# BenchRep-T: A Systematic Evaluation of T-Cell Repertoire-Based Disease Diagnostics

Alejandro Buendia*, Chiho Im*, Anshul Kundaje, Scott D. Boyd, Liel Cohen-Lavi

NeurIPS 2026 Evaluations & Datasets

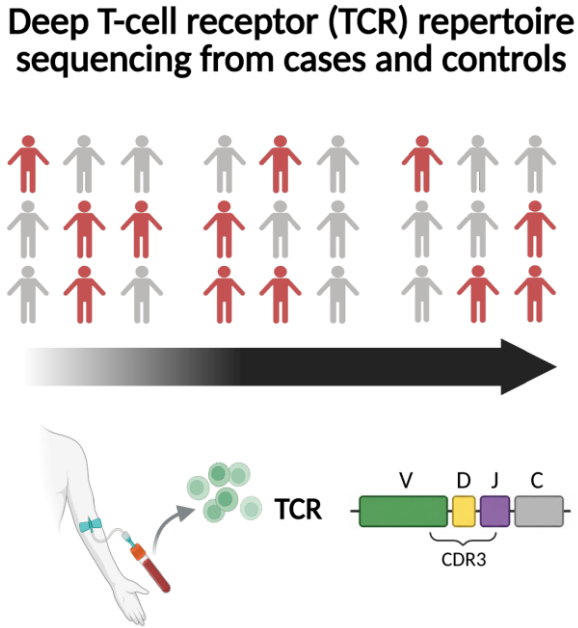
*Equal contribution



# Can a blood draw diagnose disease?

T-cell receptor (TCR) repertoires record a history of past and current immune exposures

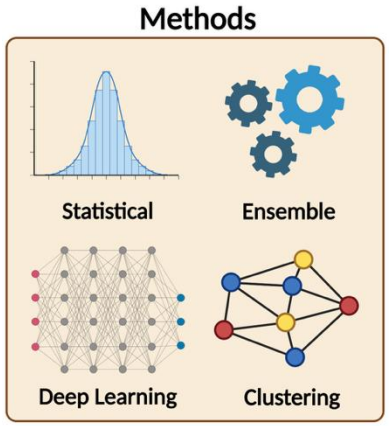
- **V(D)J recombination** gives each T cell a near-unique receptor; antigen recognition expands it into a clone
- **Each person carries $10^8$–$10^{10}$ distinct TCR clonotypes** (Greiff et al., 2020), with only a small fraction shared between individuals
- **Multiple instance learning (MIL):** each repertoire is a bag of sequences labeled only at the bag level, with disease signal in a small, unknown subset of TCRs
- **Published methods use different cohorts, preprocessing, and metrics,** and have not been rigorously benchmarked and compared directly




Stanford MEDICINE

# BenchRep-T: Benchmarking nine methods on ten cohorts

Nine methods (classified into four method families) run on identical inputs and data splits

