

# A CONTROL-CALIBRATED E-VALUE FOR FUZZY TCR SEQUENCE SEARCH OVER BIOLOGICALLY REDUNDANT REFERENCE SETS

Mikhail Shugay<sup>1,2,3\*</sup>

<sup>1</sup>Institute of Translational Medicine, Russian National Medical State University, Moscow, Russia

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

<sup>3</sup>Institute of Molecular Biology NAS RA, Erevan, Armenia

*seqtree technical appendix*

## Abstract

We derive a BLAST-style E-value [10, 1] for “hits” returned by fuzzy search over T-cell receptor (TCR) CDR3 sequences, adapted to the defining difficulty of immune repertoires: the reference set is highly *redundant*, and the redundancy is biological (convergent V(D)J recombination, public clones, clonal expansion) rather than statistical noise. The classical Karlin–Altschul theory assumes a database of independent, identically distributed letters; under that null, redundancy-driven near-matches are absurdly significant. We replace the i.i.d.-letter null with an *empirical background*  $P_0$  estimated from a matched control repertoire, retain the Poisson/Gumbel limit superstructure with an explicit non-asymptotic error bound (Chen–Stein / Le Cam [6, 13]), and handle clonal over-dispersion by collapsing to unique clonotypes. The resulting E-value is automatically deflated for hits that the generation process alone explains and is large only for antigen-driven convergence. This puts the TCRNET approach—counting sequence neighbours against a real-world control repertoire, first introduced by Ritvo et al. [17] and formalized as an annotation framework by [16]—on a rigorous, finite-sample footing, and we show the classical Karlin–Altschul E-value is its product-measure, ungapped special case. The framework is not tied to TCRs: §12 extends it to presentation-aware E-values for peptide–MHC (pMHC) epitope homology, molecular mimicry, and allele guessing, while §14 reduces it to two elementary tasks that need only its combinatorial skeleton—unique-molecular-identifier (UMI) and barcode collapse, whose chance false-merge rate is a birthday-problem calculation, and tree clustering of CDR3 nucleotide reads to correct PCR and sequencing errors.

## 1. INTRODUCTION: THE REDUNDANCY PROBLEM

Given a query CDR3  $q$  and a target set  $D$  (e.g. VDJdb), fuzzy search returns the neighbours of  $q$  within a fixed scope/budget  $\theta$ . We want a significance value for such hits. BLAST answers this for protein search with the Karlin–Altschul E-value

$$E = K m n e^{-\lambda * S}, \tag{1}$$

---

\*Correspondence: mikhail.shugay@gmail.com

where  $m$  is the query length and  $n$  the total database length (both in residues),  $S$  is the alignment score of the hit under a substitution matrix with entries  $s_{ij}$ ,  $\lambda^*$  is the unique positive root of  $\sum_{ij} p_i p_j e^{\lambda^* s_{ij}} = 1$  (the natural scale that turns scores into log-probabilities for i.i.d. letters with background frequencies  $p_i$ ), and  $K > 0$  is a prefactor—the “clumping” or edge-effect constant—fixed by the score distribution.  $E$  is the expected number of distinct alignments scoring at least  $S$  by chance; the number of such alignments is asymptotically Poisson, so  $\Pr(\text{at least one}) = 1 - e^{-E}$  [10, 11]. The whole construction rests on the database being a string of i.i.d. letters.

Immune repertoires violate the i.i.d. assumption catastrophically. CDR3s are produced by V(D)J recombination, whose generation probability  $P_{\text{gen}}$ , first derived and inferred from sequence repertoires by Murugan et al. [15], is sharply non-uniform; convergent recombination makes some sequences enormously over-represented; clonal expansion and public clones create exact and near duplicates. A query in a common, high- $P_{\text{gen}}$  region of sequence space has many neighbours *for purely generative reasons*. An i.i.d. null would flag these as wildly significant, which is biologically meaningless. The signal we actually want is the opposite: *more* neighbours than the background generative process predicts, the hallmark of antigen-driven selection.

Our approach: define the null by an empirical background distribution  $P_0$  that carries the generative and baseline-sharing redundancy but no antigen-driven enrichment, estimate the null neighbourhood mass from a matched control repertoire (the user-supplied `isalgo/airr_control` set), and calibrate the E-value against it. This is a rigorous, finite-sample generalization of Karlin–Altschul to a non-i.i.d., biologically structured null, and the statistical formalization of TCRNET-style neighbour counting against a real control [17, 16].

## 2. SETUP AND NOTATION

Let  $\Sigma$  be the amino-acid alphabet and  $\mathcal{X} = \bigcup_{L \geq 0} \Sigma^L$  the space of CDR3 sequences. The search engine defines, for a query  $q$  and budget  $\theta \geq 0$ , a non-negative score  $s_\theta(q, x)$  and a *ball*

$$B_\theta(q) = \{x \in \mathcal{X} : s(q, x) \leq \theta\}, \quad s(q, q) = 0, \quad s \geq 0. \quad (2)$$

The score need not be a metric: with a substitution matrix it is the squared-distance penalty  $\text{pen}(a, b) = s_{aa} + s_{bb} - 2s_{ab}$  summed along whatever correspondence the scoring model admits (plus gap costs, §11); in unit-cost mode it is an edit count. Both define a legitimate ball. The batched search path enumerates that ball by branch-and-bound over a trie: it is *not* an aligner and returns nothing outside the budget. A full affine alignment of one specific pair is a separate, on-demand call. We work with two background laws on  $\mathcal{X}$ : the *realized-repertoire background*  $P_0$  (what a healthy, unselected repertoire instantiates) and the *generation law*  $P_{\text{gen}}$  (the V(D)J model). Let

$$\pi_0(q, \theta) = P_0(B_\theta(q)) = \sum_{x \in B_\theta(q)} P_0(x), \quad \pi_{\text{gen}}(q, \theta) = P_{\text{gen}}(B_\theta(q)). \quad (3)$$

A control sample  $C = (C_1, \dots, C_M)$  and a target set  $D$  are given. Crucially, all counts are over *distinct clonotypes*: the engine deduplicates hits by reference id, so

$$n_S(q, \theta) = \#\{x \in S \text{ distinct} : x \in B_\theta(q)\}, \quad S \in \{C, D\}. \quad (4)$$

Write  $N = |D|$  (unique clonotypes; see §5).

LEMMA 1 (Scope monotonicity). *The balls nest,  $B_\theta(q) \subseteq B_{\theta'}(q)$  for  $\theta \leq \theta'$  (since  $s \geq 0$  and the cut is by a single threshold). Hence  $\pi_0(q, \cdot)$ ,  $n_S(q, \cdot)$ , the intensity  $\lambda(q, \cdot)$  and the E-value  $E(q, \cdot)$  are all non-decreasing in the scope/budget  $\theta$ , and the closest-hit score  $S_{\min}(q)$  of §9 is the smallest  $\theta$  with  $n_D(q, \theta) > 0$ . This justifies sweeping  $\theta$  to trace an E-value curve per query.*

ASSUMPTION 1 (Exchangeability under  $H_0$ ). Under the null, the unique clonotypes of  $D$  are exchangeable with marginal law  $P_0$ .

ASSUMPTION 2 (Independent control draws). The unique clonotypes of  $C$  are i.i.d. (or exchangeable)  $\sim P_0$ .

ASSUMPTION 3 (Background match).  $C$  and  $D$  share the background  $P_0$  (same generation + sampling process, matched chain, species and length composition). All validity is conditional on Assumption 3.

### 3. NULL HYPOTHESIS AND ESTIMATOR HIERARCHY

DEFINITION 1 (Per-query null).  $H_0(q)$ : the neighbours of  $q$  in  $D$  arise from  $P_0$  with no antigen-driven excess, i.e. each  $x \in D$  satisfies  $\mathbb{E}[\mathbf{1}(x \in B_\theta(q))] = \pi_0(q, \theta)$ . The alternative  $H_1(q)$  posits excess mass  $\pi_D(q, \theta) > \pi_0(q, \theta)$ .

LEMMA 2 (Self-match exclusion / punctured null). *When the query is itself a database member ( $q \in D$ , as in a VDJdb-vs-VDJdb scan), the count  $n_D(q, \theta)$  contains the exact self-match (and any exact duplicates of  $q$ ), which are deterministic identity hits, not random draws from  $P_0$ . Including them biases both the observed count and the null. The correct neighbour statistic is the punctured count over the distance-positive ball,*

$$n_D^>(q, \theta) = \#\{x \in D : 0 < s(q, x) \leq \theta\}, \quad (5)$$

*with null intensity  $\lambda^>(q, \theta) = (N - m_q) \pi_0^>(q, \theta)$ , where  $m_q$  is the multiplicity of  $q$  in  $D$  and  $\pi_0^> = P_0(B_\theta(q)) - P_0(\{x : s(q, x) = 0\})$  removes the point mass at exact matches. The control estimator is punctured identically,  $\hat{\pi}^> = n_C^>(q, \theta)/M$ , so the deterministic identity term cancels in the calibrated E-value. (For  $q \notin D$  the puncture is vacuous and  $n_D^> = n_D$ .)*

REMARK 1 (Consistency of the puncture, and when *not* to use it). The puncture is valid *only if applied to both sides*: the E-value  $E = (N/M) n_C$  estimates  $N\pi_0$  for one and the same ball, so dropping the  $s = 0$  point mass from the target count requires estimating the punctured mass  $\pi_0^>$  from the punctured control count  $n_C^>$ . Doing so does change the numeric E-value (it shrinks by the removed exact-match mass,  $E^> = (N/M)n_C^> \leq E$ ), but it leaves the *inference* unbiased: the exact-match term is deterministic and enters observed count and null intensity identically, so it carries no signal and its removal neither creates nor destroys significance for the genuine neighbours. Puncturing only one side (target but not control, or vice versa) *does* bias the test and must be avoided.

