

# PRESENTATION-AWARE E-VALUES, CROSS-ALLELE PSEUDOSEQUENCE DIFFUSION, AND PROTEOME-SCALE PEPTIDE MATCHING FOR PEPTIDE–MHC ANALYSIS

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

*mhcmatch technical appendix*

## Abstract

This appendix develops the mathematical and statistical theory of `mhcmatch`, the applied peptide–MHC layer built on the `seqtree` fuzzy-search substrate. We take as given the control-calibrated E-value of the `seqtree` appendix (`evaluate.tex`) and specialize it to MHC *presentation*, whose null is allele-conditional and anchor-constrained. We restate the anchor / TCR-facing decomposition and the per-allele presentation-aware E-value (the *forward* restriction problem), then the *reverse* problem—peptide to presenting allele—as a neighbour-vote posterior with a binomial-tail confidence E-value, together with its non-binder filter and class-II promiscuity treatment. The central contribution is a *cross-allele diffusion* model: each allele is represented by its 34-residue groove pseudosequence, and per-anchor preference distributions are smoothed by kernel-weighted empirical-Bayes shrinkage toward groove-similar alleles, with the per-anchor relevance of each groove position *learned* from data (a mutual-information feature importance that separates, e.g., the MHC-I B-pocket governing P2 from the F-pocket governing PΩ). This rescues rare alleles with few peptides—lifting the `seqtree` limitation that distinct alleles are distinct, never-pooled nulls—and we give its hierarchical-Bayes interpretation, effective sample size, limiting cases, and bias–variance trade-off. We then treat proteome-scale scanning (sliding-window presentation with multiple-testing control; near-exact neoantigen source identification), motif logos with length distributions, the composition rules for the planned stability / affinity / cleavage / expression / immunogenicity predictors, and the benchmark protocol against NetMHCpan, NetMHCIIpan, MixMHCpred and MixMHC2pred.

## 1. SCOPE AND RELATION TO THE SUBSTRATE

`seqtree` provides a payload-agnostic fuzzy-search core and a control-calibrated E-value: for a query  $q$  and a scope  $\theta$ , with an empirical background  $P_0$  estimated from a matched control of size  $M$  and

---

\*Correspondence: mikhail.shugay@gmail.com

a target set of size  $N$ , the expected number of chance neighbours is  $\widehat{E}(q, \theta) = (N/M) n_C(q, \theta)$ , with  $p_{\text{any}} = 1 - e^{-\widehat{E}}$  and  $p_{\text{enrich}} = \mathbb{P}(\text{Poisson}(\widehat{E}) \geq n_D)$  (`seqtree` appendix, Def. of the E-value; Poisson/Chen–Stein bound; Karlin–Altschul [9, 1] as the i.i.d. special case). We do not re-derive that theory; we *specialize* it.

For peptide–MHC the relevant background is not V(D)J generation but *presentation*, which is allele-conditional and constrains the anchor positions. `mhcmatch` owns four developments on top of the substrate: (i) the productionized forward/reverse presentation E-value (§2–3); (ii) the cross-allele *pseudosequence diffusion* model that pools statistics across groove-similar alleles (§4, the headline); (iii) proteome-scale scanning and near-exact source identification (§5); and (iv) motif logos (§6). Planned predictors and the comparison protocol are §7–8; limitations are §9.

## 2. ANCHOR DECOMPOSITION AND THE PER-ALLELE PRESENTATION E-VALUE

A presented peptide of length  $L$  factorizes  $\sigma = (\sigma_A, \sigma_T)$  into MHC-facing *anchors*  $A(a, L)$  and TCR-facing positions  $T$ . For class I the buried anchors are {P2, PΩ} (the B- and F-pockets); for the class II 9-mer core they are {P1, P4, P6, P9}. *Presentation* fixes  $\sigma_A$ ; *recognition* reads  $\sigma_T$ . Dolton et al. [7] exhibit one HLA-A02:01 TCR cross-reacting with three epitopes through a shared central motif with the anchors irrelevant, which motivates two complementary maskings, each realized by a per-position weight vector  $w$  over a `seqtree PositionalMatrix`:

$$\text{(recognition)} \quad w_i = \mathbb{1}[i \in T]; \quad \text{(presentation)} \quad w_i = \mathbb{1}[i \in A(a, L)]. \quad (1)$$

In `mhcmatch` these are exposed by `Store.decompose`, which returns both readouts with the masked positions written as the spare symbol `X`.

**FORWARD PROBLEM..** For a candidate restriction allele  $a$  with an anchor-masked presented control  $C_a$  of size  $M_a$  and target peptidome of size  $N_a$ , the presentation-aware E-value is the substrate E-value on allele-restricted, anchor-conditioned indices,

$$\widehat{E}(q, \theta; a) = \frac{N_a}{M_a} n_{C_a}(\sigma_T, \theta), \quad p_{\text{enrich}} = \mathbb{P}(\text{Poisson}(\widehat{E}) \geq n_{D_a}^T), \quad (2)$$

i.e. significance is computed on the TCR-facing readout only, so shared anchors (presentation, not recognition) do not inflate within-allele hits. The analytic null for rare  $a$  is taken up in §4.

## 3. REVERSE PROBLEM: PEPTIDE TO PRESENTING ALLELE

The forward E-value answers “is  $q$  a binder of a *given*  $a$ ?”. Restriction prediction asks the reverse: which alleles present  $q$ ? We flip the weights in (1) to the *presentation* mode and index reference peptides by their anchor signature (`layout.presentation_features`; class I N/C-pocket residues P1 P2 P3, P(Ω–1), PΩ, class II core P1 P4 P6 P9). For a query we widen the scope on its signature until it has  $n \in [10, 100]$  non-exact neighbours, and tally their alleles,  $k_a$  votes out of  $n$ .

**DEFINITION 1** (Ranking and confidence). The *ranking* score is the neighbour vote fraction  $\hat{p}(a | q) = k_a/n$ , a posterior over the panel. The *confidence* score is the per-allele enrichment over the panel

background frequency  $f_a$ ,

$$E_a = -\log_{10} \mathbb{P}(\text{Binomial}(n, f_a) \geq k_a), \quad (3)$$

the upper binomial tail; a peptide binds  $a$  iff  $E_a \geq -\log_{10} \alpha$  and  $k_a > 0$ .

**REMARK 1** (Why vote fraction, not enrichment, ranks). On a skewed panel the dominant true allele has a large expected count  $nf_a$ , so the enrichment (3) *penalises* it; the vote fraction does not. Hence  $\hat{p}$  ranks and  $E_a$  gates. This separation, validated on `pmhc_data` (per-(peptide, allele) ROC-AUC 0.90–0.99, `seqtree/bench/bench_mhc_guess.py`), is what `Store.restriction` implements.

**CLASS-II REGISTER TRICK..** The class-II core register inside a longer peptide is unknown; full deconvolution clusters it [3, 2]. We use a one-pass proxy: pick the single 9-mer window whose P1 is a large hydrophobic and whose P4/P6/P9 avoid Pro/Gly, and take that core’s {P1,P4,P6,P9}. Committing to one register makes an allele’s peptides share a consistent signature and is what lifts class-II ROC-AUC into the 0.9 range.

**NON-BINDER FILTER AND PROMISCUITY..** A peptide that binds *no* panel allele has small  $\max_a E_a$  (real-vs-random rejection); one that does not bind a *specific*  $a$  has  $E_a < -\log_{10} \alpha$ . The global statistic  $E_{\text{glob}} = \sum_a 10^{-E_a}$  is the expected number of panel alleles presenting  $q$  by chance. Class II is genuinely multi-label (an open groove with loosely specified pockets presents one peptide on many alleles), so the panel positive set is *every* observed restricting allele and the non-binder test is “binds none of the panel”.

