DeepPROTECTNeo: A Deep learning-based Personalized and RV-guided Optimization tool leveraging 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
Pralay Mitra, 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

Abstract

The development of personalized cancer vaccines relies heavily on accurately identifying neoepitopes capable of eliciting strong immune responses. T cell receptor (TCR)-epitope interactions are fundamental to cancer immunotherapy. Despite significant advances, computational approaches primarily focus on epitope-major histocompatibility complex (MHC) binding, often overlooking the critical contribution of T-cell receptor (TCR) binding particularly in the context of diverse affinity determinants crucial for binding groove of the peptide, amino acid permutations at binding sites, anchor motifs and hydrophobicity determining the quality of anti-tumour immune responses. Here, we present DeepPROTECTNeo, a unified deep learning framework to transform raw patient-specific whole-exome or whole-genome sequencing data into prioritized neoepitope candidates, within a single seamless pipeline. DeepPROTECTNeo integrates genomic variant detection, HLA typing, high-affinity peptide-MHC binding prediction, variant-driven TCR repertoire mining followed by a hybrid transformer-Convolutional Neural Network dual-branch feature extractor with explicit cross-attention based Deep Learning (DL) model for TCR-epitope binding prediction. Our biologically informed architecture inspired by reverse vaccinology (RV) pipeline fuses bidirectional (Long Short-Term Memory) LSTM sequence features, deep convolutional attention over evolutionary and physicochemical descriptors, advanced gated fusion of the physicochemical features and TCR numbered contextual embedding enabling precise modelling of TCR-epitope interactions and yielding interpretable representations at residue level. DeepPROTECTNeo substantially outperforms state-of-the-art methods, achieving a mean AUROC of 0.7234 and mean AUPRC of 0.7659 in zero-shot TCR-peptide binding scenarios on our held-out five-fold cross validation dataset. While benchmarked against challenging viral and mutational independent datasets, our model captures critical clinical features and demonstrates high binding scores across the datasets.

DeepPROTECTNeo also identified 18 out of 34 validated neo-epitopes exhibiting high TCR affinity from a patient-specific cancer cohort, establishing itself as a platform to reliably prioritize actionable neoepitopes directly from clinical sequencing data and escalate the process of personalized cancer immunotherapy.

Keywords

Neo-epitope; Dual branch Deep Learning; Cross-Attention; TCR-epitope Binding; Cancer therapy

1 Main

The adaptive immune response is initiated in secondary lymphoid organs when foreign antigens, subsequent to host exposure through infectious or non-infectious means, are captured and proteolytically processed by antigen-presenting cells (APCs) into shorter peptide fragments. Upon loading onto major histocompatibility complex (MHC) molecules and presentation at the APC surface, these peptides are designated as epitopes, reflecting their role as the specific determinants that engage T cell recognition[1]. Naïve T cells interrogate peptide–MHC (pMHC) complexes through their clonotypic T cell receptors (TCRs). Productive recognition initiates intracellular signalling cascades that drive T cell activation, clonal expansion, and differentiation into effector populations[2]. The steady progress in cancer immunotherapy has leveraged these mechanisms, enabling personalized treatments where T cells target malignant cells via TCR recognition of tumour-derived peptides presented on MHC molecules[3–5]. In particular, the identification of tumour-specific neoantigens i.e. peptides arising from somatic mutations exclusive to cancer cells and absent from normal tissues have transformed personalized therapeutic strategies[6,7]. Unlike tumour-associated antigens (TAAs), neoantigens are uniquely expressed by cancer cells, making them highly attractive targets for vaccine development and adoptive T cell therapies[8–10].

Next-generation sequencing (NGS) techniques particularly whole-exome sequencing (WES) and Whole Genome Sequencing (WGS) enable us to capture tumour-specific somatic variants followed by RNA-Seq to confirm their expression[11,12]. Translating these raw data into clinically actionable targets, however, requires a multi-step workflow including precise HLA genotyping (e.g., ArcasHLA[13], OptiType[14]), in silico pMHC binding predictions (e.g., NetMHCpan[15], MHCflurry[16]), followed by T-cell receptor (TCR) peptide (pTCR) binding

which is often neglected by traditional MHC-peptide binding approaches[17,18]. Despite these advances, the clinical applicability of existing methods mostly confines to multiple input files at different stages of processing and separate workflows for TCR profiling, HLA typing and variant calling[19,20], pMHC and pTCR binding, thereby constraining clinical applicability.

Recently various deep-learning frameworks have been developed for TCR-epitope binding prediction. Still, substantial challenges remain in aligning these computational advances with the nuanced realities of immunobiology and clinical implementation. Early CNN models like ImRex[21], NetTCR-2[22], TITAN[23] primarily focus on local motifs often failing to capture long-range dependencies. Recurrent and autoencoder-based frameworks including ERGO-AE[24], DeepTCR[25] focus on mapping CDR3β and peptide sequences into meaning latent spaces, yet they offer only marginal gains in zero-shot settings. The introduction of transformer-based approaches like ATM-TCR[26] and TEPCAM[27] yield strong in-domain accuracy but still reverting to near-random predictions on truly novel antigens. Hybrid pMHC-TCR models such as pMTnet[28] and TCRGP[29] added peptide-MHC binding priors and even docking-derived constraints, limited by the dependency on paired α/β-chain and HLA information. A parallel progress in self-supervised pre-training produced TCR-specific language models like BertTCR[30] and STAPLER[31] trained on tens of millions of unlabelled sequences, alongside LSTM-based CatELMo[32] yet they face trade-offs between generality and specificity. Given this development in DL-based techniques, the ultimate challenge in TCR-peptide specificity modelling arises from the CDR3β loop itself: it mediates the majority of peptide contacts yet exhibits staggering sequence diversity, context-dependent binding modes and, in most high-throughput assays, lacks paired α-chain information[33,34].

We introduce DeepPROTECTNeo (https://cosmos.iitkgp.ac.in/DeepPROTECTNeo/) a unified pipeline for personalized neoantigen prioritization streamlining the entire workflow from processing raw sequencing reads through somatic variant calling, HLA inference,

peptide–MHC filtering, and downstream TCR-epitope binding analysis (Fig. 1). At its core, DeepPROTECTNeo employs a dual-branch, context-aware Transformer-CNN architecture: a sequence branch leveraging bidirectional LSTMs with learned positional encodings to capture long-range dependencies in peptide and TCR CDR3β sequences, and a convolutional branch encoding 102 physicochemical and BLOSUM-derived descriptors via deep residual networks with channel-wise self-attention (Fig. 2A, B). A gated fusion module integrates these complementary representations, while a cross-attention mechanism with rotary embeddings captures fine-grained residue-level interactions, enabling both robust prediction of TCR-epitope binding and interpretability by recovering key anchor residues and novel contact motifs and progressively capturing a clear separation through propagation across different layers (Fig. 2C). Beyond achieving state-of-the-art predictive performance across multiple benchmarking cohorts, DeepPROTECTNeo demonstrates clinical relevance when evaluated on the TESLA consortium's experimentally validated neoepitope dataset, where it consistently separates binders from non-binders and recovers the majority of confirmed epitopes. These results establish DeepPROTECTNeo as a powerful framework to accelerate translational pipelines for patient-specific vaccine and adoptive T-cell therapy design, as detailed in the sections that follow.

2 Results

2.1 DeepPROTECTNeo Improves Binding Performance on Strict TCR Split Data

We investigated DeepPROTECTNeo's CDR3β-epitope binding prediction capabilities using a stringent TCR-split strategy that mirrors authentic clinical immunotherapy scenarios. We curated a unified dataset (Supplementary S1) consisting of 140992 unique CDR3β-epitope interacting pairs among 130303 diverse TCR repertoire and 1903 epitopes (see CDR3β and epitope length distribution and their overlap across the datasets in Fig. 2D). Finding

experimentally validated non-binding interaction pairs possesses a significant challenge in this domain, we relied on computational strategy to overcome this by generating negative pairs through an epitope shuffling process while preserving underlying sequence distribution characteristics maintaining biological plausibility without introducing systematic biases. The paired TCR-epitope data was split into 5 equal mutually disjoint folds based on TCR sequences. For benchmarking, we carefully selected six state-of-the-art, single-chain TCR-epitope predictors including ATM-TCR[26], ERGO-AE[24], ERGO-LSTM[24], TEINet[35], epiTCR[36] and NetTCR-2.0[22] on the basis of their ability to operate using only CDR3β input, thus matching DeepPROTECTNeo's single-chain paradigm (See Supplementary S2 for details about training). To ensure a fair and rigorous comparison, we retrained each of these models from scratch on our unified dataset using their default parameter settings. We evaluated all models using key classification metrics Balanced Accuracy (BA), macro-averaged F1-score, AUROC and AUPRC (Supplementary S3) to capture both class-wise prediction quality and overall ranking performance. In particular, BA was preferred over raw accuracy because of its ability to equally weight true positive and true negative rates crucial for wide distribution of TCR sequences in the training data. Combined ROC and precision–recall analyses (Fig. 3A, B) reveal that DeepPROTECTNeo achieves a mean±s.d. AUROC of 0.7856±0.0023 and AUPRC of 0.7932±0.0024, significantly outperforming all other models by a margin of 4 percent point gain in both the metrics. Strong early-retrieval performance evidenced by high precision at the leftmost (low-recall) region of the precision-recall curves suggests that the top-ranked predictions are densely populated with true binders. Moreover, the narrow-shaded bands (mean±1s.d. across five TCR-split folds) quantify consistent performance over the folds: DeepPROTECTNeo's inter-fold s.d. is just 0.0023 for AUROC and 0.0024 for AUPRC, whereas other models particularly epiTCR shows high variability (0.1203 for AUROC and 0.1240 for AUPRC) suggesting inter-fold performance fluctuations, making its

top predictions highly unreliable. Paired two-tailed t-tests confirm that DeepPROTECTNeo's improvements for both the metrices over ATM-TCR, ERGO-AE, ERGO-LSTM, TEINet and NetTCR-2 are highly significant ($p<1\times10^{-10}$).

To complement the continuous-ranking evaluation in earlier, we demonstrated a threshold-based assessment by showing fold-averaged metrics under a uniform classification cutoff (Fig. 3C). DeepPROTECTNeo achieves a BA of 0.6973±0.0020, outperforming the next best (NetTCR-2: 0.6678±0.0025) by 4.4-%; a precision of 0.6981±0.0021, a 0.8-% gain over closest ATM-TCR (0.6923±0.0100); a recall of 0.6973±0.0020, around 3% above NetTCR-2 (0.6678±0.0025); and an F1 of 0.6969±0.0020, 3.1-% higher than NetTCR-2 (0.6662±0.0030). All pairwise tests for BA, recall and F1 show highly significant improvements over ATM-TCR ($p=5.8\times10^{-4}$ to 1.3×10^{-3}), ERGO-AE ($p=1.4\times10^{-4}$ to 1.6×10^{-4}), ERGO-LSTM ($p=8.9\times10^{-10}$ to 3.9×10^{-9}), TEINet ($p=5.5\times10^{-7}$ to 5.9×10^{-6}) and NetTCR-2 ($p=9.9\times10^{-5}$ to 1.8×10^{-4}). Precision gains were likewise significant versus ERGO-AE ($p=6.7\times10^{-3}$), ERGO-LSTM (1.1×10^{-5}), TEINet (8.9×10^{-3}) and NetTCR-2 (2.4×10^{-4}). We demonstrated the raw score distributions for binders against non-binder pairs (Fig. 3D) and Coefficient of Variation (CV) affirms that DeepPROTECTNeo yields a clear gap between binder (mean 0.611±0.193, CV 0.316) and non-binder (mean 0.385±0.194, CV 0.505) predictions with moderate inter-fold variability. Whereas, ATM-TCR despite having a high binder (mean 0.885±0.257, CV 0.290) lacks clear non-binder scores (mean 0.614±0.422, CV 0.688), while ERGO-LSTM shows zero discrimination (both classes mean≈0.547±0.02, CV≈0.036) and TEINet's binder scores fluctuate widely (CV 0.512 for binder and 0.4 for non-binder). NetTCR-2 also displays high variability in both classes (CVs 0.373/0.637). Although epiTCR achieves the lowest CVs (0.215/0.285), its small mean separation (0.539 vs 0.454) hinders its discriminative power. Together, these results demonstrates that

DeepPROTECTNeo achieves best balance of strong binder, non-binder separation and consistent score stability across folds.

