Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

BioAnchorFormer: Protein-Aware Negative Sampling and Anchor-Biased Transformers for Chicken as Model Avian MHC-I Peptide Binding Prediction

Sarathkumar Devaraj

sxd1910@mavs.uta.edu

The University of Texas at Arlington

Shouyi Wang

The University of Texas at Arlington

Hung-Yueh Yeh

United States Department of Agriculture

Stephen N. White

United States Department of Agriculture

Research Article

Keywords: MHC-I, peptide binding prediction, BioAnchorFormer, NC positional encoding, protein

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11554038/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

BioAnchorFormer: Protein-Aware Negative Sampling and Anchor-Biased Transformers for Chicken as Model Avian MHC-I Peptide Binding Prediction

Sarathkumar Devaraj[1,*] Shouyi Wang[1] Hung-Yueh Yeh[2] Stephen N. White[2]

[1]Department of Industrial, Manufacturing and Systems Engineering,
The University of Texas at Arlington, Arlington, TX 76019, USA

[2]US National Poultry Research Center, Poultry Microbiological Safety & Processing Research,
USDA Agricultural Research Service, Athens, GA 30605, USA

[*]**Corresponding author:** Sarathkumar Devaraj, sxd1910@mavs.uta.edu

Abstract

Background: Major Histocompatibility Complex class I (MHC-I) peptide binding prediction is essential for rational vaccine design in poultry, yet computational tools for avian species lag far behind their human and murine counterparts. Existing approaches are hampered by two key methodological weaknesses: biologically implausible random negative sampling and peptide-level data splits that permit sequence-homology leakage between training and test sets.

Methods: We constructed a chicken BF2*21:01 dataset of 63,112 peptides (3,788 known positives, lengths 5–50; 59,324 negatives) from 2,797 unique proteins using a protein-aware negative sampling strategy that draws hard negatives from unmasked regions of positive-source proteins and diverse negatives from unrelated proteins. A two-stage experimental design was employed: (i) model development across eight progressive stages spanning 19 configurations and (ii) a final unbiased outer-fold evaluation of the three shortlisted models using frozen hyperparameters and 5-fold StratifiedGroupKFold splits with max 200 epochs and patience 20.

Results: In the final unbiased outer-fold evaluation with frozen hyperparameters, BioAnchorFormer-Ultimate (C6-BAF-Ultimate, proposed) achieves PR-AUC 0.750 ± 0.013 and MCC 0.708 ± 0.005 on the full dataset, the most consistent performance across outer folds. Restricted to canonical 8–11 mer peptides, performance rises to PR-AUC 0.857 ± 0.021 and MCC 0.798 ± 0.016, confirming that the full-dataset figures are conservative estimates driven partly by non-canonical-length noise.

Conclusions: BioAnchorFormer constitutes the first dedicated, leakage-controlled predictor for avian MHC-I, established through protein-grouped evaluation that prevents sequence-homology inflation. On the full 8–15 mer dataset the proposed model outperforms deep-sequence baselines;

on the canonical 8–11 mer subset, models trained and evaluated exclusively on canonical lengths (Stage 9e; 3 seeds × 5 folds) show statistically equivalent performance, with C4-BAF-OneHead achieving the highest MCC (0.800 ± 0.017) and BioAnchorFormer variants exhibiting lower cross-seed variance than the BiLSTM baseline. Interpretability analysis reveals reproducible terminal-anchor attention patterns and physicochemical weighting consistent with known MHC-I biochemistry. The full pipeline, dataset, and model weights are publicly available at https://avianepitope.org.

Keywords: MHC-I; peptide binding prediction; BioAnchorFormer; NC positional encoding; protein-aware negative sampling; avian immunology; deep learning; vaccine design; interpretability

1 Background

The adaptive immune system is a crucial vertebrate defense system. Jawed vertebrates including birds have both B cell and T cell branches of their immune responses [1]. While B cells produce antibodies for extracellular defense, cytotoxic T cell immunity enables defense against intracellular pathogens, including not only viruses but also intracellular bacteria, parasites, fungi, and more ([2,3]).

The Major Histocompatibility Complex class I (MHC-I) pathway is the primary mechanism by which nucleated cells present intracellular peptide fragments to cytotoxic T lymphocytes (CTLs) [4]. Peptides of typically 8–11 residues are processed in the cytoplasm, loaded onto MHC-I molecules in the endoplasmic reticulum, and displayed on the cell surface for surveillance by $CD8^+$ T cells. Stable peptide–MHC-I binding is a prerequisite for CTL recognition, making the computational identification of binding peptides (T-cell epitopes) a cornerstone of rational vaccine design [5,6]. While there are broad functional similarities across vertebrates including birds [7], there are important structural and functional differences between more widely studied mammals and birds [8,9].

In the domestic chicken (*Gallus gallus*), MHC-I diversity and its relationship to disease resistance—particularly against Marek's disease virus (MDV), Newcastle disease virus (NDV) and influenza—have attracted growing research interest [1,10–12]. However, the computational infrastructure for avian MHC-I epitope prediction lags markedly behind its human and murine counterparts. The Immune Epitope Database (IEDB) contains far fewer experimentally validated avian peptides, and

species-specific tools remain limited [13]. Among the known chicken BF2 alleles, only BF2*21:01 has been characterised by immunopeptidomics at a scale sufficient for machine learning; experimental peptide data for structurally studied alleles such as BF2*13:01, BF2*15:01, and BF2*19:01 are limited to tens to low hundreds of mass-spectrometry-eluted peptides [14,15]—one to two orders of magnitude below the training threshold for allele-specific deep learning models. Developing accurate predictors for MHC-I is therefore timely and practically important for avian vaccine design in general and for poultry applications in particular.

Two methodological problems recur in the existing literature. First, many published datasets generate negative (non-binding) peptides by *random sequence sampling*: arbitrary amino-acid combinations or shuffled fragments labelled as non-binders. Random negatives are biologically unrealistic because they do not reflect the actual distribution of non-binding peptides arising from natural protein processing. In vivo, the peptidome is shaped by proteasomal cleavage and TAP transport; naive random sampling ignores these constraints and may include sequences that are genuinely immunogenic. Second, most published train/test splits are constructed at the peptide level, meaning that a test peptide may originate from the same source protein as a training peptide, constituting a form of sequence-homology leakage [16–18]. Because neighbouring peptides from the same protein share sequence context, this constitutes a form of data leakage that inflates apparent generalisation.

We address both problems in a unified framework. The key contributions of this work are:

- A **protein-aware negative sampling strategy** that maps experimentally validated binders back to their source proteins, masks all confirmed binding windows, and draws negatives from the unmasked residue space in two tiers: *hard negatives* (from the same protein as positive peptides, representing structurally proximal non-binding contexts) and *diverse negatives* (from unrelated proteins), yielding a biologically realistic 1:15.7 positive-to-negative ratio.
- A novel **NC Positional Encoding** (NCEnc) that applies sinusoidal positional encoding bidirectionally—once from the N-terminus and once from the C-terminus—so that the functionally critical P2 and PΩ anchor positions receive consistent representations regardless of peptide length.
- **BioAnchorFormer**, a hybrid CNN–Transformer model whose self-attention scores are augmented by a learnable position-pair anchor bias matrix applied to a dedicated single head, and

a pairwise physicochemical similarity term with learned per-property weights.

- A two-stage evaluation design: (i) eight progressive development stages spanning **19 model configurations** from classical baselines through ensemble methods, all evaluated under **protein-grouped StratifiedGroupKFold** (zero protein overlap between folds), and (ii) a final **outer-fold evaluation with frozen hyperparameters** to report unbiased performance estimates for the top-three shortlisted models.
- **Interpretability analysis** showing that the learned anchor bias matrix and physicochemical weights recover known BF2*21:01 binding motif biology without positional supervision.

2 Related Work

2.1 Conventional MHC-I Binding Prediction

NetMHC [19] pioneered data-driven MHC-I prediction using position-specific scoring matrices and neural networks. NetMHCpan-4.1 [20] extended this with pan-allele prediction via pseudo-sequence representation and integration of eluted ligand data, and represents the current state-of-the-art for well-studied human alleles. These tools lose reliability for alleles with sparse experimental coverage, as is typical for chicken MHC-I, because the pseudo-sequence alignment may not faithfully encode novel allele-specific groove geometry.

2.2 Deep Learning Approaches

CNNs have proven effective at capturing local binding-motif patterns in peptide sequences [21], while recurrent architectures (LSTM [22]; BiLSTM [23]) capture sequential dependencies. Transformer-based models [24], including protein language models such as ESM-2 [25] and the MHC-specific BigMHC [26], exploit large-scale pre-training to yield richer residue embeddings. Despite their success on human and murine data, the benefit of MHC-specific inductive biases—particularly anchor position awareness and physicochemical compatibility modelling—for avian MHC-I has not been systematically studied.

2.3 Negative Sampling and Data Leakage

Common negative-sampling strategies include random amino-acid generation [27], binder shuffling, or experimentally tested non-binders; none account for the source-protein context of natural peptide

processing. The general principle that negatives should be drawn from real protein-sequence context — rather than random or shuffled sequences — and that mismatched negatives inflate apparent performance is well established in binding and antimicrobial-peptide prediction [18]. Our protein-aware strategy is an adaptation of this recognised best practice to the specific constraints of chicken as model avian data.

Separately, peptide-level cross-validation splits permit sequence-homology leakage when neighbouring peptides from the same source protein are split across train and test sets, inflating reported generalisation. This has been explicitly demonstrated in the pMHC and T-cell receptor (TCR) literature, where protein-grouped "hard splits" substantially reduce benchmark scores relative to peptide-level splits [16, 17]. Our pipeline addresses this at both stages: protein-grouped StratifiedGroupKFold for all cross-validation splits, with zero protein overlap verified programmatically.

3 Methods

3.1 Dataset Construction

3.1.1 Positive Peptides

Experimentally validated MHC-I binding peptides for *Gallus gallus* (allele BF2*21:01) were retrieved from IEDB [28] (https://www.iedb.org) and two published chicken immunopeptidomics studies [13, 29]. MS-eluted ligands with confirmed mass-spectrometry support were retained as positives. After deduplication, 3,788 positive peptides (3,731 with mapped source proteins and 57 unmatched) were retained.

3.1.2 Protein Mapping and Binding-Window Masking

Each positive peptide was located within its full-length source protein sequence by exact substring search; source protein sequences were retrieved from locally downloaded UniProt FASTA files comprising 43,712 *Gallus gallus* protein entries: manually curated Swiss-Prot entries (reviewed, experimentally supported) and computationally predicted TrEMBL entries (unreviewed, including putative and hypothetical proteins), with UniParc used as fallback for unresolved accessions. All confirmed positive binding windows (start to end position) were masked in the source protein sequence.

3.1.3 Protein-Aware Negative Sampling

Negatives were drawn in two complementary tiers from the unmasked residue space of the same UniProt *Gallus gallus* proteome (43,712 entries) used for protein mapping:

Hard negatives (same-protein) are sampled from unmasked windows of proteins that also contain binding peptides. They are structurally proximal to confirmed binders in sequence context and represent the most challenging decision-boundary cases.

Diverse negatives (unrelated-protein) are sampled from proteins that contain no known binding peptides, providing broad coverage of non-binding sequence space.

3.1.4 Dataset Statistics

The final dataset contains 63,112 peptides: 3,788 positives (3,731 mapped; 57 unmatched), 53,931 same-protein hard negatives, and 5,393 diverse negatives from unrelated proteins (Table 1). The overall positive-to-negative ratio is approximately 1:15.7, representing the natural sparsity of immunogenic peptides in the proteome. Peptides span 2,797 unique source proteins. Peptide lengths range from 5 to 50 residues (median 10, mean 11.3). TThe positive set includes peptides ranging from 5 to 50 amino acids, spanning both within and outside the canonical MHC-I groove length range (8–11 aa); 5–7 mers and peptides >15 aa were retained from IEDB as reported by mass-spectrometry experiments but represent a small fraction of positives and may include annotation artefacts or processed precursors rather than direct eluted ligands. The subset restricted to canonical 8–15 mer lengths (used for the NetMHCpan comparison in Stage 10) yields consistently higher performance ($\sim$8 pp PR-AUC higher than the full-length set), suggesting that non-canonical length entries act as mild noise in the full-set evaluation. Sequences longer than $L_{\max} = 30$ are truncated at the 99th-percentile cutoff. Descriptive statistics (counts, ratios, length distributions) were computed using Python 3.11 with pandas 2.0 and NumPy 1.24.