### 3.1. WHERE THE E-VALUE IS USED: SIGNIFICANCE VS RANKING

The E-value is a calibrated *significance*—an expected chance count and its Poisson tail (§2)—not a discriminative score, and the two answer different questions. *Ranking* (the diffused anchor log-odds, or the neighbour vote) *orders* candidates and is the better discriminator: for allele prediction the log-odds ranks best (§8), and for telling binders from random non-binders the max-over-panel presentation score reaches AUROC  $\approx 0.80$  (real held-out peptides vs  $10^4$  corpus-frequency random peptides). The *E-value* answers “is this many hits more than chance against the right background?”—it is what one *reports* when significance and multiplicity matter:

- **Molecular mimicry and self/proteome search** (§5): for a neoantigen, `find_mimics` reports the expected number of TCR-facing homologs in a self or pathogen set against the *per-allele presented* background, with  $p_{\text{enrich}}$  and the rule of three for empty controls. Rank cannot say whether a shared motif is surprising; the E-value deflates motifs common in the allele’s peptidome, elevates surprisingly shared ones, and lets one scan whole proteomes under FWER/FDR control. This is the principal reason to carry an E-value.
- **Binder vs non-binder** uses *both*: the diffused log-odds (or its max over the panel) to *rank / discriminate*, and the per-allele enrichment  $E_a$ —or the global  $E_{\text{glob}}$  for “binds nothing”—to *gate* with a calibrated threshold and control error across a panel or proteome. Empirically the score is the better *discriminator* (enrichment-vs-background penalizes data-rich alleles and needs neighbours), while the E-value supplies the defensible cut-off and the multiple-testing control the bare score lacks. They compose: rank by score, then threshold by  $E$  / FDR.

## 4. MHC PSEUDOSEQUENCE MODEL AND CROSS-ALLELE DIFFUSION

`seqtree` deliberately treats distinct alleles as distinct, never-pooled nulls. This is correct but wasteful: rare alleles carry few presented peptides, so both (2) and (3) are noisy exactly where data are thin—the equity problem of Glynn et al. [8]. `mhcmatch` lifts the limitation by making the allele a *sequence* the same engine can compare, and diffusing statistics along groove similarity.

**PSEUDOSEQUENCE..** Each allele  $a$  is represented by its NetMHCpan-style 34-residue groove pseudosequence  $s_a \in \Sigma^{34}$  [14]: the polymorphic, peptide-contacting positions (class I  $\alpha_1/\alpha_2$ ; class II  $\alpha_1 + \beta_1$ ). Per-allele binding can be predicted from  $s_a$  alone [8, 4], so groove distance is a meaningful allele distance.

### 4.1. ANCHOR-FACTORED SIMILARITY KERNEL

The presentation motif factorizes by pocket: anchor  $j$ ’s residue preference is governed by the groove positions lining pocket  $j$ , and different pockets are spatially separated (the class-I B-pocket sets P2, the F-pocket sets PΩ). We therefore use a *per-anchor* kernel rather than one global allele distance. For anchor  $j$  let  $w_j \in \mathbb{R}_{\geq 0}^{34}$  weight the groove positions, and define

$$d_j(a, b) = \sum_{p=1}^{34} w_{j,p} \delta(s_{a,p}, s_{b,p}), \quad K_j(a, b) = \exp(-d_j(a, b)/h_j), \quad (4)$$

where  $\delta(x, y) = s(x, x) + s(y, y) - 2s(x, y)$  is the BLOSUM62 Gram distance (zero on identity, small for conservative substitutions, large for dissimilar residues; normalized so an average substitution costs  $\approx 1$ ), evaluated through `seqtree`’s substitution matrix (`SubstitutionMatrix.penalty`, exposed for this use). Bandwidth  $h_j > 0$ ; ambiguous positions X are skipped. Replacing  $\delta$  with the identity indicator  $\mathbb{1}[x \neq y]$  gives the plain Hamming variant, and  $w_j \equiv 1$  a single un-factored kernel.

**GENERATIVE ALTERNATIVE (FISHER KERNEL)..** The kernel can instead be read off an explicit generative groove model—a per-position multinomial with the MI weights (5) as position relevance (the Bayes-net skeleton left by DPI pruning, §4.4)—whose Fisher score  $U_a[p, r] = w_p (\mathbb{1}[s_a[p]=r] - \bar{p}_p(r))$  yields a Fisher kernel  $\langle U_a, U_b \rangle$ . Empirically this generative kernel tracks the BLOSUM one closely (top-5 neighbour-set Jaccard 0.76) yet predicts held-out modal anchors no better (leave-one-allele-out accuracy 0.43 vs 0.46 for BLOSUM; `bench/fisher_kernel.py`). Because the BLOSUM Gram distance is itself a substitution log-odds,  $\exp(-\delta)$  already *is* a likelihood kernel—one with biochemical substitution structure and no panel fit—so we keep it as the default and treat the Fisher kernel as a validated equivalent rather than an improvement.

**DEFINITION 2** (Learned per-anchor relevance). Let  $A_j$  be the allele’s modal residue at peptide anchor  $j$  (from its presented set) and  $S_p$  the groove residue at pseudosequence position  $p$ , both viewed as categorical variables over the allele panel. The relevance of position  $p$  to anchor  $j$  is the mutual information

$$w_{j,p} \propto \mathbb{I}(S_p; A_j) = \sum_{x,y} \hat{\mathbb{P}}(S_p=x, A_j=y) \log_2 \frac{\hat{\mathbb{P}}(S_p=x, A_j=y)}{\hat{\mathbb{P}}(S_p=x) \hat{\mathbb{P}}(A_j=y)}, \quad (5)$$

normalized to mean one over  $p$ . Equivalently  $w_j$  may be fit by maximizing the held-out predictive log-likelihood of the shrinkage model below under an  $\ell_1$ /entropy penalty, which selects each anchor’s pocket positions. (5) is the “feature importance” that says which groove residues drive P2 versus P $\Omega$ .

#### 4.2. KERNEL-SHRINKAGE POOLING

Write  $\theta_{a,j}$  for allele  $a$ ’s empirical residue distribution at anchor  $j$  over its  $n_a$  peptides. We shrink it toward groove-similar alleles:

$$\hat{\theta}_{a,j} = \frac{n_a \theta_{a,j} + \sum_{b \neq a} K_j(a, b) n_b \theta_{b,j}}{n_a + \sum_{b \neq a} K_j(a, b) n_b}, \quad \text{ESS}_{a,j} = n_a + \sum_{b \neq a} K_j(a, b) n_b. \quad (6)$$

**PROPOSITION 1** (Hierarchical-Bayes reading and limits). *(6) is the posterior mean of a Dirichlet–multinomial in which the neighbour-weighted counts form the prior pseudocounts:  $\hat{\theta}_{a,j} = \mathbb{E}[\theta \mid \text{counts}]$  with concentration  $\text{ESS}_{a,j}$ . As  $h_j \rightarrow 0$ ,  $K_j(a, b) \rightarrow 0$  for  $b \neq a$  and  $\hat{\theta}_{a,j} \rightarrow \theta_{a,j}$  (no pooling); as  $h_j \rightarrow \infty$ ,  $K_j(a, b) \rightarrow 1$  and  $\hat{\theta}_{a,j} \rightarrow (\sum_b n_b \theta_{b,j}) / \sum_b n_b$  (the global pool).*

(Both limits are implemented and unit-tested in `Pseudoseq.shrink`.)

