DeepPROTECTNeo: A Deep learning-based Personalized and RV-guided Optimization tool for TCR Epitope interaction using Context-aware Transformers

Debraj Das[a,1], Soumyadeep Bhaduri[a,1], Avik Pramanick[b], Pralay Mitra[b*]

[a] Centre for Computational and Data Sciences, Indian Institute of Technology Kharagpur, West Bengal, India-721302

[b] Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, West Bengal, India-721302

[1]Authors have equal contributions

*Correspondence to
PralayMitra, Ph.D.
Department of Computer Science and Engineering,
Indian Institute of Technology Kharagpur
West Bengal - 721302, India
Email: pralay@cse.iitkgp.ac.in
Phone: +91-3222-282344
ORCID ID: 0000-0003-4119-3788

Highlights

- DeepPROTECTNeo is an all-in-one platform for neo-epitope identification from raw sequencing reads
- Incorporate a novel transformer-based architecture for peptide-TCR binding affinity prediction and outperforming current methods
- Accepts flexible input formats from users to provide comprehensive analysis and has a separate module for predicting peptide-TCR binding
- Successfully identified 24 true neoepitopes among 37 reported in TESLA using only SNVs, indels

Abstract

The development of personalized cancer vaccines relies heavily on accurately identifying neoepitopes capable of eliciting strong immune responses. Traditional methods primarily focus on predicting interactions between major histocompatibility complex (MHC) molecules and peptides, often neglecting the crucial step of T-cell receptor (TCR) binding in the context of anchor motif and hydrophobicity, which are critical variables in the elicitation of robust tumor-specific responses. T cells perpetually monitor lymphoid and peripheral organs for antigens, such as peptides or lipids, that are presented by peptide-MHC (pMHC) molecules in other cells. This oversight limits the efficacy of these vaccines in triggering a robust adaptive immune response. We introduce a stand-alone web-server DeepPROTECTNeo (https://cosmos.iitkgp.ac.in/DeepPROTECTNeo/) that processes Whole Exome Sequencing (WES) or Whole Genome Sequencing (WGS) data and mitigate the process of extracting patient-specific neoantigen candidate leveraging genomic and variational TCR information directly, a pivotal component in the development of adaptive immunity and personalized cancer vaccines. Along with traditional MHC binding and B-cell receptor (BCR) recognition, we introduce a context-aware transformer block leveraging cross-attention mechanisms to enhance the analysis of TCR-epitope interactions, thereby advancing neoepitope discovery with unparalleled precision and facilitating reverse vaccinology (RV)-guided multi-epitope vaccine design. With extensive hyperparameter tuning, our model shows exceptional generalizability in predicting TCR-peptide binding on independent test data with AUC 0.7234 and AUPRC 0.7659, substantially outperforming state-of-the-art methods. Furthermore, our pipeline successfully identified 372 neoepitopes out of 532 unique neoepitopes presented in the TESLA benchmark

dataset, with 24 true neoantigens, demonstrating its high efficacy and precision in neoepitope discovery.

Keywords

Neoantigen prediction, Whole Exome/Whole Genome data, Transformer, Deep Learning, Vaccine Research, Comprehensive web-server

1. Introduction

The steady progress in cancer immunotherapy has expedited the formulation of personalized treatment approaches, especially via the identification of tumor-specific neoantigens resulting from somatic mutations that are specific to cancer cells and not present in normal tissues. Neoantigens, unlike tumor-associated antigens (TAAs), are unique to cancer cells and do not occur in healthy tissues[1]. This specificity renders them highly promising targets for personalized therapies, such as vaccines and adoptive T-cell therapies. Naïve single positive mature T cells leave the thymus and travel to peripheral lymphoid organs, including the spleen and lymph nodes. Then T cells interact with foreign peptides displayed by the MHC molecules of antigen-presenting cells (APCs), such as macrophages, dendritic cells, and B cells, in the context of pathogenic infection[2]. The intricate mechanisms of recognition and presentation establish a distinct specificity that enables accurate immune recognition while preventing tolerance, thereby promoting strong and enduring anti-tumor responses. Next-generation sequencing (NGS) technologies have markedly improved the identification of neoantigens by facilitating genome-wide mutation detection. Accurately predicting which neoantigens will elicit effective immune responses is challenging due to the multiple steps involved, including the identification of human leukocyte antigen (HLA) types, prediction of peptide-MHC binding, and evaluation of T-cell receptor (TCR) interactions[3]. Current computational tools predominantly emphasize MHC binding affinity, neglecting vital downstream interactions that are critical for immune activation. Furthermore, numerous tools necessitate various input files and separate workflows for HLA typing and TCR profiling, thereby constraining clinical applicability.

Here, we present DeepPROTECTNeo, a standalone pipeline leveraging Deep Learning (DL) techniques designed for the prediction and evaluation of neoantigens directly from WES/WGS

data. DeepPROTECTNeo offers a comprehensive workflow by conducting HLA typing, identifying TCR complementary-determining region beta (CDR3β), predicting peptide-MHC binding, and assessing TCR-epitope interactions, all from a single input format. This approach consolidates tools and data formats, thereby simplifying the workflow for researchers and clinicians. DeepPROTECTNeo demonstrates significant advancement for rational vaccine design through the integration of transformers[4], a feature extraction component designed to capture amino acid-level characteristics and physicochemical properties to predict the interaction of epitope-CDR3β region. In addition to T-cell binding predictions, our extensive pipeline offers B-cell epitope scoring and immunogenicity evaluations for projected CDR3β-epitope pairs, facilitating the design of multi-epitope vaccines that integrate T-cell and B-cell responses, thereby contributing to a robust personalized immunotherapy strategy. Extensive evaluation on clinical database like TESLA[5], DeepPROTECTNeo excels in the identification of validated neoantigens, exhibiting high sensitivity (identified 24 out of 37 true neoepitopes) and minimal false-negative rates. Leveraging advanced DL techniques and building on RV principles, our pipeline exhibits a robust system for neoantigen prediction, poised to multi-epitope vaccine design and significantly enhance patient-specific cancer immunotherapy.