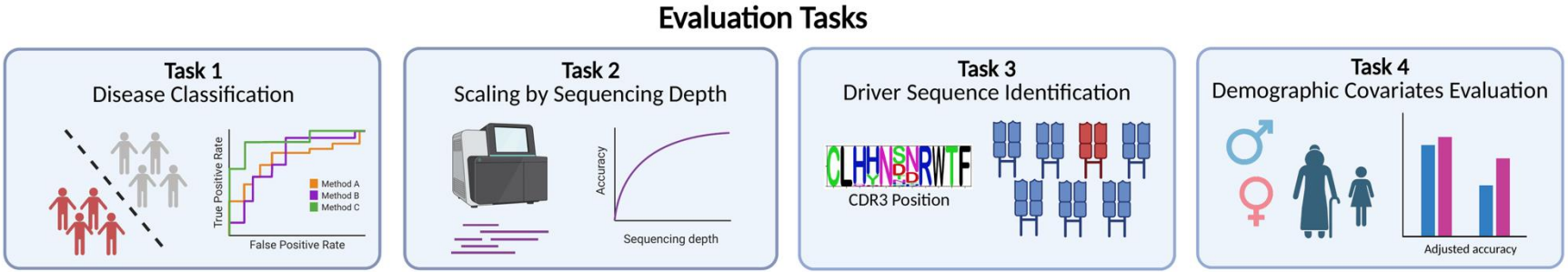



| Method | V/J gene | CDR3 sequence similarity | Aggregation strategy | Depth filter |
|---|---|---|---|---|
| *Statistical* | | | | |
| Emerson et al. | ✓ | Global | Clonotype burden | – |
| Ostmeyer et al. | – | Local | Max motif score | – |
| *Ensemble* | | | | |
| VJ-kmer (LogReg) | ✓ | Local | Frequency vector | – |
| VJ-kmer (XGBoost) | ✓ | Local | Frequency vector | – |
| Mal-ID | ✓ | Global | Varied | $\geq 10^3$ |
| *Deep Learning* | | | | |
| DeepRC | – | Local | Attention pooling | $\leq 10^4$ |
| DeepTCR | ✓ | Local | Weighted pooling | – |
| ABMIL | ✓ | Local | Attention pooling | $\leq 10^4$ |
| *Clustering* | | | | |
| GIANA | V only | Global | Cluster mean | $\leq 10^4$ |

- 10 public cohorts: **2,352 TCRβ repertoires** from 2,264 individuals
- **8 conditions**: HIV, lupus (SLE), COVID-19, influenza, tuberculosis (TB) progression, rheumatoid arthritis (RA), cytomegalovirus (CMV), and type 1 diabetes (T1D)
- **Two sequencing assays**: cDNA with paired-end Illumina MiSeq (Mal-ID) and immunoSEQ genomic DNA (Adaptive)
- Fixed, participant-disjoint 3-fold splits reused for every task

Stanford MEDICINE

# BenchRep-T: Four complementary evaluation tasks




Can methods reliably distinguish disease cases from controls?

Can methods classify disease under downsampled sequencing depth?

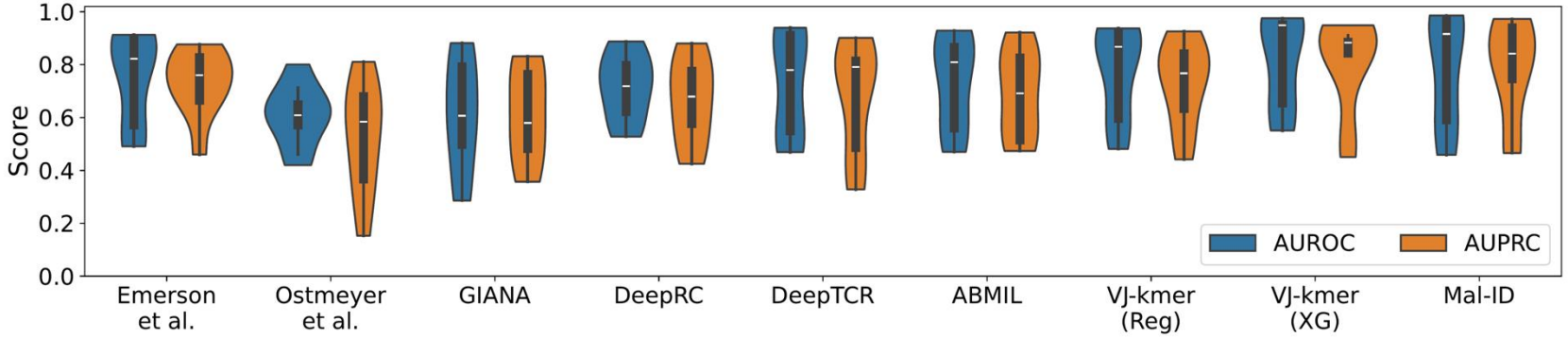
Do methods prioritize antigen-specific TCRs near the top of each repertoire?

Are methods detecting disease or demographic differences between cases and controls?

