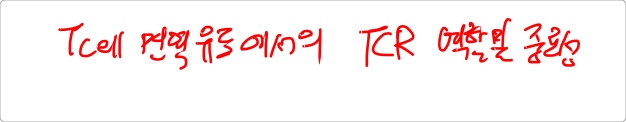
**Notes in ‘Predicting TCR epitope specificty’**

**Notes in Workspace:**

Object Group

Excerpt: The adaptive immune system implements various complex mechanisms for surveillance against both pathogens and pathological cells arising in our body. To initiate an adequate adaptive immune response, a peptide, called epitope, must first be bound by the major histo- compatibility complex (MHC) class I or II molecule expressed on the surface of a nucleated cell or a professional antigen-presenting cell, respectively. The peptide-MHC (pMHC) com- plex is then presented to T cells which can recognize the complex via T cell receptor (TCR) proteins, consequently leading to T cell activation and proliferation by clonal expansion [1]. During clonal expansion, a fraction of T cells gain a long-living memory phenotype and there- fore a clonal population of T cells with identical TCR rearrangements remain for years against the recognized antigen [2], thus forming a potentially decodable immunological signature. Learning these signatures could have implications in broad range of clinical applications including infectious diseases, autoimmunity and tumor immunology. *(Predicting recognition between T cell receptors and epitopes with TCRGP , p.2)*

Excerpt: [Image] *()*

Excerpt: Antigen recognition is one of the key factors of T cell-mediated immu- nity. T cells interact via a dimeric surface protein, the T-cell receptor (TCR), with an antigen presented on a major histocompatibility complex (MHC) located on the surface of antigen-presenting cells. This presenting cell can be experimentally modeled via an MHC multimer with an immobilized antigen (pMHC). *(Predicting antigen specificity of single T cells based on TCR CDR3 regions , p.1)*

Excerpt: The randomness of the recombination process ensures the production of many different TCR proteins, with each T cell clone expressing a particular TCR, and allows the recognition of different epitopes. Epitope binding by the TCR is a critical step for the activation of targeted immune responses. *(Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires , p.1)*

Excerpt: Cytotoxic T cells (CTLs) scan MHC class I-peptide complexes presented on the cell surface of nucleated cells. CTLs are able to recognize and kill infected or malfunctioning cells, e.g. cancer cells (1). Given the central role of the CTLs in the immune system, it is of paramount importance to understand the interaction between the T cell receptor (TCR) of the CTLs and their cognate peptide-MHCI targets. A peptide recognised in this context is referred to as a T-cell epitope.  *(NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks , p.1)*

Object Group

Excerpt: [Image] *()*

Excerpt: TCRs are formed by a pair of α- and β-chains (90-95% of T cells) or γ and δ-chains (5-10%) and V(D)J recombination happens in each locus independently. It is estimated that this rearrangement can result in the range of 1018 different TCR genes [5, 6] which provides enormous diversity for epitope-specific T cell repertoires. *(Predicting recognition between T cell receptors and epitopes with TCRGP , p.2)*

Excerpt: However, not all MHC presented peptides are immunogenic. In order to predict which MHC restricted peptides do become T cell epitopes, the interaction between a TCR and its cognate target needs to be better understood. The TCR must be able to make contacts with the peptide as well as the MHC molecule to trigger an immune response. TCR and MHC interactions were reviewed by Gruta et al. (9).  *(NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks , p.2)*

Object Group

Excerpt: [Image] *()*

Excerpt: Nevertheless, profiling of epitope-specific TCRs remains exhaustive as they require sample-consuming experiments with distinct pMHC-multimers for each epitope of interest. Therefore, there is a great need for models that examine which epitopes a TCR can recognize or to which TCRs an epitope can bind to [15]. *(Predicting recognition between T cell receptors and epitopes with TCRGP , p.2)*

Excerpt: TCRs binding the same MHC-peptide may share similarities. Thus, while for public clones the task of deciphering the relation between a peptide and the TCR binding is based on tallying the candidate public TCR, for most highly cross-reactive TCRs, a probabilistic approach is required. *(Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs , p.2)*

Excerpt: [Image] *()*

Object Group

Excerpt: Curated databases of experimentally verified TCR-peptide interactions have recently been launched, such as VDJdb, IEDB, and McPAS [16–18]. Such data sources enable more comprehensive, data-driven analysis of TCR- *(Predicting recognition between T cell receptors and epitopes with TCRGP , p.2)*

