RESEARCH ARTICLE SUMMARY

IMMUNOLOGY

Disease diagnostics using machine learning of B cell and T cell receptor sequences

Maxim E. Zaslavsky†, Erin Craig†, Jackson K. Michuda, Nidhi Sehgal, Nikhil Ram-Mohan, Ji-Yeun Lee, Khoa D. Nguyen, Ramona A. Hoh, Tho D. Pham, Katharina Röltgen, Brandon Lam, Ella S. Parsons, Susan R. Macwana, Wade DeJager, Elizabeth M. Drapeau, Krishna M. Roskin, Charlotte Cunningham-Rundles, M. Anthony Moody, Barton F. Haynes, Jason D. Goldman, James R. Heath, R. Sharon Chinthrajah, Kari C. Nadeau, Benjamin A. Pinsky, Catherine A. Blish, Scott E. Hensley, Kent Jensen, Everett Meyer, Imelda Balboni, Paul J. Utz, Joan T. Merrill, Joel M. Guthridge, Judith A. James, Samuel Yang, Robert Tibshirani, Anshul Kundaje*‡, Scott D. Boyd*‡

INTRODUCTION: Conventional clinical diagnosis relies on physical examination, patient history, laboratory testing, and imaging, but makes little use of the receptors on B cells and T cells that reflect current and past exposures and responses. Microbial pathogen detection underpins infectious disease diagnosis. Other conditions are more challenging: Autoimmune diseases can require a combination of imaging studies and testing for autoantibodies and other laboratory abnormalities in the blood that may not yield a definitive disease classification. This process can be lengthy and may be complicated by initial misdiagnoses and ambiguous or overlapping symptoms between conditions.

B cell receptors (BCRs) and T cell receptors (TCRs) allow these immune cells to recognize and respond to specific antigens on pathogens and sometimes the body's own tissues. The genes encoding BCRs and TCRs are generated by random recombination of segments in the genome of individual cells during their development, and have potential as a diverse set of sequence biomarkers associated with immune system activity. BCR and TCR populations change after exposure to pathogens, after vaccination, and in response to autoantigens in autoimmune conditions, reflecting clonal expansion and selection of B cells and T cells during immune responses. Sequencing and interpreting BCR and TCR genes could provide a single diagnostic test for simultaneous assessment of many diseases.

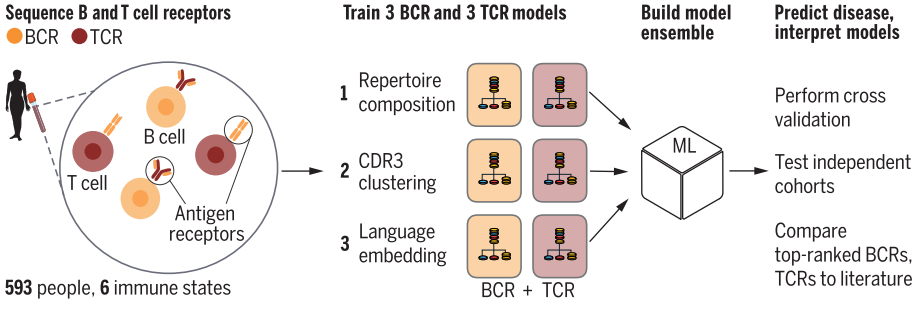
RATIONALE: We designed experimental protocols and a data analysis framework for identifying human BCR heavy chain and TCR beta chain features characteristic of infectious and immunological disorders or elicited by therapeutic or prophylactic interventions such as vaccination. Our method, named MAchine Learning for Immunological Diagnosis (Mal-ID), combines traditional immunological analyses, such as shared sequence detection between individuals with the same condition, with more complex features derived from artificial intelligence (AI) models of protein sequences, called protein language models. Although AI systems can be difficult to interpret, we developed ways to understand how the model makes its diagnostic predictions.

RESULTS: We generated large datasets of both BCR heavy chain and TCR beta chain sequences from the same individuals, spanning six disease or immune response states, to train and evaluate the Mal-ID model. Mal-ID accurately identified immune status from blood samples of 542 individuals with COVID-19, HIV, lupus, type 1 diabetes, recent flu vaccination, and healthy controls, achieving a multiclass area under the receiver operating characteristic curve (AUROC) of 0.986 on data not used for training. Combining features from both B cell and T cell receptor data led to the highest classification performance, but even with only BCR sequences, we still achieved high classification performance (0.959 AUROC in an expanded cohort adding 51 individuals for whom only BCR data were available).

Despite the model being trained to classify multiple heterogeneous classes, it can be specialized for detecting a particular condition. When applied to specifically distinguish patients with lupus from other patients and healthy controls, the classifier achieved 93% sensitivity and 90% specificity. This performance relative to current tests indicates the potential for BCR and TCR sequence analysis to detect clinically relevant signals.

We used the model to provide insights into the biologically relevant features that enable accurate disease classification. Examining which sequence categories contributed most to predictions, we confirmed that the patterns discovered from the data matched established immunological knowledge. Additionally, we assessed whether the model identified individual immune receptors known to be associated with disease. The model assigned higher COVID-19 association scores to sequences from an external database of SARS-CoV-2 binding BCRs, compared with sequences from healthy donors. We also verified that batch effects and demographic factors such as age, sex, and ancestry were not responsible for disease classification performance, and the model performed well when tested on external datasets from other laboratories.

CONCLUSION: This pilot study demonstrates that immune receptor sequencing data can distinguish a range of disease states and extract biological insights without prior knowledge of antigen-specific receptor patterns. With further validation and extension, Mal-ID could lead to clinical tools that harness the vast information contained in immune receptor populations for medical diagnosis. ■



From blood to disease classification with immune receptor sequencing. B and T cell receptors sequenced from 593 individuals were analyzed through receptor population or "repertoire" composition; clustering CDR3 sequence regions, which determine receptor antigen specificity; and protein language modeling. Disease was classified from this BCR and TCR information with a high multiclass AUROC score in cross validation experiments.

The list of author affiliations is available in the full article online.
*Corresponding author. Email: akundaje@stanford.edu (A.K.); sboyd1@stanford.edu (S.D.B.)
†These authors contributed equally to this work.
‡These authors contributed equally to this work.
Cite this article as M. E. Zaslavsky *et al.*, *Science* **387**, eadp2407 (2025). DOI: 10.1126/science.adp2407

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.adp2407

RESEARCH ARTICLE

IMMUNOLOGY

Disease diagnostics using machine learning of B cell and T cell receptor sequences

Maxim E. Zaslavsky[1]†, Erin Craig[2]†, Jackson K. Michuda[2], Nidhi Sehgal[3,4], Nikhil Ram-Mohan[5], Ji-Yeun Lee[4], Khoa D. Nguyen[4], Ramona A. Hoh[4], Tho D. Pham[4,6], Katharina Röltgen[7,8], Brandon Lam[4], Ella S. Parsons[9], Susan R. Macwana[10], Wade DeJager[10], Elizabeth M. Drapeau[11], Krishna M. Roskin[12,13], Charlotte Cunningham-Rundles[14], M. Anthony Moody[15,16,17], Barton F. Haynes[16,17,18], Jason D. Goldman[19,20], James R. Heath[21,22], R. Sharon Chinthrajah[9], Kari C. Nadeau[23,24], Benjamin A. Pinsky[4,25], Catherine A. Blish[25], Scott E. Hensley[11], Kent Jensen[25], Everett Meyer[25], Imelda Balboni[26], Paul J. Utz[25], Joan T. Merrill[10,27,28], Joel M. Guthridge[10], Judith A. James[10], Samuel Yang[5], Robert Tibshirani[2,29], Anshul Kundaje[1,3]*‡, Scott D. Boyd[4,9]*‡

Clinical diagnosis typically incorporates physical examination, patient history, various laboratory tests, and imaging studies but makes limited use of the human immune system's own record of antigen exposures encoded by receptors on B cells and T cells. We analyzed immune receptor datasets from 593 individuals to develop MAchine Learning for Immunological Diagnosis, an interpretive framework to screen for multiple illnesses simultaneously or precisely test for one condition. This approach detects specific infections, autoimmune disorders, vaccine responses, and disease severity differences. Human-interpretable features of the model recapitulate known immune responses to severe acute respiratory syndrome *coronavirus* 2, influenza, and human immunodeficiency virus, highlight antigen-specific receptors, and reveal distinct characteristics of systemic lupus erythematosus and type-1 diabetes autoreactivity. This analysis framework has broad potential for scientific and clinical interpretation of immune responses.

Modern medical diagnosis relies heavily on laboratory testing for cellular or molecular abnormalities. For example, detection of pathogenic microorganisms in patients with appropriate clinical history and physical examination findings can indicate infectious disease (*1*). For autoimmune diseases such as systemic lupus erythematosus (SLE), multiple sclerosis, or type-1 diabetes (T1D), there is no single pathogenic agent to detect and therefore a combination of diagnostic approaches is used, integrating data from the patient history, physical examination, imaging studies, testing for autoantibodies and other laboratory abnormalities, and exclusion of other conditions. This process can be lengthy and is often complicated by initial misdiagnoses and ambiguous symptoms (*2*, *3*).

Diagnostic medicine currently makes minimal use of data from the adaptive immune system's B cell receptors (BCR) and T cell receptors (TCR) that provide antigen specificity to immune responses. The genes encoding these receptors are randomly rearranged from gene segments in the germline DNA during the development of each B cell or T cell to yield a diverse repertoire of receptor specificities for antigens. In response to pathogens, vaccines, and other stimuli, the repertoires of BCRs and TCRs change in composition by clonal expansion of antigen-specific cells, introduction of additional somatic mutations into BCR genes, and selection processes that further reshape lymphocyte populations. Self-reactive lymphocytes can also clonally proliferate and cause autoimmune diseases or other immunological pathologies. Sequencing of BCRs and TCRs from an individual has the potential to provide a single diagnostic test allowing simultaneous assessment for many infectious, autoimmune, and other immune-mediated diseases (*4*, *5*).

Receptor repertoire sequencing already contributes to diagnosis and treatment response monitoring in the specialized case of lymphocyte malignancies where the BCR or TCR is a marker of the cancer cells (*6*, *7*). Moreover, prior research suggests that BCR sequencing can distinguish between some antibody-mediated pathologies (*8*). Challenges to broader application of these methods in clinical diagnoses include low frequencies of antigen-specific B cells and T cells in many patients, the high diversity of immune receptor genes produced by gene rearrangement during lymphocyte development, and somatic hypermutations that accumulate in BCRs following B cell stimulation, leading to complex data in which only a fraction of sequences are informative (*9*, *10*). Other limitations are technical factors including varying experimental protocols for sequence library preparation, and differences in patient demographics or past exposures that may influence the responses to a given antigen (*11*), suggesting a need for systematic collection of larger datasets.

Previous investigations of disease or vaccination-related immune repertoires have identified, with varying degrees of success, highly similar receptor amino acid sequences or motifs in people with the same exposures, addressing relatively few immune response types (*12*–*20*). In contrast to direct matching of the primary amino acid sequences, other studies have used alternative encodings of amino acid biochemical properties, such as charge and polarity, to improve detection of receptor groups of potentially similar antigen specificity (*21*).

Recently, numerical feature representations of BCRs or TCRs derived from neural network methods, including protein language models and variational autoencoders, have been applied to immune state classification and for predictive applications such as therapeutic antibody optimization (*22*–*29*). Probabilistic models of receptor gene segment recombination and selection processes have also been

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [3]Department of Genetics, Stanford University, Stanford, CA, USA. [4]Department of Pathology, Stanford University, Stanford, CA, USA. [5]Department of Emergency Medicine, Stanford University, Stanford, CA, USA. [6]Stanford Blood Center, Stanford, CA, USA. [7]Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Allschwil, Switzerland. [8]University of Basel, Basel, Switzerland. [9]Sean N. Parker Center for Allergy and Asthma Research, Stanford University, Stanford, CA, USA. [10]Department of Arthritis and Clinical Immunology, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA. [11]Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [12]Department of Pediatrics, University of Cincinnati, College of Medicine, Cincinnati, OH, USA. [13]Divisions of Biomedical Informatics and Immunobiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [14]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [15]Department of Pediatrics, Duke University, Durham, NC, USA. [16]Duke Human Vaccine Institute, Duke University, Durham, NC, USA. [17]Department of Immunology, Duke University, Durham, NC, USA. [18]Department of Medicine, Duke University, Durham, NC, USA. [19]Swedish Center for Research and Innovation, Swedish Medical Center, Seattle, WA, USA. [20]Division of Allergy and Infectious Diseases, University of Washington, Seattle, WA, USA. [21]Institute for Systems Biology, Seattle, WA, USA. [22]Department of Bioengineering, University of Washington, Seattle, WA, USA. [23]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [24]Division of Allergy and Inflammation, Beth Israel Deaconess Medical Center, Boston, MA, USA. [25]Department of Medicine, Stanford University, Stanford, CA, USA. [26]Department of Pediatrics, Stanford University, Stanford, CA, USA. [27]Department of Medicine, Grossman School of Medicine, New York University, New York, NY, USA. [28]Lupus Foundation of America, Washington, DC, USA. [29]Department of Statistics, Stanford University, Stanford, CA, USA.

***Corresponding author. Email: akundaje@stanford.edu (A.K.); sboyd1@stanford.edu (S.D.B.)**

†These authors contributed equally to this work.

‡These authors contributed equally to this work.

applied to better understand immune receptor generation and expansion in response to antigenic stimuli (*30*). Very few studies have attempted to integrate BCR and TCR data for diagnostic purposes, however, and it remains unclear to what extent immune receptor repertoire sequence data are sufficient for generalized and accurate infectious or immunological disease classification.

To address these challenges, we developed and validated MAchine Learning for Immunological Diagnosis (Mal-ID), which combines three machine learning representations for both BCR and TCR repertoires to detect infectious or immunological diseases in patients.

Results

Integrated repertoire models of immune states

For Mal-ID, we used three models per gene locus [BCR immunoglobulin heavy chain (IgH) and TCR beta chain (TRB)] to recognize immune states (Fig. 1 and fig. S1). IgH and TRB gene rearrangements are the most diverse and informative components of BCRs and TCRs because they are assembled from three different germline gene segment types: variable (V), diversity (D), and joining (J). The subsequence spanning the end of the V segment to the beginning of the J segment encodes the key antigen-binding complementarity-determining region 3 (CDR3). The VDJ rearrangements of IgH become joined to constant region genes to encode different isotypes including immunoglobulin M (IgM), IgD, IgG, and IgA which have different functional properties. In antigen-stimulated B cells, additional somatic hypermutation (SHM) sequence changes contribute to VDJ diversity and antigen binding affinity. In Mal-ID, each model focused on different aspects of immune repertoires shared between individuals with the same immune state or diagnosis: gene segment frequencies and IgH SHM rates in each isotype (model 1), highly similar CDR3 sequence clusters (model 2), and inferred potential structural or binding similarity based on embeddings of CDR3 sequences generated with the Evolutionary Scale Modeling-2 (ESM-2) protein language model (*31*) (model 3). Outputs from the three BCR and three TCR models were combined into a final prediction of immune status with a logistic regression ensemble model that could resolve potential errors of individual predictors (*32*). The trained program took an individual's peripheral blood BCRs and TCRs as input and predicted the probability of each disease on record (Fig. 1C). Full details of the modeling approach are provided in the materials and methods.