**PROPOSITION 2** (Bias–variance and bandwidth). *Treat each  $\theta_{b,j}$  as an unbiased estimate of allele  $b$ ’s truth  $\theta_{b,j}^*$  with variance  $\propto 1/n_b$ . The pooled estimator has variance  $O(1/\text{ESS}_{a,j})$  and bias  $\|\sum_b \kappa_b (\theta_{b,j}^* - \hat{\theta}_{a,j})\|$ , where  $\kappa_b = K_j(a, b) n_b / \text{ESS}_{a,j}$  is the borrowed mass. The mean-squared error is minimized by borrowing more (smaller effective  $h_j$  penalty) when  $n_a$  is small and groove neighbours are close, and less when  $n_a$  is large; calibrating  $h_j$  by cross-validated predictive likelihood realizes this trade-off automatically.*

**BOUNDED-CONCENTRATION VARIANT (THE PRACTICAL DEFAULT)..** The counts-weighted form (6) lets one deep neighbour ( $K_j n_b \gg n_a$ ) overwrite a rare allele’s own peptides. We therefore default to a fixed prior strength  $\tau$ : shrink toward the kernel-weighted neighbour mean  $m_{a,j} = (\sum_b K_j(a, b) n_b \theta_{b,j}) / \sum_b K_j(a, b) n_b$ ,

$$\hat{\theta}_{a,j} = \frac{n_a \theta_{a,j} + \tau m_{a,j}}{n_a + \tau}, \quad (7)$$

so borrowing adds at most  $\tau$  pseudocounts and vanishes automatically as  $n_a$  grows (Prop. 2). Empirically (`bench/bench_diffusion.py`; human, allele-split held-out rank AUC with the BLOSUM-scored, MI-weighted kernel (4),  $\tau = 10$ ,  $h = 2$ , the per-allele score being the anchor log-odds over the N/C pocket footprint {P1,P2,P3,P( $\Omega$ –1),P $\Omega$ }) the diffusion lifts the AUC of *rare* MHC-I alleles ( $\leq 30$  peptides) from 0.94 to 0.96 on the high-confidence ( $\geq 2$ -publication) set while leaving *medium* (31–199) and *frequent* alleles unchanged ( $\approx 0.97$ ); the gain decays smoothly as an allele’s own data grows and frequent alleles are never harmed. The class II AUC gains are smaller throughout ( $\leq 0.02$ , rare  $0.77 \rightarrow 0.77$ ), consistent with the harder class II problem (Fig. 1). Exclusion is *per-pMHC and benchmark-only*: only the held-out (epitope, allele) pair is dropped from training—the same epitope presented by another allele is a distinct, legitimate pMHC and is kept, so the diffusion may borrow that co-presentation (the real cross-allele transfer we measure); the production model trains and votes on all data. The richer pocket footprint matters: with the two primary anchors alone ({P2,P $\Omega$ }) the full-set rare gain nearly vanishes, whereas the auxiliary pocket-proximal positions

make per-allele estimation sparser—so diffusion has more to rescue. The data-rescue is safe: it never lowers a well-sampled allele’s AUC.

**ALLELE PREDICTION (TOP- $k$ , PROMISCUITY-AWARE, CROSS-VALIDATED)..** Cast as the reverse problem—for each held-out pMHC ( $p, a$ ) rank all panel alleles for  $p$  and ask whether the held-out allele  $a$  is recovered in the top- $k$ . We use *5-fold* cross-validation, pooling each pMHC’s single out-of-fold prediction, capped per allele so rare alleles are represented and frequent ones do not dominate (`bench/tune_diffusion.py`; per-panel tables in `bench/results/`). Because peptides are promiscuous we score *recovery@5* split by allele rarity. For MHC-I the diffusion lifts *rare*-allele *recovery@5* from 0.47 to 0.75 (short-list) and 0.30 to 0.44 (full tier), at a small frequent cost ( $\approx -0.02$ ) and roughly flat overall top-5 ( $\approx 0.85/0.64$ ). The *ranker* is the diffused anchor log-odds itself, not the neighbour vote: on a novel peptide the vote tally has few same-allele signature neighbours, so vote-first ranking buries the true allele (top-5 0.87, rare *recovery@5* 0.25 on a single split) whereas the diffused log-odds scores every allele directly. A grid sweep selects  $h = 2, \tau = 10$ , BLOSUM (marginally over identity), as optimal. Class II remains hard (*recovery@5*  $\approx 0.47$ , diffusion near-neutral, top-5  $\approx 0.25$ ) and mouse is too few-allele to be conclusive (high CV variance): both are targets for the structure-based extension. Top-1 is *not* the right yardstick here (promiscuity makes several alleles correct); top-5/*recovery@5* is.

**PER-LOCUS CALIBRATION..** A single global ( $h, \tau$ ) need not be optimal for every locus. A per-locus grid (`tune_diffusion.py --by-locus`) shows the loci do differ—HLA-B tolerates wider borrowing ( $h=2$ ) than HLA-A/C ( $h=0.5$ ), and most prefer a weaker prior ( $\tau=5$ )—but the per-locus rare sets on a single split are small and the estimates noisy (HLA-A’s apparent *recovery@5* of 1.0 is a tiny-sample artifact). We therefore keep the cross-validated global  $h=2, \tau=10$  as the production default and treat validated per-locus tuning (a CV grid per locus, needing more data) as a calibration refinement rather than promoting single-split optima.

**ZERO-SHOT TRANSFER (LEAVE-ONE-ALLELE-OUT)..** The sharpest test of the rescue removes *every* peptide of a target allele from training and scores its held-out peptides from groove neighbours alone (`bench/transfer.py`). On 30 human MHC-I alleles the diffused model reaches real-vs-random AU-ROC 0.95 with *no* data for the target allele (the data-free raw model is uninformative, 0.22), and transfer stays strong even when the nearest trained groove neighbour is only moderately similar (0.94 at kernel  $< 0.5$  vs 0.96 at  $\geq 0.5$ ): a genuinely novel allele is predicted purely from its pseudosequence—the limiting case of the rare-allele rescue.

**BAYESIAN READING (IN PLAIN TERMS)..** The shrinkage (7) is exactly a Bayesian posterior mean. Treat each allele’s anchor- $j$  preference  $\theta_{a,j}$ —the probabilities of each amino acid in pocket  $j$ —as *unknown*. Before seeing allele  $a$ ’s own peptides we expect it to resemble its groove neighbours, so we put a prior centred on the kernel-weighted neighbour mean  $m_{a,j}$  with strength  $\tau$  (measured in pseudo-peptides); the allele’s  $n_a$  observed peptides then update that belief. With the conjugate Dirichlet–multinomial model

$$\theta_{a,j} \sim \text{Dirichlet}(\tau m_{a,j}), \quad \sigma_{i,j} \mid \theta_{a,j} \sim \text{Categorical}(\theta_{a,j}), \quad \theta_{a,j} \mid \{\sigma_{i,j}\} \sim \text{Dirichlet}(\tau m_{a,j} + c_{a,j}), \quad (8)$$

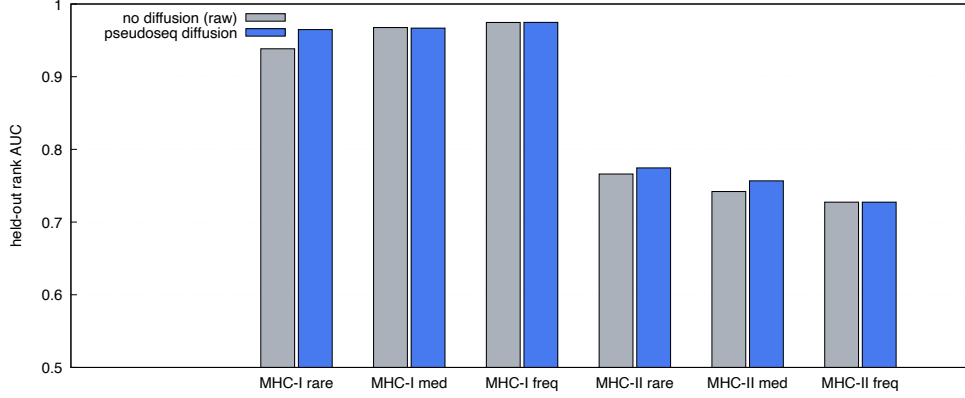


Figure 1: Held-out rank AUC for rare ( $\leq 30$  peptides), medium (31–199) and frequent ( $\geq 200$ ) alleles, without (raw, grey) and with (blue) pseudosequence diffusion, allele-split. MHC-I on the  $\geq 2$ -publication set, MHC-II on the full set. Diffusion lifts rare alleles and leaves medium and frequent alleles essentially unchanged in both classes—the gain decays smoothly as an allele’s own data grows. `bench/bench_diffusion.py`.

where  $c_{a,j}(r)$  counts peptides of  $a$  with residue  $r$  at pocket  $j$  ( $\sum_r c_{a,j}(r) = n_a$ ), the posterior mean is *precisely* (7). A data-rich allele ( $n_a \gg \tau$ ) ignores the prior and keeps its own motif; a data-poor one ( $n_a \ll \tau$ ) leans on its neighbours. The bandwidth  $h$  (how far similarity reaches) and the strength  $\tau$  (how much to borrow) are the only knobs (Fig. 2).

