TITANiAN: Robust Prediction of T-cell Epitope Immunogenicity using Adversarial Domain Adaptation Network

Jeonghyeon Kim[1], Nuri Jung[1], Jayyoon Lee[2], Nam-Hyuk Cho[3], Jinsung Noh[4*], Chaok Seok[1,4*]

[1*]Department of Chemistry, Seoul National University, Gwanak-gu, 08826, Seoul, Republic of Korea.
[2]Graduate School of Data Science, Seoul National University, Gwanak-gu, 08826, Seoul, Republic of Korea.
[3]Department of Biomedical Sciences, Seoul National University College of Medicine, Jongro-gu, 03080, Seoul, Republic of Korea.
[4*]Galux Inc., Gwanak-gu, 08738, Seoul, Republic of Korea.

*Corresponding author(s). E-mail(s): js.noh@galux.co.kr; chaok@snu.ac.kr;
Contributing authors: jasonkjh@snu.ac.kr; jnooree@snu.ac.kr; lee.jayyoon@snu.ac.kr; chonh@snu.ac.kr;

Abstract

T-cell immunogenicity, the ability of peptide fragments to elicit T-cell responses, is a critical determinant of the safety and efficacy of protein therapeutics and vaccines. While deep learning offers promise for *in silico* prediction, the scarcity of comprehensive immunogenicity data remains a major challenge. We present TITANiAN, a novel multi-domain deep learning framework that leverages adversarial domain adaptation to integrate diverse immunologically relevant data sources, including MHC presentation, pMHC binding affinity, TCR-pMHC interaction, T-cell activation, and source organism information. Validated through rigorous leakage-controlled benchmarks, TITANiAN demonstrates exceptional performance in its primary fine-tuned task (TITANiAN-IM), predicting T-cell activation for specific peptide-MHC pairs. Remarkably, it also predicts ADA-inducing potential of therapeutic antibodies without requiring MHC inputs. This success attributed to TITANiAN's biologically grounded

and data-driven multi-domain pre-training, with component contributions further confirmed by extensive ablation studies. In addition, TITANiAN supports broader immunogenicity-related tasks such as molecular binding prediction, enabled by its flexible pretraining architecture. TITANiAN's consistent and robust performance highlights its potential to advance the development of safer and more effective vaccines and protein therapeutics. The model is publicly accessible via a web server: https://galaxy.seoklab.org/design/titanian/.

Keywords: T-cell immunogenicity, T-cell epitope, Deep learning, Adversarial multi-domain learning, Immunotherapy, Vaccine, Anti-drug antibody

1 Introduction

T-cell epitope immunogenicity, the ability of short peptide fragments to trigger an immune response by T cells, is a critical factor in adaptive immunity, influencing both protective and pathogenic immune responses.[1–3] Understanding the precise immunogenicity of T-cell epitopes has broad implications for various biomedical applications, including the design of effective immunotherapies against cancer, the development of vaccines for infectious diseases [1, 3, 4], and strategies to predict and mitigate unwanted immunogenicity in protein-based drugs.[2, 5–7]

Traditional experimental methods for identifying T-cell epitopes, such as screening overlapping peptides derived from protein antigens, are costly, time-consuming, and often fail to pinpoint the exact boundaries and immunogenic potency of epitopes. [8, 9] While *in silico* prediction methods offer a more efficient alternative for T-cell epitope discovery, most current approaches rely on limited data integration, focusing primarily on individual aspects of the T-cell recognition process, such as MHC-peptide binding or TCR-pMHC interaction. [10, 11] This focus on isolated components, rather than the full integrated process, limits both predictive accuracy and generalizability, hindering the development of reliable tools for practical immunogenicity assessment and therapeutic design.

To overcome these limitations, we introduce TITANiAN, a novel deep learning framework designed to integrate diverse immunogenicity-related data and learn the complex determinants governing T-cell epitope immunogenicity. By employing a modular multi-task learning architecture, TITANiAN leverages information from multiple domains, including MHC-peptide interactions, TCR-peptide-MHC binding, and T-cell activation, to build a more comprehensive and predictive model of T-cell immunogenicity.

2 Results

2.1 Model Architecture and Training Strategy

To effectively integrate diverse resources pertinent to T-cell immunogenicity and capture its complex determinants, TITANiAN employs a modular multi-task learning architecture. This architecture comprises a shared integrative encoder that learns

domain-invariant features common across tasks, alongside task-specific encoders that capture complementary, domain-specific information (Fig. 1b, d, f). To promote disentanglement, orthogonality constraints are applied between encoders, encouraging the separation of shared and unique representations. Furthermore, an adversarial loss is employed to guide the integrative encoder toward producing domain-invariant representations by training it to confuse a discriminator tasked with identifying the data source. This mechanism mitigates the risk of overfitting to dataset-specific artifacts—such as peptide length discrepancies between MHC class I and II—and instead promotes learning of fundamental, immunogenicity-relevant features. The model training follows a two-stage process: multi-task pre-training and task-specific fine-tuning.

During the initial multi-task pretraining phase, TITANiAN is trained simultaneously on diverse immunologically relevant datasets, each aligned with a specific task designed to consolidate broad biological knowledge. Peptide–MHC interaction data, representing the early steps of antigen presentation, are relatively abundant. For example, eluted ligand mass spectrometry (MS) data—such as those from IEDB [12], comprising 557,151 class I and 206,055 class II entries and commonly used by models like BigMHC [11] and NetMHCpan-4.1 [10]—inform "Task-EL (I, II)", a masked amino acid prediction task modeling peptide processing and MHC presentation (Fig. 1a). In parallel, peptide–MHC binding affinity data—129,973 class I and 516,463 class II measurements, foundational to tools such as MHCflurry-2.0 [13]—support "Task-MHC (I, II)", which predicts binding affinities to capture interaction strength (Fig. 1a). Subsequently, the critical interaction between T-cell receptors (TCRs) and peptide–MHC complexes is addressed through "Task-TCR", leveraging 65,030 curated data points from IEDB and literature sources (Fig.1c). This task enhances the model's ability to capture downstream immune recognition beyond what is modeled by interaction-focused methods such as DLpTCR[14]. To incorporate concepts of "selfness", TITANiAN introduces auxiliary tasks using human-derived protein sequences from the OAS dataset [15] and UniProt [16]. These include "Task-Hu-CLS" (a human vs. non-human classification task, inspired by Hu-mAb [17]) and "Task-Hu-MLM" (a species-conditioned masked language modeling task, following ideas from AbNatiV [18] and Bio-phi [19]) (Fig. 1e).

Following pretraining, the shared integrative encoder is fine-tuned using T-cell activation data to enable final immunogenicity prediction. This data—representing the endpoint of the antigen recognition cascade as measured by functional assays such as ELISPOT—are sourced from repositories including IEDB [12], MHC BN [20], the NCI database [21], and PRIME-2.0 [22]. Despite its relatively limited size (12,782 data points), this dataset enables direct optimization of the model through "Task-IM" (Fig. 1h).

By leveraging rich contextual representations learned during pretraining, TITANiAN achieves robust immunogenicity prediction—setting it apart from models such as PRIME-2.0 [23], which also predict T-cell activation but lack comparable multi-domain pretraining. During inference, input flexibility is maintained: predictions can be made using just the peptide–MHC pair, or optionally include TCR sequence or species label, depending on the application context.

2.2 Biological Motivation and Justification for Multi-domain Pretraining

To further support the rationale behind our multi-domain training strategy and to gain deeper insights into the relationship between pretraining and fine-tuning datasets, we analyzed label consistency for peptides shared across both stages. Specifically, we identified peptides that appeared in both the pretraining datasets (Task-EL, Task-MHC, and Task-TCR) and the fine-tuning dataset (Task-IM) by detecting exact matches of 9-mer fragments, and then compared their corresponding biological labels (Fig.2a). Unlike the training/test split procedure—where overlapping fragments with fewer than two mismatches were excluded—this analysis focused strictly on exact 9-mer matches. This stricter criterion was used because even a single amino acid substitution can significantly alter immunogenicity outcomes, as demonstrated in Fig.6. Permitting mismatches in this context could obscure the true correspondence between labels and confound interpretation of inter-dataset consistency.

This analysis yielded two key insights that underscore the complexity of T-cell immunogenicity and the necessity of a multi-domain learning framework.

First, we frequently observed label inversion—cases in which peptides that exhibited strong MHC binding failed to induce significant cytokine release in in vitro activation assays. This observation highlights a fundamental limitation of models trained solely on isolated steps of the immunogenicity cascade. It also indicates that the presence of shared peptides across datasets does not inherently result in trivial data leakage; rather, the model must learn to navigate and integrate non-linear, biologically complex relationships to accurately map intermediate signals to final immunogenic outcomes.

Second, while the correlation between intermediate and final labels was relatively weak, it was not negligible. We observed a Cramer's V of 0.0339 with a statistically significant p-value of 0.0265, indicating a small but meaningful association between the datasets (Section 4.2). This finding suggests that even subtle signals from intermediate tasks can provide valuable information for immunogenicity prediction—provided they are integrated effectively within a multi-domain framework.

Taken together, these observations—the complex, non-linear relationship between intermediate processes and final immune activation, along with the presence of weak yet significant inter-task correlations—strongly support our multi-domain learning strategy. This approach is specifically designed to capture and exploit such intricate biological dependencies, thereby enhancing the robustness and accuracy of immunogenicity prediction.

2.3 Preventing Data Leakage: Curation and Splitting Strategy

Since data leakage poses a critical challenge in developing and evaluating deep learning models—particularly in multi-domain learning frameworks that integrate data from diverse sources—we implemented a stringent curation and splitting strategy to ensure robust and leakage-free evaluation.

To evaluate the risk of data leakage, we first analyzed the biological relationships and the hierarchical structure of information overlap across datasets (Fig. 2b,c). As

expected, the T-cell activation dataset shares content with other activation-related datasets, introducing a risk of intra-task leakage. Moreover, given the sequential nature of the immunogenicity cascade, this dataset inherently incorporates information from upstream processes such as peptide–MHC binding and TCR–peptide–MHC interactions, giving rise to potential inter-task leakage.

Similarly, the TCR–peptide–MHC binding dataset subsumes information found in peptide–MHC binding datasets. However, the reverse is not true: peptide–MHC binding datasets do not contain information about downstream processes such as TCR recognition or T-cell activation. Recognizing this asymmetry, we systematically evaluated potential leakage paths from training to benchmark datasets and took explicit steps to eliminate both intra-task and inter-task data leakage.

In all identified cases, we ensured that no peptides—and not even highly similar sequences (i.e., 9-mer fragments with fewer than two mismatches)—were shared between datasets, thereby preserving the integrity of our evaluations.

In line with our stringent data curation and splitting strategy to prevent potential leakage across diverse data sources, we applied a similarly conservative approach to the human–non-human dataset. Although prior studies have suggested that human versus non-human classification provides limited predictive power for immunogenicity [17–19], we treated this dataset as potentially encompassing the entire immunogenicity process, analogous to T-cell activation datasets. This cautious stance was adopted to guard against even subtle forms of leakage, as the dataset may implicitly encode aggregate immunogenicity-related features that transcend individual biological processes.

To construct a robust, leakage-free training set, we applied dataset-specific filtering criteria to each data source. For the peptide–MHC binding and eluted ligand MS datasets, we excluded peptides containing non-canonical amino acids or exceeding 20 residues in length, in accordance with our model architecture, which accepts inputs up to 20 residues. Shorter peptides were padded to this length to standardize input dimensions. This design accommodates the natural variation in peptide lengths across tasks, particularly the longer peptides commonly found in Task-MHC II.

We also removed entries lacking well-defined MHC allele annotations, which are essential for generating accurate pseudo-sequences [10]. A similar filtering strategy was applied to the TCR–peptide–MHC dataset, excluding peptides based on length or non-standard amino acids. In addition, we discarded samples lacking TCR CDR sequences or associated with low-confidence scores (e.g., ¡1 in VDJdb [24]) to further enhance dataset quality.

For the Human versus Non-human dataset, designed to help the model learn species-specific features, we applied a tailored filtering strategy. Each peptide was segmented into overlapping 9-mer fragments, and the frequency of each fragment was computed separately for human and non-human sequences. Peptides containing any 9-mer that appeared in both species were excluded, ensuring that only peptides uniquely associated with either human or non-human samples were retained.

The T-cell activation dataset, used during the fine-tuning stage, was subjected to the same filtering criteria as the peptide–MHC dataset. Following these dataset-specific preprocessing steps, we finalized the pretraining and fine-tuning datasets. To

further minimize the risk of data leakage, we applied an additional stringent filtering step: each peptide in both the training and benchmark sets was decomposed into overlapping 9-mer fragments, and any training sample was excluded if any of its 9-mers matched a benchmark 9-mer with fewer than two mismatches. This filtering was performed using `cd-hit` clustering [25], based on the rationale that core T-cell epitopes are often localized within 9-mer regions, even when precise epitope boundaries are uncertain. A more detailed description of this procedure is provided in Section 4.3.

Notably, this stringent level of dataset separation exceeds the standards adopted by widely used tools such as NetMHCpan [10], which excludes only exact peptide matches, and BigMHC [11], which allows overlapping peptide sequences as long as the MHC allele differs. Our more conservative filtering strategy reflects the critical emphasis we placed on eliminating even subtle forms of data leakage, thereby ensuring robust, unbiased, and generalizable model performance.

Finally, after constructing a rigorously curated and leakage-free dataset, we split the training data into training and validation subsets using a 9:1 ratio. To further ensure independence and prevent potential information leakage, the split was performed using 9-mer-level clustering, guaranteeing that no overlapping core sequence fragments were shared between subsets. This deliberate strategy was designed to simulate a realistic and challenging evaluation setting for hyperparameter tuning, thereby enabling a more reliable assessment of the model's true generalization performance.

2.4 TITANiAN-IM: T-cell Epitope Immunogenicity Prediction for Specific Peptide-MHC Pairs

Understanding the T-cell epitope immunogenicity of peptide–MHC pairs is critical for the identification of neoantigens in cancer immunotherapy and the design of vaccines against infectious diseases. Experimental assays such as ELISPOT are valuable but face significant limitations in high-throughput applications due to the complexities of antigen production [26] and the labor-intensive nature of manual spot counting [27]. In this context, computational prediction models serve as a powerful alternative, enabling substantial reduction in the number of peptide candidates that require experimental validation.

TITANiAN predicts T-cell epitope immunogenicity by combining an integrative encoder with the Task-IM decoder, as illustrated in Figure 3a. This configuration, referred to as TITANiAN-IM, produces an immunogenicity score from the input peptide and MHC pseudo-sequence, which captures amino acid residues that contact the peptide [10]. A variant model, titanian-im, shares the same architecture but is trained solely on the fine-tuning dataset, without leveraging pretraining of the integrative encoder.

Model performance was assessed using four key metrics: Mean PPVn (mean positive predictive value among the top n predictions), Precision, PR-AUC (Area Under the Precision–Recall Curve), and AUROC (Area Under the Receiver Operating Characteristic Curve). TITANiAN-IM, along with leading state-of-the-art methods—including BigMHC[11], NetMHCpan-4.1[10], MHCflurry-2.0[13], MHCnuggets-2.4.0[4], MixMHCpred-2.2, PRIME-2.0[23], TranspHLA[28], and HLAthena[29]—was

rigorously evaluated on two independent benchmark datasets: (a) a neoantigen discovery set (833 peptides: 178 positives, 655 negatives) and (b) an infectious disease vaccine discovery set (2,093 peptides: 1,488 positives, 605 negatives), as shown in Figure4a,b.

To ensure that performance evaluations reflect genuine generalization rather than memorization, we applied a strict data leakage prevention strategy: any training sequence sharing a 9-mer fragment with fewer than two mismatches compared to any benchmark peptide was excluded. This rigorous filtering procedure guarantees a robust and fair assessment of model performance.

To assess the robustness and statistical significance of each model's performance, we conducted extensive bootstrap analysis and hypothesis testing. Specifically, we computed 95% confidence intervals using 1,000 bootstrap samples for each metric. Where applicable, we further compared TITANiAN-IM to the models with the next lower and next higher performance ranks, as well as to its baseline variant, titanian-im, to assess the significance of pairwise differences in performance. We additionally performed stratified bootstrap analysis across subgroups such as HLA allele types and peptide lengths. These results demonstrate not only the superior performance of TITANiAN-IM, but also its consistency across biologically meaningful strata. For implementation details, see Section 4.6.1.

Precision, which reflects the ability to correctly identify immunogenic samples, is particularly important in neoantigen selection and is highlighted in Figure 3e. Mean PPVn, which evaluates ranking quality among top predictions, is emphasized in Figure 3f. PR-AUC is presented alongside these metrics to ensure that high Precision and PPVn values are not achieved at the expense of Recall, thereby reflecting a balanced and reliable prediction performance.

In the neoantigen discovery benchmark (Figure 4a), TITANiAN-IM significantly outperforms the non-pretrained variant titanian-im ($p < 0.001$) across all four metrics, underscoring the critical importance of pretraining. While BigMHC-IM achieves a comparable mean PPVn (0.424 vs. 0.439), TITANiAN-IM shows substantially higher Precision (0.553 vs. 0.327), a difference that is both practically meaningful and statistically significant ($p < 0.001$). Consequently, TITANiAN-IM also attains a higher PR-AUC compared to BigMHC-ELIM (0.316 vs. 0.307; $p < 0.001$), reflecting its strong Recall performance. Additionally, TITANiAN-IM exhibits a higher AUROC (0.568) than MHCnuggets-2.4.0 (0.557), with the difference being statistically significant ($p < 0.001$). We attribute the enhanced performance in neoantigen identification, particularly the significant improvement observed in the Precision metric, to our multi-domain pre-training scheme which spans broader biological mechanisms of T-cell immunogenicity. We hypothesize that the limited precision of many existing methods stems from their focus on specific facets of the immunogenicity process (e.g., only MHC binding). In contrast, TITANiAN's pre-training integrates signals across the entire cascade, enabling it to achieve exceptionally high Precision compared to these methods. This result highlights the effectiveness of a holistic pre-training approach that captures the multifaceted nature of neoantigen immunogenicity.

In the infectious disease vaccine benchmark (Figure 4b), TITANiAN-IM again significantly outperforms the non-pretrained variant titanian-im ($p < 0.001$) and

surpasses all existing methods across all metrics. It achieves a Precision of 0.916 (+0.125) and a mean PPVn of 0.927 (+0.156) compared to BigMHC-IM (0.791 and 0.771, respectively), with both differences being statistically significant ($p < 0.001$). Furthermore, TITANiAN-IM attains higher PR-AUC (0.904 vs. 0.760) and AUROC (0.784 vs. 0.569) compared to BigMHC-IM and BigMHC-EL, respectively, demonstrating both statistical and practical superiority. Notably, the performance margin observed here is even greater than in the neoantigen discovery task, suggesting that our model is particularly well-suited for identifying pathogen-derived epitopes. We further attribute this strong performance to the incorporation of the human–nonhuman discrimination domain during pretraining. Unlike neoantigens, which require subtle differentiation from self, infectious disease epitopes originate from clearly foreign pathogens. The human–nonhuman task enables the model to internalize broad features that distinguish self from non-self at the sequence level, thereby enhancing its ability to identify pathogen-derived epitopes as immunogenic. This alignment between pretraining domain and the biological nature of the benchmark task likely contributes to TITANiAN-IM's superior generalization.

The error bars in Figure 4, representing 95% confidence intervals computed via bootstrap analysis, reflect the stability of each model's performance. TITANiAN-IM exhibits narrower or comparable error bars relative to other models, indicating robust and consistent performance. Notably, compared to the non-pretrained variant titanian-im, TITANiAN-IM demonstrates substantially smaller error bars across all metrics, highlighting its enhanced stability in addition to superior predictive accuracy.

Figure 5 presents the mean performance across peptide lengths and HLA allele types for both the neoantigen and infectious disease vaccine benchmarks. For TITANiAN-IM, bootstrap-based 95% confidence intervals were computed and shaded to indicate the variability in performance.

As shown in Figure 5a, TITANiAN-IM consistently outperforms other methods across all tested peptide lengths in the neoantigen discovery set. A similar trend is observed in the infectious disease vaccine benchmark (Figure 5c), where TITANiAN-IM maintains strong performance across peptide lengths and metrics. We emphasize that, unlike many existing methods which impose restrictions on peptide length—such as NetMHCpan (8–14 amino acids) and PRIME-2.0 (8–11 amino acids)—or exhibit performance degradation with increasing peptide length, as observed in BigMHC, our model supports peptides up to 20 amino acids in length and demonstrates robust performance across varying peptide lengths. We attribute this robustness to the pretraining phase, during which the model was exposed to peptides of diverse lengths—an advantage not attainable using the Task-IM fine-tuning dataset alone. This effect is particularly relevant for MHC class II molecules, which accommodate a broad range of peptide lengths due to their open-ended binding grooves, as represented in Tasks MHC-II and EL-II.

Figure 5b, d illustrates TITANiAN-IM's performance across different HLA allele types in both the neoantigen and infectious disease vaccine benchmarks. While the model demonstrates robust and stable performance in both settings—with results largely within the 95% confidence intervals of competing methods—it shows more consistent superiority across all HLA types in the infectious disease benchmark. In the

neoantigen benchmark, strong performance is observed for HLA-A, but performance for HLA-B and HLA-C, though superior to other models, is less pronounced. This difference in robustness is likely due to disparities in label distribution (Table D2). The infectious disease set more closely mirrors the fine-tuning training data, whereas the neoantigen set contains substantially fewer positive samples, particularly for certain HLA alleles. This mismatch may limit the generalization capacity of the model despite pretraining, especially for underrepresented types like HLA-C.

Figure 6 illustrates representative use cases of TITANiAN-IM. The model distinguishes highly similar peptides bound to the same HLA allele (Figure 6a–c) and the same peptide presented by different HLA alleles (Figure 6d–f). For instance, a single amino acid mutation in the SARS-CoV-2 spike glycoprotein (Figure 6a) results in a sharp contrast in predicted immunogenicity (0.9669 vs. 2.95e-05). Similar trends are seen in antigens from Hepatitis D virus and HIV-1 (Figures 6b,c). Conversely, the same peptide yields drastically different scores when presented by HLA-B35:01 versus HLA-A01:01 (Figure 6d). These results demonstrate TITANiAN-IM's capacity to capture immunogenic variation arising from minimal peptide alterations and HLA context differences.

2.5 TITANiAN-IM: Application to Anti-Drug Antibody Level Prediction for Therapeutic Antibodies

Minimizing the immunogenicity of therapeutic antibodies, as measured by anti-drug antibody (ADA) production, is crucial for ensuring drug safety and efficacy. While selfness-based prediction methods have been commonly employed to estimate ADA risk, TITANiAN-IM predicts ADA levels by identifying overlapping 9-mer peptide segments with immunogenicity scores exceeding 0.5, without requiring MHC pseudo-sequences as input (Figure 3b). Given the objective of detecting all potential immunogenic regions within an antibody sequence, Recall is emphasized as the primary evaluation metric (Figure 3g).

We evaluated the predictive performance of TITANiAN-IM, its non-pretrained baseline (titanian-im), and several established selfness-based methods—including AbNatiV [18], T20 [30], Z-score [31], IgReconstruct [32], AbLSTM [33], Germline content [34], Hu-mAb [17], MG Score [35], and OAsis [19]—on the ADA-level benchmark dataset of FDA-approved antibody drugs (size: 216; positive: 169; negative: 47) [17–19]. Following Prihoda et al., an antibody was classified as immunogenic if more than 20% of its constituent peptide fragments exhibited an immunogenicity score above the 0.5 threshold. To maintain data integrity and prevent leakage, all training peptides sharing 9-mer fragments with less than 2 mismatches to any peptide in the benchmark set were excluded prior to evaluation.

To ascertain the statistical significance of TITANiAN-IM's performance enhancements, we employed bootstrapping with 1,000 resamples to estimate 95% confidence intervals for each metric and utilized paired Wilcoxon signed-rank tests with Bonferroni correction for hypothesis testing. Where applicable, we further compared TITANiAN-IM to the models with the next lower and next higher performance ranks, as well as to its baseline variant, titanian-im, to assess the significance of pairwise differences in performance. This rigorous analysis confirmed that TITANiAN-IM

consistently and significantly outperformed all other evaluated methods across the evaluated metrics (all corrected p-values < 0.001).

Notably, the non-pretrained baseline, titanian-im, failed to identify any immunogenic antibodies, classifying all sequences as non-immunogenic. Consequently, it achieved Recall and PR-AUC scores of 0 and an AUROC of 0.5, equivalent to random guessing (Figure 4c). In stark contrast, TITANiAN-IM achieved a Recall of 0.799, outperforming all other models, including the next-best model, IgReconstruct (0.751), with a statistically significant difference ($p < 0.001$). While achieving high Recall, TITANiAN-IM maintained strong ranking and classification performance, achieving a PR-AUC of 0.847—comparable to T20 (0.864)—and an AUROC of 0.641, slightly higher than OASis identity (0.635, $p < 0.001$). Furthermore, the error bars associated with TITANiAN-IM were comparable to those of other models, indicating that its superior performance is accompanied by stable and consistent behavior. Furthermore, the error bars associated with TITANiAN-IM were comparable to or narrower than those of other models, indicating that its superior performance is accompanied by stable and consistent behavior.

