Pan-Vertebrate MHC Class I Generalization and Exploratory Pan-Class Transfer in MHCflurry 2.5

Zero-Shot Binding Affinity and Antigen Presentation Across 223 Vertebrate Species, Audited Livestock and Companion-Animal Benchmarks, and an Experimental Class I → Class II Structural Boundary Test

MHCflurry Contributors (OpenVax & PIRL-UNC) · **Date:** November 2026
Software: `mhcflurry 2.5.x` (frozen `2.3.0` weights), `mhcseqs 2.9–2.14`, `mhcgnomes 3.79–3.96`
Web Companion: `www.openvax.org/mhcflurry/whitepaper_25x_cross_species.html`

Abstract & Key Findings

MHCflurry 2.5 integrates the optional `mhcseqs` MHC sequence database and structural aligner alongside `mhcgnomes` allele-name parsing, expanding complete Class I binding-groove coverage nearly five-fold—from 47 vertebrate species (10 trained species / 308 alleles in `2.3.0`) to **223–231 vertebrate species** (26,156–26,549 complete grooves across 14 clades)—**without retraining or altering existing `2.3.0` model weights**. Using rigorous negative controls (**50-allele mismatched real-allele null distributions**, **30-groove column-permuted pseudosequences**, and **within-peptide residue-scrambled positive controls**), we evaluate how frozen `2.3.0` weights generalize across newly added clades, engineered pocket mutants, audited cattle/dog MS cohorts, and an exploratory Class I → Class II transfer test:

- **Broad zero-shot Class I transfer verified against null distributions:** Across 15 `hitlist` unseen-species and binding-pocket mutant cohorts (0 training rows), frozen `2.3.0` weights achieve **0.77–0.98 AUROC** across Carnivora (Cat `0.977`, Giant panda `0.883`), Perissodactyla (Horse `0.967`), Aves (Mallard duck `0.893`–`0.938`), Marsupialia (Brushtail possum `0.869`, Tasmanian devil `0.806 +F`), and engineered pocket mutants (Pig `R156A` `0.951`, $p = 0.020$, beating 50/50 mismatched real alleles; Cat `W167S` `0.976`). Holding amino-acid composition constant via **within-peptide scrambled positives** confirms true positional anchor decoding ($+0.20$ to $+0.29$ AUROC over mismatched alleles on cat, duck `Anpl-UAA*76`, possum, cynomolgus, and pig `R156A`), while cleanly exposing non-specific or indel-limited cases (rabbit `RLA-A*01`, frog `Xela-UAAg`, bat `Ptal-N*01:01`).
- **Cross-clade transfer of human-trained antigen processing:** On Tasmanian devil (*Sarcophilus harrisii*, ∼160 Myr diverged from placentals), MHCflurry's human-trained processing model alone achieves **0.750 AUROC** with real marsupial flanking sequences and lifts combined Presentation from **0.749 → 0.806 AUROC** ($+0.057$, $+30\%$ relative PPV@N; Affinity beats 50/50 size-matched 11-allele mismatched draws, `0.749` vs. `0.695 [0.674, 0.718]`).
- **Mandatory three-table overlap audit on Cattle `BoLA-I` & Dog `DLA-88`:** Auditing against all three bundled `2.3.0` training tables catches **100% in-sample overlap** on `hitlist` Dog `DLA-88` and small IEDB `BoLA-I` cohorts, while isolating **21,332 strictly unseen flanked Cattle `BoLA-I` MS ligands** (`PXD008151`) that reach **0.902–0.960 Presentation (+F) AUROC** (`0.833`–`0.922` on real vs. scrambled ligands vs. `0.604`–`0.731` null) and confirm strong ultra-few-shot transfer (4–12 training rows).
- **Clean negative result on zero-shot Class I → Class II transfer:** Mapping 36/39 Class II `DR/DQ/DP` groove positions onto frozen Class I weights yields **0.551 AUROC** (9-mer core)—indistinguishable from a **30-allele mismatched Class I null (0.549)**, below **column-permuted grooves (0.571)**, and far below `NetMHCIIpan-4.3 BA` (**0.923**). On **within-peptide scrambled Class II binders**, frozen Class I collapses to **0.502** (coin flip) while `NetMHCIIpan-4.3 BA` holds **0.725** ($+0.223$), proving that pan-class prediction requires **joint pan-class training**.
- **Bit-identical training parity:** Locking existing trained alleles via `--lock-existing` reproduces legacy `clustalo` pan-allele training with **Max** $|\Delta| = 0.0$ and **Pearson** $r = 1.000000$ across all 308 trained alleles.

1. Companion Packages, Key Terms, and MHCflurry 2.5 Architecture

Table 1: Companion packages, biological terms, and negative-control designs used in this whitepaper.

Term / Package	Plain-English Meaning
`mhcseqs`	Optional companion library providing curated MHC protein sequences across > 220 vertebrate species and a structural aligner mapping any MHC protein onto standard IMGT numbering.
`mhcgnomes`	Companion MHC allele-name parser used by MHCflurry to normalize naming conventions across species (e.g., `HLA-A0201` $\rightarrow$ `HLA-A*02:01`, plus UniProt-backed names for non-model vertebrates).
`hitlist`	Open benchmark corpus of curated MHC peptide-binding and mass-spectrometry elution datasets paired with matched species reference proteomes.
39-aa pseudosequence	The 39 amino-acid positions of an MHC class I protein that line its peptide-binding pocket (groove) and directly contact the bound peptide.
Flanking sequences (`+F`/`noF`)	The source-protein amino acids immediately upstream (`n_flank`) and downstream (`c_flank`) of a peptide, used by MHCflurry's antigen-processing network to estimate cleavage and transport likelihood.
AUROC & PPV@N	Ranking metrics against 10:1 length-matched background decoys. AUROC (0.5 = random, 1.0 = perfect) measures overall separation; PPV@N is precision among the top N predictions (N = true ligands).
Mismatched real-allele null ($N = 50$)	Scoring the identical test peptides with 50 diverse, unrelated real Class I alleles (excluding the target species; size-matched k-allele random draws for multi-allelic haplotypes). Yields a true null distribution (median, 5–95% band, empirical p-value) without the motif bias of a single control allele.
Column-permuted grooves ($N = 30$)	Synthetic 39-aa pseudosequences created by independently shuffling each of the 39 columns across real Class I alleles, keeping per-position residue frequencies in-distribution while breaking pocket identity.
Within-peptide scrambled positives	Scoring real positive ligands against $10\times$ residue-shuffled copies of those exact same ligands (identical length and amino-acid composition). Eliminates hydrophobicity bias so AUROC above 0.50 measures true positional anchor recognition.
Closed vs. open groove	MHC class I holds peptides in a *closed* pocket pinning both ends (favoring 8–11mers); MHC class II has an *open* groove where an internal 9-mer core binds while both ends of a longer peptide (15+ aa) overhang freely.
Groove insertion	Extra amino acids inserted into the peptide-binding pocket of some non-model species (e.g., a 3-aa insertion in bat `Ptal-N*01:01`) that cannot be represented in a fixed 39-position vector.

In MHCflurry 2.0–2.3, building the 39-residue pseudosequence table required downloading bulk IPD-IMGT/HLA, IPD-MHC, and UniProt FASTA files and running external `clustalo` multiple-sequence alignments. That workflow covered 47 vertebrate species (20,205 alleles, with training measurements concentrated in only 10 species and 308 alleles) and could not register novel heavy chains, non-model UniProt sequences, or engineered binding-pocket mutants at inference time.

MHCflurry 2.5 resolves these limitations without changing any trained `2.3.0` weights:

1. **Direct structural extraction via `mhcseqs`:** Installing `mhcflurry[mhcseqs]` delegates signal-peptide trimming, binding-domain extraction, and mature-numbered 39-aa pseudosequence extraction to `mhcseqs`'s structural aligner. Users can register any raw vertebrate Class I heavy chain or mature-numbered point mutation at runtime (`add_allele_from_sequence(...,mutations=("W167S",))`) or extend an entire predictor (`extend_alleles_from_mhcseqs()`).
2. **Automatic insertion warnings:** When `mhcseqs` $\geq$ `2.9` encounters extra amino acids inserted inside a peptide-binding pocket (which a fixed 39-position vector must skip), MHCflurry exposes a structured `GrooveAlignment` object and logs an explicit warning rather than silently returning an out-of-register pseudosequence.
3. **Sequence-backed naming for non-model species:** Through `mhcgnomes` (`AccessionedAllele` support in `mhcflurry` $\geq$ `2.5.9`), sequence-backed vertebrate identifiers (e.g., shark `TriakisScyllium-UAA*uniprot:046817`, Tasmanian devil `Saha-classI*uniprot:G3V1L5`) flow cleanly through `normalize_allele_name` and `Class1AffinityPredictor` under `normalize_alleles=True` (26,156 alleles across 223 canonical vertebrate species, including 2,884 accessioned entries and 0 unparsed rows), while blocking bare UniProt accessions from misparsing as mammalian alleles.
4. **Zero-regression parity gate:** Across 19,886 overlapping non-`X` alleles in `2.3.0`, `mhcseqs` achieves **99.45% exact 39/39 match** (19,777/19,886) and **99.98% per-residue identity** (775,385/775,554). The updated `downloads-generation/allele_sequences/GENERATE.sh` uses `--lock-existing` plus `verify_trained_allele_parity.py` to enforce 0 character differences across all 308 trained alleles.

Complete Class I coverage expands from 47 species in `2.3.0` to 227 species in `mhcseqs 2.12` (231 union, 26,549 complete grooves, **+184 net new vertebrate species** across 14 clades; **Figure 1**). Experimental validation data across these clades in `hitlist` span proteome-flanked cellular MS elutions, direct ligands, experimental positive/negative

binding assays, and pocket mutants (**Figure 2**).