2.2 Contributions of Architectural Components on Binding Predictions

We have performed grid search for achieving a stable model configuration over a range of hyperparameters, followed by sensitivity analysis revealing its robustness to hyperparameter variations, with only modest fluctuations across top configurations. All primary metrics (F1, accuracy, precision, recall) showed minimal variance (standard deviation≤0.004), highlighting architectural stability. The skip connection dimension (skip_dim) exhibited the strongest positive correlation with all performance metrics (Spearman $\rho \approx 0.40$-0.45), confirming its importance for effective information transfer and gradient flow (see Fig. 4). CNN embedding dimension (cnn_dim) and the number of CNN blocks had moderate positive effects, supporting the extraction of multi-scale physicochemical features. The dimensionality of the stacked CNN 1D residual blocks, and the CNN dimension shows its superiority over the other model parameters when compared on validation loss, showcasing its importance in projecting the features to the forward layers which is further confirmed using F scores derived using ANOVA (Supplementary S4). The removal of the feature enriched skip connections led to a drop in accuracy and F1, confirming their role in stabilizing learning and preventing feature degradation, albeit with a slight trade-off in AUROC and AUPRC. Removal of cross-attention resulted in a minor decrease in all metrics, underlining its value for modelling inter-sequence dependencies. The complete model, integrating both skip connections and cross-attention, consistently achieved the highest scores across all evaluation metrics (Supplementary S5). DeepPROTECTNeo architecture demonstrates strong resilience to hyperparameter changes, with skip connections and cross-attention emerging as the most influential components for maximizing predictive accuracy and generalization. The model's design ensures reliable and stable performance across a range of settings, validating its

suitability for robust TCR-epitope binding prediction in diverse clinical and biological contexts.

2.3 DeepPROTECTNeo Demonstrates Improved Binding Performance on SARS-CoV-2 data

To evaluate DeepPROTECTNeo's generalizability to real world TCR-epitope diversity, we compared its performance on the ImmuneCODE[37] dataset against the six binding models as discussed earlier. ImmuneCODE is a repertoire of experimentally validated TCR-epitope interaction pairs derived from over 1000 SARS-CoV-2-infected individuals. All the six models were trained on our unified dataset over 5 stratified folds, to retain the uniformity over the predictions. To prevent data leakage, we removed all the interacting pairs from ImmuneCODE that were also present on our unified dataset finally retaining 18811 interactions. Negative pairs (non-binding TCR-epitope pairs) were generated using random recombination of the positive pairs, yielding a balanced 1:1 dataset for performance evaluation. DeepPROTECTNeo consistently outperforms state-of-the-art baselines across all key metrics (Fig. 5A). Aggregated over five folds, we achieved the highest BA (0.6237±0.0042), F1-macro (0.6223±0.0039), AUROC (0.6752±0.0047) and AUPRC (0.6766±0.0049), with paired *t*-test confirming significant gain over nearest best performing model for all the metrices: BA (ATM-TCR; Score: 0.5664±0.0179, $p=4.57\times10^{-3}$), F1-macro (ERGO-AE: 0.5258±0.0298, $p=3.92\times10^{-3}$), AUROC (ATM-TCR; Score: 0.6623±0.0077, $p=8.74\times10^{-3}$), AUPRC (ATM-TCR; Score: 0.6590±0.0049, $p=1.93\times10^{-3}$). Furthermore, DeepPROTECTNeo exhibits the lowest cross-fold variability (e.g. AUROC: $\sigma=0.0047$ vs the closes ATM-TCR $\sigma=0.0077$), demonstrating its robust generalization over the fold in zero-shot setting. Due to masking mis-calibration of high-aggregated scores[38], we further evaluated the reliability of our binding predictions (Fig. 5B). DeepPROTECTNeo's curve closely follows the diagonal line more tightly than any competitor, yielding the lowest Brier

score of 0.2267 and minimal Expected Calibration Error (ECE) of 0.0218. Using Hosmer–Lemeshow test DeepPROTECTNeo's χ^2 statistic (836) was recorded orders of magnitude smaller than all the baselines (χ^2 score>1700 for every other model), demonstrating its superior probability estimation. We examined per-epitope ΔAUC across all epitopes with violin plots (Fig. 5C) capturing the full range of gains and losses over all the epitopes. For every model, the one-sample t-test rejects the null hypothesis of zero mean difference against ATM-TCR (mean ΔAUC=+0.027, IQR=0.101; $p=1.03\times10^{-2}$), ERGO-AE (+0.065, IQR=0.131; $p=1.09\times10^{-}$), ERGO-LSTM (+0.066, IQR=0.158; $p=1.61\times10^{-}$), TEINet (+0.080, IQR=0.138; $p=2.13\times10^{-}$), epiTCR (+0.091, IQR =0.133; $p=1.48\times10^{-}$) and NetTCR-2 (+0.086, IQR=0.119; $p=2.85\times10^{-1}$). Building on the bi-directional consistency in Fig. 5C, 40-epitope waterfall (Fig. 5D) reveals that DeepPROTECTNeo markedly outperforms its closest competitor ATM-TCR on its top 20 gains (median ΔAUC=+0.174, IQR=0.111; maximum +0.420), while its 20 worst peptides show a median loss of only –0.143 (IQR=0.120; minimum –0.420, a near-symmetric profile that highlights widespread, consistent improvements. Fig. 5E shows DeepPROTECTNeo's raw score distributions, highlighting its smooth, well-calibrated confidence and superior binder/non-binder separation for reliable thresholding. Our binding prediction model demonstrated the largest separation between true binders and non-binders (Cliff's δ=0.35, U test, p≈0.00), with scores distributed almost symmetrically around the decision boundary (Skew test; skewness score=0.071, $p=5.64\times10^{-36}$), indicating only a minimal right-tail bias. In contrast, ATM-TCR achieved a similar effect size (δ=0.32) but exhibited pronounced left skew (skewness score=–1.731, $p<10^{-36}$), with nearly half of predictions saturating at the upper limit, limiting threshold resolution. ERGO-AE showed moderate discrimination (δ=0.15) and mild left skew (skewness score=–0.852, p≈0.00), while ERGO-LSTM, epiTCR and TEINet yielded effect sizes near zero, indicating negligible separability. NetTCR-2 displayed only a marginal effect

(δ=0.07) and uneven binder score distribution (skewness score=–0.049, p=0.798). These comparisons highlight that our framework not only quantifies relative ranking ability but also characterises the underlying score distribution, revealing model-specific biases that influence threshold-based performance.

2.4 DeepPROTECTNeo Outperforms Baselines on Mutation and Viral Binding Data

Next, we investigated the performance of our model on two challenging TCR-epitope interaction data namely Mutation and Viral. Mutation data is sourced from studies by ePytope-TCR[39,40] team and it comprises of 4244 TCR-epitope pairs containing six TCRs against 132 single amino acid mutations of the neo-epitope *VPSVWRSSL*, and 172 mutations of the human CMV epitope *NLVPMVATV* in which experimentally validated epitopes were systematically altered by single amino acid substitutions at key positions. Notably, this dataset comprises three TCR repertoires with annotated activation scores through exchanging the amino acid residues and providing a continuous activation score using NFAT (Nuclear Factor of Activated T cells) reporter expression in flow cytometry for each combination of TCR-epitopes. Viral dataset consists of 8932 TCR 9-mer epitope interactions, sourced from the same study by ePytope-TCR[40] , originated from a vaccine cohort study by Kocher et al.,[41] and the BEAM-T pipeline example datasets provided by 10x Genomics originally drawn from SARS-CoV-2, HHV-1, EBV and IAV from 14 donors. For benchmarking, same five-fold strategy was followed here for all the models. In the Mutation dataset DeepPROTECTNeo emerges as the top performing model (Fig. 6A) on every key metrices achieving the highest BA (0.5634±0.0183), AUROC (0.5978±0.0153), AUPRC (0.3562±0.0102) and F1-macro (0.5626±0.0172). The closest competitor in balanced accuracy is ATM-TCR (0.5368±0.0163), while ERGO-AE registers the second-highest AUROC (0.5678±0.0258) and F1-macro (0.5328±0.0191) and ATM-TCR being the closest in terms of AUPRC (0.3338±0.0203). For the Viral dataset (Fig. 6B), DeepPROTECTNeo leads

in BA (0.5656±0.0191) and AUROC (0.6091±0.0202), with ATM-TCR being a close second in BA (0.5602±0.0060) and NetTCR-2 in AUROC (0.6029±0.0262). We also compared our model's performance based on the given trained weights for different category of models (including pMHC-TCR, epitope-TCR α/β models) following the strategy and metrices provided by ePytope-TCR[40] (Supplementary S6), and observed DeepPROTECTNeo emerged as one of the top performing models on both of these datasets in every metrics.

To dissect the biophysical pattern encoded by DeepPROTECTNeo, we performed an exhaustive single-residue scan at the two central "hotspot" positions P4 and M5 across 20 amino-acid variants for each of 26 experimentally characterized TCRs on Mutation data (Fig. 6C). Each block is anchored by the unmodified peptide (gold border at P4→P; cyan border at M5→M) to serve as an internal baseline. In the P4 block, a subset of "P4-tolerant" clones (e.g. *CASSYQTGASYGYTF*, *CASSTPGSLNEQFF*) retains uniformly high scores (≈0.7-0.9) even when Proline is replaced by charged or bulky residues, whereas "P4-sensitive" receptors (e.g. *CASSFGQGVTQYF*, *CASTVGGTLQYF*) collapse to <0.2 upon any deviation from Proline, revealing strict dependence on that hydrogen-bonding donor. Conversely, in the M5 block charged or polar swaps (Asp, Lys) uniformly abrogate binding (scores<0.2), while hydrophobic substitutions (Val, Leu, Ile) preserve or enhance affinity (≈ 0.6–0.8), recapitulating the classical hydrophobic TCR pocket. Strikingly, many clones that tolerate P4 substitutions become exquisitely specific at M5 (and vice versa), shows that even adjacent side chains sculpt distinct complementarity-determining surfaces. We computed the Spearman rank correlation between model scores and experimentally measured activation scores (binarized by the thresholds of 66.09% and 40.0% for the neo-antigen and CMV datasets as given in the source study[40]) across all TCR-epitope pairs (Fig. 6D) on Mutation data. DeepPROTECTNeo achieved the strongest monotonic relationship, Spearman's $\rho=0.157$ ($p=2.4\times10^{-125}$), outperforming every other competing method: ERGO-AE ($\rho=0.116$),

ATM-TCR (ρ=0.098), NetTCR-2 (ρ=0.102), epiTCR (ρ=0.074), TEINet (ρ=0.071) and ERGO-LSTM (ρ=0.072). Further, we evaluated how DeepPROTECTNeo generalizes across diverse viral targets by computing per-epitope AUROCs for fourteen distinct 9-mer peptides on Viral data. In the heatmap (Fig. 6E) each row represents a model while each column represents a viral epitope, ranging from the immunodominant CMV peptide *NLVPMVATV* to rarer sequences like *FPQSAPHGV* and *CTELKSLDY* and cell values are five-fold average AUROCs. DeepPROTECTNeo (first row in the heatmap) maintains remarkably consistent performance, with AUROCs spanning only 0.53 to 0.74 and exceeding 0.60 on the majority of under-represented epitopes. By contrast, competing methods such as ERGO-AE and ATM-TCR exhibit greater volatility on frequent epitopes (e.g. ATM-TCR hits 0.78 on *NLVPMVATV*) but often fall below 0.55 on low-prevalence epitopes. Even NetTCR-2 and epiTCR while being moderately robust, show wider dips around 0.48-0.52 for several targets. Together, these reveals DeepPROTECTNeo's unique ability to capture both the fine-grained biophysical dependencies of TCR-epitope contacts and the broad, epitope-agnostic patterns required for real-world immunological prediction.