We applied Mal-ID to 16.2 million BCR heavy chain clones and 23.5 million TCR beta chain clones systematically collected from peripheral blood samples of 593 individuals, including patients diagnosed with Covid-19 (n = 63), human immunodeficiency virus (HIV) infection (n = 95) (*13*), SLE (n = 86), and T1D (n = 92), as well as influenza vaccination recipients (n = 37) and healthy controls (n = 220) (table S1). In total, 542 individuals had paired IgH and TRB sequence data. All datasets used a standardized sequencing protocol to minimize batch effects. To evaluate generalizability, patients were strictly separated into training, validation, and testing sets (fig. S2). Any repeated samples from the same individual were kept grouped together during this division process, to ensure that data from the same individual did not leak between training and testing steps. We trained separate models per cross-validation fold and report averaged classification performance. As described below, we further tested

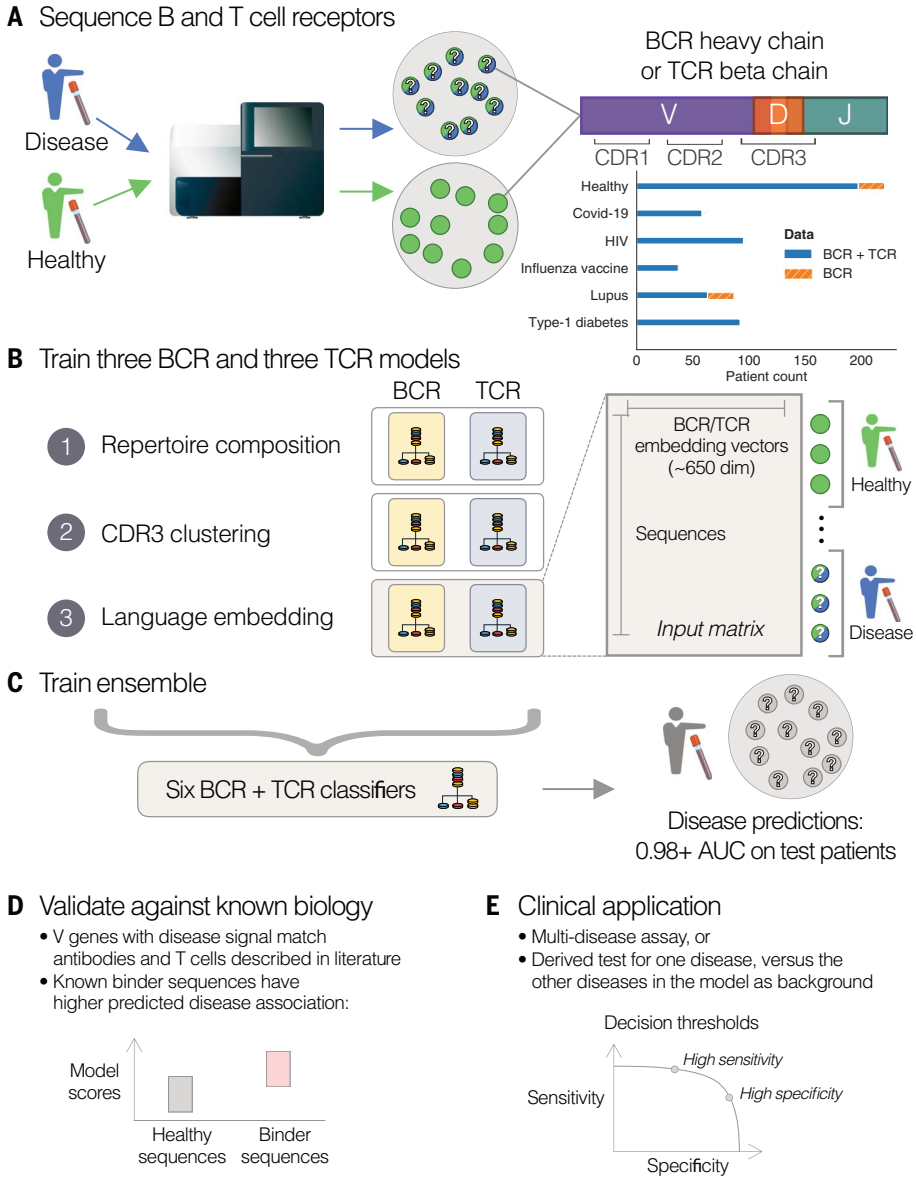


Fig. 1. Mal-ID framework. (**A**) BCR heavy chain and TCR beta chain gene repertoires are amplified and sequenced from blood samples of individuals with different disease states. Question marks indicate that most sequences from patients are not disease-specific. (**B**) Machine learning models are trained to predict disease using several immune repertoire feature representations. These include protein language models, which convert each amino acid sequence into a numerical vector. (**C**) An ensemble disease predictor is trained using three BCR and three TCR base models. The combined model predicts disease status of held-out test individuals. (**D**) For validation, the disease prediction model allows introspection of which V genes carry disease-specific signals, which can be validated against prior literature. Within each V gene, previously published BCR and TCR sequences known to be disease-associated can be tested for whether they have higher disease association. (**E**) The final trained model can be applied as either a multidisease assay or a one-disease diagnostic test. The same model will achieve a range of sensitivities and specificities depending on the chosen decision threshold.

for the potential contribution of batch effects and demographic differences to diagnostic accuracy.

The ensemble approach distinguished six specific disease states in 550 paired BCR and TCR samples from 542 individuals with a multiclass area under the receiver operating characteristic curve (AUROC) score of 0.986 (Fig. 2A). AUROC represents the probability of correctly ranking a randomly chosen positive example higher than a randomly chosen negative example (*33*). In our multiclass setting, it is computed and averaged across all disease label pairs, weighted by their frequencies. Other performance metrics are provided in table S2.

Mal-ID outperformed previously reported classification approaches on our evaluation dataset. The CDR3 clustering model, similar to convergent or public sequence discovery approaches in the literature, achieved only 0.89 AUROC for BCR and 0.80 AUROC for TCR (Fig. 2B). Another approach based on exact sequence matches, originally reported for TCR sequences (*12*), achieved 41% accuracy for BCR data and found no hits in 40% of samples (fig. S3). Identical sequences across individuals were expected to be rare for IgH because of somatic hypermutation, but the HIV class was an exception. For TCR data, the exact matches technique almost always found hits but achieved only 42% accuracy and 0.75 AUROC, and it predicted that almost all samples belong to either the Covid-19 class or the healthy class (fig. S3). Mal-ID's AUROC of over 0.98 represents a major increase in diagnostic accuracy.

The three-model approach discriminated between autoimmune diseases, viral infections, and influenza vaccine recipient samples collected at day 7 after vaccination, when B cells responding to the vaccine are usually at peak frequencies (*34*). The different BCR and TCR components of the ensemble model contributed to varying degrees for classification of each immunological condition (Fig. 2B and fig. S4). TCR sequencing provided more relevant information for lupus and T1D, whereas Covid-19, HIV, and influenza had clearer BCR signatures. Combined BCR and TCR data performed best (table S2). Alone, the repertoire composition model 1 and protein language embedding model 3 classifiers performed better on average than the CDR3 clustering in model 2. The TCR CDR3 clustering model was the weakest, potentially because the model did not account for patient human leukocyte antigen (HLA) genotypes that alter the protein sequences of the cell surface complexes that present peptide antigens for TCR recognition. Model 2 identified relatively few public TCR clusters (those of highly similar sequences identified in more than one individual) meeting the model's significance threshold for enrichment in Covid-19 patients, and for T1D, relatively few public BCR or TCR clusters were chosen (table S3). The combination of models 1, 2, and 3 generally had best performance, but pairing

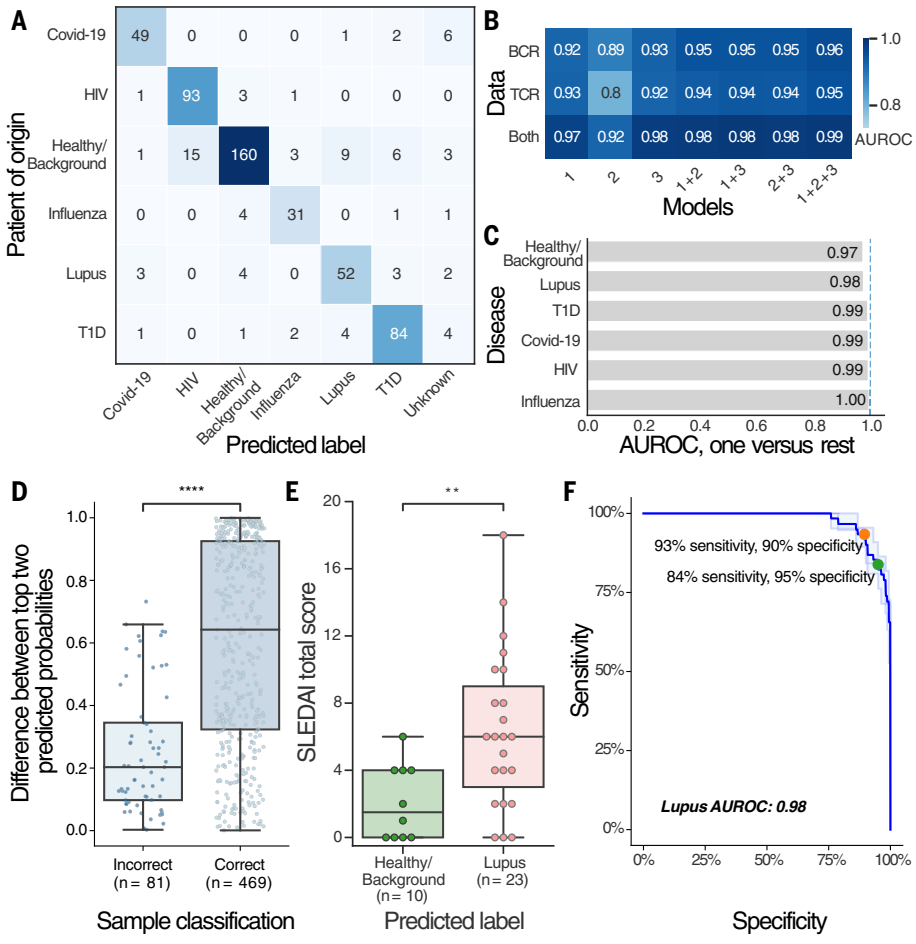


Fig. 2. Mal-ID classifies disease using IgH and TRB sequences. (**A**) Disease classification performance on held-out test data by the ensemble of three B cell repertoire and three T cell repertoire machine learning models, combined over all cross-validation folds. The number of predictions (values in boxes) for each combination of true and predicted labels is shown, for a total of n = 550 paired BCR and TCR samples. (**B**) Disease classification performance, calculated as multiclass one-versus-one AUROC scores, divided column-wise by model architecture (individual base models or ensembles of base models) and row-wise by whether BCR data, TCR data, or both were incorporated. Model 1 refers to the repertoire composition classifier, model 2 refers to the CDR3 clustering classifier, and model 3 refers to the protein language model classifier. The CDR3 clustering models abstain from prediction on some samples whereas the other models do not abstain; to make the scores comparable, abstentions were forcibly applied to the other models. The BCR-only results also include BCR-only patient cohorts (n = 66 samples) not present in TCR-only or BCR+TCR evaluation. (**C**) AUROC scores for each class versus the rest from the full ensemble architecture including models 1, 2, and 3 with both BCR and TCR data. (**D**) Difference of probabilities of the top two predicted classes for correct versus incorrect ensemble model predictions. A higher difference implies that the model is more certain in its decision to predict the winning disease label, whereas a low difference suggests that the top two possible predictions were a toss-up. Results were combined across all cross-validation folds. Each box represents the interquartile range (IQR) between the 25th and 75th percentiles of the data, with the line inside the box representing the median value. Whiskers extend to the farthest values within 1.5 times the IQR from the edges of the box. Data points represent individual samples, with total sample number n indicated below each boxplot. One-sided Wilcoxon rank-sum test: P-value 1.599×10^{-15}, U-statistic 6052. (**E**) SLEDAI clinical disease activity scores for adult lupus patients who were either classified correctly or misclassified as healthy by the BCR-only ensemble model, used here because the adult lupus data was primarily BCR-only. SLEDAI scores were only available for some patients. Boxes represent data interquartile ranges with median lines, and whiskers show data extremes up to 1.5 times the IQR from the box. Data points represent individual samples, with total sample number n indicated below each boxplot. One-sided Wilcoxon rank-sum test: P-value 4.242×10^{-3}, U-statistic 48. (**F**) Sensitivity versus specificity, averaged over three cross-validation folds, for a lupus diagnostic classifier derived from the pan-disease classifier. Two possible decision thresholds are highlighted. *P < 0.05, **P < 0.01, ***P < 0.001, and ****P < 0.0001.

models 1 and 3 performed as well for many classes (fig. S4), suggesting that CDR3 clustering may not be required for classification or is encompassed by the protein language model results.

In practice, decision thresholds to categorize patient samples into disease categories can be chosen depending on the consequences of different types of errors, the performance metrics to be optimized, and the priority given to different diseases. We illustrated how the estimated AUROCs translated to explicit misclassification rates for a few different case studies. When we assigned each patient to the immune state with the highest predicted probability, Mal-ID achieved 85.3% accuracy (Fig. 2A). Among misclassified repertoires, 2.9% lacked sequences belonging to model 2 CDR3 clusters, making the CDR3 clustering component abstain from prediction. The remaining 11.8% had inconclusive predictions (Fig. 2D). Many misclassifications involved healthy donors predicted as having an illness, indicating that the model selecting classification labels based on the highest prediction probabilities resulted in more false positive than false negative results. Some of these errors may also have been caused by healthy control individuals not being screened for definitive absence of all the diseases in our panel. However, 92.9% of sick patients and vaccine recipients were identified as not being in a healthy/baseline immune state, and 87.5% had their particular immune state properly classified. Adult lupus patients were the most challenging disease category to classify (Fig. 2C). Unlike the pediatric lupus cohort, the adults were on therapy, which can influence immune repertoires (*8*). Most adult lupus patient samples had BCR data only. Based on this more limited data, a subset of patients was predicted as healthy (fig. S5). However, misclassified patients had lower SLE disease activity index (SLEDAI) scores (*35*) (Fig. 2E), indicating better-controlled or quiescent disease in response to treatment, which likely influenced the model's tendency to classify them as immunologically healthy. Compared with the 85.3% overall accuracy achieved by the model using BCR and TCR data together, the BCR-only and TCR-only versions of Mal-ID had 74.0 and 75.1% accuracy (table S2), respectively, further highlighting the benefit of analyzing BCR and TCR data jointly when they are available.