Excerpt: peptide interactions, and allow the use of statistical machine learning techniques for the afore- mentioned tasks. Yet only a few computational methods for predicting recognition between TCRs and epitopes [10, 19–22] and for clustering similar TCRs [9, 23, 24] have been published. In addition to supervised and unsupervised methods for predicting TCR-epitope interactions, computational methods and web services such as [25] have also been proposed to predict the structure of TCRs based on their amino acid sequences. *(Predicting recognition between T cell receptors and epitopes with TCRGP , p.3)*

Excerpt: Only recently, through concerted data collection efforts (Borrman et al, 2017; Shugay et al, 2018; Vita et al, 2019) and newly emerging high-throughput technologies that allow the sequenc- ing of the TCR while probing the T-cell specificity (Klinger et al, 2015; Bentzen et al, 2016), have large enough data sets become avail- able to begin modeling the TCR–pMHC interaction through machine- learning methods (Zvyagin et al, 2020). Current methods to predict the likelihood of binding of TCRs to specific antigens use linear posi- tion-specific scoring matrices (Glanville et al, 2017), Gaussian processes (preprint: Jokinen et al, 2019), or random forests (Gielis et al, 2018). A second set of methods attempts to directly model the TCR–pMHC inter- action with neural networks in order to generalize across unseen TCR– antigen pairs (preprint: Jurtz et al, 2018). *(Predicting antigen specificity of single T cells based on TCR CDR3 regions , p.1)*

Excerpt: A large number of published assays that involve enriching a T cell sample for cells specific to an antigen of interest and sequencing their TCRs predate high-throughput sequencing era. Such data, while being sparsely organized until recent, carries an important source of information that can be applied to annotate TCR sequences and provide additional informa- tion for high-throughput TCR repertoire profiling studies. Recent efforts in summarizing such data gave rise to databases listing TCRs with known antigen specificity: McPAS-TCR (Tickotsky et al. 2017) and VDJdb (Shugay et al. 2018; Bagaev et al. 2019) databases contain ~ 20 k and ~ 55 k specific TCR sequences as of October 2019. Since 2018, im- mune epitope database (IEDB) also provides TCR and anti- body sequences linked with certain antigens having ~ 30 k TCR and ~ 2 k antibody sequences as of October 2019 *(An overview of immunoinformatics approaches and databases linking T cell receptor repertoires to their antigen specificity , p.1)*

Excerpt: Important steps have been made in this direction by Glanville et al. (4) and Dash et al. (18), who detected the clear signature of short amino acid motifs in the CDR3 region of TCRβ and TCRα in response to specific peptides presented by specific MHC molecules. This work was then extended by recent efforts that combined these motifs with machine learning to predict peptide-specific TCRs vs. naïve TCRs, using Gaussian Processes (19) or Random Forest (20), or predicting TCR-epitope binding with Convolutional Neural Networks (21, 22). These methods significantly outperform random classification in the distinction of TCR binding a specific peptide and random TCRs. *(Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs , p.2)*

Excerpt: Recently developed high throughput sequencing methods are likely to change this situation and are already contributing increasing amounts of data (13, 14). Among those methods are the MIRA assay published by Klinger et al. (15) and the TCR barcoding technique published by Bentzen et al. (16). Additionally, two recent publications by Glanville et al.(17) and Dash et al. (18) have made more high throughput data available. Furthermore, these works describe clustering algorithms able to group TCRs by their epitope specificity. In particular, the work by Glanville et al. suggests that relatively simple sequence-based models can be used to classify and define specificity groups shared by TCRs and individuals. This is in line with earlier work by Roomp and Domingues (19). Several structure-based approaches for modelling the structure and interactions of the TCR:p:MHC system have likewise been proposed including structural modeling (20, 21) and structure based prediction of TCR:p:MHC interactions (22). The IEDB (23) as well as the VDJdb (24) collect sequenced TCRs with known specificity published in peer reviewed articles, thereby providing a useful data resource to the community.  *(NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks , p.2)*

Object Group

Excerpt: [Image] *()*

Excerpt: The probabilistic formulation of GPs allows robust model inference already from small data sets, which is a great benefit as currently there exists very limited amounts of reported TCR-epitope interactions in curated databases. As the space of all TCRs that can recognize a certain epitope is potentially very large, it is important to avoid overfitting to the limited sample of TCRs that is available. I *(Predicting recognition between T cell receptors and epitopes with TCRGP , p.3)*

