IMMUNEEMBEDDING: CROSS-MODAL RECOVERY OF THE HUMAN IMMUNE STATE FROM PAIRED TCR AND BCR REPERTOIRES

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding which immune measurements can be predicted from adaptive receptor repertoires is central to representing individual immune state. We introduce ImmuneEmbedding, a modular participant-level representation and evaluation framework linking T-cell receptor (TCR) and B-cell receptor (BCR) repertoires to cellular and molecular measurements. The framework combines interpretable repertoire statistics with abundance-weighted ESM-2 sequence embeddings and evaluates a supervised structural bottleneck as a complementary representation. We study 99 clinical, cellular, transcriptomic, proteomic, and serum targets in COMBAT using donor-grouped cross-validation and donor-cluster bootstrap intervals. On 136 paired-repertoire timepoints, matched linear readouts cover 14 targets with combined structural inputs, compared with 11 for TCR and seven for BCR; three paired-input targets are absent from both single-chain coverage sets. On an expanded set of 140 timepoints, structural features cover 16 targets overall, while combining structure and sequence increases serum coverage to four, compared with three for structure and two for sequence alone. Comparisons with k-mer features, released JL-GloVe embeddings, and attention-based multiple-instance learning reveal distinct target profiles, with TCR attention reaching six serum targets. Across the evaluated ImmuneEmbedding configurations and both analysis sets, 38 targets meet the nominal coverage criterion, including 14 serum analytes. Separate cohort-calibrated TCR transfer experiments achieve cytomegalovirus serostatus AUROCs of 0.874 in confirmation donors and 0.719 across studies with supervised augmentation. These findings map repertoire-associated immune variation and show how receptor pairing, sequence representation, and aggregation shape the measurements accessible from an individual's repertoire.

1 INTRODUCTION

An individual's immune state is expressed across lymphocyte populations, receptor sequences, transcriptional programs, and circulating proteins. Adaptive immune receptor sequencing captures both the molecular identities of lymphocyte receptors and the abundance distribution of the clones carrying them. These observations connect sequence-level variation with population structure within one assay. A participant-level representation preserving both kinds of information could support prediction of immune measurements acquired through other technologies, making an available repertoire useful across downstream questions in systems immunology.

Repertoire-based prediction has established clinically relevant signals. T-cell receptor beta-chain repertoires identify cytomegalovirus (CMV) exposure and HLA-associated variation (Emerson et al., 2017). Mal-ID integrates B-cell and T-cell receptor information for multidiagnostic classification with independent-cohort evaluation (Zaslavsky et al., 2025). JL-GloVe learns reusable TCR embeddings from co-occurrence across repertoires (Chapfuwa et al., 2025). These studies motivate a broader question: which contemporaneous cellular and molecular measurements can be predicted from a repertoire, and which aspects of its representation support those predictions?

Answering this question requires an evaluation spanning the measurements that constitute immune state. Serum cytokines, cell composition, and transcriptional variation probe different properties of a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

representation. Clonal concentration, receptor-gene usage, B-cell maturation, and receptor sequence content may consequently contribute to different outcomes. Sample recovery and sequencing depth can also carry predictive information, making explicit depth comparisons important. The COMBAT blood atlas provides a setting for this analysis through joint repertoire, cellular, transcriptomic, proteomic, serum, and clinical profiling (COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium, 2022; Wang et al., 2023).

We introduce ImmuneEmbedding to connect participant-level repertoire structure and sequence content to this heterogeneous target panel. Pairing denotes TCR and BCR repertoires observed at the same participant-timepoint. Interpretable structural features and pooled protein-language-model embeddings define assay-derived coordinates, while a multitask bottleneck evaluates supervised compression of the structural information. We define immune-state recovery as out-of-sample discrimination or rank association with measured outcomes. Separate TCR experiments examine transfer with cohort-rank calibration.

The study contributes this modular representation together with a target-resolved account of its predictive scope. Matched readouts characterize the measurements gained and lost when receptor modalities are combined. Block comparisons distinguish structural, sequence, and supervised information, and sequence baselines evaluate alternative aggregation strategies. The resulting map contains 38 nominally covered targets, including 14 serum analytes. Structural summaries achieve the broadest overall coverage among individual configurations in the expanded COMBAT evaluation, whereas TCR attention reaches the most serum targets. These differences connect representation design to the immune measurement of interest.

2 Related work

Repertoire learning maps variable-sized collections of receptor sequences to participant-level outcomes. DeepRC uses attention-based multiple-instance learning to emphasize informative sequence patterns, while DeepTCR learns sequence-associated concepts for TCR analysis (Widrich et al., 2020; Sidhom et al., 2021). The immuneML ecosystem supports standardized encoding and evaluation workflows (Pavlović et al., 2021). Reusable representations include JL-GloVe's co-occurrence vectors and AIRIVA's generative latent factors (Chapfuwa et al., 2025; Pradier et al., 2023). Mal-ID combines repertoire composition, sequence clustering, and language-model predictions across TCR and BCR data for disease classification (Zaslavsky et al., 2025). Our evaluation addresses contemporaneous molecular and cellular measurements through participant-level statistics, pooled sequences, and target-specific probes. The implemented attention comparator follows multiple-instance learning (Ilse et al., 2018) with pretrained ESM-2 instances and count-aware pooling.

Cross-modal biological models provide a complementary perspective. MOFA learns factors across heterogeneous omics views, totalVI jointly models RNA and protein, and MultiVI integrates partially observed single-cell modalities (Argelaguet et al., 2018; Gayoso et al., 2021; Ashuach et al., 2023). Here, the available repertoire defines the prediction input and the other assays define the target panel. ESM-2 provides pretrained sequence coordinates (Lin et al., 2023), and their weighted pooling is a permutation-invariant set operation (Zaheer et al., 2017). Marginal rank-to-normal transformations offer an established calibration mechanism (Liu et al., 2009). We evaluate these choices at the participant level, separating within-COMBAT prediction from cohort-conditioned external transfer.

3 ImmuneEmbedding

ImmuneEmbedding represents each participant-timepoint through its repertoire structure and sequence content. For participant-timepoint i and receptor modality $m \in \{T, B\}$, write

$$\mathcal{R}_i^{(m)} = \{(s_{ij}^{(m)}, a_{ij}^{(m)}, v_{ij}^{(m)}, j_{ij}^{(m)})\}_{j=1}^{K_i^{(m)}}, \qquad p_{ij}^{(m)} = \frac{a_{ij}^{(m)}}{\sum_k a_{ik}^{(m)}}, \tag{1}$$

where s is a CDR3 or junction amino acid sequence, a its observed abundance, and v, j its receptor-gene annotations. AIRR-standardized fields provide the annotation vocabulary (Vander Heiden et al., 2018). The COMBAT structural pipeline uses the supplied TRB clone identifier and groups IGH

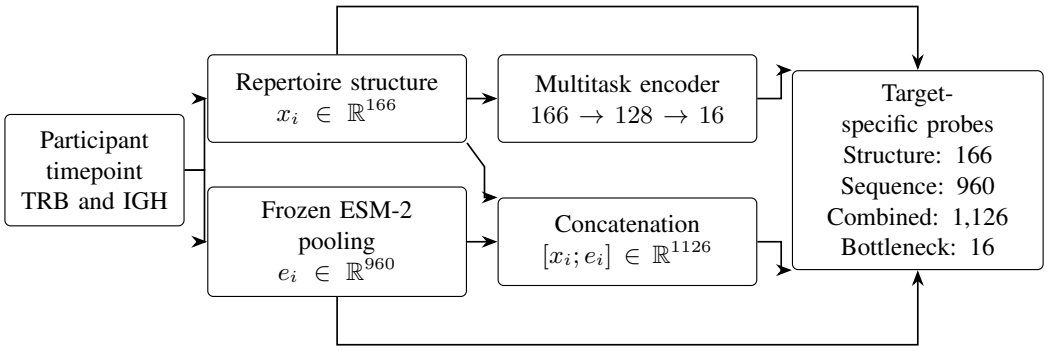


Figure 1: **Representations evaluated in COMBAT.** Repertoire structure and pooled sequence content provide 166 and 960 coordinates. Their concatenation supplies 1,126 inputs. A separate structural multitask encoder produces a 16-dimensional bottleneck used alone by its downstream probe. Imputation and standardization for linear models are fitted within each training fold. Cohort-rank calibration is evaluated in the separate external-transfer experiments.

records by junction sequence and normalized V call. Appendix B specifies the feature definitions and required companion-chain annotations.