Table 1: Dataset composition for BF2*21:01 binding prediction

Sample Type	Count	Label
Positive (mapped)	3,731	1
Positive (unmatched)	57	1
Hard negative (same protein)	53,931	0
Diverse negative (unrelated protein)	5,393	0
Total	**63,112**	

3.2 Input Representation

Sequence encoding. Each peptide is tokenised over a vocabulary of 22 tokens: 20 standard amino acids (1-indexed), one padding token (index 0), and one unknown token (index 21). Sequences are padded or truncated to a maximum length $L_{\max} = 30$ (99th-percentile peptide length). Transformer-based models prepend a learnable [CLS] token, giving input length $L_{\max} + 1 = 31$.

Physicochemical feature matrix. In parallel, each peptide is represented as a per-residue physicochemical matrix of shape $L_{\max} \times 4$, where the four channels encode: (1) Kyte–Doolittle hydrophobicity [30] normalised to $[-1, 1]$ by dividing by 4.5 (maximum absolute value), (2) formal charge at pH 7 ($\{-1, 0, +1\}$) [31], (3) polarity (binary, $\{0, 1\}$) [31], and (4) residue molecular weight normalised by the maximum (Trp = 204 Da, range $[0, 1]$) [32]. All four channels therefore lie in $[-1, 1]$ or $[0, 1]$, providing comparable scales for the learned per-property weights w_k. Padding rows are zero. This matrix serves as input to the bio-bias term in BioAnchorFormer.

3.3 Baseline Models (Stages 1–3)

Stage 1 — Classical and neural baselines. Four classical models (Logistic Regression [33], SVM [34], Random Forest [35], XGBoost [36]) are trained on a 31-dimensional biochemical vector (20-dim amino-acid composition, 4-dim physicochemical means, 7-dim residue-group fractions). Three neural models (CNN, BiLSTM, vanilla Transformer) are trained on sequence tokens. The CNN uses multi-kernel 1-D convolution with kernel sizes (3, 5, 7), 128 filters per kernel, global max pooling, and two fully connected layers. The BiLSTM uses two bidirectional LSTM layers with hidden size 128. The vanilla Transformer uses standard sinusoidal positional encoding, 4 attention heads, 3

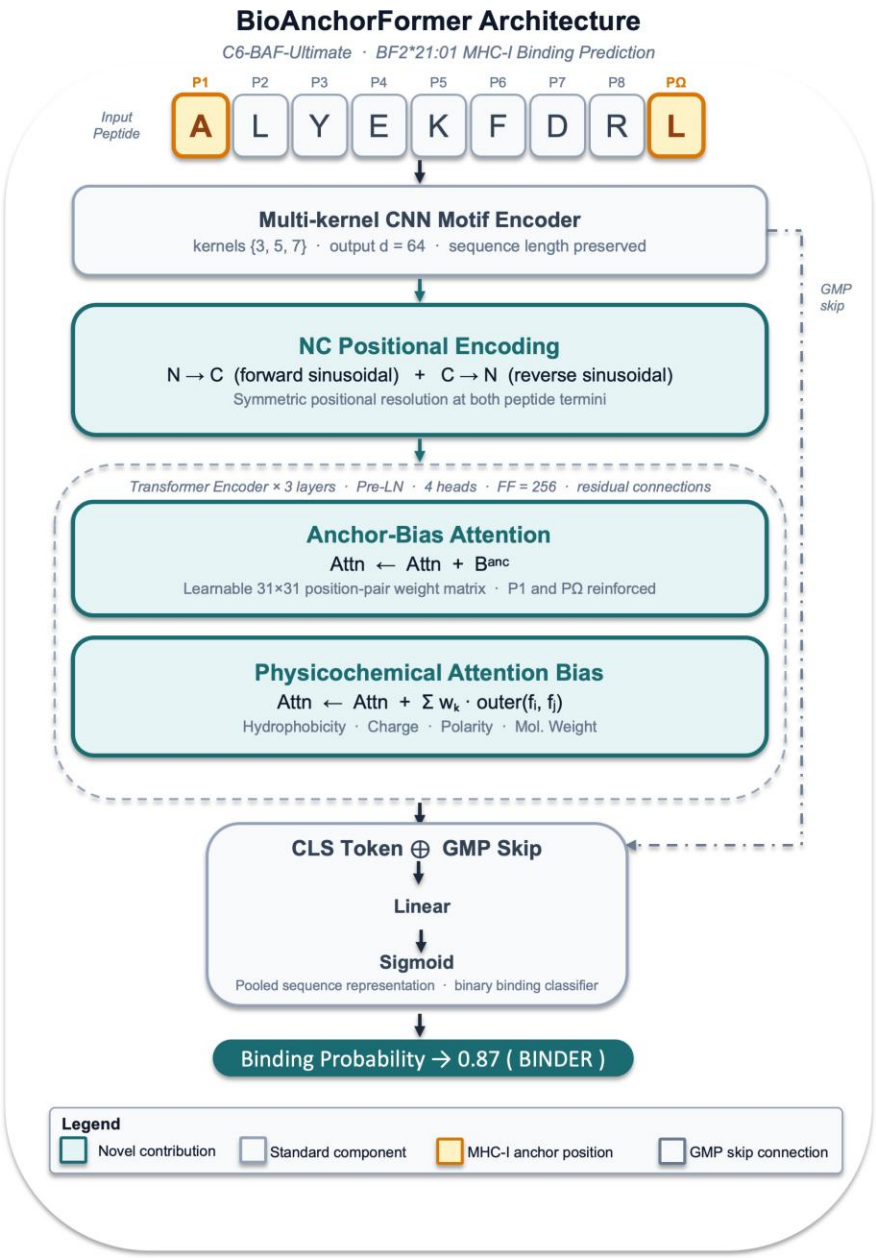

Figure 1: **BioAnchorFormer (C6-BAF-Ultimate) architecture.** Peptide tokens enter the multi-kernel CNN motif encoder (kernels 3, 5, 7), whose output ($d = 64$, sequence length preserved) is prepended with a learnable [CLS] token. NC Positional Encoding superimposes forward (N-terminal) and reverse (C-terminal) sinusoidal encodings so that anchor positions receive consistent representations regardless of peptide length. Within each of the three Transformer encoder layers (Pre-LN, 4 heads, FF = 256), self-attention scores are augmented by two additive biases: a learnable anchor-bias matrix $\mathbf{B}^{\text{anc}} \in \mathbb{R}^{31\times31}$ applied to head 0 only, and a pairwise physicochemical outer-product bias with learned per-property weights. A Global Max Pooling (GMP) skip connection from the CNN output is concatenated to the [CLS] embedding before the binary binding classifier.

encoder layers, feed-forward dimension 256, and a [CLS] classifier; all neural models use embedding dimension $d = 64$.

Stage 2 — Bio-feature fusion. CNN+Bio, BiLSTM+Bio, and Transformer+Bio append the 31-dimensional biochemical vector to the pooled sequence representation before the classifier head.

Stage 3 — CNN–Transformer hybrids. A shared CNN encoder (kernels (3, 5, 7), $d = 64$, sequence length preserved) precedes a Transformer encoder. Three variants are tested: plain CNNTransformer, CNNTransformer+Anchor (learnable position-pair bias only), and CNNTransformer+Bio (pairwise physicochemical bias only).

3.4 NC Positional Encoding

MHC-I peptides are constrained at both termini: the N-terminal anchor P2 and the C-terminal anchor PΩ form critical hydrogen bonds with conserved MHC-I groove residues across virtually all alleles [37]. Standard sinusoidal positional encoding assigns different absolute indices to PΩ of a 9-mer (index 9) and a 12-mer (index 12), despite their identical structural roles.

NC Positional Encoding resolves this by superimposing two complementary sinusoidal encodings:

$$\mathrm{NC}(x, t) = \mathrm{PE}_{\mathrm{fwd}}(t) + \mathrm{PE}_{\mathrm{rev}}(L_x - t + 1), \tag{1}$$

where L_x is the actual (non-padded) peptide length, $\mathrm{PE}_{\mathrm{fwd}}$ is the standard sinusoidal forward encoding, and $\mathrm{PE}_{\mathrm{rev}}$ is its mirror applied from the C-terminus. The forward component encodes N-terminal context (P1, P2, …) and the reverse component encodes C-terminal context (…, PΩ−1, PΩ) consistently across all peptide lengths.

Position indexing convention. Throughout this paper, peptide positions follow biological 1-based numbering: P1 is the N-terminal residue, P2 the second residue (primary anchor in canonical MHC-I), and PΩ the C-terminal residue. In the model's input tensor, the [CLS] token occupies sequence index 0, and the peptide begins at index 1 (= P1). NC encoding is applied only to peptide positions P1–PL_x; the CLS token receives no positional encoding and is updated solely through attention.

3.5 BioAnchorFormer (Stages 4–8)

BioAnchorFormer extends the CNNTransformer hybrid with two additive attention biases. The combined pre-softmax attention score is:

$$\text{score}(i, j) = \frac{Q_i K_j^\top}{\sqrt{d_k}} + \mathbf{B}^{\text{anc}}_{i,j} + \sum_{k=1} w_k \cdot p_{ik} \cdot p_{jk}, \tag{2}$$

where $\mathbf{B}^{\text{anc}} \in \mathbb{R}^{31\times 31}$ is a learnable anchor-bias matrix capturing universal position-pair importance, p_{ik} is the k-th physicochemical property of residue i, and $w_k \geq 0$ with $\sum_k w_k = 1$ are per-property weights learned via softmax. In the final model (C6-BAF-Ultimate), $\mathbf{B}^{\text{anc}}$ is applied to head 0 only, isolating positional interactions from content-based QK attention.

Stage 4 evaluates: BioAnchorFormer (full bias, standard PE), CNNTransformer+NCEnc (NC encoding only), and BioAnchorFormer+NCEnc (full combination).

Stage 5 ablates individual training-recipe improvements on top of BioAnchorFormer+NCEnc:

- **S5-LargeEnc**: doubled capacity ($d = 128$, 4 encoder layers).
- **S5-GMP**: global max-pool skip connection from the CNN output.
- **S5-FocalLoss**: Focal Loss [38] with $\gamma = 2$.
- **S5-CosineLR**: cosine annealing with 5-epoch linear warmup.
- **S5-LabelSmooth**: label smoothing with $\varepsilon = 0.05$.
- **S5-Augment**: conservative amino-acid swap augmentation ($p_{\text{aug}} = 0.10$ per residue).
- **BioAnchorFormer-Ultimate**: all improvements combined (Figure 1).

Stage 6 systematically ablates architectural components (NC encoding variants, anchor-bias parameterisations including the dedicated single-head variant S6-AB-OneHead, and physicochemical bias formulations). Stage 7 freezes optimal decision thresholds from the best Stage-6 models; no new architectures are trained at this stage. Stage 8 evaluates deeper variants, ESM-2 augmentation [25], and a soft-voting ensemble.

3.6 Loss Functions and Optimisation (Development Stages)

During model development (Stages 1–8), neural models use BCEWithLogitsLoss with class-frequency-based positive weighting unless otherwise specified. The Adam optimiser is used with initial learning rate $\eta = 5 \times 10^{-4}$, weight decay 10^{-4}, and ReduceLROnPlateau scheduling (patience 12, factor 0.5). Models are trained for up to 60 epochs with early stopping (patience 12) and batch size 128.

3.7 Experimental Setup (Development Stages)

Data splitting. Stages 1–6 share a common set of pre-computed 4-fold inner StratifiedGroupKFold assignments (grouped by protein accession number, zero protein overlap between training and validation partitions; random seed 42), generated by phase1_fold_split.py on the development partition of outer fold 1 (rows with outer_fold $\neq 1$, approximately 80% of the dataset). The same fold_split.csv indices are shared identically across all Stage 1–6 model configurations.

Stage 8 (deep-variant exploration with S8-Deep and S8-ESM2) was conducted independently using a 5-fold StratifiedKFold on the full dataset; those results serve as exploratory context and the Stage-8 models were not carried forward to the final outer-fold evaluation.

Operating thresholds are selected per fold by maximising F1 on the validation set during development. In Stage 9 (outer-fold evaluation), the frozen primary threshold is MCC-optimal: a grid search over $\vartheta \in \{0.01, 0.02, \ldots, 0.99\}$ (99 candidate values, step 0.01) was performed on the inner-fold out-of-fold predictions from Stage 6, and the threshold $\hat{\vartheta}$ that maximised Matthews Correlation Coefficient (MCC) was selected and frozen prior to any exposure to the outer-fold test partitions. This threshold is held fixed across all five outer folds. The policy change from F1- to MCC-optimal thresholding is deliberate: MCC is a more balanced metric under strong class imbalance ($\approx 1:15.7$) and directly measures binary classification quality, which is the evaluation objective in Stage 9. F1 was used for development threshold selection because it converges faster during hyper-parameter search and is less sensitive to threshold initialisation.

Evaluation metrics. Given the ~1:15.7 class imbalance, Precision–Recall Area Under the Curve (PR-AUC) is the primary metric, as it is sensitive to classifier performance on the minority (positive) class. We also report F1, Receiver Operating Characteristic Area Under the Curve (ROC-AUC), and Matthews Correlation Coefficient (MCC). All metrics are reported as mean ± standard deviation

across folds.

