Repertoire-Level Diversity, Not *De Novo* Public Clonotypes, Drives Cytomegalovirus Serostatus Prediction from TCRβ Immunosequencing at Modest Cohort Size: A Leakage-Free Machine-Learning Analysis of the Emerson (2017) Discovery Cohort

K-Dense Web
K-Dense Research
correspondence: contact@k-dense.ai

July 10, 2026

Abstract

Background. Cytomegalovirus (CMV) is a ubiquitous, lifelong herpesvirus that leaves a durable imprint on the T-cell receptor (TCR) repertoire. Emerson and colleagues (2017) showed that CMV serostatus can be inferred from peripheral-blood TCRβ immunosequencing using public, CMV-associated CDR3β clonotypes, given a large discovery cohort (∼640 subjects). We asked how far this signal survives when the cohort and per-subject sequencing depth are reduced, and which feature family carries the reliable signal under those constraints.

Methods. From the Emerson (2017) discovery repertoires (immuneACCESS; distributed via the DeepRC resource) we selected a demographically matched, balanced subset of 200 subjects (100 CMV+, 100 CMV−; 50 female / 50 male per group; per-pair age difference = 0; no sequencing-depth confound, Mann–Whitney $p = 0.81$). We built three feature families: (i) a *de novo* **burden score** (count of CMV-associated clones present per subject), (ii) a sparse binary presence matrix of the **top-200 CMV-associated CDR3β** clonotypes, and (iii) seven **global repertoire summary statistics** (Shannon entropy, Simpson diversity, clonality, CDR3 length mean/SD, maximum clone frequency, top-100 cumulative frequency). Evaluation used strict subject-level 5-fold cross-validation (no subject in both train and test; each subject tested once). Crucially, all label-dependent steps—public-clone selection by one-sided hypergeometric (Fisher) enrichment ($p < 10^{-4}$, incidence ≥ 4) and the top-N cut—were performed *inside training folds only*. We report AUROC, AUPRC against the 0.50 CMV+ prevalence baseline, and sensitivity at 90% specificity, and we trace a depth-subsampling degradation curve by multinomial down-sampling to 5,000–100,000 templates per subject.

Results. At full depth, the *de novo* public-clonotype models were near chance (burden AUROC 0.525; L1 0.510; L2 0.532), whereas the global-repertoire baseline reached **AUROC** 0.766**, AUPRC** 0.736 **(baseline** 0.50**), and sensitivity** 0.31 **at an achieved** 90% **specificity**. On average only 9.4 public clones per fold passed $p < 10^{-4}$ (union of 764 across folds; one clone, `CASSPGLAGVYNEQFF`, recurred in all five folds). CMV+ repertoires were significantly less diverse and more clonally expanded (Simpson $p = 1 \times 10^{-12}$; clonality $p = 5 \times 10^{-11}$). The global-repertoire signal was strikingly depth-robust—AUROC ≈ 0.75–0.77 across a 60-fold depth range down to 5,000 templates—while public-clonotype enrichment collapsed to 0 significant clones at ≤ 25,000 templates.

Conclusions. At $n = 200$, reliable CMV information lives in *global repertoire structure* (diversity/clonality), not in *de novo* public clonotypes, which require both large cohorts and

deep repertoires to generalize. Diversity-based features are far more sample- and depth-efficient, tolerating an order-of-magnitude reduction in sequencing depth with negligible loss. This is a statement about *scale*, not a refutation of the public-clonotype paradigm: the public-TCR signal is expected to recover with the larger cohorts and deep repertoires of the original ∼640-subject study. Per-subject predictions, discovered clonotypes, and code are released for full reproducibility.

Keywords: T-cell receptor repertoire; cytomegalovirus; immunosequencing; machine learning; cross-validation; repertoire diversity; public TCR; data leakage.

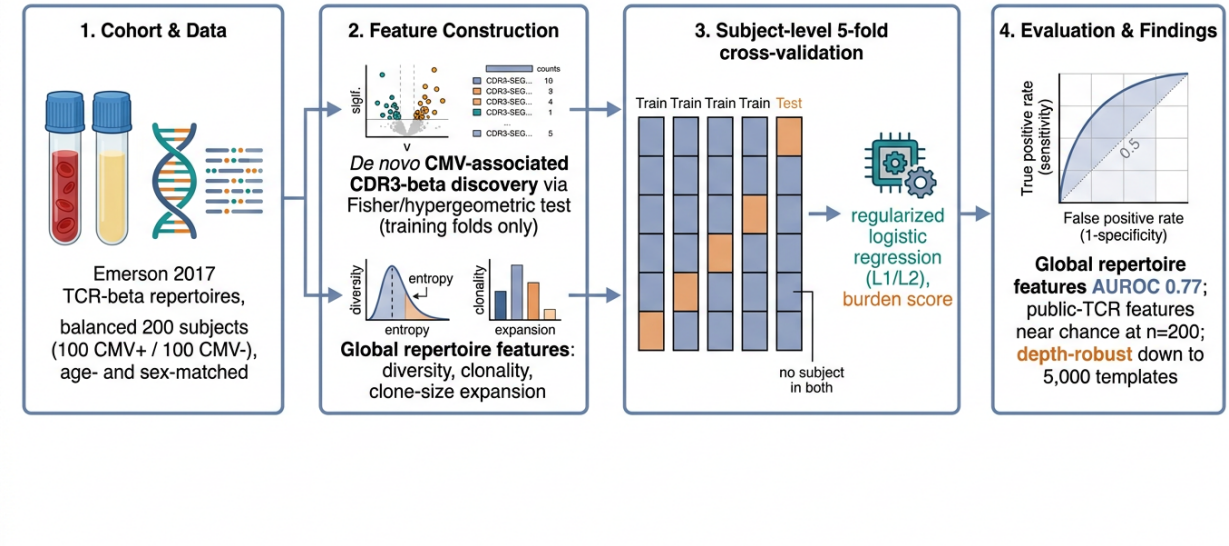

Figure 1: Graphical abstract. The analytical workflow: (1) a balanced, demographically matched subset of 200 subjects (100 CMV+/100 CMV−) is drawn from the Emerson (2017) TCRβ discovery repertoires; (2) two feature families are constructed—*de novo* CMV-associated CDR3β clonotypes discovered by Fisher/hypergeometric enrichment (in training folds only) and global repertoire summaries (diversity, clonality, clone-size expansion); (3) strict subject-level 5-fold cross-validation trains regularized logistic-regression and burden-score classifiers with no subject shared between train and test; (4) evaluation shows that global repertoire features reach AUROC 0.77 while public-TCR features remain near chance at $n = 200$, and the global signal is robust down to 5,000 templates per subject.

