Accurate and scalable multi-disease classification from adaptive immune repertoires

Running title: Scalable Multi-Disease Classification

Natnicha Jiravejchakul[1,2], Ayan Sengupta[3], Songling Li[1,2], Debottam Upadhyaya[3], Mara A. Llamas-Covarrubias[4], Florian Hauer[3], Soichiro Haruna[2], and Daron M. Standley[1,2]

[1]Department of Systems Immunology, Immunology Frontier Research Institute (IFReC), The University of Osaka, 3-1 Yamadaoka, Suita, Osaka, 565-0871, Japan

[2]Department of Genome Informatics, Research Institute for Microbial Diseases (RIMD), The University of Osaka, 3-1 Yamadaoka, Suita, Osaka, 565-0871, Japan

[3]MyImmune Corporation, 251 Little Falls Drive, Wilmington, DE 19808, United States

[4]Department of Human Host Defense, Comprehensive Education and Research Center for Infectious Diseases (CiDER), 1-10 Yamadaoka, Suita, Osaka, 565-0871, Japan

*Corresponding author: standley@biken.osaka-u.ac.jp

Summary

Background Machine learning models trained on paratope-similarity networks have shown superior accuracy compared with clonotype-based models in binary disease classification. However, the computational demands of paratope networks hinder their use on large datasets and multi-disease classification.

Methods We reanalyzed publicly available T cell receptor (TCR) repertoire data from 1,421 donors across 15 disease groups and a large control group, encompassing approximately 81 million TCR sequences. To address computational bottlenecks, we replaced the paratope-similarity network approach (Paratope Cluster Occupancy or PCO) with a new pipeline Fast Approximate Clustering Techniques (FACTS) which comprises of four main steps: (1) high-dimensional vector encoding of sequences; (2) efficient clustering of vectors; (3) donor-level feature construction from cluster distributions; and (4) gradient-boosted decision tree

classification for multi-class disease prediction.

Findings FACTS processed 10^7 sequences in under 120 CPU hours. Using only TCR data, and evaluated with 3-fold cross-validation, it achieved a mean ROC AUC of 0.99 across 16 disease classes. Compared with the recently reported Mal-ID model, FACTS achieved higher donor-level classification accuracy for BCR (0.840 vs. 0.740), TCR (0.862 vs. 0.751), and combined BCR+TCR datasets (0.898 vs. 0.853). FACTS also preserved biologically meaningful signals, as shown by unsupervised t-SNE projections revealing distinct disease-associated and age-associated clusters.

Interpretation Paratope-based encoding with FACTS-derived features provides a scalable and biologically grounded approach for adaptive immune receptor (AIR) repertoire classification. The resulting classifier achieves superior multi-disease diagnostic performance while maintaining interpretability, supporting its potential for clinical and population-scale health profiling.

Funding This study was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI [JA23H034980] and by the Japan Agency for Medical Research and Development (AMED) [JP21am0101001].

Research in context

Evidence before this study

T and B cell receptor (TCR and BCR) repertoires encode lifelong immunological memory and antigen-specific responses, making them valuable for disease diagnosis and prediction. Existing machine learning (ML) models for adaptive immune receptor (AIR) repertoires often rely on clonotype-based representations, which limit shared receptor detection between donors and thus reduce cross-individual disease signature detection. Most models also lack robust multi-disease, population-scale performance. Our previous work showed that representing repertoires as paratope-similarity networks increased the fraction of shared receptors and improved disease classification. However, their computational complexity has limited their scalability for the large datasets required in multi-disease classification.

Added value of this study

We introduce a unified ML framework integrating paratope similarity with scalable sequence encoding. Applied to TCR repertoires from 1,421 donors across 15 diseases and one control group, the approach maintained high performance while efficiently processing 81 million sequences on standard CPU infrastructure. Compared to Mal-ID, our paratope-encoded method achieved significantly higher donor-level accuracy and revealed biologically meaningful disease- and age-associated patterns.

Implications of all the available evidence

Our paratope-encoded, scalability-enhanced model offers high accuracy, and interpretability for multi-disease classification, bringing AIR repertoire-based diagnostics closer to clinical translation and potentially guiding precision immunotherapy and immune-based therapeutic discovery for a wide range of disease.

1 Introduction

B and T lymphocytes generate diverse antigen receptors through somatic V(D)J gene recombination[1,2]. The resulting B-cell receptors (BCRs) and T-cell receptors (TCRs), collectively known as adaptive immune receptors (AIRs), record a lifelong molecular record of antigen encounters. AIRs can be profiled through a routine blood draw, followed by standard sequencing, offering a minimally invasive, hypothesis-free source of disease biomarkers that can detect immune perturbations before clinical symptoms appear. AIR repertoire analysis, therefore, holds promise for early and personalized disease detection, prognosis, and the discovery of immune-based therapeutics[3,4].

In recent years, machine learning (ML) has emerged as a powerful approach for identifying disease-specific signatures from AIR repertoires, as reviewed by Mhanna *et al*.[5]. While binary classifiers have been effective in distinguishing specific disease from health, such models do not reflect the complexity of the immune system, which responds to diverse pathogens and disease conditions. Multi-disease classification, as exemplified by the recently introduced Mal-ID program on 6 disease/healthy classes, is thus both more realistic and clinically relevant[6]. However, training such ML models requires larger and more heterogeneous datasets, increasing computational demands and making efficiency a critical factor for clinical translation.

Most existing ML models rely on clonotype-based repertoire representations in which sequences that share the same V-gene and possess identical, or highly similar, complementarity-determining region 3 (CDR3) sequences are grouped together[3,7]. Due to the immense diversity of BCRs and TCRs, the proportion of clonotypes shared between donors is extremely low. Studies report approximately 1–6% overlap in IGH clonotypes and around 8–13% in TCRβ clonotypes across healthy donors[8,9]. This sparsity limits the generalizability and sensitivity of clonotype-centric approaches.