Stanford MEDICINE

# Task 1: Disease classification overall

AUROC and AUPRC reported for nine within-cohort binary classification tasks

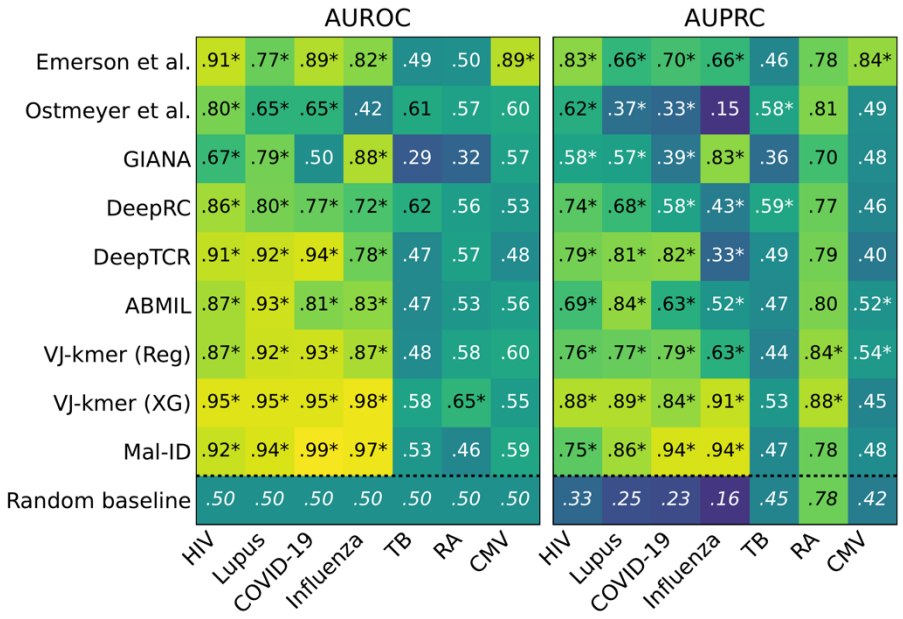



- **Ensembles ranked at or near the top:** VJ-kmer (XG) had the highest mean AUROC (0.81) and AUPRC (0.80), followed by Mal-ID (0.79, 0.78)
- Ostmeyer et al. and GIANA were lowest (mean AUROC 0.61)
- Matched-budget hyperparameter tuning and leveraging pretrained encoders (ESM-2, TCR-BERT, TCR-VALID) for the deep MIL methods did not close the gap to the ensembles

Stanford MEDICINE

# Task 1: Disease classification by task

Seven within-cohort settings; * = 95% bootstrap CI above the random baseline

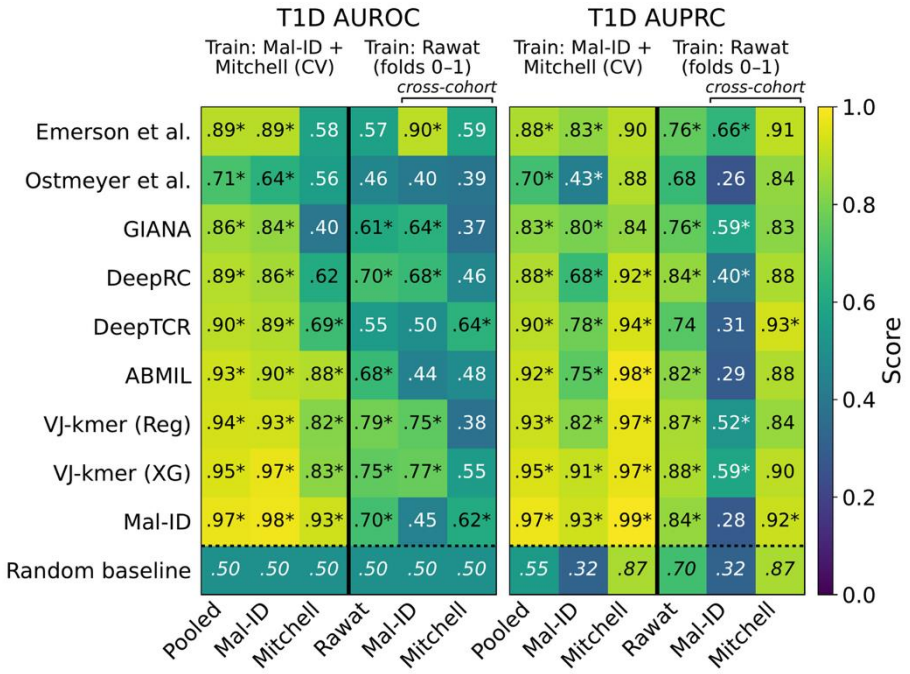



- **VJ-kmer (XG)** was best on HIV (0.95), lupus (0.95), and RA (0.65); **Mal-ID** was best on COVID-19 (0.99)
- **Only Emerson et al. performed well on CMV** (0.89 vs. ≤0.60 for all others), the cohort it was developed on
- **RA and TB were the hardest tasks**: weak repertoire signal and small cohorts (91 RA, 140 TB specimens) limit performance for these conditions

