LlamaAffinity: A Predictive Antibody-Antigen Binding Model Integrating Antibody Sequences with Llama3 Backbone Architecture

Delower Hossain[1,4], Ehsan Saghapour[2,4], Kevin Song[3,4], Jake Y. Chen[1,2,4*]

[1] Department of Computer Science, The University of Alabama at Birmingham, Birmingham, AL, US.
[2] Department of Biomedical Informatics and Data Science, School of Medicine, The University of Alabama at Birmingham, Birmingham, AL, US.
[3] Department of Biomedical Engineering, The University of Alabama at Birmingham, Birmingham, AL, US
[4] Systems Pharmacology AI Research Center, The University of Alabama at Birmingham, AL, US.

*corresponding author
jakechen@uab.edu

Keywords: Antibody-Antigen Binding Affinity; Therapeutic Antibody; LLM, Deep Learning, Protein design and AI

Abstract- Antibody-facilitated immune responses are central to the body's defense against pathogens, viruses, and other foreign invaders. The ability of antibodies to specifically bind and neutralize antigens is vital for maintaining immunity. Over the past few decades, advancements in bioengineering have significantly accelerated the development of therapeutic antibodies. These antibody-derived drugs have shown remarkable efficacy, particularly in treating Cancer, SARS-CoV-2, autoimmune disorders, infectious diseases, and other diseases. The complexity and cost of wet-lab affinity measurements remain significant obstacles to the rapid discovery of antibodies. Although recent AI-driven approaches have shown promising efficacy, these challenges persist, reinforcing the need for scalable and highly accurate computational models. Advances in Artificial Intelligence have significantly transformed in silico biomedical research. In this article, we introduce LlamaAffinity, an advanced binding affinity prediction model built on the open-source Llama 3 backbone that was trained using antibody sequence data from the Observed Antibody Space (OAS) database. The proposed approach significantly improved over existing state-of-the-art (SOTA) approaches (AntiFormer, AntiBERTa, AntiBERTy) across multiple evaluation metrics. Specifically, the model achieved an accuracy of 0.9640, an F1-score of 0.9643, a precision of 0.9702, a recall of 0.9586, and an AUC-ROC of 0.9936. Moreover, this strategy demonstrated higher computational efficiency, with a five-fold cumulative training time of only 0.46 hours, which is significantly lower than that of previous studies. LlamaAffinity defines a new benchmark for antibody-antigen binding affinity prediction, achieving advanced performance in the fields of immunotherapies and immunoinformatics. Furthermore, it can effectively assess binding affinities following novel antibody design, accelerating the discovery and optimization of therapeutic candidates.

1 Introduction

Antibodies are Y-shaped proteins produced by the immune system to detect and neutralize harmful foreign substances, known as antigens, such as bacteria and viruses. Antibodies are categorized into five primary classes: IgG (Immunoglobulin G), IgA (Immunoglobulin A), IgM, IgE, and IgD, each serving unique functions within the immune response. IgG is the most abundant and commonly used in therapeutics, while IgM is the first responder to infection. Therapeutic antibodies [1][2] have revolutionized treatment in oncology (breast, lung, and bladder cancers), autoimmune diseases, and infectious diseases. They offer high specificity with fewer side effects compared to traditional drugs. Recent trends include bispecific antibodies, antibody-drug conjugates (ADCs) [3] [4], and nanobodies. The FDA has approved at least 13 antibody drug conjugates (ADCs), including treatments for triple-negative metastatic breast cancer (MBC) and the HR-positive, HER2-negative subtypes, as reported by AXIS Pharma. By 2025, three additional ADCs developed by AstraZeneca, Daiichi Sankyo, and AbbVie are anticipated to receive FDA approval (Biopharma PEG).

Therapeutic antibodies are among the top-performing biotherapeutics, with four ranking among the top ten best-selling drugs in 2021 [5]. The global antibody drug conjugate (ADC) market was valued at $7.35 billion in 2022 and is projected to surpass $28 billion by 2028, reflecting substantial growth (Biopharma portal report). AI-driven antibody design [6][7][8][9][10] can significantly accelerate discovery and development, overcoming the time, labor, and cost limitations of traditional methods. Specifically, studies identified the following foundational antibody design models AntiBERTa [11], AntiBERTy [12], IgFold [13], AbLang [14], AbGPT [15]. AbbVie and BigHat Biosciences launched a $355 million collaboration on December 5, 2023, to leverage BigHat's AI-driven platform for next-generation oncology and neuroscience antibody therapeutics [2]. Additionally, in December 2023, AstraZeneca and AbbVie each signed deals exceeding $200 million to collaborate with Absci and BigHat Biosciences, respectively. Both partnerships aim to leverage AI-driven antibody design platforms to accelerate the development of next-generation therapeutics [2].