Disease-specific classifiers can also be trained or derived from the pan-disease model. For example, by labeling lupus predictions as positives and others as negatives, we extracted a lupus diagnosis model, which is clinically relevant because of a lack of a sensitive and specific lupus test (*3*). Adjusting the decision threshold for high lupus sensitivity, our model achieved 97% sensitivity and 86% specificity, or 84% sensitivity and 95% specificity when optimized for specificity (Fig. 2F). Balanced performance of 93% sensitivity and 90% specificity was also possible. This proof-of-concept result suggests that a classifier based on the Mal-ID framework could be developed into a multidisease test or be specialized for detecting a particular condition.

Limited impact of batch effects on classification

To assess Mal-ID's generalizability, we trained a model on all data (fig. S2) and then tested on Covid-19 patient and healthy donor repertoires from other BCR or TCR studies with similar complementary DNA (cDNA) sequencing protocols. Mal-ID predicted disease in two BCR external cohorts (*36*, *37*) with perfect 1.0 AUROC; all seven Covid-19 patients received higher Covid-19 predicted probabilities than did the six healthy donors. However, accuracy was 69% by assignment to the immune state with highest probability: one Covid-19 patient was misclassified as T1D and three healthy donors were misclassified as lupus or T1D (fig. S6A and table S4). As the base rates of disease have changed in this evaluation dataset containing only Covid-19 patients and healthy donors, the decision thresholds were tuned using a small portion of the external cohorts. After this tuning, the adjusted BCR model reached 100% accuracy in the remaining evaluation data (fig. S6B). Similar tuning could thus be performed for clinical contexts with varying disease prevalence.

In TCR external cohorts of 17 Covid-19 patients and 39 healthy donors (*38*–*40*), Mal-ID achieved 0.99 AUROC and 68% accuracy based on highest-probability class assignment, which rose to 90% accuracy after threshold tuning (fig. S6, C and D, and table S4). Almost all Covid-19 patients and healthy donors evaluated (excluding those used for tuning, to avoid train-test leakage) were correctly identified, with the exception of three healthy donors (out of 28) who were misclassified as Covid-19 and one Covid-19 patient (out of 12) who was misclassified as healthy. Low accuracy prior to tuning was caused by misclassifications of Covid-19 patients as lupus due to model 2, which also performed poorly on our primary TCR data as noted above. Disabling model 2 led to no Covid-19 patients misclassified as lupus and 89% accuracy without tuning (along with 0.97 AUROC). High performance on external published cDNA-derived datasets suggested that Mal-ID learned generalizable disease-related signals, even when only BCR or only TCR data were available. The classification framework could also be retrained for other sequencing modalities, including TCR genomic DNA-templated sequencing data from Adaptive Biotechnologies. Observing gene segment usage distinct from cDNA data as previously reported (*11*) (fig. S7A), we trained Mal-ID to successfully separate six immune states in 1365 samples: common variable immunodeficiency, Covid-19, HIV, rheumatoid arthritis, T1D, and healthy controls. These studies were conducted by different labs, introducing the possibility of batch effects, and were restricted to TCR data (table S5). Mal-ID classified these disease classes with 0.97 AUROC and 88% accuracy (fig. S7B), indicating that Mal-ID could learn disease signals across sequencing modalities and can scale to over 150 million sequences. As in the primary Mal-ID dataset, misclassifications often involved healthy individuals being predicted as sick, but 96% of sick patients were correctly identified as having an illness. The Covid-19 and healthy data came from studies that were divided into multiple cohorts; for example, the Emerson *et al.* 2017 study of healthy individuals included an original cohort and an independent validation cohort (*12*). Therefore, we also trained Mal-ID with these cohort divisions preserved. Holding out entire Covid-19 and healthy cohorts from the training process, we saw that Mal-ID accurately classified the independent cohorts with 1.0 AUROC and 98% accuracy (fig. S7C).

To test for batch effects in our primary data, we retrained Mal-ID holding out an entire Covid-19 cohort of 10 patients (denoted as "group B" in table S1), whose sequence libraries were generated from PBMCs (primarily composed of lymphocytes and monocytes) unlike the primary Covid-19 dataset derived from whole-blood PAXgene RNA tubes, which contain RNA from all cell types in the blood. We also held out 13 healthy samples that were resequenced in a separate replicate batch, following independent cDNA generation and PCR amplification from the original RNA sample ("group K" in table S1). All held-out Covid-19 samples and healthy samples (pooling the original and replicate data) were correctly classified. When we split each healthy donor's replicates, both replicates were correctly classified for 9 of 13 healthy donors with 97% or higher correlation between predicted class probabilities, while two individuals had replicates with abstention from classification and two had divergent classification for each replicate (fig. S8). Classification abstentions resulted from two replicates matching no class-associated CDR3 clusters, which was likely caused by these replicates having fewer IgH clones than the rest due to limited sequencing depth. We repeated this test, retraining Mal-ID while holding out an independent cohort of five lupus patients and two healthy controls ("group G" in table S1), which were collected in whole-blood PAXgene RNA tubes unlike the remaining lupus cohorts used for training. Four out of five lupus patients and two out of two healthy individuals were correctly classified, in line with the overall accuracy of Mal-ID. The accurate classification of completely independent cohorts and consistent scoring of healthy replicate samples increases the likelihood that Mal-ID learned true biological signal rather than batch effects.

Limited impact of age, sex, and race on classification

Patient demographics also influence the immune repertoire (*39*–*41*). To evaluate how extraneous

covariates may affect classification, we attempted to predict age, sex, or ancestry from the immune repertoires of healthy individuals. Although sex could not be accurately determined, sequences carried relatively weak ancestry signals (0.78 AUROC, table S6). Ancestry separation was visible in gene segment usage (fig. S9A), potentially from germline IgH and TRB locus differences, shaping of TCR repertoires by HLA alleles that differ between ancestry groups, and different environmental exposures in the African ancestry individuals living in Africa in the data (*41*). Consistent with potential influences of HLA genotype, Mal-ID's TCR components had less accuracy in distinguishing HIV patients and healthy controls from the African cohort. The corresponding IgH repertoires were more distinct (fig. S10), highlighting the advantages of combining BCR and TCR data.

Previous studies noted age-related changes in gene expression, cytokine levels, and immune cell frequencies (*42*). We observed a modest age signal in healthy IgH and TRB sequences, achieving 0.75 AUROC for distinguishing age 50 and up, excluding 19% of samples that matched no CDR3 age clusters (54% accuracy including abstentions; table S6). Age signatures may correspond to imprinting effects from childhood exposure to viruses such as influenza (*43*) or to autoreactivity increasing with age (*44*). Pediatric samples had especially distinct TCR beta V (TRBV) gene usage (fig. S9B), and Mal-ID identified them with perfect 1.0 AUROC when it made predictions (table S6), though accuracy was 55% due to 45% abstention. Despite substantial differences in the remaining samples, age effects did not interfere with disease classification: Mal-ID accurately distinguished pediatric patients and controls (fig. S5C). The high model 2 abstention rates indicated relatively few age-associated CDR3 sequence clusters and showed that unsupervised clustering will not necessarily choose clusters that correspond to age or other desired axis of variation. Also, we restricted Mal-ID's scope to B cell populations shaped by antigenic stimulation: somatically hypermutated IgD/IgM and class switched IgG/IgA isotypes. Studying naïve B cells may reveal additional age, sex, or ancestry effects.

To assess whether demographic differences between disease cohorts drove our classification results, we attempted to predict disease state from age, sex, and ancestry alone, ignoring sequence data. Ages by cohort were: T1D median 14.5 years (range 2 to 74); SLE median 18 years (range 7 to 71); influenza vaccine recipient median 26 years (range 21 to 74); HIV median 31 years (range 19 to 64); healthy control median 34.5 years (range 8 to 81); Covid-19 median 48 years (range 21 to 88) (table S1). The percentage of females in each cohort was 50% (healthy controls), 52% (Covid-19), 57% (influenza vaccine recipient), 64% (HIV), and 85% (SLE), consistent with higher prevalence of SLE in females (*45*). The ancestries and geographical locations of participants also differed between cohorts. Notably, 89% of individuals in the HIV cohort lived in Africa (*13*). Using only age, sex, or ancestry, disease AUROCs were 0.68, 0.59, and 0.79, respectively. A classifier with all three features achieved 0.85 AUROC, substantially lower than the 0.98 AUROC from Mal-ID retrained with demographics alongside sequence features (table S7 and fig. S11, A and B).

To further evaluate whether the disease signal was derived primarily from BCR and TCR sequences, we also tested the demographics-only classifier on the external cDNA datasets. For TCR, it achieved 0.48 AUROC and 50% accuracy (fig. S6F), compared with Mal-ID's 0.99 AUROC and 68% accuracy before threshold tuning (table S4). For BCR, the demographics-only classifier achieved 1.0 AUROC, identical to the standard Mal-ID model, because the external Covid-19 patients were all Asian whereas the healthy controls were Caucasian or African American. Nevertheless, accuracy was 58% with demographic features (fig. S6E), compared with 69% with Mal-ID before tuning (table S4). Demographic covariates, therefore, did not explain model performance on external validation data. As an additional test to confirm that predictions were not driven by demographics, we retrained with age, sex, and ancestry effects regressed out from the ensemble model's feature matrix. Classification performance for individuals with known demographics dropped slightly from 0.98 AUROC to 0.96 AUROC after decorrelating sequence features from demographic covariates (table S7 and fig. S11C), suggesting that age, sex, and ancestry had modest impacts on disease classification.

Language model recapitulates immunological knowledge

To better understand the factors contributing to the high accuracy of Mal-ID classification, we asked which biological patterns identified each disease. Model 3 revealed which receptor sequences contributed most to disease predictions because BCRs or TCRs were scored individually, then aggregated into patient predictions. Separate models generated sequence predictions specialized for each BCR immunoglobulin heavy chain V (IGHV) gene and isotype combination in the BCR case, or for each TRBV gene in TCR data (materials and methods). We calculated Shapley importance (SHAP) values (*46*) for the disease probabilities derived from each sequence category, which served as features for making model 3's patient predictions. V genes and isotypes were given priority in the aggregation model based on their prevalence in patients and on containing sequences distinct from other immune states by CDR3 features. According to V gene category contributions to disease predictions, our model's classifications aligned with established immunological knowledge from data such as antigen-specific B cell and T cell isolation and receptor sequencing (supplementary text). For example, particular BCR V genes IGHV1-24 and IGHV2-70 were prioritized for Covid-19 prediction, IGHV4-34 and IGHV4-59 had greater weight for lupus, IGHV1-2 and IGHV4-34 for HIV, and IGHV3-23 for influenza (Fig. 3). We also decomposed the lupus and T1D SHAP values into TRBV gene prioritization clusters corresponding to patient age (figs. S12 and S13). In our lupus cohort, age was associated with treatment status as the adults were on treatment while the pediatric cohort was treatment-naïve, indicating that differences in gene usage may also depend on treatment.

Different diseases showed varied association of IGHV gene usage in the context of particular BCR heavy chain isotypes. Covid-19 prediction prioritized IgG (Fig. 3A), as expected from prominent IgG expression by SARS-CoV-2–specific B cells (*47*, *48*). Although IgA contributions were minimal for Covid-19, HIV, and influenza predictions, IgA was informative for lupus, consistent with disease-associated IgA autoantibodies described in the literature (*49*), as well as for T1D, along with other isotypes (Fig. 3, D and E). The HIV model favored mutated IgM/D (Fig. 3B). Influenza predictions were primarily driven by IgG and mutated IgM/D signals (Fig. 3C). B cell isotype usage varied by person and across disease cohorts (fig. S14), but the model also considered distinct disease signal enrichment within each isotype to determine its priority. Other Mal-ID components were not influenced by isotype sampling variation: Model 1 quantified each isotype group separately and model 2 was blind to isotype information. To be sure that differences in isotype proportions between patient cohorts were insufficient to predict disease, we attempted to predict disease from a sample's isotype proportions without any sequence information, achieving only 0.68 AUROC compared with Mal-ID's AUROC of >0.98.

Having validated that V gene segments and isotypes prioritizations for disease identification matched the literature, we assessed whether the multidisease Mal-ID model could distinguish reported SARS-CoV-2 binding BCRs (*50*) from healthy donor sequences, despite having been trained for patient classification rather than sequence classification (supplementary text). Model 3 assigned higher Covid-19 probabilities to reported binders compared with healthy sequences for IGHV1-24, IGHV2-70, and other key V genes, with AUROC ranging up to 0.78 across IGHV genes and area under the precision-recall curve (AUPRC) up to 6.9 times higher than the baseline (Fig. 4, E to G). Model 2 Covid-19–associated clusters identified some known binders, with up to 100% precision in IGHV1-24 and IGHV3-53, among others, but with low recall (Fig. 4, A to D). The higher ranking of

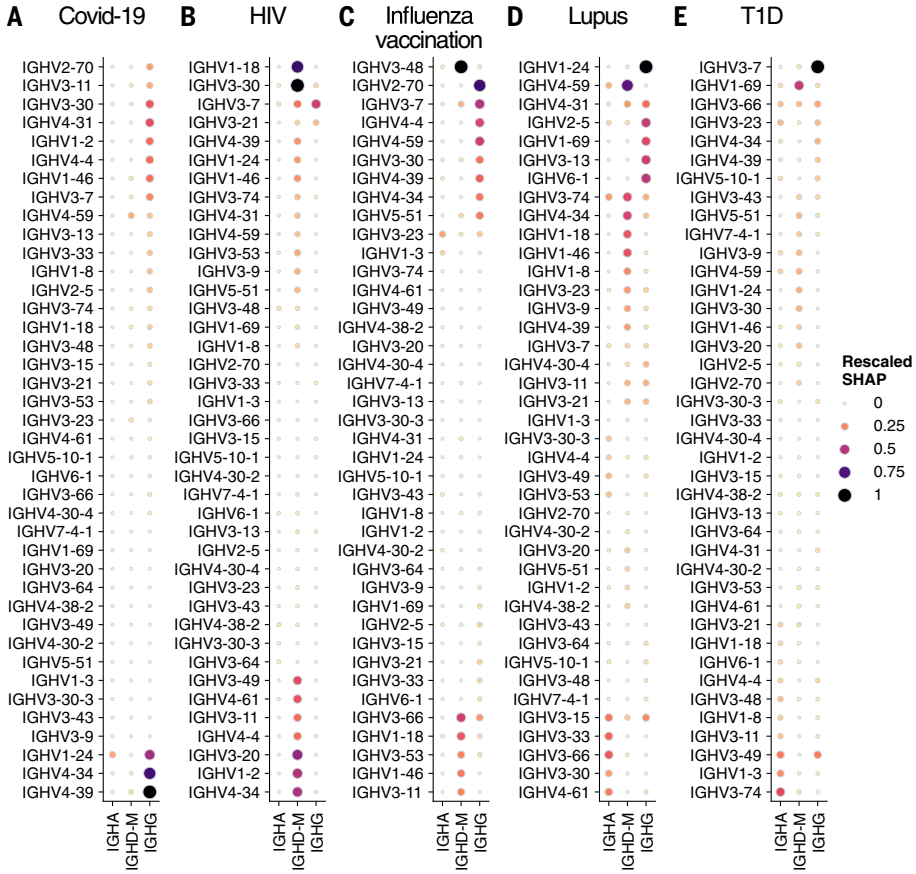


Fig. 3. Disease-associated IGHV genes and isotypes prioritized by model 3 using protein language embeddings. SHAP values quantifying the contribution of average sequence predictions from each IGHV gene and isotype category to model 3's prediction of a sample's disease state are plotted for (**A**) Covid-19 (averaged over n = 14 positive samples), (**B**) HIV (n = 21 positive samples), (**C**) influenza vaccination (n = 8 positive samples), (**D**) lupus (n = 22 positive samples), and (**E**) T1D (n = 22 positive samples).

experimentally validated, disease-specific sequences from separate cohorts suggested that the models learned antigen-specific sequence patterns within important IGHV genes that recapitulated biological knowledge gained during the international research effort in response to the Covid-19 pandemic, despite the enormous diversity of immune receptor sequences, and despite being trained without knowledge of which Covid-19 patient BCRs were specific for SARS-CoV-2 antigens. Only a fraction of peripheral blood B and T cell receptor sequences from Covid-19 patients are thought to be directly related to the SARS-CoV-2 viral antigen–specific immune response (*51*, *52*). However, cDNA sequencing may emphasize plasmablasts with high RNA copy counts, and excluding naïve B cells may highlight antigen-experienced B cells during training.