Stanford MEDICINE

# Task 1: Methods struggled to transfer to new cohorts

Zero-shot evaluation for type 1 diabetes using three independent cohorts

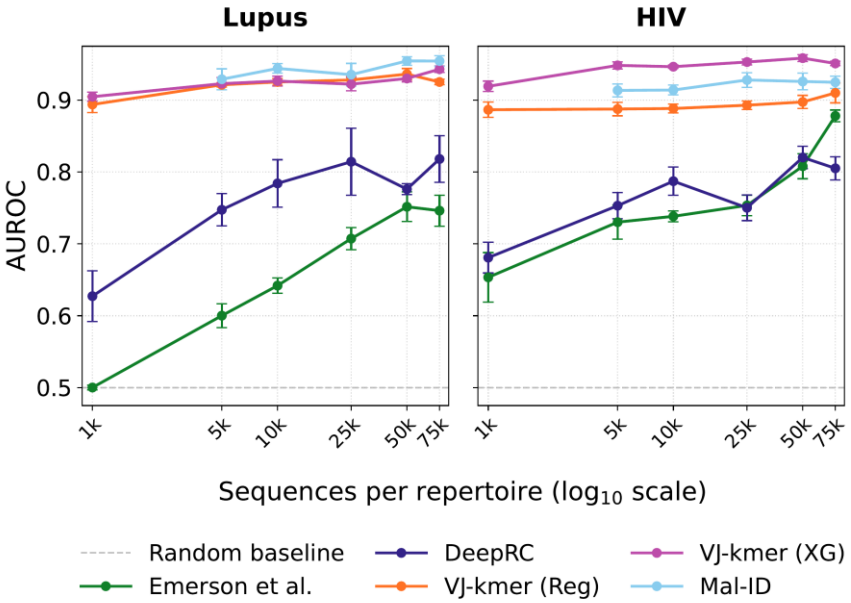



- **Within cohorts:** every method beat chance on pooled Mal-ID + Mitchell CV (Mal-ID best, AUROC 0.97)
- **When trained on Rawat and applied zero-shot:** 5 of 9 methods transferred above chance to Mal-ID, led by Emerson et al. (AUROC 0.90)
- **Transfer to Mitchell largely failed** despite its shared assay platform with Rawat (best AUROC 0.64)
- The transfer leaders were near chance on held-out Rawat, and no method beat chance on both cohorts

Stanford MEDICINE

# Task 2: Disease classification under subsampled depth

AUROC after downsampling lupus and HIV repertoires; mean ± SE over 5 nested replicates