Representative cases demonstrating ADA-level prediction by TITANiAN-IM are illustrated in Figure 6g–j. These examples include comparisons with experimentally identified deimmunized amino acids (DAAs, shown in cyan). TITANiAN-IM successfully recovers 70–100% of DAAs in Etaracizumab, Bevacizumab, Pembrolizumab, and Omalizumab, with predicted immunogenic residues highlighted in green. These predictions are based on overlapping 9-mer regions with immunogenicity scores above 0.5. Furthermore, the distribution of immunogenicity scores before and after deimmunization clearly distinguishes immunogenic from non-immunogenic variants, validating TITANiAN-IM's ability to capture critical immunogenicity-related signals.

Importantly, this competitive predictive performance is achieved without the need for MHC pseudo-sequences, a key advantage attributed to the robust pretraining on diverse, multi-domain datasets, including the MHC-independent Task-Hu. These compelling findings underscore TITANiAN-IM's significant potential for accurate and interpretable immunogenicity prediction in the crucial domain of therapeutic antibody development.

2.6 TITANiAN-IM: Dataset Ablation Study

To dissect the contribution of each pre-training data source and its corresponding task to TITANiAN's predictive capabilities, we performed a comprehensive ablation study. For each ablated variant, we systematically excluded a specific dataset and its associated task during the pre-training phase. Concurrently, the task-specific encoder and decoder corresponding to the ablated task were removed from the model architecture, while the shared integrative encoder and the remaining task-specific modules were retained. Following this modified pre-training, all ablated models underwent the identical fine-tuning procedure as the full TITANiAN-IM model, utilizing the integrative encoder and the Task-IM decoder on the same fine-tuning dataset.

This ablation process yielded four key variants: TITANiAN-IM Hu (excluding Task-Hu and human-derived protein data), TITANiAN-IM MHC I (excluding MHC

class I binding and eluted ligand MS datasets), TITANiAN-IM MHC II (excluding MHC class II binding and eluted ligand MS datasets), and TITANiAN-IM TCR (excluding TCR–peptide–MHC binding data).

The impact of these ablations was evaluated across three distinct settings (Fig. 7): (a) the neoantigen discovery set (MHC class I-associated), (b) the infectious disease vaccine discovery set (MHC class I-associated), and (c) the FDA-approved drug ADA-level benchmark set (closely linked to MHC class II immunogenicity). We note that (a) the neoantigen discovery dataset and (b) the infectious disease vaccine discovery dataset consist of experimental data measuring $CD8^+$ T-cell activation, which is associated with MHC class I pathways. In contrast, (c) the FDA-approved drug ADA-level benchmark dataset captures the prevalence of anti-drug antibody (ADA) induction in patients administered with therapeutic antibodies. Since antibody production is driven by B-cell activation, which is in turn supported by $CD4^+$ T-cell activation, this dataset is primarily associated with MHC class II pathways. For each model variant, we visualized performance using the five metrics reported earlier (Mean PPVn, Precision, PR-AUC, AUROC, and Recall), incorporating error bars representing 95% confidence intervals (derived from bootstrapping) and p-value annotations (paired Wilcoxon signed-rank test with Bonferroni correction) to rigorously assess the statistical significance of performance differences relative to the full TITANiAN-IM model.

The ablation study results confirm the meaningful contribution of all pre-training tasks to TITANiAN's overall performance, albeit with varying degrees of influence depending on the specific downstream prediction task. Notably, all comparisons between the full model and the ablated variants revealed statistically significant performance reductions ($p < 0.001$ after Bonferroni correction). Among the pre-training tasks, 'Task-Hu' exhibited the broadest impact across all benchmarks, likely attributable to the high sequence diversity of the human protein data and its weak yet informative signals related to self/non-self discrimination, a crucial aspect of immunogenicity. Interestingly, ablating MHC class I-related tasks resulted in substantial performance degradation on the $CD8^+$ T-cell associated benchmarks (Fig. 7a,b) but had a comparatively limited effect on the $CD4^+$ T-cell prediction task (Fig. 7c). A reversed pattern was observed for the ablation of MHC class II-related datasets, indicating that TITANiAN effectively learns and utilizes information relevant to both MHC class I and MHC class II T-cell epitopes. While TCR-related data also played a significant role, its ablation led to a relatively smaller performance decline, potentially due to the model architecture's current limited focus on explicitly embedding the high complexity and variability of TCR CDR regions.

2.7 TITANiAN-MHC: Predicting Peptide-MHC Binding

Predicting peptide-MHC interactions is a crucial initial step in identifying potential immunogenic peptides for applications such as cancer immunotherapy and vaccine development. While the T-cell epitope immunogenicity prediction results presented in Figures 4a and b demonstrate the utility of *in silico* methods, they also highlight that these approaches are not yet infallible, underscoring the need for complementary strategies. One such strategy involves pre-filtering peptide candidates based on

predicted MHC binding affinity using models trained on peptide-MHC interaction data. This pre-selection process can significantly reduce the search space for downstream immunogenicity assays. Therefore, in the context of peptide-MHC binding prediction, Precision and Mean PPVn are considered key metrics for evaluating the accuracy of identifying true binders. To ensure that high Precision and Mean PPVn are not achieved at the expense of Recall, we also evaluate the PR-AUC. Additionally, AUROC is included as a general measure of the classifier's overall performance in distinguishing between binders and non-binders.

As depicted in Figure 3c, TITANiAN-MHC utilizes the shared integrative encoder alongside the dedicated Task-MHC encoder and decoder. TITANiAN-MHC operates solely on the parameters learned during the pre-training phase, without undergoing any additional task-specific fine-tuning.

To evaluate the performance of TITANiAN-MHC I and II in predicting peptide-MHC binding, we employed the IEDB benchmark set [12] with data collected after December 25, 2020 (for MHC class I: size = 624, positives = 344, negatives = 280) and after February 19, 2021 (for MHC class II: size = 723, positives = 242, negatives = 481). Consistent with our rigorous data leakage prevention strategy, we excluded any training sequence that shared 9-mer fragments with less than 2 mismatches with any peptide in these benchmark sets. This stringent filtering ensures that the reported performance reflects genuine generalization capabilities rather than the memorization of highly similar sequences. Furthermore, we carefully curated the benchmark sets to encompass a diverse range of HLA alleles. Notably, the class I benchmark includes the HLA-E allele (HLA A - 368, B - 84, E - 172) to specifically assess each model's ability to generalize to alleles that are underrepresented in the training data.

To rigorously evaluate the statistical significance of TITANiAN-MHC's performance gains, we employed bootstrap analysis with 1,000 resamples to estimate 95% confidence intervals for each evaluation metric. Additionally, we conducted paired Wilcoxon signed-rank tests with Bonferroni correction to assess the statistical significance of the observed differences. Specifically, we compared TITANiAN-MHC against the models ranked immediately below and above it in performance to pinpoint the significance of improvements relative to its closest competitors.

The performance of TITANiAN-MHC I in predicting MHC class I binding was benchmarked against a panel of established methods, including NetMHCpan2.8 [36], SMM [37], ANN4.0 [38], ARB [39], SMMPMBEC [40], IEDB Consensus [41], NetMHCcons [42], PickPocket [43], mhcflurry 1.2.0 [44], ANN3.4 [45], NetMHCpan3.0 [46], and various versions of NetMHCpan 4.x [10]. As illustrated in Figure 8a, TITANiAN-MHC I demonstrates a competitive performance profile in comparison to these established models. It achieves a Mean PPVn of 0.770, which is slightly lower than IEDB Consensus (0.804) but comparable to several NetMHCpan versions, while showing statistically significant improvements over models with lower performance ($p < 0.001$). Although its precision (0.619) is lower than that of NetMHCpan 4.0 EL (0.775), its PR-AUC (0.725) is comparable to NetMHCcons (0.813) and notably higher than NetMHCpan 4.0 EL, indicating a favorable balance between precision and recall ($p < 0.001$). The AUROC of TITANiAN-MHC I (0.674) is similar to NetMHCpan 4.1 EL (0.691) and surpasses several earlier models, including NetMHCpan 4.0

BA ($p < 0.001$). Importantly, these results were achieved without any task-specific fine-tuning, relying solely on the knowledge acquired during pretraining.

Figure 8b presents a performance comparison of TITANiAN-MHC II against several established MHC class II binding prediction methods, including NN-align [47], NetMHCIIpan-3.1 [48], SMM-align [49], Tepitope (Struniolo)[50], IEDB Consensus[51], and NetMHCIIpan-4.0 EL [10]. Overall, TITANiAN-MHC II demonstrates broadly comparable performance across the evaluated metrics. Specifically, it achieves a Mean PPVn of 0.613, which is higher than NN-align 2.3 (0.598; $p < 0.001$) and ranks third among the compared methods, following NetMHCIIpan-4.0 EL (0.750) and NetMHCIIpan-4.1 EL (0.766). While its precision (0.349) and PR-AUC (0.521) are lower than those of the leading models such as NetMHCIIpan-4.0 EL (precision: 1.000) and NN-align 2.3 (PR-AUC: 0.781), TITANiAN-MHC II still outperforms several other methods, including IEDB Consensus (precision: 0.338; $p < 0.001$). Its AUROC (0.657) is lower than NetMHCIIpan-4.0 EL (0.817) but significantly higher than Tepitope (0.500; $p < 0.001$). The relatively larger performance gap observed in MHC class II prediction may be attributed to the inherent variability in peptide lengths and binding modes associated with MHC class II molecules.

2.8 TITANiAN-TCR: Predicting TCR-Peptide-MHC Binding

Predicting peptide-MHC-TCR binding holds significant potential for applications such as rational TCR design in Adoptive Cell Transfer (ACT)-based therapy and the analysis of infectious disease cohorts [52]. For these applications, the ability to accurately identify high-confidence interactions is crucial, making Precision and Mean PPVn key performance metrics. To provide a comprehensive evaluation of the classifier's capabilities, we also assess PR-AUC and AUROC. In contrast to the immunogenicity prediction task discussed in Section 2.4, the relatively balanced distribution of positive and negative samples in this dataset allows us to emphasize AUROC as a similarly important metric alongside PR-AUC.

TITANiAN-TCR, which comprises the shared integrative encoder and the dedicated Task-TCR encoder/decoder (as illustrated in Figure 3d), predicts TCR-peptide-MHC binding by leveraging solely the parameters learned during the pre-training phase, without any task-specific fine-tuning. We evaluated TITANiAN-TCR on two distinct benchmark sets: the Panpep Zero-shot test set (size: 786, positive: 393, negative: 393) and the ACT test set (size: 26, positive: 13, negative: 13) [52]. Notably, we place a particular emphasis on the ACT test set, as it contains experimental data directly relevant to the development of ACT-based immunotherapy. To assess its performance, TITANiAN-TCR was compared against several task-specific methods, including Pan-pep [52], DLpTCR [14], pMTnet [53], and ERGO2 [54].

As shown in Figure 8c, our evaluation on the zero-shot test set revealed that TITANiAN-TCR achieves performance comparable to Pan-pep and significantly outperforms all other methods. It attained a Mean PPVn of 0.656, second only to Pan-pep (0.801), and significantly higher than the third-best model, ERGO2 (0.500; $p < 0.001$). A similar trend was observed in Precision (0.650 vs. 0.717 for Pan-pep), significantly exceeding DLpTCR (0.505; $p < 0.001$). Its PR-AUC of 0.709 closely matched Pan-pep (0.747), while significantly surpassing ERGO2 (0.500; $p < 0.001$). The AUROC

of TITANiAN-TCR (0.630) also exceeded all models except Pan-pep (0.739), with significant differences observed (vs. DLpTCR: 0.504; $p < 0.001$). To assess the statistical significance of these results, we employed bootstrap analysis (1,000 resamples, 95% confidence intervals) and paired Wilcoxon signed-rank tests with Bonferroni correction, comparing TITANiAN-TCR against the closest performing models.

While TITANiAN-TCR exhibited lower performance than Pan-pep on the zero-shot test set, likely due to Pan-pep's meta-learning optimization for such scenarios, TITANiAN-TCR demonstrated a more stable and highly competitive performance on the ACT-based therapy discovery benchmark, achieving results closely comparable to the top-performing Pan-pep and significantly outperforming other methods. On the ACT-based therapy discovery set (Figure 8d), TITANiAN-TCR achieved a Mean PPVn of 0.800, comparable to Pan-pep (0.883), and substantially higher than DLpTCR (0.621; $p < 0.001$). Precision was identical to Pan-pep (0.778) and significantly higher than all other methods ($p < 0.001$). PR-AUC was also comparable (0.745 vs. 0.819), indicating that TITANiAN-TCR maintains strong recall while offering a superior balance; its PR-AUC was significantly higher than DLpTCR (0.595; $p < 0.001$). Notably, TITANiAN-TCR achieved the highest AUROC among all evaluated methods on the ACT dataset (0.769 vs. Pan-pep's 0.751; $p < 0.001$), indicating a statistically significant edge in distinguishing binders from non-binders in this specific metric. However, when considering other metrics, the performance between TITANiAN-TCR and Pan-pep remained largely comparable. The statistical significance of these ACT results was also confirmed using the aforementioned bootstrap and Wilcoxon signed-rank tests.

These strong and consistent results are particularly noteworthy given that the Task-TCR module was designed as an auxiliary component rather than a primary focus. Unlike other models that allocate substantial parameters to learn detailed representations from TCR sequences, TITANiAN-TCR primarily focuses on embedding peptide-MHC pairs—its integrative encoder, the core of our pipeline, receives only peptide-MHC inputs, not TCR information—while incorporating only minimal TCR-specific processing. Nevertheless, it demonstrates that even with this lightweight design, the model can achieve a reasonable level of performance in TCR-based prediction tasks.

3 Discussion

In this study, we introduced TITANiAN, a comprehensive deep learning framework designed to predict T-cell epitope immunogenicity by integrating diverse biological information. TITANiAN demonstrated exceptional performance via its fine-tuned module (TITANiAN-IM) across its primary prediction tasks: it achieved state-of-the-art results in predicting T-cell activation for specific peptide-MHC pairs, often outperforming existing methods, and also delivered top-tier performance in assessing therapeutic antibody immunogenicity without requiring MHC input. TITANiAN also supports additional immunogenicity-related predictions, such as molecular binding tasks, enabled by its multi-domain pre-training scheme. This broad effectiveness across different facets of immunogenicity underscores the power of our pre-training strategy,

which integrates a curated array of datasets through a modular multi-task architecture with adversarial domain adaptation. The model's robustness was further supported by rigorous data curation to prevent leakage and by comprehensive ablation studies, which confirmed the individual contribution of each pretraining task. TITANiAN is publicly available via a web server (https://galaxy.seoklab.org/design/titanian/) to facilitate broad accessibility and application.

TITANiAN's capabilities could be further enhanced by addressing several areas. First, incorporating additional data types pertinent to immunogenicity, such as transporter-associated antigen processing (TAP) data or peptide-protein docking information, could enrich the model's biological understanding. However, the limited availability of such data currently poses a constraint. Expanding the framework with more diverse public or proprietary datasets represents a promising future direction. Second, while TITANiAN demonstrates strong performance, further hyperparameter optimization could yield incremental improvements. However, the extensive computational resources required for pre-training (e.g., 10 GPU days on an RTX A6000) currently constrain exhaustive optimization and ablation experiments (like exploring optimal task combinations).

Beyond benchmarking, TITANiAN's ability to provide accurate and comprehensive immunogenicity predictions, integrating signals from binding to activation and selfness, positions it as a powerful tool to drive biomedical discovery. We anticipate its application will accelerate key processes such as the de-immunization of protein therapeutics, the rational design of vaccines, and the identification of novel cancer neoantigens, ultimately contributing to the development of safer and more effective treatments.

4 Methods

4.1 Dataset and task

We compiled all publicly available data, as detailed in Table D1. For the EL task, we collected MHC eluted ligand MS data from IEDB [12] using the following filtering criteria: "Epitope Structure: Linear Sequence", "Include Positive Assays", "No T cell assays", "No B cell assays", "Host: Homo Sapiens (human)", and "MHC Assays: MHC ligand elution assay". Then the dataset underwent prefiltering and removal of any overlapping sequences with the benchmark set, as described in Section 2.3. We note that we did not allow any overlapping with the benchmark set in either the training set or validation set by using the stringent filtering method outlined in Section 4.3. This resulted in 557,151 data points for HLA class I and 206,055 for HLA class II. As all entries were positive, we employed a Masked Language Modeling task by masking the peptide and conditioning on the MHC pseudo sequences [55]. The dataset was divided into training and validation sets in a 9:1 ratio, ensuring that peptides in the validation set did not overlap with those in the training set using the stringent filtering described in Section 4.3.

For the MHC task, we utilized IEDB [12] to obtain the peptide-MHC binding affinity dataset with similar filtering criteria: "Epitope Structure: Linear Sequence", "Include Positive Assays", "Include Negative Assays", "No T cell assays", "No B cell

assays", "Host: Homo Sapiens (human)", and "MHC Assays: MHC binding assay". Then the dataset underwent prefiltering and removal of overlapping sequences with the benchmark set as described in Section 2.3. We note that we did not allow any overlapping with the benchmark set in either the training or validation set by applying the stringent filtering method as described in Section 4.3. This process yielded 129,973 data points for HLA class I and 516,463 for HLA class II. Data were labeled based on the "Assay - Qualitative Measurement" column: entries containing "positive" were labeled as positive, and those containing "negative" were labeled as negative. This resulted in 57,976 positive and 71,997 negative data points for HLA class I, and 66,424 positive and 450,039 negative data points for HLA class II. We framed the task as a simple binary classification problem, using peptide and MHC sequences as input. The dataset was split into training and validation sets in a 9:1 ratio, ensuring no overlap between peptides in the training and validation sets using the stringent filtering described in Section 4.3.

For the TCR sequence, we utilized positive sets gathered from Deng et al., which compiled information from various public data sources, including VDJdb [24], IEDB [12], McPAS-TCR [57], ImmuneCODE [58], TBAdb [59], and 10X Genomics [60]. Preprocessing followed the exact methods described by Deng et al.. The negative set for training was gathered from Dean et al., which includes TCR CDR3β sequences derived from healthy donors, a common method to create negative sequences [52]. From this, we selected CDR3β sequences between 15 and 25 amino acids in length to align with the positive dataset's sequence length of 20 amino acids. Non-functional CDR3β sequences were randomly sampled and shuffled with epitopes from the positive sequences to create a balanced 1:1 ratio of positive to negative classes. We note that only CDR3β sequences are utilized in our study, as CDR3α sequences are difficult to obtain. Moreover, several existing *in silico* methods for predicting TCR–peptide–MHC interactions also rely solely on CDR3β sequences [14, 52]. Then the dataset underwent prefiltering and removal of any overlap with the benchmark set as described in Section 2.3. We note that we did not allow any overlapping with the benchmark set in either the training or validation set by applying the stringent filtering method described in Section 4.3. This process resulted in 65,030 positive and 65,030 negative data points, totaling 130,060 sequences. We framed this as a binary classification task using peptide and TCR CDR3β sequences as input. The dataset was split into training and validation sets in a 9:1 ratio, ensuring no overlap between peptides in the training and validation sets using the stringent filtering described in Section 4.3.

For the Hu-MLM and Hu-CLS tasks, datasets of human-originated and non-human-originated protein sequences were obtained from UNIPROT [16] and OAS [15]. The sequences were segmented into 9-mer peptides using a sliding 9-mer window applied across each protein sequence. The frequency of each peptide in human and non-human proteins was then counted and the fraction was calculated as in Equation 1. To train the model to distinguish between human and non-human peptides, we selected peptides that i) appeared more than five times and ii) had a fraction value of exactly 0 (entirely non-human) or 1 (entirely human). We adopted this criterion as peptides with fewer than five occurrences may not reliably represent species specificity, and

peptides found in both human and non-human sequences could introduce ambiguity in training.

$$\textbf{Fraction} = \frac{\text{Human Frequency}}{\text{Human Frequency} + \text{Non-human Frequency}} \tag{1}$$

For UNIPROT, given that 400,000 peptides had a fraction of 1, we balanced the dataset by including an equivalent number of non-human peptides. For OAS, we applied the same balancing strategy, resulting in sets of 40,000,000 peptides with a fraction of 1 and 40,000,000 peptides with a fraction of 0. Due to the large size of the dataset, we reduced it to 4,000,000 peptides per class for computational efficiency. We performed a Masked Language Modeling (MLM) task, masking peptides and predicting the masked regions conditioned on whether the sequence was from human or not. Additionally, we conducted a binary classification to distinguish whether the input sequence was human or not. Then the dataset underwent prefiltering and removal of any overlapping sequences with the benchmark set as described in Section 2.3. We note that we did not allow any overlapping with the benchmark set in either the training or validation set, applying the stringent filtering method as detailed in Section 4.3. The dataset was split into training and validation sets in a 9:1 ratio. To create a hard split validation set, peptides were ordered by frequency, with the last 10

For the task IM-I, a cytokine release dataset with MHC class I information, we utilized the PRIME-1.0 and PRIME-2.0 datasets [22]. The PRIME datasets were preprocessed following the same methodology as described by Albert et al.. Then the dataset underwent prefiltering and removal of overlapping sequences with the benchmark set as outlined in Section 2.3. We note that we did not allow any overlapping with the benchmark set in either the training or validation set, applying the stringent filtering method as described in Section 4.3. This process resulted in 1,512 positive and 5,163 negative data points, totaling 6,675. The dataset was split into training and validation sets in a 9:1 ratio, ensuring that peptides in the training set did not overlap with those in the validation set using the stringent filtering described in Section 4.3.

For the task IM-pep, a cytokine release dataset without MHC information, we utilized the datasets from Müller et al.. Since the number of negative samples was exceedingly high, we assumed that using all negatives would impair the model's ability to learn a meaningful positive signal, so we randomly sampled 1,000 negative sequences. Then the dataset underwent prefiltering and removal of overlapping sequences with the benchmark set as outlined in Section 2.3. We note that we did not allow any overlapping with the benchmark set in either the training or validation set, applying the stringent filtering method as detailed in Section 4.3. This process resulted in 142 positive and 996 negative samples, totaling 1,138. The dataset was split into training and validation sets in a 9:1 ratio, ensuring that peptides in the training set did not overlap with those in the validation set using the stringent filtering described in Section 4.3.

For the task IM-II, a cytokine release dataset with MHC class II information, we used datasets from IEDB [12] and MHCBN [20] to gather data on the immunogenicity of peptide-HLA class II pairs. These datasets were preprocessed following the methodology described by Kim et al., resulting in 2,956 positive and 2,013 negative

data points, totaling 4,969. The model was trained on this dataset using a binary classification task. The dataset was split into training and validation sets in a 9:1 ratio, ensuring that peptides in the training set did not overlap with those in the validation set using the stringent filtering method as outlined in Section 4.3.

4.2 Cramer's V analysis of label concordance

To evaluate the association between labels from datasets representing intermediate biological processes (e.g., peptide–MHC binding, TCR–pMHC interaction) and immunogenicity labels (cytokine release), we computed *Cramer's V* statistics based on contingency tables constructed from overlapping peptides.

Peptide pairs with exact match of 9-mer fragment between datasets were identified, and the corresponding binary labels from the source (training) and target (test) datasets were extracted. For each dataset group defined by the training task identifier, we generated a 2×2 contingency table by cross-tabulating the number of peptides with each combination of train and test labels. To ensure uniformity across groups, the contingency table was reindexed to explicitly include all possible label pairs (0, 1), with missing entries filled as zeros.

Cramer's V was calculated using the formula:

$$V = \sqrt{\frac{\chi^2}{n \cdot (\min(r, c) - 1)}}$$

where χ^2 is the chi-square statistic obtained from the contingency table, n is the total number of overlapping peptide pairs, and r, c are the number of rows and columns, respectively. The associated p-value was derived from the chi-square test using SciPy's `chi2_contingency` function.

This analysis was performed both on the combined dataset and on subsets corresponding to individual training tasks, each representing distinct biological processes such as p-MHC binding affinity, p-MHC elution, or TCR interaction. The results were visualized using heatmaps to display the distribution of label concordance.

In the overall dataset, we observed a weak but statistically significant association (Cramer's $V = 0.0339$, $p = 0.0265$), suggesting the presence of minor but informative signals between intermediate biological steps and immunogenicity. These results support the hypothesis that, despite label inversion, weak signals embedded in upstream datasets may contribute to the development of more robust immunogenicity prediction models.

4.3 Sequence identity filtering using CD-HIT

To rigorously prevent data leakage between training and benchmark datasets, we applied a peptide-level identity filtering strategy using CD-HIT [25], based on 9-mer fragment identity. This process was designed under the assumption that the immunogenic signal of a peptide is often localized within a core 9-mer.