3.8 Model Shortlisting and Hyperparameter Freezing (Stage 8)

Following the eight development stages, the three highest-performing model configurations were shortlisted for final evaluation: **BioAnchorFormer-Ultimate** (abbreviated **C6-BAF-Ultimate**; proposed model; NCEnc + single-head anchor bias + bio bias + GMP skip; weighted BCE loss), **BioAnchorFormer with Single-Head Anchor Bias** (abbreviated **C4-BAF-OneHead**; NCEnc + single-head anchor bias; hinge loss), and **BioAnchorFormer without Anchor Bias** (abbreviated **D3-NoAnchor**; NCEnc + bio bias + GMP skip, no anchor bias; weighted BCE loss). Hyperparameters — including learning rate, dropout, weight decay, and decision thresholds — were frozen at their development-stage values and were not re-tuned on any outer-fold data. Two thresholds were frozen per model: a *primary threshold* (MCC-optimal on development data) and a *laboratory threshold* (maximum precision at recall ≥ 0.70, for screening applications). The frozen configurations are recorded in configs/frozen_config.json (provided in the repository) and are immutable from this point forward.

3.9 Outer-Fold Evaluation Protocol (Stage 9)

To obtain unbiased performance estimates, we applied the three frozen models to a 5-fold outer evaluation using StratifiedGroupKFold with grouping by accession number, ensuring that peptides from the same source protein appear in only one outer fold. For each outer fold k ($k = 1, \ldots, 5$):

1. The model is trained from scratch on the outer-training partition (folds $\neq k$), using a stratified 80/20 split of that partition for early-stopping validation.
2. Training uses a maximum of 200 epochs with patience 20 on validation PR-AUC, batch size 256, and the frozen optimiser settings.
3. The trained model is evaluated on the outer test fold k at both the frozen primary and laboratory thresholds.

Results are reported as mean ± standard deviation across the 5 outer folds. All models are trained with random seed 42 (fixed across all outer folds via the frozen configuration). Model sizes: C6-BAF-Ultimate 199,750 parameters; C4-BAF-OneHead 197,570 parameters; D3-NoAnchor 198,789 parameters. For reference, the three deep-sequence baselines (evaluated in Stage 9b) have 46,465

(B1-CNN), 44,801 (B2-BiLSTM), and 153,665 (B3-Transformer) parameters. All experiments were run on a single NVIDIA GeForce RTX 4090 (24 GB VRAM). The 15 outer-fold training runs for the three shortlisted models completed in approximately 21 minutes; the 15 baseline outer-fold runs (Stage 9b) required a further 16 minutes (37 minutes total for all 30 outer-fold training runs).

3.10 Comparison with NetMHCpan-4.1 (Stage 10)

To contextualise BioAnchorFormer against the current state-of-the-art general-purpose tool, we ran NetMHCpan-4.1 [20] on the 8–15 mer subset of the outer-fold evaluation peptides. This comparison is explicitly out-of-distribution for NetMHCpan-4.1: the tool does not have BF2*21:01 in its pre-defined allele library. We supplied the experimentally determined 301-amino-acid mature heavy chain sequence as a custom allele via the -hlaseq flag. The %Rank statistic is not computed for user-defined alleles (reported as -99.9) and is therefore not used. We used Score_EL (EL score, higher = stronger binder) and Aff(nM) (binding affinity, lower = stronger binder) as ranking signals; the Aff(nM) model is expected to produce degenerate output for an unsupported allele, and results confirm this (ROC-AUC = 0.500, MCC = 0.000). This comparison is intended to demonstrate the absence of a usable general-purpose predictor for BF2*21:01, not as a head-to-head performance claim.

3.11 Interpretability Analysis (Stage 11)

To assess whether BioAnchorFormer captures biologically meaningful signals, we extracted and visualised the learned parameters from the five outer-fold checkpoints of C6-BAF-Ultimate and C4-BAF-OneHead.

Anchor bias matrix. The 31×31 parameter matrix $\mathbf{B}^{\text{anc}}$ was extracted from each fold checkpoint and averaged. The mean bias matrix (cropped to the CLS + P1–P16 region) was plotted as a heatmap using a diverging colour scale centred at zero.

Physicochemical feature weights. The four learned weights w_k (Equation 2) were extracted after softmax normalisation from each fold checkpoint and averaged. The mean ± SD across folds is compared to the uniform baseline ($\frac{1}{4} = 0.25$ per property, i.e. the weight each of the four properties would receive if the softmax output were uniform).

Attention weight extraction. For a representative set of 300 true-positive 9-mers and 300

true-negative 9-mers from all outer folds, we re-ran inference with the flags need_weights=True and average_attn_weights=False to obtain per-head, per-layer attention weights (B, H, S, S) from each of the three encoder layers. The CLS-to-position attention profile (mean over all heads and layers) was compared between binders and non-binders to test whether the model attends preferentially to anchor positions.

4 Results

4.1 Stage 1: Classical and Neural Baselines

Table 2 shows that classical models (PR-AUC 0.17–0.35) have limited discriminative power when operating on the 31-dimensional biochemical vector alone, confirming that aggregate composition features cannot capture the residue-level specificity required for MHC-I binding. Among neural baselines, CNN (PR-AUC 0.784 ± 0.028, MCC 0.705 ± 0.038) and BiLSTM (PR-AUC 0.774 ± 0.028, MCC 0.699 ± 0.025) perform comparably, both substantially outperforming the vanilla Transformer (PR-AUC 0.656 ± 0.016, MCC 0.597 ± 0.022). The Transformer's weaker performance at this stage is consistent with the known difficulty of learning useful self-attention patterns for short, constrained peptide sequences without architecture-specific inductive biases.

Table 2: Stage 1 baseline model performance (4-fold protein-grouped StratifiedGroupKFold)

Model	F1	ROC-AUC	PR-AUC	MCC
LR	0.246 ± 0.010	0.770 ± 0.011	0.171 ± 0.004	0.195 ± 0.014
SVM	0.364 ± 0.006	0.863 ± 0.009	0.280 ± 0.011	0.328 ± 0.008
Random Forest	0.398 ± 0.013	0.875 ± 0.011	0.353 ± 0.021	0.361 ± 0.015
XGBoost	0.369 ± 0.016	0.865 ± 0.010	0.322 ± 0.011	0.332 ± 0.016
CNN	0.719 ± 0.037	0.956 ± 0.008	0.784 ± 0.028	0.705 ± 0.038
BiLSTM	0.715 ± 0.025	0.955 ± 0.007	0.774 ± 0.028	0.699 ± 0.025
Transformer	0.620 ± 0.020	0.936 ± 0.007	0.656 ± 0.016	0.597 ± 0.022

4.2 Stage 2: Bio-Feature Fusion

Table 3 shows that appending the 31-dimensional biochemical vector yields small but consistent gains across all architectures. CNN+Bio (PR-AUC 0.804 ± 0.019, MCC 0.728 ± 0.011) improves modestly

over plain CNN, suggesting that sequence tokens already capture most discriminative biochemical signal but that explicit summary statistics provide complementary information. Transformer+Bio (PR-AUC 0.680 ± 0.037) shows a larger relative improvement over vanilla Transformer (+3.7 pp in PR-AUC), confirming that explicit biochemical priors partially compensate for the Transformer's weaker sequence-only performance.

Table 3: Stage 2 bio-feature fusion performance (4-fold protein-grouped StratifiedGroupKFold)

Model	F1	ROC-AUC	PR-AUC	MCC
CNN+Bio	0.743 ± 0.008	0.962 ± 0.004	0.804 ± 0.019	0.728 ± 0.011
BiLSTM+Bio	0.731 ± 0.021	0.960 ± 0.005	0.792 ± 0.015	0.715 ± 0.023
Transformer+Bio	0.626 ± 0.027	0.943 ± 0.010	0.680 ± 0.037	0.604 ± 0.030

4.3 Stages 3–4: Hybrid and Novel Architectures

Table 4 reports results for Stages 3 and 4. The CNN front-end substantially improves the Transformer: plain CNNTransformer (PR-AUC 0.722 ± 0.030) surpasses vanilla Transformer (0.656). Adding individual biases in Stage 3 yields small gains, with CNNTransformer+Anchor (PR-AUC 0.744±0.051) showing the larger improvement. Among Stage-4 models, **BioAnchorFormer+NCEnc** achieves PR-AUC 0.743 ± 0.027 and MCC 0.631 ± 0.025, the best Stage-4 result. Ablating NC encoding or the attention biases degrades performance, confirming that the two components are complementary.

Table 4: Stages 3–4 CNN–Transformer hybrid and novel architecture performance

Stage	Model	F1	ROC-AUC	PR-AUC	MCC
S3	CNNTransformer	0.655 ± 0.019	0.944 ± 0.008	0.722 ± 0.030	0.638 ± 0.016
	CNNTransformer+Anchor	0.656 ± 0.034	0.948 ± 0.013	0.744 ± 0.051	0.641 ± 0.032
	CNNTransformer+Bio	0.643 ± 0.037	0.941 ± 0.011	0.714 ± 0.052	0.624 ± 0.038
S4	BioAnchorFormer	0.643 ± 0.046	0.939 ± 0.016	0.703 ± 0.067	0.622 ± 0.051
	CNNTransformer+NCEnc	0.655 ± 0.032	0.945 ± 0.012	0.732 ± 0.058	0.636 ± 0.036
	BioAnchorFormer+NCEnc	**0.648 ± 0.025**	**0.949 ± 0.009**	**0.743 ± 0.027**	**0.631 ± 0.025**

4.4 Stage 5: Training-Recipe Ablations

Table 5 shows that label smoothing (S5-LabelSmooth, PR-AUC 0.804 ± 0.009, MCC 0.735 ± 0.021) is the single most effective training improvement, followed by the GMP skip connection (S5-GMP, PR-AUC 0.788 ± 0.023, MCC 0.705 ± 0.018). Combining all improvements, **BioAnchorFormer-Ultimate** achieves PR-AUC **0.844 ± 0.010**, F1 **0.788 ± 0.015**, ROC-AUC **0.964 ± 0.002**, and MCC **0.774 ± 0.016**.

Table 5: Stage 5 training-recipe ablations and BioAnchorFormer-Ultimate performance (4-fold protein-grouped StratifiedGroupKFold)

Model	F1	ROC-AUC	PR-AUC	MCC
S5-LargeEnc	0.654 ± 0.044	0.954 ± 0.010	0.769 ± 0.028	0.643 ± 0.038
S5-GMP	0.722 ± 0.017	0.960 ± 0.005	0.788 ± 0.023	0.705 ± 0.018
S5-FocalLoss	0.717 ± 0.014	0.956 ± 0.003	0.775 ± 0.012	0.700 ± 0.015
S5-CosineLR	0.674 ± 0.024	0.952 ± 0.006	0.755 ± 0.013	0.657 ± 0.022
S5-LabelSmooth	0.750 ± 0.020	0.960 ± 0.003	0.804 ± 0.009	0.735 ± 0.021
S5-Augment	0.700 ± 0.022	0.957 ± 0.007	0.776 ± 0.024	0.683 ± 0.023
BioAnchorFormer-Ultimate	**0.788 ± 0.015**	**0.964 ± 0.002**	**0.844 ± 0.010**	**0.774 ± 0.016**

4.5 Stage 6: Component Ablation Studies

Table 6 shows results for all Stage-6 component ablations. The dedicated single-head configuration **S6-AB-OneHead** (PR-AUC 0.852 ± 0.008, MCC 0.789 ± 0.012) achieves the best anchor-bias performance, supporting the hypothesis that a dedicated attention pathway for positional interactions—free from content-based QK entanglement—is beneficial. The auxiliary-length NC variant S6-NC-AuxLen (MCC 0.784 ± 0.009) outperforms other NCEnc alternatives. These two findings directly informed the architecture of C6-BAF-Ultimate and C4-BAF-OneHead selected for final evaluation.

