frontiers | Frontiers in Immunology

TYPE Review
PUBLISHED 07 September 2023
DOI 10.3389/fimmu.2023.1228873

OPEN ACCESS

EDITED BY
Dinler Amaral Antunes,
University of Houston, United States

REVIEWED BY
Victor Greiff,
University of Oslo, Norway
Antonella Prisco,
National Research Council (CNR), Italy
Prashant Dogra,
Houston Methodist Research Institute, United States
Martiela Vaz De Freitas,
University of Houston, United States

*CORRESPONDENCE
Jason T. George
jason.george@tamu.edu

RECEIVED 25 May 2023
ACCEPTED 17 August 2023
PUBLISHED 07 September 2023

CITATION
Ghoreyshi ZS and George JT (2023) Quantitative approaches for decoding the specificity of the human T cell repertoire. *Front. Immunol.* 14:1228873. doi: 10.3389/fimmu.2023.1228873

Quantitative approaches for decoding the specificity of the human T cell repertoire

Zahra S. Ghoreyshi[1] and Jason T. George[1,2,3]*

[1]Department of Biomedical Engineering, Texas A&M University, College Station, TX, United States, [2]Engineering Medicine Program, Texas A&M University, Houston, TX, United States, [3]Center for Theoretical Biological Physics, Rice University, Houston, TX, United States

T cell receptor (TCR)-peptide-major histocompatibility complex (pMHC) interactions play a vital role in initiating immune responses against pathogens, and the specificity of TCRpMHC interactions is crucial for developing optimized therapeutic strategies. The advent of high-throughput immunological and structural evaluation of TCR and pMHC has provided an abundance of data for computational approaches that aim to predict favorable TCR-pMHC interactions. Current models are constructed using information on protein sequence, structures, or a combination of both, and utilize a variety of statistical learning-based approaches for identifying the rules governing specificity. This review examines the current theoretical, computational, and deep learning approaches for identifying TCR-pMHC recognition pairs, placing emphasis on each method's mathematical approach, predictive performance, and limitations.

KEYWORDS

TCR, pMHC, binding prediction, protein-protein interaction, machine learning, deep learning

1 Introduction

The adaptive immune system has the remarkable responsibility of recognizing and eliminating foreign threats, which requires discriminating self from non-self-signatures. T lymphocytes, or T cells, are the cellular mediators of adaptive immunity and accomplish this feat by using their heterodimeric T cell receptors (TCRs). TCRs recognize short peptides bound to and presented by class I and II major histocompatibility complex (MHC) molecules on the cell surface (pMHC). TCR diversity is generated by genetic rearrangement through a V(D)J recombination process (1) capable of generating a staggering diversity of TCRs (estimates range from $\sim 10^{20}$ to $\sim 10^{61}$ possible receptors) (2). It is this total diversity together with the relative sparsity of realized samples that complicates the development of inferential modeling procedures capable of predicting TCR-pMHC specificity when test systems differ moderately from training samples (3, 4). Solving this problem would have numerous immunological implications that range from identifying improved antigen vaccines to facilitating optimal selection of adoptive T cell therapy for cancer patients (5, 6).

T cell responses occur when their TCRs bind pMHC with an interaction that 'appears' to the T cell as 'non-self'. In order to avoid detection of abundant self-signatures, T cell

precursors (thymocytes) undergo central tolerance to a set of self-signatures via a process called thymic negative selection (7), wherein each thymocyte is exposed to a diverse set of self-antigens, and TCR recognition of any of these self-antigents results in deletion. In addition to central tolerance, a variety of peripheral tolerance mechanisms exist to prevent self-recognition, including T cell anergy, suppression by regulatory T cells (Tregs), and tolerance induction through peripheral antigen exposure (7, 8). Collectively, these mechanisms ensure that mature T cells are selectively responsive to non-self antigens while maintaining a state of immunological self-tolerance.

Owing to the sheer complexity of the adaptive immune response, a number of theoretical and computational models have been explored focusing on various aspects of the problem. These efforts have benefited from the availability of advanced structural characterization techniques, such as X-ray crystallography (9), NMR spectroscopy (10), and cryoelectron microscopy (11), for validation. Moreover, recent advances in high-throughput approaches (12, 13) have significantly increased the available data on which inferential learning-based models can be constructed. Consequently, a number of computational models have been developed to address the need for reliable TCR specificity prediction between a collection of known TCR sequences and putative antigen targets.

In this review, we outline the recent theoretical and computational approaches to TCR-pMHC specificity prediction, emphasizing their strengths and limitations, and offer perspective on the future direction of this exciting modeling effort. In our description of the informatics-based strategies for TCR-pMHC prediction, we discuss current methodology and challenges in four main areas: modeling of TCR-pMHC complex interactions based on (1) sequence-based approaches (Figure 1A) (2) structure-based approaches (Figure 1B), (3) deep learning approaches, and (4) hybrid approaches (Figure 1C).

As we delve deeper into the complex matrix of TCR-pMHC interactions, it becomes essential to illuminate the interplay between specificity and cross-reactivity, two critical factors that significantly shape the predictive modeling landscape. Recognizing the delicate balance between these parameters not only enriches our understanding of various modeling strategies but also refines our approach to interpreting the multi-faceted nature of TCR-pMHC interactions. Cross-reactivity, a fundamental characteristic of TCRs, is an essential consideration in predicting TCR-pMHC interactions. This attribute of TCRs enables them to interact with a myriad of peptide antigens, providing our immune system with its remarkable breadth of response. However, this same characteristic poses a significant challenge in understanding T cell-based therapeutics and similarly in TCR-pMHC specificity prediction. Single TCRs are intrinsically capable of binding to multiple peptide antigens at once, complicating predictions of specificity and off-target effects (14, 15). Traditional prediction methods, which mostly rely on peptide sequence similarities and biochemical similarity, often fail to capture cross-reactivity's complex nuances. As such, more comprehensive approaches have emerged. For instance, tools like

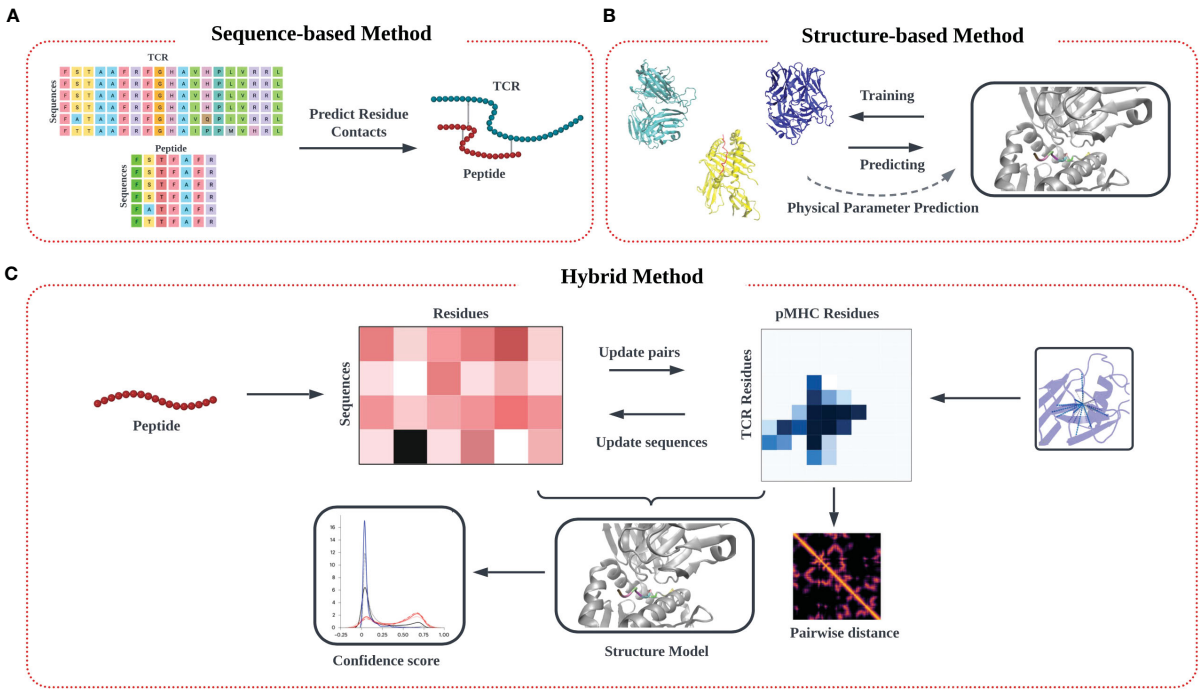

FIGURE 1
Modeling approach to TCR-pMHC prediction based on input data. **(A)** Models trained purely on TCR and peptide sequence input data feature multiple sequence alignment (MSA) on input data matrices to identify patterns, followed by identification of potential interaction pairs using various algorithms techniques. **(B)** Models trained on input structural data models commonly aim to identify the TCR-pMHC binding interface, along with associated information on binding affinity. Predictions are often made by determining similarity in secondary structure in the interfacial region of the binding interface, from which physical parameters like binding affinity and kinetic data (K_D, K_{off}) can be estimated. **(C)** A third 'hybrid' category of inferential model synergistically combines sequence and structural data in the training step.

CrossDome (16) aim to predict potential off-target toxicities by leveraging multiomics data from healthy tissues, structural information on TCR-pMHC interactions, and amino acid (AA) biochemical properties. This integrative methodology aids in generating statistically supported predictions that assist in risk assessments and enhance prediction specificity.

2 Random energy models of TCR-pMHC interactions

An early quantitative understanding of TCR-pMHC specificity began with the development of random energy models that aimed to explain known properties of the interaction, including TCR specificity versus degeneracy and the potential for self/non-self discrimination. We present here an overview of affinity-driven models, which characterize TCR-pMHC interactions by their free energy of binding. Additional models have considered the effects of kinetic features of the TCR-pMHC interaction (17–19), including on- and off-rates (20), TCR-pMHC binding lifetime (21), and the role of catch vs. slip bonds in TCR activation (22). Both binding affinity and kinetics are likely important for determining the overall outcome of a TCR-pMHC interaction (23–25). Significantly, these approaches can effectively explain the kinetic proofreading aspect of absolute ligand discrimination in a manner that is robust to antigen concentration (24, 26). However, due to the abundance of data on TCR-pMHC binding affinity (via estimated k_D values), we focus our discussion here on affinity-based models.

Early approaches modeled affinity-driven TCR-pMHC interactions using paired strings (27, 28), as detailed in a study by Perelson (29), and review various computational models for receptor representation and properties (30). In these models, interacting TCRs and peptides are represented by AA strings of length N. It is assumed that the total TCR-pMHC binding energy can be represented by the sum of individual pairs of interacting AAs:

$$E(t,q) = \sum_{i=1}^{N} J(t_i, q_i)\,. \tag{1}$$

In this case, $E(t,q)$ is the free energy of interaction between receptor t and antigen q. $J(t_i,q_i)$ is the interaction energy between the i^{th} AAs on the hypervariable (CDR3) region of the TCR (t_i) and the peptide (q_i), respectively, and N is the length of the variable regions of the TCR. Using this framework, researchers formulated digit string representations capable of explaining the large degree of alloreactivity observed in post-thymic selection T cell repertoires (31). The initial string model (27) used the set of bounded integers to represent the 'complementarity' between AA pairs, $t_i,q_i \in \{1,2,\ldots, K\}$, with $J(t_i,q_i) = |t_i - q_i|$, and has been applied to successfully model thymic selection and predict empirically observed T cell alloreactivity rates (27, 28, 32).

Chakraborty and colleagues extended this modeling framework (33) by substituting abstract digit string with experimentally observed AA interactions. This was achieved by replacing $J(t_i,q_i)$ with a pairwise AA potential - chosen to be the Miyazawa–Jernigan energy matrix (34). This modeling framework demonstrated that thymic negative selection favors TCR AAs with moderate interaction strengths to avoid T cell deletion due to high energy interactions with a small set of thymic self-peptide (33, 35). When applied to understand the selective pressures imposed on TCR recognition in the setting of HIV, this framework showed how the peptide binding characteristics of a particular HLA allele restriction resulted in enhanced recognition of viral epitopes (36).

Subsequent modeling efforts have investigated how thymic selection impacts the recognition of tumor-associated antigens using the above framework applied to diagonalized TCR-pMHC interactions (4). Here, the diagonalized interaction assumption simplifies the TCR-pMHC binding interface into a set of one-to-one contacts between the AA residues of the peptide and binding pockets of the TCR. This framework demonstrated that post-selection TCRs may capably recognize single AA differences in point-mutated self-peptides at nearly the same rate as unrelated foreign antigens. This work was then extended to describe the effects of non-diagonal interactions (37), which allow for multiple pairwise TCR-peptide AA contacts. These intricate contacts, identified from the proximity of TCR-peptide AAs in known crystal structures, represented by a TCR-peptide contact map. From this, subtle variations in TCR-peptide binding recognition profiles manifest in variable weights of interaction assigned to each of the peptide AA positions. An extension to (33) considered non-uniformity of these weights (37). The contact map $W = (W_{ij})$ contains weights W_{ij} for interactions between t and q in a given structure, and an associated AA interaction coefficient,

$$E(t,q) = \sum_{i,j} W_{ij} J(t_i, q_j) \tag{2}$$

Using this framework, non-uniformities in contact maps, which are highly variable for given MHC allele variants, can result in high-contacting peptide AA positions. At these positions, single-amino acid changes in wild-type peptides, such as cancer neoantigens or single nucleotide polymorphism peptide variants, result in an enhanced difference in the binding interaction that may ultimately enhance or break immunogenicity. Intriguingly, these statistical models predict a high likelihood of point-mutated self-peptide recognition, which suggests that thymic selection is more akin to a T cell memorization task directed at a list of important self-antigens to avoid rather than an intricately curated list of self-peptides whose tolerance confers wider immune protection. In order to improve forecasting capabilities and reduce the experimental efforts required to search for meaningful TCR-pMHC pairs, advanced machine learning algorithms have been incorporated into biophysical and probabilistic models to create data-driven and trainable predictive models. Figure 2 provides a non-exhaustive summary of recent open-source computational methods.