This exclusion is a **benchmark device**, not a default for applications. In the VDJdb-vs-VDJdb benchmark the queries are drawn from the target, so every query carries a guaranteed trivial self-hit that would otherwise inflate every count uniformly; puncturing removes it. In a real annotation task the query is a *novel* sequence scored against a reference database ( $q \notin D$ ), where an exact database match is the strongest and most informative hit and must be kept. Hence `seqtree.values` leaves `exclude_exact=False` by default and the benchmark sets it `True`.

The estimand is the per-query Poisson intensity  $\lambda(q, \theta) = N \pi_0(q, \theta)$  (read as  $\lambda^\triangleright$  with the puncture of Lemma 2 whenever  $q \in D$ ). Two estimators of  $\pi_0$  target *different* nulls and must not be conflated.

- **Control / Monte-Carlo (primary):**  $\hat{\pi}(q, \theta) = n_C(q, \theta)/M$ , unbiased for  $P_0(B_\theta(q))$ , with  $M\hat{\pi} \sim \text{Binomial}(M, \pi_0)$  under Assumption 2. It captures the *realized* background, including public-clone sharing and finite-repertoire convergence.
- **Generation / analytic (cross-check):**  $\hat{\pi}_{\text{gen}}(q, \theta) = \sum_{x \in B_\theta(q)} P_{\text{gen}}(x)$ , computed by enumerating the (small, for small  $\theta$ ) ball with the engine and weighting by the V(D)J generation probability of the Murugan et al. model [15]. It targets the pure generation null  $P_{\text{gen}}(B_\theta(q))$ , which omits selection and sampling.

REMARK 2 (Selection factor and the thymic correction).  $P_{\text{gen}}$  is a *pre-selection* law; only a fraction of generated receptors survive thymic and peripheral selection. Elhanati et al. [9] model this with a per-sequence *selection factor*  $Q(\sigma) \geq 0$  on the recombination outcome  $\sigma = (\vec{a}, V, J)$ , inferred by maximum likelihood, giving the post-selection law

$$P_0(\sigma) = \frac{1}{Z} Q(\sigma) P_{\text{gen}}(\sigma), \quad Z = \sum_{\sigma} Q(\sigma) P_{\text{gen}}(\sigma) = 1 \quad (\langle Q \rangle_{P_{\text{gen}}} = 1). \quad (6)$$

The normalization  $\langle Q \rangle = 1$  means  $Q$  *redistributes* mass without a global rescale; separately, the physical thymic acceptance fraction  $\alpha \lesssim 15\%$  is sequence-independent, and selection cuts diversity by  $\approx 6$  bits [9]. Two consequences for the E-value, and they are what let us ignore both quantities. (i) The empirical control  $P_0$  already *is* the post-selection law of (6), so  $\hat{\pi}$  needs no  $Q$  and no  $\alpha$ ;  $Q$  enters only the analytic estimator, where one substitutes  $Q P_{\text{gen}}$  for  $P_{\text{gen}}$ . (ii) Being sequence-independent,  $\alpha$  *cancels* in every ratio and in  $\hat{\pi}$ , which calibrates against the control's own size  $M$ . It would matter only for an *absolute* naive-frequency estimate  $f(\sigma) = \alpha Q(\sigma) P_{\text{gen}}(\sigma)$  — e.g. when the  $\hat{\pi}_{\text{gen}}$  fallback for a rare query (§7) is read as an expected cell count rather than a probability. The repertoire-specific calibration of  $Q$  is a `vdjmatch` concern; here we need only that it exists and that the control absorbs it.

LEMMA 3 (The two nulls differ). *In general  $P_0 \neq P_{\text{gen}}$ : thymic and peripheral selection deplete some motifs while finite-sample public-clone sharing enriches others, so neither  $\pi_0 \leq \pi_{\text{gen}}$  nor the reverse holds universally. Hence  $\hat{\pi}_{\text{gen}}$  is used as a variance-reducing control variate and as a fallback for queries too rare for the control (§7), not as a substitute for  $\hat{\pi}$ .*

#### 4. POISSON APPROXIMATION WITH AN EXPLICIT ERROR BOUND

Fix  $q, \theta$ . For the unique clonotypes  $x_1, \dots, x_N$  of  $D$  set  $X_i = \mathbb{1}(x_i \in B_\theta(q))$ ,  $p_i = \mathbb{E}X_i = \pi_0$ ,  $W = \sum_i X_i$ ,  $\lambda = \sum_i p_i = N\pi_0$ , and let  $Z \sim \text{Poisson}(\lambda)$ . We use the following standard objects.  $\mathcal{L}(W)$  denotes the *law* (probability distribution) of  $W$ . The *total-variation distance* between two laws  $\mu, \nu$  on  $\mathbb{Z}_{\geq 0}$  is

$$d_{\text{TV}}(\mu, \nu) = \sup_{A \subseteq \mathbb{Z}_{\geq 0}} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{k \geq 0} |\mu(k) - \nu(k)|, \quad (7)$$

so a bound on  $d_{\text{TV}}$  bounds the error of *every* event probability simultaneously. A family of *dependency neighbourhoods* is a choice, for each index  $i$ , of a set  $B_i \ni i$  such that  $X_i$  is independent of (or nearly independent of)  $\{X_j : j \notin B_i\}$ ; intuitively  $B_i$  collects the clonotypes whose ball-membership is statistically coupled to  $x_i$ 's (here, those sharing a motif). The residual  $b_3$  below measures exactly how far that near-independence falls short.

**THEOREM 1** (Chen–Stein bound [5, 6]). *For any dependency neighbourhoods  $\{B_i \ni i\}$ ,*

$$d_{\text{TV}}(\mathcal{L}(W), \text{Poisson}(\lambda)) \leq b_1 + b_2 + b_3, \quad (8)$$

with  $b_1 = \sum_i \sum_{j \in B_i} p_i p_j$ ,  $b_2 = \sum_i \sum_{j \in B_i, j \neq i} \mathbb{E}[X_i X_j]$ , and  $b_3 = \sum_i \mathbb{E}|\mathbb{E}[X_i - p_i \mid \sigma(X_j : j \notin B_i)]|$ .

**COROLLARY 1** (Le Cam regime [13]). *If the collapsed clonotypes are independent under  $H_0$ , take  $B_i = \{i\}$ ; then  $b_2 = b_3 = 0$  and*

$$d_{\text{TV}}(\mathcal{L}(W), \text{Poisson}(\lambda)) \leq \sum_i p_i^2 = N\pi_0^2 = \lambda\pi_0. \quad (9)$$

*The bound is small precisely in the regime of interest: a rare ball  $\pi_0 \ll 1$  with moderate  $\lambda$  gives error  $\leq \lambda\pi_0 \rightarrow 0$ .*

**COROLLARY 2** (Void and tail probabilities). *With  $w = n_D(q, \theta)$  observed,*

$$p_{\text{any}}(q, \theta) = \mathbb{P}(W \geq 1) = 1 - e^{-\lambda} + O(N\pi_0^2), \quad (10)$$

$$p(q, \theta) = \mathbb{P}(Z \geq w) = 1 - \sum_{k < w} \frac{e^{-\lambda} \lambda^k}{k!}, \quad |\mathbb{P}(W \geq w) - p(q, \theta)| \leq b_1 + b_2 + b_3. \quad (11)$$

*Both follow from Theorem 1: the void probability is the event  $A = \{0\}$  and the tail is  $A = \{w, w + 1, \dots\}$ , and by (7) the error on any single event is at most  $d_{\text{TV}} \leq b_1 + b_2 + b_3$  (so  $O(N\pi_0^2)$  in the independent regime, Corollary 1).*

**REMARK 3** (Where biology enters). Convergent recombination makes the dependency neighbourhood  $B_i$  nontrivial:  $j \in B_i$  when  $x_i$  and  $x_j$  share a high- $P_{\text{gen}}$  motif. The term  $b_2 = \sum_i \sum_{j \in B_i} \mathbb{E}[X_i X_j]$  is the excess *pairwise* ball co-occupancy. Under  $H_0$  it is controlled by the pairwise ball mass, estimable from the control by counting *pairs* of control sequences both in  $B_\theta(q)$ ; the Poisson regime holds when this estimate is  $\ll \lambda$ . The very same  $b_2$  is inflated under  $H_1$  (antigen-driven clusters co-occupy the ball), so it is simultaneously the null error term and the quantity carrying the signal.