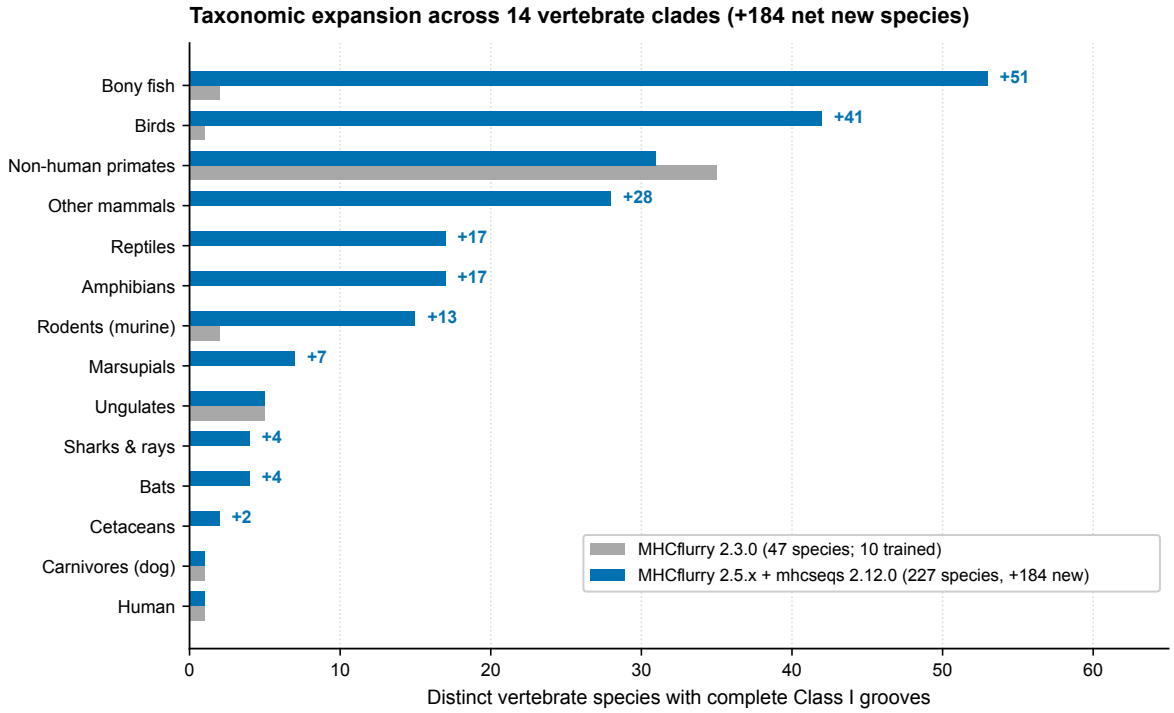

Figure 1: Taxonomic expansion of complete MHC class I binding grooves across 14 vertebrate clades. Distinct vertebrate species with complete 39-aa pseudosequences increase from 47 in MHCflurry 2.3.0 (10 trained) to 227 in `mhcseqs 2.12` (231 union; +184 net new vertebrate species), led by bony fish (+51), birds (+41), other mammals (+28), reptiles (+17), and amphibians (+17).

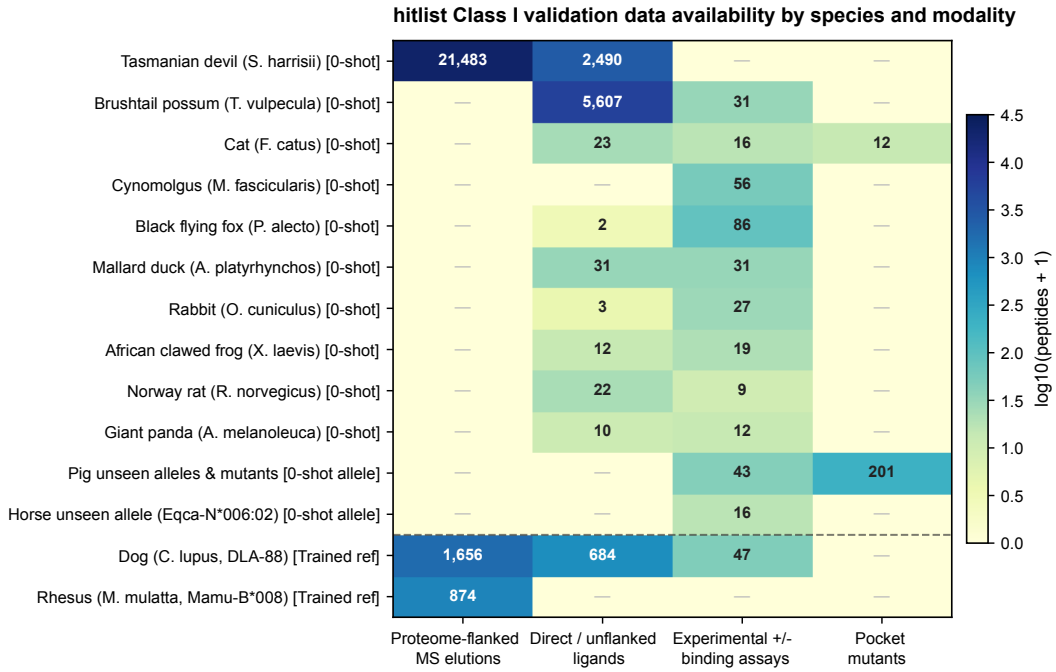

Figure 2: `hitlist` Class I validation data availability by species and experimental modality. Log-scaled peptide counts across proteome-flanked MS elutions, direct/unflanked ligands, experimental +/− binding assays, and engineered pocket mutants for 10 unseen vertebrate species, 2 unseen allele/mutant cohorts in trained species, and 2 trained MS reference controls (below dashed line).

2. Zero-Shot Class I Affinity and Antigen Presentation Across Unseen Clades

2.1 Zero-Shot Presentation (+ Real Species Proteome Flanks) on MS Elutions

MHCflurry's presentation model combines pan-allele binding affinity with a species-agnostic (human-trained) convolutional antigen-processing predictor that inspects each peptide and its upstream/downstream source-protein flanking sequences. Using cellular MS immunopeptidomics datasets in `hitlist` with verified species reference proteomes (*Sarcophilus harrisii*, *Canis lupus*, *Macaca mulatta*), we paired mapped eluted ligands against 10:1 length-matched decoys sampled from the same host proteome with symmetric real flanking sequences, benchmarking frozen `2.3.0` weights against size-matched mismatched real-allele nulls (**Figures 3–4** and **Table 2**).

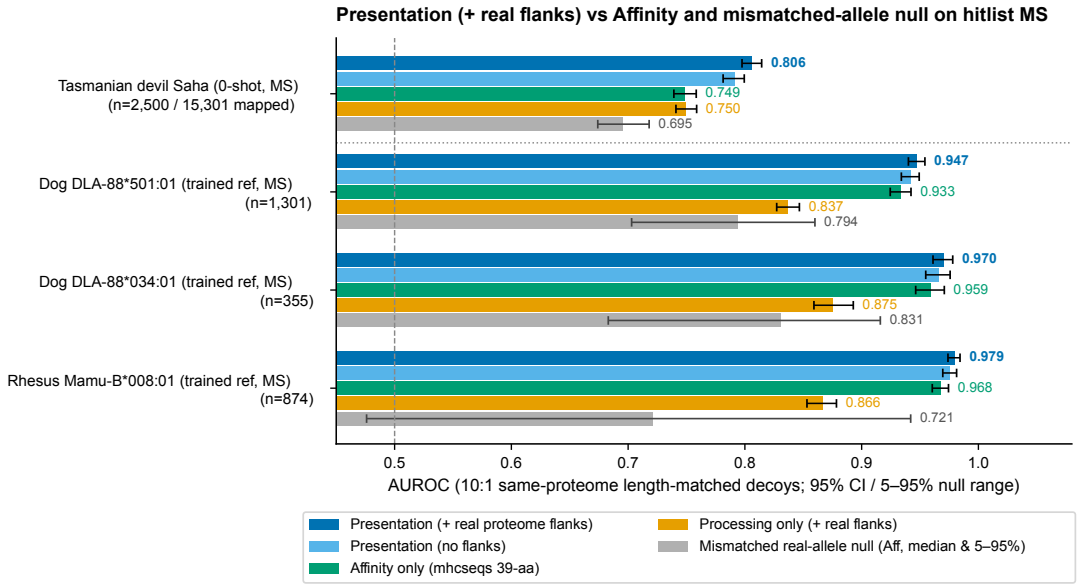

Figure 3: Zero-shot Presentation (+ real species proteome flanks) versus Affinity and mismatched-allele null AUROC on `hitlist` MS elutions. AUROC (95% bootstrap CI) against 10:1 same-proteome length-matched decoys for zero-shot Tasmanian devil (*Sarcophilus harrisii*, 0 training rows, 11 UniProt grooves, 15,301 mapped ligands) alongside trained Dog (`DLA-88`) and Rhesus macaque (`Mamu-B`) MS reference controls. Gray bars show the size-matched mismatched real-allele Affinity null (median and 5–95% range). Across ∼160 Myr to marsupials, human-trained processing alone reaches 0.750 AUROC and lifts Presentation from 0.749 to 0.806 (+0.057).

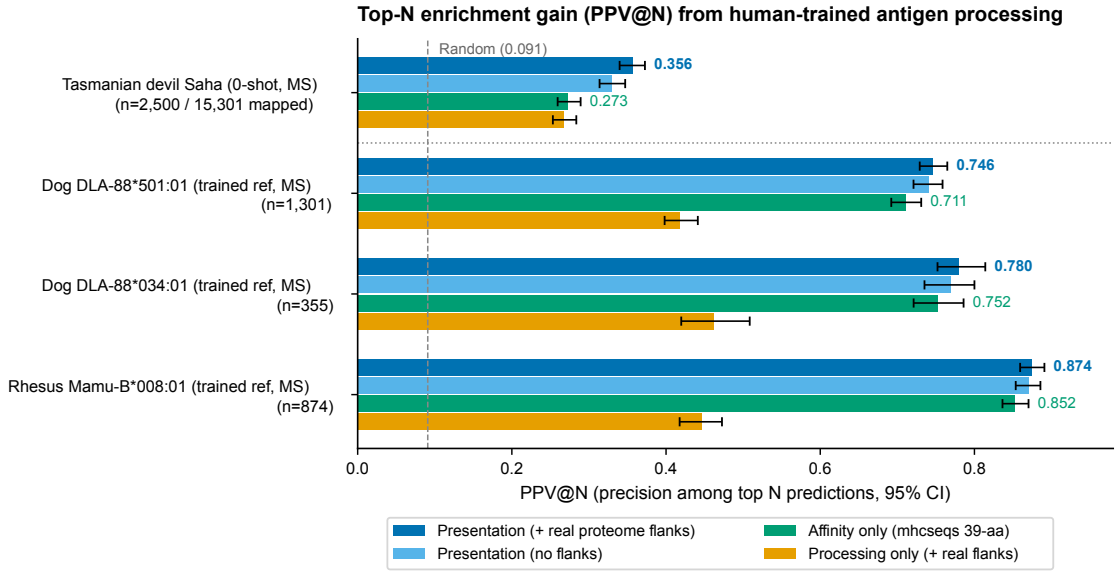

Figure 4: Top-N enrichment gain (`PPV@N`) from human-trained antigen processing on `hitlist` MS elutions. Precision among the top N predictions (dashed line: 0.091 random baseline at 10:1 decoy ratio). Adding human-trained processing and real *S. harrisii* flanks lifts Tasmanian devil `PPV@N` from 0.273 to 0.356 (+30% relative enrichment).