1 Introduction

Human cytomegalovirus (CMV; human herpesvirus 5) is among the most prevalent chronic infections of humankind. Meta-analyses of serosurveys place global seroprevalence at roughly 60–90%, rising with age and inversely with socioeconomic status, such that the majority of adults worldwide carry the virus for life [1, 2]. In immunocompetent hosts primary infection is usually subclinical and is followed by lifelong latency punctuated by subclinical reactivation [3]. The clinical weight of CMV is therefore concentrated at the extremes of immune competence: it is the classical "troll of transplantation," a leading cause of morbidity, graft dysfunction, and indirect immunomodulatory

effects after solid-organ and hematopoietic-cell transplantation [4, 5], and it is the single most important infectious cause of congenital neurological disability worldwide [6, 7]. Even in healthy ageing, chronic CMV is a dominant driver of T-cell immunosenescence, progressively expanding large clonal populations of terminally differentiated $CD8^+$ T cells at the expense of naive diversity [8, 9]. Knowing an individual's CMV serostatus is thus clinically consequential—most acutely for donor/recipient matching and risk stratification in transplantation.

Because CMV chronically stimulates the adaptive immune system, it reshapes the T-cell receptor (TCR) repertoire in ways that are, in principle, directly observable in blood. The TCRβ chain, whose hypervariable complementarity-determining region 3 (CDR3β) is generated by V(D)J recombination, provides an essentially unique molecular barcode for each T-cell clone. High-throughput immunosequencing (immunoSEQ) quantifies millions of these clonotypes per sample with reproducible, bias-controlled template counts [10, 11]. Two repertoire-level consequences of chronic CMV are well documented. First, CMV drives *memory inflation*: CMV-specific $CD8^+$ clones accumulate to extraordinary frequencies and persist for decades, so that seropositive repertoires are, in aggregate, more *clonally expanded* and *less diverse* [8, 9]. Second, because certain CDR3β rearrangements are generated convergently across individuals and are selected by shared CMV epitopes presented on common HLA alleles, CMV leaves a reproducible set of *public* clonotypes that recur across seropositive donors [11, 12].

The landmark study of Emerson et al. [13] exploited the second phenomenon to build a diagnostic classifier. Using deep TCRβ immunosequencing of a discovery cohort of $\sim$640 subjects with known serostatus, they identified public CDR3β sequences statistically enriched among CMV+ donors and combined their incidence into a burden-style classifier that predicted serostatus with approximately 0.89 sensitivity and specificity, replicated in an independent validation cohort. This established a powerful and now widely emulated paradigm: disease-associated public TCR signatures can serve as blood-based diagnostics of immune history [12, 13]. The same occurrence-pattern framework was subsequently generalized to infer HLA type and other exposures [12], and modern deep-learning architectures (e.g., attention-based multiple-instance learning) have been applied to the identical CMV benchmark [14].

A practical question, however, is under-examined: *how much of this predictive power depends on the scale of the experiment*—the number of labelled subjects available for discovery and the sequencing depth of each repertoire? The public-clonotype approach is statistical: a candidate CDR3β can only be flagged as CMV-associated if it is seen in enough donors to reach significance, and it can only be *observed* in a donor if that donor's repertoire is sequenced deeply enough to sample it. Both requirements degrade gracefully in theory but may fail abruptly in practice. Public clonotypes are individually rare, so their detection is doubly sample-limited—by cohort size and by per-subject depth. In contrast, *global* repertoire summary statistics such as Shannon entropy [15], Simpson diversity, and clonality are aggregate estimands that are comparatively stable and can be estimated from a modest number of templates [16]. Which family of features carries the reliable CMV signal when data are limited is therefore not obvious a priori and has direct implications for the design of cost-constrained repertoire diagnostics.

In this work we re-analyse the Emerson (2017) discovery repertoires under deliberately constrained conditions and with a rigorously leakage-free evaluation protocol. Our contributions are fourfold. (1) We assemble a *balanced, demographically matched* subset of 200 subjects in which sex is identically distributed, ages are matched pair-for-pair, and sequencing depth is not confounded with serostatus, removing the most common demographic and technical shortcuts to prediction. (2) We implement a strict *subject-level* cross-validation pipeline in which *de novo* public-clonotype discovery—candidate

filtering and one-sided hypergeometric (Fisher) enrichment testing—is performed entirely within training folds, eliminating the selection-on-the-full-dataset leakage that silently inflates repertoire-classifier performance. (3) We directly compare *de novo* public-clonotype features (a burden score and L1/L2-regularized logistic regression on the top-200 associated clones) against a global-repertoire baseline, reporting AUROC, AUPRC relative to prevalence, and sensitivity at 90% specificity. (4) We quantify robustness with a *depth-subsampling degradation curve*, multinomially down-sampling every repertoire to fixed depths from 5,000 to 100,000 templates and re-running the entire pipeline at each depth. The central, reproducible finding is that at this cohort size the reliable CMV signal resides in global repertoire structure, not in *de novo* public clonotypes—a conclusion with concrete consequences for how, and how cheaply, TCR-based CMV assays can be built.

2 Methods

2.1 Data source and provenance

The Emerson (2017) discovery cohort was generated by bias-controlled multiplex-PCR immunoSEQ of peripheral-blood TCRβ CDR3 regions [10, 13]. The canonical raw data are distributed through the Adaptive Biotechnologies immuneACCESS portal (`Emerson-2017-NatGen`; doi:10.21417/B7001Z), which is session/login-gated and does not expose scriptable per-file URLs. We therefore used the openly hosted, pre-processed DeepRC distribution of the identical dataset [14], comprising a subject-level metadata table (subject ID, sex, age, ethnicity, known CMV serostatus, HLA typing) and a single HDF5 container holding all 785 subjects' CDR3 amino-acid sequences with per-clone template counts. The pre-processed distribution retains CDR3 amino-acid sequence and template count but does not retain V/J gene calls; the required information for CMV classification (clonotype identity and abundance) is preserved. The HDF5 was downloaded once; only the 200 selected repertoires were extracted to per-subject tab-separated files (columns `cdr3_amino_acid`, `templates`).

