DOES REGISTER-LENGTH FLEXIBILITY HELP PEPTIDE–MHC-II CLASSIFICATION? A CONTROLLED ABLATION ON THE CHICKEN ALLELE BLB2

Anonymous authors
Paper under double-blind review

ABSTRACT

Discrete latent-alignment models are usually given a fixed-length candidate structure to select over. We ask what happens when that support is widened to match documented heterogeneity in the target structure itself, using peptide–MHC class II (MHC-II) ligand classification as a test bed. Most MHC-II predictors represent the binding core as a canonical nine-residue window and treat only its position as a latent register variable. The chicken allele *Gaga-BLB2*002:01* is a useful test case because a structural study establishes that it can anchor a decamer core (Halabi et al., 2021), while a separate computational motif-deconvolution study reports its ligand pool splits nearly evenly between 9-residue and 10-residue annotated cores (Racle et al., 2023), a split our own compiled dataset reproduces (2,134 9-mers, 2,050 10-mers among 4,184 positives). We train a register variable, supervised by an auxiliary cross-entropy loss against the annotated core alongside the primary classification loss, to range jointly over both candidate lengths, and compare it under a parameter-matched ablation that masks the 10-mer candidates out. Holding architecture, data, and training procedure fixed across five outer-fold held-out evaluations (single seed, one allele), the two variants are statistically indistinguishable on ranking metrics (pooled ROC-AUC 0.920 for both), and the restricted model is, if anything, slightly higher on PR-AUC (0.597 versus 0.583; a paired bootstrap places the gap at 0.013, 95% CI [0.006, 0.021]). We cannot attribute that gap to candidate-length support alone, since masking also removes auxiliary supervision from the 10-mer-annotated half of positives. Only the flexible model can agree with a length-flexible annotation at all: it matches the annotated core length on 74.3% of held-out positives (a 51.0% majority-class floor) and reproduces the exact annotated core substring on 30.7%. We also find that a large share of what our dataset labels "BLB2-positive, 9-mer core" may in fact be a co-expressed allele's ligands misassigned by the motif-deconvolution pipeline, a hypothesis Halabi et al. themselves supply and that predicts the largest effect we observe: 9-mer-annotated positives score far lower (PR-AUC $\approx$0.34–0.35) than 10-mer-annotated positives (PR-AUC $\approx$0.58–0.59) at matched prevalence. Register flexibility, in this architecture and on this single allele, is therefore better read as a lever on whether a model's internal explanation can engage with heterogeneous annotations at all, not as a classification-accuracy improvement. Part of the negative result may itself be an artifact of imperfect training labels rather than of the inductive bias being tested.

1 INTRODUCTION

Predicting whether a peptide is presented by an MHC class II (MHC-II) molecule is a core computational immunology problem, relevant to vaccine antigen selection and to understanding autoimmune and infectious-disease susceptibility. The canonical picture treats a peptide's binding core as a fixed 9-residue window whose *position* is a latent variable. NetMHCIIpan (Reynisson et al., 2020), MixMHC2pred (Racle et al., 2019), and DeepMHCII (You et al., 2022) all resolve position this way, but none treats core *length* as a first-class variable. Racle et al. (2023) did already run a

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

length-flexible motif search for the allele we study, re-running their deconvolution tool with a 10-mer motif alongside the default 9-mer one, so a length-flexible *analysis* predates this paper even though no predictor's *architecture* models core length jointly.

The chicken class II molecule *Gaga-BLB2*002:01* (BLB2) is a useful test case, and the evidence for its length heterogeneity comes from two different kinds of study. Structurally, Halabi et al. (2021) found that, of the eluted peptide species they characterized from the dominantly expressed B2-haplotype molecule (BL2*02:01), 50 fit a decamer motif, 6 fit a nonamer motif (attributed to the co-expressed BL1*02/BLB1, not BL2*02:01 itself), and 7 fit neither; they do not claim BL2*02:01 binds exclusively via a decamer core. Separately, Racle et al. (2023) applied motif deconvolution to mass-spectrometry data labeled *Gaga-BLB2*002:01* and found close to half of ligands assigned a 9-residue core and half a 10-residue core, a population-level estimate our own dataset reproduces (2,134 9-mers, 2,050 10-mers of 4,184 positives, Section 4).

This distinction matters beyond citation hygiene. Halabi et al. describe a mechanism that supplies a concrete alternative explanation for our own dataset's split: if our ligand list is prone to the same kind of pipeline confusion, part of what we label "BLB2-positive, 9-mer core" could actually be BLB1 ligands. We test this in Section 4 and find a $13\times$ label-overlap enrichment consistent with the hypothesis – not proof, but a live competing explanation for part of this paper's central result (Section 6).

The machine-learning question we ask is this: when a target's annotated core length is heterogeneous, is it better to let a register variable range jointly over both candidate lengths, scored by attention and supervised by an auxiliary loss against the annotated core (Section 3), or to force every peptide onto a fixed-length core? This is a question about inductive bias in discrete latent-alignment models, studied outside immunology as structured and variational attention (Kim et al., 2017; Deng et al., 2018), where supervising alignment against known correspondences is an established variant (Yang & Tang, 2022) and constraining an alignment variable's support is a recognized way to encode inductive bias (Raffel et al., 2017). BLB2 lets us compare flexible versus fixed candidate-length support under a controlled ablation, for one allele, species, architecture, and training seed (Section 9).