The structural vector $x_i \in \mathbb{R}^{166}$ contains 84 TCR and 82 BCR coordinates describing repertoire size, clonal organization, gene usage, and maturation-related properties. With natural logarithms, entropy and normalized clonality are

$$H_i = -\sum_j p_{ij} \log p_{ij}, \qquad C_i = 1 - \frac{H_i}{\log K_i} \quad (K_i > 1). \tag{2}$$

The vector includes clonality, top-1/10/100 clone mass, clone counts, count-weighted CDR3 length mean and standard deviation, and V/J frequencies. Clone counts and sequence lengths are aligned by clone identifier before weighting. IGH summaries include mutation, isotype, and hydrophobicity annotations. Available TRA, light-chain, and cell annotations extend the schema in this single-cell resource.

To encode sequence content, we use the frozen 12-layer, 35-million-parameter ESM-2 model `esm2_t12_35M_UR50D`. The final hidden states are averaged over non-padding tokens, including the model's boundary tokens, to give $\phi(s) \in \mathbb{R}^{480}$. Repertoire pooling and concatenation are

$$e_i^{(m)} = \sum_j p_{ij}^{(m)} \phi(s_{ij}^{(m)}), \qquad u_i = [x_i; e_i^{(T)}; e_i^{(B)}] \in \mathbb{R}^{1126}. \tag{3}$$

Pooling preserves sequence composition weighted by observed abundance and is invariant to record ordering and uniform abundance rescaling. Structural-only, sequence-only, and concatenated vectors define three fixed representations.

A separate supervised configuration compresses the structural vector through a $166 \to 128 \to 16$ network. For outer-training donors $\mathcal{I}_{\mathrm{tr}}$, observed-target mask o_{it}, and target-appropriate loss ℓ_t, the masked objective is

$$\mathcal{L} = \frac{\sum_{i \in \mathcal{I}_{\mathrm{tr}}, t \in \mathcal{T}_b} o_{it}\, \ell_{\mathrm{BCE}}(y_{it}, g_t(h_\theta(x_i)))}{\sum_{i \in \mathcal{I}_{\mathrm{tr}}, t \in \mathcal{T}_b} o_{it}} + \frac{\sum_{i \in \mathcal{I}_{\mathrm{tr}}, t \in \mathcal{T}_r} o_{it}\, \ell_{\mathrm{MSE}}(\widetilde{y}_{it}, g_t(h_\theta(x_i)))}{\sum_{i \in \mathcal{I}_{\mathrm{tr}}, t \in \mathcal{T}_r} o_{it}}, \tag{4}$$

where $\mathcal{T}_b$ and $\mathcal{T}_r$ denote binary and continuous targets; an empty loss family contributes zero. Continuous targets are standardized using observed outer-training values only. All held-out outcomes are excluded from both scaling and the encoder loss. A target-specific linear probe uses the resulting 16 coordinates alone. Training uses full-batch AdamW for 300 epochs, hidden-layer dropout 0.3, learning rate 10^{-3}, and weight decay 10^{-2}. Figure 1 distinguishes these evaluated configurations.

Cohort calibration is evaluated separately for external transfer. If r_{ik} is the average rank of feature k among n_c samples in cohort c, its calibrated coordinate is

$$z_{ik} = \Phi^{-1}\left(\frac{r_{ik} - 1/2}{n_c}\right). \tag{5}$$

The external cohort contributes unlabeled covariates to its own transformation, defining a covariate-transductive setting. The Emerson supervised extension appends a separately trained residual representation to these calibrated inputs. COMBAT uses training-fold imputation and scaling without rank calibration.

4 EXPERIMENTAL DESIGN

The COMBAT panel comprises 99 outcomes: nine clinical or demographic targets, nine cell fractions, 51 serum analytes, and ten principal components (PCs) each from Luminex, plasma proteomics, and bulk transcriptomics. Six clinical outcomes are binary and 93 targets are continuous. Cell fractions share the single-cell sampling platform with VDJ measurements. The remaining 81 targets belong to molecular-assay families, including correlated analytes and derived PCs.

The primary set contains 136 timepoints from 120 donors with both TRB and IGH. The expanded set contains 140 timepoints from 124 donors with at least one modality, adding two TCR-only and two BCR-only timepoints. Both sets contain ten donors with two timepoints and three with three. Target availability gives 97/99 serum observations, 91/93 severity observations, and 61/62 molecular-PC observations in the paired/expanded sets. Comparisons use their designated set; the aggregate map retains the originating sample definition (Appendix A).

Linear readouts use training-fold median imputation and standardization followed by balanced logistic regression ($C = 1$) or ridge regression ($\alpha = 10$). Missing structural or ESM values are imputed from the training fold without an indicator; all-missing training columns are discarded. Histogram gradient boosting uses native missing-value handling, 200 iterations, learning rate 0.05, depth 3, and L2 regularization 1. Paired structural predictors use concatenation, averaging of modality predictions, or nested three-fold donor-grouped selection. Depth controls use four cell/count coordinates. Appendix B specifies the selection and training protocols.

Sequence baselines comprise abundance-weighted amino acid 3-mers, released JL-GloVe vectors, and attention-MIL. Three-mer vectors have 8,000 coordinates per modality. JL-GloVe supplies 100-dimensional TCR embeddings and is adapted through CDR3-core matching. Missing 3-mer or JL-GloVe blocks are zero-filled. Attention-MIL uses the top 500 clonotypes per chain as ESM-2 instances, separate 64-dimensional chain projections, and within-chain attention with a $\log(1+\mathrm{count})$ score prior. The paired head receives both pooled vectors and chain-level abundance summaries. Empty bags are masked to zero. Regression predictions are converted from training-fold standardized outputs to original target units before pooling across folds. Appendix C gives the attention equations and optimization settings.

Cross-validation assigns all observations from a donor to one of five folds. Binary tasks use stratified grouped splits; continuous tasks use grouped splits. Fixed-feature binary probes average predictions over seeds 8128, 1, and 2, while bottleneck and attention models use the single split seed 8128. We measure AUROC for binary targets and pooled out-of-fold Spearman correlation for continuous targets. For COMBAT and SOTON, 95% percentile intervals resample donors with replacement and include all valid rows of each selected donor. This fixed-prediction cluster bootstrap preserves repeated-measurement bundles, using 2,000 binary or 1,000 regression resamples. Model fitting is held fixed during resampling (Efron, 1979).