2.2 Cohort selection and demographic balancing

To eliminate demographic and technical confounds that could be exploited as prediction shortcuts, we constructed a balanced, matched subset (seed = 42). Eligibility required a known binary CMV status, presence and validity flags in the HDF5, a known sex, and a numeric age. The selection funnel was 785 $\rightarrow$ 760 (known CMV) $\rightarrow$ 686 (valid samples) $\rightarrow$ 611 eligible. From the eligible pool we fixed exactly 50 female and 50 male subjects in each CMV group, giving identical sex distributions across groups. Within each sex stratum, ages were matched by solving an optimal assignment (`scipy.optimize.linear_sum_assignment`) that minimizes the total absolute age difference between CMV+ and CMV− subjects; the 50 lowest-cost pairs per sex were retained. The resulting cohort of 200 subjects (100 CMV+, 100 CMV−) has an *exact* pairwise age match (per-pair age difference = 0; group age 39.35 ± 11.95 years, range 10–65) and identical sex composition (Table 1). Balance tests confirmed no residual confound: sex χ^2 $p = 1.0$; age Welch t-test $p = 1.0$ and Mann–Whitney $p = 1.0$; and, critically, sequencing-depth (unique clones per subject) Mann–Whitney $p = 0.81$, so serostatus is not confounded with depth (Figure 2). All 200 repertoire files passed quality control (non-empty, required columns present, valid 20-letter amino-acid alphabet, template counts ≥ 1). The subsampling scheme is thus a *demographically matched balanced design at 1:1 prevalence*, chosen so that any predictive performance reflects immunological signal rather than age, sex, or depth artefacts.

2.3 Repertoire representation

Per-subject files were streamed once to build fold-independent, label-agnostic artefacts. A pooled vocabulary of 23,666,730 unique CDR3β amino-acid sequences was indexed across the cohort, and two sparse subject $\times$ clonotype matrices were constructed in compressed-sparse-row format: a binary *incidence* matrix (presence/absence) and a *frequency* matrix (template count normalized by the subject's total). Because these matrices involve no cross-subject label statistic, their construction cannot leak label information. In parallel, seven *global repertoire features* were computed per subject: Shannon entropy [15] and Simpson diversity of the clone-frequency distribution; clonality ($1 - H/\ln(\text{richness})$); template-weighted CDR3 length mean and standard deviation; the maximum single-clone frequency; and the cumulative frequency of the top-100 clones (an expansion index).

2.4 Strict subject-level cross-validation

Every subject contributes exactly one repertoire. We used `StratifiedKFold` ($n_{\text{splits}} = 5$, `shuffle=True`, `random_state=42`) on the 200 subjects, stratified by serostatus, yielding folds of 160 training (80 CMV+) and 40 test subjects. A runtime assertion verified zero train/test subject overlap in every fold, and a post-hoc check confirmed that each subject was assigned to a test fold exactly once. All performance metrics are computed on the concatenated out-of-fold (OOF) predictions, so no subject is ever scored by a model that saw it in training.

2.5 *De novo* discovery of CMV-associated clonotypes (training folds only)

Within each training fold, and using *only* the 160 training subjects, we identified public candidate clonotypes present in at least four training subjects (incidence ≥ 4). Each candidate was tested for enrichment among CMV+ training subjects with a one-sided hypergeometric test (equivalent to a one-sided Fisher exact test): for a clonotype observed in k of the training subjects, of whom x are CMV+, the enrichment p-value is the upper-tail probability $P(X \geq x)$ under $X \sim \text{Hypergeom}(N = 160, K = 80, n = k)$, computed vectorially over $\sim$850,000 candidate clonotypes per fold. We report the count passing $p < 10^{-4}$ and define the fold's **CMV-associated TCR set** as the *top-N* $= 200$ *most significant clonotypes* by ascending training p-value. N was fixed *a priori*, not tuned on test performance, to avoid optimistic bias. The $p < 10^{-4}$ threshold is used only for *reporting* the number of significant clonotypes; the classifier feature set is defined by the fixed top-N ranking, so no multiple-testing correction affects the models themselves. Because both the incidence filter and the enrichment test use training subjects only, the discovered set contains no test-fold information.

2.6 Feature families and classifiers

For each fold we constructed three feature families and trained the corresponding classifiers, all standardized inside a scikit-learn `Pipeline` fit on training data only [17]:

1. **Burden score** — the *count* of the fold's 200 CMV-associated clonotypes present in a subject (an incidence-based, Emerson-style score that generalizes better than a frequency sum dominated by a few high-abundance clones); a single-feature L2 logistic regression.
2. **Sparse top-N public-clonotype matrix** — the 200-dimensional binary presence vector of the CMV-associated clonotypes, classified with both **L1 (Lasso)** and **L2 (Ridge)** logistic regression [18–20].

3. **Global repertoire baseline** — L2 logistic regression on the seven global summary features.

All logistic-regression models used the `saga` solver with an explicit `l1_ratio` (1.0 for L1, 0.0 for L2), $C = 1.0$, and `max_iter`= 5000 [17]. Test-fold class probabilities were retained as OOF predictions.

2.7 Evaluation metrics

On the aggregated OOF predictions ($n = 200$) we computed the area under the receiver-operating characteristic curve (AUROC), the area under the precision–recall curve (AUPRC; average precision), and sensitivity at the threshold achieving $\geq 90\%$ specificity. Because CMV+ prevalence is exactly 0.50 by design, the AUPRC no-skill baseline is 0.50; reporting AUPRC against this prevalence is the appropriate complement to AUROC [21, 22].

2.8 Depth-subsampling degradation curve

To quantify the dependence on per-subject sequencing depth, we defined a depth grid of 5,000, 10,000, 25,000, 50,000, and 100,000 templates plus a full-depth baseline. At each target depth D, every subject's repertoire was down-sampled by multinomial sampling of D templates from the subject's observed clone-frequency distribution ($\mathbf{c} \sim \text{Multinomial}(D, \mathbf{p})$, $\mathbf{p}$ = templates/total); clones drawing zero counts drop out. Subjects whose total template count was already $\leq D$ were kept as-is (no upsampling); given the cohort minimum of 66,268 templates, this occurred only at $D = 100{,}000$, where 4 of 200 subjects were retained at full depth. A freshly seeded generator (`default_rng(42)`) was used at each depth for order-independent, bit-reproducible results. At every depth the *entire pipeline* was re-run from scratch on the down-sampled counts: a fresh incidence matrix over the shared vocabulary, recomputed global features, the *identical* 5-fold splits, training-fold *de novo* discovery, and all four classifiers with OOF evaluation (plus per-fold means $\pm$ standard error for error bars). Down-sampling and vocabulary construction are label-agnostic and all label-dependent statistics remain inside training folds, so the protocol is leakage-free at every depth.

2.9 Reproducibility and software

All steps use a fixed seed (42). Analyses were implemented in Python 3.12 with NumPy [23], SciPy [24], scikit-learn [17], pandas, h5py, and matplotlib. The full-depth re-implementation inside the depth pipeline reproduces the primary cross-validation results bit-for-bit, confirming pipeline identity. Per-subject predictions, discovered clonotypes, performance tables, and all scripts are released with this paper.

