НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2025 Том 25 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2025 Vol. 25 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2025-25-5-888-901

Incorporating negative examples into Hidden Markov Model-based classification of peptide sequences

Valeriia A. Polezhaeva[1]✉, Denis A. Kleverov[2]✉, Anatoly A. Shalyto[3], Maxim Artyomov[4]

[1,3,4] ITMO University, Saint Petersburg, 197101, Russian Federation
[2,4] Washington University in St. Louis. School of Medicine. Department of Pathology and Immunology, Saint Louis, 631110, USA

[1] polezhaevalera@yandex.ru✉, https://orcid.org/0009-0001-7469-7440
[2] denklewer@gmail.com✉, https://orcid.org/0009-0002-1362-486X
[3] shalyto@mail.ifmo.ru, https://orcid.org/0000-0002-2723-2077
[4] martyomov@pathology.wustl.edu, https://orcid.org/0000-0002-1133-4212

Abstract
Hidden Markov Models (HMMs) trained to identify binding regions in peptide sequences have demonstrated the ability to uncover shared amino acid patterns in peptides bound to major histocompatibility complex molecules. In this work, we present an enhanced approach for predicting peptide binding using an ensemble of HMMs. Building on a previously proposed method, we extend it to a classification setting by incorporating both binding (positive) and non-binding (negative) peptide sequences. Our strategy involves training two sets of models on these distinct datasets and selecting ensemble members based on conditional probability estimates. The method was evaluated across six alleles of major histocompatibility complex using two model architectures: simplified architecture with 9 states representing the peptide binding core region and two cycle-states for the amino acids outside this region, and extended architecture, in which each cycle state was replaced by 9 additional states. Models evaluated in comparison with the state-of-the-art MixMHC2pred predictor. Results show a statistically significant improvement in prediction accuracy. Notably, incorporating non-binding peptides during training improved performance in several cases, highlighting the importance of background sequence information in distinguishing binding-specific patterns.

Keywords
peptide binding, Hidden Markov Models, binding prediction, negative examples, epitope prediction, major histocompatibility complex, binding motifs, machine learning

For citation: Polezhaeva V.A., Kleverov D.A., Shalyto A.A., Artyomov M. Incorporating negative examples into Hidden Markov Model-based classification of peptide sequences. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 5, pp. 888–901. doi: 10.17586/2226-1494-2025-25-5-888-901

УДК 004.85

Классификация пептидных последовательностей с использованием скрытых марковских моделей, учитывающих отрицательные примеры

Валерия Александровна Полежаева[1]✉, Денис Анатольевич Клеверов[2]✉, Анатолий Абрамович Шалыто[3], Максим Артемов[4]

[1,3,4] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[2,4] Университет Вашингтона в Сент-Луисе. Медицинская Школа. Отдел патологии и иммунологии, Сент-Луис, 63110, США

[1] polezhaevalera@yandex.ru✉, https://orcid.org/0009-0001-7469-7440
[2] denklewer@gmail.com✉, https://orcid.org/0009-0002-1362-486X
[3] shalyto@mail.ifmo.ru, https://orcid.org/0000-0002-2723-2077
[4] martyomov@pathology.wustl.edu, https://orcid.org/0000-0002-1133-4212

Аннотация

Введение. Скрытые марковские модели могут применяться к задаче идентификации ядра связывания пептида с молекулами главного комплекса гистосовместимости, выявляя общие аминокислотные паттерны анализируемых последовательностей. Представлен усовершенствованный подход к решению этой задачи на основе ансамбля скрытых марковских моделей. Ранее предложенный авторами метод адаптирован к задаче классификации пептидов на два класса: связывающиеся и не связывающиеся. **Метод.** Разработанный подход включает в себя обучения двух типов моделей: первый тип — обученный с использованием связывающихся пептидов (положительных примеров данных), второй — не связывающихся пептидов (отрицательных примеров данных). Отбор моделей в ансамбль и классификация последовательностей выполнялась на основе оценки условной вероятности между полученными моделями. **Основные результаты.** Модифицированная стратегия обучения ансамбля моделей протестирована для шести различных аллелей главного комплекса гистосовместимости с использованием двух архитектур моделей. В первом случае использовалась упрощенная структура с девятью состояниями модели, соответствующими ядру связывания пептида, и двумя состояниями-циклами для аминокислот вне этого ядра. Во втором случае применялась расширенная схема, где состояния-циклы заменялись девятью дополнительными состояниями. Оценка эффективности моделей производилась в сравнении с современным методом MixMHC2pred, в ходе которой обученные модели продемонстрировали статистически значимое повышение точности предсказаний класса пептидов. **Обсуждение.** Разработанная стратегия обучения моделей учитывает как связывающиеся, так и не связывающиеся с комплексом пептиды, позволяет повысить точность предсказания класса связывания скрытыми марковскими моделями даже в условиях ограниченного объема положительных данных. Повышение предсказания в этом случае достигается за счет использования фонового распределения аминокислотных последовательностей, полученного из отрицательной выборки.

Ключевые слова

связывание пептидов, скрытые марковские модели, отрицательная выборка, предсказание связывающей способности, предсказание эпитопов, главные комплексы гистосовместимости, связывающие мотивы, машинное обучение

Ссылка для цитирования: Полежаева В.А., Клеверов Д.А., Шалыто А.А., Артемов М. Классификация пептидных последовательностей с использованием скрытых марковских моделей, учитывающих отрицательные примеры // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 888–901 (на англ. яз.). doi: 10.17586/2226-1494-2025-25-5-888-901

Introduction

A key characteristic of the immune system is its ability to distinguish foreign (non-self) targets from the body own tissues (self), a process primarily mediated by T cells. T cells play a central role in recognizing antigenic peptides presented by the Major Histocompatibility Complex (MHC) of Antigen-Presenting Cells (APCs). The interaction between T Cell Receptors (TCRs) and peptide-MHC (pMHC) complexes enables T cells to identify and respond to foreign pathogens while maintaining tolerance to self-antigens [1].

Identifying peptide sequences involved in this immune response, known as neoepitopes, is a critical step in cancer vaccine development [2]. Peptides that are likely to be bound and presented by MHC molecules and subsequently recognized by immune cells can serve as components of personalized vaccination strategies. These peptides can stimulate immune responses, even in cases where natural immune activation does not occur [3]. They exhibit high binding affinity — the strength of their interaction with MHC molecules [4].

MHC molecules exhibit specificity in peptide binding, governed by the structure of their peptide-binding grooves. The strength and stability of MHC-peptide interactions significantly influence antigen presentation and T cell activation [4].

MHC class I molecules primarily bind peptides of 9 amino acids in length, though they can accommodate peptides ranging from 8 to 15 amino acids. Their binding site consists of two helices forming a closed pocket, with interactions mainly occurring at the peptide terminal residues. The central region may play a comparatively lesser role [5, 6]. In contrast, MHC class II molecules bind longer peptides, typically ranging from 12 to 25 amino acids, with preference for 15-mers. Their open-ended binding groove allows the epitope to extend beyond the binding site, although the main stabilizing interaction with a binding groove is mediated by a nine-amino-acid region, known as the binding core. Therefore, despite the variability in peptide length, the fixed length core of MHC class II molecules plays a critical role in determining the specificity and stability of MHC-peptide interactions. Moreover, each individual can express up to six MHC class I and II types (alleles), but population-level diversity is vast — over 20,000 Class I and 8,000 Class II variants exist [5]. Each allele has unique peptide-binding preferences, generating distinct binding motifs.

From a computational perspective, selecting peptide candidates for vaccination can be formulated as a classification problem in machine learning. A model processes an input protein sequence and outputs a numerical score reflecting the likelihood of a peptide successfully passing through different stages of antigen processing.

This classification distinguishes:

— Binders — peptides that successfully bind MHC and undergo antigen processing;

— Non-binders — peptides that fail to bind or be processed effectively.

In this approach, protein sequences are encoded using standard amino acid labels (ACDEFGHIKLMNPQRSTVWY, where ‘A’ represents alanine and ‘Y’ represents tyrosine).

Hidden Markov Models (HMMs) offer a flexible and interpretable approach for identifying binding cores and predicting MHC-peptide binding affinity. Unlike neural networks, which often struggle with interpretability, HMMs can effectively model peptide sequences of different lengths by capturing underlying probabilistic dependencies [7].

Recent studies [5, 8, 9] have demonstrated the effectiveness of HMMs in binding affinity analysis, particularly due to their online learning capabilities, which allow incremental updates to transition and emission probabilities without requiring full retraining. This adaptability makes HMMs well-suited for evolving peptide datasets.