Contributions. (1) We isolate candidate-length flexibility in an existing register-scoring mechanism under a parameter-matched ablation, rather than proposing a new mechanism (Sections 3, 6). (2) We run a subgroup analysis testing whether flexibility helps specifically on 10-mer-annotated peptides, with a paired bootstrap interval on the observed gap (Section 6). (3) We evaluate how well the register branch's supervised task generalizes to held-out annotations, with an explicit accounting of the resulting circularity (Section 7). (4) We identify and partially quantify a labeling-contamination hypothesis that competes with our own central result (Sections 4, 6).

2 Related Work

MHC-II ligand and epitope prediction. NetMHCIIpan-4.0 (Reynisson et al., 2020) and MixMHC2pred (Racle et al., 2019) are the two most widely used MHC-II binding predictors. Both are pan-allele and resolve binding-core *position* as a latent variable while assuming a fixed 9-residue core. Per Racle et al. (2023), MixMHC2pred-2.0's architecture was built around 9-residue cores; the BLB2 9/10-mer split it reports came from re-running its deconvolution tool with an additional 10-mer motif, not a predictor change. DeepMHCII (You et al., 2022) introduces a binding-interaction convolution integrating every candidate core with the MHC pseudo-sequence, the closest prior architecture in spirit to a register-scoring mechanism, but it too operates over fixed-length, 9-mer candidates. We are not aware of a prior MHC-II predictor whose *architecture* scores candidate cores of more than one length jointly; our contribution is the controlled evaluation of length-flexible modeling as an architectural choice.

Chicken/avian MHC-II and BLB2 binding. The chicken MHC has one dominantly expressed class II locus but plays an outsized role in disease-resistance genetics, notably Marek's disease. Section 1 summarizes the two sources of evidence for BLB2's length heterogeneity: Halabi et al. (2021)'s structural determination of the decamer mode (not claiming exclusivity) and Racle et al. (2023)'s separate, computational 9/10-mer split reproduced by our dataset (Section 4). Goryanin et al. (2026) later showed structurally that a related allele, BL2*21:01, can bind a single peptide

in both a 9-mer and 10-mer conformation, evidence about BL2*21:01, not BL2*02:01. Our raw ligand list has the column structure of a motif-deconvolution pipeline in the style of Racle et al. (2019; 2023), the same kind of pipeline that, per Halabi et al., can assign a co-expressed molecule's nonamer ligands to the wrong allele; we check this contamination risk directly in Section 4.

Register-aware and latent-alignment models. Attention mechanisms that marginalize over multiple candidate alignments rather than committing to one are studied outside immunology as structured and variational attention (Kim et al., 2017; Deng et al., 2018), building on standard attention (Vaswani et al., 2017). Supervising such a distribution against a known correspondence, as our auxiliary register loss does, is an established variant (Yang & Tang, 2022), and constraining an alignment variable's support is itself a recognized way to encode inductive bias (Raffel et al., 2017). A shared cross-attention encoder has been used jointly for peptide–HLA-I and peptide–TCR binding (Yu et al., 2025); our register module is a self-attention scorer over within-peptide candidate windows, a different mechanism despite the shared framing.

3 METHODS

Both models compared here fuse branches by plain concatenation. The codebase also supports a learned per-allele gate as an alternative fusion mode for other model variants, mentioned only for completeness since it plays no role in any result reported here.

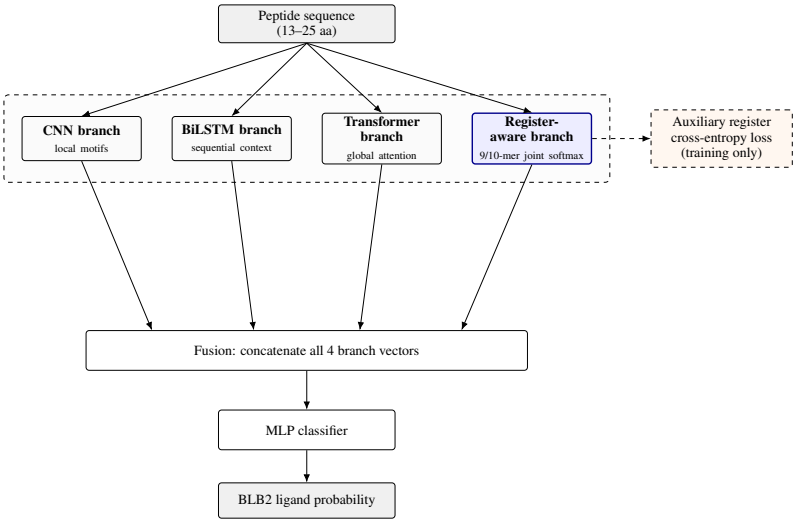

Figure 1: The BLB2 model as implemented. Each branch produces one pooled feature vector; the four are concatenated before the final classifier. In the register-aware branch (highlighted), the 33-way soft register distribution supplies the attention weights for that branch's pooled output and is additionally supervised by the dashed auxiliary loss (orange) against the annotated core. The 9-mer-only ablation (Section 6) masks this branch's 10-mer candidates before that softmax, leaving every other branch unchanged.