3 TCR-pMHC specificity prediction methods

The TCR-pMHC specificity prediction methods leverage advanced computational techniques to predict the highly specific interactions between TCRs and pMHC that are crucial to the initiation and effectiveness of adaptive immune responses. The

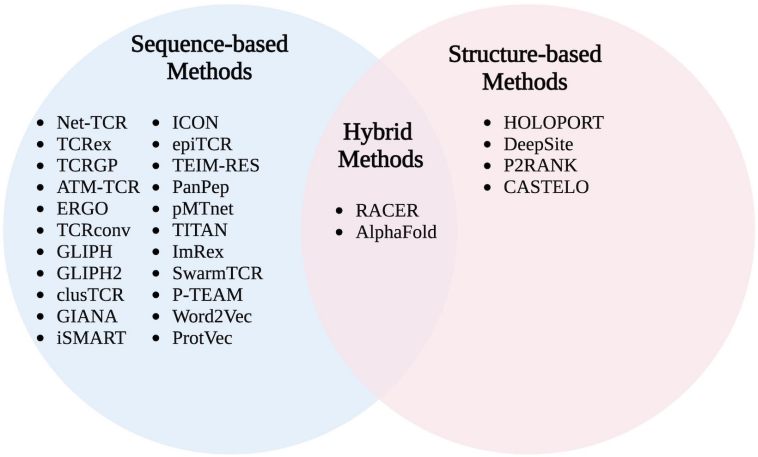

FIGURE 2
List of commonly used inference-based models of TCR-pMHC specificity partitioned by approach: sequence-based, structure-based, and hybrid.

methods discussed in this review can be classified into sequence-based, structure-based, and hybrid models. Each category uses different strategies to evaluate TCR-pMHC specificity, with sequence-based models relying on sequence similarities, structure-based models using three-dimensional structural information, and hybrid models combining these strategies. Various models have been developed to accomplish this, and their performance was assessed based on model generalization and robustness. However, the study found that these models struggle with generalization to peptides not seen in the training data and that their performance fluctuates depending on data size and balance, indicating that predicting TCR-pMHC binding specificity remains a significant challenge. Thus, additional high-quality data and innovative algorithmic approaches are necessary for further advancements.

3.1 Sequence-based approaches

Arguably the greatest challenge in predicting TCR specificity arises from the diversity of possible TCR and peptide combinations relative to those that can be studied or even realized in a single individual (38, 39). To experimentally identify relevant TCR-pMHC pairs, pMHC tetramers are often used to experimentally identify TCRs that interact with sufficient binding affinity. The affinity-based screening of TCRs can be done in a high-throughput manner (40). In addition to theoretical modeling, inferential statistical learning provides a complementary approach for studying this problem by imputing known (non-) examples of favored TCR-peptide interactions interactions. These computational models (Figure 1A) can be distinguished based on whether or not previously identified TCR-pMHC interactions are used in training, which are given by supervised and unsupervised learning approaches, respectively.

3.1.1 Supervised learning

Sequence-based prediction models refer to machine learning algorithms designed to learn a predictive function that identifies informative features and from them accurately predicts the cognate epitope of an input T cell receptor (TCR) with unknown specificity. Features are learned based on a set of known examples and non-examples of TCR-peptide pairs provided in a training dataset. Capietto et al. (41) demonstrated that peptide mutation positions matter in neoantigen prediction pipelines, and the use of this feature led to improved neoantigen ranking. Other studies found that immunogenic peptides have more hydrophobic AAs at TCR interaction sites and that AAs molecular weight, size, and charge are useful for TCR-pMHC complexes (42–44).Numerous computational tools now exist that use known TCR and peptide sequences to predict TCR-epitope interactions by binary classification, including such as NetTCR (45), TCRex (46), TCRGP (47), and ATM-TCR (48). Furthermore, there has been a progression towards developing TCR-epitope binding prediction models that are not limited to specific peptides, such as SwarmTCR (49), ERGO (50), pMTnet (51), ImRex (52), TITAN (53), and TCRconv (54) which instead utilize known binding TCRs to train the models. While these models work well with peptides having an abundance of known TCR interactions, they often struggle in predicting behavior for peptides having few known interactions or those not included in the training data, attributed to the large diversity of the interaction space. Many approaches consider TCRs and peptides as linear sequences of AAs, while others include a description of their three-dimensional orientations during the TCR-pMHC interaction.

Each of the above cases utilizes different machine learning procedures to achieve a variety of descriptions of TCR-pMHC specificity. NetTCR implements convolutional neural networks (CNN) (55, 56) in conjunction with multiple dense layers (Section 4) to learn the interactions between TCRs and epitopes associated with the most common human MHC allele variant, HLA-A*02:01. This method faces limitations due to the vastness of the TCR-pMHC space and insufficient experimental training data, which challenge current computational algorithms. ImRex, and TITAN are state-of-the-art TCR-epitope binding prediction models utilizing CNNs. ImRex, inspired by image processing CNNs, transforms CDR3 and epitope

sequences into interaction maps. It considers the pairwise differences of selected physicochemical properties of the AAs in the sequences, making the maps interpretable as multi-channel images, and predicts TCRepitope binding through a multi-layer CNN. ImRex's strength rests in its ability to recognize TCR-specific epitopes from unseen sequences that resemble the training data, improving its generalization performance for TCRepitope recognition. On the other hand, TITAN employs a one-dimensional CNN with a contextual attention mechanism (Section 4). It separately feeds encoded CDR3 and epitope sequences into convolutional layers, uses context attention layers for each, and concatenates attention weights. Significantly, TITAN extends beyond simple encoding by employing SMILES sequences for epitopes at the atomic level. In combination with transfer learning, this sophisticated method expands the input data space and enhances model performance. TITAN's attention heatmaps provide insights into biological patterns and suggest that data scarcity in epitopes can implicitly treat them as distinct classes, which could impact unseen epitope performance in complex models. Both models apply feature attribution extraction methods to explore underlying biological patterns.

TCRex utilizes a series of decision trees (57), which combines with classifiers and ensemble regression trees, to build an epitope-specific prediction model. While this method has been successfully applied for data classification and regression, several challenges involve a lack of generalizability to non-HLA A*02:01 cases and susceptibility to overfitting. EpiTCR (58), which uses the Random Forest algorithm, also integrates several crucial elements for increased precision, can significantly mitigate these issues. These include sequence encoding based on BLOSUM62, zero-padding to maintain sequence uniformity, and utilization of peptide-presenting MHC data in the predictions, providing a comprehensive approach to TCR-peptide specificity prediction. This is done by utilizing a large dataset from various public databases (over 3 million), which are encoded by a flattened BLOSUM62 matrix, and is known for their high sensitivity and specificity in detecting such interactions.

Transitioning from this approach, another promising methodology is presented by the Predicting T cell Epitope-specific Activation against Mutant versions (P-TEAM) model (59). P-TEAM, a Random Forest-based model, proficiently predicts the effect of epitope point mutations on T cell functionality. It provides quantitative predictions for altered peptide ligands and unseen TCRs, showcasing high performance and potential applicability in immunotherapy development. Several bioinformatic approaches, including SwarmTCR, predict antigen specificity from sequences. These tools optimize CDRs to enhance TCR specificity predictions using labeled sequence data. With robust performance on both single-cell and bulk sequencing data, it offers biologically interpretable weights, providing crucial insights into immune responses related to various conditions. TCRconv, a state-of-the-art deep learning model, is designed for predicting TCRs and epitope interactions. Using a protein language model and convolutional networks, it identifies contextualized motifs, improving the accuracy of TCR-epitope predictions. TCRconv, applied to COVID-19 patients' TCR repertoires, provides enhanced understanding of T cell dynamics and disease phenotypes, highlighting potential applications in infectious diseases, autoimmunity, and tumor immunology. TCRGP relies on analyzing both α and β chains of the T cell receptor (TCR) in a Gaussian process method in order to determine which CDRs are crucial for epitope recognition. ERGO incorporates Long Short-Term Memory (LSTM) (60) and autoencoder (AE) (61) models (Section 4) on a variety of inputs, including the TCR$\alpha\beta$ sequences and the VJ genes for each TCR (50). Similarly, ATM-TCR predicts the affinity between TCR and epitope based on a computational model. According to this model, AA residues within TCR and epitope sequences are considered in the context of how they interact with each other using a multi-head self-attention network (Section 4) (62). The TCR-pMHC binding prediction network (pMTnet) approach is a computational model that applies transfer learning (63) (Section 4), a form of deep learning that leverages knowledge gained from prior tasks, to predict TCR binding specificities for neoantigens and other antigens presented by class I MHCs. This model is directed at addressing the challenge of predicting TCR-antigen pairing and has demonstrated significant advancements over previous methods. As validated through the characterization of TCR-pMHC interactions in human tumors, pMTnet provides superior predictive accuracy. It effectively distinguishes between neoantigens and self-antigens, evaluates TCR binding affinities, and calculates a neoantigen immunogenicity effectiveness score (NIES). This ability allows for a comprehensive analysis of tumor neoantigens' role in tumor progression and immunotherapy treatment response, emphasizing the method's important contribution to understanding immunogenic tumor antigens and their relationship to T cell proliferation. On the other hand, pMTnet exhibits inferior performance in the few-shot settings and fails to recognize TCR binding to novel peptides in zero-shot settings, despite their robust performance in settings with an ample volume of known TCR binding data. Conversely, the effective implementation of the TITAN model, trained specifically with COVID-19 data using peptides with sparse known binding TCRs, is limited by the breadth of available empirical data, thus posing a significant hurdle to the prediction of TCR interactions with new or rare peptides.

Another perspective by Meysman et al. (64) compares TCR-pMHC binding approaches and concludes by stressing the need for an independent benchmark. The authors find improvements in predictions accuracy when including CDR1/2 information, but leave open a complete investigation of the impact of data imbalance with respect to the biological context of included training examples, size, and overtraining on model performance. The lack of standardization for these factors complicates TCR-pMHC binding prediction and comparative benchmarks. Subsequent evaluations by Meysman's group underscored the unpredictability of unseen epitope predictions, reinforcing the call for advanced models and rigorous, standardized evaluation protocols (65).

In this application, T cell receptor (TCR) sequences are transformed into numerical representations through a process called encoding. Encoding methodologies commonly utilize physicochemical properties or one-hot encoding, which is a technique where each unique AA in a sequence is represented by

its own unique binary code, making each one distinct for computational models. The immuneML (66) platform extends the capabilities of earlier methods like DeepRC (67), GLIPH2 (68), and TCRdist (3) to train and evaluate machine learning classifiers at the receptor level, and it accomplishes this by incorporating a variety of encoding methods for sequence data, including *k*-mer frequency decomposition, one-hot encoding, disease-associated sequence encodings, and repertoire distance encodings, facilitating comprehensive sequence analyses. This platform offers a variety of models, encompassing K-Nearest Neighbours (KNN), logistic regression, random forests, and the TCRdist classifier, among others, providing a versatile toolkit for receptor analysis. The inclusion of the TCRdist classifier allows for meaningful distance measurements between receptors, taking into account the unique characteristics of TCRs, like their exceptional variability and adaptability in recognizing different antigens. These models, therefore, provide a versatile toolkit for receptor analysis. The enhanced reproducibility, transparency, and interoperability offered by ImmuneML effectively overcomes traditional challenges in Adaptive Immune Receptor Repertoires (AIRR) machine learning. The review (69) provides a comprehensive overview of other advanced methods and computational tools emerging in this area of research, which facilitate a more complete and nuanced understanding of T cell receptor sequences and their functional implications. These include V(D)J recombination (70), single cell sequencing (71), multimodal experiments (72), flow cytometry (73), mass cytometry (CyTOF) (74), RNA sequencing, feature barcoding, and cell hashing.