Table 2: Zero-shot Presentation (+ real proteome flanks) versus Affinity and size-matched mismatched-allele null on `hitlist` MS elutions. All evaluations use 10:1 length-matched decoys with real flanking sequences drawn from the matching host proteome and frozen `2.3.0` weights.

Dataset (`hitlist` MS)	Status	Aff.	Null Med [5–95%]	Proc. (+F)	Pres. (+F)	PPV@N (Aff→+F)
Tas. devil `Saha` (11 gr., 15k)	**0-shot Marsup.**	0.749	0.695 [0.674, 0.718]	**0.750**	**0.806** (+0.057)	**0.273→0.356** (+30%)
Dog `DLA-88*501:01` (1,301)	Trained ref†	0.933	0.794 [0.703, 0.860]	0.837	**0.947** (+0.014)	0.711→0.746
Dog `DLA-88*034:01` (355)	Trained ref	0.959	0.831 [0.683, 0.916]	0.875	**0.970** (+0.011)	0.752→0.780
Rhesus `Mamu-B*008:01` (874)	Trained ref	0.968	0.721 [0.476, 0.942]	0.866	**0.979** (+0.011)	0.852→0.874

†As established by our mandatory three-table audit in Section 4, IEDB `DLA-88*501:01` peptides overlap `2.3.0`'s bundled affinity training table; these rows serve strictly as trained positive-control ceilings, not held-out canine benchmarks.

2.2 Experimental Positive/Negative Binding Assays and 15-Cohort Zero-Shot Ranking

Non-MS datasets in `hitlist`—including in-vitro refolding assays, purified-MHC fluorescence screens, synthetic random-library elution panels, and viral/mutant binding assays—lack cellular proteasome/TAP processing and have minimal host-proteome mapping ($19/448 = 4.2\%$ mapped in `binding.parquet`; 0% for the 5,607 synthetic random-library `Trvu-UB*01:01` ligands). For these cohorts, **Affinity** is the primary metric, evaluated both on experimental positive-versus-negative binding assays with no synthetic decoys (**Figure 5**) and on 10:1 length-matched background ligand ranking (**Figure 6** and **Table 3**), benchmarked against a **50-allele mismatched real-allele null distribution** (stratified across `HLA-A/B/C` and non-human Class I alleles, excluding the target species).

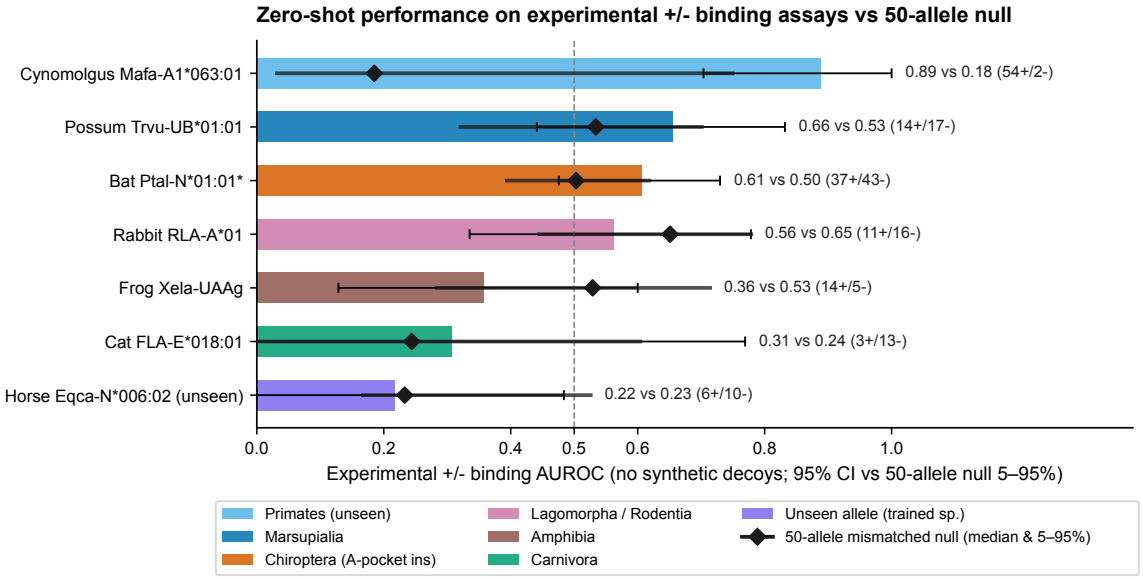

Figure 5: Zero-shot performance on experimental positive-versus-negative binding assays (no synthetic decoys) versus 50-allele mismatched null. AUROC (95% bootstrap CI) on measured binders versus measured non-binders from `hitlist/binding.parquet`, colored by clade, alongside the 50-allele mismatched real-allele null (black diamond = median; horizontal bar = 5–95% range).

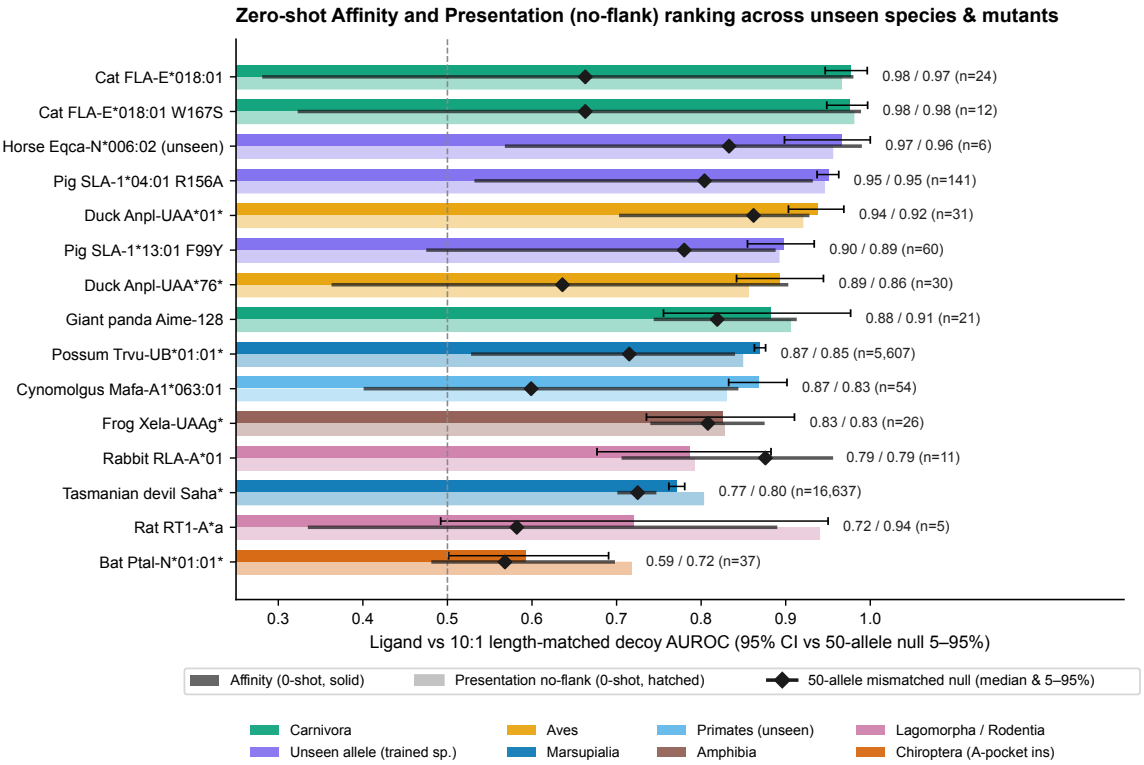

Figure 6: Zero-shot Affinity and no-flank Presentation ligand ranking across 15 `hitlist` unseen-species and pocket-mutant cohorts. Solid bars show zero-shot Affinity AUROC (95% CI); hatched bars show no-flank Presentation AUROC; black diamonds and whiskers mark the 50-allele mismatched real-allele null (median and 5–95% range; size-matched 11-allele draws for Tasmanian devil Saha). Asterisks (∗) indicate alleles carrying internal groove insertions or deletions flagged by `mhcseqs`.

Table 3: Complete zero-shot evaluation with proper negative controls across 15 `hitlist` unseen-species and binding-pocket mutant cohorts. Every target allele has 0 training rows in `2.3.0`. Mismatched null reports median [5–95%] across 50 diverse real Class I alleles (excluding target species; size-matched 11-allele draws for Saha), with empirical one-sided p-value. Real vs. Scrambled compares real ligands against 10× within-peptide residue-shuffled copies (target vs. mismatched null median).

