A Transformer-Based Multi-Modal Framework for Antigen Prediction and Vaccine Target Discovery

Ruoyu Wang

Shenzhen College of International Education, Guangzhou, China

chiharubot@gmail.com

Keywords: Antigen prediction; Vaccine design; Multi-modal learning; Transformer; Cross-attention fusion

Abstract: Vaccines are still the strongest defense against viral infections. However, finding effective antigens through traditional methods takes time and heavy experimental effort. Many computational approaches try to speed up this process, but most depend only on sequence data. They often overlook the structural and biochemical relationships that are key to understanding immunogenicity. In this study, we introduce a transformer-based multi-modal model for antigen prediction. The model combines three sources of information: physicochemical properties, amino acid sequences, and 3D structural features. A cross-attention fusion module connects these features and allows the model to learn how they interact. This design helps the system capture complex biological signals and improves recognition accuracy. We tested the model on bacterial and viral antigen datasets. It consistently performed better than traditional machine learning and single-modality deep learning methods. Across all metrics—accuracy, F1-score, and AUC—it showed strong improvements. The model can also locate highly immunogenic, surface-exposed fragments. Overall, it offers a clear, interpretable, and efficient computational tool for vaccine target discovery.

1. Introduction

For the prolonged rivalry between humans and viruses, antiviral vaccines have consistently served as one of the most effective defensive measures. This is particularly evident in combating novel or rare diseases, such as Japanese Encephalitis (JE) and Tick-Borne Encephalitis (TBE), where the development of vaccination solidifies an essential role in pandemic mitigation and public health safeguarding. Those without, for instance, Eastern Equine Encephalitis Virus (EEV) and West Nile Virus (WNV), thus serve as a latent threat to the human population. Nonetheless, traditional vaccination research and development are extensive, usually encom-passing the following steps: initially, to identify the immune response-inducing fragment of the viral protein (antigen screening), followed by animal testing and clinical trials, ultimately allowing mass production and employment. Given as such, the process not only takes multiple years, but also requires the input of a colossal amount of manpower and material resources. Antigen screening, hence within the developmental process, serves as the decisive and critical first step. Should there be earlier, accelerated acknowledgement of the proteins or peptides with immunogenicity, subsequent experiment and test stages would undergo a large increase in efficiency. However, traditional screening methods predominantly rely on expensive and time-consuming experimental approaches.

In the past few years, with advances in artificial intelligence (AI), more and more studies have begun applying computational methods to enhance the optimization of vaccine target screening efficiency. Notably, the application of machine learning (ML) in reverse vaccinology has greatly helped streamline the automation and accuracy of target identification. For instance, Ong et al. (2020) [1] created the Vaxign-ML tool, an integration of ML and reverse vaccinology, which successfully predicted the antigenicity of several SARS-CoV-2 proteins with high precision, including the conventional S protein and the poorly investigated nsp3 protein, thereby expanding the range of possible vaccine targets. Bravi (2024) [2] elucidated the practical applications of ML in

  DOI: 10.25236/iwmecs.2025.058

vaccine target screening and indicated that ML algorithms can rapidly identify potential B-cell and T-cell epitopes, thus enabling significant decision-making during vaccine design. Mugunthan et al. (2023) [3] demonstrated the significant potential of combining reverse vaccinology and ML through the successful prediction of Mycoplasma gallisepticum multi-epitope vaccine structures using various computational tools. Despite these advancements in data-driven target prediction, several limitations remain. First, most existing works focus mainly on the sequence level with little involvement of three-dimensional (3D) structural data, so potential spatial conformation impacts on antigenicity are not taken into account. Second, traditional ML models (e.g., support vector machines, random forests) still rely on manually crafted physicochemical properties, limiting their adaptive learning capacity and ability to capture high-order functional semantics of proteins comprehensively. Therefore, there is a strong need to create more expressive modeling frameworks that combine protein sequence, physicochemical properties, and 3D structural data coherently to improve the accuracy of immunogenicity prediction.

