Continued domain-specific pre-training of protein language models for pMHC-I binding prediction

Sergio E. Mares[*]
Ariel Espinoza Weinberger[†]
Nilah M. Ioannidis[‡]

Abstract

Predicting peptide–major histocompatibility complex I (pMHC-I) binding affinity remains challenging due to extreme allelic diversity (∼30,000 HLA alleles), severe data scarcity for most alleles, and noisy experimental measurements. Current methods particularly struggle with underrepresented alleles and quantitative binding prediction. We test whether domain-specific continued pre-training of protein language models is beneficial for their application to pMHC-I binding affinity prediction. Starting from ESM Cambrian (300M parameters), we perform masked-language modeling (MLM)-based continued pre-training on HLA-associated peptides (epitopes), testing two input formats: epitope sequences alone versus epitopes concatenated with HLA heavy chain sequences. We then fine-tune for functional IC_{50} binding affinity prediction using high-quality quantitative data.

Key Results: Continued pre-training on epitope sequences improves model performance in data-scarce settings, with pretrained models achieving higher Spearman correlation than non-pretrained baselines at 250 training peptides before converging at larger training sizes. After continued pre-training and fine-tuning, our resulting model (ESMCBA) achieves a Spearman correlation of 0.61 on a held-out test set for predicting binding affinity across 24 common HLA alleles, competing with NetMHCpan (0.56), MHCflurry (0.49), and other state-of-the-art predictors.

Limitations: The benefits of continued pre-training are most pronounced at moderate data availability (250–1500 peptides), with diminishing returns as training data increases beyond 3000 peptides, where pretrained and non-pretrained models converge to similar performance. Additionally, the method requires substantial computational resources and performance remains fundamentally limited by the inherent noise and experimental heterogeneity in binding affinity measurements from diverse assay protocols.

Impact: This work demonstrates that domain-specific continued pre-training improves data efficiency for protein language models on specialized prediction tasks, particularly in low-resource settings. The finding that epitope-specific pre-training provides the largest performance gains at 250–1500 training examples has important implications for neoantigen vaccine prioritization, where many clinically relevant HLA alleles lack extensive binding data. More broadly, this study establishes a methodological framework for applying continued pre-training to other specialized biological prediction tasks where task-specific data is scarce but related unlabeled sequences are abundant.

[*] Center for Computational Biology, University of California, Berkeley, CA, USA

[†] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA

[‡] Center for Computational Biology and Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA; Chan Zuckerberg Biohub, San Francisco, CA, USA

20th Conference for Machine Learning in Computational Biology (MLCB 2025).

1 Introduction

Protein language models (PLMs) trained on large protein corpora have become foundational tools for structure and function prediction [Rives et al., 2021, Lin et al., 2023]. However, most applications of these models to downstream tasks involve a standard fine-tuning approach. In natural language processing, domain-specific continued pre-training—where models undergo additional unsupervised training on task-relevant data before supervised fine-tuning—often yields substantial performance gains [Gururangan et al., 2020]. Whether this strategy translates effectively to protein modeling remains largely unexplored.

We investigate this question using peptide–major histocompatibility complex class I (pMHC-I) binding affinity prediction. This task represents an important test case for several reasons. First, accurate pMHC-I binding prediction is critical for vaccine design and personalized immunotherapy [Vita et al., 2019], making performance improvements directly clinically relevant. Second, the task suffers from extreme data scarcity and imbalance: while humans express approximately 30 000 different HLA class I alleles, quantitative binding data exist for fewer than 200 alleles, with most alleles having fewer than 2000 measured peptide binding affinities. Third, the task requires joint modeling of highly polymorphic HLA chains and diverse peptide sequences, creating a stringent benchmark for cross-sequence generalization.

Current pMHC-I binding predictors face three fundamental challenges. **Allelic diversity:** The extreme polymorphism of HLA genes creates a long-tail distribution where most alleles lack sufficient training data for robust supervised learning. **Experimental bias:** Mass spectrometry-based datasets systematically over-represent peptides with experimental biases, creating training distributions skewed toward specific motifs while under-sampling strong binders [Bruno et al., 2023]. **Label heterogeneity:** Binding measurements come from diverse experimental protocols (competitive binding, mass spectrometry, radioactive binding) with varying quality and interpretation, complicating model training and evaluation.

1.1 Hypothesis and Approach

We hypothesize that domain-specific continued pre-training can improve protein language model representations of peptide sequences bound to MHC-I, improving downstream performance on binding affinity prediction. Specifically, we test whether additional masked-language modeling pre-training on HLA-associated peptides—before supervised fine-tuning—enables models to learn generalizable binding motifs across alleles.

Starting from the 300M-parameter ESM Cambrian model (ESMC) [Nijkamp and Team, 2024, Hayes et al., 2025], we implement a two-stage training protocol:

Stage 1 (Unsupervised): Continued masked-language modeling pre-training on two domain-specific corpora: (i) epitope sequences alone and (ii) epitopes concatenated with their corresponding HLA heavy chains.

Stage 2 (Supervised): Fine-tuning of the continued pre-training models for half-maximal inhibitory concentration (IC_{50}) binding affinity prediction. To mitigate experimental bias, we train exclusively on high-quality functional antagonist assays, avoiding mass spectrometry data.

We evaluate our approach—termed ESMCBA (ESM Cambrian Binding Affinity)—on the hypothesis that continued pre-training should: (1) improve performance over baseline ESM models without additional pre-training, (2) enhance data efficiency for low-resource alleles, and (3) match or exceed current state-of-the-art predictors.

1.2 Related Work

Early pMHC-I binding predictors relied on position-weight matrices and linear models. Modern neural approaches, including MHCflurry [O'Donnell et al., 2020], HLAthena [Sarkizova et al., 2020], MHCnuggets [Shao et al., 2020], NetMHCpan [Reynisson et al., 2020], and HLApollo [Thrift et al., 2024], have achieved substantial improvements. However, all of these methods train on the same types of experimental datasets and thus inherit the systematic biases present in mass spectrometry-derived training data [Bruno et al., 2023]. Recent work has begun exploring protein language models for immunological applications. However, these efforts have primarily focused on standard feature-extraction approaches without investigating domain-specific continued pre-training Thrift et al. [2024]. Our work fills this gap by systematically evaluating whether additional unsupervised learning on immunological sequences can improve downstream task performance.

In this work, we demonstrate that domain-specific continued pre-training significantly enhances pMHC-I binding prediction, with the greatest performance gains observed in data-scarce settings. We provide a systematic analysis of how continued pre-training influences pLMs across varying data availability conditions, revealing that the benefit of epitope-specific pre-training is most pronounced at moderate training sizes (250–1500 peptides). At higher data availability (>3000 peptides), both pretrained and non-pretrained models converge to similar performance, indicating that continued pre-training primarily improves data efficiency rather than maximum achievable performance. Finally, we establish a methodological framework for applying continued pre-training to specialized biological prediction tasks, demonstrating that domain adaptation is most valuable when task-specific training data is scarce but related unlabeled sequences are abundant.