We repeated the test with influenza known binders (*53*), finding that both models again prioritized binding sequences in key IGHV genes (supplementary text). However, enrichment was more muted, ranging up to 0.65 AUROC and 4.0-fold change over baseline AUPRC for model 3. The relatively lower scores may be because the reference influenza-specific antibodies were derived from studies using a small sampling of all the influenza antigens that have been reported over past decades and were not derived from responses to the annual vaccine of the same year as the samples analyzed in our study. Differences in response to flu infection versus vaccination may also contribute to the relatively lower known binder enrichment scores; unlike the Covid-19 case in which the models were trained with data from patients, our influenza training data was limited to vaccinated individuals whereas the known binders studied were derived from both infected and vaccinated individuals.

Finally, evaluating SARS-CoV-2–specific TCRs (*54*), model 2 performed poorly, consistent with the relatively low model 2 TCR patient classification performance described earlier, and model 3 scores had weak enrichment for known binders, up to 0.56 AUROC and 1.30-fold change in AUPRC in any TRBV gene (supplementary text). Compared with IgH, TRB known binders may have had less enrichment for higher model 3 ranks over healthy sequences because the interactions between TCR and genetically diverse HLA molecules that present peptide antigens to T cells during T cell stimulation could introduce additional differences between cohorts and between participants within cohorts. In addition, activation of T cells upon peptide stimulation in culture may have resulted in some bystander clone activation not involved in the antigen-specific response. Further, unlike the IgH classification, the TCR analyses did not exclude naïve T cells that could contain low frequencies of SARS-CoV-2–specific clones in unexposed individuals. This moderate performance for antigen-specific sequence identification nevertheless led to high patient diagnosis performance; aggregating many complementary classifiers has been previously shown to be capable of producing a more accurate ensemble classifier (*55*). Also, to produce patient diagnosis predictions from TCR data, sequence-level predictions were aggregated simply by calculating average predicted probabilities after filtering out a percentage of low information content sequences (materials and methods). The strength of the patient predictions achieved by averaging many sequences indicated that diseases may alter immune repertoires by affecting a larger proportion of clones than those that explicitly bind antigens from the stimulus. Therefore, another possible explanation for the moderate enrichment in predicted probabilities for SARS-CoV-2 binding TCRs over healthy TCRs is that the classifier may have learned additional patterns other than those of TCRs that directly bind to the virus.

Discussion

In this study, we asked whether immune receptor sequencing could accurately determine a person's disease or immune response state, based on pathogenic exposures and autoreactivity shaping the immune system's collection of antigen-specific adaptive immune receptors. The three-part machine learning analysis framework we applied to well-characterized datasets of six distinct immunological states classified immune responses with performance of 0.986 AUROC, leveraging both B and T cell signals in 542 individuals. We ensured that models were never trained on data from a patient and then evaluated on other data from the same person. Faced with highly diverse repertoires containing tens to hundreds of thousands of distinct sequences, the Mal-ID ensemble of classifiers learned disease-specific patterns and prioritized meaningful sequences for prediction of specific viral infections and autoimmune diseases. These signatures of specific disease types overrode more modest differences detectable between individuals differing by sex, age, or ancestry. Mal-ID generalized to sequencing data from other laboratories and experimental protocols after additional tuning. Our architecture

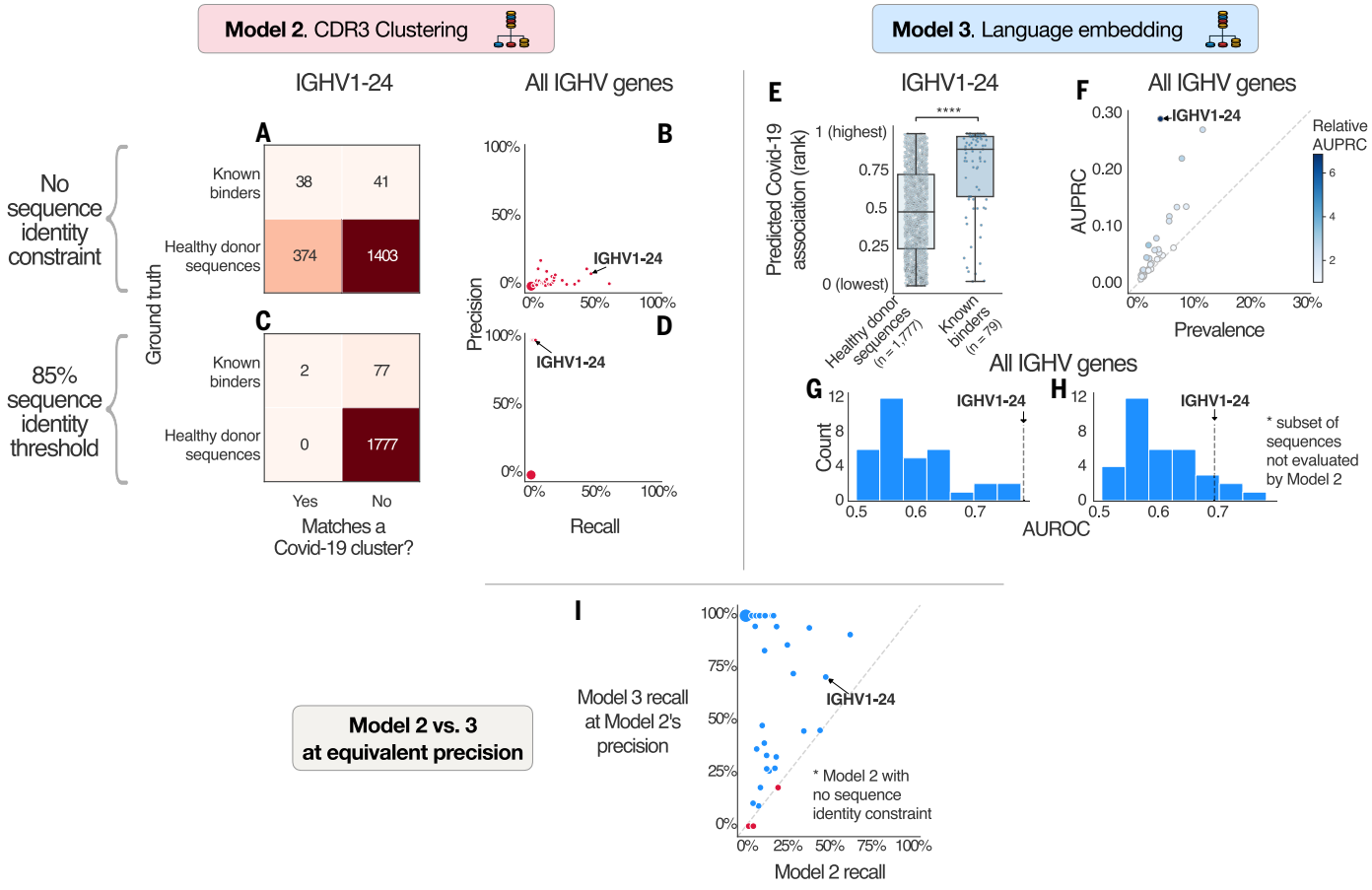


Fig. 4. Models 2 and 3 learn SARS-CoV-2 antigen-specific sequence patterns from Covid-19 patient data and can distinguish between known SARS-CoV-2–specific antibody sequences and healthy donor sequences. For this comparison, validated SARS-CoV-2-binding sequences from the CoV-AbDab database (*50*) and a subset of healthy donor sequences were held out from training. Known binder detection using model 2 or model 3 predictions of sequence association to disease was evaluated separately for each IGHV gene; performance is shown for IGHV1-24 and compared across IGHV genes. (**A**) to (**D**) Model 2 identifies a conservative set of public clones enriched in Covid-19 patients which match some known binders. In (A) and (C), the number of predictions (values in boxes) for each combination of true and predicted labels is shown for a total of n = 1856 sequences that use IGHV1-24. Model 2's precision and recall across IGHV genes is shown, with binding predictions determined: (A) and (B) based on shared IGHV gene, IGHJ gene, and CDR3 length with any Covid-19 cluster identified in model 2's training procedure; or (C) and (D) with an additional 85% CDR3 sequence identity threshold. (**E**) to (**H**) Model 3 ranks known binders higher than healthy sequences based on predicted Covid-19 probability (E), with relative AUPRC ranging up to 6.9 times higher than baseline prevalence (F) and AUROC up to 0.78 across IGHV genes (G). Permutation test in (E) to assess whether IGHV1-24 known binders have higher ranks than healthy donor sequences, with consistent labels maintained during the permutation process across sequences from each healthy donor: P-value 0. In (E), boxes represent interquartile ranges (IQR) with median value lines superimposed; whiskers extend to data points within 1.5 times the IQR from the box edges; and data points represent individual sequences using IGHV1-24, with total sequence number n indicated below each boxplot. (H) Model 3 maintains reasonable performance (AUROC up to 0.75) for sequences that are not evaluated by model 2's clustering (sequences for which model 2 identified no SARS-CoV-2 clusters with matching IGHV gene, IGHJ gene, and CDR3 length). (**I**) At equivalent precision, model 3 generally exhibits higher recall than model 2, identifying more true binders but with increased false positives. IGHV genes where model 3 has higher recall than model 2 are shown in blue. For each IGHV gene, recall was calculated for models 2 and 3 at model 2's precision shown in (B), with no sequence identity constraint applied during matching to model 2 clusters. Data points represent n = 34 individual V genes in (B), (D), (F), (G), (H), and (I). Point size indicates number of identical values plotted at a particular location for (B), (D), and (I). *P < 0.05, **P < 0.01, ***P < 0.001, and ****P < 0.0001.

scaled to population-level data; in this study, we demonstrated its use for over 1350 samples at a time with external datasets.

Key innovations for Mal-ID's performance are the trio of analysis models to extract signals from B and T cell receptor repertoires, as well as the way they are combined, fusing aggregate repertoire composition properties, detection of important sequence groups, and language model interpretations of individual sequences. The components are complementary; integrating these models outperformed them individually and suggested that they capture different patterns. Combining BCR and TCR repertoire data provided more accurate classification than either receptor type alone, potentially reflecting variation in the roles of B cell and T cell responses in different diseases. For example, T1D is considered to be predominantly T cell–mediated (*56*) and our T cell–only model indeed distinguished T1D from other classes better than our B cell–only model, but combining both signals further increased T1D detection performance. Similarly, lupus could be classified by either B or T cell information alone, which is supported by the prominence of autoantibodies in this condition and the known contributions of T cells to the pathology of SLE (*57*), but it was best classified by the combination of B cell and T cell models.

These results confirmed that B and T cell information considered together in immune response analysis provided a more complete description of the immune state.

The CDR3 clustering and language model components of our model assessed which receptor sequences have highest predicted disease association. Sequences independently validated to be pathogen-associated were distinguished from healthy donor sequences in the Covid-19 and influenza analyses, confirming that Mal-ID learned receptor sequence patterns used in the immune response to disease and vaccination. Disease category labels on individual sequences were not required to train these models. Additionally, the model architecture revealed which sequence categories contributed most to predictions of each disease—which V genes and isotypes were important building blocks for the BCRs and TCRs deployed by the immune system. We confirmed that V genes reported in prior literature carry high weight in the Mal-ID prediction process. This would be consistent with Mal-ID learning biologically meaningful sequence features rather than fitting to dataset-specific artifacts. Our analysis also highlighted several V genes as characteristic ones not previously associated with individual disease conditions, posing hypotheses that can be tested in future research. Unlike comparisons limited to patients with one disease versus healthy individuals, which may flag generic inflammatory responses, the multiclass modeling approach in this study can pinpoint immune responses specific to each disease type. With appropriate clinical validation, a model trained with the Mal-ID framework could be deployed either as an assay to distinguish several infectious and autoimmune diseases simultaneously, or as a diagnostic test for one particular disease. For translation of these results to clinical practice, acceptable sensitivity and specificity values will need to be determined based on the clinical context.

In this study, we emphasized the use of empirical data from a large cohort of patients with consistently collected IgH and TRB immune receptor sequencing data. Such data come with potential concerns about batch effects and confounders that we attempted to address. We used standardized receptor sequencing protocols and bioinformatic analysis for all samples and determined that models based on demographic covariates could not categorize patient immune status as accurately as IgH and TRB signatures. We withheld patient cohorts from the primary analysis and confirmed they were properly classified in a validation step. Performance on completely independent cohorts from other laboratories further showed that Mal-ID generalizes to independent data and does not fit to latent, unknown hidden variables.

The Mal-ID framework appeared to capture fundamental principles of immune responses and generalize to separate clinical cohorts. The task of differentiating Covid-19, HIV infection, lupus, T1D, and healthy was employed as a demonstration of the methodology's potential. Additional testing will be needed to establish appropriate cutoffs in clinical studies for sensitivity and specificity for particular diseases with diverse and variable prevalence, and further evaluate optimal sample volumes and sequencing depth. Any results from this methodology will need to be interpreted in light of other clinical assessment and laboratory testing of patients. Other important topics to address will be the potential for multiple conditions or comorbidities in the same patient, the development of models for different severities or subtypes of a particular disease, the value of using other kinds of lymphocyte-containing specimens such as tissue biopsies, and the possibility of identifying evidence for diseases not included in prior models, such as ones that may occur in future pandemics.