Input and branches. The model's only input is a single peptide sequence (13–25 residues). BLB2 uses a dedicated single-output model, never modeled jointly with BLB1, with no shared MHC-sequence input. Figure 1 shows the architecture. The peptide is encoded independently by a CNN branch (LeCun et al., 1998), a BiLSTM branch (Hochreiter & Schmidhuber, 1997; Schuster & Paliwal, 1997), a plain Transformer branch (Vaswani et al., 2017) (each producing a pooled embedding), and the register-aware branch described next. Branch outputs are concatenated and passed to an MLP that outputs a single ligand-probability logit.

Register-aware branch. For a peptide of length L (up to 25), every valid 9-mer window ($N_9 = 17$ start positions) and every valid 10-mer window ($N_{10} = 16$ start positions) is enumerated, giving $N = 33$ candidate registers, each a (length, start-position) pair with a core-relative positional encoding, an N-flank/core/C-flank segment embedding, and a length embedding. Sharing the same

10 position buckets across both lengths assumes the i-th position of a 9-mer core roughly matches the i-th (rescaled) position of a 10-mer core, a simplification since Halabi et al. (2021)'s decamer arises from a localized backbone distortion rather than a uniform stretch; we do not implement the alternative of tracking the actual distortion site. A Transformer encoder scores each candidate once, and a joint softmax over all 33 candidates converts these scores into a single *soft register distribution* (candidates that do not fit inside the peptide, or that an ablation has masked out, are set to $-\infty$ first). This distribution weights a pooled representation concatenated with the other branches, so it reaches the classifier through pooling even though it is never itself concatenated as a feature vector, and is separately supervised by an auxiliary cross-entropy loss against the slot matching the annotated core. Peptides without a valid annotated core, and, under the 9-mer-only ablation, peptides whose annotated core falls outside the allowed length, are excluded from this auxiliary loss but still contribute to the primary loss. Classification consumes the full soft distribution, while our interpretability analysis (Section 7) scores only its arg-max slot, a mismatch we return to there.

The flexible-vs-fixed manipulation. The 9-mer-only ablation masks all 16 ten-mer slots to $-\infty$ unconditionally, for every peptide, before the softmax, via a `restrict_core_lengths` flag. Both variants have identical parameter counts, since the masked slots still exist as parameters and are simply never selected, so any performance difference is not a difference in model capacity. This also withholds the auxiliary register loss from every peptide whose annotated core is a 10-mer (2,050 of 4,184 positives), a confound we detail in Section 6.

Training. We use Adam (Kingma & Ba, 2015), learning rate 10^{-3}, batch size 64, and a fixed seed of 42 (no multi-seed averaging), with dropout (Srivastava et al., 2014) in every branch (rate in Table 5). The loss is a weighted sum of the primary classification loss and the auxiliary register loss, with auxiliary weight 0.3 for both models; this was fixed by an earlier, target-independent decision and not swept here, so our results are conditional on it (Section 9). We tuned three primary loss variants, plain BCE, mildly class-weighted BCE, and margin-augmented BCE, while a fourth, focal loss, was excluded from this sweep by that same decision. Learning-rate tuning and model ensembling are not used anywhere in this project. Per-branch hyperparameters, epoch budgets, and early-stopping criteria are in Table 5 (Appendix A.2). Final-model training took 24–26 seconds per outer fold for the flexible model and 31–34 for the ablation, one GPU, no distributed training.

4 Dataset and Experimental Setup

Data. Our master dataset of 62,590 peptide–protein pairs is built from a curated chicken MHC-II ligand list plus computationally generated negatives, with per-target binary labels for BLB1 and BLB2. This paper concerns BLB2 only: 4,184 positive and 58,406 negative peptides, a roughly 14:1 imbalance that is why we treat PR-AUC (Davis & Goadrich, 2006), not accuracy or ROC-AUC alone, as the primary metric (ROC-AUC and MCC (Matthews, 1975) are reported alongside it throughout). The negatives are source-protein decoys, so they are putative rather than experimentally verified non-binders; the 14:1 ratio is a design choice of this project, not a property of the underlying biology, so our PR-AUC values are not directly comparable to figures reported under a different negative-sampling scheme elsewhere.

Core-length distribution. All 4,184 BLB2 positives carry a non-null annotated `Core`: 2,134 (51.0%) are 9 residues and 2,050 (49.0%) are 10 residues long. (Class-balance and core-length distribution figures are in Appendix A.1, moved there for space.) BLB1 and the other alleles we examined are overwhelmingly 9-mer-annotated; BLB2 is the outlier. As in Section 2, this `Core` column is annotated or inferred by the source ligand list's own motif-deconvolution pipeline, not experimentally verified per peptide.

Checking the BLB1-contamination hypothesis. Section 1 raised a concrete alternative explanation for this split: a motif-deconvolution pipeline run over a mixed elution pool can assign a co-expressed molecule's nonamer ligands to the wrong allele (Halabi et al., 2021), and our raw ligand list has exactly that kind of pipeline's column structure. Our dataset carries independent BLB1/BLB2 labels per peptide: among the 2,134 9-mer-annotated BLB2 positives, 196 (9.2%) are also labeled BLB1-positive; among the 2,050 10-mer-annotated BLB2 positives, only 14 (0.7%) are, a $13\times$ enrichment. This is a lower bound (it only catches peptides independently confirmed positive for both alleles), but it is a sizeable signal consistent with Halabi et al.'s mechanism: an unknown share of our 9-mer-annotated "BLB2 positives" may not be primarily BLB2 ligands. We return to this in Section 6,

where the 9-mer subgroup's classification difficulty is far higher than the 10-mer subgroup's, the pattern this hypothesis predicts.