For configuration a, target t, and interval lower endpoint L_{at}, nominal coverage is

$$I_{at} = \mathbf{1}\{L_{at} > b_t\}, \qquad N_{a,f} = \sum_{t \in \mathcal{T}_f} I_{at}, \qquad b_t = \begin{cases} 0.5 & \text{AUROC,} \\ 0 & \text{Spearman.} \end{cases} \tag{6}$$

Individual configurations supply the primary comparisons. Family-level coverage is the union of their qualifying target sets. The aggregate ImmuneEmbedding map combines paired structural configurations with the four expanded-set representations. These targetwise counts are unadjusted for multiplicity, and selecting among predictors adds uncertainty beyond the reported fixed-prediction intervals (Cawley and Talbot, 2010; Benjamini and Hochberg, 1995).

External CMV experiments evaluate 132 confirmation and 76 Keck donors (Emerson et al., 2017). A calibration experiment uses 192 development donors for both groups. A separate official JL-GloVe comparison uses 192 development donors for confirmation and all 324 eligible cohort-1 donors for

Table 1: **Matched-input coverage on the paired set.** The 136 timepoints come from 120 donors. Entries count targets whose 95% donor-cluster bootstrap lower bound exceeds the metric-specific null. Linear readouts use logistic or ridge regression according to target type. The boosting breakdown appears in Appendix D.

Assay family	Targets	TCR	BCR	TCR+BCR	Depth
Clinical/demographic	9	4	2	4	6
Cell composition	9	3	4	5	7
Serum analytes	51	1	0	2	4
Luminex PCs	10	0	1	1	0
Proteome PCs	10	2	0	1	1
Transcriptome PCs	10	1	0	1	1
Total, linear	99	11	7	14	19
Total, boosting	99	10	7	14	17

Table 2: **Representation comparisons on the expanded set.** The 140 timepoints come from 124 donors. Each row represents one configuration under the same nominal donor-cluster criterion. JL-GloVe uses TCR input; 3-mers use both modalities. The supervised probe uses the 16-dimensional bottleneck alone. Molecular targets include serum analytes and the three PC families.

Configuration	All / 99	Molecular / 81	Serum / 51
Structure, linear	16	8	3
ESM-2, linear	5	4	2
Structure + ESM, linear	14	7	4
Supervised bottleneck, linear	14	4	0
JL-GloVe, linear	4	4	4
JL-GloVe, boosting	8	6	5
3-mer, linear	2	2	1
3-mer, boosting	13	4	1
Attention-MIL, TCR	13	6	6
Attention-MIL, BCR	6	2	1
Attention-MIL, paired	6	2	1

Keck, with tuned logistic regression. Additional paired classifications use 35 Binder participants and 60 SOTON tissue specimens from 30 donors, with depth-excluded features and donor grouping. Resource accessions and retained external protocols appear in Appendices A and F.

5 RESULTS

Pairing TCR and BCR structural features changes both the number and identity of predictable targets. Under the matched linear readout, paired inputs cover 14 targets, compared with 11 for TCR and seven for BCR (Table 1). Interferon-alpha, IL-15, and Luminex PC5 are reached by the paired input and absent from both single-chain coverage sets. The single-chain union contains 16 targets: pairing adds three and loses five. With boosting, paired inputs cover 14 targets, compared with ten for TCR and seven for BCR; seven paired targets lie outside the 15-target single-chain union and eight union targets are lost. Thus, receptor pairing supplies target-specific predictions under both readouts while changing which single-chain associations are retained.

Depth alone covers 19 targets with the linear readout and 17 with boosting. Linear paired inputs reach five molecular targets versus six for depth, and two serum targets versus four. The advantage of the structural representation is therefore concentrated in particular measurements, with sample recovery carrying substantial predictive information. These findings motivate examining the molecular targets individually alongside the coverage totals.

The expanded-set block comparison shows a different distribution of predictive information across representations (Table 2). Structure alone covers 16 targets, pooled ESM-2 covers five, structure

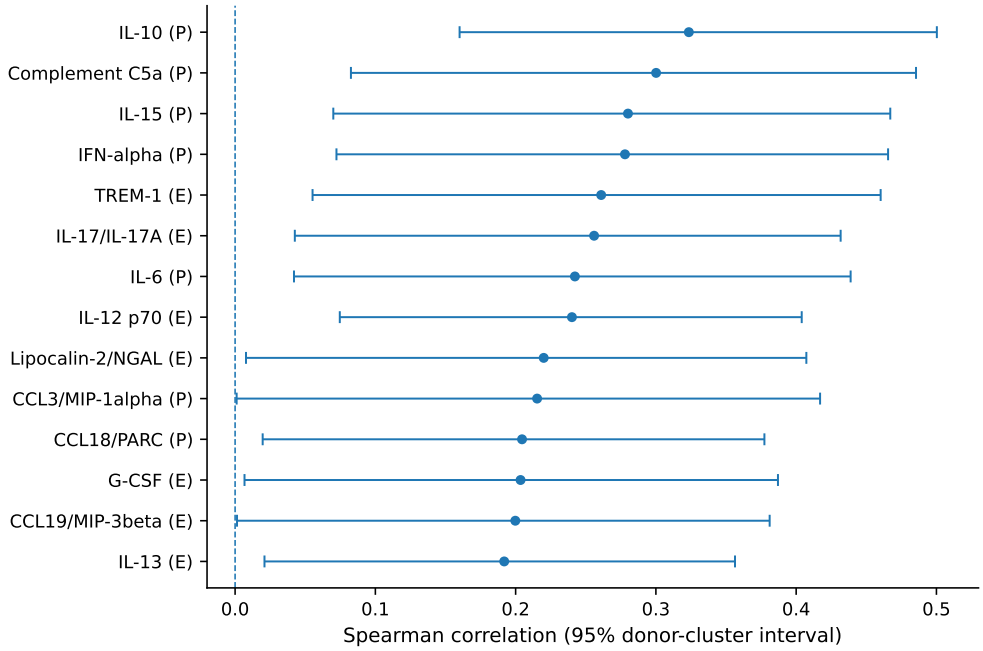


Figure 2: **Serum measurements predicted from repertoire representations.** Points show the highest qualifying estimate for each of 14 serum targets across ImmuneEmbedding configurations, with 95% donor-cluster bootstrap intervals. P denotes 97 timepoints from 82 donors and E denotes 99 from 84. Selection is targetwise and nominal; Table 11 identifies each predictor.

plus ESM covers 14, and the supervised bottleneck covers 14. Structural features reach three transcriptomic PCs, two Luminex PCs, and three serum analytes. Adding sequence features changes this profile and increases serum coverage to four, compared with two for ESM-only. For severe or critical versus mild COVID-19, the combined representation yields AUROC 0.760 [0.648, 0.861], compared with 0.736 [0.634, 0.836] for structure and 0.644 [0.506, 0.766] for sequence on the same 93 timepoints from 79 donors.

The structural bottleneck emphasizes clinical and cellular outcomes, covering six clinical targets and four cell fractions. It also reaches four molecular PCs and no serum analytes under the nominal criterion. Its severity AUROC is 0.718 [0.606, 0.828]. Structural summaries consequently have the highest overall individual-configuration coverage in this benchmark; adding sequence or supervision redistributes that coverage across outcomes.

The alternative sequence models provide further molecular associations. Released JL-GloVe covers four targets with linear prediction and eight with boosting, giving 11 across both readouts. The corresponding 3-mer counts are two and 13, with a union of 15. TCR attention-MIL reaches 13 targets overall and six serum analytes, the broadest serum coverage among individual configurations. Its BCR and paired variants each reach six targets overall and one serum analyte. Across the three variants, attention covers 18 targets, seven outside the aggregate ImmuneEmbedding map. Structure has the broadest overall coverage and attention the broadest serum coverage, illustrating the dependence of representation performance on the assay family.