Excerpt: With the currently available epitope-specific TCR sequence data we have been able to come this far, but as more data becomes available with modern high-throughput techniques pre- sented recently [40, 41], new possibilities will rise. *(Predicting recognition between T cell receptors and epitopes with TCRGP , p.15)*

Excerpt: We expand on these efforts but also consider the current limitation in the number of available anti- gens in training data sets. *(Predicting antigen specificity of single T cells based on TCR CDR3 regions , p.1)*

Excerpt: Given the fact that most TCR sequencing projects focus on characterizing the CDR3 region of the T cell receptor beta chain sequence, we chose to train a model based on this part of the TCR only. It is clear that this potentially has limited the predictive power of our model, and that future extension of the model would benefit from being trained on paired T cell sequence data  *(NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks , p.12)*

Excerpt: ERGO randomly initializes our amino-acid embeddings and trains the embeddings with the model parameters. Using word- embedding algorithms such as Word2Vec (40) or GloVe (41) can give a good starting point to the embeddings. Special options for amino-acids pre-trained embeddings include the use of BLOSUM matrix (42) or Kidera-factors-based manipulations (43). As pre-trained embedding usually provides better model results, we plan to further test such encodings. *(Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs , p.7)*

Excerpt: Due to the small amount of training peptides, the model can however at present only be applied to the limited set of peptides included in the training data. However as more data becomes available, we expect the predictive power of the model to increase, and allow for accurate predictions also for uncharacterized peptides as has been observed earlier for the pan-specific prediction models of peptide-MHC interactions (44).  *(NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks , p.14)*

Excerpt: These results indicate that current models are therefore not neces- sarily bound by technical limitations but rather by a lack of suitable training data. As we anticipate the amount of avail- able MHC-peptide-TCR data to increase in the future, *(On the feasibility of mining CD8+ T cell receptor patterns underlying immunogenic peptide recognition, p.8)*

Object Group

Excerpt: [Image] *()*

Excerpt: Advances in natural language processing (NLP) have shown that self-supervised learning is a powerful tool for extracting information from unlabeled sequences [5–7], which raises a tantalizing question: can we adapt NLP-based techniques to extract useful biological information from massive sequence datasets? *(Evaluating Protein Transfer Learning with TAPE , p.2)*

Excerpt: Semi-supervised learning tries to jointly leverage information in the unlabeled and labeled data, with the goal of maximizing performance on the supervised task. One successful approach to learning from unlabeled examples is self-supervised learning, which in NLP has taken the form of next token prediction [5], masked token prediction [6], and next sentence classification [6]. Analogously, there is good reason to believe that unlabelled protein sequences contain significant information about their structure and function [2, 4]. Since proteins can be modeled as sequences of discrete tokens, we test both next token and masked token prediction for self-supervised learning. *(Evaluating Protein Transfer Learning with TAPE , p.4)*

Excerpt: Recent advances in natural language processing have enabled techniques to train complex models that understand semantics from text without labels (self-supervised learning) (Peters et al., 2018; Devlin et al., 2019). Such models are trained to predict words masked out in a sentence or to predict the next word or sentence following some context. *(BERTMHC, p.2)*

Object Group

Excerpt: [Image] *()*

Excerpt: Similar techniques have also been applied to proteins (Rao et al., 2019; Nambiar et al., 2020; Heinzinger et al., 2019). Since these models do not require labels to train, they can be trained on very large corpora of protein sequences across many species. One example of these models is the TAPE model (Rao et al., 2019), which was trained with 31 million protein sequences from the Pfam database (El-Gebali et al., 2019). The model has been shown to be helpful in a variety of downstream tasks such as remote protein homology prediction (Fox et al., 2013) and stability prediction (Rocklin et al., 2017). Detailed analysis of the model has shown that it captures long-range interactions in 3D structure (Vig et al., 2020). It is highly relevant to explore whether pretrained protein sequence models can be helpful for MHC–peptide binding and presentation prediction, especially for MHC class II where much less data is available. *(BERTMHC, p.2)*

Excerpt: Here, we focus on developing models for predicting MHC–peptide binding and presentation for MHC class II. We show that models taking advantage of self-supervised pretraining from large corpora of protein sequences can achieve better performance on both binding and presentation prediction tasks. We found pretraining to be extremely valuable in the case where training data is limited. *(BERTMHC, p.2)*