## 5. CLONAL REDUNDANCY AND OVER-DISPERSION

**PROPOSITION 1** (Collapsing restores Poisson). *Let the raw target carry clonotypes with multiplicities (clone sizes)  $m_x$ . Counting reads/cells in the ball gives a compound-Poisson total  $T = \sum_{k=1}^K m_k$  with  $K \sim \text{Poisson}(\lambda)$  and  $m_k$  i.i.d.  $\sim G$ , so  $\mathbb{E}T = \lambda\mu_G$  and  $\text{Var } T = \lambda\mathbb{E}[m^2]$ , with over-dispersion index  $\text{Var } T / \mathbb{E}T = \mathbb{E}[m^2] / \mu_G \geq 1$ . Collapsing to unique clonotypes is the projection  $G \equiv 1$ , which removes multiplicity-driven over-dispersion and returns the Poisson count  $W$  of §4. We therefore deduplicate  $C$  and  $D$  to unique clonotypes by default.*

PROPOSITION 2 (Negative-binomial robustness check). *If multiplicities must be modelled (e.g. read-level tests) and  $G$  is geometric,  $T$  is negative-binomial; report the NB tail  $\mathbb{P}(\text{NB} \geq w)$  with mean  $\lambda\mu_G$  and dispersion estimated from observed clone sizes. Under  $H_1$  antigen-driven clones are stochastically larger, so power is retained; the collapsed Poisson test remains the assumption-light default.*

PROPOSITION 3 (tf-idf is self-information weighting). *Weighting each target hit  $x$  by its background self-information  $w(x) = -\log P_0(\{x\}\text{-ball})$  makes the expected per-hit contribution constant under  $H_0$ ; the inverse-document-frequency weight is exactly the inverse background ball mass, and the term frequency is the clone multiplicity. In the rare regime the control-set  $E$ -value satisfies  $E \approx e^{-\sum_x \text{idf}(x)}$ , so the “control-set” and “tf-idf” approaches to redundancy are one object.*

## 6. THE E-VALUE AND MULTIPLE TESTING

DEFINITION 2 (E-value). For a query family  $\mathcal{Q}$ , the expected number of background hits is

$$E_{\text{tot}}(\theta) = \mathbb{E}_{H_0}[\#\text{hits}] = \sum_{q \in \mathcal{Q}} N \pi_0(q, \theta) = \sum_{q \in \mathcal{Q}} \lambda(q, \theta). \quad (12)$$

The per-query specialization  $\mathcal{Q} = \{q\}$  gives the BLAST-convention E-value  $E(q, \theta) = N\pi_0(q, \theta)$ , estimated by

$$\widehat{E}(q, \theta) = \frac{N}{M} n_C(q, \theta), \quad p_{\text{any}} = 1 - e^{-\widehat{E}}. \quad (13)$$

PROPOSITION 4 (Assumption-free expectation). *Equation (12) holds by linearity of expectation regardless of any dependence among hits (clonal, convergent, or across  $\mathcal{Q}$ ). Consequently  $\mathbb{P}(\#\text{false hits} \geq 1) \leq E_{\text{tot}}$  by Markov’s inequality, and  $E_{\text{tot}}$  bounds the expected number of false discoveries. This robustness—no independence needed for the mean—is why the E-value, not the Poisson tail, is the primary report.*

PROPOSITION 5 (Family-wise and false-discovery control). *Two thresholding regimes for a family of  $|\mathcal{Q}|$  tested queries:*

1. E-value / Bonferroni (FWER). *Reporting every query with  $\widehat{E}(q, \theta) \leq \alpha/|\mathcal{Q}|$  controls the family-wise error rate at level  $\alpha$ : by Proposition 4 the expected number of false positives is  $\sum_q \widehat{E} \leq \alpha$ , and  $\mathbb{P}(\geq 1 \text{ false positive}) \leq \alpha$  by Markov. No independence is required. A fixed E-value cutoff (e.g.  $\widehat{E} \leq 1$ , the BLAST default) is the  $\alpha = |\mathcal{Q}|$  case and bounds the expected count of false positives by 1.*
2. p-value / Benjamini–Hochberg (FDR). *Using the per-query enrichment p-values  $p(q, \theta) = \mathbb{P}(Z \geq n_D^>(q, \theta))$  from Corollary 2, the Benjamini–Hochberg procedure [7]—sort  $p_{(1)} \leq \dots \leq p_{(|\mathcal{Q}|)}$ , reject the  $k$  largest with  $p_{(k)} \leq \frac{k}{|\mathcal{Q}|}\alpha$ —controls the false discovery rate at  $\alpha$  under positive dependence of the test statistics, the relevant regime here (convergent clusters induce positive correlation).*

REMARK 4 (The *within*-query seed family). Proposition 5 corrects multiplicity *across* queries. A second family arises *within* a single query when significance is read off shared  $k$ -mer seeds rather than off the ball: a junction of length  $L$  carries  $L - 2F - k + 1$  overlapping central seeds  $w_1, \dots, w_s$  (flank

width  $F$ ), and “ $q$  shares *some*  $w_j$  with  $x$ ” is a union of  $s$  dependent events. For one *pre-specified* seed the estimator is the ordinary (13) with the ball replaced by the seed’s occurrence set,

$$\widehat{E}_{\text{seed}}(w) = \frac{N}{M} n_C(w), \quad n_C(w) = \#\{x \in C : w \subseteq x\}, \quad (14)$$

and no conditioning correction is needed: under  $H_0$  the target draw  $x$  is independent of  $q$ , so  $P_0(w \subseteq x \mid w \subseteq q) = P_0(w \subseteq x)$  — the query’s content is fixed, not a second random draw, and the probability must *not* be squared.

For the family, do not correct: **count the union**. The event  $\bigcup_j \{w_j \subseteq x\}$  is a single measurable set, so  $\widehat{E}_{\bigcup} = (N/M) \#\{x \in C : \exists j, w_j \subseteq x\}$  is exact, inherits Propositions 4 and 10 verbatim, and absorbs the seed dependence with no constant. The Boole bound  $\widehat{E}_{\bigcup} \leq \sum_j \widehat{E}_{\text{seed}}(w_j)$  is admissible and mildly conservative: empirically it is  $\approx 16\%$  loose at the median and  $36\%$  in the mean at  $k = 5$ , because overlapping string seeds gather largely disjoint background sets (19.9% of gathered control sequences share  $\geq 2$  of a query’s seeds). Prefer the exact union count when it is available. Implemented in `seqtree.seeds (SeedIndex.evaluate, union_value)`.

Two cautions. (i)  $n_C(w)$  counts control *sequences containing*  $w$ , not occurrences. (ii)  $P_0(w)$  cannot be substituted by a per-position or low-order Markov background: the residual  $\text{KL}(P_{\text{data}} \parallel \hat{P})$  over central  $k$ -mers of the human TRB control is 0.43/1.60/4.48 bits at  $k = 4/5/6$  even after a second-order fit, and *grows* with  $k$ , because the templated D segment injects long germline runs. Seed significance must be counted.

**REMARK 5** (Seed weighting versus ball weighting). Proposition 3 already places self-information at the level of the whole ball. Seed self-information —  $\log P_0(w)$  is the same object restricted to a sub-feature, so the two must not be applied on top of one another. In particular, a per-position weight profile fed to a `PositionalMatrix` shapes the ball’s *geometry* (which substitutions to tolerate where) and carries no evidence at all: the engine’s score is a non-negative penalty with  $\text{pen}(a, a) = 0$ , so a positional weight multiplies mismatch cost only and a *match* — however surprising — contributes exactly zero. All statistical evidence stays in the control-calibrated  $\widehat{E}$  of (13) and (14).

**PROPOSITION 6** (Detectability / minimum cluster size). *Under  $H_1(q)$  let the antigen-driven excess add  $k$  neighbours beyond the background mean  $\lambda = \lambda^>(q, \theta)$ , so  $n_D^> \approx \lambda + k$ . The enrichment test at  $E$ -value cutoff  $\widehat{E} \leq \alpha$  rejects when  $n_D^> \geq w_\alpha$ , the smallest  $w$  with  $\mathbb{P}(\text{Poisson}(\lambda) \geq w) \leq \alpha$ . For small  $\lambda$  (the typical rare-ball regime),  $w_\alpha$  grows only logarithmically,  $w_\alpha \approx \frac{\log(1/\alpha)}{\log \log(1/\alpha) - \log \lambda}$  by the Poisson right tail, so a cluster of a handful of convergent neighbours is already detectable; for moderate  $\lambda$  the Gaussian approximation gives the familiar  $k \gtrsim z_{1-\alpha} \sqrt{\lambda}$ . The control size enters only through the resolution of  $\hat{\lambda}$  (§7):  $M$  must be large enough that the sampling noise of  $\widehat{E}$  is below the excess  $k$  being claimed.*

## 6.1. EPITOPE DETECTION COMPLEXITY

Proposition 6 concerns one query; in practice one samples a depth- $n$  repertoire and asks how much of an epitope-specific response is recoverable. Let an epitope’s TCR repertoire  $R_e$  have  $K$  unique

clonotypes and within-set scope- $\theta$  neighbour graph with degree distribution  $\{d_x\}_{x \in R_e}$  and neighbour density  $\rho = \frac{1}{K(K-1)} \sum_x d_x = \bar{d}/(K-1)$  (the probability that two random members of  $R_e$  are within  $\theta$ ).