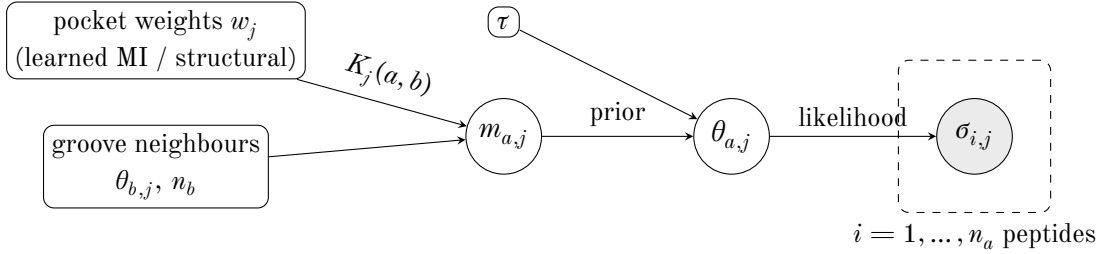


Figure 2: Graphical model behind the diffusion shrinkage, fit independently per pocket  $j$ . The per-pocket groove weights  $w_j$ —learned by mutual information (5) or measured from structures (§ 4.4)—set the BLOSUM-scored similarity kernel  $K_j(a, b)$  (4); the kernel-weighted neighbour mean  $m_{a,j}$  and the strength  $\tau$  form a Dirichlet *prior* on the allele’s pocket- $j$  residue distribution  $\theta_{a,j}$ ; the  $n_a$  observed anchor residues  $\sigma_{i,j}$  (shaded) are the *likelihood*. The posterior mean is the shrinkage estimator (7): rare alleles borrow from groove-similar neighbours, well-sampled alleles do not.

**POOLED NULL AND RESTORED E-VALUE..** For a rare allele the analytic presented null is built from the shrunk anchor marginals: assuming approximate position independence in the anchor block (the maximum-entropy presented distribution of [15] conditioned on the motif),  $\widehat{P}_0(\sigma_A | a) \approx \prod_j \hat{\theta}_{a,j}(\sigma_{A,j})$ , which feeds a usable  $\widehat{E}$  in (2)–(3) where the raw per-allele control was too small. The estimator is a smoothed null: its bias is controlled by the borrowed mass on dissimilar alleles (Prop. 2), vanishing as the panel around  $a$  becomes dense and groove-tight. In the production reverse problem

(Store.restriction, diffuse=True) the shrunk anchor log-odds both *rank*s the alleles and, with the neighbour enrichment (Def. 1), *gates* binders (binder iff vote-significant or anchor log-odds  $> 0$ ). On held-out (novel) peptides the anchor log-odds is the better ranker: the vote tally needs same-allele signature neighbours, which are sparse for a genuinely new peptide, so vote-first ranking buries the true allele (top-5 0.87, rare recovery@5 0.25), whereas ranking by the diffused log-odds scores every allele directly (top-5 0.95, rare recovery@5 0.88; §8). Vote thus serves the *confidence gate*, not the ranking.

### 4.3. CLUSTERING AND PROMISCUITY

Thresholding  $K_j$  (or the un-factored  $K$ ) yields single-linkage allele *clusters* of functionally similar alleles (Pseudoseq.cluster). A peptide presented across a cluster is *promiscuous*; its promiscuity is the spread of its positive alleles over the clustering, and the cluster is the natural unit over which to pool nulls when per-allele data are thin (especially class II). This is the principled cross-allele similarity that the substrate omits. Fig. 3 renders this as an allele network—edges are the MI-weighted, BLOSUM-scored kernel (4) above a soft/hard threshold, communities are greedy-modularity clusters [6]—which recover the classical HLA supertypes (A2, A3, B7, B27, B44, ...), the structural basis of cross-allele promiscuity. Mouse panels are few and groove-divergent, so they form no such communities. The kernel communities are well separated (partition modularity  $Q = 0.94$  human MHC-I, 0.90 class II) and respect allele structure—about half of same-two-field-family allele pairs co-cluster, the remainder splitting only because distant family members fall below the edge threshold (bench/promiscuity\_graph.py); a quantitative map to a curated supertype set is the natural external-data validation.

An empirical counterpart (Figs. 4 and 5, split by species) builds the same network from *observed* co-presentation—edges are the overlap coefficient  $|P_a \cap P_b| / \min(|P_a|, |P_b|)$  of two alleles’ presented-peptide sets—instead of predicted groove similarity. Its communities agree with the supertypes, while the extra cross-community edges are genuinely promiscuous peptides shared across distinct grooves; it covers alleles that lack a pseudosequence (203 vs 166 human MHC-I) and even links the otherwise groove-divergent mouse alleles. Quantifying predicted-vs-observed agreement, the Jaccard overlap of the two edge sets (over the alleles common to both, human) is 0.19 for MHC-I (66/343 edges) and 0.09 for MHC-II (42/458): the groove kernel captures the high-confidence supertype core, while observed sharing is broader and noisier. The two views are complementary rather than redundant.

### 4.4. PER-POCKET DECOMPOSITION

The anchor-factored weights (5) make the pocket structure explicit: for each peptide anchor we learn an independent weight vector  $w_j$  over the 34 groove positions, per class and species (Fig. 6, all four panels). Raw MI is inflated by *linkage* between groove positions—they co-vary across alleles by descent and structural co-evolution, so a position correlated with the true pocket residue also scores high and the profile smears. We therefore prune indirect links by the data-processing inequality (ARACNE [12]): position  $p$ ’s edge to pocket  $j$  is dropped when another position  $q$  is more informative about both the pocket and  $p$  ( $I(p; A_j) \leq \min(I(q; A_j), I(p; q))$ ), leaving the *direct* pocket positions.



After pruning, MHC-I (human) P $\Omega$ /F-pocket collapses to a single dominant groove position ( $\sim 19$ ) and P2/B-pocket to a small N-terminal cluster ( $\sim 7-8$ ), recovering the structural pocket layout from presented-peptide statistics alone (Fig. 6). For MHC-II the relevance sits on the polymorphic  $\beta_1$  segment and is near-zero on the largely monomorphic  $\alpha_1$  (sharpest for DR, whose DRA  $\alpha$  chain is invariant). The kernel itself keeps the *un-pruned* MI weights (the fuller signal gives marginally better neighbour matching—rare-allele recovery@5 0.88 vs 0.85 pruned); the prune is for interpretation. Mouse panels are data-limited (8–13 alleles), noisier but with the same gross localization. (Class-II alleles are keyed by the  $\alpha$ – $\beta$  pair, cores from the register trick; its diffusion gain is small,  $0.72 \rightarrow 0.74$  AUC.)