Recently, vaccine design and antigen prediction have become a common use of deep learning technology. The preliminary survey of deep learning-aided epitope recognition and vaccine construction techniques by Bhattacharya et al. (2025) [4] systematically explained how AI-based prediction systems are slowly overtaking the traditional methods that require experiments to be carried out. As an example, the Vaxi-DL model offered by Kamal Rawal and colleagues (2021) [5] uses fully connected neural networks to predict the physicochemical properties and fundamental biological properties of protein sequences, and its effectiveness in predicting antigens is relatively high. Nevertheless, the approach can only use one-dimensional sequences as input and ignores the spatial conformation information in the protein structures. It also does not model the dependency relationship among remote residues, which could be the most important factor in antigen recognition. In addition, the general state of these methods is that various source characteristics (including sequence physicochemical properties and amino acid composition) are treated as independent inputs of equivalent importance, without deep fusion strategies, and are unable to reflect interactions and synergies among these characteristics. Accordingly, even beyond their advances, there exist two key gaps: first, the inability to present information about protein structure to model features; and second, the lack of integration and communication between various feature modalities. To this end, a universal model framework that integrates protein sequence data, biochemical behavior, and three-dimensional spatial architecture is necessary to conduct a comprehensive excavation of the latent determinants that impact antigenicity, thereby presenting more biologically interpretable and generalizable predictive models that can be applied to subsequent vaccine development.

Addressing these limitations—such as the inability to capture long-range dependencies, the neglect of spatial structural features, and the lack of effective fusion across different feature modalities—this paper proposes a transformer-based multi-encoder antigen prediction model. By integrating protein sequence information, physicochemical properties, and three-dimensional spatial structures, the model enhances the accuracy of antigen identification. The proposed framework consists of three parallel encoders, each dedicated to processing a distinct input feature modality. Specifically, the physicochemical feature encoder extracts biochemical properties of each amino acid (e.g., hydrophobicity, polarity, and surface accessibility) obtained via the PyPro toolkit; the sequence feature encoder processes raw amino acid sequences through an embedding-based transformer to capture antigen-related patterns; and the structural feature encoder leverages AlphaFold-predicted tertiary structural features—such as residue-wise distance maps and contact maps—to effectively model spatial dependencies among residues. Each encoder consists of 2–4 transformer layers, employing self-attention mechanisms to model intra-modal dependencies. Prior to encoding, a linear projection aligns all features into a unified latent space, ensuring cross-modality compatibility. Finally, a Cross-Modal Attention Fusion Module is introduced to dynamically align, interact, and integrate features from all modalities. Unlike conventional feature concatenation, our attention-based fusion adaptively learns inter-modal importance weights, enhancing the model's ability to prioritize informative representations and improve discriminative

power for antigenic epitope prediction.

We evaluated systematically the proposed multimodal fusion model using several public anti-gen protein data sets, and compared comprehensively with mainstream existing approaches,i.e., traditional Machine Learning models (Support Vector Machines, Random Forests),representa-tive Deep Neural Networks (Multilayer Perceptrons, Convolution Neural Networks, Recurrent Neural Networks, Long Short-Term Memory network), and conventional Transformer encoder model.Our findings show that our approach produces a better quality according to all scores. Even more significantly, if we mimic predictions on the set of antigens that had not yet been pub-lished containing EEE virus, WNV, and YF virus (that are antibody candidates), we confidently can confirm that our model predicts structurally accessible, highly-immunogenic fragments, thus showing that it has the potential to be used in practice to screen real antibody candi-dates.This enables a more efficient and data-driven toolchain for downstream vaccine design of particular utility for virus species undersampled, difficult to experiment with, or indifferent to the common methods. This work shows strong generalisation capability and interpretability on several practical cases, providing reliable computational assistance for the search for targets for vaccines against emerging mosquito-borne viruses. Additionaly, we can utilize it for high throughput prediction screening, with which experiment time and money in vaccine preclinical development will be dramatically shortened. Our framework is shown in Fig.1.

Our contributions are summarized as follows:

• We introduce a transformer-based multi-encoder model that fuses protein sequence, physicochemical, and structural modalities, offering a combinatorial representation for the antigen prediction.

• We introduce a cross-modal attention fusion module that adaptively captures inter-modal dependencies and enhances the interpretability of antigenicity prediction.