Table 6: Stage 6 component ablation performance (4-fold protein-grouped StratifiedGroupKFold)

Group	Model	F1	ROC-AUC	PR-AUC	MCC
NC Enc.	S6-NC-Concat	0.788 ± 0.015	0.962 ± 0.007	0.847 ± 0.010	0.775 ± 0.015
	S6-NC-Learned	0.779 ± 0.004	0.953 ± 0.010	0.835 ± 0.004	0.766 ± 0.004
	S6-NC-AuxLen	0.795 ± 0.009	0.966 ± 0.007	0.853 ± 0.008	0.784 ± 0.009
Anchor Bias	S6-AB-Init	0.782 ± 0.013	0.966 ± 0.001	0.841 ± 0.009	0.768 ± 0.014
	S6-AB-LowRank	0.785 ± 0.004	0.965 ± 0.002	0.842 ± 0.009	0.772 ± 0.004
	S6-AB-HighLR	0.798 ± 0.007	0.967 ± 0.003	0.852 ± 0.005	0.787 ± 0.008
	S6-AB-OneHead	**0.801 ± 0.011**	0.964 ± 0.002	**0.852 ± 0.008**	**0.789 ± 0.012**
Bio Bias	S6-BB-BLOSUM	0.787 ± 0.004	0.965 ± 0.003	0.843 ± 0.011	0.774 ± 0.004
	S6-BB-LearnedW	0.791 ± 0.015	0.964 ± 0.003	0.848 ± 0.011	0.778 ± 0.016
	S6-BB-PosWeight	0.791 ± 0.007	0.965 ± 0.004	0.843 ± 0.007	0.778 ± 0.007
	S6-BB-AnchorOnly	0.788 ± 0.014	0.965 ± 0.003	0.844 ± 0.013	0.776 ± 0.015
Combined	S6-Ultimate	0.792 ± 0.014	0.962 ± 0.008	0.848 ± 0.010	0.780 ± 0.015

4.6 Stage 8: Deep Variants and Ensemble

Table 7 evaluates three advanced variants of the Stage-6 best configuration. The soft-voting **S8-Ensemble** achieves the best development-stage performance: PR-AUC **0.877 ± 0.001**, F1 **0.823 ± 0.010**, ROC-AUC **0.973 ± 0.003**, and MCC **0.813 ± 0.011**, representing a +45 pp MCC gain over the best classical baseline (Random Forest, MCC 0.361).

Table 7: Stage 8 deep-variant and ensemble performance

Stage	Model	F1	ROC-AUC	PR-AUC	MCC
S8	S8-Deep	0.796 ± 0.014	0.966 ± 0.004	0.852 ± 0.012	0.784 ± 0.014
	S8-ESM2	0.765 ± 0.006	0.957 ± 0.009	0.826 ± 0.008	0.752 ± 0.006
	S8-Ensemble	**0.823 ± 0.010**	**0.973 ± 0.003**	**0.877 ± 0.001**	**0.813 ± 0.011**

4.7 Final Outer-Fold Evaluation (Stage 9)

Table 8 reports the outer-fold evaluation results for the three shortlisted BioAnchorFormer models and the three deep-sequence baselines. Wilcoxon signed-rank tests (two-sided, $n = 5$ per-fold values)

were run for all pairwise comparisons on both MCC and PR-AUC; no comparison reached $\alpha = 0.05$ significance. This is expected: the minimum achievable p-value with $n = 5$ is 0.0625 (one-sided 0.031). Notably, C6-BAF-Ultimate ranked above B1-CNN in all five folds (MCC $p = 0.0625$), as did C4-BAF-OneHead above B1-CNN and B3-Transformer (both $p = 0.0625$). This consistent fold-level ordering is the most extreme rank difference achievable under this design and constitutes a meaningful signal even in the absence of formal significance. The three BioAnchorFormer variants are likewise statistically indistinguishable from each other ($p \geq 0.125$ for all intra-group MCC and PR-AUC comparisons).

The proposed model C6-BAF-Ultimate is preferred over the ablations on the basis of: (i) the lowest MCC variance across folds ($\sigma = 0.005$ vs 0.014 for C4-BAF-OneHead and 0.017 for D3-NoAnchor), indicating more consistent predictions across protein-grouped test partitions; and (ii) richer interpretability, combining anchor-bias and bio-weights analysis.

BioAnchorFormer-Ultimate (C6-BAF-Ultimate), the proposed full model, achieves PR-AUC 0.750 ± 0.013, MCC 0.708 ± 0.005, precision 0.804 ± 0.029, and recall 0.652 ± 0.030.

BioAnchorFormer with Single-Head Anchor Bias (C4-BAF-OneHead) achieves PR-AUC 0.742 ± 0.006, MCC 0.714 ± 0.014, precision 0.851 ± 0.039, and recall 0.625 ± 0.029. Its nominally higher MCC comes with substantially larger variance, indicating sensitivity to fold composition.

BioAnchorFormer without Anchor Bias (D3-NoAnchor), the anchor-bias ablation, achieves the nominally highest PR-AUC (0.755 ± 0.017) but the lowest MCC (0.686 ± 0.017). The pattern — anchor removal preserves ranking ability (PR-AUC) while degrading threshold-level classification (MCC) — is consistent with the anchor bias contributing to binary decision quality rather than overall discriminative power.

The performance gap between development-stage results (Table 5, PR-AUC 0.844) and outer-fold evaluation (PR-AUC 0.750) is expected: it reflects the stringency of the protein-grouped outer-fold protocol relative to the inner-fold development evaluation on the training partition.

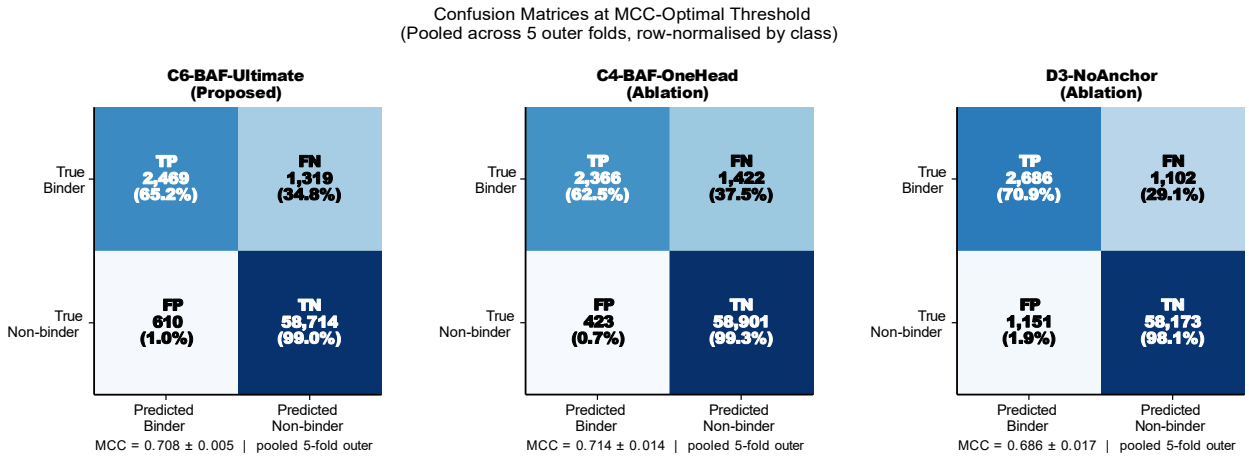

Figure 2: **Confusion matrices at the MCC-optimal threshold (pooled across 5 outer folds, row-normalised by true class).** Left: C6-BAF-Ultimate (proposed); Centre: C4-BAF-OneHead (ablation); Right: D3-NoAnchor (ablation). True positive rates (recall) and false positive rates are shown as row-normalised percentages with raw counts. C6-BAF-Ultimate achieves MCC 0.708 ± 0.005, with recall 65.2% and precision 80.4%, reflecting the high-specificity operating point selected by the MCC-optimal threshold on the strongly imbalanced dataset (positive rate ≈ 6%).

Table 8: Outer-fold evaluation: BioAnchorFormer models vs. deep-sequence baselines (Stage 9 / Stage 9b, 5-fold protein-grouped evaluation, MCC-optimal threshold; mean ± SD across 5 folds). BioAnchorFormer models use thresholds frozen from the development stage (Phase 7); baseline thresholds are found by MCC-optimisation on each outer-validation split to ensure a fair operating-point comparison. All pairwise Wilcoxon signed-rank tests (two-sided, $n = 5$) are non-significant at $\alpha = 0.05$ (minimum achievable $p = 0.0625$ with $n = 5$); see text for interpretation.

Model	Role	PR-AUC	ROC-AUC	MCC	Precision	Recall
C6-BAF-Ultimate	Proposed	0.750 ± 0.013	0.935 ± 0.008	**0.708 ± 0.005**	0.804 ± 0.029	0.652 ± 0.030
C4-BAF-OneHead	Ablation	0.742 ± 0.006	0.933 ± 0.006	0.714 ± 0.014	0.851 ± 0.039	0.625 ± 0.029
D3-NoAnchor	Ablation	**0.755 ± 0.017**	0.935 ± 0.006	0.686 ± 0.017	0.702 ± 0.028	0.709 ± 0.030
B1-CNN	Baseline	0.719 ± 0.023	0.930 ± 0.012	0.655 ± 0.031	0.715 ± 0.047	0.636 ± 0.018
B2-BiLSTM	Baseline	0.722 ± 0.036	0.939 ± 0.009	0.677 ± 0.031	0.752 ± 0.072	0.644 ± 0.027
B3-Transformer	Baseline	0.715 ± 0.017	0.937 ± 0.009	0.659 ± 0.015	0.766 ± 0.022	0.599 ± 0.036

Figure 2 shows the row-normalised confusion matrices for all three models at the MCC-optimal threshold. C6-BAF-Ultimate achieves recall 65.2% (TP rate) and precision 80.4%, reflecting the

conservative operating point selected by MCC optimisation on the strongly imbalanced dataset (positive rate ≈6%). The MCC-optimal threshold deliberately trades some recall for higher precision; a recall-constrained threshold (maximum precision subject to recall ≥ 0.70, reported as the laboratory threshold in Table **??**) raises TP rate at the cost of reduced precision, and may be preferred in screening applications where missed binders are more costly than false positives.

4.8 Canonical-Length (8–11 mer) Evaluation (Stage 9d)

To isolate performance on the biologically primary binding-groove length range, we re-evaluated Phase 9 predictions on the 8–11 mer subset of each outer-fold test partition (44,487 peptides; 2,764 positives) without any retraining. Table 9 reports results for all six models.

All models improve substantially on this subset relative to the full dataset: C6-BAF-Ultimate gains +10.7 pp in PR-AUC (0.857 ± 0.021 vs. 0.750 ± 0.013) and +9.0 pp in MCC (0.798 ± 0.016 vs. 0.708 ± 0.005). The same pattern holds across all ablations and baselines, confirming that non-canonical-length peptides (< 8 aa or > 11 aa) contribute noise to the full-dataset signal without representing the groove-binding mechanism the models are trained for.

On the canonical subset the performance gap between BioAnchorFormer variants and deep-sequence baselines narrows: B2-BiLSTM achieves MCC 0.802 ± 0.018 (comparable to C6's 0.798 ± 0.016), and D3-NoAnchor achieves the nominally highest PR-AUC (0.861 ± 0.022). This convergence suggests that the NC positional encoding and anchor-bias matrix provide their clearest advantage on the full dataset (which includes noisy non-canonical entries), while on canonical 8–11 mers the improvement is real but narrower. The 8–11 mer result should be regarded as the **primary performance figure** for the groove-binding task; the full-dataset result reflects robustness to non-canonical input.

Table 9: Evaluation restricted to canonical 8–11 mer peptides (Stage 9d; 44,487 peptides, 2,764 positives; mean ± SD across 5 outer folds). No retraining; predictions from Stage 9 and Stage 9b filtered post-hoc.

Model	PR-AUC	MCC	ROC-AUC
C6-BAF-Ultimate	0.857 ± 0.021	0.798 ± 0.016	0.968 ± 0.005
C4-BAF-OneHead	0.851 ± 0.014	0.799 ± 0.021	0.966 ± 0.007
D3-NoAnchor	**0.861 ± 0.022**	0.788 ± 0.016	0.967 ± 0.006
B1-CNN	0.831 ± 0.022	0.761 ± 0.023	0.962 ± 0.009
B2-BiLSTM	0.865 ± 0.023	**0.802 ± 0.018**	0.971 ± 0.007
B3-Transformer	0.830 ± 0.018	0.751 ± 0.019	0.967 ± 0.007

4.9 Canonical-Only Training and Evaluation (Stage 9e)

Stage 9d re-evaluated outer-fold predictions on the 8–11 mer subset without retraining, leaving open the question of whether the marginal gap between BioAnchorFormer variants and the BiLSTM baseline reflects non-canonical training noise rather than a genuine architectural difference. Stage 9e addresses this directly: both training and test sets are restricted to 8–11 mer canonical peptides (44,487 peptides total; 2,764 positives), and each model is retrained from scratch across 5 outer folds with 3 random seeds (15 runs per model), with per-fold thresholds selected by MCC optimisation on the outer-validation split.