2. Methods

We describe the key components of the DeepPROTECTNeo pipeline, beginning with data preprocessing, followed by variant calling and annotation followed by variant calling, variant annotations and neoepitope prediction in this section (Figure 1). Each step is designed to streamline neoantigen identification and enhance immunogenicity evaluation for personalized cancer immunotherapy.

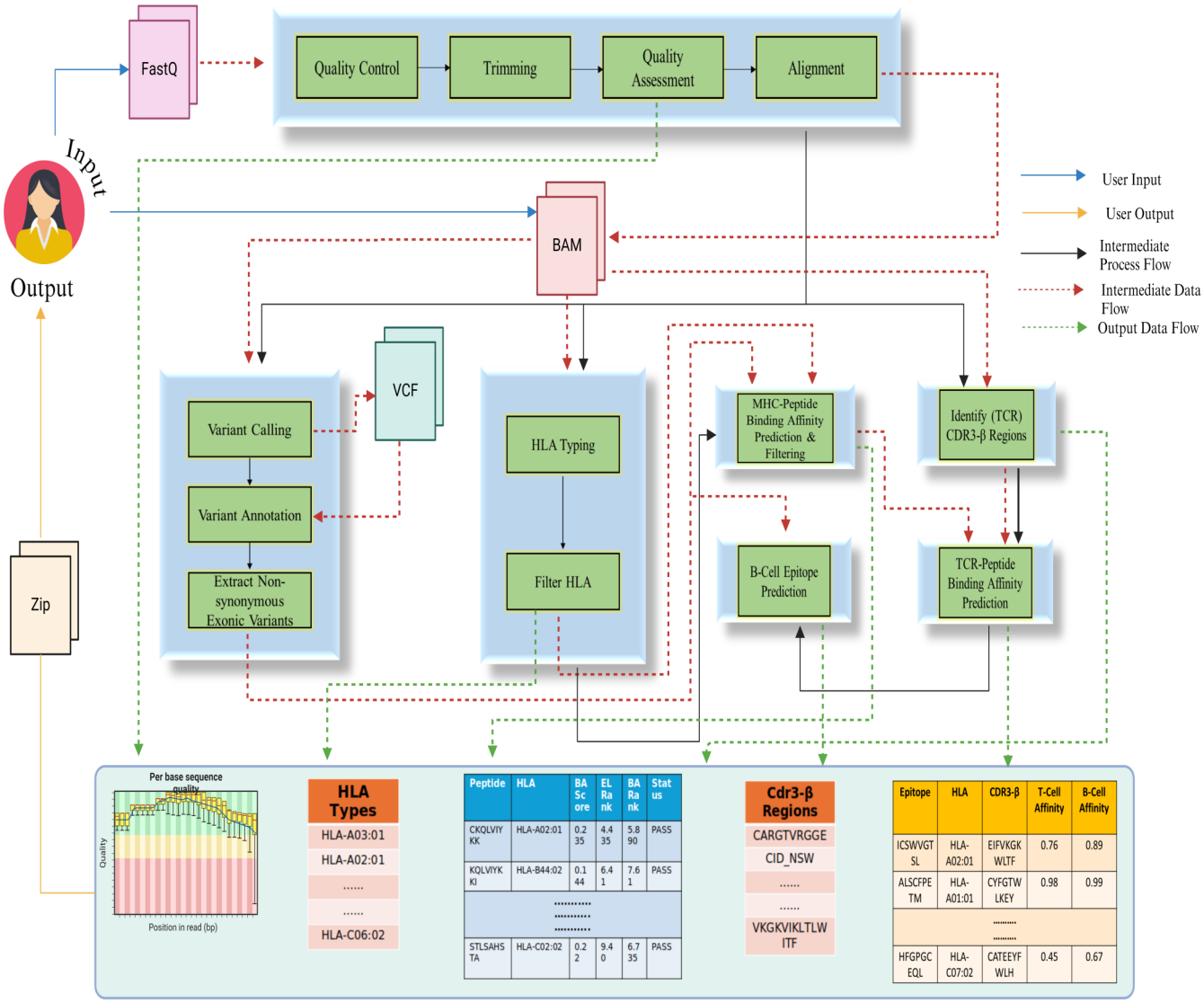

Figure 1. An overview of DeepPROTECTNeo pipeline illustrating the blueprint of an end-to-end workflow. Starting from user WES/WGS data input, we demonstrate the process of data processing, variant identification and finally the neoepitope selection through MHC binding, TCR, BCR interaction. Also, we have shown the final components of the output of our pipeline using a separate block in the figure.