• We show the performance of our model on multiple benchmark datasets, as well as practical utility in predicting putative antigenic fragments for EEEV, WNV and Yellow Fever Virus.

• We believe that our framework offers a generalizable and biologically interpretable com-putational tool to facilitate antigen discovery and guide data driven vaccine design.

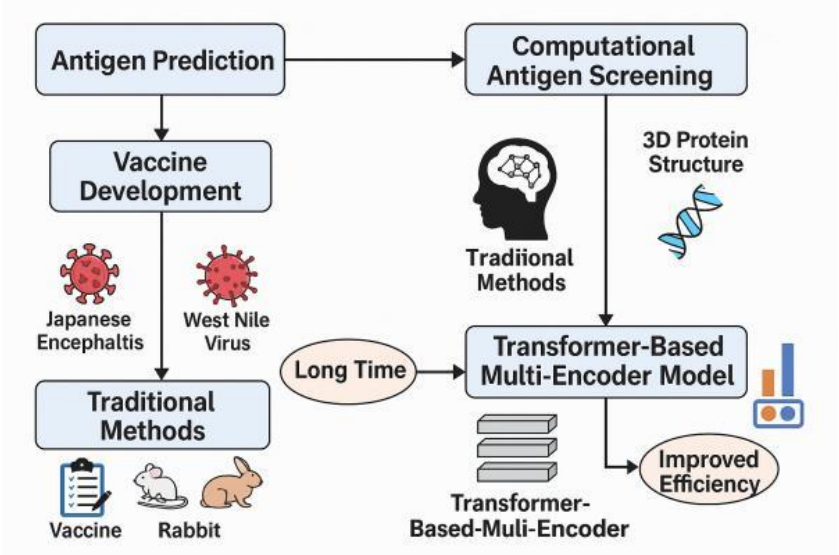

Figure 1 Overall research workflow of the proposed antigen prediction study.

2. Method

2.1. Overview

The overall workflow of the proposed transformer-based multi-encoder antigen prediction framework is illustrated in Fig. 2. We seek to formulate our model of antigenic determinants that makes use of multiple sources of information, that we collectively call heterogeneous. Specifically, these include the data describing one-shot epitopes, data detailing the evolutionary history of

antigens, and complementary structural data such as β-structures for antigenic determinants. (1) physicochemical properties of amino acids, (2) sequential dependencies in primary structure, and (3) three-dimensional spatial relationships among residues. We independently process each feature modality using a single transformer encoder, then fuse them into one, single modality, discriminative representation using a cross-modal attention mechanism for antigenicity prediction.

Given a protein sample P = {a1, a2,..., aN } composed of N amino acids, we extract three distinct types of features: physicochemical properties (Fphy ∈ RN×d1), sequence-based embed-dings (Fseq ∈ RN×d2), and structure-derived descriptors (Fstr ∈ RN×d3). Each of these feature matrices is linearly projected to a shared dimension d, ensuring compatibility across modalities. These projected features are then independently processed using dedicated transformer encoder modules to capture intra-modal representations. Subsequently, the learned representations from each modality are integrated via a Cross-Modal Attention Fusion mechanism, yielding a unified representation (Hfuse). This fused representation is then input to a multilayer perceptron (MLP) for the task of binary antigenicity classification.

2.2. Transformer Encoder

Transformer encoder is the central calculation part of this framework and it provides an effective method of measuring both long and short-range interactions between amino acid residues. Transformer. The self-attention mechanism used in transformers is unlike the conventional recurrent or convolutional model, in that every residue is free to directly draw attention to all the others, which makes this architecture specifically well-placed to model sequence-based as well as structure-informed features in proteins.