3 Results

3.1 A confound-free, balanced 200-subject cohort

The matched design produced a cohort with no exploitable demographic or technical shortcut to serostatus (Table 1, Figure 2). Sex was identically distributed (50 female / 50 male per group; χ^2 $p = 1.0$), ages were matched pair-for-pair (group mean 39.35 ± 11.95 years; Welch $p = 1.0$; Mann–Whitney $p = 1.0$), and sequencing depth did not differ by serostatus (unique clones per subject: CMV+ median 201,412, CMV− median 202,400; Mann–Whitney $p = 0.81$). Across the

cohort, per-subject sequencing was deep (total templates: median 310,408, range 66,268–634,423; unique clones: median 186,488, range 49,700–402,248), pooling to a vocabulary of 23.67 million unique CDR3β sequences. Any predictive signal in this cohort therefore reflects immunological differences rather than demographic or depth artefacts.

Table 1: Balanced, demographically matched cohort (Emerson 2017 discovery subset). Age is matched pair-for-pair (per-pair difference = 0); sequencing depth is not confounded with serostatus.

Group	n	Female	Male	Age (mean ± SD)	Age range
CMV+	100	50	50	39.35 ± 11.95	10–65
CMV−	100	50	50	39.35 ± 11.95	10–65

Balance tests: sex χ^2 $p = 1.0$; age Welch t-test $p = 1.0$, Mann–Whitney $p = 1.0$; depth (unique clones/subject) Mann–Whitney $p = 0.81$. QC: 200/200 files passed.

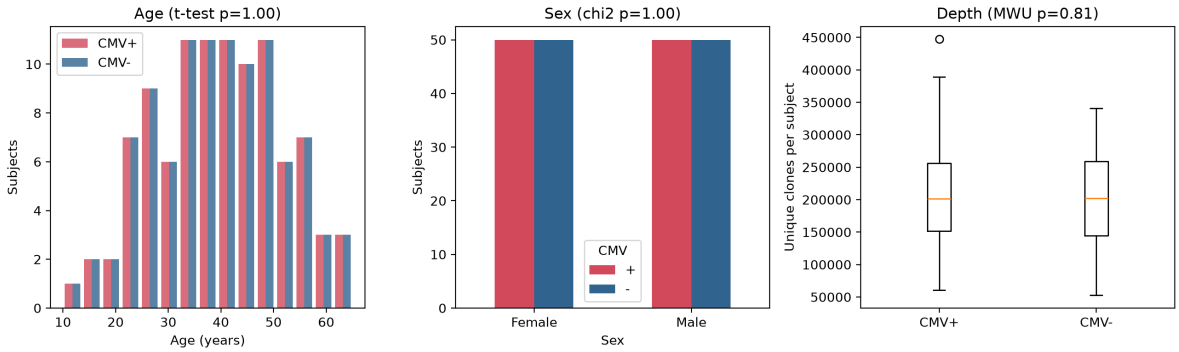

Figure 2: The 200-subject cohort is balanced on age, sex, and sequencing depth. Left: near-identical age histograms for CMV+ (red) and CMV− (blue) subjects (Welch t-test $p = 1.00$, reflecting exact pairwise age matching). Middle: identical sex composition (50 female / 50 male per group; χ^2 $p = 1.00$). Right: overlapping distributions of unique clones per subject by serostatus (Mann–Whitney $p = 0.81$), showing that sequencing depth is not a confounder.

3.2 *De novo* public clonotypes are sparse and barely generalize

De novo discovery inside training folds recovered, on average, 9.4 public clonotypes per fold at $p < 10^{-4}$ (per-fold counts 18, 8, 12, 4, 5), from ∼850,000 candidate public clonotypes screened per fold (Table 2). Across the five folds, 764 distinct clonotypes entered at least one fold's top-200 set, but reproducibility was low: only 6 clonotypes were selected in all five folds and 13 in four (Table 6). The single most reproducible clonotype, `CASSPGLAGVYNEQFF`, was selected in every fold (minimum training $p = 1.7 \times 10^{-5}$).

Table 2: Per-fold *de novo* discovery (training subjects only). Each fold trains on 160 subjects (80 CMV+) and tests on 40. "Significant" clonotypes pass $p < 10^{-4}$; the CMV-associated set is the top-200 by ascending training p-value.

Fold	Train (n)	Test (n)	Candidate public clonotypes	Significant ($p < 10^{-4}$)	Feature set size
1	160	40	851,675	18	200
2	160	40	848,437	8	200
3	160	40	861,755	12	200
4	160	40	873,250	4	200
5	160	40	856,314	5	200
		Mean significant clonotypes / fold		**9.4**	

3.3 Global repertoire structure, not public clonotypes, predicts serostatus

The central result is a sharp dissociation between feature families (Table 3, Figure 3). The three *de novo* public-clonotype models performed at or near chance on held-out subjects: the burden score reached AUROC 0.525 (AUPRC 0.535; sensitivity 0.13 at 90% specificity), L1 (Lasso) AUROC 0.510, and L2 (Ridge) AUROC 0.532. In stark contrast, the **global-repertoire baseline reached AUROC** 0.766**, AUPRC** 0.736 **(vs. the** 0.50 **prevalence baseline), and sensitivity** 0.31 **at 90% specificity**—a +0.24 AUROC margin over the best public-clonotype model. The ROC and precision–recall curves (Figure 3) show the global model dominating across the entire operating range, while all three public-clonotype models track the chance diagonal and the prevalence baseline.

Table 3: Out-of-fold classification performance ($n = 200$; CMV+ prevalence $= 0.50$). The global-repertoire baseline strongly outperforms all *de novo* public-clonotype models. AUPRC no-skill baseline $= 0.50$.

Model	Features	AUROC	AUPRC	Sens. @ 90% Spec.
Burden score (LR)	count of CMV-assoc. clones present	0.525	0.535	0.13
Sparse TCR L1 (Lasso)	binary presence of top-200 clonotypes	0.510	0.502	0.07
Sparse TCR L2 (Ridge)	binary presence of top-200 clonotypes	0.532	0.536	0.16
Global repertoire	**diversity / clonality / expansion**	**0.766**	**0.736**	**0.31**