Unseen Species / Allele (Clade)	Aff / Pres	50-Allele Null [5–95%]	Beats (p)	Col-Perm	Scrambled (Null)
Cat `FLA-E*018:01` ($n = 24$)	**0.977** / 0.966	0.663 [0.283, 0.978]	46/50 (0.098)	0.662	**0.957** (0.703)
Cat `FLA-E*018:01 W167S` ($n = 12$)	**0.976** / **0.981**	0.663 [0.325, 0.987]	45/50 (0.118)	0.677	**0.957** (0.664)
Horse `Eqca-N*006:02` ($n = 6$)	**0.967** / 0.956	0.833 [0.570, 0.988]	41/50 (0.196)	0.876	**0.919** (0.821)
Pig `SLA-1*04:01 R156A` ($n = 141$)	**0.951** / 0.946	0.804 [0.534, 0.930]	**50/50 (0.020)**	0.807	**0.890** (0.688)
Duck `Anpl-UAA*01*` ($n = 31$)	**0.938** / 0.920	0.862 [0.705, 0.926]	**50/50 (0.020)**	0.900	**0.876** (0.804)
Pig `SLA-1*13:01 F99Y` ($n = 60$)	**0.898** / 0.892	0.780 [0.477, 0.886]	**49/50 (0.039)**	0.740	**0.852** (0.706)
Duck `Anpl-UAA*76*` ($n = 30$)	**0.893** / 0.856	0.636 [0.365, 0.901]	46/50 (0.098)	0.637	**0.848** (0.620)
Giant panda `Aime-128` ($n = 21$)	**0.883** / **0.906**	0.819 [0.746, 0.911]	37/50 (0.275)	0.869	**0.866** (0.769)
Possum `Trvu-UB*01:01*` ($n = 5{,}607$)	**0.869** / 0.849	0.715 [0.530, 0.838]	**49/50 (0.039)**	0.725	**0.754** (0.542)
Cynomolgus `Mafa-A1*063` ($n = 54$)	**0.868** / 0.830	0.599 [0.403, 0.842]	48/50 (0.059)	0.645	**0.763** (0.536)
Clawed frog `Xela-UAAg*` ($n = 26$)	0.826 / 0.828	0.808 [0.742, 0.873]	36/50 (0.294)	0.832	0.812 (0.769)
Rabbit `RLA-A*01` ($n = 11$)	0.787 / 0.792	0.876 [0.708, 0.954]	8/50 (0.843)	0.905	0.734 (0.819)
Tas. devil Saha* (11 gr., $n = 23{,}973$)	**0.772** / **0.806**‡	0.725 [0.703, 0.745]	**50/50 (0.020)**	0.732	0.676 (0.668)
Norway rat `RT1-A*a` ($n = 5$)	0.720 / **0.940**	0.582 [0.337, 0.888]	36/50 (0.294)	0.522	0.752 (0.548)
Bat `Ptal-N*01:01*` (3-aa ins, $n = 37$)	0.593 / 0.717	0.568 [0.483, 0.696]	31/50 (0.392)	0.597	0.540 (0.529)

‡0.803 noF; 0.806 with real *S. harrisii* flanks. Separate IEDB feline MS cohorts `FLA-E*003:01` ($n = 1{,}612$) and `FLA-K*007:01` ($n = 491$) achieve 0.830 and 0.880 AUROC zero-shot (see Figure 9).

3. Groove Specificity, Positional Anchor Controls, Motifs, and Training Parity

3.1 Mismatched-Allele Nulls and Within-Peptide Scrambled Controls

A single fixed human control allele (such as `HLA-A*02:01`) is noisy because it represents one specific hydrophobic P2/P9 groove: alleles with overlapping motifs look artificially weak against it, while alleles with anti-correlated motifs look artificially strong. Benchmarking each unseen cohort against **two complementary null controls** resolves this cleanly (**Figures 7–8**):

1. **Mismatched real-allele & column-permuted groove nulls (Figure 7):** Fourteen of fifteen cohorts exceed the 50-allele mismatched median, and five large/clean cohorts achieve empirical $p \leq 0.05$ (49–50/50 unrelated real alleles beaten): Pig `SLA-1*04:01 R156A` ($p = 0.020$), Duck `Anpl-UAA*01` ($p = 0.020$), Tasmanian devil `Saha` ($p = 0.020$ against size-matched 11-allele draws), Brushtail possum `Trvu-UB*01:01` ($p = 0.039$), and Pig `SLA-1*13:01 F99Y` ($p = 0.039$), with Cynomolgus `Mafa-A1*063:01` at 48/50 ($p = 0.059$, and $p = 0.020$ on experimental $+/-$ binding). Column-permuted synthetic grooves ($N = 30$, open squares) closely track the real-allele null.
2. **Within-peptide residue-scrambled positive controls (Figure 8):** Comparing real ligands against $10\times$ residue-shuffled copies of those exact ligands holds length and amino-acid composition strictly constant. Strong positive lifts over mismatched alleles confirm true positional anchor decoding on Cat `FLA-E*018:01 W167S` (0.957 vs. 0.664, $+0.293$), Cat `FLA-E*018:01` (0.957 vs. 0.703, $+0.254$), Duck `Anpl-UAA*76` (0.848 vs. 0.620, $+0.228$), Cynomolgus `Mafa-A1*063:01` (0.763 vs. 0.536, $+0.227$), Brushtail possum `Trvu-UB*01:01` (0.754 vs. 0.542, $+0.212$), and Pig `SLA-1*04:01 R156A` (0.890 vs. 0.688, $+0.202$).
3. **Honest caveats exposed:** Rabbit `RLA-A*01` sits below the 50-allele null (0.787 vs. 0.876, 8/50; scrambled 0.734 vs. 0.819), African clawed frog `Xela-UAAg` ($\sim$350 Myr, 2 indels) sits inside the null (0.826 vs. 0.808, $p = 0.294$), and Bat `Ptal-N*01:01` (3-aa `MDL` pocket insertion) collapses to 0.540 vs. 0.529 null on scrambled peptides.

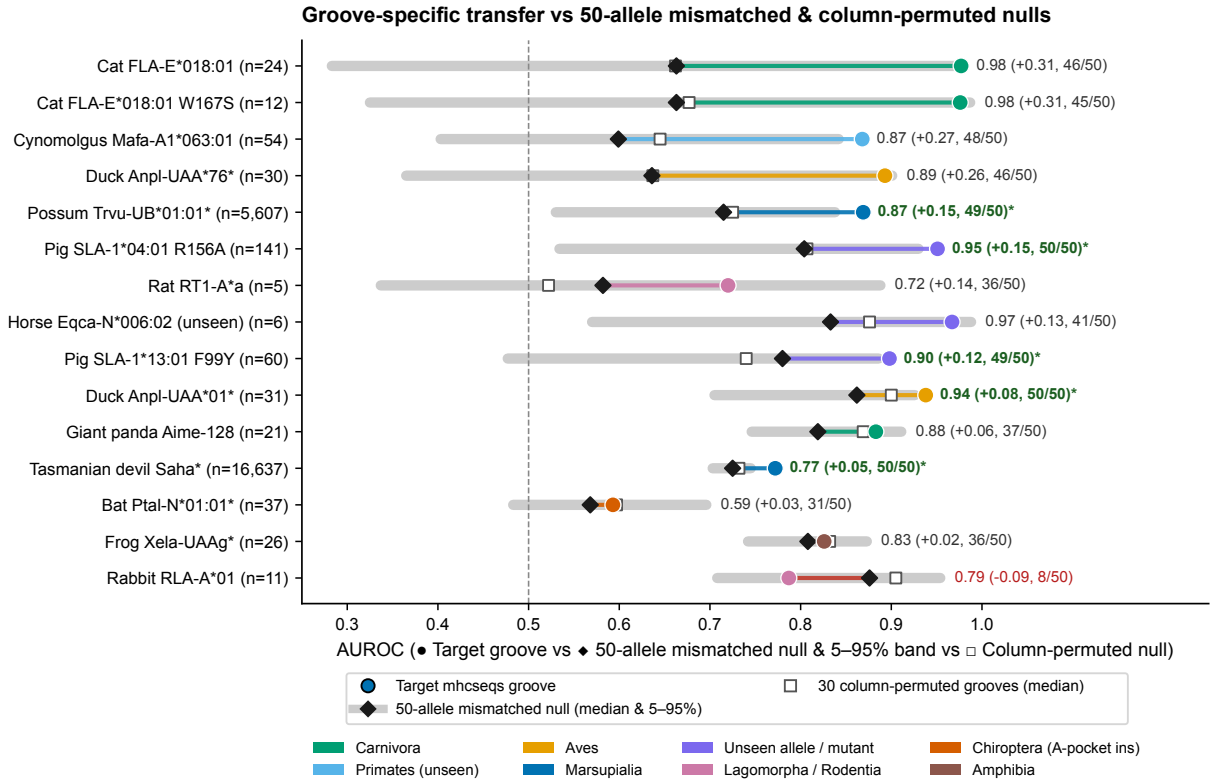

Figure 7: Groove-specific zero-shot transfer versus 50-allele mismatched real-allele and 30 column-permuted pseudosequence nulls. Colored circles show target-allele zero-shot AUROC; gray bars and black diamonds mark the 50-allele mismatched real-allele null (5–95% band and median); open squares mark the median of 30 column-permuted synthetic grooves. Asterisks (∗) denote empirical $p \leq 0.05$.

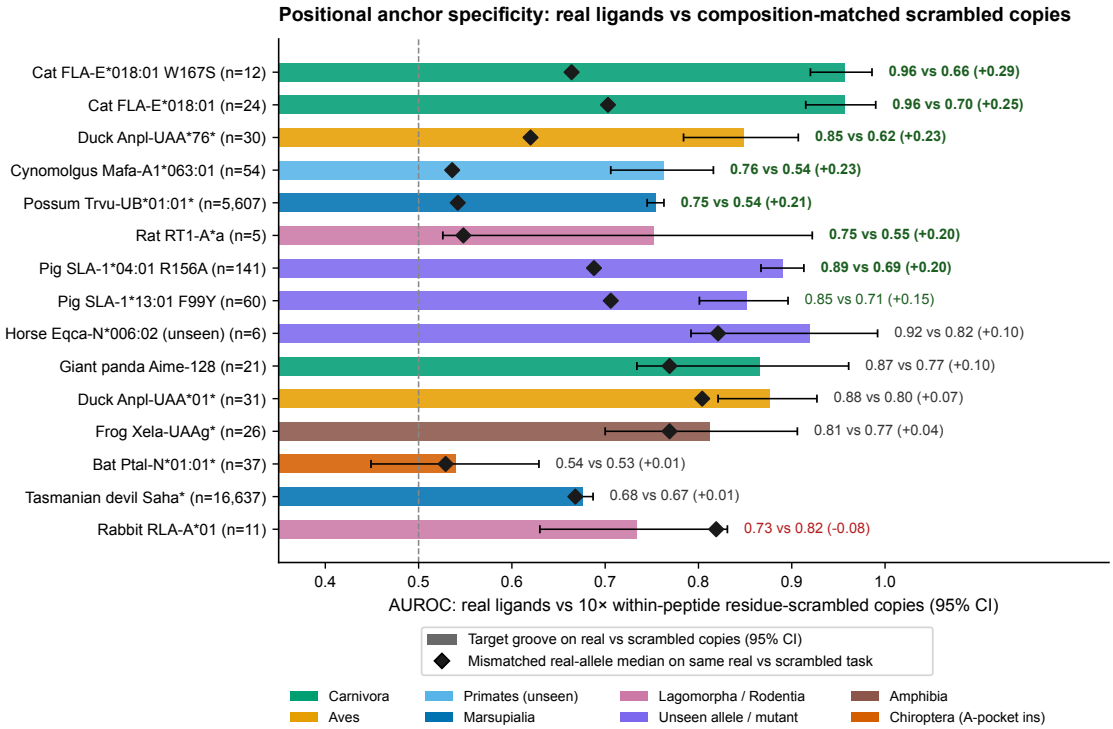

Figure 8: Positional anchor specificity: real ligands versus $10\times$ within-peptide residue-scrambled copies. Colored bars show target-groove AUROC (95% bootstrap CI) when separating real ligands from their own composition-matched shuffled copies; black diamonds show the mismatched real-allele median on the exact same task, sorted by lift Δ.