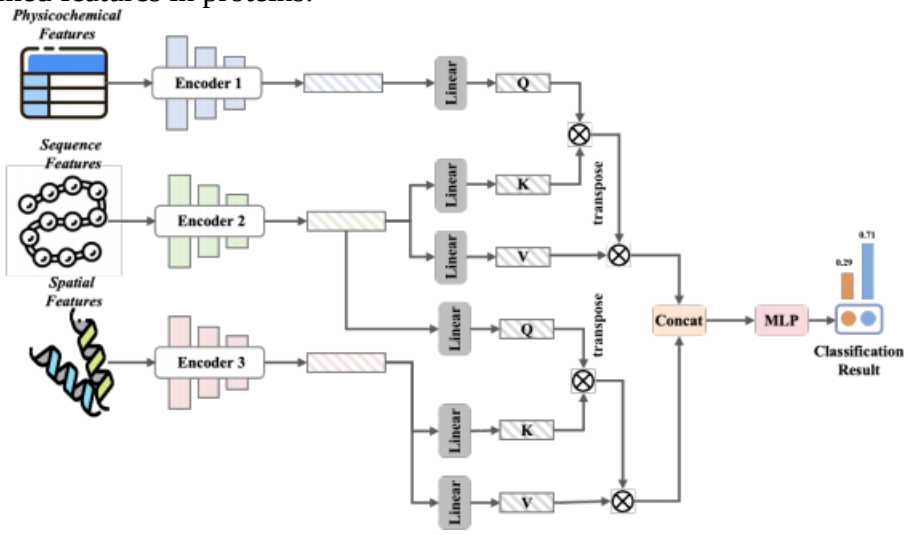

Figure 2 Overall architecture of the proposed transformer-based multi-encoder antigen prediction model.

All three encoder types—physicochemical, sequence, and structural—utilize the same standard transformer encoder design. As depicted in Figure 2, each encoder comprises a stack of L identical layers. Each layer contains two primary components: (1) a Multi-Head Self-Attention (MHSA) mechanism and (2) a Position-wise Feed-Forward Network (FFN). These are inter-leaved with residual connections and layer normalization to ensure effective information flow and stable training.

The three types of encoders, namely, physicochemical, sequence and structural, all use the identical standard transformer encoder design. (shown in Fig. 3) which are identical. Every layer consists of two major items, namely (1) a Multi-Head Self-Attention (MHSA) and (2) a Position-wise Feed-Forward Network (FFN). They are mixed with the residual connections and layer normalization to guarantee the efficient information flow and stable training.

Given an input representation H ∈ RN×d, the self-attention mechanism projects it into three matrices: queries (Q), keys (K), and values (V), each of dimension RN×dk, through learned linear

transformations:

$$Q = HW_Q,\ K = HW_K,\ V = HW_V \tag{1}$$

where $W_Q, W_K, W_V \in R^{d \times d_k}$ are trainable weight matrices. The scaled dot-product attention is then computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2}$$

The attention mechanism allows the model to weigh interactions between residues dynamically, capturing both short-and long-distance relationships critical to antigenicity.

In the multi-head formulation, h independent attention heads operate in parallel, each learning distinct feature subspaces. Their outputs are concatenated and projected back to the original dimensionality:

$$\text{MHSA}(H) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O \tag{3}$$

where $W_O \in R^{hd_k \times d}$ is a learnable projection matrix. This multi-head mechanism enhances model expressiveness and enables the simultaneous capture of multiple interaction patterns among residues.

The second element of every encoder block is a two-layer position-wise feed-forward net-work (FFN) that smooths the contextualized residue representations generated by MHSA. It uses two non-linear activation (usually GELU) linear transformations between them:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \tag{4}$$

where $W_1 \in R^{d \times d_{ff}}$ and $W_2 \in R^{d_{ff} \times d}$ are trainable parameters, and d_{ff} is the hidden dimension. This component enables feature transformation and abstraction, complementing the relational modeling of the self-attention layer.

To stabilize training and facilitate gradient flow, each sublayer is wrapped with residual connections and layer normalization:

$$\tilde{H} = \text{LayerNorm}(H + \text{MHSA}(H)),\quad H' = \text{LayerNorm}(\tilde{H} + \text{FFN}(\tilde{H})) \tag{5}$$

These operations ensure that each encoder layer refines rather than overwrites prior repre-sentations, promoting stable convergence and improved generalization. Dropout layers are also applied to attention weights and feed-forward outputs to mitigate overfitting, especially important for small-scale antigen datasets.

