

Prototype embeddings for immune receptors: geometry, statistics, and the TCREMP coordinate system

Mikhail Shugay^{1,2,3*}

¹Institute of Translational Medicine, Russian National Medical State University, Moscow, Russia

²Department of Genomics of Adaptive Immunity, Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry
RAS, Moscow, Russia

³Institute of Molecular Biology NAS RA, Yerevan, Armenia

Appendix T — TCREMP embedding geometry

Abstract

TCREMP [1] embeds an immune receptor as the vector of its alignment distances to a fixed set of *prototype* receptors sampled from the V(D)J recombination model [2, 3]. This appendix gives the mathematics behind that construction. We first develop the general theory of *prototype* (landmark) embeddings of a metric space — their Hölder/Lipschitz-continuity and isometry properties, the negative-type condition that makes the alignment dissimilarity an honest squared Hilbert distance [4] with metric $\rho = \sqrt{d}$, the energy functional that makes embedding-space Euclidean distance track pairwise alignment distance, and a landmark sample-complexity bound (§T.1–§T.2). We then specialise to immune receptors: taking the prototype measure to be the generation distribution P_{gen} turns each coordinate into a Monte-Carlo sketch of a receptor’s relationship to the generative repertoire, which explains why the choice of prototype source scarcely matters and yields the observed Gamma / extreme-value distribution laws (§T.3). The three-component structure of the TCR embedding and paired-chain concatenation are read off as a product metric, with the germline blocks provably low-rank (§T.4); somatic hypermutation is a bounded perturbation that explains why IGH is the hardest chain (§T.5); and the pushforward of P_{gen} onto embedding space is the null model for density-based background subtraction (§T.6). Lifting from single receptors to whole repertoires, a sample is summarised as one fixed vector — a kernel mean, a coverage-standardised diversity profile, and a second moment (§T.7) — that also serves as a neural-network modality (§T.8); and the embedding’s approximate invertibility, the privacy it implies, and the open frontier close the account (§T.9–§T.11). Every quantitative claim is reproduced in-repo (§T.11, Table 4) by `mir/bench/theory.py`.

PART I

THE SINGLE-RECEPTOR EMBEDDING

*Correspondence: mikhail.shugay@gmail.com

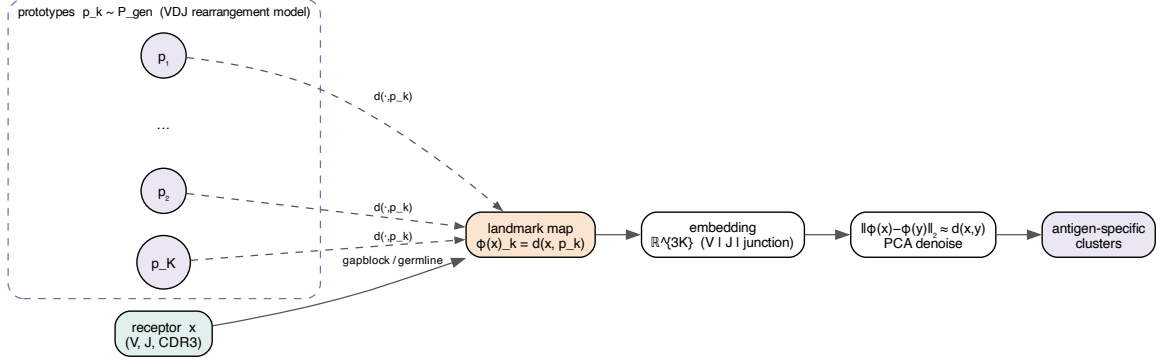


Figure 1: The TCREMP prototype-embedding pipeline. A receptor x is mapped to the vector of its alignment distances to prototypes p_k sampled from the recombination model P_{gen} ; Euclidean distance in the resulting $\mathbb{R}^{3K}$ coordinate system approximates pairwise alignment distance (Prop. 1–3), and specificity groups become compact clusters.

Mapping one receptor to a point: the geometry of distance-to-prototype coordinates, the metric behind the alignment dissimilarity, and the immune-receptor biology those coordinates encode.

T.1. FOUNDATIONS OF PROTOTYPE EMBEDDINGS

SETUP. Let X be a space of receptor sequences carrying a symmetric dissimilarity d with $d(x, x) = 0$ and $d \geq 0$ (verified for the alignment dissimilarity in §T.2). Crucially d is a *negative-type semimetric*: it is the squared distance $d = \|\psi(\cdot) - \psi(\cdot)\|^2$ of a Hilbert feature map ψ (§T.2), so it does *not* obey the triangle inequality and is *not* a metric. The induced genuine metric is $\rho := \sqrt{d}$ (Schoenberg, Prop. 5); write $\text{diam} := \sup_{a, b \in X} \rho(a, b) = \sup_{a, b} \sqrt{d(a, b)}$ for its (finite) diameter. Fix an ordered set of *prototypes* $P = \{p_1, \dots, p_K\} \subset X$. The **prototype embedding** is the landmark map

$$\varphi : X \rightarrow \mathbb{R}^K, \quad \varphi(x) = (d(x, p_1), d(x, p_2), \dots, d(x, p_K)), \quad (1)$$

and the induced **embedding distance** is $D(x, y) = \|\varphi(x) - \varphi(y)\|_2$. Figure 1 shows the pipeline. In TCREMP the prototypes are drawn from P_{gen} , d is a BLOSUM alignment dissimilarity, and φ carries three coordinate blocks per prototype (§T.4); the foundational results below hold for any (X, d, P) .

CONTINUITY OF THE COORDINATES. The single structural fact from which everything else follows is a modulus of continuity for each coordinate of φ . Because the shipped coordinates $\varphi(x)_k = d(x, p_k)$ are *squared* distances, they are not 1-Lipschitz in d ; they are Hölder- $\frac{1}{2}$ in d — equivalently Lipschitz in the induced metric $\rho = \sqrt{d}$.

PROPOSITION 1 (Hölder coordinates and the upper bound). *For every prototype p and all $x, y \in X$,*

$$|d(x, p) - d(y, p)| \leq 2 \text{diam} \sqrt{d(x, y)}, \quad (2)$$

i.e. the coordinates are Hölder- $\frac{1}{2}$ in d with constant 2diam — equivalently 1-Lipschitz in the induced metric $\rho = \sqrt{d}$, $|\rho(x, p) - \rho(y, p)| \leq \rho(x, y)$. Consequently

$$\begin{aligned} \|\varphi(x) - \varphi(y)\|_\infty &\leq 2 \text{diam} \sqrt{d(x, y)}, \\ D(x, y) = \|\varphi(x) - \varphi(y)\|_2 &\leq 2 \text{diam} \sqrt{K d(x, y)} = 2 \text{diam} \sqrt{K} \rho(x, y). \end{aligned} \quad (3)$$

Proof. Because $\rho = \sqrt{d}$ is a genuine metric (Prop. 5), the reverse triangle inequality gives $|\rho(x, p) - \rho(y, p)| \leq \rho(x, y)$. Factoring the difference of squares, $|d(x, p) - d(y, p)| = |\rho(x, p) - \rho(y, p)| (\rho(x, p) + \rho(y, p)) \leq \rho(x, y) \cdot 2 \text{diam} = 2 \text{diam} \sqrt{d(x, y)}$, since $\rho(x, p), \rho(y, p) \leq \text{diam}$. Taking the maximum over $p \in P$ gives the ℓ^∞ bound; summing K squares each $\leq (2 \text{diam})^2 d(x, y)$ and taking the root gives the ℓ^2 bound. $\square$

Proposition 1 is the rigorous half of the headline “embedding distance $\leq$ alignment distance” claim: no pair is ever *farther* apart in embedding space than $2 \text{diam} \sqrt{K}$ times the metric $\rho(x, y) = \sqrt{d(x, y)}$ — the coordinate map is Lipschitz in ρ , Hölder- $\frac{1}{2}$ in d . The reverse — that close-in-embedding implies close-in-alignment — is not automatic and is where the prototypes must do work.

EXACTNESS IN THE LIMIT. With enough prototypes the ℓ^∞ bound is tight.

PROPOSITION 2 (Kuratowski isometry). *Take $P = X$. The metric coordinates $x \mapsto (\rho(x, p))_p$ embed (X, ρ) isometrically into $(\ell^\infty, \|\cdot\|_\infty)$: $\sup_p |\rho(x, p) - \rho(y, p)| = \rho(x, y)$. For the shipped squared coordinates $\varphi(x) = (d(x, p))_p$ this is not an isometry, but the pairwise dissimilarity is still recovered exactly — at $p \in \{x, y\}$, $|d(x, p) - d(y, p)| = d(x, y)$ — so $d(x, y) \leq \|\varphi(x) - \varphi(y)\|_\infty \leq 2 \text{diam} \sqrt{d(x, y)}$.*

Proof. For the metric coordinates, the upper bound $\sup_p |\rho(x, p) - \rho(y, p)| \leq \rho(x, y)$ is the reverse triangle inequality (Prop. 1), and $p = y$ attains it: $|\rho(x, y) - 0| = \rho(x, y)$. For the squared coordinates, $p = y$ gives $|d(x, y) - d(y, y)| = d(x, y)$ (lower bound) and the upper bound is Prop. 1. $\square$

The construction of Eq. (1) is thus a Kuratowski/Fréchet embedding — isometric in the induced metric ρ . The supplementary experiments that use the sequences themselves as prototypes (§ S1–S2 of [1]) are exactly this regime: each pairwise $d(x, y)$ appears verbatim as the x - and y -prototype coordinates, read out in ℓ^2 rather than ℓ^∞ . Passing from ℓ^∞ to ℓ^2 costs the $\sqrt{K}$ factor of Eq. (3) but makes the coordinates amenable to PCA and Euclidean clustering.

WHY ℓ^2 DISTANCE TRACKS ALIGNMENT DISTANCE. Fix a prototype-sampling measure μ on X (in TCREMP, $\mu = P_{\text{gen}}$) and draw $p_1, \dots, p_K$ i.i.d. from μ .

PROPOSITION 3 (Energy functional). *Define the μ -energy dissimilarity $\bar{D}_\mu^2(x, y) := \mathbb{E}_{p \sim \mu} [(d(x, p) - d(y, p))^2]$. Then $\frac{1}{K} D(x, y)^2 \rightarrow \bar{D}_\mu^2(x, y)$ almost surely as $K \rightarrow \infty$, and $\bar{D}_\mu$ is a μ -dependent semimetric with $0 \leq \bar{D}_\mu^2(x, y) \leq (2 \text{diam})^2 d(x, y) = 4 \text{diam}^2 d(x, y)$.*

Proof. $\frac{1}{K} D^2 = \frac{1}{K} \sum_k (d(x, p_k) - d(y, p_k))^2$ is an empirical mean of i.i.d. terms bounded by $(2 \text{diam})^2 d(x, y)$ (Prop. 1); the strong law gives the limit and the bound, and symmetry plus the triangle inequality for $\|\cdot\|_2$ make $\bar{D}_\mu$ a semimetric. The limit $\bar{D}_\mu^2$ is unchanged from the naïve normalisation; only the constant in the range differs. $\square$

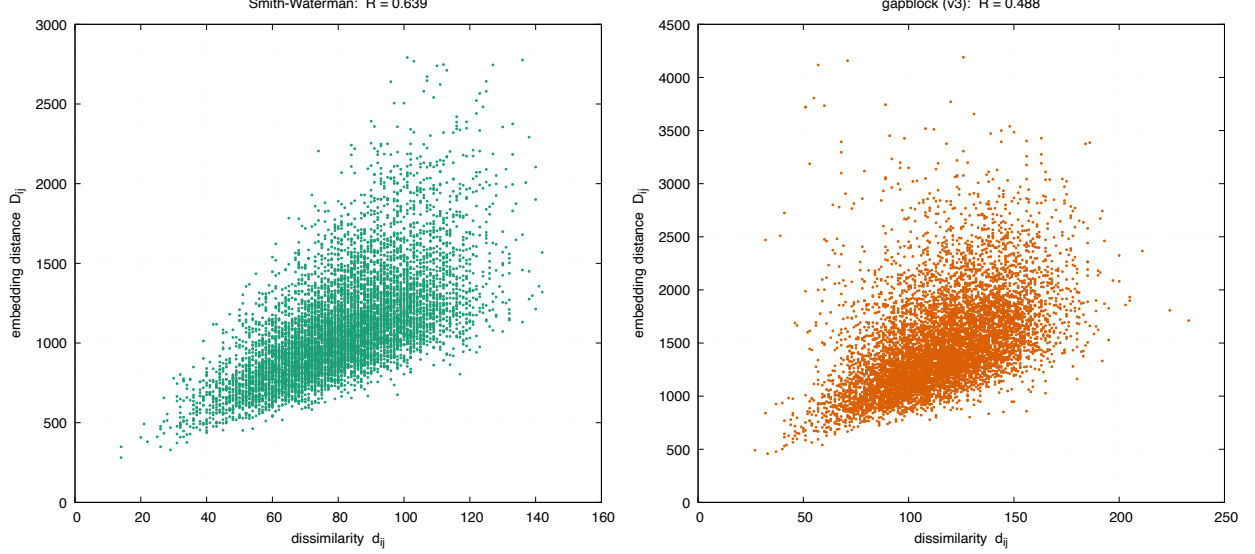


Figure 2: Embedding-space Euclidean distance D_{ij} against pairwise dissimilarity d_{ij} for $n = 2500$ CDR3 β used as their own prototypes (Prop. 3; reproduces suppl. S2). Left: Smith–Waterman d ; right: the `gapblock` dissimilarity shipped in `v3`. Generated by `mir/bench/theory.py::s2_dissimilarity_distance_correlation`.

Proposition 3 says D estimates a fixed monotone-in- d functional $\bar{D}_\mu$: two receptors that are close under d have similar distance profiles to the whole of μ and hence small $\bar{D}_\mu$, while dissimilar receptors have systematically different profiles. This is the mechanism behind supplementary Fig. S2 — the empirical Pearson correlation between D_{ij} and d_{ij} (Fig. 2, $R = 0.64$ under Smith–Waterman, 0.49 under the `gapblock` dissimilarity of § T.2) — and behind S2c/d, where D_{ik} concords with means of d_{ik} over the prototype set: those means are Monte-Carlo estimates of $\bar{D}_\mu$.

HOW MANY PROTOTYPES. The $\sqrt{K}$ gap of Eq. (3) is worst-case; for the *random* landmark map the relative fluctuation of $\frac{1}{K}D^2$ around $\bar{D}_\mu^2$ is $O(K^{-1/2})$ by Prop. 3, so a Johnson–Lindenstrauss-style argument [5] on the bounded coordinates gives the following.

PROPOSITION 4 (Landmark sample complexity). *Let $S \subset X$ with $|S| = n$. For $K = O(\varepsilon^{-2} \log n)$ i.i.d. prototypes, with high probability every pair $x, y \in S$ satisfies $|\frac{1}{K}D(x, y)^2 - \bar{D}_\mu^2(x, y)| \leq \varepsilon (2 \text{diam})^2 d(x, y)$ simultaneously.*

Proof. Fix a pair $x, y \in S$. Prototype $p_k \sim \mu$ contributes $Y_k = (d(x, p_k) - d(y, p_k))^2$, and the Hölder bound of Prop. 1 gives $0 \leq Y_k \leq (2 \text{diam})^2 d(x, y)$, so the Y_k are i.i.d. and confined to an interval of length $R := (2 \text{diam})^2 d(x, y)$ with mean $\mathbb{E}_\mu[Y] = \bar{D}_\mu^2(x, y)$ (Prop. 3). Their average is $\frac{1}{K} \sum_k Y_k = \frac{1}{K} D(x, y)^2$, so Hoeffding’s inequality yields $\Pr[|\frac{1}{K} D(x, y)^2 - \bar{D}_\mu^2(x, y)| > \varepsilon R] \leq 2 \exp(-2K\varepsilon^2)$. A union bound over the $\binom{n}{2} < n^2$ pairs caps the total failure probability at $2n^2 \exp(-2K\varepsilon^2)$, which tends to 0 as soon as $K \geq C\varepsilon^{-2} \ln n$, i.e. $K = \Theta(\varepsilon^{-2} \log n)$: the Hölder constant enters only the error scale R , not the rate. $\square$

The logarithmic dependence on n is the theoretical counterpart of Fig. S4 of the published sup-

plement [1], where the UMAP of an antigen-specific set stabilises once $K \gtrsim 100$ prototypes are used: the embedding geometry saturates at a prototype count that grows only logarithmically with the data. Prototype embeddings are the *empirical* / Nyström feature map of the alignment kernel, and Landmark MDS [6, 7] and the dissimilarity representation [8] are the classical instances; the PCA step of §T.4 is Landmark MDS on φ .

T.2. THE ALIGNMENT DISSIMILARITY: A NEGATIVE-TYPE SEMIMETRIC

THE GRAM CONSTRUCTION. TCREMP scores two sequences by an alignment similarity s and converts it to a dissimilarity by

$$d(a, b) = s(a, a) + s(b, b) - 2s(a, b). \quad (4)$$

This is exactly the squared distance $\|\psi(a) - \psi(b)\|^2$ induced by a feature map ψ whenever s is an inner product, $s(a, b) = \langle \psi(a), \psi(b) \rangle$. The v3 implementation (`mir/distances/junction.py`) reads d off directly: `seqtree's` `blosum62()` is already the Gram transform of Eq. (4), and `gapblock.score_matrix` returns d (zero diagonal, symmetric, non-negative) — a squared Hilbert distance, hence a *negative-type semimetric* rather than a metric; the induced genuine metric is $\rho = \sqrt{d}$. It selects the best of contiguous gap-block placements at offsets (3, 4, −4, −3).

DEFINITION 1. A symmetric d with zero diagonal is of *negative type* (equivalently, $-d$ is conditionally positive definite) if $\sum_{i,j} c_i c_j d(x_i, x_j) \leq 0$ for all finite $\{x_i\}$ and all $\{c_i\}$ with $\sum_i c_i = 0$.

PROPOSITION 5 (When Eq. (4) is an honest distance). *d of Eq. (4) is of negative type if and only if the centred similarity matrix s is positive semidefinite, in which case $\sqrt{d}$ embeds isometrically into a Hilbert space and $d = \|\psi(\cdot) - \psi(\cdot)\|^2$ for the associated feature map ψ .*

Proof. Suppose d is of negative type. Fix a base point x_0 and form the centred kernel $\tilde{s}(a, b) = \frac{1}{2}(d(a, x_0) + d(b, x_0) - d(a, b))$. Given any coefficients $c_1, \dots, c_m$, set $c_0 = -\sum_{i \geq 1} c_i$ so that $\sum_{i \geq 0} c_i = 0$; expanding, $\sum_{i,j \geq 1} c_i c_j \tilde{s}(x_i, x_j) = -\frac{1}{2} \sum_{i,j \geq 0} c_i c_j d(x_i, x_j) \geq 0$ by negative type, so $\tilde{s}$ is positive semidefinite. As a PSD kernel $\tilde{s}$ is a Gram matrix, $\tilde{s}(a, b) = \langle \psi(a), \psi(b) \rangle$ for a Hilbert-space map ψ , and $\|\psi(a) - \psi(b)\|^2 = \tilde{s}(a, a) + \tilde{s}(b, b) - 2\tilde{s}(a, b) = d(a, b)$, so $\sqrt{d}$ embeds isometrically. Expanding this square is Eq. (4) with $s(a, b) = \langle \psi(a), \psi(b) \rangle$, identifying negative type of d with positive semidefiniteness of the centred similarity. Conversely, if s is PSD then Eq. (4) is manifestly a squared Hilbert distance, hence of negative type. This is Schoenberg's theorem [4]; [9, 10] give bi-Lipschitz refinements. $\square$

REMARK 1 (Which coordinates ship). TCREMP embeds with the *squared* dissimilarity, $\varphi(x)_k = d(x, p_k)$, not with the metric $\rho(x, p_k) = \sqrt{d(x, p_k)}$. The two induce different geometries; a ρ -coordinate variant is available (`metric="sqrt"` in `mir/distances/junction.py`) and was benchmarked, but the squared coordinates cluster at least as well on the VDJDB retention/F1 metrics, so v3 ships d . The price is only theoretical bookkeeping: the coordinate map is then Hölder- $\frac{1}{2}$ in d rather than 1-Lipschitz (Prop. 1), which propagates a constant 2 diam (the ρ -diameter) through the geometry results but changes none of their limits or rates.

ALIGNMENT SCORES NEED CORRECTION. General pairwise-alignment similarities (Smith–Waterman, BLAST) are *not* positive semidefinite, so the centred similarity of Eq. (4) can fail to be PSD and d can violate $d \geq 0$ or negative type; canonicalising an alignment score into a negative-type dissimilarity (so that $\rho = \sqrt{d}$ is a genuine metric) requires an explicit correction [11]. In v3 this correction is the non-negativity clamp inside `seqtree`’s Gram transform — a projection of the score onto the nearest conditionally-negative-definite cone — which restores the negative type that Prop. 5, and through it Prop. 1–2, rely on. The choice of s (BLOSUM62 here; tcrBLOSUM or a VDJ-specific matrix are alternatives) selects *which* metric ρ ; all are admissible provided the clamp restores negative type.