3.2 Empirical vs. Zero-Shot Sequence Logos and Insertion Failure Mode

Comparing empirical 9-mer MS ligands against the top 1% of 30,000 random 9-mers scored zero-shot (**Figure 9**) explains both why transfer succeeds and where fixed 39-aa pseudosequences hit their physical ceiling:

- **Feline & Marsupial anchor recovery:** With zero feline training data, `FLA-E*003:01` recovers both its empirical P2-Pro and P9-Arg/Lys anchor motif ($n = 1{,}610$ 9-mers). Across $\sim$160 Myr to marsupials, Brushtail possum `Trvu-UB*01:01` ($n = 5{,}597$) and Tasmanian devil `Saha` recover aromatic/hydrophobic P9 anchors.
- **Single-residue pocket-mutant sensitivity:** Extracting mature-numbered point mutations via `mhcseqs` improves AUROC over the corresponding wild-type groove on mutant-specific ligands for both Pig `SLA-1*04:01 R156A` (ΔAUROC $= +0.010$, 95% paired CI $[+0.007, +0.013]$) and Cat `FLA-E*018:01 W167S` (ΔAUROC $= +0.006$, 95% paired CI $[+0.000, +0.015]$).
- **Honest failure mode — Bat `Ptal-N*01:01` (3-aa `MDL` pocket insertion):** Black flying fox `Ptal-N*01:01` carries three extra amino acids (`MDL`) inserted at position 51 inside the N-terminal binding pocket (driving empirical P1-Asp and P9-Pro specificity) alongside `Y171H`. Because a fixed 39-aa vector must omit those three inserted residues, predicted affinities flatten to $\sim$26–30 μM ($1.1\times$ separation; 0.540 on real vs. scrambled ligands). `mhcseqs` $\geq$ `2.9` automatically detects and warns on every such insertion.

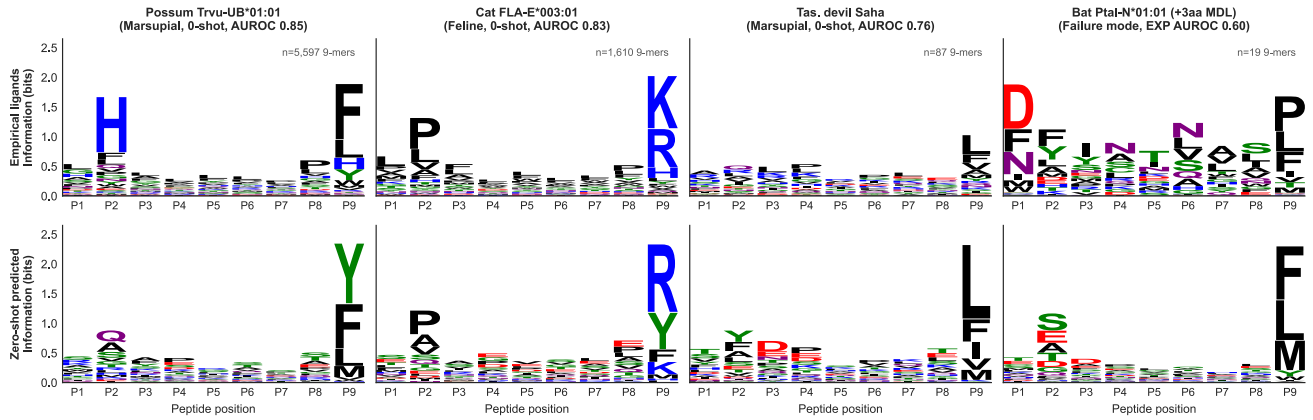

Figure 9: Empirical versus zero-shot predicted 9-mer binding motifs. Observed 9-mer ligands (top row) compared against the top 1% zero-shot predicted 9-mers out of 30,000 random background peptides (bottom row) for Brushtail possum `Trvu-UB*01:01`, Cat `FLA-E*003:01`, Tasmanian devil `Saha`, and the Bat `Ptal-N*01:01` 3-aa insertion failure mode.

3.3 Controlled A/B Pan-Allele Training Parity (`--lock-existing`)

In a controlled A/B pan-allele training run ($[1024, 1024]$ architecture, 30,760 train / 7,690 held-out test across all 307 trained alleles, seed 42; **Figure 10** and **Table 4**), `--lock-existing` reproduced legacy `clustalo` training bit-for-bit while adding all $+184$ net new vertebrate species.

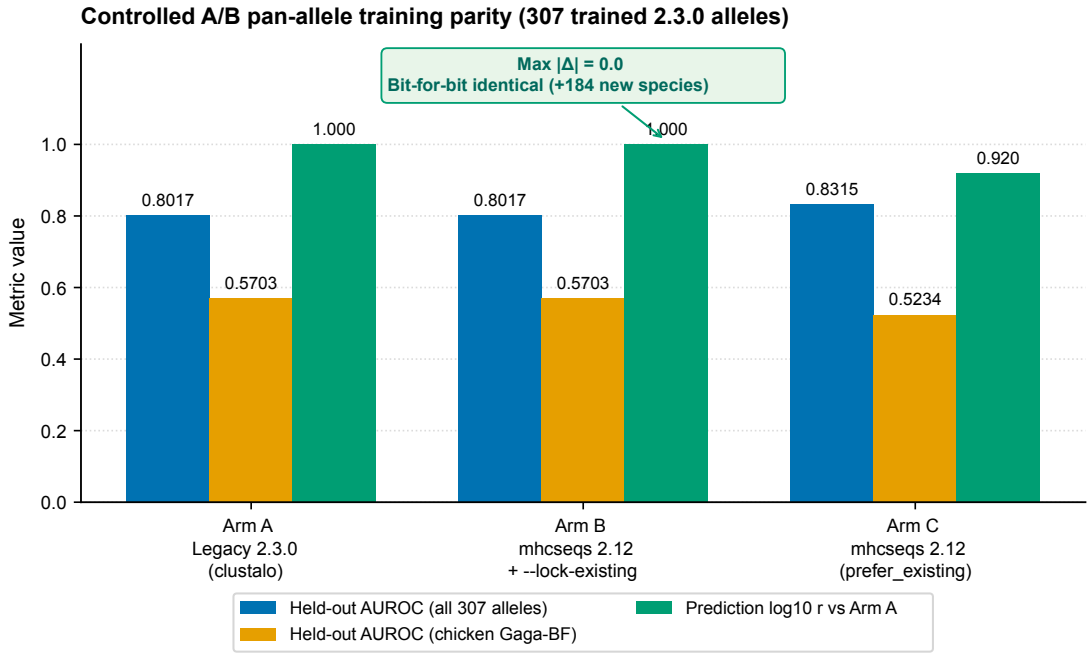

Figure 10: Controlled A/B pan-allele training parity across all 307 trained `2.3.0` alleles. Locking existing trained alleles verbatim (Arm B: `mhcseqs + --lock-existing`) reproduces legacy `clustalo` training (Arm A) bit-for-bit (Max $|\Delta| = 0.0$, $\log_{10} r = 1.000000$, AUROC $= 0.80169$) while expanding the table by $+184$ vertebrate species.

Table 4: Controlled A/B pan-allele training parity validating the `GENERATE.sh --lock-existing` migration.

Arm	Pseudosequence Table	Held-out AUROC	Chicken Gaga	Max $\lvert\Delta\rvert$ vs A	$\log_{10} r$ vs A
Arm A	Legacy `2.3.0` `clustalo`	0.80169	0.57025	0.0	1.000000
Arm B	`mhcseqs` + `--lock-existing`	**0.80169**	**0.57025**	**0.0**	**1.000000**
Arm C	`mhcseqs` (`prefer_existing`, unlocking 7 `Gaga-BF`)	0.83151	0.52342	non-zero	0.920

4. Audited Livestock and Companion-Animal Validation: Cattle `BoLA-I` and Dog `DLA-88`

Evaluating released pan-allele weights on refreshed public corpora carries a subtle contamination hazard: checking only `curated_training_data.affinity.csv.bz2` misses mass-spectrometry ligands embedded inside the actual bundled training tables of `models_class1_presentation/models/`—specifically `affinity_predictor_train_data.csv.bz2` (811,134 rows), `processing_predictor_with_flanks_train_data.csv.bz2` (399,392 rows), and `processing_predictor_no_flank_train_data.csv.bz2` (399,392 rows).

Auditing the refreshed November 2026 `hitlist` Cattle and Dog cohorts against all three tables (**Figure 11** and **Table 5**) immediately separated in-sample cohorts from genuinely unseen data:

- **Excluded (100% in-sample):** All 1,959 `DLA-88*501:01` peptides (PMID 27893789), all 50 `DLA-88*508:01` peptides (PMID 29475511), and all 48 IEDB Cattle `BoLA-I` peptides from PMID 36423003 already appear verbatim in `affinity_predictor_train_data.csv.bz2`. This confirms issue `#490`: current public IEDB canine MS data cannot support held-out `DLA-88` accuracy claims.
- **Strictly unseen Cattle `BoLA-I` (`PXD008151`, 21,332 flanked ligands):** Stripping every exact peptide present in any of the three `2.3.0` training tables leaves 5,811 monoallelic `BoLA-6*013:01` ligands and 15,521 multi-allelic haplotype ligands mapped to the *Bos taurus* UniProt proteome (51,509 proteins). Evaluated against $10\times$ unseen, length-matched bovine decoys with symmetric real 5-aa flanks (**Figure 12**), all three cohorts achieve **0.902–0.960 Presentation (+F) AUROC**—well above size-matched mismatched real-allele nulls (`0.691`–`0.752`)—and **0.833–0.922** on real vs. residue-scrambled bovine ligands (vs. `0.604`–`0.731` null). Real bovine flanks lift AUROC on every cohort ($+0.003$ to $+0.006$) and lift multi-allelic PPV@N by $+4.1\%$ and $+4.2\%$.
- **Ultra-few-shot transfer (4–12 training rows):** In haplotype ablations (**Figure 13**), adding `BoLA-3*002:01` (only 12 training rows in `2.3.0`) lifts the 2-allele haplotype by $+0.041$ Affinity AUROC ($+0.029$ Presentation +F) and claims 42.8% of unseen ligands as top binder; `BoLA-4*024:01` (only 6 training rows) reaches **0.904 Presentation (+F) AUROC** alone, and `BoLA-2*025:01` (4 training rows) reaches **0.850**.