Additionally, HMMs have been explored for refining binding core identification, improving affinity estimation accuracy. However, comparative studies on different HMM architectures remain limited, and existing research often lacks detailed analyses of structural variations in these models. While some studies [5] have evaluated HMM performance for specific MHC classes, a broader comparison across multiple alleles is still needed to fully understand their potential in MHC-peptide interaction modeling.

In this study, we extend the methodology of Hidden Markov-based peptide analysis models by adapting a previously proposed approach to the machine learning task of peptide classification. Our method involves the construction of an ensemble of predictors, where each predictor is built from two independently trained HMMs: one using binder peptides and the other using non-binders, which serve as a representation of the background distribution of the peptide space. Predictors are included in the final ensemble based on conditional probability estimates derived from these models. In the following sections, we first outline the models training and validation procedures followed by ensemble construction pipeline. We then present the results of our comparative experiments across multiple MHC alleles, demonstrating improved classification performance relative to state-of-the-art MixMHCpred predictor and identifying optimal model architectures and training parameters.

Development of the Hidden Markov Model classification ensemble

Training dataset generation

Experimental data. MHC class II peptide data used to train the models were obtained from the Immune Epitope Database (IEDB) [10]. Peptides were categorized as binders based on IEDB quantitative binding data, specifically those with IC50 ≤ 500 nM or binding quality annotations such as *positive*, *positive-high*, or *positive-intermediate*. Peptides not meeting these criteria were categorized as non-binders. All protein sequences were encoded using standard 20-letter amino acid notation (ACDEFGHIKLMNPQRSTVWY), allowing direct computational processing. To ensure robust evaluation we applied four-fold cross validation [11], splitting the dataset into four equal parts. In each fold, three subsets were used for training and one for validation, resulting in four independent training and evaluation cycles per experiment.

Random sequence generation. Not all MHC alleles are equally well characterized. For some alleles, the amount of available binding data is limited [12], and existing datasets are often biased toward artificially mutated peptides derived from known binders [13]. As a result, experimental peptide datasets may not accurately represent the true background distribution of natural human peptides.

To address this, we generated an additional dataset consisting of random peptide sequences sampled from the human proteome, using data from the GENCODE database [14]. These peptides represent the natural background distribution of human sequences that could realistically occur in vivo. Given that MHC molecules bind fewer than 1 % of possible peptides [1], this random peptide set not only offers a biologically plausible background but also serves as a reasonable approximation of a negative dataset for model training and evaluation.

HMM Models training procedure

Following the training procedure outlined in a previous study [5], we train HMM by iteratively updating its three key components:

1) transition probability matrix, which defines the probabilities of transitions between hidden states;
2) emission probability matrix, which describes the probabilities of amino acid occurrences in each state;
3) initial state probability vector, which determines the probability of a sequence starting from a specific state.

We consider two standard algorithms for HMM training and optimization: the Viterbi [15] and Baum–Welch [16] algorithms. The Viterbi algorithm updates model parameters based on the single most probable state path for each sequence, preserving the interpretability and uniqueness of individual states. In contrast, the Baum–Welch algorithm averages over all possible state paths, which can dilute motif specificity. To prioritize interpretability and enable detailed analysis of binding core alignments and sequence motifs, we employ the Viterbi algorithm in this work.

A key contribution of our approach is the introduction of dedicated non-binder models, trained using negative peptide examples. To our knowledge, this represents a novel direction in MHC-peptide modeling. For each MHC allele, a single predictor consists of two separate HMMs: one trained on confirmed binder peptides, and another on non-binders. The non-binder models can either be trained using experimentally validated negative data from IEDB or initialized using amino acid frequencies derived from randomly sampled peptides from the human proteome.

Training parameters

Model architecture. We evaluated two architectures of models in this study. The first, referred to as the base model, consists of two loop states (representing peptide start and end fragments) and a central body of 9 states corresponding to the peptide binding core (Fig. 1, *a*). The second architecture, referred to as the **extended model**, introduces an additional set of 9 states which replace loop states (Fig. 1, *b*). Both include 9 states for the binding core (C_1, …, C_9). However, in the base model, the peptide terminal amino acids transition into loop regions (T_1, T_2), whereas the extended model includes extended set of states representing terminal amino acids (T_1, …, T_{18}). This design enables the model to capture not only the amino

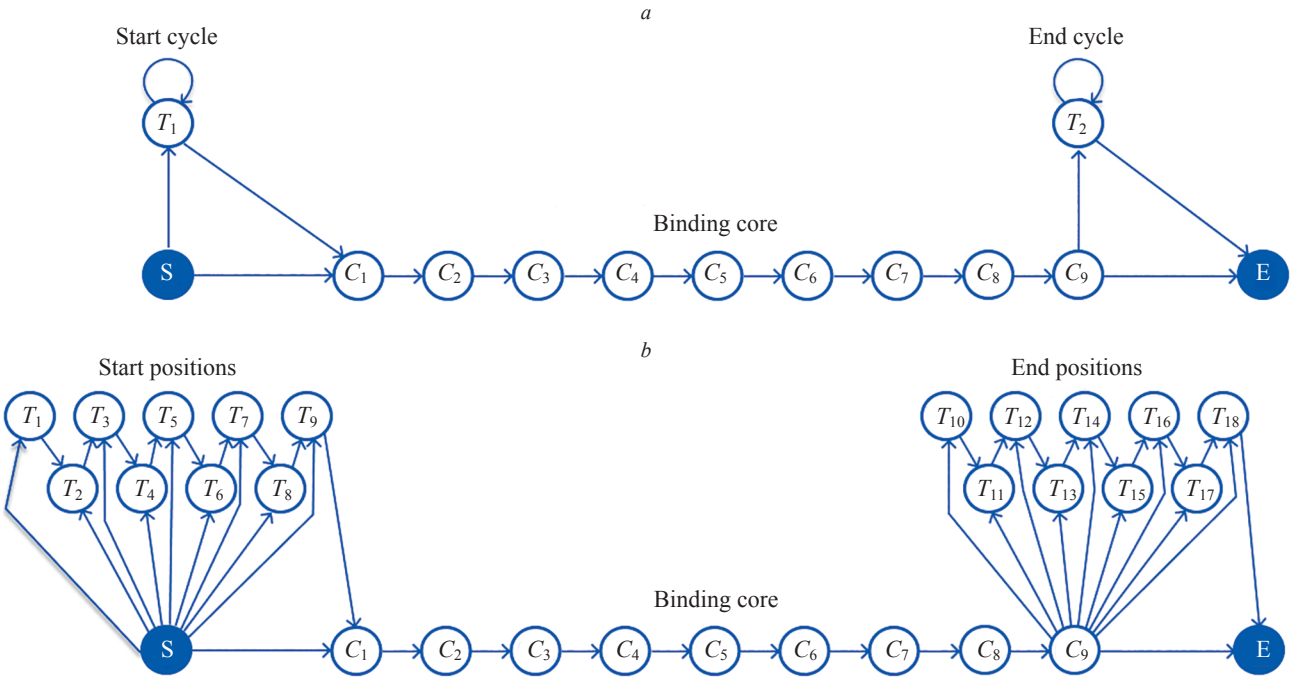

Fig. 1. Architecture of HMM: "Base model" with 2 loop states (T_1 and T_2) (*a*); Extended architecture with 9 start and end states (T_1 … T_{18}) (*b*)

acid distributions within the binding core, but also the properties of the flanking regions (known as N-terminal and C-terminal) of the peptide, which do not directly participate in binding. Each emission probability in the initial states was initialized with a random value drawn from a uniform distribution over the standard 20-letter amino acid alphabet.

Anchor state parameter. Anchor states are positions within the model that play a key role in binding prediction, typically corresponding to conserved residues within the binding core. In these states, high emission probabilities are assigned to a limited set of amino acids that occur most frequently during training. To improve model specificity, we restricted evaluated probabilities to the top four to 9 most common amino acids, as suggested in previous work [17]. This selective focus allows the model to emphasize biologically relevant sequence features while minimizing noise. To evaluate the impact of this parameter, we trained models using both architectures with anchor state values ranging from four to 9 and assessed their effect on classification performance.

Building an ensemble of models

To avoid the risk of suboptimal convergence — a known limitation of the Viterbi algorithm commonly used in HMM training — we implemented an ensemble-based strategy to improve the prediction stability. For each fold in the cross-validation procedure, 50 independent training runs were performed with different random initializations. We selected the top 70 % based on their validation performance. For each peptide, we then computed the final log-probability score by taking the median of its scores across this selected ensemble. This was done separately for both the binder and non-binder models, ensuring that the resulting scores are robust to the randomness of individual training runs.