Materials and Methods

Modeling approach

We performed high-throughput immune receptor repertoire sequencing on peripheral blood RNA from 63 Covid-19, 95 chronic HIV-1, 86 SLE, and 92 T1D patients, along with 217 healthy controls and 37 influenza vaccination recipients. We did not consider other immunological conditions such as allergy in patient classification. Over 16 million B cell receptor heavy chain and 23 million T cell receptor beta chain clones were PCR-amplified with immunoglobulin and T cell receptor gene primers and sequenced as previously described (*13*, *58*). Each IgH isotype was amplified in a separate PCR reaction. We annotated V, D, and J gene segments with IgBLAST v1.3.0, keeping productive rearrangements only (*59*). Then we grouped nearly identical sequences within the same person into clones using single linkage clustering, as described previously (*13*). Using the clonal lineage groupings to deduplicate the dataset, we kept one copy of each clone per isotype, for each replicate of a sample from a patient. Among BCR sequences, we analyzed class-switched IgG or IgA isotype sequences, and non-class-switched IgD or IgM isotype sequences that were still antigen-experienced (with at least 1% somatic hypermutation).

We divided individuals into three stratified cross-validation folds, each split into a training set and a test set (fig. S2). Each individual was assigned to one test set. Some patients had multiple samples; all were grouped together for the cross-validation divisions. The splits were respected across the training of the complete Mal-ID pipeline. The architecture includes three base models, which are each trained for BCR and TCR data, and an ensemble model where all base models are combined.

Model 1: Overall repertoire composition

The first machine learning model uses an individual's IgH or TRB repertoire composition to predict disease status. Prior studies have reported immune status classification using deviations in B cell or T cell V(D)J recombination gene segment usage from healthy individuals (*16*, *60*). Certain V gene segments may be more prevalent among antigen-responding V(D)J rearrangements than in the population of immune receptors in naïve lymphocytes, and these gene segments increase in frequency as antigen-specific cells become clonally expanded (*47*, *61*), which can be seen in our data (fig. S7A). We previously identified class-switched IgH sequences with low somatic mutation (SHM) frequencies as prominent features of acute infection with Ebola virus or SARS-CoV-2, consistent with naïve B cells recently having class-switched during the primary response to infection (*47*, *61*). V gene usage changes and other repertoire changes have also been described in chronic infectious or immunological conditions (*8*, *13*). Therefore, we trained a logistic regression model with V/J gene counts, along with somatic hypermutation rate for IgH data, as features.

Model 2: Convergent clustering of antigen-specific sequences by edit distance

The second classifier detects highly similar CDR3 amino acid sequences shared between individuals with the same diagnosis, an approach we and others have previously reported (*12*–*15*). The CDR3s are the highly variable regions of IgH and TRB that often determine antigen binding specificity. For each locus, we clustered CDR3 sequences with the same V gene, J gene, and CDR3 length that had high sequence identity, allowing for some variability created by somatic hypermutation in B cell receptors. A new sample's sequences can then be assigned to nearby clusters with the same constraints. We selected clusters enriched for sequences from subjects with a particular disease, using Fisher's exact test and setting a significance threshold based on cross-validation with data derived from different individuals. The same significance threshold was used for all immune conditions tested. These clusters represent candidate sequences predictive of a specific disease across individuals. To score a new sample, we assigned its sequences to the identified predictive clusters. For each sample, we counted how many clusters associated with each disease were matched, and used these counts as features in a logistic regression model to predict immune status.

Model 3: Immune receptor sequence features extracted from a large language model

Small changes to immune receptor amino acid sequences can alter receptor structure and function, while different structures with divergent primary amino acid sequences can bind the same target epitope (*62*). We used a protein

language model, which transforms BCR and TCR amino acid sequences into a lower-dimensional representation, to estimate functional similarities between sequences that extend beyond sequence alignment. Specifically, we used ESM-2, a self-supervised model trained to predict masked amino acids from the remaining sequence context of a protein, learning complex statistical relationships between residues in each sequence and encoding functional and evolutionary relationships across sequences (*31*). Prior autoencoder models, which also convert immune receptor sequences to a latent representation, have enabled classification and clustering of functionally related sequences (*26*, *28*). However, ESM-2 is a large language model with substantially more parameters that is trained on a much larger compendium of over 65 million proteins across the tree of life, which allows it to learn richer latent representations that encode properties of a broad diversity of protein structures and functions (*31*). We developed machine learning models with a two-stage training strategy to predict patient-level disease status based on ESM-2-derived representations of their immune repertoire. First, we trained machine learning models to map ESM-2 derived 640-dimensional latent representations of each receptor sequence from each patient sample to a surrogate disease state corresponding to the disease state of the patient. Each model is specialized to one IGHV gene and isotype combination in the BCR case, or to one TRBV gene in the TCR case. Somatic hypermutation rate was used as an additional feature in the BCR case (hypermutation does not occur in TCRs). Then we trained a second-stage model that aggregates predicted probabilities of disease state of all sequences in a patient sample, again grouped by IGHV gene and isotype or by TRBV gene, to predict disease state at the patient level.

Ensemble of B and T cell models

Finally, we combined all three classifiers (overall repertoire composition, clustering by edit distance, and language model representation) for IgH and three for TRB into the final Mal-ID ensemble predictor of disease (fig. S1). As with the individual component models in Mal-ID, we trained a separate metamodel for each cross-validation group, maintaining strict separation of each individual's data into training, validation or test datasets.

B and T cell receptor repertoire sequencing

We assembled immune receptor repertoires from 63 Covid-19, 95 chronic HIV-1, 86 SLE, and 92 T1D patients, along with 217 healthy controls and 37 influenza vaccination recipients. Disease and demographic metadata are listed in table S1 in aggregate and for every individual. Venipuncture blood was collected in PAXgene Blood RNA Tubes or Tempus Blood RNA tubes, or isolated as PBMCs; the sample type is also enumerated in table S1. Ethics approvals for study of the sample sets were provided by Stanford University institutional review board (IRB) protocols #8629, #13952, #35453, #48973, #55650, and #55689; Oklahoma Medical Research Foundation IRBs #05-04, #06-12, #09-21, and #11-53; Providence St. Joseph Health IRB study number STUDY2020000175; University of Pennsylvania IRB #849398; and Duke University for the dataset previously deposited under SRA BioProject PRJNA486667. Informed consent was obtained from study participants. Most non-Covid-19 cohort samples were collected before the emergence of SARS-CoV-2, except for the influenza vaccine cohort and some of the diabetes cohort and associated healthy controls. Covid-19 samples were collected early in the pandemic. Among Covid-19 patients, we excluded mild cases, samples prior to seroconversion, and patients known to be immunosuppressed. These filters limited model training data to active disease samples to improve our chances of learning patterns for the disease-specific minority of receptor sequences. However, we wanted to avoid creating an artificially simple classification problem from filtering to trivially separable immune states. To this end, we included both treatment-naïve and treated SLE patients, and our HIV cohort included patients regardless of whether they generated broadly neutralizing antibodies to HIV. Had we instead restricted our analysis to HIV-infected individuals who produce broadly neutralizing antibodies, we may have created a more easily separable HIV class, due to the unusual characteristics of those antibodies (*13*).

Across these diverse immune states, over 16.2 million B and 23.5 million T cell receptor clones were sampled, PCR amplified with immunoglobulin and T cell receptor gene primers, and sequenced as previously described (*13*, *58*). Briefly, we amplified T cell receptor beta chains and each immunoglobulin heavy chain isotype in separate PCR reactions using random hexamer-primed cDNA templates, and performed paired-end Illumina MiSeq sequencing. To reduce the potential for batch effects, data collection followed a consistent protocol. Only IgH sequencing was performed for some older cohorts processed before the study was extended to include TRB sequencing. Paired-end reads were merged with FLASH (Fast Length Adjustment of SHort reads) v1.2.11. Samples were demultiplexed by matching barcodes to the sample reads, and the barcodes and primers were trimmed. We annotated V, D, and J gene segments and junctional bases with IgBLAST v1.3.0, keeping productive rearrangements only (*59*). Sequences with poor IGHV matches (IgBLAST IGHV segment alignment score less than 200) or poor TRBV matches (IgBLAST TRBV segment match alignment score less than 80) were removed. Using IgBLAST's identification of mutated nucleotides, we calculated the fraction of the IGHV gene segment that was mutated in any particular sequence; this is the somatic hypermutation rate (SHM) of a B cell receptor heavy chain. On the other hand, T cell receptors are known not to exhibit somatic hypermutation in humans. We also restricted our dataset to CDR-H3 and CDR3β segments with eight or more amino acids; otherwise the edit distance clustering method below might group short but unrelated sequences. Sequence data are deposited at the Sequence Read Archive under BioProject accession numbers PRJNA486667, PRJNA491287, and PRJNA1147802. Processed data are deposited on the Synapse platform at https://www.synapse.org/Synapse:syn61987835, both in Adaptive Immune Receptor Repertoire (AIRR) Rearrangement Schema format and in an internal format (*63*).

We grouped nearly identical sequences within the same person into clones, as described previously (*13*). To do so, for each individual, we grouped all nucleotide sequences from all samples (including samples at different timepoints) across all isotypes, and ran single-linkage hierarchical clustering to infer clonal lineages. This process iteratively merged sequence clusters from the same individual with matching IGHV/TRBV genes, IGHJ/TRBJ genes, and CDR-H3/CDR3β lengths, and with any cross-cluster pairs having at least 95% CDR3β sequence identity by string substitution distance, or at least 90% CDR-H3 identity, which allows for BCR somatic hypermutation (*13*).

We used the clonal lineage groupings to deduplicate the dataset. For each replicate of a sample from a patient, we kept one copy of each clone per isotype, choosing the sequence with the highest number of RNA reads. Similarly, we kept one copy of each TCRβ clone. Any replicates with fewer than 100 IgG, 100 IgA, and 500 IgD or IgM clones, or with fewer than 500 TRB clones, were rejected.

Among BCR sequences, we kept only class-switched IgG or IgA isotype sequences, and non-class-switched but still antigen-experienced IgD or IgM sequences with at least 1% SHM. By restricting the IgD and IgM isotypes to somatically hypermutated BCRs only, we ignored any unmutated cells that had not been stimulated by an antigen and were irrelevant for disease classification. The selected non-naïve IgD and IgM receptor sequences were combined into an IgM/D group.

On average, any two patients had 0.0003% IgH and 0.166% TRB sequence overlap, underscoring the enormous diversity of T cell receptor and especially B cell receptor sequences, as would be expected from random sequence generation by the V(D)J recombination process followed by additional BCR somatic hypermutation.

Cross-validation

We divided individuals into three stratified cross-validation folds, each split into a training set and a test set (fig. S2). Each individual was assigned

to one test set. Some patients had multiple samples; all were grouped together for the cross-validation divisions. The splits were respected across the training of the complete Mal-ID pipeline. Stratified cross-validation preserved the global imbalanced disease class distribution in each fold. We also carved out a validation set from each training set. What remained of the training set was further subdivided into two parts we call "train-1" and "train-2". The repertoire classification, CDR3 clustering, and language model base classifiers were trained on the training set and evaluated on the validation set. Then using the base models with highest validation set performance, the ensemble model was trained on the validation set, and then evaluated on the test set. In the case of multi-stage models like models 2 and 3, the sequence classification stage was fit on the train-1 set, then the patient level aggregation stage was fit on train-2. When we used logistic regression classification models, regularization hyperparameters were tuned with additional nested cross-validation. This training process happens separately for each fold; in other words, one collection of models is trained using fold 1's training, validation, and test sets, then a separate set of models is trained using fold 2's training, validation, and test sets, and so on. On average in any fold, we observed 0.05% of IgH and 5.3% of TRB sequences shared between any pair of the train, validation, and test sets.

Since any single repertoire contains many clonally related sequences, but is very distinct from other people's immune receptors, we made sure to place all sequences from an individual person into only the training, validation, or the test set, rather than dividing a patient's sequences across the three groups. Otherwise, the prediction strategies evaluated here could appear to perform better than they actually would on brand-new patients. Given the chance to see part of someone's repertoire in the training procedure, a prediction strategy would have an easier time of scoring other sequences from the same person in a held-out set. Had we not avoided this pitfall, models may also have been overfitted to the particularities of training patients. For the minority of individuals with multiple samples, we accordingly made sure that, in each cross-validation fold, all samples from the same person were grouped together into one of the training, validation, or test sets, as opposed to being spread across multiple sets. This principle was also respected for all nested cross-validation.

Finally, for the purpose of external cohort validation, we repeated the model training procedures with a "global" fold designed to incorporate all the data, by having only a training set and a validation set but no test set (fig. S2). Repertoires from independent external studies are used in place of the test set at evaluation time.

Evaluation metrics

Models were trained with the python-glmnet implementation of logistic regression (with multinomial loss and regularization strength tuned through cross-validation), as well as with the scikit-learn implementations of random forests (with 100 trees) and support vector machines (in "each class versus the rest" mode, with linear kernel and default regularization strength hyperparameter C=1.0). In all cases, we used prevalence-balanced class weights inversely proportional to input class frequencies. Predicted labels from all test sets were concatenated for global accuracy evaluation. Performance metrics that take predicted class probabilities as input, including AUROC and AUPRC, were computed separately for each fold, because probabilities may be on different scales in each fold and should not be combined into a global AUROC or AUPRC score. For overall performance, we report multiclass AUROC and AUPRC calculated in a one-versus-one fashion, taking the class size-weighted average of the binary AUROCs/AUPRCs calculated for each pair of classes, allowing each class a turn to be the positive class in the pair. For each disease class's individual performance, we report multiclass AUROC calculated in a one-versus-rest fashion. The AUROC and AUPRC measures do not reflect classification abstention, because abstained samples have no predicted class probabilities and cannot be included in the computation of metrics that use predicted probabilities. On the other hand, every abstention hurts label-based metrics like accuracy: each abstention counts as a prediction error. All analyses were performed and plotted with software versions python v3.9.17, numpy v1.24.3, pandas v1.5.3, scipy v1.11.1, scikit-learn v1.2.2, python-glmnet v2.2.1, pytorch v2.0.1, bio-transformers v0.1.17, matplotlib v3.7.1, and seaborn v0.12.2.

Model 1: Disease classifier using overall BCR or TCR repertoire composition features

For each sample, we created IgG, IgA, IgM/D, and TRB summary feature vectors by tallying IGHV/TRBV gene and IGHJ/TRBJ gene usage, counting each clone once. We ranked IGHV or TRBV genes by training set prevalence and excluded the bottom half, to avoid overfitting to minute differences in rare V gene proportions between cohorts. To account for different total clone counts across samples, we normalized total counts to sum to one per sample. Then we log-transformed and Z-scored (i.e., subtracted the mean and divided by the standard deviation, to achieve zero mean and unit variance) the matrix representing how counts are distributed across V-J gene pairs. Finally, we performed a PCA to reduce the count matrix to fifteen dimensions. All transformations were computed on each training set and applied to the corresponding validation and test sets. In addition, for each sample's subset of BCR sequences belonging to each isotype, we calculated the median sequence somatic hypermutation rate and the proportion of sequences that are somatically hypermutated (with at least 1% SHM). Only BCRs have somatic hypermutation, so we did not include mutation rate features of TCRs. In total, we arrived at 51 features across IgG, IgA, and IgM/D (fifteen count matrix principal components and two mutation rate features per isotype) for the IgH repertoire composition model, and 15 features for the TRB repertoire composition model.