Paratope-based representations address this limitation by focusing on receptor residues likely to interact with antigens or peptide–MHC complexes[10,11]. This avoids combining categorial germline gene annotations with continuous amino acid sequences and extends analysis beyond CDR3 to include CDR1 and CDR2, enabling statistically robust sequence comparisons. Notably,

paratope clusters are typically antigen- and epitope-specific. In prior work, grouping BCRs by paratope similarity (80% sequence identity and 90% alignment coverage) yielded clusters in which more than 95% of antibodies recognized the same antigen, with epitope-specific neutralizing antibodies co-clustering as well[12]. Paratope similarity networks were shown to provide superior features for binary disease classification, outperforming state-of-the-art clonotype-based models such as DeepRC and immuneML[13]. However, constructing such networks requires all-against-all similarity comparison, limiting scalability to datasets with fewer than 10^6 sequences.

In this study, we developed a scalable paratope-based feature construction pipeline to enable multi-disease classification across large AIR datasets. We replaced computationally intensive paratope similarity graphs with an efficient sequence encoding and clustering strategy called Fast Approximate Clustering Techniques (FACTS). To validate FACTS, we reanalyzed publicly available TCR repertoire data from 1,421 donors across 15 disease groups and one large control group. Compared to our previous method, Paratope Cluster Occupancy (PCO), FACTS dramatically reduced computational time while maintaining high classification performance, achieving a mean three-fold ROC AUC of 0.99. On the six-class Mal-ID benchmark, FACTS achieved comparable TCR-only accuracy (86.2%) to the Mal-ID model using both BCR and TCR data (85.3%). Moreover, unsupervised embedding projections revealed disease- and demographic-associated clusters, supporting FACTS as a robust, scalable, and interpretable framework for repertoire-based diagnostics with potential for clinical and population-scale applications.

2 Methods

Dataset curation and preprocessing

Adaptive immune receptor (AIR) sequencing data from a total of 1,421 donors were obtained from publicly available studies spanning 16 disease groups, including healthy individuals, patients with autoimmune disease, infectious disease, cancer, and vaccinated donors. Data for healthy donors, human immunodeficiency virus (HIV) infection, influenza vaccination, coronavirus disease 2019 (COVID-19), systemic lupus erythematosus (SLE), and type 1 diabetes (T1D) were sourced from the Mal-ID public dataset[6] (Supplementary Table 1). Samples labeled as "HIV-negative" were excluded from the healthy donor group in our study. Data for the remaining 10 disease groups, as well as additional healthy donors, were obtained from 12 published studies[14-25] (Supplementary Table 1). No original sequencing data were generated in this study, and ethical approval for secondary analysis was not required by the authors' institution.

Paratope extraction

In the first step of the FACTS pipeline, paratope regions were extracted from raw AIR sequences using ANARCI[26], with residue numbering assigned according to the IMGT scheme[27]. For each BCR and TCR sequence, paratopes were defined as all positions within CDR1, CDR2, and CDR3, concatenated to form a single pseudo-sequence used in all downstream encoding and clustering steps.

Sequence encoders

Following paratope extraction, each pseudo-sequence was transformed into a fixed length embedding using a protein language model. A proprietary encoder for TCR sequence was developed, adopting a Transformer-based encoder-only architecture trained on large-scale, publicly available AIR datasets. Training employed the masked language modelling (MLM) objective from BERT[28], in which random amino acid tokens are masked, and the model is optimized to predict the missing residues. Novel training strategies were implemented to incorporate structural features relevant to receptor–antigen interactions. For the TCR encoder model, weights were initialized from ESM-2[29] and fine-tuned on TCR datasets using paratope-focused sequence inputs. The resulting TCR encoder generated dense vector embeddings for

each paratope pseudo-sequence, which were subsequently used for clustering and classification. By default, FACTS uses IgBert[30] for encoding BCR sequences.

Clustering of sequence embeddings

To group similar paratope embeddings, MiniBatchKMeans, a memory- and time-efficient variant of k-means, was applied using the scikit-learn[31] implementation with k ranging from 100 to 1,000. A batch size of 1,024 was used, and initial centroids were seeded with the k-means++ initializer to improve stability. Clustering quality was assessed across the k range using inertia (within-cluster sum of squares) and silhouette coefficient scores[32]. MiniBatchKMeans scales approximately as O(N log k) per pass through the dataset, enabling efficient clustering of tens of millions of paratope embeddings without prohibitive computational cost.

Construction of donor-level features

Following clustering, donor-level feature vectors were generated by aggregating cluster membership data. For each donor, the frequency distribution of clustered paratope embeddings was computed as the proportion of sequences assigned to each cluster. Inter-cluster mutual information was also calculated. This produced a fixed-length vector representing each donor's immune repertoire, which served as input for downstream classification.

Classification models and statistical analysis

All classification models were trained using the XGBoost[33] implementation in soft probability mode. Model performance was assessed using three-fold cross-validation with donor-level stratification to prevent donor overlap between folds. Evaluation metrics included accuracy, ROC AUC, macro-averaged F1 score (macro F1), sensitivity, and specificity. Paired AUC values were compared using the Wilcoxon signed-rank test. Ninety-five percent confidence intervals (CIs) for performance metrics were calculated via bootstrap resampling (n = 10,000 iterations).

