ANNUAL REVIEWS

Annual Review of Chemical and Biomolecular Engineering

Applications of Machine and Deep Learning in Adaptive Immunity

Margarita Pertseva,[1,2] Beichen Gao,[1] Daniel Neumeier,[1] Alexander Yermanos,[1,3,4] and Sai T. Reddy[1]

[1]Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland; email: sai.reddy@ethz.ch

[2]Life Science Zurich Graduate School, ETH Zurich and University of Zurich, 8006 Zurich, Switzerland

[3]Department of Pathology and Immunology, University of Geneva, 1205 Geneva, Switzerland

[4]Department of Biology, Institute of Microbiology and Immunology, ETH Zurich, 8093 Zurich, Switzerland

Annu. Rev. Chem. Biomol. Eng. 2021. 12:39–62

First published as a Review in Advance on April 14, 2021

The *Annual Review of Chemical and Biomolecular Engineering* is online at chembioeng.annualreviews.org

https://doi.org/10.1146/annurev-chembioeng-101420-125021

Copyright © 2021 by Annual Reviews. All rights reserved

ANNUAL REVIEWS CONNECT

www.annualreviews.org

- Download figures
- Navigate cited references
- Keyword search
- Explore related articles
- Share via email or social media

Keywords

immune repertoire, T cell receptor, TCR, B cell receptor, BCR, major histocompatibility complex, MHC, neural networks, machine learning, deep learning

Abstract

Adaptive immunity is mediated by lymphocyte B and T cells, which respectively express a vast and diverse repertoire of B cell and T cell receptors and, in conjunction with peptide antigen presentation through major histocompatibility complexes (MHCs), can recognize and respond to pathogens and diseased cells. In recent years, advances in deep sequencing have led to a massive increase in the amount of adaptive immune receptor repertoire data; additionally, proteomics techniques have led to a wealth of data on peptide–MHC presentation. These large-scale data sets are now making it possible to train machine and deep learning models, which can be used to identify complex and high-dimensional patterns in immune repertoires. This article introduces adaptive immune repertoires and machine and deep learning related to biological sequence data and then summarizes the many applications in this field, which span from predicting the immunological status of a host to the antigen specificity of individual receptors and the engineering of immunotherapeutics.

INTRODUCTION

Immune Receptors in Adaptive Immunity

The ability of the adaptive immune system to recognize foreign pathogens and diseased cells is driven by lymphocytes: B and T cells, which can identify specific molecular structures (antigens or epitopes) on foreign pathogens. Specificity to these antigenic epitopes is achieved through a group of adaptive immune receptors belonging to the immunoglobulin superfamily: B cell receptors (BCRs) and their secreted version antibodies and T cell receptors (TCRs). BCRs and TCRs cover a large functional sequence space, enabling them to recognize the myriad of different pathogens and antigenic determinants (epitopes) an individual gets exposed to throughout life. Because these receptors belong to the immunoglobulin superfamily, both constitute a disulfide bond–linked heterodimer of either typically a heavy and light chain (BCR) or an α and β chain (in some cases γ and δ chain; TCR) and are thus structurally related. BCRs bind to antigen directly, whereas TCRs recognize antigen presented as peptides on major histocompatibility complexes (MHCs).

Adaptive Immune Receptor Repertoires

For both BCRs and TCRs, amino acid sequence diversity is largely confined to variable region domains, which consequently drives their antigen recognition. Owing to their common structural protein ancestry, primary variable region receptor diversity results from the imprecise somatic recombination [also called V(D)J recombination] of distinct germline gene segments by the enzyme RAG recombinase, which joins different variable (V), diversity (D), and joining (J) segments together during the cellular development of lymphocytes (1–3) (**Figure 1**). For each segment, multiple variants/alleles exist in the respective genomic loci. The variable regions of heavy and β chains are formed by recombining a V, D, and J segment, whereas light and α chains use only a V and J segment (**Figure 1**). It is estimated that these recombination events in humans can generate a theoretical diversity of 5×10^{13} naïve BCRs and 10^{18} α:β TCRs (4). In addition to V(D)J recombination, which takes place in the bone marrow and the thymus, B cells, unlike T cells, can further undergo secondary diversification in the peripheral lymphoid tissues (primarily spleen and lymph nodes) through somatic hypermutation (5, 6) and class-switch recombination (7), which represent highly regulated enzymatic processes that lead to BCRs with higher affinity to target antigen (8). In most cases, B and T cell clones are defined based on the site of junctional recombination known as complementarity determining region 3 (CDR3); CDR3 of the heavy (CDRH3) and β (CDRβ3) chains are the main paratope components, which means they contribute to much of the binding specificity of BCRs and TCRs, respectively. The clonal population of lymphocytes and their recombined BCRs and TCRs in an individual represents their immune repertoire. Recent developments in deep sequencing technologies have enabled unprecedented insight into the diversity and distribution of immune repertoires. Immune repertoire sequencing provides the ability to monitor the dynamic changes in the immune repertoire landscape, ranging from clonal expansion to germline recombination and somatic hypermutation (**Figure 1**).

B cell receptors and antibody repertoires. Unlike traditional methods of antibody analysis (i.e., serological binding assays), targeted deep sequencing of BCR variable regions in combination with a carefully chosen experimental design (9) can capture a wealth of quantitative information on repertoires, including clonal selection and expansion, clonal diversity, clonal convergence, and clonal evolution via somatic hypermutation. BCR repertoire sequencing has been used to shed light on basic questions in immunobiology and development across various species (10–12). Furthermore, BCR repertoire sequencing has also been used for medical and biotechnological

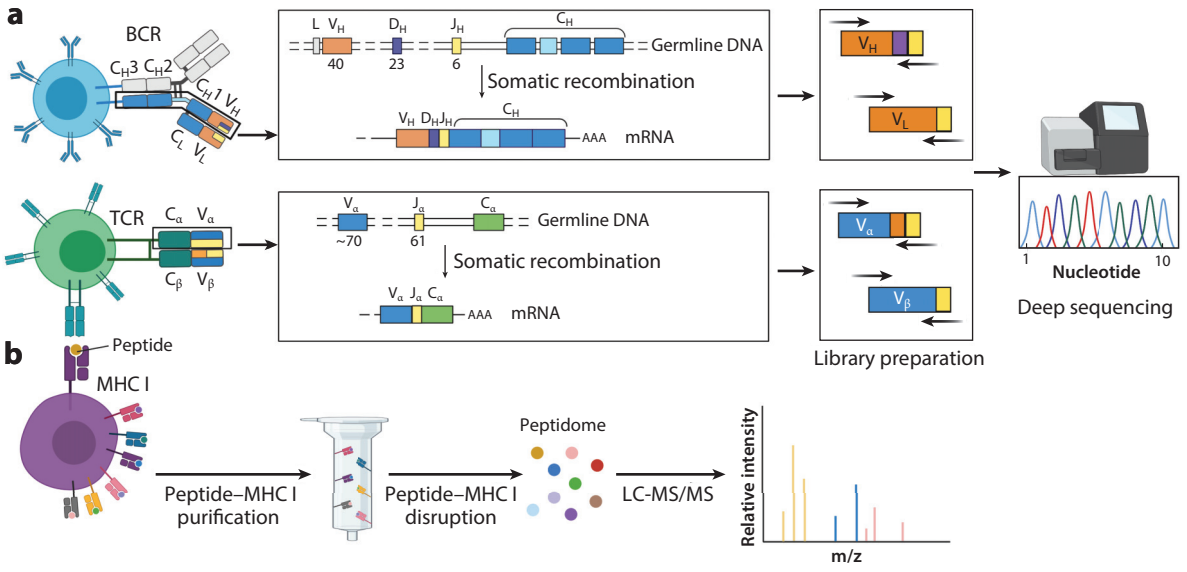

Figure 1

The immune receptor repertoire and immune peptidome. (*a*) B cell receptors (BCRs) and T cell receptors (TCRs) generate diversity in their variable regions via the process of V(D)J recombination, which assembles different variable (V), diversity (D), and joining (J) segments together to create distinct variable region combinations, which are then followed by the constant regions (C_L/C_H1–3) that comprise the rest of the receptor. In BCR variable heavy (V_H) and TCR variable beta (V_β) chains, all three V, D, and J genes are used, whereas in the BCR variable light (V_L) and TCR variable alpha (V_α) chains, only V and J genes are used. Library preparation from recombined genomic DNA or mRNA of variable regions is coupled to deep sequencing to enable quantitative analysis of the immune receptor repertoire. Key interrogated features relate to clonal selection and expansion, clonal evolution by somatic hypermutation, germline recombination, repertoire architecture, and more. (*b*) Major histocompatibility complexes (MHCs) present peptides to TCRs on T cells. Each cell carries multiple MHC alleles on its surface, and each allele presents multiple peptides forming a cell peptidome—a sum of all peptides presented on a surface by MHCs. Peptidomes can be obtained by purifying peptide–MHC complexes with subsequent peptide elution. The resulting peptide pool sequences then can be identified via mass-spectrometry methods (i.e., liquid chromatography-mass spectrometry, LC-MS/MS). Figure adapted from Cell Receptors (Alpha and Beta Chains) and B Cell Receptors (Light and Heavy Chains) by BioRender.com (2020), retrieved from **https://app.biorender.com/biorender-templates**.

purposes, such as vaccine profiling (13–15) and discovery of monoclonal antibodies (16, 17). Substantial recent progress in the field of droplet microfluidics has enabled single-cell sequencing of lymphocytes, therefore providing detailed insights into the landscape of natively paired heavy-/light-chain BCR repertoires (18) and their associated phenotypic properties. Several sophisticated strategies have been devised to directly link paired BCR or TCR sequences to antigen specificity (19–22).