GAP-BLOCK VERSUS SMITH–WATERMAN. The v3 junction dissimilarity replaces a full Smith–Waterman optimisation by the best of four predetermined contiguous gap-block placements. When the optimal CDR3 alignment inserts a single contiguous gap inside the hypervariable core — the generic case once the conserved Cys/Phe anchors are fixed — the gap-block score equals the Smith–Waterman optimum; the two diverge only for multi-gap alignments of very unequal lengths. Empirically the induced coordinate systems are strongly but not perfectly aligned (Fig. 2: the same D – d correlation is $R = 0.64$ for Smith–Waterman and $R = 0.49$ for `gapblock`), so v3 embeddings are a *new, versioned* coordinate system rather than a bit-for-bit reproduction of the published one.

T.3. IMMUNE RECEPTORS AND THE RECOMBINATION MODEL

THE PROTOTYPE MEASURE IS P_{gen} . The diversity of an immune repertoire is astronomically large, but it is not arbitrary: it is the image of a low-entropy stochastic process, the V(D)J recombination model P_{gen} [2, 12, 3]. TCREMP takes the prototype-sampling measure of §T.1 to be exactly this model, so each coordinate $\varphi(x)_k = d(x, p_k)$, $p_k \sim P_{\text{gen}}$, measures the alignment distance from x to a typical product of recombination.

PROPOSITION 6 (Distance-profile functional). *For $p \sim P_{\text{gen}}$ the empirical distribution of $\{\varphi(x)_k\}_{k=1}^K$ converges weakly to the law of $d(x, p)$. Hence $\varphi(x)$ is a Monte-Carlo sketch of the map $x \mapsto \text{Law}_{p \sim P_{\text{gen}}} d(x, p)$, and two receptors with the same distance-profile-to- P_{gen} are asymptotically embedding-indistinguishable.*

Proof. The coordinates are i.i.d. evaluations of the fixed measurable map $p \mapsto d(x, p)$ under P_{gen} ; the Glivenko–Cantelli theorem gives uniform convergence of their empirical law to that of $d(x, p)$. $\square$

This reframes what the embedding “learns”: not a language model of receptor sequences — which the entropy of P_{gen} (~ 40 bits for TRB) makes futile — but each receptor’s geometric relationship to the generative distribution itself. It also predicts that the embedding is insensitive to how the prototypes are obtained, provided they cover the same support.

PROPOSITION 7 (Prototype-source robustness). *Let μ, ν be two prototype measures that are mutually absolutely continuous with bounded density ratio $r_{\mu\nu} = d\mu/d\nu$ on the support of the repertoire. Then the energy dissimilarities of Prop. 3 satisfy $\bar{D}_\mu^2(x, y) = \mathbb{E}_\nu[r_{\mu\nu}(p) (d(x, p) - d(y, p))^2]$, so the two embeddings induce distances that agree up to the χ^2 discrepancy between μ and ν ; as $\chi^2(\mu \parallel \nu) \rightarrow 0$ the pairwise-distance vectors become perfectly correlated.*

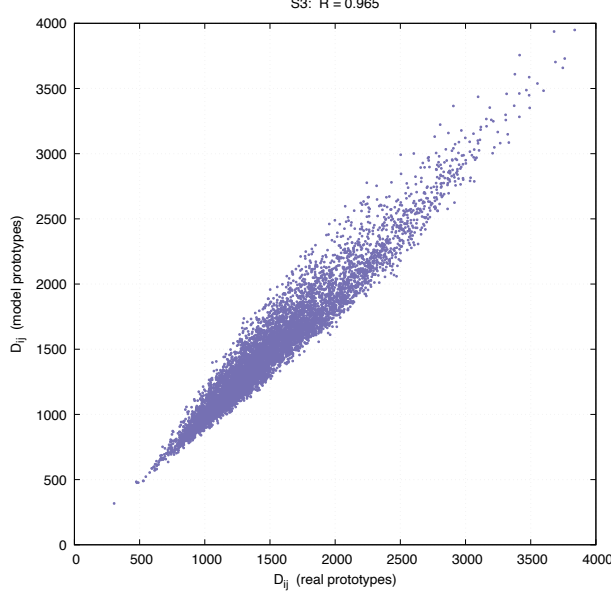


Figure 3: Pairwise embedding distances built from real (Britanova [13]) versus model (OLGA/Murugan [2, 3]) prototypes agree ($R = 0.97$; suppl. S3), confirming Prop. 7. `mir/bench/theory.py::prototype_source_correlation`.

Proof. By Prop. 3, $\bar{D}_\mu^2(x, y) = \mathbb{E}_\mu[u_{xy}(p)]$ with $u_{xy}(p) = (d(x, p) - d(y, p))^2$. Writing the μ -expectation as a ν -expectation with importance weight $r_{\mu\nu} = d\mu/d\nu$ gives $\bar{D}_\mu^2(x, y) = \mathbb{E}_\nu[r_{\mu\nu}(p) u_{xy}(p)]$, the first claim. Its discrepancy from the ν -embedding is $\bar{D}_\mu^2 - \bar{D}_\nu^2 = \mathbb{E}_\nu[(r_{\mu\nu} - 1) u_{xy}]$, and Cauchy–Schwarz bounds this by $\sqrt{\mathbb{E}_\nu[(r_{\mu\nu} - 1)^2]} \sqrt{\mathbb{E}_\nu[u_{xy}]} = \sqrt{\chi^2(\mu\|\nu)} \sqrt{\mathbb{E}_\nu[u_{xy}]}$, using $\mathbb{E}_\nu[(r_{\mu\nu} - 1)^2] = \mathbb{E}_\nu[r_{\mu\nu}^2] - 1 = \chi^2(\mu\|\nu)$. Hence the two families of pairwise distances differ by a factor set by $\chi^2(\mu\|\nu)$; as $\chi^2(\mu\|\nu) \rightarrow 0$ the perturbation vanishes uniformly and their correlation tends to one. $\square$

A real healthy-donor repertoire (Britanova *et al.* [13]) and the synthetic recombination model (Murugan *et al.* [2]) are, after thymic selection, close in exactly this sense, which is why supplementary Fig. S3 finds their embedding distances correlated at $R = 0.96$ — reproduced here at $R = 0.97$ (Fig. 3).

DISTRIBUTION LAWS. The random construction also fixes the shape of the two histograms that dominate any embedding analysis.

PROPOSITION 8 (Dissimilarity and distance laws). *Under mild mixing of per-position alignment contributions: (i) the pairwise dissimilarity d_{ij} — a sum of many non-negative, bounded per-column penalties — is asymptotically Gaussian by a central-limit tendency, with the positivity constraint favouring its max-entropy positive counterpart, the Gamma; the two laws coincide as the number of aligned columns grows and are empirically indistinguishable on $v3$ coordinates. (ii) The embedding distance D_{ij} follows a generalized extreme value (Fréchet–Gumbel) law rather than a Gaussian, because $D = \|\varphi(x) - \varphi(y)\|_2$ is dominated by its largest coordinate differences, which are extreme-*

value statistics over the K prototypes. Part (ii) is the sharp, discriminating claim; part (i) is a shape statement only.

Proof. (i) By Eq. (4), $d(a, b) = \sum_{\ell} \pi_{\ell}$ is a sum over aligned columns of bounded, non-negative Gram penalties π_{ℓ} . For sequences of comparable length the columns are many and weakly dependent, so a Lindeberg central-limit tendency drives d toward a Gaussian; the constraint $d \geq 0$ and bounded support add only a mild right-skew, for which the max-entropy positive law with matched mean and variance is the Gamma. As the number of aligned columns grows the shape parameter diverges and the Gamma converges to the Gaussian, so on real-repertoire (v3) coordinates the two fits are empirically indistinguishable (Fig. 4, left); part (i) is therefore a statement about *shape*, not a model-selection claim. (ii) The coordinates of $D(x, y) = \|\varphi(x) - \varphi(y)\|_2$ are K i.i.d. bounded differences $d(x, p_k) - d(y, p_k)$. Its ℓ^{∞} part equals $d(x, y)$ (Prop. 2), a maximum over the K prototypes; by the Fisher–Tippett–Gnedenko theorem [14] the only non-degenerate limit of a normalised maximum of i.i.d. variables is the generalized-extreme-value family, so D ’s upper tail is GEV, not Gaussian. In high dimension $\|\cdot\|_2$ concentrates and its fluctuations are carried by the same extreme coordinates [15], so the GEV law governs D itself (Fig. 4, right). $\square$

The v3 numbers bear this out (Fig. 4): the Gamma and Gaussian fits to d_{ij} are statistically indistinguishable (Kolmogorov–Smirnov 0.014 versus 0.013; $\Delta_{\text{AIC}} \approx 10$ over $\sim n^2/2$ pairs, well inside the noise), so d_{ij} is simply a unimodal, weakly right-skewed bulk. The discriminating result is the tail law of D_{ij} : the GEV fit beats the Gaussian decisively (Kolmogorov–Smirnov 0.021 versus 0.078). The fitted shape $\xi \approx 0$ places v3 at the Gumbel–Fréchet boundary, marginally lighter-tailed than the $\xi = +0.11$ of the published Smith–Waterman construction — a metric-dependent detail, not a qualitative difference. (The paper’s supplement reports Gamma $>$ Normal for d_{ij} ; on the tighter real-repertoire junction-length distribution of the arda-native prototypes the skew shrinks and the two collapse together.)

WHY NOT LEARN THE REPRESENTATION? A natural alternative to Eq. (1) is to train a self-supervised sequence model — a masked or autoregressive “receptor language model” — and use its latent as the embedding. The recombination model shows why this cannot improve on the alignment geometry for specificity. Two properties of P_{gen} are decisive: it is *low-complexity* and it is *antigen-blind*.

First, the repertoire process has little learnable structure. P_{gen} factorises as a shallow Bayesian network — germline gene choice, a nearest-neighbour (dinucleotide Markov) insertion chain, and independent trimming [2, 12] — so its *predictive information*, the mutual information between a CDR3 prefix and its remainder (the excess entropy of the process), is bounded by the description length of that network: a few bits, and short-range. Natural language carries large predictive information — syntax, semantics, world knowledge — that deep representations discover and compress; a CDR3 carries almost none beyond its germline templating and local insertion statistics. A likelihood-trained model therefore converges to a sufficient statistic for P_{gen} itself and has nothing further to represent.

Second, and decisively, P_{gen} has no antigen variable: generation precedes and is independent of antigen selection. Let A be the antigen a receptor X recognises, unobserved during self-supervised training.

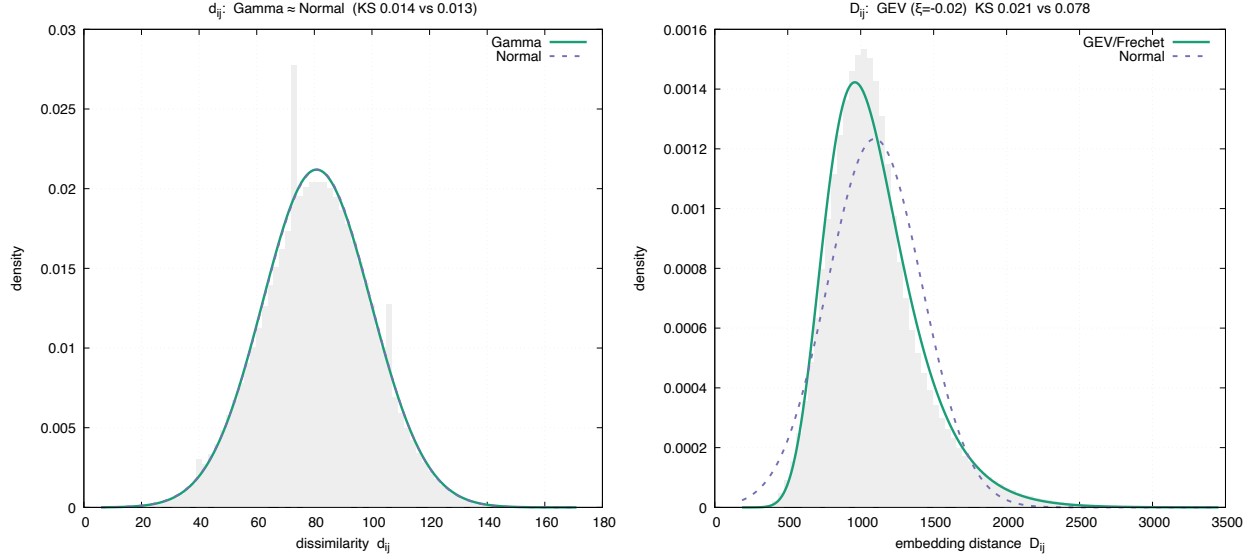


Figure 4: Distribution laws (Prop. 8; reproduces the published suppl. S1 [1]). Left: d_{ij} is a unimodal, only weakly skewed bulk — Gamma and Gaussian fit near-identically (ks 0.014 vs. 0.013; the curves overlies the histogram and each other). Right: the discriminating law — D_{ij} follows a GEV/Fréchet law (ks 0.021) that the Gaussian (0.078) misses on the heavy right tail. `mir/bench/theory.py::fit_distributions`.

PROPOSITION 9 (No unsupervised separation of specificity). *Let a representation f be fitted by any objective that is a functional of the unlabelled generation marginal P_{gen} alone. Then the training signal carries zero information about the antigen partition, $I(\text{objective}; A) = 0$; the population objective is invariant under relabelling of antigens, and the alignment of $f(X)$ with A is fixed by the inductive bias of f (its architecture, tokenisation, and metric), not by learning.*

Proof. By hypothesis the objective is a functional of the unlabelled generation marginal P_{gen} — equivalently of the i.i.d. sample $X_1, \dots, X_m \sim P_{\text{gen}}$ — and never accesses A . Because generation precedes selection, A enters the joint law only through the post-generation selection channel, which the training data do not observe; hence the training data and A are independent, $I(\text{training data}; A) = 0$. For any representation f measurable with respect to that data the data-processing inequality gives $I(f; A) \leq I(\text{training data}; A) = 0$. The population objective is therefore a functional of P_{gen} that does not reference the antigen partition, so it is invariant under any permutation ς of antigen labels: if f^* is optimal, so is $\varsigma \circ f^*$. No minimiser can prefer an antigen-separating representation over an antigen-mixing one on the strength of the data, so whatever alignment $f(X)$ has with A is fixed by the inductive bias of f , not learnt. $\square$

Specificity separation in a receptor embedding is thus an *inductive-bias* phenomenon, not a learned one: it is present exactly insofar as the chosen geometry matches the biophysics of recognition. The prototype embedding builds in precisely the right bias — the alignment metric that Fig. S5 of the published supplement [1] shows tracks structural (C α RMSD) similarity — with no training

and no antigen labels. A learned network has only three options: (i) recover P_{gen} , which does not separate antigens (Prop. 9); (ii) be supervised by scarce antigen labels, abandoning the unsupervised premise and the unlimited P_{gen} -sampling that makes the approach attractive; or (iii) regress onto the alignment geometry itself — which is exactly the neural forward codec (`mir/ml/encoder.py`), a network trained to reproduce φ . Option (iii) is not a competitor to the prototype embedding but an accelerator of it: the network learns to *compute* the distance functional quickly, not to discover a better coordinate, and its embeddings are only as good as the φ it approximates. In this precise sense the alignment-to-reference geometry is the canonical unsupervised coordinate for a repertoire whose only structure is an antigen-blind generative model, and prototypes are its minimal, assumption-free realisation.

SPECIFICITY IS HOST-CONDITIONED AND, IN THE LIMIT, COVERAGE-COMPLETE. Proposition 9 concerns unlabelled objectives; one might hope to inject the missing antigen signal from a labelled database. Three facts make even that route converge back to P_{gen} .

(1) *The label is not a function of the receptor.* What one observes as antigen-specific expansion in a donor is not a map $X \mapsto \text{antigen}$ but the host-restricted shadow $A_h(X) = B(X) \cap (\text{HLA}_h \times \text{Exposure}_h)$, where $B(X)$ is the large, cross-reactive set of pMHC that X can engage: which peptides are presented is set by the HLA haplotype, which are encountered by pathogen history. Both are host latent variables, and donors with identical infections hold sparse, *largely disjoint* receptor sets that nonetheless cover the same antigens [16]. Specificity is thus not a stable, sequence-intrinsic function: the same X carries different labels in different hosts, so the map is non-identifiable from sequence across a population.

(2) *Labels cover a vanishing slice.* Databases such as VDJDB [17] annotate a measure-zero corner of the antigen $\times$ receptor space ($\geq 20^9$ peptides per MHC class, thousands of alleles, $> 10^{15}$ possible receptors), so a supervised model saturates at database coverage and collapses on unseen epitopes — it learns the covered epitopes, not specificity.

(3) *The completeness limit is P_{gen} .* Pursuing full coverage does not escape the loop, because the recombination model is itself the immune system’s coverage code: the repertoire is (near-)optimally tuned to the pathogen distribution, exploiting cross-reactivity so that essentially every antigen has cognate precursors [16, 2].

PROPOSITION 10 (Coverage fixed point). *Assume coverage-completeness: there is $p_0 > 0$ such that for almost every antigen a , $\Pr_{x \sim P_{\text{gen}}}[x \text{ binds } a] \geq p_0$. Then the binding relation is a dense, many-to-many graph whose antigen-marginal is bounded away from 0 and 1 uniformly; averaging the specificity map over the whole antigen space carries no receptor-discriminative information beyond generation propensity, so as the labelled antigen set expands to the full space the population optimum of any supervised specificity representation converges to the antigen-blind P_{gen} -sufficient statistic of Prop. 9.*

Proof. Coverage-completeness gives $\Pr_{x \sim P_{\text{gen}}}[x \text{ binds } a] \geq p_0 > 0$ for almost every antigen a , and cross-reactivity gives the dual bound, each receptor binding a $\Theta(1)$ fraction of $\mathcal{A}$. Thus the binding indicator $b(x, a)$ has antigen-marginal $\mathbb{E}_{x \sim P_{\text{gen}}}[b(x, a)]$ bounded strictly inside $(0, 1)$ uniformly in a . For any receptor statistic $f(X)$ the specificity information integrated over antigens,

$\int_{\mathcal{A}} I(f(X); b(X, a)) da$, is then bounded by the entropy of a Bernoulli whose parameter is pinned away from 0 and 1 and whose across-receptor variation is driven by the clone-frequency structure P_{gen} imposes, not by any low-dimensional antigen axis. Hence as the labelled antigen set grows to the whole space, the population optimum of a supervised specificity representation is determined by generation propensity alone and converges to the P_{gen} -sufficient statistic of Prop. 9. $\square$

The circle closes: an unsupervised model recovers P_{gen} and cannot separate antigens (Prop. 9); a supervised model learns only its labelled slice (fact 2) of a host-conditioned, non-identifiable label (fact 1); and the ideal of learning *all* specificity returns to P_{gen} (Prop. 10), because P_{gen} is the coverage code. Sampling prototypes from P_{gen} is therefore not an arbitrary reference choice — it discretises the very coverage design that antigen breadth is built on, so distance-to-prototypes coordinates a receptor by its position in that coverage geometry, the maximal host-independent, specificity-relevant structure a sequence can carry.

T.4. TCRs, THREE-COMPONENT EMBEDDINGS, AND PAIRED CHAINS

THE PRODUCT METRIC. TCREMP emits three distances per prototype — a V-germline, a J-germline, and a junction distance (or CDR1, CDR2, junction) — interleaved into $\mathbb{R}^{3K}$ (`mir/embedding/tcremp.py`). Writing $\varphi = (\varphi_V, \varphi_J, \varphi_{\text{junc}})$ for the three blocks, the block structure of the Euclidean norm gives the exact decomposition

$$D(x, y)^2 = D_V(x, y)^2 + D_J(x, y)^2 + D_{\text{junc}}(x, y)^2, \quad (5)$$

so the embedding metric is the orthogonal sum of the per-component metrics.

PROPOSITION 11 (Paired-chain concatenation). *For a paired receptor (α, β) embedded by concatenating the per-chain maps, $D_{\text{pair}}(x, y)^2 = D_{\alpha}(x, y)^2 + D_{\beta}(x, y)^2$. More generally, concatenating landmark embeddings of independent factors realises the ℓ^2 product of the factor metrics.*

Proof. Concatenation stacks coordinate blocks; the squared ℓ^2 norm of a stacked vector is the sum of the squared block norms, which is Eq. (5) extended across chains. $\square$

Paired $\alpha\beta$ TCR embedding (validated in [1]) and, prospectively, the concatenation of epitope- and MHC-groove coordinates into a single TCR-pMHC vector are therefore principled: each modality contributes an orthogonal block and distances add in quadrature.

THE GERMLINE BLOCKS ARE LOW-RANK. The V and J coordinates depend on the sequence only through its germline gene identity, which explains the redundancy that PCA removes.

PROPOSITION 12 (Germline rank bound). *Let G_V be the number of distinct V alleles present in a dataset S . The V-block matrix $\Phi_V \in \mathbb{R}^{|S| \times K}$ with rows $\varphi_V(x)$ has at most G_V distinct rows, hence $\text{rank } \Phi_V \leq G_V$. The same holds for J with G_J , and $G_V, G_J \ll K$.*

Proof. $\varphi_V(x)_k = d_V(v(x), v(p_k))$ depends on x only through its V allele $v(x)$ (the germline distances are a fixed allele $\times$ allele table, `mir/distances/germline.py`); receptors sharing a V allele have identical V-rows, so there are at most G_V distinct rows and the rank is bounded by G_V . $\square$

A human TRB repertoire has $G_V \lesssim 60$, $G_J \lesssim 15$, against K in the thousands: two-thirds of the raw coordinates live in a subspace of dimension $\leq G_V + G_J$. The `StandardScaler`→PCA step (`mir/embedding/pca.py`) recovers this compact basis — Landmark MDS on φ — and an ANOVA of the leading components against gene usage recovers the germline axes, while the junction block carries the residual $\Theta(K)$ -dimensional specificity signal.