2 Results

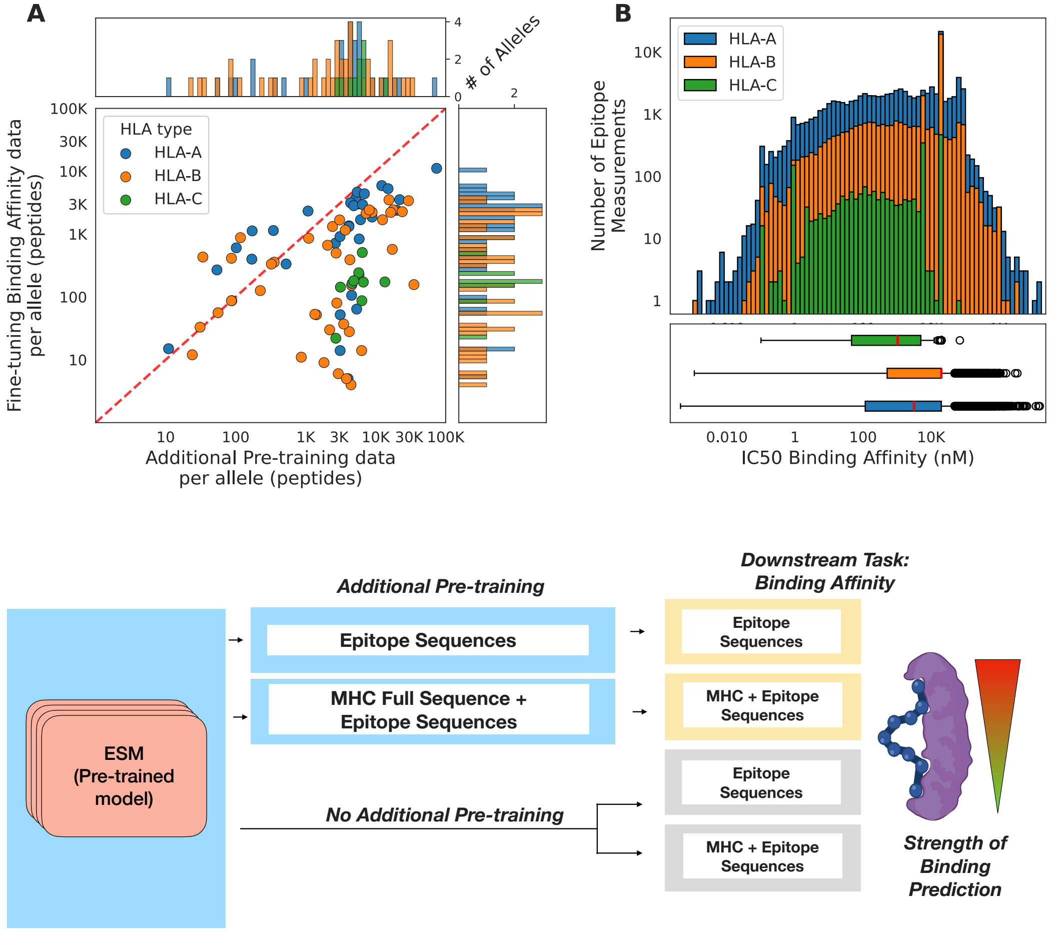


Figure 1: (A) Distribution of available pMHC-I training data across 121 HLA alleles from classes A, B, and C, showing the available data between additional pre-training data and fine-tuning binding affinity data per allele. (B) IC_{50} binding affinity distribution (nM) across HLA types. (C) Two-stage training workflow: unsupervised continued pre-training on epitope sequences (E) or HLA+epitope concatenations (H+E), followed by supervised fine-tuning for binding affinity prediction.

2.1 IC_{50} data is scarce across alleles

We first extracted quantitative peptide–MHC binding affinity measurements from the Immune Epitope Database (IEDB) across available HLA-A, HLA-B, and HLA-C alleles, filtering entries to exclude sequences containing non-canonical residues. We show (Fig. 1a) that most alleles have fewer than 2000 peptide measurements, highlighting the data scarcity for many alleles. We also observe substantial variability and notable outliers in the measured IC_{50} binding affinities (Fig. 1b).

2.2 Continued pre-training captures HLA-specific binding motifs

To evaluate whether masked language modeling can learn allele-specific epitope binding patterns, we performed continued pre-training on epitope sequences stratified by HLA allele. The analysis reveals how the model captures HLA-specific preferences through position-specific and residue-level patterns across four representative alleles: HLA-A*02:01, HLA-B*51:01, HLA-B*57:01, and HLA-C*04:01 (Fig. 2).

The training data exhibits strong positional bias, with training data showing that over 40% of epitopes at certain positions are dominated by a handful of amino acids specific to each allele (Fig. 2a-d, left panels), reflecting the structural constraints of HLA-peptide binding. Position-wise accuracy analysis demonstrates that the model learns to accurately predict residues at P9, the canonical anchor residues for pMHC-I binding, with statistically significant improvements (***p < 0.05, permutation test with Bonferroni correction) at these positions across all four alleles (Fig. 2a-d, middle panels). This indicates that continued pre-training forces the model to generate embeddings that capture the critical anchoring sites essential for pMHC-I stability.

Furthermore, the per-residue negative log likelihood analysis reveals that models learn allele-specific residue preferences (Fig. 2a-d, right panels): HLA-A*02:01 shows lowest perplexity for aliphatic residues leucine (L) and valine (V), consistent with canonical anchoring residues; HLA-B*51:01 captures its preference for proline (P) and isoleucine (I) at anchor positions; HLA-B*57:01 learns the distinctive tryptophan (W) preference; and HLA-C*04:01 identifies aspartic acid (D) and phenylalanine (F) as key residues. These findings demonstrate that the model successfully learns the physicochemical and structural constraints that govern allele-specific peptide binding.

2.3 Continued pre-training provides data efficiency in low-data regimes

To assess the impact of continued pre-training on binding affinity prediction, we compared eight training regimes across multiple HLA alleles, varying both the pre-training and encoding strategy (pre-trained vs. non-pretrained, epitope-only vs. HLA-specific encoding) and the fine-tuning configuration (0 layers vs. 30 layers unfrozen).

After filtering models with large enough test sets, we used 10 HLA alleles for this evaluation. When pooling performance across 10 HLA alleles (Fig. 3, Table S2), pretrained epitope models with 30 trained layers (PT E 30) consistently outperform their non-pretrained counterparts (Non-PT E 30) across all training sizes, though confidence intervals overlap substantially. At 250 training peptides, PT E 30 achieves $\rho = 0.463$ (95% CI: 0.384–0.538) compared to $\rho = 0.382$ (95% CI: 0.314–0.445) for Non-PT E 30, representing a 21% relative improvement ($\Delta\rho = 0.081$). This performance gap narrows systematically as training data increases: at 750 peptides ($\Delta\rho = 0.049$), 1500 peptides ($\Delta\rho = 0.031$), and 3000 peptides ($\Delta\rho = 0.013$), before nearly vanishing at 5000 peptides where both regimes converge to $\rho \approx 0.57$ (PT E 30: 0.569, 95% CI: 0.509–0.625; Non-PT E 30: 0.567, 95% CI: 0.501–0.623; $\Delta\rho = 0.002$). The consistent pattern—where pretrained models show their largest advantage at low data sizes—supports the hypothesis that pretraining provides inductive biases most valuable in data-scarce regimes (250–1500 peptides), where models must generalize from limited allele-specific examples. As training data becomes abundant (>3000 peptides), sufficient allele-specific signal enables even non-pretrained models to learn effective binding patterns, diminishing the relative benefit of pretraining.

The role of explicit MHC sequence encoding reveals a more complex picture. Models with no trained layers (0L) using HLA sequence information perform poorly regardless of pretraining status (PT H 0: $\rho = 0.223$ at 5000 samples; Non-PT H 0: $\rho = 0.201$), substantially underperforming epitope-only frozen models (PT E 0: $\rho = 0.354$; Non-PT E 0: $\rho = 0.338$). However, when fine-tuning 30 layers, HLA+epitope models achieve performance comparable to epitope-only models: at 5000 samples,