Because the self-attention mechanism itself is permutation-invariant, positional encodings are introduced to retain the order information of residues. We employ sinusoidal positional encoding defined as:

$$\text{PE}(p, 2i) = \sin\left(\frac{p}{10000^{2i/d}}\right),\quad \text{PE}(p, 2i+1) = \cos\left(\frac{p}{10000^{2i/d}}\right) \tag{6}$$

where p denotes the residue position and i the embedding dimension index. This encoding provides continuous and interpretable position information that generalizes to unseen sequence lengths.

After passing through L encoder layers, the output representations H_{phy}, H_{seq}, and H_{str} from each modality encoder encode high-level contextualized information, which is then forwarded to the cross-modal attention fusion module for interaction and integration. This hierarchical encoding process allows the model to jointly learn antigenic determinants from biochemical, sequential, and structural perspectives, providing a robust foundation for accurate antigenicity prediction.

2.3. Multi-Modal Fusion Module

The physicochemical, sequential and spatial features are individually extracted using their respective encoders in this study. Basic concatenation or averaging following independent encoding will not usually suffice to reflect the underlying correlations between the various modalities. To

deal with this we design a multi-modal fusion module based on cross-attention to allow interaction of various modalities dynamically and deeply.

This process does not only combine the complementary information between physicochemical and spatial features, but also permits the sequence representation to dynamically choose useful regions by dynamically choosing the attention-weighted advice.

Specifically, let the physicochemical features be denoted as $F_p \in R^{n_p \times d}$, the sequence features as $F_q \in R^{n_q \times d}$, and the spatial features as $F_s \in R^{n_s \times d}$, where n_p, n_q, n_s represent the lengths of different modalities and d is the feature dimension. During fusion, the model treats the physicochemical features as keys (K) and values (V), while the sequence and spatial features serve as queries (Q) to compute cross-modal dependencies through the self-attention mechanism:

$$Qp = FpWQ(p),\ Ks = FsWK(s),\ Vs = FsWV(s),$$

$$Qt = Fq\ WQ(t),\ Ks = FsWK(s),\ Vs = FsWV(s),$$

$$\text{Attn}_p = \text{softmax}\left(\frac{Q_p K_s^\top}{\sqrt{d_k}}\right) V_s,\ \text{Attn}_t = \text{softmax}\left(\frac{Q_t K_s^\top}{\sqrt{d_k}}\right) V_s. \quad (7)$$

Here, Attn_p represents the attention output of the physicochemical modality over the spatial modality, and Attn_t denotes the attention output of the sequence modality over the spatial modality. Through this mechanism, non-sequential modalities can selectively attend to spatially relevant information embedded in the structural representation, thereby enhancing feature complementarity and avoiding redundancy.

The fused representation is constructed by concatenating the attention-enhanced features with the original sequence features:

$$F_{fusion} = \text{Concat}(\text{Attn}_p, F_q, \text{Attn}_t) \quad (8)$$

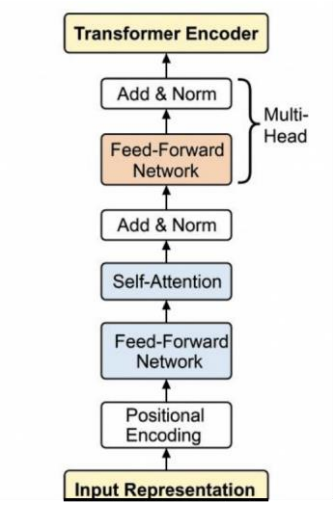

Figure 3 Architecture of the Transformer encoder block.

The fused feature is then projected into a shared latent space and passed through a multilayer perceptron (MLP) for antigenicity prediction:

$$\hat{y} = \sigma(\text{MLP}(F_{fusion})), \quad (9)$$

where $\sigma(\cdot)$ denotes the sigmoid activation function.

This mechanism not only integrates the complementary information between physicochemical and spatial features, but also allows the sequence representation to dynamically select relevant regions under the attention-weighted guidance. Consequently, the proposed fusion module achieves efficient and interpretable multi-modal integration, ensuring cooperative in-formation exchange across modalities and providing a biologically meaningful representation space for antigenicity prediction.