In novel antibody development, an antibody's effectiveness largely depends on its interaction with the target antigen, whereas binding affinity is a key indicator of the strength of this interaction. Higher binding affinity generally correlates with greater therapeutic success, making it a critical focus in antibody engineering. Despite intensive research on antibody affinity systems leveraging AI, studies have shown that current approaches, such as AntiFormer [16], MVSF-AB [17], AbAgIntPre [18], AttABseq [19], and CSM-AB [20], face limitations in performance and generalizability. While Generative AI (GenAI) holds significant promise, no dedicated studies have yet explored its potential for antibody‑antigen binding affinity prediction utilizing leading large language model (LLM) backbone architectures.

In this article, we propose LlamaAffinity, a novel predictive model built on the LLaMA 3 backbone [21] architecture, which integrates antibody sequence data based on the Observed Antibody Space (OAS) dataset [22]. Our approach outperformed the prior state-of-the-art (SOTA) method (AntiFormer) [16] across multiple evaluation metrics. Specifically, the model achieved an accuracy of 0.9640, an F1-score of 0.9643, a precision of 0.9702, a recall of 0.9586, and an AUC-ROC of 0.9936 for the affinity prediction task. Additionally, the proposed LlamaAffinity approach offers improved computational efficiency compared to existing strategies.

2 Data and Methods

The section serves as a procedure for building the LlamaAffinity model to classify antibody binding affinity. It involves data curation and preparation, Llama backbone architecture, training, and the model performance evaluation phase.

2.1 *Dataset*

We employed the Observed Antibody Space (OAS) dataset [22], which was curated from the official AntiFormer GitHub repository (link). Since AntiFormer is the current state-of-the-art (SOTA) model in this domain, we applied the same dataset for a fair comparison. This cohort comes pre-tokenized using the ProtBERT tokenizer (BertTokenizer) from the transformer library, with a vocabulary size of 30. It includes key attributes such as input_ids, attention_mask, and token_type_ids, representing tokenized antibody sequences. Each sample is labeled with antigen-binding affinity: low affinity (label 0) and binder (label 1). The antibody sequences comprise both heavy and light chains concatenated into a single sequence. The dataset was split into five folds using StratifiedKFold cross-validation to evaluate model performance.

2.2 *Model Architecture*

Large Language Models (LLMs) have transformed and redefined the modern era of artificial intelligence. The proposed model, LlamaAffinity, is developed using the LLaMA-3 [21] backbone architecture, which is recognized as a spectacular LLM. It takes as input token IDs, padding masks, and a batch size specification of 4. Additionally, the model configuration includes four transformer layers followed by a GlobalAveragePooling layer and fully connected dense layers with softmax activation (Equation 2). The training was conducted using the Adam optimizer with a learning rate of 0.0001; the loss function employed was Sparse Categorical Cross-entropy, as shown in Equation 1. We optimized the hyperparameter tuning based on our previous simulations of drug and protein design with the same approach. However, additional key hyperparameters are summarized in Table 1. It's important to note that this custom training (removing previous massive irrelevant default pretrained weights) utilizing the backbone of Llama3 provides that outstanding performance.

Table 1. Key hyperparameters and their values used in the LlamaAffinity model architecture.

Parameter	Value
num_layers	4
num_query_heads	12
hidden_dim	384
intermediate_dim	192
vocabulary_size	30
num_key_value_heads	12
rope_max_wavelength	100,000
rope_scaling_factor	1
layer_norm_epsilon	1e-6
dropout	0.1

Equation 1: Sparse Categorical Cross-Entropy

$$\text{Loss} = -[\text{y} \cdot \log(\text{p}) + (1 - \text{y}) \cdot \log(1 - \text{p})] \quad (1)$$

Equation 2: Softmax Activation

$$\text{softmax}(z_i) = \frac{e^{z_i - \max(\boldsymbol{z})}}{\sum_{j=1}^{C} e^{z_j - \max(\boldsymbol{z})}} \quad (2)$$

2.3 *Training Phase*

We conducted the simulation with five-fold cross-validation to evaluate model robustness and generalization using a StratifiedKFold split. Notably, AntiFormer conducted the same cross-validation approach. The training was accomplished on Colab GPU, using over 10 epochs against each fold. The total training time across all five folds was approximately 27.48 minutes, averaging about 5.496 minutes per fold, which emphasizes the model's scalability and generalizability for large-scale applications.