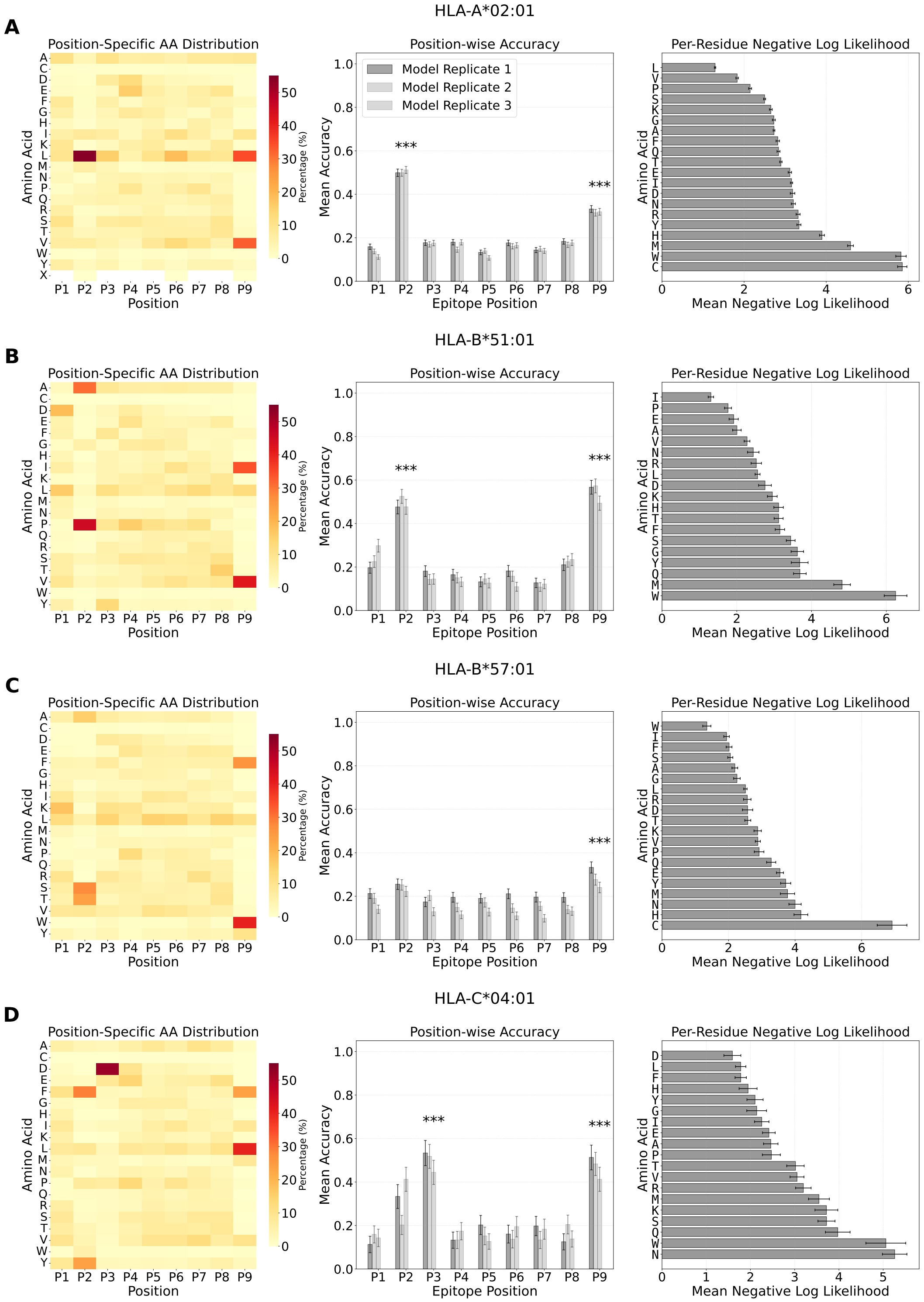


Figure 2: Position-specific and per-residue analysis of masked language modeling performance across four HLA alleles. Each row represents analysis for (A) HLA-A*02:01, (B) HLA-B*51:01, (C) HLA-B*57:01, and (D) HLA-C*04:01. The heatmap on the left shows position-specific amino acid distribution heatmap showing the percentage of each amino acid at positions P1-P9 in length-9 epitopes from training data. The bar graph in the middle shows position-wise prediction accuracy across epitope positions P1-P9 for three model replicates). The bar graph on the right shows per-residue negative log likelihood for each amino acid using the best model, sorted by mean value.

Non-PT H 30 reaches $\rho = 0.570$ (95% CI: 0.511–0.628) and PT H 30 achieves $\rho = 0.570$ (95% CI: 0.487–0.643), matching the performance of their epitope-only counterparts. This suggests that with sufficient architectural flexibility (30 trainable layers), models can learn to extract relevant binding information from concatenated HLA+epitope sequences, but this additional complexity provides no consistent advantage over epitope-only encoding. The poor performance of frozen HLA-encoding models (H 0) indicates that pretrained representations from full-length proteins do not inherently capture peptide-MHC binding groove interactions without substantial fine-tuning.

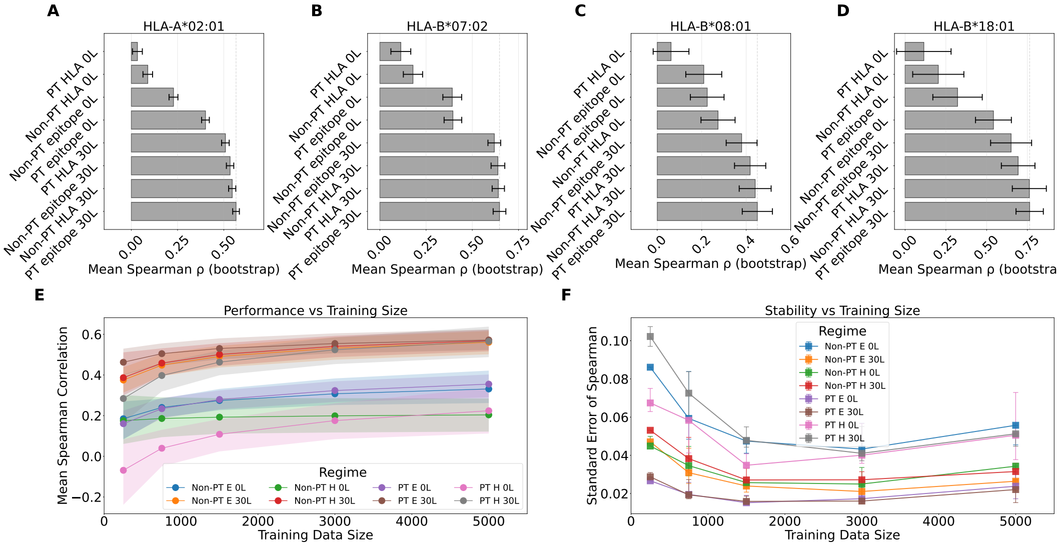


Figure 3: Comparison of training regimes across HLA alleles and training data sizes. (A-D) Mean Spearman ρ with bootstrap standard errors for different training regimes on four HLA alleles: HLA-A*02:01, HLA-B*07:02, HLA-B*08:01, and HLA-B*18:01. Each bar represents a different training regime (PT = pre-training, E = epitope, H = HLA, 0L/30L = number of unfrozen layers, Non-PT = no pre-training). (E) Performance vs training size showing mean Spearman correlation pooled across 10 HLA alleles as a function of training data proportion. Lines represent different training regimes with shaded regions indicating standard error. (F) Stability vs training size showing the standard error of Spearman correlation (replicate variability) across training data proportions. Lines represent median SE with shaded regions showing interquartile range (25th-75th percentile) across HLAs.

Several important caveats apply when interpreting these pooled results. First, aggregating performance across 10 alleles with diverse binding specificities may obscure allele-specific effects. Each HLA allele has distinct peptide-binding preferences governed by polymorphisms in the binding groove, and optimal architectures or training strategies may vary by allele. The substantial overlap in confidence intervals across all training sizes and regimes (Table S2) reflects considerable heterogeneity in how individual HLA alleles respond to different modeling choices. Pooling averages over this heterogeneity, potentially diluting signals where specific configurations excel for particular alleles.

Second, the comparable performance of epitope-only models (E 30L) and HLA+epitope models (H 30L) at high training sizes carries an important implication for allele-specific prediction: epitope sequence alone appears sufficient to learn binding patterns when training dedicated models for individual alleles. Despite encoding explicit MHC sequence information, HLA+epitope architectures with 30 trained layers (PT H 30 and Non-PT H 30, both $\rho = 0.570$ at 5000 samples) provide no advantage over epitope-only models (PT E 30: $\rho = 0.569$, Non-PT E 30: $\rho = 0.567$). This contrasts sharply with pan-allelic prediction approaches, where MHC sequence or pseudo-sequence features are typically essential to distinguish between alleles. Our results demonstrate that when training dedicated models for individual alleles, binding preferences can be effectively captured from epitope sequences alone, simplifying model architecture and reducing input dimensionality without sacrificing predictive performance. This finding challenges the conventional wisdom that explicit MHC sequence context is necessary for accurate binding prediction, at least in the allele-specific modeling paradigm.

Replicate-level stability (Fig. 3f) improves with training data for both regimes, with median standard errors decreasing from ∼0.07 at 250 samples to ∼0.03 at 5000 samples, indicating more consistent model performance with larger datasets.

2.4 Comparison to state-of-the-art models

To contextualize the performance of our approach within the broader landscape of pMHC-I binding prediction methods, we evaluated ESMCBA alongside several established models across two distinct evaluation paradigms: quantitative binding affinity prediction and qualitative immunogenicity classification. These complementary evaluations reveal how model performance varies depending on the prediction task and data characteristics, highlighting that no single approach universally excels across all scenarios.