Using these median log-probabilities, we computed the final classification score — the estimated probability that a peptide is a binder — using the following conditional probability formula:

$$Score = \frac{\exp\left(\frac{\ln P_b}{L}\right)}{\left(\exp\left(\frac{\ln P_b}{L}\right) + \exp\left(\frac{\ln P_{nb}}{L}\right)\right)},$$

where P_b — probability to generate peptide for binder model; P_{nb} — probability to generate peptide for non-binder model; L — the length of peptide.

Statistical evaluation of model performance

To assess performance differences during hyperparameter optimization of our model, we used DeLong's nonparametric test for comparing correlated Receiver Operating Characteristic (ROC) curves [18]. The test was implemented via the delong_test function from the MLstatkit Python package [19]. We evaluated significance between different configurations by comparing their predicted probabilities for each peptide against ground truth labels from testing datasets.

It was applied in the following comparisons:

— negative examples dataset choice — all combinations of anchor state parameters and architectures were compared between models trained on confirmed non-binders; randomly initialized negative examples;

— model architecture comparisons — all combinations of anchor state parameters and negative datasets were tested between the base and extended architectures.

For benchmarking against the state-of-the-art MixMHX2pred tool, we performed a paired t-test on ROC Curve (ROC AUC) scores across all data splits, comparing

our best-performing model with MixMHX2pred. The test was implemented using SciPy's ttest_rel function [20].

Results

Experiments Design

To evaluate the effectiveness of our method, we conducted the following experiments.

1. The anchor state parameter tuning. We evaluated the performance of base models using different values of the anchor state parameters to identify the optimal setting for each MHC allele.
2. Effect of non-binder dataset choice. We assessed the impact of incorporating experimentally confirmed non-binders into training by comparing ensembles built with two types of non-binder models. This allowed us to determine the most suitable negative dataset for each allele.
3. Architecture comparison. We compared ensemble performance between the base and extended model architectures to assess how model design affects prediction accuracy.
4. Overall, method evaluation. We benchmarked our HMM-based classification approach against the state-of-the-art MixMHC2pred tool, using identical validation data to ensure a fair performance comparison.

Anchor state parameter tuning

Fig. 2 presents boxplots of ROC AUC scores for the base architecture across different alleles, with varying values of the anchor state parameter. For most of the alleles we indicated non-decreasing trend. As the value of the parameter increased, the model performance remained the same or increased too. However, for HLA-DRB4*01:01 allele there was an upward trend in ROC AUC as the anchor state increases, although variability remains high. This suggests that a universal optimal value cannot be determined, and instead, the anchor state parameter must be fine-tuned individually for each allele to achieve the best predictive performance.

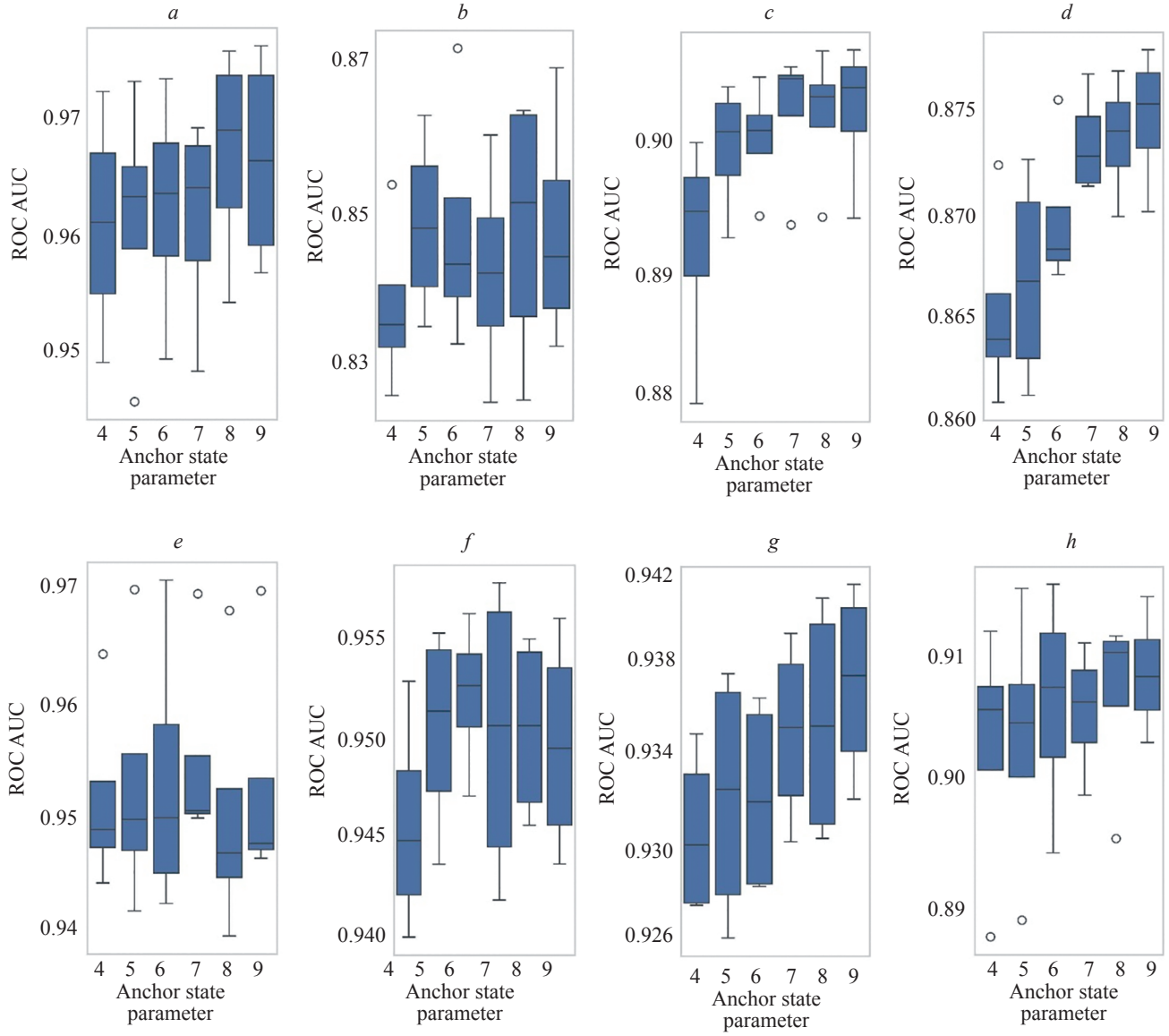

Fig. 2. ROC AUC score for base architecture for different values of anchor state parameter: HLA-DRB3*02:02 (*a*); HLA-DRB4*01:01 (*b*); HLA-DRB1*07:01 (*c*); HLA-DRB1*15:01 (*d*); HLA-DRB1*12:01 (*e*); HLA-DRB3*01:01 (*f*); HLA-DRB1*03:01 (*g*); HLA-DRB1*11:01 (*h*)

Negative examples dataset choice

In this experiment, we compared the impact of training and scoring with different types of negative datasets on the performance of different models. Specifically, five out of eight models using the base architecture and six out of eight models with the extended architecture performed better when using non-binders as the background distribution. Incorporating non-binders as negative dataset during training led to an improvement of ROC AUC scores (Fig. 3). Demonstrated performance enhancement appears to depend on the characteristics of the non-binders used in training. When the non-binders exhibited a more consistent and normal pattern — particularly with peptides of length 15 — the models achieved better performance. However, when the non-binders resembled random sequences, as observed in the HLA-DRB1*15:01 and HLA-DRB1*03:01 alleles, performance decreased, and random scoring yielded better results. This suggests that carefully selecting non-binders, rather than using completely random sequences, helps the HMMs learn more relevant distinctions between binders and non-binders, ultimately leading to improved predictive performance.

The length distribution of non-binder peptides (Fig. 5) further supports this observation, as it highlights variations in non-binder quality. A more uniform length distribution (e.g., predominantly 15-mers) correlates with improved model performance, whereas fragmented or highly variable distributions — indicative of less reliable non-binders — may contribute to reduced predictive accuracy.