2.5 A Case-Study on Patient-specific Clinically Validated Cancer Cohort

To evaluate DeepPROTECTNeo's clinical relevance, we applied it to five TESLA subjects-three with melanoma (P1, P2, P3) and two with NSCLC (P12, P16) each containing matched WES and validated neoepitope data from the TESLA consortium[42]. After filtering out fusions and splicing events due to insufficient validation evidence, our analytical pipeline effectively identified 350 candidates out of the 532 potential peptide–MHC complexes evaluated by TESLA using NetMHCpan-based filtering based on SNV/indels only. Among these 350 candidates, 24 were among the 34 experimentally validated immunogenic epitopes in the TESLA dataset. We further analysed these 350 candidates by pairing each epitope with RNA-Seq derived TCR sequences from the same patient for TCR binding (detailed patient-wise

distribution of detected validated and non-validated epitope is given Supplementary S7). We obtained binding score for all possible TCR-epitope combinations, and each epitope was summarized using the 95th percentile of its TCR binding scores to captures strong interaction signals while minimizing the impact of outliers. A TCR-specificity criterion of 0.5 was applied to the 95th percentile scores, resulting in the retention of 252 peptide candidates as high-confidence neoepitopes (Supplementary S8). The core discriminative principle of DeepPROTECTNeo, in which retained epitopes consistently exhibit higher 95th percentile TCR scores than non-retained ones (Fig. 7A). This separation was statistically profound (Mann-Whitney U=24696, $p=8.1\times10^{-48}$, Cliff's $\delta=1.0$), demonstrating complete non-overlap of score distributions and substantiating the choice of the 0.5 threshold for epitope retention.

Further, we demonstrated DeepPROTECTNeo's 95th-percentile TCR specificity scores for 24 experimentally validated neoepitopes across five patients (Fig. 7B), ranging from 0.431 to 0.863 (median 0.578). For retaining peptides having high TCR specificity, we determined the cutoff score in two steps. First, for each epitope we summarized its predicted TCR-epitope interactions by taking the 95th percentile of all binding scores ("TCR_95P_Score"), followed by collecting those epitopes with TCR_95P_Score≤0.50 and computed the 95th percentile of that subset, yielding a final retention cutoff of 0.48 (Supplementary S9), which retained 3/7 in P1, 2/3 in P2, 9/10 in P3, 2/2 in P12, 2/2 in P16, resulting 18/24 (75%) overall (Supplementary S7). The retained versus non-retained distributions were perfectly separated (Mann–Whitney U=108, $p=1.5\times10^{-5}$, Cliff's $\delta=1.00$), confirming maximal discrimination at this cutoff. In summary, DeepPROTECTNeo calls 252 of 532 peptides, capturing 18/34 true neoepitopes (52.9% sensitivity) and rejecting 264/498 negatives (53.0% specificity), for a BA of 53.0% and ROC AUC of 0.526. Although specificity is modest, having experimental validation of only 34 epitopes and the strong TCR support for 252 candidates demands further experimental studies. By comparison, TSNAD v2.0's[43] MHC-only filter achieves

11.8% sensitivity and 96.0% specificity (BA 53.9%, AUC 0.487), while DeepNeo's[43] dual deep-learning model reaches 32.4% sensitivity and 87.2% specificity (BA 59.8%, AUC 55.6%), underscoring how patient-derived TCR data in DeepPROTECTNeo uniquely boosts recall without sacrificing discrimination. To investigate the statistical significance of these binding scores, we identified retained epitopes show a strong positive correlation between the number of supporting TCRs and their 95th-percentile score (ρ=0.77, p=2.8×10□□) (Fig. 7C), whereas non-retained peptides exhibit no such relationship, reinforcing that true immunogenicity arises from convergent, high-affinity TCR engagement rather than stochastic binding. Further, we examined per-patient score distributions (Fig. 7D), showing that validated epitopes consistently receive higher DeepPROTECTNeo scores in 4 out of 5 patients. Notably, P2 (p=1.9×10□□^{3}), P3 (p=1.1×10□^{1}□□), and P16 (p=8.2×10□^{23}□) showed extremely significant differences, and P1 reached moderate significance (p=2.0×10□□).

To assess how DeepPROTECTNeo's 95th percentile TCR binding scores (TCR_95P_Score) complement standard neoantigen features, we trained logistic regression (LR) and random forest (RF) models incorporating features such as MHC affinity, stability, hydrophobicity, foreignness, and the TCR_95P_Score. LR significantly outperformed RF (AUC=0.81 vs. 0.72). Crucially, TCR_95P_Score had the highest odds ratio (~1.50), surpassing foreignness, agreptopicity and MHC-binding features, indicating it is the strongest independent predictor of epitope validation, supporting the orthogonality and additive value of TCR-aware information in neoantigen prediction (Supplementary S10). UMAP projections of TCR–epitope interactions reveal distinct TCR clusters (Fig. 7E), particularly in P1, where five major clusters were observed (largest: 418 TCRs), indicating clonal expansion. TCRs binding validated epitopes showed significantly lower CDR3β entropy in 4 of 5 patients, P1 (3.25 vs. 3.03, p=4.0×10□^{123}), P2 (3.28 vs. 3.14, p=1.6×10□^{1}□), P3 (3.25 vs. 3.03, p=6.7×10□□□),

and P16 (3.24 vs. 3.04, $p=1.7\times10^{-2}$)—suggesting repertoire convergence toward immunogenic epitopes. Only P12 showed no significant difference (p = 0.49). These patterns validate DeepPROTECTNeo's capacity to capture immune selection dynamics. Together, these comprehensive results suggest DeepPROTECTNeo significantly enhances both the sensitivity and specificity of immunogenic epitope prioritization suggesting a critical advancement for designing effective cancer vaccines.

2.6 DeepPROTECTNeo Reveals Descriptor Hierarchies and Structural Contacts of TCR–epitope Binding

Using our model's interpretability, we investigate a clustered heatmap of normalized importance ranks for 82 common peptide descriptors across ImmuneCODE, Mutation, and Viral test sets (See Fig. 8A). Applying agglomerative hierarchical clustering with Euclidean distance and the average-linkage criterion to attention-derived ranks (Supplementary S11), reveals three clear descriptor families. Across all three cohorts DeepPROTECTNeo's attention scores consistently captured 14 descriptors (PRIN1/2, SVGER7/8, ProtFP2/3, AF3/5) to normalized importance >0.8, reflecting their universal role in capturing global physicochemical gradients[44], mid-range sequence-order couplings[45], and short-motif signatures. Conversely, the Mutation dataset specifically engaged specialized descriptors such as BLOSUM7, T3, VSTPV1, ST3 with moderate significance (0.5–0.7), aligning with their responsiveness to individual amino-acid alterations at canonical anchor sites[46,47]. Similarly, viral epitopes within the Viral set demonstrated an enrichment of hydrophobic and aromatic-frequency characteristics (VHSE4, ProtFP5, VHSE8, ProtFP7), suggesting that these attributes encapsulate conserved glycoprotein motifs[48] critical for the identification of CMV/EBV/Influenza. At the lower end, electronic Z-scale descriptors (E1–E4), 3D-shape WHIM features (MSWHIM3), and low-order SVGER terms (SV2/3) consistently recorded below 0.1 across all datasets, indicating that detailed electronic or inferred tertiary-structure

proxies provide negligible predictive value in a sequence-only context[44,49]. Pairwise Spearman correlations between ImmuneCODE, Mutation, and Viral importance ranks were all very high (ρ = 0.90-0.94, $p<10^{-3}$), indicating a common core of key descriptors with only a few dataset-specific outliers (Supplementary S12).

Further, we mapped DeepPROTECTNeo's epitope-side attention onto two representative TCR–epitope cocrystal structures (PDB ID: 3QDJ, 4MJI; Fig. 8B). In 3QDJ (left panel), five sub-8Å contacts: A2–F97 (7.40Å), L7–S94 (7.11Å), L7–L95 (3.55Å), L7–S96 (3.62Å) and L7–F97 (5.81Å) colocalize with the highest attention at P2 and P7. Distances of 3.5-3.6Å between L7 and L95/S96 reflect tight van der Waals packing and hydrogen bonding in the CDR3 loop[48], while the 5-7Å separation to F97 denotes stabilizing aromatic edge-to-face interactions[50]. Similarly, in 4MJI (right panel), four contacts under 8Å: I5–G98 (7.58Å), S7–L95 (3.62Å), S7–G98 (7.98Å) and I8–L95 (4.33Å), align with pronounced attention at mid-peptide positions, underscoring the model's sensitivity to sequence-order couplings and aromatic-frequency signals[51]. By recovering these sub-8Å hydrogen bonds, van der Waals contacts, and aromatic stacking purely from sequence attention, DeepPROTECTNeo captures the very molecular grammar that underlies TCR-peptide specificity and by extension, the key features leveraged in reverse-vaccinology epitope selection[52]. Together, the cross-dataset heatmap (Fig. 8A) and structural projections (Fig. 8B) demonstrate that DeepPROTECTNeo's sequence-only attention mechanism distils a minimal, biologically meaningful set of descriptors using interpretable feature prioritization with direct recapitulation of key noncovalent contacts, while offering both mechanistic insight into the molecular grammar of TCR recognition and robust predictive generalization across diverse epitope datasets.

3 Discussion

DeepPROTECTNeo is an end-to-end framework for neoantigen prioritization, integrating raw WES/WGS preprocessing, HLA typing, external MHC-peptide affinity scoring via NetMHCpan and a unique CDR3β-epitope binding prediction module within a unified pipeline. Two-branch structure of DeepPROTECTNeo combines backbone-constrained knowledge with sequence-derived determinants, helping to learn both long-range dependencies and local spatial correlations that are important for TCR-epitope pairing. Our model yields state-of-the-art predictive accuracy, strong early-retrieval of true binders and minimal inter-fold variability, while offering mechanistic insights that align with decades of structural immunology. DeepPROTECTNeo generalizes robustly across all three benchmarks and exhibits strong quantile concordance, achieving the highest alignment between ImmuneCODE and Viral (Pearson r = 0.99), substantial agreement between ImmuneCODE and Mutation (r = 0.91), and robust, if slightly lower, concordance between Mutation and Viral (r = 0.89) demonstrating both broad generalization and dataset specific calibration of predictive confidence (Supplementary S13). Central to DeepPROTECTNeo is its multi-scale interpretability: attention-guided motif extraction and descriptor-based embeddings converge on a concise set of biochemical signatures like global physicochemical axes (PRIN), sequence-order couplings (SVGER) and short-motif fingerprints (ProtFP) which consistently map onto atomic contacts within 8Å, thereby reconstructing the core TCR-peptide contacts purely from sequence.

Although DeepPROTECTNeo achieves strong discrimination of true binders, opportunities remain to expand its biological scope and contextual understanding. Incorporating broader TCR chain information (like α), unifying MHC-peptide prediction within an end-to-end framework, enriching antigen-processing context, and grounding negative sampling in structural or experimental evidence could further align the model with immunological

complexity. Continued advances in single-cell sequencing, deep structural modelling and structure-informed representation learning offer promising avenues to further enhance its biological fidelity and generalisation.

In conclusion, DeepPROTECTNeo represents a substantial advance in computational neoepitope prioritization for immunotherapy by integrating well-calibrated predictive model with mechanistic interpretability. Its multi-scale fusion and attention-based architecture enables accurate, biologically grounded feature prioritization, distilling a minimal set of descriptors that recapitulate key noncovalent contacts underlying TCR recognition. This combination of domain-informed design, strict cross-dataset benchmarking, and interpretable sequence-only modelling delivers both sensitivity and specificity gains in epitope ranking, supporting more effective cancer vaccine design. Moreover, the architecture provides a flexible foundation for integrating α/β TCR pairing, detailed antigen-processing events, experimentally derived negatives, and validated structural representations, positioning DeepPROTECTNeo as a platform for next-generation immunoinformatics.

4 Method

4.1 Data Processing

DeepPROTECTNeo begins with processing of raw WES/WGS FASTQ files using FastQC[53] for base quality, GC content and read length distributions, then adapters and low quality bases are trimmed using Trimmomatic[54]. High quality reads are aligned to the GRCh38/hg38 reference genome via HISAT2[55], and the resulting BAM files are processed with SAMtools[56] for filtering unmapped or low quality reads, sorting, marking duplicates, realigning indels and recalibrating base qualities to produce high fidelity alignments. Somatic variants (SNVs and indels) are called using GATK4 Mutect2[57] and annotated with ANNOVAR[58] against refGene[59], ensGene[60], gnomAD[61], ClinVar[62] and COSMIC[63], retaining

only exonic, non synonymous mutations for neoepitope generation. Patient HLA class I alleles are inferred directly from these BAMs using arcasHLA[13], while CDR3β repertoires are reconstructed in parallel with TRUST4[64]. Each candidate neoepitope spanning an identified mutation is then scored for peptide–MHC binding affinity with NetMHCpan[15], and high confidence neoantigens are selected via threshold-based filtering on binding affinity score, eluted ligand rank and binding affinity rank.