2.1 Data Preprocessing

The DeepPROTECTNeo begins with preprocessing of WES/WGS files containing raw read counts to retain only high-quality sequencing reads. For assessing the quality of raw reads, we used FastQC[6] to evaluate metrics like base quality scores, GC content and sequence length

distribution enabling us to detect potential sequencing errors. Low-quality reads are filtered using Trimmomatic[7] by removing adaptors and trim sequences, retaining only high-quality reads for downstream analysis. The cleaned reads are then aligned to the hg38 reference genome using HISAT2[8], a splice-aware aligner optimal for RNASeq data, enabling precise mapping to splice junctions. Produced Binary Alignment Map (BAM) files were further refined utilizing several utilities provided by SAMtools[9] to remove low-quality and unmapped reads, correct mate-pairs, sort coordinates, eliminate PCR duplicates with *markdup*, left-align indels with *bamleftalign*, and recalibrate mapping quality with *calmd* for accurate downstream processing. We performed variant calling on the processed BAM files using GATK4 Mutect2[10], which employs a Bayesian model and machine learning filters to detect tumor-specific somatic mutations with high specificity, to identify somatic mutations that could serve as potential neoantigen candidates. To assess the functional relevance of these mutations, we employed ANNOVAR[11] for annotating the variants (SNVs, indels) using refGene[12], ensGene[5], gnomAD[13], ClinVar[14] and COSMIC[15] databases. COSMIC's extensive mutation database enabled us to gain critical insights into mutations associated with cancer. This step elucidated the process of identifying mutational effects, prevalent across diverse populations with potential clinical significance. Only exonic and nonsynonymous variants with the capacity to alter protein structures and impact immunogenicity were extracted for further analysis. Pre-processed BAM file was used to identify patient specific HLA types using arcasHLA[16] and we utilized TRUST4 for determining CDR3β regions.

2.2 MHC-Peptide Binding Prediction

A critical stage in the pipeline is the prediction of MHC binding, where tumor-derived peptides (neoepitopes) are assessed for their binding potential to MHC molecules. This step is essential as it enables T cells to recognize and potentially attack tumor cells. We use NetMHCpan[17], a state-of-the-art predictor that evaluates peptide-MHC binding affinity by incorporating binding data from mass spectrometry-based studies. NetMHCpan[17] evaluates each peptide's binding affinity to the HLA alleles identified earlier via arcasHLA[16]. A high-binding affinity score flags these peptides as probable neoantigen candidates. An in-house filtering script was used to refine the neoepitope list, setting flexible affinity thresholds on Binding Affinity Score (BA Score), Eluted Ligand Rank (EL Rank), and Binding Affinity Rank (BA Rank) to retain only high-confidence candidates with strong immunogenic potential.

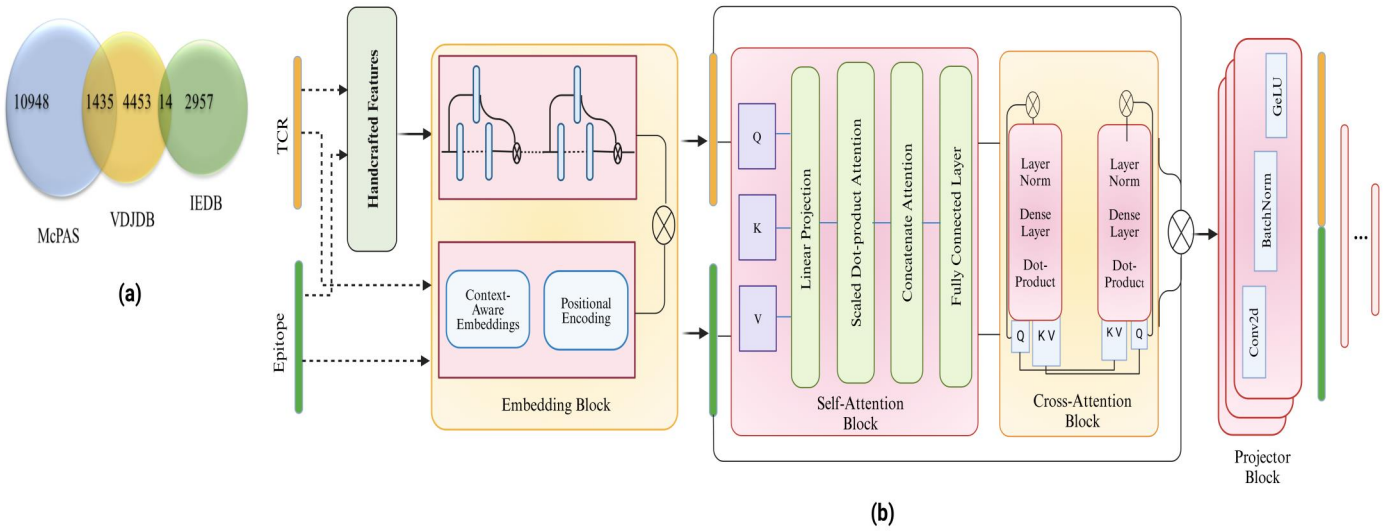

Figure 2. *Overview of our Transformer module for peptide-TCR binding affinity prediction. (a) Shows the collected peptide-TCR pairs from various databases. (b) Cross-attention based transformer architecture*

2.3 TCR-Peptide Binding Prediction

The TCR-peptide binding prediction module of DeepPROTECTNeo is a cutting-edge DL framework, exhibiting high precision, sensitivity, and interoperability in predicting TCR-epitope interactions. Unlike traditional approaches that focus solely on MHC-peptide binding affinity, DeepPROTECTNeo integrates a range of computational techniques to capture complex biological interactions of TCR and peptides. This section provides a detailed description of the model's architecture, covering data preprocessing, negative data sampling, embedding layers, and attention mechanisms used in our study.