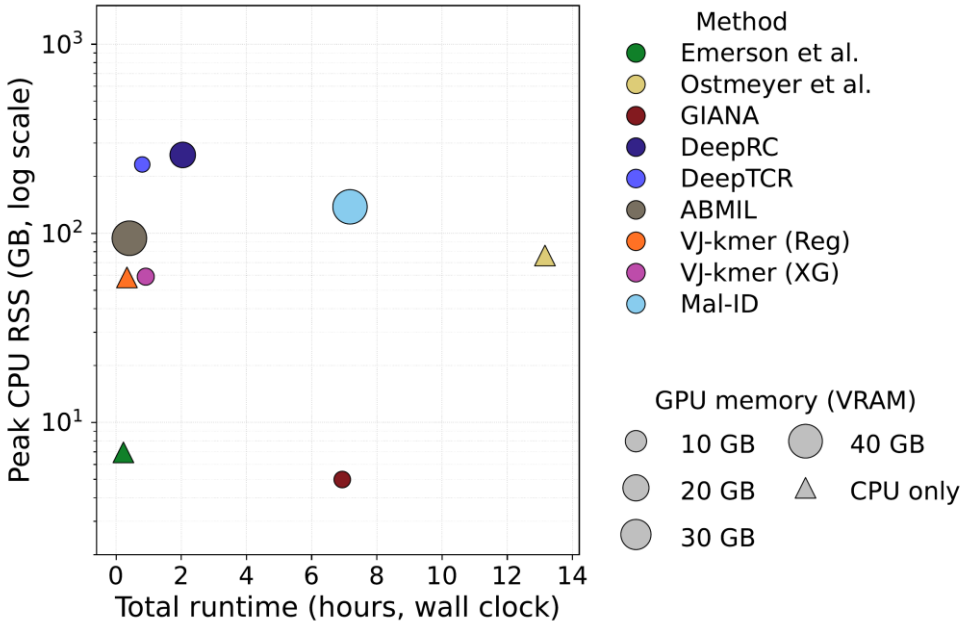



- **VJ-kmer and Mal-ID dropped <0.04 AUROC** from 75k down to 1k sequences (5k for Mal-ID)
- V/J-gene usage is stable under subsampling and alone reaches high AUROC on the Mal-ID cohorts
- **Emerson et al. was most dependent on depth:** at chance for 1k sequences on lupus, still improving at 75k on HIV
- DeepRC was intermediate, with a non-monotonic trend

Stanford MEDICINE

# Task 2: Memory usage and runtime vary widely for methods

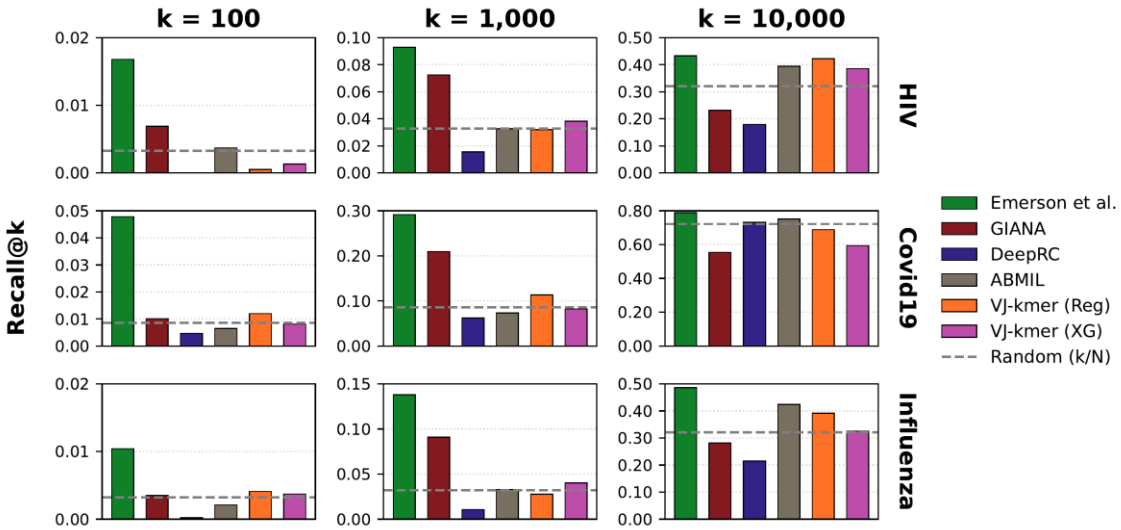
End-to-end runtime and peak memory usage (CPU RSS and GPU VRAM) for one HIV fold