T cell receptor repertoires. Similar to that of BCRs, deep sequencing of TCR repertoires provides a quantitative framework to understand and harness this information to address questions in fundamental immunology, as well as applications in molecular diagnostics and immunotherapeutics. TCR repertoire sequencing has been instrumental for profiling clonal selection across a variety of T cell populations, including effector, memory, and exhausted cytotoxic $CD8^+$ T cells and a wide range of helper $CD4^+$ T cell subsets (Th1, Th2, Th17, T follicular helper cells) (23–26). These studies have been pivotal in understanding how clones either can coexist or are exclusive to certain T cell subsets, thereby informing selective pressures shaping cell-mediated adaptive

immunity. In addition to profiling distinct T cell subsets, new quantitative insight on T cell responses during various infection, immunization, and disease conditions has been obtained from TCR repertoire sequencing. In most of these studies, TCR repertoire parameters are described based on clonal expansion, germline gene usage, and a range of sequence and biophysical properties (polarity, acidity, and sequence length) (25–28). TCR repertoire sequencing has similarly demonstrated a usefulness in more applied settings, with substantial promise in T cell engineering, immunodiagnostics, and TCR discovery (29–32).

Immune repertoire databases. To organize the output of immune repertoire sequencing experiments, the Adaptive Immune Receptor Repertoire (AIRR) community was founded to bring together academic researchers, industry partners, data experts, and others to manage immune repertoire sequencing data in a standardized fashion (33). The AIRR community defined minimal experimental information guidelines for data set annotation, such as V(D)J usage, species, diagnosis, and standard file formats for the annotated data, thus supporting high workflow reproducibility and greater ease for meta-analysis across data sets. The AIRR Data Commons API is also available for easy access, querying, and implementation across several immune repertoire sequencing data repositories (34). The current main repositories are the iReceptor and Observed Antibody Space (OAS) databases, the latter of which is maintained by the Oxford Protein Informatics Group, who have also established the Structural T-Cell Receptor Database for curated sets of TCR sequences and their confirmed structures (35–37). Other independent repositories of interest include VDJServer, a platform that offers a complete analysis workflow for preprocessing, annotation, and characterization of BCR sequences, and the Pan Immune Repertoire Database (PIRD), which collects annotated TCR and BCR sequences from the China National GeneBank (38, 39). For further databases of interest, the AntiBodies Chemically Defined (ABCD) database offers manually curated sequences of antibodies and their known targets; the VDJdb aggregates antigen specificities of TCR sequences from published T cell specificity assays as well as TCR motifs to be used in specificity prediction; and the PIRD contains a database of TCRs and BCRs with confirmed specificity toward specific antigens or diseases (39–42).

Major Histocompatibility Complexes and Antigen Presentation

MHC I and MHC II molecules are located on cell surfaces and present peptides to cytotoxic $CD8^+$ T cells and helper $CD4^+$ T cells, respectively. Both MHC types share similar structure; however, the MHC I groove is closed and accommodates only short peptides (8–10 amino acids in length), whereas the open groove of MHC II binds longer peptides (from 13 to 25 amino acids) (43, 44). MHC I is expressed on every nucleated cell and presents fragments of intracellular proteins. In contrast, MHC II expression is restricted to antigen-presenting cells such as dendritic cells, macrophages, and B cells, which take up extracellular proteins by phagocytosis or endocytosis and present their peptide fragments to T cells through the MHC II pathway. The difference in the source of the peptides is reflected in MHC I and MHC II pathway dissimilarities that are described below.

In brief, the classical MHC I pathway starts with the degradation of defective cytosolic and nuclear proteins by the proteasome protein complex (4). Resulting peptides are released back to the cytosol, where some of them are translocated to the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP) complex. In the ER, high-affinity peptides are bound to MHC I molecules prefolded on the ER membrane (4). Once the peptide–MHC I complex is formed, it is transported to a cell surface through the Golgi apparatus. The MHC II pathway is initiated when antigen-presenting cells uptake extracellular proteins into endosomes,

where they are degraded by pH-dependent proteases. MHC II molecules are preassembled in the ER and, together with the other proteins necessary for peptide loading, are transported to endosomes (4). In endosomes, MHC II is loaded with digested protein fragments and then transferred to the surface to present its ligands to $CD4^+$ T cells.

Importantly, MHC I and II molecules are both polygenic and polymorphic, meaning cells express multiple MHC alleles, and thousands of alleles are found in the human population. Moreover, MHC molecules are extremely promiscuous and, in theory, can bind more than a million different peptides (45). Such complexity of peptide presentation by MHC has thus made it challenging to obtain comprehensive peptide–MHC data sets and predict peptide presentation on MHC.

There are two primary ways to characterize peptide presentation on MHC, either by measuring the affinity of the interaction of synthetic peptide–MHC or by mass spectrometry (MS) from cells expressing peptide–MHC (**Figure 1**). Full details on these methods and emerging technologies based on cellular display of recombinant peptide–MHC libraries can be found in Joglekar & Li's recent review (46). Regardless of the method used, the vast majority of data from experimental peptide–MHC binding assays are collected and stored in the Immune Epitope Database (47). As of July 2020, this database contains in total ~1,630,000 MHC ligand data points, from which ~1,300,000 were identified via MS and more than 300,000 were obtained from MHC binding assays.

MACHINE AND DEEP LEARNING ON BIOLOGICAL SEQUENCES

Introduction to Machine Learning

Machine learning (ML) is a category of algorithms that identify and learn patterns in data without being explicitly programmed. ML and statistics are closely related fields, and the difference between them is often quite subtle. Statistics is focused more on making conclusions about a population based on a given sample, whereas ML specializes in discovering patterns in existing (training) data that then can be used to make predictions on new (test) data (48). To make a statistical model, usually, some preexisting knowledge is required to assume a data-generating distribution. In ML, no rigid assumptions must be made, and thus it becomes possible to find complex nonlinear dependencies (48).

ML can perform a variety of tasks, such as classification, regression (predicting a numerical value), clustering, and outlier detection (49). Classification and regression are the most common tasks in the ML field that are applied to predict an outcome on the new data. Classification models aim to assign a label to a given data sample, such as to predict cell type based on a gene expression profile. Regression tasks are similar, but instead of a label, a model predicts numerical value, for example, antibody affinity to an antigen of interest given the antibody sequence. Clustering is another common problem and refers to finding and grouping similar data points within a data set, for example, clustering TCR sequences associated with antigen specificity. A variety of ML algorithms exist for each of these tasks, such as support vector machines (SVMs) and decision trees (random forests) for classification and regression, and *k*-means for clustering (49) (**Figure 2**). Independent of the ML algorithm, input data should be described through the vector of numerical variables or features that are curated manually. Feature examples include an amino acid composition of a protein or its physicochemical properties or functional motifs. Not all features are equally informative for a given task; thus, feature selection is performed to discard nonrelevant features from an existing set. Moreover, several features can be combined by a linear function or transformed into a lower dimension, a process termed feature extraction. Feature selection and extraction can be used together or interchangeably depending on a task. The quality and relevance of the curated features play a pivotal role in ML model performance, and

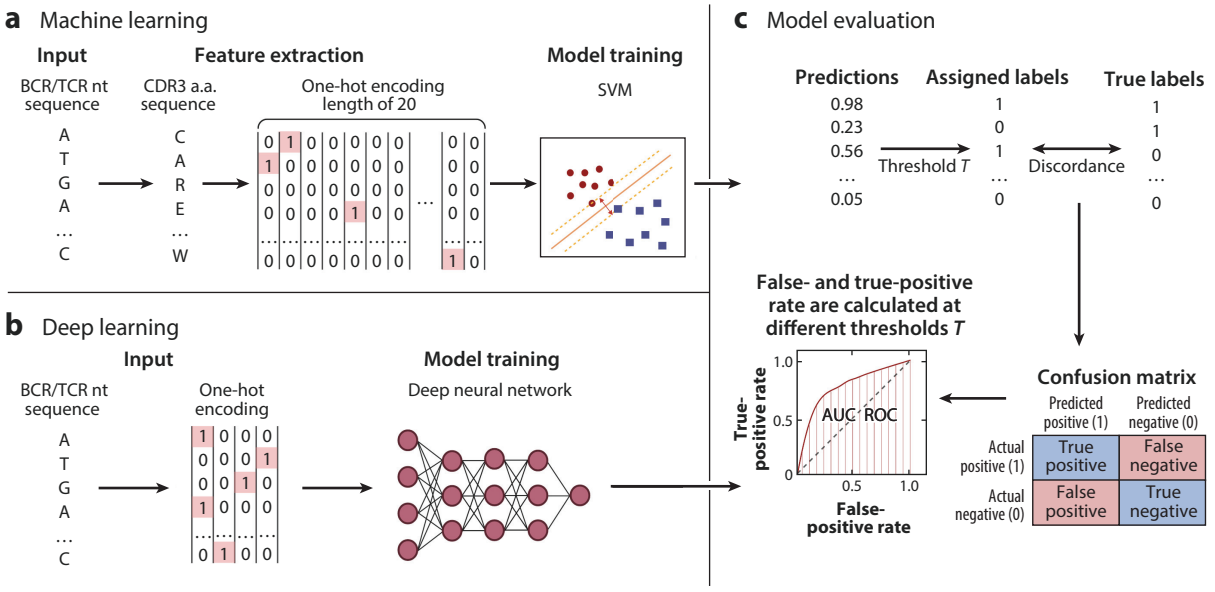

Figure 2

Machine learning (ML) and deep learning (DL) model implementations. (*a*) ML models are based on a set of algorithms that discover patterns in the data without making rigid assumptions about their distribution. To train ML models, data must be described through numerical variables or features relevant for the given learning task. An example of relevant features are CDR3 sequences, which can be used for model training. A common ML classification model is a support vector machine (SVM), which optimizes the decision surface that separates two classes. (*b*) DL is a subfield of ML that takes advantage of multilayered neural networks to extract relevant data variables in an automated manner. Thus, the fundamental difference between the two is that ML requires a feature extraction step. In the context of immune receptor repertoires, data consist of nucleotide (nt) or amino acid (a.a.) sequences, which can be converted into numerical vectors using one of the encoding schemes. In one-hot encoding, each nt or a.a. is assigned to a unique position marked with a one in an otherwise all-zeros vector. DL models are based on neural networks composed of the layers of nodes connected by weights or parameters. (*c*) During the training, the SVM and neural network parameters are tuned to reduce the difference between true and predicted outcome on a given training data set. ML and DL models are validated on new unseen data (test data), and their performance is estimated by different metrics that are calculated based on the discordance between predicted and true labels. One of the widely used metrics is an area under a receiver operating characteristic curve (AUC ROC) plotted in false- and true-positive rate coordinates and calculated at multiple thresholds for label assignments. Figure adapted from images created with BioRender.com.

feature-selection techniques are an active area of research. For more details on ML algorithms, please refer to the valuable resource by Murphy (49).