**PROPOSITION 7** (Detection curve from the degree law). *Draw  $n$  clonotypes i.i.d. from  $R_e$ . A node of full-set degree  $d_x$  retains in expectation  $d_x(n-1)/(K-1)$  of its neighbours (hypergeometric sampling). Against the near-empty background ball ( $\lambda \approx 0$ , so  $w_\alpha$  is  $O(1)$ , Proposition 6), the node is detected once this exceeds a level  $d_{\min}(\alpha) = O(1)$ , i.e. at sampling depth*

$$n_x^* \approx 1 + d_{\min} \frac{K-1}{d_x}, \quad (15)$$

and the detectable fraction at depth  $n$  is

$$\varphi(n) = \frac{1}{K} \#\{x : d_x \geq d_{\min} \frac{K-1}{n-1}\}, \quad (16)$$

fixed entirely by the degree law. Equivalently the expected number of within-sample neighbour pairs is  $\binom{n}{2}\rho$ , so the first significant pair appears near  $n \approx \sqrt{2/\rho}$ . The detection complexity of  $R_e$ —the depth to recover a target fraction of the response—is therefore set by the upper tail of  $\{d_x\}$  (equivalently by  $\rho$  and the largest cluster): a repertoire dominated by one large convergent cluster is detected at small  $n$ , a diverse repertoire of many near-singletons requires deep sampling.

**REMARK 6** (Worked example: A\*02 NLV vs GIL). Measured on VDJdb TRB / HLA-A\*02 repertoires against a  $10^6$ -sequence OLGA background at scope  $\theta = 1$  substitution: **GIL** (GILGFVFTL, influenza M1;  $K = 5236$ ) has  $\rho = 3.4 \times 10^{-4}$ , max degree 52, and one dominant component of 896 (17% of the set); **NLV** (NLVPMVATV, CMV pp65;  $K = 13044$ ) has  $\rho = 2.8 \times 10^{-5}$  ( $\approx 12 \times$  sparser), max degree 22, and a largest component of only 152 (1.2%). Equation (16) then predicts—and the subsampled Benjamini–Hochberg significant fraction confirms (Fig. 2, bench/bench\_epitope.py)—that GIL is  $\sim 20$ – $30\%$  recovered by  $n \sim 10^3$  sampled TCRs while NLV stays below 5% even at  $n \sim 5 \times 10^3$ . The two epitopes have detection complexities differing by more than an order of magnitude purely from repertoire structure, with no change to the search or the background.

## 7. HOW LARGE MUST THE CONTROL BE?

Since  $M\hat{\pi} \sim \text{Binomial}(M, \pi_0)$ ,  $\text{Var } \hat{\pi} = \pi_0(1-\pi_0)/M$  and the relative error is  $\text{CV}(\hat{\pi}) \approx (M\pi_0)^{-1/2}$ .

**PROPOSITION 8** (Resolution). *To resolve a target  $E$ -value  $E^* = N\pi_0$  to relative error  $\rho$  requires*

$$M \gtrsim \frac{1}{\rho^2 \pi_0} = \frac{N}{\rho^2 E^*}. \quad (17)$$

Resolving  $E^* \sim 1$  to 10% thus needs  $M \sim 100 N$ .

**PROPOSITION 9** (Empty-ball regime). *If  $n_C = 0$ , the point estimate  $\hat{\pi} = 0$  is degenerate; use the rule of three  $\pi_0 \lesssim 3/M$  (95%), or a Beta( $n_C + a, M - n_C + b$ ) posterior, which propagates control uncertainty into a Poisson–Gamma (negative-binomial) posterior-predictive  $p$ -value for  $n_D$ . The implementation reports the rule-of-three upper bound  $\hat{E} \leq 3N/M$  when  $n_C = 0$ .*

When  $M$  is inadequate for a rare  $q$ , the analytic  $\hat{\pi}_{\text{gen}}$  is exact per query for any  $M$  and serves as the fallback (Lemma 3).

REMARK 7 (Inverting the E-value: a cutoff is a per-query object). Scope monotonicity (Lemma 1) makes  $n_C(q, \theta)$  non-decreasing in  $\theta$ , so (13) inverts. Because engine scores are integers the inversion is exact rather than a root-find: sort the control-hit scores of  $q$  and set

$$\theta^*(q) = \max\{\theta \leq \theta_{\max} : \widehat{E}(q, \theta) \leq E^*\} = \min(\theta_{\max}, s_{(k+1)}(q) - 1), \quad k = \lfloor E^*M/N \rfloor, \quad (18)$$

with  $s_{(j)}(q)$  the  $j$ -th smallest control score and  $\theta^* = \theta_{\max}$  when fewer than  $k + 1$  control hits exist. One control scan at the ceiling  $\theta_{\max}$  supplies  $\theta^*$  for every query (`seqtree.threshold_for_evalue`).

Two consequences deserve emphasis. First, Proposition 9 makes  $E^* < 3N/M$  *unattainable*: with  $M$  control sequences even an empty ball certifies no better than  $3N/M$ , so the implementation returns  $-1$  rather than a cutoff it cannot honour. On the bundled  $M = 250,000$  control against an  $N = 19,246$  target this floor is  $E^* \geq 0.231$ ; certifying  $E^* = 0.01$  would require  $M \geq 3N/E^* \approx 5.8 \times 10^6$ .

Second, and less obviously, a *fixed* score cutoff is not a calibrated one.  $P_0$  is dense near germline and sparse among rare junctions, so a single  $\theta$  purchases a common query far more chance neighbours than a rare one, and (13) is precisely the correction. The effect is not academic: building a neighbour graph on human TRB by joining CDR3s at a fixed single-gap-block score  $\leq 60$ , 31.7% of size-matched *random control* junctions land in a connected component of size  $\geq 5$  — structure generated by the threshold, not by specificity. Re-drawing the same graph with edges required to satisfy  $s(a, b) \leq \min(\theta^*(a), \theta^*(b))$  at  $E^* = 0.05$  removes all of it: 2.334 edges per node for co-specific TCRs against 0.021 for the control, which forms no connected component of size three. The control arm’s realized edge rate landing on  $E^*$  is the check that the calibration is honest.

## 8. COMPOSITION AND LENGTH ARE HANDLED AUTOMATICALLY

PROPOSITION 10. *Because  $\pi_0(q, \theta) = P_0(B_\theta(q))$  is computed for the specific query  $q$ , the control estimator  $n_C(q, \theta)/M$  conditions on  $q$ ’s length and composition automatically: the same biases that make  $q$  common make  $n_C$  large. This is the finite-sample, composition-exact analogue of the Karlin–Altschul  $K$   $mn$  length normalization, which is needed precisely because the i.i.d. background is query-independent. The only caveat is statistical: rare  $q$  require adequate  $M$  (§7), else fall back to  $\hat{\pi}_{\text{gen}}$ .*

## 9. THE CLOSEST HIT: AN EXTREME-VALUE LAW

THEOREM 2 (From Poisson to Gumbel). *Let  $\lambda(q, t) = NP_0(B_t(q))$ . By the Poisson approximation applied at each radius,*

$$\mathbb{P}(S_{\min}(q) > t) \approx e^{-\lambda(q, t)} = e^{-NP_0(B_t(q))}. \quad (19)$$

*If  $\log P_0(B_t(q)) \approx a + \beta t$  (log-linear ball-mass growth, the generic regime), then the best score  $Y = -S_{\min}$ , centred at  $u_N = (\log N + a)/\beta$ , obeys  $\mathbb{P}(Y - u_N \leq y) \rightarrow \exp(-e^{-\beta y})$ , a Gumbel law with scale  $1/\beta$ . Here  $\beta$  is the empirical ball-mass log-slope (regress  $\log n_C(q, t)$  on  $t$ ), not the Karlin–Altschul  $\lambda^*$ . (For lattice scores the Gumbel carries the usual periodic correction.)*

## 10. RELATION TO KARLIN–ALTSCHUL

**THEOREM 3** (KA is the product-measure, ungapped case). *If  $P_0 = \otimes_{\ell} p$  is a product measure and  $s$  is the ungapped additive score, then  $P_0(B_{\theta}(q))$  factorizes and, by Cramér’s theorem,  $-\frac{1}{|q|} \log P_0(B_{\theta}(q)) \rightarrow \lambda^*$ , the Karlin–Altschul parameter solving  $\sum_{ij} p_i p_j e^{\lambda^* s_{ij}} = 1$ . The intensity  $\lambda(q, \theta) = NP_0(B_{\theta}(q))$  then reduces to  $E = K m n e^{-\lambda^* S}$  with  $K$  the Poisson-clumping constant, recovering [10, 11, 2].*

Thus the present framework generalizes Karlin–Altschul in three ways: (i) the product measure  $\otimes p$  is replaced by the empirical/generative background  $P_0$ ; (ii) gaps and matrix-weighted balls are admitted via the engine’s score; (iii) the asymptotic constants  $K, \lambda^*$  are replaced by a finite- $N$ , finite- $M$  non-asymptotic error bound (Theorem 1).