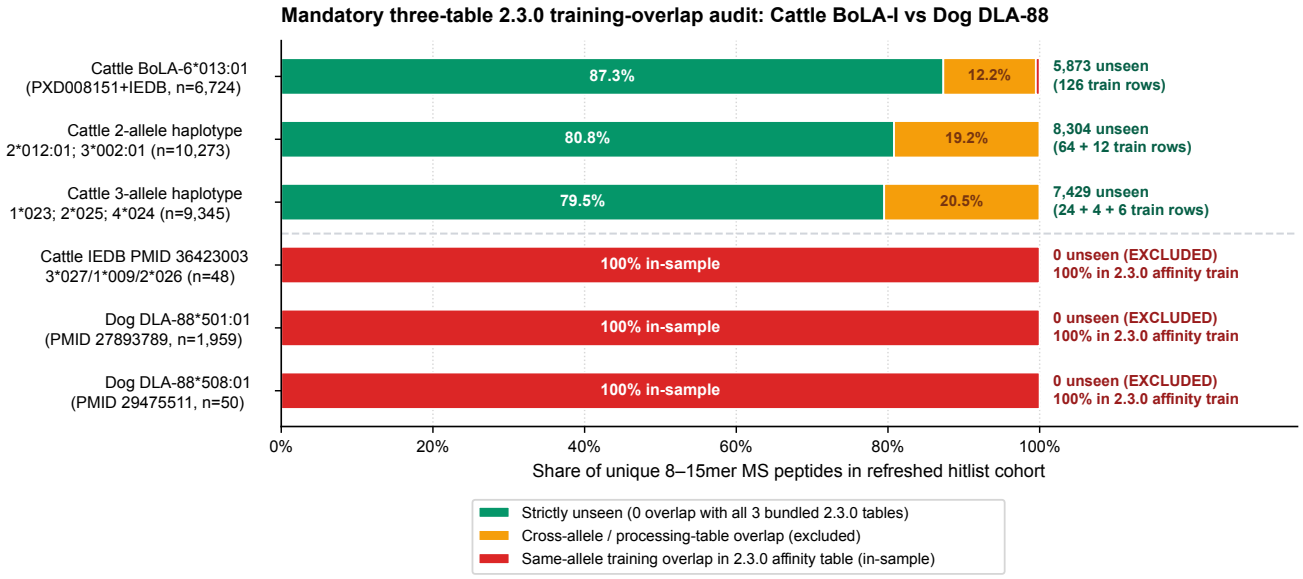

Figure 11: Mandatory three-table `2.3.0` training-overlap audit for Cattle `BoLA-I` and Dog `DLA-88`. Every `hitlist` Dog `DLA-88` MS peptide and small IEDB Cattle `BoLA-I` cohort (PMID 36423003) is 100% in-sample and excluded from held-out claims, whereas 79.5%–87.3% (21,606 unique; 21,332 flanked) of `PXD008151` Cattle `BoLA-I` ligands have zero overlap with any bundled `2.3.0` training table.

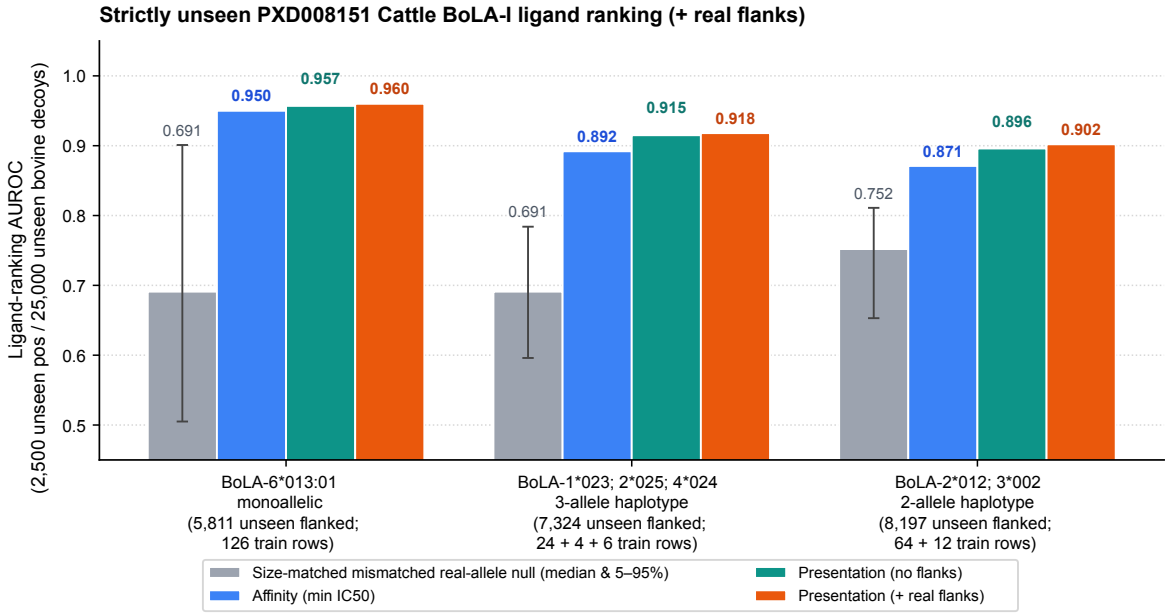

Figure 12: Strictly unseen `PXD008151` Cattle `BoLA-I` ligand ranking (+ real *Bos taurus* flanks). Evaluated on 2,500 strictly unseen flanked ligands per cohort against 25,000 (10×) unseen bovine proteome decoys with real 5-aa N/C flanks, alongside size-matched mismatched real-allele null distributions (median and 5–95% whiskers).

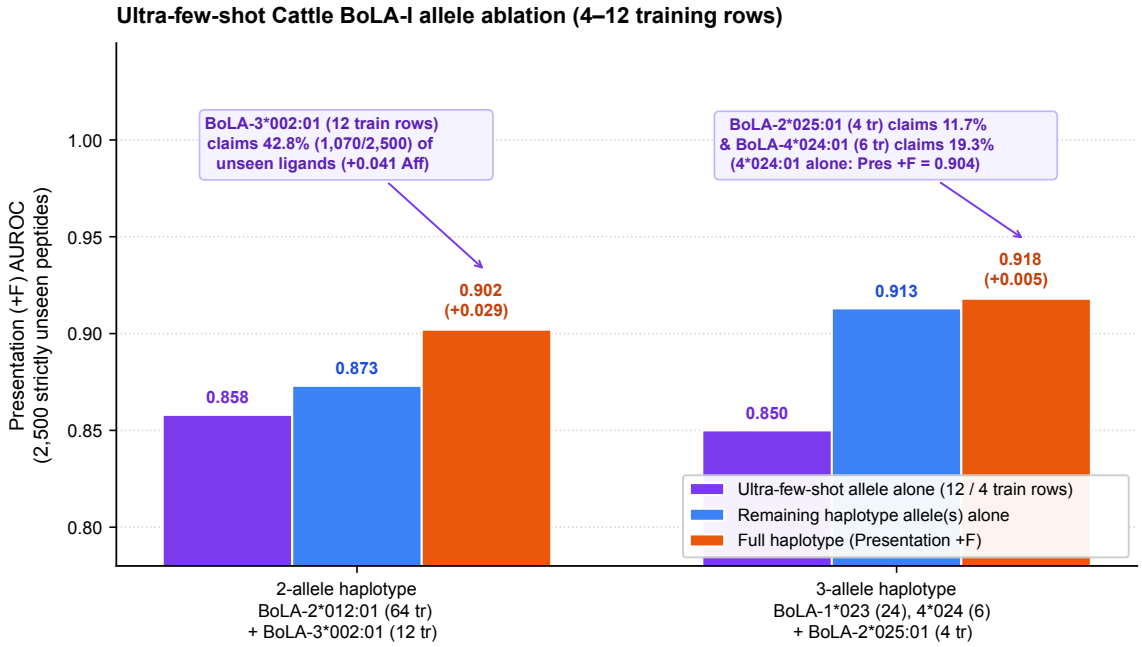

Figure 13: Ultra-few-shot Cattle `BoLA-I` allele ablation (4–12 training rows). Presentation (+F) AUROC on 2,500 strictly unseen bovine ligands per haplotype comparing ultra-few-shot alleles alone, remaining haplotype alleles alone, and the full haplotype.

Table 5: Three-table overlap audit and strictly unseen Cattle `BoLA-I` (`PXD008151`) benchmark. Unseen cohorts evaluate 2,500 strictly unseen flanked positives against 25,000 (10×) unseen *Bos taurus* proteome decoys with real 5-aa N/C flanks, plus size-matched mismatched nulls and within-peptide scrambled controls.

Cohort (Source)	Train Rows	Unseen (%)	Aff.	Pres. (+F)	Null [5-95%]	Scrambled (Null)	PPV@N (noF→+F)
Cattle `BoLA-6*013:01`	126	5,873 (87.3%)	0.950	**0.960**	0.691 [0.505, 0.901]	**0.922** (0.721)	0.726→0.725
Cattle `2*012; 3*002`	64 + 12	8,304 (80.8%)	0.871	**0.902**	0.752 [0.653, 0.811]	**0.833** (0.731)	0.538→**0.560**
Cattle `1*023; 2*025; 4*024`	24+4+6	7,429 (79.5%)	0.892	**0.918**	0.691 [0.596, 0.784]	**0.857** (0.604)	0.570→**0.594**
Cattle `BoLA` (PMID 36423003)	21/14/13	0 (0.0%)	*Excluded – 100% in-sample in `2.3.0` bundled affinity table*				
Dog `DLA-88*501:01` (27893789)	1,963	0 (0.0%)	*Excluded – 100% in-sample (1,959/1,959 overlap; confirms #490)*				
Dog `DLA-88*508:01` (29475511)	96	0 (0.0%)	*Excluded – 100% in-sample (50/50 overlap; confirms #490)*				

5. Exploratory Boundary Study: Zero-Shot Class I → Class II Transfer

Because MHC Class I (one heavy chain forming both halves of the binding pocket) and Class II (separate α and β chains each contributing one half) share the same overall 3D fold, `mhcseqs.align_domain_to_imgt(...)` maps all four domains onto one shared structural coordinate system. Mapping MHCflurry's 39 Class I contact positions directly populates **36 of 39 positions** across Class II heterodimers, with only 3 Class I-only positions (mature 16, 19, 59) padded with `X`.