Introduction to Deep Learning

ML can be very powerful for data analysis and predictions. However, one of its major limitations is that the feature extraction step can be tedious and often requires domain-specific knowledge (50). This becomes increasingly difficult when large amounts of data are generated containing hundreds of potential features. This bottleneck has been overcome by the emergence of a subclass of ML known as deep learning (DL). DL uses a class of algorithms that find a relevant set of features required to perform a particular task in a more automated manner (50). DL algorithms are based on artificial neural networks that are organized in hierarchical layers that transform input in a nonlinear fashion (**Figure 2**). Owing to the layered structure, the computed result of one layer acts as an input to the next layer, resulting in an increasingly abstract data representation (50). Such architecture allows DL algorithms to discover complex patterns in high-dimensional data,

which has led to breakthroughs in image and speech recognition as well as improved performance over ML on a variety of problems. One disadvantage of many ML algorithms, especially those that utilize neural networks and DL, is the lack of interpretability (51).

A wide spectrum of DL and neural network architectures have been developed that differ in areas such as activation and loss functions, weight sharing, and connectivity between nodes. Convolutional neural networks (CNNs) have been employed extensively for image processing and capture local patterns in the data independently of their position. This property also proved to be valuable for many biological problems, as CNNs have been applied to classify binding sites of transcription factors (52) or microRNA targets (53). Another major class of DL models are recurrent neural networks (RNNs), which have made a breakthrough in speech and language processing tasks and found their application in a variety of biological tasks, including protein engineering (54). DL architectures are also used to reduce data dimensionality or generate novel sequences, with variational autoencoders (VAEs) and generative adversarial networks (GANs) being the main algorithms applied to such tasks. Full coverage of DL models is outside the scope of this review; the interested reader could refer to several additional resources (50, 51, 55).

Training and Testing Machine and Deep Learning Models

In biology, ML and DL models are widely applied to omics data and drug discovery; examples include ML models that predict the function of genomic noncoding variants (56), transcription factor binding sites (57), protein structure (58), small-molecule drug inhibitors (59), and novel antibiotics (60). Independent of the application, building an ML model typically includes the following steps: (*a*) collection and processing (cleaning) of training data, (*b*) extraction of meaningful variables (features), (*c*) model training, and (*d*) model evaluation (testing). To explain how ML algorithms work, we briefly go through each of these steps.

Training data. The size and quality of training data are at the heart of all ML and DL models and largely determine their performance, robustness, and accuracy. The structure of training data also dictates which type of learning can be performed: supervised or unsupervised (51). Supervised learning requires labeled data, meaning that each data point is paired with a certain output, such as class or quantitative value. For example, a library of antibody variants with known amino acid sequences (e.g., features) are stratified to binders and nonbinders (two output classes). Supervised learning algorithms, such as SVMs or random forest, are then used to predict an outcome on new data, such as whether a new antibody sequence is a binder or nonbinder to target antigen. On the contrary, in unsupervised learning, only input features are given, and an algorithm must make sense of data without guidance. Typical unsupervised learning tasks are data clustering and dimensionality reduction performed with methods such as principal component analysis or *k*-means clustering (51). Note that the division between supervised and unsupervised learning is formal, and many algorithms can do both tasks as well as perform semisupervised learning; the interested reader is directed to several other valuable reviews on applications of ML and DL in biology (55, 61–63).

Features. The next essential step after data collection is feature selection and extraction. This step is necessary to remove redundant (e.g., highly correlated) features and select the most relevant ones for the specific learning task. Reducing the number of features speeds up learning time by decreasing computational burden as well as simplifying the model. An example of different features given to an algorithm could be distinct ways of antibody sequence representation: whole sequence or only CDR3 region or frequency of amino acid substrings present in a sequence.

Model training. Model training, supervised or unsupervised, refers to iteratively adjusting model parameters or weights. It is based on a loss function that estimates the performance of the current model and an optimization method that gradually corrects weights toward better performance based on a loss. In supervised learning, the loss function is usually defined as a difference between true and predicted labels. In unsupervised learning, a definition of the loss function is not as straightforward, as there are no true labels. Nevertheless, one can still determine that the model performs well if it simplifies the original data representation while keeping as much information as possible and, thus, define a loss based on this criterion. Therefore, even though loss and optimization methods are implemented differently in unsupervised learning, the algorithms can still be thought of in these terms.

Model evaluation and testing. Once a model is trained, it must be evaluated on a new set of test data. The reason for this is that ML and DL models could have so many parameters that they could memorize training data. However, such a model would have little value when applied to a new data set owing to its low generalization ability and poor real-world performance. This phenomenon, referred to as overfitting, represents a fundamental challenge in ML and DL. To avoid bias in model evaluation, the data set is usually split into two parts: training and test set, with no data overlap between the sets. A common practice of random data split may not be a valid approach in biology (64). For example, training data based on protein sequences should consider sequence homology when determining training and test splits, as related proteins (defined by their sequence or structure similarity) should be located to the same split.

Metrics to evaluate model performance depend on several factors, such as problem type (regression or classification) and the proportion of the classes in a data set. The most commonly applied metrics for classification problems are metrics based on the confusion matrix values (true positive, false positive, true negative, and false negative), for instance, precision and recall (true positive rate). Another widely used metric is an area under receiver operating characteristic curve (AUC ROC). AUC ROC is plotted in true positive rate–false positive rate coordinates that are calculated at different confidence thresholds for label assignment. AUC ROC values of 0.5 indicate a random classifier, and a value of 1.0 indicates a perfect classifier. As a rule of thumb, several performance metrics should be computed to get a complete picture of the model performance.

Encoding Biological Sequence Data

To train any algorithm, input data must first be transformed into numerical vectors. For example, an image can be encoded into numbers through the color intensity values of its pixels. In biology, typical data sets consist of sequences (either nucleotide or amino acid) or protein structures. Protein structures are rich in information that can be used to predict protein function and possible interaction partners. However, the number of available structures, especially for BCRs and TCRs, is minuscule in comparison to the number of known sequences. Therefore, in this review, we focus primarily on the numerical encoding of biological sequences and ML models trained on sequence data (**Figure 2**).

One-hot encoding. One-hot encoding represents a simple way to encode categorical values such as the 20 canonical amino acids (49). In one-hot encoding, each category (amino acid length) is converted into a vector of length equal to the number of categories (20 amino acids). For the given amino acid residue, 19 of the categories will be filled with a 0, whereas a 1 will be used for the one category with the corresponding amino acid. One-hot encoding is widely used to transform categorical values and provides a good baseline; however, it is not computationally efficient, because it

is sparse and high dimensional (65). High dimensionality of encoded input is also associated with model overfitting, as the number of features exceeds the number of data points, a phenomenon known as the curse of dimensionality (65). Additionally, one-hot encoding treats all amino acids equally without taking their physical and biochemical properties into account.

_k_-mer encoding. In *k*-mer encoding, the sequences are split into substrings of length *k*, and a frequency of every unique *k*-mer is calculated (66). Therefore, each sequence is encoded as a vector indicating the frequency of *k*-mers it is composed of. One drawback of this method is again the sparsity and high dimensionality of the encoding as the number of possible unique *k*-mers grows exponentially with the increased *k*. For example, a DNA sequence encoded through the frequency of its pentameric substrings would be converted into a vector of length $4^5 = 1{,}024$ (number of possible unique DNA pentamers). This effect is even more dramatic for protein encoding with the alphabet size of 20.

One-hot and *k*-mer encodings are applied to both nucleotide and amino acid sequences. Below we describe a few methods specific to amino acid sequence data.

Amino acid composition. A sequence is encoded as a vector of 20 in which every position corresponds to the frequency of a given amino acid in a given sequence. The advantage of this method is that all sequences are transformed into same-length dense vectors; however, the positional amino acid information is lost, and preserving it is often important for many biological applications.

Physicochemical properties–based and evolutionary-based encoding. This type of encoding is based on amino acid descriptors, such as their hydrophobicity or charge or evolutionary relationships. A large number of amino acid descriptors are available; for example, the current version of the AAindex database has a collection of ∼700 indices based on biochemical properties, substitution matrices, or pairwise interaction propensity (67). The five Atchley factors describing amino acid volume, charge, polarity, structure, and codon diversity (68), as well as BLOSUM and PAM matrices (69, 70) characterizing amino acid mutation propensity, are among the most well-known and utilized descriptors. Sequence encoding through the amino acid descriptors is usually the method of choice; however, it is often challenging to determine which descriptor would perform best on a given task, and thus testing multiple descriptors is recommended.

Learned encodings. An alternative approach to encoding based on certain properties (biochemical, structural) is embedding learned from raw data via ML and DL algorithms. Several learned protein representations have been proposed built on the architectures initially made for language processing tasks (54, 71–73). Learned embeddings, as a rule of thumb, are low dimensional, providing higher computational efficiency while performing on par with the traditional encoding schemes (72, 74). Nevertheless, as with amino acid descriptors, it is challenging to predict which learned embedding would work for a particular problem. Two recent studies proposed to learn amino acid encodings during the training, e.g., to initialize an encoding vector randomly and update its values through the training step to reduce error (74, 75). In this way, amino acid embeddings are assumption free and tuned for a particular given problem.

Similar to how there is no magic bullet algorithm that performs best on any task, there is no one-size-fits-all amino acid encoding. Interestingly, recent studies have shown that although the encoding choice might be important for linear methods, it seems to be surprisingly less critical for more complex architectures (74, 75). Only a minor difference in performance was present for randomized versus biochemical properties–based encoding, indicating high flexibility in the nonlinear models (74, 75). ElAbd et al. (74) also showed that the dimensionality of encoding seems

to be more crucial than the encoding per se, so for optimal performance, one should explore different encoding dimensions. Although intriguing, these results must be confirmed by further studies, such as those comparing a diverse range of models and encoding methods.