Across paired structural and expanded representation configurations, the aggregate map contains 38 targets: six clinical, seven cellular, 14 serum, four Luminex PCs, four proteomic PCs, and three transcriptomic PCs. The paired structural family covers 28 targets and the expanded representation family covers 27, with 17 shared target labels. The resulting 25 molecular targets link repertoire inputs to independently measured assay variation (Appendix E).

The serum predictions include IL-10 at Spearman 0.323 [0.160, 0.500], complement C5a at 0.300 [0.083, 0.485], IL-15 at 0.280 [0.070, 0.467], and interferon-alpha at 0.278 [0.072, 0.466] from

Table 3: **CMV prediction on held-out donor groups.** Values are AUROC [95% bootstrap CI]. The first block uses the same 192 development donors for both evaluations. The official JL-GloVe comparison uses 192 development donors for confirmation and 324 cohort-1 donors for Keck, with cross-validation-tuned logistic regression.

Representation	Confirmation (132)	Keck (76)
Cohort-calibration experiment		
Uncalibrated	0.821 [0.745, 0.895]	0.679 [0.547, 0.808]
Rank-calibrated	0.857 [0.787, 0.917]	0.706 [0.574, 0.819]
Rank + residual	0.874 [0.810, 0.927]	0.719 [0.592, 0.834]
Official embedding comparison		
JL-GloVe, mean	0.651 [0.550, 0.742]	0.586 [0.452, 0.708]
JL-GloVe, count-weighted	0.583 [0.472, 0.688]	0.493 [0.358, 0.627]
Structure	0.814 [0.737, 0.889]	0.641 [0.510, 0.767]
Reference axes + structure	0.803 [0.717, 0.875]	0.641 [0.516, 0.758]
JL-GloVe + structure	0.797 [0.722, 0.869]	0.700 [0.570, 0.813]

paired structural predictors. Structure plus ESM predicts TREM-1 at 0.261 [0.055, 0.460], while ESM-only predicts IL-17A at 0.256 [0.043, 0.432] and IL-12 p70 at 0.240 [0.075, 0.404]. Other mapped measurements include IL-6, Lipocalin-2/NGAL, G-CSF, CCL19, and IL-13 (Figure 2). Their rank associations connect receptor population measurements to inflammatory and innate-immune variation, with target-specific predictors contributing to the aggregate map.

The external TCR experiments provide a separate assessment of transfer (Table 3). Cohort-rank calibration raises confirmation AUROC from 0.821 to 0.857 and Keck AUROC from 0.679 to 0.706. Residual augmentation yields 0.874 and 0.719, with paired increments of 0.017 $[-0.009, 0.043]$ and 0.013 $[-0.032, 0.061]$. These intervals describe uncertainty in the smaller supervised contribution. In the official JL-GloVe comparison, structural features reach 0.814 and 0.641 versus 0.651 and 0.586 for mean-pooled embeddings. Adding JL-GloVe to structure gives 0.797 in confirmation and 0.700 in Keck, showing cohort-dependent contributions from co-occurrence information.

Additional paired classifications show similarly endpoint-dependent effects. Binder yields AUROC 0.820 for TCR, 0.667 for BCR, and 0.797 for paired inputs. SOTON yields 0.767, 0.748, and 0.786, respectively; the paired estimate has a donor-cluster interval of [0.640, 0.913]. Paired differences span zero in both datasets (Appendix F). Together with the COMBAT comparisons, these results locate the value of receptor pairing at the level of particular outcomes.

6 Discussion

ImmuneEmbedding connects repertoire representation to measurements of cellular and molecular immune state. The cross-modal analysis establishes predictive associations across distinct assays and relates their distribution to the information preserved by a representation. Paired structural inputs reach measurements absent from both single-chain coverage sets. Sequence features add serum targets when combined with structure, while a supervised structural bottleneck concentrates coverage in clinical and cellular measurements. These findings give the measurement of interest a central role in evaluating repertoire representations.

The serum associations provide biological context for this representation-level comparison. Interferon and inflammatory programs vary with COVID-19 severity in independent work and in COMBAT (Hadjadj et al., 2020; COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium, 2022). Predicting these measurements from receptor features connects adaptive receptor populations with broader inflammatory state. Shared disease context, cell composition, and immune history may contribute to this coupling. The present rank associations provide assay-level evidence for the connection and identify molecular targets for subsequent biological investigation.

The baseline comparison also clarifies the value of retaining simple and learned summaries. Structure has the broadest overall individual-model coverage in the expanded set, whereas TCR attention reaches the most serum analytes. Different aggregation rules thus preserve different predictive

information. Pairing introduces useful targets while losing others, and supervision changes the balance across assay families. These observations support evaluating modality and aggregation choices against intended outcomes. The separate CMV experiments further show that cohort calibration can materially affect transfer estimates, making the reference population part of the predictor's specification.

The scope of inference follows from the study design. Molecular recovery is evaluated within COMBAT; external transfer concerns TCR-based CMV serostatus. The two COMBAT sets overlap and differ in receptor availability. Donor-cluster intervals preserve repeated measurements but condition on fixed out-of-fold predictions. Coverage remains nominal across correlated targets and, for the aggregate map, multiple predictors. Depth is strongly predictive, so the associations do not by themselves identify receptor-specific causal mechanisms. Larger paired cohorts with repeated molecular assays would support external molecular validation and fixed-reference deployment.

7 Conclusion

ImmuneEmbedding provides a participant-level framework for relating TCR and BCR repertoire structure and sequence content to measured immune variation. Donor-grouped analyses map 38 nominally covered targets, including 14 serum analytes, across the evaluated configurations and analysis sets. Matched comparisons and block analyses identify target-dependent contributions from pairing, sequence features, and aggregation, while separate CMV experiments assess cohort-conditioned transfer. The representation and evaluation together provide an empirical basis for selecting repertoire summaries according to the immune measurement of interest.

Ethics statement

This study analyzes human immune-profiling datasets described in the cited source studies. Secondary use is governed by the original consent, ethics approvals, repository conditions, and data-use agreements. Repertoire-derived representations may reveal infection history, health status, and demographic information. Appropriate access controls and evaluation across populations remain important for subsequent use. The reported predictors are research analyses; prospective clinical utility has not been established.

Reproducibility statement

Sections 3 and 4 specify the representations and evaluation protocol. Appendices A–F detail sample definitions, preprocessing, training, baseline configurations, and target-level results. The supplementary archive contains the metric exports used in this manuscript, feature and target dictionaries, and deterministic table-generation code. Model and evaluation source files document the implemented analyses. The saved author-held out-of-fold predictions, donor partitions, and target-scaler records permit independent reconstruction of the reported statistics; individual-level clinical records remain subject to source data-use conditions. Trained model checkpoints were not retained. The reported bootstrap intervals are conditional on the fixed predictions and the nominal target criterion is unadjusted for multiplicity.

AI use statement

Generative AI tools assisted with code generation, analysis and interpretation and language editing.

References

Ricard Argelaguet, Britta Velten, Damien Arnol, et al. Multi-omics factor analysis: a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14:e8124, 2018. doi: 10.15252/msb.20178124.

Tal Ashuach, Mariano I. Gabitto, Rohan V. Koodli, et al. MultiVI: deep generative model for the integration of multimodal data. *Nature Methods*, 20:1222–1231, 2023. doi: 10.1038/s41592-023-01909-9.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57:289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010. URL https://jmlr.org/papers/v11/cawley10a.html.