## 11. THE SCORE MODEL: ONE GAP BLOCK, PLACED BY A PRIOR

Point (ii) above is where the ball meets biology, and it is not free. A junction’s length varies because of V/J trimming and N-addition — *one* contiguous indel event, not a scatter of them. The engine therefore restricts the alignment underlying  $s(q, x)$  to a single gap block of length  $d = |m - n|$  placed at some  $i \in [0, \min(m, n)]$ , and takes the best such placement. The optimum is a prefix/suffix sum, hence  $O(\min(m, n))$  rather than  $O(mn)$ .

**THE RESTRICTION IS NEARLY FREE, AND IT IS PROTECTIVE..** Against a model-independent structural oracle — iterative superposition followed by an *unrestricted* affine dynamic program, free to open any number of gap blocks anywhere — the true residue correspondence between two crystal junctions is a single contiguous block in 95.2–100% of 3,049 pairs drawn from 199 unique sequences, for both chains and every  $d = 1, \dots, 4$ ; forcing one block costs no median C $\alpha$ -RMSD. On sequence pairs that are genuinely related (one indel plus zero to two substitutions) the single-block score equals the unrestricted affine optimum on 98.8% of pairs at a gap cost calibrated to the matrix scale. On *unrelated* pairs affine undercuts it every time, by a median of 106 penalty units: unrestricted gapping does not discover a better alignment there, it manufactures one. Since  $B_{\theta}(q)$  is defined by the score, a score that invents similarity inflates  $\pi_0$  and  $n_D$  alike, and the E-value cannot repair what the metric destroyed.

**GAP COSTS LIVE ON THE MATRIX SCALE..** With  $\text{pen}(a, a) = 0$  the Gram transform puts a typical BLOSUM62 mismatch at 14. A gap cost of order unity therefore makes gaps an order of magnitude cheaper than substitutions and every alignment degenerates to gaps. Costs must be  $O(\text{median mismatch})$ . Correspondingly the affine cost must be guarded at  $d = 0$ : the naive  $g_{\text{open}} + (d - 1)g_{\text{ext}}$  is *negative* there whenever  $g_{\text{open}} < g_{\text{ext}}$ , which would violate  $s \geq 0$  and with it (2), Lemma 1, and admissible trie pruning.

**A PRIOR, NOT THE SCORE, PLACES THE BLOCK..** A substitution score alone selects the structurally correct block position about as often as chance: a hard central pin agrees with the score-only choice on only 10.6% of pairs. The engine therefore admits a *gap prior*  $\Lambda(i, d, m)$  added to each candidate placement before the minimisation. It must satisfy exactly two conditions, and no others:

$$\Lambda \geq 0 \quad (\text{admissible pruning}), \quad \Lambda(\cdot, 0, \cdot) = 0 \quad (\text{so } s(q, q) = 0, \text{ hence (2)}). \quad (20)$$

Monotonicity in  $d$  is *not* required and does not hold for the central prior  $\Lambda = \lambda|i + d/2 - m/2|$ : growing a leading block drags its midpoint toward the centre, so the penalty falls. Empirically the block sits at the loop apex — Cys-offset 6 for both TRA and TRB — and does *not* drift with  $d$ , up to  $d = 4$ . The central prior recovers that position exactly 42.4% (TRA) and 30.1% (TRB) of the time. A rule placing the block instead in the germline-untemplated span, where the insertion generatively occurred, recovers it 0.4% and 19.8% of the time and is rejected: the site at which nucleotides were inserted and the site at which a backbone accommodates an extra residue are different places.

Constraining the placement is what buys precision. Compared at a *matched* false-positive rate — each rung given the cutoff (18) at which its own ball admits  $E^*$  chance neighbours, since a freer rung finds lower scores and a fixed budget would reward it for that — retrieval precision on the length-different fraction of VDJdb human TRB same-epitope pairs (2,000 queries,  $E^* = 0.1$ ) is 0.414 for a hard central pin and 0.336 for a central prior, against 0.176 when the score is free to choose among  $L + 1$  placements and 0.156 when it chooses among five plausible ones. Extra freedom to place the block does not help; it costs.

REMARK 8 (A placement rule is a column frame). Pairwise-optimal placement is not transitive. Embedding a set of unequal-length sequences into a common  $W$ -column frame yields a consistent column index — and hence a position weight matrix — if and only if the block start is constant in  $d$ . Under the central prior the start is  $\lfloor (W - d)/2 \rfloor$ , which drifts, and two shorter members of the frame are then related by *two* blocks rather than one. The unique transitive rule pins the block to a fixed column  $c$ : left-anchor the first  $c$  residues on the conserved Cys, right-anchor the remainder on the Phe. Any per-position weight profile fed to a positional matrix (Remark 5, §12) presupposes such a frame.

## 12. EPITOPES: PRESENTATION-AWARE E-VALUES AND MOLECULAR MIMICRY

For TCR CDR<sub>3</sub> the background is V(D)J generation; for epitopes (MHC-presented peptides) it is *presentation*, which is allele-conditional and constrains the anchor positions. We adapt the framework with  $P_0 := P_0^{\text{pep}}(\cdot | a)$  per allele  $a$ , anchor-masked.

This section states only what the general framework needs in order to specialize; the pMHC layer itself — pocket decomposition, cross-allele pseudosequence diffusion, structural validation, and the comparison against trained predictors — is developed in the companion `mhcmatch` appendix, which defers its derivations to this document. The two agree; we do not restate them here.

**ANCHOR MASKING..** A presented peptide factorizes  $\sigma = (\sigma_A, \sigma_T)$  into MHC-facing anchors  $A(a, L)$  (class I {P2, P $\Omega$ }; class II core {P1, P4, P6, P9}) and TCR-facing positions  $T$ . Presentation fixes  $\sigma_A$ ; recognition reads  $\sigma_T$  — Dolton et al. [8] exhibit one HLA-A02:01 TCR cross-reacting with three epitopes through a shared central motif, the anchors irrelevant. Masking  $\mu_a(\sigma) = \sigma_T$  gives the TCR-facing ball  $B_\theta^T(q) = \{\sigma : \sigma_A \in \text{motif}(a), s_T(q_T, \sigma_T) \leq \theta\}$ , with  $s_T$  zeroing the anchors via a per-position matrix (which presupposes a fixed frame, Remark 8). Significance must be computed on  $\mu_a$  alone: shared anchors are presentation, not recognition, and would inflate every within-allele hit.

**PER-ALLELE E-VALUE..** Null mass  $\pi_0^{\text{pep}}(q, \theta; a) = P_0^{\text{pep}}(B_\theta^T(q) \mid a)$ . With a per-allele anchor-masked presented control  $C_a$  of size  $M_a$  and a target presented peptidome of size  $N_a$ ,

$$\widehat{E}(q, \theta; a) = \frac{N_a}{M_a} n_{C_a}(q_T, \theta), \quad p_{\text{any}} = 1 - e^{-\widehat{E}}, \quad p_{\text{enrich}} = \mathbb{P}(\text{Poisson}(\widehat{E}) \geq n_D^T), \quad (21)$$

i.e. `evaluates` (Def. 2) on anchor-masked, allele-restricted indices. The analytic fallback for rare alleles uses the maximum-entropy presented distribution of Tiffeau-Mayer et al. [18],  $P_0^{\text{pep}} \propto \exp(\sum_i h_i(\sigma_i) + \sum_{i < j} J_{ij}(\sigma_i, \sigma_j))$ , conditioned on  $\sigma_A \in \text{motif}(a)$ ; its 2-point couplings make the co-occupancy term  $b_2$  (Thm 1) generically nonzero, so the E-value mean stays robust (Prop. 4) while the tail takes the  $b_2$  / negative-binomial correction (Prop. 2).

