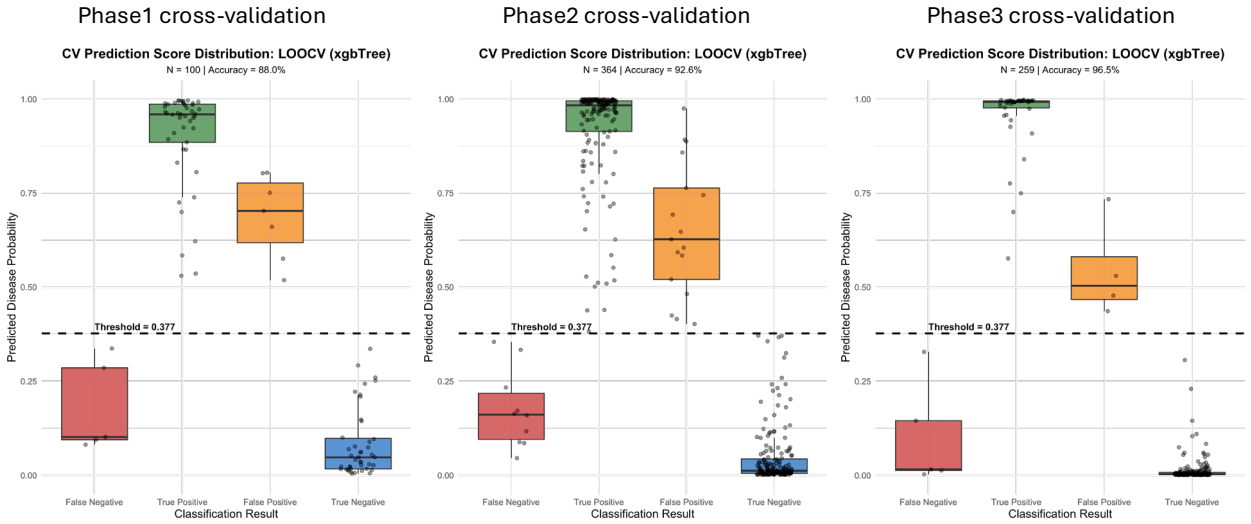
Supplemental Figure 2: Probability scores from 15 classifiers for Phase 2b, left and Phase 3b, right. The boxplots represent the probability scores for each of the models. A perfect classifier result would score close to 1 or 0 for disease and healthy, respectively. The t-test between the scores for disease and healthy for XGB (best model) was $p=2.8x10^{-120}$ (Phase 2b) and for LDA (worst model) $p=3.1x10^{-51}$. For the Phase 2a data the XGB p-value between disease and healthy calls from the probability scores is $p=2.2x10^{-107}$ and $p=1.36x10^{-70}$ for LDA.