All peptides from both the training and benchmark sets were processed into overlapping 9-mer fragments using a sliding window approach. For peptides shorter than 9

residues, the full-length sequence was retained as a single fragment. Each fragment was assigned a unique label (e.g., `train_0_0`, `test_1_3`) and saved in FASTA format. The combined set of training and test 9-mers was clustered using CD-HIT with a sequence identity threshold of 0.8 (`-c 0.8`, to ensure less than 2 mismatches is tolerated for 9-mer fragment) and word size of 2 (`-n 2`).

From the resulting cluster assignments, we identified test peptides as overlapping if *any* of their 9-mer fragments were clustered together with a fragment derived from the training set. These peptides were excluded from the final benchmark set to eliminate potential memorization or indirect data contamination.

This strict filtering step ensured that the benchmark set contained no sequences with high similarity to any training peptide at the 9-mer level. The overall procedure is illustrated in Fig. 2b.

4.4 Architecture of TITANiAN

TITANiAN employs a convolutional neural network (CNN)-based model to represent protein amino acid sequences, following the encoder and decoder architecture proposed by Yang et al.. While most protein language models utilize transformers [64, 65] or LSTMs [11, 66], we chose a CNN-based model. This decision aligns with the characteristics of T-cell immunogenicity tasks, where long protein sequences are fragmented into 9-mer peptides. This eliminates the need for global sequence attention and instead emphasizes local peptide features. We utilized kernel size 3 to catch the local feature of peptides. Similar approaches have been successfully adopted in various methods [67, 68], demonstrating that CNN-based models can perform comparably well to transformer or LSTM-based architectures.

We employed the ByteNet architecture, a CNN-based encoder-decoder framework, as the core components of our model. We utilized the ByteNet encoder architecture to model the integrative encoder and task-specific encoder, except for conditioned sequence masked region prediction tasks. For the tasks Hu-MLM and EL, we used the conditioned ByteNet Decoder, which was slightly adjusted by adding an attention module to incorporate the condition into the embedding. For other tasks, we used the conventional ByteNet Decoder. As depicted in Figure 9, the embeddings generated by ByteNet were utilized as queries and values in the attention network, with each condition acting as the key.

By combining these components, we designed the integrative encoder, task-specific encoder/decoder, as shown in Figure 9. The integrative encoder consists of three ByteNet models that embed peptide, MHC class I pseudo sequence, and MHC class II pseudo sequence. These embeddings are combined using a gated attention mechanism which calculates attention scores as described in Equation 3. These scores are then used as weights to compute a weighted sum of the embeddings, as shown in Equation 2. For tasks that do not require MHC pseudo sequence (e.g., Hu-CLS), the output of the integrative encoder defaults to the peptide embedding.

$$e_{\text{final}} = \sigma(W^T e_{\text{proj}}) \odot e_{\text{proj}} \tag{2}$$

$$\text{where } e_{\text{proj}} = W^T \odot e \tag{3}$$

This gated attention mechanism is also applied to the task-specific encoder when multiple inputs are needed. The embeddings from the integrative encoder and the task-specific encoder are then concatenated and fed into the task-specific decoder, which is a simple MLP layer.

Lastly, the framework includes a discriminator for adversarial training of integrative encoder. The discriminator is implemented as a simple MLP layer that takes the integrative encoder embedding as input and outputs the data source classification probability. This component enables adversarial learning, where the model is trained using a gradient reversal mechanism to improve robustness and generalization.

4.5 Training scheme

We applied a multi-domain adversarial learning scheme [69], a method widely used in other biomedical applications [70, 71]. The training of TITANiAN involved three types of loss functions: task-specific loss, orthogonal loss, and adversarial loss.

Task-specific Loss. For each task, we used task-specific loss as described in Equations (4) and (5).

$$\text{loss}_{\text{task}} = -(y \log(p) + (1-y)\log(1-p)) \text{ for tasks MHC, TCR, Hu-CLS, and fine-tuning} \tag{4}$$

$$\text{loss}_{\text{task}} = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \text{ for tasks EL, Hu-MLM} \tag{5}$$

For binary classification tasks (e.g., MHC, TCR, and Hu-CLS), we used binary cross-entropy loss as in Equation (4), where y represents the label and p the predicted probability. For sequence masked region prediction tasks (e.g., EL and Hu-MLM), we employed cross-entropy loss to predict masked and mutated amino acids, as shown in Equation (5). In this context, c represents tokens, including amino acid codes, as well as stop, gap, mask, and start tokens. Here, $p_{o,c}$ is the model's probability for each token, and $y_{o,c}$ denotes the correct token. We masked 80% and mutated 10% of the sequence, following Devlin et al..

Orthogonal Loss. To differentiate the task-specific encoder output from the integrative encoder, we applied an orthogonal loss to make the embeddings orthogonal, as described in Equation (6).

$$\text{loss}_{\text{orthogonal}} = |e_{\text{integrative}} \cdot e_{\text{task}}| \tag{6}$$

In this equation, $e_{\text{integrative}}$ represents the output embedding from the integrative encoder, and e_{task} denotes the output embedding from the task-specific encoder.

Adversarial loss. To ensure the integrative encoder captures common information across data sources, we applied cross-entropy loss to the discriminator, similar to Equation (5), and adversarial loss to the integrative encoder, as defined in Equation (7).

$$\text{loss}_{\text{adv}} = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}), \tag{7}$$

In this equation, c represents the task index, $p_{o,c}$ is the model's probability for each class, and $y_{o,c}$ is the true class label.

The total loss was optimized by linearly combining these loss components, as described in Equation (8).

$$\text{loss} = \sum_{i}^{\text{task}} (\text{loss}_{\text{task}_i} + \text{loss}_{\text{orthogonal}_i}) + \alpha \text{loss}_{\text{adv}} \tag{8}$$

The coefficient α was set to 5 after testing values of 1, and 5. All models and training processes were implemented using PyTorch [72]. The training utilized the Adam optimizer [73] and a cosine annealing learning rate scheduler [74] were employed, with a learning rate of 1e-5, a batch size of 40, gradient accumulation of 4, a model dimension of 280, and 6 layers, with a total training duration of 10 days on an RTX A6000 GPU.

For fine-tuning, we employed binary cross-entropy loss as described in Equaion 4, using the same optimizer and scheduler as in pretraining. The fine-tuning settings included a learning rate was of 5e-5, a batch size of 128, and a total of 100 epochs.

4.6 Benchmark method

4.6.1 Bootstrapping and statistical evaluation

To assess the reliability and statistical significance of each model's performance, we applied a resampling-based evaluation framework using bootstrapping and non-parametric hypothesis testing.

Bootstrap confidence intervals. For each model and each metric—AUROC, PR-AUC, precision (PPV), recall, and mean PPVn—we generated 1,000 bootstrap replicates by sampling with replacement from the full test dataset. For each bootstrap sample, we computed the metric of interest, and used the resulting empirical distribution to estimate the 95% confidence interval via percentile estimation (2.5th to 97.5th percentiles). Metrics involving ranking (e.g., PPVn) were computed by sorting predictions and averaging top-n precision scores across the range of n.

Paired statistical testing. To assess whether performance differences between models were statistically significant, we used the Wilcoxon signed-rank test on the paired bootstrap distributions of each metric. We used the TITANiAN variants as the reference model in each task-specific setting. To account for multiple comparisons across baseline models, we applied the Bonferroni correction to the resulting p-values.

Ablation analysis. The same bootstrapping and statistical testing procedure was applied to the ablation study. For each ablated variant of TITANiAN-IM, we computed confidence intervals and conducted paired comparisons against the original model to quantify the contribution of each pretraining dataset and module. Performance differences and statistical significance were visualized to highlight the impact of each ablation.

Stratified bootstrap analysis. To evaluate model robustness across biologically meaningful subgroups, we additionally performed stratified bootstrapping based on HLA allele type (e.g., A, B, C for class I; DP, DQ, DR for class II) and peptide length (e.g., 8–11-mers). For each subgroup, bootstrap resampling was repeated within

the subgroup, and the metric distributions were independently computed. Stratified results were visualized as line plots with shaded confidence intervals.

All statistical analyses and visualizations were implemented using Python, with the `scikit-learn`, `scipy`, and `matplotlib` libraries.

4.6.2 Benchmark label

Since we defined all the tasks in TITANiAN as classification tasks, continuous measurements and scoring outputs were preprocessed into binary classification labels or standardized probability-like values depending on the context of each dataset.

For the ADA (anti-drug antibody) evaluation dataset, immunogenicity values indicated how many individuals exhibited an ADA response to a given antibody. We assigned binary labels such that any non-zero immunogenicity value was labeled as 1 (immunogenic), while 0.0 was labeled as 0 (non-immunogenic). To enable fair comparison with other models, which predominantly output humanness or human-likeness scores, we transformed those scores to align with the immunogenicity scale.

Scores where higher values indicated human-likeness (e.g., AbNatiV, Hu-mAb, Germline content) were inverted by:

$$\text{Transformed Score} = 1 - x \tag{9}$$

Percentile-based scores such as T20 were converted using:

$$\text{Transformed Score} = 1 - \frac{x}{100} \tag{10}$$

Z-score-based outputs (e.g., Z-score, IgReconstruct, AbLSTM) were standardized by applying the cumulative distribution function (CDF) of the standard normal distribution:

$$\text{Transformed Score} = 1 - \Phi(x) \quad \text{for Z-score, IgReconstruct} \tag{11}$$

$$\text{Transformed Score} = \Phi(x) \quad \text{for AbLSTM} \tag{12}$$

depending on the original directionality of the score, where $\Phi(x)$ denotes the standard normal CDF.

Likewise, OASis identity scores (loose to strict), were transformed as:

$$\text{Transformed Score} = 1 - x \tag{13}$$

For benchmarking TITANiAN-MHC, peptide-MHC binding affinity values (IC50) were converted into binary labels and normalized scores. Specifically, for entries where the "Measurement type" column indicated a non-binary value, the IC50 value was converted to a binary label: peptides with IC50 $\leq$ 500 nM were labeled as 1 (binder), and those with IC50 $>$ 500 nM were labeled as 0 (non-binder). This threshold is commonly used to define peptide-MHC binding and IEDB guideline [12]. In addition, for each MHC binding prediction model, predicted IC50 values were transformed using the following function:

$$\text{Transformed Score} = 1 - \sigma(x - 500) \tag{14}$$

where σ is the sigmoid function and x is the raw IC50 value. This transformation ensures that stronger predicted binding is mapped to a score closer to 1, facilitating consistent comparison with binary classifiers. Non-numeric entries (e.g., "-") were excluded from the transformation. The entire procedure was implemented in a Python script provided in the supplementary materials.

These label and score transformation procedures allowed all datasets and models to be consistently evaluated under a unified classification framework.

4.6.3 Embedding quality benchmark

To evaluate the quality of the embeddings generated by the integrative encoder, we first converted them into 2D tensors using t-SNE [75], implemented in scikit-learn [76]. The quality of these embeddings was then quantified using two widely recognized metrics: the Silhouette Score (SI) and Adjusted Mutual Information (AMI), which are commonly used to assess embedding and clustering quality [77–79]. The SI is calculated using Equation (15).

$$\text{SI} = \frac{b - a}{\max(a, b)} \tag{15}$$

where a represents the mean intra-cluster distance, and b denotes the mean nearest-cluster distance. The AMI quantifies the mutual information by Equation (16).

$$\text{AMI}(U, V) = \frac{\text{MI}(U, V) - \text{E}\{\text{MI}(U, V)\}}{\max\{\text{H}(U), \text{H}(V)\} - \text{E}\{\text{MI}(U, V)\}} \tag{16}$$

where MI(U,V) denotes mutual information between two distributions U and V, H(U) represents entropy of distribution U, and E(MI(U,V)) denotes expected mutual information between two distributions U and V.

Declarations

4.7 Funding

This work was supported by the National Research Foundation of Korea (NRF) grants (2020M3A9G7103933) funded by the Korea government (MIST).

4.8 Conflict of interest

The authors declare the following competing interests: Jeonghyeon Kim, Jinsung Noh, and Chaok Seok are inventors on pending domestic Patent 10-2024-0126554, co-filed by Seoul National University and Galux Inc., related to machine learning for predicting T-cell immunogenicity.

4.9 Data availability

All datasets used for training, validation, and benchmarking of TITANiAN are available at https://figshare.com/articles/dataset/TITANiAN_benchmark_result/

26323582. Data were sourced from publicly available databases, including IEDB[12], VDJdb[24], MCPAS-TCR[57], ImmuneCODE[58], TBAdb[59], 10X Genomics[60], OAS[80], UNIPROT[16], PRIME[23], MHCBN[20], BigMHC[11], Biophi[19],and Panpep[52].

4.10 Code availability

The TITANiAN code repository, which includes the model architecture, training code, inference code, and analysis code, is available at https://github.com/seoklab/immuno_mtl. Additionally, a user-friendly web server is provided for easy access and utilization of the model at https://galaxy.seoklab.org/design/titanian/.

4.11 Author contribution

Jeonghyeon Kim, Jinsung Noh, and Chaok Seok conceived the study and conducted the experiments. Jeonghyeon Kim, Nuri Jung, Jayyoon Lee, Namhyuk Cho, Jinsung Noh, and Chaok Seok analyzed the results. Jeonghyeon Kim drafted the manuscript, while Nuri Jung implemented the web server. All authors—Jeonghyeon Kim, Nuri Jung, Jayyoon Lee, Namhyuk Cho, Jinsung Noh, and Chaok Seok—contributed to revising the manuscript.

Appendix A ROC curve for benchmark results

Figure A1 (left) displays the Receiver Operating Characteristic (ROC) curve for the benchmark detailed in the main text, with the navy line indicating the ROC curve of a random classifier as a reference for each model's performance. Figure A2 (right) presents the Precision-Recall curve for the benchmarks mentioned in the main text. From the top, the curves represent: TITANiAN-IM benchmark result for Neoantigen discovery set," TITANiAN-IM benchmark result for Infectious disease set," TITANiAN-IM benchmark result for ADA level test," TITANiAN-MHC I benchmark result for IEDB benchmark set," TITANiAN-MHC II benchmark result for IEDB benchmark set," TITANiAN-TCR benchmark result for zero-shot set," and "TITANiAN-TCR benchmark result for ACT-based therapy discovery set."

Appendix B PPVn based on the number of top predictions

Figure B3 depicts the line plot of PPV vs. Top N, with a dashed parallel line representing the random classifier to provide a reference for each model's PPV vs. Top N curve. From the top left, the curves represent: TITANiAN-IM benchmark result for Neoantigen discovery set," TITANiAN-IM benchmark result for Infectious disease set," TITANiAN-MHC I benchmark result for IEDB benchmark set," TITANiAN-MHC II benchmark result for IEDB benchmark set," TITANiAN-TCR benchmark result for zero-shot set," and TITANiAN-TCR benchmark result for ACT-based therapy discovery set."

Appendix C Optimizing the coefficient for the adversarial loss

To determine the adversarial loss coefficient α, we conducted experiments to evaluate how adjusting α affects the performance of our final model, TITANiAN-IM. We pretrained each model under identical conditions, varying only α, and monitored the validation loss during the fine-tuning steps. As depicted in Figure C4, a larger α value results in lower validation loss and higher validation set AUROC, indicating that a larger α value enhances the model's generalizability in predicting immunogenicity. We attribute this to the fact that a higher adversarial loss coefficient enforces more rigorous alignment from different data sources in the integrative encoder.

Appendix D t-SNE plot of TITANiAN-IM embedding on various datasets

We also generated a t-SNE plot of the test set embeddings on vairous test set, shown in Figure D5, to demonstrate that TITANiAN-IM's performance is not solely dependent on the decoder's classification but is also influenced by the effective embeddings from pretraining and fine-tuning. The embeddings were taken from the output tensor of the integrative encoder on neoantigen discovery set and infectious disease vaccine discovery set. According to figure D5 a), for the neoantigen discovery set, the highlighted region of the figure shows that immunogenic peptides cluster with minimal non-immunogenic peptide points. Additionally, the Silhouette score (SI) and Adjusted Mutual Information (AMI) values are 0.132 and 0.017, respectively, indicating that the embedding is well-clustered compared to random embeddings, which have SI values of -1.0 and AMI values of 0.0. These analyses suggest that TITANiAN-IM's embeddings are effective in discriminating between immunogenic and non-immunogenic peptides. Moreover, the highlighted region of the figure D5 b) , which is t-SNE plot on infectious disease vaccine discovery set, shows that non-immunogenic peptides cluster with minimal immunogenic peptide points. The Silhouette score (SI) and Adjusted Mutual Information (AMI) values are 0.203 and 0.045, respectively, indicating that the embedding is well-clustered compared to random embeddings, which have SI values of -1.0 and AMI values of 0.0. These analyses suggest that TITANiAN-IM's embeddings are effective in discriminating between immunogenic and non-immunogenic peptides also in infectious disease vaccine discovery set.

Figure D5 c) shows the t-SNE plot of TITANiAN-IM's embeddings for the ADA benchmark set. To obtain the embedding of each antibody drug sequence, we applied max pooling to each peptide's sequence. Since immunogenicity can be triggered by a single problematic sequence, max pooling was used to derive the protein's overall embedding. The highlighted region of the figure demonstrates that TITANiAN-IM effectively creates meaningful embeddings regarding the immunogenicity of protein sequences. The Silhouette score (SI) and Adjusted Mutual Information (AMI) values are 0.057 and 0.045, respectively, indicating that the embedding is well-clustered compared to random embeddings, which have SI values of -1.0 and AMI values of 0.0.This

indicates that TITANiAN's performance does not solely depend on the decoding classifier.

References

[1] Hudson, D., Fernandes, R.A., Basham, M., Ogg, G., Koohy, H.: Can we predict t cell specificity with digital biology and machine learning? Nature Reviews Immunology **23**(8), 511–521 (2023)

[2] Weber, C.A., Mehta, P.J., Ardito, M., Moise, L., Martin, B., De Groot, A.S.: T cell epitope: friend or foe? immunogenicity of biologics in context. Advanced drug delivery reviews **61**(11), 965–976 (2009)

[3] Schaap-Johansen, A.-L., Vujović, M., Borch, A., Hadrup, S.R., Marcatili, P.: T cell epitope prediction and its application to immunotherapy. Frontiers in Immunology **12**, 712488 (2021)

[4] Shao, X.M., Bhattacharya, R., Huang, J., Sivakumar, I.A., Tokheim, C., Zheng, L., Hirsch, D., Kaminow, B., Omdahl, A., Bonsack, M., *et al.*: High-throughput prediction of mhc class i and ii neoantigens with mhcnuggets. Cancer immunology research **8**(3), 396–408 (2020)

[5] Brummelen, E.M., Ros, W., Wolbink, G., Beijnen, J.H., Schellens, J.H.: Antidrug antibody formation in oncology: clinical relevance and challenges. The oncologist **21**(10), 1260–1268 (2016)

[6] Duke, B.R., Mitra-Kaushik, S.: Current in vitro assays for prediction of t cell mediated immunogenicity of biotherapeutics and manufacturing impurities. Journal of Pharmaceutical Innovation **15**, 202–218 (2020)

[7] Jawa, V., Cousens, L.P., Awwad, M., Wakshull, E., Kropshofer, H., De Groot, A.S.: T-cell dependent immunogenicity of protein therapeutics: preclinical assessment and mitigation. Clinical immunology **149**(3), 534–555 (2013)

[8] Habel, J.R., Nguyen, T.H., Sandt, C.E., Juno, J.A., Chaurasia, P., Wragg, K., Koutsakos, M., Hensen, L., Jia, X., Chua, B., *et al.*: Suboptimal sars-cov-2- specific cd8+ t cell response associated with the prominent hla-a* 02: 01 phenotype. Proceedings of the National Academy of Sciences **117**(39), 24384–24391 (2020)

[9] Le Bert, N., Tan, A.T., Kunasegaran, K., Tham, C.Y., Hafezi, M., Chia, A., Chng, M.H.Y., Lin, M., Tan, N., Linster, M., *et al.*: Sars-cov-2-specific t cell immunity in cases of covid-19 and sars, and uninfected controls. Nature **584**(7821), 457–462 (2020)

[10] 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

research **48**(W1), 449–454 (2020)

[11] Albert, B.A., Yang, Y., Shao, X.M., Singh, D., Smith, K.N., Anagnostou, V., Karchin, R.: Deep neural networks predict class i major histocompatibility complex epitope presentation and transfer learn neoepitope immunogenicity. Nature machine intelligence **5**(8), 861–872 (2023)

[12] Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A., Peters, B.: The immune epitope database (iedb): 2018 update. Nucleic acids research **47**(D1), 339–343 (2019)

[13] O'Donnell, T.J., Rubinsteyn, A., Laserson, U.: Mhcflurry 2.0: improved pan-allele prediction of mhc class i-presented peptides by incorporating antigen processing. Cell systems **11**(1), 42–48 (2020)

[14] Xu, Z., Luo, M., Lin, W., Xue, G., Wang, P., Jin, X., Xu, C., Zhou, W., Cai, Y., Yang, W., *et al.*: Dlptcr: an ensemble deep learning framework for predicting immunogenic peptide recognized by t cell receptor. Briefings in Bioinformatics **22**(6), 335 (2021)

[15] Kovaltsuk, A., Leem, J., Kelm, S., Snowden, J., Deane, C.M., Krawczyk, K.: Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. The Journal of Immunology **201**(8), 2502–2509 (2018)

[16] Consortium, U.: Uniprot: a worldwide hub of protein knowledge. Nucleic acids research **47**(D1), 506–515 (2019)

[17] Marks, C., Hummer, A.M., Chin, M., Deane, C.M.: Humanization of antibodies using a machine learning approach on large-scale repertoire data. Bioinformatics **37**(22), 4041–4047 (2021)

[18] Ramon, A., Ali, M., Atkinson, M., Saturnino, A., Didi, K., Visentin, C., Ricagno, S., Xu, X., Greenig, M., Sormanni, P.: Assessing antibody and nanobody nativeness for hit selection and humanization with abnativ. Nature Machine Intelligence, 1–18 (2024)

[19] Prihoda, D., Maamary, J., Waight, A., Juan, V., Fayadat-Dilman, L., Svozil, D., Bitton, D.A.: Biophi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. In: MAbs, vol. 14, p. 2020203 (2022). Taylor & Francis

[20] Lata, S., Bhasin, M., Raghava, G.P.: Mhcbn 4.0: A database of mhc/tap binding peptides and t-cell epitopes. BMC research notes **2**, 1–6 (2009)

[21] Müller, M., Huber, F., Arnaud, M., Kraemer, A.I., Altimiras, E.R., Michaux, J., Taillandier-Coindard, M., Chiffelle, J., Murgues, B., Gehret, T., *et al.*: Machine learning methods and harmonized datasets improve immunogenic neoantigen

prediction. Immunity **56**(11), 2650–2663 (2023)

[22] Schmidt, J., Smith, A.R., Magnin, M., Racle, J., Devlin, J.R., Bobisse, S., Cesbron, J., Bonnet, V., Carmona, S.J., Huber, F., et al.: Prediction of neo-epitope immunogenicity reveals tcr recognition determinants and provides insight into immunoediting. Cell Reports Medicine **2**(2) (2021)

[23] Gfeller, D., Schmidt, J., Croce, G., Guillaume, P., Bobisse, S., Genolet, R., Queiroz, L., Cesbron, J., Racle, J., Harari, A.: Improved predictions of antigen presentation and tcr recognition with mixmhcpred2. 2 and prime2. 0 reveal potent sars-cov-2 cd8+ t-cell epitopes. Cell Systems **14**(1), 72–83 (2023)

[24] Goncharov, M., Bagaev, D., Shcherbinin, D., Zvyagin, I., Bolotin, D., Thomas, P.G., Minervina, A.A., Pogorelyy, M.V., Ladell, K., McLaren, J.E., *et al.*: Vdjdb in the pandemic era: a compendium of t cell receptors specific for sars-cov-2. Nature methods **19**(9), 1017–1019 (2022)

[25] Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W.: Cd-hit: accelerated for clustering the next-generation sequencing data. Bioinformatics **28**(23), 3150–3152 (2012)

[26] Perfetto, S.P., Ambrozak, D., Nguyen, R., Chattopadhyay, P., Roederer, M.: Quality assurance for polychromatic flow cytometry. Nature protocols **1**(3), 1522–1530 (2006)

[27] Janetzki, S., Price, L., Schroeder, H., Britten, C.M., Welters, M.J., Hoos, A.: Guidelines for the automated evaluation of elispot assays. Nature protocols **10**(7), 1098–1115 (2015)

[28] Chu, Y., Zhang, Y., Wang, Q., Zhang, L., Wang, X., Wang, Y., Salahub, D.R., Xu, Q., Wang, J., Jiang, X., *et al.*: A transformer-based model to predict peptide–hla class i binding and optimize mutated peptides for vaccine design. Nature Machine Intelligence **4**(3), 300–311 (2022)

[29] Sarkizova, S., Klaeger, S., Le, P.M., Li, L.W., Oliveira, G., Keshishian, H., Hartigan, C.R., Zhang, W., Braun, D.A., Ligon, K.L., *et al.*: A large peptidome dataset improves hla class i epitope prediction across most of the human population. Nature biotechnology **38**(2), 199–209 (2020)