4.2 TCR-epitope Interaction prediction

DeepPROTECTNeo employs a dual-branch architecture that fuses sequence and physicochemical information for TCR–epitope binding prediction. The token branch embeds CDR3 and peptide sequences into 240-d vectors with learned positional encodings, then applies BiLSTM and attention pooling. Simultaneously, the skip branch processes 102-d descriptors through 32 Conv1D + SE blocks (channel width $C = 64$, kernel size 5) and optional self-attention, pooling intermediate and final maps into a 30-d skip summary. A gated fusion merges each 240-d token feature with its 30-d skip, a 6-head rotary cross-attention captures inter-sequence context, and the concatenated 720-d vector is refined by 6-head self-attention and a 24-layer ReZero RetNet before final MLP classification.

Let $T = (t_1, \dots, t_n), P = (p_1, \dots, p_m)$ be the CDR3β and epitope sequences over the 20 -amino-acid alphabet. We learn $f: (T, P) \rightarrow \{0,1\}$ to predict binding. Internally, T and P are mapped to continuous representations via embedding, positional encoding and feature projection thereby processed through parallel branches, fused and cross-attended, refined by a ReZero network and finally classified by a two-way MLP.

4.2.1 Token Branch

To convert discrete tokens into continuous representations, we use a shared embedding matrix of dimension $d = 240$, and we add learned positional encodings so the model knows each residue's absolute position along the sequence. Concretely, for the TCR:

$$\mathrm{X}^T = [E(t_1), \dots, E(t_L)] \in \mathrm{R}^{L\times d}, \mathrm{X}^T = \mathrm{X}^T + \mathrm{PosEnc}(T) \in \mathbb{R}^{n\times d}, \quad 1$$

and similarly for the epitope $\mathrm{X}^P \in \mathbb{R}^{m\times d}$. The positional embeddings ($\mathrm{PosEnc}\,(\mathrm{T/P}) \in \mathbb{R}^{L\times d}$) are trained jointly with the rest of the network, letting the model learn which positions are most informative.

X is fed through a two-layer bidirectional LSTM (hidden size $d/2$ per direction). Denote by

$$\vec{h}_i = LSTM_{\rightarrow}(x_i, \vec{h}_{i-1}), \overleftarrow{h}_i = LSTM_{\leftarrow}(x_i, \overleftarrow{h}_{i+1}), \quad 1$$

the forward and backward states at position i. We concatenate to form $\mathrm{h}_i = [\vec{\mathrm{h}}_i; \overleftarrow{\mathrm{h}}_i] \in \mathbb{R}^d$ and collect $\mathrm{H} = (\mathrm{h}_1, \dots, \mathrm{h}_L) \in \mathbb{R}^{L\times d}$ for P and T.

4.2.2 Skip Branch: DeepCNNAttnProjector and SkipGate

Handcrafted peptide descriptors from both TCR sequences and epitopes $F \in \mathbb{R}^{B\times 102}$ are first reshaped to $H^{(0)} = F.\mathrm{unsqueeze}(1) \in \mathbb{R}^{B\times 1\times 102}$. This tensor is passed through a stack of 32 residual Conv1D blocks. At block i, the input $H^{(i-1)}$ undergoes two Conv1D-BatchNorm-GELU steps and is added back to a skip connection:

$$Y = \mathrm{GELU}\left(\mathrm{BN}_1\left(\mathrm{Conv1D}\left(H^{(i-1)}\right)\right)\right), Z = \mathrm{BN}_2\left(\mathrm{Conv1D}(Y)\right), H^{(i)} \quad 2$$
$$= \mathrm{GELU}\left(Z + W_{\mathrm{skip}}^{(i)} H^{(i-1)}\right),$$

Here $W_{\mathrm{skip}}^{(i)}$ is the identity unless a 1×1 projection is needed to match dimensions. After 32 blocks, a 1×1 convolution reduces channels to $D = 60$:

$$U = Conv\,1 \times 1\left(H^{(32)}\right) \in \mathbb{R}^{B \times D \times L}, \quad 3$$

A channel attention module then re-weights these channels by first computing

$$z = GAP(U) \in \mathbb{R}^{B \times D}, s = \sigma\left(W_2 ReLU(W_1 z)\right) \in \mathbb{R}^{B \times D}, \quad 4$$

and scaling $U_{\mathrm{CA}} = U \odot s$.unsqueeze(-1). A multi-head self-attention layer refines this to U_{SA}. A final GAP produces the 60 -dim feature $t_{\text{final}} = \mathrm{GAP}(U_{\mathrm{SA}}) \in \mathbb{R}^{B \times 60}$.

Each intermediate feature map $H^{(30)}$ and $H^{(31)}$ is globally average pooled to $\mathbb{R}^{B \times 64}$, feature map t_{final} to $\mathbb{R}^{B \times 60}$. Each resulting vector is then passed through its own linear to produce three 30-dim skip vectors.

$$skip_j = W_j GAP\left(H^{(30+j)}\right) \in \mathbb{R}^{B \times 30}, j = 1,2;\ skip_{main} = W_m t_{final} \in \mathbb{R}^{B \times 30}, \quad 5$$

Concatenating these three 30-dim vectors yields $S \in \mathbb{R}^{B \times 90}$, followed by a final ReLU-MLP "SkipGate" reduces to the 30-dim summary

$$\text{skip}_{\text{summary}} = ReLU\left(W_g S + b_g\right) \in \mathbb{R}^{B \times 30} \quad 6$$

ready for fusion with the token-branch features.

4.2.3 Gated Fusion

Let $\mathbf{F}^t \in \mathbb{R}^{B \times d_{\text{model}}}$ be the batch of token-branch features and $\mathrm{F}^s \in \mathbb{R}^{B \times d_{\text{skip}}}$ the batch of skip-branch summaries. We first project both into the common d_{model} -dimensional space:

$$\tilde{F}^t = \tilde{F}^t W_t + b_t, \tilde{F}^s = \tilde{F}^s W_s + b_s, \quad 7$$

where $W_t, W_s \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ and $b_t, b_s \in \mathbb{R}^{d_{\text{model}}}$. Next, we compute an element-wise gate $\mathrm{g} \in (0,1)^{B \times d_{\text{model}}}$ from their concatenation:

$$g = \sigma\left([\tilde{F}^t; \tilde{F}^s] W_g + b_g\right), \quad 8$$

With $W_g \in \mathbb{R}^{2d_{\text{model}} \times d_{\text{model}}}$ and $b_g \in \mathbb{R}^{d_{\text{model}}}$. Finally, the fused feature $\mathbf{F}_{\text{fused}} \in \mathbb{R}^{B \times d_{\text{model}}}$ is a convex combination:

$$F_{\text{fused}} = g \odot \tilde{F}^t + (1 - g) \odot \tilde{F}^s, \tag{9}$$

allowing the network to dynamically balance sequence-level and physicochemical information on a per dimension basis.

4.2.4 Rotary Positional Embedding and Cross-Attention

Rotary Positional Embeddings (RoPE) was used to capture relative offsets by rotating each head's query and key vectors in the complex plane by an angle dependent on token position p. Concretely, for a d_k-dimensional vector x_p, we split even and odd indices and rotate:

$$\begin{aligned} \tilde{x}_{p,even} &= x_{p,even} cos\ \theta_p - x_{p,odd} sin\ \theta_p, \\ \tilde{x}_{p,odd} &= x_{p,even} sin\ \theta_p + x_{p,odd} cos\ \theta_p, \\ RoPE(x)_p &= \left[\tilde{x}_{p,even}; \tilde{x}_{p,odd}\right] \in \mathbb{R}^{d_k}, \end{aligned} \tag{10}$$

where θ_p is a fixed (or learned) rotation per position. Applying the same RoPE to both queries and keys ensures that their dot-product depends only on the relative distance $\Delta = p - q$.

Given fused TCR and peptide features $\mathrm{F}_f^T, \mathrm{F}_f^P \in \mathbb{R}^{B \times d_{\text{model}}}$, we project into multi-head queries, keys, and values:

$$Q_T = RoPE\left(F_f^T W_Q\right), K_P = RoPE\left(F_f^P W_K\right), V_P = F_f^P W_V, \tag{11}$$

and compute cross-attention in both directions:

$$A_{T \to P} = softmax\left(Q_T K_P^\top / \sqrt{d_k}\right) V_P, A_{P \to T} = softmax\left(Q_P K_T^\top / \sqrt{d_k}\right) V_T, \tag{12}$$

We then symmetrize by averaging:

$$F_{\text{cross}} = \frac{1}{2}\left(A_{T \to P} + A_{P \to T}\right) \in \mathbb{R}^{B \times d_{\text{model}}}, \tag{13}$$

This bidirectional cross-attention captures inter-sequence dependencies and relative positional relationships in a single, unified operation.

4.2.5 RetNet and Self-Attention

Let $Z^{(0)} = \left[F_f^T; F_f^P; F_{\text{cross}}\right] \in \mathbb{R}^{B\times 3d_{\text{model}}}$, denote the concatenated TCR-fused, peptide-fused, and cross-attended features. This tensor is pro two parallel streams. In the Enhanced RetNet stream, we apply L ReZero-style residual layers

$$Z^{(l+1)} = Z^{(l)} + \alpha_l ReLU\left(W_l LayerNorm\left(Z^{(l)}\right)\right),\ \alpha_l|_{\text{init}} = 0, l = 0, \dots, L-1 \qquad 14$$

Initializing each $\alpha_l = 0$ ensures that the network begins as an identity mapping, providing improved gradient flow and training stability, so that after L layers we get $F_{\text{retnet}} = Z^{(L)} \in \mathbb{R}^{B\times 3d_{\text{model}}}$.

Concurrently, the Extra Self-Attention stream applies a single multi-head attention block to the original $Z^{(0)}$ to obtain $F_{\text{SA}} = \text{SelfAttn}\left(Z^{(0)}\right) \in \mathbb{R}^{B\times 3d_{\text{model}}}$.

The final representation is obtained by element-wise summation of these two pathways:

$$F_{\text{final}} = F_{\text{retnet}} + F_{SA} \in \mathbb{R}^{B\times 3d_{\text{model}}}, \qquad 15$$

which is then fed into the downstream MLP followed by sigmoid for final classification.

4.3 Loss

To mitigate over-confidence and tolerate label noise, we train with a label-smoothed cross-entropy. Given a batch of N examples with ground-truth labels $y_i \in \{0,1\}$, the network produces logits $\mathbf{z}_i = (z_i^0, z_i^1)$ and probabilities

$$p_i^k = \frac{exp\ \left(z_i^k\right)}{exp\ \left(z_i^0\right) + exp\ \left(z_i^1\right)}, k \in \{0,1\}, \qquad 16$$

We then define a "soft" target distribution $\mathrm{q}_i = (q_i^0, q_i^1)$ by smoothing the one-hot label:

$$q_i^k = \begin{cases} 1-\varepsilon, & k = y_i, \\ \dfrac{\varepsilon}{K-1}, & k \neq y_i, \end{cases} \varepsilon = 0.1, \tag{17}$$

Hence, the final loss averaged over the batch is

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{1} q_i^k \, log \, p_i^k, \tag{18}$$

This label smoothing reduces the model's tendency to become over-confident on the training labels, improving generalization and calibration.

Data Availability

The raw data were downloaded from publicly available databases (McPAS-TCR[65], VDJdb[66], IEDB[67]). All the processed data used in this study for model building, validations and trained model weights are available via Zenodo (https://doi.org/10.5281/zenodo.16751524)[68]. For aligning raw sequencing files, we used Human Genome assembly GRCh38 available via HISAT2 Download page (https://daehwankimlab.github.io/hisat2/download/)[55]. We downloaded and installed NetMHCPan-4.1 for MHC-peptide binding which is available via DTU Health Tech (https://services.healthtech.dtu.dk/services/NetMHCpan-4.1/)[15]. Variant annotation using GATK4 Mutect2 required Homo Sapiens Assembly FASTA for best practices short variant discovery from WGS which was obtained via Google Cloud bucket. For annotating identified variations, all the reference databases were directly downloaded from ANNOVAR[58] website (https://annovar.openbioinformatics.org/en/latest/user-guide/download/). 3D crystal structure of TCR-MHC-peptide complex (PDB ID: 3QDJ and 4MJI) were downloaded PDB database (https://www.rcsb.org/). For clinical validation of our model, we used TESLA cohort which is available via Synapse (https://www.synapse.org/) with ID syn23446508. The complete DeepPROTECTNeo workflow is available as a web-server via https://cosmos.iitkgp.ac.in/DeepPROTECTNeo/.

Code Availability

The source code is available via GitHub (https://github.com/debraj-55555/DeepPROTECTNeo) under MIT License and via Zenodo (https://doi.org/10.5281/zenodo.16751524)[68].