STRUCTURAL GROUNDING. That the metric is biologically meaningful, not merely combinatorial, is shown by Fig. S5 of the published supplement [1]: embedding distances between TCR:pMHC complexes correlate with the C α RMSD of their crystal structures. Formally this asserts a Lipschitz map from the sequence metric (X, ρ) into the structural metric, so that specificity groups [18] — receptors binding one epitope — occupy small-diameter regions of embedding space, and clustering recovers them (Table S1; `mir/bench/metrics.py`).

T.5. IGH AND SOMATIC HYPERMUTATION

B-cell receptors differ from TCRs in one decisive respect: after recombination they undergo somatic hypermutation (SHM), a context-dependent point-mutation process [19]. Write M_m for an operator applying m substitutions to a germline-rearranged sequence.

PROPOSITION 13 (Bounded embedding drift). *There is a constant c (the maximal per-substitution Gram penalty) such that the dissimilarity drifts linearly, $d(x, M_m x) \leq c m$. The embedding distance then drifts sublinearly, $D(x, M_m x) \leq 2 \text{diam} \sqrt{K d(x, M_m x)} \leq 2 \text{diam} \sqrt{K c m} = O(\sqrt{m})$, and is in every case bounded.*

Proof. Each substitution changes at most one aligned column of Eq. (4), by at most the bounded penalty range c ; m substitutions change the squared dissimilarity by at most $d(x, M_m x) \leq c m$. The Hölder- $\frac{1}{2}$ coordinates of Prop. 1 propagate this to $D \leq 2 \text{diam} \sqrt{K d(x, M_m x)}$, giving the $\sqrt{m}$ growth. (Summing single-substitution steps through the embedding-space triangle inequality gives the coarser linear envelope $D \leq 2 \text{diam} \sqrt{K c m}$; the $\sqrt{m}$ bound is tighter for large m .) $\square$

Proposition 13 has two consequences. First, a clonal lineage under affinity maturation traces a bounded path in embedding space — sublinear in the accumulated mutation load, and well-approximated by a line over the biological SHM range (T5, linear $R^2 \approx 0.97$ – 0.99) — so the embedding preserves lineage proximity — the substrate for density and trajectory analysis (§T.6). Second, it explains why IGH is the hardest chain for the neural codec (forward cosine 0.79, exact reconstruction 0.16, against $\geq 0.93/0.50$ for TCR): (i) IGH CDR3s are long, so the fixed-length-40 gap-block tokenisation (`mir/ml/tokenize.py`) truncates an appreciable fraction of the variable middle, discarding information; (ii) SHM inflates the junction distance by $\Theta(m)$, so germline-rearrangement prototypes sampled from P_{gen} systematically under-cover the mutated support (the density ratio $r_{\mu\nu}$ of Prop. 7 becomes large and the robustness guarantee degrades); and (iii) the per-receptor entropy is higher. The remedy the theory suggests is to separate the SHM-robust germline distance (V/J, Prop. 12) from the mutation-inflated junction distance and to augment the prototype set with a mutation model, i.e. to sample prototypes from P_{gen} pushed through M_m rather than from P_{gen} alone.

PART II

FROM RECEPTORS TO REPERTOIRES

Analysing many receptors at once: the density of the receptor cloud in embedding space, the embedding of a whole repertoire as a single vector, and its form as a neural-network modality.

T.6. DENSITY GEOMETRY AND BACKGROUND SUBTRACTION

The embedding turns a repertoire into a point cloud in $\mathbb{R}^{3K}$ (or its PCA projection to $p \approx 65$ coordinates), and the natural null model is the pushforward of the generation measure. Antigen-driven selection deposits, on top of this null, over-dense *ridges* of convergent receptors; TCRNET and ALICE [20, 21, 22] detect them by counting near-neighbours on a Hamming-1 graph. We recast the whole procedure as *density-ratio estimation* on embedding space — graph-free — and give it what an honest treatment requires: a model of the noise (§ T.6.1), the estimator and its rate (§ T.6.2), the exact local test (§ T.6.3–§ T.6.4), calibration of a systematically wrong null (§ T.6.5), the differential control that fixes it (§ T.6.6), the depth it requires (§ T.6.7), the ridge geometry it estimates (§ T.6.8), and how clone sizes enter (§ T.6.9).

DEFINITION 2. Let $f_{\text{gen}} = \varphi_{\#} P_{\text{gen}}$ be the density of embedded prototypes and f_{obs} the density of an observed repertoire on embedding space. The *enrichment* is the density ratio $E(z) = f_{\text{obs}}(z)/f_{\text{gen}}(z)$; convergent selection is the region $E(z) > 1$.

NOTATION. A receptor (clonotype) σ is mapped to a point $z = \varphi(\sigma) \in \mathbb{R}^p$ (the prototype embedding, standardised and PCA-projected to $p \approx 65$ coordinates). P_{gen} is the V(D)J generation measure and π_{obs} the sampled-repertoire law; their pushforwards under φ are the densities $f_{\text{gen}} = \varphi_{\#} P_{\text{gen}}$ and $f_{\text{obs}} = \varphi_{\#} \pi_{\text{obs}}$, and $E = f_{\text{obs}}/f_{\text{gen}}$ is their ratio. The selection factor $Q(\sigma) \geq 0$ reweights P_{gen} into π_{obs} (Eq. (6)), with normaliser $Z_Q = \mathbb{E}_{P_{\text{gen}}}[Q]$ and embedding-space average $q(z) = \mathbb{E}[Q \mid \varphi(\sigma) = z]$. We draw N observed and M background points; a ball $B(z, r)$ of radius r contains n_{obs} observed and n_{bg} background points, with target null occupancy $\lambda_0 = \mathbb{E}[n_{\text{obs}} \mid H_0]$; r_1 is the median embedding distance moved by a single CDR3 amino-acid substitution, and a_j (§ T.6.9) the read count of clonotype j . Symbols $\mathbb{E}$ (expectation), Var , $\text{CV} = \sqrt{\text{Var}}/\mathbb{E}$, and Φ^{-1} (standard-normal quantile) are standard.

T.6.1. THE REPERTOIRE AS A MIXTURE, AND WHY THE RAW P_{gen} NULL IS MISCALIBRATED

The observed repertoire is not the generation model. A sampled sequence has survived thymic and peripheral selection, so the observed law is a reweighting of P_{gen} [23, 24],

$$\pi_{\text{obs}}(\sigma) = P_{\text{gen}}(\sigma) \frac{Q(\sigma)}{Z_Q}, \quad Z_Q = \mathbb{E}_{P_{\text{gen}}}[Q], \quad (6)$$

with $Q \geq 0$ the (non-constant) selection factor; the scalar Q of ALICE is its mean-field reduction. Let $q(z) = \mathbb{E}[Q(\sigma) \mid \varphi(\sigma) = z]$ be the local mean selection field.

PROPOSITION 14 (Pushforward ratio is the selection field). $E(z) = q(z)/Z_Q$.

Proof. $f_{\text{obs}}(z) = \int \delta(q(\sigma) - z) P_{\text{gen}}(\sigma) Q(\sigma) / Z_Q d\sigma = q(z) f_{\text{gen}}(z) / Z_Q$. $\square$

So $E \equiv 1$ only if selection is neutral. Two immunological facts make q sharply peaked: *convergent recombination* makes public receptors recur across individuals in proportion to their generation frequency [25, 26, 27], and *thymic selection* is positively correlated with those same generation biases [23], amplifying rather than flattening the hierarchy — a rich-get-richer field concentrated on the convergent, germline-proximal support. The cost is a null that is biased *high* exactly where clones sit.

PROPOSITION 15 (Size-biased bulk fold; the flood). *Averaging the fold over observed clones,*

$$\mathbb{E}_{z \sim f_{\text{obs}}} [E(z)] = \frac{\mathbb{E}_{f_{\text{gen}}} [q^2]}{\mathbb{E}_{f_{\text{gen}}} [q]^2} = 1 + \text{CV}^2(q) \geq 1, \quad (7)$$

strict whenever $\text{Var}(q) > 0$.

Proof. $\mathbb{E}_{f_{\text{obs}}} [E] = \int (q/Z_Q) f_{\text{obs}} = \int (q/Z_Q)^2 f_{\text{gen}} = \mathbb{E}[q^2]/\mathbb{E}[q]^2 = 1 + \text{Var}(q)/\mathbb{E}[q]^2$. $\square$

A test that assumes $E \equiv 1$ therefore rejects across the entire region where q/Z_Q clears the fold threshold. This is measurable: on a VDJDB benchmark in which known antigen-specific ridges are admixed with an unrelated repertoire as noise, a raw P_{gen} background flags 43% of the noise clones as “enriched”, whereas a matched control background (§T.6.6) flags only 1% — a 46-fold gain in signal-to-noise. The 43% is the selection inflation $1 + \text{CV}^2(q)$, *not* a 43% non-null fraction. The remedy is therefore to recentre the null empirically (§T.6.5) or to cancel q with a matched control (§T.6.6). Note q/Z_Q cannot exceed 1 everywhere — both densities integrate to one, so $\mathbb{E}_{f_{\text{gen}}} [q/Z_Q] = 1$ — so it exceeds 1 only on the observed support, which is exactly where clones are tested.

REMARK 2 (Ridges are heterogeneous). The signal is a sum of epitope-indexed ridges, $f_{\text{sig}} = \sum_e w_e g_e$, with $w_e = \sum_{\sigma \in S_e} \pi_{\text{obs}}(\sigma)$ the precursor frequency of epitope e ; the elevation g_e/f_{gen} is *larger* for low- P_{gen} (private) motifs at equal observed count. Inheriting the heavy tail of P_{gen} , the weights w_e span about two orders of magnitude across epitopes and six within one [28]. The polyclonality of a response tracks this precursor frequency: on VDJDB the immunodominant HLA-A*02 influenza epitope GILGFVFTL resolves into 32 distinct convergent mountains, while the CMV A*02 epitope NLVPMVATV gives only 4. No global height threshold separates signal from background across such a range; significance must be assessed locally, and the constructions below are engineered so that it is.

T.6.2. ENRICHMENT AS A DENSITY RATIO: THREE ESTIMATORS AND THEIR RATE

Estimating E as a quotient of two kernel-density estimates is doubly bad: it pays the density-estimation rate twice and blows up where f_{gen} is small. Direct estimation avoids both, and rests on one moment identity [29, 30].

LEMMA 1 (Least-squares characterisation). $\mathbb{E}_{f_{\text{gen}}} [gE] = \mathbb{E}_{f_{\text{obs}}} [g]$ for all $g \in L^2(f_{\text{gen}})$, and $E = \arg \min_g \{ \frac{1}{2} \mathbb{E}_{f_{\text{gen}}} [g^2] - \mathbb{E}_{f_{\text{obs}}} [g] \}$.

Proof. $\int g(f_{\text{obs}}/f_{\text{gen}})f_{\text{gen}} = \int g f_{\text{obs}}$; the objective equals $\frac{1}{2}\mathbb{E}_{f_{\text{gen}}}[g-E]^2$ up to a constant, minimised uniquely at $g = E$. $\square$

Because the empirical objective uses the two clouds only through sample means, the same target admits three consistent estimators. **(i) Balloon** count-ratio (§ T.6.4), local and carrying an exact test. **(ii) Relative ratio** (RuLSIF) $E_\alpha = f_{\text{obs}}/(\alpha f_{\text{obs}} + (1-\alpha)f_{\text{gen}}) \in [0, 1/\alpha]$, fit in a kernel basis [31]; it is bounded, hence valley-stable, and inverts to $E = (1-\alpha)E_\alpha/(1-\alpha E_\alpha)$, with α the water level — a denominator denser than f_{gen} caps the runaway public-convergence fold at $1/\alpha$. **(iii) Classifier** logit: labelling observed points $y=1$ (prior $\pi = N/(N+M)$) and background $y=0$, a calibrated classifier η satisfies

$$\text{logit } \eta(z) = \log E(z) + \log \frac{\pi}{1-\pi} \quad [32, 33], \quad (8)$$

so any discriminator — featurised by the neural forward codec φ (`mir/ml/encoder.py`) — recovers $\log E$ up to the known imbalance offset and scales to 10–100M receptors; normalising flows [34] give the two densities explicitly when wanted. The shared offset $\log(N/M)$ is why the control-background and P_{gen} -background detectors estimate the *same* object (§ T.6.3).

PROPOSITION 16 (Direct estimation beats plug-in). *If E is bounded and lies in a class of bracketing entropy $H(\varepsilon) \asymp \varepsilon^{-2\gamma}$, $0 < \gamma < 1$, the regularised least-squares estimator attains $\|\hat{E} - E\|_{L^2(f_{\text{gen}})} = O_p(n^{-1/(2+2\gamma)})$, $n = \min(N, M)$, which is minimax for the likelihood ratio [35]; the plug-in quotient pays the density-estimation rate $n^{-s/(2s+d)}$ per factor plus a small-denominator variance blow-up.*

Proof. By Lemma 1 the population risk of any candidate g is $J(g) = \frac{1}{2}\mathbb{E}_{f_{\text{gen}}}[g-E]^2$ up to an additive constant, so the excess risk of the empirical minimiser $\hat{E}$ over a class $\mathcal{G}$ equals $\frac{1}{2}\|\hat{E} - E\|_{L^2(f_{\text{gen}})}^2$. Empirical-process theory bounds this by an approximation term (the regularisation bias) plus a stochastic term set by the local modulus of the empirical process indexed by $\mathcal{G}$, i.e. by the entropy integral $\int_0^\delta \sqrt{H(\varepsilon)} d\varepsilon \asymp \delta^{1-\gamma}$ when $H(\varepsilon) \asymp \varepsilon^{-2\gamma}$. Choosing the regularisation so that bias and stochastic term are of the same order and solving the resulting fixed point for the critical radius δ gives $\delta \asymp n^{-1/(2+2\gamma)}$; a matching lower bound follows from Fano’s inequality on a δ -packing of $\mathcal{G}$, so the rate is minimax [35]. Boundedness of E keeps the class envelope finite in $L^\infty(f_{\text{gen}})$, which the bound requires. The plug-in quotient $\hat{f}_{\text{obs}}/\hat{f}_{\text{gen}}$ has no such envelope: where $\hat{f}_{\text{gen}} \rightarrow 0$ the ratio’s variance diverges — exactly the instability the bounded relative ratio $E_\alpha \leq 1/\alpha$ of § T.6.2 removes by construction. $\square$

At $p \approx 65$ the nonparametric rate is slow; treat $\hat{E}$ as a locally smoothed, not pointwise-exact, field. The three estimators cross-validate: the classifier gives a fast global E , the balloon supplies calibrated significance, the relative estimator delineates smooth ridge boundaries; disagreement flags estimation error, not signal.

T.6.3. LOCAL SIGNIFICANCE: THE EXACT POINT-PROCESS TEST

Model the two clouds as independent Poisson point processes of intensities $Nf_{\text{obs}}, Mf_{\text{gen}}$ [36]. In a ball $B = B(z, r)$ the counts are Poisson with means $\int_B Nf_{\text{obs}}, \int_B Mf_{\text{gen}}$, independent across disjoint sets (Kingman restriction).

PROPOSITION 17 (Poisson exact test — ALICE). *Under $H_0 : f_{\text{obs}} = f_{\text{gen}}$, with the leave-one-out, Laplace-smoothed null mean*

$$\mu_0 = \frac{(N-1)(n_{\text{bg}}+1)}{M+1}, \quad p = \Pr[\text{Poisson}(\mu_0) \geq n_{\text{obs}}], \quad (9)$$

the enrichment p -value is exact given μ_0 ; for an empty background ball the rule-of-three bound $\mu_0 = 3N/M$ applies. This is (9) as implemented in `mir/density.py`.

Proof. Poisson tail on the restricted process; $(N-1)/(M+1)$ with $+1$ in the numerator is the rule-of-succession (Beta(1, 1)) estimate of the ball probability with self-exclusion, and $3N/M$ is the 95% upper bound on a Poisson mean from zero events. $\square$

PROPOSITION 18 (Conditional binomial — TCRNET). *For independent processes with in-ball means $\mu_{\text{obs}}, \mu_{\text{bg}}$, conditioning on the total $T = n_{\text{obs}} + n_{\text{bg}}$ gives $n_{\text{obs}} \mid T \sim \text{Binomial}(T, \mu_{\text{obs}}/(\mu_{\text{obs}} + \mu_{\text{bg}})) \xrightarrow{H_0} \text{Binomial}(T, N/(N+M))$, the local volume cancelling. This nuisance-free two-sample test [37] is the binomial branch used with a real control background.*

PROPOSITION 19 (ALICE is the rare-event limit of TCRNET). *With $p_{\text{bg}} = (n_{\text{bg}}+1)/(M+1) \sim \lambda_0/M \ll 1$, the Le Cam / Chen–Stein bound [38] gives $d_{\text{TV}}(\text{Binomial}(T, p_{\text{bg}}), \text{Poisson}(Tp_{\text{bg}})) \leq Tp_{\text{bg}}^2$, so the two exact tests agree to $O(\lambda_0 p_{\text{bg}})$.*

Proposition 19 is the theoretical reason a generated P_{gen} background (Poisson) and a real control (binomial) return the *same* enriched set — benchmarked at Jaccard 0.86 against a chance 0.30; ALICE is a special case of TCRNET, and the residual 0.14 is the convergence-field over-call the differential control removes (§T.6.6). One caveat: clonal expansion makes raw counts super-Poisson, so counting must be over *distinct* clonotypes and the null recentred (§T.6.5).

T.6.4. THE BALLOON ESTIMATOR AND THE DETECTION FLOOR

Set the radius $R(z)$ to the distance to the k -th *background* neighbour, $k = \text{round}(\lambda_0 M/N)$. The shared ball volume cancels from the ratio, which is what makes the estimator usable in high dimension [39, 40, 41].

PROPOSITION 20 (Volume-free ratio; homogenised null). *The balloon densities $\hat{f}_{\text{obs}} = n_{\text{obs}}/(NV_p R^p)$, $\hat{f}_{\text{gen}} = n_{\text{bg}}/(MV_p R^p)$ give*

$$\hat{E}(z) = \frac{n_{\text{obs}}/N}{n_{\text{bg}}/M}, \quad k = \text{round}(\lambda_0 M/N) \implies \mathbb{E}[n_{\text{obs}} \mid H_0] \approx \lambda_0 \text{ for all } z. \quad (10)$$

The null is $\text{Poisson}(\lambda_0)$ independent of the local density $f_{\text{gen}}(z)$; the leading rate is set by the intrinsic dimension $d^ \ll p$ of the receptor manifold, not by p .*

Proof. Fixing k background points fixes the enclosed background mass $\approx k/M$, so the expected observed count is $N \cdot k/M = \lambda_0$; $V_p R^p$ cancels between numerator and denominator, removing the explicit dimension. $\square$

Fixing λ_0 (rather than the radius) is exactly what Remark 2 demands: with precursor frequencies spanning orders of magnitude, a fixed radius would be empty in sparse regions and saturated in dense ones, whereas fixing the null occupancy makes every ball testable on its own scale.

PROPOSITION 21 (Minimum detectable enrichment). *Under H_1 with local fold E , $n_{\text{obs}} \sim \text{Poisson}(E\lambda_0)$, and a one-sided level- q test is cleared at 50% power once*

$$E^*(q, \lambda_0) = 1 + \frac{c(q)}{\sqrt{\lambda_0}}, \quad c(q) = \Phi^{-1}(1 - q); \quad (11)$$

under Benjamini–Hochberg over N tests $c \approx \sqrt{2 \ln N}$. Thus E^ falls only as $\lambda_0^{-1/2}$ while the radius grows as λ_0^{1/d^*} (essentially flat in high d^*) and must stay $\lesssim r_1$, the median one-substitution embedding drift, to resolve a motif ridge; balancing these gives the default $\lambda_0 = 3$ (detectable ~ 2.3 -fold) with the range $[1, 5]$.*

Proof. By Proposition 20 the null count is $n_{\text{obs}} \sim \text{Poisson}(\lambda_0)$, of mean and variance λ_0 . The one-sided level- q test rejects at $n_{\text{obs}} \geq \tau_q$ with $\tau_q = \min\{k : \Pr[\text{Poisson}(\lambda_0) \geq k] \leq q\}$, and the normal approximation gives $\tau_q \approx \lambda_0 + c(q)\sqrt{\lambda_0}$, $c(q) = \Phi^{-1}(1 - q)$. Under H_1 the count has mean $E\lambda_0$, so the test attains 50% power once $E\lambda_0 \geq \tau_q$, i.e. once $E \geq 1 + c(q)/\sqrt{\lambda_0}$ — Eq. (11). Demanding power $1 - \beta$ replaces $c(q)$ by $c(q) + \Phi^{-1}(1 - \beta)$, and under Benjamini–Hochberg the effective per-test level tightens to $\approx q/N$, whence $c \approx \Phi^{-1}(1 - q/N) \approx \sqrt{2 \ln N}$. The exact Poisson tail at $q=0.05$ has critical counts $\tau = 4, 7, 10$ for $\lambda_0 = 1, 3, 5$, giving $E^* = \tau/\lambda_0 \approx 4.0, 2.3, 2.0$. Finally a ball holding λ_0 of the M background points encloses mass $\lambda_0/M \propto R^{d^*}$ on the intrinsic-dimension- d^* manifold, so $R \propto \lambda_0^{1/d^*}$ — nearly flat for large d^* , so raising λ_0 costs almost no resolution until R reaches the motif width r_1 , which pins the operating point at $\lambda_0 = 3$. $\square$

Pinning $R \approx r_1$ makes the ball the continuous analogue of the Hamming-1 shell, and $E(z)$ is its $r \rightarrow 0$ limit, $E(z) = \lim_{r \rightarrow 0} [n_{\text{obs}}(B_r)/(N-1)]/[n_{\text{bg}}(B_r)/M]$. This explains the measured continuous–discrete concordance: the Spearman correlation between the balloon count and the discrete Hamming-1 count is $\rho = 0.37, 0.34, 0.11, -0.05$ at $r = 0.5, 1, 2, 3 \times r_1$ — high at one substitution (where the ball coincides with the graph neighbourhood), decaying past it as the ball integrates several mode basins. The continuous test *generalises* the graph test rather than reproducing it.