2.4 *Model Evaluation*

The LlamaAffinity performance was evaluated using the following metrics: accuracy, F1-score, precision, recall, and ROC AUC. Additionally, confusion matrix analysis revealed strong and accurate positive rates, along with low misclassification rates, with only 3.04% predicted false positives and 4.14% predicted false negatives (Figure 2).

3 **Results**

This section reveals the performance of the proposed LlamaAffinity model. The results exhibit the model's effectiveness and generalizability. To illustrate, Table 2 represents the outcomes of 5-fold cross-validation, while Table 3 compares the model's performance with several state-of-the-art (SOTA) approaches. Specifically, Table 2 presents the consistent performance of LlamaAffinity across all five folds, with an average accuracy of 0.9640, F1-score of 0.9643, precision of 0.9702, recall of 0.9586, and an exceptional ROC AUC of 0.9936. The top accuracy of 0.9725 was achieved by both Fold 3 and Fold 4. Fold 3 also recorded the highest ROC AUC (0.9969) (Figure 1), while Fold 4 led in F1-score and recall. Conversely, Fold 2 attained the highest precision (0.9847). Notably, the minimal variation in scores across all folds underscores the model's robustness. However, in terms of efficiency, the total training time for all five folds was approximately 27 minutes and 47 seconds, with each fold averaging around 5 minutes and 49 seconds, demonstrating the model's scalability for large-scale applications.

Table 2: Cross-validation results of the LlamaAffinity model across five folds.

Fold	Accuracy	F1_score	Precision	Recall	ROC AUC	Training (Minutes)
0	0.9550	0.9548	0.9694	0.9406	0.9886	5.9600
1	0.9525	0.9533	0.9510	0.9557	0.9913	4.9559
2	0.9675	0.9674	0.9847	0.9507	0.9961	5.9042
3	0.9725	0.9728	0.9752	0.9704	0.9969	5.4559
4	0.9725	0.9730	0.9706	0.9754	0.9951	5.2030
Average	0.9640	0.9643	0.9702	0.9586	0.9936	5.496

Table 3 further displays the performance comparison for LlamaAffinity, which yields the highest average accuracy (0.9640), F1-score (0.9643), and ROC AUC (0.9936) among all evaluated models. It outperformed AntiFormer (ROC AUC: 0.9660) and AntiBERTa (ROC AUC: 0.9340) while requiring significantly less training time (0.46 hours vs. 0.76 hours and 2.97 hours, respectively). In contrast, baseline models such as the 6-layer Transformer showed considerably lower accuracy (0.7865) and ROC AUC (0.7930), highlighting the strength of the Llama3 backbone in capturing intricate antibody-antigen interactions.

Additionally, the confusion matrix provides evidence of LlamaAffinity's robustness, correctly classifying 96.96% of low-affinity samples and 95.86% of binder samples (Figure 2). Misclassification rates remained low, with only 3.04% false positives and 4.14% false negatives, reflecting a strong balance between precision and recall (Figure 2).

Table 3: Comparison of antibody affinity prediction models across performance metrics and training time.

Model	Accuracy	F1-Score	Precision	Recall	ROC AUC	Training (hours)
Transformer-6 L	0.7865	0.7590	0.8060	0.7990	0.7930	0.38
Transformer-12 L	0.8011	0.7890	0.8310	0.8180	0.8290	0.63
AntiBERTy	0.8321	0.8510	0.9110	0.8910	0.9400	1.46
AntiBERTa	0.8796	0.8570	0.9080	0.9090	0.9340	2.97
AntiFormer	0.9169	0.8820	0.9630	0.9250	0.9660	0.76
LlamaAffinity	**0.9640**	**0.9643**	**0.9702**	**0.9586**	**0.9936**	**0.46**

Note: SOTA model scores obtained from the AntiFormer article.