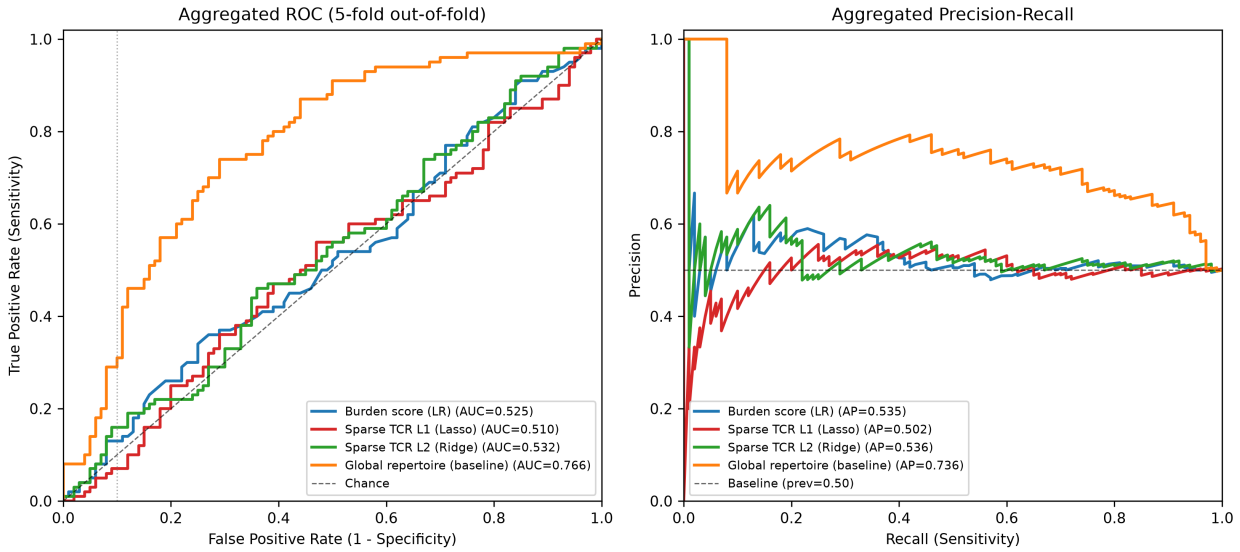

Figure 3: Aggregated out-of-fold ROC (left) and precision–recall (right) curves. The global-repertoire baseline (orange; AUROC 0.766, AP 0.736) dominates across all operating points, whereas the burden score (blue), L1 (red), and L2 (green) public-clonotype models track the chance diagonal (ROC) and the prevalence baseline (PR, 0.50). Curves are computed on concatenated out-of-fold predictions from strict subject-level 5-fold cross-validation.

The reason for the public-clonotype failure is visible in the burden-score distributions (Figure 4): the count of CMV-associated clonotypes present is nearly indistinguishable between CMV+ and CMV− subjects (one-sided Mann–Whitney $p = 0.28$). Although in-fold training burden separates the classes almost perfectly (training AUROC ≈ 0.97), this separation collapses to chance on held-out subjects—the hallmark of selection on noise when only $\sim$9 clonotypes clear significance from a 160-subject training fold.

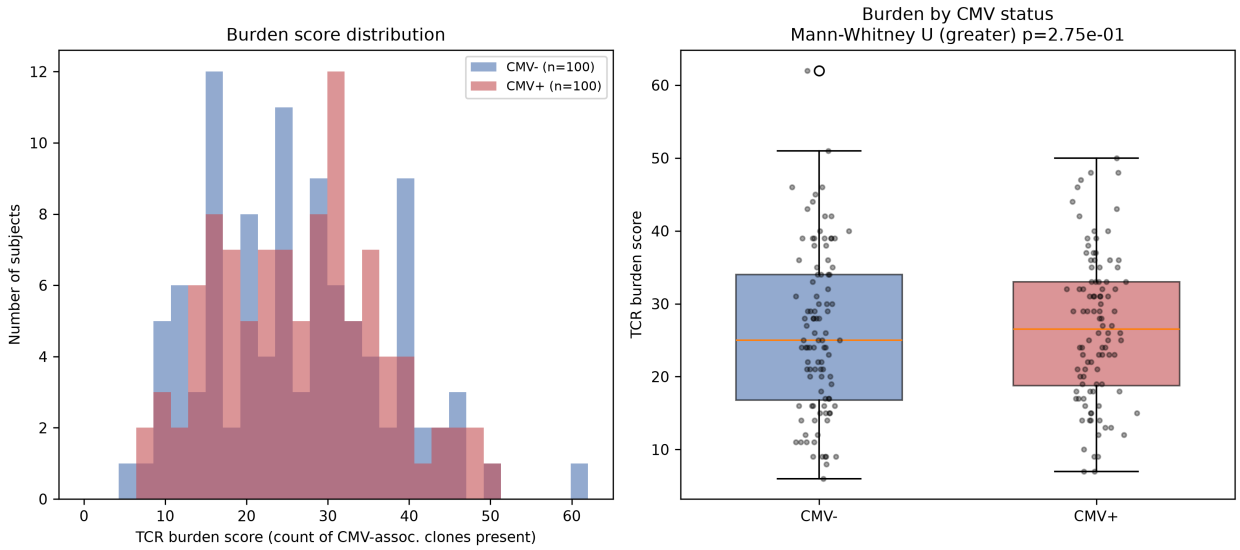

Figure 4: The *de novo* burden score does not separate serostatus out-of-fold. Left: overlapping histograms of the out-of-fold burden score (count of CMV-associated clonotypes present) for CMV− (blue) and CMV+ (red) subjects. Right: box-and-strip plots by serostatus; the distributions are statistically indistinguishable (one-sided Mann–Whitney U $p = 0.28$), explaining the near-chance AUROC of the burden classifier.

3.4 The predictive signal is a diversity/clonality shift

Consistent with CMV-driven memory inflation, CMV+ repertoires were significantly less diverse and more clonally expanded than CMV− repertoires, and these differences drive the global-baseline performance (Table 4). CMV+ subjects had lower Shannon entropy (11.04 vs. 11.47; $p = 4 \times 10^{-5}$) and Simpson diversity (0.9971 vs. 0.9993; $p = 1 \times 10^{-12}$), and higher clonality (0.086 vs. 0.048; $p = 5 \times 10^{-11}$), maximum clone frequency (0.029 vs. 0.012; $p = 8 \times 10^{-12}$), and top-100 cumulative frequency (0.138 vs. 0.075; $p = 1 \times 10^{-11}$). By contrast, mean CDR3 length ($p = 0.12$) and the number of unique clones ($p = 0.92$) did not differ—confirming that the signal is a redistribution of clonal *abundance* (expansion), not of repertoire size or sequence length, and that it is not an artefact of depth.

Table 4: Global repertoire features by CMV serostatus. CMV+ repertoires are less diverse and more clonally expanded. p-values are two-sided Mann–Whitney U tests ($n = 100$ per group).

Feature	CMV− (mean)	CMV+ (mean)	Mann–Whitney p
Shannon entropy	11.47	11.04	4×10^{-5}
Simpson diversity	0.9993	0.9971	1×10^{-12}
Clonality ($1 - H/\ln R$)	0.048	0.086	5×10^{-11}
Max. clone frequency	0.012	0.029	8×10^{-12}
Top-100 cumulative freq.	0.075	0.138	1×10^{-11}
Mean CDR3 length	14.49	14.48	0.12 (n.s.)
Unique clones / subject	183,375	186,459	0.92 (n.s.)