Excerpt: In this work, we have demonstrated that self-supervised pretraining for a transformer model leads to state-of-the-art performance for MHC class II binding and presentation prediction. *(BERTMHC, p.6)*

Excerpt: The self-supervised pretraining step was not specific to downstream tasks, and was trained from generic protein sequences that may have very distinct biochemical properties compared to the sequences of our task. Nevertheless, the pretraining step was very beneficial. Preliminary work has been done to interpret the models trained from protein sequences with self-supervised learning (Heinzinger et al., 2019; Vig et al., 2020). This is a promising direction for future research in order to better understand what the models have learned and how that can guide better treatment decisions. *(BERTMHC, p.6)*

Object Group

Excerpt: [Image] *()*

Excerpt: We use the pretrained bidirectional encoder representations from transformers (BERT, Devlin et al., 2019) from the TAPE repository to model the input amino acid sequences. The TAPE model was trained with self-supervised learning from a dataset of over 31 million protein sequences. Briefly, taking unlabelled protein sequence as input, the TAPE model was trained with two tasks. One task is bidirectional next-token prediction (predicting p(xi|x1, x2, ..., xi−1) and p(xi|xi+1, xi+2, ..., xL)), and the other task is masked-token prediction (predicting p(xmasked|xunmasked)). The model has 12 layers with 12 self- attention heads (Equation 1) in each layer, which enables the model to learn long distance interactions. For an input amino acid sequence z = (z1, z2, ..., zL), the output of the model are L continuous vectors of dimension 768 corresponding to the input amino acids *(BERTMHC, p.3)*

Excerpt: Losses: We examine two self-supervised losses that have seen success in NLP. The first is next- token prediction [44], which models p(xi | x1, …, xi−1). Since many protein tasks are sequence-to-sequence and require bidirectional context, we apply a variant of next-token prediction which additionally trains the reverse model, p(xi | xi+1, …, xL), providing full context at each position (assuming a Markov sequence). The second is masked-token prediction [6], which models p(xmasked | xunmasked) by replacing the value of tokens at multiple positions with alternate tokens. Protein-specific loss: In addition to self-supervised algorithms, we explore another protein-specific training procedure proposed by Bepler et al. [11]. They suggest that further supervised pretraining of models can provide significant benefits. In particular, they propose supervised pretraining on contact prediction and remote homology detection, and show it increases performance on secondary structure prediction. Similar work in computer vision has shown that supervised pretraining can transfer well to other tasks, making this a promising avenue of exploration  *(Evaluating Protein Transfer Learning with TAPE , p.8)*

Object Group

Excerpt: [Image] *()*

Excerpt: Datasets. In our experiments, we use a data set collected by Dash et al. [10] (from hereon referred to as Dash data), which contains epitope-specific paired TCRα and TCRβ chains for three epitopes from humans and for seven epitopes from mice, see Table 3 for details. We also gather a new data set (VDJdb data) from VDJdb (http://https://vdjdb.cdr3.net, downloaded 9th October 2018) [16], which is a database that contains TCR sequences with known antigen specificity. Every entry in VDJdb has been given a confidence score between 0 and 3 (0: critical information missing, 1: medium confidence, 2: high confidence, 3: very high confidence). We constructed our data set so that we selected all epitopes that have at least 50 TCRβ sequences with a confidence score at least 1 and found 22 such epitopes. Twenty of these epitopes are MCH class I restricted and two are MHC class II restricted and have varying HLA-types, see Table 3 for details. VDJdb also contains TCRα sequences, but since these are not in general paired with the corresponding TCRβ sequences, we chose to only experiment with the TCRβ sequences. The VDJdb data and the Dash data have some overlap for TCRs specific to three epitopes: In the VDJdb data 34 (27 unique) of the 413 (242 unique) TCRs for pp65495-503, 30 (27) of 299 (152) for BMLF1280-288, and 74 (61) of 239 (138) for M158-66 can also be found from the Dash data. For the training and testing of the models, we also required some background TCRs that we do not expect to recognize the epitopes in our data sets. For this purpose, we randomly sampled the required amount of TCRs from sets of background TCRs constructed by Dash et al. [10]. They report that the human α- and β-chains have been obtained from Howie et al. [52], who have collected blood from two healthy adults. The mouse α-chains they have gath- ered from short read archive (SRA) projects SRP010815 [53], and SRP059581 [54] and mouse β-chains from SRA projects SRP059581 [54], SRP015131 [55], and SRP004475. To create paired α- and β-chains they randomly paired the unpaired α- and β-chains from the reper- toires for the corresponding organism. *(Predicting recognition between T cell receptors and epitopes with TCRGP , p.16)*