We tested whether these structurally aligned 39-aa Class II pseudosequences allow **frozen Class I MHCflurry `2.3.0` weights** to predict quantitative Class II 15-mer binding affinity zero-shot across 15 HLA-DR, HLA-DQ, and HLA-DP alleles from `hitlist` (500 unique 15-mers per allele; 7,500 total IC50 nM pairs; binder threshold IC50 $\leq$ 500 nM), benchmarked head-to-head against `NetMHCIIpan-4.3` (BA and EL), a **30-allele mismatched real Class I null distribution**, **20 column-permuted grooves**, and **within-peptide residue-scrambled Class II binders** (**Figures 14–16** and **Table 6**).

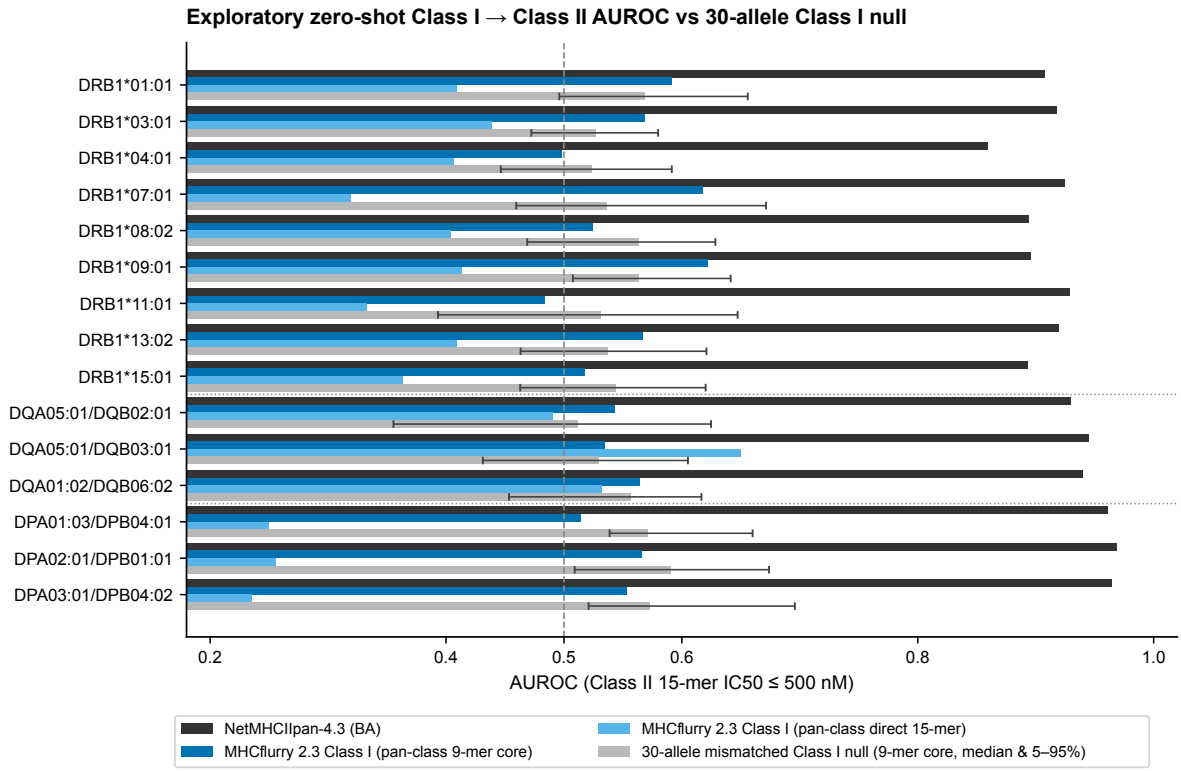

Figure 14: Exploratory zero-shot Class I → Class II AUROC versus 30-allele mismatched Class I null across 15 HLA-DR, DQ, and DP alleles. Binary binder classification AUROC (IC50 $\leq$ 500 nM; 500 15-mers per allele). Frozen Class I MHCflurry with 9-mer core scanning averages 0.551—indistinguishable from the 30-allele mismatched Class I null (0.549, gray bars with 5–95% whiskers) and far below native `NetMHCIIpan-4.3 BA` (0.923). Direct 15-mer scoring drops to 0.235–0.255 on HLA-DP.

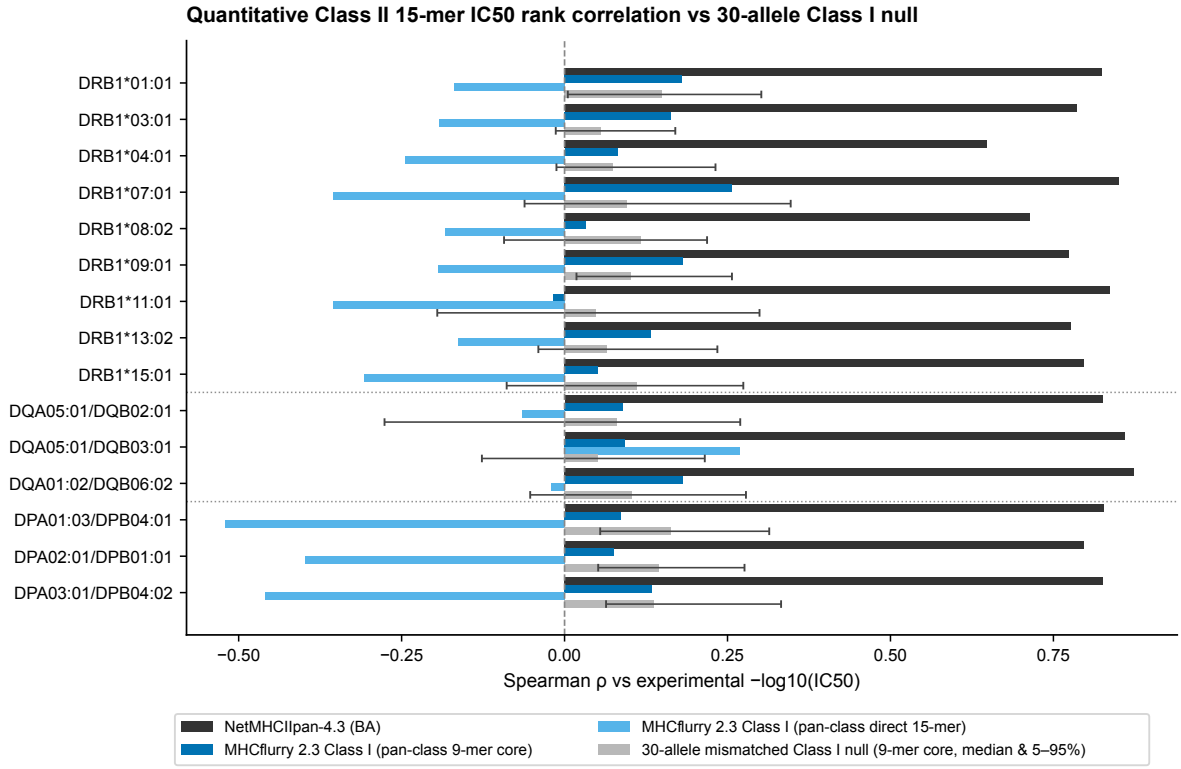

Figure 15: Quantitative Class II 15-mer IC50 rank correlation (Spearman ρ) versus 30-allele mismatched Class I null across 15 alleles. Continuous rank correlation against experimental $-\log_{10}(\text{IC50})$. Native `NetMHCIIpan-4.3 BA` achieves mean $\rho = +0.801$, whereas frozen Class I 9-mer core scanning reaches $+0.115$ (matching the 30-allele mismatched Class I null at $+0.100$, below 20 column-permuted grooves at $+0.146$) and direct 15-mer scoring is systematically anti-correlated ($\rho = -0.224$, down to -0.521 on HLA-DP).

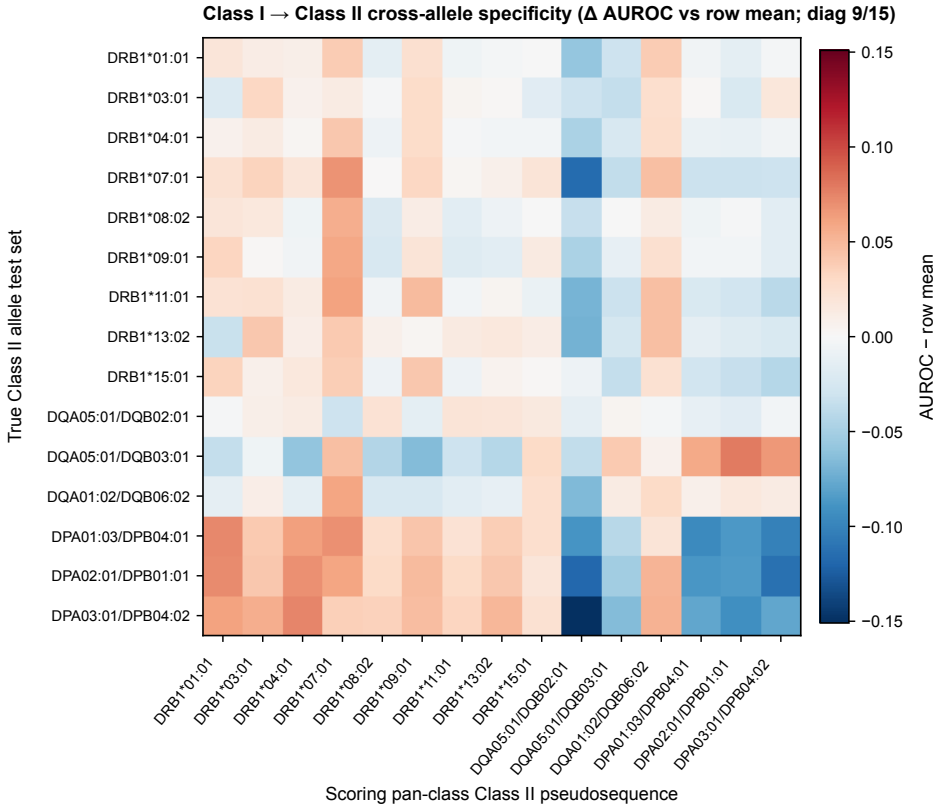

Figure 16: 15×15 Class I → Class II cross-allele specificity matrix. Row-centered AUROC (AUROC−row mean) when scoring each Class II allele test set (rows) with every pan-class Class II pseudosequence (columns) under 9-mer core scoring. Strong vertical column stripes rather than a diagonal ridge (mean diagonal 0.551 vs. off-diagonal 0.556, $\Delta = -0.005$) confirm zero Class II allele specificity.