MACHINE AND DEEP LEARNING ON IMMUNE RECEPTOR REPERTOIRES

Machine Learning for Immune Repertoires

With increasing amounts of deep sequencing data on immune repertoires available, the use of ML and DL is emerging as a promising method to detect complex patterns underlying repertoire dynamics, clonal selection, and antigen specificity. The exquisite ability of BCRs and TCRs to recognize a large variety of antigens with changes in a few regions (e.g., CDRH3, CDRβ3) implies there may be high-order dependencies within positional sequences; thus, it may be advantageous to incorporate structural information. Recent efforts using structural modeling approaches can be found in other reviews, such as that from Graves et al. (76). Here, we instead focus on sequence-based ML and DL models that take advantage of immune repertoire deep sequencing and are being used to classify immune status, predict antigen specificity, and engineer immune receptor drug candidates (**Figure 3**).

Immune Repertoire Classification

Given that antigen exposure causes specific patterns of clonal expansion and convergence of the immune repertoire, ML approaches offer a practical strategy to perform repertoire comparisons (11). For example, one can study the convergence of repertoires, which is when two different individuals converge to similar BCR or TCR sequences in response to the same antigen, by identifying common sequence patterns (features). In two landmark studies, Dash et al. (77) and Glanville et al. (78) identified convergence based on TCR sequence patterns associated with antigen specificity using either a CDR-weighted distance metric (TCRdist) or clustering based on several sequence features (GLIPH). In both cases, TCR sequence patterns were used to build a classification system that could predict antigen specificity with high accuracy. An updated version, GLIPH2, has shown a substantial improvement in the percent of clustered sequences (36% versus 15% of all sequences clustered in a first version) and can process millions of sequences (79). A recent tool, iSMART, has evaluated clustering of tumor-infiltrating T cells based on CDRβ3 and demonstrated that grouped TCRs had signatures of activation (80). All of these approaches are based on a similarity distance, which is a predefined measurement of the similarity of sequences (e.g., 80% similarity in CDR3 is often used to define clonal groups) between TCRs based on metrics such as modified pairwise alignment of CDRs (77, 80) and work in a supervised manner. However, unsupervised approaches for repertoire sequence clustering would be highly useful considering that specificity is known for only a miniscule percentage of sequenced data. Meysman et al. (81) investigated and benchmarked methods for TCR clustering based on explicit distance metrics (Levenshtein distance, trimer similarity, alignment) as well as on the unsupervised DBSCAN algorithm. They concluded that the performance of simple Levenshtein clustering is equivalent to that of more advanced methods. However, all methods struggled to cluster TCRs binding to the same peptide–MHC. This is because TCRs of shared specificity can be as diverse as TCRs targeting different epitopes (81, 82). That study considered only CDRβ3 sequences, and including an alpha chain as well as V(D)J information might improve performance in the future. Studying the architecture of the BCR repertoire sequence space, Miho et al. (83) revealed that clustering by sequence-similarity networks revealed a high degree of architectural and network similarity between individuals, despite the fact that

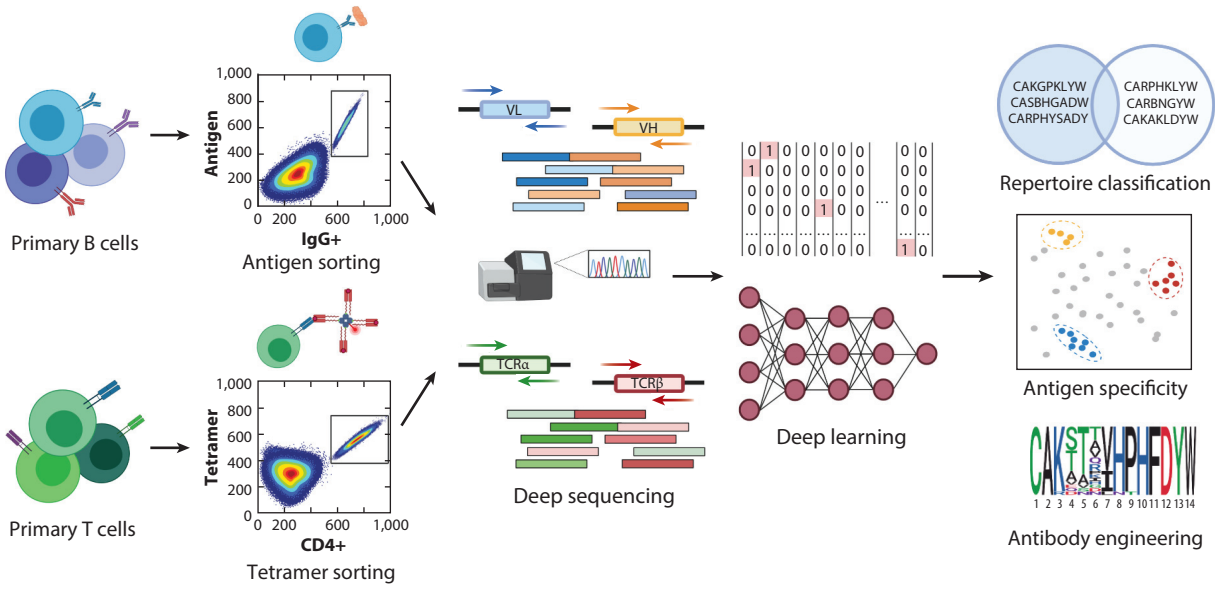

Figure 3

Immune repertoire sequence data generation for deep learning (DL) applications. Starting from a pool of primary cells, B cells or T cells are labeled with antigen and sorted by flow cytometry. Deep sequencing of B or T cell receptor variable regions (V_L/V_H or V_α/V_β, respectively) is then performed to generate immune repertoire training data. Sequences are then encoded and used to train DL models for applications such as repertoire classification, antigen specificity clustering, or antibody engineering. In repertoire classification, DL algorithms seek to use patterns such as clonal expansion and convergence to identify features that are important to a specific disease state. Ensemble DL models are also used that incorporate features such as V gene usage, somatic hypermutation, and isotype to improve classifier performance. Sequence-based approaches to antigen specificity employ DL methods such as variational autoencoders and maximum-entropy modeling to identify clusters of highly convergent antigen-associated sequences. In antibody engineering applications, DL models are trained to predict antigen specificity based on antibody sequence, which are then used to in silico screen large libraries of antibody sequences for additional antigen-specific variants with favorable properties (e.g., drug developability). Figure adapted from images created with BioRender.com.

they had highly divergent antibody sequences. This highlights the need for sequence-based approaches to capture high-dimensional patterns, rather than just using simple sequence alignment and similarity thresholds.

Initial methods of repertoire classification sought to learn underlying structures using single-algorithm approaches. For example, an unsupervised method called ALICE has been established to identify TCRs participating in an immune response; in this approach, TCRs that have more neighbors (defined as max 1 amino acid difference in CDRβ3) than expected by a statistical model are defined as involved in the immune response (84). The method can be applied to repertoire data to detect public and private TCRs associated with a certain disease or condition. Similarly, Emerson et al. (85) have developed a statistical classifier based on the presence of known cytomegalovirus (CMV)-associated TCR sequences in repertoire data to diagnose the CMV status of patients. By employing an SVM-based approach trained on the compositional information of CDRH3 sequences, Greiff et al. (86) could predict public and private clones within human and murine repertoires with 80% accuracy. However, the performance of their SVM model was also highly dependent on the size of the training data set, with a thousandfold increase in training data from 10^2 to 10^5, improving final prediction accuracy by 25%. Despite this, once trained, the SVM classifier was highly robust, achieving high performance on both BCR and TCR repertoires, as

well as across species and with different antigen exposures. In another application, a classifier was built to diagnose relapsing remitting multiple sclerosis, an extremely difficult-to-diagnose autoimmune disease, from other neurological diseases (87). By training a log-likelihood-based classifier on Atchley factor–encoded CDRH3 sequences from patient data, the authors achieved an accuracy of 87% on leave-one-out cross-validation and 72% on previously unseen data from a larger, separate study—far outperforming the accuracy of magnetic resonance imaging–based multiple sclerosis diagnosis techniques (68).

More recent approaches have begun implementing ensemble approaches, which draw upon the strengths of many different models to improve feature learning. To address the challenges of sparse and heterogeneous data sets often found in clinical studies, Tomic et al. (88) developed the Sequential Iterative Modelling "Overnight" (SIMON) ML system. In short, SIMON creates optimal, more complete data sets from sparse clinical data and then trains a large array of 128 different ML models before selecting the best model for feature selection and exploratory analysis. Using SIMON, the authors could identify immune signatures predictive of influenza vaccination from five separate clinical studies of seasonal influenza vaccination. In another study, Konishi et al. (89) incorporated features such as V/J-frame patterns, CDR lengths, number of somatic hypermutations, and physicochemical properties of amino acid sequences in the CDRs to build an ensemble model that combines inputs from linear, Bayesian, and CNN classifiers. Their ensemble model was then used to classify normal versus cancerous tissue based on the localized BCR sequences with an AUC value of 0.826. Besides encoded amino acid sequences, the ensemble model also identified other significant discriminative features, such as the number of somatic hypermutations between tumor and normal tissue. In fact, a recent publication comparing BCR repertoires between six different immune-mediated diseases also revealed that other features, such as isotype and V-gene usage patterns, are also important for discriminating between disease states (90), which could be captured by more complex ensemble models in the future. By incorporating a greater array of features into the final discriminator, ensemble-based approaches hold great promise for accurate repertoire classification.

Finally, repertoire classification is a multiple-instance learning problem, where in an immune repertoire of millions of sequences, only a few are true sequences that indicate its class. Therefore, it is important for a discriminator to isolate patterns important to the disease state, or immune status, rather than other confounding factors, such as genetic background, environmental factors, or immune history. To do so, several groups have leveraged attention-based classifiers that can identify true discriminator sequences within a repertoire (91–93). The most recent, DeepRC, developed by Widrich and colleagues (93), leverages the exponential memory capacity of modern Hopfield networks to greatly improve the storage capacity of the model's attention mechanism, which, combined with CNN-based sequence embedding, allows it to efficiently and accurately extract motifs and residues within repertoires that contribute toward prediction of a disease class. Tests on both real-world CMV data (85) and simulated data show that DeepRC outperforms a panel of SVM, K-nearest neighbors, logarithmic regression, and previous multiple-instance learning–based models, especially when detecting sequence motifs with very low witness rates.