Comparing model architectures

Among the eight evaluated models, the extended architecture exhibited significantly improved performance in two cases (HLA-DRB1*15:01 and HLA-DRB1*03:01 alleles) (Fig. 6). Interpretability of the model can be accessed via graph alignments for both base and extended models (Fig. 7, *a*, *b*). In this graph edges represent transitions between states of the models (with respective probabilities depicted as weights for edges) and nodes represent a single state of the models, which correspond to a single position within a peptide. To highlight the importance of the given amino acid we transformed the probability to information value of the amino acid given the emission probabilities of the state (the bigger the letter — the higher importance) (Fig. 7, *c*). Both the baseline and extended models accurately identified the binding core (Fig. 7, *a*, *b*); however, the extended architecture provided enhanced interpretability due to its more detailed terminal states. The optimal parameter combination — comprising anchor state configuration, scoring function (comparison to non-binders versus random peptides), and model architecture — favored the second approach in more cases (Fig. 6). This suggests that the selected configuration may exhibit greater sensitivity to variations in other parameters.

Comparing the best HMM models to the state-of-the-art MixMHX2pred tool

Best-performing HMM models consistently outperform the state-of-the-art MixMHC2pred tool [21] in terms of ROC AUC score across multiple alleles. In particular, significant improvements are observed for HLA-DRB1*03:01, HLA-DRB1*07:01, HLA-DRB1*12:01, HLA-DRB1*15:01, and HLA-DRB3*02:02 (Fig. 8). Even in cases where the improvement is less pronounced, such as HLA-DRB1*11:01 and HLA-DRB3*01:01, our models perform comparably to MixMHC2pred.

Discussion

Building upon the methodology outlined in a previous study [5], we developed a rigorous HMMbased training and validation pipeline with a strong emphasis on model evaluation using the area under the receiver operating characteristic curve (ROC AUC) as the primary performance metric. A central enhancement in our approach was the incorporation of non-binding peptides as a dataset of negative examples during training. Specifically, we trained two separate HMMs: one on experimentally validated binding peptides and another on non-binding peptides, which represent a background distribution of the possible peptides space.

The inclusion of negative examples enabled a more nuanced estimation of conditional probabilities, allowing us to rank test peptides based on their relative likelihood of being true binders, and solve the machine learning classification problem. Strikingly the HMM-based approach led to a clear and consistent improvement in ROC AUC scores across multiple alleles, demonstrating the value of this dual-model approach for enhancing predictive power as seen from the comparison with the state-of-the art MixMHC2pred predictor.

These findings highlight the efficacy of HMMs in MHC-peptide binding prediction and underscore the importance of incorporating negative sampling and specialized anchor states to enhance discriminative power. Notably, we observed that the quality and representativeness of non-binding peptides significantly influence the performance of models trained on non-binders, suggesting that careful selection of negative examples is critical.

Future work could explore further refinement of models architectures and training procedures, particularly improving the quality and handling of non-binder datasets, as models tend to be dependent on the quality of this data.

Additionally, one open challenge is the ability of such models to generalize to alleles exhibiting multiple peptide-binding motifs. For these alleles, models considering a single motif show reduced predictive power [22]. Addressing this limitation may require adapting the current approach to a multiclass classification framework or incorporating unsupervised motif clustering strategies.

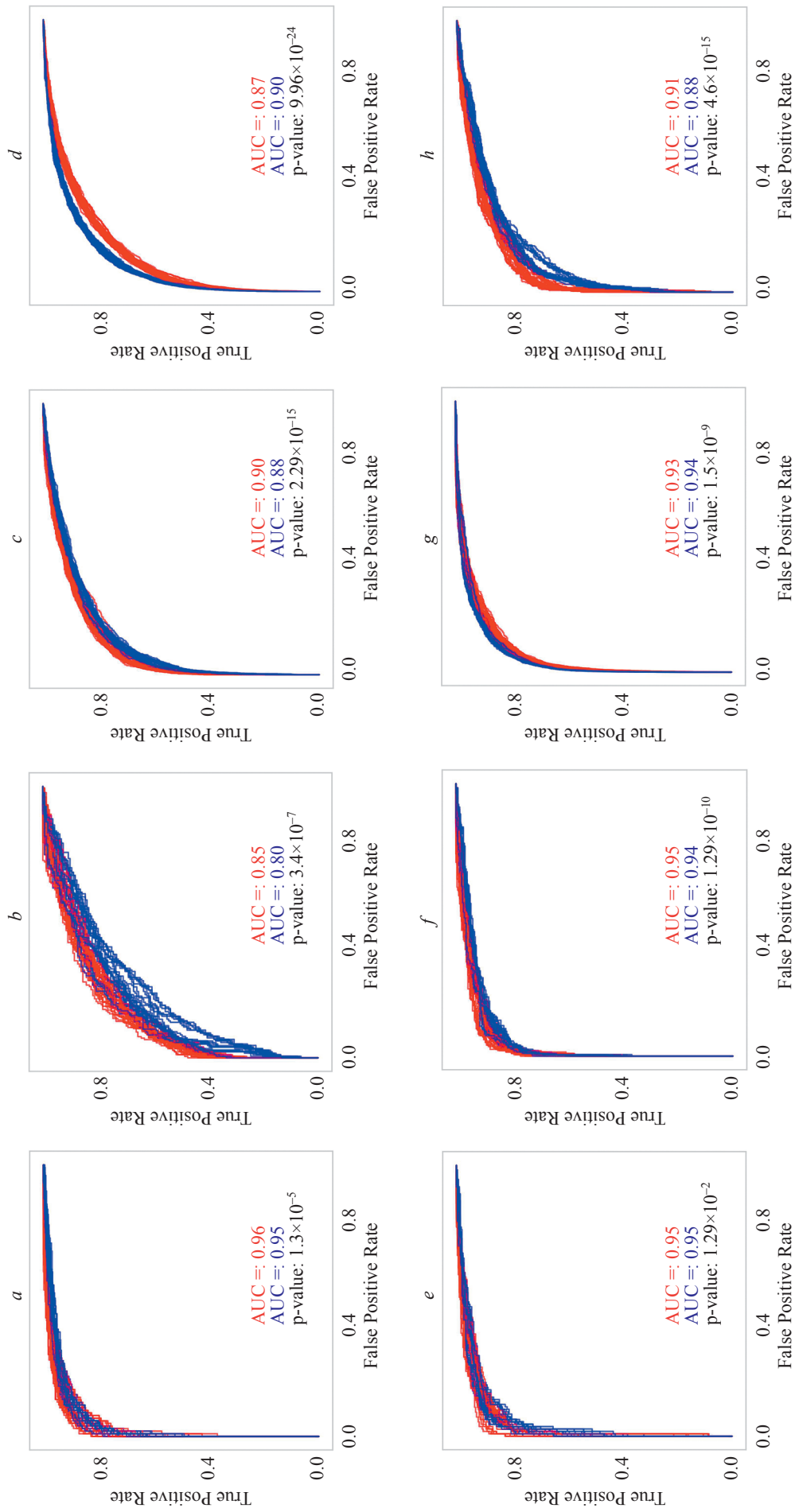

Fig. 3. ROC curves comparing non-binders (blue) vs. random (red) peptides scoring in base architecture models. P-values are used to assess the statistical significance of the difference in ROC AUC scores between the two types of test datasets: Non-binders (red) and Randoms (blue) calculated using T-test: HLA-DRB3*02:02 (*a*); HLA-DRB4*01:01 (*b*); HLA-DRB1*07:01 (*c*); HLA-DRB1*15:01 (*d*); HLA-DRB1*12:01 (*e*); HLA-DRB3*01:01 (*f*); HLA-DRB1*03:01 (*g*); HLA-DRB1*11:01 (*h*)