T.6.5. EMPIRICAL-NULL CALIBRATION: THE WATER LEVEL

By Proposition 15 the theoretical P_{gen} null is wrong by a global factor; before any multiplicity control it must be recentred to the data. This is Efron’s empirical null [42], and it precedes FDR.

PROPOSITION 22 (Enrichment is an inverse local FDR). *In the two-groups model $f = \pi_0 f_0 + \pi_1 f_1$ with a calibrated null f_0 , the local false-discovery rate is $\text{fdr}(z) = \pi_0 f_0(z)/f(z) = \pi_0/E(z)$; declaring $E(z) \geq 1/\alpha$ is a conservative $\text{fdr} \leq \alpha$ rule [42]. With the raw null $f_0 = f_{\text{gen}}$ one instead has $\text{fdr} = \pi_0 s/E$ — off by the selection inflation $s = 1 + \text{CV}^2(q)$ of (7).*

PROPOSITION 23 (Water level = one-sided empirical-null recentring). *Let the true background be $f_{\text{bg}}(z) = s f_{\text{gen}}(z)$ on the tested support with unknown inflation $s > 1$. If the ridge fraction is below $\frac{1}{2}$, then*

$$c = \max\left(\frac{\text{median}_i n_{\text{obs},i}}{\text{median}_i \mu_{0,i}}, 1\right), \quad \mu_0 \leftarrow c \mu_0 \implies \text{median}_i \text{fold}_i \approx 1, \quad (12)$$

and c is a robust (breakdown $\frac{1}{2}$) consistent estimate of s .

Proof. Write $\log \hat{E} = \log s + \varepsilon$ with $\varepsilon \approx 0$ on the bulk and > 0 on ridges; if ridges occupy under half the tests, $\text{median}(\log \hat{E}) = \log s$, so (12) estimates s and $\max(\cdot, 1)$ enforces one-sidedness (never deflate below the P_{gen} floor). This is Efron’s empirical-null location correction [42], equivalently the median-of-ratios size factor of DESeq [43]. Being a median it is bounded above by the mean $1 + \text{CV}^2(q)$ of (7). $\square$

This is the `calibrate="median"` step of `mir/density.py`: because the signal is $< 5\%$ in a naive repertoire it barely moves the median, so recentring the bulk to $\text{fold} \approx 1$ is robust. After recentring, overlapping balls are positively associated, so Benjamini–Hochberg controls the FDR under positive-regression dependence [44, 45]. The failure mode is sharp: a single scalar c removes only a *constant* inflation. A spatially varying selection field $s(z)$ — or a tetramer-sorted sample with ridge fraction $\geq \frac{1}{2}$ — breaks the median, and the only fix is a real control.

T.6.6. THE DIFFERENTIAL CONTROL: CANCELLING CONVERGENCE POINTWISE

PROPOSITION 24 (Shared-field cancellation). *Let case and control share the selection/convergence field, $f_{\text{case}} = (q/Z_Q)f_{\text{gen}}(1 + a_1)$ and $f_{\text{ctrl}} = (q/Z_Q)f_{\text{gen}}(1 + a_0)$, with $a_i \geq 0$ the condition-specific antigen expansion. Then*

$$R(z) = \frac{f_{\text{case}}(z)}{f_{\text{ctrl}}(z)} = \frac{1 + a_1(z)}{1 + a_0(z)} \perp q, Z_Q, f_{\text{gen}}, \quad (13)$$

independent of the shared field; under H_0 ($a_1 \equiv a_0$) $R \equiv 1$ with no water-level offset, even where $q(z)$ varies spatially. The binomial test of Proposition 18 is its finite-sample plug-in; ALICE is the special case control $\equiv f_{\text{gen}}$ after recentring.

Proof. Direct division cancels the shared factor $(q/Z_Q)f_{\text{gen}}$; in the small-ball limit the case/control odds converge to $(N_{\text{case}}/N_{\text{ctrl}})R(z)$, a Radon–Nikodym change of reference from the generation model to the control. $\square$

PROPOSITION 25 (Control-absent motifs are maximally identifiable). *If the control lacks a motif on a region A ($f_{\text{ctrl}} \rightarrow 0, f_{\text{sig}} > 0$) then $R \rightarrow \infty$ on A (odds-ratio $\rightarrow$ relative-risk regime [46]); under the pure-control anchor $\text{ess inf}_z (f_{\text{sig}}/f_{\text{ctrl}}) = 0$ the mixing weight is identified as $1 - \pi = \text{ess inf}_z R(z)$ [47].*

This is why the recommended course of action prefers a biological control (pre- vs post-vaccination, patient vs healthy, B27 $\pm$) over the raw P_{gen} null, and it is what the benchmarks show. The public CASSVGL[YF]STDTQYF (TRBV9/TRBJ2-3) motif is control-absent in ankylosing spondylitis, so $R \rightarrow \infty$ on its ridge: it surfaces among enriched B27+ synovial hits (9) and is absent

in B27— (0). In yellow-fever vaccination the shared naive field cancels, isolating the expansion: day-15 versus day-0 enriched hits matching the LLW/A*02 reference are $63 > 35$ (of 129 vs 79 distinct such clonotypes present). Cancellation requires case and control to share q — match HLA, depth, and cell subset; a contaminated control attenuates R toward 1, making the differential test conservative, not anti-conservative.

T.6.7. SAMPLING DEPTH: WHY THE FULL REPERTOIRE

Subsampling a repertoire with retention probability t is Poisson thinning: counts scale to $\text{Poisson}(t\mu)$ [36].

PROPOSITION 26 (Thinning cost and detection boundary). *With signal mass Λ_e and null mean λ_0 , the test’s non-centrality $\delta = \Lambda_e/\sqrt{\lambda_0}$ thins as $\delta_t = \sqrt{t}\delta$. Writing depth $D = tN$ and $\lambda_0 \propto Dw_e$, the minimum detectable precursor frequency is*

$$w_e^{\min}(D) = \frac{c(\alpha, \beta, \lambda_0)}{D}, \quad (14)$$

and a clone is even observed only with probability $1 - e^{-Dw_e}$ [48].

Proof. Poisson thinning with retention probability t maps the intensity of every region from μ to $t\mu$ [36]. At depth $D = tN$ a ridge of precursor frequency w_e contributes an expected in-ball signal count $\propto Dw_e$ against a null of mean λ_0 and standard deviation $\sqrt{\lambda_0}$, so by the Gaussian approximation of Proposition 21 its power is $\Phi(Dw_e/\sqrt{\lambda_0} - \Phi^{-1}(1-\alpha))$, reaching $1-\beta$ exactly when $Dw_e \geq (\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta))\sqrt{\lambda_0}$; that is $w_e \geq w_e^{\min}(D) = c/D$ with $c = (\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta))\sqrt{\lambda_0}$, which is Eq. (14). Independently, the clone is *present* at all only if its count $\text{Poisson}(Dw_e)$ is nonzero, an event of probability $1 - e^{-Dw_e}$ [48]. $\square$

Antigen clusters are rare in *absolute* number, so subsampling by s inflates the detection floor by $1/s$ and drops sparse ridges below it — and at reduced depth low- w_e clusters are often not sampled at all. Hence the full repertoire must be processed; save memory with junction-only embedding or a PCA-fit cap, never by subsampling. Because w_e is heavy-tailed (Remark 2), the boundary $Dw_e = c$ is crossed at a different depth for every epitope — one more reason significance is local, never a global height.

T.6.8. RIDGE DELINEATION AND PER-RIDGE TESTS

The antigen-specific groups are the high components of E : its *density ridges* [49] and upper-level-set components [50, 51].

PROPOSITION 27 (Ridge extraction as level-set estimation). *The connected components of $\{z : E(z) > \tau\}$ form the density cluster tree; distinct epitope motifs are distinct components appearing at different levels, so no single global τ certifies all of them, but each is certified locally. The FDR-significant hit set $\{q < \alpha, \text{fold} > 1\}$ estimates one level set, and DBSCAN with $\varepsilon = r_1$, $\text{minPts} \in \{2, 3\}$ links its points into motifs — its core-point rule is itself a superlevel-set threshold [52, 51], so clustering and the Poisson test stay mutually consistent; $\varepsilon \gg r_1$ over-merges, the same regime where the continuous-discrete concordance collapses.*

Each motif is then labelled: a two-sample Poisson test versus a matched control for a condition, and, for an epitope, the binomial test of [22] — successes = receptors in the ridge matching the epitope, trials = ridge size, probability = the repertoire-wide annotation rate — with Benjamini–Hochberg across ridges. Because each ridge is judged against its own local background, the height heterogeneity of Remark 2 is handled automatically; background subtraction removes the smooth selection field and the surviving convergent groups are the antigen-specific clusters.

T.6.9. CLONAL ABUNDANCE: SCORING WITH CLONE SIZES, ROBUSTLY

The tests so far count *distinct* clonotypes, discarding each clone’s read count a_j . This is deliberate — read counts are heavy-tailed (roughly Zipf or log-normal), so a single hyperexpanded clone would dominate a raw read-weighted count — but it throws away real signal. Two independent signatures mark an antigen-driven response: *breadth*, a group of similar sequences produced by convergent diversification (measured by the neighbour count n_{obs}), and *depth*, a clonotype present at high copy number through clonal expansion (measured by a_j). A textbook antigen-specific cluster shows both, but the corners matter: a convergent cluster of singletons has breadth without depth, and a hyperexpanded *orphan* — large a_j , few or no neighbours, the case [22] recommends inspecting individually — has depth without breadth. A complete detector reads both channels, and uses size to boost the score without letting its tail rule the outcome.

A VARIANCE-STABILISED WEIGHTED COUNT. Replace the in-ball count by a weighted mass

$$S(z) = \sum_{j \in B(z,r)} g(a_j), \quad (15)$$

with g non-decreasing and *concave* — for instance $g(a) = \log(1 + a)$ or the Anscombe root $g(a) = \sqrt{a + \frac{3}{8}}$. Concavity is precisely robustness: the leverage of clone j on S is the marginal $g'(a_j)$, which *decreases* with size, so a size-1000 clone amid size-1 neighbours contributes a bounded increment, not a thousandfold one. The endpoints are the read count $g(a) = a$ (constant leverage, Zipf-dominated) and the shipped distinct count $g(a) = 1$ (zero leverage, no size information); g is a tunable sensitivity–robustness dial between them.

PROPOSITION 28 (Compound-Poisson null for the weighted count). *Under H_0 the in-ball clones are a $\text{Poisson}(\lambda_0)$ -sized sample with independent abundances $A \sim \nu$ (the null size law), so*

$$\mathbb{E}[S \mid H_0] = \lambda_0 \mathbb{E}_\nu[g(A)], \quad \text{Var}[S \mid H_0] = \lambda_0 \mathbb{E}_\nu[g(A)^2], \quad (16)$$

with dispersion index $\text{Var}/\mathbb{E} = \mathbb{E}_\nu[g(A)^2]/\mathbb{E}_\nu[g(A)]$, equal to 1 for $g \equiv 1$ (a Poisson null) and $\gg 1$ for $g(a) = a$ under a heavy-tailed ν . A negative-binomial (Gamma–Poisson) tail matched to these two moments gives a calibrated one-sided p -value; a variance-stabilising g keeps the null close to Poisson, so the exact test of §T.6.3 still applies.

Proof. Write $S = \sum_{i=1}^K g(A_i)$ with $K \sim \text{Poisson}(\lambda_0)$ and $A_i \sim \nu$ i.i.d. and independent of K . Wald’s identity gives $\mathbb{E}[S] = \mathbb{E}[K] \mathbb{E}[g(A)]$, and the law of total variance gives $\text{Var}[S] = \mathbb{E}[K] \text{Var}[g(A)] + \text{Var}[K] \mathbb{E}[g(A)]^2 = \lambda_0 (\text{Var}[g(A)] + \mathbb{E}[g(A)]^2) = \lambda_0 \mathbb{E}[g(A)^2]$, using $\mathbb{E}[K] = \text{Var}[K] = \lambda_0$. The dispersion index follows, and for $g \equiv 1$ it is 1, recovering the $\text{Poisson}(\lambda_0)$ null of Proposition 20. $\square$

THE ORPHAN CHANNEL. A hyperexpanded clone with few neighbours is invisible to S yet significant on its own. Under neutral sampling at depth D its size is $\text{Poisson}(Df_j)$ with f_j its frequency, so its abundance p -value against the null size law,

$$p_j^{\text{size}} = \Pr_{A \sim \nu} [A \geq a_j], \quad (17)$$

tests expansion independently of any neighbourhood. Combining the breadth and depth channels by Fisher’s rule, $-2(\log p_j^{\text{nbr}} + \log p_j^{\text{size}}) \sim \chi_4^2$ under the (approximately independent) null, recovers the orphan when $p^{\text{nbr}} \approx 1$ but p^{size} is small, and multiplicatively sharpens a cluster that is both convergent and expanded.

REMARK 3 (Rarefaction sets the size law). The null size law ν and the breadth/depth balance are read from the frequency-of-frequencies spectrum — the counts $f_1, f_2, \dots$ of singletons, doubletons, and larger clones [48]. The Good–Turing mass on unseen clones is $\approx f_1/D$, and the singleton fraction is a saturation gauge: a repertoire heavy in singletons is undersampled, so breadth carries the signal, whereas a saturated repertoire lets depth dominate. Singletons and doubletons are the convergence substrate; the 3+ and hyperexpanded classes are where the weight g and the orphan test earn their keep.

This abundance layer leaves the geometry untouched: S replaces n_{obs} in the same balloon, the same recentring (§ T.6.5) and FDR apply, and the orphan test is a per-clonotype side-channel. Clone sizes then boost the score exactly where expansion co-occurs with convergence, while the concavity of g keeps the Zipf tail from ruling the result. The shipped `mir/density.py` uses $g \equiv 1$; the weighted and orphan channels are the abundance-aware extension.

T.6.10. COURSE OF ACTION

The theory fixes every parameter, giving a graph-free reincarnation of TCRNET/ALICE (`mir/density.py`):

1. **Background.** With a matched control, use it and the binomial test (Prop. 24); otherwise generate $M \geq 5N$ samples from P_{gen} and use the Poisson test. Prefer a control whenever one exists.
2. **One basis.** Embed both clouds through one prototype set and one PCA rotation (Lemma 1); $E(z)$ is meaningless across mismatched bases.
3. **Adaptive ball.** Per observed z , radius = distance to the k -th background neighbour, $k = \text{round}(\lambda_0 M/N)$, $\lambda_0 = 3$ (Prop. 20, 21); count $n_{\text{obs}}, n_{\text{bg}}$; form μ_0 by (9).
4. **Recentre** the P_{gen} null by (12) (Prop. 23); a differential control needs none.
5. **Test** the exact upper tail (Poisson or binomial), Benjamini–Hochberg $q < 0.05$, fold > 1 ; the detection floor is $E^* \approx 1 + c(q)/\sqrt{\lambda_0}$.
6. **(Optional) clone sizes.** Swap n_{obs} for the variance-stabilised weighted count $S(z) = \sum g(a_j)$ and add the per-clonotype orphan test (§ T.6.9, Prop. 28) to score expansion alongside convergence.

7. **Cluster** hits by DBSCAN, $\varepsilon = r_1$ (Prop. 27), and test each motif for label enrichment. Process the full repertoire (Prop. 26).

Every default traces to a theorem rather than to tuning; the same estimand $E = f_{\text{obs}}/f_{\text{gen}}$ underlies the graph tests it replaces, the neural classifier, and the smooth relative-ratio field.

T.7. SAMPLE-LEVEL (REPertoire) EMBEDDING

A single receptor is a point $z = \varphi(\sigma)$; a whole *repertoire* is a cloud of such points, weighted by clone size. Section T.6 asked where that cloud is locally over-dense; here we build the complementary object — one fixed vector $\Phi(S)$ summarising the entire cloud, so repertoires can be compared, regressed against phenotype, and clustered. Four requirements pull against one another: **(a)** invariant to clonotype order, **(b)** robust to sequencing depth, **(c)** faithful to diversity (richness, Shannon, unobserved species), and **(d)** able to carry the rare clonotype *interactions* that are HLA-linked. The weight on (b) is set by the intended data: low-coverage bulk RNA-seq, which recovers all chains but only 10^2 – 10^4 clonotypes per chain, yet exists for hundreds of thousands of clinically annotated samples. That $\sim 10^3$ receptors already fingerprint an individual and separate identical twins [53, 54], that repertoire diversity falls with age [55, 56], and that the CMV response is HLA-restricted [57] are the empirical anchors the theory must reproduce. We represent a sample by its empirical *measure* on embedding space (§T.7.1), whose kernel mean is the depth-robust backbone (§T.7.2); we show why depth-robustness and diversity are antagonistic and how coverage standardisation reconciles them (§T.7.3), that similarity-aware diversity is the squared norm of that same mean (§T.7.4), that interactions live in its second moment (§T.7.5), how HLA conditioning enters (§T.7.6), that the whole construction is the codebook-free limit of the graph–cluster–histogram heuristic (§T.7.7), how the two technical nuisances — sequencing *depth* and *batch* — are decoupled from biological signal (§T.7.8) and normalised across heterogeneous RNA-seq kits, read lengths, and the tissue depth–infiltration confound (§T.7.9), how the whole compresses to a model-plus-signal code (§T.7.10), and how that signal traces back to the exact clonotypes and motifs driving a phenotype (§T.7.11).

DEFINITION 3 (The sample measure). A sample (repertoire) is a multiset $S = \{(\sigma, a_\sigma)\}$ of clonotypes σ with read counts $a_\sigma \geq 1$, depth $N = \sum_\sigma a_\sigma$ and frequencies $f_\sigma = a_\sigma/N$. With the clonotype map $z_\sigma = \varphi(\sigma) \in \mathbb{R}^p$ of §T.1 it induces the weighted empirical measure

$$\rho_S = \sum_{\sigma \in S} w_\sigma \delta_{z_\sigma}, \quad w_\sigma = \frac{g(a_\sigma)}{\sum_\tau g(a_\tau)}, \quad (18)$$

with g the non-decreasing concave weight of §T.6.9 (Eq. (15); $g(a) = \log(1 + a)$, the Anscombe root [58], $g \equiv 1 = \text{presence}$, or $g(a) = a = \text{reads}$); concavity tames the heavy-tailed (Zipf / power-law) clone-size distribution [59, 60] so a single hyperexpanded clone cannot dominate Φ . A *sample-level embedding* is any fixed vector $\Phi(S)$ that is a functional of ρ_S alone; sample dissimilarity is the maximum mean discrepancy (MMD) between two such measures.

NOTATION. S is a sample, σ a clonotype, a_σ its read count, f_σ its frequency, N the depth, $z_\sigma = \varphi(\sigma)$ its embedding point (§T.1), and ρ_S the sample measure of Eq. (18) with weights w_σ (summing

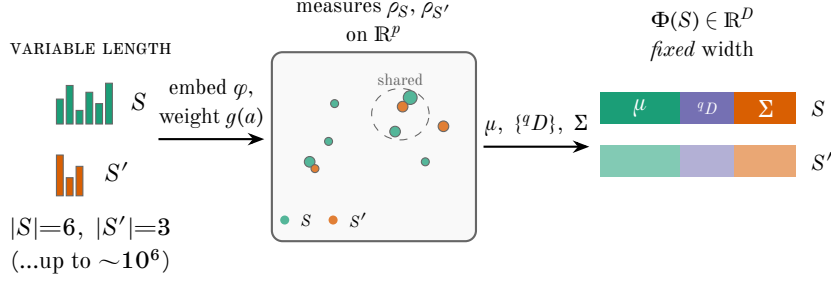


Figure 5: The sample-level embedding turns a *variable-length* repertoire into a *fixed-width* vector. Two repertoires of different clonotype counts ($|S|=6$, $|S'|=3$; real values span 10^2 – 10^6) become two size-weighted measures $\rho_S, \rho_{S'}$ on the *same* embedding space — overlapping where the repertoires share clonotypes, their distance being the MMD — and are then summarised by the three blocks (mean μ , diversity profile $\{^q D\}$, second moment Σ) into vectors of identical dimension D . The list length is never a coordinate; it survives only as the richness $^0 D$ inside the diversity block.

to 1). k is a bounded normalised characteristic kernel on $\mathbb{R}^p$ ($k(z, z) = 1$; Gaussian unless stated), $\psi : \mathbb{R}^p \rightarrow \mathbb{R}^D$ its random-Fourier feature map (Eq. (19)), and $\mu_\rho = \int \psi d\rho = \sum_\sigma w_\sigma \psi(z_\sigma)$ the kernel mean embedding. $^q D = (\sum_\sigma f_\sigma^q)^{1/(1-q)}$ is the Hill number of order q (with $^1 D = \exp(-\sum_\sigma f_\sigma \ln f_\sigma)$), $S_{\text{obs}} = ^0 D$ the observed richness, f_1, f_2 the singleton/doubleton counts, $\hat{C}$ the Good–Turing sample coverage, and n_{eff} the effective sample size (Eq. (21)). P is the individual’s true receptor law, of which S is a depth- N sample; $\varphi, P_{\text{gen}}, r_1$ are as in § T.6.