Excerpt: In this study, we worked on data sets from public databases IEDB (Vita et al, 2019) and VDJdb (Shugay et al, 2018) and on a public data set from a single-cell pMHC-based T-cell specificity experi- ment (10x Genomics, 2019). IEDB and VDJdb contain pairs of binding T-cell receptors (TCRs) and antigens. *(Predicting antigen specificity of single T cells based on TCR CDR3 regions , p.10)*

Excerpt: Three TCR-peptide datasets were used in the binding prediction task. McPAS-TCR dataset was downloaded from http:// friedmanlab.weizmann.ac.il/McPAS-TCR/ and VDJdb dataset was downloaded from https://vdjdb.cdr3.net/, both in November 2019. We used a dataset of cancer neoantigen peptides and their matching TCRs, published by Zhang et al. (31). A set of cancerous peptides was made for extracting TCRs matching to these peptides also in McPAS-TCR and VDJdb databases. We extended the original cancer dataset to include all TCRs- cancerous peptide pairs in all datasets. The data were processed into TCR-peptide pair files, using only TCRβ chains and valid TCR and peptide sequences. The TCR autoencoder was trained on a data which was derived from a prospective clinical study (NCT00809276) by Kanakry et al. (46) The dataset is freely available at the Adaptive database (www.adaptivebiotech.com) that provides open access to a variety of datasets of TCRs next-generation sequencing. *(Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs , p.7)*

Excerpt: A dataset of TCR beta chain CDR3 sequences and corresponding cognate peptide targets was downloaded from the IEDB (23) in April 2018. Only peptides presented by HLA-A\*02:01 were selected. The training data consisted of 9015 unique data points, spanning 91 peptides and 8920 TCR sequences. Further, an additional dataset generated using the MIRA assay was kindly provided by Klinger et al. (15). This dataset consisted of 379 unique data points, spanning 16 peptides and 379 TCR sequences derived from 5 donors. Since these data sets contain only positive interactions, negative data examples were generated by creating internal wrong combinations of TCRs and peptides, i.e. combining TCR sequences with peptides different from their cognate target. These combinations were made by extracting the list of peptide targets from the positive data set (keeping duplicates if a peptide was found to interact with multiple TCRs), and next pairing each TCR with a peptide different from the cognate target randomly drawn from this list of peptide targets. In this way, a data set with 50% positive and 50% negative data points was obtained. *(NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks , p.3)*

Excerpt: A positive training dataset was constructed containing human epitope-specific TCR beta chain sequences collected from the manually curated catalog of pathology-associated T cell receptor sequences [McPAS-TCR (10)] (11 339 TCR-pathology combinations) and the VDJ database [VDJdb (9)] (17 792 TCR- epitope combinations) on November 18, 2018. To train the prediction models, both the CDR3 beta amino acid sequence and the V/J genes were gathered. The following quality filtering steps were applied for the McPAS-TCR dataset: (1) retaining epitope-specific TCRs determined with peptide-MHC tetramers or peptide stimulation and (2) removal of TCRs with missing information (i.e., CDR3 beta sequence, V/J genes or the specific epitope), reconstructed J genes, V/J genes with special characters that could not be matched to known V/J genes, CDR3 beta sequences with lower case amino acids or non-amino acid characters, TCRs with additional quality remarks and TCRs from studies using mouse strains. The dataset was further extended with 550 TCR-epitope combinations found in the published literature (13, 16–18). Standard TCRex filtering steps were carried out (Supplemental Material S1). *(Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires , p.2)*

Object Group

Excerpt: [Image] *()*

Excerpt: Background TCRs. To estimate background frequencies of the different V and J genes, generate background TCRs for use in assessing the significance of CDR3 motifs, and as negative samples for discrimination tests, we relied on the following high-throughput repertoire profiling experiments: for the mouse α chain, short read archive (SRA) projects SRP010815 (ref. 10) and SRP059581 (ref. 11); for the mouse β chain, SRA projects SRP059581 (ref. 11), SRP015131 (ref. 12), and SRP004475; for the human α and β chains, the study of Howie et al.13. For back- ground frequency comparisons we took the minimum normalized JS *(Quantifiable predictive features define epitope- specific T cell receptor repertoires , p.7)*