Benchmarking sequence encoders

The proprietary MyImmune TCR encoder was benchmarked against open-source protein and immune receptor language models, including TCRBert[34], ESM-2[29], ProtTrans[35], SCEPTR[36], and UniTCR[37]. For BCR sequences, six publicly available encoders; IgBERT[30], AntiBERTa2[38],

194 AntiBERTy[39], ESM-2[29], ProtTrans[35], and AbLang[40], were evaluated. Details of each encoder are
195 provided in Supplementary Table 2. Each model was integrated into our pipeline using identical
196 clustering and classification procedures, differing only in the embedding step. Benchmarking
197 was performed on TCR and BCR datasets from the Mal-ID cohort[6] (499 donors for TCR, and
198 550 donors for BCR dataset, across five disease classes and one healthy control group.
199
200 **3 Results**
201
202 **Scalable paratope-based classification pipeline**
203 We developed a scalable machine learning framework to classify AIR repertoire data using
204 paratope-derived sequence features. Previously, these features were computed by combining
205 paratope clusters with a paratope similarity graph, denoted Paratope Cluster Occupancy (PCO)[13]
206 (Figure 1A). To avoid the all-against-all comparisons required for graph construction, we
207 reversed the workflow: sequences were first encoded as vectors, then efficiently clustered using
208 these vector representations. We term this approach Fast Approximate Clustering Techniques
209 (FACTS) (Figure 1B).
210
211 The pipeline begins with the extraction of paratope regions from BCR or TCR sequences,
212 represented as pseudo-sequences composed of the concatenated CDR1, CDR2, and CDR3. These
213 sequences are encoded, and the resulting embeddings are clustered using k-means (k = 50-500).
214 A silhouette coefficient score[32] is used to determine a good starting point for choosing the
215 number of clusters k. Donor-level feature vectors are constructed from cluster frequency
216 distributions and inter-cluster mutual information. These features are then used to train gradient-
217 boosted decision tree models for multi-class disease classification.
218
219 The use of FACTS qualitatively improved the scalability compared with PCO (Figure 2). Two
220 PCO implementations were tested: one using an exact all-against-all comparison[13] and one using
221 an MMSEQS-based approximation[41]. Even with MMSEQS acceleration, PCO runtimes
222 increased quadratically with input size ($R^2 = 0.9997$). In contrast, FACTS scaled linearly when
223 using a small number of clusters: at k = 50, runtime increased to 115 ms per 1,000 sequences (R^2
224 $= 0.9957$); at k = 500, runtime rose to 1,158 ms per 1,000 sequences ($R^2 = 0.9195$), reflecting the

225 added cost of inter-cluster comparisons. Processing 5 million sequences required approximately
226 1.9 CPU-hours using FACTS, compared to about 324 CPU-hours with MMSEQS-accelerated
227 PCO. This demonstrated an approximate 170-fold improvement in efficiency and scaled linearly
228 with increasing data size.
229
230 **Encoded paratope features achieve robust multi-disease TCR repertoire classification**
231 To assess classification performance, we reanalyzed publicly available TCR repertoire data
232 comprising 1,421 donors across 16 disease classes (Supplementary Table 1). The classifier
233 achieved a mean ROC AUC of 0.99 in three-fold cross-validation (Figure 3A). Per-disease
234 AUCs were consistently high: the lowest was for healthy controls (0.94), followed by influenza-
235 vaccinated donors (0.97), with all other disease classes greater than 0.98 (Figure 3B). The model
236 achieved high sensitivity (0.9867 ± 0.0007) with moderately high specificity (0.7785 ± 0.0203)
237 (Figure 3C; Supplementary Figure 1).
238
239 The dataset included most of the Mal-ID benchmark, enabling direct comparison with the Mal-
240 ID framework (499 donors for TCR and 550 donors for BCR dataset, five disease classes, and
241 one healthy control group). Using overall donor-level accuracy, our method outperformed Mal-
242 ID for all input types: BCR (84% vs 74%), TCR (86% vs 75.1%), and combined BCR+TCR
243 (89.8% vs 85.3%) (Figure 3D). Our TCR-only model achieved accuracy comparable to Mal-ID's
244 combined-input model, which represents a significant simplification in the input requirements for
245 FACTS. ROC AUCs for the Mal-ID subset were as follows: 0.9718 (BCR), 0.9747 (TCR), and
246 0.9833 (BCR+TCR) (Supplementary Figure 2).
247
248 **MyImmune encoder for multi-disease TCR classification**
249 We next evaluated alternative encoders within the FACTS pipeline using the Mal-ID dataset.
250 All encoders followed the same classification procedure, differing only in embedding generation.
251 For BCR sequences, among publicly available BCR encoders, IgBERT achieved the highest
252 performance (accuracy: 82.0% ± 3.7%; ROC AUC: 0.9736 ± 0.0092). Our TCR encoder,
253 referred to as MyImmune, achieved the highest performance for TCR data (accuracy: 86.2% ±
254 6.2%; ROC AUC: 0.9775 ± 0.0097), outperforming TCRBert, ProtTrans, ESM-2, and SCEPTR

(Table 1). These results support the use of paratope-based classification and demonstrate that MyImmune, combined with FACTS, enables highly accurate TCR-based disease classification.

Encoded paratope representations capture biologically meaningful structures of TCR repertoires

To assess biological relevance, we visualized donor-level embeddings from encoded TCR paratope sequences using t-SNE. In the full dataset (1,421 donors, 16 disease classes), projections showed clear separation between disease groups, indicating that paratope-based features capture disease-specific immune signatures (Figure 4A; Supplementary Figure 3). A batch effect was observed in the autoimmune hepatitis (AIH) dataset, likely reflecting differences in experimental protocols across publicly available studies (Supplementary Figure 4).

Restricting analysis to the Mal-ID TCR subset (n = 499 donors), which was collected and processed using a standardized protocol, removed obvious batch effects and maintained strong disease group separation (Figure 4B). The t-SNE projection also revealed older individuals ($Z < 5$) and younger individuals ($Z > 5$) clustered separately. Statistically significant age differences were found within both healthy controls and SLE donors (Figure 4C), suggesting that paratope features capture clinically relevant variation beyond disease labels. Similar, although not statistically significant, trends were observed in other disease groups.

4 Discussion

This study presents a scalable, accurate, and interpretable framework for multi-disease classification using AIR repertoires. By introducing the FACTS pipeline with paratope-focused sequence representations, we achieved major improvements in processing speed compared with our previous paratope similarity network approach, without compromising accuracy. This improvement in efficiency is critical as the computational cost precluded the use of the PCO approach on larger datasets like the Mal-ID benchmark. Applied to an even larger TCR dataset of 1,421 donors spanning 16 disease classes, the FACTS framework demonstrated strong classification performance. In direct comparison with Mal-ID on its six-class benchmark subset, donor-level accuracy for single-chain inputs (BCR or TCR) was markedly higher with our

284 approach, suggesting that paratope-based features may have distinct advantages over clonotype-
285 based features.