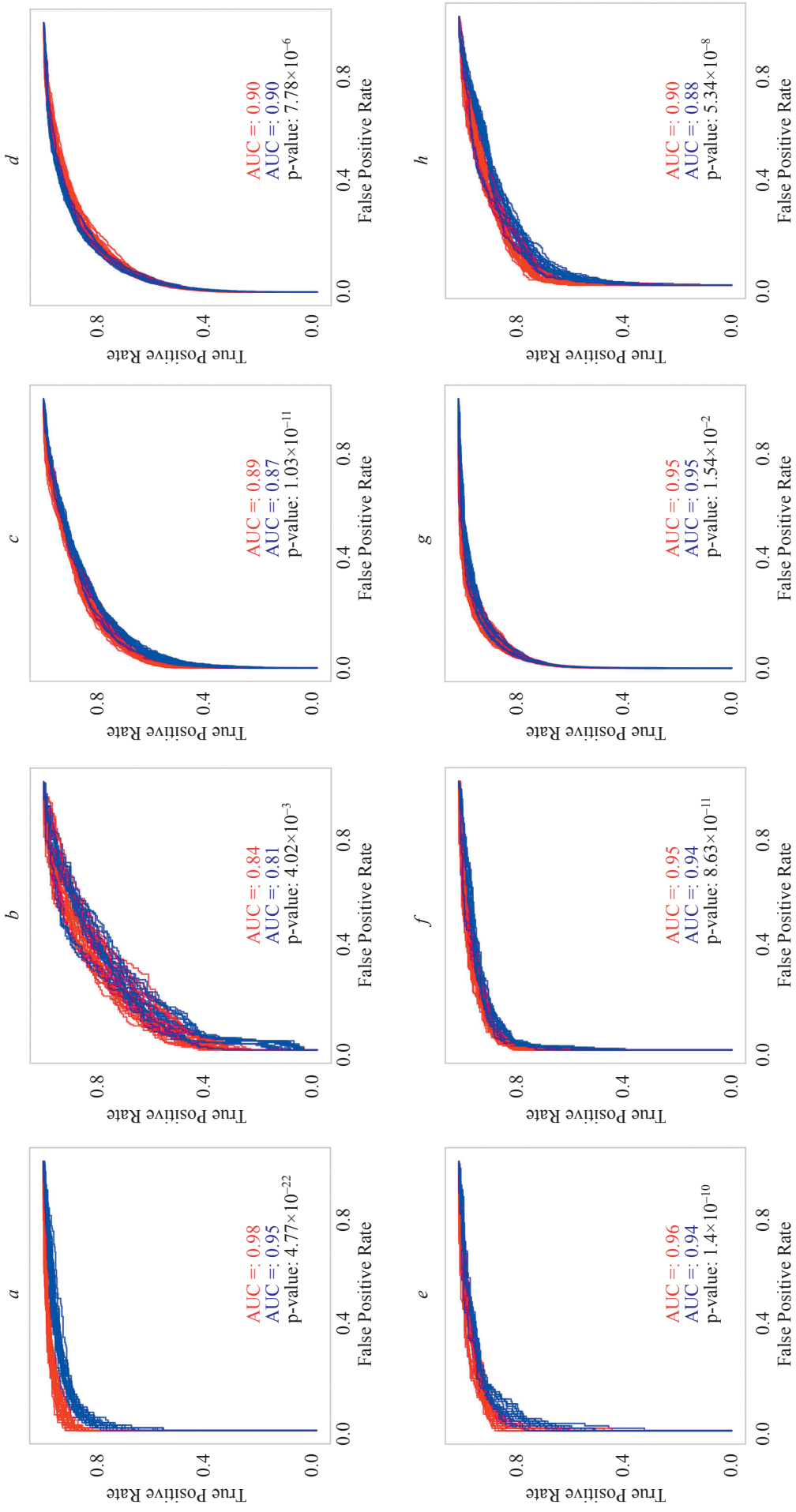

Fig. 4. ROC curves comparing non-binders (blue) vs. random (red) peptides scoring in extended architecture models. P-values are used to assess the statistical significance of the difference in ROC AUC scores between the two types of test datasets: Non-binders (red) and Randoms (blue) calculated using T-test: HLA-DRB3*02:02 (*a*); HLA-DRB4*01:01 (*b*); HLA-DRB1*07:01 (*c*); HLA-DRB1*15:01 (*d*); HLA-DRB1*12:01 (*e*); HLA-DRB3*01:01 (*f*); HLA-DRB1*03:01 (*g*); HLA-DRB1*11:01 (*h*)

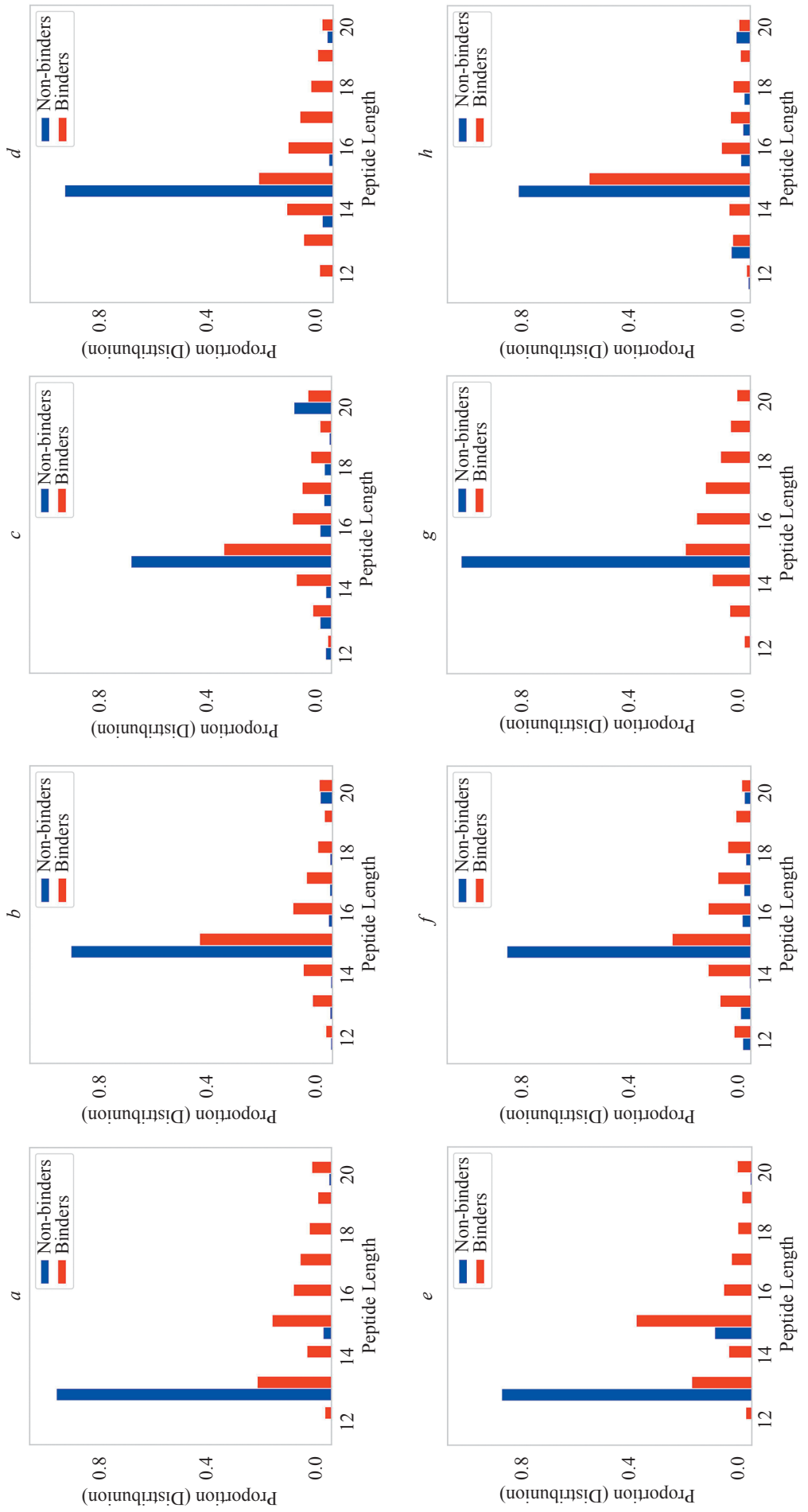

Fig. 5. Length distribution of non-binders. The predominance of 15-mer sequences suggests higher-quality non-binders, making them suitable for negative sampling when training HMMs: HLA-DRB1*03:01 (*a*); HLA-DRB1*07:01 (*b*); HLA-DRB1*11:01 (*c*); HLA-DRB1*12:01 (*d*); HLA-DRB1*15:01 (*e*); HLA-DRB3*01:01 (*f*); HLA-DRB3*02:02 (*g*); HLA-DRB4*01:01 (*h*)