Splits and comparison arms. Peptides are grouped by source protein and split into 5 outer folds for nested cross-validation, with zero protein overlap between any outer-test fold and its development partition, verified programmatically. Architecture selection, loss tuning, and threshold selection use only outer round 1's inner-fold OOF predictions; the other four outer folds stay held out until each architecture's single final evaluation. We originally planned five comparison arms; since our implemented architecture has no cross-attention module separate from the register branch, two collapse onto one model. The classical baseline uses a 26-dimensional hand-crafted feature vector: 20 amino-acid composition fractions, 5 mean physicochemical properties (hydrophobicity, charge, polarity, molecular weight, isoelectric point), and peptide length, scored by a random forest (Breiman, 2001).

Table 6 (Appendix A.3) reports the full architecture screening at this untuned resolution: the best single-branch baseline (a BiLSTM (Hochreiter & Schmidhuber, 1997; Schuster & Paliwal, 1997), PR-AUC 0.517) and the classical baseline (0.240) both trail the fusion model without a register branch (0.520), which trails the full flexible register model (0.535). The 9-mer-only ablation is not included, since it was never run through this untuned step; it is compared separately, at tuned resolution, in Section 6.

5 RESULTS

Final held-out evaluation. Table 2 (Section 6) reports the frozen flexible-register model's final 5-outer-fold held-out performance (loss *weighted_bce_mild*, threshold 0.68, both fixed using only outer round 1's inner-fold data) alongside the 9-mer-only ablation, since the two are directly comparable. Pooled PR-AUC is 0.583 and pooled ROC-AUC is 0.920. At this operating point the model misses close to three-fifths of true BLB2 ligands (pooled recall 0.401); for a downstream use like epitope mapping or antigen design (Section 8) rather than a binary presentation call, this miss rate would matter as much as the headline PR-AUC. Pooled PR-AUC (0.583) falls below the per-fold mean (0.604 ± 0.014) because pooling concatenates five folds whose score distributions and prevalence are not identical; we report both for that reason.

External reference: MixMHC2pred. MixMHC2pred v2.0.2 (Racle et al., 2019) assumes a 9-residue core and was never extended to variable-length cores, so applying it to BLB2 is a zero-shot, out-of-distribution evaluation, not a like-for-like comparison: outscoring it is the expected direction, not informative on its own. It could not score 4,700 of the 62,590 BLB2 peptides (all 22–25 residues, well above the dataset's 17-residue median), so Table 1 reports our frozen model both on the full dataset and on exactly the peptides MixMHC2pred could score, alongside MixMHC2pred's own numbers on that matched subset (the two rows for our model are nearly identical, so the exclusion does not materially change the comparison). MixMHC2pred's chicken training data may also overlap our peptide set, so we treat this as a directional reference point, not a generalization test.

Table 1: Comparison against MixMHC2pred v2.0.2. MixMHC2pred could not score 4,700 of the 62,590 BLB2 peptides (all 22–25 residues); the matched subset restricts both models to the 57,890 peptides MixMHC2pred could score.

Model	n	PR-AUC	ROC-AUC
Our model (full dataset)	62,590	0.583	0.920
Our model (matched subset)	57,890	0.595	0.923
MixMHC2pred v2.0.2 (matched subset)	57,890	0.488	0.842

6 ABLATION: 9-MER-ONLY VS. FLEXIBLE 9/10-MER REGISTER

This is the paper's central experiment: a masking change only (Section 3). The resulting model is parameter-for-parameter identical to the flexible model and goes through the same loss-tuning, threshold-selection, and 5-fold evaluation pipeline, writing to separate output paths. As noted in

Section 3, this experiment cannot separate candidate-length support from the accompanying loss of auxiliary-supervision volume; the decisive control that would resolve this is not run here.

Loss-tuned inner-fold OOF comparison. At matching resolution, both variants tuned across the same three losses on the same outer-round-1 inner folds, the 9-mer-only ablation's best loss is plain BCE (PR-AUC 0.560, MCC 0.504) and the flexible model's is mildly class-weighted BCE (PR-AUC 0.552, MCC 0.496); the full table is Table 7 in Appendix A.3.

Final 5-outer-fold held-out comparison. Table 2 extends this with the ablation's own independent 5-fold run, using its own threshold, 0.37, chosen by the same MCC-optimal procedure on its own OOF predictions and never reusing the flexible model's threshold of 0.68. Two models trained on the same data landing on thresholds that differ by nearly a factor of two, despite the indistinguishable ranking performance shown next, is itself a calibration difference between the variants.

Table 2: BLB2 final 5-outer-fold held-out evaluation, flexible model vs. 9-mer-only ablation. "Per-fold" = mean ± SD across the 5 independent outer-test folds; "Pooled" = metrics on the union of all 5 folds' predictions.