Acknowledgements

This work is supported by the Department of Science and Technology and Biotechnology, Government of West Bengal, Kolkata - 700064 (Sanction Letter No: 2053(Sanc.)/STBT-11012(31)/1/2024-ST SEC Dated: 31-01-2024) and IIT KHARAGPUR AI4ICPS I HUB

FOUNDATION, IIT KHARAGPUR Campus, West Bengal - 721302 (Sanction Letter No: TRP3RD70001, Dated: 20-02-2024). Additionally, Debraj (PMRF Id: 2402784) thanks the Ministry of Education, Government of India, for the Prime Minister Research Fellowship.

Declaration of competing interest

The authors declare that they have no known competing interests.

Author contributions

D.D. and S.B. conceived the study. D.D., S.B, P.M. contributed to the conceptualization and methodology. D.D. and S.B. contributed to data curation, model implementation, validation and analysis with baselines, website development. D.D. along with S.B. conducted the validation on clinical data and neo-epitope extraction methodology design. A.P. supported the web development and data analysis. P.M. D.D., S.B. wrote the manuscript and prepared figures. P.M. provided supervision and contributed to interpretation of predictive results. All authors contributed to the original draft and critically revised the manuscript.

References

1. Blum, J. S., Wearsch, P. A. & Cresswell, P. *Pathways of Antigen Processing*. *Annual Review of Immunology* vol. 31 (2013).
2. Smith-Garvin, J. E., Koretzky, G. A. & Jordan, M. S. *T Cell Activation*. *Annual Review of Immunology* vol. 27 (2009).
3. Rosenberg, S. A. & Restifo, N. P. Adoptive cell transfer as personalized immunotherapy for human cancer. *Science (1979)* **348**, 62–68 (2015).
4. Chen, D. & Mellman, I. Oncology meets immunology: The cancer-immunity cycle. *Immunity* **39**, 1–10 (2013).
5. Shah, K., Al-Haidari, A., Sun, J. & Kazi, J. U. T cell receptor (TCR) signaling in health and disease. *Signal Transduct Target Ther* **6**, (2021).
6. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science (1979)* **348**, 69–74 (2015).

7. Yarchoan, M., Johnson, B. A., Lutz, E. R., Laheru, D. A. & Jaffee, E. M. Targeting neoantigens to augment antitumour immunity. *Nat Rev Cancer* **17**, 209–222 (2017).

8. Yarchoan, M., Johnson, B. A., Lutz, E. R., Laheru, D. A. & Jaffee, E. M. Targeting neoantigens to augment antitumour immunity. *Nat Rev Cancer* **17**, 209–222 (2017).

9. Ott, P. A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217–221 (2017).

10. Sahin, U. & Türeci, Ö. Personalized vaccines for cancer immunotherapy. *Science (1979)* **359**, 1355–1360 (2018).

11. Yadav, M. *et al.* Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515**, 572–576 (2014).

12. Robbins, P. F. *et al.* Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat Med* **19**, 747–752 (2013).

13. Orenbuch, R. *et al.* ArcasHLA: High-resolution HLA typing from RNAseq. *Bioinformatics* **36**, (2020).

14. Szolek, A. *et al.* OptiType: Precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310–3316 (2014).

15. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. 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).

16. O'Donnell, T. J. *et al.* MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Syst* **7**, 129-132.e4 (2018).

17. Dash, P. *et al.* Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).

18. Luksza, M. *et al*. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* **551**, 517–520 (2017).

19. Glanville, J. *et al*. Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).

20. Dash, P. *et al*. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).

21. Moris, P. *et al*. Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Brief Bioinform* **22**, (2021).

22. Montemurro, A. *et al.* NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. *Commun Biol* **4**, (2021).

23. Weber, A., Born, J. & Rodriguez Martínez, M. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics* **37**, (2021).

24. Springer, I., Besser, H., Tickotsky-Moskovitz, N., Dvorkin, S. & Louzoun, Y. Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs. *Front Immunol* **11**, (2020).

25. Sidhom, J.-W., Larman, H. B., Pardoll, D. M. & Baras, A. S. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nat Commun* **12**, (2021).

26. Cai, M., Bang, S., Zhang, P. & Lee, H. ATM-TCR: TCR-Epitope Binding Affinity Prediction Using a Multi-Head Self-Attention Model. *Front Immunol* **13**, (2022).

27. Chen, J. *et al*. TEPCAM: Prediction of T-cell receptor–epitope binding specificity via interpretable deep learning. *Protein Science* **33**, (2024).

28. Lu, T. *et al*. Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nat Mach Intell* **3**, 864–875 (2021).

29. Jokinen, E., Huuhtanen, J., Mustjoki, S., Heinonen, M. & Lähdesmäki, H. Predicting recognition between T cell receptors and epitopes with TCRGP. *PLoS Comput Biol* **17**, (2021).

30. Zhang, M. *et al.* BertTCR: a Bert-based deep learning framework for predicting cancer-related immune status based on T cell receptor repertoire. *Brief Bioinform* **25**, (2024).

31. Kwee, B. P. Y. *et al.* STAPLER: Efficient learning of TCR-peptide specificity prediction from full-length TCR-peptide data. Preprint at https://doi.org/10.1101/2023.04.25.538237 (2023).

32. Zhang, P., Bang, S., Cai, M. & Lee, H. Context-Aware Amino Acid Embedding Advances Analysis of TCR-Epitope Interactions. Preprint at https://doi.org/10.7554/eLife.88837.1 (2023).

33. Rossjohn, J. *et al. T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules*. *Annual Review of Immunology* vol. 33 (2015).

34. Arstila, T. P. *et al*. A direct estimate of the human αβ T cell receptor diversity. *Science (1979)* **286**, 958–961 (1999).

35. Jiang, Y., Huo, M. & Li, S. C. TEINet: a deep learning framework for prediction of TCR–epitope binding specificity. *Brief Bioinform* **24**, (2023).

36. Pham, M.-D. N. *et al.* epiTCR: a highly sensitive predictor for TCR–peptide binding. *Bioinformatics* **39**, (2023).

37. Nolan, S. *et al*. 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) doi:10.21203/rs.3.rs-51964/v1.

38. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. in *34th International Conference on Machine Learning Icml 2017* vol. 3 2130–2143 (2017).

39. Drost, F. *et al*. Predicting T cell receptor functionality against mutant epitopes. *Cell Genomics* **4**, (2024).

40. Drost, F. *et al*. Benchmarking of T cell receptor-epitope predictors with ePytope-TCR. *Cell Genomics* (2025) doi:10.1016/j.xgen.2025.100946.

41. Kocher, K. *et al*. Quality of vaccination-induced T cell responses is conveyed by polyclonality and high, but not maximum, antigen receptor avidity. Preprint at https://doi.org/10.1101/2024.10.30.620795 (2024).

42. Wells, D. K. *et al*. Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. *Cell* **183**, (2020).

43. Kim, J. Y., Bang, H., Noh, S.-J. & Choi, J. K. DeepNeo: A webserver for predicting immunogenic neoantigens. *Nucleic Acids Res* **51**, W134–W140 (2023).

44. Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M. & Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J Med Chem* **41**, 2481–2491 (1998).

45. Dubchak, I., Muchnik, I., Holbrook, S. R. & Kim, S.-H. Prediction of protein folding class using global description of amino acid sequence. *Proc Natl Acad Sci U S A* **92**, 8700–8704 (1995).

46. Birnbaum, M. E. *et al*. Deconstructing the peptide-MHC specificity of t cell recognition. *Cell* **157**, 1073–1087 (2014).

47. Wooldridge, L. *et al*. A single autoimmune T cell receptor recognizes more than a million different peptides. *Journal of Biological Chemistry* **287**, 1168–1177 (2012).

48. Rudolph, M. G., Stanfield, R. L. & Wilson, I. A. *How TCRs Bind MHCs, Peptides, and Coreceptors*. *Annual Review of Immunology* vol. 24 (2006).

49. Abelin, J. G. *et al.* Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* **46**, 315–326 (2017).

50. Hunter, C. A. & Sanders, J. K. M. The Nature of π-π Interactions. *J Am Chem Soc* **112**, 5525–5534 (1990).

51. Vig, J. *et al.* BERTOLOGY MEETS BIOLOGY: INTERPRETING ATTENTION IN PROTEIN LANGUAGE MODELS. in *Iclr 2021 9th International Conference on Learning Representations* (2021).

52. Masignani, V., Rappuoli, R. & Pizza, M. Reverse vaccinology: A genome-based approach for vaccine development. *Expert Opin Biol Ther* **2**, 895–905 (2002).

53. Wingett, S. W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res* **7**, (2018).

54. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, (2014).

55. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).

56. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, (2009).

57. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. *bioRxiv* (2019).

58. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, (2010).

59. National Center for Biotechnology Information. RefSeq: NCBI Reference Sequence Database. *U.S. National Library of Medicine* Preprint at (2018).

60. Harrison, P. W. *et al.* Ensembl 2024. *Nucleic Acids Res* **52**, (2024).

61. Gudmundsson, S. *et al.* Variant interpretation using population databases: Lessons from gnomAD. *Hum Mutat* **43**, 1012–1030 (2022).

62. Landrum, M. J. *et al.* ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44**, (2016).

63. Bamford, S. *et al.* The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer* **91**, (2004).

64. Song, L. *et al.* TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. *Nat Methods* **18**, 627–630 (2021).

65. Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E. & Friedman, N. McPAS-TCR: A manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* **33**, (2017).

66. Shugay, M. *et al.* VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res* **46**, D419–D427 (2018).

67. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2024 update. *Nucleic Acids Res* **53**, D436–D443 (2025).

68. Das, D. Data for DeepPROTECTNeo: A Deep learning-based Personalized and RV-guided Optimization tool for TCR Epitope interaction using Context-aware Transformers . Preprint at https://doi.org/10.5281/zenodo.16751524 (2025).

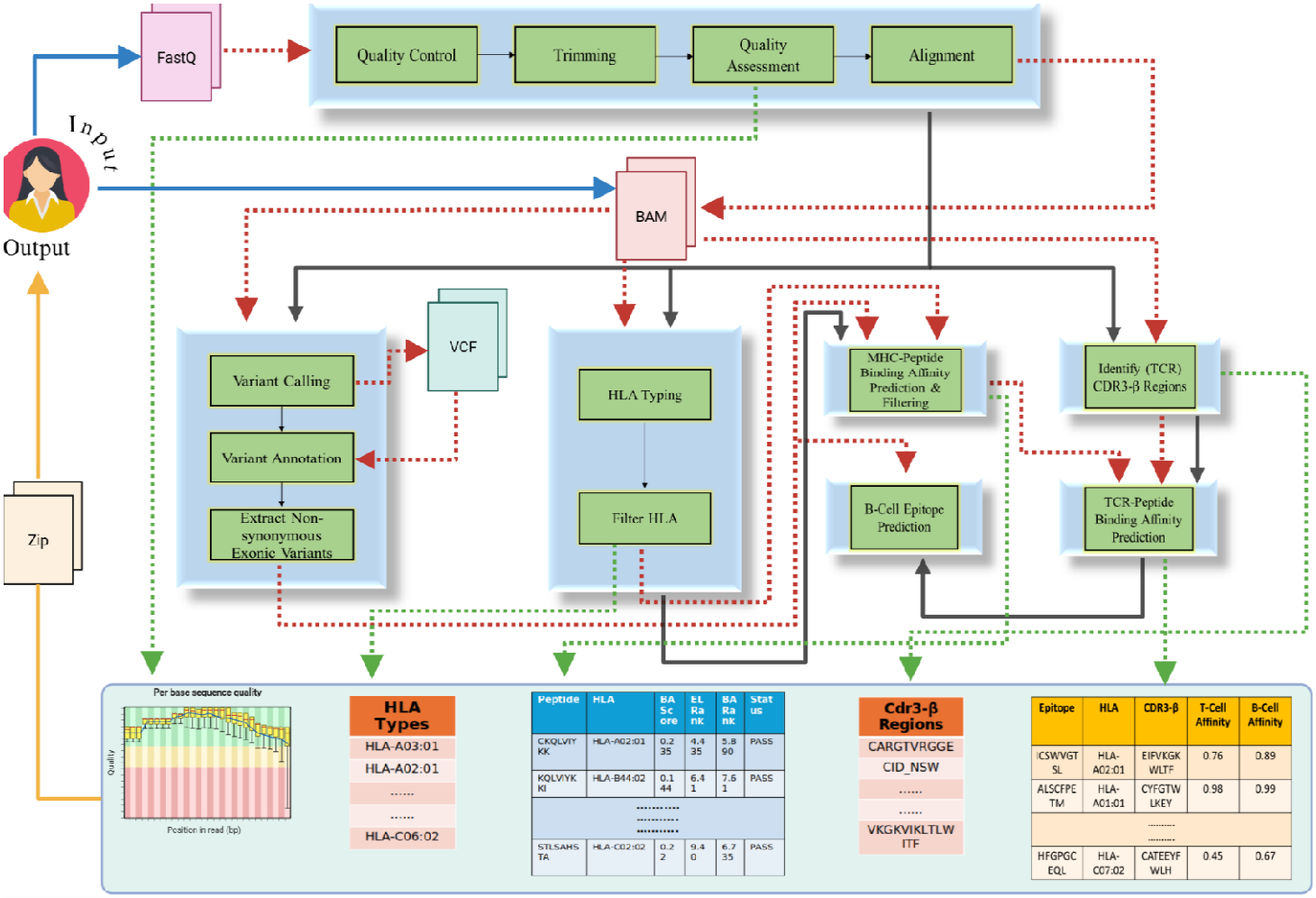

Figure 1. Integrated DeepPROTECTNeo neoantigen discovery workflow. Solid blue arrows denote user inputs (raw FASTQ or BAM or candidate CDR3β–epitope pairs); solid black arrows show core processing steps—QC/trimming/alignment → variant calling & annotation → HLA typing → CDR3β extraction. Candidate peptides are scored for MHC binding (NetMHCpan) and B-cell epitopes, then matched with CDR3β sequences for DeepPROTECTNeo binding prediction. Red dashed lines trace intermediate data transfers (VCFs, HLA calls, CDR3β lists); green dashed lines indicate final data outputs (QC reports, HLA alleles, neoepitope and repertoire tables, binding scores); solid orange arrow marks the ZIP archive delivered to the user.