3.5 Global features are depth-robust; public clonotypes are depth-limited

The depth-subsampling analysis (Table 5, Figure 5) sharpens the dissociation. As per-subject depth fell from full (∼185,000 unique clones on average) through 100,000, 50,000, 25,000, 10,000, to 5,000 templates (mean unique clones 63.1k, 36.3k, 19.9k, 8.5k, 4.4k), the global-repertoire baseline held essentially constant at AUROC $\approx$ 0.75–0.77, degrading by only ∼0.015 AUROC even at a 60-fold depth reduction. Diversity and expansion summaries are stable estimands that survive aggressive subsampling. As a validation of the depth pipeline, its full-depth row in Table 5 reproduces the primary cross-validation results of Table 3 bit-for-bit.

The *de novo* public-clonotype models behaved oppositely. The number of clonotypes reaching $p < 10^{-4}$ per fold collapsed to **0** at depths $\leq$ 25,000 templates, rose only to 0.2 at 50,000 and 1.4 at 100,000, and reached 9.4 only at full depth—because shallow sequencing destroys the public-clone sharing on which enrichment testing depends. Accordingly, the burden, L1, and L2 models sat at or below chance throughout the subsampled range (AUROC 0.39–0.54) and only began to approach chance-plus at 100,000/full depth as public sharing re-emerged. The public-clonotype approach thus requires *both* a large cohort *and* deep repertoires, and its performance is monotonically worse at shallower depth.

Table 5: Depth-subsampling: out-of-fold AUROC by per-subject template depth. The global baseline is depth-robust; public-clonotype models remain near chance and their supply of significant clonotypes collapses at shallow depth.

Depth (templates)	Burden	L1 (Lasso)	L2 (Ridge)	**Global**	Sig. clonotypes/fold
5,000	0.425	0.481	0.420	**0.752**	0.0
10,000	0.388	0.439	0.431	**0.751**	0.0
25,000	0.400	0.483	0.431	**0.756**	0.0
50,000	0.428	0.395	0.409	**0.769**	0.2
100,000	0.465	0.541	0.479	**0.768**	1.4
Full	0.525	0.510	0.532	**0.766**	9.4

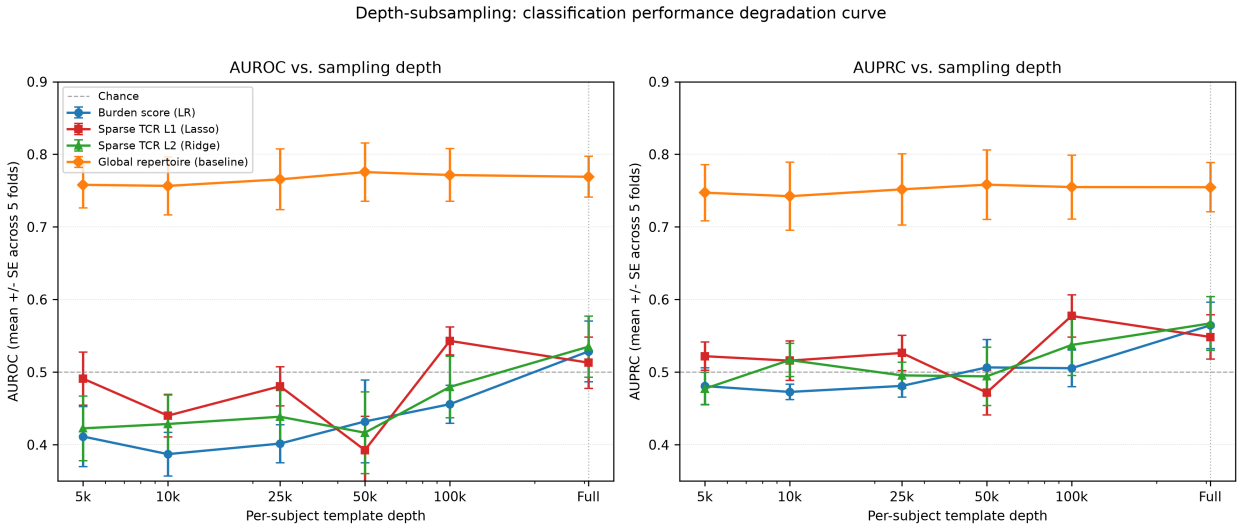

Figure 5: Performance-versus-depth degradation curves for AUROC (left) and AUPRC (right). Points are out-of-fold means with ±SE error bars across the five folds; the x-axis is per-subject template depth (log scale). The global-repertoire baseline (orange) is essentially flat at AUROC ≈ 0.75–0.77 across a 60-fold depth range, whereas the burden (blue), L1 (red), and L2 (green) public-clonotype models remain near the chance line (0.5) throughout, rising only slightly as depth (and thus public-clone sharing) increases.

3.6 Number of features the model relies on

The reported classifiers rely on strikingly few effective features. The winning global-repertoire model uses only **seven** summary statistics per subject. The *de novo* public-clonotype models nominally use a 200-clonotype feature set per fold, but of these only ∼9.4 per fold are statistically significant at $p < 10^{-4}$, and the union of ever-selected clonotypes across all folds is 764 with only 6 recurring in all five folds (Table 6). In other words, the signal that *does* generalize is carried by a seven-dimensional diversity/clonality descriptor, whereas the tens-to-hundreds of *de novo* clonotype features carry essentially no transferable information at this cohort size. Because several of the seven global features (Simpson diversity, clonality, maximum-clone and top-100 cumulative frequency) are correlated by construction, they effectively encode a single lower-dimensional *clonal-expansion* axis; the L2 penalty makes the model robust to this collinearity without inflating its effective dimensionality.

Table 6: Most reproducible *de novo* CMV-associated CDR3β clonotypes (selected in ≥ 4 of 5 folds), ranked by fold recurrence then mean training p-value.