T.7.1. THE REPERTOIRE AS A MEASURE

PROPOSITION 29 (Order-invariance and sufficiency of the measure). *A statistic $\Phi(S)$ is invariant to every reordering of the clonotype list if and only if it factors through the measure ρ_S ; moreover every continuous permutation-invariant Φ admits the sum-decomposition $\Phi(S) = \varrho(\sum_\sigma w_\sigma h(z_\sigma))$ for some maps h, ϱ .*

Proof. ρ_S is symmetric in the list order by construction, so any $\Phi = \Psi(\rho_S)$ is invariant. Conversely ρ_S records exactly the distinct points and their weights, so it is a sufficient statistic for the unordered multiset and an order-invariant Φ can depend on S only through it. The sum-decomposition is the Deep Sets representation of permutation-invariant functions [61]: pooling a per-element map by a weighted sum is universal, and Eq. (18) is exactly such a pool. $\square$

This fixes the shape of every candidate embedding — a per-clonotype feature $h(z_\sigma)$, size-weighted-summed, then read out. The only design choice is h ; the next result gives the canonical one.

WHY VARIABLE LENGTH IS NOT A PROBLEM. Repertoires have wildly different sizes — from $\sim 10^2$ clonotypes in a shallow RNA-seq sample to $\sim 10^6$ in deep amplicon sequencing — so the raw inputs live in $\bigsqcup_{n \geq 1} (\mathbb{R}^p)^n / \mathfrak{S}_n$, lists of *any* length n taken up to reordering, on which no fixed-dimensional map is even defined. Passing to the measure ρ_S (Fig. 5) is exactly the quotient that removes *both* order (Prop. 29) and length: ρ_S is one point of the fixed space $\mathcal{P}(\mathbb{R}^p)$ of probability measures whatever n is,

and $\mu_{\rho_S} \in \mathbb{R}^D$ inherits that fixed dimension. Cardinality is not thrown away — it re-enters, cleanly, as the richness $^0D = S_{\text{obs}}$ inside the diversity block, so the embedding *sees* how many clonotypes there are without letting that count set its dimension. Two repertoires differing only in depth (one a subsample of the other) land at nearby vectors (Prop. 30); two differing in biology land far apart.

T.7.2. THE KERNEL MEAN EMBEDDING IS THE DEPTH-ROBUST BACKBONE

Take $h = \psi$, the random-Fourier map of k : with ω_j drawn from the spectral density of k and $b_j \sim \text{Unif}[0, 2\pi]$,

$$\psi(z) = \sqrt{2/D} (\cos(\omega_1^\top z + b_1), \dots, \cos(\omega_D^\top z + b_D)), \quad \mathbb{E}[\psi(z)^\top \psi(z')] = k(z, z') \quad [62]. \quad (19)$$

The backbone is the size-weighted mean feature — the (random-feature) kernel mean embedding of ρ_S [63, 64],

$$\Phi_1(S) = \sum_{\sigma} w_{\sigma} \psi(z_{\sigma}) = \mu_{\rho_S}, \quad \|\Phi_1(S) - \Phi_1(S')\| \approx \text{MMD}_k(\rho_S, \rho_{S'}). \quad (20)$$

PROPOSITION 30 (Consistency and depth-robustness). *Let S be a depth- N sample from a population law P with mean embedding $\mu_P = \mathbb{E}_{\sigma \sim P} \psi(z_{\sigma})$. Then $\Phi_1(S)$ is a consistent estimator of μ_P , and*

$$\mathbb{E} \|\Phi_1(S) - \mu_P\| \leq \frac{1}{\sqrt{n_{\text{eff}}}}, \quad n_{\text{eff}} = \left(\sum_{\sigma} w_{\sigma}^2 \right)^{-1}, \quad (21)$$

with high-probability deviations of the same order. Because the weights are frequencies, Φ_1 is invariant to the sequencing scale N up to this $O(n_{\text{eff}}^{-1/2})$ fluctuation.

Proof. The canonical feature $k(z, \cdot)$ has unit norm, $\|k(z, \cdot)\|_{\mathcal{H}} = \sqrt{k(z, z)} = 1$ (normalised kernel), and $\Phi_1 = \sum_{\sigma} w_{\sigma} \psi(z_{\sigma})$ is a convex combination, so $\mathbb{E} \Phi_1 \rightarrow \mu_P$ as the sampled clonotypes fill P (Glivenko–Cantelli for the mean map, as in Prop. 6). Viewing Φ_1 as a weighted average of independent unit-norm vectors and applying the bounded-difference (McDiarmid) inequality to $\|\Phi_1 - \mu_P\|$, replacing one clonotype moves Φ_1 by at most $2w_{\sigma}$, so $\mathbb{E} \|\Phi_1 - \mu_P\|^2 \leq \sum_{\sigma} w_{\sigma}^2 = 1/n_{\text{eff}}$; Jensen gives the root bound and McDiarmid the concentration. $n_{\text{eff}} = (\sum_{\sigma} w_{\sigma}^2)^{-1}$ is Kish’s effective sample size, and the MMD reading of the distance is [65]; RFF realises it in D fixed coordinates (Eq. (19)). $\square$

Equation (21) is the formal content of “depth-robustness”: the backbone converges at rate $n_{\text{eff}}^{-1/2}$, and n_{eff} is set not by the raw depth N but by how evenly the sample’s mass is spread — a diversity number (next subsection). Reading a phenotype (age, CMV) off Φ_1 is a *distribution regression*, consistent at a rate governed by n_{eff} and the kernel [66, 64].

REMARK 4 (The backbone is an empirical characteristic function). For a reader from probability, Φ_1 is a familiar object in disguise. The kernel is shift-invariant, so by *Bochner’s theorem* $k(z, z') = \int_{\mathbb{R}^p} e^{i\omega^\top(z-z')} d\Lambda(\omega)$ for a spectral measure Λ , and the random-Fourier map of Eq. (19) simply samples the complex exponential $z \mapsto e^{i\omega^\top z}$ at frequencies $\omega_j \sim \Lambda$ (the $\cos(\omega^\top z + b)$ form is the real-inner-product bookkeeping). Each coordinate of the pooled backbone is therefore

$$\sum_{\sigma} w_{\sigma} e^{i\omega_j^\top z_{\sigma}} = \int e^{i\omega_j^\top z} d\rho_S(z) = \widehat{\rho_S}(\omega_j), \quad (22)$$

the *empirical characteristic function* of the clonotype cloud at a random frequency: Φ_1 is a finite sketch of ρ_S 's characteristic function. This is the moment-generating function's bounded cousin — the characteristic function is the MGF evaluated on the imaginary axis, $\hat{\rho}(\omega) = \mathbb{E} e^{i\omega^\top z}$ against the MGF's $\mathbb{E} e^{t^\top z}$ — and the swap is exactly what a repertoire summary needs. Clone sizes are heavy-tailed (Zipf, §T.6.9), so an MGF-style $e^{t^\top z}$ feature would be dominated by the single largest clone and need not even converge; the characteristic function is bounded ($|e^{i\omega^\top z}| = 1$), always exists, is robust to that hyperexpanded clone, and is a *complete* invariant (Lévy's uniqueness theorem). This is precisely why the kernel must be *characteristic* (Notation): the term is inherited from the characteristic function, and the mean map μ_{ρ_S} is injective for the same reason $\widehat{\rho_S}$ determines its law. Two results below are then classical Fourier statistics: the sample distance is a spectrally-weighted L^2 distance between characteristic functions, $\text{MMD}_k^2(\rho_S, \rho_{S'}) = \int |\widehat{\rho_S}(\omega) - \widehat{\rho_{S'}}(\omega)|^2 d\Lambda(\omega)$ [64, 65], and the diversity block $\|\mu_{\rho_S}\|^2 = \int |\widehat{\rho_S}(\omega)|^2 d\Lambda(\omega)$ is the *energy* of that same characteristic function (§T.7.4, Prop. 32). The depth-robustness of Prop. 30 is then just the convergence rate of an empirical characteristic function.

REMARK 5 (Use the unbiased MMD across depths). The plug-in squared distance $\|\Phi_1(S) - \Phi_1(S')\|^2$ is the *biased* MMD V -statistic: it carries a self-term $\sum_\sigma w_\sigma^2 k(z_\sigma, z_\sigma) = \sum_\sigma w_\sigma^2 = 1/n_{\text{eff}}$ (the kernel diagonal, $k(z, z) = 1$). This term is not a constant offset — it *varies* with the sample's effective size, and in a cohort where diversity tracks the phenotype (older donors have lower n_{eff}) it *correlates with the label*, manufacturing a spurious divergence. The remedy is the diagonal-removed *unbiased* estimator [65], $\langle \mu, \mu \rangle_u = (\|\mu\|^2 - \sum_\sigma w_\sigma^2)/(1 - \sum_\sigma w_\sigma^2)$; the biased form is admissible only when n_{eff} is matched across samples (a fixed downsample). The artefact is real: the diversity–divergence coupling in the aging cohort drops from -0.15 (biased) to -0.05 (unbiased) once the diagonal is removed (`benchmark_repertoire_agediverge.py`).

REMARK 6 (Aging divergence is a re-expression of the clonality decline, not a new axis). Repertoires *do* grow more mutually dissimilar with age: an exact-overlap divergence — $\log F$ against age has $\rho \approx 0.70$ at 500k reads (invisible at 40k, so it is a *deep*-repertoire signal living in private clonal expansions). But the part independent of scalar diversity is null — partial $\rho(\text{age, divergence} \mid {}^1D) \approx 0.07$ (n.s.) across 40k/250k/500k — so the embedding adds no directional aging axis beyond the Hill profile of §T.7.3, consistent with the thesis *diversity for how-even*. This also cautions against reading a single overlap number: the sign of the diversity–divergence relation is metric-dependent, ≈ 0 for the unbiased kernel mean (a diverse repertoire sits near the shared P_{gen} baseline) but -0.68 for frequency-weighted overlap (clonal expansion destroys F -overlap). “More diverse $\Rightarrow$ less overlap” holds only for unweighted-richness overlap; abundance weighting flips its age-sign (`benchmark_repertoire_agediverge.py`).

T.7.3. DEPTH AND DIVERSITY ARE ANTAGONISTIC; COVERAGE STANDARDISATION RECONCILES THEM

The mean Φ_1 buys depth-robustness by averaging, which erases the rare tail — exactly where diversity lives. The tension is a theorem, not an accident.

PROPOSITION 31 (The antagonism and its resolution). (i) *The effective sample size of the backbone is a Hill diversity number: under read weighting $n_{\text{eff}} = {}^2D$ (inverse Simpson), under presence weighting*

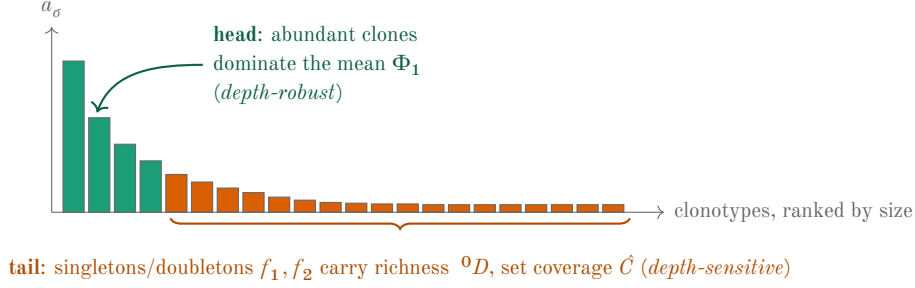


Figure 6: The depth $\leftrightarrow$ diversity antagonism, on the (Zipf/power-law [59]) clone-size curve. The frequency-weighted mean Φ_1 is governed by the abundant *head* and barely moves under subsampling (depth-robust, Prop. 30); diversity lives in the long *tail* of rare clones, whose singleton count f_1 is the most depth-sensitive quantity in the sample. Coverage standardisation (Prop. 31) compares samples at equal completeness $\hat{C}$ rather than equal depth, letting the tail be read without the depth confound.

$n_{\text{eff}} = {}^0D$ (richness), with a concave g interpolating, so $n_{\text{eff}} \in [{}^2D, {}^0D]$. (ii) Richness is the most depth-sensitive functional of S : a depth- N sample undercounts asymptotic richness by the Good–Turing missing mass, $\mathbb{E}[S_{\text{obs}}] = {}^0D_{\infty} - \Theta(f_1/N)$ [48, 67, 68], so it moves under subsampling precisely where Φ_1 is stable. (iii) Comparing Hill numbers at a common coverage $\hat{C}$ rather than a common depth N removes the confound: the coverage-standardised profile $\{{}^0D, {}^1D, {}^2D\}_{|\hat{C}^*}$ is depth-comparable across samples [69, 70, 71, 72].

Proof. (i) With $w_{\sigma} = f_{\sigma}$, $\sum_{\sigma} w_{\sigma}^2 = \sum_{\sigma} f_{\sigma}^2 = 1/{}^2D$, so $n_{\text{eff}} = {}^2D$; with $w_{\sigma} \equiv 1/S_{\text{obs}}$ (presence), $\sum_{\sigma} w_{\sigma}^2 = 1/S_{\text{obs}}$ and $n_{\text{eff}} = S_{\text{obs}} = {}^0D$. A concave g yields weights between these, hence $n_{\text{eff}} \in [{}^2D, {}^0D]$. (ii) The expected number of species seen at depth N falls short of the asymptotic richness by the mass on unseen species, which Good–Turing estimates as $\approx f_1/N$: $\Theta(1/N)$, carried by singletons, the noisiest counts, whereas the frequency-weighted mean is $O(f_1)$ -insensitive. (iii) Sample coverage $\hat{C} = 1 - (f_1/N)(N-1)f_1/[(N-1)f_1 + 2f_2]$ estimates the observed population mass; rarefying or extrapolating every sample to a common $\hat{C}^*$ equalises completeness, and Hill numbers evaluated there are independent of the raw depths [71]. $\square$

The diversity block is therefore $\Phi_2(S) = \{{}^0D, {}^1D, {}^2D\}$ at a shared coverage, with 1D tracking the age-related decline [55]. It imports ecology’s *solved* depth-robustness for exactly the information the mean discards.

T.7.4. SIMILARITY-AWARE DIVERSITY IS THE SQUARED NORM OF THE MEAN MAP

Blocks 1 and 2 are not independent: up to the kernel, the diversity block already sits inside the backbone.

PROPOSITION 32 (The mean-map norm is an order-2 similarity diversity). $\|\mu_{\rho_S}\|^2 = \sum_{\sigma, \tau} w_{\sigma} w_{\tau} k(z_{\sigma}, z_{\tau})$ is the Rao quadratic entropy of ρ_S under similarity k — the order-2 similarity-sensitive diversity of Leinster and Cobbold [73]. As the bandwidth $\rightarrow 0$ ($k \rightarrow \delta$) it tends to $\sum_{\sigma} w_{\sigma}^2 = 1/n_{\text{eff}}$, recovering plain inverse Simpson; at finite bandwidth it discounts clonotypes close in φ -space, so convergent near-duplicate receptors are not double-counted.

Proof. By Eq. (19), $\|\mu_\rho\|^2 = \langle \sum_\sigma w_\sigma \psi(z_\sigma), \sum_\tau w_\tau \psi(z_\tau) \rangle = \sum_{\sigma,\tau} w_\sigma w_\tau k(z_\sigma, z_\tau)$, Rao’s quadratic entropy with similarity matrix $Z_{\sigma\tau} = k(z_\sigma, z_\tau)$, whose order-2 Hill form is the Leinster–Cobbold diversity. With $k \rightarrow \delta$, $Z \rightarrow I$ and the sum collapses to $\sum_\sigma w_\sigma^2 = 1/n_{\text{eff}}$; under read weighting that is $1/2D$. A finite bandwidth makes near-duplicates $\sigma \approx \tau$ contribute $k \approx 1$ off-diagonal, damping their joint weight. $\square$

This is the unification: the depth-robust backbone Φ_1 and a φ -aware diversity are the first moment and the squared length of the *same* mean embedding. Diversity that respects sequence similarity costs nothing beyond the backbone.

T.7.5. INTERACTIONS LIVE IN THE SECOND MOMENT

Any functional of the marginal measure ρ_S is blind to *which* clonotypes co-occur — yet co-occurrence is the signature of shared selection.

PROPOSITION 33 (Second-order structure and its necessity). *No functional of the first moment μ_{ρ_S} distinguishes two repertoires with the same marginal clonotype measure but different joint co-occurrence; the compressed second moment $\Sigma_S = \sum_\sigma w_\sigma \psi(z_\sigma) \psi(z_\sigma)^\top$ (a codebook-free Fisher vector [74, 75]) does. Public, HLA-restricted clusters [57, 76, 18] are exactly such second-order structure.*

Proof. μ_{ρ_S} depends on ρ_S only through its first moment $\int \psi d\rho$, so two samples with equal marginals but different pairwise co-occurrence share it and no function of it separates them. The second moment $\Sigma_S = \int \psi \psi^\top d\rho_S$ is the uncentred covariance of the feature cloud — the Gaussian Fisher-vector / second-order pooling statistic, codebook-free here because ψ is a fixed random basis rather than a learned GMM. Convergent HLA-restricted responses deposit correlated mass in specific φ -regions, which Σ_S (and not μ_{ρ_S}) records. $\square$

In the characteristic-function language of Rem. 4 this is transparent: since $\cos a \cos b = \frac{1}{2}[\cos(a-b) + \cos(a+b)]$, an entry of Σ_S reads $\widehat{\rho}_S$ at the *sum and difference* frequencies $\omega_j \pm \omega_k$, whereas Φ_1 reads it only at the single frequencies ω_j . The first moment samples the characteristic function; the second moment samples its frequency mixing — and frequency mixing is exactly the spectral signature of co-occurrence, which is why the step from Φ_1 to Σ_S is the step from marginals to interactions.

Two routes exploit this and mirpy treats them as co-equal. The unsupervised route ships Σ_S (or its leading spectrum) as a third block. The supervised route replaces the fixed pool by a learned permutation-invariant set network — Set Transformer inducing points or DEEPRC attention [77, 78] — whose heads learn the public-cluster detectors directly.

T.7.6. HLA CONDITIONING

PROPOSITION 34 (HLA-restricted overlap). *Let an antigen response occupy an embedding ridge R presented only by HLA type h . Its contribution to $\langle \mu_{\rho_S}, \mu_{\rho_{S'}} \rangle$ is proportional to $\mathbf{1}[h \in \text{HLA}(S)] \mathbf{1}[h \in \text{HLA}(S')]$: two samples share the ridge’s mass only when both carry h . Hence CMV^+ repertoires are close on the CMV ridge only within an HLA-matched stratum, and the appropriate distance is an HLA-stratified MMD (or plain MMD with HLA appended as a covariate).*

Proof. A response-specific receptor enters ρ_S only if its clone was selected, which requires the presenting allele h ; so the ridge- R mass of μ_{ρ_S} vanishes unless $h \in \text{HLA}(S)$. The ridge term of $\langle \mu_{\rho_S}, \mu_{\rho_{S'}} \rangle = \sum_{\sigma, \tau} w_{\sigma} w'_{\tau} k(z_{\sigma}, z_{\tau})$ is a product of the two ridge masses, hence carries the indicator product. Stratifying the MMD by HLA match restores specificity — the mechanism behind the CMV/HLA signal of [57]. $\square$

REMARK 7 (The imprint spans MHC class II, and is strongest in TRA). On four-digit-typed data the second-moment / co-occurrence block (Prop. 33) detects HLA carriage across loci and *both* MHC classes — 15/17 alleles directional, class II 8/9, top DRB1*07:01 at AUC 0.758 — extending the class-I, β -chain result of [76] to class II. On a paired cohort the α chain carries the stronger imprint (DRB1*15: α 0.81 vs β 0.75; A*02 0.64/0.54; B*07 0.70/0.65): TRA’s lower junctional diversity makes HLA-restricted public α clones more shared, hence more detectable. Naive $\alpha + \beta$ concatenation sits *between* the chains — the noisier β dilutes — so HLA inference should use α (or an α -weighted distance), not an unweighted concat. The chain asymmetry is a testable prediction of the junctional-diversity argument (benchmark_repertoire_covidhla.py, _covidpaired.py).

T.7.7. THE CODEBOOK-FREE LIMIT OF THE GRAPH-CLUSTER-HISTOGRAM HEURISTIC

A common recipe builds a global clonotype graph, clusters it into K classes, and represents a sample as its histogram over classes; its weakness is the arbitrary K and clustering rule. The kernel mean embedding is the principled limit of exactly this recipe.

PROPOSITION 35 (Soft assignment removes K). *A hard-assignment bag-of-clusters (VLAD/Fisher) histogram over K codewords converges, as $K \rightarrow \infty$ with soft (RBF) assignment, to the kernel mean embedding μ_{ρ_S} of Eq. (20); the random-Fourier map ψ realises this limit in D fixed coordinates with no clustering step.*

Proof. Soft assignment bins mass by $c_k(z) = k(z, \mu_k) / \sum_l k(z, \mu_l)$ over codewords $\{\mu_k\}$; as the codebook densifies, $\sum_{\sigma} w_{\sigma} c_k(z_{\sigma})$ is a Nadaraya–Watson estimate of ρ_S against the kernel, whose infinite-codebook limit is $\int k(\cdot, z) d\rho_S = \mu_{\rho_S}$. RFF approximates k by $\psi^{\top} \psi$ (Eq. (19)), so μ_{ρ_S} is the D -vector Φ_1 , computed without choosing K or clustering. VLAD/Fisher vectors are the finite- K , first/second-order truncations of this limit [75, 74]. $\square$

The answer to “which K , which clustering” is therefore: none. The soft infinite-codebook limit is a closed form (the RFF mean embedding), and the graph-cluster-histogram methods are its truncations.