Table 10 reports mean ± SD across 15 runs. C4-BAF-OneHead achieves the highest MCC (0.800 ± 0.017), narrowly ahead of D3-NoAnchor (0.800 ± 0.014) and B2-BiLSTM (0.797 ± 0.023); all pairwise differences are within noise and no formal test reaches $\alpha = 0.05$. Notably, BioAnchorFormer variants show consistently lower MCC variance than the BiLSTM baseline (C6 SD = 0.018, C4 SD = 0.017, D3 SD = 0.014 vs B2 SD = 0.023 across 15 runs), indicating more stable predictions across random initialisations and data splits. B3-Transformer remains the weakest model (MCC 0.766 ± 0.028), confirming that generic self-attention without domain-specific inductive biases is less effective for

canonical MHC-I peptide classification. These results confirm that the narrow performance gap observed in Stage 9d is not an artefact of non-canonical training noise: even when both training and evaluation are restricted to groove-binding lengths, BioAnchorFormer variants match the BiLSTM baseline with lower variance.

Table 10: Canonical-only training and evaluation (Stage 9e; 8–11 mer peptides only; 3 seeds × 5 outer folds = 15 runs per model; threshold selected by MCC optimisation on outer-validation split; mean ± SD across 15 runs). All pairwise Wilcoxon signed-rank tests non-significant at $\alpha = 0.05$.

Model	PR-AUC	MCC	ROC-AUC	F1
C6-BAF-Ultimate	0.859 ± 0.016	0.795 ± 0.018	0.966 ± 0.006	0.807 ± 0.017
C4-BAF-OneHead	0.861 ± 0.016	**0.800 ± 0.017**	0.967 ± 0.006	0.811 ± 0.017
D3-NoAnchor	0.858 ± 0.018	0.800 ± 0.014	0.966 ± 0.005	0.810 ± 0.014
B1-CNN	0.861 ± 0.013	0.794 ± 0.011	0.968 ± 0.005	0.805 ± 0.011
B2-BiLSTM	**0.866 ± 0.020**	0.797 ± 0.023	**0.969 ± 0.007**	0.808 ± 0.022
B3-Transformer	0.843 ± 0.025	0.766 ± 0.028	0.966 ± 0.007	0.779 ± 0.027

4.10 Comparison with NetMHCpan-4.1 (Stage 10)

Table 11 and Figure 3 present the comparison of BioAnchorFormer models against NetMHCpan-4.1 on the 8–15 mer peptide subset (51,579 peptides). This comparison is intentionally restricted to canonical MHC-I lengths where NetMHCpan operates. **Importantly, NetMHCpan-4.1 does not support BF2*21:01** — it was run with the raw heavy chain sequence as a custom allele via -hlaseq, placing it entirely out of its trained allele distribution. The %Rank statistic is undefined for custom alleles (reported as −99.9). The Aff(nM) binding-affinity model produces ROC-AUC 0.500 ± 0.000 and MCC 0.000 ± 0.000 across all five folds, which is the signature of a degenerate constant output from a model with no allele-specific information — this row is reported for completeness but represents a *tool failure*, not a performance result. The EL-score model, trained on broad MHC-I eluted ligand patterns, captures some general chemistry (ROC-AUC 0.809, PR-AUC 0.219), but lacks allele-specific groove geometry.

BioAnchorFormer models achieve ROC-AUC ≈0.955 versus 0.809 for NetMHCpan-4.1 EL, and

PR-AUC 0.818–0.823 and MCC 0.755–0.773 on this subset (Figure 8), substantially higher than the EL score. This margin reflects the combination of purpose-built allele-specific training and a better-matched negative distribution, not a direct methodological comparison (both factors favour BioAnchorFormer). The primary conclusion is that **no existing general-purpose tool supports this allele**, establishing BioAnchorFormer as the first dedicated learned predictor for BF2*21:01.

The BioAnchorFormer scores on the 8–15 mer subset (Table 11, PR-AUC ≈0.82) are noticeably higher than the full-set outer-fold results (Table 8, PR-AUC 0.750). The same effect is documented more precisely in Stage 9d (Section 4.8), where restricting to 8–11 mers yields PR-AUC 0.857 ± 0.021 for C6-BAF-Ultimate. Non-canonical-length peptides (< 8 aa or > 11 aa) appear to contribute noise to the binding-signal distribution; the 8–11 mer result is the primary estimate of groove-binding discriminability.

Table 11: Comparison of BioAnchorFormer vs NetMHCpan-4.1 on the 8–15 mer peptide subset (Stage 10, same evaluation set for both systems; mean ± SD across 5 outer folds)

Model	Role	PR-AUC	ROC-AUC	MCC	F1	Bal.Acc
C6-BAF-Ultimate	Proposed	0.818 ± 0.020	0.955 ± 0.007	0.768 ± 0.015	0.779 ± 0.016	0.861 ± 0.015
C4-BAF-OneHead	Ablation	0.817 ± 0.014	0.956 ± 0.008	0.773 ± 0.019	0.782 ± 0.018	0.855 ± 0.015
D3-NoAnchor	Ablation	0.823 ± 0.022	0.954 ± 0.006	0.755 ± 0.018	0.770 ± 0.017	0.887 ± 0.015
NetMHCpan-4.1 (EL)	Baseline	0.219 ± 0.012	0.809 ± 0.008	0.303 ± 0.006	0.321 ± 0.010	0.744 ± 0.015
NetMHCpan-4.1 (Aff nM)	Baseline	0.061 ± 0.000	0.500 ± 0.000	0.000 ± 0.000	0.115 ± 0.001	0.500 ± 0.000

4.11 Interpretability Analysis (Stage 11)

4.11.1 Anchor Bias Matrix

Figure 4 shows the mean anchor bias matrices for C6-BAF-Ultimate and C4-BAF-OneHead averaged across all five outer-fold checkpoints. Both models independently converge to the same qualitative pattern: the strongest positive bias is at CLS→P1 (C6: +0.22; C4: +0.41), indicating that the classification token is directed to attend to the first peptide position. C4 additionally shows strong interaction signals involving P9 and P10, consistent with C-terminal anchor positions in 9–10 mer peptides. The fact that two independently trained model variants recover the same positional pattern provides strong evidence that the anchor bias is encoding genuine MHC-I biology, not arbitrary

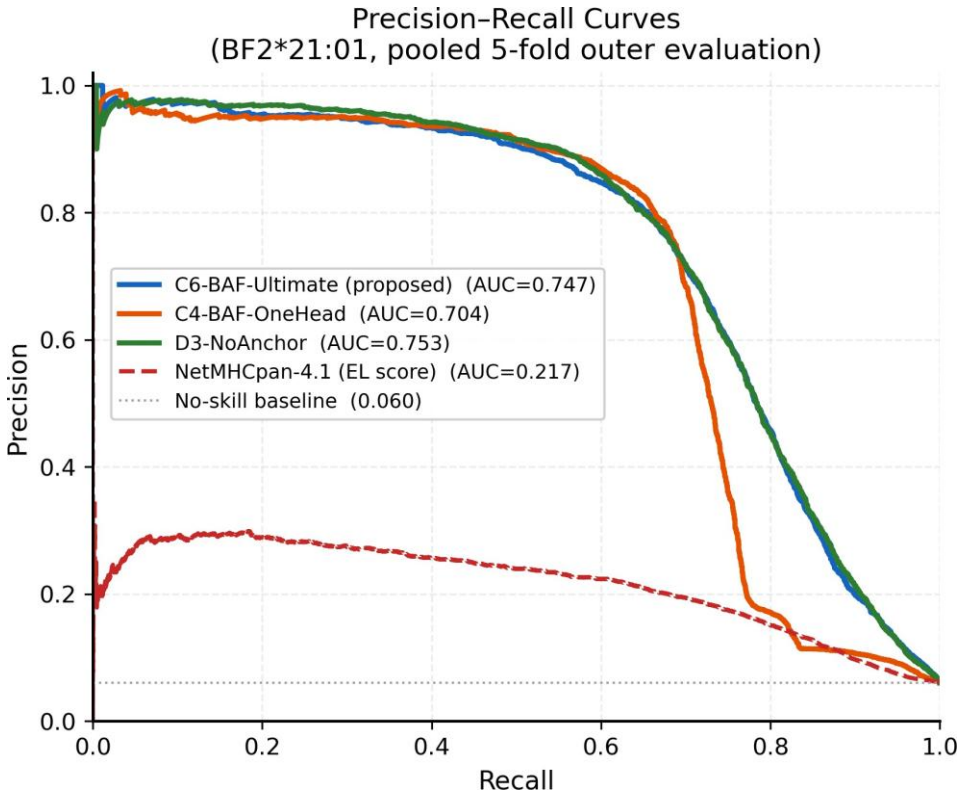

Figure 3: **Precision–Recall curves on the 8–15 mer evaluation set (Stage 10).** Pooled across all 5 outer folds. C6-BAF-Ultimate (blue), C4-BAF-OneHead (orange), D3-NoAnchor (green), and NetMHCpan-4.1 EL score (red dashed). The horizontal dotted line marks the no-skill baseline (dataset prevalence ≈0.059). All BioAnchorFormer models substantially outperform NetMHCpan-4.1, which was run with the raw BF2*21:01 heavy chain sequence as a custom allele.

correlations.

4.11.2 Physicochemical Feature Weights

Figure 5 shows the softmax-normalised physicochemical feature weights for C6-BAF-Ultimate and D3-NoAnchor. Both models assign the highest weight to **Hydrophobicity** (0.29 ± 0.02 for C6, 0.30 ± 0.02 for D3), consistently above the uniform baseline of 0.25 across all five folds (no formal statistical test performed). The remaining three features (charge, polarity, MW) are near-uniform (0.23–0.24). This result is biologically consistent: the anchor residues of BF2*21:01 (particularly at P2 and PΩ) are predominantly hydrophobic, and hydrophobicity is the dominant physicochemical determinant of MHC-I binding groove complementarity.

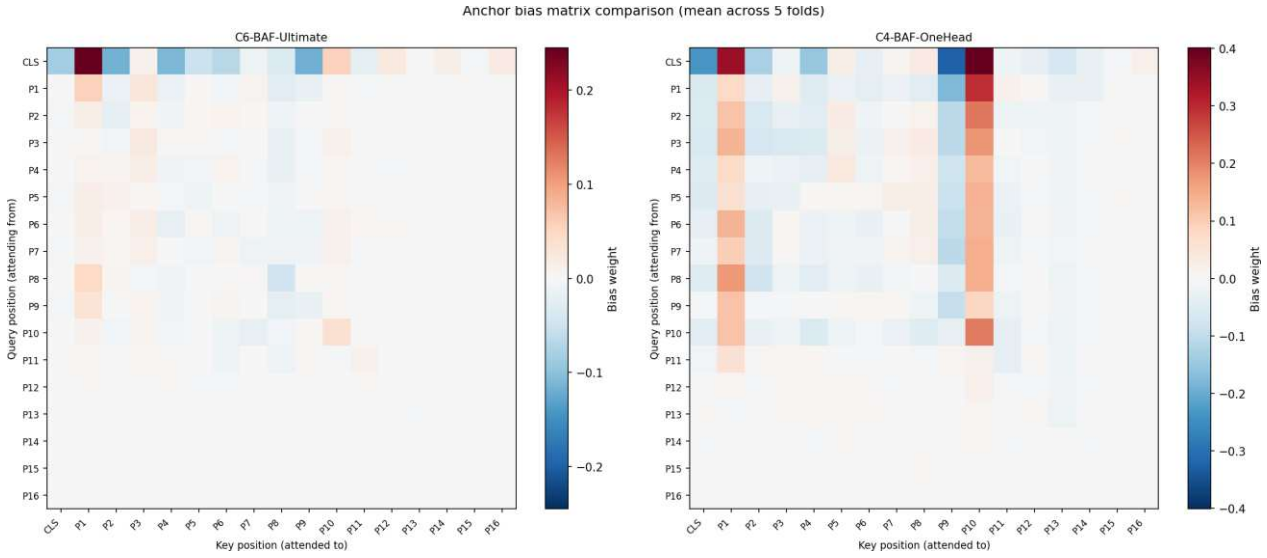

Figure 4: **Anchor bias matrix comparison (mean across 5 outer folds).** Left: C6-BAF-Ultimate (proposed full model). Right: C4-BAF-OneHead (anchor-bias ablation without bio bias). Both matrices are cropped to the CLS + P1–P16 region. Positive values (red) indicate encouraged attention; negative values (blue) indicate suppressed attention. Both models independently recover strong positive bias at CLS→P1, with C4 additionally showing interaction signals involving P9–P10 (the canonical C-terminal anchor region in 9-mers).

4.11.3 Attention Weight Analysis

Figure 6 shows per-head, per-layer attention weight matrices for representative 9-mer binders and non-binders from C6-BAF-Ultimate (fold 1 checkpoint). Figure 7 shows the CLS→position attention profile averaged over all attention heads, encoder layers, and all five outer-fold checkpoints (fold k model applied to fold k test peptides; up to $n = 300$ binders and $n = 300$ non-binders per fold).