CDR3β (amino acid)	Folds selected	Mean training p	Min. training p
CASSPGLAGVYNEQFF	5	3.6×10^{-4}	1.7×10^{-5}
CASSPDRVGQETQYF	5	4.4×10^{-4}	3.3×10^{-5}
CASSLTSGGAGNEQFF	5	5.6×10^{-4}	3.3×10^{-5}
CASSIGPLEHNEQFF	5	7.8×10^{-4}	1.6×10^{-4}
CASSLGGHNSPLHF	5	8.0×10^{-4}	3.1×10^{-5}
CASSRDRSTDTQYF	5	8.0×10^{-4}	1.8×10^{-4}
CASSLSVGEQYF	4	2.8×10^{-4}	1.0×10^{-5}
CASSFDYSGNTIYF	4	3.7×10^{-4}	7.2×10^{-5}
CASSPRVNEQFF	4	4.9×10^{-4}	4.7×10^{-5}
CASRPGTGSYNEQFF	4	4.9×10^{-4}	4.2×10^{-5}

4 Discussion

4.1 Principal findings

Re-analysing the Emerson (2017) discovery repertoires under a deliberately constrained, confound-free design, we find a clean dissociation between two families of TCRβ features. At a balanced cohort of 200 subjects, *de novo* public CMV-associated clonotypes—the exact signal exploited by the original diagnostic—carry almost no information that transfers to held-out subjects (AUROC 0.51–0.53), whereas seven *global* repertoire summary statistics predict serostatus at AUROC 0.766, AUPRC 0.736, and sensitivity 0.31 at 90% specificity. Moreover, the global signal is remarkably depth-robust: it survives an order-of-magnitude reduction in per-subject sequencing depth (to 5,000 templates) with negligible loss, while public-clonotype discovery fails entirely at shallow depth. The generalizable signal is a diversity/clonality shift—CMV+ repertoires are more clonally expanded and less diverse—directly mirroring the immunological phenomenon of CMV memory inflation [8, 9].

4.2 Why the *de novo* public-clonotype model underperforms here—and why this is not a contradiction of Emerson (2017)

Our public-clonotype result may appear to conflict with the high accuracy reported by Emerson et al. [13], but the two are fully consistent once cohort scale is considered. The public-clonotype paradigm is doubly sample-limited. First, each candidate clonotype must be seen in enough donors to reach statistical significance; with only 160 training subjects per fold, only $\sim$9 clonotypes clear $p < 10^{-4}$, and a handful of enriched-by-chance clonotypes selected from $\sim$850,000 candidates do not generalize—classic selection on noise, evident in the collapse from near-perfect in-fold training separation to out-of-fold chance. Second, each clonotype must be *observed* in a donor, which requires deep sequencing; our depth curve shows public-clone sharing (and hence significant discovery) vanishing below 25,000 templates. Emerson and colleagues succeeded precisely because they had both a large discovery cohort ($\sim$640 subjects) and deep repertoires, giving the statistical power to identify and stably weight thousands of CMV-associated clonotypes. Our finding is therefore a quantitative statement about *scaling*: the public-clonotype approach needs both a large cohort and deep repertoires, and it degrades gracelessly when either is reduced. This is consistent with the broader literature in which attention-based deep models are trained and evaluated on the *full* Emerson cohort [14] and with occurrence-pattern analyses that leverage very large donor panels [12].

Critically, our leakage-free protocol is what makes this dissociation visible. A common but subtle error in repertoire classification is to select CMV-associated clonotypes using the *entire* labelled dataset before cross-validation; because the selection has already "seen" the test subjects, the resulting burden score appears highly predictive, yielding optimistic AUROCs that do not reflect prospective performance. By confining all label-dependent steps to training folds, we report the honest generalization gap. This underscores a general methodological point for immune-ML: feature selection must live inside the cross-validation loop, and evaluation must be strictly subject-level [21, 22].

4.3 Biological interpretation

The features that do generalize have a clear immunological reading. Chronic CMV drives lifelong accumulation of large CMV-specific CD8$^+$ clones—memory inflation—so seropositive repertoires

concentrate more of their template mass in fewer expanded clones [8, 9]. This is exactly what our data show: CMV+ subjects have higher clonality, higher maximum-clone and top-100 cumulative frequencies, and correspondingly lower Shannon and Simpson diversity, all at $p < 10^{-4}$ to $p < 10^{-12}$, with no change in the number of unique clones or CDR3 length. Because these are population-level properties of the clone-size distribution rather than the presence of any particular rare sequence, they are estimable from a modest template sample and are robust to subsampling [15, 16]. The dissociation we report is thus not merely a statistical curiosity but reflects two genuinely different observables of the same biology: the *identity* of expanded clones (public clonotypes; rare, sample-hungry) versus the *shape* of the clone-size distribution (diversity/clonality; aggregate, sample-efficient).

4.4 Limitations

Several limitations qualify our conclusions. (1) By design we analyse a balanced, 1:1-prevalence subset of 200 subjects; while this removes demographic and depth confounds and yields clean effect estimates, real-world CMV prevalence is higher and variable, so our AUPRC (against a 0.50 baseline) should be re-benchmarked at the deployment prevalence [21]: because CMV seroprevalence is typically 60–90%, the operational PR baseline (and achievable precision) would generally be *higher* than reported here, whereas the balanced-design AUROC is prevalence-invariant. (2) We use a pre-processed distribution that retains CDR3 amino-acid sequence and template count but not V/J gene calls; amino-acid-level collapse can merge distinct rearrangements and, if anything, *inflates* apparent public-clone sharing, so the near-chance public-clonotype result is if anything conservative. Full V(D)J-resolved clonotypes would increase the specificity of public-clone matching and might modestly raise the public-clonotype ceiling, but would not change the depth/sample-size scaling behaviour that is our central point. (3) Our global-feature model is intentionally simple (L2 logistic regression on seven interpretable descriptors); more expressive models or additional features (e.g., V-gene usage, network/convergence metrics) could raise the ceiling further. (4) The cohort is drawn from a predominantly North American bone-marrow-donor population with a specific HLA distribution; because CMV-associated public clonotypes are HLA-restricted, both the public-clonotype and (to a lesser degree) diversity signals may transfer imperfectly to other populations [12, 13]. (5) We did not perform experimental validation (e.g., peptide-MHC multimer binding) of the discovered clonotypes; the reproducible clonotypes in Table 6 are candidate CMV-associated sequences, not confirmed CMV-specific TCRs.

4.5 Implications and future directions

Our results carry a concrete design message for TCR-based CMV diagnostics under resource constraints. When labelled cohorts are modest, diversity/clonality descriptors are far more sample- and depth-efficient than *de novo* public-clonotype burden: per-subject sequencing depth can be cut by an order of magnitude with negligible loss for a global-features model, directly lowering assay cost. In practice, a hybrid strategy is attractive: use global repertoire features as a robust, cheap first-line signal, and layer in *pre-curated*, externally validated public-clonotype catalogs (built once on very large cohorts) rather than rediscovering clonotypes *de novo* in every small study [12, 13]. Future work should (i) evaluate this hybrid on independent, HLA-diverse validation cohorts; (ii) characterize the sample-size scaling of public-clonotype discovery explicitly, tracing the cohort size at which *de novo* burden overtakes global features; (iii) extend the depth-subsampling framework to attention-based deep models [14] to test whether learned representations are more depth-robust than enrichment-based discovery; and (iv) benchmark performance at realistic, imbalanced CMV

prevalence. More broadly, the leakage-free, subject-level, depth-aware evaluation template used here is directly transferable to other repertoire-based classification tasks, where the same public-versus-global trade-off is likely to recur.