3. Experiment

3.1. Dataset Construction

This research focuses on two major categories of pathogens, bacteria and viruses, to have an effective dataset in training and testing the proposed antigen prediction model(Fig.4). In the dataset of each type of pathogen, the samples are comprised of antigenic proteins (positive samples) and non-antigenic proteins (negative samples). Peer-reviewed literature and the Protegen database [6], which contains only experimentally validated immunogenicity of protein sequences, were used as primary sources of positive samples, however. In order to encourage generalizability of the model to different backgrounds of pathogens, we chose bacteria and viruses representing a broad range of different species.

Culture of negative samples was performed according to a strict filtering plan in order to exclude possible antigenic bias. The randomly selected protein sequences of each pathogen were about 100 entries in the UniProt database [7]. The BLAST tool [8] was used to delete redundant sequences which had more than 90 percent similarity with any positive sample in the present study. Out of the rest of the sequences, only those less than 30 percent similar to any known antigenic protein were selected as final negative samples to exclude the possibility of missing potential immunogenicity.

The primary amino acid sequences provided physicochemical properties, which include molecular weight, isoelectric point (pI), amino acid composition, hydrophobicity, polarity, charge, and aromaticity. We used the toolkit of BioPython [9] (version 2.3.3) and ProPy [10] (version 2.1.2) to compute the following descriptors and normalized them before they could be fed into the model. The above properties refer to immunological properties such as stability, solubility, and electrostatic distributions.

Other biochemical and sequence-based attributes that we considered include the statistics of k-mers (e.g., tri-peptide frequency), position-specific scoring matrices (PSSM), predicted secondary structures (e.g., α-helices and β-sheets), and protein annotations (e.g., Gene Ontology [11] and Pfam domains [12]).

All protein sequences were structurally predicted with the help of the AlphaFold2 application [13] to include spatial data points in them. The resulting PDB files were transformed into 3D coordinate graphs, by which pairwise atomic distance matrices, adjacency matrices, and contact maps were computed to depict the spatial relations between residues. Further structural features such as solvent accessibility, dihedral angles, and relative residue positions were retrieved to obtain the geometrical patterns of antigenic epitopes. These structural features enable the model to identify conformational epitopes that maybe exposed in space but discontinuous in sequence, thus providing complementary information from a different perspective to the sequential and physicochemical representations.

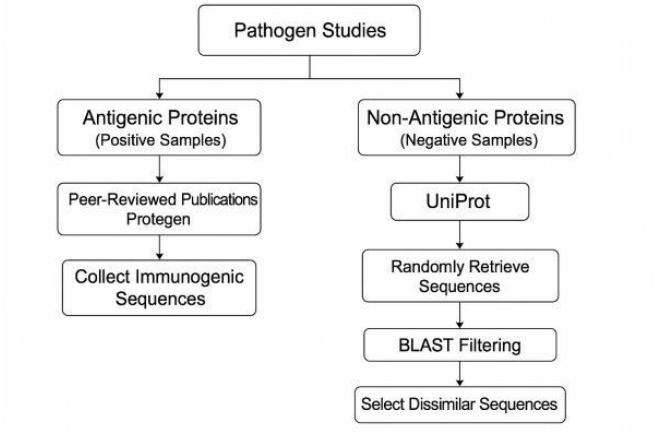

Figure 4 Workflow of dataset construction for antigen prediction.

3.2. Baseline Models

Under this section, to establish the validity of the suggested multi-modal antigen prediction framework, we compared it to some of the commonly used classification models as baselines,

including both the old style (machine learning) and the new style (deep learning). In case of the traditional machine learning, we employed Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Decision Tree (GBDT) as an example. These algorithms have been extensively used in bioinformatics applications and can learn discriminant boundaries in the feature space of lower dimensions, so as to give valid grounds of comparison of the approach suggested. To obtain deep learning baselines, we used an implementation of a Convolutional Neural Network (CNN). to determine the localized feature extraction performance in antigenicity prediction. Besides, to test the potential of the architecture known as the Transformer under single-modality setup, we designed a unimodal Transformer architecture that can perform with sequence features only. The combination of this set-up enables us to determine the disparity between the suggested multi-modal fusion architecture. and its sequence-only analog, thus controlling out the role of cross-modal feature integration. Systematic comparison across these models of the baselines, we critically analyze the benefits and enhancements that our approach has made to antigenicity prediction activities.