2.3.1 Data Collection and Processing

Building a robust model for TCR-epitope interaction prediction is challenging due to the lack of unified ground-truth databases, which necessitates merging diverse datasets from reliable sources to address this limitation. To mitigate this process, we collected positive pairs from McPAS-TCR[18], VDJdb[19], and IEDB[20] (see Figure 2a) containing experimentally validated TCR-epitope binding pairs. Given that these databases lack negative interacting pairs, we employed several negative sampling strategies to ensure a balanced dataset, essential for training an effective classifier. These strategies include reference TCR sampling, random TCR sampling, random epitope sampling and unified epitope sampling, with the latter chosen due to its superior performance in preserving epitope distribution and improving model generalization. Different ratios of positive to negative samples were tested, with a 1:1, 1:3 ratio. Our model with imbalance dataset has a higher ACC value compared to the 1:1 split but the F1 score of this model comes out to be a little compromised as there is a slight difference in the precision and recall. We have found out unified epitope strategy to be most effective.

2.3.2 Transformer Architecture

Prediction of TCR-epitope binding is of complex nature due to the high diversity of CDR3β regions and the limited availability of experimentally validated TCR-epitope binding data, hindering generalization to unseen epitope. Our TCR-epitope binding module is a transformer based architecture, structured to learn the complex patterns of TCR-epitope interactions in a context-specific manner. Our study aims to develop a robust framework for encoding CDR3β and epitopes through a feature extractor block that incorporates multiple tunable ResNet[21] blocks and skip connections. The Peptide package[22] facilitates amino acid encoding and encompasses essential features, including BLOSUM indices, Cruciani properties, Kindera factors, among others provided by the package. We have integrated an additional embedding block, initially randomized, with positional encoders alongside the feature extractor, which is a critical component of our architecture. After this transformation the data undergoes a series of attention mechanisms that encodes the local and the global features and understands the context similar to natural languages, paying more attention to the tokens having more importance, in our context it is the residue that decides the ammino acid and its language. Our cross-attention module captures the inter token dependency, deciphering the cross-contexts of the binding sites capturing inter sequence dependency of the tokens. Skip connections are concatenated with the output from the attention modules and subsequently processed through a series of convolution block before being projected to the forward projection block for final classification. Skip connections are projected in the forward block along with the context rich transformer output to retain any informations that are lost in the previous attention blocks.

Handcrafted feature and embedding block: We tokenize our TCR and epitope pairs into numerical representations using an amino acid vocabulary. Each sequence pair pass down the embedding blocks to learn specific numerical representations and project the query to the key value pairs to learn specific contexts related to the sequence and its amino acid residue (see Figure 2b, Embedding Block). After carefully analyzing both the training and the test data we have restricted the maximum length of the CDR3β to 20 amino acids and the epitopes to 11 considering their distribution in the databases with the most occurring sequence being a 9 mer ("KLGGALQAK") for the epitopes. We encode our CDR3β and the peptide into numerical representations where our vocab remains to be amino acids and X for all invalid characters. Each sequence thereby pair , passes down the embedding blocks to learn specific numerical representations and project the query to the key value pairs to learn specific contexts related to the sequence and its amino acid residue. In addition to sequence embeddings, our transformer module incorporates peptide-specific descriptors which capture the global and the features from the amino acid in hierarchal way also mimicking the reverse vaccinology[23] guided featurization. We convert each pair into a 102-dimensional vector that encapsulates key physicochemical properties, including hydrophobicity, charge, and polarity among other local features.

Deep CNN Projector with Residual Connections: We use a deep CNN[24] projector, composed of multiple deep residual blocks. Each block includes convolutional layers with GELU activation[25] functions and batch normalization, allowing the model to capture nonlinear relationships among descriptors while maintaining stable gradient flow. Our ResNet[21] blocks are tunable and the number of blocks of becomes a critical component while tuning the hyper parameters. The inclusion of residual connections helps to prevent overfitting, even in imbalanced datasets. A global average pooling layer at the end of the CNN Projector reduces spatial

dimensions, distilling each sequence's biochemical profile into a concise format for the attention layer[26].

Self-Attention and Cross-Attention Mechanisms: The integral part of our architecture are self-attention and cross-attention mechanisms[4,27], which enable the model to establish inter- and intra-sequence dependencies, essential for modeling complex TCR-epitope interactions. The Multi-Head Self-Attention (SA)[4] layer processes TCR and epitope embeddings independently, allowing each sequence to capture self-dependencies and focus on critical regions within itself. This is followed by the Cross-Attention (CA)[27] module (See Figure 2b, Attention and Cross-Attention Block), which refines the interaction between TCR and epitope representations and queries come from the epitope block manipulates the key value pair in the CDR3β blocks attending the critical inter block dependency, identifying shared motifs or conserved positions critical for binding specificity. The query Q, key K and value V cross-attends the peptide and TCR tokens and the attention score is calculated using the following equation:

$$\text{Attention}_{TCR\rightarrow pep}\left(Q_{\text{TCR}}, K_{\text{pep}}, V_{\text{pep}}\right) = \text{softmax}\left(\frac{Q_{\text{TCR}} K_{\text{pep}}^{T}}{\sqrt{d_k}}\right) V_{\text{pep}} \quad 1$$

$$\text{Attention}_{\text{pep}\rightarrow TCR}\left(Q_{\text{pep}}, K_{\text{TCR}}, V_{\text{TCR}}\right) = \text{softmax}\left(\frac{Q_{\text{pep}} K_{\text{CCR}}^{T}}{\sqrt{d_k}}\right) V_{\text{TCR}} \quad 2$$

Impact of Skip Connections on the Forward Projector Block: After cross-attention module, the TCR and epitope embeddings along with skip-connected peptide descriptors are fused into a unified feature vector. This vector is then passed through a convolutional block with multiple layers and a fully connected feed-forward network (FFN), which compresses the combined features into a two-dimensional output representing binding specificity.