We fit separate logistic regression linear models on the 51-dimensional (17 × 3 isotypes) BCR and 15-dimensional TCR feature vectors from each sample to predict disease. Features were standardized to zero mean and unit variance. We repeated this feature engineering and model training procedure on each cross-validation fold separately. The best performing models, according to average validation set AUROC across three cross-validation folds for the disease classification task on our primary dataset, were elastic net logistic regression with an L1/L2 regularization ratio of 0.25 for BCR and lasso, L1-regularized logistic regression for TCR.

Model 2: Disease classifier by clustering CDR-H3 sequences with edit distance

We performed single-linkage clustering on CDR3β sequences from T cells with identical TRBV genes, TRBJ genes, and CDR3β lengths, and separately on CDR-H3 sequences from B cells with identical IGHV genes, IGHJ genes, and CDR-H3 lengths, as described previously (*13*). Nearest-neighbor clusters were iteratively merged if any cross-cluster pairs had high sequence identity: at least 90% for CDR3β or 85% for CDR-H3, allowing for somatic hypermutation in B cells, as measured by string substitution distance (normalized Hamming distance). Clustering was performed on the train-1 data sets. This process was run separately for each cross-validation fold.

Filter to BCR and TCR disease-specific enriched clusters

For each sequence cluster found in the train-1 portion of a cross-validation fold's training set, we performed a Fisher's exact test using a two-by-two contingency table denoting how many unique people have a particular disease and have some receptor sequences fall into the cluster. In other words, each cluster's *P*-value from the Fisher's exact test denotes the cluster's enrichment for a particular disease. This approach is consistent with prior work that selects a set of disease-specific enriched sequences, then counts exact matches to this sequence set in new samples (*12*). Given a *P*-value threshold, the full list of training set clusters was filtered to clusters specific for each disease type. We performed all the following featurization and model fitting steps for *P*-values ranging from 0.0005 to 0.05, then selected the *P*-value that

led to the highest train-2 set performance as measured by the Matthews correlation coefficient (MCC) score, a classification performance metric that is well-suited to imbalanced datasets (*64*). The final chosen *P*-values differed depending on the cross-validation fold and the receptor type (i.e., BCR or TCR).

Compute BCR and TCR cluster membership feature vectors for each sample

For each selected enriched cluster, we created a cluster centroid: a single consensus sequence. Recall that each cluster member is a clone from which only the most abundant sequence was sampled. Rather than having each cluster member contribute equally to the consensus centroid sequence, contributions at each position were weighted by clone size, the number of unique BCR or TCR sequences originally part of each clone. Sequences from a sample were then matched to these predictive cluster centroids. In order to be assigned, a sequence must have the same IGHV/TRBV gene, IGHJ/TRBJ gene, and CDR-H3/CDR3β length as the candidate cluster, and must have at least 85% (BCR) or 90% (TCR) sequence identity with the consensus sequence representing the cluster's centroid. After assigning sequences to clusters, we counted cluster memberships across all sequences from each sample. Cluster membership counts were arranged as a feature vector for each sample: A sample's count for a particular disease was defined as the number of disease-enriched clusters into which some sequences from the sample were matched. This featurization captures the presence or absence of convergent T cell receptor or immunoglobulin sequences (separated by locus, but without regard for IgH isotypes).

Fit and evaluate model for each locus

Features were standardized, then used to fit separate BCR and TCR logistic regression models mapping from cluster counts to patient diagnosis. The models were fit on each train-2 set and evaluated on the corresponding validation set. The best performing models, according to average validation set AUROC across three cross-validation folds for the disease classification task on our primary dataset, were ridge logistic regression for BCR and lasso logistic regression for TCR.

We abstained from prediction if a sample had no sequences fall into a predictive cluster; this indicated no evidence was found for any particular class. Abstentions hurt accuracy and MCC scores, but were not included in the AUROC calculation, since no predicted class probabilities are available for abstained samples. Fewer than 3% of samples resulted in abstention (table S2).

Comparison to exact matches approach

Briefly, Emerson *et al.* classified cytomegalovirus (CMV) exposure by counting the number of TRB sequences that were exact matches to a CMV-associated list derived from a training set of CMV+ and CMV- individuals (*12*). CMV-associated sequences were determined with a Fisher's exact test using a two-by-two contingency table denoting how many unique people are CMV+ and have a particular sequence; the threshold on Fisher's exact test *P*-values was selected by cross-validation.

We reimplemented this method for the Mal-ID dataset to compare the "exact sequence matches" featurization of Emerson *et al.* against the "fuzzy matches" featurization of the CDR3 clustering component of Mal-ID. The binary classification generative model used in Emerson *et al.* after the featurization step does not translate to our multiclass disease classification problem, so we instead used the same classification framework as the CDR3 clustering model: Each sample's feature vector consisted of the number of disease-specific hits for each disease, normalized by the total size of the sample. Additionally, we ensured that both models had a consistent approach to abstention. The CDR3 clustering model abstains on samples that had zero matches to any disease-associated cluster; similarly, our implementation of Emerson *et al.* in the multiclass problem abstains on samples that had zero matches to any disease-associated sequence (i.e., there is no evidence of disease). Just as when training the CDR3 clustering model, the exact matches featurization and model fits were performed for different *P*-value thresholds, then the best threshold was chosen by optimizing performance on the second part of the training set (train-2) using the MCC score. Therefore, the Emerson *et al.* and CDR3 clustering models are trained the same way in this comparison, differing only in whether the featurization step finds exact sequence matches or fuzzy matches.

Model 3: Disease classifier using language model embeddings

The analysis pipeline for classifying disease with language model embeddings of sequences is complex, but necessarily so because it aggregates individual sequence data to generate patient-level predictions.

Generate embeddings

We embedded the CDR-H3/CDR3β segments of each receptor sequence with the 30-layer, 150-million-parameter ESM-2 neural network (*31*), using the bio-transformers v0.1.17 implementation. A final 640-dimensional vector representation was calculated by averaging ESM-2's hidden state over the original protein's length dimension.

Train sequence-level disease classifier for each sequence category

First, we trained classification models to map sequences to disease labels—one model per fold and per sequence category, defined as an IGHV gene and isotype pair for BCR sequences or a TRBV gene for TCR sequences. As input data, we used ESM-2 embeddings (standardized to zero mean and unit variance), along with somatic hypermutation rate in the BCR case. To train the individual-sequence-level model, we labeled each sequence with the patient's immune status or disease category. These labels should be considered noisy: We do not know which of a patient's sequences are truly associated with their disease. Since we have no true sequence labels, we also cannot evaluate classification performance for the sequence-level classifier directly. These sequence-level classifiers were trained on the train-1 set of each cross-validation fold.

Aggregate sequence predictions within each sequence category

We combined predictions for individual BCR or TCR sequences into a patient sample-level prediction by the following procedure. Given a sample with n BCR (or TCR) sequences, we first scored each sequence with the corresponding sequence model. For example, we applied the IGHV3-53, IgG model to input sequences arising from the IGHV3-53 gene segment and the IgG isotype. Each sequence now has a vector of k predicted probabilities, with one value for each of the k disease classes. These values are only comparable between sequences that were scored by the same model, as models for different sequence groups are not guaranteed to have matching calibration. Therefore, we next aggregated predicted class probabilities among sequences from the same sequence category, one IGHV gene and isotype (or one TRBV gene) at a time. To calculate the aggregate probability for each of the k classes, we used one of the following methods:

- Mean
- Median
- Trimmed mean: Remove the lowest 10% of sequence-level probabilities before calculating the mean.
- Entropy thresholded mean: Before taking the mean, remove any sequences whose predicted class probability vectors had high entropy, indicating they carry little information that could indicate a particular disease class. A sequence with probabilities of $1/k$ for all k classes would have the highest possible entropy. We removed sequences whose entropy was within either 10% or 20% of this maximal value.

This procedure gives the final k-dimensional predicted disease class probabilities vector for each sequence category in each sample. For example, it computes P(Covid19) among IGHV1-24/IgG sequences, P(HIV) among IGHV1-24/IgG sequences, and so on; then similarly P(Covid19) among IGHV3-53/IgA sequences, P(HIV) among IGHV3-53/IgA sequences, and so forth.

Map from aggregate predictions for each sequence category to a sample prediction

Using the aggregated sequence-level predictions, we make a final prediction for the sample with a

second-stage model. This model was fitted in a one-versus-rest fashion, and the submodel for each class was trained only with features corresponding to that class. For example, the Covid-19-vs-rest model was provided P(Covid-19) in IGHV1-24/IgG, P(Covid-19) in IGHV3-53/IgG, and so on, but not P(HIV), P(Influenza), P(Lupus), P(T1D), or P(Healthy). This design prohibits unwanted feature leakage: deciding whether a sample is from a Covid-19 patient should rely only on sequence-level probabilities for the Covid-19 class, not any other classes. Also, we incorporated features for only the top 50% of IGHV or TRBV genes to avoid having far more features than samples for this second-stage model, and because rare V genes may not be present in all samples. Therefore, the number of features in this second-stage model for the BCR case was half the number of IGHV genes, times three isotype categories: IgG, IgA, and IgM/D excluding naïve B cells with <1% somatic hypermutation. For TCR, which has no isotype subdivisions, the number of features was half the number of TRBV genes. Each sample's features were reweighed according to sequence category frequencies. In the BCR case, frequencies were computed separately for each isotype to account for technical variation in isotype frequencies between sequencing runs. The aggregation model was trained on the train-2 set in each cross-validation fold.

Evaluate classifier

We evaluated the pipeline by computing sample-level classification performance on the validation set using AUROC scores. (The one-versus-rest model predicted probabilities are not necessarily calibrated against each other, so we did not evaluate accuracy or other metrics determined by the comparison of predicted class probabilities for selecting a winning label). For the BCR case, the highest validation set performance on our primary dataset was achieved by a pipeline consisting of random forest sequence-level models, followed by a random forest second-stage model using mean aggregation. In the TCR case, the best pipeline used one-versus-rest ridge logistic regression sequence-level models, with a random forest second-stage model using mean aggregation after an entropy cutoff at 20% below the maximal entropy value (table S8). To evaluate feature contributions to predictions of each disease class, we ran Tree SHAP on each one-class-versus-rest random forest aggregation model, and averaged the SHAP feature importance values across positive class instances from the train-2 data used to train the aggregation model. SHAP values were rescaled from 0 to 1. Alternatively, to find SHAP clusters, we performed Louvain clustering (resolution 1.0) on the full SHAP value matrix in which rows represent positive class examples and columns represent features, then calculated average SHAP values within each cluster.

Ensemble metamodel

After training repertoire composition, CDR3 clustering, and language model embedding models on each fold's training set, we combined the classifiers with an ensemble strategy. We used the base model versions with highest validation set performance; different base model versions performed best on the validation sets in our primary dataset compared to when Mal-ID was retrained on other datasets, such as the Adaptive Biotechnologies genomic DNA data. For each fold, we ran all trained base classifiers on the validation set, and concatenated the resulting predicted class probability vectors from each base model. We carried over any sample abstentions from the CDR3 clustering model (the other models do not abstain). Finally, we trained a ridge logistic regression classification metamodel to map the combined predicted probability vectors to validation set sample disease labels. We evaluated this metamodel on the held-out test set. To evaluate individual model component contributions, we refit the metamodel with subsets of features, such as only those features derived from models 1 and 2.

Batch effect evaluation using language model embeddings

Having integrated many datasets in this study, we sought to test whether our disease classification performance was driven by technical differences between batches of library preparation or sequencing instrument run. It would be expected in any study of human cohorts to identify some batch effects, given the difficulty of collecting identical samples in identical manner, at identical severity and timepoints, from patients suffering from diseases that appear in different populations at different frequencies. Notably, the IgH data collected for individual participants in this study were typically based on multiple Illumina MiSeq sequencer runs and were combined prior to analysis. Many of our sequencing run batches included only one disease type, but batches that included both diseased and healthy controls from the same population permitted accurate classification of the disease or healthy state, for example, with classification of HIV-infected patients and healthy controls that were sequenced together in the same batch, or SLE patients and healthy controls sequenced in the same batch.

Acknowledging that there were biological differences between many sequencing batches that were enriched for a particular disease state, and that several sequencer runs were performed for some sample sets, we evaluated the potential impact of these batch differences using the language model embeddings of BCR and TCR repertoires from the disease types found in multiple batches: Covid-19 patients, SLE patients, and healthy donors. We applied the *k*-nearest-neighbor batch effect test (kBET) metric from the single cell sequencing literature (*65*). kBET measures whether cells from many batches are well-mixed by comparing the batch label distribution among each cell's neighbors to the global distribution. In place of cells described by gene expression vectors, we have sequences described by language model embedding features. We measured kBET for every disease in every test set fold and in both BCR and TCR data. For example, we constructed a k-nearest neighbors graph (k = 50) with all BCR sequences from Covid-19 patients in test fold 1. We performed chi-squared tests for the difference between the batch label distribution among each sequence's 50 nearest neighbors and the expected distribution from the total number of sequences belonging to each batch in the entire graph. After multiple hypothesis correction with a significance threshold of $P = 0.05$, we measured the number of sequences for which we could reject the null hypothesis that the local neighborhood batch distribution is the same as the global batch distribution. Aggregating these results by disease across gene loci and folds, we see that the null hypothesis is rejected for only 18.2% of sequences on average, suggesting that the sequence data in the graph are well mixed according to batch (table S9). The average rejection rate is higher for Covid-19 BCR sequences at 44.1%, which may be influenced by disease severity differences between cohorts (table S1). Time point differences between batches may also influence kBET metrics for acute diseases like Covid-19. At earlier time points, Covid-19 patient repertoires may include more healthy background sequences, leading to a different batch overlap graph in comparison to how batches compare after clonal expansion of Covid-19 responding sequences. Overall, these results suggest that most sequences have well-mixed batch proportions amongst their nearest neighbors.

Validation on external cohorts

The best test of whether our model has learned true biological signal as opposed to batch effects is whether our model generalizes to unseen data from other cohorts. For the purposes of evaluating external cohorts, rather than using models trained on our cross-validation divisions of the data, we trained a set of "global" models incorporating all Mal-ID data without holding out a test set (fig. S2). To train the ensemble metamodel, we still held out a validation set, with a ratio of training set to validation set size equivalent to the ratio used in the cross-validation regime.