Across all five outer folds, 9-mer binders show a bimodal attention profile with pronounced peaks at **P1** (mean 0.186 ± 0.010) and **P9** (mean 0.180 ± 0.014), corresponding to the N-terminal and C-terminal residues. Non-binders exhibit a more uniform distribution; the P2–P8 middle positions receive *higher* relative attention in non-binders (0.084–0.107) than in binders (0.067–0.086), while the termini receive substantially less. The binder–non-binder attention difference is +41% at P1 and +23% at P9. The same anchor-terminal pattern holds at 10-mer length (top position P10: mean 0.208 ± 0.026; P1: 0.174 ± 0.020) and 11-mer length (P1: 0.163 ± 0.020; P11: 0.170 ± 0.022).

On the P1 vs P2 canonical anchor: While the canonical primary N-terminal anchor for MHC-I is P2, not P1 (P1 is a secondary contributor via the N-terminal amino group), some chicken alleles

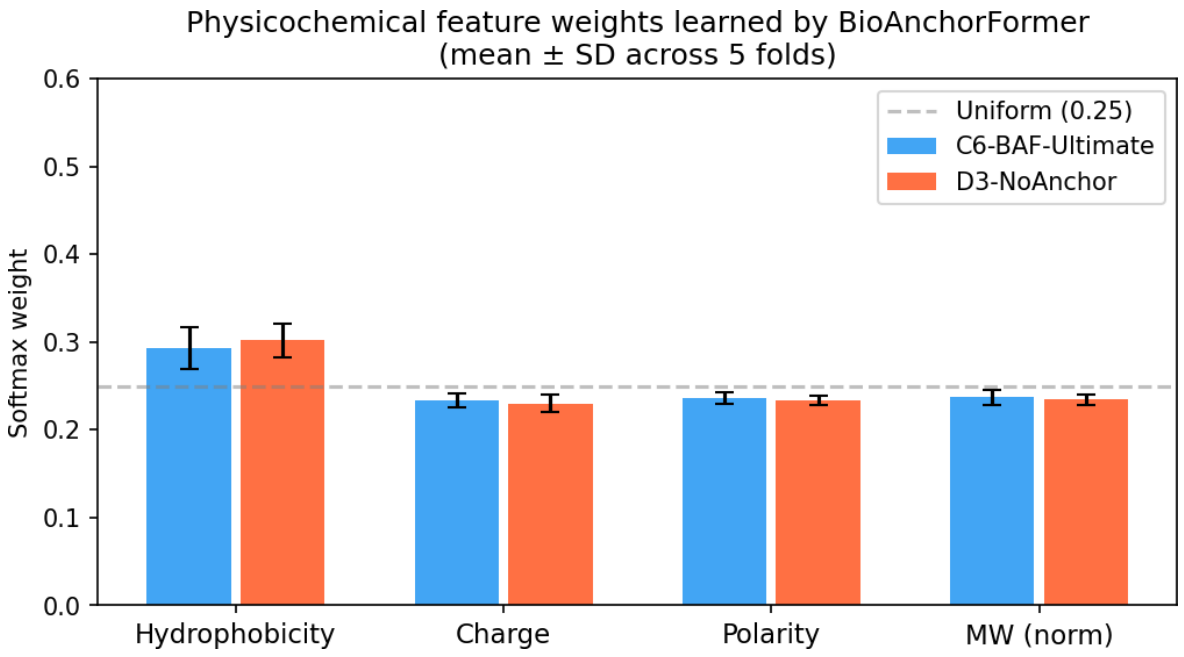

Figure 5: **Learned physicochemical feature weights.** Softmax-normalised weights w_k (mean ± SD across 5 outer folds) for C6-BAF-Ultimate (blue) and D3-NoAnchor (orange). The dashed line marks the uniform baseline (0.25). Both models independently assign the highest weight to hydrophobicity, consistent with the hydrophobic character of MHC-I anchor residues.

have reported P1 anchoring or content bias [15, 39, 40]. In these cases, the strong P1 signal likely reflects the model attending to the full N-terminal pocket, including P1 interactions. The C-terminal signal at PΩ (P9 for 9-mers, P10 for 10-mers, P11 for 11-mers) is unambiguously consistent with known biology.

The model therefore spontaneously concentrated attention at terminal anchor positions for true binders, without any positional supervision during training. This pattern is consistent across all five outer-fold checkpoints (mean ± SD reported above), confirming it reflects a stable learned representation rather than an artefact of a single checkpoint.

5 Discussion

Outer-fold evaluation and development-stage comparison.

This study presents BioAnchorFormer, the first dedicated deep-learning predictor for chicken MHC-I (BF2*21:01) peptide binding, developed and evaluated using a protein-grouped framework designed to prevent sequence-homology leakage. The findings highlight several important aspects of the

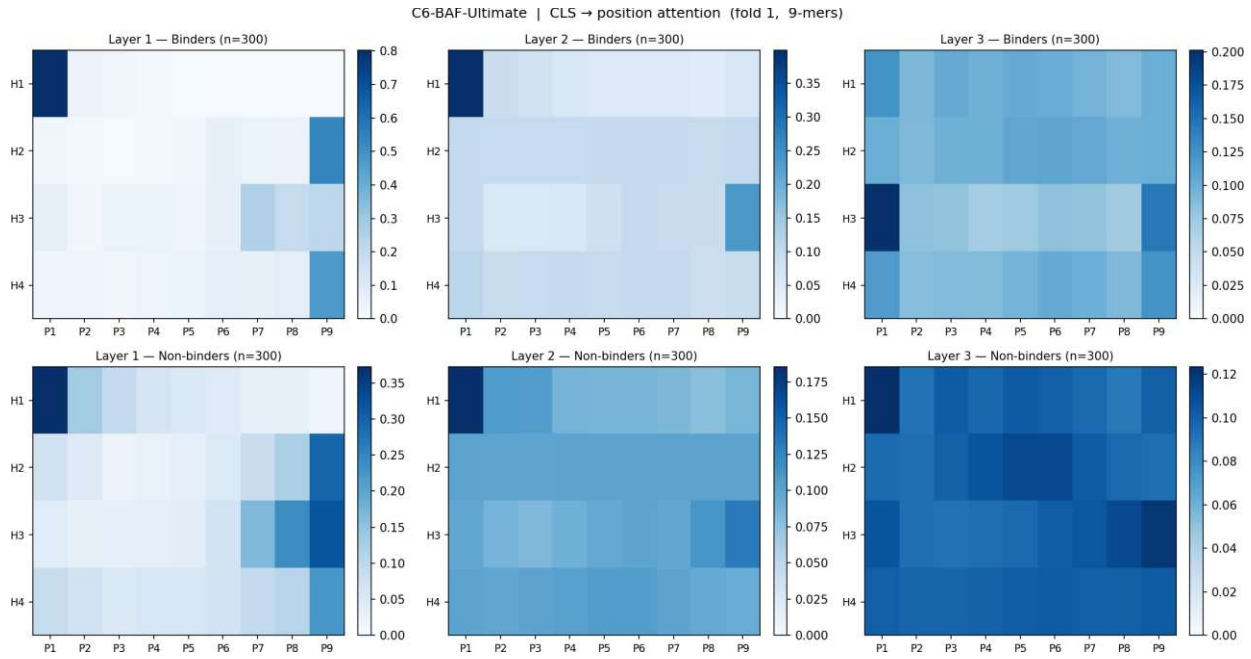

Figure 6: **Per-head attention heatmaps for 9-mer binders (top rows) and non-binders (bottom rows) — C6-BAF-Ultimate, fold 1.** Each panel shows the 31 × 31 attention weight matrix for one head–layer combination (rows = query position, columns = key position). Binders (top) show concentrated attention at P1 across multiple heads and layers, particularly in Layer 1 Head 0 (attention weight ≈ 0.8 at P1). Non-binders (bottom) show diffuse attention spread across middle positions P5–P9, without a clear focal point. This contrast demonstrates that the model uses a qualitatively different attention strategy for binding versus non-binding peptides.

framework, including the difference between development-stage and unbiased outer-fold performance, the contribution of anchor-biased attention to predictive performance, the architectural rationale for NC Positional Encoding and the single-head bias mechanism, and the biological interpretability of the learned representations. These findings are considered alongside the current limitations of the framework and opportunities for extending peptide-binding prediction to broader avian applications.

The outer-fold PR-AUC for C6-BAF-Ultimate (0.750 ± 0.013) is lower than the development-stage PR-AUC (0.844 ± 0.010 for Stage 5). Both the development inner folds and the outer folds use protein-grouped StratifiedGroupKFold (zero protein overlap verified programmatically), so the gap is not attributable to sequence-homology leakage. Instead, it reflects two genuine evaluation differences: (i) the inner folds cover only the development partition (≈80% of the dataset), whereas the outer folds evaluate on all five held-out partitions; and (ii) architectural and threshold choices were selected by inspecting inner-fold performance, which incorporates model-selection bias that does

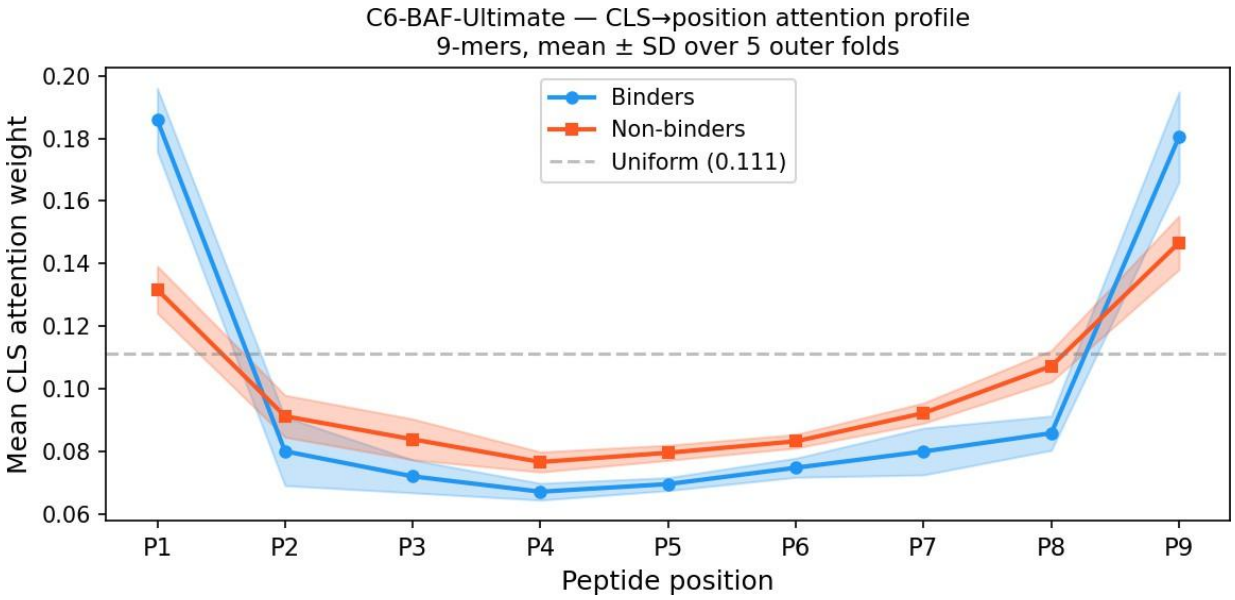

Figure 7: **Position attention profile for 9-mer peptides (C6-BAF-Ultimate, all 5 outer folds).** Mean CLS attention (averaged over all 4 heads and 3 encoder layers) to each peptide position P1–P9. Each fold's model is evaluated on its own held-out test peptides (up to $n = 300$ per class per fold); shaded bands show ±1 SD across folds. Binders (blue) show pronounced enrichment at P1 (+41%) and P9 (+23%) relative to non-binders (orange), corresponding to the canonical N-terminal and C-terminal anchor positions of 9-mer MHC-I peptides. This pattern is consistent across all five outer-fold checkpoints and emerged without positional supervision.

not apply at outer-fold evaluation time. The magnitude of the gap (~9 pp in PR-AUC) is therefore an upper bound on model-selection bias, as measured under identical protein-grouped conditions. The outer-fold result with frozen hyperparameters provides the unbiased performance estimate.