2.3.3 Hyperparameter Tuning and Optimization

Hyperparameter tuning is essential for balancing expressivity and computational efficiency in DeepPROTECTNeo. Using grid search, critical parameters such as the number of attention heads, convolutional kernel sizes, and the number of CNN residual blocks were optimized. For instance, a hidden dimension of 64 with 8 attention heads provided an effective balance, allowing the model to capture both local and global motifs (Supplementary Table S1).

In summary, DeepPROTECTNeo is built upon advanced embedding, feature extraction, and attention techniques, and is well-suited for predicting TCR-epitope interactions in immunotherapy and vaccine design. Its alignment-free method for capturing TCR and epitope interactions through a fully integrated pipeline beginning with WES/WGS data offers a powerful tool for researchers in cancer immunology and beyond.

In addition to T-cell epitope prediction, DeepPROTECTNeo also incorporates B-cell epitope scoring using NetBCE[28] to support multi-epitope vaccine development, which is critical for eliciting comprehensive immune responses. This unified framework, integrating both T-cell and B-cell receptor affinity predictions, provides a versatile platform for personalized vaccine design, facilitating broad-spectrum immune activation. The model's comprehensive output allows users to combine results with advanced models like AlphaFold Multimer[29] to gain structural insights into how mutations affects binding interactions unrevealing receptor binding affinity of the neoepitopes.

3. Results

DeepPROTECTNeo has achieved a superior performance and generalization on ImmuneCODE[30] (containing 28,303 unique TCR–epitope pairs) benchmark dataset. Driven primarily by its feature projector block, multi-head self and cross attention modules, skip connections, and alignment-free architecture, our model allows dynamic alignment of TCR and epitope sequences enhancing context-aware interactions that are crucial for immune recognition. We have compared our model with the methods that have used additional alignment for the CDR3β regions. Our method turns out to be unique in contextualizing these features without the need of any alignment additionally, upon analyzing the attention maps (see Figure 3c) we have revealed the significance of the anchor motifs that is captured in the attention maps. The cross-attention mechanism in DeepPROTECTNeo provides multi-dimensional perspectives on TCR-epitope interactions, enhancing specificity and improving robustness in identifying true binding pairs. Skip connections contribute significantly by preserving critical local and global features including incorporated RV-like features as discussed earlier. Our performance was evaluated across multiple metrics, including precision-recall curves (PRC), area under the receiver operating characteristic curve (AUROC), F1 score, precision, and recall. On extensive hyperparameter tuning, we achieved a validation ACC of 0.76, F1 of 0.68 and AUPRC of 78 with the following hyperparameters d_model 64, n_heads 8, num_blocks 5, kernel_size 5, channels 64, learning_rate 0.0005, self_sa 8, self_ca 4 (Supplementary S1). Furthermore, we have compared our model with several state-of-the-art DL-based methods such as ATM-TCR[31], TITAN[32], ERGO[33], TEINet[34] outperforming them and achieved an ACC of 0.61, F1 of 0.61 and AUPRC of 0.67 (a detailed comparison is shown in Figure 3b). While our model completely outperformed the previously mentioned methods, we were unable to directly run and compare TEPCAM[35] due to limitations

in reproducing their code. However, our approach (ACC 0.62, AUC 0.68) achieved performance metrics comparable to TEPCAM (ACC of 0.62, AUC 0.67) in our imbalance dataset evaluated on

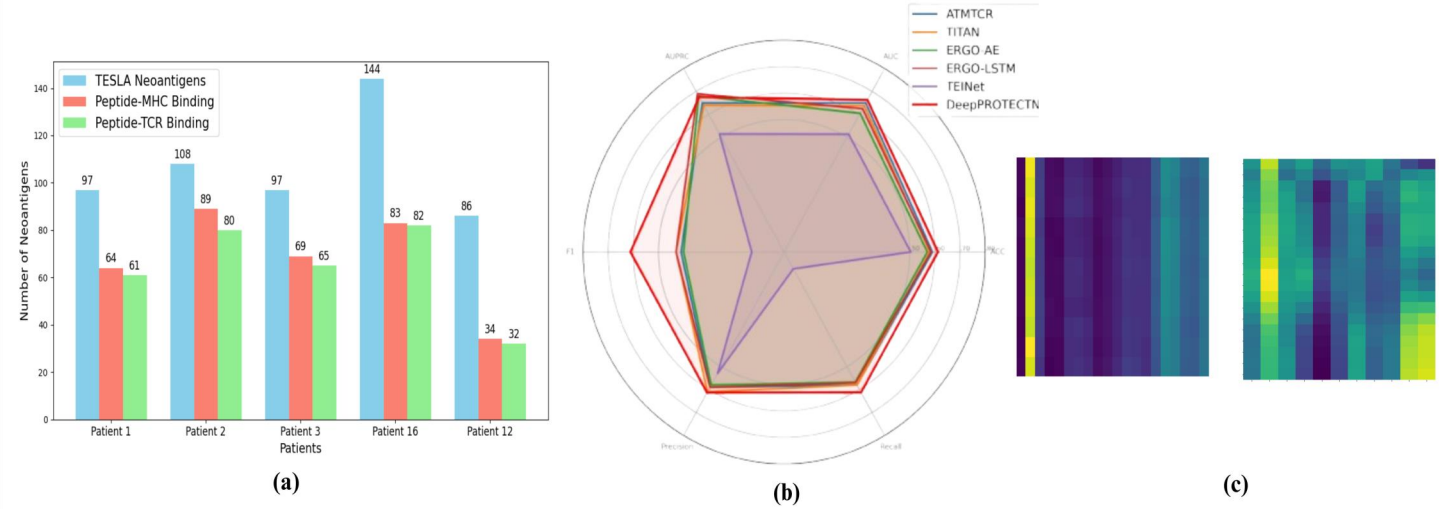

Figure 3. *Performance of our method (a) Detected neoepitopes against TESLA (b) peptide-TCR binding metric comparison with state-of-the-art DL models*

the ImmuneCODE benchmark dataset. Our model consistently demonstrated high specificity and accuracy, successfully identifying true binding pairs. The use of peptide-specific descriptors contributes to the model's generalizability, surpassing models that rely solely on sequence-based features.