4.6 Conclusion

Under a rigorous, confound-free, leakage-free evaluation of a balanced 200-subject subset of the Emerson (2017) discovery cohort, cytomegalovirus serostatus is predictable from peripheral-blood TCRβ repertoires primarily through *global* repertoire structure—a CMV-driven diversity/clonality shift—rather than through *de novo* public CMV-associated clonotypes, which require both large cohorts and deep repertoires to generalize. The global signal reaches AUROC 0.77 and is robust to a 60-fold reduction in sequencing depth, making it the more practical foundation for cost-constrained repertoire diagnostics.

Data and code availability

The Emerson (2017) discovery repertoires and CMV serostatus labels are available via the Adaptive Biotechnologies immuneACCESS resource (doi:10.21417/B7001Z) and, in pre-processed form, via the DeepRC distribution [13, 14]. Per-subject out-of-fold predictions (`subject_predictions.csv`), discovered CMV-associated clonotypes (`discovered_cmv_tcrs.csv`), performance tables, and all analysis scripts (data acquisition, cohort matching, matrix construction, cross-validation, and depth subsampling) are released with this manuscript. All results are reproducible with fixed seed 42.

Author contributions

K-Dense Web conceived and performed the analysis, implemented the pipeline, generated the figures, and wrote the manuscript.

Conflicts of interest

The author declares no competing interests.

Acknowledgements

We thank the authors of Emerson et al. [13] and the maintainers of the DeepRC distribution [14] for making the discovery-cohort repertoires and CMV serostatus labels openly available.

References

[1] Michael J. Cannon, D. Scott Schmid, and Terri B. Hyde. Review of cytomegalovirus seroprevalence and demographic characteristics associated with infection. *Reviews in Medical Virology*, 20(4):202–213, 2010. doi: 10.1002/rmv.655.

[2] Mohammad Zuhair, Gloria S. A. Smit, Graham Wallis, Faiz Jabbar, Charlotte Smith, Brecht Devleesschauwer, and Paul Griffiths. Estimation of the worldwide seroprevalence of cytomegalovirus: A systematic review and meta-analysis. *Reviews in Medical Virology*, 29 (3):e2034, 2019. doi: 10.1002/rmv.2034.

[3] Paul Griffiths, Ilona Baraniak, and Matthew Reeves. The pathogenesis of human cytomegalovirus. *The Journal of Pathology*, 235(2):288–297, 2015. doi: 10.1002/path.4437.

[4] Raymund R. Razonable and Atul Humar. Cytomegalovirus in solid organ transplantation. *American Journal of Transplantation*, 13(S4):93–106, 2013. doi: 10.1111/ajt.12103.

[5] Camille N. Kotton, Deepali Kumar, Angela M. Caliendo, Shirish Huprikar, Sunwen Chou, Lara Danziger-Isakov, and Atul Humar. The third international consensus guidelines on the management of cytomegalovirus in solid-organ transplantation. *Transplantation*, 102(6): 900–931, 2018. doi: 10.1097/TP.0000000000002191.

[6] Sheetal Manicklal, Vincent C. Emery, Tiziana Lazzarotto, Suresh B. Boppana, and Ravindra K. Gupta. The "silent" global burden of congenital cytomegalovirus. *Clinical Microbiology Reviews*, 26(1):86–102, 2013. doi: 10.1128/CMR.00062-12.

[7] Aileen Kenneson and Michael J. Cannon. Review and meta-analysis of the epidemiology of congenital cytomegalovirus (cmv) infection. *Reviews in Medical Virology*, 17(4):253–276, 2007. doi: 10.1002/rmv.535.

[8] Sven Koch, Anis Larbi, Dennis Ozcelik, Rafael Solana, Cecile Gouttefangeas, Sebastian Attig, Anders Wikby, Jan Strindhall, Claudio Franceschi, and Graham Pawelec. Cytomegalovirus infection: a driving force in human T cell immunosenescence. *Annals of the New York Academy of Sciences*, 1114:23–35, 2007. doi: 10.1196/annals.1396.043.

[9] Paul Klenerman and Annette Oxenius. T cell responses to cytomegalovirus. *Nature Reviews Immunology*, 16(6):367–377, 2016. doi: 10.1038/nri.2016.38.

[10] Harlan S. Robins, Paulo V. Campregher, Santosh K. Srivastava, Abigail Wacher, Cameron J. Turtle, Orsalem Kahsai, Stanley R. Riddell, Edus H. Warren, and Christopher S. Carlson. Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood*, 114(19): 4099–4107, 2009. doi: 10.1182/blood-2009-04-217604.

[11] Harlan S. Robins, Santosh K. Srivastava, Paulo V. Campregher, Cameron J. Turtle, Jessica Andriesen, Stanley R. Riddell, Christopher S. Carlson, and Edus H. Warren. Overlap and effective size of the human CD8+ T cell receptor repertoire. *Science Translational Medicine*, 2 (47):47ra64, 2010. doi: 10.1126/scitranslmed.3001442.

[12] William S. DeWitt, Anajane Smith, Gary Schoch, John A. Hansen, Frederick A. Matsen, and Philip Bradley. Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *eLife*, 7:e38358, 2018. doi: 10.7554/eLife.38358.

[13] Ryan O. Emerson, William S. DeWitt, Marissa Vignali, Jenna Gravley, Joyce K. Hu, Edward J. Osborne, Cindy Desmarais, Mark Klinger, Christopher S. Carlson, John A. Hansen, Mark Rieder, and Harlan S. Robins. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature Genetics*, 49(5): 659–665, 2017. doi: 10.1038/ng.3822.

[14] Michael Widrich, Bernhard Schäfl, Milena Pavlović, Hubert Ramsauer, Lukas Gruber, Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, and Günter Klambauer. Modern Hopfield networks and attention for immune repertoire classification. *Advances in Neural Information Processing Systems*, 33:18832–18845, 2020.

[15] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[16] Qian Qi, Yi Liu, Yong Cheng, Jacob Glanville, David Zhang, Ji-Yeun Lee, Richard A. Olshen, Cornelia M. Weyand, Scott D. Boyd, and Jörg J. Goronzy. Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences*, 111(33): 13139–13144, 2014. doi: 10.1073/pnas.1409155111.

[17] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[18] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

[19] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. doi: 10.1111/j.1467-9868.2005.00503.x.

[20] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01.

[21] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

[22] Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 233–240, 2006. doi: 10.1145/1143844.1143874.

[23] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, et al. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. doi: 10.1038/s41586-020-2649-2.

[24] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020. doi: 10.1038/s41592-019-0686-2.