We downloaded data from other BCR and TCR Covid-19 patient and healthy donor repertoire studies with cDNA sequencing (*36*–*40*, *66*). Among acute Covid-19 cases, we selected active disease timepoint samples at least two weeks after symptom onset, after which time we would expect seroconversion (*47*). We reprocessed sequences through the same version of IgBLAST

and IgBLAST reference data used for the primary Mal-ID cohorts, to ensure consistent gene nomenclature. This was not possible for the Britanova *et al.* datasets (*39*, *40*) because the raw sequences were unavailable, so we used their gene calls and confirmed the naming was consistent with our training data, especially for indistinguishable TRBV genes TRBV6-2/6-3 and TRBV12-3/12-4. We embedded productive CDR3 sequences with the language model, then processed the downloaded repertoires through the entire Mal-ID model architecture. We also tuned class decision thresholds to adapt the model to the new base rates of disease in the data. Specifically, we held out several external cohort samples and reweighted their predicted class probabilities to optimize the MCC score. After this procedure, the winning label for each sample is chosen based on the class with highest predicted probability after class weights are applied. If a class had its probabilities reweighted by 1/5, for example, the model must be five times more confident to choose that class label. This procedure affected only the confusion matrix, accuracy, and other metrics based on predicted labels.

Additionally, we retrained Mal-ID after downloading TCR repertoire data collected with the Adaptive Biotechnologies genomic DNA sequencing protocol (table S5). This data was reprocessed with the same IgBLAST version as above, for consistency.

Predicting demographic information from healthy subject repertoires

We repeated the model training process to predict age, sex, or ancestry instead of disease. Input data was limited to healthy controls to avoid learning any disease-specific patterns. To cast this as a classification problem, age was discretized either into deciles, as a binary "under 50 years old" / "50 or older" variable, or as a binary "under 18 years old" / "18 or older" variable. Only one healthy control individual was over 80 years old, therefore our data do not assess repertoire changes at more extreme older ages. We excluded the healthy individual over 80 years old from the analysis.

For each of the demographic prediction tasks, we trained the full BCR+TCR Mal-ID architecture on all cross-validation folds. We note that we did not explicitly introduce data from allelic variant typing in germline IGHV, IGHD, or IGHJ gene segments or in HLA genes into our models, but such data could be expected to increase detection of ancestry in such datasets.

Evaluating predictive power of potential demographic confounding variables

We retrained the entire Mal-ID disease-prediction set of models on the subset of individuals with known age, sex, and ancestry (as above, we excluded any individuals over 80 years old). Additionally, we regressed out those demographic variables from the feature matrix used as input to the ensemble step. Specifically, we fit a linear regression for each column of the feature matrix, to predict the column's values from age, sex, and ancestry. The feature matrix column was then replaced by the fitted model's residuals. This procedure orthogonalizes or decorrelates the metamodel's feature matrix from age, sex, and ancestry effects. We regressed out covariates at the metamodel stage because it is a sample-level, not sequence-level model, and age/sex/ancestry demographic information is tied to samples rather than sequences.

Separately, we also trained models to predict disease from either age, sex, or ancestry information encoded as categorical dummy variables. Here, no sequence information was provided as input. Finally, we trained metamodels with both demographic features and sequence features, along with interaction terms between the demographic and sequence features to allow for interaction effects. Comparing the performance of these models to the demographics-only models shows the added value of adding sequence information.

Model ranking of known antigen-specific sequences

We downloaded the 13 June 2023 version of CoV-AbDab (*50*), and reprocessed these B cell receptor heavy chain sequences through the same version of IgBLAST used for our primary cohorts to ensure consistent V gene nomenclature. However, CoV-AbDab contains amino acid sequences, rather than nucleotide sequences as in our internal data, so we used the protein version of IgBLAST ("igblastp") and quantified somatic hypermutation based on the percentage of mutated amino acids. We filtered to antibody sequences known to bind to SARS-CoV-2 (including weak binders, but excluding sequences shown to selectively bind certain viral variants but not others), and only kept sequences from human patients or vaccinees. We clustered the selected SARS-CoV-2 binders with identical IGHV gene, IGHJ gene, and CDR-H3 lengths and at least 95% sequence identity, using single linkage clustering as in the pipeline for our primary cohorts. As a result, several related sequences were combined and replaced by a consensus sequence. This preprocessing was repeated for influenza-specific antibody sequences from human patients and vaccinees (*53*), excluding H5N1 and H7N9 vaccine or infection data because those strains are not included in the seasonal flu vaccine that our classifier was trained to distinguish.

Similarly, we downloaded the ImmuneCode MIRA database (*54*), version 002.1, and reprocessed these T cell receptor beta chain sequences with our pipeline's standard IgBLAST version for consistent V gene nomenclature. As above, we filtered to productive sequences from patients with acute Covid-19, and also to only the TRBV genes present in our dataset, as any others would not be compatible with the sequence model, which uses V gene segment identity as a feature. Among the remaining SARS-CoV-2 associated sequences, we deduplicated those with identical TRBV genes, TRBJ genes, and CDR3β sequences.

We scored the external databases of known binder sequences using models 2 and 3 trained on the global fold. Isotype designations were not available in the BCR antigen-specific datasets; we applied our IgG sequence models because many antigen-specific B cells in Covid-19 have been reported to express IgG (*47*, *48*, *67*). Correspondingly, we compared to IgG sequences from healthy donors in the global fold's validation set, which were held out from training. To perform the statistical test shown for a particular V gene (e.g., IGHV1-24 for the Covid-19 analysis), we conducted a one-sided permutation test to assess whether known binder sequences had higher model 3 predicted Covid-19 class probabilities compared to sequences from healthy individuals. The permutation test ensured that all sequences originating from each healthy donor individual retained their grouping (i.e., had consistent binder/non-binder labels) throughout the process of performing 1000 label permutations. Since the known binders have low prevalence and since permutation affects the prevalence, we computed the AUPRC fold change over baseline prevalence in each permutation, then calculated the p-value as the proportion of permutations whose AUPRC fold change was greater than the observed AUPRC fold change in the original data.

REFERENCES AND NOTES

1. C. L. Charlton *et al.*, Practical Guidance for Clinical Microbiology Laboratories: Viruses Causing Acute Respiratory Tract Infections. *Clin. Microbiol. Rev.* **32**, CMR.00042-18 (2018). doi: 10.1128/CMR.00042-18; pmid: 30541871
2. R. Milo, A. Miller, Revised diagnostic criteria of multiple sclerosis. *Autoimmun. Rev.* **13**, 518–524 (2014). doi: 10.1016/j.autrev.2014.01.012; pmid: 24424194
3. A. Kavanaugh, R. Tomar, J. Reveille, D. H. Solomon, H. A. Homburger, American College of Pathologists, Guidelines for clinical use of the antinuclear antibody test and tests for specific autoantibodies to nuclear antigens. *Arch. Pathol. Lab. Med.* **124**, 71–81 (2000). doi: 10.5858/2000-124-0071-GFCUOT; pmid: 10629135
4. S. C. A. Nielsen, S. D. Boyd, Human adaptive immune receptor repertoire analysis-Past, present, and future. *Immunol. Rev.* **284**, 9–23 (2018). doi: 10.1111/imr.12667; pmid: 29944765
5. R. A. Arnaout, E. T. L. Prak, N. Schwab, F. Rubelt, Adaptive Immune Receptor Repertoire Community, The Future of Blood Testing Is the Immunome. *Front. Immunol.* **12**, 626793 (2021). doi: 10.3389/fimmu.2021.626793; pmid: 33790897
6. J. J. M. van Dongen *et al.*, Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: Report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* **17**, 2257–2317 (2003). doi: 10.1038/sj.leu.2403202; pmid: 14671650
7. T. Ching *et al.*, Analytical evaluation of the clonoSEQ Assay for establishing measurable (minimal) residual disease in acute lymphoblastic leukemia, chronic lymphocytic leukemia, and multiple myeloma. *BMC Cancer* **20**, 612 (2020). doi: 10.1186/s12885-020-07077-9; pmid: 32605647
8. R. J. M. Bashford-Rogers *et al.*, Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature* **574**,

122–126 (2019). doi: 10.1038/s41586-019-1595-3; pmid: 31554970
9. V. Greiff, G. Yaari, L. G. Cowell, Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Curr. Opin. Syst. Biol.* **24**, 109–119 (2020). doi: 10.1016/j.coisb.2020.10.010
10. S. D. Boyd, J. E. Crowe Jr., Deep sequencing and human antibody repertoire analysis. *Curr. Opin. Immunol.* **40**, 103–109 (2016). doi: 10.1016/j.coi.2016.03.008; pmid: 27065089
11. P. Barennes *et al*., Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. *Nat. Biotechnol.* **39**, 236–245 (2021). doi: 10.1038/s41587-020-0656-3; pmid: 32895550
12. R. O. Emerson *et al*., Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.* **49**, 659–665 (2017). doi: 10.1038/ng.3822; pmid: 28369038
13. K. M. Roskin *et al*., Aberrant B cell repertoire selection associated with HIV neutralizing antibody breadth. *Nat. Immunol.* **21**, 199–209 (2020). doi: 10.1038/s41590-019-0581-0; pmid: 31959979
14. P. Dash *et al*., Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017). doi: 10.1038/nature22383; pmid: 28636592
15. J. Glanville *et al*., Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017). doi: 10.1038/nature22976; pmid: 28636589
16. X. Liu *et al*., T cell receptor β repertoires as novel diagnostic markers for systemic lupus erythematosus and rheumatoid arthritis. *Ann. Rheum. Dis.* **78**, 1070–1078 (2019). doi: 10.1136/annrheumdis-2019-215442; pmid: 31101603
17. W. D. Chronister *et al*., TCRMatch: Predicting T-Cell Receptor Specificity Based on Sequence Similarity to Previously Characterized Receptors. *Front. Immunol.* **12**, 640725 (2021). doi: 10.3389/fimmu.2021.640725; pmid: 33777034
18. S. Eliyahu *et al*., Antibody repertoire analysis of hepatitis C virus infections identifies immune signatures associated with spontaneous clearance. *Front. Immunol.* **9**, 3004 (2018). doi: 10.3389/fimmu.2018.03004; pmid: 30622532
19. M. Safra *et al*., A somatic hypermutation-based machine learning model stratifies individuals with Crohn's disease and controls. *Genome Res.* **33**, 71–79 (2023). doi: 10.1101/gr.276683.122; pmid: 36526432
20. D. H. May *et al*., Identifying immune signatures of common exposures through co-occurrence of T-cell receptors in tens of thousands of donors, bioRxiv 2024.03.26.583354 [Preprint] (2024); doi: 10.1101/2024.03.26.583354
21. J. Ostmeyer, S. Christley, I. T. Toby, L. G. Cowell, Biophysicochemical Motifs in T-cell Receptor Sequences Distinguish Repertoires from Tumor-Infiltrating Lymphocyte and Adjacent Healthy Tissue. *Cancer Res.* **79**, 1671–1680 (2019). doi: 10.1158/0008-5472.CAN-18-2292; pmid: 30622114
22. J. Leem, L. S. Mitchell, J. H. R. Farmery, J. Barton, J. D. Galson, Deciphering the language of antibodies using self-supervised learning. *Patterns* **3**, 100513 (2022). doi: 10.1016/j.patter.2022.100513; pmid: 35845836
23. J. A. Ruffolo, J. J. Gray, J. Sulam, "Deciphering antibody affinity maturation with language models and weakly supervised learning" in *Machine Learning for Structural Biology Workshop* (NeurIPS, 2021).
24. T. H. Olsen, I. H. Moal, C. M. Deane, AbLang: An antibody language model for completing antibody sequences. *Bioinform. Adv.* **2**, vbac046 (2022). doi: 10.1093/bioadv/vbac046; pmid: 36699403
25. D. Prihoda *et al*., BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *MAbs* **14**, 2020203 (2022). doi: 10.1080/19420862.2021.2020203; pmid: 35133949
26. J.-W. Sidhom, H. B. Larman, D. M. Pardoll, A. S. Baras, DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nat. Commun.* **12**, 1605 (2021). doi: 10.1038/s41467-021-21879-w; pmid: 33707415
27. M. Widrich *et al*., Modern Hopfield Networks and Attention for Immune Repertoire Classification. *Adv. Neural Inf. Process. Syst.* **33**, 18832–18845 (2020). doi: 10.1101/2020.04.12.038158
28. S. Friedensohn *et al*., Convergent selection in antibody repertoires is revealed by deep learning, bioRxiv 2020.02.25.965673 [Preprint] (2020); doi: 10.1101/2020.02.25.965673
29. S. Dvorkin, R. Levi, Y. Louzoun, Autoencoder based local T cell repertoire density can be used to classify samples and T cell receptors. *PLOS Comput. Biol.* **17**, e1009225 (2021). doi: 10.1371/journal.pcbi.1009225; pmid: 34310600
30. Z. Sethna *et al*., Population variability in the generation and selection of T-cell repertoires. *PLOS Comput. Biol.* **16**, e1008394 (2020). doi: 10.1371/journal.pcbi.1008394; pmid: 33296360
31. Z. Lin *et al*., Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023). doi: 10.1126/science.ade2574; pmid: 36927031
32. O. Sagi, L. Rokach, Ensemble learning: A survey. *WIRES* **8**, e1249 (2018). doi: 10.1002/widm.1249
33. D. J. Hand, R. J. Till, A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Mach. Learn.* **45**, 171–186 (2001). doi: 10.1023/A:1010920819831
34. R. J. Cox *et al*., An early humoral immune response in peripheral blood following parenteral inactivated influenza vaccination. *Vaccine* **12**, 993–999 (1994). doi: 10.1016/0264-410X(94)90334-4; pmid: 7975853
35. M. Petri *et al*., Combined oral contraceptives in women with systemic lupus erythematosus. *N. Engl. J. Med.* **353**, 2550–2558 (2005). doi: 10.1056/NEJMoa051135; pmid: 16354891
36. S. I. Kim *et al*., Stereotypic neutralizing V_H antibodies against SARS-CoV-2 spike protein receptor binding domain in patients with COVID-19 and healthy individuals. *Sci. Transl. Med.* **13**, eabd6990 (2021). doi: 10.1126/scitranslmed.abd6990; pmid: 33397677
37. B. Briney, A. Inderbitzin, C. Joyce, D. R. Burton, Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566**, 393–397 (2019). doi: 10.1038/s41586-019-0879-y; pmid: 30664748
38. A. S. Shomuradova *et al*., SARS-CoV-2 Epitopes Are Recognized by a Public and Diverse Repertoire of Human T Cell Receptors. *Immunity* **53**, 1245–1257.e5 (2020). doi: 10.1016/j.immuni.2020.11.004; pmid: 33326767
39. O. V. Britanova *et al*., Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J. Immunol.* **192**, 2689–2698 (2014). doi: 10.4049/jimmunol.1302064; pmid: 24510963
40. O. V. Britanova *et al*., Dynamics of Individual T Cell Repertoires: From Cord Blood to Centenarians. *J. Immunol.* **196**, 5005–5013 (2016). doi: 10.4049/jimmunol.1600005; pmid: 27183615
41. C. T. Watson *et al*., Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am. J. Hum. Genet.* **92**, 530–546 (2013). doi: 10.1016/j.ajhg.2013.03.004; pmid: 23541343
42. A. Alpert *et al*., A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring. *Nat. Med.* **25**, 487–495 (2019). doi: 10.1038/s41591-019-0381-y; pmid: 30842675
43. K. M. Gostic, M. Ambrose, M. Worobey, J. O. Lloyd-Smith, Potent protection against H5N1 and H7N9 influenza via childhood hemagglutinin imprinting. *Science* **354**, 722–726 (2016). doi: 10.1126/science.aag1322; pmid: 27846599
44. J. J. Goronzy, C. M. Weyand, Immune aging and autoimmunity. *Cell. Mol. Life Sci.* **69**, 1615–1623 (2012). doi: 10.1007/s00018-012-0970-0; pmid: 22466672
45. C. E. Weckerle, T. B. Niewold, The unexplained female predominance of systemic lupus erythematosus: Clues from genetic and cytokine studies. *Clin. Rev. Allergy Immunol.* **40**, 42–49 (2011). doi: 10.1007/s12016-009-8192-4; pmid: 20063186
46. S. M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4765–4774 (2017).
47. S. C. A. Nielsen *et al*., Cell Clonal Expansion and Convergent Antibody Responses to SARS-CoV-2. *Cell Host Microbe* **28**, 516–525.e5 (2020). doi: 10.1016/j.chom.2020.09.002; pmid: 32941787
48. J. M. Dan *et al*., Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. *Science* **371**, eabf4063 (2021). doi: 10.1126/science.abf4063; pmid: 33408181
49. H. R. Waterman *et al*., Lupus IgA1 autoantibodies synergize with IgG to enhance plasmacytoid dendritic cell responses to RNA-containing immune complexes. *Sci. Transl. Med.* **16**, eadl3848 (2024). doi: 10.1126/scitranslmed.adl3848; pmid: 38959329
50. M. I. J. Raybould, A. Kovaltsuk, C. Marks, C. M. Deane, CoV-AbDab: The coronavirus antibody database. *Bioinformatics* **37**, 734–735 (2021). doi: 10.1093/bioinformatics/btaa739; pmid: 32805021
51. C. Kreer *et al*., Longitudinal Isolation of Potent Near-Germline SARS-CoV-2-Neutralizing Antibodies from COVID-19 Patients. *Cell* **182**, 1663–1673 (2020). doi: 10.1016/j.cell.2020.08.046; pmid: 32946786
52. J. Braun *et al*., SARS-CoV-2-reactive T cells in healthy donors and patients with COVID-19. *Nature* **587**, 270–274 (2020). doi: 10.1038/s41586-020-2598-9; pmid: 32726801
53. Y. Wang *et al*., An explainable language model for antibody specificity prediction using curated influenza hemagglutinin antibodies. *Immunity* **57**, 2453–2465.e7 (2024). doi: 10.1016/j.immuni.2024.07.022; pmid: 39163866
54. S. Nolan *et al*., A large-scale database of T-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. Res Sq, rs.3.rs-51964 [Preprint] (2020); doi: 10.21203/rs.3.rs-51964/v1
55. D. H. Wolpert, Stacked generalization. *Neural Netw.* **5**, 241–259 (1992). doi: 10.1016/S0893-6080(05)80023-1
56. J.-F. Bach, Insulin-dependent diabetes mellitus as an autoimmune disease. *Endocr. Rev.* **15**, 516–542 (1994). doi: 10.1210/edrv-15-4-516; pmid: 7988484
57. A. Suárez-Fueyo, S. J. Bradley, G. C. Tsokos, T cells in Systemic Lupus Erythematosus. *Curr. Opin. Immunol.* **43**, 32–38 (2016). doi: 10.1016/j.coi.2016.09.001; pmid: 27636649
58. Q. Qi *et al*., Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 13139–13144 (2014). doi: 10.1073/pnas.1409155111; pmid: 25157137
59. J. Ye, N. Ma, T. L. Madden, J. M. Ostell, IgBLAST: An immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, W34–40 (2013). doi: 10.1093/nar/gkt382; pmid: 23671333
60. A. M. Sevy, C. Soto, R. G. Bombardi, J. Meiler, J. E. Crowe Jr., Immune repertoire fingerprinting by principal component analysis reveals shared features in subject groups with common exposures. *BMC Bioinformatics* **20**, 629 (2019). doi: 10.1186/s12859-019-3281-8; pmid: 31801472
61. C. W. Davis *et al*., Longitudinal Analysis of the Human B Cell Response to Ebola Virus Infection. *Cell* **177**, 1566–1582.e17 (2019). doi: 10.1016/j.cell.2019.04.036; pmid: 31104840
62. C. Marks, C. M. Deane, How repertoire data are changing antibody science. *J. Biol. Chem.* **295**, 9823–9837 (2020). doi: 10.1074/jbc.REV120.010181; pmid: 32409582
63. S. Boyd, Mal-ID (Synapse, 2024); https://doi.org/10.7303/SYN61987835.
64. D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020). doi: 10.1186/s12864-019-6413-7; pmid: 31898477
65. M. Büttner, Z. Miao, F. A. Wolf, S. A. Teichmann, F. J. Theis, A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019). doi: 10.1038/s41592-018-0254-1; pmid: 30573817
66. B. D. Corrie *et al*., iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol. Rev.* **284**, 24–41 (2018). doi: 10.1111/imr.12666; pmid: 29944754
67. D. Mathew *et al*., UPenn COVID Processing Unit, Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. *Science* **369**, eabc8511 (2020). doi: 10.1126/science.abc8511; pmid: 32669297
68. M. Zaslavsky, Maximz/Malid: August 2024 Release, Zenodo (2024); https://zenodo.org/records/13357613.