A persistent challenge for these approaches is that such models often make incorrect predictions because of limited validated pMHC-TCR interaction data (38, 39, 75). According to Deng et al.'s study, the effectiveness of these models is significantly affected by the data balance and size (75). Furthermore, the models exhibited limitations in generalizing to untrained peptides, emphasizing the need for improved data collection and algorithmic improvements. Similarly, epitope binding affinity models such as TCRGP and TCRex cannot be used to investigate novel or understudied systems since they require that a new model be constructed for each epitope once there are a sufficient number of identified cognate TCRs. NetTCR, ERGO, and ATM-TCR models are all capable of predicting novel or rare epitopes, but they perform poorly overall as evaluated by the area under the receiver-operating characteristic curve (ROC-AUC) metric. A promising approach that addresses this issue is the Pan-Peptide Meta Learning (PanPep) (76), a meta-learning method combined with a Neural Turing Machine (NTM) (77). Meta-learning allows for the model to learn from a set of tasks and then apply insights gained to predict binding specificity for new and unknown tasks, such as predicting binding specificity for neoantigens or exogenous peptides. Using a NTM adds external memory to the system, ensuring retention of learned tasks and thereby improving prediction accuracy for TCR binding specificity to unknown peptides. Despite meta-learning's effectiveness when there are few examples available, its reliance on labeled data can limit its application. A powerful alternative to manual labeling, unsupervised learning is capable of extracting meaningful patterns from unlabeled data.

3.1.2 Unsupervised learning

In contrast to supervised learning, unsupervised learning does not rely on the availability of known TCR-peptide pairs, instead learning to group TCR, antigen, or HLA inputs based on statistical variation inherent in their sequences. Consequently, a number of approaches have attempted to train unsupervised models. The GLIPH (Grouping Lymphocyte Interactions by Paratope Hotspots) method utilizes high throughput data analysis to identify distinct TCR sequences that recognize the same antigen based on motifs shared in their CDR3 sequences (78), has a significant place in TCR-antigen interaction studies, and enhances TCR specificity prediction when combined with other resources, such as the V(D)J database. This clustering assists in pinpointing known TCR specificities. It's crucial, though, to understand the inherent constraints of GLIPH, including challenges in managing large datasets and the model's reliance on other resources for direct antigen interaction predictions. Understanding the specificity of the T cell repertoire in this context requires the identification of related systems from a small training subset in a high-dimensional space.

Given the absence of *a priori* identified specificity groups, to clustering methods may outperform traditional supervised classification schemes. In comparison to randomly grouped clones, TCRs within the cluster exhibited highly correlated gene expression and shared a common specificity. TCRs are clustered by GLIPH based on two similarity indexes: 1) global similarity, which refers to the difference between CDR3 sequences up to one AA, and 2) local similarity, which refers to the fact that two TCRs share a common CDR3 motif of 2, 3, or 4 AAs (enriched relative to that of a random subsampling of unselected repertoires). Moreover, the GLIPH algorithm, by adeptly recognizing shared motifs within the CDR3 of TCR sequences, has a significant place in TCR-antigen interaction studies and enhances TCR specificity prediction when combined with other resources, such as the V(D)J database. This combination assists in pinpointing known TCR specificities. It's crucial, though, to understand the inherent constraints of GLIPH, including challenges in managing large datasets and its reliance on other resources for direct antigen interaction predictions

TCR sequences with a shared epitope specificity carry motifs that are statistically enriched in the peptides they mutually recognize. A method that builds on GLIPH to include motif-based clustering (GLIPH2) (68) is fast but lacks specificity, while clusTCR (79) is faster because it encodes CDR3 sequences using physiochemical features by representing them as integers with an assigned Hamming distance. This comes with the tradeoff of lacking TCR variable gene information, and thus clusTCR has lower clustering purity defined as the proportion of items in a cluster that belong to the most common category or group. To address this challenge, the Geometric Isometry-based TCR Alignment Algorithm (GIANA) (80) transforms CDR3 sequences using the Nearest Neighbor (NN) search in high-dimensional Euclidean space to solve the problem of sequence alignment and clustering. In these methods, similar features are found among TCRs recognizing the same target. Similar TCRs can be grouped/clustered by predicting which targets they will recognize in this

way. Various additional factors, including alignment of T cell receptors and identification of TCR-antigen interactions using high-throughput pMHC binding data, are also considered in other methods (13, 81).

The identification of Tumor-Associated Antigen (TAA)-TCR specificity as a subset of the overall predictive task has historically been challenging in large part owing to the fact that a majority of TAAs are point-mutated self-peptide. Given the tremendous clinical value of reliably predicting TCR-TAA specificity, several algorithms have been developed for this specifically. In one case, called TCR Repertoire Utilities for Solid Tumors (TRUST), assembles hypervariable CDR3 regions of TCRs, and then applies a clustering method called immuno-Similarity Measurement using Alignments of Receptors of T cells (iSMART) to group TCRs based on their antigen specificity (81). Despite these advances, no systematic evaluation of these methods has been conducted on large and noisy datasets, and experiments to reduce nonspecific multimer binding, validate correct folding, and improve signal-to-noise ratios are still required. Integrated COntext-specific Normalization (ICON) (13) is a notable development in this field that identifies TCRantigen interactions in high-throughput pMHC binding experiments. The experimental approach consists of initial filtering of T cells based on single-cell RNA-seq, followed by background noise estimation via single-cell dCODE-Dextramer-seq, and then lastly TCR identification via paired $\alpha\beta$ single cell TCR sequencing. The TCRAI neural network predicts and characterizes these interactions and in doing so reveals conserved motifs and binding mechanisms. The combination of ICON and TCRAI leads to the discovery of novel subgroups of TCRs that interact with a given pMHC via diverse mechanisms.

Although many clustering-based approaches have been developed, conventional clustering methods usually perform poorly on high-dimensional data often as a result of inefficiencies in the defined similarity measures (82–84). On large-scale datasets required for studying TCR-pMHC specificity, these methods are generally computationally formidable. Consequently, raw data is often mapped into a more suitable feature space where existing classifiers can separate the generated data more easily, followed by dimensionality reduction and feature transformation. A number of existing transformation methods have been applied to this problem, including linear methods like Principal Component Analysis (PCA) (13) as well as non-linear strategies such as kernel methods and spectral methods (85). Clustering methods often encounter difficulties when dealing with complex structures owing to the fact that their clustering criteria are based on simplified criteria, in contrast to Feed-Forward neural networks and Deep Neural Networks (DNNs), that provide highly non-linear transformations of data that can be used to cluster the data. Further advancements in artificial intelligence have led to deep learning surpassing other statistical methods in many domains (86–91). Unlike traditional machine learning algorithms, which often struggle to assimilate complex features from data, the success of deep learning relies on understanding and interpreting data, which occurs by first learning simple patterns at initial levels of the algorithm and complex patterns at higher ones (92).

Several developed approaches utilize deep learning in an unsupervised manner, and although we discuss the specifics of the deep learning algorithms in Section 4, we will touch briefly on several here. One area where unsupervised deep learning applied to identify meaningful sequences includes the application of Natural Language Processing (NLP) algorithms based on word embedding, such as Word2Vec (93) and ProtVec (94). These algorithms offer a novel approach to understanding the relationship between TCR sequences and antigen binding. By leveraging the concept of word embedding from NLP, they are capable of capturing semantic or functional similarities among TCR sequences, much like similar words in a language (95). Therefore, if two TCR sequences share common motifs, it suggests they may bind to similar antigens. Consequently, these algorithms are valuable tools in immunoformatics, converting raw TCR sequence data into a format conducive to modeling and predicting TCR-antigen interactions. Word2Vec interprets non-overlapping 3-mer sequences of AAs, while ProtVec represents proteins as the sum of overlapping sequence fragments of length k. These approaches had several limitations, including limited interpretability due to the lack of biophysical meaning of three-residue segments of protein sequences, and overlapping models often do not out-perform non-overlapping models (94). Recurrent Neural Networks(RNN) (96) were proposed to improve these initial schemes. The RNN model is a sequence-based representation method averaging over the representations of each residue to produce a fixed-length real representation of arbitrary-length protein sequences. This scheme is further improved by implementing a transformer, which differs from RNNs by its incorporation of parallel task assignment. Models based on the transformer were found to be superior to traditional LSTM-based approaches (a variety of RNNs introduced in Section 3.1 and discussed further in Section 4) (60) when applied to tasks such as TCR-pMHC interactions, protein docking, and protein structure prediction, since in these cases the RNN model struggles to capture long-range relationships and does not include parallelizability (97, 98).

More recently, AlphaFold, an artificial intelligence system developed by DeepMind that predicts protein structure using primary sequence information, has been applied to the TCR-pMHC specificity problem (99). This method is a transformer model that utilizes an attention mechanism in order to operate within each row of a Multiple Sequence Alignment (MSA), which generally the alignment of multiple protein sequences of similar length to maximize the positional correspondence of homologous residues across these sequences. This attention mechanism (100) allows the model to focus on specific parts of the sequence, providing a more comprehensive understanding of the relationship between residues and protein folding. The ultimate output is in the form of an accurate 3-dimensional structure that can be assessed for binding specificity. We note that because of this, AlphaFold is a pure sequenced-based prediction model since no structural data is used as input (101).

Lastly, AEs and Variational Autoencoders (VAEs) (102), which stochastically map the input space to the latent space, have surpassed former techniques in the field of sequence-based representation. In contrast, the VAE model is designed to capture

the dynamics of the peptide-MHC binding process and to identify per-residue binding contributions by providing a stochastic map between the input and latent space. VAEs in peptide-MHC binding optimization have great potential for advancing the design of vaccines and immunotherapies (103). One recent study, TCR–Epitope Interaction Modelling at Residue Level (TEIM-Res) (104), uses the sequences of TCRs and epitopes as inputs to predict pairwise residue distances and contact sites. An epitope feature vector generated by an AE is fed into an interaction extractor for global epitope information. Using this approach, the method was able to predict TCR-epitope interactions at the residue level, outperforming existing models and demonstrating versatility in mutation and binding pattern analyses.

In addition to supervised and unsupervised learning methodologies, negative data plays a crucial role in enhancing model accuracy and preventing overfitting. By providing contrasting data, negative data aids in identifying patterns and trends in positive data, leading to a more enriched learning process (105). However, while useful for TCR–epitope binding prediction, this study also uncovers the potential pitfalls of its application. The bias it introduces can lead to a dip in model performance in practical scenarios. For instance, the PanPep model was observed to underperform with shuffled negative data. As a result, it is imperative to seek more effective strategies to preserve model practicality while also enhancing applicability, including the uniform employment of a negative sampling strategy during both the training and testing phases (106).

3.2 Structure-based approaches

Whereas sequence-based approaches contain no explicit spatial information on the interacting system, several alternative strategies have leveraged structural knowledge of the TCR-pMHC interaction to aid in understanding specificity. When available, structural templates couple primary sequence data with significant spatial information of the interacting pairs, thereby enabling sophisticated computational methods for representing and analyzing structures (Figure 1B). Additionally, two models have been developed to explain the T cell's ability to discriminate between self and non-self pMHCs that utilize the identification of a specific conformational change in the TCR complex and kinetic thresholding (23–25). Direct measurements of signaling molecules and pMHC-TCR ligand interactions are used to develop a model that accounts for the characteristics of T cell signaling in response to antigens.

Despite a large abundance of protein crystal structures (currently over one million in the Protein Data Bank), the number of identified TCR-pMHC crystal structures is quite limited (on the order of hundreds of TCR-pMHC complexes), likely due to the difficulty in producing these complexes in large quantities and in conditions suitable for crystallization. Computational methods for structure representation and analysis include Molecular Dynamics (MD) simulations, homology modeling (107), machine learning, alchemical free energy perturbation (108), and hybrid approaches.

MD simulations have also been used to establish a detailed, all-atom description to better understand TCR-pMHC specificity (109, 110). MD analysis provides an in-depth, mechanistic understanding of TCR and pMHC interactions. However, due to the high computational cost of these approaches, an MD-derived understanding of TCR-pMHC specificity is at present restricted to a small collection of TCRs and peptides in a given analysis. Nonetheless, these insights are critical to predicting TCR-pMHC specificity, as they allow for an understanding of the molecular behaviors and relationships that underpin this complex biological interaction. In this way, MD simulations effectively bridge the gap between fundamental biophysical interactions and the computational prediction of TCR-pMHC binding. This approach begins by generating an initial structure, which can be achieved through side-chain substitution, homology modeling (107), and ligand-protein docking (111), and proceeds using time-dependent simulations of atomic motions in the system, MD simulations account for both the main-chain conformational flexibility and the solvation and entropy effects. The simulation protocols themselves can be accelerated through the use of coarsegraining, increased masses (112), virtual sites (113), n-bead models (114), or the movement of rigid protein regions (115). A variety of pertinent features, including RMSD, RMSF, Solvent-Accessible (SASA), PCA, and hydrogen bonds can be analyzed based on MD simulations, and geometric approaches (116) have also been developed to analyze the binding orientation between the heavy and light chains of antibodies and the TCR α and β chains. Collectively, this approach can provide highly detailed information on the dynamics of TCR-pMHC systems. However, the high computational cost of performing full MD simulations limits feasible analyses to several TCR-pMHC pairs (117).

Molecular Mechanics (MM) provides a complementary approach to study the bound TCR-pMHC complex using molecular docking techniques. The molecular docking process has two key applications: binding mode prediction and virtual screening. The former involves optimizing the 3D conformation of a peptide when it binds to its target receptor, while the latter entails evaluating a vast number of potential peptides to identify those that can bind to the target receptor (118). In studying the TCR-pMHC interaction, both MD and MM approaches are both challenged by cases having significant peptide flexibility, since a peptide with more flexible bonds can adopt more conformations. In addition to the position and orientation of the peptide inside the receptor's binding cleft, docking methods must consider these alternative conformations in order to determine the most suitable binding mode.