**TWO DIRECTIONS, ONE WEIGHTING..** Zeroing the anchors and peaking the weights on the central hotspot gives the *recognition* distance  $d_{\text{TCR}}$ ; a self or bacterial peptide is a candidate mimic of a neoantigen iff it is presented by a compatible allele, lies within  $\theta$ , and is significant against that allele’s presented background — which deflates motifs common in the peptidome and elevates surprisingly shared ones, connecting the cross-reactivity distance of [14] and the close-to-self shell of [18] to a calibrated E-value. Flipping the weights ( $w_i = 1$  on anchors) scores the binding motif instead and answers the reverse problem, peptide to restricting allele, by voting the alleles of a peptide’s anchor-signature neighbours. One subtlety is load-bearing and easy to get backwards: enrichment against background *penalises* the dominant true allele on a skewed panel, because its expected count is large. The vote fraction  $\hat{p}(a \mid q) = k_a/n$  is therefore the ranking statistic; the per-allele E-value  $E_a$  is the confidence gate, and the global  $E_{\text{glob}} = \sum_a E_a$  counts how many panel alleles would present the peptide by chance. Both directions, and the class-II *register trick* — commit to the single 9-mer window whose P1 is large-hydrophobic and whose P4/P6/P9 avoid Pro/Gly, rather than pooling windows or deconvolving by clustering [4, 3], which is what lifts class-II ROC-AUC into the 0.9 range — are specified and productionized in `mhcmatch`.

**TWO REFERENCE SETS..** `pmhc_data` ships in two tiers. The *full* set is every IEDB-positive epitope–allele assay; the *shortlist* keeps only epitope–allele pairs supported by  $\geq 2$  publications — a higher-confidence but smaller subset that suppresses single-report artefacts at the cost of thinner mouse and class-II panels. Table 1 gives the scope of both. The ROC/PR below are computed on the *full* set; the shortlist is the conservative alternative and is reported here for scope only.

		full (IEDB+)		shortlist ( $\geq 2$ refs)	
class	species	epitopes	alleles	epitopes	alleles
MHC-I	human	437,850	203	137,627	107
MHC-I	mouse	57,892	9	6,903	8
MHC-II	human	198,921	19	28,859	11
MHC-II	mouse	12,018	13	2,564	8

Table 1: Scope of the two `pmhc_data` tiers by class and species: unique epitopes and distinct restricting alleles. The benchmark guessing panel (Table 2) further restricts to standard-amino-acid epitopes and to alleles with  $\geq 100$  (human) /  $\geq 40$  (mouse) peptides.

**EMPIRICAL (bench/bench\_mhc\_guess.py)..** On the full `pmhc_data` set (VDJdb/IEDB), evaluated as a per-(peptide, allele) binary task — exactly how `vdjmatch` scores TCR-antigen specificity — and split by species, the vote-fraction ranking recovers the restricting allele with ROC-AUC  $\approx 0.90$ – $0.99$  and PR-AUC far above the prevalence baseline, for both classes, both species, and both dataset tiers (Table 2, Figs. 3–4; the ROC/PR *curves* are shown for the full set). The confidence E-value additionally separates real peptides from length-matched random “noise” (noise-rejection AUROC column). This works only with *presentation* (anchor) features: the TCR-facing homology used for cross-reactivity carries no allele information (AUROC  $\approx 0.5$ ), confirming that MHC restriction lives in the anchors, not the TCR-facing surface.

class	species	set	alleles	top-1 acc	ROC-AUC	PR-AUC (base)	noise AUROC
MHC-I	human	full	129	0.58	0.92	0.47 (0.012)	0.69
MHC-I	human	$\geq 2$ refs	82	0.72	0.97	0.71 (0.015)	0.64
MHC-I	mouse	full	7	0.69	0.90	0.71 (0.205)	0.62
MHC-I	mouse	$\geq 2$ refs	4	0.83	0.95	0.88 (0.265)	0.37
MHC-II	human	full	11	0.70	0.94	0.72 (0.100)	0.55
MHC-II	human	$\geq 2$ refs	8	0.57	0.91	0.64 (0.130)	0.54
MHC-II	mouse	full	9	0.89	0.98	0.92 (0.113)	0.39
MHC-II	mouse	$\geq 2$ refs	2	0.98	0.99	0.99 (0.500)	0.46

Table 2: Allele guessing from presentation-signature neighbours, per-(peptide, allele) ROC/PR, for both `pmhc_data` tiers (full and the  $\geq 2$ -publication shortlist). About 600 held-out query peptides per class $\times$ species group; the guessing panel keeps alleles with  $\geq 100$  (human) /  $\geq 40$  (mouse) peptides. Ranking by neighbour vote fraction gives ROC-AUC 0.90–0.99 and PR-AUC well above the prevalence baseline (in parentheses); the higher-confidence shortlist matches or exceeds the full set on ROC despite far smaller panels. The confidence E-value separates presented peptides from length-matched random noise (last column). Mouse class-II panels are tiny (full 9 alleles, shortlist 2) and dominated by I-A<sup>b</sup>, inflating both baseline and scores.

**NON-BINDERS AND CLASS-II PROMISCUITY..** The same per-allele E-value rejects non-binders at two granularities: a peptide binding *no* MHC has a large best-over-panel E-value (the noise-rejection column of Table 2), while a peptide that merely fails to bind a *specific a* has  $E_a > \alpha$ . Rank by vote fraction, threshold by  $E_a$ . Class-II binding is far more degenerate — an open groove and loosely specified pockets let one peptide be presented by many alleles — so allele assignment there is genuinely multi-label, which is how the benchmark scores it (positives = every observed restricting allele), why class-II PR baselines sit above class-I, and why a class-II non-binder filter must test “binds none of the panel” rather than “binds exactly one.”  $E_{\text{glob}}$  is the natural statistic.

### 12.1. IMPLEMENTATION AND LIMITATIONS

Implemented in `seqtree.pmhc` (anchor-masked TCR-facing k-mer homology via the C++ `KmerIndex` seed-and-gather; `find_mimics`; `assign_allele`) and `seqtree.pmhc_evalue`; the positive control is the recovery of the A02:01 trio of [8] as mutual TCR-facing homologs. Anchor positions are parametrized (presets per class, overridable); class-II homology is register-agnostic (all-window k-

mers), while the reverse allele-guessing problem uses the register trick. Alternative CDR<sub>3</sub> build strategies, e.g. the bilateral run-peeling decomposition of [12], plug into the same layer. *Limitation:* we do not model MHC-groove similarity across alleles — distinct alleles are distinct nulls, searched independently and never pooled; “compatible allele” is a user input, and validity is conditional on a faithful per-allele presented control. The `seqtree` code here is a *reference implementation and benchmark* of the E-value methodology; production allele guessing and non-binder filtering, with an optimized index, cross-allele pocket modelling and tuned thresholds, are built downstream in `vdjmatch` (TCR–antigen specificity) and `mhcmatch` (peptide–MHC), whose appendices specialize this one rather than repeat it.

### 13. PRACTICAL DEFAULTS AND ALGORITHM

1. Deduplicate  $C$  and  $D$  to unique clonotypes (Proposition 1).
2. Build a `seqtree` index of  $C$ ; for each query compute  $n_C$  and  $n_D$  at scope  $\theta$  via batched search. When the query may itself be in  $D$  or  $C$ , use the punctured counts  $n^>$  that drop distance-zero (exact/self) hits (Lemma 2).
3. Report  $\hat{E} = (N/M) n_C^>$  (Eq. (13)),  $p_{\text{any}} = 1 - e^{-\hat{E}}$ , and  $p_{\text{enrich}} = \mathbb{P}(\text{Poisson}(\hat{E}) \geq n_D^>)$ ; use the rule of three when  $n_C^> = 0$  (Proposition 9).
4. Across a query family, threshold on  $\hat{E}$  for FWER control or apply Benjamini–Hochberg to the  $p_{\text{enrich}}$  for FDR control (Proposition 5).
5. Validate the Poisson regime via the pairwise co-occupancy estimate of  $b_2$ ; if inflated, use the negative-binomial check (Proposition 2).
6. Size the control by Eq. (17); for rare queries fall back to the model-based  $\hat{\pi}_{\text{gen}} = P_{\text{gen}}(B_\theta(q))$  (the Murugan et al. generation model [15]), optionally reweighted per sequence by the Elhanati et al. selection factor  $Q(\sigma) \langle\langle Q \rangle\rangle_{P_{\text{gen}}} = 1$ , Eq. (6)) for the post-selection null — no global thymic scalar enters, as the average acceptance  $\alpha \lesssim Q_{\text{max}}^{-1} \approx 1/7$  is sequence-independent and cancels in the calibrated  $\hat{\pi}$  (Remark 2).

This is implemented in `seqtree.evalues` (with `exclude_exact` for the punctured counts), a thin layer over batched search; the control loader `seqtree.load_control` supplies a deduplicated background.

### 14. RELATED APPLICATIONS

The construction is not specific to antigen-specificity. The same neighbour-counting machinery—and, in places, only its combinatorial skeleton—serves several adjacent tasks.

**PEPTIDE–MHC ANALYSIS.** The principal extension is the presentation-aware pMHC layer of §12: anchor-masked, TCR-facing homology for cross-reactivity and molecular mimicry, and the inverse anchor-weighted reading for allele guessing, each calibrated against its own per-allele null.