Paidamoyo Chapfuwa, Ilker Demirel, Lorenzo Pisani, Javier Zazo, Elon Portugaly, H. Jabran Zahid, and Julia Greissl. Scalable universal T-cell receptor embeddings from adaptive immune repertoires. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wyF5vNIsO7.

Scott Christley, Walter Scarborough, Eddie Salinas, et al. VDJServer: A cloud-based analysis portal and data commons for immune repertoire sequences and rearrangements. *Frontiers in Immunology*, 9:976, 2018. doi: 10.3389/fimmu.2018.00976.

Brian D. Corrie, Nishanth Marthandan, Bojan Zimonja, et al. iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunological Reviews*, 284:24–41, 2018. doi: 10.1111/imr.12666.

COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium. A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell*, 185:916–938.e58, 2022. doi: 10.1016/j.cell.2022.01.012.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7:1–26, 1979. doi: 10.1214/aos/1176344552.

Ryan O. Emerson, William S. DeWitt, Marissa Vignali, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature Genetics*, 49:659–665, 2017. doi: 10.1038/ng.3822.

Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L. Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18:272–282, 2021. doi: 10.1038/s41592-020-01050-x.

Jérôme Hadjadj, Nader Yatim, Laura Barnabei, et al. Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science*, 369:718–724, 2020. doi: 10.1126/science.abc6027.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Zeming Lin, Halil Akin, Roshan Rao, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379:1123–1130, 2023. doi: 10.1126/science.ade2574.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328, 2009. URL https://jmlr.org/papers/v10/liu09a.html.

Sean Nolan, Marissa Vignali, Mark Klinger, et al. A large-scale database of T-cell receptor beta (TCR beta) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2, 2020. URL https://doi.org/10.21203/rs.3.rs-51964/v1. Research Square preprint.

Tobias H. Olsen, Fergus Boyles, and Charlotte M. Deane. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31:141–146, 2022. doi: 10.1002/pro.4205.

Milena Pavlović, Lonneke Scheffer, Keshav Motwani, et al. The immuneML ecosystem for machine learning analysis of adaptive immune receptor repertoires. *Nature Machine Intelligence*, 3:936–944, 2021. doi: 10.1038/s42256-021-00413-z.

Melanie F. Pradier, Niranjani Prasad, Paidamoyo Chapfuwa, Sahra Ghalebikesabi, Maximilian Ilse, Steven Woodhouse, Rebecca Elyanow, Javier Zazo, Javier Gonzalez Hernandez, Julia Greissl, and Edward Meeds. AIRIVA: A deep generative model of adaptive immune repertoires. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219 of *Proceedings of Machine Learning Research*, pages 588–611, 2023. URL https://proceedings.mlr.press/v219/pradier23a.html.

Christoph Schultheiß, Lisa Paschold, Donjete Simnica, et al. Next-generation sequencing of T and B cell receptor repertoires from COVID-19 patients showed signatures associated with severity of disease. *Immunity*, 53:442–455.e4, 2020. doi: 10.1016/j.immuni.2020.06.024.

P. Shinde, F. Soldevila, J. Reyna, et al. A multi-omics systems vaccinology resource to develop and test computational models of immunity. *Cell Reports Methods*, 4:100731, 2024. doi: 10.1016/j.crmeth.2024.100731.

P. Shinde, L. Willemsen, M. Anderson, et al. Putting computational models of immunity to the test: An invited challenge to predict B. pertussis vaccination responses. *PLOS Computational Biology*, 21:e1012927, 2025. doi: 10.1371/journal.pcbi.1012927.

John-William Sidhom, H. Benjamin Larman, Drew M. Pardoll, and Alexander S. Baras. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nature Communications*, 12:1605, 2021. doi: 10.1038/s41467-021-21879-w.

Jason A. Vander Heiden, Susanna Marquez, Nishanth Marthandan, et al. AIRR community standardized representations for annotated immune repertoires. *Frontiers in Immunology*, 9:2206, 2018. doi: 10.3389/fimmu.2018.02206.

D. Wang et al. COMBATdb: a database for the COVID-19 Multi-Omics Blood ATlas. *Nucleic Acids Research*, 51(D1):D896–D905, 2023. doi: 10.1093/nar/gkac1019.

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. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/da4902cb0bc38210839714ebdcf0efc3-Abstract.html.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1703.06114.

M. E. Zaslavsky et al. Disease diagnostics using machine learning of B cell and T cell receptor sequences. *Science*, 387(6736):eadp2407, 2025. doi: 10.1126/science.adp2407.

A Data resources and evaluation sets

COMBAT integrates repertoire, cellular, molecular, and clinical measurements from a shared blood-atlas study (COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium, 2022; Wang et al., 2023). Let $\mathcal{I}_T$ and $\mathcal{I}_B$ index eligible TRB and IGH participant-timepoints. The paired set $P = \mathcal{I}_T \cap \mathcal{I}_B$ contains 136 timepoints from 120 donors; the expanded set $E = \mathcal{I}_T \cup \mathcal{I}_B$ contains 140 from 124. Each set contains ten donors observed twice and three observed three times. All other donors contribute one observation. Target availability within each set gives Table 4; missing targets are excluded from that target's evaluation. The expanded set is used as defined for each representation, including single-modality baselines with their stated missing-input handling.

Table 4: **Outcome-specific COMBAT sample sizes.** P requires both TRB and IGH; E requires at least one. Repeated timepoints share a fold and a bootstrap cluster.

Outcome	P timepoints	P donors	E timepoints	E donors
Age, BMI, sex, mortality	136	120	140	124
Time since onset	126	110	130	114
COVID-19 vs healthy	101	87	103	89
Severe/critical vs mild	91	77	93	79
Sepsis vs healthy	34	32	35	33
Influenza vs healthy	21	21	22	22
Cell composition	95	95	97	97
Serum analytes	97	82	99	84
Molecular PC families	61	52	62	53

The clinical binary targets are COVID-19 versus healthy controls, sepsis versus healthy controls, influenza versus healthy controls, severe/critical versus mild COVID-19, sex, and 28-day mortality. Age, body-mass index, and time since onset are continuous targets. Cell composition comprises nine percentages from the single-cell profiling platform. Serum analytes retain the source assay labels, including names ending in "high" or "low", and are evaluated as continuous measurements. Ten supplied processed components are used for each of the bulk-transcriptome, plasma-proteome, and Luminex PC families. These are fixed source outcomes; the repertoire readout does not refit the source PC decomposition. Correlation among components, analytes, and clinical measurements is retained in the interpretation of nominal coverage.

The processed COMBAT resources are associated with Zenodo record `10.5281/zenodo.5139561` and EGA study `EGAS00001005493`. The structural pipeline requires at least ten eligible clonotypes for each included receptor modality. TRB records are restricted to called cells with valid junction sequences after removing the `TRB_` prefix. IGH structural records require valid junction sequences. The structural validation alphabet includes the 20 standard amino acids and an asterisk; the ESM extraction accepts canonical amino acids only. The supplementary dictionaries specify input fields, coordinate order, and target definitions.

The Emerson resource originally comprises 640 cohort-1 donors (Emerson et al., 2017); the present eligible analysis subset has 324 cohort-1 donors, split into 192 development and 132 confirmation donors, plus 76 Keck donors. Data are associated with immuneACCESS record `10.21417/B7001Z`. The Binder paired-repertoire analysis contains 35 participants with a single active/recovered label and is associated with Schultheiß et al. (2020) and iReceptor accession `IR-Binder-000001` (Corrie et al., 2018). SOTON contains tumour and adjacent-normal specimens from 30 donors, accessed through VDJServer accession `SOTON_iSci_11311R1` (Christley et al., 2018). These resources address different prediction settings and retain their own feature definitions.