**STRUCTURAL VALIDATION..** The learned weights are confirmed by structure. We thread the pseudosequence onto 372 pMHC crystal structures (the *Canonical2026* TCR:pMHC set) with a fast C++ fitting aligner (`tcren._align`; no mmseqs/arda,  $\sim 0.1$  s/structure) and count peptide-anchor  $\leftrightarrow$  groove-position heavy-atom contacts (`bench/structural_pockets.py`). Class is assigned per structure by which pseudosequence fits best—an MHC-I 34-mer onto a single chain, or an MHC-II 34-mer onto the  $\alpha_1 + \beta_1$  chain-pair—rather than a  $\beta_2$ m/chain-length heuristic, which fails because TCR variable domains ( $\sim 110$  aa) and class-II groove domains ( $\sim 85$  aa) overlap  $\beta_2$ m’s size and class-II crystals are often domain-split; this yields 279 MHC-I and 93 MHC-II structures. For MHC-I the contacts recover the learned layout from physics alone (Fig. 7)—P2 contacts pseudosequence positions  $\sim 7-8$  (B-pocket), P $\Omega$  positions  $\sim 15-17$  (F-pocket), overlapping the learned maxima—and substituting these *structural* contact-frequency weights into the kernel (`weights="structural"`) gives near-identical rare-allele recovery@5 (0.72 vs 0.75 learned, 5-fold CV): a data-independent prior agreeing with the data-learned one. For MHC-II the four core pockets P1/P4/P6/P9 map to distinct  $\alpha_1 + \beta_1$  pseudosequence segments, and structural and learned weights give indistinguishable (and near-neutral) recovery@5 (0.464 vs 0.465)—confirming that the small class-II diffusion gain is intrinsic to the harder class-II problem, not an artifact of how the groove weights are estimated. A convex *blend* (`weights="blend"`, structure as a prior on the learned weights,  $\frac{1}{2}$  structural +  $\frac{1}{2}$  learned, renormalized) is the principled empirical-Bayes choice when the learned weights are data-starved; on the current panels it gives indistinguishable recovery@5 (0.462 vs 0.465 learned for MHC-II, 5-fold CV), reinforcing the same conclusion—more class-II ligand data, not a better weight estimator, is what the harder class needs.

## 5. PROTEOME-SCALE MATCHING

**PRESENTATION SCAN..** To find all presented peptides in a protein we slide every binding-length window (class I 8–11; class II  $\geq 13$  with the register-anchored core) and evaluate (3) per window and allele. With  $W$  windows and  $|\mathcal{A}|$  panel alleles the test is multiple: under the null the expected number of windows called for allele  $a$  is  $\approx W\alpha$ , so we control the family either by Bonferroni / max-statistic thresholds (FWER) or by Benjamini–Hochberg on the  $p_{\text{enrich}}$  across the  $W \times |\mathcal{A}|$  grid (FDR) [5]. `scan_protein(correction="bonferroni"|"bh")` applies exactly this control over the voted (*window, allele*) tests—Bonferroni for FWER, Benjamini–Hochberg for FDR—and re-flags binders at the corrected threshold; “correction=None” keeps the per-window  $\alpha$ .

**SAME-MHC VERSUS TCR-FACING SEARCH..** At proteome scale two notions of “similar peptide” are distinguished by which mask of (1) is indexed: *same-MHC* similarity indexes the presentation signature (peptides likely presented by the same allele), *TCR-facing* similarity indexes the anchor-masked central motif (similar T-cell recognition, the basis of cross-reactivity). Both run on the C++ KmerIndex seed-and-gather.

**MOLECULAR MIMICRY..** A self or foreign peptide  $p$  is a candidate cross-reactive mimic of a neoantigen  $v$  iff (i) it is presented by a compatible allele, (ii) the TCR-facing distance  $d_{\text{TCR}}(p, v) = \sum_i w_i \text{pen}(p_i, v_i) \leq \theta$  (anchors zeroed), and (iii) the hit is significant against the per-allele presented background,  $\hat{E} \leq \alpha$ . Condition (iii) deflates motifs common in the allele’s peptidome and elevates surprisingly shared ones. This connects to the neoantigen-quality program of Łuksza et al.: the quality  $Q = R \times D$  [11, 10] multiplies a recognition / non-self term  $R$  (the cross-reactivity distance to known antigens) by a self-discrimination term  $D$  (differential MHC binding), and to the close-to-self “shell” picture of [15]; `mhcmatch` supplies a calibrated E-value for the  $R$  component.

**NEAR-EXACT SOURCE IDENTIFICATION..** Given a neoantigen we identify the self peptide it derives from—its parent protein and the mutated position—by full-sequence (unmasked)  $\leq m$ -mismatch search over all length- $L$  windows of the reference proteome. Under an i.i.d. background of composition  $\{q_c\}$  the expected number of chance windows within  $m$  substitutions of a query in a proteome of  $R$  residues is

$$\mathbb{E}[\text{\#chance}] \approx R \sum_{i=0}^m \binom{L}{i} \bar{q}^{L-i} (1 - \bar{q})^i, \quad \bar{q} = \sum_c q_c^2, \quad (9)$$

the Karlin–Altschul/birthday count specialized to the Hamming ball (the empirical-background variant reuses `seqtree` §KA). For  $m \in \{0, 1\}$  and  $L \geq 8$ , (9) is  $\ll 1$ , so a hit is essentially certainly the true source rather than coincidence—`Proteome.find_source` returns it with the differing residue flagged.

## 6. MOTIF LOGOS AND LENGTH DISTRIBUTIONS

For an allele’s presented set (class I peptides of a fixed length; class II register-anchored 9-mer cores) the per-position information content is

$$I_i = \log_2 20 - H_i - e(n_i), \quad H_i = - \sum_c f_{i,c} \log_2 f_{i,c}, \quad (10)$$

with  $f_{i,c}$  the residue frequency at position  $i$  and  $e(n_i)$  the small-sample entropy correction for  $n_i$  counts (negligible for the deep `pmhc_data` alleles). The logo draws letter  $c$  at height  $f_{i,c} I_i$  bits; columns at the anchors stand tall (e.g. A02:01 shows P2=L,  $P\Omega \in \{V, L\}$ ), the TCR-facing columns are flat. The length distribution is reported as a companion histogram. Columns may be confidence-weighted by  $n$  (or by the shortlist  $\geq 2$ -publication support). `logo.motif` returns the numeric logo; `logo.render` draws it.

## 7. PLANNED PREDICTORS AND THEIR COMPOSITION

Presentation is necessary but not sufficient for immunogenicity. Each planned predictor scores a distinct biophysical or population term and *composes* with the presentation evidence on the log scale;

for peptide  $\sigma$  and allele  $a$  the combined log-odds is

$$\log \frac{\mathbb{P}(\text{immunogenic})}{\mathbb{P}(\text{not})} = \beta_0 + \beta_{\text{pres}} (-\log p_{\text{enrich}}) + \beta_{\text{aff}} \log \frac{1}{\text{IC}_{50}} + \beta_{\text{stab}} t_{1/2} + \beta_{\text{clv}} c_{\text{Cterm}} + \beta_{\text{expr}} \log e_{\text{gene}} + \beta_{\text{imm}} \rho_{\text{TCR}}, \quad (11)$$

with coefficients fit on labelled outcome data (the user will supply tuning/benchmark sets). The terms: *affinity*  $\text{IC}_{50}$  and *stability*  $t_{1/2}$  are the quantitative complements of the binary presentation E-value, regressable on the same anchor features plus the pseudosequence; *cleavage*  $c_{\text{Cterm}}$  is a position-specific C-terminal generation probability (with optional N-terminal trimming); *expression*  $e_{\text{gene}}$  and variant frequency enter as Bayesian priors multiplying the presentation odds; *immunogenicity*  $\rho_{\text{TCR}}$  combines physicochemical TCR-facing features with a TCR precursor-frequency estimate in the  $Q = R \times D$  spirit of [11]. The precursor-frequency estimator is a TCR-side quantity and is deferred to a separate package, consumed here only as a score. Each term is a roadmap Phase-2 milestone specified by this subsection.