[30] Gao, S.H., Huang, K., Tu, H., Adler, A.S.: Monoclonal antibody humanness score and its applications. BMC biotechnology **13**, 1–12 (2013)

[31] Abhinandan, K., Martin, A.C.: Analyzing the "degree of humanness" of antibody sequences. Journal of molecular biology **369**(3), 852–862 (2007)

[32] Schmitz, S., Soto, C., Crowe Jr, J.E., Meiler, J.: Human-likeness of antibody biologics determined by back-translation and comparison with large antibody variable gene repertoires. In: MAbs, vol. 12, p. 1758291 (2020). Taylor & Francis

[33] Wollacott, A.M., Xue, C., Qin, Q., Hua, J., Bohnuud, T., Viswanathan, K., Kolachalama, V.B.: Quantifying the nativeness of antibody sequences using long short-term memory networks. Protein Engineering, Design and Selection **32**(7), 347–354 (2019)

[34] Lazar, G.A., Desjarlais, J.R., Jacinto, J., Karki, S., Hammond, P.W.: A molecular immunology approach to antibody humanization and functional optimization. Molecular immunology **44**(8), 1986–1998 (2007)

[35] Clavero-Álvarez, A., Di Mambro, T., Perez-Gaviro, S., Magnani, M., Bruscolini, P.: Humanization of antibodies using a statistical inference approach. Scientific reports **8**(1), 14820 (2018)

[36] Koşaloğlu-Yalçın, Z., Lanka, M., Frentzen, A., Logandha Ramamoorthy Premlal, A., Sidney, J., Vaughan, K., Greenbaum, J., Robbins, P., Gartner, J., Sette, A., *et al.*: Predicting t cell recognition of mhc class i restricted neoepitopes. Oncoimmunology **7**(11), 1492508 (2018)

[37] Peters, B., Sette, A.: Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. BMC bioinformatics **6**, 1–9 (2005)

[38] Andreatta, M., Nielsen, M.: Gapped sequence alignment using artificial neural networks: application to the mhc class i system. Bioinformatics **32**(4), 511–517 (2016)

[39] Bui, H.-H., Sidney, J., Peters, B., Sathiamurthy, M., Sinichi, A., Purton, K.-A., Mothé, B.R., Chisari, F.V., Watkins, D.I., Sette, A.: Automated generation and evaluation of specific mhc binding predictive tools: Arb matrix applications. Immunogenetics **57**, 304–314 (2005)

[40] Kim, Y., Sidney, J., Pinilla, C., Sette, A., Peters, B.: Derivation of an amino acid similarity matrix for peptide: Mhc binding and its application as a bayesian prior. BMC bioinformatics **10**, 1–11 (2009)

[41] Moutaftsi, M., Peters, B., Pasquetto, V., Tscharke, D.C., Sidney, J., Bui, H.-H., Grey, H., Sette, A.: A consensus epitope prediction approach identifies the breadth of murine tcd8+-cell responses to vaccinia virus. Nature biotechnology **24**(7), 817–819 (2006)

[42] Karosiene, E., Lundegaard, C., Lund, O., Nielsen, M.: Netmhccons: a consensus method for the major histocompatibility complex class i predictions. Immunogenetics **64**, 177–186 (2012)

[43] Zhang, H., Lund, O., Nielsen, M.: The pickpocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to mhc-peptide binding. Bioinformatics **25**(10), 1293–1299 (2009)

[44] O'Donnell, T.J., Rubinsteyn, A., Bonsack, M., Riemer, A.B., Laserson, U., Hammerbacher, J.: Mhcflurry: open-source class i mhc binding affinity prediction. Cell systems **7**(1), 129–132 (2018)

[45] Lundegaard, C., Lamberth, K., Harndahl, M., Buus, S., Lund, O., Nielsen, M.: Netmhc-3.0: accurate web accessible predictions of human, mouse and monkey mhc class i affinities for peptides of length 8–11. Nucleic acids research **36**(suppl_2), 509–512 (2008)

[46] Nielsen, M., Andreatta, M.: Netmhcpan-3.0; improved prediction of binding to mhc class i molecules integrating information from multiple receptor and peptide length datasets. Genome medicine **8**, 1–9 (2016)

[47] Nielsen, M., Lund, O.: Nn-align. an artificial neural network-based alignment algorithm for mhc class ii peptide binding prediction. BMC bioinformatics **10**, 1–10 (2009)

[48] Andreatta, M., Karosiene, E., Rasmussen, M., Stryhn, A., Buus, S., Nielsen, M.: Accurate pan-specific prediction of peptide-mhc class ii binding affinity with improved binding core identification. Immunogenetics **67**, 641–650 (2015)

[49] Nielsen, M., Lundegaard, C., Lund, O.: Prediction of mhc class ii binding affinity using smm-align, a novel stabilization matrix alignment method. BMC bioinformatics **8**, 1–12 (2007)

[50] Sturniolo, T., Bono, E., Ding, J., Raddrizzani, L., Tuereci, O., Sahin, U., Braxenthaler, M., Gallazzi, F., Protti, M.P., Sinigaglia, F., *et al.*: Generation of tissue-specific and promiscuous hla ligand databases using dna microarrays and virtual hla class ii matrices. Nature biotechnology **17**(6), 555–561 (1999)

[51] Wang, P., Sidney, J., Dow, C., Mothé, B., Sette, A., Peters, B.: A systematic assessment of mhc class ii peptide binding predictions and evaluation of a consensus approach. PLoS computational biology **4**(4), 1000048 (2008)

[52] Gao, Y., Gao, Y., Fan, Y., Zhu, C., Wei, Z., Zhou, C., Chuai, G., Chen, Q., Zhang, H., Liu, Q.: Pan-peptide meta learning for t-cell receptor–antigen binding recognition. Nature Machine Intelligence **5**(3), 236–249 (2023)

[53] Lu, T., Zhang, Z., Zhu, J., Wang, Y., Jiang, P., Xiao, X., Bernatchez, C., Heymach, J.V., Gibbons, D.L., Wang, J., *et al.*: Deep learning-based prediction of the t cell receptor–antigen binding specificity. Nature machine intelligence **3**(10), 864–875 (2021)

[54] Springer, I., Besser, H., Tickotsky-Moskovitz, N., Dvorkin, S., Louzoun, Y.: Prediction of specific tcr-peptide binding from large dictionaries of tcr-peptide pairs. Frontiers in immunology **11**, 534364 (2020)

[55] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[56] Deng, L., Ly, C., Abdollahi, S., Zhao, Y., Prinz, I., Bonn, S.: Performance comparison of tcr-pmhc prediction tools reveals a strong data dependency. Frontiers in Immunology **14**, 1128326 (2023)

[57] 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**(18), 2924–2929 (2017)

[58] Nolan, S., Vignali, M., Klinger, M., Dines, J.N., Kaplan, I.M., Svejnoha, E., Craft, T., Boland, K., Pesesky, M., Gittelman, R.M., 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. Research square (2020)

[59] Zhang, W., Wang, L., Liu, K., Wei, X., Yang, K., Du, W., Wang, S., Guo, N., Ma, C., Luo, L., *et al.*: Pird: pan immune repertoire database. Bioinformatics **36**(3), 897–903 (2020)

[60] Genomics: A new way of exploring immunity–linking highly multiplexed antigen recognition to immune repertoire and phenotype. Tech. rep (2019)

[61] Dean, J., Emerson, R.O., Vignali, M., Sherwood, A.M., Rieder, M.J., Carlson, C.S., Robins, H.S.: Annotation of pseudogenic gene segments by massively parallel sequencing of rearranged lymphocyte receptor loci. Genome medicine **7**, 1–8 (2015)

[62] Kim, J.Y., Cha, H., Kim, K., Sung, C., An, J., Bang, H., Kim, H., Yang, J.O., Chang, S., Shin, I., *et al.*: Mhc ii immunogenicity shapes the neoepitope landscape in human tumors. Nature Genetics **55**(2), 221–231 (2023)

[63] Yang, K.K., Fusi, N., Lu, A.X.: Convolutions are competitive with transformers for protein sequence pretraining. Cell Systems **15**(3), 286–294 (2024)

[64] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., *et al.*: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**(6637), 1123–1130 (2023)

[65] Ferruz, N., Schmidt, S., Höcker, B.: Protgpt2 is a deep unsupervised language model for protein design. Nature communications **13**(1), 4348 (2022)

[66] Wu, J., Wang, W., Zhang, J., Zhou, B., Zhao, W., Su, Z., Gu, X., Wu, J., Zhou, Z., Chen, S.: Deephlapan: a deep learning approach for neoantigen prediction considering both hla-peptide binding and immunogenicity. Frontiers in immunology **10**, 2559 (2019)

[67] Li, G., Iyer, B., Prasath, V.S., Ni, Y., Salomonis, N.: Deepimmuno: deep learning-empowered prediction and generation of immunogenic peptides for t-cell immunity. Briefings in bioinformatics **22**(6), 160 (2021)

[68] Yang, X., Zhao, L., Wei, F., Li, J.: Deepnetbim: deep learning model for predicting hla-epitope interactions based on network analysis by harnessing binding and immunogenicity information. BMC bioinformatics **22**(1), 231 (2021)

[69] Schoenauer-Sebag, A., Heinrich, L., Schoenauer, M., Sebag, M., Wu, L.F., Altschuler, S.J.: Multi-domain adversarial learning. arXiv preprint arXiv:1903.09239 (2019)

[70] Orbes-Arteaga, M., Varsavsky, T., Sudre, C.H., Eaton-Rosen, Z., Haddow, L.J., Sørensen, L., Nielsen, M., Pai, A., Ourselin, S., Modat, M., *et al.*: Multi-domain adaptation in brain mri through paired consistency and adversarial learning. In: Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data: First MICCAI Workshop, DART 2019, and First International Workshop, MIL3ID 2019, Shenzhen, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings 1, pp. 54–62 (2019). Springer

[71] Wang, X., Cheng, Y., Yang, Y., Yu, Y., Li, F., Peng, S.: Multitask joint strategies of self-supervised representation learning on biomedical networks for drug discovery. Nature Machine Intelligence **5**(4), 445–456 (2023)

[72] Ketkar, N., Moolayil, J., Ketkar, N., Moolayil, J.: Introduction to pytorch. Deep learning with python: learn best practices of deep learning models with PyTorch, 27–91 (2021)

[73] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[74] Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)

[75] Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)

[76] Kramer, O., Kramer, O.: Scikit-learn. Machine learning for evolution strategies, 45–53 (2016)

[77] Shahapure, K.R., Nicholas, C.: Cluster quality analysis using silhouette score. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pp. 747–748 (2020). IEEE

[78] Lovmar, L., Ahlford, A., Jonsson, M., Syvänen, A.-C.: Silhouette scores for assessment of snp genotype clusters. BMC genomics **6**, 1–6 (2005)

[79] Romano, S., Bailey, J., Nguyen, V., Verspoor, K.: Standardized mutual information for clustering comparisons: one step further in adjustment for chance. In: International Conference on Machine Learning, pp. 1143–1151 (2014). PMLR

[80] Olsen, T.H., Boyles, F., Deane, C.M.: Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Science **31**(1), 141–146 (2022)

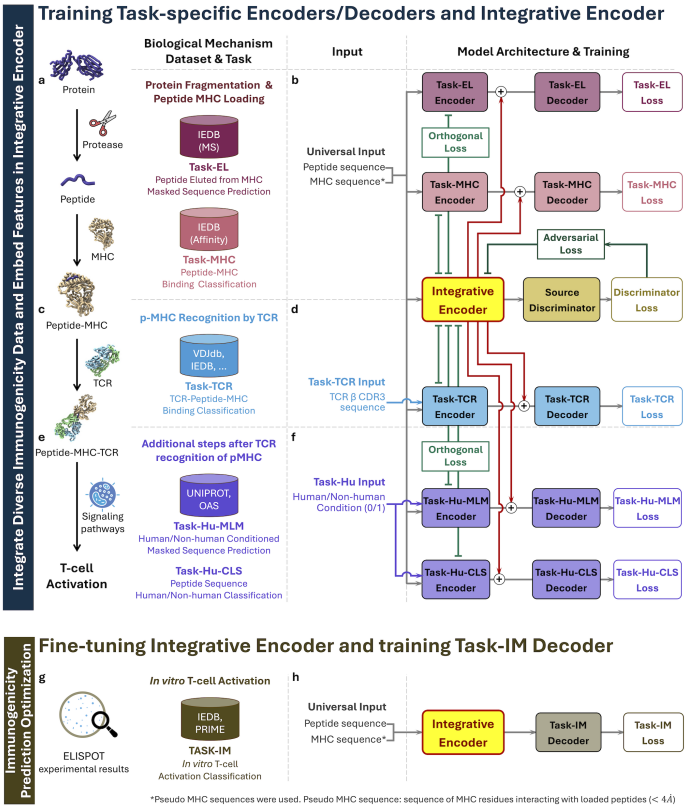


Fig. 1 Schematic overview of the TITANiAN framework architecture and two-phase training strategy. TITANiAN adopts a modular multi-task learning architecture comprising a shared integrative encoder (yellow box) that learns domain-invariant features common across tasks, alongside task-specific encoders that extract complementary, domain-specific information. Corresponding task-specific decoders use these representations to generate predictions. Training proceeds in two phases:(a–f) Pretraining phase — The integrative and task-specific modules are jointly trained on diverse datasets spanning multiple immunogenicity-relevant tasks: MHC presentation and peptide loading prediction (Task-EL I, II; a, b), peptide–MHC binding affinity prediction (Task-MHC I, II; a, b), TCR–peptide–MHC binding prediction (Task-TCR; c, d), and peptide "selfness" learning (Task-Hu-CLS/MLM; e, f). This phase incorporates three learning signals: (i) task-specific losses from each decoder head, (ii) orthogonality constraints (light green lines) to enforce disentangled representations between the integrative and task-specific encoders, and (iii) an adversarial loss (dark green line) guided by a source discriminator (gold box), which encourages the integrative encoder to produce domain-invariant representations. This strategy enhances generalizability and prioritizes learning of fundamental immunogenic features shared across datasets. (g–h) Fine-tuning phase — The pre-trained integrative encoder is further fine-tuned using a smaller T-cell activation dataset, with a dedicated Task-IM decoder, to directly optimize for the final immunogenicity prediction.

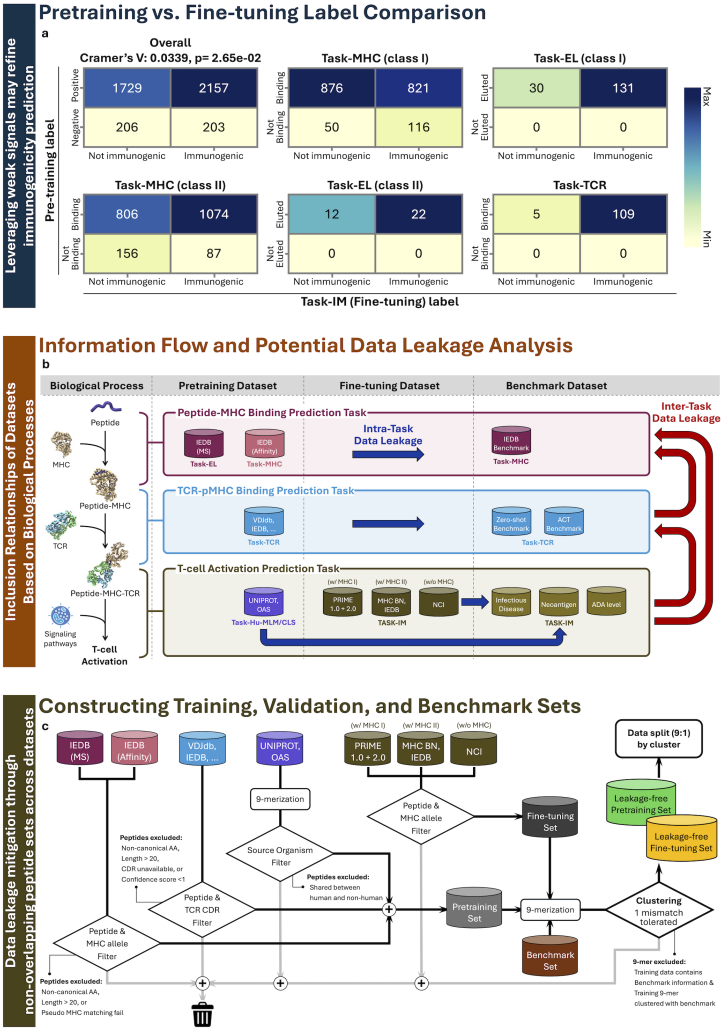


Fig. 2 Data-driven Motivation and Curation Strategy for Multi-Domain Learning. (a) Label concordance analysis between intermediate tasks and final immunogenicity. Comparison of labels from intermediate biological datasets (e.g., peptide-MHC binding) and cytokine-based immunogenicity assays reveals frequent label inversion, supporting the need for direct immunogenicity prediction. Despite a weak overall association (Cramer's V = 0.0339, p = 0.0265), the correlation is statistically significant, suggesting intermediate tasks provide informative, albeit complex, signals. (b) Data inclusion hierarchy. Schematic illustrating the biological information hierarchy among datasets. T-cell activation data inherently captures upstream information (peptide–MHC binding, TCR–pMHC recognition), while the reverse does not hold. This asymmetry guided our data leakage assessment strategy. (c) Dataset curation and splitting workflow. Flowchart outlining the pipeline for dataset filtering and leakage removal. This involved tailored preprocessing for each dataset, followed by stringent 9-mer based similarity filtering using `cd-hit` to ensure no overlap (removed peptides with less than 2 mismatches) between training and benchmark sets. The resulting clustered split enables robust, leakage-free model evaluation.

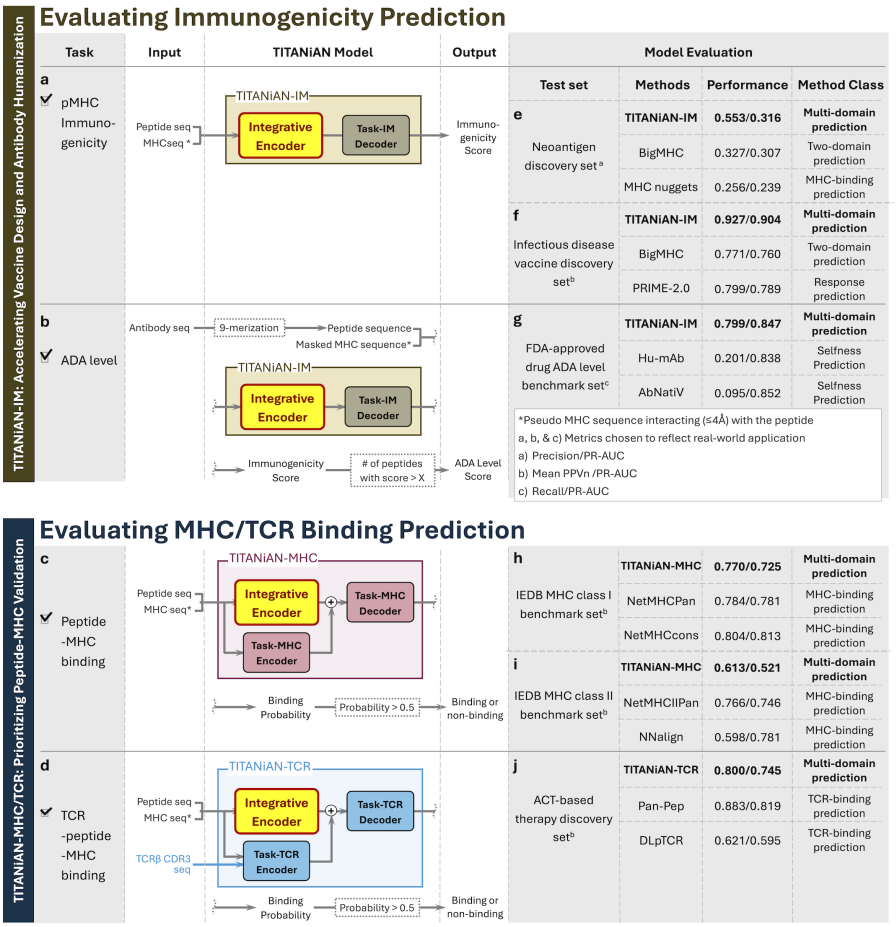


Fig. 3 TITANiAN module usage for prediction tasks (a-d) and performance comparison across benchmarks (e-j). TITANiAN utilizes different module configurations for specific tasks. (a) For T-cell epitope immunogenicity prediction and (b) ADA level prediction, TITANiAN-IM uses the integrative encoder output with the Task-IM decoder. (c) For peptide-MHC binding (TITANiAN-MHC) and (d) TCR-peptide-MHC binding (TITANiAN-TCR), features from the shared integrative encoder and the respective task-specific encoder are concatenated before being processed by the task-specific decoder. Performance summaries are presented for TITANiAN-IM assessed on (e) the neoantigen discovery set, (f) the infectious disease vaccine discovery set, and (g) the FDA-approved ADA level benchmark; TITANiAN-MHC evaluated on (h) the IEDB MHC class I and (i) IEDB MHC class II benchmarks; and TITANiAN-TCR tested on (j) the ACT-based therapy discovery set. Performance metrics appropriate for each benchmark task are specified within the panels.

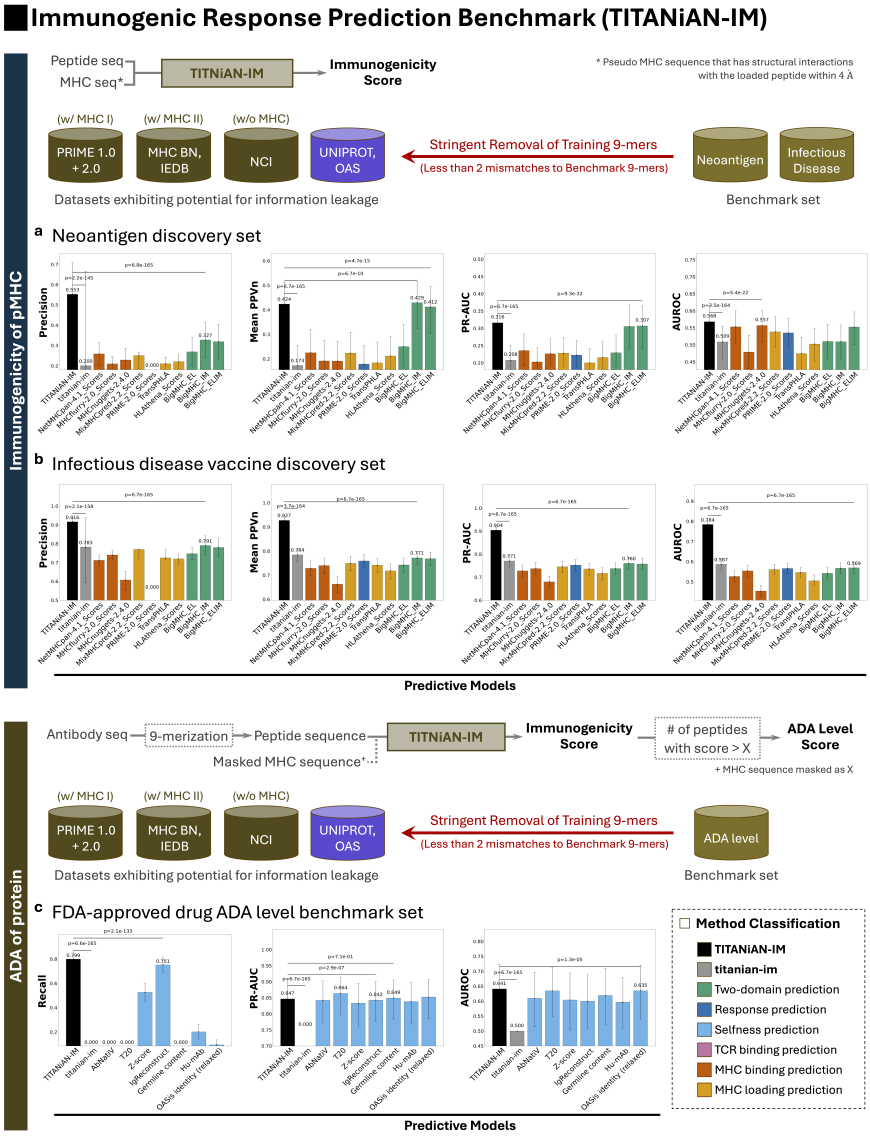


Fig. 4 Performance comparison of TITANiAN-IM on immunogenicity prediction benchmarks. Performance comparison of TITANiAN-IM with competing prediction methods (color-coded by method type, see legend) for T-cell epitope immunogenicity prediction on (a) the neoantigen discovery set and (b) the infectious disease vaccine discovery set, evaluated using Precision, Mean PPVn, PR-AUC, and AUROC. (c) Performance comparison for predicting ADA generation levels in therapeutic antibodies, evaluated using Recall, PR-AUC, and AUROC. The performance of a baseline variant (titanian-im), trained without multi-domain pre-training but sharing the same architecture, is included for comparison. Error bars represent 95% confidence intervals estimated via bootstrapping; statistical significance (p-values) was assessed using paired Wilcoxon signed-rank tests with Bonferroni correction where applicable.