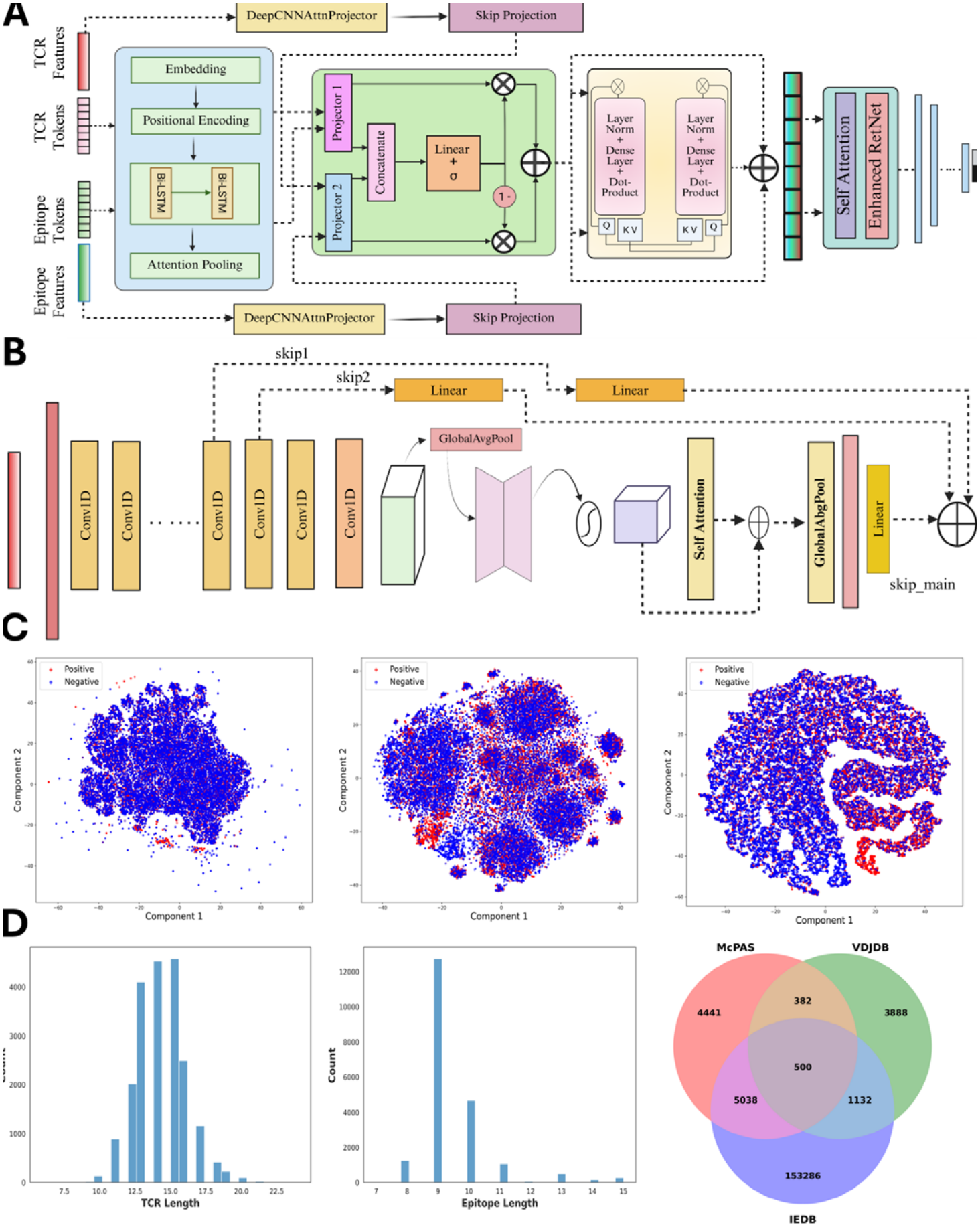

Figure 2. DeepPROTECTNeo architecture, latent□space embeddings and dataset. A)

Paired TCR (20 aa) and epitope (11 aa) tokens plus 102-dim physicochemical descriptors enter two streams: a token branch that embeds into □^(B×L×240), applies learned positional encodings, a 2-layer BiLSTM and attention pooling to yield 240-dim features; and a skip branch that runs 32 Conv1D + Channel Attention (CA) blocks (C=64, ks=5) with optional self-attention over descriptors, pools to a 60-dim vector, extracts two intermediate maps and t_final into three 30-dim skip vectors, and gates them into a single 30-dim summary. A GatedFusion per side produces t_fused, p_fused ∈ □^(B×240); a 6-head rotary cross-attention yields x_fused ∈ □^(B×240); concatenation [t_fused; p_fused; x_fused] ∈ □^(B×720) is refined by 6-head self-attention plus a 24-layer ReZero RetNet and classified by a 2-way MLP. **B)** Skip branch detail. Shows Conv1D block outputs (B×64×51), 1×1 Conv→CA, global average pooling to t_final (B×60), two intermediate skip-map projections and t_final projections to 30-dim, concatenation to 90 dim, and gating to 30 dim. **C)** t-SNE embeddings. Positive (red) vs negative (blue) test pairs projected into 2D using features from (left) concatenated fusion vector [t_fused; p_fused; x_fused] before extra self-attention, (middle) cross-attention x_fused before extra self-attention, and (right) final post-RetNet outputs—demonstrating progressively clearer class separation. **D)** Histogram of TCR β-chain CDR3 lengths distribution (left), histogram of epitope sequence lengths (middle), and Venn diagram of unique TCR–epitope pair overlap among IEDB (Class I & II), McPAS, and VDJdb (right).

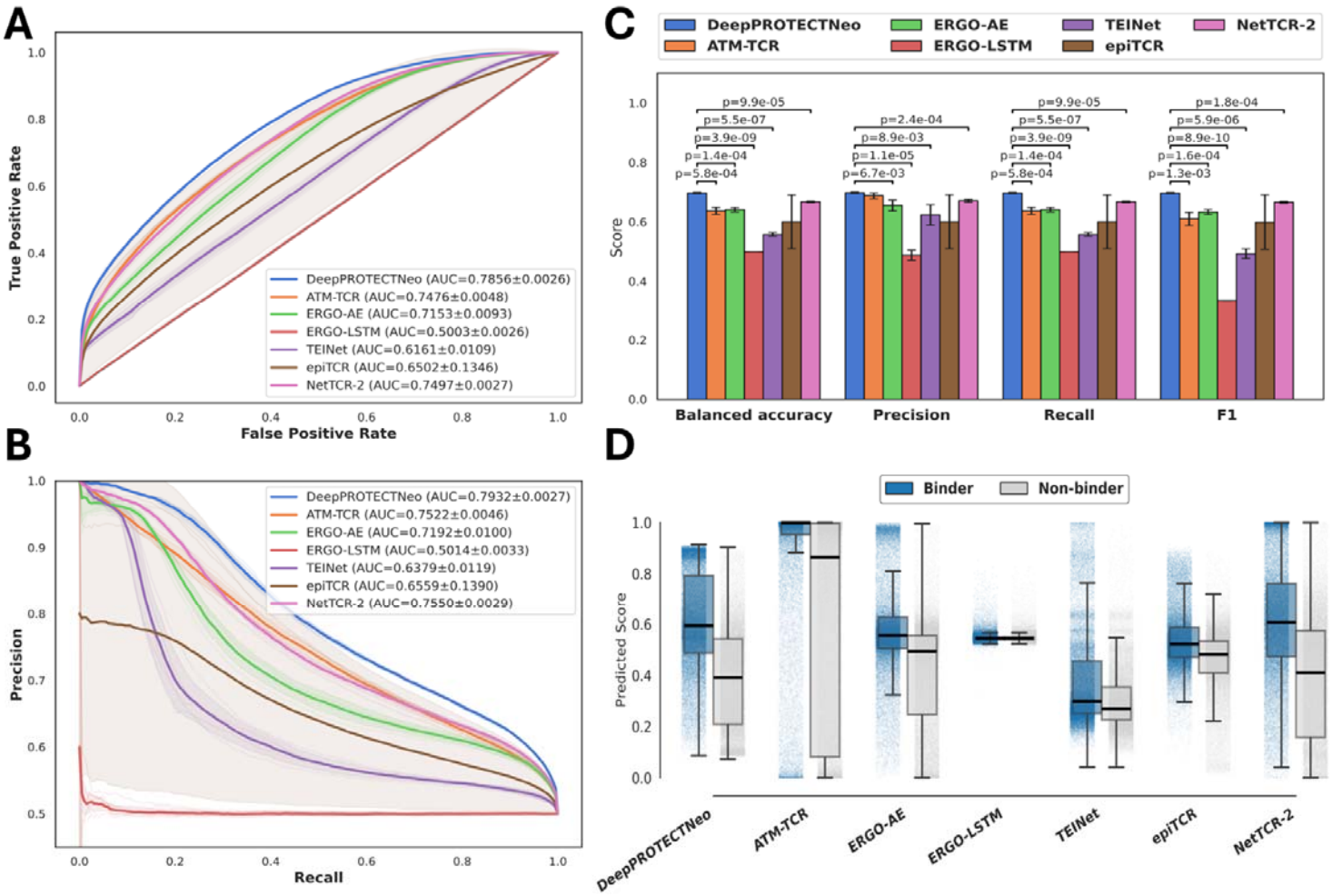

Figure 3. Benchmarking DeepPROTECTNeo against CDR3β-only TCR–epitope predictors under stringent TCR-split over 5 folds. **A)** Mean receiver-operating-characteristic (ROC) curves (true□positive rate versus false□positive rate) for DeepPROTECTNeo (blue), ATM-TCR (orange), ERGO-AE (green), ERGO-LSTM (red), TEINet (purple), epiTCR (brown) and NetTCR-2 (pink) computed over five independent folds. Shaded bands denote ±1□s.d. across folds; legend reports mean□AUROC□±□s.d. **B)** Mean precision–recall (PR) curves (precision versus recall) for the same models and folds, with shaded bands = ±1□s.d. and mean□AUPRC□±□s.d. in the legend. **C)** Fold-averaged balanced accuracy, macro-precision, macro-recall and macro-F1 at a fixed decision threshold. Bars = mean over five folds; error bars = ±1□s.d. Pairwise significance versus DeepPROTECTNeo was determined by paired two-tailed t-tests (exact p-values shown for significant ones). **D)** Distribution of per-interaction predicted scores for binder (blue) and

non-binder (grey) classes across all models. Boxplots show median (central line), 25th–75th percentiles (box limits) and 1.5×□IQR whiskers; overlaid jittered points (30% subsample) illustrate score density. All analyses are based on a unified dataset of 140,992 unique CDR3β–epitope pairs evaluated under five TCR-split folds.