Finally, there is a need for ground truth data sets to benchmark ML and DL tools for immune repertoire classifier evaluation (94). In particular, it is essential to be able to separate true sequence enrichment from the generation probabilities of repertoires and other confounding factors referenced previously. To fulfill this need, various bioinformatic tools have been created for in silico repertoire generation (95, 96). One notable example is immuneSIM, an open-access software package that generates standardized ground truth immune repertoires to be used for comparative benchmarking analysis (97).

Machine Learning to Predict the Antigen Specificity of Immune Receptors

Traditional methods of antibody specificity prediction are based on antibody–antigen structures, obtained either experimentally or through modeling. Tools such as the PEASE and BepiPred-2.0 (random forest based), AntibodyInterfacePrediction (SVM based), AG-Fast-Parapred (RNN based), and the recently updated proABC-2 (CNN based) all seek to apply features learned from antibody–antigen structures to predict new paratopes and epitopes based on given antibody or antigen sequences (98). Although all of these tools have been used for fast and relatively effective prediction of epitopes given the paratope sequence, and vice versa, they are all restricted by scarce antibody–antigen crystal structure data; consequently, they often fail at predicting antibody–antigen binding with novel sequences.

In contrast, sequence-based approaches for specificity prediction are not constrained by the need for readily available crystal structures and are emerging as a promising approach. The potential of sequence-only ML approaches for paratope–epitope pairing is supported by Akbar et al.'s (99) recent discovery of a compact vocabulary of antibody–antigen interaction motifs through use of antibody–antigen structural data. Their findings show that paratope–epitope interactions are a priori predictable and the motifs for antibody–antigen specificity could be learned, paving the way for DL-guided approaches for in silico antibody specificity predictions. Because the current motifs Akbar et al. discovered were derived from structural data, more large-scale, high-throughput methods of antibody–antigen receptor pair generation, similar to library-on-library screens, may enable the identification of more motifs that underlie the rules of antibody–antigen interaction and specificity (100). This can be addressed by single-cell sequencing methods, in which investigators link specific BCR sequences to their antigen specificities at the time of sequencing, leading to discoveries of antibodies specific against HIV-1 and peanut allergen (101, 102). In terms of predicting TCR specificity, De Neuter et al. (103) developed a random forest classifier to predict specificity to two HIV-1 epitopes presented on HLA-B*08 and demonstrated that it is possible to predict TCR specificity based solely on sequence data. Later, Gielis et al. (104) expanded the training set of the model to 18,679 TCR–epitope pairs to detect epitope-specific TCR sequences within the repertoires. With this method, they were able to detect epitope-specific clones with high specificity but low sensitivity.

More recently, statistical inference and ML have also begun to emerge as promising methods of predicting antibody–antigen specificity. In 2016, Asti et al. (105) developed a maximum-entropy model to predict antibody–antigen specificity of sequences from the immune repertoires of individuals. By exploiting the nature of affinity maturation and clonal expansion, they attempted to predict and map the fitness landscape based on finding clusters of comparatively highly mutated sequences. Recently, by using a DL approach, our group observed that convergent selection occurs in immunized mice on a much larger scale than previously known (106). In this study, clustering of these repertoires was performed using VAEs and revealed multiple convergent, antigen-associated sequence patterns that could be used to build an SVM classifier of antigen exposure, as well as to generate synthetic antibody sequences that maintained antigen binding. By leveraging somatic hypermutation and repertoire convergence, attempts to discover meaningful clusters in the repertoire sequence space appear to be a promising direction for finding novel antigen-specific sequences.

Engineering Antibodies with Machine Learning

Common antibody design and engineering workflows employ computational model-based approaches, which rely on predicting affinities through modeling the antibody–antigen interaction interface. Sormanni et al. (107) provided an in-depth summary of some of the common

methods, such as OptMAVEn, OptCDR, AbDesign, and RosettaAntibodyDesign (107–111). These approaches have been used successfully for the affinity maturation of antibodies against several different targets, such as interleukin-17, insulin, and HIV (111–113). Furthermore, newer ML-based models trained on both sequence and structural features also appear to show promise in predicting changes in binding affinity in response to multipoint mutations (114). However, once again, as with most structural-based approaches, these engineering and affinity maturation methods are heavily dependent on the presence of structural data, which are quite rare.

In recent years, by leveraging increased deep sequencing capacity, sequence-based DL approaches have emerged for antibody design and engineering. Deep mutational scanning is a protein engineering method that uses directed evolution and deep sequencing (115). It has recently emerged as a powerful way to obtain high-throughput sequence-function data and can be used to develop predictive models for protein structure and evolution (116, 117). In one prominent example, Mason et al. (118) used deep mutational scanning–guided library design on the CDRH3 of the therapeutic antibody trastuzumab and combined this with a mammalian cell directed evolution platform to generate deep sequencing data of antigen-binding variants. These sequencing data were used to train neural networks that could accurately predict the binding status of antibodies based on their protein sequence. The authors also showed that their DL models outperform structural modeling software in predicting binding ability and can synthesize novel binding sequences from a much larger sequence space. Liu et al. (119) proposed an ensemble model titled Ens-Grad for sequence optimization and novel binder sequence synthesis trained on phage display data. Using Ens-Grad, the authors were able to apply antibody features for antigen specificity learned by the ensemble model for in silico affinity maturation of new input seed sequences, showing that Ens-Grad can generalize into the unseen antibody space. However, in contrast to the method Mason et al. (118) described, Ens-Grad cannot engineer antibodies focused on a specific epitope, as that feature is not in the initial training data and therefore not learned as a discriminator. The authors noted that enriched sequences form isolated clusters of distinct sequence families, which may correspond to specific epitopes; therefore, it may be possible in the future to include epitope prediction as a step in the ensemble model.

In addition to engineering CDRs for affinity maturation, efforts are also underway to improve antibody developability and humanization. For developability, ensemble models for aggregation and random forest regression models for hydrophobicity have shown promise in rapidly identifying liabilities in antibody libraries before and after selection in the discovery process (120, 121). Recently, a new tool, the Therapeutic Antibody Profiler, has been developed for prediction of five more developability characteristics based on antibody variable domain sequences (122). With respect to humanization, Clavero-Álvarez et al. (123) used a multivariate Gaussian model trained on human and mouse variable heavy and light sequences from the Immunogenetics database to predict sequence humanness. The final model was used not only to assign a score to sequences based on their degree of humanness (defined as the multivariate Gaussian score) but also to perform in silico humanization of murine antibody sequences. Focusing on capturing long-range and higher-order dependencies between residues in a human repertoire, Wollacott et al. (124) employed an RNN to quantify sequence nativeness. Trained on variable region sequences from the OAS database, the final model could not distinguish human from mouse, chicken, and llama sequences (AUC $>$ 0.97), but the authors also showed the model can select germline sequences more compatible with CDRs from nonhuman sources.

Although significant progress has been made in efficient assessment of antibody developability and humanization, what is lacking are generative models capable of synthesizing novel sequences given developability parameters. Very recently, steps have been taken to address this missing

piece. Amimeur et al. (125) employed a GAN model trained on sequences from the OAS to discriminate between synthetic and natural human repertoire sequences, following which the GAN was used as a generative model to synthesize novel libraries of sequences with humanlike properties for expression in Chinese hamster ovary cells (CHO) and phage-display systems. Because other generative modeling approaches, such as VAEs, have also been used for in silico generation of antibody sequences (106), it is foreseeable that DL will make a profound impact on therapeutic antibody discovery and development.

Machine Learning for the Major Histocompatibility Complex Antigen Presentation Pathway

Several recent reviews have described and evaluated the ML algorithms for the prediction of MHC antigen presentation (126, 127). Here we describe the main points and tools designed to predict MHC presentation.

Predicting Antigen Processing and Presentation

A variety of tools have incorporated proteasome cleavage sites and the probability of TAP transport into the MHC peptide prediction pipeline (128–131). The ML algorithms used to predict MHC pathway processing steps include the stabilization matrix method (129), hidden Markov models (131), and shallow neural networks (128, 130). Early studies demonstrated that integration of proteasomal cleavage and TAP transportation increases predictive performance (129). However, later results have shown that predicting proteasome cleavage and TAP transport does not contribute to optimal sensitivity (130). This was explained by the high selectivity of peptide loading to MHC in comparison to all other steps of the MHC I pathway (132, 133), as well as proteasome and peptidase promiscuity (134). Nevertheless, with the accumulation of the large body of peptidomics data, the field has been revived. Several studies stated that the integration of the cleavage signals into the MHC I peptide prediction model helped significantly improve performance, though the contribution was not dominant (135, 136). Recently, Laserson and colleagues (137) took one step forward and built a separate antigen processing classifier to discriminate identified MHC I peptides from unobserved peptides, showing a substantial increase in the percentage of true-positive predictions (also termed positive predictive value).

Although several tools exist for prediction of MHC I peptide processing, it has been more challenging to develop equivalent tools for the MHC II pathway owing to its higher complexity, such as antigen degradation by numerous proteases (138) and limited data availability. Only recently, motifs located upstream and downstream of peptide termini were integrated into models for MHC II ligand predictions (139–142), demonstrating a measurable boost in performance.

Predicting Peptide–Major Histocompatibility Complex Binding

The field of peptide–MHC binding predictions is expanding rapidly, with a multitude of papers published annually. In this review, we focus on the leading tools according to recent benchmark studies (126, 143, 144), which are the NetMHC, MixMHCpred, and MHCflurry suites. Most of these models rely on data generated through MS performed on peptides eluted from MHC complexes of lysed cells. However, because cells express multiple MHC alleles, one of the main challenges of MS data on eluted peptides is determining which MHC allele they originate from. One of the experimental solutions is the development of monoallelic cell lines that provide unambiguous data (135, 145, 146); however, this is not compatible with primary cells from patient samples (i.e., tumor cells). Therefore, several methods have been proposed to address this problem

computationally, including GibbsCluster, NNAlign_MA, and MixMHCp (145–147), all of which have shown comparable performance.