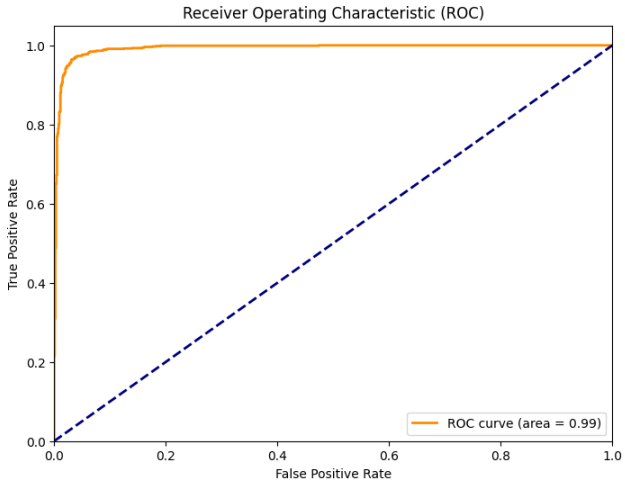

Figure 1. LlamaAffinity ROCAUC curve

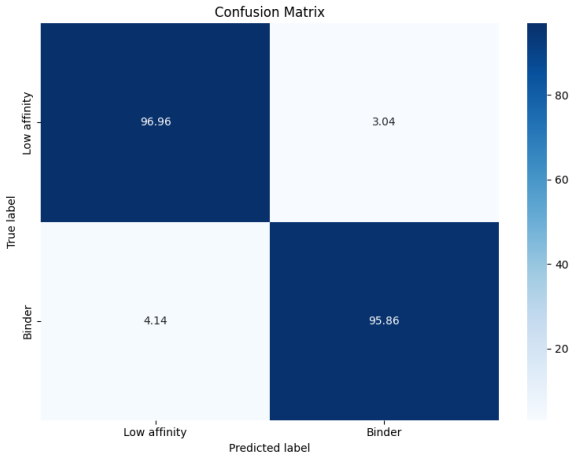

Figure 2. LlamaAffinity Confusion Matrix

4 Conclusion

This study presents LlamaAffinity, a novel model for predicting antibody-antigen binding affinity, built upon the Llama 3 architecture and antibody sequence inputs. The performance is benchmarked against several state-of-the-art models, including AntiBERTa and AntiFormer. LlamaAffinity achieved the highest performance across all evaluation metrics, with an accuracy of 0.9640, an F1-score of 0.9643, and an AUC-ROC of 0.9936, while maintaining a relatively low training time. The findings prove effectiveness and high computational efficiency. The proposed model enhances current prediction capabilities and provides practical utility for evaluating as binders for downstream novel antibody design pipelines, marking a significant step forward in immunoinformatics and protein engineering research. For future work, conducting case studies would be an ideal way to validate the practical applicability of LlamaAffinity in real-world scenarios. For instance, evaluating its performance in predicting high-affinity binders for targets such as the SARS-CoV-2 spike protein or HER2 in breast cancer would provide valuable insights upon training with more diverse instances of antibody-antigen pairs.

Conflict of interest

The authors expressed no potential conflicts of interest.

Funding (optional)

This work was partially supported by the National Institutes of Health (NIH) under grant number UMITR004771-01.

Availability of data and software code (optional)

Our code is available at the following URL: GitHub

References

[1] X. Lyu et al., "The global landscape of approved antibody therapies," Antibody Therapeutics, vol. 5, no. 4, pp. 233–257, Sep. 2022, doi: 10.1093/abt/tbac021.

[2] B. Dealmakers, "Antibody design enters the AI era," Biopharma Dealmakers, May 2024, doi: 10.1038/d43747-024-00030-w.

[3] P. Gogia, H. Ashraf, S. Bhasin, and Y. Xu, "Antibody–Drug Conjugates: A review of approved drugs and their clinical level of evidence," Cancers, vol. 15, no. 15, p. 3886, Jul. 2023, doi: 10.3390/cancers15153886.

[4] K. Liu et al., "A review of the clinical efficacy of FDA-approved antibody–drug conjugates in human cancers," Molecular Cancer, vol. 23, no. 1, Mar. 2024, doi: 10.1186/s12943-024-01963-7.

[5] Kim J, McFee M, Fang Q, Abdin O, Kim PM. "Computational and artificial intelligence-based methods for antibody development," Trends in pharmacological sciences. 2023 Mar 1;44(3):175-89.