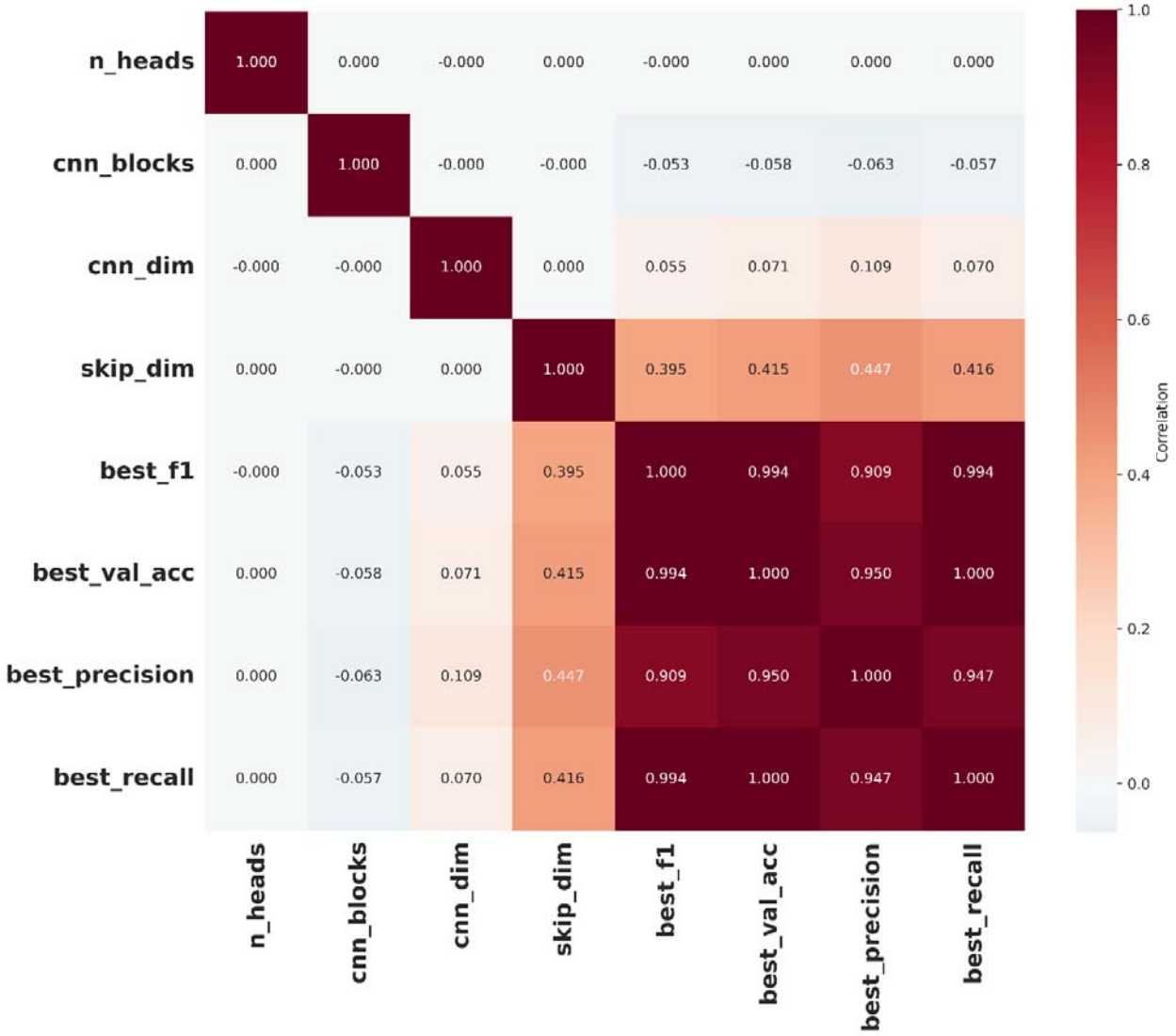

Figure 4. Correlation heatmap displaying relationships between model hyperparameters (n_heads, cnn_blocks, cnn_dim, skip_dim) and performance metrics (best_f1, best_val_acc, best_precision, best_recall), highlighting strong positive correlations between key evaluation scores and skip_dim, while revealing little or no correlation between hyperparameters and performance outcomes. The colour scale represents the degree of linear association, with darker shades indicating higher correlations.

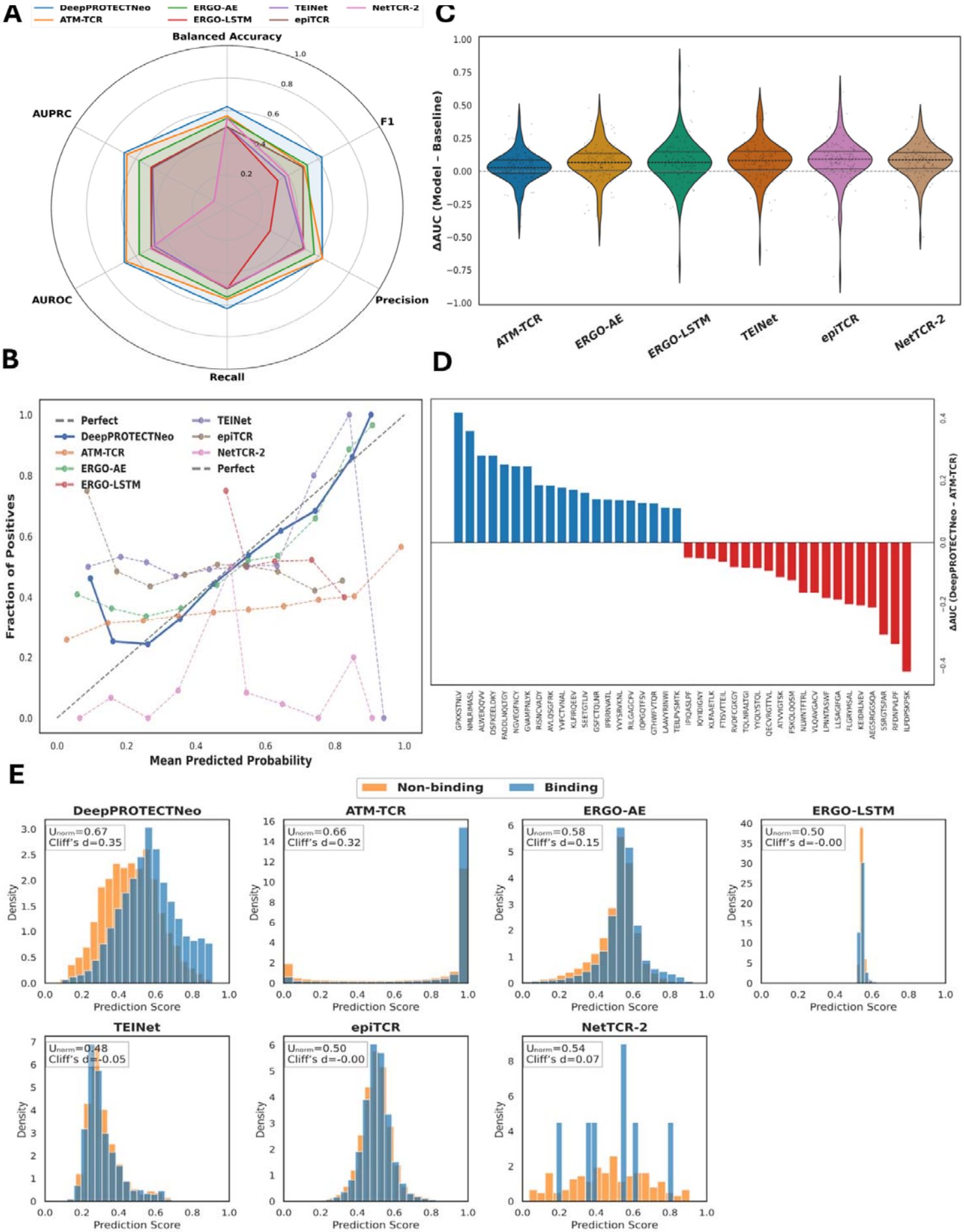

Figure 5. Performance evaluation of DeepPROTECTNeo on the ImmuneCODE

benchmark. A) Radar plot of six key metrics balanced accuracy, F1, precision, recall, AUROC and AUPRC, showing the fold□averaged mean for DeepPROTECTNeo (blue) and six comparator models. **B)** Reliability (calibration) diagram: observed binder fraction versus mean predicted probability in ten uniform bins for each model; the gray dashed line is perfect calibration. **C)** Violin plots of per□epitope ΔAUC (DeepPROTECTNeo minus each baseline) with individual points and median (dashed line). **D)** Waterfall of the 20 epitopes with the largest ΔAUC gains (blue; median +0.174, IQR 0.111, max +0.420) and losses (red; median −0.143, IQR 0.120, min −0.420) compared to the closest ATM-TCR. **E)** Overlaid histograms of prediction scores for true binders (blue) and non□binders (orange), annotated with normalized Mann–Whitney U ($U_{norm} = U/(m \cdot n)$, where U is the test statistic and m, n are sample sizes) and Cliff's δ, illustrating DeepPROTECTNeo's superior class separation ($U_{norm} = 0.67$, $d = 0.35$).