## 8. BENCHMARK AND EVALUATION METHODOLOGY

We evaluate as a per-pMHC task: the unit is the (epitope, allele) pair. *Exclusion is per-pMHC and benchmark-only*—we hold out test (epitope, allele) pairs and drop *only those pairs* from training; the same epitope presented by a *different* allele is a distinct, legitimate pMHC and stays (so the model may borrow that co-presentation, exactly the cross-allele transfer the diffusion is meant to exploit). We never exclude an epitope “separately” across alleles, and the production model trains and votes on all data—the split exists only to score generalization. Matching counts are *punctured* for self-identity only (a query never counts itself, the  $n^>$  convention of the `seqtree` appendix). *Metrics* are top-1 and top-5 accuracy and per-rarity *recovery@5* (the multi-label, promiscuity-aware score—top-1 is misleading when a peptide binds several alleles), the per-(peptide,allele) ROC/PR-AUC, and a *non-binder* baseline AUROC: the presentation score (max over the panel) of real held-out peptides versus a matched set of 10,000 random peptides drawn at the corpus amino-acid frequency and length distribution. The latter establishes the floor for non-binder rejection (real-vs-random AUROC  $\approx 0.80$  for human MHC-I, diffusion-neutral) and is the negative class for the multi-class (allele + non-binder) confusion matrix realized in Fig. 8 (`bench/confusion.py`). There each held-out peptide is assigned its top-1 allele’s *locus* when the diffused score clears a gate calibrated to a 5% non-binder false-positive rate, else *non-binder*. The model assigns the correct locus where it commits (precision 0.62–0.65 for HLA-A/B/C) and rejects 95% of random peptides by construction, but a single panel-max score gate cannot simultaneously reject non-binders and retain rare positives—top-1 recall falls to 0.17–0.32 at 5% FPR—because maximizing the per-allele log-odds over  $\sim 100$  alleles inflates the family-wise false-positive rate. This is exactly why binder/non-binder calling composes the per-allele score with the global  $E_{\text{glob}}$  gate (§3.1) rather than thresholding one score. *Comparison:* against NetMHCpan-4.1 and NetMHCIIpan-4.0 [14], MixMHCpred [4], and MixMHC2pred [13] on matched alleles and shared eluted-ligand / assay test sets, with rank-threshold calibration and an explicit *rare-allele subgroup* analysis (the equity axis of [8]) where the pseudosequence diffusion is expected to help most. The benchmark datasets and the paper live in a dedicated repository; this section fixes the protocol so it stays consistent with the predictors defined here.

**PERFORMANCE..** Built on the C++ `KmerIndex`, the toolbox is fast enough for proteome-scale use (`bench/bench_speed.py`, one core, human MHC-I shortlist panel of 107 alleles): a one-off store build of 1.4 s; allele restriction at  $\sim 1900$  peptides/s with diffusion (rank all alleles) and  $\sim 4000$ /s by vote; a 596-aa protein scan in 0.3 s; TCR-facing similarity search of  $2 \times 10^4$  peptides in 40 ms; and proteome source lookup in  $\sim 60$  ms, at  $\sim 0.6$  GB peak.

## 9. ASSUMPTIONS AND LIMITATIONS

(i) The analytic pooled null assumes approximate position independence in the anchor block; real co-occupancy adds a two-point ( $b_2$ ) correction (the negative-binomial tail of the `seqtree` appendix) that keeps the mean robust while inflating the tail. (ii) Kernel validity rests on a faithful pseudosequence; ambiguous (X) positions are skipped and locus-aware normalization of allele names is required (class II especially). (iii) The pooled E-value is a smoothed estimator (Prop. 2)—its bias must be reported alongside the gain for rare alleles. (iv) The class-II register trick is a one-pass proxy for full deconvolution [3, 2]. (v) “Compatible allele” in mimicry is a user input; cross-reactivity significance is conditional on a faithful per-allele presented control. These are the open items tracked in `ROADMAP.md`.

## 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] 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.
- [3] Massimo Andreatta, Bruno Alvarez, and Morten Nielsen. GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Research*, 45(W1):W458–W463, 2017.
- [4] Michal Bassani-Sternberg, Chloé Chong, Philippe Guillaume, Marthe Solleder, HuiSong Pak, Philippe O. Gannon, Lana E. Kandalaft, George Coukos, and David Gfeller. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allosteric regulating HLA specificity. *PLOS Computational Biology*, 13(8):e1005725, 2017.
- [5] 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.
- [6] Aaron Clauset, M. E. J. Newman, and Christopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.
- [7] 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.

- [8] Eric Glynn, Dario Gherzi, and Mona Singh. Toward equitable major histocompatibility complex binding predictions. *Proceedings of the National Academy of Sciences USA*, 122(8):e2405106122, 2025.
- [9] 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.
- [10] 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.
- [11] Marta Łuksza, Zachary M. Sethna, Luis A. Rojas, Jayon Lihm, Barbara Bravi, Yuval Elhanati, Kevin Soares, Masataka Amisaki, Anton Dobrin, David Hoyos, Pablo Guasp, Abderezak Zeboudj, Rebecca Yu, Adrienne Kaya Chandra, Theresa Waters, Zagaa Odgerel, Joanne Leung, Rajya Kappagantula, Alvin Makohon-Moore, Amber Johns, Anthony Gill, Mathieu Gigoux, Jedd Wolchok, Taha Merghoub, Michel Sadelain, Erin Patterson, Remi Monasson, Thierry Mora, Aleksandra M. Walczak, Simona Cocco, Christine Iacobuzio-Donahue, Benjamin D. Greenbaum, and Vinod P. Balachandran. Neoantigen quality predicts immunoediting in survivors of pancreatic cancer. *Nature*, 606(7913):389–395, 2022.
- [12] Adam A. Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(S1):S7, 2006.
- [13] Julien Racle, Justine Michaux, Georg Alexander Rockinger, Marion Arnaud, Sara Bobisse, Chloe Chong, Philippe Guillaume, George Coukos, Alexandre Harari, Camilla Jandus, Michal Bassani-Sternberg, and David Gfeller. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nature Biotechnology*, 37(11):1283–1286, 2019.
- [14] Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Research*, 48(W1):W449–W454, 2020.
- [15] 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.