The Nielsen lab (148) developed the NetMHC tool, with the latest versions being NetMHCpan-4.1 and NetMHCIIpan-4.0 for MHC I and MHC II ligand predictions, respectively. NetMHCpan tools are pan-specific, meaning that peptide ligands are predicted for multiple human and animal alleles as well as any custom MHC of known sequence. The latest NetMHCpan versions are an ensemble of the feed-forward neural networks trained on both binding affinity and eluted peptide MS data. Different data types can be combined in a model architecture proposed by Jurtz et al. (149). This neural network has two separate outputs for each data type but shared input and hidden layers. Such a training approach demonstrated improvement over methods trained on only one data type.

MHCflurry, implemented by O'Donnell et al. (150), is also based on neural networks and combines both affinity data and eluted peptide data; however, this is achieved through modified loss function. The latest version, MHCflurry2.0, adds an antigen processing prediction step on top of the binding model. This antigen processing model learns to discriminate between predicted strong binders originating from the same protein that were either present in the MS data set (hits) or not observed (decoys) (137). Predictions from both models are then combined via logistic regression to give a composite score. The authors showed that this method leads to a dramatic increase in the positive predictive value of MHC ligand predictions.

Another line of tools, MixMHCpred, is based on probabilistic modeling and was developed by Bassani-Sternberg & Gfeller (145). MixMHCpred is trained solely on eluted peptide data and uses allele-specific position weight matrices for prediction of peptide binding. To decipher MHC I peptidomics data, Bassani-Sternberg et al. (151) took advantage of the shared motifs across data sets with shared HLA alleles, which allowed them to assign motifs and predict ligands more precisely. Identification of the MHC II motifs is more challenging in comparison to the MHC I motifs owing to the longer peptides and flexible position of the binding core on a peptide. To take these factors into account, Racle et al. (142) proposed a specific motif deconvolution algorithm called MoDec. MoDec is a probabilistic algorithm that allows flexible binding core position and simultaneously learns MHC II motifs, weights of position matrices, and allele-specific preferences of the binding core position (142). Deciphered motifs are then used to train MixMHC2pred, which also integrates motifs of peptide N and C termini. In contrast to neural network tools such as NetMHCpan and MHCflurry, the MixMHCpred suite is based on a simple linear method of position weight matrices. However, its precision is equivalent to neural network–based approaches, suggesting that the peptide binding to MHC may simply have a linear complexity.

Overall, the performance of the MHC I and MHC II peptide binding prediction tools has substantially improved in accuracy in recent years, largely driven by the increase in MS data and ML models. Nevertheless, MHC II prediction accuracy is substantially lower owing to the higher complexity of the problem: MHC II open groove accommodates ligands of variable lengths, and the position of the binding core is flexible. Also, details of MHC II pathways are poorly studied in comparison to MHC I, and less MHC II ligand data are available; thus, obtaining more data and improving tools for MHC II ligand prediction would be an important direction for future work. Other challenges are the binding prediction for peptides with post-translational modifications and prediction of peptide immunogenicity (binding to TCRs) as well as immunodominance. The factors governing these phenomena are still not fully understood, so further research and efforts are needed to build ML models capable of predicting defining features of MHC ligands eliciting a strong immune response.

CONCLUSIONS AND FUTURE DIRECTIONS

This review highlights the substantial progress that has been made in applying ML and DL to unravel the complexity of immune receptor repertoires. The field has been catalyzed by the recent exponential growth of data on BCR and TCR repertoires from deep sequencing experiments, as well as peptide–MHC data from MS experiments. This has made it possible to encode these molecular sequence data and use them for training of ML and DL models. To date, a wide variety of model architectures have been implemented, spanning from the simple (SVMs, random forest, logistic regression) to the complex (CNNs, RNNs, VAEs). Future research in this field will benefit from comparing more model architectures and establishing guidelines for model selection for immune repertoires, as such standardization has been beneficial for ML and DL in other applications (i.e., image classification and speech recognition). Progress in ML and DL for adaptive immunity still depends on the generation of large-scale and high-quality training data; although major progress has been made, there is still a major lack of immune repertoire data, which refers to BCR sequences with known antigen specificity and TCR sequences with known peptide–MHC specificity. Therefore, advanced experimental approaches, such as single-cell sequencing, recombinant library screening, and antigen binding and function assays, must continue to develop to generate repertoire data with defined antigen specificity. The long-term trajectory of this field is very promising, as immunology, like other fields of biology, is going through a transformation in which it is merging with computational and data science; thus, ML and DL are poised to lead to important advances in the basic understanding of adaptive immunity as well as applications such as prediction of immune status and specificity, discovery, and development of immunotherapeutics.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was funded by the ERC Starting Grant Antibodyomics and ETH Research Grant (to S.T.R.). Figures included in this review were created with BioRender.com.

LITERATURE CITED

1. Brack C, Hirama M, Lenhard-Schuller R, Tonegawa S. 1978. A complete immunoglobulin gene is created by somatic recombination. *Cell* 15(1):1–14
2. Alt FW, Rosenberg N, Casanova RJ, Thomas E, Baltimore D. 1982. Immunoglobulin heavy-chain expression and class switching in a murine leukaemia cell line. *Nature* 296(5855):325–31
3. Tonegawa S. 1983. Somatic generation of antibody diversity. *Nature* 302(5909):575–81
4. Murphy K, Weaver C. 2016. *Janeway's Immunobiology*. New York: Garland Sci. 9th ed.
5. Weigert MG, Cesari IM, Yonkovich SJ, Cohn M. 1970. Variability in the lambda light chain sequences of mouse antibody. *Nature* 228(5276):1045–47
6. Jacob J, Kelsoe G, Rajewsky K, Weiss U. 1991. Intraclonal generation of antibody mutants in germinal centres. *Nature* 354(6352):389–92
7. Liu YJ, Malisan F, de Bouteiller O, Guret C, Lebecque S, et al. 1996. Within germinal centers, isotype switching of immunoglobulin genes occurs after the onset of somatic mutation. *Immunity* 4(3):241–50
8. Mesin L, Ersching J, Victora GD. 2016. Germinal center B cell dynamics. *Immunity* 45(3):471–82

9. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. 2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* 32(2):158–68
10. Weinstein JA, Jiang N, White RA 3rd, Fisher DS, Quake SR. 2009. High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324(5928):807–10
11. Greiff V, Menzel U, Miho E, Weber C, Riedel R, et al. 2017. Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B cell development. *Cell Rep.* 19(7):1467–78
12. Briney B, Inderbitzin A, Joyce C, Burton DR. 2019. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* 566(7744):393–97
13. Jiang N, He J, Weinstein JA, Penland L, Sasaki S, et al. 2013. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* 5(171):171ra19
14. Laserson U, Vigneault F, Gadala-Maria D, Yaari G, Uduman M, et al. 2014. High-resolution antibody dynamics of vaccine-induced immune responses. *PNAS* 111(13):4928–33
15. Lavinder JJ, Wine Y, Giesecke C, Ippolito GC, Horton AP, et al. 2014. Identification and characterization of the constituent human serum antibodies elicited by vaccination. *PNAS* 111(6):2259–64
16. Doria-Rose NA, Schramm CA, Gorman J, Moore PL, Bhiman JN, et al. 2014. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* 509(7498):55–62
17. Reddy ST, Ge X, Miklos AE, Hughes RA, Kang SH, et al. 2010. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat. Biotechnol.* 28(9):965–69
18. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, et al. 2015. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* 21:86–91
19. Setliff I, Shiakolas AR, Pilewski KA, Murji AA, Mapengo RE, et al. 2019. High-throughput mapping of B cell receptor sequences to antigen specificity. *Cell* 179(7):1636–46.e15
20. Wang B, DeKosky BJ, Timm MR, Lee J, Normandin E, et al. 2018. Functional interrogation and mining of natively paired human VH:VL antibody repertoires. *Nat. Biotechnol.* 36(2):152–55
21. Gilman MSA, Castellanos CA, Chen M, Ngwuta JO, Goodwin E, et al. 2016. Rapid profiling of RSV antibody repertoires from the memory B cells of naturally infected adult donors. *Sci. Immunol.* 1(6):eaaj1879
22. Gérard A, Woolfe A, Mottet G, Reichen M, Castrillon C, et al. 2020. High-throughput single-cell activity-based screening and sequencing of antibodies using droplet microfluidics. *Nat. Biotechnol.* 38(6):715–21
23. Jiang X, Wang S, Zhou C, Wu J, Jiao Y, et al. 2020. Comprehensive TCR repertoire analysis of $CD4^+$ T-cell subsets in rheumatoid arthritis. *J. Autoimmun.* 109:102432
24. Brenna E, Davydov AN, Ladell K, McLaren JE, Bonaiuti P, et al. 2020. $CD4^+$ T follicular helper cells in human tonsils and blood are clonally convergent but divergent from non-Tfh $CD4^+$ cells. *Cell Rep.* 30(1):137–52.e5
25. Ritvo P-G, Saadawi A, Barennes P, Quiniou V, Chaara W, et al. 2018. High-resolution repertoire analysis reveals a major bystander activation of Tfh and Tfr cells. *PNAS* 115(38):9604–9
26. Maceiras AR, Almeida SCP, Mariotti-Ferrandiz E, Chaara W, Jebbawi F, et al. 2017. T follicular helper and T follicular regulatory cells have different TCR specificity. *Nat. Commun.* 8:15067
27. Welten SPM, Yermanos A, Baumann NS, Wagen F, Oetiker N, et al. 2020. $Tcf1^+$ cells are required to maintain the inflationary T cell pool upon MCMV infection. *Nat. Commun.* 11:2295
28. Yermanos A, Sandu I, Pedrioli A, Borsa M, Wagen F, et al. 2020. Profiling virus-specific Tcf1+ T cell repertoires during acute and chronic viral infection. *Front. Immunol.* 11:986
29. Woodsworth DJ, Castellarin M, Holt RA. 2013. Sequence analysis of T-cell repertoires in health and disease. *Genome Med.* 5(10):98
30. Kirsch IR, Watanabe R, O'Malley JT, Williamson DW, Scott L-L, et al. 2015. TCR sequencing facilitates diagnosis and identifies mature T cells as the cell of origin in CTCL. *Sci. Transl. Med.* 7(308):308ra158
31. Thomas S, Mohammed F, Reijmers RM, Woolston A, Stauss T, et al. 2019. Framework engineering to produce dominant T cell receptors with enhanced antigen-specific function. *Nat. Commun.* 10:4451
32. Guo X-ZJ, Dash P, Calverley M, Tomchuck S, Dallas MH, Thomas PG. 2016. Rapid cloning, expression, and functional characterization of paired αβ and γδ T-cell receptor chains from single-cell analysis. *Mol. Ther. Methods Clin. Dev.* 3:15054