	Flexible (main model)		9-mer-only (ablation)	
Metric	Per-fold (mean ± SD)	Pooled	Per-fold (mean ± SD)	Pooled
PR-AUC	0.604 ± 0.014	0.583	$\mathbf{0.607 \pm 0.018}$	**0.597**
ROC-AUC	0.924 ± 0.003	0.920	0.923 ± 0.004	0.920
MCC	0.506 ± 0.033	0.503	$\mathbf{0.523 \pm 0.034}$	**0.521**
Precision	$\mathbf{0.717 \pm 0.081}$	**0.694**	0.711 ± 0.059	0.698
Recall	0.400 ± 0.101	0.401	$\mathbf{0.426 \pm 0.080}$	**0.426**
Specificity	0.987 ± 0.008	0.987	0.987 ± 0.005	0.987
F1	0.502 ± 0.057	0.508	$\mathbf{0.526 \pm 0.056}$	**0.529**
Balanced accuracy	0.694 ± 0.047	0.694	$\mathbf{0.707 \pm 0.037}$	**0.706**

We lead with the threshold-free metrics, since they give the cleanest comparison. Pooled ROC-AUC is 0.920 for both models to three decimal places (Table 2), the strongest evidence that the two variants are indistinguishable as rankers. Pooled PR-AUC favors the ablation, 0.597 versus 0.583. A paired bootstrap (Efron & Tibshirani, 1994) (2,000 resamples, stratified by class, resampling from the pooled predictions) on this difference gives a point estimate of 0.013 in the ablation's favor, 95% CI $[0.006, 0.021]$, excluding zero. This quantifies sampling variability given these two already-trained models, not retraining variance with a different seed, which we still have no estimate of (Section 9); within that scope, the ablation's small PR-AUC edge is not simply evaluation noise. MCC, precision, recall, F1, and balanced accuracy are consistent but not unanimous: the ablation is numerically higher on all but precision, which favors the flexible model (0.717 ± 0.081 versus 0.711 ± 0.059), though all five are computed at a per-model MCC-optimal threshold and are secondary to PR-AUC and ROC-AUC. Put together, the flexible register does not measurably improve ligand-classification ranking for BLB2, and its small PR-AUC shortfall is not attributable to evaluation noise alone, though we still cannot attribute it specifically to candidate-length support rather than to the accompanying loss of auxiliary-supervision volume.

Negative controls. Two negative-control models, one with shuffled register targets and one with a randomly initialized register scorer, trained through the identical procedure, top out around 0.525–0.526 PR-AUC at the same tuned resolution, below the flexible model's 0.552 and the ablation's 0.560. Both real register-aware models score above both controls, consistent with the register mechanism contributing real signal in either configuration; we have not run a significance test on this gap.

Does 10-mer support help specifically where it should? The aggregate comparison cannot rule out a subgroup effect that is simply outweighed by the majority 9-mer subgroup. We check this directly with a standalone script, run inference-only against the same held-out predictions with no retraining, recomputing PR-AUC restricted to 9-mer-annotated and to 10-mer-annotated positives

separately, in both cases against the identical, full negative pool, after first confirming both models were evaluated on the exact same 62,590 rows.

Table 3: PR-AUC by annotated core-length subgroup (each subgroup's positives vs. the full, identical 58,406-negative pool; same held-out folds for both models).

Model	PR-AUC, 9-mer subgroup (n=2,134)	PR-AUC, 10-mer subgroup (n=2,050)
Flexible (main model)	0.339	0.576
9-mer-only (ablation)	0.351	0.595

The answer is no. The ablation is not worse than the flexible model on the 10-mer subgroup either, if anything marginally higher there too (0.595 vs. 0.576), consistent with the aggregate result. Table 3 also contains the largest effect anywhere in this paper, and it is not the flexible-vs-ablation comparison: at near-identical prevalence, the 9-mer subgroup scores PR-AUC 0.339–0.351 for both models, while the 10-mer subgroup scores 0.576–0.595, a gap of roughly 0.23, an order of magnitude larger than anything separating the two register variants. This is precisely the pattern the BLB1-contamination check in Section 4 predicts: if a non-trivial share of "BLB2-positive, 9-mer-annotated" peptides are mislabeled ligands of another allele, a good BLB2 classifier should score that subgroup poorly. We cannot fully separate this from the alternative that 9-mer cores are simply harder to recognize regardless of label quality, but the size and direction of the gap, combined with the $13\times$ label-overlap enrichment, makes label contamination a live competing explanation for part of this paper's central negative result. This does not make the 9-mer-only model's structural inability to ever predict a 10-mer register a non-issue for interpretability, but it is not a measured cost to classification accuracy.

7 Interpretability: Register and Core-Length Recovery

This is not an unsupervised or emergent alignment: the register branch is trained with an explicit auxiliary cross-entropy loss against the slot matching the annotated core. Each of the 33 candidate slots encodes a core length and a start position together, so one 33-way training signal supervises both at once, and the model's arg-max prediction at inference determines both simultaneously. We evaluate how well that prediction generalizes to peptides the model was not trained on: for every held-out positive with an annotated `Core`, we run inference only (*src/training/analyze_core_predictions.py*) and compare the predicted slot against the annotation, pooling across all 5 outer-test folds so each positive is scored by a fold-model that never saw it during training. We report core-length accuracy, register-start accuracy conditional on a correctly predicted length, sensitivity for 9-mer and 10-mer cores separately, and the exact core-sequence match rate, for both models.

Table 4: BLB2 core-length confusion matrix (annotated vs. predicted, all 4,184 held-out positives with a valid annotated core, pooled across the 5 outer-test folds).