REMARK 8 (Optimal transport as the geometry-faithful alternative). The mean embedding compares repertoires by their first moment; when the geometry of the rare tail matters more than depth-robustness, *linear optimal transport* [79, 80] gives the natural competitor — a fixed Euclidean vector (the barycentric transport map from a fixed reference measure to ρ_S) whose distance approximates the Wasserstein metric on embedding space. It is more tail-faithful but pays an $n^{-1/p}$ empirical-convergence price (the ambient embedding dimension p), so it trades the depth-robustness of Prop. 30 for geometry; mirpy keeps it as an add-on, not the backbone.

T.7.8. DECOUPLING DEPTH AND BATCH FROM BIOLOGICAL SIGNAL

A phenotype signal — age, CMV status — rides on two technical nuisances that also move Φ but are not biology: *depth* (how deeply the sample was sequenced) and *batch* (which platform, primer set, and pipeline produced it). The measure view separates all three. Model the observed sample as the true repertoire measure passed through depth thinning and then a batch distortion,

$$\rho_S^{\text{obs}} = \mathcal{B}_{\text{batch}}(\text{Thin}_N(\rho_S^{\text{true}})), \quad (23)$$

where Thin_N is Poisson/hypergeometric subsampling to depth N (§T.6.7, Prop. 26) and $\mathcal{B}_{\text{batch}}$ reweights the measure by a bounded, batch-specific field $b(z) > 0$ (primer efficiency, GC/length amplification bias) shared by every sample run in that batch.

DEPTH IS ESTIMATION VARIANCE, NOT BIAS. Thinning does not change the target μ_P ; it only lowers the effective sample size, so by Prop. 30 it contributes a fluctuation of size $O(n_{\text{eff}}^{-1/2})$ and nothing systematic. Reporting diversity at a common coverage rather than a common depth (Prop. 31) removes the residual depth dependence of the diversity block. Depth is therefore a bounded, standardisable *variance*.

BATCH IS A SHARED SHIFT THAT CANCELS IN CONTRASTS. Batch is a genuine bias, but a structured one.

PROPOSITION 36 (Within-batch contrasts are batch-free). *Under Eq. (23) with batch field b shared by all samples in a batch: (i) the raw mean embedding is confounded, $\mu_S^{\text{obs}} = \int \psi b d\rho_S^{\text{true}} / \int b d\rho_S^{\text{true}}$; but (ii) any within-batch contrast is invariant to b : the density ratio of two same-batch samples has b cancel, $\rho_A^{\text{obs}} / \rho_B^{\text{obs}} = \rho_A^{\text{true}} / \rho_B^{\text{true}}$, and to first order the mean difference $\mu_A^{\text{obs}} - \mu_B^{\text{obs}}$ removes the shared shift.*

Proof. (i) is the pushforward of ρ_S^{true} under the reweighting b . (ii) In the ratio $\rho_A^{\text{obs}} / \rho_B^{\text{obs}}$ the common factor b occurs in numerator and denominator and cancels pointwise. Writing the multiplicative bias in the feature mean as an additive log-scale shift, $\mu_S^{\text{obs}} = \mu_S^{\text{true}} + \delta_{\text{batch}} + o(\|b - 1\|)$ with δ_{batch} identical for every sample in the batch, so the within-batch difference cancels δ_{batch} . This is the shared-field cancellation of Prop. 24 with the selection field q there replaced by the batch field b . $\square$

Proposition 36 dictates the design (Fig. 7). *Prefer a within-batch or paired design* — run case and control in the same batch — so the biological contrast is batch-free by construction. Where that is impossible, two fallbacks apply: *covariate adjustment*, residualising Φ on batch indicators or stratifying the MMD by batch exactly as it is stratified by HLA in Prop. 34; and *reference normalisation*, dividing each sample’s density by the shared P_{gen} pushforward f_{gen} (§T.6), which removes the multiplicative bias common to the generation background. In every case the operative object is a *contrast*, never a raw position.

REMARK 9 (When batch is confounded with the phenotype). The cancellation of Prop. 36 needs the batch to be *balanced* across the contrast — case and control both present in each batch (Fig. 7a). If instead batch is *confounded* with the phenotype — all cases sequenced in one run, all controls in another — the batch shift δ_{batch} and the biological effect v are aliased: the observed group difference is $v + \delta_{\text{batch}}$

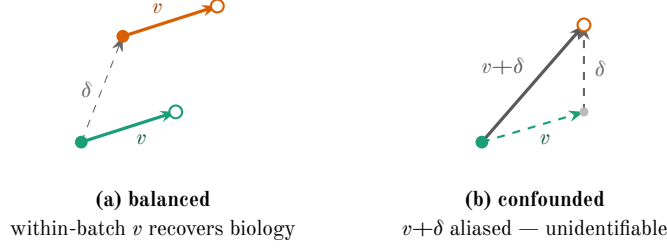


Figure 7: Batch cancellation needs a balanced design (Prop. 36, Rem. 9). **(a)** When both batches contain case and control, the case→control vector v (biology) is identical across batches while their absolute positions differ by the batch shift δ , so a within-batch (paired) contrast recovers v free of δ . **(b)** When batch is confounded with the phenotype — cases in one run, controls in another — the observed difference is $v + \delta$ with no within-batch pair to split it (the true v and δ , dashed, are unobservable): they are *aliased* and unidentifiable. Filled = case, open = control; teal = batch 1, orange = batch 2. Structure partly rescues (b) — δ lives mainly in the V/J -usage and depth subspace, adjustable without unmixing the CDR3 signal (Rem. 9).

and no contrast can separate them (Fig. 7b). This is a design failure, unidentifiable in general. It is *partly* rescued by structure: a batch acts mainly through V/J primer-amplification bias and sequencing depth, which live in the low-rank germline subspace (§T.4, Prop. 12) and the depth/diversity direction — largely orthogonal to the junction/CDR3-motif directions that carry antigen specificity. Projecting out the V/J -usage and depth components (or conditioning the MMD on V/J usage) removes the bulk of a confounded batch effect while sparing the CDR3 signal, and a spike-in or technical replicate calibrates the residual. Full confounding, though, cannot be undone after the fact: the only complete remedy is to balance batches across the phenotype at design time.

THE DECOMPOSITION. Depth and batch together give

$$\Phi_{\text{obs}}(S) = \underbrace{\Phi_{\text{bio}}(S)}_{\text{signal}} + \underbrace{\varepsilon_{\text{depth}}}_{O(n_{\text{eff}}^{-1/2}), \text{ Prop. 30}} + \underbrace{\delta_{\text{batch}}}_{\text{batch-shared, Prop. 36}}, \quad (24)$$

so a phenotype is read by a within-batch (or batch-adjusted) contrast, standardised to a common coverage, and tested against a label-permutation null — depth entering only as the variance the permutation test already accounts for, and batch removed by the contrast.

REMARK 10 (Batch is real, first-order, and cancellable: a validated recipe). Proposition 36 is borne out on nine real sequencing runs of a paired-chain cohort. The embedding Φ encodes the batch strongly — a one-vs-rest batch classifier reaches AUC 0.78 — and *residualising Φ on the batch indicator collapses it to chance* (0.03). The two regimes of Rem. 9 then separate cleanly: a batch-*orthogonal* biological signal (HLA) is *preserved* through the correction (0.60 → 0.61), while a batch-*confounded* one (COVID status, with some runs all-healthy) collapses to its honest within-batch value (0.66 → 0.41) — the naive number was riding the confound. This dictates a four-step cookbook, to run before any biological reading: (1) *detect* — a batch one-vs-rest AUC $\gg$ chance; (2) *quantify* — the MMD ratio (same-phenotype-across-batch : cross-phenotype-within-batch), ≈ 1.05 here, i.e.

batch as large as biology; (3) *correct* — residualise Φ on the batch indicator, or contrast within-batch / stratify the MMD by batch (Prop. 36); (4) *verify* — the batch AUC falls to chance while a known batch-orthogonal signal survives (`benchmark_repertoire_covidbatch.py`).

T.7.9. NORMALISING HETEROGENEOUS RNA-SEQ: DEPTH, KIT, READ LENGTH, AND THE TISSUE INFILTRATION CONFOUND

The target modality — bulk RNA-seq, 10^2 – 10^4 clonotypes per chain but 10^5 samples with clinical metadata (§ T.7) — is heterogeneous on four axes that Eq. (23) folded into one batch field: sequencing *depth*, library *kit*/primer chemistry, *read length*, and, in tissue uniquely, the T-cell *content* itself. The measure factorisation sorts these into three fates — *quotiented for free*, *corrected on an anchor*, and *kept as biology* — and the sorting is not a matter of taste: it follows from whether a nuisance scales every clonotype alike or reshapes the profile. Write the recovered read count of clonotype σ in sample S as

$$a_\sigma \approx \underbrace{g_{\text{kit}}(\sigma)}_{\text{kit gain}} \cdot \underbrace{R\theta c}_{\sigma\text{-independent scalars}} \cdot p_\sigma \cdot \xi_\sigma, \quad \sum_\sigma p_\sigma = 1, \quad (25)$$

with R the sequenced library size, θ the T-cell fraction (*infiltration*), c the mean receptor transcripts per T cell, p_σ the true clonal frequency *within* the T-cell compartment, $g_{\text{kit}}(\sigma) > 0$ the kit- and length-dependent amplification gain, and ξ_σ multinomial sampling noise. The recovered TCR depth is $N = \sum_\sigma a_\sigma \approx R\theta c \sum_\tau g_{\text{kit}}(\tau) p_\tau$. The one structural fact is that R, θ, c do *not* depend on σ whereas $g_{\text{kit}}(\sigma)$ does, and everything below is that distinction applied.

PROPOSITION 37 (Frequencies quotient every scalar; only the kit gain survives). *Under Eq. (25) the frequency representation $f_\sigma = a_\sigma / \sum_\tau a_\tau$ cancels every σ -independent factor,*

$$f_\sigma = \frac{g_{\text{kit}}(\sigma) p_\sigma}{\sum_\tau g_{\text{kit}}(\tau) p_\tau}, \quad (26)$$

so depth R and infiltration θ (and c) vanish identically while the σ -dependent gain g_{kit} does not. Consequently a frequency-weighted $\Phi(S)$ is exactly invariant to sequencing depth and to infiltration, and the only multiplicative nuisance it still carries is the kit/length gain g_{kit} .

Proof. $R\theta c$ pull out of numerator and of the denominator sum and cancel; $g_{\text{kit}}(\sigma)$ and p_σ are σ -indexed and remain. Invariance of Φ_1 to N is Prop. 30; the identical argument gives invariance to θ and c , which occupy the same scalar position in Eq. (25). $\square$

This is the organising statement, and it is almost embarrassingly simple: depth and infiltration are the *same kind of thing* — a scalar amount of material — so frequency normalisation kills both with no model, while the kit gain is a *different kind of thing* — a reshaping of the profile — that no rescale can remove. The three remaining axes each take one consequence.

DEPTH: STANDARDISE COVERAGE, DO NOT CHASE DEPTH. Once frequencies quotient R , the only residue of depth is estimation variance (Prop. 30) and a diversity bias that coverage standardisation removes (Prop. 31). Do *not* down-sample deep samples to a common *depth* — that discards real observations to match the shallowest sample. Standardise to a common *coverage* C^* and floor samples

whose $\hat{C}$ falls below it. More finely, depth acts through a per-clonotype *detection probability*: a clonotype of frequency f is seen at depth s with probability $\approx 1 - (1 - f)^{s/\bar{s}}$ (relative to a reference depth $\bar{s}$), so depth enters as an exponent — a censoring, not a rescaling. This is why an ad-hoc inverse-depth reweighting $w_i \propto 1/s_i$ has *no* calibrated null, whereas the Poisson–binomial detection model does: it is the sample-level face of the exact point-process test of §T.6.3, and it is what a co-occurrence or enrichment statistic should be built on.

INFILTRATION: ESTIMATE IT, CARRY IT, NEVER NORMALISE IT AWAY. In tissue the count total conflates two things frequency normalisation deliberately cannot see, and the reflex to “normalise for depth” destroys a prognostic signal.

PROPOSITION 38 (Infiltration is identifiable, and is a channel not a normaliser). *The recovered TCR depth $N \approx R\theta c$ confounds technical depth R with biological infiltration θ : a shallowly sequenced sample and a T-cell-poor tissue are indistinguishable from N alone. But θ is separately identifiable from the TCR read fraction*

$$\hat{\theta} \propto \frac{N_{\text{TCR}}}{N_{\text{total}}} = \frac{R\theta c}{R} = \theta c, \quad (27)$$

in which the library size R divides out. Hence: (i) estimate infiltration from the read fraction (or a deconvolution signature [81]) and carry it as an explicit scalar channel; (ii) build $\Phi(S)$ on frequencies, so repertoire shape (clonality, focusing) is orthogonal to amount θ (Prop. 37); (iii) the depth–infiltration confound in N alone is resolved by Eq. (27), and what stays irreducible is only the joint low- R , low- θ regime, where n_{eff} collapses and Φ is coverage-limited.

Proof. Eq. (25) gives $N = R\theta c \sum gp$; dividing by the total mapped reads $N_{\text{total}} \approx R$ removes R , leaving θc up to the kit constant $\sum gp$. Orthogonality of shape to θ is Prop. 37. When both R and θ are small N is small, so $n_{\text{eff}} \leq N$ is small and the $O(n_{\text{eff}}^{-1/2})$ bound of Prop. 30 is loose — an uncertainty to propagate, not a bias to subtract. $\square$

This answers “depth and infiltration are not distinguishable” directly: from the count total N alone they indeed are not, but the read *fraction* $N_{\text{TCR}}/N_{\text{total}}$ separates them, being a technical-over-technical ratio in which depth cancels and infiltration does not [82]. Down-sampling to a common depth — the amplicon reflex — would erase exactly the prognostic tumour-infiltrating-lymphocyte signal that makes tissue RNA-seq worth sequencing. Infiltration is not noise; it is a low-dimensional biological channel that sits *beside* $\Phi(S)$: θ says how much T cell, Φ says of what shape, and the two are complementary. The correction burden is regime-dependent. In peripheral blood, where lymphocyte density per unit volume is roughly constant, N is *primarily technical* and θ a light adjustment; in a tumour biopsy N folds in the TIL fraction across a $\sim 10^4$ -fold range, and separating θ from depth is essential. The same count total that is a disposable nuisance in blood is a biological readout in tissue — which is exactly why it must be routed to a channel, not to a normaliser.

KIT AND READ LENGTH: CORRECT THE σ -DEPENDENT GAIN ON AN ANCHOR. The surviving gain $g_{\text{kit}}(\sigma)$ is not arbitrary. Read length enters through a length-recovery curve, and chemistry through a V/J bias — both confined to the low-rank germline subspace, not the CDR3-motif directions.

PROPOSITION 39 (Read-length dropout is length-censoring). *Read length L enters g only through a recovery curve $r_L(\ell)$ decreasing in junction length ℓ : $g_{\text{kit}}(\sigma) = r_L(\ell_\sigma) g_{\text{chem}}(\sigma)$. This is missing-not-at-random on the length marginal — it depresses the long-CDR3 tail and drops the clonotypes there — and it is confined to the length/ V - J subspace (Prop. 12). Two corrections restore comparability: (i) restrict to the length range commonly recoverable across kits (the intersection of supports) — lossy but assumption-free; or (ii) reweight surviving clonotypes by $1/r_L(\ell)$ (Horvitz–Thompson), r_L calibrated on samples sequenced at several read lengths.*

Both the chemistry bias and the residual of a confounded batch (Rem. 9) live in the low-rank germline-plus-depth subspace, orthogonal to the CDR3-motif directions that carry specificity (Prop. 12). So the correction is a *subspace* operation, not a global rescale: estimate the batch shift on an anchor and project it out of the V/J -usage and length coordinates, leaving the motif coordinates untouched. The anchor is what makes the shift identifiable, and there are three sources, strongest first. A *technical replicate* or a shared *reference sample* run on every kit measures δ_{batch} directly, whether subtracted empirical-Bayes (ComBAT-style [83]) or aligned as clouds (Harmony-style [84]). *Negative-control features* — the germline V/J usage *predicted by* the rearrangement model P_{gen} , which is biology-invariant — play the role of RUV’s housekeeping genes, letting the batch direction be estimated from features that carry no signal [85]. Weakest, batch *indicators* alone suffice only under the balanced design of Prop. 36. With no anchor and batch confounded with the phenotype the shift is unidentifiable (Rem. 9): the projection removes its bulk, but only balancing batches across the phenotype at design time is a complete fix.

REMARK 11 (Stratified conditional testing beats regression adjustment). When a covariate moves the marginal frequencies alongside depth — age (thymic involution lowers naive output), HLA, CMV status — adjust by *stratification* rather than by residualising on a linear term. Bin samples by $(\lfloor \log_{10} s \rfloor \times \text{covariate band})$ and test the phenotype association conditional on the strata with a Cochran–Mantel–Haenszel test [86], estimating the marginal frequencies $\hat{f}_k$ *within* each stratum. Conditioning on depth *and* the covariate jointly is strictly more stringent than adjusting for either: a signal that survives is neither a depth artefact nor the age-related clonal contraction that would otherwise masquerade as disease-specific focusing. The test degrades gracefully to the exact point-process form (§T.6.3) when a stratum is too small for the normal approximation.

COURSE OF ACTION. Table 1 is the recipe — each nuisance treated by its nature, not by a single omnibus “normalisation”. The discipline reduces to one line: *scalars are quotiented by frequencies, shape distortions are projected out on an anchor, and the one scalar that is biology — infiltration — is kept as its own channel.*

T.7.10. COMPRESSION: THE BACKGROUND IS A MODEL, THE SIGNAL IS A LIST

A deep repertoire lists 10^5 – 10^6 clonotypes, but its *information* is far smaller: most of it is the near-neutral background that the recombination model already predicts, and only a sparse signal — the antigen-driven, convergent *metaclonotypes* — is genuinely new. This invites a two-part (description-length) code [87], and the mixture of §T.6.1 makes it precise: $f_{\text{obs}} = \pi f_{\text{sig}} + (1 - \pi) f_{\text{bg}}$ with background $f_{\text{bg}} \approx \varphi_{\#}(P_{\text{gen}} \cdot Q/Z_Q)$, the pushforward of the selection-reweighted generation law.

Table 1: Normalising a heterogeneous RNA-seq cohort. Each nuisance is handled by whether it scales clonotypes uniformly (σ -independent, quotiented by frequencies) or reshapes the profile (σ -dependent, corrected on an anchor) — with infiltration the one scalar deliberately *not* removed.

Nuisance	Nature	Treatment	Basis
depth R	σ -indep. technical	frequencies; coverage-standardise diversity	Prop. 37, 31
infiltration θ	σ -indep. <i>biological</i>	estimate from read fraction; carry as a channel	Prop. 38, Eq. (27)
kit / primer chemistry	σ -dep. shape	anchored subspace correction (V/J usage); spare CDR ₃	Prop. 36, Rem. 9
read length	σ -dep. censoring	common-range restrict / inverse-recovery reweight	Prop. 39
pipeline / batch	σ -dep. shift	within-batch contrast; P_{gen} -reference normalise	Prop. 36

PROPOSITION 40 (The background is model-compressible). *The background sub-repertoire contributes to $\Phi(S)$ only through the deterministic functional $\mu_{\text{bg}} = \int \psi d\varphi_{\#}(P_{\text{gen}}Q/Z_Q)$ of the model (P_{gen}, Q) , up to an $O(n_{\text{eff}}^{-1/2})$ sampling fluctuation (Prop. 30); its diversity block is the model’s coverage-standardised Hill profile. Hence storing $(P_{\text{gen}}, Q, N, \pi)$ and the enriched set determines $\Phi(S)$ to that precision at a description length independent of the number of naive clonotypes.*

Proof. The background clonotypes are draws from $P_{\text{gen}}Q/Z_Q$; their frequency-weighted mean feature converges to μ_{bg} with fluctuation $O(n_{\text{eff}}^{-1/2})$ by Prop. 30, and their Hill numbers to the model’s (Prop. 31). Neither depends on *which* naive clonotypes were drawn, only on the model and the depth, so the naive list carries no information about Φ beyond the sampling noise. $\square$

The code is therefore

$$\Phi(S) \approx \underbrace{(1 - \pi) \mu_{\text{bg}}(P_{\text{gen}}, Q)}_{\text{model: } O(1) \text{ description}} + \pi \underbrace{\sum_{\sigma \in \text{signal}} w_{\sigma} \psi(z_{\sigma})}_{\text{explicit metaclonotypes}} : \quad (28)$$

store (i) the shared model plus the per-sample scalars $(N, \pi, {}^qD)$ — a constant-size summary of the naive bulk — and (ii) the sparse list of enriched metaclonotypes above the density water-level (§ T.6), the incompressible remainder (Fig. 8). This is “compress the predictable, keep the surprising”: the naive part, being the image of a ~ 40 -bit process (§ T.3), collapses onto the model, while the signal — typically $< 5\%$ of clonotypes — is retained verbatim. Across a cohort the model is shared, so only $(N, \pi, {}^qD, \text{signal})$ vary per sample; the codec of § T.9 is the per-sequence layer beneath this per-sample one. The embedding as a whole is not a compressor — § T.9 shows the per-sequence code is an *expansion*; the compression here is at the *ensemble* level, in not storing what the model generates.