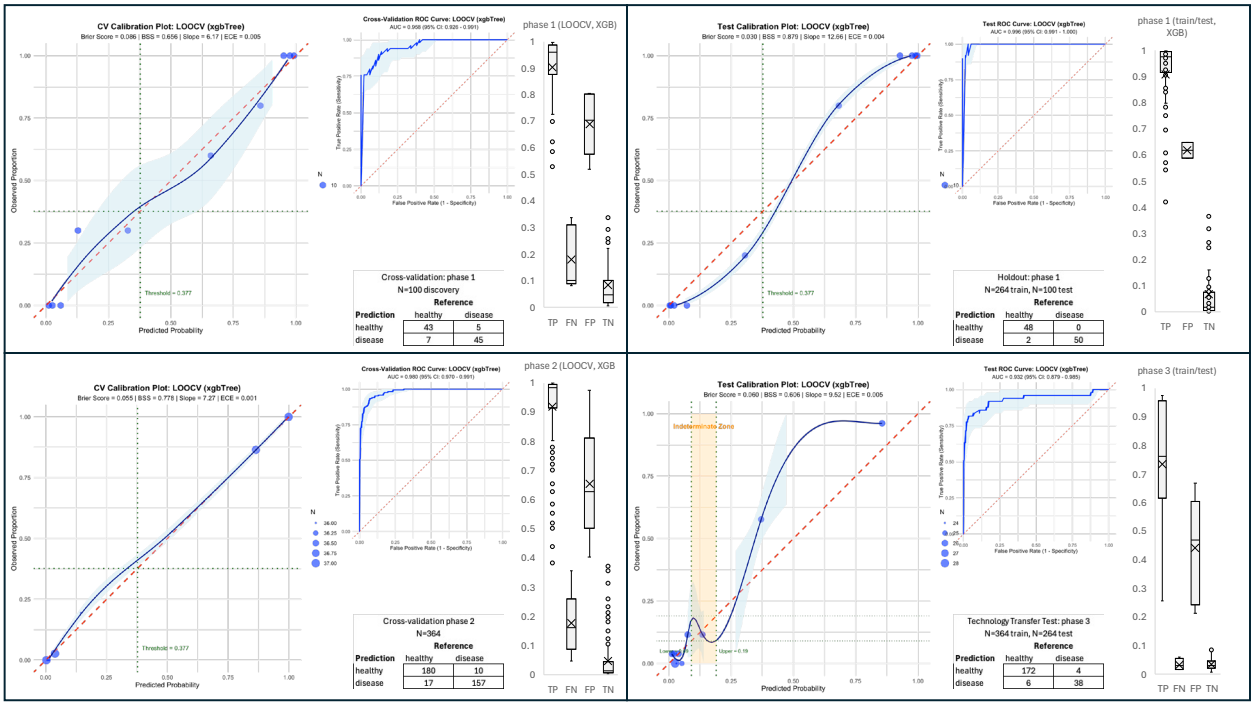
Supplemental Figure 3: Youden's index was computed for the Phase 1 cohort. The optimal performance for both sensitivity and specificity is computed using Youden's index (sensitivity-specificity-1) to allow for the best decision threshold for the discovery cohort data. Once the 0.377 index was computed, it was locked for all further analyses.



Supplemental Figure 4: Feature selection. Five classifiers (C5.0, gbm, rf, svm, xgb) optimized AUC using 3-20 features with LOOCV. Line graphs show individual models plus the average denoted 'Ensemble'. Phase 1 data (N=100) reveals no clear set of features <20 that consistently improves the cross-validation AUC. High-ranked features from this method include OspC1-IgG, OppA2-IgG and OspA-IgG.



Supplemental Figure 5: Phase 1 (N=100), phase 2 (N=364) and Phase 3 (N=259) data were cross validated using the locked 0.377 decision threshold and the locked model, XGBoost. The relationship between the probability scores for TP, TN, FP and FN are a graphical representation of the quality of the scores. Probability of disease should lie closer to 0 for healthy and 1 for disease; values closer to the decision threshold of 0.377 represent lower confidence. The boxplots represent the 75th and 25th percentiles, the end of the whiskers extend to the 95th and 5th percentiles, and the circles represent outliers.



Supplemental Figure 6: Calibration and discrimination plots for four validation phases. Upper left—calibration plot with 10 bins (perfect = 45° diagonal). Vertical green line marks 0.377 threshold. Dot size indicates sample count; position shows

predicted vs. observed probabilities. Metrics: Brier score (0=perfect), BSS (1=perfect), ECE (lower=better). Upper right—ROC curve with AUC and 95% CI (DeLong's method, blue bands). Lower plots—confusion matrix (left) and probability distributions for TP/TN/FP/FN (right).

Supplemental Figure 7: SHAP ranking of features for Phase 1, Phase 2(b) and Phase 3(b) cross validation studies. SHAP provides a contribution per feature to the decision-making process of the XGBoost model, here we average the values per prediction to arrive at a cumulative information value and rank for all 20 features.