The field of MM utilizes simulation-based prediction methods, which involve tracking the time evolution of a molecular system through the use of an energy potential. The quality of the potential, or score function, plays a crucial role in protein structural modeling, as it describes the potential energy landscape of a protein. Score functions may also contain knowledge-based terms to distinguish native from non-native conformations. MD or Monte Carlo (MC) simulations with advanced force fields or score functions can accurately reproduce the statistical behavior of biomolecules. The MM-based task of learning a force field with predictive utility has

recently been augmented by incorporating deep learning-based approaches. These approaches represent each atom's chemical environment through graph convolutions (119) and by doing so aim to enhance the accuracy and reliability of MM predictions, through the capture of complex atomistic relationships in local and global chemical environments and generation of transferable, interpretable features that facilitate end-to-end learning. These approaches can broadly be categorized into two categories: graph-based and fingerprint-based.

Graph-based approaches construct a mathematical graph of molecules, containing atoms as nodes and chemical bonds as edges. They maintain structural and chemical information and preserve topological complexity to facilitate more detailed and complex molecular structural analyses, which can be used to predict chemical reactions and molecular docking. In contrast, fingerprint-based approaches represent molecules as binary digits. While these methods provide a computationally efficient, fixed-length representation, they simplify the molecular structure and may lose fine-grain detail about the exact structure and topology. Dual methods that combine both strategies also exist and have been applied to studying the TCR-pMHC interaction. Collectively, these approaches have been shown to enhance the accuracy of molecular modeling in describing simple molecular pairs and possess potential for describing more complex biological processes, including protein complex interactions. The current methodology for computational Protein-Protein Interaction (PPI) prediction is largely based on deep learning methods.

One example of a dual methodology is a multiscale graph construction of HOLOPROT (120), which connects surface to structure and sequence, demonstrates the utility of hierarchical representations for binding and function prediction. Using geometric deep learning and mesh CNN (55, 56) embed protein surface patches into fingerprints for fast scanning and binding site identification, eliminating the need for hand-crafted or expensive pre-computed features. Importantly, these methods do not perform structural blind docking, which involves determining the binding site, orientation, and location of the two molecules, and internal conformational deformations during binding. Consequently, they capture and predict molecular interactions based on effective molecular representations and efficient learning algorithms, without explicitly simulating binding dynamics.

Another example includes Graph Deep Learning (GDL) methods. While they are reliant on known structural data, GDL approaches offer unique advantages in capturing the complex, non-linear relationships between features, making them potentially valuable for predicting protein structures (121), interactions (122), and functions (123). AlphaFold has revolutionized PPIs modeling with its sophisticated end-to-end approach, which outperforms traditional docking methods. In order to accurately model complex interactions, such as T cell receptor-antigen complexes, further enhancements are needed. This challenge might be addressed by building upon AlphaFold or integrating it with geometric deep learning (124).

Notably, AlphaFold's prowess lies in its ability to deduce a protein's 3D configuration from its primary amino acid sequence. From this, AlphaFold's EvoFormer module learns complex patterns of AA interactions and predicts the distances and orientations of those interactions in 3D space, with the goal of essentially providing an estimated structural representation. Moreover, it uses a structure-based method for refining the coordinates of all heavy atoms within a protein (101). Because AlphaFold can generate detailed structural predictions from primary sequence information alone, its use in identifying relevant TCR-pMHC interactions is particularly intriguing. A recent approach utilizes a modified version of AlphaFold to resolve correct and incorrect peptide epitopes in TCR-pMHC interactions (125). This study suggested that supervision is required for appropriately applying the AlphaFold approach to TCR-pMHC systems: In comparison to the default AlphaFold (126), AlphaFold-Multimer (99), designed specifically to interrogate protein-protein structural complexes, more capably predicts TCR-pMHC binding specificity at a lower computational cost and higher accuracy.

4 Deep learning approaches

Deep learning, a machine learning subclass, is dramatically transforming the exploration and comprehension of TCR specificity. Machine learning excels in pattern recognition and prediction, making it versatile in applications like predicting cell types or antibody affinity based on gene expression profiles. However, the laborious feature extraction process, particularly with vast, feature-rich data, is a limitation. Deep learning alleviates this with an automated approach for feature extraction. Its layered structure facilitates capturing complex, high-dimensional data patterns, despite its interpretability challenges. CNN and RNN, two key Deep learning models, find varied biological applications, from image processing to protein engineering. Deep learning is poised to revolutionize TCR specificity understanding, and possesses the potential for ushering in the design of optimized immune treatment strategies.

4.1 Deep learning architecture

In contrast with the computational approaches discussed in detail thus far, which use physical equations and modeling to predict data, machine learning algorithms infer a relationship between inputs and outputs by learning from a set of hypotheses. This can be described by a collection of K training samples that may contain features x in an input space $\mathcal{X}$ (e.g. AA sequences), and corresponding labels y in output spaces $\mathcal{Y}$ (e.g. pairwise residue distances), where $\{x_j, y_j\}_{j=1}^{N}$ are sampled independently and identically (i.i.d) from some joint distribution. Additionally, an identified function $f : \mathcal{X} \rightarrow \mathcal{Y}$ maps inputs to labels, and a corresponding loss function $l : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ measures how far $f(x)$ deviates from its corresponding label y. In supervised learning, the goal is to find a function f that minimizes the expected loss, $\mathbb{E}_{(x,y)\sim p}[l(f(x),y)]$, for (x, y) jointly sampled from. Parameterization of the hypothesis class depends on the allowable choice of the network f in some allowable space $\mathcal{F}$.

Data analysis and deep learning predictions often overcome the traditional challenges of feature extraction in ML by recognizing

relevant features, constructing hierarchical representations, handling large datasets effectively, providing end-to-end learning, and facilitating transfer learning, overcoming the limitations of classical approaches. High-dimensional data tasks can be efficiently handled with deep learning algorithms using hierarchical artificial neural networks. However, the interpretability of neural networks and deep learning can be a problem, due to their complexity, non-linearity, and the lack of physical interpretation and transparency due to their black-box nature. We will describe in detail the use of several common architectures (Figure 3), such as CNNs, RNNs, VAEs, and Generative Adversarial Networks (GANs), which have been developed for different applications, including biological problems such as cancer immunology (127).

4.1.1 Convolutional neural networks

CNNs are a subtype of deep learning network architecture that have historically performed well on two-dimensional data with grid-like topologies, including images, and this approach is also applicable to other problems requiring shift-invariance or covariance (128). In order to capture this translational invariance, CNNs use convolutional kernels (feature extraction) for layer-wise affine transformations. There are three factors involved in the learning process of a CNN: sparse interaction, parameter sharing, and equivariant representation (129). CNNs utilize convolutional layers for sparse interaction, enabling efficient processing of high-dimensional data while reducing computational demands. Parameter sharing across input data locations decreases required parameters, enhancing training and inference efficiency. Lastly, equivariant representation ensures the network's output remains invariant to input transformations, promoting generalization across diverse input variations. CNN has been applied to predict protein residue distance maps based on AA sequences (130) (Figure 3A). Convolutional operation $*$ with respect to the Kernel W and 2D data X (in this case, represented by residue-residue distance maps from AA sequences) can be expressed as

$$(X*W)(i,j) = \sum_m \sum_n X(m,n)W(i-m, j-n), \tag{3}$$

Where $(X * W)(i,j)$ denotes the convolution output at position (i,j), and $X(m,n)$ and $W(i-m, j-n)$ represent the value of the input X at position (m,n) and the parameter of the kernel at position $(i-m, j-n)$, respectively. One important variation on this general scheme that is relevant to the TCR-pMHC problem, called Residual Network (ResNet) (131), includes skip-connections between layers to recover spatial information lost during down-sampling. AlphaFold is one example of such an approach that uses ResNets to predict inter-residue distance maps of primary AA sequences (132).

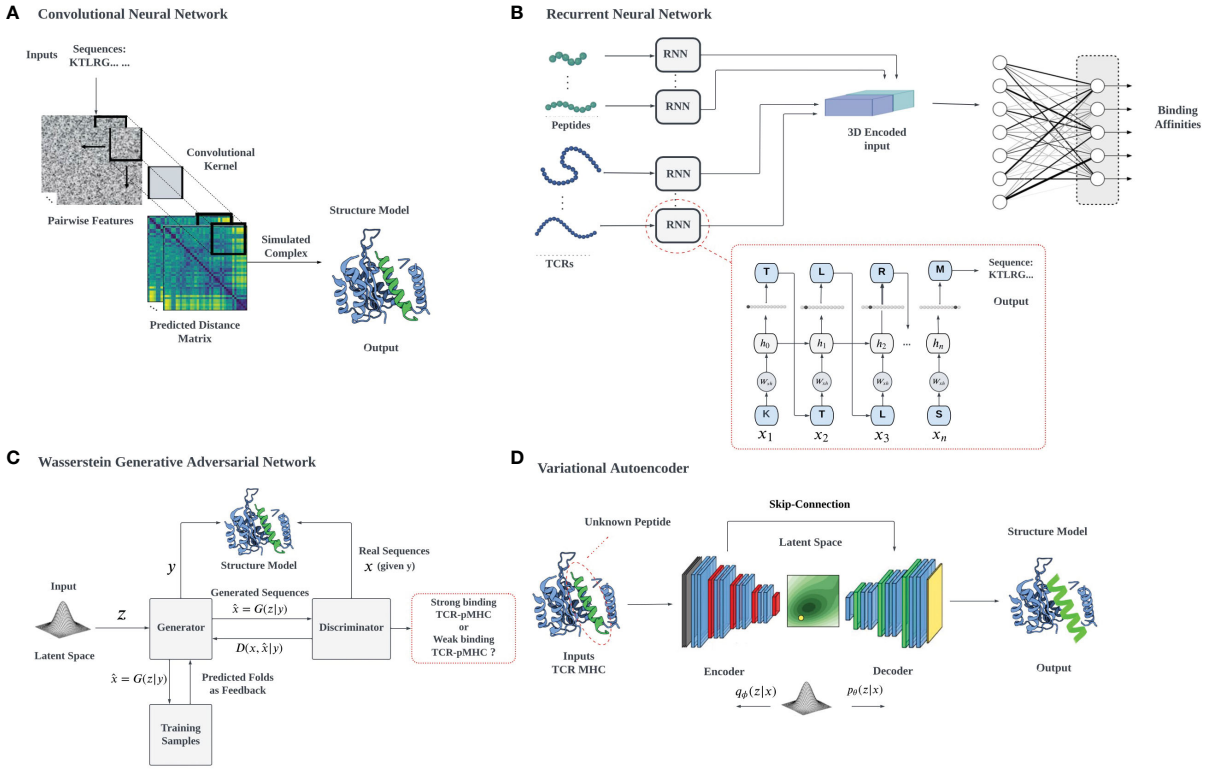

FIGURE 3
A schematic illustration of various deep learning architectures employed for TCR-pMHC interaction prediction: **(A)** 2D CNN-based prediction of TCR-pMHC interactions: The pairwise features of protein sequences are encapsulated in a 2D matrix representation, which serves as input for the 2D CNN. The CNN systematically samples the entire protein pairwise feature space, processing the data to facilitate the learning of TCR-pMHC interactions, **(B)** RNNs utilize auto-regressive learning to generate sequences, which can be applied in the context of TCR-pMHC interaction prediction, **(C)** In the GAN framework, a mapping from a prior distribution to the design space can be obtained through adversarial training, enabling the generation of novel TCR-pMHC interaction predictions, **(D)** VAEs can be jointly trained on protein sequences and their properties to construct a latent space that correlates with the properties of interest, for example, the TCR binding capacity of unevaluated target peptides.

CNNs can also be used to treat 3D protein structure prediction as a computer vision problem by voxelizing a given structure. One example is DeepSite (133), which uses voxelized representations of different atom types and deep CNNs to predict binding sites. Despite DeepSite's potential to capture more interactions using voxelized representations and larger datasets, its performance appears lower than an alternative, template-free machine learning method (P2Rank) that applies clustering to score regions of a protein's solvent accessible surface to identify candidate binding pockets (134). This discrepancy is possibly due to the CNN approach requiring even larger training dataset or differences in training set distributions. Yet another method employs a CNN-based segmentation model inspired by U-Net to predict binding sites in a single step (135). In general, U-Net is a CNN architecture originally designed to segment biomedical images. It utilizes symmetric encoder-decoder structures with skip connections between mirrored layers in both encoding and decoding paths, which allows accurate localization and the preservation of detailed information. In this method, a three-dimensional grid is generated around the protein, and each voxel within the grid is assigned a probability of being part of a binding pocket. The U-Net-inspired approach offers a more streamlined prediction process compared to traditional methods and has shown improved performance in detecting binding when compared to DeepSite, another prominent tool in the field. Overall, both P2Rank and U-Net-inspired methods offer unique advantages for the identification and prediction of protein-ligand binding sites.