3.3. Experimental Results

To obtain a comprehensive vision on the predictive performance of proposed multi-modal antigen prediction framework, we conducted comparative experiments against traditional ma-chine learning and deep learning baseline models, including Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), Convolutional Neural Network (CNN), and the unimodal Transformer model based solely on sequence features. All models were trained and evaluated under identical experimental settings using five-fold cross-validation.

Table 1 Performance comparison between the proposed method and baseline models on the antigenicity prediction task. The best results are highlighted in bold.

Model	Accuracy	Precision	Recall	F1-score	AUC
SVM	0.812	0.796	0.781	0.788	0.842
RF	0.825	0.809	0.794	0.801	0.856
GBDT	0.837	0.821	0.806	0.813	0.865
CNN	0.851	0.839	0.824	0.831	0.877
Unimodal Transformer	0.868	0.854	0.839	0.846	0.890
ours	**0.910**	**0.896**	**0.902**	**0.899**	**0.937**

As shown in Table 1, the proposed Multi-Modal Transformer significantly outperforms all baseline models across all evaluation metrics. Compared with the unimodal Transformer, our method achieves a 4.2% improvement in F1-score and a 4.7% increase in AUC, that the integration of multi-modal features (physicochemical, sequential and structural) can significantly improve the discriminative capabilities of the model. Traditional machine learning methods (SVM, RF, and GBDT) achieve reasonable performance, they, however, are highly dependent on handcrafted feature vectors and thus cannot learn complex and high-level semantics of biological sequences.

Furthermore, CNN and unimodal Transformer models show better adaptability to sequence data; however, they are unable to integrate spatial or biochemical relationships, that are vital in the correct prediction of conforming epitopes. In contrast, our proposed multi-modal attention fusion mechanism allows effective interaction across modalities, thereby capturing deeper biochemical–structural dependencies and improving both recall and precision.

These results clearly demonstrate that integrating complementary modalities provides a more biologically informative representation of antigenic determinants. We conclude that proposed method can better generalize on both bacterial and viral datasets. Further experiments are presented to evaluate contributions from each modality and to validate the ability of the proposed fusion method, respectively. We conducted several ablation tests removing different components from the complete model in a principled manner.

3.4. Ablation Study

In order to further evaluate the contribution of various modalities and verify the success-fulness

of the suggested fusion technique, we conducted a set of ablation experiments. This was a systematic deprivation of main elements of the entire model. First, we examined the impact of drop-out of structural features (w/o Structural Features). Here the model was trained on the basis of physicochemical and sequence data only. The second test involved an option that had removed the physicochemical features (w/o Physicochemical Features) but retained the sequence and spatial representations. Lastly, we looked at the case when we eliminated the cross-attention fusion mechanism (w/o Cross-Attention Fusion). In this case, the model substituted the adaptive attention-based fusion with a simple concatenation of features operation. The entire implementation of the suggested multi-modelled scheme is referred to as the Full Model (ours).

Table 2 Ablation results of the proposed multi-modal transformer on the antigenicity prediction dataset. The best performance is highlighted in bold.

Model Variant	Accuracy	Precision	Recall	F1-score	AUC
w/o Structural Features	0.874	0.861	0.853	0.857	0.893
w/o Physicochemical Features	0.882	0.867	0.861	0.864	0.902
w/o Cross-Attention Fusion	0.895	0.881	0.876	0.878	0.915
Full Model (ours)	**0.910**	**0.896**	**0.902**	**0.899**	**0.937**

Table 2 of this paper illustrates that, the eradication of any modality or the mechanism of fusion leads to an apparent deterioration of the performance, proving that all sources of features supply complementary information to the model. The omission of structural features leads to the greatest decrease in F1score, which implies that spatial conformation data is important in antigenic epitopes recognition. On the same note, the elimination of physicochemical characteristics results in a significant decrease in AUC, which indicates that biochemical descriptors, including hydrophobicity and polarity, complement successfully the representation acquired on structural and sequential inputs. This also reduces the overall accuracy when the cross-attention fusion is replaced by direct concatenation. it shows the need to have adaptive feature alignment in the learning of inter-modal dependencies. These findings affirm that, the suggested multi-modal attention fusion mechanism allows efficient and interpretable multi-modal integration, resulting in a better antigenicity prediction performance.