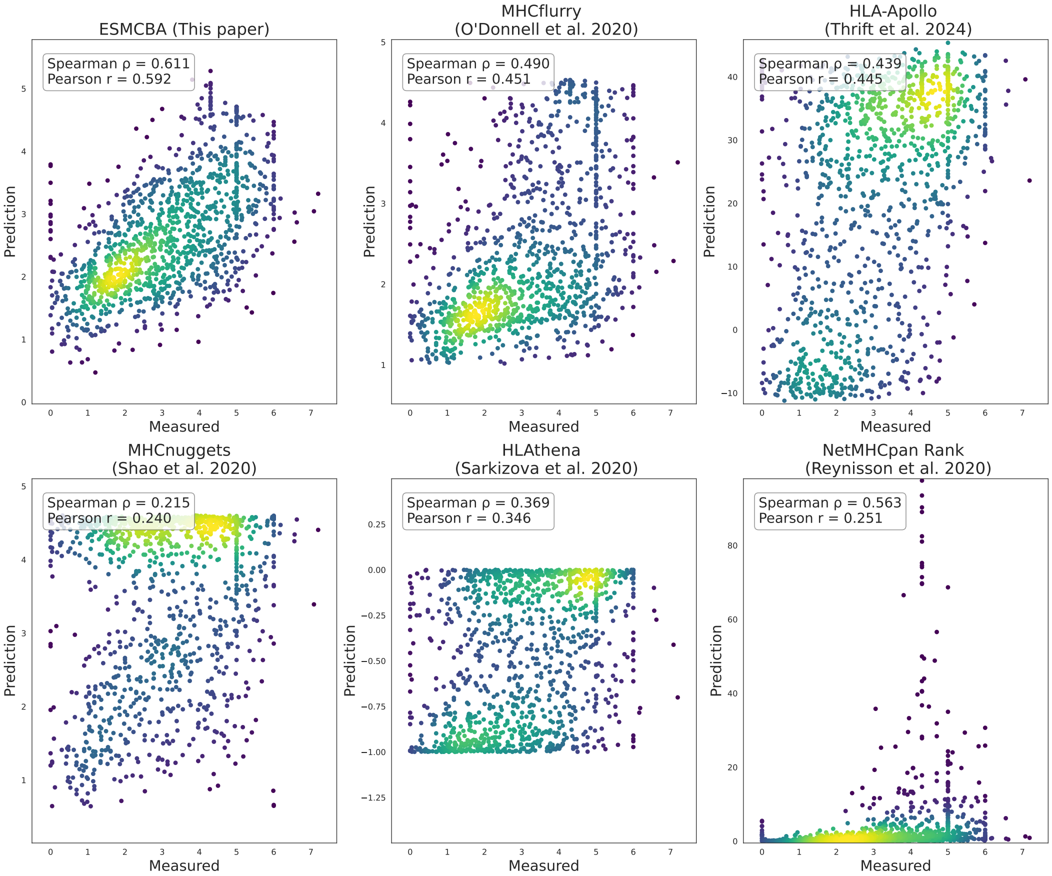


Figure 4: Predicted vs. measured pMHC-I binding affinity on a temporally held-out IEDB test set (2020 to 2025; $n = 1{,}203$ pMHC pairs). The x-axis is ground-truth $\log_{10}(\mathrm{IC}_{50}\ \mathrm{nM})$ and the y-axis is the model's native prediction without rescaling. Spearman ρ and Pearson r are annotated per panel. Each panel shows a different model in its native output scale: NetMHCpan (percentile rank), MHCflurry (IC_{50} in nM), MHCnuggets ($\log_{10}$ IC_{50}), HLA-Apollo (raw logits), HLAthena (score 0–1), and ESMCBA (this work; model score).

2.4.1 Quantitative binding affinity prediction on held-out temporal data

We first assessed the ability of each model to predict IC_{50} binding affinities for 1,203 peptide–allele pairs deposited in IEDB between 2020 and 2025, ensuring strict temporal separation from all training data (Fig. 4). For inference we used the ESMCBA (PT E 30L) as it demonstrated the best performance across training data sizes and exhibited the high stability. On this held-out test set, ESMCBA achieves the highest Spearman correlation (ρ = 0.611, Pearson r = 0.592), followed by NetMHCpan Rank (ρ

= 0.563, r = 0.251) and MHCflurry (ρ = 0.490, r = 0.451). HLApollo (ρ = 0.439), HLAthena (ρ = 0.369), and MHCnuggets (ρ = 0.215) show progressively lower correlations.

Several factors may contribute to the observed performance differences across models. ESMCBA's strong performance (ρ = 0.611) suggests that continued pre-training on epitope sequences, combined with protein language model embeddings, effectively captures binding-relevant features that generalize to temporally held-out data. NetMHCpan Rank shows competitive Spearman correlation (ρ = 0.563) but exhibits a notably lower Pearson correlation (r = 0.251), due to its ranked output. MHCflurry (ρ = 0.490), despite being explicitly trained on IC_{50} measurements using neural networks optimized for regression, performs moderately on this particular test set, possibly reflecting differences in training data composition, allele representation, or the specific characteristics of peptides deposited in IEDB between 2020 and 2025.

The performance of HLApollo (ρ = 0.439), HLAthena (ρ = 0.369), and MHCnuggets (ρ = 0.215) varies considerably, which may reflect differences in model architecture choices, training data availability at the time of their development, or the specific alleles and peptide lengths represented in this evaluation set. It is important to emphasize that this evaluation represents a single snapshot of model generalization on recently deposited data, and performance may vary depending on allele representation, peptide length distribution, and assay conditions in the test set. The temporal split ensures that models are evaluated on truly unseen data, but it also introduces potential distribution shifts if recent IEDB submissions differ systematically from older entries (e.g., increased representation of non-canonical alleles, novel experimental protocols, or evolving data quality standards).

2.4.2 Qualitative immunogenicity classification across diverse assay types

To assess how well each model generalizes beyond quantitative IC_{50} labels, we assembled a held-out set of 18,239 peptide–allele pairs carrying qualitative annotations from IEDB. To prevent data leakage, we verified that none of these epitopes appeared in the training sets of ESMCBA or MHCflurry. These labels—*Negative*, *Positive-Low*, *Positive-Intermediate*, *Positive-High*, or *Positive*—originate from diverse experimental protocols including mass spectrometry, competitive binding assays, and radioactive binding assays, and are therefore considerably noisier than the binding affinity measurements used in quantitative evaluations. For evaluation, we used the class labels as described in IEDB (e.g., Negative vs. Positive, Positive-Low vs. Positive-High) and computed the per-allele AUROC (Fig. 5).

Across the six binary classification tasks, we observe that model rankings differ substantially from the quantitative evaluation. For the broadest classification task (Negative vs. Positive), most models achieve high AUROC values (>0.8), with ESMCBA, HLAthena, HLAthena, MHCflurry, and NetMHCpan performing comparably. However, when distinguishing between more granular binding categories (e.g., Positive-Intermediate vs. Positive-High, Positive-Low vs. Positive-Intermediate), performance becomes more variable across models, with median AUROCs ranging from 0.5 to 0.8. This variability likely reflects the inherent noise in qualitative assays, which report categorical outcomes rather than continuous measurements, and the difficulty of discriminating between motifs.

Notably, the performance hierarchy observed in quantitative binding affinity prediction does not directly translate to the qualitative classification setting. Models that excel at IC_{50} regression may not necessarily dominate in categorical pMHC-Interaction tasks, and vice versa. ESMCBA maintains consistent and competitive performance across both evaluations, suggesting that the language model embeddings capture generalizable features of pMHC recognition that are robust to differences in labeling schemes and experimental protocols.

3 Discussion

3.1 Improved workflow for pMHC binding prediction

Our study demonstrates the benefits of extending domain-specific continued pre-training from natural language processing to protein modeling, specifically in the data-scarce landscape of immunology. Continued pre-training on domain-specific sequences improves predictive performance over traditional pMHC predictors, many of which are data-hungry or learn experimental biases such as those present in mass spectrometry data.