4.1.2 Recurrent neural networks

RNNs are neural networks that operate on sequential data (96), such as time series data, written text (i.e., NLP), and AA sequences (Figure 3B). The RNN algorithm can be represented by in the following mathematical setup, where a hidden state $h^{(n)}$ is recursively solved using an initial value $h^{(0)}$ and sequential data $[x^{(1)},x^{(2)},\ldots,x^{(N)}]$, via

$$h^{(n)} = z^{(n)}(x^{(n)}, x^{(n-1)}, \ldots, x^{(2)}, x^{(1)}) = g(h^{(n-1)}, x^{(n)}; \theta)\,.$$

Here, θ represents the RNN parameters, which include the weights and biases associated with the network's connections, learned during the training process. The function g represents the update function describing the transformation from one position to another and utilizes the previous hidden state $h^{(n-1)}$, current input $x^{(n)}$, and θ parameters to produce the updated hidden state $h^{(n)}$. $z^{(n)}$ represents the cumulative transformation for position n. The hidden state vector contains all previously observed information at position i. Using this approach, sequential data of variable length can be fed to an RNN. This approach can be susceptible to a vanishing gradient, complicating optimization, and the 'explosion problem' (the error signal decreases or increases exponentially during training), potentially affecting the predictive accuracy and robustness of TCR-pMHC model. Specifically, the recurrence relation $h^{(n)} = g(h^{(n-1)},x^{(n)};\theta)$ in this context becomes especially vulnerable. When back-propagating through time over multiple steps, the gradient with respect to the loss function L, which measures the discrepancy between predicted and actual outcomes, can either shrink or grow exponentially. This behavior is due to the repeated multiplication by the weight matrix, as described by $\partial L/\partial h^{(n-t)}$. If the network's weights, in the context of TCR-pMHC modeling, are not properly initialized or regularized, it can lead to gradients significantly diverging from the ideal range. Consequently, LSTM networks (60), which are commonly used to mitigate this (136). An example of an LSTM approach in the context of specific TCR-Peptide binding prediction is using embedding vectors of AAs to construct a single vector, which can then be used as an LSTM (137) to learn long-range interactions within AA sequences; however, their efficacy depends on the formulation of the problem, the dataset characteristics, and the network architecture. In some situations, alternative deep learning approaches, such as CNNs and transformers, may be more applicable.

As an alternative to the recurrent network architecture, the attention mechanism is a method that can be used to improve the information processing ability of the neural networks (100). This mechanism is inspired by human biological systems that process large amounts of information by focusing on distinct parts and works by preventing the system from processing available information simultaneously (62). Attention-based models have several advantages over RNN models, including their parallelizability and ability to capture long-range relationships. The transformer model (62), a groundbreaking deep learning architecture is characterized by its self-attention mechanism, which enables the processing of input sequences in parallel rather than sequentially, distinguishing it from traditional attention mechanisms that typically rely on recurrent or convolutional layers. AlphaFold-Multimer (99) is one example of a transformer model that employs the attention-based model to generate models of TCR-pMHC interaction, which can then be used to distinguish correct peptide epitopes from incorrect ones with substantial accuracy. In directing these approaches to TCR-pMHC data in the future, these methods could be particularly helpful for predicting a target residue or the desired residue-specific properties of a target residue from the AA sequence of a protein. For example, transformer-based models have already been used to generate protein sequences conditioned on target structure and learn protein sequence data to predict protein-protein binding interfaces (138).

4.1.3 Variational autoencoder

The AE neural network is an unsupervised learning algorithm based on backpropagation that sets its target values equal to the input values (61). This is typically accomplished by mapping input to latent space in the encoder and reverse mapping in the decoder (Figure 3D). The latent space's dimension is less than the dimension of the original input and is constrained in some way (for example, by sparsity). In this framework, one assumes a set of unlabeled training vectors, $\{x^{(1)},x^{(2)},x^{(3)},\ldots\}$, where $x^{(i)} \in \mathbb{R}^n$. AE attempts to approximate the identity function in order to produce output y that is similar to x with respect to a loss function L: n

$$\theta = \underset{y^{(i)}\in R^n}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} L(x^{(i)}, y^{(i)}) \tag{4}$$

In one AE application directed at TCR-pMHC interaction prediction (139), researchers predicted PPIs from AA sequences

in order to identify key antigenic features to gain a more detailed understanding of the underlying immune recognition process.

VAEs (102) build on AEs by providing a stochastic mapping between the input space and a lower dimensional latent space, which is particularly useful when the input space follows a complex distribution. The latent space distribution typically takes a much simpler functional form, such as a multivariate Gaussian. Variational Inference (VI) (140) is a machine learning technique used in VAEs that approximates complex probability densities through optimization, allowing for efficient learning and data compression in the transformed latent space. Comparatively, it is faster than classical methods, such as Markov chains and MC sampling. In the VI method, the stochastic encoder is trained so that it approximates the true posterior distribution $p_\theta(z|x)$ of the representation z given the data x with parameters θ, by means of the inference model $q_\phi(z|x)$ with parameters ϕ, and weights parameterized by the data. In contrast, a decoder gives an estimate of the data given the representation, $p_\theta(x|z)$. However, direct optimization is not computable; thus, training is done by maximizing the evidence lower bound (ELBO), $\Lambda_{\theta,\phi}(x)$, which gives a lower bound on the log-likelihood of the data:

$$\Lambda_{\theta,\phi}(x) = \mathbb{E}_{z\sim q_\phi(z|x)} \log p_\theta(x|z) - D_{KL}(q_\phi(z|x)||p_\theta(z|x)) \quad (5)$$

where in general $\mathbb{E}_{z\sim q_\phi(z|x)} log\, p_\theta(z|x)$ represents the expected value of a function $log\, p_\theta(z|x)$ with respect to the conditional distribution $q_\phi(z|x)$, which measures the average value of the function $log\, p_\theta(z|x)$ when considering all possible values of z, weighted by the probabilities assigned to them via $q_\phi(z|x)$. $D_{KL}(q_\phi||p_\theta)$ is the Kullback-Leibler divergence quantifying the distance between two distributions q_ϕ and p_θ, which represents the similarity of the latent space distribution with the target distribution $p(z)$. An example of VAE prediction in the TCR-pMHC interaction prediction field includes the CASTELO approach, which was used in combination with MD simulations to identify mutated versions of a known WT peptide that lead to enhancements in TCR-pMHC binding (103). Future applications of VAE-based prediction schemes will likely make an impact on describing TCR-pMHC interactions in combination with other preexisting strategies.

4.1.4 Generative adversarial networks

GANs (141) are an emerging technique for both semi-supervised and unsupervised learning (142) that provide a method to obtain deep representations without the necessity to employ extensive training data annotations. In contrast to VAEs, GANs are trained through adversarial games between two models or networks (Figure 3): a generator network, G, which maps from latent space $\mathbb{R}^{|z|}$ of dimension $|z|$, to the space of data, $G: \mathcal{G}(z) \rightarrow \mathbb{R}^{|x|}$, where $z \in \mathbb{R}^{|z|}$ is a sample from latent space or simple distribution $p_z(z)$ (e.g. Gaussian), $x \in \mathbb{R}^{|x|}$ is a data-point, D is a discriminator function that maps an example to the probability that the example belongs to the real data distribution rather than the generator distribution (fake data), $D: \mathcal{D}(x) \rightarrow (0, 1)$. This game-based setup trains the generator model $G \in \mathcal{G}$ by maximizing the error rate of the discriminator, D, so that the discriminator is fooled. On the other hand, the discriminator $D \in \mathcal{D}$ is trained to recognize fooling attempts. It is expressed as the following objective (143):

$$\min_{G\in\mathcal{G}} \max_{D\in\mathcal{D}} V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]. \quad (6)$$

In training, this loss function is optimized stochastically. Both the generator and discriminator are trainable via Standard Gradient Descent (SGD) algorithms. The discriminator can be updated M times for every generator update. After training, synthetic data is created using only the generator network.

GANs have been making rapid progress in continuous domains, but mode collapse and instabilities can occur when training this GAN objective and has made analyzing discrete sequences a significant challenge. One variation, referred to as the Wasserstein GAN (WGAN) (144, 145), introduces a penalty to constrain the gradients of the discriminator's output, resulting in a more stable and trainable model. While GANs utilize a sigmoid function in the last layer for binary classification, the WGAN approach removes this function to approximate the Wasserstein distance (146), using Lipschitz discriminators: namely, that for discriminator function D there exists a constant L such that $|D(x) - D(y)| \leq L||x - y||$ for any two points x and y in the input space. This ensures that the gradient of the discriminator's output with respect to its input is bounded by some constant $K: \| \nabla(D(x)) \| \leq K$.

GANs can be used in protein modeling to produce new protein-like folds by learning the distribution of protein backbone distances. In one application, one network, G, generates folds, while a second network, D, distinguishes generated folds from fake ones (147). While WGAN models have not yet been widely applied to study TCR-pMHC specificity, they have been used to generate genomic sequence data (148). While their optimization behavior is generally well behaved, WGANs can exhibit undesired behavior in some applications. For example, in generating sequences containing particular motifs in the above application, in some cases, a strong motif match appeared twice in the same generated sequence because the final predictor score was insensitive to the presence of two motifs (the best match is used). Biologically, such sequences can be undesirable. Other technical issues that impact GAN approaches include unstable objective functions, mode collapse, variable length structure generation, conditioning difficulty, and the need to sample from a distribution instead of predicting a single output (149), and various approaches (144, 145, 150–153) have attempted to address these issues. GANs have influenced the field of sequence design, both when conditioning structural information (154) and when not (155, 156).

Diffusion models, an alternative to GANs, address many of these issues. Diffusion models are a class of latent variable models modeling the data generation process as iterative denoising of a random prior. They use a specific parameterization of the approximate posterior distribution that can be interpreted as an unobserved fixed prior diffusing to the observed posterior distribution (157). The diffusion model addresses some limitations of GANs by enabling explicit density estimation, reducing the mode collapse problem often seen in GANs, and providing more stable training procedures.

Due to several key differences, data-driven generative modeling methods have not had the same impact in the protein modeling setting as in the image generation setting. The first difference between proteins and images is that proteins cannot be represented on a

discretized grid that is amenable to the straightforward application of generative models. Inconsistencies in the predictions of the pairwise distance matrix of a protein's atoms lead to nontrivial errors when optimization routines are used to recover the final 3D structure when using existing models (158). Furthermore, proteins are not naturally oriented in a canonical manner like images. Therefore, rotationally invariant methods must account directly for this factor of variation in model weights. This reduces the amount of effective model capacity that can be dedicated to structural variation.

5 Hybrid approaches

In modeling natural systems, the exponential family of pairwise models is an important class of distributions to consider, which enjoys mathematically interpretable forms and is sufficiently general to include many of the common distributions, such as Gaussian, Poisson, and Bernoulli distributions (159). Additionally, pairwise models are commonly used in the statistical physics community for the analysis of categorical sequence data. There have been many successful applications of pairwise models such as the Ising model (160) or the generalized Potts model (91). One of the open questions in this area is how to train such models when additional higher-order interactions are present in the data that cannot be included in a pairwise model. Hybrid models addressed these issues, which combine a pairwise model with a neural network and can significantly improve pairwise interaction reconstruction. These hybrid approaches can often demonstrate performance improvements over alternative methods. We will focus on one particular example of a hybrid model recently developed to characterize systems-level TCR-pMHC specificity.

The Rapid Coarse-Grained Epitope TCR (RACER) (161, 162) model utilizes high-throughput TCR and peptide data, crystal structures, and a pairwise energy model to accurately predict TCR-peptide binding affinities. In this approach, supervised machine learning is applied to pre-identified TCR-peptide structures (45, 137) and experimental data to derive a coarse-grained, chemically accurate energy model of the TCR-pMHC interaction. While deep learning algorithms can implicitly capture higher-order interactions, they may still be limited by the availability of sequences. To mitigate this, RACER uses pairwise potentials to reduce the requirement for extensive sequence data. The optimization framework employed by RACER utilizes the AWSEM force field (163) to represent direct PPIs:

$$V_{direct} = \sum_{\substack{i \in \text{TCR} \\ j \in \text{peptide}}} \gamma(a_i, a_j)\Theta_{ij}^{I} \tag{7}$$

Where $\gamma_{ij}(a_i, a_j)$ denotes the pairwise interaction between one of 20 AA residues a_i and a_j at positions i and j in the index TCR and peptide, respectively. Θ_{ij}^{I} describes a sigmoidally decreasing 'switching function' that inversely weights each pairwise interaction based on inter-residue distance. In this model, TCR-peptide accurately assessed in a computationally efficient manner across entire immune repertoires using supervised machine learning to differentiate strong and weak binding pairs, assisting in identifying T cells specific to tumor antigens and enhancing cancer immunotherapy. Of course, the compromise for requiring fewer training sequences is the added requirement of a reasonable structure for the system of interest.

As we mentioned in Section 4, AlphaFold-Multimer (99), developed by DeepMind, can also be categorized as a hybrid model since this approach uses both sequence and structural information in training and predicting steps. AlphaFold-Multimer algorithm consists of two key processing elements, the input derived from MSAs and the evaluation of interatomic distances between AAs within a protein complex structure. A distance matrix provides spatial information for each AA pair, while the MSA aspect preserves and analyzes AA conservation and covariant properties. AlphaFold-Multimer uses the attention-based model to generate models of TCR-pMHC interaction that can be used to distinguish correct peptide epitopes from incorrect ones with substantial accuracy (164). In the future, AlphaFold's ability to predict a collection of key structures could significantly enhance the predictive power of other hybrid approaches that rely on structural templates like RACER.