286 Paratopes, comprising the CDR1, CDR2, and CDR3 regions, represent AIR residues capable of
287 directly interacting with antigens. Previous work has shown that paratope sequences can reflect
288 convergent immune responses across individuals[13], and that BCR clusters are highly antigen- and
289 epitope-specific[12]. Other studies have demonstrated that antibodies binding the same epitope
290 converge in paratope space and can be identified using deep learning predictors such as
291 ParaPred[42] and structure-aware tools like Ab-Ligity[43].

292 Here, by reversing the order of clustering and similarity comparison in our previous pipeline, we
293 improved training runtime without loss of accuracy. Rather than computing pairwise similarities
294 first, we encoded sequences into vector representations, then performed clustering in this reduced
295 space. The embeddings, generated using protein language models trained on natural amino acid
296 sequences, capture biochemical and structural properties. Such embeddings have recently been
297 used in various immune repertoire analyses[30,34,38,39,43]. MiniBatchKMeans clustering was then
298 applied to group encoded paratope sequences, enabling efficient processing of up to hundreds of
299 millions of sequences. Donor-level features were derived from cluster frequency distributions
300 and inter-cluster mutual information, which were subsequently used for classification. The
301 integration of these steps produced a fast, scalable pipeline for immune repertoire-based disease
302 classification.

303 Benchmarking demonstrated that the MyImmune TCR paratope encoder further enhanced
304 classification performance. Compared with multiple publicly available TCR protein language
305 models, MyImmune consistently achieved the highest accuracy on the Mal-ID cohort,
306 reinforcing the value of paratope-aware representation learning in repertoire-based diagnostics.

307 Beyond classification accuracy, the learned feature space preserved biologically meaningful
308 structures. Unsupervised t-SNE projections revealed clustering by disease cohort and age
309 associations, particularly among healthy and SLE donors, suggesting that paratope features
310 capture clinically relevant immune signatures beyond discrete disease classification. However,
311 batch effects were evident in the full 1,421-donor dataset, especially in the AIH dataset, which
312 likely reflects variability in experimental protocols. These challenges underscore the importance

of using well-matched, prospectively collected datasets to validate findings in clinically controlled settings.

Overall, our findings show that paratope-encoded representations, when combined with the FACTS pipeline, enable accurate, scalable, and biologically meaningful classification of immune repertoires, supporting their potential for clinical translation and population-scale immune profiling.

Contributors

N.J., A.S., F.H., and D.M.S. conceived and designed the study.
A.S., S.L., D.U., and M.L. acquired the data and performed the formal analysis.
N.J., A.S., H.S., and D.M.S. contributed to the result interpretation and critical discussion.
N.J. and D.M.S. drafted the original manuscript and prepared the figures and tables.
All authors contributed to the manuscript revision, reviewed the final version, and approved it for submission.

Data sharing statement

All AIR-seq datasets used in this study are publicly available, and no individual-level data were collected specifically for this investigation. The source and access information for each dataset are detailed in Supplementary Table 1. This study was not subject to an institutional protocol and did not require participant consent. Processed donor-level feature matrices will be made available upon submission. Custom code for paratope extraction and model training is proprietary but can be shared with editors and qualified academic investigators upon request under a material transfer agreement.

Declaration of interests

A.S. and F.H. are employees and shareholders of MyImmune Corporation. D.M.S. is a shareholder in MyImmune Corporation. All other authors declare no competing interests.

Acknowledgements

We thank all members of the Systems Immunology Lab at Osaka University for their helpful discussions and support. This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant [JA23H034980], the Japan Agency for Medical Research and Development (AMED) Grant [JP21am0101001], and the Kishimoto Foundation Fellowship (N.J.).

Reference

1. Schatz DG, Swanson PC. V(D)J recombination: mechanisms of initiation. *Annu Rev Genet* 2011; **45**: 167–202.
2. Victora GD, Nussenzweig MC. Germinal centers. *Annu Rev Immunol* 2012; **30**: 429–57.
3. Emerson RO, DeWitt WS, Vignali M, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet* 2017; **49**(5): 659–65.
4. Robins HS, Campregher PV, Srivastava SK, et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 2009; **114**(19): 4099–107.
5. Mhanna V, Bashour H, Lê Quý K, et al. Adaptive immune receptor repertoire analysis. *Nature Reviews Methods Primers* 2024; **4**(1): 6.
6. Zaslavsky ME, Craig E, Michuda JK, et al. Disease diagnostics using machine learning of B cell and T cell receptor sequences. *Science* 2025; **387**(6736): eadp2407.
7. Chen Y, Ye Z, Zhang Y, et al. A Deep Learning Model for Accurate Diagnosis of Infection Using Antibody Repertoires. *J Immunol* 2022; **208**(12): 2675–85.
8. Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* 2019; **566**(7744): 393–7.
9. Nikolich-Zugich J, Slifka MK, Messaoudi I. The many important facets of T-cell repertoire diversity. *Nat Rev Immunol* 2004; **4**(2): 123–32.
10. Richardson E, Galson JD, Kellam P, et al. A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-pertussis toxoid antibodies. *MAbs* 2021; **13**(1): 1869406.
11. Xu Z, Li S, Rozewicki J, et al. Functional clustering of B cell receptors using sequence and structural features. *Molecular Systems Design & Engineering* 2019; **4**(4): 769–78.
12. Saputri DS, Ismanto HS, Nugraha DK, et al. Deciphering the antigen specificities of antibodies by clustering their complementarity determining region sequences. *mSystems* 2023; **8**(6): e0072223.