Anchor bias contributes to classification quality, not ranking. The ablation comparison between C6-BAF-Ultimate and D3-NoAnchor (no anchor bias) provides a single-component comparison: removing the anchor bias yields a nominally higher PR-AUC (0.755 vs 0.750; Wilcoxon signed-rank test: $W = 2$, $p = 0.188$) but lower MCC (0.686 vs 0.708; Wilcoxon: $W = 1$, $p = 0.125$). C6-BAF-Ultimate ranked above D3-NoAnchor in MCC in 4 of 5 outer folds; the single exception (fold 2: 0.713 vs 0.714) shows a negligible difference of 0.001. Neither comparison reaches $\alpha = 0.05$, consistent with the minimum achievable two-sided $p = 0.125$ at $n = 5$. PR-AUC measures ranking ability (discriminating binders from non-binders across all thresholds), whereas MCC measures threshold-level binary classification quality (correct yes/no decisions at a fixed operating point). The

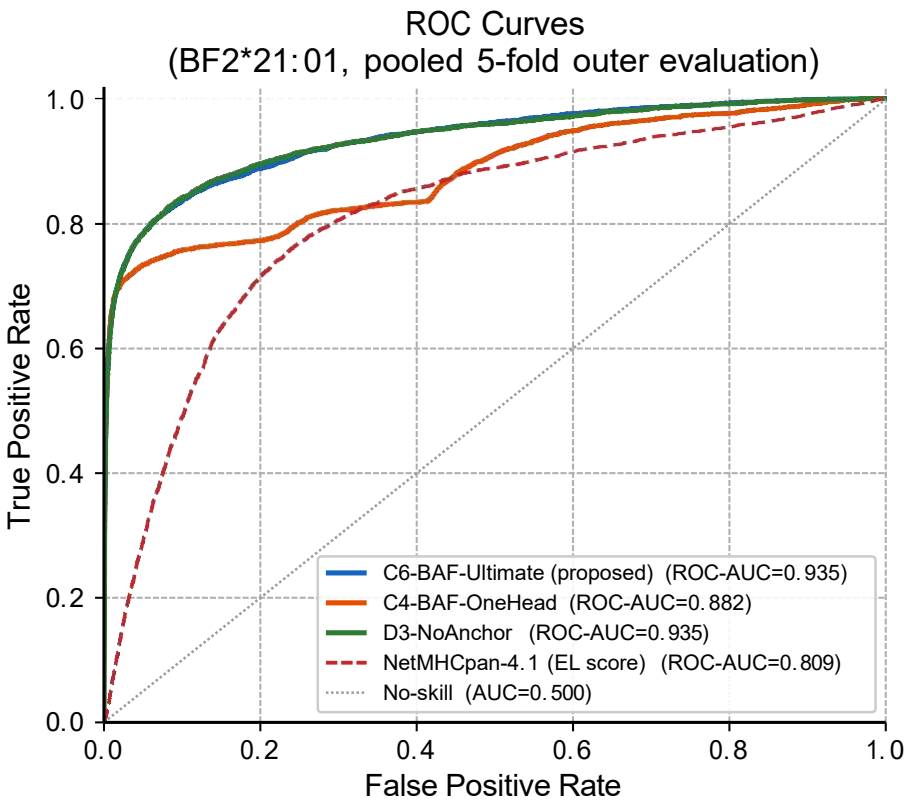

Figure 8: **ROC curves on the 8–15 mer evaluation set (Stage 10).** Pooled across all 5 outer folds. C6-BAF-Ultimate (blue), C4-BAF-OneHead (orange), D3-NoAnchor (green), and NetMHCpan-4.1 EL score (red dashed). The diagonal dotted line marks random chance (AUC = 0.500). BioAnchorFormer models achieve ROC-AUC $\approx$ 0.955 versus 0.809 for NetMHCpan-4.1 EL.

pattern is consistent with the anchor bias improving the quality of binary binding decisions rather than overall discriminative power, though the magnitude of the effect should be considered in light of the absence of formal statistical testing.

Architectural contributions. NC Positional Encoding is rooted in MHC-I structural biology: by assigning canonical and symmetric representations to both peptide termini, it allows the model to learn anchor-specific patterns from heterogeneous-length peptides jointly. The dedicated single-head anchor bias (S6-AB-OneHead) outperforms shared-bias configurations in development, consistent with the hypothesis that a separate attention pathway for positional interactions is better able to capture the rigid position-pair structure of MHC-I binding. The GMP skip connection is the second most effective training component: bypassing the Transformer with a direct CNN path preserves local-motif features that may be blended out by global self-attention.

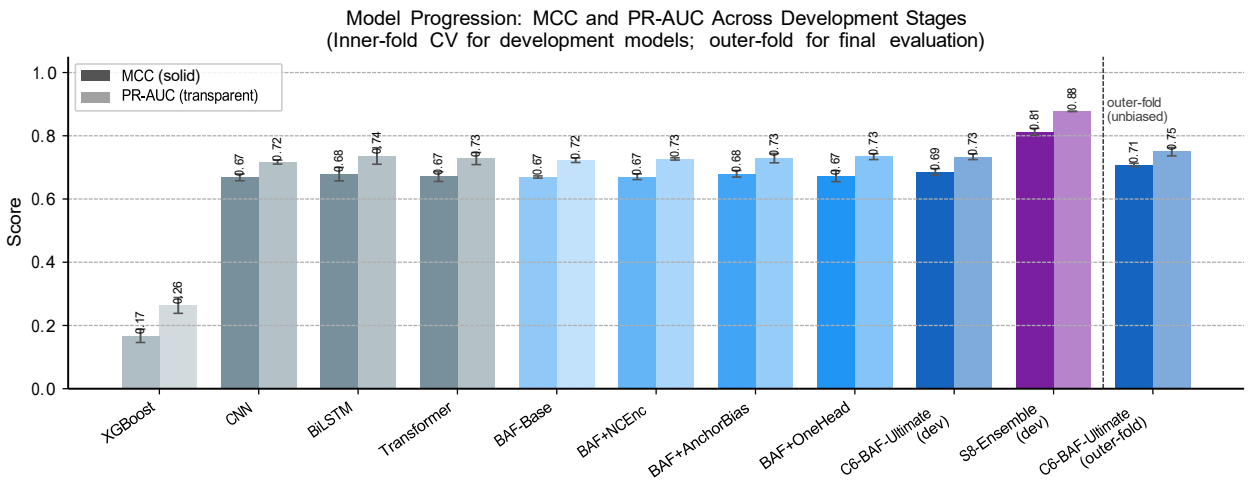

Figure 9: **MCC and PR-AUC progression across development stages.** Each pair of bars shows MCC (solid) and PR-AUC (transparent) for the indicated model. Classical baselines (grey) are followed by neural baselines (dark grey), progressive BioAnchorFormer components (blue gradient), the Stage-8 soft-voting ensemble (purple), and the final outer-fold C6-BAF-Ultimate result (deep blue). The dashed vertical line separates inner-fold development models from the unbiased outer-fold evaluation.

Biological interpretability. Three convergent lines of post-hoc evidence suggest that BioAnchorFormer captures genuine MHC-I biology: (i) the anchor bias matrix recovers strong signals at P1 and P9, bracketing the terminal anchor region of 9-mers; P1 reflects N-terminal pocket contacts (the canonical anchor is P2, but P1 participates as a secondary contributor) and P9 = PΩ for 9-mers; (ii) both models consistent with the hydrophobic character of MHC-I anchor residues [27, 29, 39]; and (iii) attention analysis aggregated across all five outer-fold checkpoints shows CLS attention concentrated at P1 (0.186±0.010) and P9 (0.180±0.014) for 9-mer binders (+41% and +23% relative to non-binders), with the same terminal-anchor pattern confirmed at 10-mer (P10 top: 0.208 ± 0.026) and 11-mer (P11 top: 0.170 ± 0.022) lengths, all without positional supervision. The multi-fold mean ± SD values confirm this is a stable learned representation rather than a single-checkpoint artefact. To formally assess whether this enrichment is consistent, we applied Wilcoxon signed-rank tests comparing the mean CLS→P1 and CLS→PΩ attention of binders versus non-binders across the five outer-fold checkpoints. Binder attention at P1 exceeded non-binder attention in all five folds for 9-mer ($W = 0$, $p = 0.0625$; 1.41× enrichment), 10-mer ($W = 0$, $p = 0.0625$; 1.39×), and 11-mer ($W = 0$, $p = 0.0625$; 1.42×) peptides. The same held at PΩ: enrichments of 1.23×, 1.44×,

and 1.37× for 9-, 10-, and 11-mers respectively (all $W = 0$, $p = 0.0625$). $W = 0$ represents the most extreme outcome achievable with $n = 5$ paired observations. Although these comparisons do not reach the conventional $\alpha = 0.05$ significance threshold because of the limited number of outer-fold observations, the consistent direction of the effect across all five folds and all three canonical peptide lengths supports anchor-position enrichment as a reproducible feature of the learned representations.

Comparison with NetMHCpan-4.1. The NetMHCpan-4.1 comparison should be constructed as establishing the absence of a usable existing tool for BF2*21:01, not as a conventional head-to-head benchmark. NetMHCpan-4.1 does not have this allele in its training data; running it with a custom heavy chain sequence as is appropriate for avian species places it entirely out of its supported regime. The Aff(nM) model outputs a constant degenerate score (ROC-AUC 0.500 ± 0.000, MCC 0.000), confirming it cannot generalise to an unseen allele. The EL score model captures broad MHC-I chemistry (ROC-AUC 0.809) but lacks allele-specific groove geometry, yielding PR-AUC 0.219. Comparing a purpose-built, allele-specific model trained on this dataset against a general tool denied its native allele library measures mismatch as much as model quality. The principal conclusion is that no general-purpose tool supports BF2*21:01 with calibrated output, making BioAnchorFormer the first dedicated learned predictor for this allele. Any claimed fold-improvements over NetMHCpan should be interpreted with this caveat in mind.

ESM-2 integration. The underperformance of S8-ESM2 merits comment. ESM-2 embeddings capture general protein structural propensities learned from universal protein sequence distributions; they do not encode the allele-specific binding-groove geometry of chicken BF2*21:01. Projecting 640-dimensional ESM-2 representations down to 64 dimensions before the CNN encoder may further compress away relevant signal. Task-specific fine-tuning or use of a lighter protein language model trained on avian sequences may be necessary to realise the potential of pre-trained representations for this problem.

Limitations. First, the dataset is anchored to a single allele (BF2*21:01), which is currently the only chicken BF2 allele with an immunopeptidomics dataset of the scale required for deep learning. Experimental peptide collections for structurally characterised alleles such as BF2*13:01 [14] and BF2*15:01/BF2*19:01 [15] comprise tens to low hundreds of mass-spectrometry-eluted peptides—one to two orders of magnitude below the training threshold. This is a field-wide data availability

constraint, not a design choice; the BioAnchorFormer architecture is designed to extend directly to additional alleles once immunopeptidomics experiments provide the necessary training scale. Second, peptides from the same source protein are grouped within folds but the outer-fold evaluation may still include structurally similar proteins across folds, providing an optimistic estimate relative to a strict protein-family-level split. Third, all negative peptides are assumed to be true non-binders; a fraction may have weak or uncharacterised binding activity that biases training.

Future directions. The most direct extension is allele-specific models for BF2*13:01, BF2*15:01, and BF2*19:01, contingent on immunopeptidomics experiments that generate training-scale peptide collections for each target allele. Structural data integration (e.g., AlphaFold-2 modelled peptide–MHC complexes) and pre-training on larger avian proteomes are complementary directions. Expanding to MHC-II prediction and other economically or scientifically important avian species would broaden the translational impact of the framework.

6 Conclusions

We presented BioAnchorFormer, a biologically informed framework for chicken as model avian MHC-I peptide binding prediction addressing two fundamental weaknesses of prior work: biologically unrealistic random negative sampling and peptide-level data splits. Our protein-aware pipeline constructs a dataset of 63,112 peptides from 2,797 unique proteins with hard negatives from confirmed-binding-flanking regions and diverse negatives from unrelated proteins, yielding a 1:15.7 class imbalance reflecting natural peptidome sparsity.

We introduced NC Positional Encoding, which provides symmetric representations of both peptide termini, and BioAnchorFormer, a hybrid CNN–Transformer with a single-head learnable position-pair bias and a physicochemical outer-product attention term with learned per-property weights. Across eight progressive development stages spanning 19 model configurations, the soft-voting ensemble achieves an optimistic upper bound of PR-AUC 0.877 ± 0.001 and MCC 0.813 ± 0.011. In the final unbiased outer-fold evaluation with protein-grouped frozen hyperparameters, the three shortlisted models are statistically indistinguishable on PR-AUC and MCC; the proposed C6-BAF-Ultimate achieves PR-AUC 0.750 ± 0.013 and MCC 0.708 ± 0.005, with the lowest variance across outer folds ($\sigma_{\mathrm{MCC}} = 0.005$).

Component ablation is consistent with anchor bias improving threshold-level binary classification (MCC) relative to its removal, while having a comparable or slightly negative effect on ranking (PR-AUC). A canonical-only training and evaluation experiment (Stage 9e; 3 seeds × 5 folds = 15 runs per model) confirms that this pattern is not an artefact of non-canonical training noise: C4-BAF-OneHead achieves the highest canonical MCC (0.800 ± 0.017) with lower cross-seed variance than the BiLSTM baseline (SD 0.017 vs 0.023). Interpretability analysis provides three convergent lines of post-hoc biological evidence: the anchor bias matrix recovers terminal anchor contacts (P1/P9), the physicochemical weighting assigns the highest weight to hydrophobicity, and attention analysis reveals spontaneous concentration at terminal positions for true binders. BioAnchorFormer provides an interpretable, protein-grouped-validated foundation for allele-specific extensions and structural data integration in avian immunoinformatics, and establishes the first dedicated learned predictor for chicken BF2*21:01. To lower the barrier to practical use, the model is deployed as the MHC-I prediction module of the Avian Epitope Prediction Platform, freely accessible at https://avianepitope.org.