		Flexible: Predicted		9-mer-only: Predicted	
		9	10	9	10
Annotated	9	1851	283	2134	0
	10	793	1257	2050	0

The flexible model's predicted slot agrees with the annotated core length on 74.3% of positives overall, against a 51.0% majority-class floor (the 9-mer-only ablation attains 51.0% for free, always predicting length 9; Table 4's "predicted 10" columns are identically zero for it by construction, a structural blind spot rather than a learned failure). Sensitivity is 86.7% on annotated 9-mers and 61.3% on annotated 10-mers. Conditional on the length already being right, register-start agreement is 41.4%, and the exact full-core-sequence match rate, agreement on both length and exact position simultaneously, is 30.7%. We report all three numbers together deliberately: 74.3% read

alone invites over-interpretation, and 30.7% is the more demanding, and arguably more informative, statistic.

There is also a circularity worth stating plainly. Per Section 2, the `Core` annotations we supervise and evaluate against are themselves the output of a length-flexible motif-deconvolution procedure, so a model built to be length-flexible is scored on agreement with a length-flexible annotator, while the restricted model is barred by construction from ever agreeing on half the cases. Read this way, the 74.3%/61.3% figures are close to definitional rather than a demonstration of independently verifiable biological learning: a model whose support matches its supervision target's support can agree with that target more often, regardless of whether the target is itself biologically accurate. The finding we can defend is narrower: flexibility buys a model whose internal explanation can engage with a heterogeneous annotation at all, not evidence that the explanation is more biologically accurate. The single unambiguous, non-circular cost of the ablation is the confusion-matrix asymmetry itself: for the 2,050 positives whose annotated core is a 10-mer, its register explanation disagrees with that annotation on 100% of them, by construction, even though its classification of those same peptides is, per Section 6, if anything slightly more accurate on aggregate than the flexible model's.

8 Discussion

Section 1 asked whether letting the register variable range over both candidate core lengths improves prediction and interpretability together. On ranking performance the two variants are indistinguishable (pooled ROC-AUC 0.920 for both), and the ablation is, if anything, slightly higher on PR-AUC, a gap our bootstrap places outside zero but cannot attribute specifically to candidate-length support rather than the accompanying loss of auxiliary supervision (Section 6). Both variants score above the negative controls, consistent with the register mechanism contributing real signal either way. On interpretability, by contrast, the two models are not close: the flexible model's register explanation agrees with the annotated core length 74% of the time and with annotated 10-mer cores most of the time (61%), while the 9-mer-only model's explanation for every 10-mer-annotated ligand disagrees by construction, though part of this gap is close to definitional given that the annotation target is itself length-flexible (Section 7).

One reading is that the signal a peptide is a BLB2 ligand does not require resolving which window matches the annotated core: the CNN, BiLSTM, and Transformer branches already carry enough of it, so the extra candidate slots may cost a little statistical efficiency rather than buying accuracy. This is not an argument against the register mechanism generally, since the negative controls show that removing learned register selection costs considerably more than restricting its candidate length does; it is an argument that length flexibility affects interpretability far more than classification ranking here, with the caveat that part of the apparent null result may reflect mislabeled 9-mer-annotated positives (Section 4) rather than the inductive bias being tested.

For the broader ML question: matching a latent alignment variable's support to annotated structural heterogeneity is not guaranteed to help the task the model is nominally trained for, even when that heterogeneity is annotated in roughly half the positive examples. What clearly does matter is whether the model's internal explanation can engage with peptides whose annotated core exceeds 9 residues at all, which bears on any downstream use of a register output beyond a binary presentation call, where the pooled recall of 0.401 (Section 5) is itself a reason for caution.

9 Limitations

Scope. Our evaluation covers a single chicken MHC-II allele, one species, one architecture, and one training seed; we make no claim that any result generalizes to human HLA-II, other chicken alleles, or even quantitatively to BLB2 under a different dataset or negative-sampling scheme. BLB1 models exist in this project and are out of scope; we report only BLB2 and note this explicitly rather than let a reader wonder if BLB1 was omitted for comparing less favorably against MixMHC2pred.

The auxiliary-supervision confound. As detailed in Sections 3 and 6, the ablation differs from the flexible model in candidate-length support and auxiliary-supervision volume at once; the decisive control that would separate them is not run here.

Label quality. The `Core` annotation is computationally inferred, not experimentally verified per peptide (Section 2), so the confusion matrix in Section 7 is partly circular for the reason given there, and, per Section 4, an unknown share of "BLB2-positive, 9-mer core" positives may be a co-expressed allele's misassigned ligands (a $13\times$ label-overlap lower bound, uncorrected for). It remains a live competing explanation for part of our central result (Section 6).

Statistical scope. The paired bootstrap in Section 6 quantifies evaluation-sampling uncertainty given our trained models, not seed-to-seed retraining variance, which we do not estimate; both comparisons are observed differences under a single training run each, not claims tested for robustness to re-training. MixMHC2pred's chicken training data may also overlap our peptide set, so that comparison is directional, not a clean generalization test.

10 Conclusion

BLB2's ligands are annotated with 9-mer and 10-mer binding cores in nearly equal proportion. Structurally, the existence of a decamer mode is established for this allele; the near-even split is a separate, computational finding, and no standard MHC-II predictor's architecture models core length as a first-class variable. We used BLB2 as a controlled test case for whether a register variable, trained with explicit supervision from the annotated core, benefits from ranging over both candidate lengths. Under a parameter-matched ablation, we found no evidence that flexible-length modeling improves ligand-classification ranking for BLB2, and a small, bootstrap-supported PR-AUC gap in the restricted model's favor that we cannot yet attribute to candidate-length support specifically rather than to auxiliary-supervision volume. Both variants scored above tuned negative controls, consistent with the register mechanism functioning as intended.