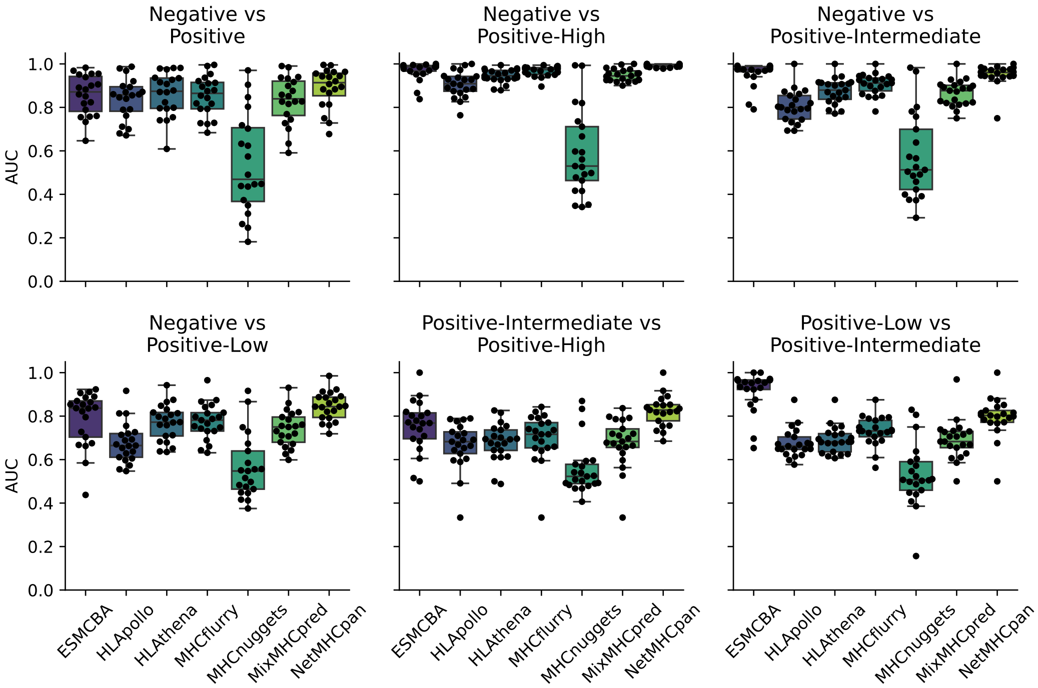


Figure 5: ROC-AUC performance across qualitative assay outcomes. Evaluations based on 18,239 qualitative entries from IEDB, excluded from quantitative training sets.

3.2 Mechanisms behind the success of continued pre-training

We propose two complementary mechanisms underlying these improvements. Firstly, pre-training on HLA-associated peptides may adjust the model's biochemical priors, better capturing residue preferences and interactions within binding peptides. Secondly, concatenating peptides with their corresponding HLA chains may facilitate learning of allele-specific binding contexts.

3.2.1 Implications for model selection

Taken together, these evaluations demonstrate that ESMCBA performs competitively with established state-of-the-art methods across diverse prediction tasks, with relative performance depending on the specific evaluation context. In the quantitative binding affinity evaluation on temporally held-out data, ESMCBA achieves the highest Spearman correlation, indicating strong generalization to recently deposited peptide–MHC measurements. In the qualitative immunogenicity classification tasks, ESMCBA performs comparably to other leading methods across multiple binary classification thresholds, demonstrating robustness to the heterogeneity inherent in categorical assay labels.

Importantly, the variability in model rankings across evaluations highlights that "state-of-the-art" performance is context-dependent and should be assessed relative to the intended application. For applications requiring precise binding affinity estimates—such as rational vaccine design where IC_{50} thresholds guide peptide selection—ESMCBA and NetMHCpan both provide strong predictive performance in our evaluation. For broader immunogenicity screening tasks involving qualitative labels from diverse experimental sources, multiple models perform comparably, suggesting that ensemble predictions combining complementary approaches may yield the most robust results.

Rather than declaring a single superior model, we advocate for task-specific benchmarking and transparent reporting of evaluation protocols, including data splits, allele coverage, and the provenance of training and test labels. The strong performance of ESMCBA across both quantitative and qualitative evaluations validates the use of protein language model embeddings for MHC-I binding prediction and demonstrates that continued pre-training on domain-specific epitope data is an effective strategy for adapting general-purpose protein language models to specialized immunological prediction tasks. These results support the hypothesis that bridging the distribution gap between general protein sequences and immunogenic epitopes through targeted pre-training enables models to learn both allele-specific binding preferences and generalizable pMHC-Interaction patterns.

3.3 Limitations and future work

Data bias is a recurrent problem in immunopeptidomics and modeling of pMHC-Interactions. IEDB serves as a cornerstone for the development of these models, yet careful considerations and practices involving noise and false positives need to be taken into account for model improvements. Our current study does not explore variations in model scale or structural supervision, nor does it thoroughly address the noisy nature of qualitative labels. Future research could leverage this work and expand the framework to MHC class II and TCR-pMHC complexes, testing whether the continued pre-training approach is scalable and effective across broader immunological scenarios.

3.4 Broader implications

Our results underscore the value of continued pre-training when applying large PLMs for biochemical prediction tasks. Modest, targeted domain-specific pre-training can result in substantial improvements, providing a practical approach for developing predictive tools essential for personalized immunotherapies and neo-antigen discovery.

4 Conclusion

We present ESMCBA as a novel allele-aware extension of the ESM protein language models, enhanced by domain-specific continued pre-training specifically on peptide–MHC sequence data. Our approach incorporates only high-quality quantitative IC_{50} measurements. By fine-tuning the task-relevant transformer layers, we significantly improve data efficiency and predictive accuracy. ESMCBA achieves a pooled Spearman correlation of 0.61 across 24 HLA alleles in training regime comparisons, competing with existing state-of-the-art predictors. The model also robustly generalizes to noisy qualitative labels, demonstrating resilience to experimental variability. Our results have important practical implications for accelerating neoantigen vaccine design cycles and facilitating large-scale screening for underrepresented alleles.

5 Methods

5.1 Data curation

We applied the following pipeline to curate quantitative binding data from IEDB:

1. Download raw IEDB entries (accessed on 16-01-2025) and filter peptides to lengths between 8 and 15 amino acids.
2. Remove all entries containing non-canonical residues.
3. Apply a $\log_{10}$ transform to IC_{50} values to stabilize variance.
4. Perform a temporal split: all peptides submitted before 2020 were used for training; peptides submitted after 2020, were held out as a test set.
5. Divide functional antagonist measurement $\log_{10}IC_{50}$ values and subsample using a Gaussian kernel centered at 10^3 nM (the approximate mean affinity across alleles), which reduces class imbalance between high-affinity and low-affinity peptides.

5.2 Overlap between pre-training and fine-tuning datasets

An important methodological consideration is the potential overlap of epitope sequences between the unsupervised continued pre-training stage and the supervised fine-tuning stage. While the same peptide sequences may appear in both datasets, this does not constitute data leakage in the traditional sense. During continued pre-training, the model performs masked language modeling on epitope sequences without access to any binding affinity labels—it learns only the motif amino acid co-occurrence within binding pMHC peptides. The model cannot memorize IC_{50} values or binding strengths it has never observed. This is analogous to standard practice in natural language processing, where language models are pre-trained on large text corpora and subsequently fine-tuned on tasks using overlapping text, but with newly introduced labels Devlin et al. [2018]. The concern in supervised learning is *label leakage*—when a model has access to test set labels during training—not sequence familiarity. In fact, exposing the model to the epitope sequence space during pre-training is beneficial: it allows the model to learn domain-specific representations of immunogenic motifs before being trained to predict binding affinity. To prevent true data leakage, we implemented strict temporal separation for the supervised evaluation: all peptides in the held-out test set (deposited in IEDB on or after 2020) were excluded from the fine-tuning training set, ensuring the model never observed their binding affinity measurements during training (Sup. Fig. S1). Therefore, while epitope sequences may recur across pre-training and fine-tuning, the absence of labels during pre-training ensures that this overlap does not compromise the validity of our evaluation.