33. Rubelt F, Busse CE, Bukhari SAC, Bürckert J-P, Mariotti-Ferrandiz E, et al. 2017. Adaptive immune receptor repertoire community recommendations for sharing immune-repertoire sequencing data. *Nat. Immunol.* 18:1274–78
34. Christley S, Aguiar A, Blanck G, Breden F, Bukhari SAC, et al. 2020. The ADC API: a web API for the programmatic query of the AIRR Data Commons. *Front. Big Data* 3:22
35. Corrie BD, Marthandan N, Zimonja B, Jaglale J, Zhou Y, et al. 2018. iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol. Rev.* 284(1):24–41
36. Kovaltsuk A, Leem J, Kelm S, Snowden J, Deane CM, Krawczyk K. 2018. Observed Antibody Space: a resource for data mining next-generation sequencing of antibody repertoires. *J. Immunol.* 201(8):2502–9
37. Leem J, de Oliveira SHP, Krawczyk K, Deane CM. 2018. STCRDab: the Structural T-Cell Receptor Database. *Nucleic Acids Res.* 46(D1):D406–12
38. Christley S, Scarborough W, Salinas E, Rounds WH, Toby IT, et al. 2018. VDJServer: a cloud-based analysis portal and data commons for immune repertoire sequences and rearrangements. *Front. Immunol.* 9:976
39. Zhang W, Wang L, Liu K, Wei X, Yang K, et al. 2020. PIRD: Pan Immune Repertoire Database. *Bioinformatics* 36(3):897–903
40. Lima WC, Gasteiger E, Marcatili P, Duek P, Bairoch A, Cosson P. 2020. The ABCD database: a repository for chemically defined antibodies. *Nucleic Acids Res.* 48(D1):D261–64
41. Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, et al. 2018. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* 46(D1):D419–27
42. Bagaev DV, Vroomans RMA, Samir J, Stervbo U, Rius C, et al. 2020. VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res.* 48(D1):D1057–62
43. Matsumura M, Fremont DH, Peterson PA, Wilson IA. 1992. Emerging principles for the recognition of peptide antigens by MHC class I molecules. *Science* 257(5072):927–34
44. Chicz RM, Urban RG, Lane WS, Gorga JC, Stern LJ, et al. 1992. Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. *Nature* 358(6389):764–68
45. Eisen HN, Hou XH, Shen C, Wang K, Tanguturi VK, et al. 2012. Promiscuous binding of extracellular peptides to cell surface class I MHC protein. *PNAS* 109(12):4580–85
46. Joglekar AV, Li G. 2020. T cell antigen discovery. *Nat. Methods.* **https://doi.org/10.1038/s41592-020-0867-z**
47. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, et al. 2019. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* 47(D1):D339–43
48. Bzdok D, Altman N, Krzywinski M. 2018. Statistics versus machine learning. *Nat. Methods* 15(4):233–34
49. Murphy KP. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press
50. LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521(7553):436–44
51. Goodfellow I, Bengio Y, Courville A. 2016. *Deep Learning*. Cambridge, MA: MIT Press
52. Wang M, Tai C, E W, Wei L. 2018. DeFine: Deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. *Nucleic Acids Res.* 46(11):e69
53. Cheng S, Guo M, Wang C, Liu X, Liu Y, Wu X. 2016. MiRTDL: a deep learning approach for miRNA target prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 13(6):1161–69
54. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. 2019. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* 16(12):1315–22
55. Angermueller C, Pärnamaa T, Parts L, Stegle O. 2016. Deep learning for computational biology. *Mol. Syst. Biol.* 12(7):878
56. Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* 12(10):931–34
57. Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33(8):831–38
58. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, et al. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* 577(7792):706–10

59. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, et al. 2019. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* 37(9):1038–40
60. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, et al. 2020. A deep learning approach to antibiotic discovery. *Cell* 180(4):688–702.e13
61. Tarca AL, Carey VJ, Chen X-W, Romero R, Drăghici S. 2007. Machine learning and its applications to biology. *PLOS Comput. Biol.* 3(6):e116
62. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. 2019. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* 20(7):389–403
63. van Engelen JE, Hoos HH. 2020. A survey on semi-supervised learning. *Mach. Learn.* 109(2):373–440
64. Jones DT. 2019. Setting the standards for machine learning in biology. *Nat. Rev. Mol. Cell Biol.* 20(11):659–60
65. Bishop CM. 1995. *Neural Networks for Pattern Recognition*. New York: Oxford Univ. Press
66. Leslie C, Eskin E, Noble WS. 2001. The spectrum kernel: a string kernel for SVM protein classification. *Pac. Symp. Biocomput.* 2002:564–75
67. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. 2008. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* 36(Database issue):D202–5
68. Atchley WR, Zhao J, Fernandes AD, Drüke T. 2005. Solving the protein sequence metric problem. *PNAS* 102(18):6395–400
69. Altschul SF. 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219(3):555–65
70. Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *PNAS* 89(22):10915–19
71. Rives A, Meier J, Sercu T, Goyal S, Lin Z, et al. 2020. *Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences*. Work. Pap.
72. Yang KK, Wu Z, Bedbrook CN, Arnold FH. 2018. Learned protein embeddings for machine learning. *Bioinformatics* 34(23):4138
73. Asgari E, Mofrad MRK. 2015. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLOS ONE* 10(11):e0141287
74. ElAbd H, Bromberg Y, Hoarfrost A, Lenz T, Franke A, Wendorff M. 2020. Amino acid encoding for deep learning applications. *BMC Bioinform.* 21:235
75. Raimondi D, Orlando G, Vranken WF, Moreau Y. 2019. Exploring the limitations of biophysical propensity scales coupled with machine learning for protein sequence analysis. *Sci. Rep.* 9:16932
76. Graves J, Byerly J, Priego E, Makkapati N, Parish SV, et al. 2020. A review of deep learning methods for antibodies. *Antibodies* 9(2):12
77. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, et al. 2017. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* 547(7661):89–93
78. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, et al. 2017. Identifying specificity groups in the T cell receptor repertoire. *Nature* 547(7661):94–98
79. Huang H, Wang C, Rubelt F, Scriba TJ, Davis MM. 2020. Analyzing the *Mycobacterium tuberculosis* immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nat. Biotechnol.* 38:1194–202
80. Zhang H, Liu L, Zhang J, Chen J, Ye J, et al. 2020. Investigation of antigen-specific T-cell receptor clusters in human cancers. *Clin. Cancer Res.* 26(6):1359–71
81. Meysman P, De Neuter N, Gielis S, Bui Thi D, Ogunjimi B, Laukens K. 2019. On the viability of unsupervised T-cell receptor sequence clustering for epitope preference. *Bioinformatics* 35(9):1461–68
82. Bentzen AK, Marquard AM, Lyngaa R, Saini SK, Ramskov S, et al. 2016. Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes. *Nat. Biotechnol.* 34(10):1037–45
83. Miho E, Roškar R, Greiff V, Reddy ST. 2019. Large-scale network analysis reveals the sequence space architecture of antibody repertoires. *Nat. Commun.* 10:1321
84. Pogorelyy MV, Minervina AA, Shugay M, Chudakov DM, Lebedev YB, et al. 2019. Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLOS Biol.* 17(6):e3000314

85. Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, et al. 2017. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.* 49(5):659–65
86. Greiff V, Weber CR, Palme J, Bodenhofer U, Miho E, et al. 2017. Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. *J. Immunol.* 199(8):2985–97
87. Ostmeyer J, Christley S, Rounds WH, Toby I, Greenberg BM, et al. 2017. Statistical classifiers for diagnosing disease from immune repertoires: a case study using multiple sclerosis. *BMC Bioinform.* 18:401
88. Tomic A, Tomic I, Rosenberg-Hasson Y, Dekker CL, Maecker HT, Davis MM. 2019. SIMON, an automated machine learning system, reveals immune signatures of influenza vaccine responses. *J. Immunol.* 203(3):749–59
89. Konishi H, Komura D, Katoh H, Atsumi S, Koda H, et al. 2019. Capturing the differences between humoral immunity in the normal and tumor environments from repertoire-seq of B-cell receptors using supervised machine learning. *BMC Bioinform.* 20:267
90. Bashford-Rogers RJM, Bergamaschi L, McKinney EF, Pombal DC, Mescia F, et al. 2019. Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature* 574(7776):122–26
91. Ostmeyer J, Christley S, Toby IT, Cowell LG. 2019. Biophysicochemical motifs in T-cell receptor sequences distinguish repertoires from tumor-infiltrating lymphocyte and adjacent healthy tissue. *Cancer Res.* 79(7):1671–80
92. Sidhom J-W, Larman HB, Ross-MacDonald P, Wind-Rotolo M, Pardoll DM, Baras AS. 2019. *DeepTCR: a deep learning framework for understanding T-cell receptor sequence signatures within complex T-cell repertoires*. Work. Pap.
93. Widrich M, Schäfl B, Pavlović M, Ramsauer H, Gruber L, et al. 2020. *Modern Hopfield networks and attention for immune repertoire classification*. Work. Pap.
94. Brown AJ, Snapkov I, Akbar R, Pavlović M, Miho E, et al. 2019. Augmenting adaptive immunity: progress and challenges in the quantitative engineering and analysis of adaptive immune receptor repertoires. *Mol. Syst. Des. Eng.* 4(4):701–36
95. Marcou Q, Mora T, Walczak AM. 2018. High-throughput immune repertoire analysis with IGoR. *Nat. Commun.* 9:561
96. Olson BJ, Moghimi P, Schramm CA, Obraztsova A, Ralph D, et al. 2019. sumrep: a summary statistic framework for immune receptor repertoire comparison and model validation. *Front. Immunol.* 10:2533
97. Weber CR, Akbar R, Yermanos A, Pavlović M, Snapkov I, et al. 2020. immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking. *Bioinformatics* 36(11):3594–96
98. Norman RA, Ambrosetti F, Bonvin AMJJ, Colwell LJ, Kelm S, et al. 2019. Computational approaches to therapeutic antibody design: established methods and emerging trends. *Brief. Bioinform.* 21(5):1549–67
99. Akbar R, Robert PA, Pavlović M, Jeliazkov JR. 2019. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. bioRxiv 759498. **https://doi.org/10.1101/759498**
100. Younger D, Berger S, Baker D, Klavins E. 2017. High-throughput characterization of protein-protein interactions by reprogramming yeast mating. *PNAS* 114(46):12166–71
101. Croote D, Darmanis S, Nadeau KC, Quake SR. 2018. High-affinity allergen-specific human antibodies cloned from single IgE B cell transcriptomes. *Science* 362(6420):1306–9
102. Setliff I, McDonnell WJ, Raju N, Bombardi RG, Murji AA, et al. 2018. Multi-donor longitudinal antibody repertoire sequencing reveals the existence of public antibody clonotypes in HIV-1 infection. *Cell Host Microbe* 23(6):845–54.e6
103. De Neuter N, Bittremieux W, Beirnaert C, Cuypers B, Mrzic A, et al. 2018. On the feasibility of mining $CD8^+$ T cell receptor patterns underlying immunogenic peptide recognition. *Immunogenetics* 70(3):159–68
104. Gielis S, Moris P, Bittremieux W, De Neuter N, Ogunjimi B, et al. 2019. Detection of enriched T cell epitope specificity in full T cell receptor sequence repertoires. *Front. Immunol.* 10:2820