B Representation and training details

The structural schema combines 19 TCR scalar summaries, 52 TRBV frequencies, 13 TRBJ frequencies, 25 BCR scalar summaries, 51 IGHV frequencies, and six IGHJ frequencies (Table 5). For TRB, abundance is the number of eligible cell records assigned to a supplied clone identifier. For IGH, the

clonotype key combines junction amino acid sequence and normalized V call. TRA summaries group junction sequences directly. Gene calls are normalized by removing the allele suffix after an asterisk and retaining the first call before comma or slash separators. Usage coordinates are fractions of eligible cell records; absent genes receive zero. The released coordinate dictionary fixes the complete column order and gene vocabulary.

Table 5: **Input dimensions.** The supervised COMBAT configuration projects structural inputs into 16 coordinates used alone by its probe. Full structural dimension is the schema size before any all-missing training columns are discarded.

Component	Coordinates
TCR summaries (including TRA)	19
TRBV usage	52
TRBJ usage	13
BCR summaries (including light chains)	25
IGHV usage	51
IGHJ usage	6
Structural representation	166
ESM-2 TRB + IGH pooling	960
Structure + ESM	1126
Structural bottleneck probe	16

For a repertoire of K clonotypes, relative counts p_j determine clonality and the mass of the largest $k \in \{1, 10, 100\}$ clones. Natural-log entropy enters Equation 2; the helper assigns clonality one for $K = 1$, although the structural eligibility threshold removes such units from the primary evaluation. Lengths l_j are aligned with counts a_j by the same clone key before computing

$$\bar{l} = \frac{\sum_j a_j l_j}{\sum_j a_j}, \qquad s_l = \left(\frac{\sum_j a_j (l_j - \bar{l})^2}{\sum_j a_j} \right)^{1/2}. \tag{7}$$

These are population-weighted moments of the sampled clonotypes. The TCR scalar block also includes total cell recovery, MAIT/iNKT annotation fractions, and the analogous eight TRA summaries where available. The depth-only input contains `tcr_n_cells`, `tcr_n_unique`, `bcr_n_reads`, and `bcr_n_unique`. Here the BCR count derives from eligible heavy-chain records.

IGH mutation summaries use $100 -$ `v_identity_HC`, with mean, sample standard deviation, and 90th percentile across eligible records. Hydrophobicity mean and sample standard deviation use the source `junction_hydrophobicity_HC` annotation. Isotype frequencies cover IGHM, IGHD, IGHG1–4, IGHA1–2, and IGHE after allele removal. Light-chain summaries contain kappa and lambda fractions and mean $100 -$ `v_identity_LC`. These fields, TRA summaries, and cell annotations require the richer assay schema present in COMBAT. The structural representation therefore describes that schema explicitly.

The ESM-2 checkpoint is `facebook/esm2_t12_35M_UR50D` (Lin et al., 2023). Canonical junction amino acid sequences are encoded in batches of 256 with the frozen model in evaluation mode. Final hidden states are averaged over attention-mask-valid tokens, including boundary tokens. Half-precision inference supplies the sequence cache. Repertoire vectors average sequence embeddings over eligible records, equivalently weighting aggregated identical sequences by their observed record counts. Two modality vectors give 960 coordinates. Missing values in structural and sequence inputs are handled within each linear predictor's outer training fold by median imputation followed by standardization; no missingness indicator is appended. All-missing training columns are removed by the imputer. Boosting instead uses its native NaN handling.

For each primary target, multitask training uses participants with that primary outcome observed and the other available outcomes among those participants. Each outer fold has its own structural imputer, feature scaler, encoder, and probe. The hidden layer uses ReLU and dropout 0.3, followed by a linear 16-dimensional bottleneck and separate heads for six binary and 93 continuous outcomes. With all 166 input columns present, the encoder and heads contain 25,123 parameters. Training uses full-batch

AdamW with learning rate 10^{-3}, weight decay 10^{-2}, and 300 epochs. The final-epoch model is used; there is no early stopping or validation-based checkpoint selection. Initialization begins from seed 8128, with the stored execution order defining subsequent random-number consumption.

Each continuous target is centered and scaled using its observed outer-training values. An all-missing training target uses mean zero and scale one and contributes no entries to the masked loss. A constant target retains its training mean and uses scale one. Binary cross-entropy and regression mean-squared error are each averaged over their observed entries and added with equal weight; a loss family with no observed entries contributes zero. No held-out target affects these transforms or the encoder loss. The downstream probe uses the bottleneck alone, with a training-fitted scaler and either balanced logistic regression ($C = 1$, 3,000 maximum iterations) or ridge regression ($\alpha = 10$) on the original primary target units.

Fixed-feature linear models use the same logistic/ridge settings. Histogram gradient boosting uses 200 iterations, learning rate 0.05, maximum depth three, L2 regularization one, and estimator seed zero. Late fusion averages the two modality-specific predictions. Nested selection compares TCR, BCR, direct concatenation, and late fusion in three donor-grouped inner folds, selects by AUROC or Spearman according to target type, and refits the selected option on the outer training data. Binary inner folds are stratified and use the outer split seed plus seven. The outer selection run uses seed 8128.

For fixed structural, sequence, and concatenated configurations, binary predictions are averaged across outer split seeds 8128, 1, and 2 before scoring. Continuous tasks use deterministic five-fold GroupKFold; repeated fixed-feature runs therefore reuse the same split. The multitask and MIL models use one outer split seed, 8128. Three-mer and JL-GloVe continuous runs use the single deterministic split. The saved fold manifests specify the partitions used by each pipeline.

Confidence intervals for COMBAT and SOTON condition on these stored predictions. Each bootstrap draw samples as many donors as occur in the target's evaluation set, with replacement, and includes every valid row of each selected donor. Selecting a donor twice includes its entire row bundle twice. The statistic remains the pooled row-level AUROC or Spearman correlation. Sampling does not replace repeated rows with donor means. Binary intervals use 2,000 draws and continuous intervals 1,000, with seed 8128 and 2.5/97.5 percentiles. Undefined single-class AUROCs or constant-vector correlations are skipped and counted; at most one draw per reported task was skipped. Paired comparisons use the same donor draws. These fixed-prediction intervals preserve donor clustering while leaving model-refitting and selection variability outside their scope.

C SEQUENCE-BASELINE CONFIGURATIONS

The 3-mer representation counts contiguous amino acid triples over eligible CDR3-beta and IGH junction records. A 20-letter alphabet gives 8,000 coordinates per modality. Counts reflect record abundance and are normalized by the total 3-mer count within each chain before concatenation. A missing chain contributes an all-zero block. Linear and boosting probes use the expanded analysis set. This representation supplies a local sequence-composition comparison to the pretrained encoders.

The released JL-GloVe artifact contains 100-dimensional vectors for 535,186 TCR clonotypes learned from 3,991 ImmuneCODE repertoires (Chapfuwa et al., 2025). Our adaptation harmonizes CDR3 boundaries and indexes the core sequence. When multiple released keys share a core, the first encountered vector is retained; V/J identity is not used as an additional matching constraint in this adaptation. Unmatched sequences do not contribute to pooling. COMBAT averages vectors over matched eligible cell records, and missing or wholly unmatched inputs produce zero vectors. The reported median matching fraction is 11.5% of eligible COMBAT sequence records and 8.7% of unique repertoire entries in Emerson. These two percentages retain their respective counting denominators. Emerson evaluates both unweighted and count-weighted pooling with the official vectors, as well as their concatenation with structural summaries.