5.3 Unsupervised Continuation pre-training

ESM Cambrian model weights were downloaded from `https://github.com/evolutionaryscale/esm`. Sequences were tokenized with the 33-character ESM vocabulary and truncated or zero-padded to a maximum length of 1,024 tokens. To adapt representations to allele-specific context, we continue pre-training with a masked-language-model (MLM) objective on peptide sequences or HLA-concatenated peptides. A linear head predicts the original amino acid at 15% of randomly selected peptide positions, while HLA residues remain visible. From the IEDB, we used positive binders as described in the qualitative labels. Training sequences were split 80:10:10 into train, validation, and evaluation sets. Peptide and HLA tokens share the same vocabulary. We introduced data augmentation by duplicating the number of sequences in our training data for this step. The unsupervised continuation was trained for 10 epochs on a single RTX 2080 Ti GPU.

5.4 Evaluation of masked language modeling performance

For the pretraining analysis (Figure 2), we performed a grid search over model hyperparameters, training each configuration on independent training and validation splits of the masked epitope dataset. For each HLA allele, we selected the top three models based on overall prediction accuracy on the held-out test set. We evaluated these models using two complementary metrics: (1) position-wise prediction accuracy, quantifying how frequently the model correctly predicted each masked amino acid residue, and (2) per-residue negative log likelihood, measuring the model's predictive confidence.

5.4.1 Standard error estimation for MLM predictions

For position-wise accuracy, we calculated the standard error at each epitope position as:

$$SE_{position} = \frac{\sigma}{\sqrt{n}}$$

where σ is the sample standard deviation of binary prediction outcomes (correct=1, incorrect=0) at that position, and n is the number of predictions at that position within the test set. This quantifies the uncertainty in estimating the true mean accuracy at each position from the finite test sample.

For per-residue negative log likelihood, we calculated the standard error across all masked occurrences of each amino acid:

$$SE_{residue} = \frac{\sigma_{NLL}}{\sqrt{m}}$$

where σ_{NLL} is the standard deviation of negative log likelihood values for amino acid type a, and m is the number of times that amino acid was masked in the test set. Error bars in Figure 2 represent these within-model standard errors, reflecting sampling uncertainty from the test set rather than model-to-model variability.

When displaying multiple models per allele (e.g., Model Replicate 1, 2, 3 in Figure 2), we show the top three performing models to illustrate the consistency of learned patterns across different training runs. The best-performing model (Model Replicate 1, shown with darker bars) was used for per-residue perplexity analysis and confusion matrix generation.

5.5 Evaluation of Model Performance across Training Regimes

To ensure valid performance estimates, we removed train-test overlap by excluding any test sequences that appeared in the training data for each model. We applied temporal cutoffs to training data, retaining sequences submitted before 2019 when available. We filtered models to include only those with at least $n_{rep} = 2$ independent replicates per HLA allele, training regime, and data proportion combination, and required test sets to contain at least $n_{test} = 20$ peptides to ensure stable correlation estimates.

For each model replicate, we computed Spearman rank correlation coefficients between predicted and measured binding affinities on the held-out test set. This resulted in a hierarchical dataset where multiple replicates (varying in random initialization and data sampling) were nested within HLA alleles, and multiple alleles were evaluated across different training regimes (pretraining status, encoding type, and number of trained layers) and training data sizes.

5.6 Statistical Transformations

To stabilize variance and satisfy modeling assumptions, we applied two transformations to our data:

1. **Fisher Z-transformation** of correlation coefficients. For each Spearman correlation r, we computed:
$$z = arctanh(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \tag{1}$$
This transformation normalizes the sampling distribution of correlations and stabilizes variance across the range of $r \in [-1, 1]$.
2. **Log-transformation** of training data size. We applied $\log(n+1)$ to the number of training examples, which linearizes power-law scaling relationships commonly observed in machine learning performance curves.

5.7 Hierarchical Mixed-Effects Model

To account for the nested structure of our data (replicates within HLA alleles) and estimate population-level scaling laws, we fit separate linear mixed-effects models for each training regime using the specification:

$$z_{ij} = \beta_0 + \beta_1 \log(n_{ij} + 1) + u_i + \epsilon_{ij} \tag{2}$$

where:

- z_{ij} is the Fisher Z-transformed Spearman correlation for replicate j of HLA i
- n_{ij} is the training data size for that replicate
- β_0 and β_1 are fixed effects representing the population-level intercept and scaling coefficient
- $u_i \sim \mathcal{N}(0, \sigma_u^2)$ is the random intercept for HLA i, capturing allele-specific baseline performance
- $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is the residual error, representing replicate-level variance

The fixed effect β_1 quantifies how model performance scales with training data size across all HLA alleles in that regime. The random effect u_i allows each HLA to have its own baseline performance while sharing the common slope. Models were fit using restricted maximum likelihood (REML) estimation with the Powell optimization algorithm (maximum 1,000 iterations).

5.8 Bootstrap Confidence Intervals

To quantify uncertainty in the population-level scaling curves, we employed a hierarchical bootstrap procedure that respects the nested data structure. For each training regime and each of five predetermined training sizes ($n \in \{250, 750, 1500, 3000, 5000\}$), we performed the following steps:

1. **Resample HLA alleles**: From the set of K unique HLA alleles, draw a bootstrap sample of K alleles with replacement. Some alleles may appear multiple times in the bootstrap sample, while others may be excluded.
2. **Construct bootstrap dataset**: For each resampled HLA allele, include all of its replicate observations across all training data sizes. This preserves the within-HLA correlation structure.
3. **Refit model**: Fit the mixed-effects model (Equation 2) to the bootstrap dataset, obtaining bootstrap estimates of the fixed effects $\beta_0^{(b)}$ and $\beta_1^{(b)}$.
4. **Generate prediction**: For the target training size n, compute the population-level prediction:

   $$z^{(b)}(n) = \beta_0^{(b)} + \beta_1^{(b)} \log(n+1) \tag{3}$$

   Transform back to the correlation scale: $r^{(b)}(n) = \tanh(z^{(b)}(n))$.
5. **Repeat**: Iterate steps 1–4 for $B = 500$ bootstrap replications.

From the bootstrap distribution $\{r^{(1)}(n), r^{(2)}(n), \dots, r^{(B)}(n)\}$, we computed the mean as the point estimate and the 2.5th and 97.5th percentiles as the 95% confidence interval for each training size.

This bootstrap approach quantifies uncertainty arising from sampling variability at the HLA level, reflecting the question: "If we had studied a different panel of HLA alleles, how would the estimated scaling law differ?"

5.9 Model Stability Assessment

To evaluate the stability of model performance across replicates, we computed within-HLA standard errors. For each HLA allele with at least two replicates at a given training size, we calculated:

$$SE_i(n) = \frac{s_i(n)}{\sqrt{m_i(n)}} \tag{4}$$

where $s_i(n)$ is the sample standard deviation of Spearman correlations across the $m_i(n)$ replicates for HLA i near training size n (within ± 0.7 log units). We then summarized these HLA-specific standard errors by computing the median and interquartile range across all HLA alleles within each regime and training size bin.

5.10 Supervised binding–affinity fine-tuning

We attach a single-unit linear head to the 300 M-parameter *ESM-Cambrian* backbone and unfreeze the 30 transformer blocks plus the final layer norm. The head receives the mean-pooled token embeddings of the last hidden layer, after a 0.3 dropout, and outputs a prediction for the binding affinity. We fine-tuned using a batch size of 12, an initial learning rate of 1×10^{-4} with linear decay, and AdamW optimization.

5.11 Benchmarking of external predictors

MHCflurry 2.1.2 produces IC_{50} values in nanomolar. HLApollo outputs raw logits that are proportional to binding likelihood, and these were used without further transformation. MHCnuggets 2.3 already reports $\log_{10}$ IC_{50}. HLAthena returns a score between 0 and 1, where larger values indicate stronger binders; we used the score as provided. For NetMHCpan 4.2, we retained its percentile rank column; smaller ranks denote stronger predicted affinity and were incorporated directly into the analyses. For every peptide–allele pair, predictions were paired with ground-truth $\log_{10}$ IC_{50} measurements (or with the qualitative labels.

5.12 Code Availability

Custom scripts and pipelines used for training and evaluation are publicly available at `https://github.com/sermare/ESMCBA`.

6 Acknowledgments

This research used the Savio computational cluster resource provided by the Berkeley Research Computing program at the University of California, Berkeley (supported by the UC Berkeley Chancellor, Vice Chancellor for Research, and Chief Information Officer). NMI is a Chan Zuckerberg Biohub San Francisco Investigator.