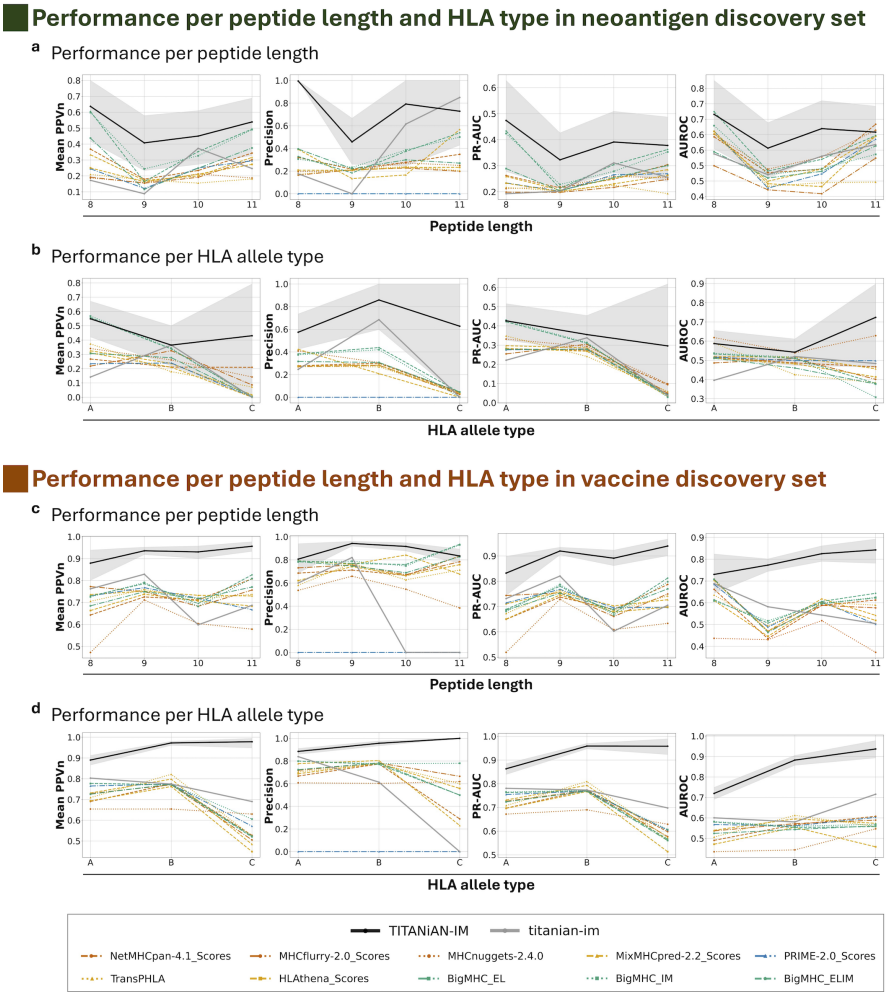


Fig. 5 Mean performance by peptide length and HLA allele type. (a) Neoantigen discovery set — peptide length: TITANiAN-IM consistently demonstrates strong performance across all tested peptide lengths, outperforming or matching competing methods.(b) Neoantigen discovery set — HLA allele type: TITANiAN-IM maintains robust performance across diverse HLA types, generally performing at a level comparable to or exceeding other models.(c) Infectious disease vaccine set — peptide length: Consistent with the neoantigen results, TITANiAN-IM sustains strong performance across all peptide lengths relative to other methods.(d) Infectious disease vaccine set — HLA allele type: TITANiAN-IM shows consistently strong and stable performance across all major HLA types, outperforming or matching alternative approaches on this benchmark.

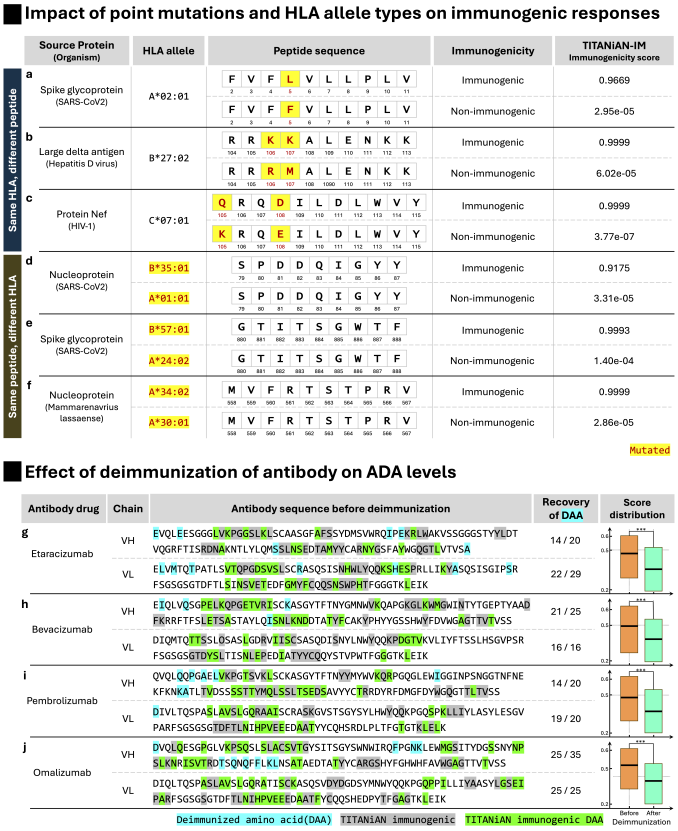


Fig. 6 Case studies demonstrating TITANiAN-IM performance. Example cases illustrate TITANiAN-IM's ability to predict: (a–f) the impact of point mutations and HLA allele types on immunogenicity, and (g–j) the effect of deimmunization on predicted ADA levels for therapeutic antibodies. (a-c) The model accurately discriminates single amino acid variations leading to significant immunogenicity changes for antigens such as (a) SARS-CoV-2 Spike glycoprotein, (b) Hepatitis D virus Large delta antigen, and (c) HIV-1 Nef protein. (d-f) It also effectively captures the influence of HLA allele differences on immunogenicity for identical peptides from (d) SARS-CoV-2 Nucleoprotein, (e) SARS-CoV-2 Spike glycoprotein, and (f) Lassa virus Nucleoprotein. (g-j) Furthermore, TITANiAN-IM identifies residues experimentally mutated for antibody deimmunization, showing distinct ADA-level score differences between sequences before and after deimmunization for (g) Etaracizumab, (h) Bevacizumab, (i) Pembrolizumab, and (j) Omalizumab. Statistically significant score differences ($p < 0.001$) are denoted by ***. Amino acid color codes and residue numbering conventions are provided within the figure panels.

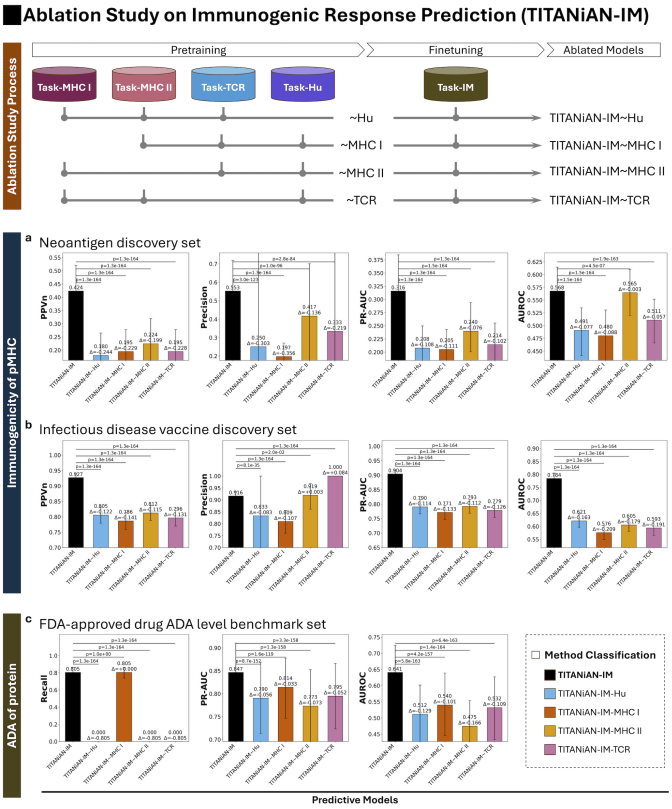


Fig. 7 Ablation study evaluating pre-training component contributions to TITANiAN-IM performance. Each variant model was generated by excluding a specific dataset and its corresponding task/module during pre-training, followed by identical fine-tuning on Task-IM. Performance was evaluated on (a) the neoantigen discovery set, (b) the infectious disease vaccine set, and (c) the ADA-level benchmark set using five key metrics. Results are shown with 95% confidence intervals (estimated via bootstrapping) and statistical significance annotations (p-values from paired Wilcoxon signed-rank test with Bonferroni correction).

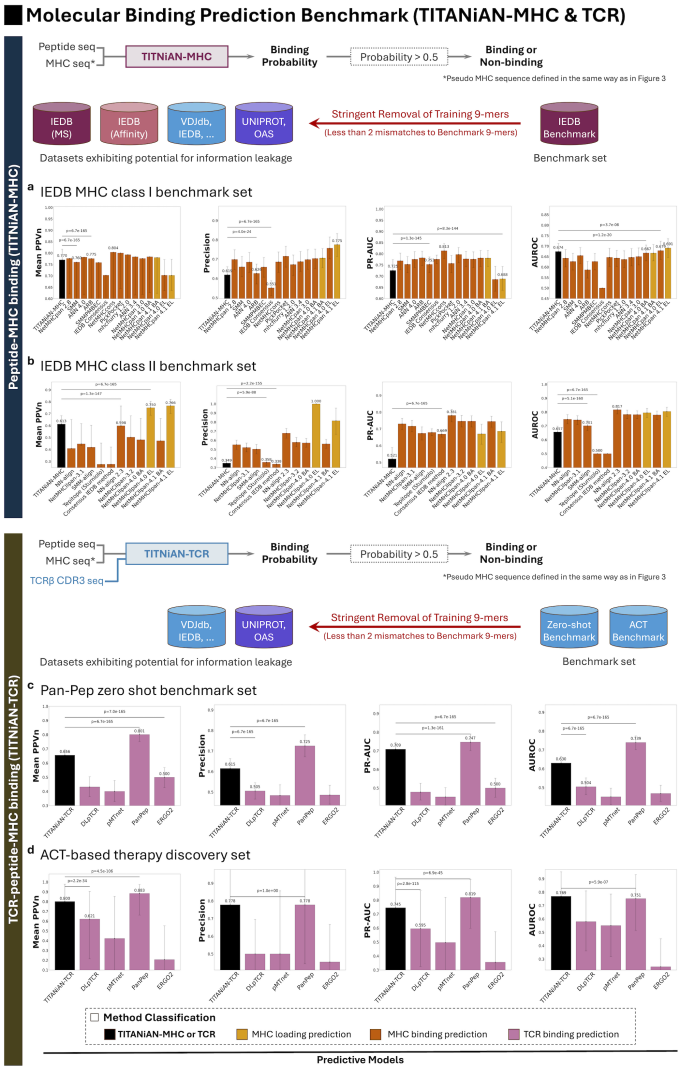


Fig. 8 Performance of pre-trained TITANiAN auxiliary modules on binding prediction benchmarks. Comparison of TITANiAN-MHC I (a), TITANiAN-MHC II (b), and TITANiAN-TCR (c, d)—utilized solely with parameters learned during multi-domain pre-training without task-specific fine-tuning—against established task-specific binding prediction methods. Evaluations were performed on benchmarks specific to each task: (a) IEDB MHC class I binding (using data post 2020-12-25), (b) IEDB MHC class II binding (post 2021-02-19), (c) PanPep Zero-shot TCR binding, and (d) ACT-based therapy discovery TCR binding. Performance was assessed using Mean PPVn, Precision, PR-AUC, and AUROC. All metrics are presented with 95% confidence intervals (bootstrapping); statistical significance (p-values) was evaluated using paired Wilcoxon signed-rank tests with Bonferroni correction.

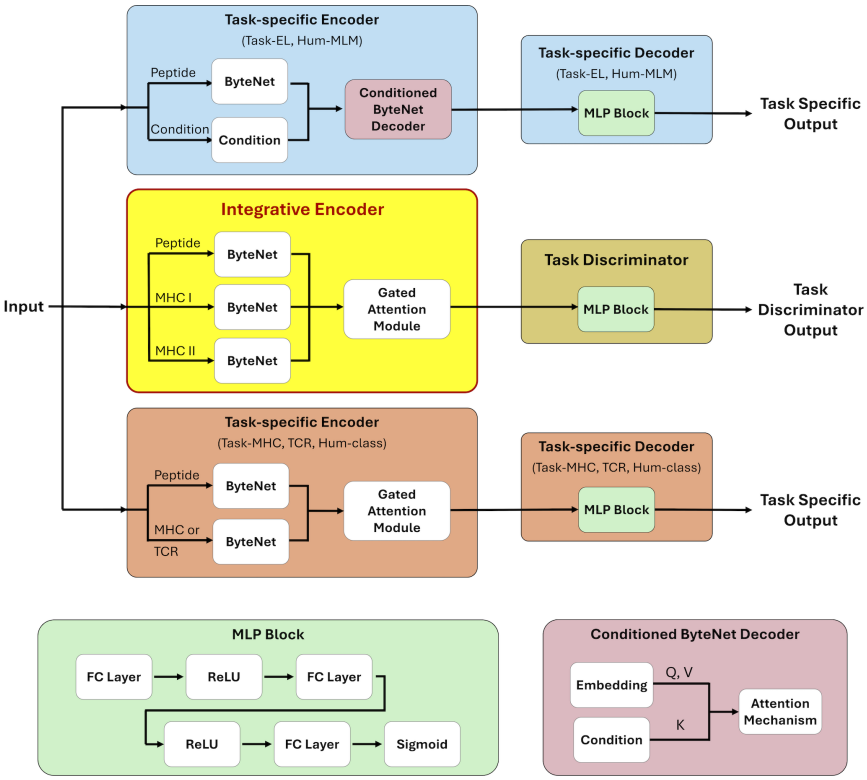


Fig. 9 Detailed architecture of TITANiAN

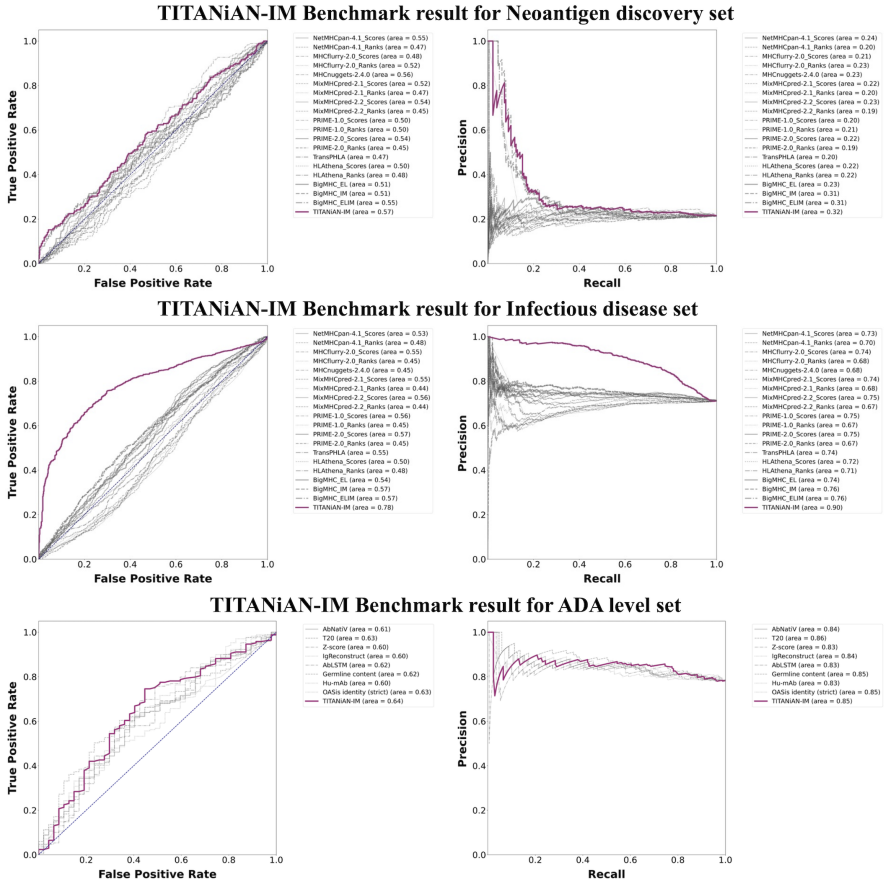


Fig. A1 Left) Receiver Operating Characteristic (ROC) curve for the benchmarks described in the main text. The navy line represents the ROC curve of a random classifier, providing a reference for each model's ROC curve performance. Right) Precision-Recall curve for the benchmarks described in the main text. From the top, each curve represents: "TITANiAN-IM benchmark result for Neoantigen discovery set", "TITANiAN-IM benchmark result for Infectious disease set", and "TITANiAN-IM benchmark result for ADA level test".

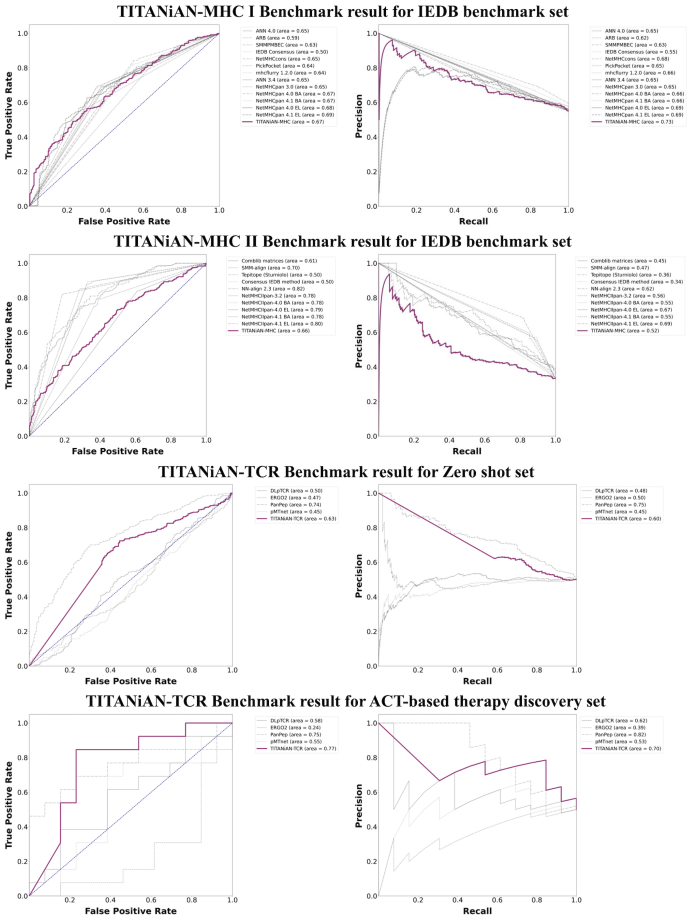


Fig. A2 Left) Receiver Operating Characteristic (ROC) curve for the benchmarks described in the main text. The navy line represents the ROC curve of a random classifier, providing a reference for each model's ROC curve performance. Right) Precision-Recall curve for the benchmarks described in the main text. From the top, each curve represents: "TITANiAN-MHC I benchmark result for IEDB benchmark set", "TITANiAN-MHC II benchmark result for IEDB benchmark set", "TITANiAN-TCR benchmark result for zero-shot set", and "TITANiAN-TCR benchmark result for ACT-based therapy discovery set"

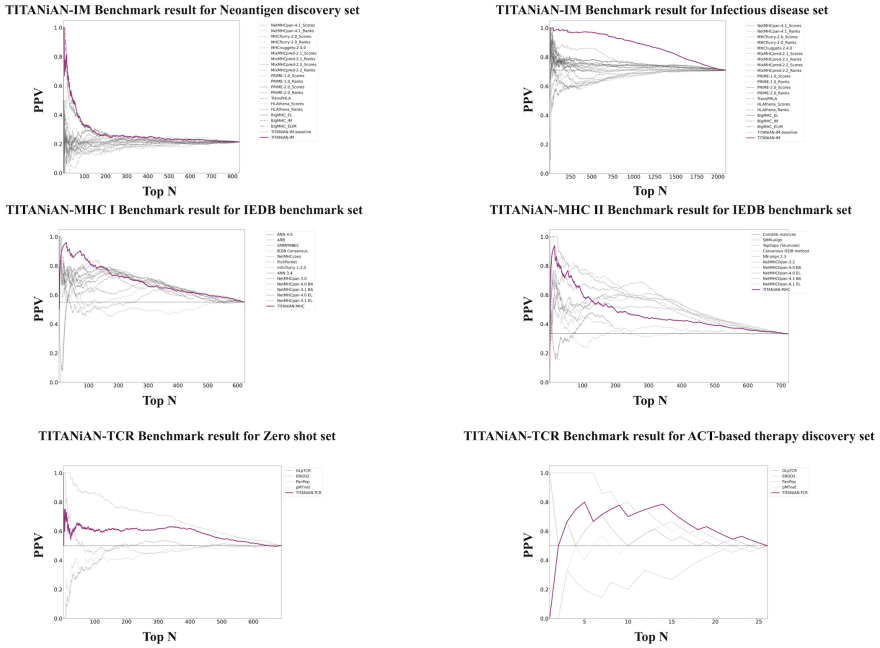


Fig. B3 Line plot of PPV vs Top N. A dashed parallel line represents the random classifier, providing a reference for each model's PPV vs. Top N curve. From the top left, each curve represents: "TITANiAN-IM benchmark result for Neoantigen discovery set", "TITANiAN-IM benchmark result for Infectious disease set", "TITANiAN-MHC I benchmark result for IEDB benchmark set", "TITANiAN-MHC II benchmark result for IEDB benchmark set", "TITANiAN-TCR benchmark result for zero-shot set", and "TITANiAN-TCR benchmark result for ACT-based therapy discovery set".

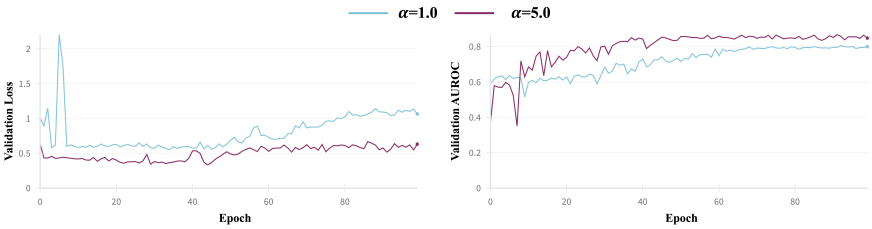


Fig. C4 Left) Epoch vs. Validation Loss curve, and Right) Epoch vs. Validation set AUROC curve in fine-tuning steps, depicting the impact of varying the adversarial loss coefficient α in the pretraining step. In both figures, higher α values result in improved performance, leading to the selection of $\alpha = 5.0$.