The attention-MIL comparator uses frozen ESM-2 instances from the top 500 clonotypes in each modality. For instance embedding $e_{ij}^{(m)}$ and count $a_{ij}^{(m)}$, the within-chain computation is

$$h_{ij}^{(m)} = \tanh(W_m e_{ij}^{(m)} + b_m), \tag{8}$$

$$q_{ij}^{(m)} = v_m^\top h_{ij}^{(m)} + c_m + \log(1 + a_{ij}^{(m)}), \tag{9}$$

$$\alpha_{ij}^{(m)} = \frac{\exp q_{ij}^{(m)}}{\sum_{k \in \mathcal{B}_i^{(m)}} \exp q_{ik}^{(m)}}, \qquad z_i^{(m)} = \sum_{j \in \mathcal{B}_i^{(m)}} \alpha_{ij}^{(m)} h_{ij}^{(m)}. \tag{10}$$

Here $h_{ij}^{(m)}$ has 64 dimensions and $\mathcal{B}_i^{(m)}$ indexes non-padding instances. Padded scores are masked before normalization and their weights are removed from the pooled sum. An empty bag yields $z_i^{(m)} = 0$. Each modality has its own projection, score function, and softmax. A count summary $t_i^{(m)} = \sum_j \log(1 + a_{ij}^{(m)})$ accompanies the pooled vector. The paired prediction head is linear in $[z_i^{(T)}; z_i^{(B)}; t_i^{(T)}; t_i^{(B)}]$. This is a specified attention-MIL adaptation (Ilse et al., 2018; Widrich et al., 2020) using ESM-2 inputs.

MIL training is full-batch for 40 epochs with Adam, learning rate 3×10^{-3}, weight decay 10^{-4}, and initialization seed 8128 reset for each fit. Binary heads optimize logit cross-entropy with positive-class weight $\max(n_-/n_+, 1)$. Regression heads optimize MSE after training-fold target standardization, using population standard deviation plus 10^{-8}. Before pooling OOF predictions across folds, outputs are restored by $\widehat{y}_i = \mu_f + \sigma_f \widehat{z}_i$, where μ_f, σ_f are the training-fold constants. Both standardized and restored predictions are retained. This places outputs from different folds on common target units before computing Spearman correlation.

D MATCHED COMPARISONS AND REPRESENTATION PROFILES

Table 6 gives the family-level boosting results on the paired set. Paired inputs reach 14 targets, compared with ten for TCR, seven for BCR, and 17 for depth. Under linear prediction the corresponding totals are 14, 11, seven, and 19. Each comparison holds the readout and eligible sample set fixed while changing the representation.

Table 6: **Matched gradient-boosting coverage.** All configurations use the paired set and target-specific donor grouping.

Assay family	Targets	TCR	BCR	TCR+BCR	Depth
Clinical/demographic	9	3	0	4	3
Cell composition	9	3	4	6	6
Serum analytes	51	1	2	4	6
Luminex PCs	10	0	1	0	0
Proteome PCs	10	2	0	0	1
Transcriptome PCs	10	1	0	0	1
Total	99	10	7	14	17

Target membership describes the effect of pairing beyond the totals. Let $U = A_T \cup A_B$ be the union of single-chain covered targets and A_{TB} the paired set. Then

$$|A_{TB}| - |U| = |A_{TB} \setminus U| - |U \setminus A_{TB}|. \tag{11}$$

The linear comparison yields $3 - 5 = -2$ and boosting yields $7 - 8 = -1$ (Table 7). Linear paired-specific targets are interferon-alpha, IL-15, and Luminex PC5. The boosting paired-specific set contains BMI, HSC and NK fractions, CCL3/MIP-1alpha, IL-10, IL-6, and TREM-1. These target-set differences describe the nominal criterion separately from a paired test of each performance difference.

The representation-block comparison is performed on the expanded set (Table 8). Structure supplies the broadest overall coverage, while concatenation with ESM increases serum coverage from three

Table 7: **Target membership under matched readouts.** Paired-specific targets lie outside the single-chain union; single-specific targets lie outside the paired set.

Readout	TCR	BCR	Single union	Paired	Paired-specific	Single-specific
Linear	11	7	16	14	3	5
Boosting	10	7	15	14	7	8

to four. The structural bottleneck concentrates its 14 targets in clinical, cellular, and molecular-component outcomes. All four block configurations together cover 27 target labels.

Table 8: **Representation-block coverage by assay family.** All configurations use the expanded set; linear probes use training-fold preprocessing.

Assay family	Targets	Structure	ESM	Structure+ESM	Bottleneck
Clinical/demographic	9	4	1	4	6
Cell composition	9	4	0	3	4
Serum analytes	51	3	2	4	0
Luminex PCs	10	2	1	0	1
Proteome PCs	10	0	1	1	2
Transcriptome PCs	10	3	0	2	1
Total	99	16	5	14	14

E CROSS-MODAL PREDICTION MAP

The paired structural family includes concatenation, late fusion, and nested modality selection under both evaluated readouts. Its 28-target set overlaps the 27-target representation family in 17 targets, giving 38 aggregate targets (Table 9). Each target retains a qualifying configuration and its originating evaluation set. The paired-set single-chain family union contains 21 targets, 17 shared with the aggregate map. Thus, 21 aggregate targets lie outside that single-chain union, four single-chain targets remain outside the aggregate, and the combined list contains 42 labels across these records.

Table 9: **Coverage across ImmuneEmbedding configurations.** P denotes the paired set and E the expanded set. The aggregate is the union of qualifying target labels across the two families, each retaining its predictor and evaluation sample.

Assay family	Targets	Structural P	Representation E	Aggregate
Clinical/demographic	9	6	6	6
Cell composition	9	7	4	7
Serum analytes	51	8	8	14
Luminex PCs	10	3	3	4
Proteome PCs	10	2	3	4
Transcriptome PCs	10	2	3	3
Total	99	28	27	38

The two ImmuneEmbedding families each contain eight serum targets, with two shared, producing 14 in the aggregate. Table 10 gives the alternative families: 18 targets for attention-MIL, 11 for released JL-GloVe, 15 for 3-mers, and 21 for the depth readout family. Seven attention-MIL targets lie outside the aggregate ImmuneEmbedding set. The family unions summarize collections of configurations, whereas Table 2 reports each individual configuration.

For each serum target, Table 11 selects the highest point estimate among configurations satisfying the nominal interval criterion. Selection preserves the originating set and predictor. P-based serum

Table 10: **Model-family coverage.** MIL pools its three modality configurations; JL-GloVe and 3-mers pool linear and boosting readouts on E. Depth pools the two readouts on P. These columns summarize different evaluated configuration families.

Assay family	Targets	MIL	JL-GloVe	3-mer	Depth
Clinical/demographic	9	3	1	3	6
Cell composition	9	7	1	6	7
Serum analytes	51	7	8	2	6
Luminex PCs	10	0	0	2	0
Proteome PCs	10	0	1	1	1
Transcriptome PCs	10	1	0	1	1
Total	99	18	11	15	21

predictions use 97 timepoints from 82 donors; E-based predictions use 99 from 84. Exact interval endpoints determine membership and displayed values are rounded to three decimals. The 14 labels span inflammatory, complement, and other soluble immune measurements; their reported uncertainty is conditional on the selected configuration.