References

P. M. Bruno, R. T. Timms, N. S. Abdelfattah, Y. Leng, F. J. N. Lelis, D. R. Wesemann, X. G. Yu, and S. J. Elledge. High-throughput, targeted mhc class i immunopeptidomics using a functional genetics screening platform. *Nature Biotechnology*, 41(7):980–992, July 2023. doi: 10.1038/s41587-022-01566-x. URL `https://doi.org/10.1038/s41587-022-01566-x`.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. pages 8342–8360, July 2020. doi: 10.18653/v1/2020.acl-main.740. URL `https://aclanthology.org/2020.acl-main.740/`.

T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert, R. Badkundri, I. Shafkat, J. Gong, A. Derry, R. S. Molina, N. Thomas, Y. A. Khan, C. Mishra, C. Kim, L. J. Bartie, M. Nemeth, P. D. Hsu, T. Sercu, S. Candido, and A. Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, Jan. 2025. doi: 10.1126/science.ads0018. URL `https://doi.org/10.1126/science.ads0018`.

Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, Mar. 2023. doi: 10.1126/science.ade2574. URL `https://doi.org/10.1126/science.ade2574`.

E. Nijkamp and E. Team. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. EvolutionaryScale Blog, 2024. URL `https://evolutionaryscale.ai/blog/esm-cambrian`.

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

B. Reynisson, B. Alvarez, S. Paul, B. Peters, and M. Nielsen. Netmhcpan 4.1 and netmhciipan 4.0 improve MHC antigen-presentation predictions by integrating mass-spectrometry and affinity data. *Nucleic Acids Research*, 48(W1):W449–W454, 2020. doi: 10.1093/nar/gkaa379.

A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus. 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. doi: 10.1073/pnas.2016239118.

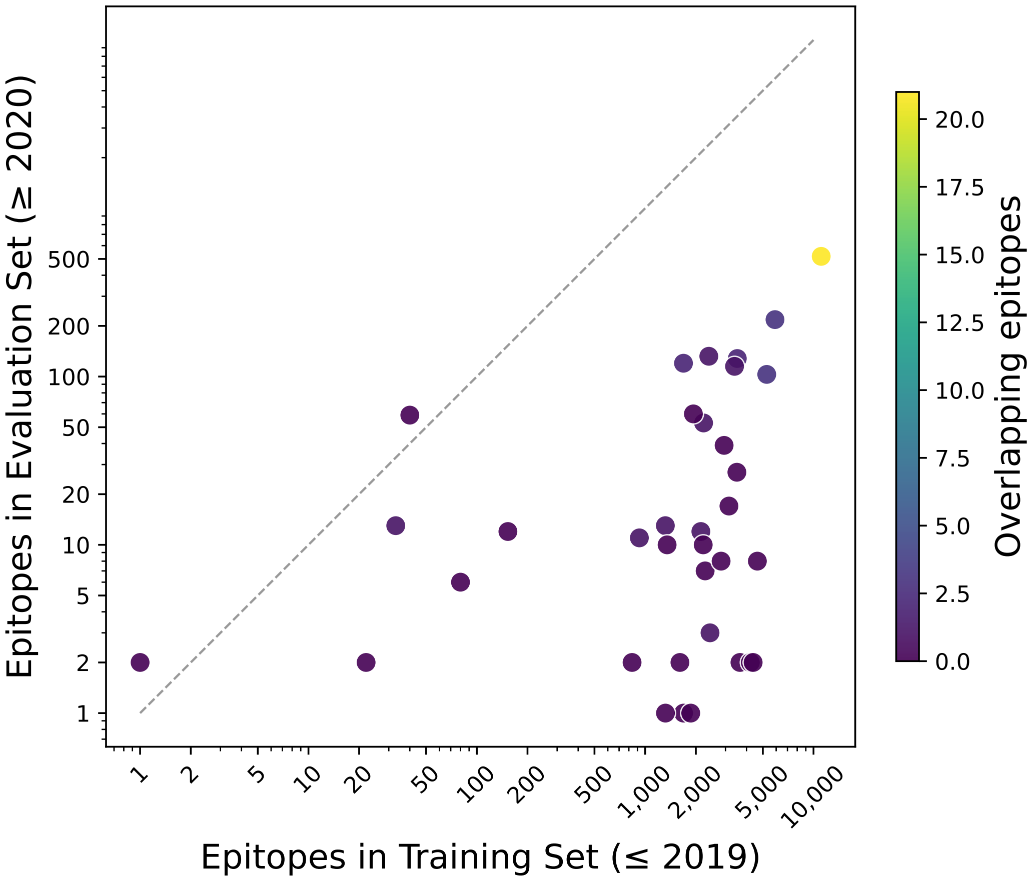
S. Sarkizova, S. Klaeger, P. M. Le, L. W. Li, G. Oliveira, H. Keshishian, C. R. Hartigan, W. Zhang, D. A. Braun, K. L. Ligon, P. Bachireddy, I. K. Zervantonakis, J. M. Rosenbluth, T. Ouspenskaia, T. Law, S. Justesen, J. Stevens, W. J. Lane, T. Eisenhaure, G. L. Zhang, K. R. Clauser, N. Hacohen, S. A. Carr, C. J. Wu, and D. B. Keskin. A large peptidome dataset improves hla class i epitope prediction across most of the human population. *Nature Biotechnology*, 38(2):199–209, Feb. 2020. doi: 10.1038/s41587-019-0322-9. URL `https://doi.org/10.1038/s41587-019-0322-9`.

X. M. Shao, R. Bhattacharya, J. Huang, I. K. A. Sivakumar, C. Tokheim, L. Zheng, D. Hirsch, B. Kaminow, A. Omdahl, M. Bonsack, A. B. Riemer, V. E. Velculescu, V. Anagnostou, K. A. Pagel, and R. Karchin. High-throughput prediction of mhc class i and class ii neoantigens with mhcnuggets. *Cancer Immunology Research*, 8(3):396–408, 2020. doi: 10.1158/2326-6066.CIR-19-0464. URL `https://doi.org/10.1158/2326-6066.CIR-19-0464`.

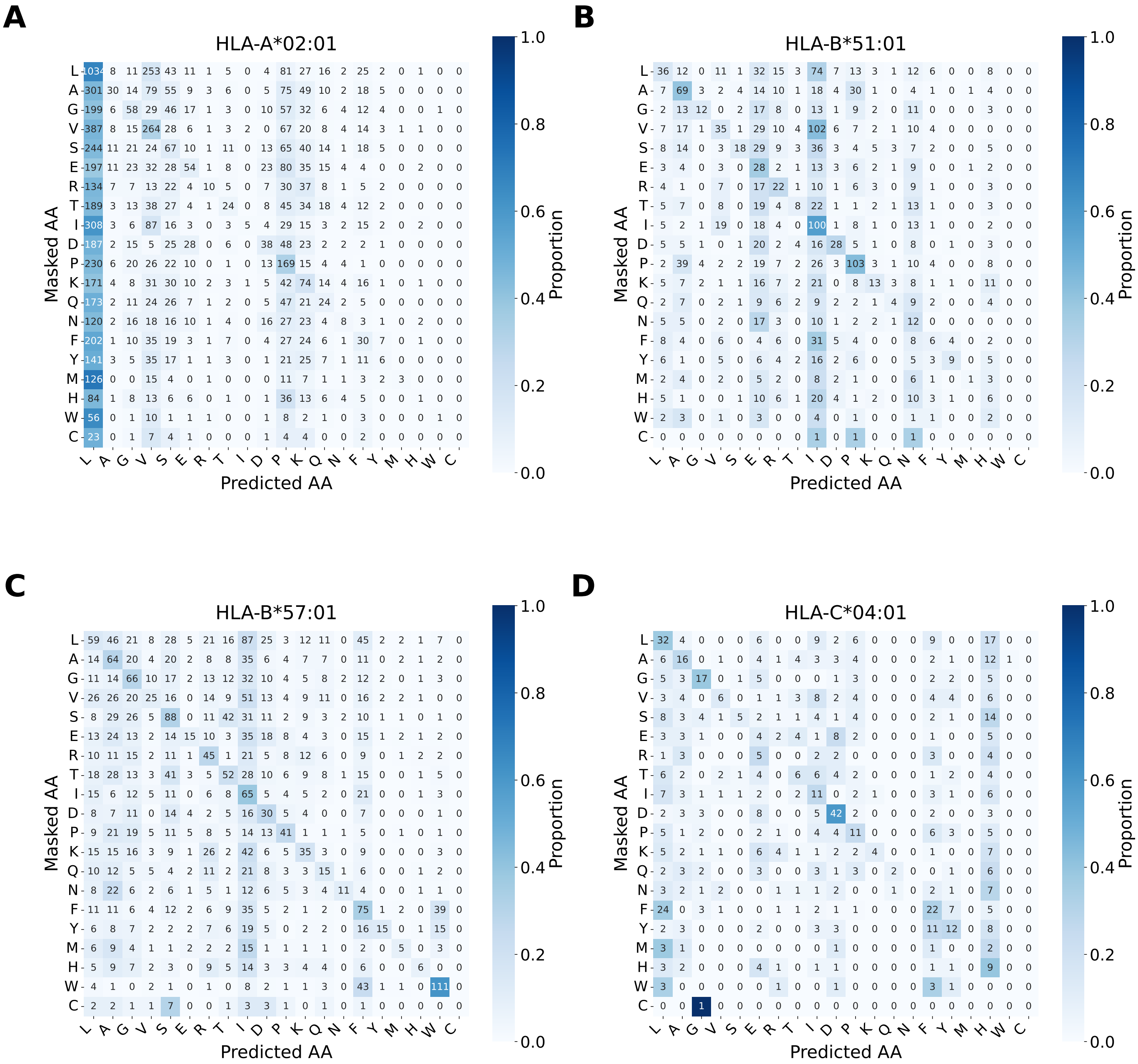
W. J. Thrift, N. W. Lounsbury, Q. Broadwell, A. Heidersbach, E. Freund, Y. Abdolazimi, Q. T. Phung, J. Chen, A.-H. Capietto, A.-J. Tong, C. M. Rose, C. Blanchette, J. R. Lill, B. Haley, L. Delamarre, R. Bourgon, K. Liu, and S. Jhunjhunwala. Towards designing improved cancer immunotherapy targets with a peptide-mhc-i presentation model, hlapollo. *Nature Communications*, 15(1):10752, 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-54887-7. URL `https://doi.org/10.1038/s41467-024-54887-7`.