105. Asti L, Uguzzoni G, Marcatili P, Pagnani A. 2016. Maximum-entropy models of sequenced immune repertoires predict antigen-antibody affinity. *PLOS Comput. Biol.* 12(4):e1004870
106. Friedensohn S, Neumeier D, Khan TA, Csepregi L, Parola C, et al. 2020. *Convergent selection in antibody repertoires is revealed by deep learning*. Work. Pap.
107. Sormanni P, Aprile FA, Vendruscolo M. 2018. Third generation antibody discovery methods: in silico rational design. *Chem. Soc. Rev.* 47(24):9137–57
108. Chowdhury R, Allan MF, Maranas CD. 2018. OptMAVEn-2.0: de novo design of variable antibody regions against targeted antigen epitopes. *Antibodies* 7(3):23
109. Pantazes RJ, Maranas CD. 2010. OptCDR: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding. *Protein Eng. Des. Sel.* 23(11):849–58
110. Lapidoth GD, Baran D, Pszolla GM, Norn C, Alon A, et al. 2015. AbDesign: an algorithm for combinatorial backbone design guided by natural conformations and sequences. *Proteins* 83(8):1385–406
111. Adolf-Bryfogle J, Kalyuzhniy O, Kubitz M, Weitzner BD, Hu X, et al. 2018. RosettaAntibodyDesign (RAbD): a general framework for computational antibody design. *PLOS Comput. Biol.* 14(4):e1006112
112. Nimrod G, Fischman S, Austin M, Herman A, Keyes F, et al. 2018. Computational design of epitope-specific functional antibodies. *Cell Rep.* 25(8):2121–31.e5
113. Baran D, Pszolla MG, Lapidoth GD, Norn C, Dym O, et al. 2017. Principles for computational design of binding antibodies. *PNAS* 114(41):10900–5
114. Myung Y, Pires DEV, Ascher DB. 2020. mmCSM-AB: guiding rational antibody engineering through multiple point mutations. *Nucleic Acids Res.* 48(W1):W125–31
115. Fowler DM, Fields S. 2014. Deep mutational scanning: a new style of protein science. *Nat. Methods* 11(8):801–7
116. Rollins NJ, Brock KP, Poelwijk FJ, Stiffler MA, Gauthier NP, et al. 2019. Inferring protein 3D structure from deep mutation scans. *Nat. Genet.* 51(7):1170–76
117. Riesselman AJ, Ingraham JB, Marks DS. 2018. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* 15(10):816–22
118. Mason DM, Friedensohn S, Weber CR, Jordi C, Wagner B, et al. 2021. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat. Biomed. Eng.* **https://doi.org/10.1038/s41551-021-00699-9**
119. Liu G, Zeng H, Mueller J, Carter B, Wang Z, et al. 2020. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics* 36(7):2126–33
120. Obrezanova O, Arnell A, de la Cuesta RG, Berthelot ME, Gallagher TRA, et al. 2015. Aggregation risk prediction for antibodies and its application to biotherapeutic development. *mAbs* 7(2):352–63
121. Jain T, Boland T, Lilov A, Burnina I, Brown M, et al. 2017. Prediction of delayed retention of antibodies in hydrophobic interaction chromatography from sequence using machine learning. *Bioinformatics* 33(23):3758–66
122. Raybould MIJ, Marks C, Krawczyk K, Taddese B, Nowak J, et al. 2019. Five computational developability guidelines for therapeutic antibody profiling. *PNAS* 116(10):4025–30
123. Clavero-Álvarez A, Di Mambro T, Perez-Gaviro S, Magnani M, Bruscolini P. 2018. Humanization of antibodies using a statistical inference approach. *Sci. Rep.* 8:14820
124. Wollacott AM, Xue C, Qin Q, Hua J, Bohnuud T, et al. 2019. Quantifying the nativeness of antibody sequences using long short-term memory networks. *Protein Eng. Des. Sel.* 32(7):347–54
125. Amimeur T, Shaver JM, Ketchem RR, Taylor JA. 2020. Designing feature-controlled humanoid antibody discovery libraries using generative adversarial networks. bioRxiv 2020.04.12.024844. **https://doi.org/10.1101/2020.04.12.024844**
126. Mei S, Li F, Leier A, Marquez-Lago TT, Giam K, et al. 2020. A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Brief. Bioinform.* 21(4):1119–35
127. Nielsen M, Andreatta M, Peters B, Buus S. 2020. Immunoinformatics: predicting peptide-MHC binding. *Annu. Rev. Biomed. Data Sci.* 3:191–215
128. Nielsen M, Lundegaard C, Lund O, Keşmir C. 2005. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* 57(1–2):33–41

129. Tenzer S, Peters B, Bulik S, Schoor O, Lemmel C, et al. 2005. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cell Mol. Life Sci.* 62(9):1025–37
130. Stranzl T, Larsen MV, Lundegaard C, Nielsen M. 2010. NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* 62(6):357–68
131. Schneidman-Duhovny D, Khuri N, Dong GQ, Winter MB, Shifrut E, et al. 2018. Predicting CD4 T-cell epitopes based on antigen cleavage, MHCII presentation, and TCR recognition. *PLOS ONE* 13(11):e0206654
132. Yewdell JW, Bennink JR. 1999. Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu. Rev. Immunol.* 17:51–88
133. Garstka MA, Fish A, Celie PHN, Joosten RP, Janssen GMC, et al. 2015. The first step of peptide selection in antigen presentation by MHC class I molecules. *PNAS* 112(5):1505–10
134. Toes RE, Nussbaum AK, Degermann S, Schirle M, Emmerich NP, et al. 2001. Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. *J. Exp. Med.* 194(1):1–12
135. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, et al. 2017. Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* 46(2):315–26
136. Sarkizova S, Klaeger S, Le PM, Li LW, Oliveira G, et al. 2020. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* 38(2):199–209
137. O'Donnell TJ, Rubinsteyn A, Laserson U. 2020. MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Syst*. 11(1):42–48.e7
138. Manoury B. 2013. Proteases: essential actors in processing antigens and intracellular toll-like receptors. *Front. Immunol.* 4:299
139. Barra C, Alvarez B, Paul S, Sette A, Peters B, et al. 2018. Footprints of antigen processing boost MHC class II natural ligand predictions. *Genome Med*. 10:84
140. Ciudad MT, Sorvillo N, van Alphen FP, Catalán D, Meijer AB, et al. 2017. Analysis of the HLA-DR peptidome from human dendritic cells reveals high affinity repertoires and nonconventional pathways of peptide generation. *J. Leukoc. Biol.* 101(1):15–27
141. Abelin JG, Harjanto D, Malloy M, Suri P, Colson T, et al. 2019. Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope prediction. *Immunity* 51(4):766–79.e17
142. Racle J, Michaux J, Rockinger GA, Arnaud M, Bobisse S, et al. 2019. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol.* 37(11):1283–86
143. Zhao W, Sher X. 2018. Systematically benchmarking peptide-MHC binding predictors: from synthetic to naturally processed epitopes. *PLOS Comput. Biol.* 14(11):e1006457
144. Paul S, Croft NP, Purcell AW, Tscharke DC, Sette A, et al. 2020. Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system. *PLOS Comput. Biol.* 16(5):e1007757
145. Bassani-Sternberg M, Gfeller D. 2016. Unsupervised HLA peptidome deconvolution improves ligand prediction accuracy and predicts cooperative effects in peptide-HLA interactions. *J. Immunol.* 197(6):2492–99
146. Alvarez B, Reynisson B, Barra C, Buus S, Ternette N, et al. 2019. NNAlign_MA; MHC peptidome deconvolution for accurate MHC binding motif characterization and improved T-cell epitope predictions. *Mol. Cell. Proteom.* 18(12):2459–77
147. Andreatta M, Lund O, Nielsen M. 2013. Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics* 29(1):8–14
148. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. 2020. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res*. 48(W1):W449–54
149. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. 2017. NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* 199(9):3360–68

150. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. 2018. MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst*. 7(1):129–32.e4
151. Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, et al. 2017. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLOS Comput. Biol.* 13(8):e1005725