We include interpretability features, such as attention maps (see Figure 3c) that highlight critical regions within TCR and epitope sequences. These maps reveal the model's focus on specific amino acid motifs, particularly anchor motifs in epitope binding pockets that stabilize TCR-MHC interactions. The skip connections help retain important features from peptide descriptors, providing deeper biological insights into binding specificity. Attention maps validate DeepPROTECTNeo's predictive capacity, and the use of peptide descriptors enhances its generalizability over models relying solely on sequence features.

Upon evaluation with the TESLA[5], a cohort of experimentally validated patient-specific neoepitopes from Synapse (https://www.synapse.org/), DeepPROTECTNeo identified 339 out of 532 unique candidate neoepitopes excluding the 8-mers, maintaining a high specificity while focusing on critical mutation types such as single nucleotide variants (SNVs) and insertions/deletions (indels). We identified 64 out of 97 epitopes for patient 1, 89 out of 108 for patient 2, 69 out of 97 for patient 3, 83 out of 144 for patient 16, and 34 out of 86 for patient 12 (Supplementary Table S2-S6). Also, 84, 63, 89, 81, 32 epitopes (Fig 3a) have shown high specificity towards TCR binding highlighting the effectiveness and reliability of our transformer module for the corresponding patients. Also, we successfully detected 24 among 37 true neoepitopes reported by TESLA[5] (Supplementary S7). DeepPROTECTNeo focuses on single nucleotide variants (SNVs) and insertions/deletions (indels), leaving out more complex mutation types like copy number alterations (CNAs), splicing variants, and structural variants (SVs) due to the difficulty of seamlessly integrating them into a web tool. Even with these exclusions, we successfully detected a large number of neoepitopes, demonstrating the robust predictive capability and dependability of our pipeline, and paving the way towards a unified platform for cutting-edge vaccinology research.

4. Web-Server Implementation

To serve the research community targeting those who does not have access to high-performance computing facilities (HPC) or the programming proficiency, we hosted DeepPROTECTNeo as a free web service. Currently, the DeepPROTECTNeo is running on a system with Intel i9 12th Gen processor, 64GB of RAM, and an A5000 24GB GPU, designed to efficiently handle complex computational workflows. The server provides a robust and user-friendly interface and is

developed with PHP, HTML, and JavaScript. In the “Service” section, users can upload paired-end FASTQ files or BAM files to run the complete analysis pipeline or upload CDR3-epitope pair files to evaluate binding affinity, allowing flexibility in selecting the desired analysis. To simplify usage, DeepPROTECTNeo offers an “Example” page with precomputed results for Patient 1 from the TESLA dataset, available for download in a zip file, helping users quickly familiarize themselves with the output format. Additionally, a detailed “Help” page provides thorough data preparation guidelines, including step-by-step instructions and explanations of key acronyms, assisting users in correctly preparing their data to fully use the platform's features. With its intuitive interface, flexible input options, and comprehensive support resources, DeepPROTECTNeo aims to provide a powerful and one-stop solution for clinical data processing in real time, performing neoantigen predictions and screening the best neoepitopes fit for eliciting a robust immune response.

5. Conclusion

DeepPROTECTNeo offers a stand-alone and integrative pipeline enabling users to analyze their raw WES/WGS data and identify neo-antigens derived from tumor-specific somatic mutation through a single platform. Our model demonstrated high performance and generalization on unseen epitopes compared to previous methods. Our transformer module predicted almost no false negatives on the 339 potential neoepitopes extracted from the TESLA dataset giving high binding values for 320 neoepitopes (see Fig 3a) with cut off greater than 0.90 for epitope-TCR binding prediction, which is a major advancement to previous predicting with low confidence positive values. Moreover we captured the regions of CDR3b-epitope interaction in our attention map which adds criticality to our analysis and adds significant biological inference in analysing and

identifying crucial residues and binding motifs. The alignment-free method, augmented by attention mechanisms, peptide-specific descriptors, and with advanced transformer modules, DeepPROTECTNeo promises to deliver exceptional precision in predicting TCR-epitope binding and neoantigen immunogenicity. Additionally, the integration of B-cell epitope scoring expands its utility to support multi-epitope vaccine design, promoting a holistic understanding to immune system activation and broadening its application across humoral and cellular immune responses and paving the way for comprehensive cancer immunotherapy solutions.

CRediT authorship contribution statement

DD: Writing – review & editing, Writing – original draft, Conceptualization, Methodology, Data curation, Validation, Analysis, Software, Illustration

SB: Writing – review & editing, Writing – original draft, Conceptualization, Implementation, Methodology, Data curation, Validation, Analysis, Illustration

AP: Software, Analysis

PM: Writing – review & editing, Writing – original draft, Conceptualization, Methodology, Supervision

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by Department of Science and Technology and Biotechnology (Please Add Project Details) and IIT KHARAGPUR AI4ICPS I HUB FOUNDATION (Please Add Project Details). Additionally, Debraj (PMRF Id: 2402784) thanks the MOE, India, for the Prime Minister research fellowship.