ACKNOWLEDGMENTS

We thank A. Balsubramani and members of the Kundaje and Boyd labs for helpful discussions. We also thank Stanford Health Care Clinical Virology Laboratory members F. Yamamoto, M. K. Sahoo, C. H. Huang, and D. Solis, as well as H. Maecker and the Stanford Human Immune Monitoring Center. We also thank the Stanford Covid-19 Biobank Study Group's members: E. J. Zudock, M. M. Hashemi, K. C. Tjandra, J. A. Newberry, J. V. Quinn, R. Mann, A. Visweswaran, T. Ranganath, J. Roque, M. Manohar, H. N. Din, K. Kumar, K. Jee, B. Noon, J. Anderson, B. Fay, D. Schreiber, N. Zhao, R. Vergara, J. McKechnie, A. Wilk, L. de la Parte, K. W. Dantzler, M. Ty, N. Kathale, A. Rustagi, G. Martinez-Colon, G. Ivison, R. Pi, M. Lee, R. Brewer, T. Hollis, A. Baird, M. Ugur, D. Bogusch, G. Nahass, K. Haider, K. Q. T. Tran, L. Simpson, M. Tal, I. Chang, E. Do, A. Fernandes, A. Lee, N. Ahuja, T. Snow, and J. Krempski. Icons were used under Creative Commons licenses: "Sample" by G. Cresnar and "Box Icon" by F. Hafizd from TheNounProject.com (CC-BY 3.0); "Blood Sample" by M. Tisch and "Classification" by S. Dürr from BioIcons.com (CC0); and "Illumina MiSeq" by DBCLS from BioIcons.com (CC-BY 4.0). **Funding:** S.D.B. was partially supported by NIH/NIAID grants R01AI130398, R01AI127877, U19AI057229, U54CA260518, and U19AI167903, and a philanthropic gift from an anonymous donor. M.E.Z. was supported

by the National Science Foundation Graduate Research Fellowship and the Stanford Bio-X Bowes Graduate Student Fellowship. E.C. was supported by the Stanford Graduate Fellowship and the Stanford Data Science Scholarship. R.T. was supported by the National Institutes of Health (NIH 5R01 EB001988-16) and the National Science Foundation (NSF 19 DMS1208164). B.F.H and M.A.M. were supported by the NIH, National Institute of Allergy and Infectious Disease (NIAID), Division of AIDS Center for HIV/AIDS Vaccine Immunology-Immunogen Discovery (UM-1 AI100645) and the Consortia for HIV/AIDS Vaccine Development (UM1 AI144371). C.C.R. was supported by the National Institutes of Health, AI 101093, AI-086037, AI-48693, and the David S Gottesman Immunology Chair. A.K. was partially supported by the Stanford School of Medicine COVID-19 Research Fund. S.Y. was supported by NIH/NIAID grants R01AI153133, R01AI137272, and 3U19AI057229–17W1 COVID SUPP 2 and a philanthropic gift from Eva Grove. C.A.B. was supported by the Burroughs Wellcome Fund Investigators in the Pathogenesis of Infectious Diseases 1016687 and U19 AI057229. S.R.M., W.D., J.M.G., J.T.M., and J.A.J. were partially supported by NIH/NIAMS AR073750 and NIH/NIAID UM1AI144292. K.J. and E.M. were supported by NIH grant NIDDK P30DK116074. K.C.N. was supported by the National Institutes of Health, U54CA260518, U19AI167903, the Sunshine Foundation, and the John Rock Professor Chair at Harvard T.H. Chan School of Public Health. P.J.U. was funded by Henry Gustav Floren Trust; Stanford Department of Medicine Team Science Program; and NIH R01 AI175771-01. The influenza vaccine clinical study was funded in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. 75N93021C00015. J.R.H. was partially supported by NIH/NCI R01 CA264090-01. R.S.C. receives grant support from the Consortium for Food Allergy Research, National Institute of Allergy and Infectious Disease, and Food Allergy Research & Education. This study was partially supported by NIH/NCI SeroNet award 1U54CA260517; funder mandates include open access publication and full data sharing.

Author contributions: M.E.Z., A.K., and S.D.B. conceived the study and wrote the manuscript with input from all authors. Data generation: N.R., T.D.P., E.S.P., S.R.M., W.D., E.M.D., C.C.R., M.A.M., B.F.H., J.D.G., J.R.H., R.S.C, K.C.N., B.A.P., C.A.B., S.E.H., K.J., E.M., I.B., P.J.U., J.T.M., J.M.G., J.A.J., and S.Y. provided blood samples, as well as clinical and demographic data annotation and analysis and helpful discussions. J.Y.L., K.D.N., R.A.H., K.R., and B.L. prepared and sequenced samples. K.M.R. designed and created the data warehouse. M.E.Z. and K.R. ran the bioinformatics pipeline. M.E.Z. processed the external cohorts. Model development: M.E.Z. and S.D.B. developed models 1 and 2. E.C., M.E.Z., J.K.M., S.D.B., R.T., and A.K. designed the model 3 sequence and patient classification stages. M.E.Z. and J.K.M. evaluated language model embeddings for model 3. M.E.Z., S.D.B., R.T., and A.K. designed the ensemble model. M.E.Z., J.K.M., and N.S. wrote the Python package. Computational analysis: M.E.Z. evaluated performance by dataset, disease, gene locus, and model component. N.S. compared SLE predictions and SLEDAI scores. J.K.M. and M.E.Z. performed the healthy donor replicate analysis. M.E.Z., R.T., and A.K. evaluated demographic influences on immune repertoires. M.E.Z., E.C., and A.K. analyzed the Shapley feature importances. M.E.Z., N.S., E.C., and R.T. analyzed the known binders. **Competing interests:** M.E.Z., E.C., J.K.M., N.S., R.T., A.K., and S.D.B. are coinventors on patent applications related to this manuscript. S.D.B. has consulted for Regeneron, Sanofi, Novartis, Genentech, Visterra, and Janssen on topics unrelated to this study and owns stock in AbCellera Biologics. A.K. is scientific co-founder of Ravel Biotechnology Inc., is on the scientific advisory board of PatchBio Inc., SerImmune Inc., AINovo Inc., TensorBio Inc. and OpenTargets, was a consultant with Illumina Inc. and owns shares in DeepGenomics Inc., Immunai Inc., and Freenome Inc. C.A.B. reports compensation for consulting and/or SAB membership from Catamaran Bio, DeepCell Inc., Immunebridge, Sangamo Therapeutics, and Revelation Biosciences on topics unrelated to this study. J.D.G. has consulted for Eli Lilly, Gilead, GSK, and Karius, and reports research support from Eli Lilly, Gilead, Regeneron, Merck, and collaborative services agreements with Adaptive Biotechnologies, Monogram Biosciences, and Labcorp (outside of this study). R.T. is a consultant for Genentech. J.A.J. has served as a consultant for AbbVie, Janssen, Novartis, and GlaxoSmithKline. J.A.J. also has unrelated patents through the Oklahoma Medical Research Foundation which the foundation has licensed to Progentec Biosciences, LLC. J.T.M. has served as a consultant for AbbVie, Alexion, Alumis, Amgen, AstraZeneca, Aurinia, Bristol Myers Squibb, EMD Serono, Genentech, Gilead, GlaxoSmithKline, Lilly, Merck, Pfizer, Provention, Remegen, Sanofi, UCB, and Zenas, and reports research support from AstraZeneca, Bristol Myers Squibb, and GlaxoSmithKline (outside of this study). K.C.N. is an inventor or coinventor on unrelated patents, is a scientific co-founder of Alladapt, BeforeBrands, IgGenix, and Latitude, owns stock in those and in Seed, Excellergy, ClostraBio, and Cour Pharmaceuticals. K.C.N. has consulted for Regeneron and Novartis on topics unrelated to this study. S.E.H. reports receiving consulting fees from Sanofi Vaccines, Lumen, Novavax, and Merck. S.E.H. is a coinventor on patents that describe the use of nucleoside-modified mRNA as a vaccine platform. J.R.H. is a consultant for Regeneron and has received research support from Merck and Gilead. R.S.C. is an advisory board member for Alladapt Immunotherapeutics, Novartis, Allergenis, Intrommune Therapeutics, Phylaxis, Genentech, and Blueprint Therapeutics, and owns stock for Intrommune Therapeutics. J.K.M. owns stock in Tempus AI. All other authors declare that they have no competing interests. **Data and materials availability:** Raw sequencing data are deposited and freely accessible at the Sequence Read Archive under BioProject accession number PRJNA1147802. Prior published datasets are listed in table S1 and are deposited under BioProject accession numbers PRJNA486667 and PRJNA491287. Processed data are deposited on the Synapse platform at https://www.synapse.org/malid, both in Adaptive Immune Receptor Repertoire (AIRR) Rearrangement Schema format and in Mal-ID internal format (*63*). All other data needed to evaluate the conclusions in the paper are present in the paper or the Supplementary Materials. The use of data was approved by Stanford University IRBs #8629, #13952, #35453, #48973, #55650, and #55689; Oklahoma Medical Research Foundation IRBs #05-04, #06-12, #09-21, and #11-53; Providence St. Joseph Health IRB study number STUDY2020000175; University of Pennsylvania IRB #849398; and Duke University for the dataset previously deposited under SRA BioProject PRJNA486667. Code is deposited with version tag release-202408 (*68*) and shared under a license that allows non-commercial use.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adp2407
Supplementary Text
Figs. S1 to S17
Tables S1 to S9
References (*69*–*100*)
MDAR Reproducibility Checklist

Submitted 3 April 2024; accepted 29 November 2024
10.1126/science.adp2407