6 Discussion

This review has presented an overview of recent efforts to predict TCR-pMHC using theoretical, computational, and deep learning approaches, emphasizing both their strengths and limitations. We have explored sequence-based, structure-based, and hybrid methodologies for predicting TCR-pMHC interactions across species, emphasizing the growing importance of these computational techniques within the field. Predicting TCR-pMHC interactions based on AA sequences offers a number of advantages, including leveraging an abundance of publicly available data and using deep learning to extract meaningful features. This representation, however, is also inherently sparse and sample-inefficient, posing challenges. A traditional method of representing AA sequences often fails to encapsulate all essential information, despite the possibility of adding physical descriptors and biological characteristics.

Structure-based models incorporate 3D information crucial for binding and signaling. Nonetheless, challenges arise from the complexity of raw 3D data and the high interdependence of variables within the structure. While graph-based and surface-based representations via Graph Neural Networks and geometric deep learning frameworks have shown promise, they require meticulous model design and implementation, and the invertibility of 2D projections to the original 3D structure is not guaranteed. Hybrid models, combining pairwise models with neural networks, effectively address the issue of higher-order interactions unaccounted for in traditional pairwise models, leading to improved performance in reconstructing pairwise interactions. Hybrid models, despite their ability to handle higher-order interactions, are limited by the requirement for well-defined system structures and extensive sequence data, and their complexity may hinder interpretability and computational efficiency.

With respect to understanding TCR-pMHC specificity, Recent modeling approaches commonly integrate deep neural network

techniques with more traditional methods like cluster analysis. To-date, successful models of TCR-pMHC interactions attempt to deliver on a subset of important objectives, including 1) Computationally efficient characterization for large-scale implementation, 2) Sensitivity in recognizing novel favorable TCR-pMHC pairs, 3) Specificity in predictions through demonstrating the identification of non-recognition pairs, 4) Accurate predictions on data that are far away from the training data, including completely new test TCRs or peptide and 5) Accurate predictions on exhaustive test data that is very close to training examples, including the classification of all point-mutations of a previously identified peptide. At present, no current model adequately addresses all of these objectives. Because of the sheer allowable diversity of TCR and peptide feature space, sparsity in available training data will be a persistent challenge in future applications.

Because of the significant clinical implications of successful models of TCR-pMHC specificity, the number of newly developed approaches is rapidly expanding. As a result, we advocate for standardization in the testing protocols. Because new models are often trained on data that is distinct from that of other previous models, comparative performance is often highly sensitive to the choice of test data. This can artificially enhance the perceived predictive utility of a new model or unreasonably diminish the ability of existing models. Comparative predictive assessments, when performed, should utilize data with neutral similarity to either model. Despite considerable progress in this domain, numerous challenges and future research directions remain. To gain valuable biological insights from TCR-pMHC binding prediction models, current limitations must be addressed and their generalizability, interpretability, and precision must be improved. Enhancing precision involves integrating diverse data modalities and high-quality sources, with special attention given to those reflecting epitope mutations. Improving generalizability entails training models on comprehensive datasets that span both known and novel epitopes, ensuring robustness across varied biological conditions. Crucially, models must be interpreted in a way that translates complex computational outputs into biologically meaningful insights, advancing our understanding of immune responses beyond mere computational contexts. Such targeted improvements will catalyze the development of potent and precise immunotherapies. TCR-pMHC interactions are expected to benefit substantially from new advances in data availability and computational techniques as the availability of high-quality data increases. As a result, innovative therapeutic approaches and tailored medical treatments will be developed based on a deeper understanding of their functions in human health and disease.

In the exhaustive analysis of various methodologies for inferring TCR specificity, our study finds no single superior approach. Rather, we propose a dynamic, integrated strategy that transcends traditional methods and embraces a confluence of techniques while remaining receptive to continual advancements. This multifaceted approach emphasizes the importance of harnessing unlabelled TCR sequences and leveraging data augmentation techniques. It also calls for the integration of both sequence- and structure-aware features, coupled with the application of cutting-edge computational techniques. Furthermore, we underscore the critical need for a collaborative ecosystem that fosters interactions among experts from disparate domains, including immunology, machine learning, and both translational and industrial sectors. Such synergy is pivotal in driving forward-thinking solutions, and we advocate for the unobstructed accessibility of successful models to promote open collaboration and accelerate progress in TCR specificity prediction.

Author contributions

JG supervised the work, ZG performed the review, ZG and JG wrote the paper. All authors contributed to the article and approved the submitted version.

Acknowledgments

JTG was supported by the Cancer Prevention Research Institute of Texas (RR210080). JTG is a CPRIT Scholar in Cancer Research.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Alt FW, Oltz EM, Young F, Gorman J, Taccioli G, Chen J. Vdj recombination. *Immunol Today* (1992) 13(8):306–14. doi: 10.1016/0167-5699(92)90043-7

2. Mora T, Walczak AM. Quantifying lymphocyte receptor diversity. In: *Systems Immunology*. (Boca Raton: CRC Press) (2018). p. 183–98.

3. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, et al. Quantifiable predictive features define epitope-specific t cell receptor repertoires. *Nature* (2017) 547(7661):89–93. doi: 10.1038/nature22383

4. George JT, Kessler DA, Levine H. Effects of thymic selection on T cell recognition of foreign and tumor antigenic peptides. *Proc Natl Acad Sci* (2017) 114(38):E7875–81. doi: 10.1073/pnas.1708573114

5. Morse MA, Gwin IIIWR, Mitchell DA. Vaccine therapies for cancer: then and now. *Targeted Oncol* (2021) 16(2):121–52. doi: 10.1007/s11523-020-00788-w

6. Yee C. Adoptive T cell therapy: addressing challenges in cancer immunotherapy. *J Trans Med* (2005) 3(1):1–8. doi: 10.1186/1479-5876-3-17

7. Klein L, Kyewski B, Allen PM, Hogquist KA. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nat Rev Immunol* (2014) 14 (6):377–91. doi: 10.1038/nri3667

8. Davis MM. Not-so-negative selection. *Immunity* (2015) 43(5):833–5. doi: 10.1016/j.immuni.2015.11.002

9. Slabinski L, Jaroszewski L, Rodrigues APC, Rychlewski L, Wilson IA, Lesley SA, et al. The challenge of protein structure determination—lessons from structural genomics. *Protein Sci* (2007) 16(11):2472–82. doi: 10.1110/ps.073037907

10. Markwick PRL. Théresè Malliavin, and Michael Nilges. Structural biology by nmr: structure, dynamics, and interactions. *PLoS Comput Biol* (2008) 4(9):e1000168. doi: 10.1371/journal.pcbi.1000168

11. Jonic S, Venien-Bryan C. Protein structure determination by electron cryomicroscopy. *Curr Opin Pharmacol* (2009) 9(5):636–42. doi: 10.1016/j.coph.2009.04.006

12. Birnbaum ME, Mendoza JL, Sethi DK, Dong S, Glanville J, Dobbins J, et al. Deconstructing the peptide-mhc specificity of T cell recognition. *Cell* (2014) 157 (5):1073–87. doi: 10.1016/j.cell.2014.03.047

13. Zhang W, Hawkins PG, He J, Gupta NT, Liu J, Choonoo G, et al. A framework for highly multiplexed dextramer mapping and prediction of t cell receptor sequences to antigen specificity. *Sci Adv* (2021) 7(20):eabf5835. doi: 10.1126/sciadv.abf5835

14. Lee CH, Salio M, Napolitani G, Ogg G, Simmons A, Koohy H. Predicting cross-reactivity and antigen specificity of T cell receptors. *Front Immunol* (2020) 11:565096. doi: 10.3389/fimmu.2020.565096

15. Antunes DA, Rigo MaurícioM, Freitas MV, Mendes MFA, Sinigaglia M, Lizée G, et al. Interpreting t-cell cross-reactivity through structure: implications for TCR-based cancer immunotherapy. *Front Immunol* (2017) 8:1210. doi: 10.3389/fimmu.2017.01210

16. Fonseca AF, Antunes DA. Crossdome: an interactive r package to predict cross-reactivity risk using immunopeptidomics databases. *Front Immunol* (2023) 14:1142573. doi: 10.3389/fimmu.2023.1142573

17. Jansson A. Kinetic proofreading and the search for nonself-peptides. *Self/nonself* (2011) 2(1):1–3. doi: 10.4161/self.2.1.15362

18. Gascoigne NRJ, Zal T, Munir Alam S. T-cell receptor binding kinetics in t-cell development and activation. *Expert Rev Mol Med* (2001) 3(6):1–17. doi: 10.1017/S1462399401002502

19. Hwang W, Mallis RJ, Lang MJ, Reinherz EL. The $\alpha\beta$ tcr mechanosensor exploits dynamic ectodomain allostery to optimize its ligand recognition site. *Proc Natl Acad Sci* (2020) 117(35):21336–45. doi: 10.1073/pnas.2005899117

20. Liu B, Chen W, Evavold BD, Zhu C. Antigen-specific tcr–pmhc catch bonds trigger signaling by fast accumulation of force-prolonged bond lifetimes. *Cell* (2014) 157(2):357. doi: 10.1016/j.cell.2014.02.053

21. Franois P, Altan-Bonnet Grégoire. The case for absolute ligand discrimination: modeling information processing and decision by immune t cells. *J Stat Phys* (2016) 162:1130–52. doi: 10.1007/s10955-015-1444-1

22. Liu B, Chen W, Evavold BD, Zhu C. Accumulation of dynamic catch bonds between TCR and agonist peptide-MHC triggers T cell signaling. *Cell* (2014) 157 (2):357–68. doi: 10.1016/j.cell.2014.02.053

23. Altan-Bonnet Grégoire, Germain RN. Modeling T cell antigen discrimination based on feedback control of digital ERK responses. *PLoS Biol* (2005) 3(11):e356. doi: 10.1371/journal.pbio.0030356

24. Teimouri H, Kolomeisky AB. Relaxation times of ligand-receptor complex formation control T cell activation. *Biophys J* (2020) 119(1):182–9. doi: 10.1016/j.bpj.2020.06.002

25. Stone JD, Chervin AS, Kranz DM. T-cell receptor binding affinities and kinetics: impact on t-cell activity and specificity. *Immunology* (2009) 126(2):165–76. doi: 10.1111/j.1365-2567.2008.03015.x

26. McKeithan TW. Kinetic proofreading in t-cell receptor signal transduction. *Proc Natl Acad Sci* (1995) 92(11):5042–6. doi: 10.1073/pnas.92.11.5042

27. Detours V, Mehr R, Perelson AS. A quantitative theory of affinity-driven t cell repertoire selection. *J Theor Biol* (1999) 200(4):389–403. doi: 10.1006/jtbi.1999.1003

28. Chao DL, Davenport MP, Forrest S, Perelson AS. The effects of thymic selection on the range of t cell cross-reactivity. *Eur J Immunol* (2005) 35(12):3452–9. doi: 10.1002/eji.200535098

29. Bauer AL, Beauchemin CAA, Perelson AS. Agent-based modeling of host–pathogen systems: The successes and challenges. *Inf Sci* (2009) 179(10):1379–89. doi: 10.1016/j.ins.2008.11.012

30. Lee HaY, Perelson AS. Computational models of B cell and t cell receptors. *In Silico Immunol* (2007), 65–81. doi: 10.1007/978-0-387-39241-7_5

31. Detours V, Perelson AS. Explaining high alloreactivity as a quantitative consequence of affinity-driven thymocyte selection. *Proc Natl Acad Sci* (1999) 96 (9):5153–8. doi: 10.1073/pnas.96.9.5153

32. Detours V, Mehr R, Perelson AS. Deriving quantitative constraints on T cell selection from data on the mature T cell repertoire. *J Immunol* (2000) 164(1):121–8. doi: 10.4049/jimmunol.164.1.121

33. Košmrlj A, Jha AK, Huseby ES, Kardar M, Chakraborty AK. How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *Proc Natl Acad Sci* (2008) 105(43):16671–6. doi: 10.1073/pnas.0808081105

34. Miyazawa S, Jernigan RL. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* (1996) 256(3):623–44. doi: 10.1006/jmbi.1996.0114

35. Košmrlj A, Chakraborty AK, Kardar M, Shakhnovich EI. Thymic selection of t-cell receptors as an extreme value problem. *Phys Rev Lett* (2009) 103(6):068103. doi: 10.1103/PhysRevLett.103.068103

36. Košmrlj A, Read EL, Qi Y, Allen TM, Altfeld M, Deeks SG, et al. Effects of thymic selection of the t-cell repertoire on hla class i-associated control of HIV infection. *Nature* (2010) 465(7296):350–4. doi: 10.1038/nature08997

37. Chau KNg, George JT, Onuchic Jose´N, Lin X, Levine H. Contact map dependence of a t-cell receptor binding repertoire. *Phys Rev E* (2022) 106(1):014406. doi: 10.1103/PhysRevE.106.014406a

38. Davis MM, Bjorkman PJ. T-cell antigen receptor genes and t-cell recognition. *Nature* (1988) 334(6181):395–402. doi: 10.1038/334395a0