References

[1] X. Zhao, X. Pan, Y. Wang, Y. Zhang, Targeting neoantigens for cancer immunotherapy, Biomark Res 9 (2021). https://doi.org/10.1186/s40364-021-00315-7.

[2] K. Shah, A. Al-Haidari, J. Sun, J.U. Kazi, T cell receptor (TCR) signaling in health and disease, Signal Transduct Target Ther 6 (2021). https://doi.org/10.1038/s41392-021-00823-w.

[3] Q. Yin, J. Tang, X. Zhu, Next-generation sequencing technologies accelerate advances in T-cell therapy for cancer, Brief Funct Genomics 18 (2019). https://doi.org/10.1093/bfgp/ely018.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, (2017). http://arxiv.org/abs/1706.03762.

[5] D.K. Wells, M.M. van Buuren, K.K. Dang, V.M. Hubbard-Lucey, K.C.F. Sheehan, K.M. Campbell, A. Lamb, J.P. Ward, J. Sidney, A.B. Blazquez, A.J. Rech, J.M. Zaretsky, B. Comin-Anduix, A.H.C. Ng, W. Chour, T. V. Yu, H. Rizvi, J.M. Chen, P. Manning, G.M. Steiner, X.C. Doan, A.A. Khan, A. Lugade, A.M.M. Lazic, A.A.E. Frentzen, A.D. Tadmor, A.S. Sasson, A.A. Rao, B. Pei, B. Schrörs, B. Berent-Maoz, B.M. Carreno, B. Song, B. Peters, B. Li, B.W. Higgs, B.J. Stevenson, C. Iseli, C.A. Miller, C.A. Morehouse, C.J.M. Melief, C. Puig-Saus, D. van Beek, D. Balli, D. Gfeller, D. Haussler, D. Jäger, E. Cortes, E. Esaulova, E. Sherafat, F. Arcila, G. Bartha, G. Liu, G. Coukos, G. Richard, H. Chang, H. Si, I. Zörnig, I. Xenarios, I. Mandoiu, I. Kooi, J.P. Conway, J.H. Kessler, J.A. Greenbaum, J.F. Perera, J. Harris, J. Hundal, J.M. Shelton, J. Wang, J. Wang, J. Greshock, J.

Blake, J. Szustakowski, J. Kodysh, J. Forman, L. Wei, L.J. Lee, L.F. Fanchi, M. Slagter, M. Lang, M. Mueller, M. Lower, M. Vormehr, M.N. Artyomov, M. Kuziora, M. Princiotta, M. Bassani-Sternberg, M. Macabali, M.R. Kojicic, N. Yang, N.M.I. Raicevic, N. Guex, N. Robine, N. Halama, N.M. Skundric, O.S. Milicevic, P. Gellert, P. Jongeneel, P. Charoentong, P.K. Srivastava, P. Tanden, P. Shah, Q. Hu, R. Gupta, R. Chen, R. Petit, R. Ziman, R. Hilker, S.A. Shukla, S. Al Seesi, S.M. Boyle, S. Qiu, S. Sarkizova, S. Salama, S. Liu, S. Wu, S. Sridhar, S.L.C. Ketelaars, S. Jhunjhunwala, T. Shcheglova, T. Schuepbach, T.H. Creasy, V. Josipovic, V.B. Kovacevic, W. Fu, W.J. Krebber, Y.H. Hsu, Y. Sebastian, Z.K. Yalcin, Z. Huang, T. Merghoub, J. Guinney, A. Kolom, C. Selinsky, A. Ribas, M.D. Hellmann, N. Hacohen, A. Sette, J.R. Heath, N. Bhardwaj, F. Ramsdell, R.D. Schreiber, T.N. Schumacher, P. Kvistborg, N.A. Defranoux, Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction, Cell 183 (2020). https://doi.org/10.1016/j.cell.2020.09.015.

[6] S.W. Wingett, S. Andrews, FastQ Screen: A tool for multi-genome mapping and quality control, F1000Res 7 (2018). https://doi.org/10.12688/f1000research.15931.2.

[7] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014). https://doi.org/10.1093/bioinformatics/btu170.

[8] D. Kim, J.M. Paggi, C. Park, C. Bennett, S.L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype, Nat Biotechnol 37 (2019) 907–915. https://doi.org/10.1038/s41587-019-0201-4.

[9] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools, Bioinformatics 25 (2009). https://doi.org/10.1093/bioinformatics/btp352.

[10] D. Benjamin, T. Sato, K. Cibulskis, G. Getz, C. Stewart, L. Lichtenstein, Calling Somatic SNVs and Indels with Mutect2, BioRxiv (2019).

[11] K. Wang, M. Li, H. Hakonarson, ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Res 38 (2010). https://doi.org/10.1093/nar/gkq603.

[12] National Center for Biotechnology Information, RefSeq: NCBI Reference Sequence Database, U.S. National Library of Medicine (2018).

[13] S. Gudmundsson, M. Singer-Berk, N.A. Watts, W. Phu, J.K. Goodrich, M. Solomonson, H.L. Rehm, D.G. MacArthur, A. O'Donnell-Luria, Variant interpretation using population databases: Lessons from gnomAD, Hum Mutat 43 (2022) 1012–1030. https://doi.org/10.1002/humu.24309.

[14] M.J. Landrum, J.M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein, D.R. Maglott, ClinVar: Public archive of interpretations of clinically relevant variants, Nucleic Acids Res 44 (2016). https://doi.org/10.1093/nar/gkv1222.