T.7.11. TRACING THE SIGNAL BACK TO CLONOTYPES AND MOTIFS

A repertoire feature that predicts a phenotype is actionable only if the exact receptor can be named: to build a diagnostic or a therapeutic against an autoimmune-associated TCR one needs its V gene and CDR3 sequence, not a summary vector. The sample embedding is lossy on the background by design (§T.7.10) but the signal is both localisable and retained.

PROPOSITION 41 (Witness attribution to metaclonotypes). *Let μ_A, μ_B be the mean embeddings of two sample groups. The MMD witness $w = \mu_A - \mu_B$ [65] scores a clonotype σ by*

$$s(\sigma) = \langle w, \psi(\varphi(\sigma)) \rangle = \sum_{a \in A} w_a k(\varphi(\sigma), z_a) - \sum_{b \in B} w_b k(\varphi(\sigma), z_b), \quad (29)$$

the kernel-smoothed differential enrichment of σ between the groups; the receptors maximising s are the response-associated ones. By the product-metric factorisation of φ (§T.4) the kernel splits over the V/J and junction blocks, so s decomposes into a germline-usage part and a CDR3-motif part — naming a (V , CDR3 k -mer) metaclonotype [18, 88].

Proof. $\langle w, \psi(\varphi(\sigma)) \rangle = \langle \mu_A - \mu_B, \psi(\varphi(\sigma)) \rangle$ expands via $\psi^\top \psi = k$ (Eq. (19)) to Eq. (29). A product kernel $k = k_V k_J k_{\text{junc}}$ on the interleaved blocks (§T.4) factorises the score, and thresholding the junction factor on shared k -mers recovers the CDR3 motif. $\square$

Three routes read the signal back, in increasing specificity. (i) *The witness* (Eq. (29)) ranks every observed clonotype by its contribution to the group difference — a linear read-off of the same embedding, no auxiliary model. (ii) *Density ridges* (§T.6): the enriched regions cluster (level sets / DBSCAN) into metaclonotypes — a V gene + CDR3 k -mer motif in the GLIPH/TCRdist sense [18, 88], the standard actionable unit. (iii) *Retention, not reconstruction*: the exact sequence is *kept*, not decoded. The two-part code of §T.7.10 retains the signal clonotypes verbatim (V call + junction string), so tracing an autoimmune cluster to its receptors is a *retrieval*. The inverse codec of §T.9 serves the complementary task — *generating* novel candidate sequences inside a ridge (to design a probe or expand a motif family) — not recovering observed ones, which are on hand. The design is thus lossy on the noise and lossless on the signal: exactly where immunological traceability demands it, the exact (V , CDR3) receptor survives, ready for a tetramer, a diagnostic, or a therapeutic.

REMARK 12 (The bulk noise floor: a long-past exposure need not be readable). The embedding reads *distributional* structure — diversity and public-clone co-occurrence — so a phenotype carried by rare *private* memory clones can sit below the bulk noise floor. A convalescent (long-resolved) SARS-CoV-2 exposure is a clean example: at RNA-seq depth it is *not* recoverable from the bulk repertoire after batch correction — classification is chance (0.49–0.52 across β/α /paired), and a from-scratch MMD witness (Prop. 41) does not rediscover the known response clones (0.37 β /0.45 α); the apparent 0.66–0.72 was the batch confound of Rem. 10. Reading such rare memory-phase clones needs a *supervised* targeted burden over a known clone panel, or a differential control in the density machinery of §T.6 — not an unsupervised bulk contrast. This is the honest boundary of the modality: *diversity for how-even, the embedding for which public clones*, but not for a handful of private ones lost in the naive background (`benchmark_repertoire_covidstatus.py`).

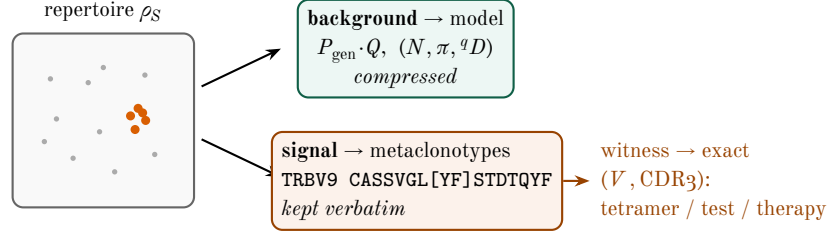


Figure 8: The two-part sample code and its traceability. The naive *background* (grey) is the image of the ~ 40 -bit generation-plus-selection process, so it is stored as the model (P_{gen}, Q) plus the per-sample scalars ($N, \pi, {}^qD$) — compressed, not enumerated (§ T.7.10, Prop. 40). The sparse *signal* (orange ridge) — antigen-driven, convergent metaclonotypes above the density water-level — is kept verbatim as a list of $(V, \text{CDR3})$ motifs. The MMD witness (Prop. 41) ranks these by phenotype association, so a response- or autoimmune-linked cluster is traced to the *exact* receptor sequences it retains — ready for a tetramer, diagnostic, or therapeutic.

COURSE OF ACTION. The default $\Phi(S)$ concatenates three blocks, each owning one requirement: the RFF kernel mean Φ_1 (backbone; (a),(b)), the coverage-standardised Hill profile Φ_2 ((c)), and the second moment Σ_S ((d)), with a learned set network as the co-equal alternative to Σ_S . Sample distance is MMD, HLA-stratified when antigen specificity is the target. Every default is derived (Table 2): frequencies not counts for scale-freeness (Prop. 30), a concave g for outlier-robust weighting (§ T.6.9), bandwidth from the one-substitution scale r_1 (§ T.6), and coverage- not depth-standardised diversity (Prop. 31). The propositions’ empirical counterparts are borne out by `mir/repertoire.py` and its benchmarks: the $n_{\text{eff}}^{-1/2}$ depth-robustness curve (Prop. 30), the $n_{\text{eff}} = {}^qD$ identity (Prop. 31), HLA-stratified separation across both MHC classes (Prop. 34, Rem. 7), and batch cancellation under residualisation (Prop. 36, Rem. 10). Three lessons temper the reading and are the operative guidance. (i) *Compare contrasts, not raw distances*: the unbiased MMD and a within-batch (or batch-residualised) design (Rem. 5, 10), since both the MMD self-term and an unbalanced batch otherwise manufacture signal. (ii) *The imprint is richest in TRA and MHC class II* (Rem. 7). (iii) *Honest negatives*: aging divergence is scalar diversity in disguise (Rem. 6), and a long-past exposure leaves no batch-robust bulk biomarker (Rem. 12). The one-line thesis — *diversity for how-even, the embedding for which clones* — holds only when read through contrasts.

T.8. SAMPLE EMBEDDINGS AS A NEURAL-NETWORK MODALITY

The sample-level embedding of § T.7 is built to be a *modality*: a fixed-width tensor that plugs into a neural network directly, or fuses with other data types in a multimodal model. This section makes the computation explicit — the architecture that maps a variable-length list of clonotype embeddings and counts to $\Phi(S)$, how it is stored, and how it composes downstream (Fig. 9).

THE ARCHITECTURE IS A SET-POOLING NETWORK. By Prop. 29 the map $\{(z_\sigma, a_\sigma)\} \mapsto \Phi(S)$ is a permutation-invariant set network of the Deep Sets form [61]: a per-clonotype feature $\psi(z_\sigma)$, a size weight $w_\sigma = g(a_\sigma)/\sum_\tau g(a_\tau)$, a sum-pool $\sum_\sigma w_\sigma \psi(z_\sigma)$, and a read-out. Two instantiations share one

Table 2: Sample-level embedding $\Phi(S)$: each choice traces to a proposition rather than to tuning.

Step	Choice	Justification
weight $g(a)$	$\log(1+a)$ / Anscombe (concave)	outlier/Zipf robustness (§ T.6.9)
pooling	frequency-weighted sum	order-invariant + scale-free (Prop. 29, 30)
backbone Φ_1	RFF kernel mean, $D \approx 1\text{--}4\text{k}$	depth-robust $O(n_{\text{eff}}^{-1/2})$; codebook-free (Prop. 30, 35)
bandwidth	r_1 (one-substitution scale)	matches the sequence metric (§ T.6)
diversity Φ_2	$\{^0D, ^1D, ^2D\}$ at common $\hat{C}^*$	depth $\leftrightarrow$ diversity resolved (Prop. 31)
interaction	Σ_S or Set-Transformer/DEEPRC	co-occurrence (Prop. 33)
distance	MMD, HLA-stratified	metric on measures; HLA restriction (Prop. 34)
n_{eff}	$(\sum_{\sigma} w_{\sigma}^2)^{-1}$, a Hill number	ties depth to diversity (Prop. 31)

computational graph. The *fixed* network takes $\psi = \text{random-Fourier features}$ and needs no training — a single forward pass computing the kernel mean Φ_1 , the diversity block, and the second moment. The *learned* network takes $\psi = \text{an MLP}$ and replaces the sum by attention pooling — a Set Transformer or DEEPRC [77, 78] whose inducing points are trained metaclonotype detectors — and is fit end-to-end to a label. Both reduce the ragged clonotype axis to the same fixed $(D,)$ output, so one can initialise from the fixed embedding and fine-tune into the learned one.

TENSOR SHAPES AND THE RAGGED AXIS. A batch is a ragged stack of samples: sample i is a matrix $Z_i \in \mathbb{R}^{n_i \times p}$ of its clonotype embeddings with a count vector $a_i \in \mathbb{R}^{n_i}$, and the per-sample n_i differ (Fig. 5). The pooling layer contracts that ragged axis — $\sum_{\sigma}$ over the n_i clonotypes — to a fixed $(D,)$ vector, giving a dense (batch, D) tensor that any downstream MLP or transformer consumes. This contraction is the whole point of the set-network view: variable length enters, fixed width leaves (§ T.7.1), with no padding or truncation of the clonotype list.

STORAGE. $\Phi(S)$ is a fixed $(D,)$ float tensor — the three blocks concatenated (mean D_{μ} , diversity, second moment). A cohort is then an ordinary (n samples, D) design matrix, cached once and reused across models. For archival or cohort-scale storage the two-part code of § T.7.10 is the compact form — the shared model plus per-sample $(N, \pi, {}^qD)$ and the sparse metaclonotype list, from which Φ is reconstituted. Training a *learned* pooler instead consumes the raw ragged $\{(Z_i, a_i)\}$; the clonotype embeddings are cheap to recompute from the sequences (§ T.1, or the forward codec of § T.9).

AS A MODALITY IN MULTIMODAL MODELS. Because $\Phi(S)$ is a fixed-width vector it drops into any fusion scheme. *Early* fusion concatenates it with other modalities — clinical covariates, bulk/single-cell transcriptome, proteomics, and the HLA genotype of Prop. 34 — into one input; *intermediate/late* fusion gives each modality its own encoder and combines them by attention or cross-attention. The repertoire branch is exactly the set-pooling network above; HLA enters both as a fusion input and, per Prop. 34, as the stratifier of the repertoire distance. The two-part code (§ T.7.10) keeps the fused model interpretable: an attribution on the repertoire branch traces back through the witness (Prop. 41) to the exact metaclonotypes (§ T.7.11).

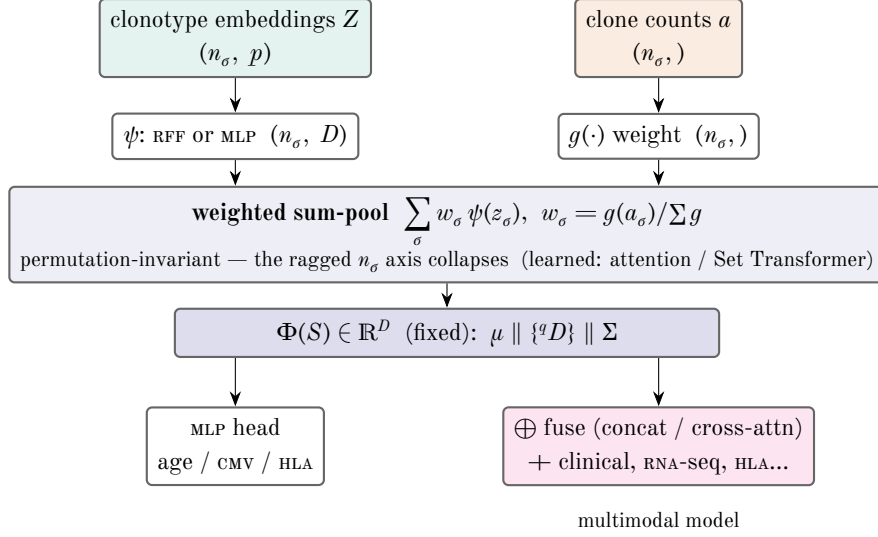


Figure 9: The sample embedding as a neural-network block. A ragged batch — sample i a matrix of n_i clonotype embeddings Z_i with counts a_i — is featurised per clonotype ($\psi = \text{RFF}$ for the fixed embedding, a learned MLP for the trained one), weighted by the variance-stabilised clone size $g(a_\sigma)$, and **sum-pooled**: the ragged n_σ axis collapses to a fixed $\mathbb{R}^D$ vector (Prop. 29, 30), so variable-length repertoires become a dense (batch, D) tensor. The learned variant swaps the sum for attention pooling (Set Transformer / DEEPRC). $\Phi(S)$ then feeds an MLP head (age / CMV / HLA) or fuses with other modalities (clinical, transcriptomic, HLA) as one branch of a multimodal model.

PART III

INVERTING THE EMBEDDING, AND ITS LIMITS

Turning the map around: recovering sequences from the embedding, the privacy that reconstruction implies, and the theoretical frontier that remains open.

T.9. CODEC INVERTIBILITY AND THE LIMITS OF LOSSLESSNESS

The forward and inverse neural codecs (`mir/ml/encoder.py`, `mir/ml/decoder.py`) turn the embedding into a learned, approximately invertible map $x \mapsto Z \mapsto \hat{x}$, raising a sharp question: *how much of the receptor does the compact code actually keep?* The answer separates cleanly into three notions of losslessness — geometric, informational, and reconstructive — that are measured independently by `mir/bench/theory.py` and answer different halves of the question. The punchline is that on real data the code is *informationally* lossless (injective) while remaining only partly *reconstructively* lossless, so every missing exact match is a limitation of the decoder and its training data, not of the code.

DEFINITION 4 (Three losslessness levels). Fix an encoder $Z: \mathcal{X} \rightarrow \mathbb{R}^m$ (a receptor mapped to its compact code) and a data set $S \subset \mathcal{X}$. (i) *Geometric* — Z preserves pairwise distance: $\text{corr}(\|Z(x) - Z(y)\|, d(x, y)) \rightarrow 1$ (the T1 claim, §T.1, validated by S2). (ii) *Informational* — Z is injective on

S : distinct receptors get distinct codes, so x is *determined* by $Z(x)$. Quantify by the collision rate at resolution ε , $\gamma(\varepsilon) = |\{x \in S : \exists y \in S, y \neq x, \|Z(x) - Z(y)\| < \varepsilon\}|/|S|$. (iii) *Reconstructive* — a decoder g recovers the sequence: exact-match $\text{EM}(g) = |\{x : g(Z(x)) = x\}|/|S|$.

THE CEILING. No decoder can beat informational losslessness: if two distinct receptors share a code (within the decoder’s resolution ε), any g maps both to one string and loses at least one, so

$$\text{EM}(g) \leq 1 - \gamma(\varepsilon) =: \text{EM}^*, \quad (30)$$

the *information ceiling*. The gap $\text{EM}^* - \text{EM}(g)$ is decoder-limited; $1 - \text{EM}^*$ is information-limited. Measuring both localises the loss.

T.9.1. DISTANCE PRESERVATION FORCES INJECTIVITY

The geometric and informational levels are not independent: the lower half of the bi-Lipschitz bound behind T1 already *implies* injectivity.

PROPOSITION 42 (Geometric $\Rightarrow$ informational losslessness). *Suppose the embedding obeys a lower bound $\|Z(x) - Z(y)\| \geq c \rho(x, y)$ for some $c > 0$ and the negative-type metric $\rho = \sqrt{d}$ (Prop. 1 and §T.2), where $\rho(x, y) > 0$ whenever $x \neq y$. Then Z is injective and $\gamma(0) = 0$, so $\text{EM}^* = 1$: the code loses no sequence information. Truncating to $m < 3K$ principal components replaces Z by ΠZ and can violate the lower bound only along the discarded $(3K - m)$ -dimensional subspace — the same subspace that makes reconstruction a hard inverse problem in §T.10 — so a collision requires two receptors to differ only within it.*

Proof. If $x \neq y$ then $\rho(x, y) > 0$, so $\|Z(x) - Z(y)\| \geq c \rho(x, y) > 0$ and $Z(x) \neq Z(y)$; injectivity gives $\gamma(0) = 0$. Under truncation $\|\Pi Z(x) - \Pi Z(y)\|^2 = \|Z(x) - Z(y)\|^2 - \|(I - \Pi)(Z(x) - Z(y))\|^2$, which is positive unless $Z(x) - Z(y)$ lies in the range of $I - \Pi$. $\square$

EMPIRICALLY THE CODE IS INJECTIVE. At the recommended operating point ($K = 2000$ prototypes, $m = 300$ PCs, 98.7% retained variance) `codec_losslessness` finds **zero** collisions over held-out real TCR β (collision rate 0, so $\text{EM}^* = 100\%$), against a median nearest-code distance of 15.4 — distinct receptors are separated by a wide margin, not sitting on the collision threshold. The truncated code is thus injective in practice, and by (30) *all* of the reconstruction shortfall below is decoder-and-data-limited, none of it information-limited. This is the precise, measurable sense in which the embedding “keeps the fine differences” that determine specificity: single-residue variants never collapse to a shared code (consistent with the bounded one-substitution drift of §T.5).

T.9.2. RATE AND DISTORTION: DEPTH SATURATES, DATA DOES NOT

Reconstructive losslessness is a rate–distortion problem [89, 87]: trade code rate (bits) against reconstruction distortion. Two floors bound it — the fixed length-40 tokenisation frame (§T.5: a receptor with > 40 junction residues cannot round-trip, so $\text{EM} \leq$ the in-frame fraction, 100% for TCR, $\approx 99.9\%$ for real IGH) and the collision floor EM^* of (30).

Table 3: Codec losslessness on real held-out TCR β ($K = 2000$ prototypes, $n = 20,000$). Ceiling $\text{EM}^* = 1 - \gamma$ is flat at 100% (injective code); exact-match rises with code rate then saturates; edit distance is the graded distortion. Reproduced by `experiments/benchmark_codec_losslessness.py`.

PCs m	code bits	retained var.	ceiling EM^*	exact-match	mean edit
50	1,600	94.0%	100%	67.4%	0.565
100	3,200	96.6%	100%	78.2%	0.328
200	6,400	98.2%	100%	85.0%	0.206
300	9,600	98.7%	100%	88.9%	0.155
500	16,000	99.3%	100%	89.5%	0.143

PROPOSITION 43 (Rate–distortion of the inverse codec). *Let $\delta(r)$ be the optimal expected reconstruction distortion (mean edit distance, or $1 - \text{EM}$) at code rate r . Then δ is non-increasing in r down to the frame/collision floor, and increasing r past the point where the retained variance already resolves the sequences yields no further gain. The binding rate is instead the training rate n : the decoder estimates a map $\mathbb{R}^m \rightarrow \mathcal{X}$ from n examples, and its excess risk decays in n .*

Sketch. Monotonicity in r is the data-processing bound: a higher-rate code refines the σ -algebra the decoder conditions on, so the Bayes-optimal reconstruction weakly improves. Saturation follows from Prop. 42: once the code is injective the target is a deterministic function of Z , and added components carry no new information about x (only noise the finite-sample decoder must fit). The n -dependence is the standard excess-risk decay of the learned g . $\square$

The autoencoder viewpoint [90] makes the same point: the inverse codec is the decoder half of an autoencoder whose *encoder* is fixed to the interpretable prototype map, so its ceiling is set by that encoder’s injectivity (Prop. 42), and its realised distortion by decoder capacity and data. Table 3 shows both: distortion falls with code rate but plateaus near $m = 300$, while the ceiling stays pinned at 100%.

DATA IS THE LEVER. Holding the code fixed at ($K = 2000$, $m \approx 300$) and growing the *training* corpus, exact-match climbs $88.5\% \rightarrow 94.1\% \rightarrow 95.8\%$ at $n = 20\text{k} \rightarrow 50\text{k} \rightarrow 100\text{k}$ (`benchmark_lossless_kpc.py`; the 88.9% at $n=20\text{k}$ in Table 3 is the same configuration under a different split) — crossing 95% purely on data, with the same one-shot decoder and no extra bits. Prototype count K saturates near 2000 ($K = 10,000$ regresses and costs an order more memory and query time), and the per-position breakdown is uniform on TCR β (conserved anchors 99.5% vs. variable middle 99.6%) because its junctions are short; the middle-concentrated loss appears only for long IGH junctions near the frame limit (§ T.5).

LOSSLESSNESS IS NOT COMPRESSION. The code is an *expansion*: 300 float PCs is $\sim 10^4$ bits against a ~ 63 -bit junction (≈ 15 residues $\times \log_2 20$). The prototype embedding exists to *linearise similarity*, not to compress; for archival exact recovery one simply stores the string (and the exact categorical $V/J/C$ calls). The inverse codec earns its keep where a differentiable, samplable sequence space is needed — generation, interpolation, density modelling — not as a lossless coder.