Stanford MEDICINE

# Task 3: Recovery of antigen-specific driver sequences

Recall@k of known antigen-specific TCRs; dashed line = random ranking

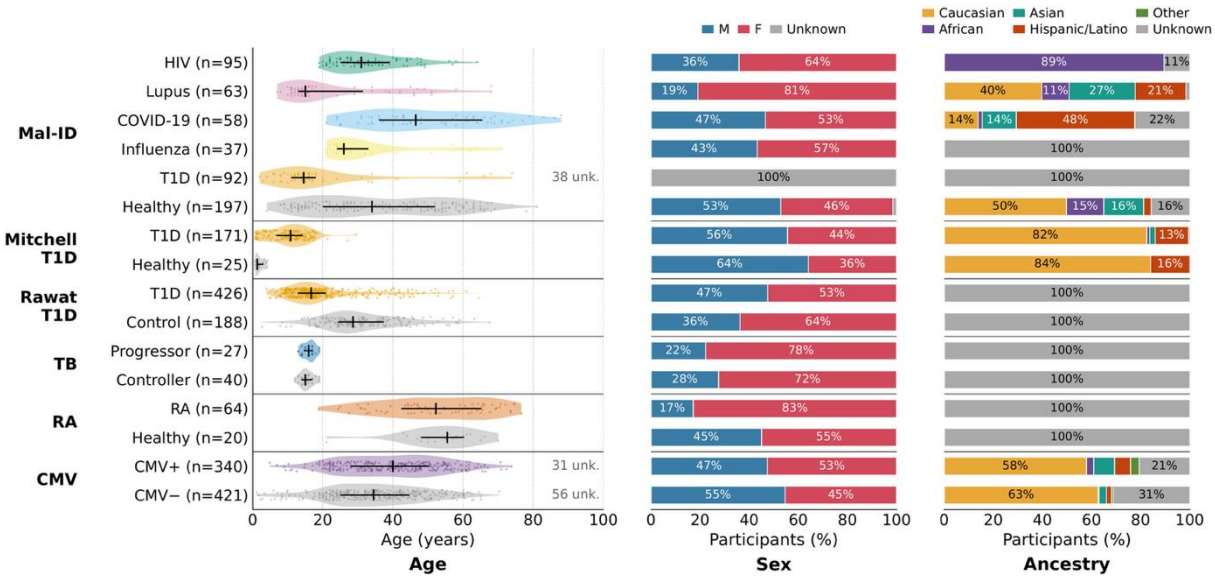



- **Driver sequences**: CDR3 sequences ≥90% similar to VDJdb references (0.02–1.4% of each repertoire)
- **Emerson et al. ranked drivers best,** clearly above random at k = 1,000 for all three diseases
- GIANA performed strongly at lower k (100 and 1,000) but below random at k = 10,000
- Emerson et al. performs particularly well: it looks for identical sequences enriched across patients, which resemble published drivers

Stanford MEDICINE

# Task 4: Existing cohorts show heavy demographic imbalances

Age, sex, and self-reported ancestry of cases and controls

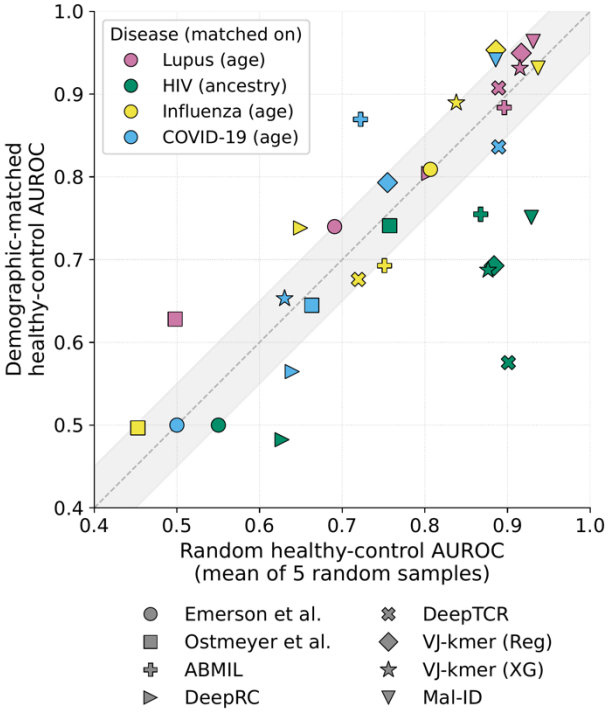



- Disease and demographic signal can be entangled through cohort selection or biology
- Some imbalances reflect disease biology: Mal-ID lupus cohort is 81% female
- Others reflect cohort selection: 89% of the Mal-ID HIV cohort reports African ancestry
- Demographics often incomplete: ancestry unrecorded for most immunoSEQ cohorts