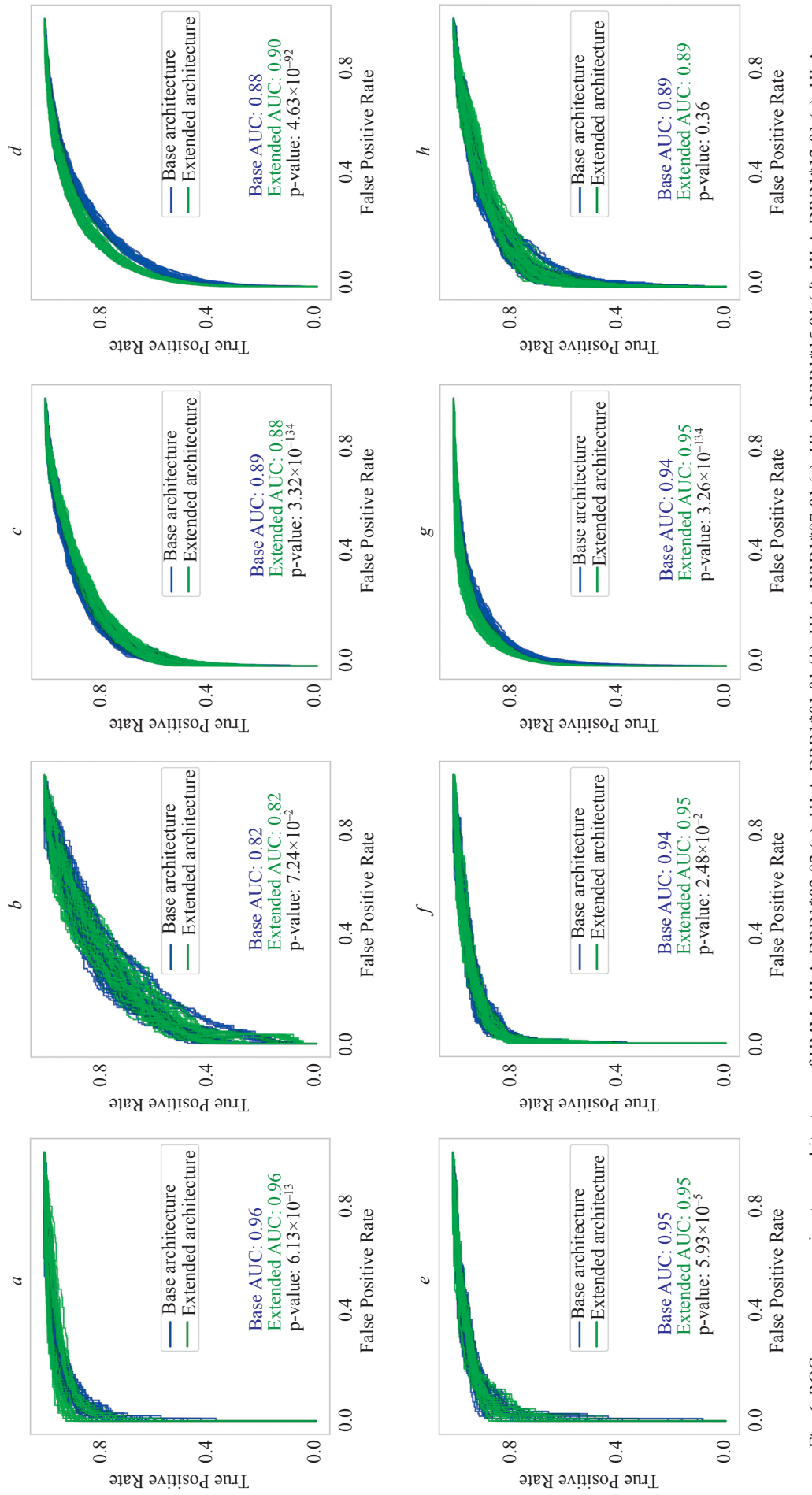

Fig. 6. ROC curves comparing two architectures of HMM: HLA-DRB1*02:02 (*a*); HLA-DRB1*01:01 (*b*); HLA-DRB1*07:01 (*c*); HLA-DRB1*15:01 (*d*); HLA-DRB1*12:01 (*e*); HLA-DRB3*01:01 (*f*); HLA-DRB3*03:01 (*g*); HLA-DRB4*11:01 (*h*)

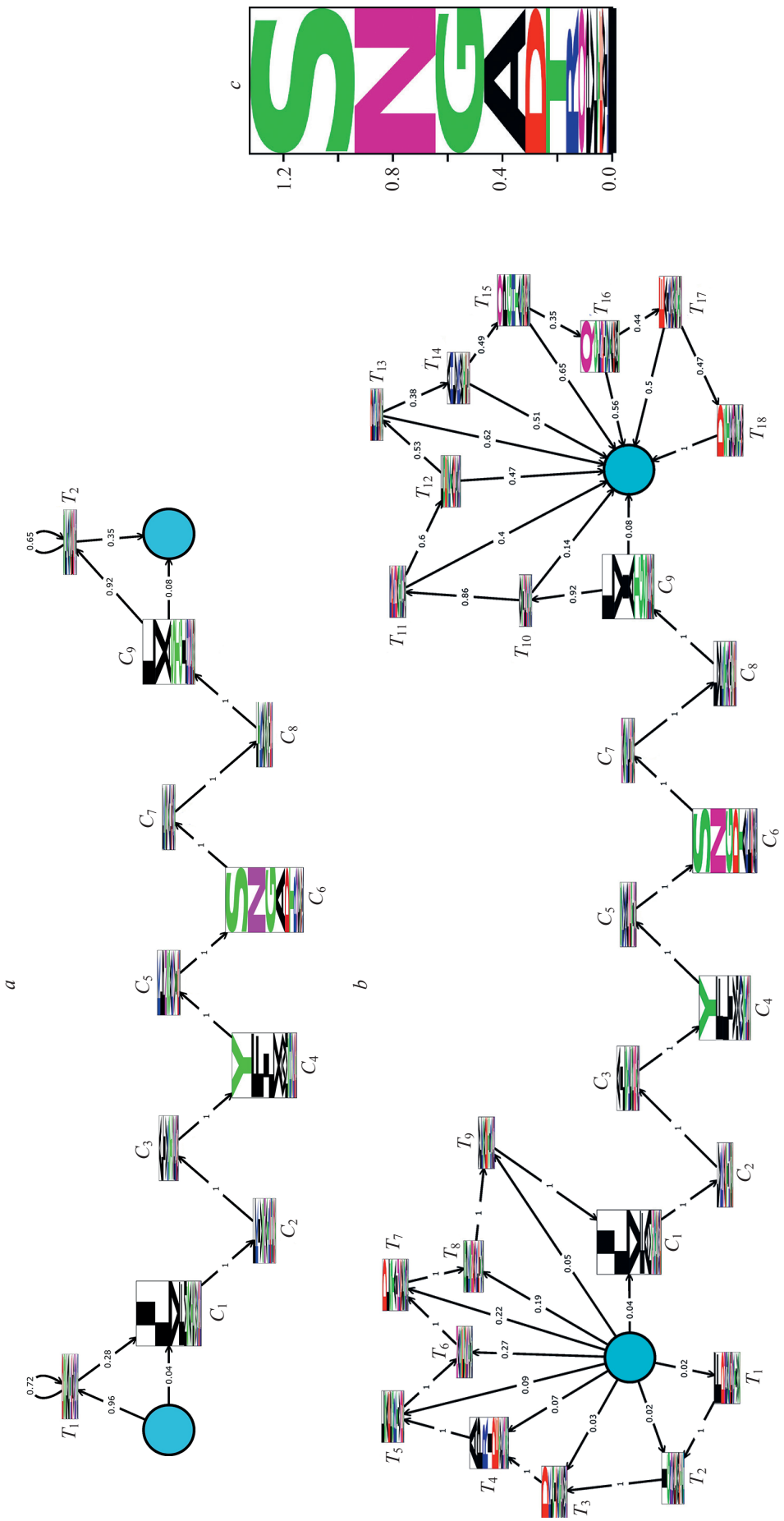

Fig. 7. The example of a state distribution plot with transition probabilities after training a model: Base architecture (*a*); Extended architecture (*b*); State distribution example represented as information content of amino acids (*c*)