R. Vita, S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette, and B. Peters. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Research*, 47 (D1):D339–D343, 2019. doi: 10.1093/nar/gky1006.

Appendix: Supplementary Figures and Tables



Supplemental Figure S1: Distribution of epitopes across training and evaluation sets for HLA alleles. Each point represents an HLA allele. Points are colored by the number of overlapping epitopes between the two sets. The dashed diagonal line indicates equal representation.



Supplemental Figure S2: Amino acid prediction confusion matrices for four HLA alleles. Heatmaps showing the proportion of correct and incorrect amino acid predictions at masked positions for (A) HLA-A*02:01, (B) HLA-B*51:01, (C) HLA-B*57:01, and (D) HLA-C*04:01. Rows represent the true masked amino acid and columns represent the predicted amino acid. Color indicates the proportion of predictions (scale: 0.0–1.0).

Model	**Hidden / FC Layers**	**Total Parameters**	**Training Size / Dataset Size**
MHCnugget	1 LSTM layer (64 units) and 1 fully connected layer (64 units)	∼26 000 per allele-specific network	Varies by MHC allele; trained on IEDB 2018 data plus extra HLAp data for some alleles
HLApollo	4 transformer encoder layers (400 dim, 16 heads) and 3 FC layers (256, 128, 1 units)	∼11.7 million[*]	953 693 unique peptide–genotype tuples across 171 HLA-I alleles
MixMHCpred 2.2	No hidden layers (position-weight matrices only)	∼90 440[**]	258 414 unique peptides, 384 070 peptide–HLA interactions, 119 HLA-I alleles
HLAthena	1 hidden layer (250 units, ReLU)	∼4.3 million	186 464 unique peptides across 95 HLA-I alleles
MHCflurry 2.0	2–3 dense layers (256–1024 units, 50 % dropout)	∼355 841[***]	713 069 peptide–MHC pairs across 171 HLA-I alleles
ESMCBA	30 transformer encoder layers (960 dim, 20 heads) + linear prediction head	∼333 million	Continued masked-language pre-training; supervised fine-tuning on peptide–MHC pairs across 121 HLA-I alleles
NetMHCpan 4.1	Ensemble of 50 neural networks, each with 1 hidden layer (56 or 66 neurons) and 2 output neurons	∼604 000 (estimated)	13 245 212 data points covering 250 distinct MHC class I molecules

Supplemental Table S1: Model architectures, parameter counts, and training data for pMHC binding affinity predictors. [*] Estimate from HLApollo publication. [**] Sum of PWM parameters. [***] Reported in MHCflurry 2.0 release notes.

Supplemental Table S2: Model Performance Across Training Sizes (Mean with 95% CI)

Group	**250**	**750**	**1500**	**3000**	**5000**
PT E 0	0.160 (0.0870, 0.225)	0.231 (0.183, 0.286)	0.277 (0.233, 0.328)	0.322 (0.270, 0.378)	0.354 (0.294, 0.417)
PT E 30	**0.463** (0.384, 0.538)	**0.504** (0.442, 0.564)	**0.529** (0.473, 0.582)	**0.552** (0.497, 0.606)	**0.569** (0.509, 0.625)
PT H 0	0.0490 (0.237, 0.104)	0.0519 (0.0590, 0.149)	0.116 (0.0353, 0.198)	0.178 (0.0830, 0.284)	0.223 (0.101, 0.363)
PT H 30	0.284 (0.160, 0.406)	0.399 (0.317, 0.470)	0.466 (0.395, 0.532)	0.528 (0.454, 0.595)	**0.570** (0.487, 0.643)
Non-PT E 0	0.179 (0.0774, 0.273)	0.240 (0.182, 0.295)	0.277 (0.221, 0.328)	0.313 (0.243, 0.381)	0.338 (0.256, 0.429)
Non-PT E 30	0.382 (0.314, 0.445)	0.455 (0.399, 0.508)	0.498 (0.440, 0.547)	0.539 (0.476, 0.592)	0.567 (0.501, 0.623)
Non-PT H 0	0.166 (0.0592, 0.280)	0.179 (0.0878, 0.272)	0.187 (0.103, 0.271)	0.196 (0.115, 0.274)	0.201 (0.123, 0.275)
Non-PT H 30	0.402 (0.304, 0.519)	0.468 (0.396, 0.542)	0.507 (0.448, 0.569)	0.545 (0.490, 0.601)	**0.570** (0.511, 0.628)

Supplemental Table S3: Predictive accuracy counts for recently submitted IEDB epitopes (2020 to 2025) across shared alleles

Allele	**Peptides tested**
HLA-A*02:01	399
HLA-A*03:01	139
HLA-A*24:02	120
HLA-A*01:01	115
HLA-B*07:02	101
HLA-B*44:02	97
HLA-A*11:01	92
HLA-B*08:01	43
HLA-A*68:01	30
HLA-B*38:01	12
HLA-B*18:01	11
HLA-A*31:01	8
HLA-B*57:01	7
HLA-A*26:01	6
HLA-B*14:02	6
HLA-B*15:01	4
HLA-A*30:01	2
HLA-B*35:01	2
HLA-B*44:03	2
HLA-B*51:01	2
HLA-C*07:01	2
HLA-A*32:01	1
HLA-B*39:06	1
HLA-B*40:01	1
Total	**1,203**

Supplemental Table S4: Number of peptides tested per allele for the ROC-AUC analysis of qualitative assay outcomes

Allele	**Peptides tested**
HLA-A*01:01	1000
HLA-A*02:01	1000
HLA-A*03:01	1000
HLA-A*11:01	1000
HLA-A*26:01	1000
HLA-A*30:01	1000
HLA-A*31:01	1000
HLA-A*68:01	1000
HLA-B*07:02	1000
HLA-B*08:01	1000
HLA-B*15:01	1000
HLA-B*18:01	1000
HLA-B*44:02	1000
HLA-B*51:01	1000
HLA-B*53:01	1000
HLA-B*57:01	1000
HLA-B*39:01	991
HLA-A*32:01	838
HLA-C*06:02	221
HLA-B*38:01	155
HLA-B*39:06	34
Total	**18,239**