Supplemental Methods:

Binomial classifier methods: We selected 15 models representing diverse classification approaches. Tree-based methods: C5.0 (boosting/pruning), Random Forest (bootstrapped ensemble), XGBoost (gradient boosting framework), and Gradient Boosted Machines (sequential tree building). Linear/Generalized models: GLMNet (regularized GLM), Linear Discriminant Analysis (maximizes between-class separation), Quadratic Discriminant Analysis (class-specific covariance), and Polynomial Discriminant Analysis (nonlinear extensions). Distance-based: K-Nearest Neighbors. Probabilistic: Naïve Bayes (assumes feature independence). Neural/Flexible: Neural Networks (multilayer perceptrons) and Multivariate Adaptive Regression Splines

(piecewise splines). Rules-based: RIPPER (if-then rules). Kernel-based: Support Vector Machines (nonlinear kernels) and SVM-Linear.

All algorithms were implemented in R (v4.4.3) using caret (v6.0-94) with seed=200 for reproducibility. To assess generalization, each algorithm underwent 5-fold CV, 10-fold CV, and LOOCV, comparing AUC, sensitivity, and specificity. Confidence intervals (95%) were calculated for fold-based methods. A python3 program was written to ensure commercial viability, using XGBClassifier, LeaveOneOut, StratifiedKFold, cross_val_predict, confusion_matrix, roc_auc_score and classification_report from sklearn. Predictions were identical to the R version, but decision probabilities differed, no case differed by >0.6%. This may be due to the difference in usage of the seed value.

Algorithm Selection

Tree-based ensemble methods demonstrated consistently superior performance across all three cross-validation schemes. For the discovery Phase 1 cohort, XGBoost achieved the highest average AUC, outperforming linear methods but comparable to Random Forest but with superior computational efficiency. Based on this systematic comparison, XGBoost was selected as the classification algorithm for subsequent model development and validations. No algorithm modifications were permitted after this selection. The hyperparameters were locked with estimators=100, learning rate=0.1, max depth=3, gamma=0, min child weight=1, subsample=0.8.

Decision Threshold Optimization Using Youden's Index

The optimal probability decision threshold was determined using Youden's index (J statistic), which maximizes the sum of sensitivity and specificity (J = Sensitivity + Specificity - 1). This

approach identifies the threshold that optimizes overall diagnostic accuracy in balanced populations without favoring either sensitivity or specificity.

The threshold optimization procedure was:

1) Generate cross-validation predictions: 10-fold cross-validation was performed on the discovery cohort, generating out-of-fold probability predictions for all 100 samples.
2) ROC curve construction: A receiver operating characteristic curve was constructed using the pROC package, evaluating all possible probability thresholds
3) Youden's index calculation: For each threshold, Youden's J statistic was calculated as J = (Sensitivity + Specificity - 1)
4) Optimal threshold selection: The threshold maximizing J was identified using the coords() function with ret="youden"

This analysis identified an optimal probability threshold of 0.377 with corresponding sensitivity of 0.92, specificity of 0.76 and accuracy of 0.84. This threshold was locked prior to further analysis.

Rationale for Youden's Index

Youden's index was selected to optimize thresholds because it balances sensitivity and specificity without requiring pre-specified misclassification costs [33], making it appropriate for moderate-to-high prevalence diagnostic scenarios and facilitating comparison with published literature. Alternative thresholds may suit different clinical contexts (screening vs. confirmation); complete ROC curves enable threshold adjustment for intended use. Supplemental Figures 3-4

show sensitivity/specificity curves, decision thresholds, and AUC plots for all phases including cross-validation.

Feature selection:

Feature selection was performed using recursive feature elimination, LASSO regularization, Boruta analysis, and ensemble ranking across classifiers. Ensemble AUC optimization provided the most informative feature ranking. However, full versus reduced feature panels showed minimal, non-significant AUC differences (Supplemental Figure 3). We therefore retained all 10 antigens with both IgG and IgM measurements.