**UMI AND BARCODE COLLAPSE: A BIRTHDAY PROBLEM..** Unique molecular identifiers and cell barcodes are random strings of length  $L$  over an alphabet of size  $A$ . Deduplication collapses identifiers lying within a small edit budget (sequencing errors)—the fuzzy search of §2 on a tiny alphabet. The statistical question is the *dual* of the E-value: not “are there more neighbours than chance predicts” but “how often do two *distinct* molecules collide by chance.” For  $n$  molecules the expected number of colliding pairs is  $\binom{n}{2}A^{-L}$  and  $\mathbb{P}(\text{any collision}) \approx 1 - \exp(-n^2/(2A^L))$ —the birthday bound—with the edit-distance- $\theta$  variant inflating  $A^{-L}$  by the ball size  $|B_\theta(q)|$ . This fixes the safe number of molecules for a given  $(A, L, \theta)$  and separates merges that are sequencing errors (one true molecule) from chance collisions (two).

**CDR3 NUCLEOTIDE ERROR CORRECTION..** PCR and sequencing errors spawn read variants one or two substitutions from a true clonotype. Clustering reads by edit distance—the seqtree engine on the nucleotide alphabet, with the conserved 3' end indexed first so shared suffixes prune early (§7)—collapses each error cloud onto its parent. Whether a low-count neighbour is an error or a genuine rare clonotype is decided by the same Poisson tail (§4), now under a per-base error-rate null in place of the control repertoire.

## REFERENCES

- [1] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.
- [2] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.
- [3] Bruno Alvarez, Birkir Reynisson, Carolina Barra, Søren Buus, Nicola Ternette, Tim Connelley, Massimo Andreatta, and Morten Nielsen. NAlign\_MA; MHC peptidome deconvolution for accurate MHC binding motif characterization and improved T-cell epitope predictions. *Molecular & Cellular Proteomics*, 18(12):2459–2477, 2019.
- [4] Massimo Andreatta, Bruno Alvarez, and Morten Nielsen. GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Research*, 45(W1):W458–W463, 2017.
- [5] Richard Arratia, Larry Goldstein, and Louis Gordon. Two moments suffice for Poisson approximations: the Chen-Stein method. *The Annals of Probability*, 17(1):9–25, 1989.
- [6] Richard Arratia, Larry Goldstein, and Louis Gordon. Poisson approximation and the Chen-Stein method. *Statistical Science*, 5(4):403–434, 1990.
- [7] 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.

- [8] Garry Dolton, Cristina Rius, Aaron Wall, et al. Targeting of multiple tumor-associated antigens by individual T cell receptors during successful cancer immunotherapy. *Cell*, 186(16):3333–3349, 2023.
- [9] 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 USA*, 111(27):9875–9880, 2014.
- [10] Samuel Karlin and Stephen F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences USA*, 87(6):2264–2268, 1990.
- [11] Samuel Karlin and Stephen F. Altschul. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Sciences USA*, 90(12):5873–5877, 1993.
- [12] Thomas Konstantinovsky and Gur Yaari. Flashback: A reversible bilateral run-peeling decomposition of strings, 2026.
- [13] Lucien Le Cam. An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics*, 10(4):1181–1197, 1960.
- [14] Marta Łuksza, Nadeem Riaz, Vladimir Makarov, Vinod P. Balachandran, Matthew D. Hellmann, Alexander Solovyov, Naiyer A. Rizvi, Taha Merghoub, Arnold J. Levine, Timothy A. Chan, Jedd D. Wolchok, and Benjamin D. Greenbaum. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature*, 551(7681):517–520, 2017.
- [15] 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 USA*, 109(40):16161–16166, 2012.
- [16] 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.
- [17] 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 USA*, 115(38):9604–9609, 2018.
- [18] Andreas Tiffeau-Mayer, Jonathan A. Levine, Christopher J. Russo, Quentin Marcou, William Bialek, and Benjamin D. Greenbaum. How different are self and nonself? *PRX Life*, 4:013027, 2026.

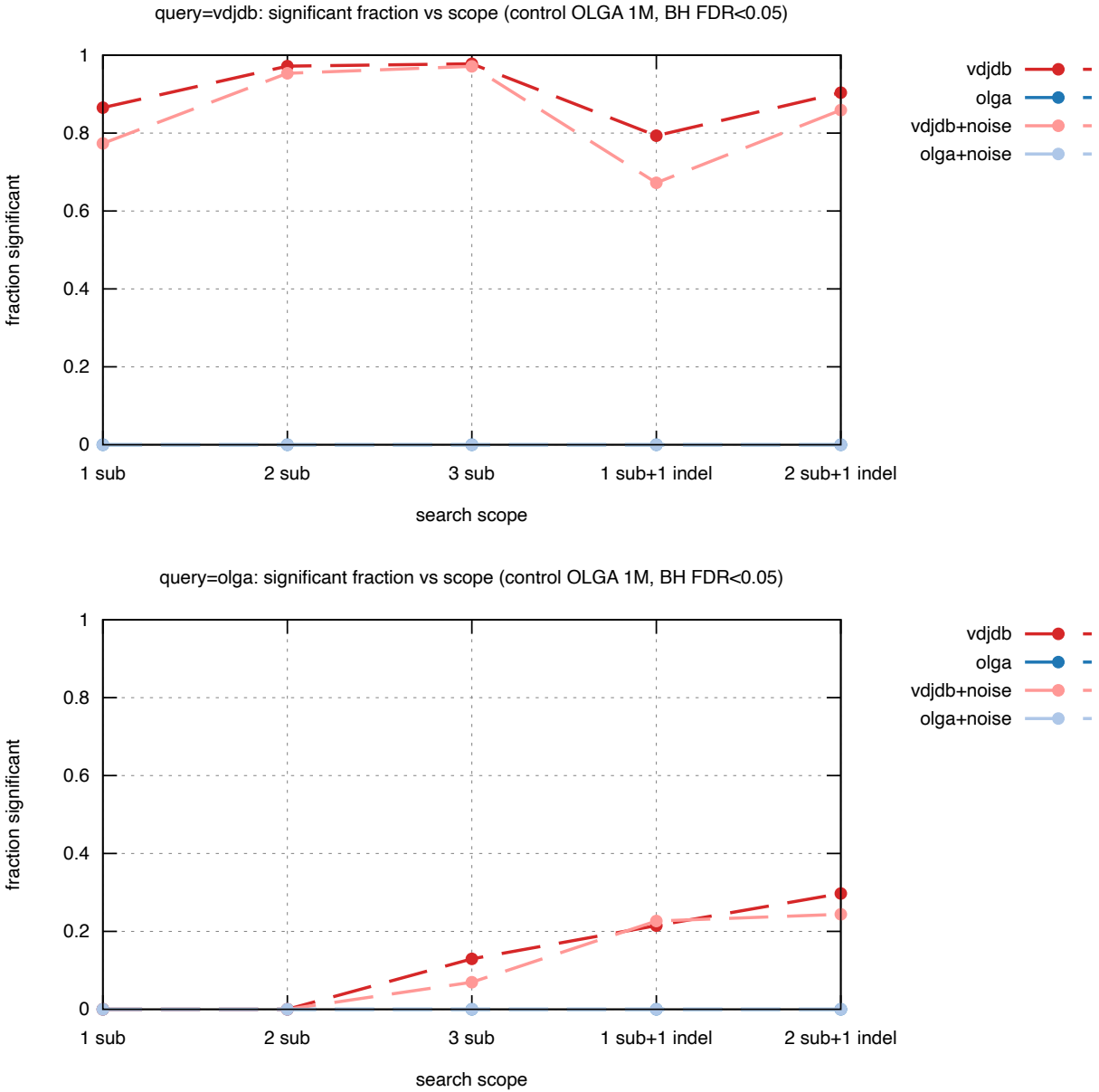


Figure 1: Control-calibrated significance across reference structure and search scope (bench/bench\_evaluate\_matrix.py): fraction of unique hits called significant (Benjamini–Hochberg  $FDR < 0.05$ ) vs scope — 1/2/3 substitutions and 1–2 substitutions + one indel — for VDJdb / OLGA / +noise references against an OLGA background, query = VDJdb (top) and OLGA (bottom), both panels on  $[0, 1]$ . Antigen-selected VDJdb is enriched and survives noise dilution; the unstructured OLGA nulls sit at zero — the multiple-testing correction is what separates them (a fixed  $E < 1$  cutoff over-calls).