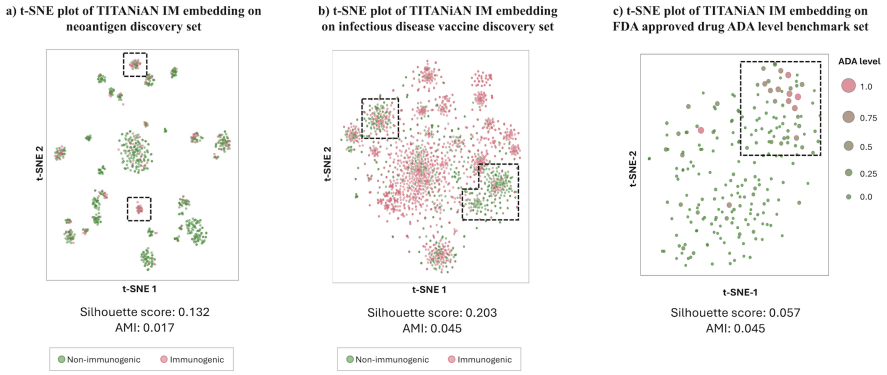


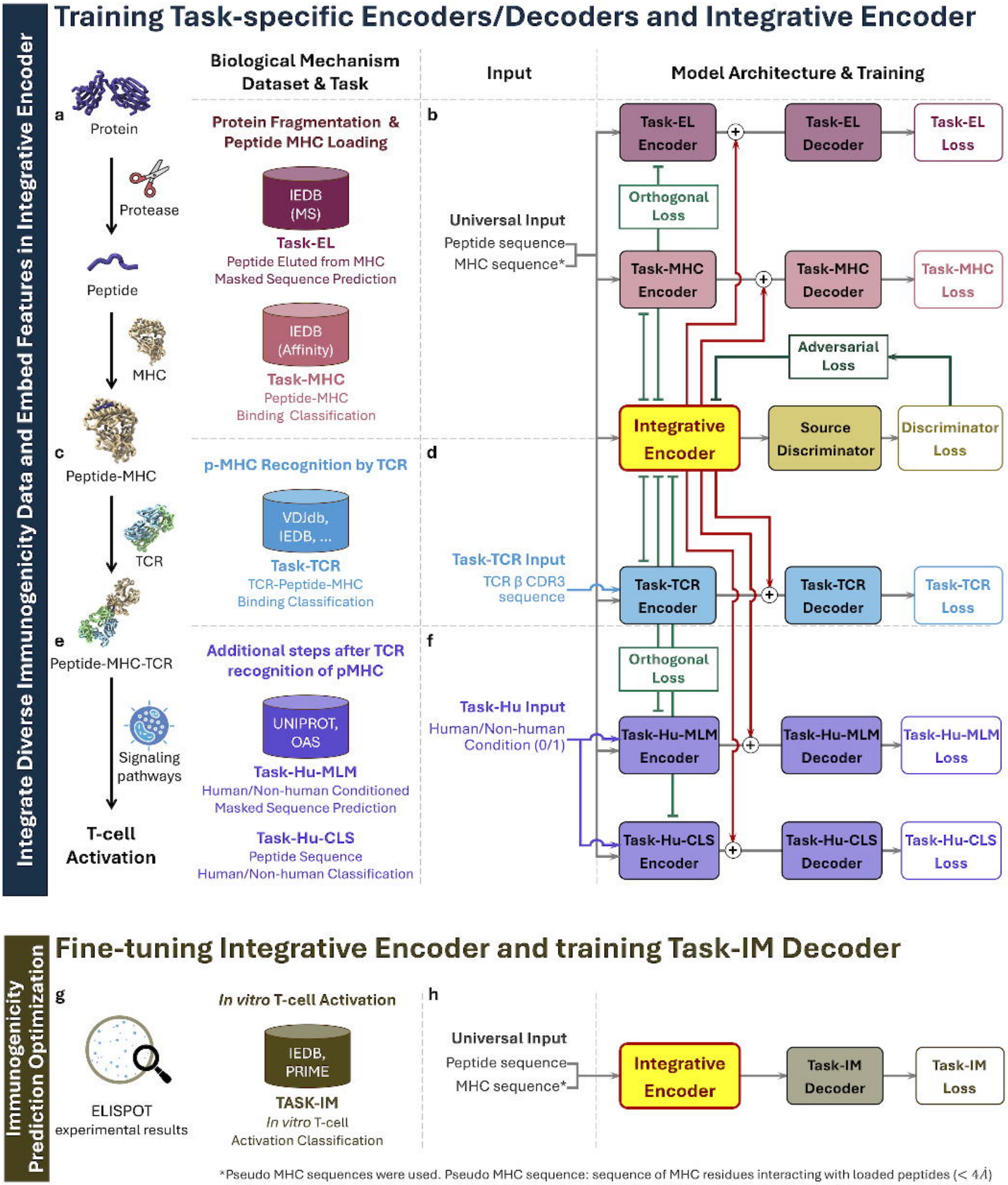
Fig. D5 t-SNE plot of TIANiAN-IM embedding on various datasets. a) A t-SNE plot of our embeddings for the neoantigen discovery set illustrates TITANiAN-IM's effectiveness in distinguishing between immunogenic and non-immunogenic peptides.b) A t-SNE plot of our embeddings for the infectious disease benchmark set illustrates TITANiAN-IM's effectiveness in distinguishing between immunogenic and non-immunogenic peptides. c) A t-SNE plot of our embeddings for the ADA level benchmark set illustrates TITANiAN-IM's effectiveness in distinguishing between high and low ADA level antibody drug sequences.

Table D1 Datasets and tasks used for training TITANiAN

Classification	Data Source	Dataset	Task name	Task explanation
Pretraining Task	IEDB	MHC I eluted ligand MS dataset	Task-EL-I	Peptide eluted from MHC masked sequence prediction
		MHC II eluted ligand MS dataset	Task-EL-II	
		MHC I peptide binding affinity dataset	Task-MHC-I	Peptide-MHC binding classification
		MHC II peptide binding affinity dataset	Task-MHC-II	
	VDJdb+IEDB+McPAS-TCR+ImmuneCODE +TBAdb+10XGenomics	TCR-Peptide-MHC specificity dataset	Task-TCR	TCR-peptide-MHC binding classification
	OAS + UNIPROT	Human nonhman classification dataset	Task-Hu-MLM	Selfness conditioned masked sequence prediction
			Task-Hu-CLS	Peptide Sequence selfness classification
Fine-tuning Task	IEDB + PRIME 1.0 + PRIME 2.0	MHC I ELISPOT dataset	Task-IM-I	ELISPOT result classification task
	IEDB+MHCBN	MHC II ELISPOT dataset	Task-IM-II	

Table D2 The dataset's distribution of positive and negative labels according to each MHC allele type

	MHC Allele Type	Positive Label	Negative Label	Total Number
Fine-tuning Set (NCI)	-	142	996	1138
Fine-tuning Set (PRIME)	A	1199	3540	4739
	B	287	1362	1649
	C	26	261	287
Fine-tuning Set (MHCBN)	DR	2956	2013	4969
Neoantigen Discovery Set	A	92	253	345
	B	77	194	271
	C	9	208	217
Infectious Disease Set	A	924	393	1317
	B	524	180	704
	C	40	32	72


Integrate Diverse Immunogenicity Data and Embed Features in Integrative Encoder
Training Task-specific Encoders/Decoders and Integrative Encoder
Biological Mechanism Dataset & Task
Input
Model Architecture & Training
a
Protein
Protease
Peptide
MHC
c
Peptide-MHC
TCR
e
Peptide-MHC-TCR
Signaling pathways
T-cell Activation
Protein Fragmentation & Peptide MHC Loading
IEDB (MS)
Task-EL
Peptide Eluted from MHC Masked Sequence Prediction
IEDB (Affinity)
Task-MHC
Peptide-MHC Binding Classification
p-MHC Recognition by TCR
VDJdb, IEDB, ...
Task-TCR
TCR-Peptide-MHC Binding Classification
Additional steps after TCR recognition of pMHC
UNIPROT, OAS
Task-Hu-MLM
Human/Non-human Conditioned Masked Sequence Prediction
Task-Hu-CLS
Peptide Sequence Human/Non-human Classification
b
Universal Input
Peptide sequence
MHC sequence*
Task-EL Encoder
Task-EL Decoder
Task-EL Loss
Orthogonal Loss
Task-MHC Encoder
Task-MHC Decoder
Task-MHC Loss
Adversarial Loss
Integrative Encoder
Source Discriminator
Discriminator Loss
d
Task-TCR Input
TCR β CDR3 sequence
Task-TCR Encoder
Task-TCR Decoder
Task-TCR Loss
f
Orthogonal Loss
Task-Hu Input
Human/Non-human Condition (0/1)
Task-Hu-MLM Encoder
Task-Hu-MLM Decoder
Task-Hu-MLM Loss
Task-Hu-CLS Encoder
Task-Hu-CLS Decoder
Task-Hu-CLS Loss
Immunogenicity Prediction Optimization
Fine-tuning Integrative Encoder and training Task-IM Decoder
g
ELISPOT experimental results
In vitro T-cell Activation
IEDB, PRIME
TASK-IM
In vitro T-cell Activation Classification
h
Universal Input
Peptide sequence
MHC sequence*
Integrative Encoder
Task-IM Decoder
Task-IM Loss
*Pseudo MHC sequences were used. Pseudo MHC sequence: sequence of MHC residues interacting with loaded peptides (< 4Å)

Leveraging weak signals may refine immunogenicity prediction

Pretraining vs. Fine-tuning Label Comparison

a

Overall
Cramer's V: 0.0339, p= 2.65e-02

Pre-training label	Not immunogenic	Immunogenic
Positive	1729	2157
Negative	206	203

Task-MHC (class I)

Pre-training label	Not immunogenic	Immunogenic
Binding	876	821
Not Binding	50	116

Task-EL (class I)

Pre-training label	Not immunogenic	Immunogenic
Eluted	30	131
Not Eluted	0	0

Task-MHC (class II)

Pre-training label	Not immunogenic	Immunogenic
Binding	806	1074
Not Binding	156	87

Task-EL (class II)

Pre-training label	Not immunogenic	Immunogenic
Eluted	12	22
Not Eluted	0	0

Task-TCR

Pre-training label	Not immunogenic	Immunogenic
Binding	5	109
Not Binding	0	0

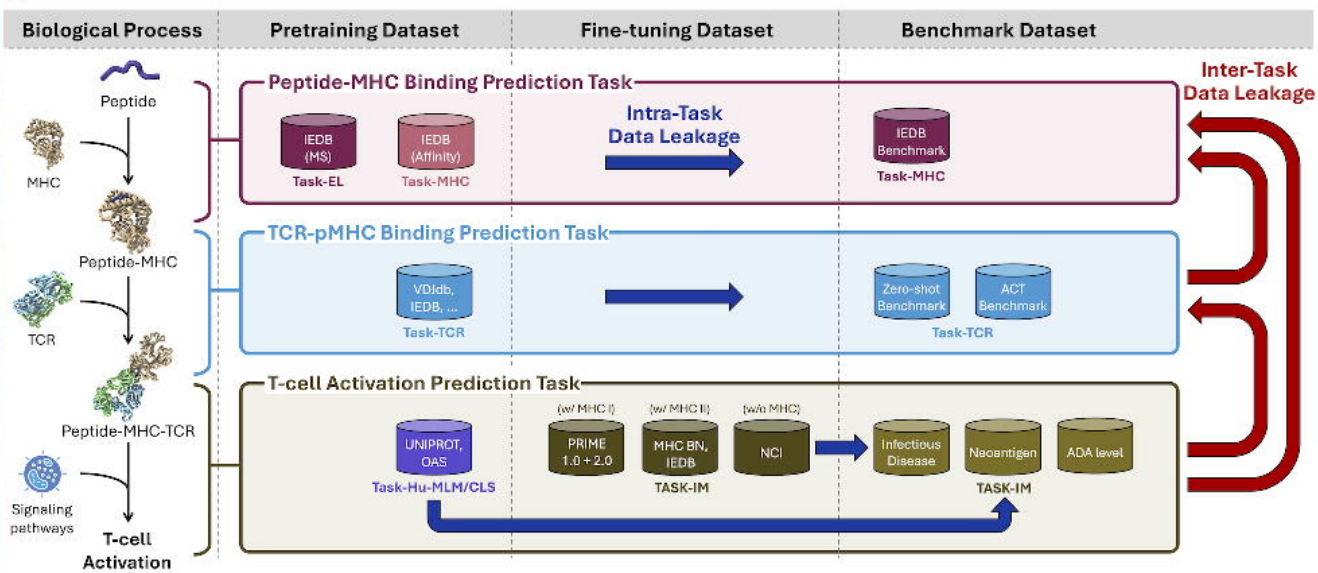
Max
Min

Task-IM (Fine-tuning) label

Inclusion Relationships of Datasets Based on Biological Processes

Information Flow and Potential Data Leakage Analysis

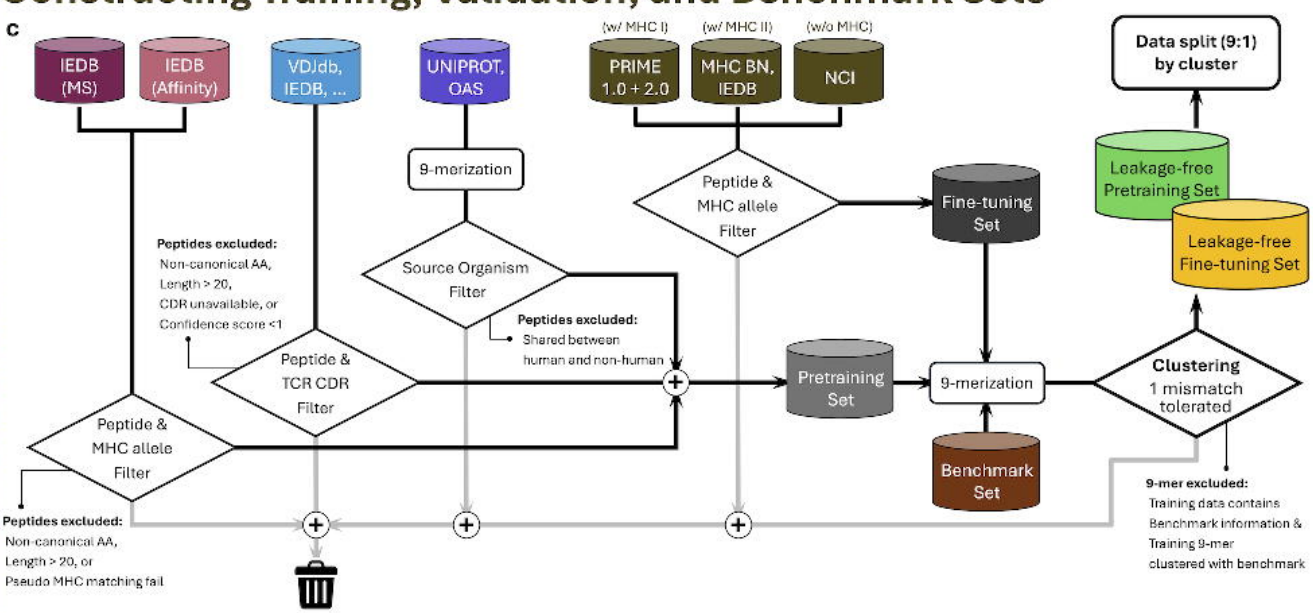
b



Data leakage mitigation through non-overlapping peptide sets across datasets

Constructing Training, Validation, and Benchmark Sets

c



TITANiAN-IM: Accelerating Vaccine Design and Antibody Humanization

Evaluating Immunogenicity Prediction

Task	Input	TITANiAN Model	Output
a ✔ pMHC Immuno-genicity	Peptide seq, MHCseq *	TITANiAN-IM: Integrative Encoder → Task-IM Decoder	Immuno-genicity Score
b ✔ ADA level	Antibody seq → 9-merization → Peptide sequence, Masked MHC sequence*	TITANiAN-IM: Integrative Encoder → Task-IM Decoder → Immunogenicity Score → # of peptides with score > X	ADA Level Score

Model Evaluation			
Test set	**Methods**	**Performance**	**Method Class**
e Neoantigen discovery set[a]	**TITANiAN-IM**	**0.553/0.316**	**Multi-domain prediction**
	BigMHC	0.327/0.307	Two-domain prediction
	MHC nuggets	0.256/0.239	MHC-binding prediction
f Infectious disease vaccine discovery set[b]	**TITANiAN-IM**	**0.927/0.904**	**Multi-domain prediction**
	BigMHC	0.771/0.760	Two-domain prediction
	PRIME-2.0	0.799/0.789	Response prediction
g FDA-approved drug ADA level benchmark set[c]	**TITANiAN-IM**	**0.799/0.847**	**Multi-domain prediction**
	Hu-mAb	0.201/0.838	Selfness Prediction
	AbNatiV	0.095/0.852	Selfness Prediction

*Pseudo MHC sequence interacting (≤4Å) with the peptide
a, b, & c) Metrics chosen to reflect real-world application
a) Precision/PR-AUC
b) Mean PPVn /PR-AUC
c) Recall/PR-AUC

TITANiAN-MHC/TCR: Prioritizing Peptide-MHC Validation

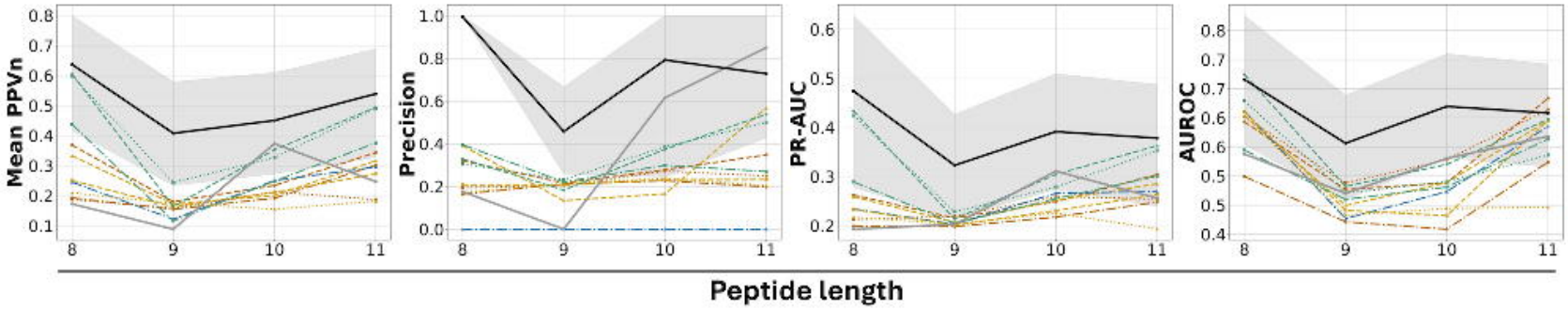
Evaluating MHC/TCR Binding Prediction

Task	Input	TITANiAN Model	Output
c ✔ Peptide -MHC binding	Peptide seq, MHC seq*	TITANiAN-MHC: Integrative Encoder, Task-MHC Encoder → Task-MHC Decoder → Binding Probability → Probability > 0.5	Binding or non-binding
d ✔ TCR -peptide -MHC binding	Peptide seq, MHC seq*, TCRβ CDR3 seq	TITANiAN-TCR: Integrative Encoder, Task-TCR Encoder → Task-TCR Decoder → Binding Probability → Probability > 0.5	Binding or non-binding

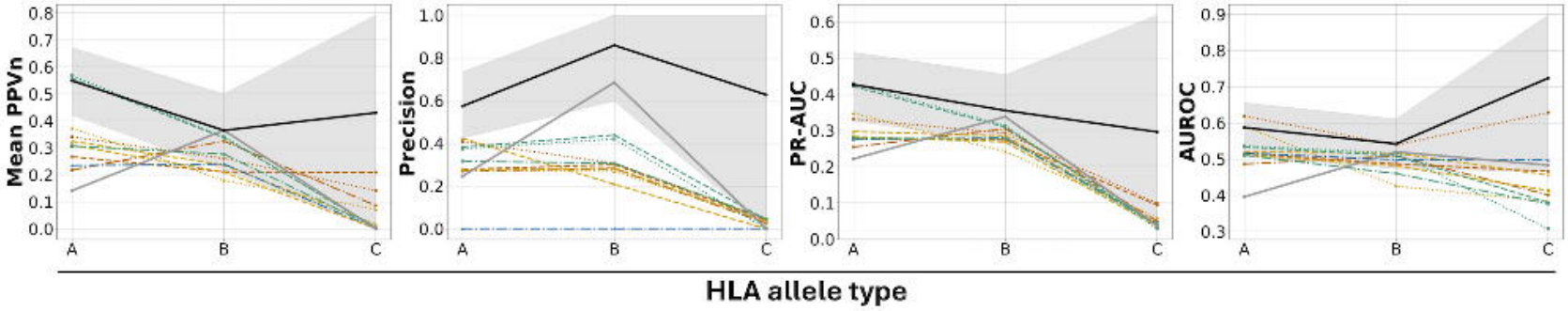
Model Evaluation			
h IEDB MHC class I benchmark set[b]	**TITANiAN-MHC**	**0.770/0.725**	**Multi-domain prediction**
	NetMHCPan	0.784/0.781	MHC-binding prediction
	NetMHCcons	0.804/0.813	MHC-binding prediction
i IEDB MHC class II benchmark set[b]	**TITANiAN-MHC**	**0.613/0.521**	**Multi-domain prediction**
	NetMHCIIPan	0.766/0.746	MHC-binding prediction
	NNalign	0.598/0.781	MHC-binding prediction
j ACT-based therapy discovery set[b]	**TITANiAN-TCR**	**0.800/0.745**	**Multi-domain prediction**
	Pan-Pep	0.883/0.819	TCR-binding prediction
	DLpTCR	0.621/0.595	TCR-binding prediction

Performance per peptide length and HLA type in neoantigen discovery set

a Performance per peptide length

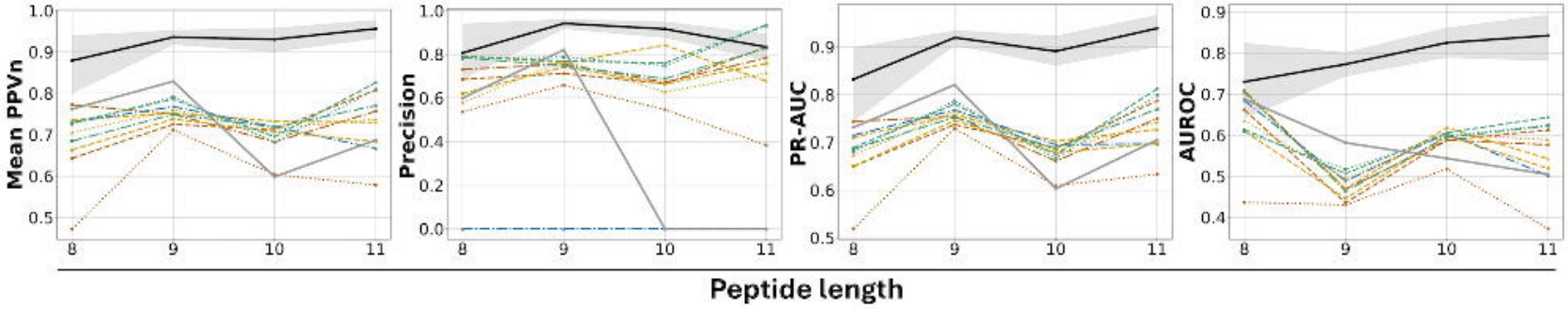


b Performance per HLA allele type

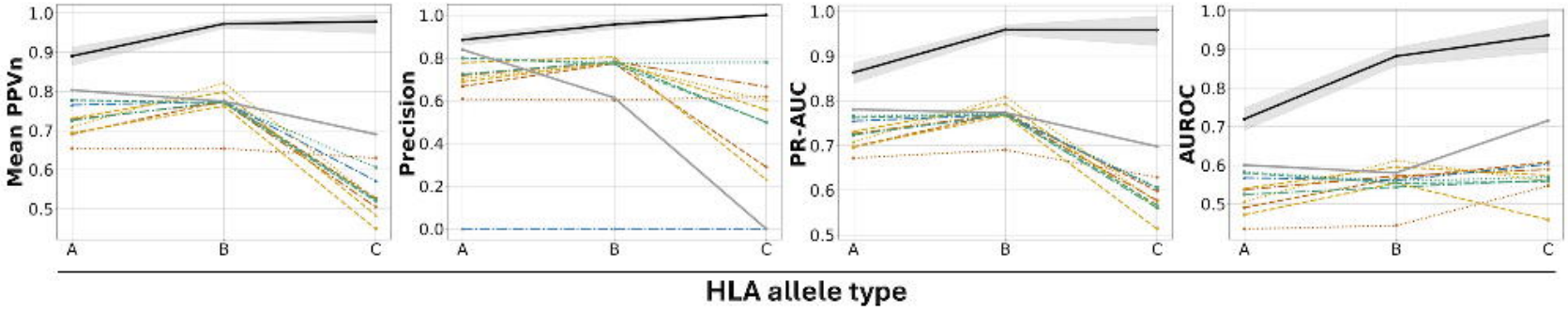


Performance per peptide length and HLA type in vaccine discovery set

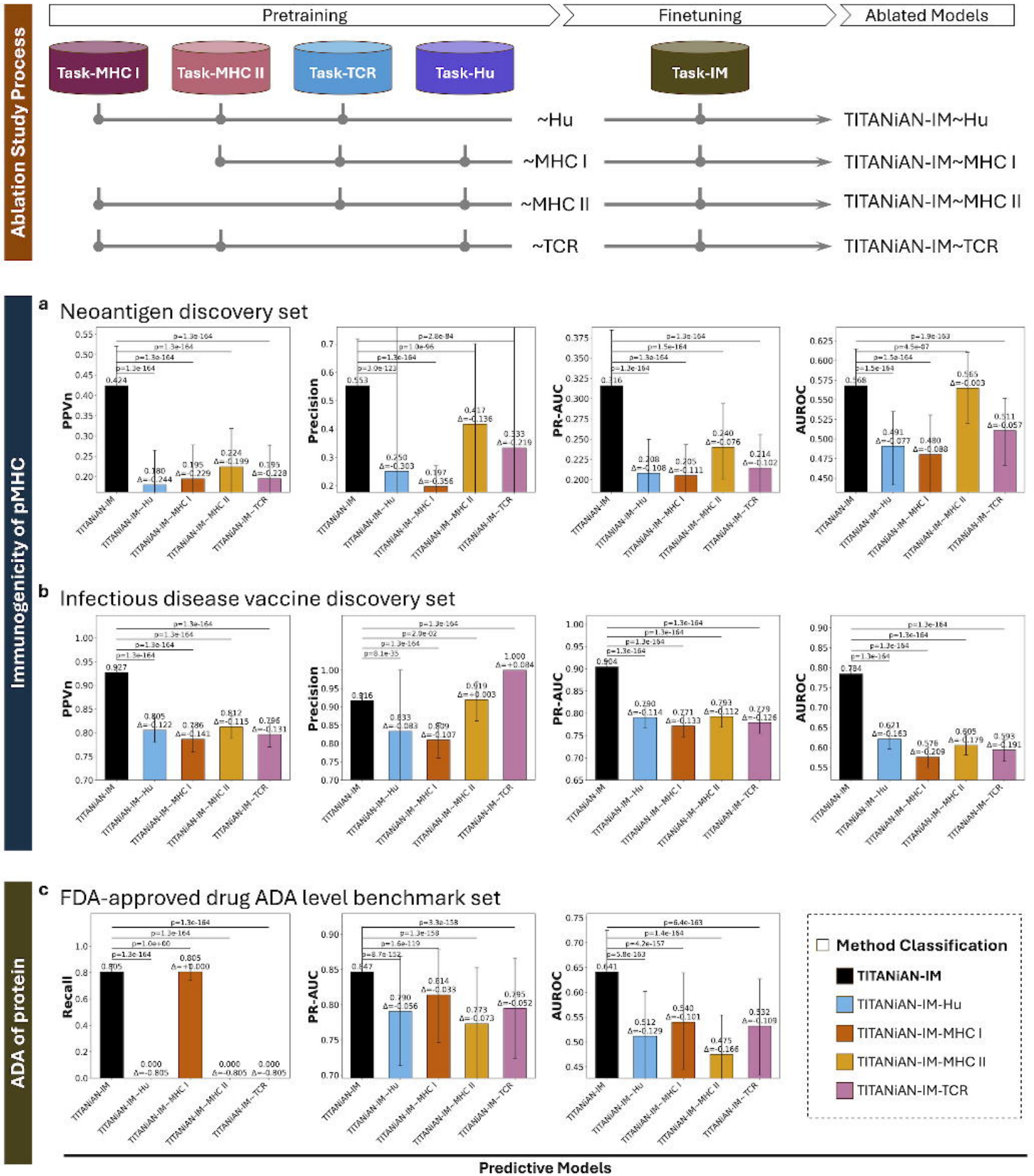
c Performance per peptide length



d Performance per HLA allele type



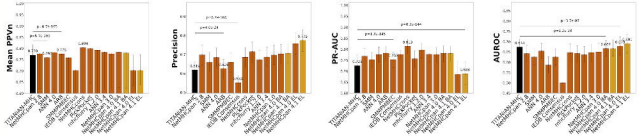
Ablation Study on Immunogenic Response Prediction (TITANiAN-IM)