[6] R.-M. Lu et al., "Development of therapeutic antibodies for the treatment of diseases," Journal of Biomedical Science, vol. 27, no. 1, Jan. 2020, doi: 10.1186/s12929-019-0592-z.

[7] L. Li et al., "Machine learning optimization of candidate antibody yields highly diverse sub-nanomolar affinity antibody libraries," Nature Communications, vol. 14, no. 1, Jun. 2023, doi: 10.1038/s41467-023-39022-2.

[8] J. Zheng, Y. Wang, Q. Liang, L. Cui, and L. Wang, "The application of machine learning on antibody discovery and optimization," Molecules, vol. 29, no. 24, p. 5923, Dec. 2024, doi: 10.3390/molecules29245923.

[9] J. Cheng, T. Liang, X.-Q. Xie, Z. Feng, and L. Meng, "A new era of antibody discovery: an in-depth review of AI-driven approaches," Drug Discovery Today, vol. 29, no. 6, p. 103984, Apr. 2024, doi: 10.1016/j.drudis.2024.103984.

[10] V. Dewaker, V. K. Morya, Y. H. Kim, S. T. Park, H. S. Kim, and Y. H. Koh, "Revolutionizing oncology: the role of Artificial Intelligence (AI) as an antibody design, and optimization tools," Biomarker Research, vol. 13, no. 1, Mar. 2025, doi: 10.1186/s40364-025-00764-4.

[11] J. Leem, L. S. Mitchell, J. H. R. Farmery, J. Barton, and J. D. Galson, "Deciphering the language of antibodies using self-supervised learning," Patterns, vol. 3, no. 7, p. 100513, May 2022, doi: 10.1016/j.patter.2022.100513.

[12] J. A. Ruffolo, J. J. Gray, and J. Sulam, "Deciphering antibody affinity maturation with language models and weakly supervised learning," arXiv (Cornell University), Jan. 2021, doi: 10.48550/arxiv.2112.07782.

[13] J. A. Ruffolo, L.-S. Chu, S. P. Mahajan, and J. J. Gray, "Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies," Nature Communications, vol. 14, no. 1, Apr. 2023, doi: 10.1038/s41467-023-38063-x.

[14] T. H. Olsen, I. H. Moal, and C. M. Deane, "AbLang: an antibody language model for completing antibody sequences," Bioinformatics Advances, vol. 2, no. 1, Jan. 2022, doi: 10.1093/bioadv/vbac046.

[15] D. Kuan and A. B. Farimani, "ABGPT: De novo antibody design via Generative Language Modeling," arXiv (Cornell University), Sep. 2024, doi: 10.48550/arxiv.2409.06090.

[16] Q. Wang et al., "AntiFormer: graph enhanced large language model for binding affinity prediction," Briefings in Bioinformatics, vol. 25, no. 5, Jul. 2024, doi: 10.1093/bib/bbae403.

[17] M. Li et al., "MVSF-AB: Accurate antibody-antigen binding affinity prediction via multi-view sequence feature learning," Bioinformatics, Oct. 2024, doi: 10.1093/bioinformatics/btae579.

[18] Y. Huang, Z. Zhang, and Y. Zhou, "AbAgIntPre: A deep learning method for predicting antibody-antigen interactions based on sequence information," Frontiers in Immunology, vol. 13, Dec. 2022, doi: 10.3389/fimmu.2022.1053617.

[19] R. Jin et al., "AttABseq: an attention-based deep learning prediction method for antigen–antibody binding affinity changes based on protein sequences," Briefings in Bioinformatics, vol. 25, no. 4, May 2024, doi: 10.1093/bib/bbae304.

[20] Y. Myung, D. E. V. Pires, and D. B. Ascher, "CSM-AB: graph-based antibody–antigen binding affinity prediction and docking scoring function," Bioinformatics, vol. 38, no. 4, pp. 1141–1143, Nov. 2021, doi: 10.1093/bioinformatics/btab762.

[21] M. Ettaleb, M. Kamel, V. Moriceau, and N. Aussenac-Gilles, "The Llama 3 herd of models," Arxiv, Nov. 2024, doi: 10.48550/arxiv.2407.21783.

[22] T. H. Olsen, F. Boyles, and C. M. Deane, "Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences," Protein Science, vol. 31, no. 1, pp. 141–146, Oct. 2021, doi: 10.1002/pro.4205.