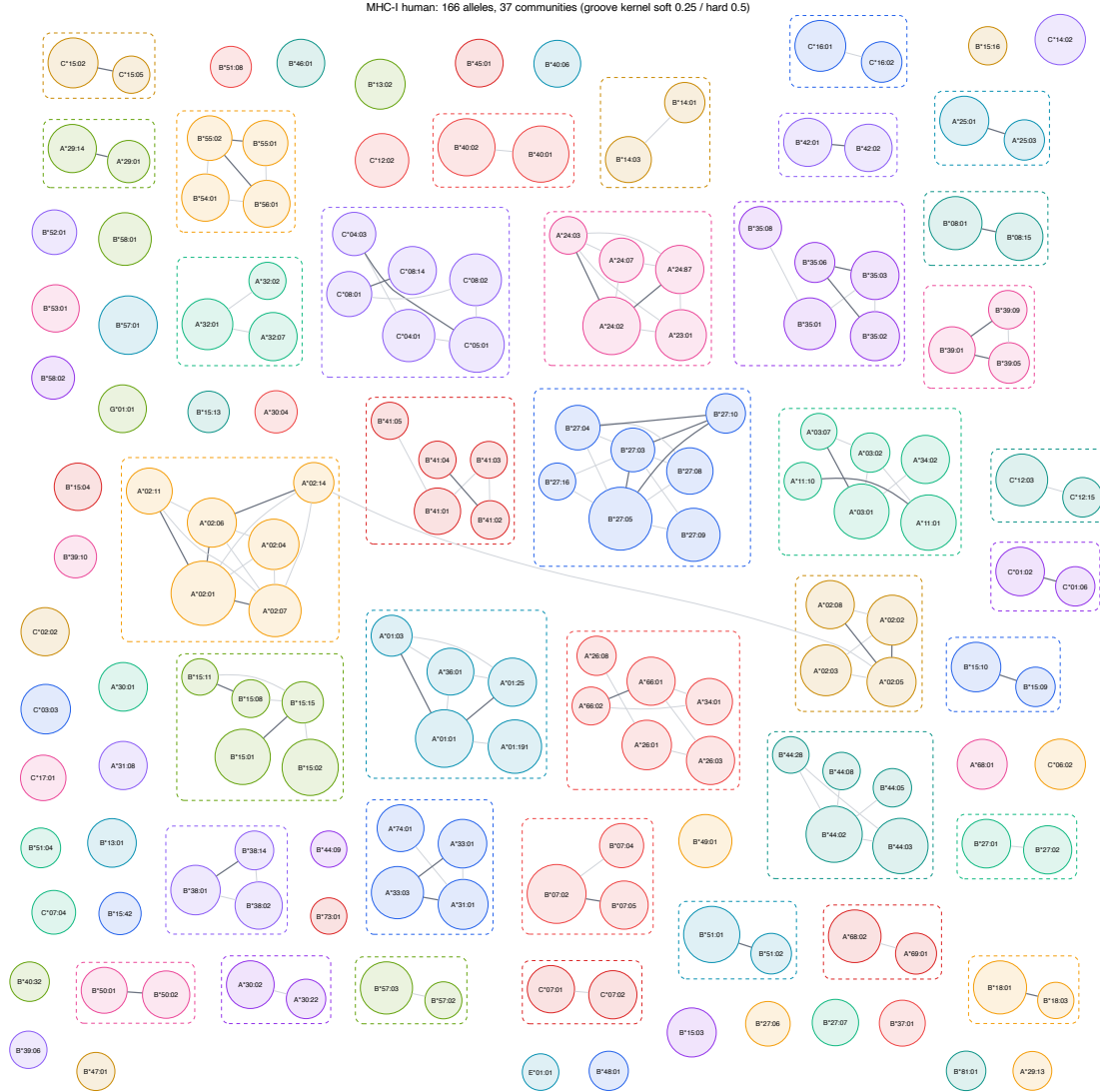


Figure 3: Human MHC-I allele network: nodes are alleles (size  $\propto$  presented-set), edges the MI-weighted BLOSUM groove kernel (4) (thin  $\geq 0.25$ , bold  $\geq 0.5$ ), dashed boxes the greedy-modularity communities. The communities recover known HLA-I supertypes. Mouse and class-II panels are generated likewise (`bench/promiscuity_graph.py`).

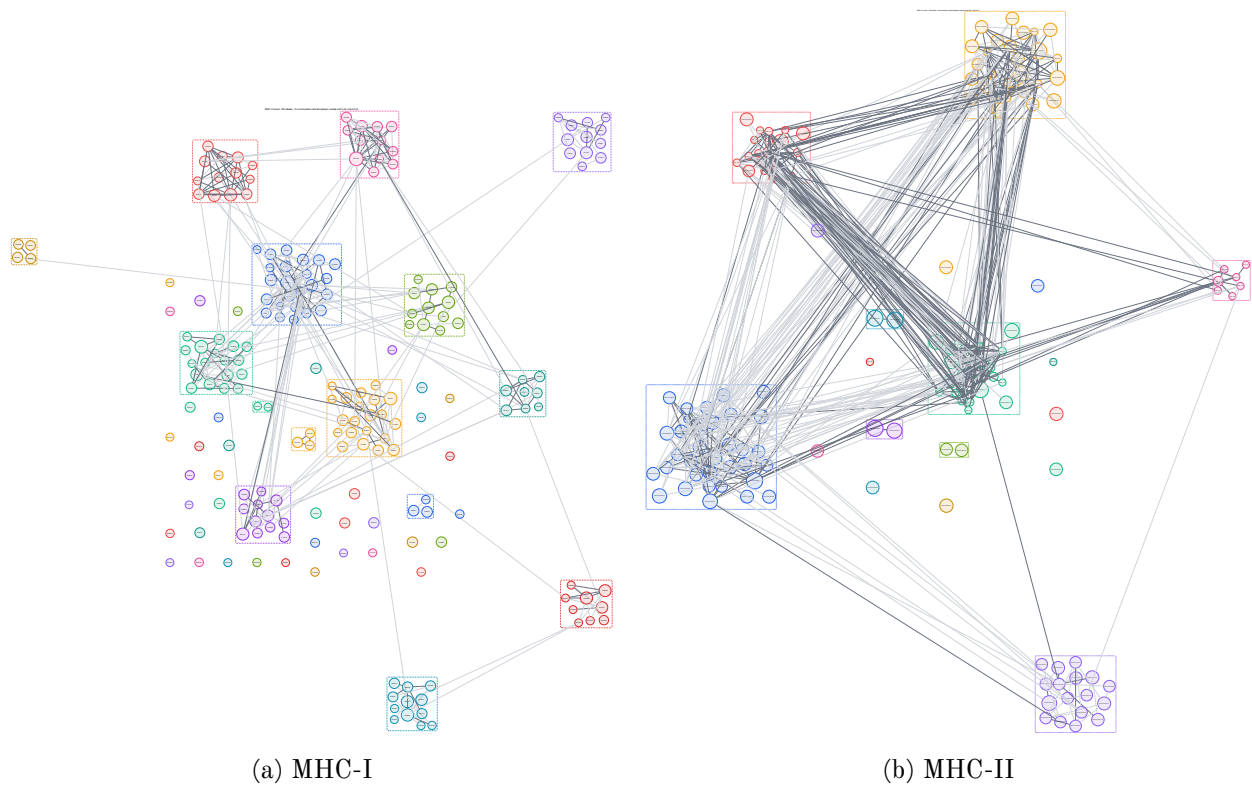


Figure 4: *Observed human* co-presentation networks (shared-epitope overlap coefficient; thin  $\geq 0.25$ , bold  $\geq 0.5$ ; greedy-modularity communities). Communities are coarser than the predicted groove network (Fig. 3) because promiscuous peptides bridge supertypes. `bench/promiscuity_graph.py`.

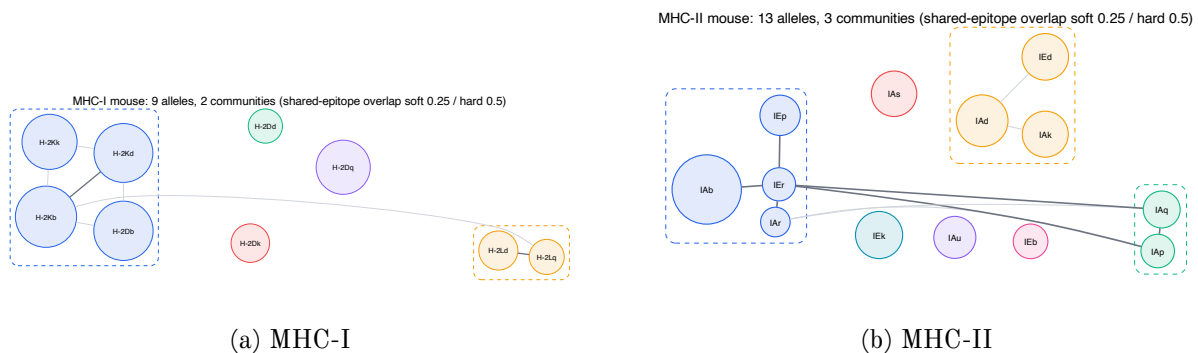


Figure 5: *Observed mouse* co-presentation networks (shared-epitope overlap coefficient; same thresholds and method as Fig. 4). Shared epitopes still link the groove-divergent mouse alleles, where the predicted groove kernel forms no communities. `bench/promiscuity_graph.py`.