39. Qi Q, Liu Yi, Cheng Y, Glanville J, Zhang D, Lee J-Y, et al. Diversity and clonal selection in the human t-cell repertoire. *Proc Natl Acad Sci* (2014) 111(36):13139–44. doi: 10.1073/pnas.1409155111

40. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The exac browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res* (2017) 45(D1):D840–5. doi: 10.1093/nar/gkw971

41. Capietto Aude-Hélène, Jhunjhunwala S, Pollock SB, Lupardus P, Wong J, Hänsch L, et al. Mutation position is an important determinant for predicting cancer neoantigens. *J Exp Med* (2020) 217(4):315–22. doi: 10.1084/jem.20190179

42. Calis JJA, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, Sette A, et al. Properties of mhc class i presented peptides that enhance immunogenicity. *PLoS Comput Biol* (2013) 9(10):e1003266. doi: 10.1371/journal.pcbi.1003266

43. Kim S, Kim HS, Kim E, Lee MG, Shin E-C, Paik S. Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Ann Oncol* (2018) 29(4):1030–6. doi: 10.1093/annonc/mdy022

44. Wells DK, van Buuren MM, Dang KK, Hubbard-Lucey VM, Sheehan KCF, Campbell KM, et al. Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. *Cell* (2020) 183 (3):818–34. doi: 10.1016/j.cell.2020.09.015

45. Jurtz VI, Jessen LE, Bentzen AK, Jespersen MC, Mahajan S, Vita R, et al. Nettcr: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. *BioRxiv* (2018), 433706. doi: 10.1101/433706

46. Gielis S, Moris P, De Neuter N, Bittremieux W, Ogunjimi B, Laukens K, et al. Tcrex: a webtool for the prediction of t-cell receptor sequence epitope specificity. *BioRxiv* (2018) 373472. doi: 10.1101/373472

47. Jokinen E, Huuhtanen J, Mustjoki S, Heinonen M, Lahdesm¨ aki.¨ H. Predicting recognition between t cell receptors and epitopes with tcrgp. *PLoS Comput Biol* (2021) 17(3):e1008814. doi: 10.1371/journal.pcbi.1008814

48. Cai M, Bang S, Zhang P, Lee H. Atm-tcr: Tcr-epitope binding affinity prediction using a multi-head self-attention model. *Front Immunol* (2022) 13. doi: 10.3389/fimmu.2022.893247

49. Ehrlich R, Kamga L, Gil A, Luzuriaga K, Selin LK, Ghersi D. Swarmtcr: a computational approach to predict the specificity of t cell receptors. *BMC Bioinf* (2021) 22(1):1–14. doi: 10.1186/s12859-021-04335-w

50. Springer I, Tickotsky N, Louzoun Y. Contribution of t cell receptor alpha and beta cdr3, mhc typing, v and j genes to peptide binding prediction. *Front Immunol* (2021) 12:664514. doi: 10.3389/fimmu.2021.664514

51. Lu T, Zhang Ze, Zhu J, Wang Y, Jiang P, Xiao X, et al. Deep learning-based prediction of the t cell receptor–antigen binding specificity. *Nat Mach Intell* (2021) 3 (10):864–75. doi: 10.1038/s42256-021-00383-2

52. Moris P, De Pauw J, Postovskaya A, Gielis S, De Neuter N, Bittremieux W, et al. Current challenges for unseen-epitope tcr interaction prediction and a new perspective derived from image classification. *Briefings Bioinf* (2021) 22(4):bbaa318. doi: 10.1093/bib/bbaa318

53. Weber A, Born J, Rodriguez Martínez Mar´ıa. Titan: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics* (2021) 37(Supplement 1): i237–44. doi: 10.1093/bioinformatics/btab294

54. Jokinen E, Dumitrescu A, Huuhtanen J, Gligorijevic V, Mustjoki S, Bonneau R, et al. Tcrconv: predicting recognition between t cell receptors and epitopes using contextualized motifs. *Bioinformatics* (2023) 39(1):btac788. doi: 10.1093/bioinformatics/btac788

55. Gainza P, Sverrisson F, Monti F, Rodola E, Boscaini D, Bronstein MM, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* (2020) 17(2):184–92. doi: 10.1038/s41592-019-0666-6

56. Sverrisson F, Feydy J, Correia BE, Bronstein MM. (2021). Fast end-to-end learning on protein surfaces, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (Virtual: IEEE). pp. 15272–81.

57. Breiman L. Random forests. *Mach Learn* (2001) 45:5–32. doi: 10.1023/A:1010933404324

58. Nguyen Pham M-D, Nguyen T-N, Tran LeS, Bui Nguyen Q-T, Nguyen TH, Quynh Pham TM, et al. epitcr: a highly sensitive predictor for tcr–peptide binding. *Bioinformatics* (2023) 39(5):btad284. doi: 10.1093/bioinformatics/btad284

59. Dorigatti E, Drost F, Straub A, Hilgendorf P, Wagner KI, Bischl B, et al. Predicting T cell receptor functionality against mutant epitopes. *bioRxiv* (2023), 2023–05. doi: 10.1101/2023.05.10.540189

60. Hochreiter S, Schmidhuber Jurgen¨. Long short-term memory. *Neural Comput* (1997) 9(8):1735–1780. doi: 10.1162/neco.1997.9.8.1735

61. Ng A, et al. Sparse autoencoder. *CS294A Lecture Notes* (2011) 72(2011):1–19. Available at: https://web.stanford.edu/class/cs294a/sparseAutoencoder.pdf.

62. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst* (2017) 30:5998–6008. doi: 10.5555/3295222.3295349

63. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *J Big Data* (2016) 3(1):1–40. doi: 10.1186/s40537-016-0043-6

64. Meysman P, Barton J, Bravi B, Cohen-Lavi L, Karnaukhov V, Lilleskov E, et al. Benchmarking solutions to the t-cell receptor epitope prediction problem: Immrep22 workshop report. *ImmunoInformatics* (2023) 9:100024. doi: 10.1016/j.immuno.2023.100024

65. Dens C, Bittremieux W, Affaticati F, Laukens K, Meysman P. Interpretable deep learning to uncover the molecular binding patterns determining tcr–epitope interactions. *bioRxiv* (2022), 2022–05. doi: 10.1101/2022.05.02.490264

66. Pavlovic M, Scheffer L, Motwani K, Kanduri C, Kompova R, Vazov N, et al. The immuneml ecosystem for machine learning analysis of adaptive immune receptor repertoires. *Nat Mach Intell* (2021) 3(11):936–44. doi: 10.1038/s42256-021-00413-z

67. Widrich M, Schafl B, Pavlovic M, Kjetil Sandve G, Hochreiter S, Greiff V, et al. Deeprc: immune repertoire classification with attention-based deep massive multiple instance learning. *BioRxiv* (2020) 038158:2020. doi: 10.1101/2020.04.12.038158

68. Huang H, Wang C, Rubelt F, Scriba TJ, Davis MM. Analyzing the mycobacterium tuberculosis immune response by t-cell receptor clustering with gliph2 and genome-wide antigen screening. *Nat Biotechnol* (2020) 38(10):1194–202. doi: 10.1038/s41587-020-0505-4

69. Valkiers S, de Vrij N, Gielis S, Verbandt S, Ogunjimi B, Laukens K, et al. Recent advances in t-cell receptor repertoire analysis: bridging the gap with multimodal single-cell rna sequencing. *ImmunoInformatics* (2022) 5:100009. doi: 10.1016/j.immuno.2022.100009

70. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* (2015) 33(5):495–502. doi: 10.1038/nbt.3192

71. Zappia L, Theis FJ. Over 1000 tools reveal trends in the single-cell rna-seq analysis landscape. *Genome Biol* (2021) 22:1–18. doi: 10.1186/s13059-021-02519-4

72. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. *Cell* (2021) 184(13):3573–87. doi: 10.1016/j.cell.2021.04.048

73. Picot J, Guerin CL, Le Van Kim C, Boulanger CM. Flow cytometry: retrospective, fundamentals and recent instrumentation. *Cytotechnology* (2012) 64:109–30. doi: 10.1007/s10616-011-9415-0

74. Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, Lou X, et al. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Analytical Chem* (2009) 81(16):6813–22. doi: 10.1021/ac901049w

75. Deng L, Ly C, Abdollahi S, Zhao Yu, Prinz I, Bonn S. Performance comparison of tcr-pmhc prediction tools reveals a strong data dependency. *Front Immunol* (2023) 14:1128326. doi: 10.3389/fimmu.2023.1128326

76. Gao Y, Gao Y, Fan Y, Zhu C, Wei Z, Zhou C, et al. Pan-peptide meta learning for t-cell receptor–antigen binding recognition. *Nat Mach Intell* (2023) 5(3):236–49. doi: 10.1038/s42256-023-00619-3

77. Graves A, Wayne G, Danihelka I. Neural turing machines. *arXiv preprint arXiv* (2014) 1410:5401. doi: 10.48550/arXiv.1410.5401

78. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, et al. Identifying specificity groups in the t cell receptor repertoire. *Nature* (2017) 547(7661):94–8. doi: 10.1038/nature22976

79. Valkiers S, Van Houcke M, Laukens K, Meysman P. Clustcr: a python interface for rapid clustering of large sets of cdr3 sequences with unknown antigen specificity. *Bioinformatics* (2021) 37(24):4865–7. doi: 10.1093/bioinformatics/btab446

80. Zhang H, Zhan X, Li Bo. Giana allows computationally-efficient tcr clustering and multi-disease repertoire classification by isometric transformation. *Nat Commun* (2021) 12(1):4699. doi: 10.1038/s41467-021-25006-7

81. Zhang H, Liu L, Zhang J, Chen J, Ye J, Shukla S, et al. Investigation of antigen-specific t-cell receptor clusters in human cancerstumor-infiltrating antigen-specific tcr clusters. *Clin Cancer Res* (2020) 26(6):1359–71. doi: 10.1158/1078-0432.CCR-19-3249

82. Reynolds DA, et al. Gaussian mixture models. *Encyclopedia biometrics* (2009) 741:659–63. doi: 10.1007/978-0-387-73003-5_196

83. Ester M, Kriegel H-P, Sander Jörg, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd* (1996) 96:226–31. doi: 10.5555/3001460.3001507

84. Arthur D, Vassilvitskii S. (2007). K-means++ the advantages of careful seeding, in: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, (SIAM, Philadelphia: Discrete Algorithms). pp. 1027–35.

85. Kumar Saini S, Hersby DS, Tamhane T, Povlsen HR, Hernandez SPA, Nielsen M, et al. Sars-cov-2 genome-wide t cell epitope mapping reveals immunodominance and substantial cd8+ t cell activation in covid-19 patients. *Sci Immunol* (2021) 6(58):eabf7550. doi: 10.1126/sciimmunol.abf7550

86. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, ZecChina R, et al. Protein 3d structure computed from evolutionary sequence variation. *PloS One* (2011) 6(12):e28766. doi: 10.1371/journal.pone.0028766

87. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci* (2011) 108(49):E1293–301. doi: 10.1073/pnas.1111471108

88. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci* (2009) 106(1):67–72. doi: 10.1073/pnas.0805923106

89. Balakrishnan S, Kamisetty H, Carbonell JG, Lee S-I, Langmead CJ. Learning generative models for protein fold families. *Proteins: Structure Function Bioinf* (2011) 79(4):1061–78. doi: 10.1002/prot.22934

90. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolutionbased residue–residue contact predictions in a sequence-and structure-rich era. *Proc Natl Acad Sci* (2013) 110(39):15674–9. doi: 10.1073/pnas.1314045110

91. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Phys Rev E* (2013) 87 (1):012707. doi: 10.1103/PhysRevE.87.012707

92. Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. *Electronic Markets* (2021) 31(3):685–95. doi: 10.1007/s12525-021-00475-2

93. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv* (2013) 1301:3781. doi: 10.48550/arXiv.1301.3781

94. Asgari E, McHardy AC, Mofrad MRK. Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (dimotif) and sequence embedding (protvecx). *Sci Rep* (2019) 9(1):1–16. doi: 10.1038/s41598-019-38746-w

95. Zhang P, Bang S, Cai M, Lee H. Context-aware amino acid embedding advances analysis of tcr-epitope interactions. *bioRxiv* (2023), 2023–04. doi: 10.1101/2023.04.12.536635

96. Medsker LR, Jain LC. Recurrent neural networks. *Design Appl* (2001) 5:64–7.

97. Fang Y, Liu X, Liu H. Attention-aware contrastive learning for predicting t cell receptor–antigen binding specificity. *Briefings Bioinf* (2022) 23(6):bbac378. doi: 10.1093/bib/bbac378

98. Lin P, Yan Y, Huang S-Y. Deephomo2. 0: improved protein–protein contact prediction of homodimers by transformer-enhanced deep learning. *Briefings Bioinf* (2023) 24(1):bbac499. doi: 10.1093/bib/bbac499

99. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with alphafold-multimer. *BioRxiv* (2021), 2021–10. doi: 10.1101/2021.10.04.463034

100. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv* (2014) 1409:0473. doi: 10.48550/arXiv.1409.0473

101. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with alphafold. *Nature* (2021) 596 (7873):583–9. doi: 10.1038/s41586-021-03819-2