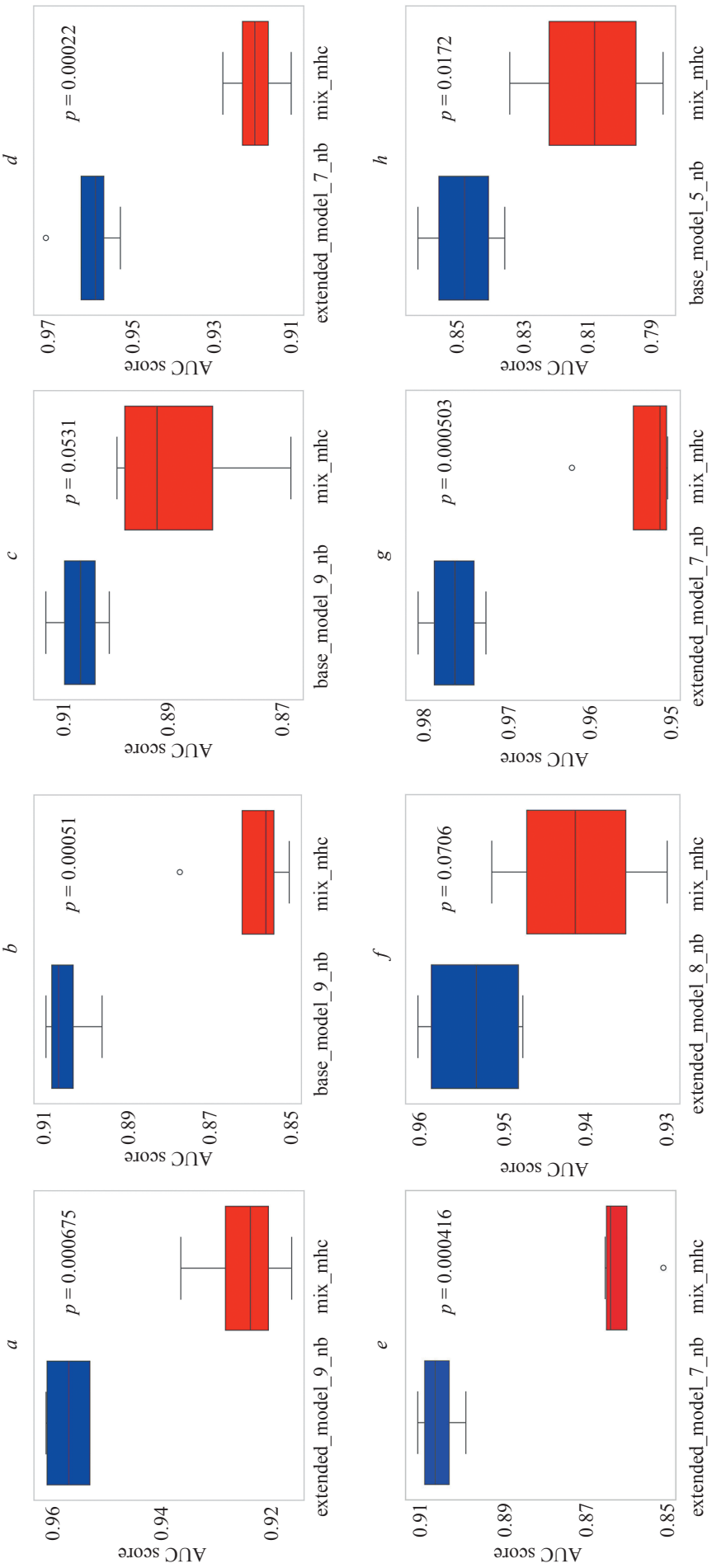

Fig. 8. Performance comparison of AUC score between fine-tuned HMM ensemble predictors with MixMHC2pred tool: HLA-DRB1*03:01 (*a*); HLA-DRB1*07:01 (*b*); HLA-DRB1*11:01 (*c*); HLA-DRB1*12:01 (*d*); HLA-DRB1*15:01 (*e*); HLA-DRB3*01:01 (*f*); HLA-DRB3*02:02 (*g*); HLA-DRB4*01:01 (*h*)

Conclusion

In this study, we demonstrated the potential of Hidden Markov Models (HMMs) to address the challenge of peptide to Major Histocompatibility Complex (MHC) binding affinity prediction. Our findings underscore the effectiveness of HMMs in predicting peptide to MHC binding affinity, offering a statistically grounded and interpretable alternative to existing methods. A key innovation was negative examples model introduction, in which separate HMMs were trained on binding and non-binding peptides. This dual-model strategy effectively captured the background distribution of proteomic sequences and led to statistically significant improvements in predictive performance, as measured by Receiver Operating Characteristic Curve (ROC AUC), compared to the current state-of-the-art, MixMHC2pred.

Our work also involved systematic optimization of anchor state parameters within the HMM architecture, guided by a conditional probability scoring framework. Performance was evaluated through extensive cross-validation and comparative analysis against models trained on either experimentally confirmed non-binders or naturally occurring peptide sequences.

References

1. Corradin G. Antigen processing and presentation. *Immunology Letters*, 1990, vol. 25, no. 1–3, pp. 11–13. https://doi.org/10.1016/0165-2478(90)90082-2
2. Abualrous E.T., Sticht J., Freund C. Major histocompatibility complex (MHC) class I and class II proteins: impact of polymorphism on antigen presentation. *Current Opinion in Immunology*, 2021, vol. 70, pp. 95–104. https://doi.org/10.1016/j.coi.2021.04.009
3. Waldman A.D., Fritz J.M., Lenardo M.J. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nature Reviews Immunology*, 2020, vol. 20, no. 11, pp. 651–668. https://doi.org/10.1038/s41577-020-0306-5
4. Wieczorek M., Abualrous E.T., Sticht J., Alvaro-Benito M., Stolzenberg S., Noé F., Freund C. Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation. *Frontiers in Immunology*, 2017, vol. 8, pp. 292. https://doi.org/10.3389/fimmu.2017.00292
5. Kleverov D.A., Shalyto A.A., Artyomov M.N. A method for constructing interpretable hidden Markov models for the task of identifying binding cores in sequences. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 5, pp. 989–1000. (in Russian). https://doi.org/10.17586/2226-1494-2023-23-5-989-1000
6. Gutiérrez S.E., Esteban E.N., Lützelschwab C.M., Juliarena M.A. Major histocompatibility complex-associated resistance to infectious diseases: the case of bovine leukemia virus infection. *Trends and Advances in Veterinary Genetics*, 2017, pp. 101–126. https://doi.org/10.5772/intechopen.68416
7. Eddy S.R. Profile hidden Markov models. *Bioinformatics*, 1998, vol. 14, no. 9, pp. 755–763. https://doi.org/10.1093/bioinformatics/14.9.755
8. Alspach E., Lussier D.M., Miceli A.P., Kizhvatov I., DuPage M., Luoma A.M., et al. MHC-II neoantigens shape tumour immunity and response to immunotherapy. *Nature*, 2019, vol. 574, no. 7780, pp. 696–701. https://doi.org/10.1038/s41586-019-1671-8
9. Kim M.W., Gao W., Lichti C.F., Gu X., Dykstra T., Cao J., et al. Endogenous self-peptides guard immune privilege of the central nervous system. *Nature*, 2025, vol. 637, no. 8044, pp. 176–183. https://doi.org/10.1038/s41586-024-08279-y
10. Vita R., Blazeska N., Marrama D., Duesing S., Bennett J., Greenbaum J., et al. The Immune Epitope Database (IEDB): 2024 update. *Nucleic Acids Research*, 2025, vol. 53, no. D1, pp. D436–D443. https://doi.org/10.1093/nar/gkae1092
11. Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009, 767 p. https://doi.org/10.1007/978-0-387-84858-7
12. Capietto A.H., Jhunjhunwala S., Pollock S.B., Lupardus P., Wong J., Hänsch L., et al. Mutation position is an important determinant for predicting cancer neoantigens. *Journal of Experimental Medicine*, 2020, vol. 217, no. 4, pp. e20190179. https://doi.org/10.1084/jem.20190179
13. Rahman K.S., Chowdhury E.U., Sachse K., Kaltenboeck B. Inadequate reference datasets biased toward short non-epitopes confound B-cell epitope prediction. *The Journal of Biological Chemistry*, 2016, vol. 291, no. 28, pp. 14585–14599. https://doi.org/10.1074/jbc.M116.729020