Stanford MEDICINE

# Task 4: Matched controls lower HIV AUROC

AUROC with random vs. demographic-matched healthy controls; gray band = within 0.05

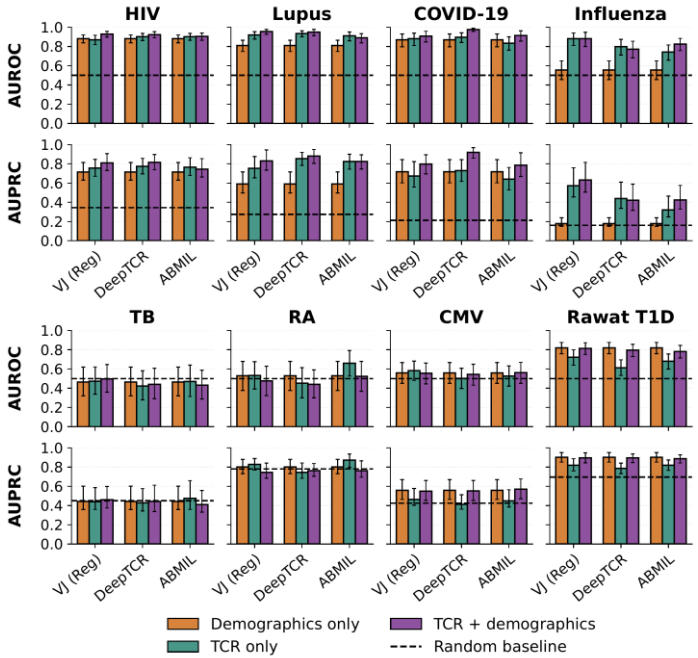



- Random healthy controls replaced with controls matched on the dominant confounder for each Mal-ID cohort: age for lupus, influenza, and COVID-19; ancestry for HIV
- **Ancestry matching lowered HIV AUROC by 0.11–0.33** for six of eight methods
- **Age matching had mixed effects**: lupus mostly unchanged or higher; influenza and COVID-19 moved both ways

Stanford MEDICINE

# Task 4: Demographics-only models can match methods

Demographics only vs. TCR only vs. TCR + demographics, on the specimens with full metadata




- **TCRs added clear signal only on lupus** (+0.10–0.13 AUROC) **and influenza** (+0.19–0.33)
- **HIV and COVID-19:** no TCR model beat the demographics-only baseline
- **Rawat T1D:** demographics alone (0.82) beat all three TCR models (0.61–0.72)
- TB, RA, and CMV: demographics uninformative (0.46–0.56), and no model differed from them significantly

Stanford MEDICINE

# Looking forward: recommendations for data generation

Many performance limitations reflect how repertoire cohorts are currently collected and shared

1. **Standardize releases:** use an AIRR-compliant schema with fixed, published data splits
2. **Include metadata:** report age, sex, ancestry, and HLA type, so confounding can be measured and controlled
3. **Share openly:** include raw and processed data with metadata and a datasheet in persistent open repositories
4. **Report sequencing depth:** methods differ in depth robustness and depth can differ between cases and controls
5. **Scale up cohorts:** collect specimens at adequate sequencing depth, more independent participants across centers, and cohorts balanced on age, sex, and ancestry
6. **Record batch:** record processing batch, sequencing run, and collection site
7. **Mix batches:** process and sequence cases and controls together so batch is not confounded with the disease label

Stanford MEDICINE

# Conclusions

- **No single method wins everywhere:** Methods should be compared across multiple diseases, cohorts, and tasks
- **Simple feature-engineered ensembles are competitive:** Gradient-boosted trees on V/J-gene usage and k-mers matched or beat deep MIL and were the most depth-robust methods
- **High within-cohort AUROC is not enough:** High within-cohort performance did not imply cross-cohort transfer, recovery of antigen-specific TCRs, or robustness to demographic confounding
- **Progress depends on data collection as much as models:** Methods are dependent on high-quality TCR repertoire datasets, collected for demographically balanced cohorts

Paper: https://biorxiv.org/content/10.64898/2026.06.09.727013v2.full.pdf
Code: https://github.com/boyd-lab/BenchRep-T
Data: https://huggingface.co/datasets/lordim/BenchRep-T

Stanford MEDICINE