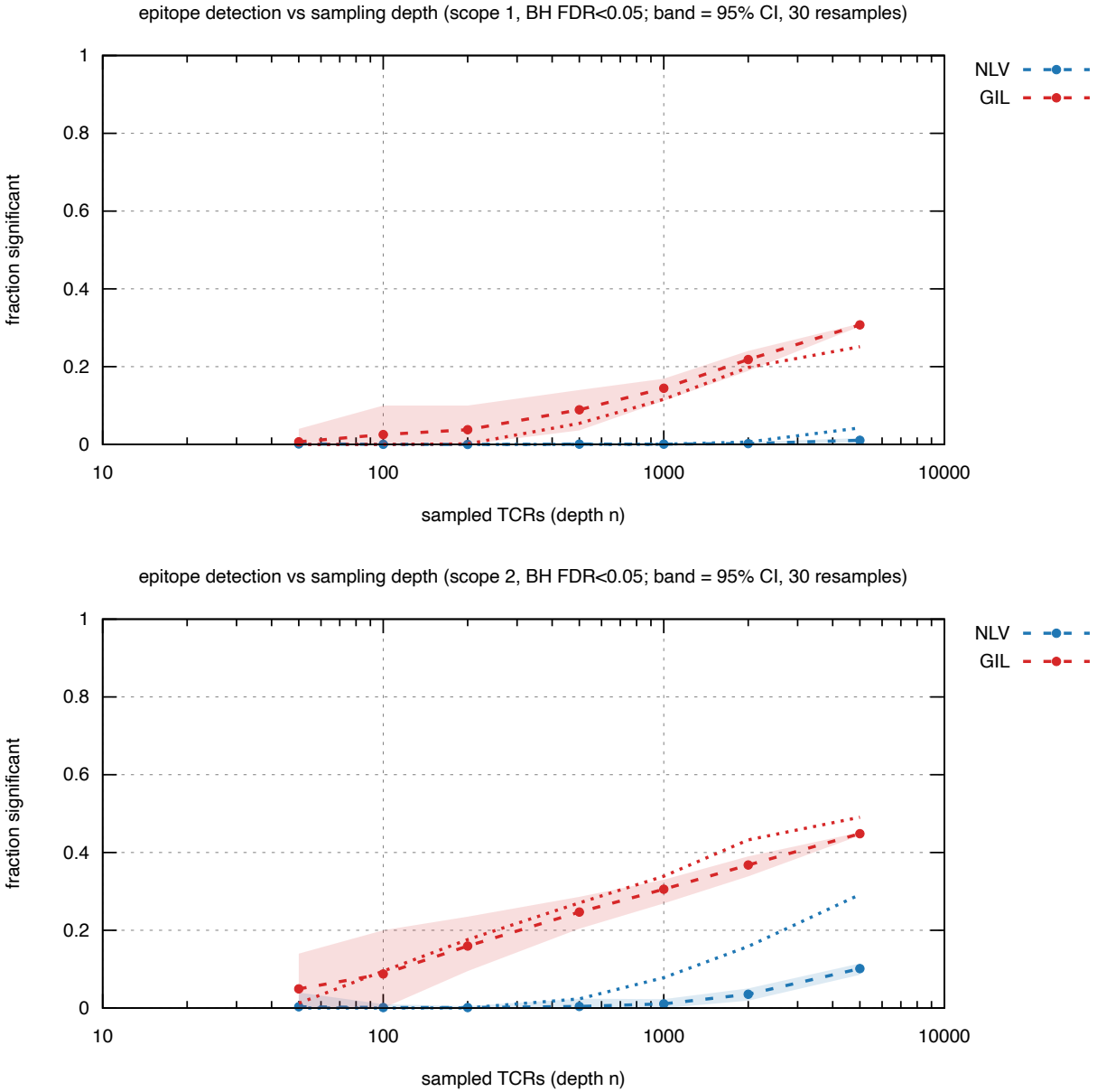


Figure 2: Epitope detection complexity (`bench/bench_epitope.py`): fraction called significant (Benjamini–Hochberg,  $FDR < 0.05$ ) versus sampled depth  $n$  for the convergent GIL and the diffuse NLV A\*02 repertoires. **Dashed** = observed with a **shaded 95% confidence band** (30 independent subsamples per depth); **dotted** = the degree-distribution prediction of Eq. (16); colour denotes the epitope (NLV / GIL). GIL’s dominant cluster is recovered an order of magnitude sooner than NLV’s diffuse one.

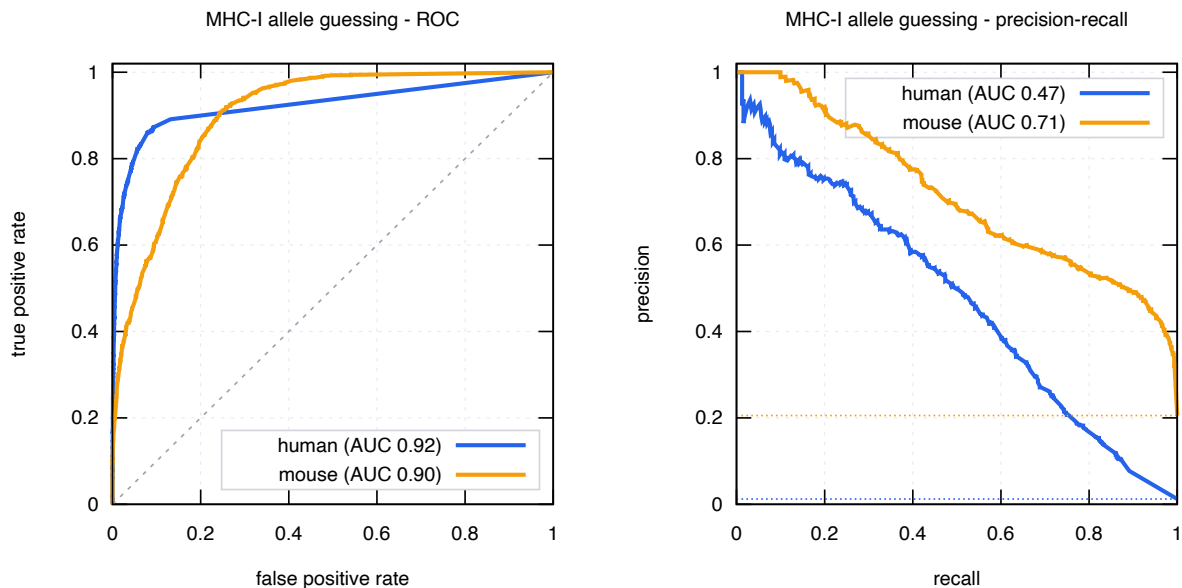


Figure 3: MHC-I allele guessing: ROC (left) and precision–recall (right) for the per-(peptide, allele) task, human and mouse separately. Diagonal = chance (ROC); dashed = prevalence baseline (PR). AUCs in legend.

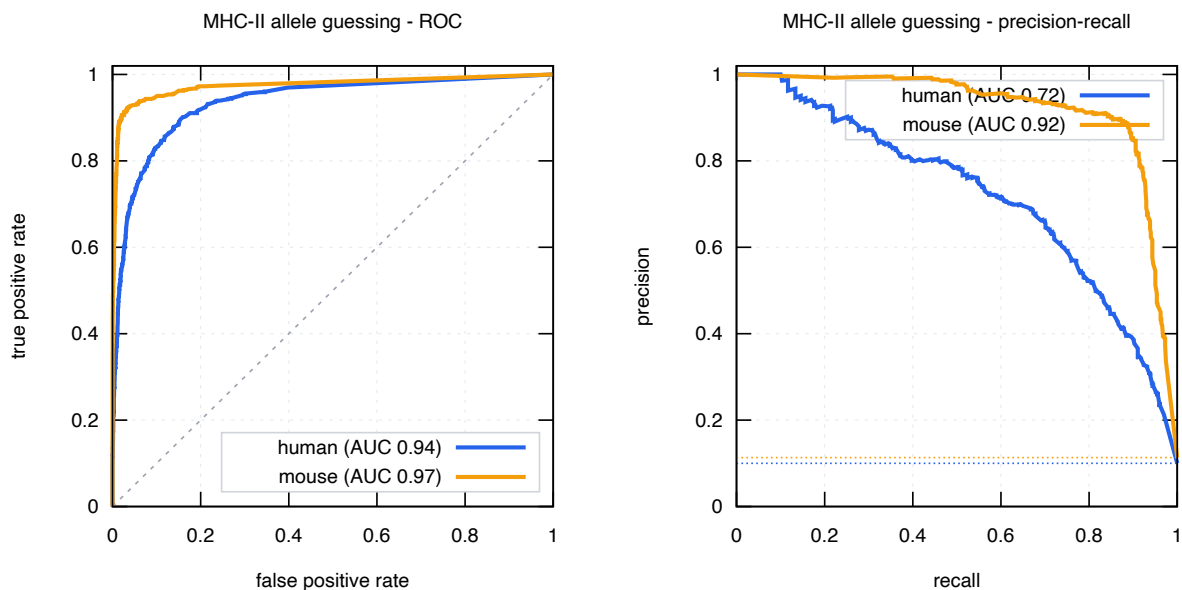


Figure 4: MHC-II allele guessing (register trick, single best 9-mer core register): ROC and precision–recall, human and mouse separately. Same conventions as Fig. 3.