13. Xu Z, Ismanto HS, Saputri DS, et al. Robust detection of infectious disease, autoimmunity, and cancer from the paratope networks of adaptive immune receptors. *Brief Bioinform* 2024; **25**(5).
14. Pothuri VS, Hogg GD, Conant L, et al. Intratumoral T-cell receptor repertoire composition predicts overall survival in patients with pancreatic ductal adenocarcinoma. *Oncoimmunology* 2024; **13**(1): 2320411.
15. Formenti SC, Rudqvist NP, Golden E, et al. Radiotherapy induces responses of lung cancer to CTLA-4 blockade. *Nat Med* 2018; **24**(12): 1845–51.
16. Wu L, Zhu J, Rudqvist NP, et al. T-Cell Receptor Profiling and Prognosis After Stereotactic Body Radiation Therapy For Stage I Non-Small-Cell Lung Cancer. *Front Immunol* 2021; **12**: 719285.
17. Zuckerbrot-Schuldenfrei M, Aviel-Ronen S, Zilberberg A, Efroni S. Ovarian cancer is detectable from peripheral blood using machine learning over T-cell receptor repertoires. *Briefings in Bioinformatics* 2024; **25**(2): bbae075.
18. Komech EA, Pogorelyy MV, Egorov ES, et al. CD8+ T cells with characteristic T cell receptor beta motif are detected in blood and expanded in synovial fluid of ankylosing spondylitis patients. *Rheumatology (Oxford)* 2018; **57**(6): 1097–104.
19. Shugay M, Bagaev DV, Turchaninova MA, et al. VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLoS Comput Biol* 2015; **11**(11): e1004503.
20. Luoma AM, Suo S, Wang Y, et al. Tissue-resident memory and circulating T cells are early responders to pre-surgical cancer immunotherapy. *Cell* 2022; **185**(16): 2918–35.e29.
21. Robert L, Tsoi J, Wang X, et al. CTLA4 blockade broadens the peripheral T-cell receptor repertoire. *Clin Cancer Res* 2014; **20**(9): 2424–32.
22. Yusko E, Vignali M, Wilson RK, et al. Association of Tumor Microenvironment T-cell Repertoire and Mutational Load with Clinical Outcome after Sequential Checkpoint Blockade in Melanoma. *Cancer Immunol Res* 2019; **7**(3): 458–65.
23. Schultheiß C, Simnica D, Willscher E, et al. Next-Generation Immunosequencing Reveals Pathological T-Cell Architecture in Autoimmune Hepatitis. *Hepatology* 2021; **73**(4): 1436–48.
24. Zhang Y, Zhu Y, Wang J, et al. A comprehensive model based on temporal dynamics of peripheral T cell repertoire for predicting post-treatment distant metastasis of nasopharyngeal carcinoma. *Cancer Immunol Immunother* 2022; **71**(3): 675–88.
25. Aterido A, López-Lasanta M, Blanco F, et al. Seven-chain adaptive immune receptor repertoire analysis in rheumatoid arthritis reveals novel features associated with disease and clinically relevant phenotypes. *Genome Biol* 2024; **25**(1): 68.
26. Dunbar J, Deane CM. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* 2016; **32**(2): 298–300.
27. Lefranc MP, Pommie C, Ruiz M, et al. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* 2003; **27**(1): 55–77.
28. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019.
29. Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023; **379**(6637): 1123–30.
30. Kenlay H, Dreyer FA, Kovaltsuk A, Miketa D, Pires D, Deane CM. Large scale paired antibody language models. *PLOS Computational Biology* 2024; **20**(12): e1012646.

31. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 2011; **12**: 2825–30.
32. Monshizadeh M, Khatri V, Kantola R, Yan Z. A deep density based and self-determining clustering approach to label unknown traffic. *Journal of Network and Computer Applications* 2022; **207**: 103513.
33. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA: Association for Computing Machinery; 2016. p. 785–94.
34. Wu KE, Yost KE, Daniel B, et al. TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-binding analyses. Proceedings of the 18th Machine Learning in Computational Biology Meeting. —: PMLR; 2023. p. 194–229.
35. Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2021; **—**(—): 1–.
36. Nagano Y, Pyo AGT, Milighetti M, et al. Contrastive learning of T cell receptor representations. *Cell Systems* 2025; **16**(1).
37. Gao Y, Dong K, Gao Y, et al. Unified cross-modality integration and analysis of T-cell receptors and T-cell transcriptomes by low-resource-aware representation learning. *Cell Genomics* 2024; **4**(5): 100553——.
38. Barton J. Enhancing Antibody Language Models with Structural Information (AntiBERTa2-CSSP). (preprint/discussion); 2023.
39. Ruffolo JA, Gray JJ, Sulam J. Deciphering antibody affinity maturation with language models and weakly supervised learning. *arXiv preprint* 2021.
40. Olsen TH, Moal IH, Deane CM. AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Advances* 2022; **2**(1): vbac046.
41. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017; **35**(11): 1026–8.
42. Liberis E, Velickovic P, Sormanni P, Vendruscolo M, Lio P. Parapred: antibody paratope prediction using convolutional and recurrent neural networks. *Bioinformatics* 2018; **34**(17): 2944–50.
43. Wong WK, Robinson SA, Bujotzek A, et al. Ab-Ligity: identifying sequence-dissimilar antibodies that bind to the same epitope. *MAbs* 2021; **13**(1): 1873478.

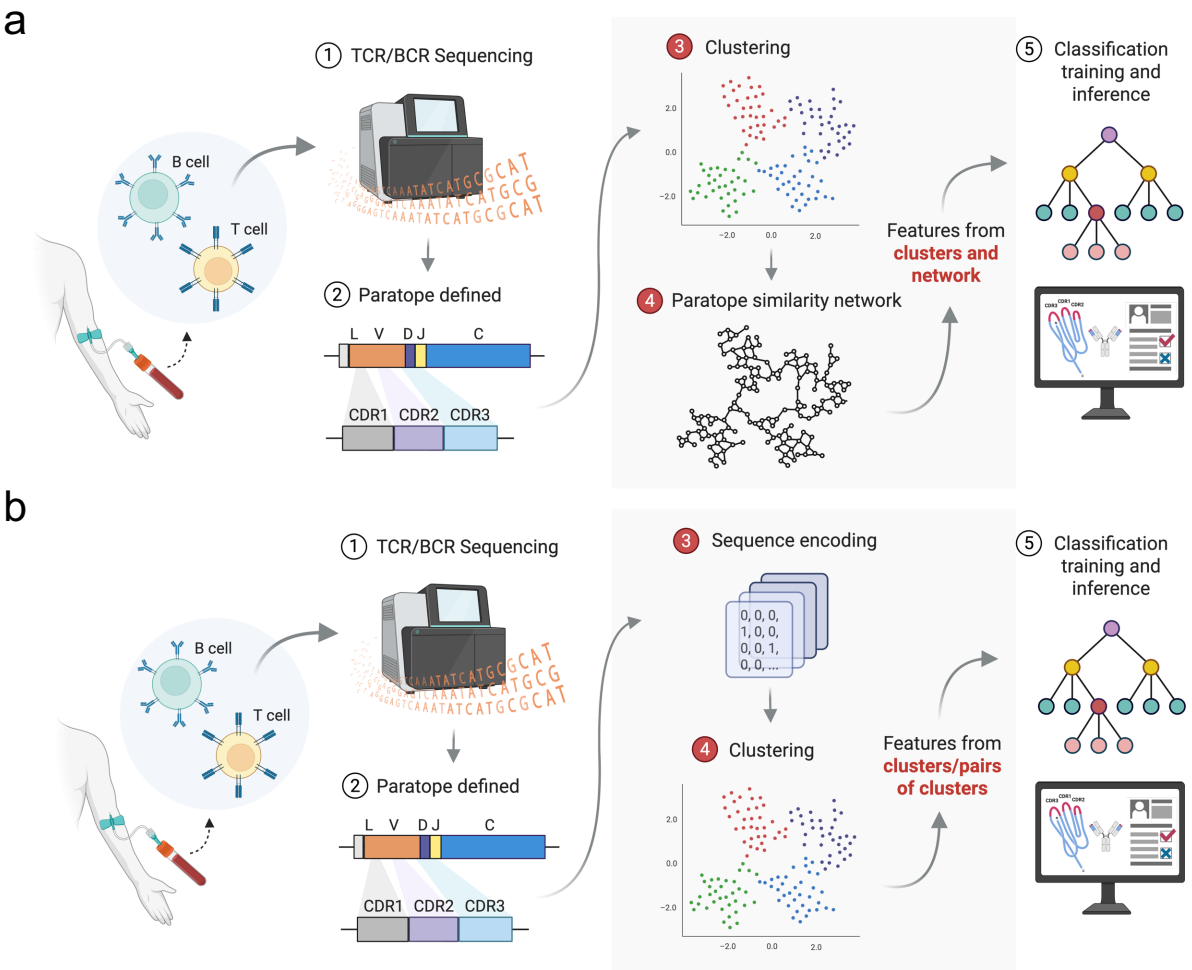

Figure 1. Scalable paratope-based classification pipelines for immune repertoire data. (**a**) The original paratope cluster occupancy (PCO) framework. After TCR/BCR sequencing, paratope regions (CDR1, CDR2, and CDR3) are defined and clustered. A paratope similarity network is constructed via all-against-all comparisons of paratope sequences, and donor-level features are derived from both clusters and networks. These features are then used to train a classification model. (**b**) The proposed fast approximate clustering techniques (FACTS). Instead of pairwise comparison, paratope sequences are first encoded into vector embeddings before clustering. Donor-level features are constructed from cluster frequencies and inter-cluster relationships. This reverse-order approach avoids the computational cost of paratope graph construction and enables scalable training of classification models.

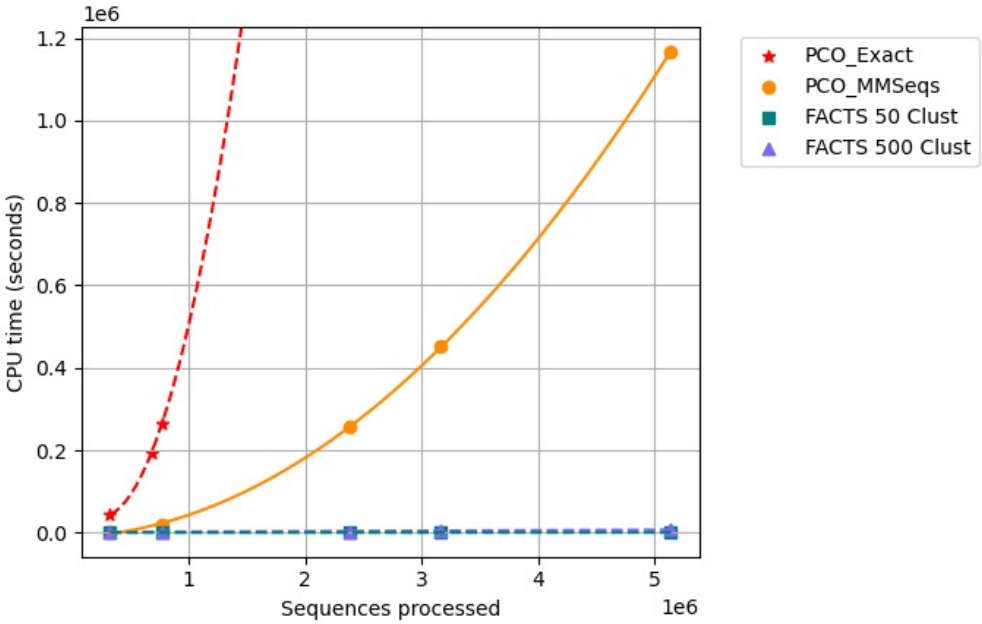

Figure 2. Runtime comparison between PCO and FACTS pipelines.
CPU time (in seconds) is plotted against the number of sequences processed for four methods. The original PCO pipeline using exact paratope similarity (PCO_Exact, red stars) shows quadratic runtime scaling with input size due to its all-against-all comparison step. A faster variant using MMseqs2 (PCO_MMSeqs, orange circles) improves efficiency but still scales nonlinearly. In contrast, the FACTS approach achieves near-linear scaling, with FACTS configured for 50 clusters (blue squares) and 500 clusters (purple triangles) showing minimal runtime increase over 5 million sequences. The FACTS pipeline achieves up to ~170-fold speedup compared to the exact PCO method, making it suitable for large-scale repertoire analysis.

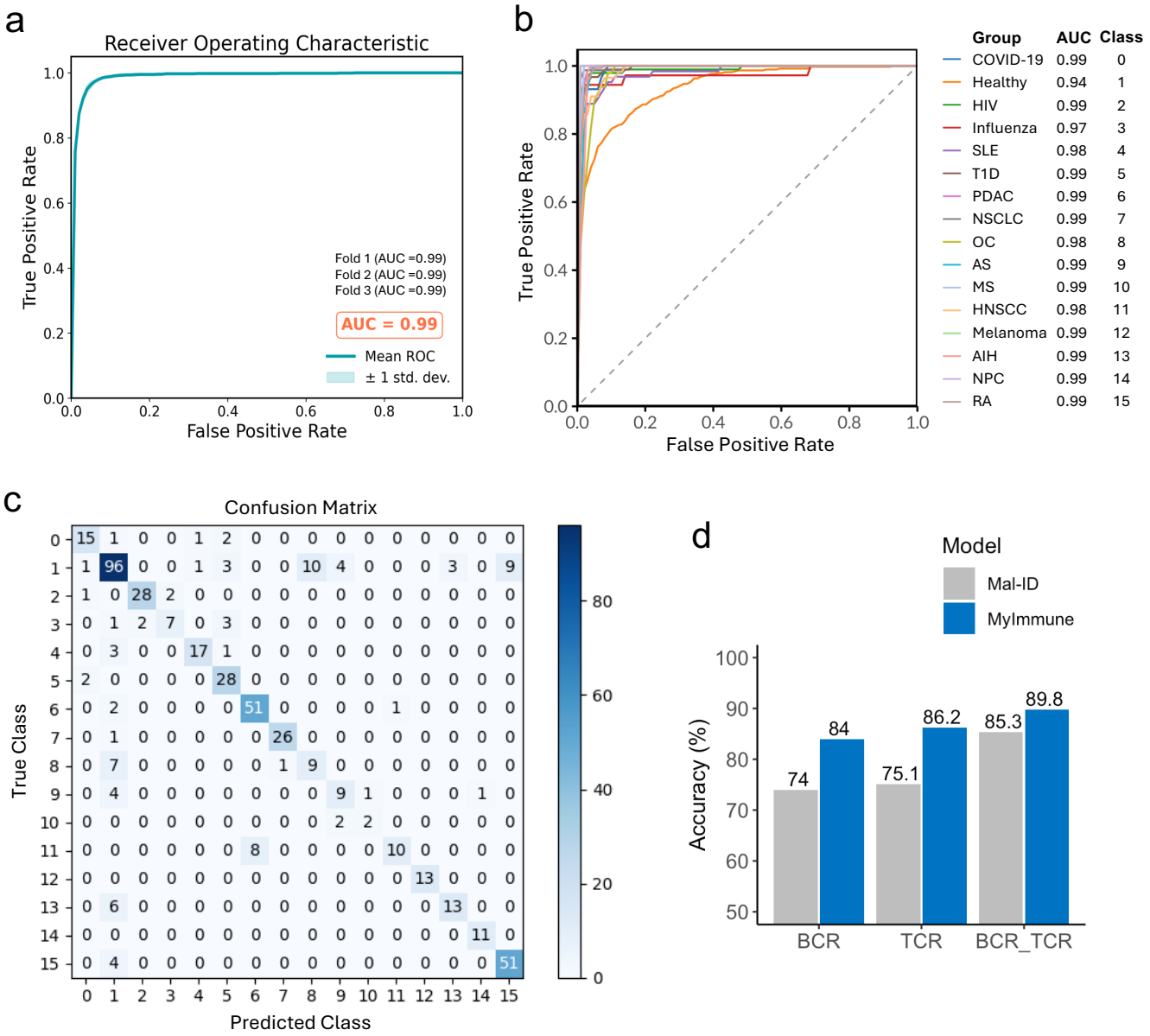

Figure 3. Encoded paratope features enable accurate multi-disease classification of TCR repertoires. (**a**) Receiver operating characteristic (ROC) curve showing overall performance of the paratope-based classifier on a 1,421-donor dataset comprising 16 disease classes, including healthy controls. The model achieved a mean ROC AUC of 0.99 with consistent performance across cross-validation folds. (**b**) One-vs-rest ROC curves for each individual class showing a robust performance across all disease classes. **(c)** Confusion matrix from one of the three cross-validation folds, illustrating donor-level classification results across the 16 disease classes. Disease labels corresponding to each class ID are provided in Supplementary Table 1. **(d)** Comparative performance between our paratope-based model ("MyImmune") and the original clonotype-based Mal-ID framework on the Mal-ID cohort, consisting of 550 donors for BCR and 499 donors for TCR dataset across six disease classes. Accuracy was computed as donor-level classification accuracy, representing the proportion of correctly classified donors.

BCR Encoders	Accuracy	AUROC	F1 score
IgBert	0.8400 ± 0.0306	0.9718 ± 0.0073	0.8386 ± 0.0386
Antiberta2	0.7920 ± 0.0497	0.9693 ± 0.0080	0.7809 ± 0.0241
AntiBERTy	0.7500 ± 0.0173	0.9555 ± 0.0108	0.7546 ± 0.0132
ESM-2 8M	0.7440 ± 0.0522	0.9501 ± 0.0158	0.7407 ± 0.0526
ProtTrans	0.7360 ± 0.0371	0.9537 ± 0.0108	0.8057 ± 0.0514
ESM-2 150M	0.7320 ± 0.0228	0.9473 ± 0.0063	0.7281 ± 0.0369
ESM-2 650M	0.7000 ± 0.0283	0.9386 ± 0.0076	0.7031 ± 0.0366
AbLang	0.6940 ± 0.0207	0.9398 ± 0.0030	0.6880 ± 0.0403

TCR Encoders	Accuracy	AUROC	F1 score
MyImmune Encoder	**0.8619 ± 0.0616**	**0.9775 ± 0.0097**	**0.8300 ± 0.0728**
TCRBert	0.8578 ± 0.0453	0.9735 ± 0.0108	0.8201 ± 0.0533
ESM-2 8M	0.8439 ± 0.0552	0.9755 ± 0.0104	0.8153 ± 0.0586
ProtTrans	0.8438 ± 0.0458	0.9715 ± 0.0110	0.8057 ± 0.0514
SCEPTR: Large model	0.8438 ± 0.0375	0.9747 ± 0.0121	0.8135 ± 0.0425
ESM-2 150M	0.8318 ± 0.0595	0.9742 ± 0.0103	0.7969 ± 0.0673
SCEPTR: Beta-chain only model	0.8317 ± 0.0282	0.9727 ± 0.0105	0.7905 ± 0.0241
SCEPTR: Default model	0.8238 ± 0.0452	0.9709 ± 0.0111	0.7902 ± 0.0597
UniTCR	0.6875 ± 0.0594	0.9299 ± 0.0121	0.6163 ± 0.0777

Table 1. Sequence encoders form multi-disease classification
Accuracy, AUROC, and F1 score (mean ± standard deviation) are shown for each encoder using the FACTS pipeline on the Mal-ID cohort (550 donors for BCR and 499 donors for TCR dataset across 6 disease classes).

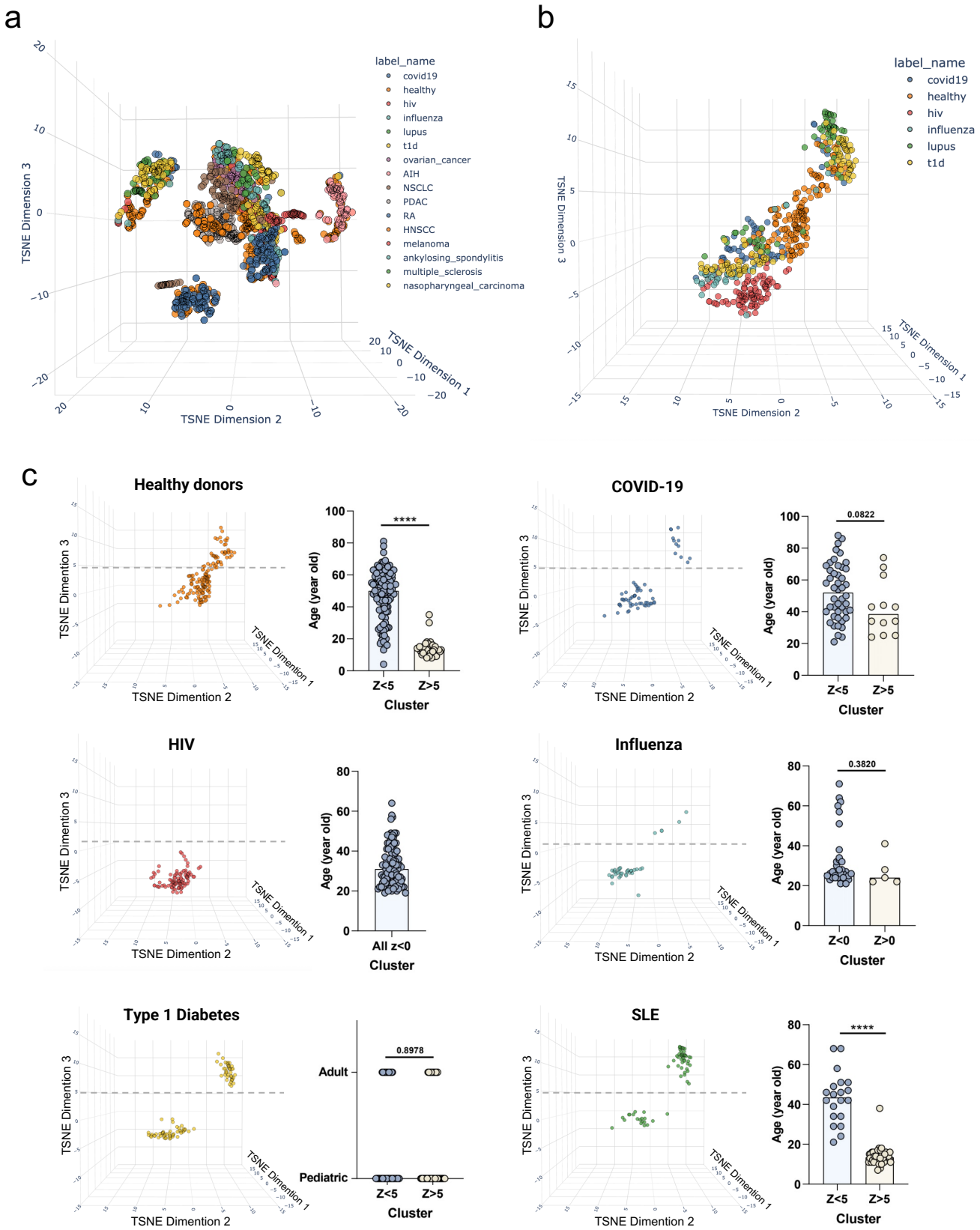

Figure 4. Encoded TCR paratope representations reveal disease- and age-associated structure. (**a**) t-SNE projection of donor-level TCR paratope embeddings across 1,421 donors from 16 disease classes. (**b**) t-SNE projection of the encoded TCR paratope sequences from the Mal-ID cohort at donor level (n = 499 donors, 6 disease classes). (**c**) Age association across clusters for each disease group in the Mal-ID cohort. Donors clustered along the third t-SNE axis (Z) based on age, with older individuals enriched in the Z < 5 region and younger donors in Z > 5. Statistical comparisons between clusters were performed using unpaired t-tests; **** indicates $p < 0.0001$.