Литература

1. Corradin G. Antigen processing and presentation // Immunology Letters. 1990. V. 25. N 1–3. P. 11–13. https://doi.org/10.1016/0165-2478(90)90082-2
2. Abualrous E.T., Sticht J., Freund C. Major histocompatibility complex (MHC) class I and class II proteins: impact of polymorphism on antigen presentation // Current Opinion in Immunology. 2021. V. 70. P. 95–104. https://doi.org/10.1016/j.coi.2021.04.009
3. Waldman A.D., Fritz J.M., Lenardo M.J. A guide to cancer immunotherapy: from T cell basic science to clinical practice // Nature Reviews Immunology. 2020. V. 20. N 11. P. 651–668. https://doi.org/10.1038/s41577-020-0306-5
4. Wieczorek M., Abualrous E.T., Sticht J., Alvaro-Benito M., Stolzenberg S., Noé F., Freund C. Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation // Frontiers in Immunology. 2017. V. 8. P. 292. https://doi.org/10.3389/fimmu.2017.00292
5. Клеверов Д.А., Шалыто А.А., Артемов М. Метод построения интерпретируемых скрытых марковских моделей для задачи поиска связываемых участков пептидов в последовательностях белков // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23. № 5. С. 989–1000. https://doi.org/10.17586/2226-1494-2023-23-5-989-1000
6. Gutiérrez S.E., Esteban E.N., Lützelschwab C.M., Juliarena M.A. Major histocompatibility complex-associated resistance to infectious diseases: the case of bovine leukemia virus infection // Trends and Advances in Veterinary Genetics. 2017. P. 101–126. https://doi.org/10.5772/intechopen.68416
7. Eddy S.R. Profile hidden Markov models // Bioinformatics. 1998. V. 14. N 9. P. 755–763. https://doi.org/10.1093/bioinformatics/14.9.755
8. Alspach E., Lussier D.M., Miceli A.P., Kizhvatov I., DuPage M., Luoma A.M., et al. MHC-II neoantigens shape tumour immunity and response to immunotherapy // Nature. 2019. V. 574. N 7780. P. 696–701. https://doi.org/10.1038/s41586-019-1671-8
9. Kim M.W., Gao W., Lichti C.F., Gu X., Dykstra T., Cao J., et al. Endogenous self-peptides guard immune privilege of the central nervous system // Nature. 2025. V. 637. N 8044. P. 176–183. https://doi.org/10.1038/s41586-024-08279-y
10. Vita R., Blazeska N., Marrama D., Duesing S., Bennett J., Greenbaum J., et al. The Immune Epitope Database (IEDB): 2024 update // Nucleic Acids Research. 2025. V. 53. N D1. P. D436–D443. https://doi.org/10.1093/nar/gkae1092
11. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009. 767 p. https://doi.org/10.1007/978-0-387-84858-7
12. Capietto A.H., Jhunjhunwala S., Pollock S.B., Lupardus P., Wong J., Hänsch L., et al. Mutation position is an important determinant for predicting cancer neoantigens // Journal of Experimental Medicine. 2020. V. 217. N 4. P. e20190179. https://doi.org/10.1084/jem.20190179
13. Rahman K.S., Chowdhury E.U., Sachse K., Kaltenboeck B. Inadequate reference datasets biased toward short non-epitopes confound B-cell epitope prediction // The Journal of Biological Chemistry. 2016. V. 291. N 28. P. 14585–14599. https://doi.org/10.1074/jbc.M116.729020

14. Mudge J.M., Carbonell-Sala S., Diekhans M., Martinez J.G., Hunt T., Jungreis I., et al. GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Research*, 2025, vol. 53, no. D1, pp. D966–D975. https://doi.org/10.1093/nar/gkae1078
15. Forney G.D. The viterbi algorithm. *Proceedings of the IEEE*, 1973, vol. 61, no. 3, pp. 268–278. https://doi.org/10.1109/proc.1973.9030
16. Rabiner L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989, vol. 77, no. 2, pp. 257–286. https://doi.org/10.1109/5.18626
17. Nielsen M., Lundegaard C., Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics*, 2007, vol. 8, pp. 238. https://doi.org/10.1186/1471-2105-8-238
18. DeLong E.R., DeLong D.M., Clarke-Pearson D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 1988, vol. 44, no. 3, pp. 837–845. https://doi.org/10.2307/2531595
19. Sun X., Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 2014, vol. 21, no. 11, pp. 1389–1393. https://doi.org/10.1109/LSP.2014.2337313
20. Virtanen P., Gommers R., Oliphant T.E., Haberland M., Reddy T., Cournapeau D., et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 2020, vol. 17, no. 3, pp. 261–272. https://doi.org/10.1038/s41592-019-0686-2
21. Racle J., Michaux J., Rockinger G.A., Arnaud M., Bobisse S., Chong C., et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nature Biotechnology*, 2019, vol. 37, no. 11, pp. 1283–1286. https://doi.org/10.1038/s41587-019-0289-6
22. Koşaloğlu-Yalçın Z., Sidney J., Chronister W., Peters B., Sette A. Comparison of HLA ligand elution data and binding predictions reveals varying prediction performance for the multiple motifs recognized by HLA-DQ2.5. *Immunology*, 2021, vol. 162, no. 2, pp. 235–247. https://doi.org/10.1111/imm.13279

14. Mudge J.M., Carbonell-Sala S., Diekhans M., Martinez J.G., Hunt T., Jungreis I., et al. GENCODE 2025: reference gene annotation for human and mouse // Nucleic Acids Research. 2025. V. 53. N D1. P. D966–D975. https://doi.org/10.1093/nar/gkae1078
15. Forney G.D. The viterbi algorithm // Proceedings of the IEEE. 1973. V. 61. N 3. P. 268–278. https://doi.org/10.1109/proc.1973.9030
16. Rabiner L.R. A tutorial on hidden Markov models and selected applications in speech recognition // Proceedings of the IEEE. 1989. V. 77. N 2. P. 257–286. https://doi.org/10.1109/5.18626
17. Nielsen M., Lundegaard C., Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method // BMC Bioinformatics. 2007. V. 8. P. 238. https://doi.org/10.1186/1471-2105-8-238
18. DeLong E.R., DeLong D.M., Clarke-Pearson D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach // Biometrics. 1988. V. 44. N 3. P. 837–845. https://doi.org/10.2307/2531595
19. Sun X., Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves // IEEE Signal Processing Letters. 2014. V. 21. N 11. P. 1389–1393. https://doi.org/10.1109/LSP.2014.2337313
20. Virtanen P., Gommers R., Oliphant T.E., Haberland M., Reddy T., Cournapeau D., et al. SciPy 1.0: fundamental algorithms for scientific computing in Python // Nature Methods. 2020. V. 17. N 3. P. 261–272. https://doi.org/10.1038/s41592-019-0686-2
21. Racle J., Michaux J., Rockinger G.A., Arnaud M., Bobisse S., Chong C., et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes // Nature Biotechnology. 2019. V. 37. N 11. P. 1283–1286. https://doi.org/10.1038/s41587-019-0289-6
22. Koşaloğlu-Yalçın Z., Sidney J., Chronister W., Peters B., Sette A. Comparison of HLA ligand elution data and binding predictions reveals varying prediction performance for the multiple motifs recognized by HLA-DQ2.5 // Immunology. 2021. V. 162. N 2. P. 235–247. https://doi.org/10.1111/imm.13279

Authors

Valeriia A. Polezhaeva — Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0001-7469-7440, polezhaevalera@yandex.ru
Denis A. Kleverov — Visiting Researcher, Washington University in St. Louis. School of Medicine. Department of Pathology and Immunology, Saint Louis, 631110, USA, sc 58741254400, https://orcid.org/0009-0002-1362-486X, denklewer@gmail.com

Anatoly A. Shalyto — D.Sc., Full Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 56131789500, https://orcid.org/0000-0002-2723-2077, shalyto@mail.ifmo.ru

Maxim Artyomov — PhD (Chemistry), Full Professor, Washington University in St. Louis. School of Medicine. Department of Pathology and Immunology, Saint Louis, 631110, USA; Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 9242717500, https://orcid.org/0000-0002-1133-4212, martyomov@pathology.wustl.edu

Авторы

Полежаева Валерия Александровна — студент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0001-7469-7440, polezhaevalera@yandex.ru
Клеверов Денис Анатольевич — приглашенный научный сотрудник, Университет Вашингтона в Сент-Луисе. Медицинская Школа. Отдел патологии и иммунологии, Сент-Луис, 63110, США, sc 58741254400, https://orcid.org/0009-0002-1362-486X, denklewer@gmail.com
Шалыто Анатолий Абрамович — доктор технических наук, профессор, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 56131789500, https://orcid.org/0000-0002-2723-2077, shalyto@mail.ifmo.ru
Артемов Максим — PhD, химические науки, профессор (исследователь), профессор, Университет Вашингтона в Сент-Луисе. Медицинская Школа. Отдел патологии и иммунологии, Сент-Луис, 63110, США; профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 9242717500, https://orcid.org/0000-0002-1133-4212, martyomov@pathology.wustl.edu

Received 06.05.2025
Approved after reviewing 29.08.2025
Accepted 20.09.2025

Статья поступила в редакцию 06.05.2025
Одобрена после рецензирования 29.08.2025
Принята к печати 20.09.2025