Flexibility does matter for interpretability: only the flexible model can agree with a 10-mer annotation at all, agreeing on 61% of annotated 10-mer cores that a fixed-9-mer model is guaranteed to disagree with, though part of this gap is definitional given that the annotation target is itself length-flexible. We also identified a competing explanation for part of our central result: a $13\times$ enrichment for co-occurring BLB1 positivity among 9-mer-annotated BLB2 positives, consistent with Halabi et al.'s account of co-expression, and with the largest effect we observe anywhere in this paper, a roughly 0.23 PR-AUC gap between the 9-mer and 10-mer subgroups. Without ruling out that confound, register-length flexibility looks like a lever on interpretability rather than classification accuracy here, a claim scoped to one allele, one architecture, and one training seed.

AI use statement

Generative AI tools were used to assist with manuscript preparation and research support. Claude (Anthropic) assisted with drafting and editing text based on author-provided methods, settings, results, and interpretations, and with selected code modifications, debugging, and implementation ideas. ChatGPT (OpenAI) assisted with identifying some relevant literature. All AI-assisted content, code, and references were independently reviewed and verified by the authors. Generative AI was not used to generate or estimate experimental results, and the authors take full responsibility for the final content.

Ethics statement

This work uses only publicly available or previously compiled chicken MHC-II ligand sequence data. No human subjects, patient data, or live animals were involved. We are not aware of a dual-use or harm concern specific to this work.

Reproducibility statement

Optimizer, learning rate, batch size, auxiliary loss weight, and seed are in Section 3; per-branch hyperparameters and training/early-stopping details are in Table 5 (Appendix A.2). Negative sampling, threshold selection, dataset construction, and fold assignment are in Sections 4–6, with zero protein-level leakage verified programmatically (an unrun redundancy check is flagged in Appendix A.2). Every number traces to a file in an anonymized manifest (Appendix A.5); code will be released as an anonymized repository at submission.

REFERENCES

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *International Conference on Machine Learning (ICML)*, pp. 233–240, 2006. doi: 10.1145/1143844.1143874.

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander M. Rush. Latent alignment and variational attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.

Anna Goryanin, Anastasia G. Cook, Shahriar Behboudi, Jim Kaufman, and Samer Halabi. A major histocompatibility complex (MHC) class II molecule that binds the same viral pathogen peptide with both nonamer and decamer core sequences for presentation to T cells. *Journal of Virology*, 100(5):e01859–25, 2026. doi: 10.1128/jvi.01859-25.

Samer Halabi, Mou Ghosh, Stefan Stevanović, Hans-Georg Rammensee, Luca D. Bertzbach, Benedikt B. Kaufer, Martin C. Moncrieffe, Bernd Kaspers, Sonja Härtle, and Jim Kaufman. The dominantly expressed class II molecule from a resistant MHC haplotype presents only a few Marek's disease virus peptides by using an unprecedented binding motif. *PLOS Biology*, 19(4):e3001057, 2021. doi: 10.1371/journal.pbio.3001057.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. Structured attention networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Brian W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, 1975. doi: 10.1016/0005-2795(75)90109-9.

Julien Racle, Justine Michaux, Georg A. 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*, 37:1283–1286, 2019. doi: 10.1038/s41587-019-0289-6.

Julien Racle, Philippe Guillaume, Julien Schmidt, Justine Michaux, Amédé Larabi, Kelvin Lau, Marta A. S. Perez, Giancarlo Croce, Raphaël Genolet, George Coukos, Vincent Zoete, Florence Pojer, Michal Bassani-Sternberg, Alexandre Harari, and David Gfeller. Machine learning predictions of MHC-II specificities reveal alternative binding mode of class II epitopes. *Immunity*, 56(6):1359–1375, 2023. doi: 10.1016/j.immuni.2023.03.009.

Colin Raffel, Minh-Thang Luong, Peter J. Liu, Ron J. Weiss, and Douglas Eck. Online and linear-time attention by enforcing monotonic alignments. In *International Conference on Machine Learning (ICML)*, 2017.

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. doi: 10.1093/nar/gkaa379.

Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997. doi: 10.1109/78.650093.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Gene-Ping Yang and Hao Tang. Supervised attention in sequence-to-sequence models for speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. doi: 10.48550/arXiv.2204.12308.

Ronghui You, Wei Qu, Hiroshi Mamitsuka, and Shanfeng Zhu. DeepMHCII: a novel binding core-aware deep interaction model for accurate MHC-II peptide binding affinity prediction. *Bioinformatics*, 38(Supplement_1):i220–i228, 2022. doi: 10.1093/bioinformatics/btac231.

Chenpeng Yu, Xing Fang, and Hui Liu. A unified cross-attention model for predicting antigen binding specificity to both HLA and TCR molecules. *Nature Machine Intelligence*, 7:278–292, 2025. doi: 10.1038/s42256-024-00973-w.

A Appendix

A.1 Dataset figures

Figures 2 and 3 show the BLB2 class-balance and core-length distributions whose numbers are reported in prose in Section 4 (4,184 positives vs. 58,406 negatives; 2,134 9-mer vs. 2,050 10-mer annotated cores). They are moved here, rather than into the main text, purely to satisfy the page limit.