Excerpt: To increase the specificity of our model it was necessary to add more TCR sequences and peptides as negative examples to the training. The negative TCR sequences were obtained by repertoire sequencing of healthy individuals and paired with the peptides in the IEDB and further self peptides identified by ligand elution assays to be presented by HLA-A\*02:01. Combining TCR sequences of healthy individuals with human self peptides should result in true negative examples in the vast majority of cases, but when combining those TCRs with the peptides present in the IEDB data, which are to a large extent well studied influenza and herpes virus epitopes (42), they might result in some false negatives.  *(NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks , p.13)*

Excerpt: For creating the negative examples for the TPP-I task, we first chose a peptide randomly from the peptides in the training set. Then, we chose five random TCRs from the training set that are not reported to bind this peptide, to create five internal wrong pairs *(Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs , p.7)*

Excerpt: We generated negative samples for both training and test sets separately by generating unobserved pairs of TCR and antigens. *(Predicting antigen specificity of single T cells based on TCR CDR3 regions , p.12)*

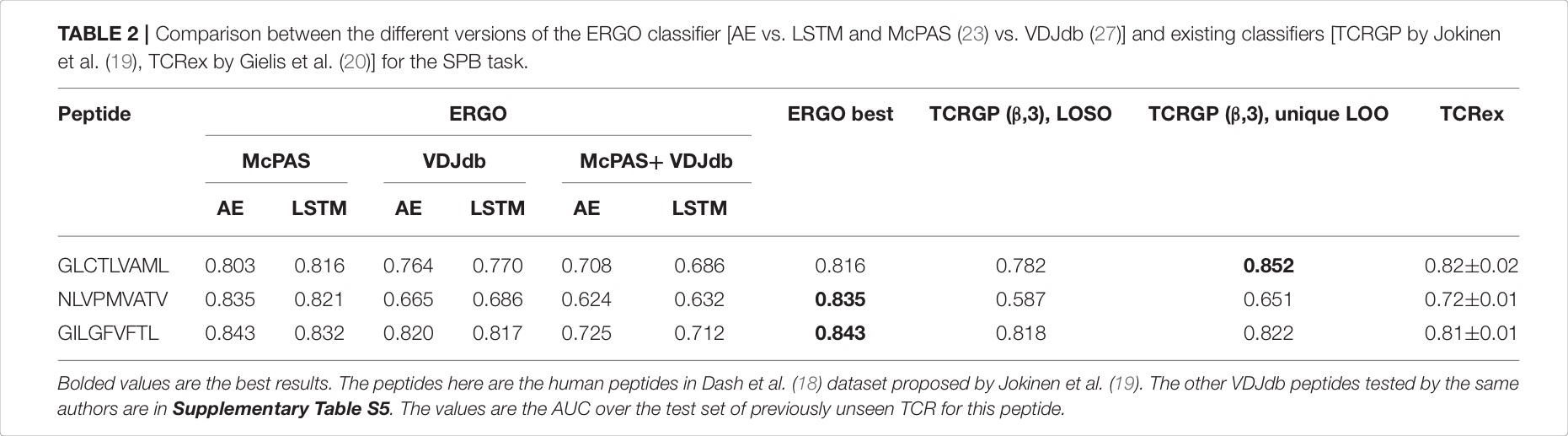
Excerpt: A set of peptides covering cancer mutations and their cor- responding non-mutated counterparts was generated to serve as a background control. *(Predicting T cell recognition of MHC class I restricted neoepitopes , p.13)*

Excerpt: Negative control data was obtained by querying the ImmuneACCESS database using the following terms: ‘hu- man’, ‘TCRB’, ‘HLA-B\*08’, ‘CD8+’ and ‘control’; which returned 66.235 TCRβ chain sequences originating from a single individual. From these, 56.023 unique, productive, in- frame sequences were withheld. *(On the feasibility of mining CD8+ T cell receptor patterns underlying immunogenic peptide recognition, p.8)*