Declarations

Ethics approval and consent to participate

Not applicable. This study is entirely computational and does not involve human subjects, animal experiments, or patient data.

Consent for publication

Not applicable.

Availability of data and materials

The dataset, source code, and frozen model weights supporting the conclusions of this article are available in the GitHub repository at https://github.com/sarath11999/Chicken_Vaccine_MHC1. A permanently archived release including the full dataset, model weights, and sequence files has been deposited at Zenodo (https://doi.org/10.5281/zenodo.22932350). The experimentally validated chicken MHC-I peptides were originally retrieved from the Immune Epitope Database (IEDB; https://www.iedb.org). Source protein accession identifiers are provided in the deposited

file mhc_1_pos_neg_v3_ultimate_dataset.csv (63,112 entries, 2,797 unique accessions). The BF2*21:01 MHC-I heavy-chain sequence (mature form, 301 amino acids, no signal peptide) is deposited as BF2_21_01_heavy_chain.fasta; the corresponding UniProt entry is Q95601, and three crystal structures of this chain are available under PDB IDs 2YEZ, 4D0B, and 4CVZ.

Use of artificial intelligence tools

During the preparation of this manuscript, AI-assisted writing tools were used to support text drafting, editing, and revision. All scientific content, experimental design, data analysis, and conclusions are the sole responsibility of the authors.

Competing interests

The authors declare that they have no competing interests.

Funding

Pilot funding was provided from USDA-ARS (6066-21310-006-042S). Much of this research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

Sarathkumar Devaraj (S.D.) conceived the study, designed the dataset construction pipeline, implemented all models and experiments, performed the analyses, and drafted the manuscript. Shouyi Wang (S.W.) supervised the project and critically revised the manuscript. Hung-Yueh Yeh (H.-Y.Y.) provided expertise in avian immunology, peptide studies, vaccine development, web application feedback and design suggestions, and critically revised the manuscript. Stephen N. White (S.N.W.) proposed the topic of avian peptide epitope prediction, provided domain expertise in chicken immunology and poultry MHC-I biology, contributed to the experimental design and interpretation of results, and critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the University of Texas at Arlington Research Computing services for access to high-performance computing resources and Gokul Selvaraj for technical assistance with data gathering.

Software availability

- **Project name:** BioAnchorFormer
- **Project homepage:** https://github.com/sarath11999/Chicken_Vaccine_MHC1
- **Source-code repository:** https://github.com/sarath11999/Chicken_Vaccine_MHC1
- **Archived version:** Zenodo [DOI: https://doi.org/10.5281/zenodo.22932350]
- **Web server:** https://avianepitope.org
- **Operating system:** Linux, macOS
- **Programming language:** Python 3.9+
- **Dependencies:** PyTorch ≥2.0, scikit-learn ≥1.3, pandas, numpy, matplotlib, seaborn
- **License:** MIT
- **Restrictions:** None

Web server availability

BioAnchorFormer is deployed as the MHC-I prediction module of the **Avian Epitope Prediction Platform** (https://avianepitope.org), a freely accessible web server requiring no account or login for core prediction functionality.

The server accepts a raw protein sequence (pasted or uploaded as FASTA), generates overlapping 8–11 mer sliding-window peptides, and scores each against the deployed BioAnchorFormer checkpoint. Results are presented as a searchable, sortable table of ranked predictions including binding scores (0–100 scale) and probabilities; a validation-tuned decision threshold is reported per prediction and updated automatically when the checkpoint is refreshed. Predictions can be exported to CSV or Excel for offline analysis.

The backend is implemented in Python with FastAPI and PyTorch; the frontend uses Next.js (React) with TypeScript. The service is containerised with Docker and served over HTTPS. Optional user accounts provide prediction history; guest users have unrestricted access to all prediction features. Email notification on job completion is supported.

List of abbreviations

AUC	Area Under the Curve
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
CTL	Cytotoxic T Lymphocyte
ESM	Evolutionary Scale Modelling
GMP	Global Max Pooling
IEDB	Immune Epitope Database
MCC	Matthews Correlation Coefficient
MHC-I	Major Histocompatibility Complex class I
NC	N-terminus + C-terminus (positional encoding)
NCEnc	NC Positional Encoding
NDV	Newcastle Disease Virus
MDV	Marek's Disease Virus
PR-AUC	Precision–Recall Area Under the Curve
ROC-AUC	Receiver Operating Characteristic Area Under the Curve
TAP	Transporter Associated with Antigen Processing

References

[1] Jim Kaufman. Unfinished business: evolution of the MHC and the adaptive immune system of jawed vertebrates. *Annual Review of Immunology*, 36:383–409, 2018.

[2] Nicolas Blanchard and Nilabh Shastri. Topological journey of parasite-derived antigens for presentation by MHC class I molecules. *Trends in Immunology*, 31(11):414–421, 2010.

[3] Jim Kaufman. Generalists and specialists: a new view of how MHC class I molecules fight infectious pathogens. *Trends in Immunology*, 39(5):367–379, 2018.

[4] Jonathan W Yewdell, Eric Reits, and Jacques Neefjes. Making sense of mass destruction: quantitating MHC class I antigen presentation. *Nature Reviews Immunology*, 3(12):952–961, 2003.

[5] Hans-Georg Rammensee, Jutta Bachmann, Niels Peter Nikolaus Emmerich, Oliver Alexander Bachor, and Stefan Stevanović. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50(3-4):213–219, 1999.

[6] Claus Lundegaard, Kasper Lamberth, Mikkel Harndahl, Søren Buus, Ole Lund, and Morten Nielsen. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Research*, 36(suppl_2):W509–W512, 2008.

[7] Andreas Hinz, Juliane Jedamzick, Valentina Herbring, Hanna Fischbach, Jana Hartmann, David Parcej, and Robert Tampe. Assembly and function of the major histocompatibility complex (MHC) I peptide-loading complex are conserved across higher vertebrates. *Journal of Biological Chemistry*, 289(48):33109–33117, 2014.

[8] Cynthia M Hess and Scott V Edwards. The evolution of the major histocompatibility complex in birds. *BioScience*, 52(5):423–431, 2002.

[9] Emily A O'Connor, Helena Westerdahl, Reto Burri, and Scott V Edwards. Avian MHC evolution in the era of genomics: phase 1.0. *Cells*, 8(10):1152, 2019.

[10] Kriengsak Boonyanuwat, Somkuan Thummabutra, Naiyana Sookmanee, Vichai Vatchavalkhu, and Virote Siripholvat. Influences of major histocompatibility complex class I haplotypes on avian influenza virus disease traits in Thai indigenous chickens. *Animal Science Journal*, 77(3):285–289, 2006.

[11] Henry D Hunt, Samadhan Jadhao, and David E Swayne. Major histocompatibility complex and background genes in chickens influence susceptibility to high pathogenicity avian influenza virus. *Avian Diseases*, 54(s1):572–575, 2010.

[12] Eric J Aston, Ying Wang, Kelsey E Tracy, Rodrigo A Gallardo, Susan J Lamont, and Huaijun Zhou. Comparison of cellular immune responses to avian influenza virus in two genetically distinct, highly inbred chicken lines. *Veterinary Immunology and Immunopathology*, 235:110233, 2021.

[13] Samer Halabi and Jim Kaufman. New vistas unfold: Chicken MHC molecules reveal unexpected ways to present peptides to the immune system. *Frontiers in Immunology*, 13:886672, 2022.

[14] Mark A Sherman, Ronald M Goto, Roger E Moore, Henry D Hunt, Terry D Lee, and Marcia M Miller. Mass spectral data for 64 eluted peptides and structural modeling define peptide binding preferences for class I alleles in two chicken MHC-B haplotypes associated with opposite responses to Marek's disease. *Immunogenetics*, 60(9):527–541, 2008.

[15] Lingxia Han, Shaolian Wu, Ting Zhang, Weiyu Peng, Min Zhao, Can Yue, Wanxin Wen, Wenbo Cai, Min Li, Hans-Joachim Wallny, et al. A wider and deeper peptide-binding groove for the class I molecules from B15 compared to B19 chickens correlates with relative resistance to Marek's disease. *The Journal of Immunology*, 210(5):668–680, 2023.

[16] Filippo Grazioli, Anja Mösch, Pierre Machart, Kai Li, Kai Ye, Martin Renqiang Min, and Michael Gribskov. On TCR binding predictors failing to generalise to unseen peptides. *PLOS Computational Biology*, 18(9):e1010411, 2022.

[17] Bjørn P.Y. Kwee, Marius Messemaker, Eric Marcus, Giacomo Oliveira, Wouter Scheper, Catherine J. Wu, Jonas Teuwen, and Ton N. Schumacher. STAPLER: Efficient learning of TCR-peptide specificity prediction from full-length TCR-peptide data. *bioRxiv*, 2023. Preprint.

[18] Katarzyna Sidorczuk, Przemysław Gagat, Filip Pietluch, Jadwiga Kała, Dawid Rafacz, Laura Bąkała, Jakub Słowik, Remigiusz Kolenda, Stefan Rödiger, Michał Gagat, et al. Benchmarks in antimicrobial peptide prediction are biased due to the selection of negative data. *Briefings in Bioinformatics*, 23(6):bbac343, 2022.

[19] Morten Nielsen, Claus Lundegaard, Peder Worning, Sine Larsen Lauemøller, Kasper Lamberth, Søren Buus, Søren Brunak, and Ole Lund. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Science*, 12(5):1007–1017, 2003.

[20] 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.

[21] Yeeleng Scott Vang and Xiaohui Xie. HLA class I binding prediction via convolutional neural networks. *Bioinformatics*, 33(17):2658–2665, 2017.

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*,

9(8):1735–1780, 1997.

[23] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

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

[25] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[26] Benjamin A Albert, Yunxiao Yang, Xiaoyu Rachel Shao, Digvijay Singh, Graham K Smyth, Terence P Speed, and Deepak L Bhatt. BigMHC: Elimination of antigen presentation and peptide–MHC binding false positives. *PLOS Computational Biology*, 19(5):e1011220, 2023.

[27] Alessandro Sette, Antonio Vitiello, Barbara Reherman, Patrick Fowler, Rainer Nayersina, Willem Martin Kast, Cornelis JM Melief, Carla Oseroff, Lai Yuan, John Ruppert, et al. The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. *The Journal of Immunology*, 153(12):5586–5592, 1994.

[28] Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Research*, 47(D1):D339–D343, 2019.

[29] Michael Harrison, Paul E Chappell, Samer Halabi, Maria Danysz, Enock M Mararo, Łukasz Magiera, Clemens Hermann, Michael J Deery, Kathryn S Lilley, Hans-Joachim Wallny, David W Avila, William Mwangi, Venugopal Nair, Susan M Lea, Nicola Ternette, and Jim Kaufman. Simplifying principles that underlie the highly complex peptide motif of the promiscuous chicken class I molecule, BF2*21:01. *eLife*, 15:RP111098, 2026.

[30] Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 1982.

[31] David L Nelson and Michael M Cox. *Lehninger Principles of Biochemistry*. W.H. Freeman and

Company, New York, 8th edition, 2021.

[32] Elisabeth Gasteiger, Christine Hoogland, Alexandre Gattiker, Séverine Duvaud, Marc R Wilkins, Ron D Appel, and Amos Bairoch. Protein identification and analysis tools on the ExPASy server. In John M Walker, editor, *The Proteomics Protocols Handbook*, pages 571–607. Humana Press, 2005.

[33] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[34] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[35] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[36] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[37] Chun Zhang, Amy Anderson, and Charles DeLisi. Structural principles that govern the peptide-binding motifs of class I MHC molecules. *Journal of Molecular Biology*, 281(5):929–947, 1998.

[38] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[39] Hans-Joachim Wallny, David Avila, Laila G Hunt, Timothy J Powell, Petra Riegert, Jan Salomonsen, and Jim Kaufman. Peptide motifs of the single dominantly expressed class I molecule explain the striking MHC-determined response to Rous sarcoma virus in chickens. *Proceedings of the National Academy of Sciences*, 103(5):1434–1439, 2006.

[40] Miriam Koch, Stacy Camp, Tjaart Collen, David Avila, Jan Salomonsen, Hans-Joachim Wallny, and Jim Kaufman. Structures of an MHC class I molecule from B21 chickens illustrate promiscuous peptide binding. *Immunity*, 27(6):885–899, 2007.