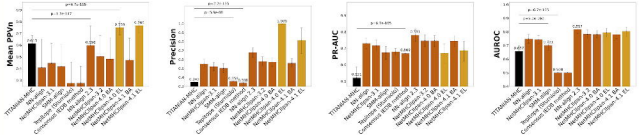
Molecular Binding Prediction Benchmark (TITANiAN-MHC & TCR)

Peptide-MHC binding (TITNIAN-MHC)

a IEDB MHC class I benchmark set

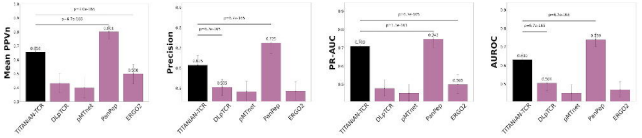


b IEDB MHC class II benchmark set

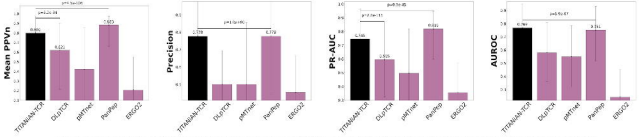


TCR-peptide-MHC binding (TITNIAN-TCR)

c Pan-Pep zero shot benchmark set

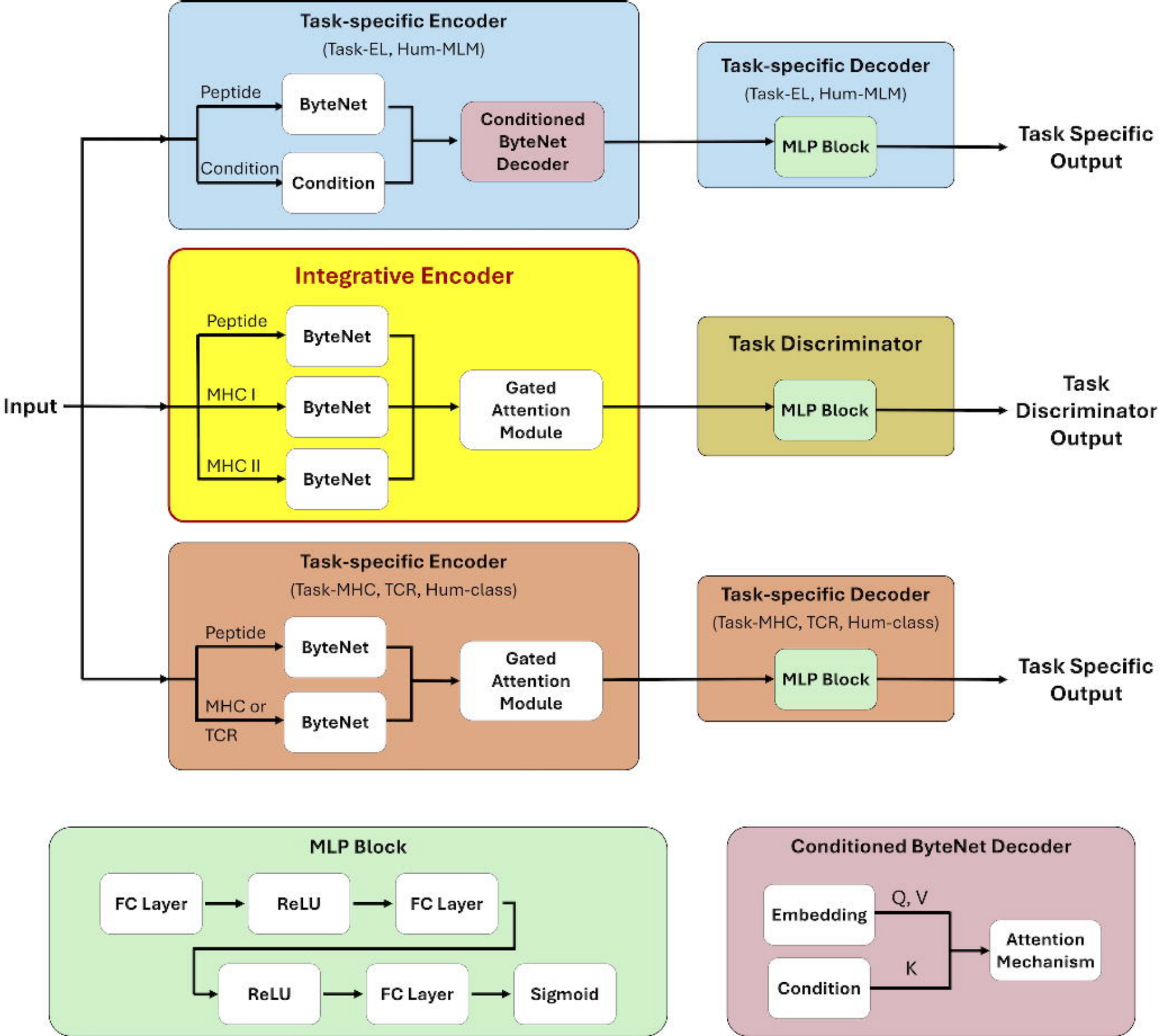


d ACT-based therapy discovery set

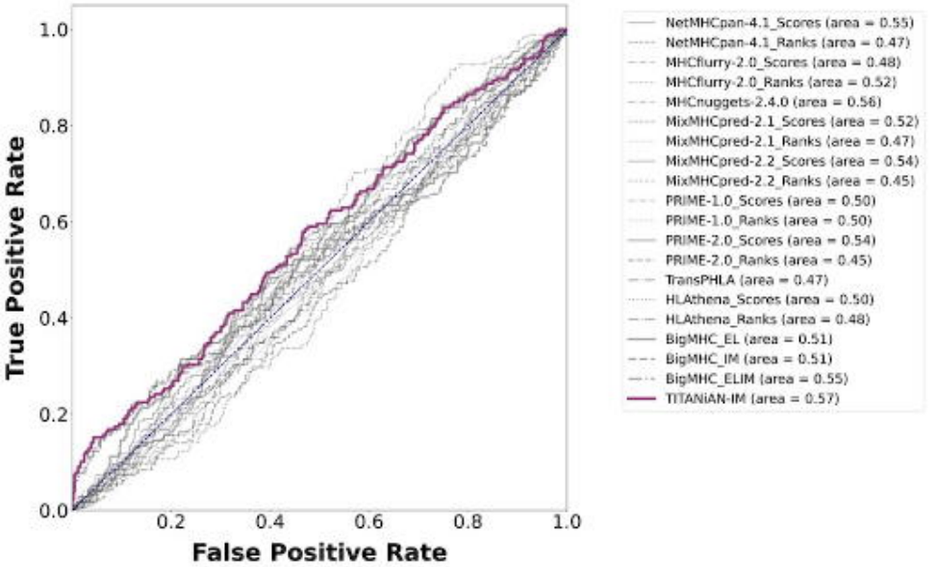


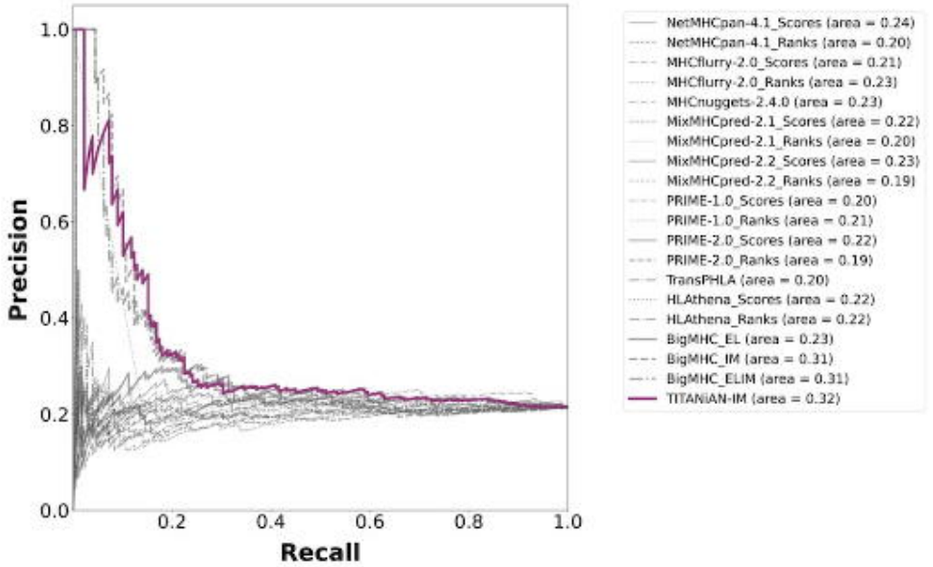
Method Classification
TITANiAN-MHC or TCR | MHC loading prediction | MHC binding prediction | TCR binding prediction

Predictive Models

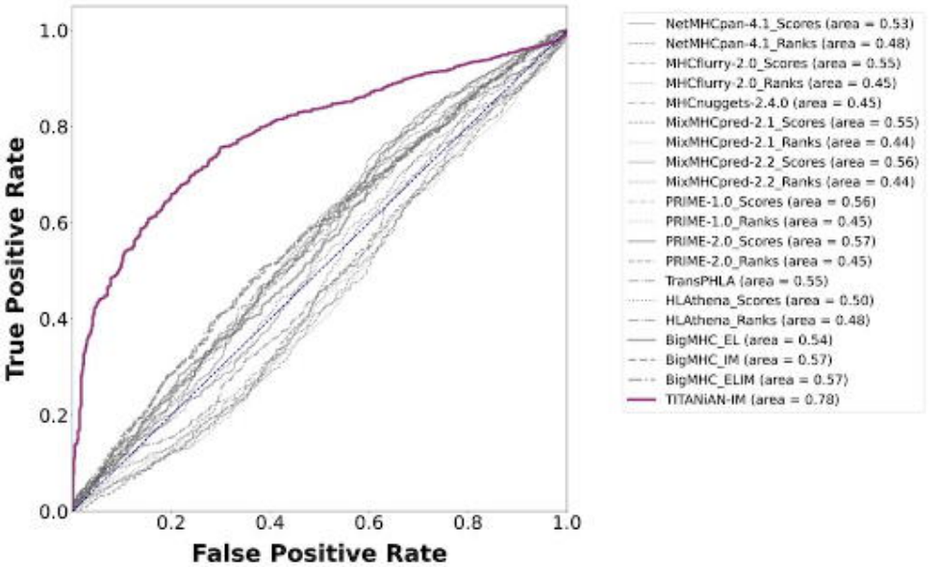

Task-specific Encoder
(Task-EL, Hum-MLM)
Peptide
ByteNet
Condition
Condition
Conditioned ByteNet Decoder
Task-specific Decoder
(Task-EL, Hum-MLM)
MLP Block
Task Specific Output
Integrative Encoder
Peptide
ByteNet
MHC I
ByteNet
MHC II
ByteNet
Gated Attention Module
Task Discriminator
MLP Block
Task Discriminator Output
Input
Task-specific Encoder
(Task-MHC, TCR, Hum-class)
Peptide
ByteNet
MHC or TCR
ByteNet
Gated Attention Module
Task-specific Decoder
(Task-MHC, TCR, Hum-class)
MLP Block
Task Specific Output
MLP Block
FC Layer
ReLU
FC Layer
ReLU
FC Layer
Sigmoid
Conditioned ByteNet Decoder
Embedding
Q, V
Condition
K
Attention Mechanism

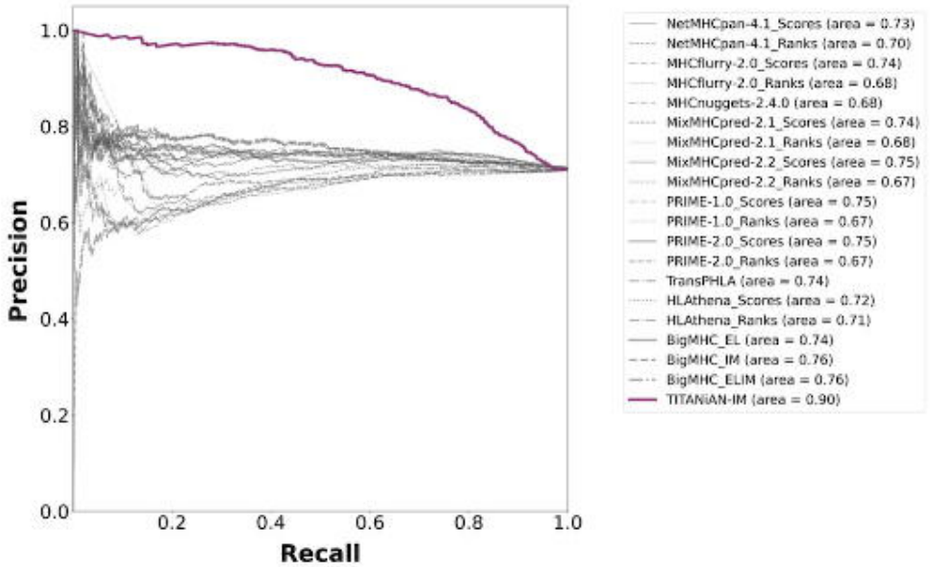
TITANiAN-IM Benchmark result for Neoantigen discovery set



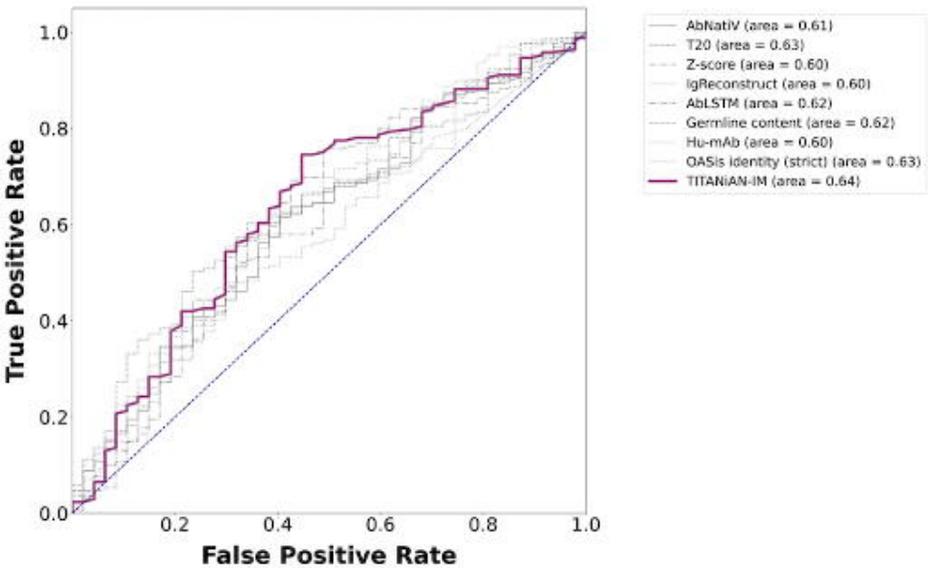


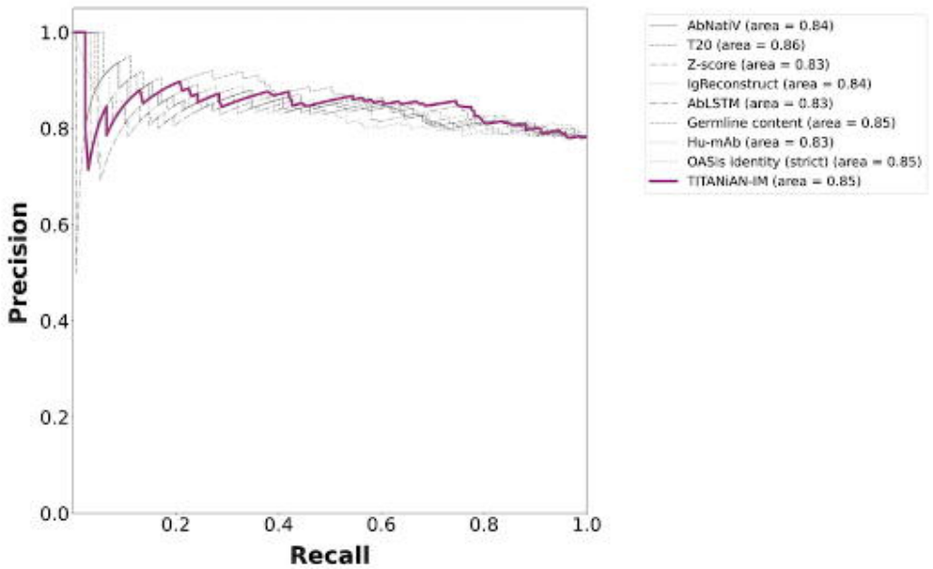
TITANiAN-IM Benchmark result for Infectious disease set



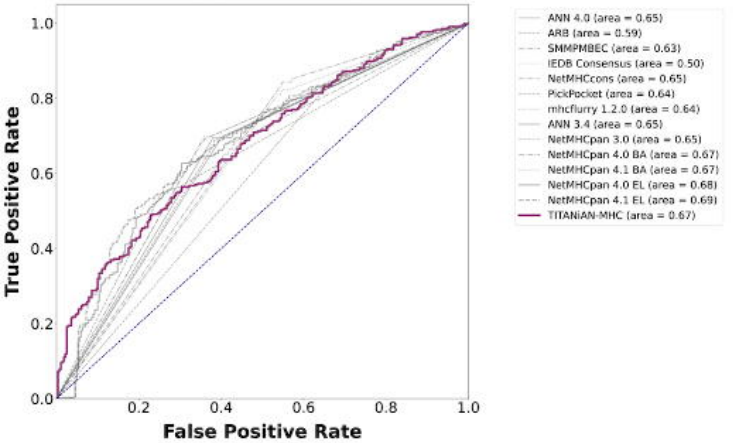


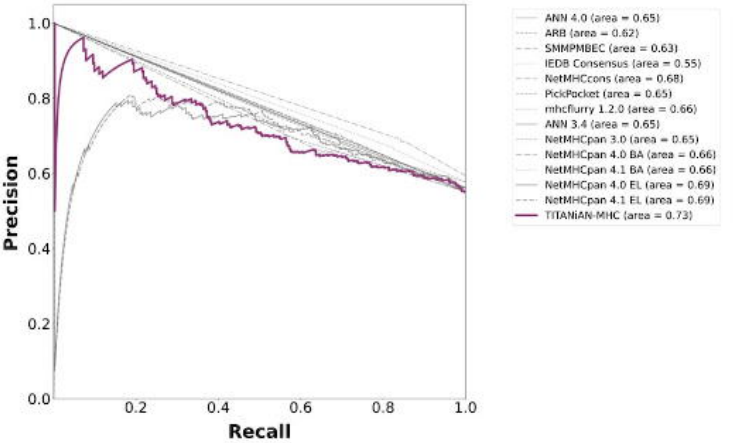
TITANiAN-IM Benchmark result for ADA level set



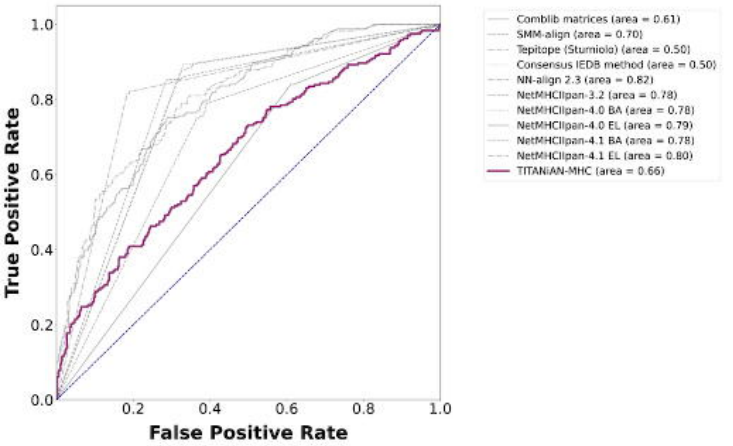


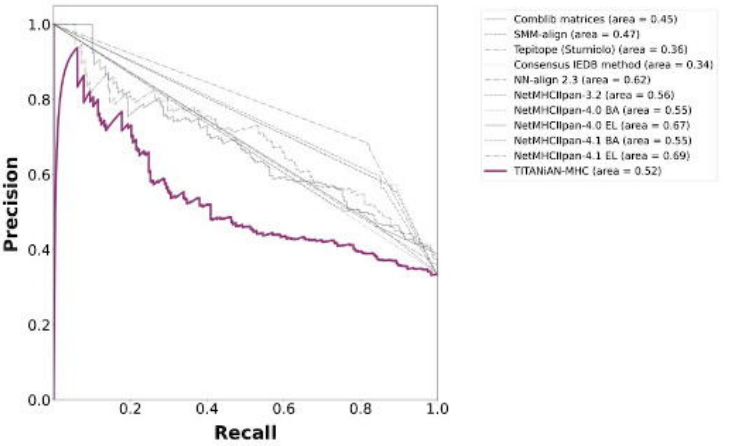
TITANiAN-MHC I Benchmark result for IEDB benchmark set



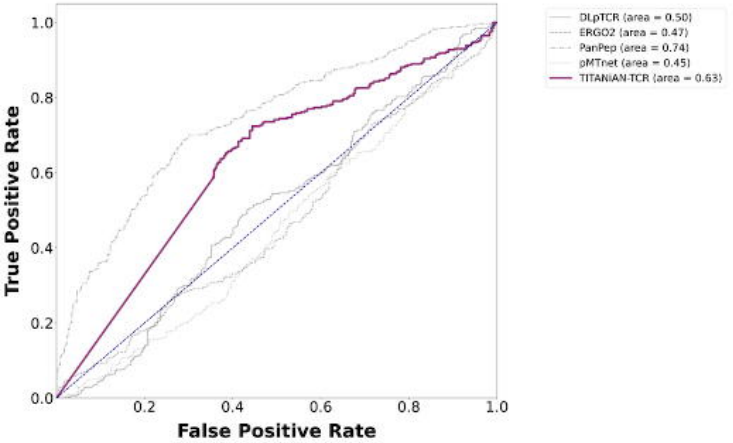


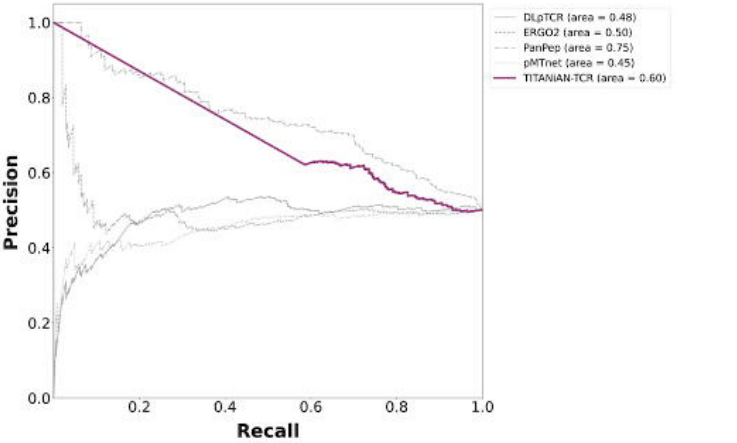
TITANiAN-MHC II Benchmark result for IEDB benchmark set