102. Kingma DP, Welling M. Auto-encoding variational bayes. *arXiv preprint arXiv* (2013) 1312:6114. doi: 10.48550/arXiv.1312.6114

103. Bell DR, Domeniconi G, Yang C-C, Zhou R, Zhang L, Cong G. Dynamics-based peptide–mhc binding optimization by a convolutional variational autoencoder: A use-case model for castelo. *J Chem Theory Comput* (2021) 17(12):7962–71. doi: 10.1021/acs.jctc.1c00870

104. Peng X, Lei Y, Feng P, Jia L, Ma J, Zhao D, et al. Characterizing the interaction conformation between t-cell receptors and epitopes with deep learning. *Nat Mach Intell* (2023) 5:1–13. doi: 10.1038/s42256-023-00634-4

105. Dens C, Laukens K, Bittremieux W, Meysman P. The pitfalls of negative data bias for the t-cell epitope specificity challenge. *bioRxiv* (2023), 2023–04. doi: 10.1101/2023.04.06.535863

106. Gao Y, Dong K, Wu S, Liu Q. Reply to: The pitfalls of negative data bias for the t-cell epitope specificity challenge. *bioRxiv* (2023) 2023–04. doi: 10.1101/2023.04.07.535967

107. Cárdenas C, Bidon-Chanal A, Conejeros P, Arenas G, Marshall S, Luque FJ. Molecular modeling of class i and ii alleles of the major histocompatibility complex in salmo salar. *J Computer-Aided Mol Design* (2010) 24:1035–51. doi: 10.1007/s10822-010-9387-8

108. Shirts MR, Mobley DL, Chodera JD. Alchemical free energy calculations: ready for prime time? *Annu Rep Comput Chem* (2007) 3:41–59. doi: 10.1016/S1574-1400(07)03004-6

109. Baker BM, Scott DR, Blevins SJ, Hawse WF. Structural and dynamic control of t-cell receptor specificity, cross-reactivity, and binding mechanism. *Immunol Rev* (2012) 250(1):10–31. doi: 10.1111/j.1600-065X.2012.01165.x

110. Borbulevych OY, Piepenbrink KH, Gloor BE, Scott DR, Sommese RF, Cole DK, et al. T cell receptor cross-reactivity directed by antigendependent tuning of peptide-mhc molecular flexibility. *Immunity* (2009) 31(6):885–96. doi: 10.1016/j.immuni.2009.11.003

111. Sousa SF, Fernandes PA, Ramos MJ. Protein–ligand docking: current status and future challenges. *Proteins: Structure Function Bioinf* (2006) 65(1):15–26. doi: 10.1002/prot.21082

112. Anton Feenstra K, Hess B, Berendsen HJC. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J Comput Chem* (1999) 20(8):786–98. doi: 10.1002/(SICI)1096-987X(199906)20:8<786::AID-JCC5>3.0.CO;2-B

113. Hess B, Kutzner C, van der Spoel D, Lindahl E. Gromacs 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* (2008) 4(3):435–47. doi: 10.1021/ct700301q

114. Minary P, Levitt M. Probing protein fold space with a simplified model. *J Mol Biol* (2008) 375(4):920–33. doi: 10.1016/j.jmb.2007.10.087

115. Sim AYL, Levitt M, Minary P. Modeling and design by hierarchical natural moves. *Proc Natl Acad Sci* (2012) 109(8):2890–5. doi: 10.1073/pnas.1119918109

116. Dunbar J, Fuchs A, Shi J, Deane CM. Abangle: characterising the VH–VL orientation in antibodies. *Protein Engineering Design Selection* (2013) 26(10):611–20. doi: 10.1093/protein/gzt020

117. Knapp B, Demharter S, Esmaielbeiki R, Deane CM. Current status and future challenges in t-cell receptor/peptide/mhc molecular dynamics simulations. *Briefings Bioinf* (2015) 16(6):1035–44. doi: 10.1093/bib/bbv005

118. Antunes DA, Abella JR, Devaurs D, Rigo Maur´ıcioM, Kavraki LE. Structure-based methods for binding mode and binding affinity prediction for peptide-mhc complexes. *Curr topics medicinal Chem* (2018) 18(26):2239–55. doi: 10.2174/1568026619666181224101744

119. Noé F, Tkatchenko A, Müller K-R, Clementi C. Machine learning for molecular simulation. *Annu Rev Phys Chem* (2020) 71:361–90. doi: 10.1146/annurev-physchem-042018-052331

120. Somnath VR, Bunne C, Krause A. Multi-scale representation learning on proteins. *Adv Neural Inf Process Syst* (2021) 34:25244–55. doi: 10.48550/arXiv.2204.02337

121. Jing B, Eismann S, Suriana P, Townshend RJL, Dror R. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv* (2020) 2009:01411. doi: 10.48550/arXiv.2009.01411

122. Dai B, Bailey-Kellogg C. Protein interaction interface region prediction by geometric deep learning. *Bioinformatics* (2021) 37(17):2580–8. doi: 10.1093/bioinformatics/btab154

123. Gligorijević V, Renfrew PD, Kosciolek T, Koehler Leman J, Berenberg D, Vatanen T, et al. Structure-based protein function prediction using graph convolutional networks. *Nat Commun* (2021) 12(1):1–14. doi: 10.1038/s41467-021-23303-9

124. Yin R, Feng BY, Varshney A, Pierce BG. Benchmarking alphafold for protein complex modeling reveals accuracy determinants. *Protein Sci* (2022) 31(8):e4379. doi: 10.1002/pro.4379

125. Bradley P. Structure-based prediction of t cell receptor: peptide-mhc interactions. *eLife* (2023) 12:e82813. doi: 10.7554/eLife.82813

126. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. *Nature* (2020) 577 (7792):706–10. doi: 10.1038/s41586-019-1923-7

127. Ward JP, Gubin MM, Schreiber RD. The role of neoantigens in naturally occurring and therapeutically induced immune responses to cancer. *Adv Immunol* (2016) 130:25–74. doi: 10.1016/bs.ai.2016.01.001

128. Arel I, Rose DC, Karnowski TP. Deep machine learning-a new frontier in artificial intelligence research. *IEEE Comput Intell magazine* (2010) 5(4):13–8. doi: 10.1109/MCI.2010.938364

129. LeCun Y, Bengio Y, Hinton G. Deep learning. *nature* (2015) 521(7553):436–44. doi: 10.1038/nature14539

130. Xie Z, Deng X, Shu K. Prediction of protein–protein interaction sites using convolutional neural network and improved data sets. *Int J Mol Sci* (2020) 21(2):467. doi: 10.3390/ijms21020467

131. He K, Zhang X, Ren S, Sun J. (2016). Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, IEEE, Las Vegas, NV: CVPR. pp. 770–8.

132. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Protein structure prediction using multiple deep neural networks in the 13th critical assessment of protein structure prediction (casp13). *Proteins: Structure Function Bioinf* (2019) 87 (12):1141–8. doi: 10.1002/prot.25834

133. Jiménez José, Doerr S, Mart´ınez-Rosell G, Rose AS, De Fabritiis G. Deepsite: protein-binding site predictor using 3d-convolutional neural networks. *Bioinformatics* (2017) 33(19):3036–42. doi: 10.1093/bioinformatics/btx350

134. Krivak´ R, Hoksza D. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J cheminformatics* (2018) 10:1–12. doi: 10.1186/s13321-018-0285-8

135. Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Improving detection of protein-ligand binding sites with 3d segmentation. *Sci Rep* (2020) 10 (1):1–9. doi: 10.1038/s41598-020-61860-z

136. Bi J, Zheng Y, Wang C, Ding Y. An attention based bidirectional lstm method to predict the binding of tcr and epitope. *IEEE/ACM Trans Comput Biol Bioinf* (2021) 19(6):3272–80. doi: 10.1109/TCBB.2021.3115353

137. Springer I, Besser H, Tickotsky-Moskovitz N, Dvorkin S, Louzoun Y. Prediction of specific tcr-peptide binding from large dictionaries of tcr-peptide pairs. *Front Immunol* (2020) 11:1803. doi: 10.3389/fimmu.2020.01803

138. Pittala S, Bailey-Kellogg C. Learning context-aware structural representations to predict antigen and antibody binding interfaces. *Bioinformatics* (2020) 36(13):3996–4003. doi: 10.1093/bioinformatics/btaa263

139. Wang Y-B, You Z-H, Li X, Jiang T-H, Chen X, Zhou Xi, et al. Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. *Mol Biosyst* (2017) 13(7):1336–44. doi: 10.1039/C7MB00188F

140. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. *J Am Stat Assoc* (2017) 112(518):859–77. doi: 10.1080/01621459.2017.1285773

141. Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: An overview. *IEEE Signal Process magazine* (2018) 35 (1):53–65. doi: 10.1109/MSP.2017.2765202

142. Zhu XJ. Semi-supervised learning literature survey. *University of Wisconsin Madison Department of Computer Sciences* (2005). Available at: http://digital.library.wisc.edu/1793/60444.

143. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. *Commun ACM* (2020) 63(11):139–44. doi: 10.1145/3422622

144. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of wasserstein gans. *Adv Neural Inf Process Syst* (2017) 30:5767–77. doi: 10.48550/arXiv.1704.00028

145. Arjovsky M, Chintala S, Bottou Leon´. Wasserstein gan. *International Conference on Machine Learning* (2017). doi: 10.48550/arXiv.1701.07875

146. Farajzadeh-Zanjani M, Razavi-Far R, Saif M, Palade V. Generative adversarial networks: a survey on training, variants, and applications. In: *Generative Adversarial Learning: Architectures and Applications*. (Cham, Switzerland: Springer International) (2022). p. 7–29.

147. Anand N, Eguchi R, Huang P-S. Fully differentiable full-atom protein backbone generation. *International Conference on Learning Representations (ICLR)* (2019). Available at: https:// openreview.net/revisions?id=SJxnVL8YOV.

148. Killoran N, Lee LJ, Delong A, Duvenaud D, Frey BJ. Generating and designing dna with deep generative models. *arXiv preprint arXiv* (2017) 1712:06148. doi: 10.48550/arXiv.1712.06148

149. Mescheder L, Geiger A. Sebastian Nowozin. Which training methods for gans do actually converge? In: *International conference on machine learning*. PMLR, Stockholm, Sweden (2018). p. 3481–90.

150. Roth K, Lucchi A, Nowozin S, Hofmann T. Stabilizing training of generative adversarial networks through regularization *Adv Neural Inf Process Syst*. (2017) 30:2018–28. doi: 10.48550/arXiv.1705.09367

151. Mirza M, Osindero S. Conditional generative adversarial nets. *arXiv preprint arXiv* (2014) 1411:1784. doi: 10.48550/arXiv.1411.1784

152. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv* (2015) 1511:06434. doi: 10.48550/arXiv.1511.06434

153. Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier gans. In: *International conference on machine learning*. New York, NY: ACM Digital Library (2017). p. 2642–51.

154. Anand N, Eguchi R, Mathews II, Perez CP, Derry A, Altman RB, et al. Protein sequence design with a learned potential. *Nat Commun* (2022) 13(1):1–11. doi: 10.1038/s41467-022-28313-9

155. Castro E, Godavarthi A, Rubinfien J, Givechian KB, Bhaskar D, Krishnaswamy S. Guided generative protein design using regularized transformers. *arXiv preprint arXiv* (2022) 2201:09948. doi: 10.1038/s42256-022-00532-1

156. Gao Z, Tan C, Li S. Alphadesign: A graph protein design method and benchmark on alphafolddb. *arXiv preprint arXiv* (2022) 2202:01079. doi: 10.48550/arXiv.2202.01079a

157. Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics. In: *International Conference on Machine Learning*. Lille: PMLR (2015). p. 2256–65.

158. Anand N, Huang P. Generative modeling for protein structures. *Adv Neural Inf Process Syst* (2018) 31:7505–16. Available at: http://papers.nips.cc/ paper/7978-generative-modeling-for-protein-structures.pdf.

159. Diaconis P. Group representations in probability and statistics. *Lecture notes-monograph Ser* (1988) 11:i–192. doi: 10.1214/lnms/1215467407

160. Chau Nguyen H, ZecChina R, Berg J. Inverse statistical problems: from the inverse ising problem to data science. *Adv Phys* (2017) 66(3):197–261. doi: 10.1080/00018732.2017.1341604

161. Lin X, George JT, Schafer NP, Ng Chau K, Birnbaum ME, Clementi C, et al. Rapid assessment of t-cell receptor specificity of the immune repertoire. *Nat Comput Sci* (2021) 1(5):362–73. doi: 10.1038/s43588-021-00076-1

162. Wang A, Lin X, Chau KN, Onuchic JN, Levine H, George JT, et al. RACER-m leverages structural features for sparse T Cell specificity prediction. *(Cold Spring Harbor Laboratory) bioRxiv* (2023) 2023–08.

163. Davtyan A, Schafer NP, Zheng W, Clementi C, Wolynes PG, Papoian GA. Awsem-md: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J Phys Chem B* (2012) 116(29):8494–503. doi: 10.1021/jp212541y

164. Abdin O, Nim S, Wen H, Kim PM. Pepnn: a deep attention model for the identification of peptide binding sites. *Commun Biol* (2022) 5(1):503. doi: 10.1038/s42003-022-03445-2