Object Group

Excerpt: [Image] *()*

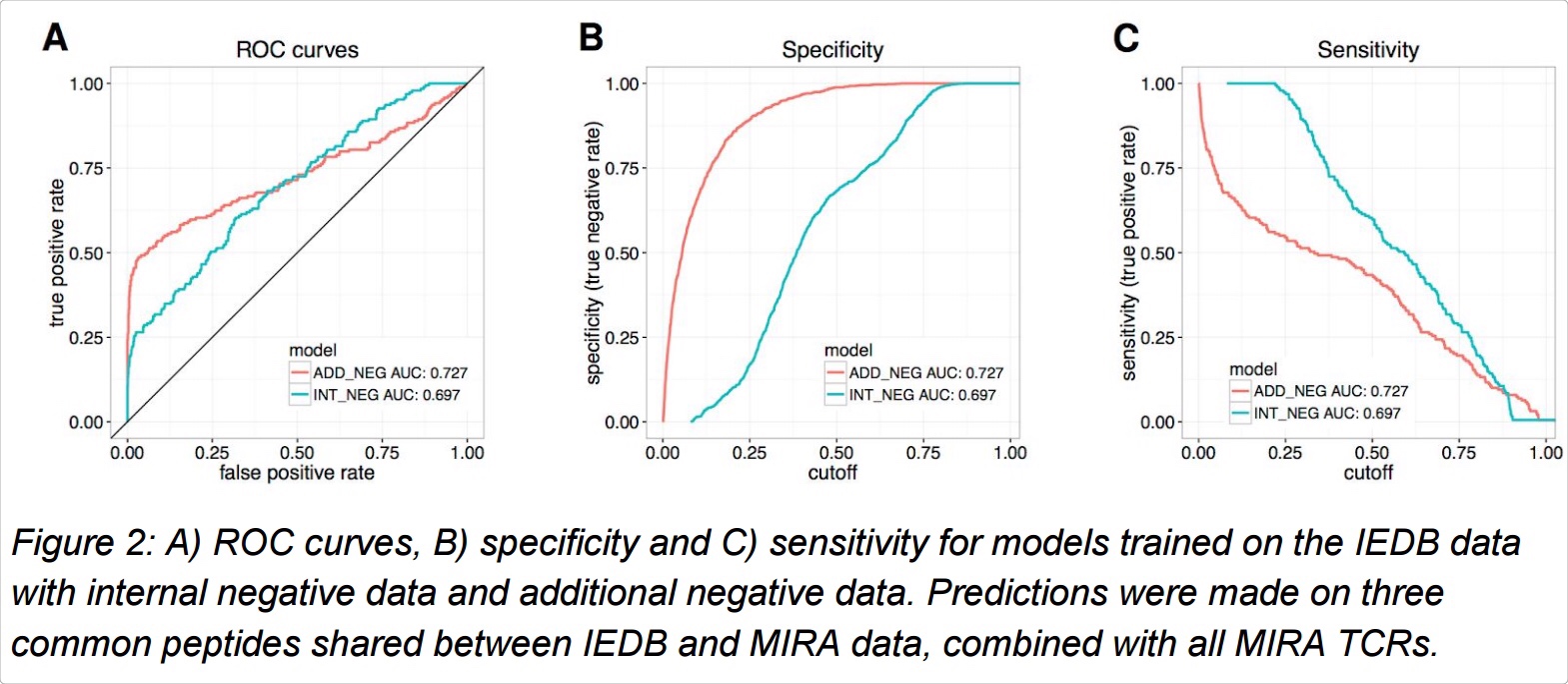
Excerpt: Additionally, we trained a classifier for COVID-19 epitope S-protein269-277 (YLQPRTFLL) with the 352 epitope-specific TCRβs from the recent study of Shomuradova et al. [34] and tested the classifier with the 415 in-frame S-protein269-277-specific TCRs from the ImmuneRACE study launched by Adaptive Biotechnologies and Microsoft (https://immunerace. adaptivebiotech.com, June 10, 2020 dataset). The immuneRACE data did not have sufficient information of the Vβ-gene for us to utilize all CDRβs, so the classifier was trained and tested only with CDR3β. Nonetheless, we achieved an AUROC score of 0.895 on the independent test data. *(Predicting recognition between T cell receptors and epitopes with TCRGP , p.13)*

Excerpt: [Image] *(Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs , p.5)*

Excerpt: In the future, ERGO may contribute to the development of TCR-based diagnostic tools. However, it can already be used for the detection of TCRs that bind specific tumor antigens. