Accurate and scalable multi-disease classification from adaptive immune repertoires

Running title: Scalable Multi-Disease Classification

Natnicha Jiravejchakul[a,b,i], Ayan Sengupta[c,i], Songling Li[a,b], Debottam Upadhyaya[c], Mara A. Llamas-Covarrubias[d], Florian Hauer[c], Soichiro Haruna[b], and Daron M. Standley[a,b]

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

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

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

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

[i]Denotes that these authors contributed equally to this work

*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 Fast Approximate Clustering Techniques (FACTS) pipeline, which is comprised of four main steps: (1) high-dimensional vector encoding of sequences; (2) efficient clustering of the resulting 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 5-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 (0.882 vs. 0.751), and combined BCR+TCR datasets (0.904 vs. 0.853) on the six-class Mal-ID benchmark. FACTS also preserved biologically meaningful signals, as shown by unsupervised t-SNE projections revealing distinct disease-associated and potentially 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], the Japan Agency for Medical Research and Development (AMED) [JP25am0101001], and the Kishimoto Foundation Fellowship.

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 biomarkers 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 between donors 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 FACTS, 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, FACTS 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 potentially age-associated patterns.

Implications of all the available evidence

FACTS 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 diseases.

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 cohorts of patients with specific disease from healthy donors, 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 than previous models[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 ML 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 five-fold ROC AUC of 0.99. On the six-class Mal-ID benchmark, FACTS using only TCR data achieved an accuracy of 88.2%, which is slightly higher than the 85.3% accuracy of the Mal-ID model that used both BCR and TCR data. 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 S1). 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 S1). 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. An encoder for TCR sequence was developed by 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. 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 "MyImmune" 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. Mathematical details of how the donor-level features are defined, was reported in the previous study[13].

Classification models and statistical analysis

All classification models were trained using the XGBoost[33] implementation in soft probability mode. Model performance was assessed using five-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

MyImmune TCR encoder was benchmarked against several 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], AntiBERTy[39], ESM-2[29], ProtTrans[35], and AbLang[40], were evaluated. Details of each encoder are provided in Supplementary Table S3. Each model was integrated into our pipeline using identical clustering and classification procedures, differing only in the embedding step. Benchmarking was performed on TCR and BCR datasets from the Mal-ID cohort[6] (499 donors for TCR, and 550 donors for BCR), across five disease classes and one healthy control group.

3 Results

Scalable paratope-based classification pipeline

We developed a scalable machine learning framework to classify AIR repertoire data using paratope-derived sequence features. Previously, these features were computed by combining paratope clusters with a paratope similarity graph, denoted Paratope Cluster Occupancy (PCO)[13] (Figure 1a). To avoid the all-against-all comparisons required for graph construction, we modified the workflow: sequences were first encoded as vectors, then efficiently clustered using these vector representations. We term this approach Fast Approximate Clustering Techniques (FACTS) (Figure 1b).

The pipeline begins with the extraction of paratope regions from BCR or TCR sequences, represented as pseudo-sequences composed of the concatenated CDR1, CDR2, and CDR3. These sequences are encoded, and the resulting embeddings are clustered using k-means (k = 50-500). A silhouette coefficient score[32] is used to determine a good starting point for choosing the number of clusters k. Donor-level feature vectors are constructed from cluster frequency

distributions and inter-cluster mutual information. These features are then used to train gradient-boosted decision tree models for multi-class disease classification.

The use of FACTS qualitatively improved the scalability compared with PCO (Figure 2). Two PCO implementations were tested: one using an exact all-against-all comparison and one using an MMSEQS[41]-based approximation[13]. Even with MMSEQS acceleration, PCO runtimes increased quadratically with input size ($R^2 = 0.9997$). In contrast, FACTS scaled linearly when using a small number of clusters: at k = 50, runtime increased to 115 ms per 1,000 sequences ($R^2 = 0.9957$); at k = 500, runtime rose to 1,158 ms per 1,000 sequences ($R^2 = 0.9195$), reflecting the added cost of inter-cluster comparisons. Processing 5 million sequences required approximately 1.9 CPU-hours using FACTS, compared to about 324 CPU-hours with MMSEQS-accelerated PCO. This demonstrated an approximate 170-fold improvement in efficiency and scaled linearly with increasing data size.

Robust multi-disease TCR repertoire classification achieved with encoded paratope features

To assess classification performance, we reanalyzed publicly available TCR repertoire data comprising 1,421 donors across 16 disease classes (Supplementary Table S1). The classifier achieved a mean ROC AUC of 0.9885 in five-fold cross-validation (Figure 3a; Supplementary Table S2). Per-disease AUCs were consistently high, with the lowest observed for healthy controls (0.95), while all other disease classes were greater than 0.98 (Figure 3b). The model obtained an accuracy of 81.77% (0.8177 ± 0.0090) with high specificity (0.9866 ± 0.0007) and moderately high sensitivity (0.7785 ± 0.0203) (Figure 3c; Supplementary Figure S1).

We further evaluated FACTS on the Mal-ID cohort to enable a direct comparison with the Mal-ID framework (499 donors for TCR and 550 donors for BCR dataset, five disease classes, and one healthy control group). We trained separate models for each input type; BCR, TCR, and combined BCR+TCR. Using overall donor-level accuracy, our method outperformed Mal-ID for all input types: BCR (84% vs 74%), TCR (88.2% vs 75.1%), and combined BCR+TCR (90.4% vs 85.3%) (Figure 3d). ROC AUCs for the Mal-ID subset were as follows: 0.9718 (BCR), 0.9781 (TCR), and 0.9824 (BCR+TCR) (Supplementary Figure S2). Importantly, our TCR-only model achieved accuracy comparable to Mal-ID's combined-input model. This represents a high-impact reduction in input requirements for FACTS. By relying only on TCR data, equivalent performance can be obtained with less data, lowering both the cost and complexity of diagnosis in practical use cases.

Benchmarking dense encoders for multi-disease classification

We next evaluated the performance of our in-house trained TCR encoder, along with multiple publicly available sequence encoders, on the MAL ID dataset. For this evaluation, we varied only the encoder, keeping the rest of the classification pipeline unchanged. For BCR only models, the pipeline with IgBERT as sequence encoder achieved the highest performance (accuracy: 82.0% ± 3.7%; ROC AUC: 0.9736 ± 0.0092), outperforming Antiberta2, AntiBERTy, ESM-2, ProtTrans, and AbLang. For TCR only models, the training pipeline with our TCR encoder achieved the highest performance (accuracy: 0.8819 ± 0.0562; ROC AUC: 0.9781 ± 0.0111; F1 score: 0.8576 ±

0.0644), outperforming TCRBert, ProtTrans, ESM-2, and SCEPTR (Table 1). These results highlight the strength of dense encoders for immune receptor modeling and demonstrate that our TCR encoder establishes a strong new baseline for this task.

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 S3). An interactive 3D tSNE plot is provided as Supplementary Table S4. From this analysis, a batch effect was observed in the autoimmune hepatitis (AIH) dataset (Supplementary Figure S4), likely reflecting differences in experimental protocols among publicly available studies, as detailed in Supplementary Note S1.

Restricting the analysis to the Mal-ID TCR subset (n = 499 donors), which was collected and processed using a standardized protocol, obvious batch effects were reduced and while strong disease group separation was maintained (Figure 4Bb; Supplementary Table S4). Our model consistently achieved high classification accuracy and ROC AUC values when applied to either the full 1,421-donor dataset or the Mal-ID subset, demonstrating the robustness of the approach across diverse data sources (Supplementary Table S2).

Beyond disease classification, the learned feature space preserved biologically meaningful structures. Unsupervised t-SNE projection revealed a clustering pattern suggestive of age associations, particularly among SLE and healthy donors, where 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). Similar, although not statistically significant, trends were observed in other disease groups, suggesting that paratope features capture clinically relevant immune signatures beyond discrete disease classification.

4 Discussion

In the field of immunology, ML has primarily been applied to molecular-level problems such as MHC allele prediction[42] and TCR clustering[43]. In the latter case, paratopes, comprising the CDR1, CDR2, and CDR3 regions, represent AIR residues capable of directly interacting with antigens. Previous work has shown that paratope sequences can reflect convergent immune responses across individuals[13], and that BCR clusters are highly antigen- and epitope-specific[12]. Other studies have demonstrated that antibodies binding the same epitope converge in paratope space and can be identified using deep learning predictors such as ParaPred[44] and structure-aware tools like Ab-Ligity[45].

More recently, ML methods have been applied to systems-level problems, including disease classification from AIR repertoire data. Previously, we showed that paratope networks enabled robust binary disease classification, outperforming clonotype-based methods. However, paratope networks require extensive computational resources,

253 precluding their application to larger datasets. In the present study, we introduce FACTS, a much more scalable,
254 framework for multi-disease classification using AIR sequences. By reversing the order of clustering and similarity
255 comparison in our previous pipeline, we improved training runtime without apparent loss in accuracy. Rather than
256 computing pairwise similarities first, we encoded sequences into vector representations, then performed clustering in
257 this reduced space. We also introduced our own MyImmune TCR paratope encoder, which out-performed alternatives
258 in the context of disease classification. The embeddings, generated using protein language models trained on natural
259 amino acid sequences, captured biochemical and structural properties. Such embeddings have recently been used in
260 various immune repertoire analyses[30,34,38,39,45]. MiniBatchKMeans clustering was then applied to group encoded
261 paratope sequences, enabling efficient processing of up to hundreds of millions of sequences. Donor-level features
262 were derived from cluster frequency distributions and inter-cluster mutual information, which were subsequently used
263 for classification.
264
265 Benchmarking demonstrated that the MyImmune TCR paratope encoder further enhanced classification performance.
266 Compared with multiple publicly available TCR protein language models, MyImmune consistently achieved the
267 highest accuracy in classification of MAL-ID dataset, reinforcing the value of paratope-aware representation learning
268 in repertoire-based diagnostics. Notably, using only TCR input, FACTS achieved accuracy comparable to a Mal-ID
269 model that used both TCR and BCR input. This finding further suggests that reliable classification may be achievable
270 using data from only one receptor type, which could significantly reduce diagnostic costs. Taken together, the
271 integration of FACTS and the MyImmune encoder produced a fast, scalable, and highly accurate pipeline for immune
272 repertoire-based disease classification.
273
274 When applied to a larger TCR dataset of 1,421 donors spanning 16 disease classes, the FACTS framework
275 demonstrated consistently strong classification performance, although batch effects were in the AIH dataset, which
276 likely reflects variability in experimental protocols across the cohorts. Despite this, FACTS achieved high
277 classification accuracy not only in the large, heterogeneous dataset, but also in the smaller, more homogeneous
278 dataset where experimental protocols were more consistent and batch effects were minimal. This consistency
279 across diverse conditions indicates that our model is robust against batch effects. These findings emphasize the
280 importance of validating diagnostic models using well-matched, prospectively collected clinical datasets under
281 standardized conditions to ensure reliability in real-world applications. Overall, our findings show that paratope-
282 encoded representations, when combined with the FACTS pipeline and MyImmune encoder, enable accurate, scalable,
283 and biologically meaningful classification of immune repertoires, supporting their potential for clinical translation and
284 population-scale immune profiling.
285
286 **Contributors**
287 N.J., A.S., F.H., and D.M.S. conceived and designed the study.
288 A.S., S.L., D.U., and M.A.L. acquired the data and performed the formal analysis.
289 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 S1. 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.

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 [JP25am0101001], 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. Hoof I, Peters B, Sidney J, et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. Immunogenetics 2009; 61(1): 1–13.

43. Sidhom J-W, Larman HB, Pardoll DM, Baras AS. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. Nature Communications 2021; 12(1): 1605.

44. 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.

45. 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.

Legends

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 reordering 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 five 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 five 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 S1. **(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.

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 $p < 0.0001$.

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 1

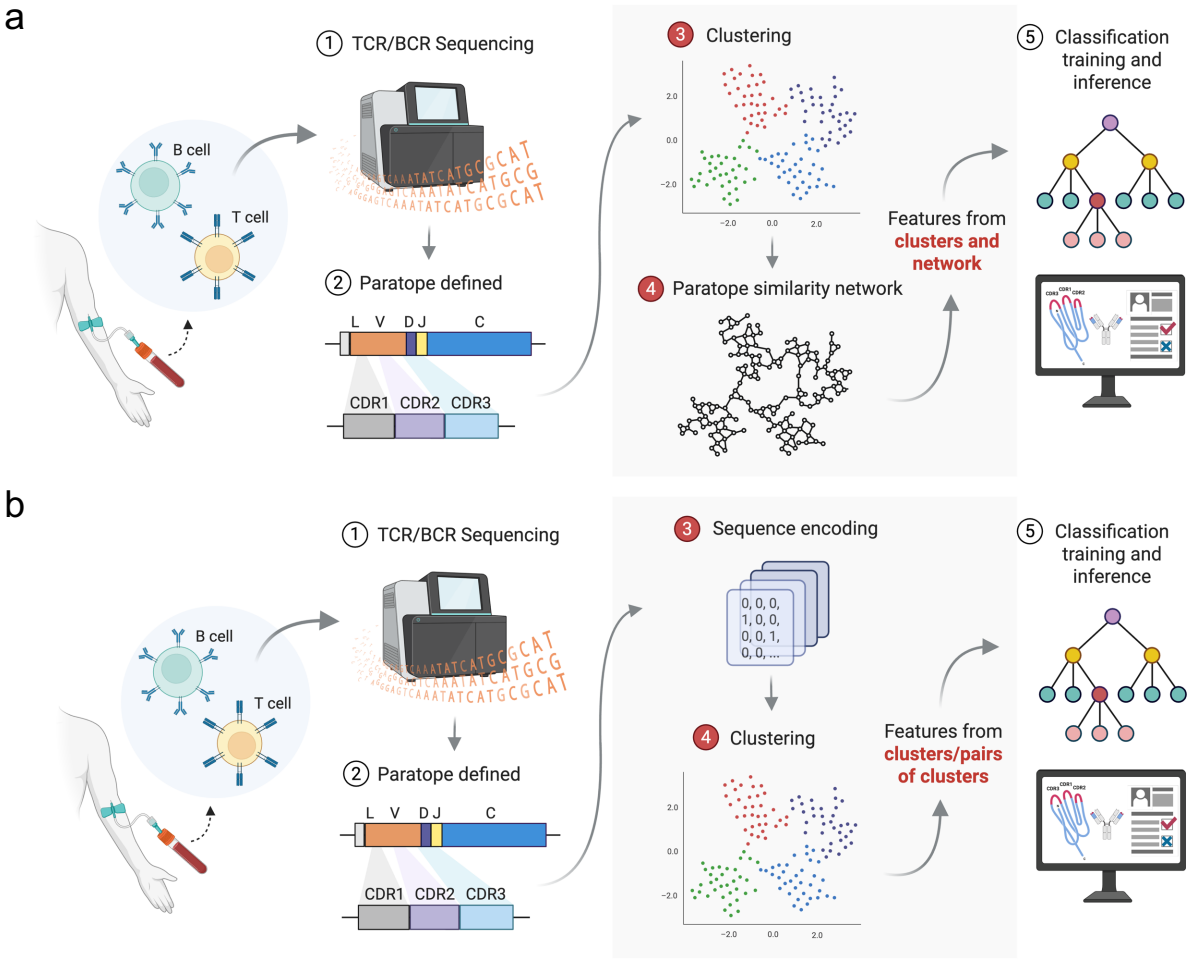
a
B cell
T cell
1 TCR/BCR Sequencing
2 Paratope defined
L V D J C
CDR1 CDR2 CDR3
3 Clustering
4 Paratope similarity network
Features from clusters and network
5 Classification training and inference
b
B cell
T cell
1 TCR/BCR Sequencing
2 Paratope defined
L V D J C
CDR1 CDR2 CDR3
3 Sequence encoding
0, 0, 0,
1, 0, 0,
0, 0, 1,
0, 0, ...
4 Clustering
Features from clusters/pairs of clusters
5 Classification training and inference

Figure 2

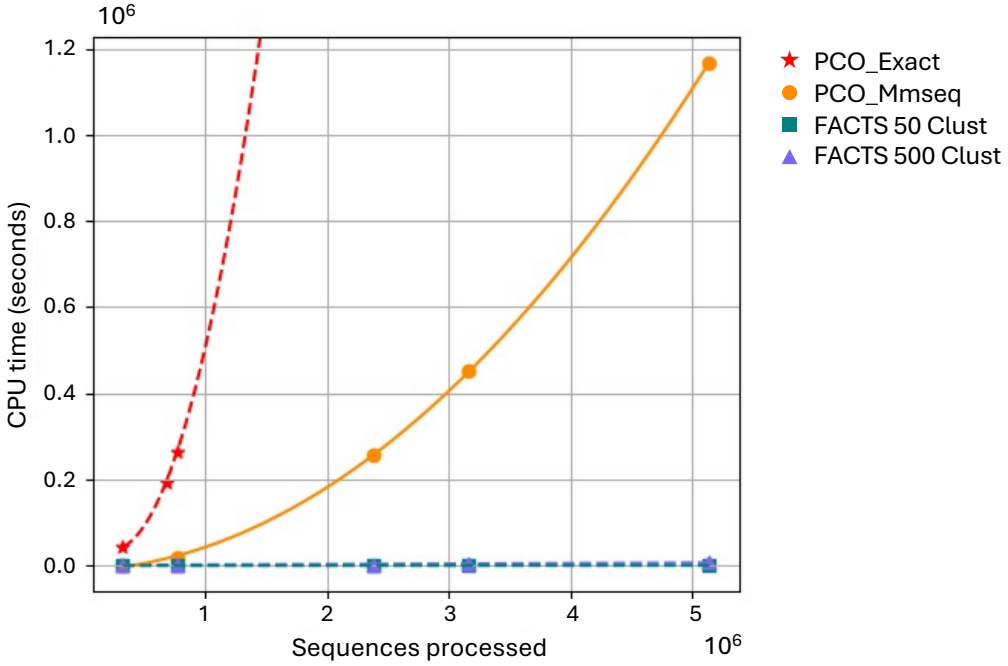
10^6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
CPU time (seconds)
1
2
3
4
5
Sequences processed
10^6
PCO_Exact
PCO_Mmseq
FACTS 50 Clust
FACTS 500 Clust

Figure 3

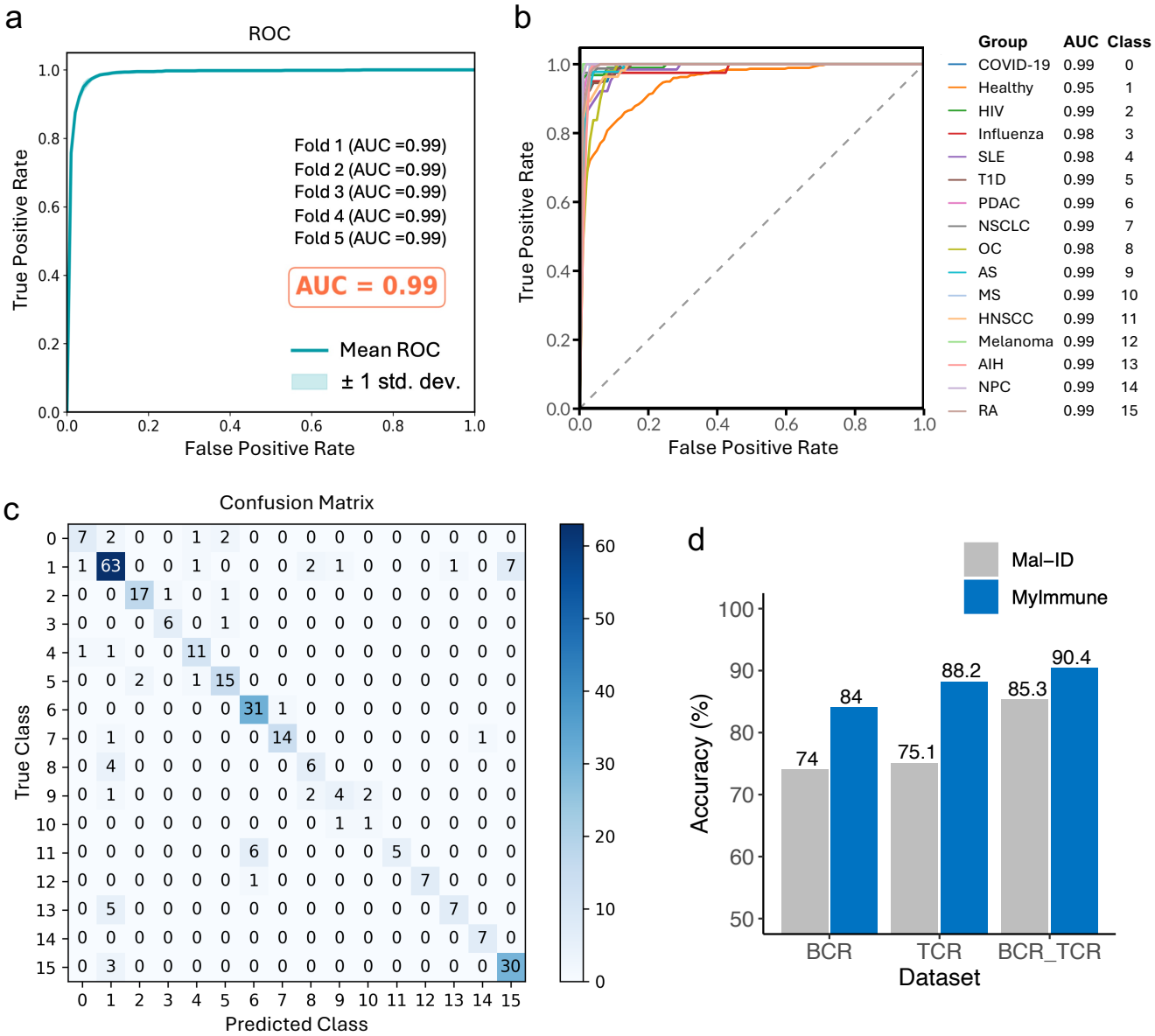
a
ROC
True Positive Rate
False Positive Rate
Fold 1 (AUC =0.99)
Fold 2 (AUC =0.99)
Fold 3 (AUC =0.99)
Fold 4 (AUC =0.99)
Fold 5 (AUC =0.99)
AUC = 0.99
Mean ROC
± 1 std. dev.
b
True Positive Rate
False Positive Rate
Group AUC Class
COVID-19 0.99 0
Healthy 0.95 1
HIV 0.99 2
Influenza 0.98 3
SLE 0.98 4
T1D 0.99 5
PDAC 0.99 6
NSCLC 0.99 7
OC 0.98 8
AS 0.99 9
MS 0.99 10
HNSCC 0.99 11
Melanoma 0.99 12
AIH 0.99 13
NPC 0.99 14
RA 0.99 15
c
Confusion Matrix
True Class
Predicted Class
d
Mal–ID
MyImmune
Accuracy (%)
74
84
75.1
88.2
85.3
90.4
BCR
TCR
BCR_TCR
Dataset

Figure 4

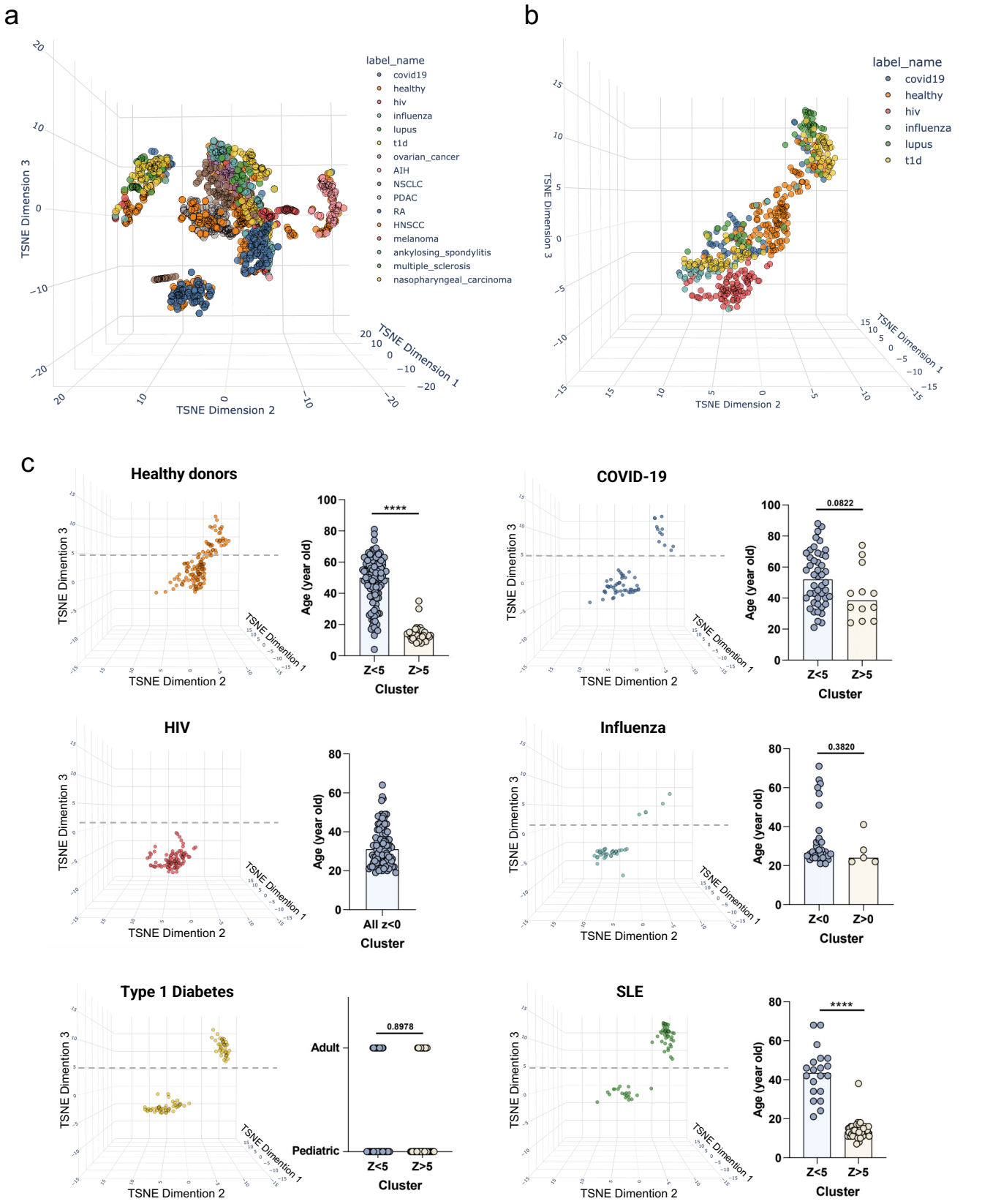
a
b
c
label_name
covid19
healthy
hiv
influenza
lupus
t1d
ovarian_cancer
AIH
NSCLC
PDAC
RA
HNSCC
melanoma
ankylosing_spondylitis
multiple_sclerosis
nasopharyngeal_carcinoma
TSNE Dimension 1
TSNE Dimension 2
TSNE Dimension 3
Healthy donors
COVID-19
HIV
Influenza
Type 1 Diabetes
SLE
Age (year old)
Cluster
Z<5
Z>5
Z<0
Z>0
All z<0
Adult
Pediatric
****
0.0822
0.3820
0.8978

Table 1

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	**0.8819 ± 0.0562**	**0.9781 ± 0.0111**	**0.8576 ± 0.0644**
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