Table 6: Per-allele Class I → Class II transfer results and proper negative controls across 15 HLA-DR/DQ/DP alleles (500 15-mers per allele; 7,500 pairs total). Scrambled Binders compares measured Class II binders (IC50 $\leq$ 500 nM) against $3\times$ within-peptide residue-shuffled copies of those exact 15-mers.

Allele	Locus	NetMHCIIpan BA	MHCf 9-mer	30-Allele C1 Null [5–95%]	MHCf 15-mer	Scrambled: MHCf / NMII
HLA-DRB1*01:01	DR	0.908	0.592	0.569 [0.496, 0.656]	0.410	0.491 / **0.638**
HLA-DRB1*03:01	DR	0.918	0.568	0.527 [0.472, 0.580]	0.438	0.533 / **0.789**
HLA-DRB1*04:01	DR	0.859	0.498	0.524 [0.446, 0.591]	0.406	0.478 / **0.774**
HLA-DRB1*07:01	DR	0.924	0.617	0.536 [0.460, 0.671]	0.319	0.526 / **0.705**
HLA-DRB1*08:02	DR	0.894	0.525	0.564 [0.469, 0.628]	0.404	0.513 / **0.750**
HLA-DRB1*09:01	DR	0.896	0.622	0.564 [0.508, 0.641]	0.413	0.526 / **0.735**
HLA-DRB1*11:01	DR	0.929	0.484	0.531 [0.393, 0.647]	0.333	0.453 / **0.741**
HLA-DRB1*13:02	DR	0.920	0.567	0.537 [0.463, 0.621]	0.409	0.494 / **0.706**
HLA-DRB1*15:01	DR	0.893	0.518	0.544 [0.463, 0.620]	0.363	0.499 / **0.717**
DQA1*05:01/DQB1*02:01	DQ	0.930	0.543	0.512 [0.355, 0.625]	0.491	0.473 / **0.748**
DQA1*05:01/DQB1*03:01	DQ	0.945	0.534	0.530 [0.431, 0.605]	0.650	0.523 / **0.655**
DQA1*01:02/DQB1*06:02	DQ	0.940	0.565	0.556 [0.453, 0.617]	0.532	0.498 / **0.734**
DPA1*01:03/DPB1*04:01	DP	0.961	0.514	0.571 [0.539, 0.660]	0.250	0.422 / **0.731**
DPA1*02:01/DPB1*01:01	DP	0.969	0.566	0.591 [0.509, 0.674]	0.255	0.572 / **0.767**
DPA1*03:01/DPB1*04:02	DP	0.964	0.553	0.573 [0.521, 0.696]	0.235	0.529 / **0.691**
Mean (15 alleles)	–	**0.923** (ρ +0.801)	**0.551** (ρ +0.115)	**0.549** (ρ +0.100)*	**0.394** (ρ -0.224)	**0.502** / **0.725**

*20 column-permuted grooves average 0.571 AUROC ($\rho = +0.146$); mismatched Class I null on scrambled binders averages 0.514.

Mechanistic Diagnosis of the Class I → Class II Negative Result

1. **Within-peptide scrambled binders prove 9-mer core scanning is pure hydrophobicity bias (0.502 vs. 0.725):** Scanning 9-mer cores lifts AUROC from 0.394 to 0.551 only because Class II 15-mer binders contain hydrophobic internal cores favored by *any* Class I groove (30-allele mismatched Class I null $= 0.549$; 20 column-permuted grooves $= 0.571$). Holding amino-acid composition constant by comparing real Class II binders against their own residue-shuffled copies collapses frozen Class I straight to **0.502** (coin flip), while native `NetMHCIIpan-4.3 BA` maintains **0.725** ($+0.223$).
2. **Closed-groove vs. open-groove geometry causes severe anti-correlation on HLA-DP (0.235–0.255, $\rho = -0.40$ to -0.52):** Class I binds 15-mers in a *closed groove* with pinned N- and C-termini and a central bulge, whereas Class II binds an *open groove* where the 9-mer core floats internally with polar/charged flanking overhangs extending past both ends.
3. **Zero diagonal specificity proves joint pan-class training is mandatory:** In **Figure 16**, mean on-diagonal AUROC (0.551) is slightly *below* mean off-diagonal AUROC (0.556, $\Delta = -0.005$; vertical column stripes dominate). While `mhcseqs`'s shared 3D structural coordinate frame provides an ideal unified feature representation, unlocking pan-class prediction requires training a future architecture jointly on both closed-groove Class I and open-groove Class II binding registers.

6. Practical Guidance, Ecosystem Data Hygiene, and Reproducibility

6.1 Checklist for Predicting on Newly Added Vertebrate Species

1. **Inspect groove alignment warnings:** Use `extract_class1_pseudosequence_with_alignment()` (or watch `mhcflurry` logs). Canonical grooves and small deletions transfer robustly across mammals and birds; multi-residue pocket insertions (e.g., bat `Ptal-N*01:01` 3-aa `MDL`) degrade fixed-39-aa accuracy and require retrained insertion-aware models.
2. **Use mismatched-allele null distributions and within-peptide scrambled controls:** Rather than comparing against a single arbitrary human allele, benchmark new non-model datasets against a diverse panel of unrelated real alleles (size-matched k-allele draws for haplotypes) and residue-scrambled copies of the positive ligands.
3. **Provide real host flanking sequences for MS/cellular tasks:** Real species proteome flanks consistently improve Presentation AUROC and PPV@N even across $\sim$160 Myr to marsupials ($+0.057$ AUROC on Tasmanian devil) and on audited bovine MS ($+4$% PPV@N).
4. **Audit all three bundled training tables before held-out claims:** Check `affinity_predictor_train_data.csv.bz2` and both `processing_predictor_*_train_data.csv.bz2` files inside `models_class1_presentation/models/`—never `curated_training_data.affinity.csv.bz2` alone.
5. **Do not apply frozen Class I weights to Class II alleles:** Use dedicated Class II predictors (e.g., `NetMHCIIpan-4.3`) until a jointly trained pan-class MHCflurry architecture is released.

6.2 Upstream Nomenclature and Corpus Hygiene Resolved During 2.5.x

Cross-species validation surfaced several ecosystem-wide issues that were filed and resolved cleanly upstream: `pirl-unc/mhcseqs#141` (canonicalization of 4 trained alleles, fixed in `2.12.5`); `pirl-unc/mhcgnomes#306` (blocking bare UniProt accessions such as medaka `H2LR16` from misparsing as mouse `H2-L`, fixed in `3.87.0`); `pirl-unc/mhcgnomes#311`, `#328`, and `mhcseqs#150` (first-class `AccessionedAllele` identifiers preserving Class I across 223 species); `pirl-unc/hitlist#977` (9 misassigned UniProt proteome UPIDs); and `pirl-unc/hitlist#989`–`#993` (data-bundle organization and provenance improvements).

6.3 Reproducible Experiment Scripts

Every figure and table in this whitepaper is reproducible from version-controlled scripts in the MHCflurry repository:

- **2.5.0 zero-shot, motifs & pocket mutants:** `scripts/experiments/release_2_5_0/benchmark_cross_specie`

`s_mhcseqs.py`

- **2.5.7–2.5.8 `hitlist` Presentation/Affinity & training parity:** `scripts/experiments/release_2_5_0/benchmark_hitlist_cross_species.py` and `plot_hitlist_cross_species_figures.py`
- **2.5.9 Class I → Class II transfer vs. `NetMHCIIpan-4.3`:** `scripts/experiments/release_2_5_0/benchmark_pan_class_ii_transfer.py`
- **2.5.13 Audited Cattle `BoLA-I` & Dog `DLA-88`:** `scripts/experiments/release_2_5_0/benchmark_cattle_dog_hitlist.py`
- **This whitepaper PDF:** `python docs/build_whitepaper_pdf.py`

7. References

[1] O'Donnell, T. J., Rubinsteyn, A., & Laserson, U. (2020). MHCflurry 2.0: Improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Systems*, 11(1), 42–48.e7. doi:10.1016/j.cels.2020.06.010

[2] O'Donnell, T. J., Rubinsteyn, A., Bonsack, M., Riemer, A. B., Laserson, U., & Hammerbacher, J. (2018). MHCflurry: Open-source Class I MHC binding affinity prediction. *Cell Systems*, 7(1), 129–132.e4. doi:10.1016/j.cels.2018.05.014

[3] Nielsen, M., Connelley, T., & Ternette, N. (2018). Improved prediction of bovine leucocyte antigens (BoLA) presented ligands by use of mass-spectrometry-determined ligand and in vitro binding data (`PXD008151`, PMID 29115832). *Journal of Proteome Research*, 17(1), 559–567.

[4] Nilsson, J. B., Kaabinejadian, S., Yari, H., Kester, M. G. D., van Balen, P., Hildebrand, W. H., & Nielsen, M. (2023). Accurate prediction of HLA class II antigen presentation across all loci using tailored data acquisition and refined machine learning (`NetMHCIIpan-4.3`). *Science Advances*, 9(47), eadj6367.

[5] Lefranc, M.-P., Pommié, C., Kaas, Q., Duprat, E., Bosc, N., Guiraudou, D., Jean, C., Ruiz, M., Da Piédade, I., Rouard, M., Foulquier, E., Thouvenin, V., & Lefranc, G. (2005). IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains (and G-DOMAINs). *Developmental & Comparative Immunology*, 29(3), 185–203.

[6] Caldwell, A., Coleby, R., Tovar, C., et al. (2018). The newly-arisen Devil facial tumour disease 2 (DFT2) reveals a mechanism for the emergence of a contagious cancer (*Sarcophilus harrisii* immunopeptidome). *eLife*, 7, e35314.

[7] Wynne, J. W., Weyer, J. L., Shiell, B. J., et al. (2016). Characterization of the MHC class I peptide repertoire of the black flying fox (*Pteropus alecto*, `Ptal-N*01:01`). *Journal of Immunology*, 196(11), 4445–4454.

[8] Rubinsteyn, A., et al. (2026). OpenVax / PIRL-UNC immunogenomics ecosystem: `mhcseqs` (github.com/pirl-unc/mhcseqs), `mhcgnomes` (github.com/pirl-unc/mhcgnomes), and `hitlist` (github.com/pirl-unc/hitlist).