REMARK 13 (Losslessness is privacy’s mirror). Prop. 42 is exactly the linkage hazard of §T.10: the injectivity that lets a key-holder invert or re-identify a code is the same property that makes the code retain the sequence. Reconstructive losslessness (a useful codec) and re-identifiability (a privacy risk) are one phenomenon seen from two sides; the differential-privacy noise of Prop. 44 degrades both together, by design.

T.10. PRIVACY: THE EMBEDDING AS A KEYED, LOSSY REPRESENTATION

Immune repertoires are individually identifying — like genomic data, a person’s TCRs are a stable, near-unique fingerprint — so whether an embedding may be shared in place of raw sequences is a practical question with a clean geometric answer. Reproducing a coordinate requires two secrets: the prototype set $P = \{p_k\}$ and the PCA rotation W fitted to the cohort (together with the centring μ and scaling). Write the released representation as $Z = (\varphi(X) - \mu) W \in \mathbb{R}^{n \times m}$ with $m \ll 3K$.

THREE BARRIERS TO RECONSTRUCTION. (i) *Lossy projection.* PCA to m components discards a $(3K - m)$ -dimensional subspace, so $\varphi \mapsto Z$ is many-to-one and non-invertible on the discarded directions *even given* W — an information-theoretic barrier no decoder can cross. (ii) *Unknown anchors.* Recovering x from $\varphi(x) = (d(x, p_k))_k$ is reconstruction from distances to *unknown* reference points in a discrete sequence space, a hard combinatorial inverse problem without P . (iii) *Data-dependent mixing.* W is estimated from the cohort, not a public transform; without it the released numbers are linear combinations of unknown prototype distances. Releasing Z alone is therefore a **keyed, lossy pseudonymisation**: the pair (P, W) is a secret key, and without it reconstruction reduces to a hard inverse problem stacked on an irreversible compression.

WHAT IT DOES NOT PROTECT. The embedding is built to preserve pairwise geometry (Prop. 1–3), and that is precisely what a linkage adversary exploits. A key-holder — or anyone able to embed a reference sequence under the same (P, W) — can (a) test membership by embedding a candidate and matching it against Z (re-identification / membership inference), (b) link the same patient across releases through preserved distances, and (c) invert approximately with a trained decoder: the neural inverse codec (`mir/ml/decoder.py`) demonstrates that φ is approximately invertible *given the key*. Distance preservation is thus double-edged — the source of the embedding’s analytic value and of its linkage risk. Keyed pseudonymisation is not anonymisation.

A DIFFERENTIAL-PRIVACY MECHANISM THE GEOMETRY HANDS US. The Hölder- $\frac{1}{2}$ property of Prop. 1 turns the embedding into a clean substrate for a formal guarantee. Replacing one receptor moves its coordinate by $\|\varphi(x) - \varphi(x')\|_2 \leq 2 \text{diam} \sqrt{K d(x, x')} \leq 2\sqrt{K} d_{\max}$ (using $\text{diam} = \sqrt{d_{\max}}$), so $X \mapsto \varphi(X)$ has bounded ℓ^2 sensitivity.

PROPOSITION 44 (Private release of the embedding). *Releasing $\tilde{Z} = (\varphi(X) - \mu)W + N$ with N Gaussian of scale $\sigma \propto \sqrt{K} d_{\max} \|W\|_2 / \varepsilon$ satisfies record-level (ε, δ) -differential privacy, and by Prop. 1 the induced distortion of any pairwise embedding distance D_{ij} is $O(\sigma\sqrt{m})$ — a privacy/utility trade-off governed entirely by the sensitivity constant $2\sqrt{K} d_{\max}$.*

Proof. Take two repertoires differing in a single record, x_i versus x'_i . Only row i of $\varphi(X)$ changes, and by the Hölder- $\frac{1}{2}$ coordinates of Prop. 1, $\|\varphi(x_i) - \varphi(x'_i)\|_2 \leq 2 \text{diam} \sqrt{K d(x_i, x'_i)} \leq 2\sqrt{K} d_{\max}$.

Composing with W , the released row moves by $\|(\varphi(x_i) - \varphi(x'_i))W\|_2 \leq 2\sqrt{K}d_{\max}\|W\|_2 =: \Delta_2$, the record-level ℓ^2 -sensitivity. The Gaussian mechanism, adding $N \sim \mathcal{N}(0, \sigma^2 I)$, is a standard (ϵ, δ) -differentially private release whenever $\sigma \geq \Delta_2\sqrt{2\ln(1.25/\delta)}/\epsilon$, i.e. for $\sigma \propto \sqrt{K}d_{\max}\|W\|_2/\epsilon$. For utility, the perturbed distance $\tilde{D}_{ij} = \|(z_i - z_j) + (N_i - N_j)\|$ obeys $|\tilde{D}_{ij} - D_{ij}| \leq \|N_i - N_j\|$ by the triangle inequality, and $N_i - N_j \sim \mathcal{N}(0, 2\sigma^2 I_m)$ has $\mathbb{E}\|N_i - N_j\|^2 = 2m\sigma^2$, so the distortion is $O(\sigma\sqrt{m})$. $\square$

RECOMMENDATION. For patient data, treat (P, W) as a managed secret, release only the truncated projection Z (never φ or the invertible residual), and — when a provable guarantee is required rather than obscurity — add the calibrated noise of Prop. 44, whose cost is a quantifiable, Lipschitz-bounded distortion of the very distances the analysis depends on.

T.11. INSIGHTS, REPRODUCTION, AND OPEN PROBLEMS

A UNIFYING VIEW. The results above place TCREMP in a well-understood corner of metric geometry. The map φ is the empirical (Nyström) feature map of the alignment kernel; its ℓ^∞ read-out is the Kuratowski isometry, its ℓ^2 read-out is Landmark MDS, and its PCA is the classical scaling of the landmark columns. Three observations may be of independent interest beyond immunology: (a) taking the landmark measure to be a *generative model* of the data (P_{gen}) rather than a random subsample makes the embedding a sketch of each point’s relation to the data-generating process, and makes prototype choice provably robust (Prop. 7); (b) high-dimensional prototype-embedding distances obey a generalized extreme value law (Prop. 8), a clean consequence of distance concentration that is easy to mistake for Gaussian; and (c) the germline rank bound (Prop. 12) shows how categorical side-information injects exactly-low-rank structure into an otherwise full-rank distance embedding.

REPRODUCTION. Every quantitative statement is regenerated by `mir/bench/theory.py` and the data/figure scripts in this directory (`gen_theory_data.py`, the `*.gp` gnuplot scripts, `diagram_embedding.dot`); Table 4 collects the numbers.

Table 4: Theory predictions and their `v3` reproduction (`mir/bench/theory.py`; $n = 2500 \text{ CDR3}\beta$). Paper values from [1], suppl. S1–S3.

Claim	Paper	mirpy (Smith–Waterman)	mirpy (<code>gapblock v3</code>)
S2 / Prop. 3: $\text{corr}(D_{ij}, d_{ij})$	0.56	0.64	0.49
S1 / Prop. 8: law of d_{ij}	Gamma $>$ Normal	Gamma $\approx$ Normal	Gamma $\approx$ Normal
S1 / Prop. 8: law of D_{ij}	GEV, $\xi = +0.11$	GEV $\gg$ Normal, $\xi \approx 0$	GEV $\gg$ Normal, $\xi \approx 0$
S3 / Prop. 7: real vs. model prototypes	0.96	—	0.97

OPEN PROBLEMS. Four questions are left open. (1) Sharp distortion constants for the lower bound complementary to Prop. 1 when prototypes are P_{gen} -sampled, i.e. a Bourgain-type [91] guarantee tuned to the recombination measure rather than to the worst case. (2) The exact extreme-value index ξ of Prop. 8 as a function of the substitution matrix and prototype count. (3) A mutation-corrected metric for IGH that makes Prop. 7 hold under SHM (§T.5). (4) Consistency and rates

for the three density-ratio estimators of §T.6 (balloon, relative [31], and classifier-based [33]) on the PCA-projected embedding, together with a per-epitope detection limit in the sense of Prop. 26 as a function of precursor frequency and motif generation probability, and calibration of the neural classifier’s significance — which would put continuous background subtraction on the same footing as the graph tests it replaces.

REFERENCES

- [1] Yulia Kremlyakova, Elizaveta K. Vlasova, Daniil Luppov, and Mikhail Shugay. TCREMP: a bioinformatic pipeline for efficient embedding of T-cell receptor sequences from immune repertoire and single-cell sequencing data. *Journal of Molecular Biology*, 437(15):169205, 2025. PMID: 40368275.
- [2] Anand Murugan, Thierry Mora, Aleksandra M. Walczak, and Curtis G. Callan. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences*, 109(40):16161–16166, 2012.
- [3] Zachary Sethna, Yuval Elhanati, Curtis G. Callan, Aleksandra M. Walczak, and Thierry Mora. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics*, 35(17):2974–2981, 2019.
- [4] Isaac J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.
- [5] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.
- [6] Vin de Silva and Joshua B. Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, Stanford University, 2004.
- [7] Christos Faloutsos and King-Ip Lin. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, pages 163–174, 1995.
- [8] Elżbieta Pełkalska and Robert P. W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. World Scientific, 2005.
- [9] Ian Doust, Sergei Sánchez, and Anthony Weston. Negative type and bi-Lipschitz embeddings into Hilbert space. *arXiv preprint*, 2023. arXiv:2309.16070.
- [10] Conditionally strictly negative definite kernels. *arXiv preprint*, 2013. arXiv:1307.1778.
- [11] Canonicalization of the E-value from BLAST similarity search: a dissimilarity measure and distance function for a metric space of protein sequences. *arXiv preprint*, 2025. arXiv:2509.06849.

- [12] Quentin Marcou, Thierry Mora, and Aleksandra M. Walczak. High-throughput immune repertoire analysis with IGoR. *Nature Communications*, 9(1):561, 2018.
- [13] Olga V. Britanova, Mikhail Shugay, Ekaterina M. Merzlyak, et al. Dynamics of individual T cell repertoires: from cord blood to centenarians. *Journal of Immunology*, 196(12):5005–5013, 2016.
- [14] Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction*. Springer Series in Operations Research and Financial Engineering. Springer, 2006.
- [15] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is “nearest neighbor” meaningful? In *Proceedings of the 7th International Conference on Database Theory (ICDT)*, volume 1540 of *Lecture Notes in Computer Science*, pages 217–235, 1999.
- [16] Andreas Mayer, Vijay Balasubramanian, Thierry Mora, and Aleksandra M. Walczak. How a well-adapted immune system is organized. *Proceedings of the National Academy of Sciences*, 112(19):5950–5955, 2015. PMID: 25918407.
- [17] Mikhail Goncharov, Dmitry Bagaev, Dmitrii Shcherbinin, et al. VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2. *Nature Methods*, 19(9):1017–1019, 2022.
- [18] Jacob Glanville, Huang Huang, Allison Nau, Olivia Hatton, Lisa E. Wagar, Florian Rubelt, Xuhuai Ji, Arnold Han, Sheri M. Krams, Christina Pettus, Nikhil Haas, Cecilia S. Lindestam Arlehamn, Alessandro Sette, Scott D. Boyd, Thomas J. Scriba, Olivia M. Martinez, and Mark M. Davis. Identifying specificity groups in the T cell receptor repertoire. *Nature*, 547(7661):94–98, 2017. PMID: 28636589.
- [19] Gur Yaari, Jason A. Vander Heiden, Mohamed Uduman, et al. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Frontiers in Immunology*, 4:358, 2013. PMID: 24298272.
- [20] Paul-Gydeon Ritvo, Ahmed Saadawi, Pierre Barennes, Valentin Quiniou, Wahiba Chaara, Karim El Soufi, Benjamin Bonnet, Adrien Six, Mikhail Shugay, Encarnita Mariotti-Ferrandiz, and David Klatzmann. High-resolution repertoire analysis reveals a major bystander activation of Tfh and Tfr cells. *Proceedings of the National Academy of Sciences*, 115(38):9604–9609, 2018. PMID: 30158170.
- [21] Mikhail V. Pogorelyy, Anastasia A. Minervina, Mikhail Shugay, et al. Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLoS Biology*, 17(6):e3000314, 2019.
- [22] Mikhail V. Pogorelyy and Mikhail Shugay. A framework for annotation of antigen specificities in high-throughput T-cell repertoire sequencing studies. *Frontiers in Immunology*, 10:2159, 2019. PMID: 31616409.

- [23] Yuval Elhanati, Anand Murugan, Curtis G. Callan, Thierry Mora, and Aleksandra M. Walczak. Quantifying selection in immune receptor repertoires. *Proceedings of the National Academy of Sciences*, 111(27):9875–9880, 2014. PMID: 24941953.
- [24] Zachary Sethna, Giulio Isacchini, Thomas Dupic, Thierry Mora, Aleksandra M. Walczak, and Yuval Elhanati. Population variability in the generation and selection of T-cell repertoires. *PLoS Computational Biology*, 16(12):e1008394, 2020. PMID: 33296360.
- [25] Vanessa Venturi, Katherine Kedzierska, David A. Price, Peter C. Doherty, Daniel C. Douek, Stephen J. Turner, and Miles P. Davenport. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proceedings of the National Academy of Sciences*, 103(49):18691–18696, 2006. PMID: 17130450.
- [26] Máire F. Quigley, Hui Yee Greenaway, Vanessa Venturi, Ross Lindsay, Kylie M. Quinn, Robert A. Seder, Daniel C. Douek, Miles P. Davenport, and David A. Price. Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. *Proceedings of the National Academy of Sciences*, 107(45):19414–19419, 2010. PMID: 20974936.
- [27] Vanessa Venturi, David A. Price, Daniel C. Douek, and Miles P. Davenport. The molecular basis for public T-cell responses? *Nature Reviews Immunology*, 8(3):231–238, 2008. PMID: 18301425.
- [28] Mikhail V. Pogorelyy, Alla D. Fedorova, James E. McLaren, Kristin Ladell, Dmitri V. Bagaev, Alexey V. Eliseev, Artem I. Mikelov, Anna E. Koneva, Ivan V. Zvyagin, David A. Price, Dmitry M. Chudakov, and Mikhail Shugay. Exploring the pre-immune landscape of antigen-specific T cells. *Genome Medicine*, 10(1):68, 2018. PMID: 30144804.
- [29] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.
- [30] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.
- [31] Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25(5):1324–1370, 2013.
- [32] Aditya Krishna Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *PMLR*, pages 304–313, 2016.
- [33] Michael U. Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361, 2012.

- [34] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. arXiv:1912.02762.
- [35] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010. arXiv:0809.0853.
- [36] J. F. C. Kingman. *Poisson Processes*. Oxford Studies in Probability. Oxford University Press, 1993.
- [37] K. Krishnamoorthy and Jessica Thomson. A more powerful test for comparing two Poisson means. *Journal of Statistical Planning and Inference*, 119(1):23–35, 2004.
- [38] Richard Arratia, Larry Goldstein, and Louis Gordon. Poisson approximation and the Chen–Stein method. *Statistical Science*, 5(4):403–434, 1990.
- [39] Don O. Loftsgaarden and Charles P. Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.
- [40] George R. Terrell and David W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20(3):1236–1265, 1992.
- [41] Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer, 2015.
- [42] Bradley Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.
- [43] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010. PMID: 20979621.
- [44] 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(1):289–300, 1995.
- [45] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.
- [46] Jerome Cornfield. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute*, 11(6):1269–1275, 1951. PMID: 14861651.
- [47] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11:2973–3009, 2010.
- [48] Irving J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3–4):237–264, 1953.

- [49] Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Nonparametric ridge estimation. *The Annals of Statistics*, 42(4):1511–1545, 2014. arXiv:1212.5156.
- [50] Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, pages 343–351, 2010.
- [51] Alessandro Rinaldo and Larry Wasserman. Generalized density clustering. *The Annals of Statistics*, 38(5):2678–2722, 2010. arXiv:0907.3454.
- [52] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press, 1996.
- [53] Ivan V. Zvyagin, Mikhail V. Pogorelyy, Marina E. Ivanova, Ekaterina A. Komech, Mikhail Shugay, Dmitry A. Bolotin, Andrey A. Shelenkov, Alexey A. Kurnosov, Dmitriy B. Staroverov, Dmitriy M. Chudakov, Yuri B. Lebedev, and Ilgar Z. Mamedov. Distinctive properties of identical twins’ TCR repertoires revealed by high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 111(16):5980–5985, 2014.
- [54] Thomas Dupic, Meriem Bensouda Koraichi, Anastasia A. Minervina, Mikhail V. Pogorelyy, Thierry Mora, and Aleksandra M. Walczak. Immune fingerprinting through repertoire similarity. *PLoS Genetics*, 17(1):e1009301, 2021.
- [55] Olga V. Britanova, Ekaterina V. Putintseva, Mikhail Shugay, Ekaterina M. Merzlyak, Maria A. Turchaninova, Dmitriy B. Staroverov, Dmitriy A. Bolotin, Sergey Lukyanov, Ekaterina A. Bogdanova, Ilgar Z. Mamedov, Yuriy B. Lebedev, and Dmitriy M. Chudakov. Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *Journal of Immunology*, 192(6):2689–2698, 2014.
- [56] Ayelet Alpert, Yishai Pickman, Michael Leipold, Yael Rosenberg-Hasson, Xuhuai Ji, Renaud Gaujoux, Hadas Rabani, Elina Starosvetsky, Ksenya Kveler, Steven Schaffert, David Furman, Oren Caspi, Uri Rosenschein, Purvesh Khatri, Cornelia L. Dekker, Holden T. Maecker, Mark M. Davis, and Shai S. Shen-Orr. A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring. *Nature Medicine*, 25(3):487–495, 2019.
- [57] 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.
- [58] F. J. Anscombe. The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, 35(3-4):246–254, 1948.

- [59] Jonathan Desponds, Thierry Mora, and Aleksandra M. Walczak. Fluctuating fitness shapes the clone-size distribution of immune repertoires. *Proceedings of the National Academy of Sciences of the United States of America*, 113(2):274–279, 2016.
- [60] Hillary Koch, Dmytro Starenki, Sara J. Cooper, Richard M. Myers, and Qunhua Li. powerter: A model-based approach to comparative analysis of the clone size distribution of the T cell receptor repertoire. *PLOS Computational Biology*, 14(11):e1006571, 2018.
- [61] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, volume 30, pages 3391–3401. Curran Associates, Inc., 2017.
- [62] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, volume 20, pages 1177–1184. Curran Associates, Inc., 2007.
- [63] Alexander J. Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto, editors, *Algorithmic Learning Theory (ALT 2007)*, volume 4754 of *Lecture Notes in Computer Science*, pages 13–31. Springer Berlin Heidelberg, 2007.
- [64] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.
- [65] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.
- [66] Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016.
- [67] Anne Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11(4):265–270, 1984.
- [68] Joseph Kaplinsky and Ramy Arnaout. Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. *Nature Communications*, 7:11881, 2016.
- [69] M. O. Hill. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2):427–432, 1973.
- [70] Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.
- [71] Anne Chao and Lou Jost. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, 93(12):2533–2547, 2012.

- [72] T. C. Hsieh, K. H. Ma, and Anne Chao. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods in Ecology and Evolution*, 7(12):1451–1456, 2016.
- [73] Tom Leinster and Christina A. Cobbold. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489, 2012.
- [74] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, Minneapolis, MN, USA, 2007. IEEE.
- [75] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3304–3311, San Francisco, CA, USA, 2010. IEEE.
- [76] 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.
- [77] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 3744–3753. PMLR, 2019.
- [78] 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 Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.
- [79] Wei Wang, Dejan Slepcev, Saurav Basu, John A. Ozolek, and Gustavo K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International Journal of Computer Vision*, 101(2):254–269, 2013.
- [80] Soheil Kolouri, Navid Naderializadeh, Gustavo K. Rohde, and Heiko Hoffmann. Wasserstein embedding for graph learning. In *International Conference on Learning Representations (ICLR)*, 2021.
- [81] Aaron M. Newman, Chih Long Liu, Michael R. Green, Andrew J. Gentles, Weiguo Feng, Yue Xu, Chuong D. Hoang, Maximilian Diehn, and Ash A. Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5):453–457, 2015. PMID: 25822800.
- [82] Dmitriy A. Bolotin, Stanislav Poslavsky, Alexey N. Davydov, Felix E. Frenkel, Lorenzo Fanchi, Olga I. Zolotareva, Saskia Hemmers, Ekaterina V. Putintseva, Anna S. Obraztsova, Mikhail Shugay, Ravshan I. Ataullakhanov, Alexander Y. Rudensky, Ton N. Schumacher, and

- Dmitriy M. Chudakov. Antigen receptor repertoire profiling from RNA-seq data. *Nature Biotechnology*, 35(10):908–911, 2017. PMID: 29020005.
- [83] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007. PMID: 16632515.
- [84] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-Ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, 2019. PMID: 31740819.
- [85] Johann A. Gagnon-Bartsch and Terence P. Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012. PMID: 22101192.
- [86] Nathan Mantel and William Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–748, 1959. PMID: 13655060.
- [87] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.
- [88] Pradyot Dash, Andrew J. Fiore-Gartland, Tomer Hertz, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, 547(7661):89–93, 2017. PMID: 28636592.
- [89] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.
- [90] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.
- [91] Jean Bourgain. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel Journal of Mathematics*, 52(1–2):46–52, 1985.