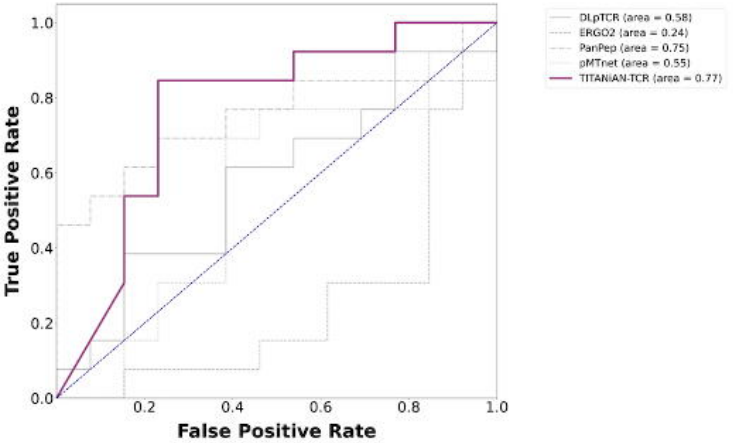


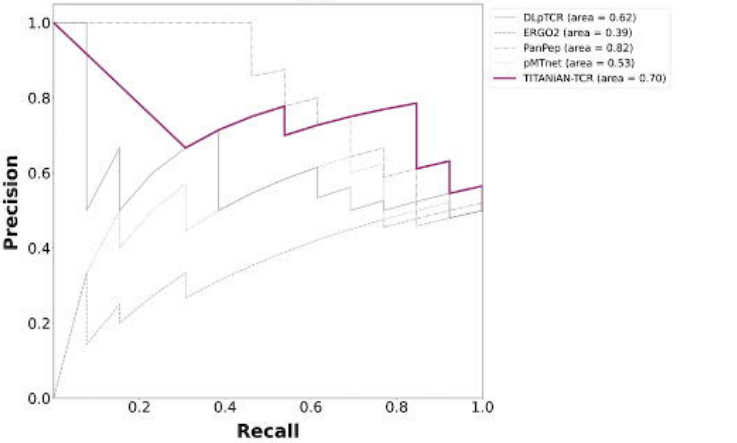
TITANiAN-TCR Benchmark result for Zero shot set



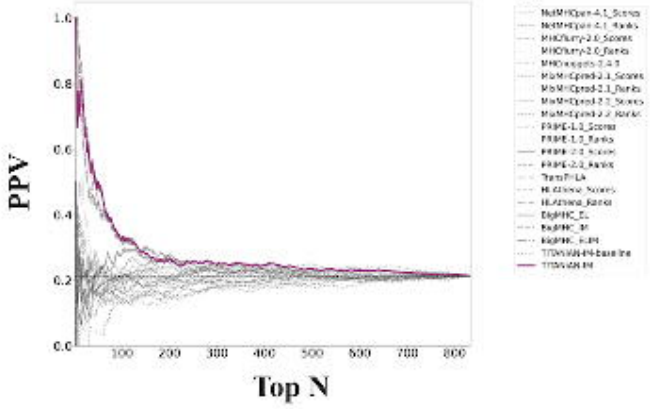


TITANiAN-TCR Benchmark result for ACT-based therapy discovery set

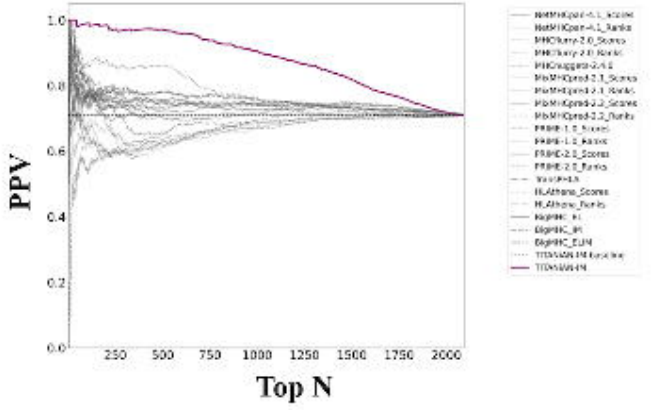




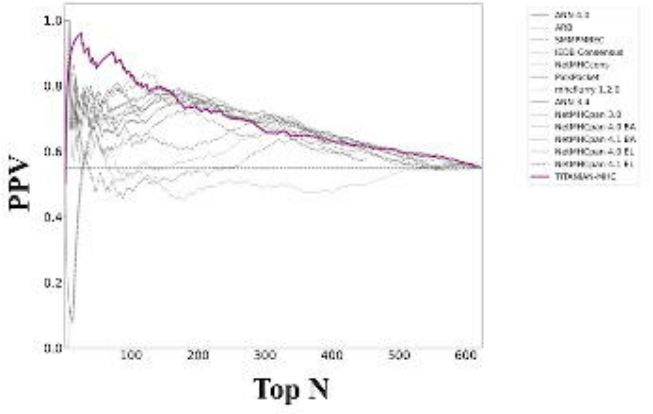
TITANiAN-IM Benchmark result for Neoantigen discovery set



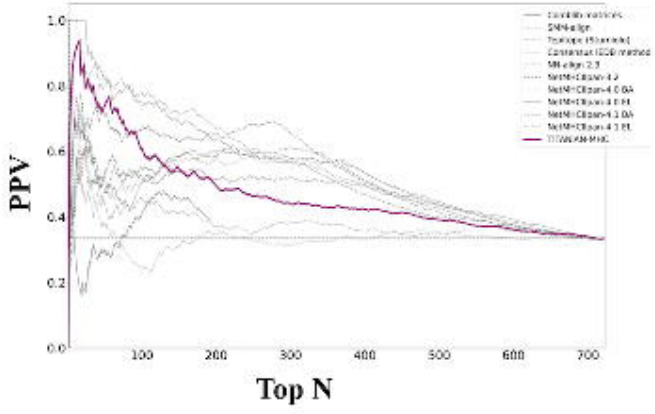
TITANiAN-IM Benchmark result for Infectious disease set



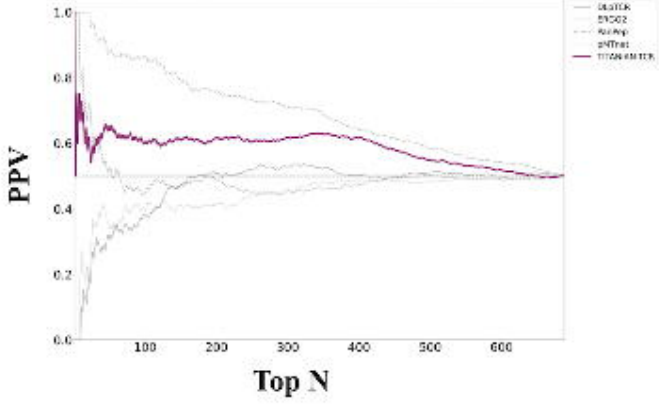
TITANiAN-MHC I Benchmark result for IEDB benchmark set



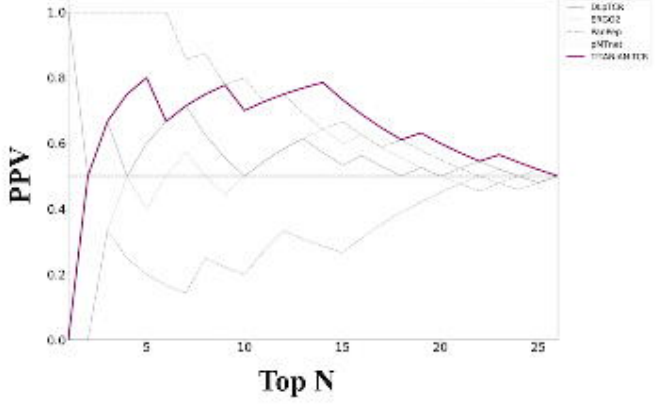
TITANiAN-MHC II Benchmark result for IEDB benchmark set

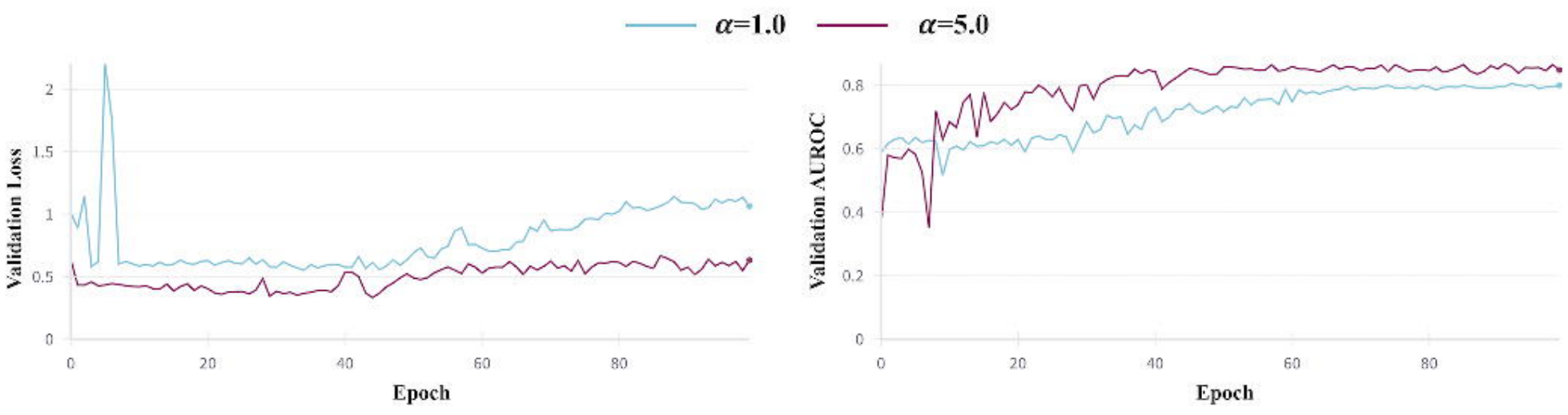


TITANiAN-TCR Benchmark result for Zero shot set

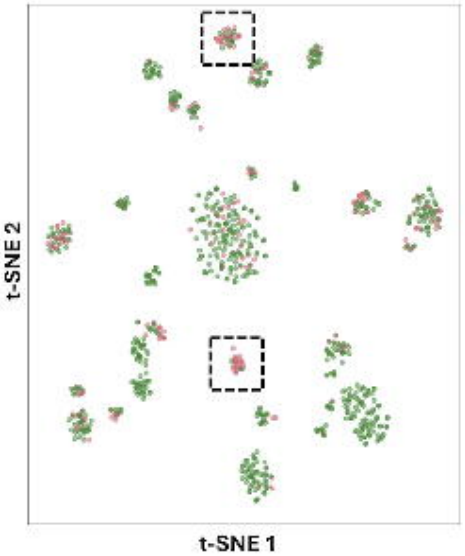


TITANiAN-TCR Benchmark result for ACT-based therapy discovery set



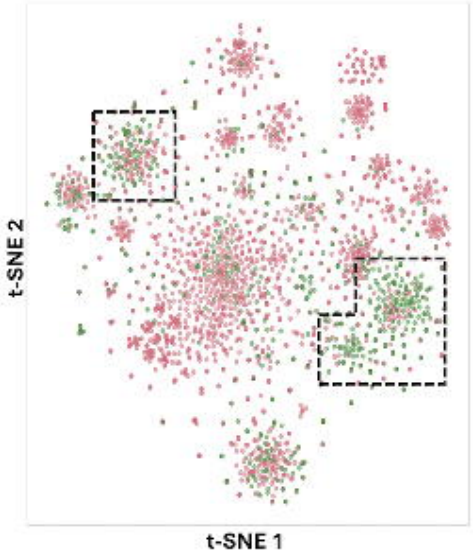

α=1.0
α=5.0
Validation Loss
Epoch
Validation AUROC
Epoch

a) t-SNE plot of TITANiAN IM embedding on neoantigen discovery set



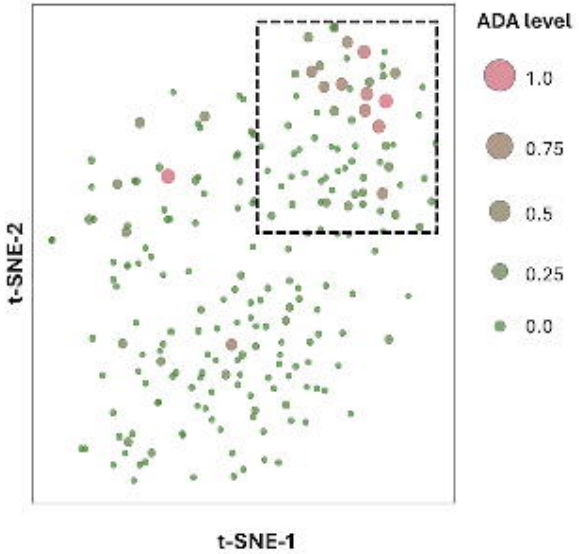
Silhouette score: 0.132
AMI: 0.017

b) t-SNE plot of TITANiAN IM embedding on infectious disease vaccine discovery set



Silhouette score: 0.203
AMI: 0.045

c) t-SNE plot of TITANiAN IM embedding on FDA approved drug ADA level benchmark set



Silhouette score: 0.057
AMI: 0.045

Non-immunogenic Immunogenic

Non-immunogenic Immunogenic

Immunogenic Response Prediction Benchmark (TITANiAN-IM)

Immunogenicity of pMHC

Peptide seq
MHC seq*
→ TITNiAN-IM → Immunogenicity Score

* Pseudo MHC sequence that has structural interactions with the loaded peptide within 4 Å

(w/ MHC I) PRIME 1.0 + 2.0
(w/ MHC II) MHC BN, IEDB
(w/o MHC) NCI
UNIPROT, OAS

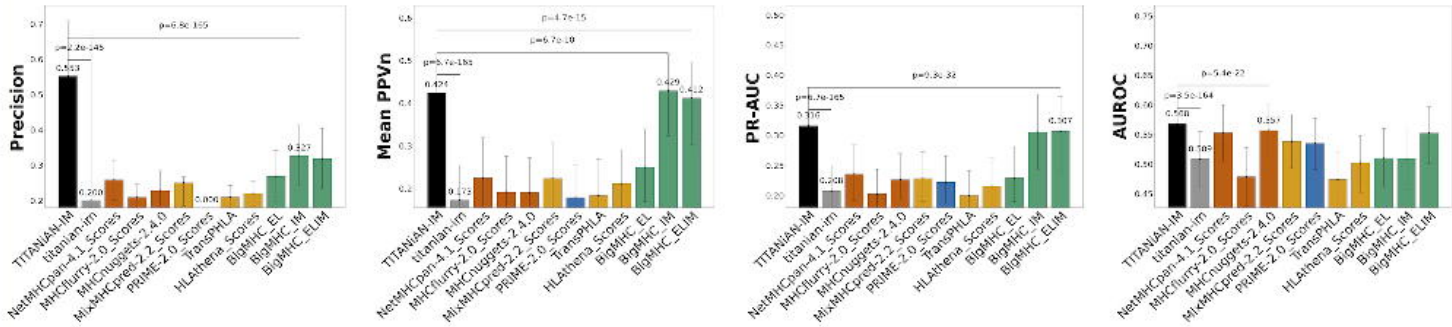
Datasets exhibiting potential for information leakage

← Stringent Removal of Training 9-mers
(Less than 2 mismatches to Benchmark 9-mers)

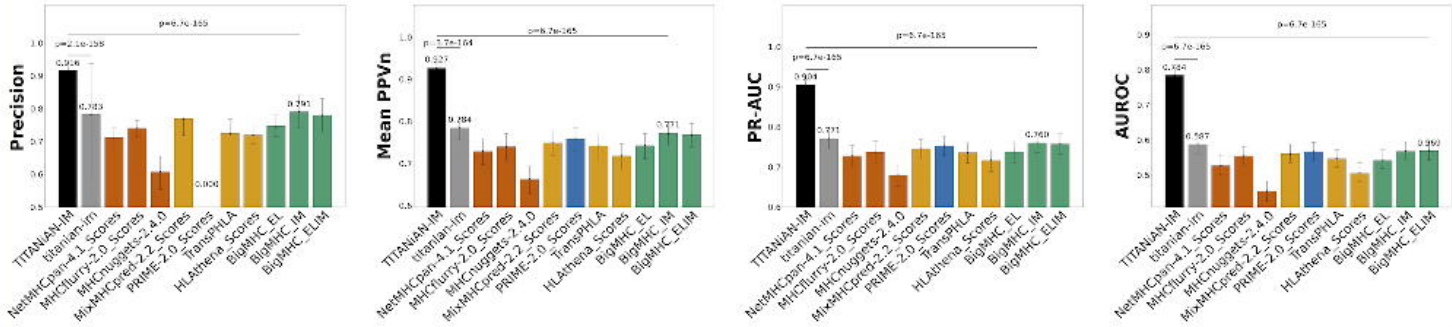
Neoantigen
Infectious Disease

Benchmark set

a Neoantigen discovery set



b Infectious disease vaccine discovery set



Predictive Models

ADA of protein

Antibody seq → 9-merization → Peptide sequence
Masked MHC sequence+
→ TITNiAN-IM → Immunogenicity Score → # of peptides with score > X → ADA Level Score

+ MHC sequence masked as X

(w/ MHC I) PRIME 1.0 + 2.0
(w/ MHC II) MHC BN, IEDB
(w/o MHC) NCI
UNIPROT, OAS

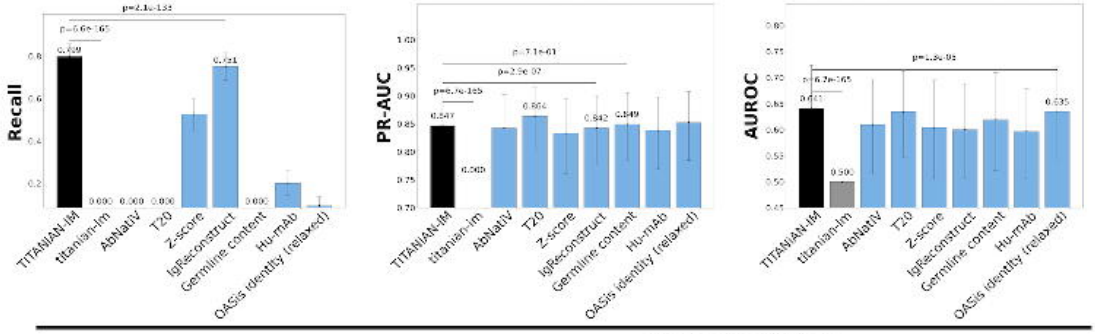
Datasets exhibiting potential for information leakage

← Stringent Removal of Training 9-mers
(Less than 2 mismatches to Benchmark 9-mers)

ADA level

Benchmark set

c FDA-approved drug ADA level benchmark set



Predictive Models

Method Classification

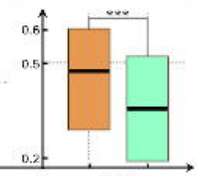
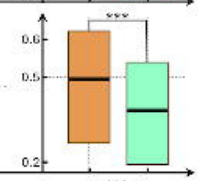
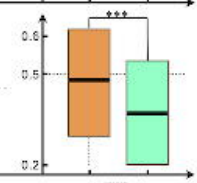
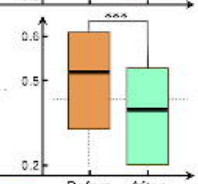
- TITANiAN-IM
- titanian-im
- Two-domain prediction
- Response prediction
- Selfness prediction
- TCR binding prediction
- MHC binding prediction
- MHC loading prediction

Impact of point mutations and HLA allele types on immunogenic responses

	Source Protein (Organism)	HLA allele	Peptide sequence	Immunogenicity	TITANiAN-IM Immunogenicity score
Same HLA, different peptide					
a	Spike glycoprotein (SARS-CoV2)	A*02:01	F V F **L** V L L P L V (2 3 4 5 6 7 8 9 10 11)	Immunogenic	0.9669
			F V F **F** V L L P L V (2 3 4 5 6 7 8 9 10 11)	Non-immunogenic	2.95e-05
b	Large delta antigen (Hepatitis D virus)	B*27:02	R R **K** **K** A L E N K K (104 105 106 107 108 109 110 111 112 113)	Immunogenic	0.9999
			R R **R** **M** A L E N K K (104 105 106 107 108 109 110 111 112 113)	Non-immunogenic	6.02e-05
c	Protein Nef (HIV-1)	C*07:01	**Q** R Q **D** I L D L W V Y (105 106 107 108 109 110 111 112 113 114 115)	Immunogenic	0.9999
			K R Q **E** I L D L W V Y (105 106 107 108 109 110 111 112 113 114 115)	Non-immunogenic	3.77e-07
Same peptide, different HLA					
d	Nucleoprotein (SARS-CoV2)	B*35:01	S P D D Q I G Y Y (79 80 81 82 83 84 85 86 87)	Immunogenic	0.9175
		A*01:01	S P D D Q I G Y Y (79 80 81 82 83 84 85 86 87)	Non-immunogenic	3.31e-05
e	Spike glycoprotein (SARS-CoV2)	B*57:01	G T I T S G W T F (880 881 882 883 884 885 886 887 888)	Immunogenic	0.9993
		A*24:02	G T I T S G W T F (880 881 882 883 884 885 886 887 888)	Non-immunogenic	1.40e-04
f	Nucleoprotein (Mammarenavrius lassaense)	A*34:02	M V F R T S T P R V (558 559 560 561 562 563 564 565 566 567)	Immunogenic	0.9999
		A*30:01	M V F R T S T P R V (558 559 560 561 562 563 564 565 566 567)	Non-immunogenic	2.86e-05

Mutated

Effect of deimmunization of antibody on ADA levels

	Antibody drug	Chain	Antibody sequence before deimmunization	Recovery of DAA	Score distribution
g	Etaracizumab	VH	EVQLEESGGGLVKPGGSLKLSCAASGFAFSSYDMSVWRQIPEKRLWAKVSSGGGSTYYLDT VQGRFTISRDNAKNTLYLQMSSLNSEDTAMYYCARNYGSFAYWGQGTLVTVSA	14 / 20	
		VL	ELVMTQTPATLSVTQPGDSVSLSCRASQSISNHWLYQQKSHESPRLLIKYASQSISGIPSR FSGSGSGTDFTLSINSVETEDFGMYFCQQSNSWPHTFGGGTKLEIK	22 / 29	
h	Bevacizumab	VH	EIQLVQSGPELKQPGETVRISCKASGYTFTNYGMNWVKQAPGKGLKWMGWINTYTGEPTYAAD FKRRFTFSLETSASTAYLQISNLKNDDTATYFCAKYPHYYGSSHWYFDVWGAGTTVTVSS	21 / 25	
		VL	DIQMTQTTSSLSASLGDRVIISCSASQDISNYLNWYQQKPDGTVKVLIYFTSSLHSGVPSR FSGSGSGTDYSLTISNLEPEDIATYYCQQYSTVPWTFGGGTKLEIK	16 / 16	
i	Pembrolizumab	VH	QVQLQQPGAELVKPGTSVKLSCKASGYTFTNYYMYWVKQRPGQGLEWIGGINPSNGGTNFNE KFKNKATLTVDSSSSTTYMQLSSLTSEDSAVYYCTRRDYRFDMGFDYWGQGTTLTVSS	14 / 20	
		VL	DIVLTQSPASLAVSLGQRAAISCRASKGVSTSGYSYLHWYQQKPGQSPKLLIYLASYLESGV PARFSGSGSGTDFTLNIHPVEEEDAATYYCQHSRDLPLTFGTGTKLELK	19 / 20	
j	Omalizumab	VH	DVQLQESGPGLVKPSQSLSLACSVTGYSITSGYSWNWIRQFPGNKLEWMGSITYDGSSNYNP SLKNRISVTRDTSQNQFFLKLNSATAEDTATYYCARGSHYFGHWHFAVWGAGTTVTVSS	25 / 35	
		VL	DIQLTQSPASLAVSLGQRATISCKASQSVDYDGDSYMNWYQQKPGQPPILLIYAASYLGSEI PARFSGSGSGTDFTLNIHPVEEEDAATFYCQQSHEDPYTFGAGTKLEIK	25 / 25	

Before After Deimmunization

Deimmunized amino acid(DAA) TITANiAN immunogenic TITANiAN immunogenic DAA