Table 11: **Serum targets in the aggregate prediction map.** Each row is the highest point estimate among qualifying ImmuneEmbedding configurations. Intervals use donor-cluster resampling of fixed predictions. P has 97 serum timepoints from 82 donors; E has 99 from 84. The selection is targetwise and unadjusted for multiplicity.

Serum analyte	Set / predictor	n	Spearman [95% CI]
IL-10	P / Concat., boosting	97	0.323 [0.160, 0.500]
Complement C5a	P / Selection, boosting	97	0.300 [0.083, 0.485]
IL-15	P / Concat., linear	97	0.280 [0.070, 0.467]
IFN-alpha	P / Selection, boosting	97	0.278 [0.072, 0.466]
TREM-1	E / Structure+ESM, linear	99	0.261 [0.055, 0.460]
IL-17/IL-17A	E / ESM, linear	99	0.256 [0.043, 0.432]
IL-6	P / Concat., boosting	97	0.242 [0.042, 0.439]
IL-12 p70	E / ESM, linear	99	0.240 [0.075, 0.404]
Lipocalin-2/NGAL	E / Structure+ESM, linear	99	0.220 [0.008, 0.407]
CCL3/MIP-1alpha	P / Concat., boosting	97	0.215 [0.001, 0.417]
CCL18/PARC	P / Fusion, linear	97	0.205 [0.020, 0.377]
G-CSF	E / Structure+ESM, linear	99	0.204 [0.007, 0.387]
CCL19/MIP-3beta	E / Structure+ESM, linear	99	0.200 [0.001, 0.381]
IL-13	E / Structure, linear	99	0.192 [0.021, 0.356]

F External-transfer and resource analyses

The Emerson cohort-calibration experiment concatenates repertoire statistics, TRBV usage, and previously constructed V/J-reference features. Feature medians and the uncalibrated scaler are fitted on 192 development donors. Equation 5 is computed separately for development, confirmation, and Keck groups after development-median imputation. The 132 confirmation and 76 Keck labels remain held out. Reference-anchored inputs and count summaries distinguish this assay-specific TCR representation from the 166-coordinate COMBAT schema.

The Emerson residual uses hidden width 64, dropout 0.4, and a 16-dimensional bottleneck, with masked CMV, sex, and age supervision restricted to development donors. Age centering and scaling use development values only. AdamW trains for 300 epochs at learning rate 10^{-3} and weight decay 2×10^{-2}. Latent vectors from initialization seeds 8128, 1, 7, 42, and 99 are averaged and appended to the calibrated features before the final probe. The paired residual-minus-rank intervals are 0.017 $[-0.009, 0.043]$ in confirmation and 0.013 $[-0.032, 0.061]$ in Keck. These estimates describe an append-based external residual, separately from COMBAT's bottleneck-only probe.

The official JL-GloVe comparison evaluates the released vectors and their pooling variants on the same evaluation groups. Its confirmation model uses the 192 development donors, and its Keck model uses all 324 eligible cohort-1 donors. Logistic regression is L2-regularized with balanced classes and a five-fold search over ten regularization values. Table 3 reports the canonical official-embedding experiment. The mean-pooling estimates are 0.651 and 0.586, and adding structure yields 0.797 and 0.700. Bootstrap intervals for these independent-donor evaluations use the original 2,000 resamples and are displayed at three-decimal precision from the retained metric records.

Binder uses one repertoire vector per participant and five stratified five-fold split seeds, 8128, 1, 2, 3, and 4. The depth-excluded configuration removes `n_reads` and `n_unique` from each of the resource's structural vectors, with the remaining coordinates computed by that resource's feature functions. No cohort-rank normalization is used. The final probabilities are averaged across seeds and the same donor resamples are used for paired differences (Table 12).

Table 12: **Binder active-versus-recovered classification.** All estimates use the same 35 participants and depth-excluded inputs. Difference intervals compare the same resampled donors.

Configuration or contrast	Estimate	95% CI
TCR	0.820	[0.653, 0.947]
BCR	0.667	[0.461, 0.843]
TCR+BCR	0.797	[0.628, 0.931]
Paired minus TCR	-0.023	[-0.131, 0.078]
Paired minus BCR	0.131	[-0.043, 0.327]

SOTON keeps tumour and normal specimens from each donor in the same stratified grouped fold and averages probabilities over the same five split seeds. The reported depth-excluded inputs remove the resource's count and unique-clonotype coordinates. Donor-cluster intervals preserve both specimens and give paired AUROC 0.786 [0.640, 0.913] (Table 13). Paired accuracy measures whether the tumour score is greater than the normal score within each donor. The observed paired-minus-TCR difference is 0.019 and paired-minus-BCR is 0.038; both paired intervals span zero.

Table 13: **SOTON tumour-versus-normal classification.** Sixty specimens from 30 donors are evaluated with donor-grouped cross-validation. Intervals and paired differences use common donor-cluster resamples. Difference point estimates are differences between the observed AUROCs. Paired accuracy is the fraction of donors for whom the tumour score exceeds the normal score.

Configuration or contrast	AUROC / difference	95% CI	Paired accuracy
TCR	0.767	[0.667, 0.867]	0.800
BCR	0.748	[0.588, 0.897]	0.767
TCR+BCR	0.786	[0.640, 0.913]	0.800
Paired minus TCR	0.019	[-0.150, 0.182]	
Paired minus BCR	0.038	[-0.009, 0.093]	

CMI-PB provides a systems-vaccinology resource and prediction challenge (Shinde et al., 2024; 2025). The retained analysis trains on earlier-year cohorts and evaluates 54 participants from the 2023 cohort using baseline tabular multi-omics inputs. It evaluates a preprocessing question separate from repertoire prediction. Rank calibration raises vaccine-history AUROC from 0.446 to 0.671 and day-28 pertactin-IgG fold-change Spearman from 0.289 to 0.582. The residual configurations yield 0.665 and 0.565, respectively (Table 14), showing endpoint-specific changes from supervision.

The retained resource analyses also include study-specific BCR classification from Observed Antibody Space (Olsen et al., 2022) and TCR evaluations in ImmuneCODE (Nolan et al., 2020). Their quantitative records are supplied separately from the corrected COMBAT benchmark. ImmuneCODE reference-anchored evaluations use a development dataset and external datasets to classify source-recorded COVID-19 positivity. Across evaluated variants, external AUROCs are approximately 0.44–0.53, despite a development estimate reaching 0.616. These results characterize the evaluated

Table 14: **CMI-PB calibration with non-repertoire inputs.** Baseline tabular multi-omics predicts vaccine history (AUROC) and day-28 pertactin-IgG fold-change (Spearman) on 54 external participants. Within each task, the three configurations share their input setting.

Endpoint	Configuration	Estimate [95% CI]
Vaccine history	Uncalibrated	0.446 [0.308, 0.589]
	Rank	0.671 [0.522, 0.808]
	Rank + residual	0.665 [0.512, 0.811]
Day-28 pertactin-IgG fold-change	Uncalibrated	0.289 [0.004, 0.543]
	Rank	0.582 [0.354, 0.768]
	Rank + residual	0.565 [0.352, 0.729]

representation and cohort setting. They provide context for task-dependent transfer alongside the CMV experiments.

A secondary COMBAT summary predicts 51 serum analytes from bulk-transcriptome PCs, plasma-proteome PCs, and cell fractions using grouped-CV ridge regression. The retained summary reports at most three targets above a point Spearman threshold of 0.2 and a median of 0.02. Its inputs, readout, and point threshold define that secondary comparison; it is not combined with the nominal interval-based repertoire coverage.