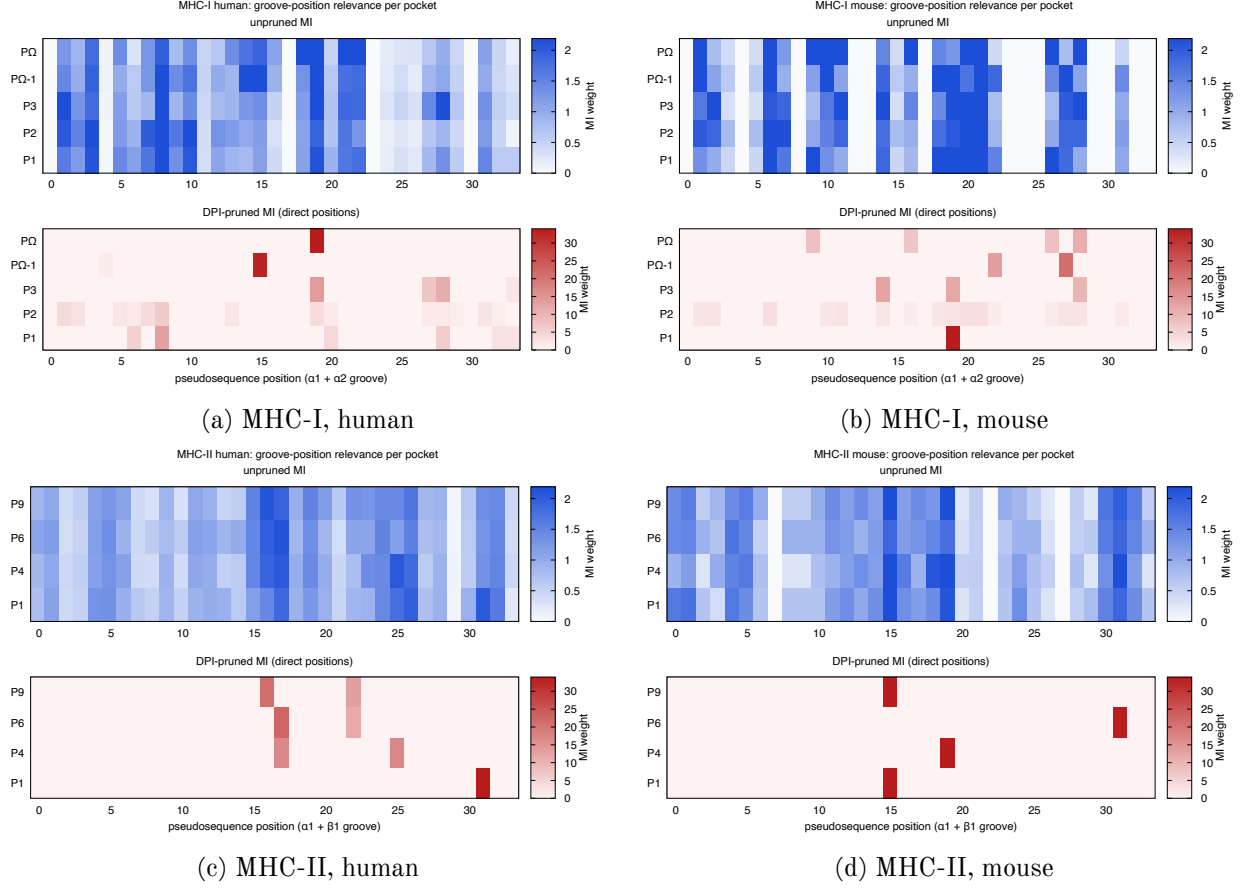


Figure 6: Mutual-information relevance  $w_{j,p}$  of each groove pseudosequence position  $p$  ( $x$ ) for each anchor pocket  $j$  (rows), per class and species. Each panel shows raw MI (*top*, blues) above the data-processing-inequality-pruned [12] weights (*bottom*, reds): raw MI is smeared by linkage between groove positions, while pruning keeps each pocket’s *direct* positions (MHC-I  $PQ \rightarrow \sim 19$ ,  $P2 \rightarrow \sim 7-8$ , disjoint); MHC-II concentrates on  $\beta_1$ . Colour scales are shared across all panels (one for raw, one for pruned) and winsorized at the 90th percentile so the bulk of the scale stays legible. `bench/make_figures.py`.



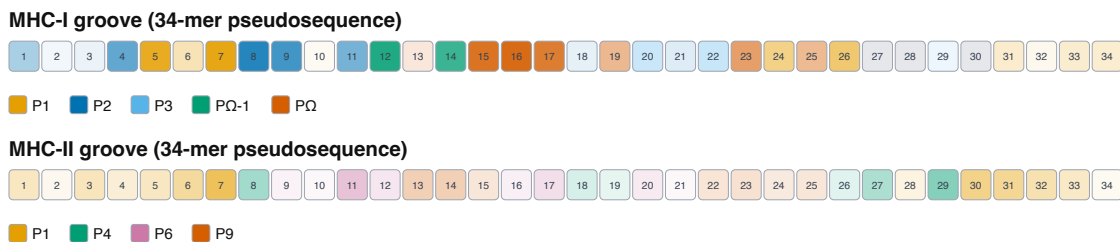


Figure 7: Structural pocket map: each of the 34 groove pseudosequence positions coloured by the peptide anchor it contacts most across the pMHC crystals (§ 4.4, 279 MHC-I / 93 MHC-II structures), with opacity proportional to that heavy-atom contact frequency; pale grey marks positions with no appreciable peptide contact. The B-pocket (P2, ~7–9) and F-pocket (PQ, ~15–17) footprints emerge directly from structure and coincide with the pruned MI maxima of Fig. 6. Rendered in the residue-square style of `tcren`’s complementarity maps. `bench/structural_figure.py`.

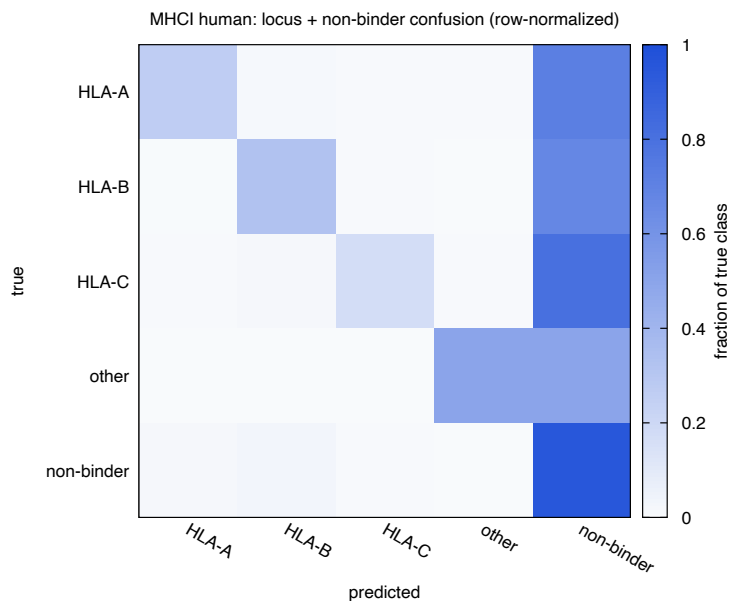


Figure 8: Human MHC-I locus + non-binder confusion (row-normalized; `bench/confusion.py`, shortlist tier). Each held-out peptide is assigned its top-1 allele’s locus when the diffused score clears a gate calibrated to a 5% non-binder false-positive rate, else *non-binder*; the bottom row is 10,000 corpus-AA random peptides (rejected at 95% by construction). Locus assignment is accurate when the model commits (precision 0.62–0.65), but the strict global gate costs top-1 positive recall—motivating the per-allele/global- $E$  composition of § 3.1 over a single score threshold.