4. Conclusion

This paper suggested that a multi-modal predictive antigenicity model can be developed using a transformer. This model is a reliable way to combine physicochemical, sequential, and structural data of proteins with a cross-attention fusion mechanism. Our model provides long-range relationships and interactions among data types of different types as opposed to traditional methods which only generate features based on sequences or manually constructed ones. It provides a biologically understandable and informationally informed vaccine target screening solution.

Extensive tests of bacterial and viral antigen datasets demonstrate that our approach is much more effective than such classical machine learning systems as SVM, RF, and GBDT, and uni-modal deep learning models as CNN and Transformer. Through multi-modal fusion, our model is able to learn complementary information across types of data, and results in steady increases in accuracy, F1-score, and AUC. The significance of the use of 3D structural information is supported by ablation studies. The importance of each of the learning components, as well as the fusion of cross-attention, is important in terms of antigenicity recognition. In addition to predictability, our model offers us the knowledge about the physicochemical and spatial contributions to the immunogenicity. This can inform vaccine design and epitope discovery.

We will use this framework in further work to model even more complex antigen antimi-crobial interactions. Our long-term goal is also to use pre-trained protein language models and graph-based encoders to enhance cross-species generalization.

References

[1] E. Ong, M. Wong, A. Huffman, and Y. He, “Vaxign-ml: supervised machine learn-ing reverse vaccinology model for improved prediction of bacterial protective antigens,” Bioinformatics, vol. 36, no. 11, pp. 3185–3191, 2020.

[2] L. Bravi, “Machine learning and reverse vaccinology: accelerating vaccine target discov-ery,” Frontiers in Immunology, vol. 15, p. 1523078, 2024.

[3] B. Mugunthan et al., “Integrating reverse vaccinology and machine learning for the pre-diction of multi-epitope vaccines in mycoplasma gallisepticum,” Computational Biology and Chemistry, vol. 105, p. 107862, 2023.

[4] R. Bhattacharya, P. Singh, and M. Kaur, “Deep learning-assisted epitope recognition and vaccine construction: a systematic review,” Briefings in Bioinformatics, 2025.

[5] K. Rawal,P. Khandelia, and V. Jaiswal, “Vaxi-dl: deep learning-based multi-feature fusion framework for antigenicity prediction and vaccine design,” Scientific Reports, vol. 11, no. 1, pp. 1–12, 2021.

[6] B. Yang, S. Sayers, Z. Xiang, and Y. He, “Protegen: a web-based protective antigen database and analysis system,” Nucleic Acids Research, vol. 39, no. suppl_1, pp. D1073– D1078, 2011.

[7] T. U. Consortium, “Uniprot: the universal protein knowledgebase in 2021,” Nucleic Acids Research, vol. 49, no. D1, pp. D480–D489, 2021.

[8] S. F. Altschul,W. Gish,W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” Journal of Molecular Biology, vol. 215, no. 3, pp. 403–410, 1990.

[9] P. J. Cock, T. Antao, J. T. Chang et al., “Biopython: freely available python tools for computational molecular biology and bioinformatics,” Bioinformatics, vol. 25, no. 11,pp. 1422–1423, 2009.

[10] D. Cao, Q. Xu, Y. Liang, Q. Zhang, and Y. Hao, “Propy: a tool to generate various modes of chou’s pseaac,” Bioinformatics, vol. 29, no. 7, pp. 960–962, 2013.

[11] M. Ashburner et al., “Gene ontology: tool for the unification of biology,” Nature Genetics, vol. 25, no. 1, pp. 25–29, 2000.

[12] R. D. Finn, A. Bateman, J. Clements et al., “Pfam: the protein families database,” Nucleic Acids Research, vol. 42, no. D1, pp. D222–D230, 2014.

[13] J. Jumper, R. Evans, A. Pritzel et al., “Highly accurate protein structure prediction with alphafold,” Nature, vol. 596, no. 7873, pp. 583–589, 2021.