[15] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, P.A. Futreal, M.R. Stratton, R. Wooster, The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website, Br J Cancer 91 (2004). https://doi.org/10.1038/sj.bjc.6601894.

[16] R. Orenbuch, I. Filip, D. Comito, J. Shaman, I. Pe'Er, R. Rabadan, ArcasHLA: High-resolution HLA typing from RNAseq, Bioinformatics 36 (2020). https://doi.org/10.1093/bioinformatics/btz474.

[17] B. Reynisson, B. Alvarez, S. Paul, B. Peters, M. Nielsen, NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data, Nucleic Acids Res 48 (2021). https://doi.org/10.1093/NAR/GKAA379.

[18] N. Tickotsky, T. Sagiv, J. Prilusky, E. Shifrut, N. Friedman, McPAS-TCR: A manually curated catalogue of pathology-associated T cell receptor sequences, Bioinformatics 33 (2017). https://doi.org/10.1093/bioinformatics/btx286.

[19] M. Shugay, D. V Bagaev, I. V Zvyagin, R.M. Vroomans, J.C. Crawford, G. Dolton, E.A. Komech, A.L. Sycheva, A.E. Koneva, E.S. Egorov, A. V Eliseev, E. Van Dyk, P. Dash, M. Attaf, C. Rius, K. Ladell, J.E. McLaren, K.K. Matthews, E.B. Clemens, D.C. Douek, F. Luciani, D. van Baarle, K. Kedzierska, C. Kesmir, P.G. Thomas, D.A. Price, A.K. Sewell, D.M. Chudakov, VDJdb: a curated database of T-cell receptor sequences with known antigen specificity, Nucleic Acids Res 46 (2018) D419–D427. https://doi.org/10.1093/nar/gkx760.

[20] R. Vita, S. Mahajan, J.A. Overton, S.K. Dhanda, S. Martini, J.R. Cantrell, D.K. Wheeler, A. Sette, B. Peters, The Immune Epitope Database (IEDB): 2018 update, Nucleic Acids Res 47 (2019). https://doi.org/10.1093/nar/gky1006.

[21] M. Razavi, S. Mavaddati, H. Koohi, ResNet deep models and transfer learning technique for classification and quality detection of rice cultivars, Expert Syst Appl 247 (2024). https://doi.org/10.1016/j.eswa.2024.123276.

[22] D. Osorio, P. Rondón-Villarreal, R. Torres, Peptides: A package for data mining of antimicrobial peptides, R Journal 7 (2015). https://doi.org/10.32614/rj-2015-001.

[23] R. Rappuoli, Reverse vaccinology, Curr Opin Microbiol 3 (2000) 445–450. https://doi.org/10.1016/S1369-5274(00)00119-3.

[24] A. Saxena, An Introduction to Convolutional Neural Networks, Int J Res Appl Sci Eng Technol 10 (2022). https://doi.org/10.22214/ijraset.2022.47789.

[25] D. Hendrycks, K. Gimpel, Gaussian Error Linear Units (GELUs), (2016). http://arxiv.org/abs/1606.08415.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Adv Neural Inf Process Syst, 2017.

[27] M. Gheini, X. Ren, J. May, Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural

Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2021: pp. 1754–1765. https://doi.org/10.18653/v1/2021.emnlp-main.132.

[28] H. Xu, Z. Zhao, NetBCE: An Interpretable Deep Neural Network for Accurate Prediction of Linear B-cell Epitopes, Genomics Proteomics Bioinformatics 20 (2022). https://doi.org/10.1016/j.gpb.2022.11.009.

[29] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, Protein complex prediction with AlphaFold-Multimer, BioRxiv (2022).

[30] S. Nolan, M. Vignali, M. Klinger, J.N. Dines, I.M. Kaplan, E. Svejnoha, T. Craft, K. Boland, M. Pesesky, R.M. Gittelman, T.M. Snyder, C.J. Gooley, S. Semprini, C. Cerchione, M. Mazza, O.M. Delmonte, K. Dobbs, G. Carreño-Tarragona, S. Barrio, V. Sambri, G. Martinelli, J.D. Goldman, J.R. Heath, L.D. Notarangelo, J.M. Carlson, J. Martinez-Lopez, H.S. Robins, A large-scale database of T-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2., Res Sq (2020). https://doi.org/10.21203/rs.3.rs-51964/v1.

[31] M. Cai, S. Bang, P. Zhang, H. Lee, ATM-TCR: TCR-Epitope Binding Affinity Prediction Using a Multi-Head Self-Attention Model, Front Immunol 13 (2022). https://doi.org/10.3389/fimmu.2022.893247.

[32] A. Weber, J. Born, M. Rodriguez Martínez, TITAN: T-cell receptor specificity prediction with bimodal attention networks, Bioinformatics 37 (2021). https://doi.org/10.1093/bioinformatics/btab294.

[33] I. Springer, H. Besser, N. Tickotsky-Moskovitz, S. Dvorkin, Y. Louzoun, Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs, Front Immunol 11 (2020). https://doi.org/10.3389/fimmu.2020.01803.

[34] Y. Jiang, M. Huo, S.C. Li, TEINet: a deep learning framework for prediction of TCR–epitope binding specificity, Brief Bioinform 24 (2023). https://doi.org/10.1093/bib/bbad086.

[35] J. Chen, B. Zhao, S. Lin, H. Sun, X. Mao, M. Wang, Y. Chu, L. Hong, D.Q. Wei, M. Li, Y. Xiong, TEPCAM: Prediction of T-cell receptor–epitope binding specificity via interpretable deep learning, Protein Science 33 (2024). https://doi.org/10.1002/pro.4841.