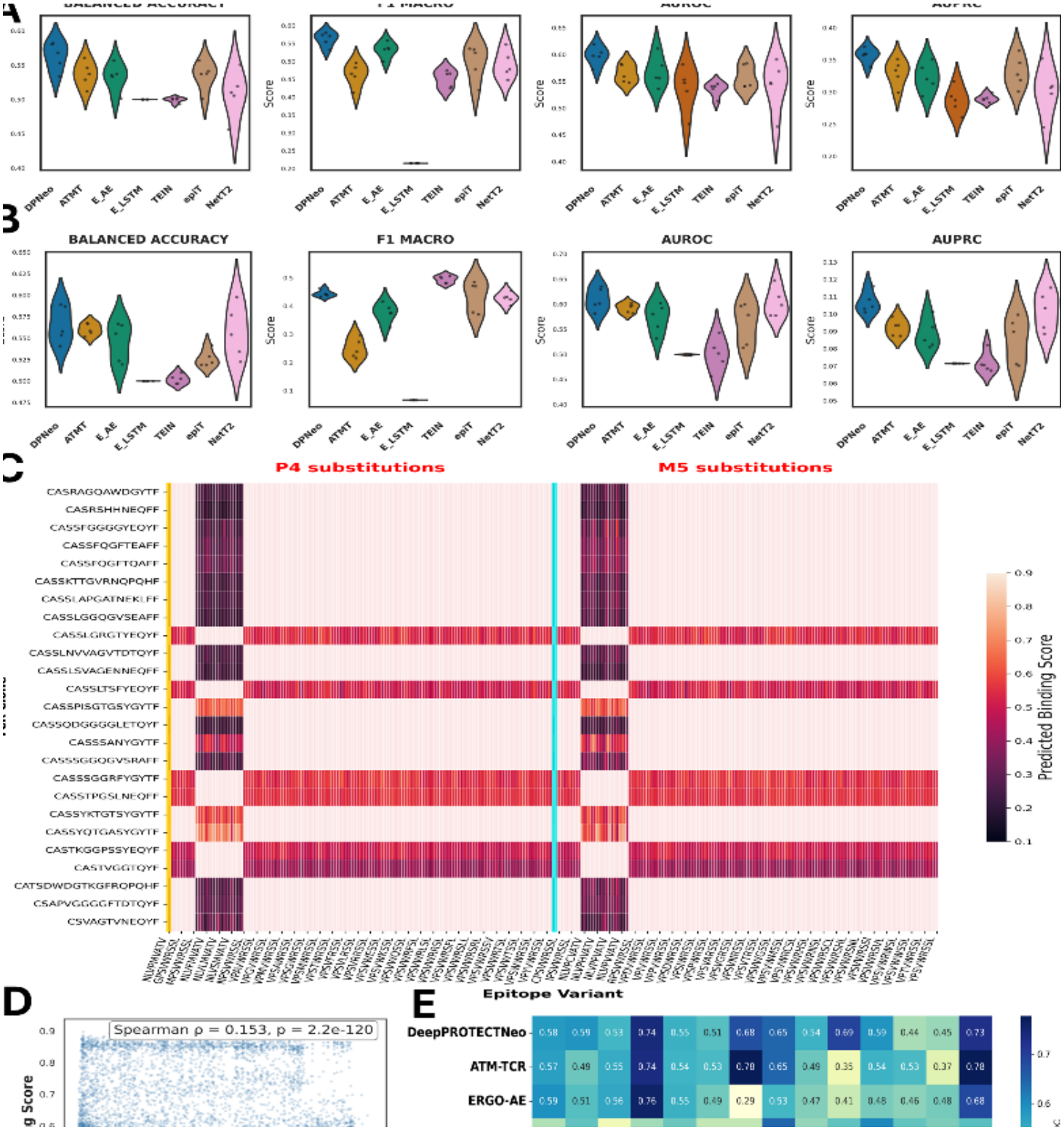

Figure 6. Comprehensive Performance Assessment and Biophysical Insights of

DeepPROTECTNeo on Mutation and Viral datasets. A) Violin plots of five-fold balanced accuracy, F1-macro, AUROC and AUPRC on the engineered mutation dataset (4244 TCR–peptide pairs across 26 TCR clones). DeepPROTECTNeo (DPNeo) is shown as the leftmost (blue) violin in each panel, followed by ATM-TCR (ATMT), ERGO-AE (E_AE), ERGO-LSTM (E_LSTM), TEINet (TEIN), epiTCR (epiT) and NetTCR-2 (NetT2); individual fold scores are overlaid as black dots. **B)** Corresponding performance on the Viral dataset (8932 TCR–epitope interactions from CMV, EBV, and influenza). **C)** Two-anchor mutation-scan heatmap on Mutation data: DeepPROTECTNeo's mean predicted binding for every P4 (left) and M5 (right) substitution across 26 TCR clones. Wild-type columns are outlined in gold (P4→P) and cyan (M5→M); colour scale from low (dark) to high (pale) affinity. **D)** Scatterplot of predicted binding versus experimental activation scores on Mutation data, with Spearman's $\rho = 0.153$ and two-tailed $p = 2.2 \times 10^{-12}$ indicated, demonstrating a significant monotonic relationship. **E)** Epitope-wise AUROC heatmap on Viral data: five-fold average AUROCs for seven models across fourteen viral 9-mer peptides; deeper blue denotes superior discrimination.

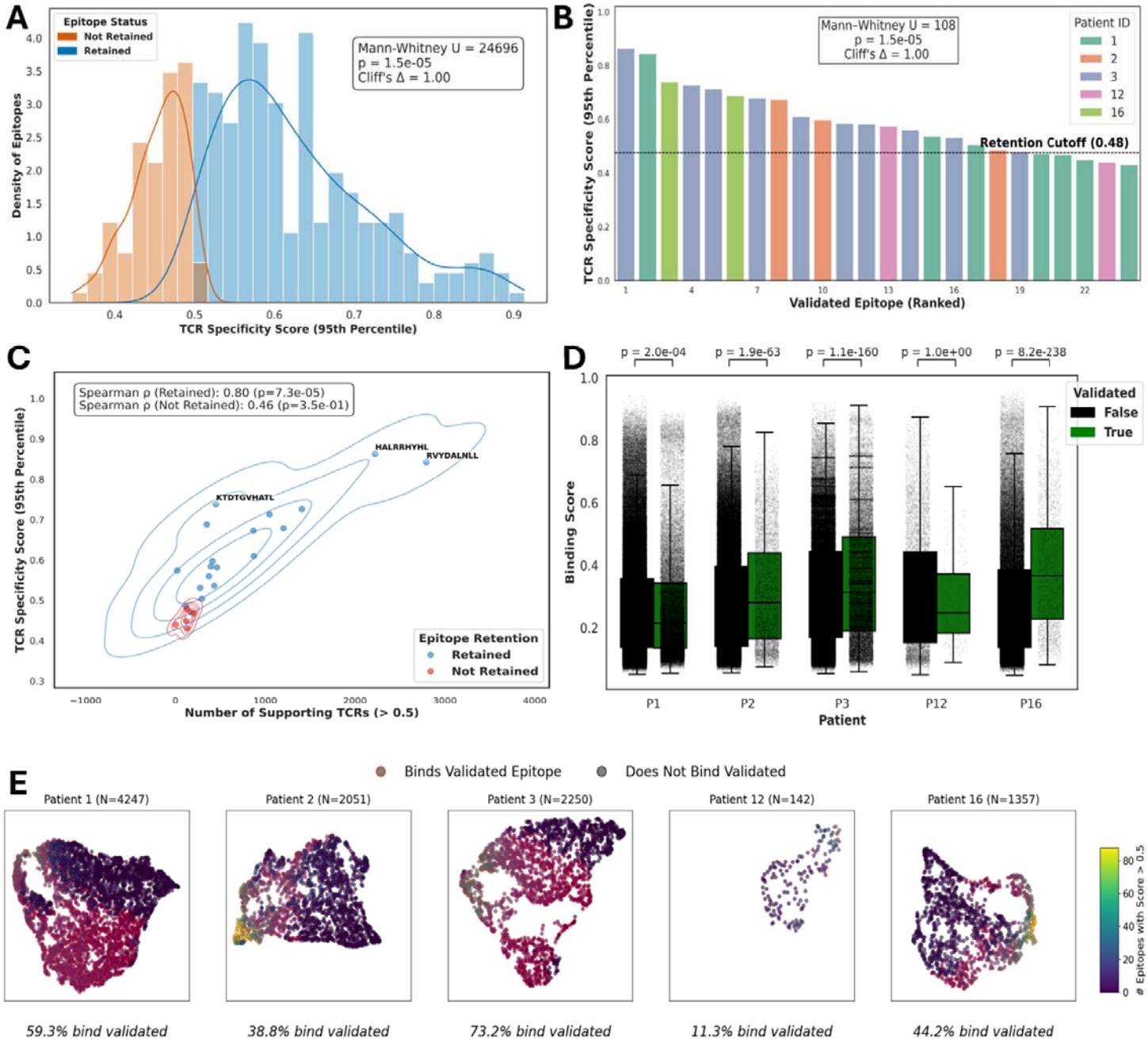

Figure 7. **DeepPROTECTNeo captures immunogenic epitopes in TESLA patients via TCR-aware prediction. A)** Distribution of DeepPROTECTNeo-predicted TCR specificity scores (95th percentile per epitope) for validated (blue) and non-validated (orange) neoepitopes across five TESLA patients. Validated epitopes displayed significantly higher scores (Mann–Whitney U = 24,696, p = 8.1×10□□□, Cliff's δ=1.00), supporting strong immunogenic signal capture. **B)** Ranking of 24 validated epitopes by TCR specificity score,

stratified by patient. A threshold of 0.48 (dashed line) retained 18 of 24 epitopes (70.83% recall) with a large effect size (Cliff's δ= 1.00). **C)** Association between epitope TCR specificity and the number of strongly interacting TCRs (*score > 0.5*) for 24 validated epitopes. A strong positive Spearman correlation was observed for retained epitopes (ρ = 0.80, p = 7.3×10□□), but not for non-retained ones (ρ = 0.46, p = 0.35), indicating that immunogenic peptides tend to recruit larger numbers of high-affinity TCRs. Contour density overlays show distinct spatial clustering patterns between retained (blue) and non-retained (red) epitopes. The top three epitopes by TCR specificity score: KTDTGVHATL, HALRRHYHL, and RYVDALNLL are annotated, illustrating the immunological prominence of these sequences within the cohort. **D)** Per-patient comparison of DeepPROTECTNeo binding scores between validated and non-validated epitopes. Validated epitopes consistently exhibited significantly higher scores across four patients (Mann–Whitney U test; p < 2×10□□; Bonferroni-corrected), with P3 (p = 1.1×10□^{1}□□) and P16 (p = 8.2×10□^{23}□) showing the strongest effect. **E)** UMAP projections of patient-specific TCRs coloured by number of epitopes bound with high affinity (*score > 0.5*). TCRs binding at least one validated epitope (red edges) formed enriched spatial clusters, particularly in P1 and P3, with validated binding rates ranging from 11.3% (P12) to 73.2% (P3).

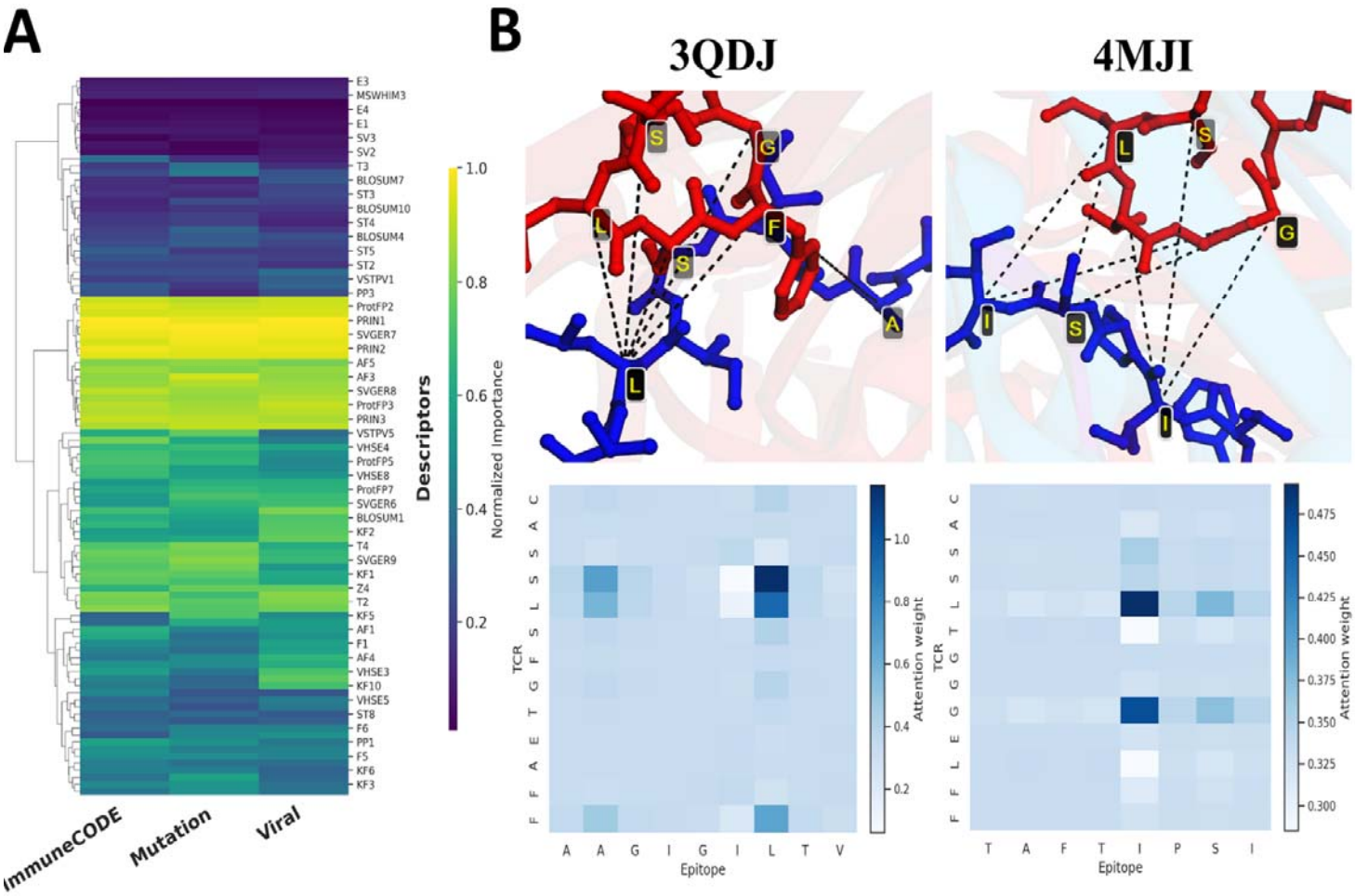

Figure 8. Cross□dataset descriptor prioritization and structural attention mapping. A) Non-zero–variance descriptors from ImmuneCODE, Mutation and Viral sets were ranked by attention importance, merged, normalized, and clustered (Ward's linkage). The bright yellow block highlights 14 core features (PRIN1/2, SVGER7/8, ProtFP2/3, AF3/5) consistently prioritized across all datasets; mid-green bands denote dataset-specific specialists; deep-purple tails are low-variance or pruned descriptors. **B)** Top: Two representative TCR–peptide cocrystal structures (PDB 3QDJ, 4MJI) with peptide side chains colored by normalized attention (blue: epitope, red: TCR) and dashed lines indicating sub-8□Å contacts. Bottom: Corresponding sequence□level attention heatmaps, recapitulating the structural hotspots captured by DeepPROTECTNeo.