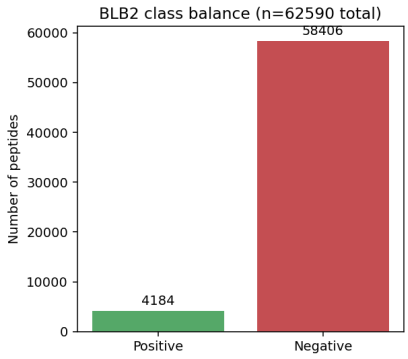

Figure 2: BLB2 class balance: 4,184 positive vs. 58,406 negative peptides.

A.2 Hyperparameters and training details

Table 5 lists every branch-level architecture hyperparameter, plus epoch budgets and early-stopping criteria, for the two models compared in this paper (both models share these values identically, since the ablation changes only the register candidate mask).

Known gap. Splits are grouped by source protein to prevent protein-level leakage (Section 4), but we did not run a sequence-identity-based redundancy check across folds, for example for near-identical peptides shared between homologous or paralogous proteins placed on opposite sides of a split. We flag this as an open reproducibility item rather than claim it is handled.

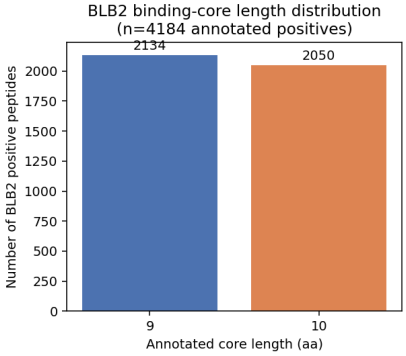

Figure 3: BLB2 annotated binding-core length distribution among 4,184 positive peptides: 2,134 (51.0%) nine-residue cores vs. 2,050 (49.0%) ten-residue cores.

Table 5: Full hyperparameter listing. "Screening/tuning" columns describe the inner-fold model-selection phase; "Final retrain" describes training the frozen model once on the full outer-development partition before the single outer-test prediction (Section 3).

Hyperparameter	Value
Token embedding dim (per branch, independent)	64
CNN branch: filters / kernel sizes	128 / (3, 5, 7)
BiLSTM branch: hidden dim / layers	128 / 2 (bidirectional)
Transformer branch: layers / heads / FF dim	2 / 4 / 256
Register branch Transformer: layers / heads / FF dim	4 / 4 / 256
Register branch parallel CNN: filters / kernel sizes	64 / (3, 5, 7)
Register length embedding dim	8
Fusion projection dim (per branch, before concat)	128
Dropout (all branches)	0.3
Optimizer	Adam
Learning rate	10^{-3}
Batch size	64
Seed	42 (single seed, no averaging)
Auxiliary register loss weight	0.3
Screening/tuning: epoch cap / patience	100 / 10 (early stopping)
Screening/tuning: selection metric	smoothed validation PR-AUC
Final retrain: epoch budget	2 (median best-epoch across 4 inner folds; per-fold values [2,3,2,5] for the flexible model)
Final retrain: early stopping	none (fixed epoch budget, no validation split)

A.3 Architecture screening and loss-tuning tables

Table 6 is the full architecture screening referenced in Section 4; Table 7 is the loss-tuned inner-fold comparison referenced in Section 6. Both are moved here, rather than into the main text, purely to satisfy the page limit; the final 5-outer-fold held-out numbers that matter most for this paper's central comparison remain in the main text (Table 2).

A.4 Note on the full architecture-screening leaderboard

Table 6 in Appendix A.3 reports the four models relevant to this paper's core comparison. The complete, machine-generated leaderboard across all 25 catalog models evaluated during architecture screening (BLB2, inner-fold OOF, outer round 1) is not reproduced in this PDF, since it is a large,

Table 6: BLB2 architecture screening, inner-fold OOF, outer round 1, all trained with the default (untuned) BCE loss. All rows use the same folds, negative sampling, and training procedure. PR-AUC is the primary metric. The flexible model (row 4) and the 9-mer-only ablation are not directly comparable at this untuned resolution; the ablation was evaluated only at tuned resolution (Table 7).

#	Arm	Model (catalog name)	PR-AUC	MCC
1	Best single-branch baseline	*bilstm*	0.5172	0.4607
1b	Best classical baseline	*random_forest*	0.2404	0.0888
2/3	Fusion, no register (arms 2 & 3 collapse – see text)	*e9_full_noregister*	0.5199	0.4811
4	+ register, flexible 9/10-mer (main model)	*e10_full_nogate*	0.5352	0.4838

Table 7: BLB2 ablation, loss-tuned inner-fold OOF (outer round 1, best of 3 losses per model).

Model	Best loss	PR-AUC	MCC
e10_full_nogate_9mer_only (ablation)	*bce*	**0.5600**	**0.5044**
e10_full_nogate (flexible, main model)	*weighted_bce_mild*	0.5520	0.4964

machine-generated table rather than a small hand-curated one; it is included in the anonymized code repository referenced in Appendix A.5.

A.5 File manifest

Every number in this paper traces to a specific output file. The full mapping, source file, exact value, and which table or sentence it supports, is maintained as a single manifest document (*RESULT_SOURCES.md*) rather than repeated inline in the main text. The scripts that produced the ablation, the subgroup analysis, the matched-subset external comparison, and the bootstrap interval are listed there by name; we intend to release both the manifest and the underlying code as an anonymized repository at submission time, per ICLR's anonymous-supplement policy, since none of it is meaningfully checkable by a reviewer as free-text file names alone.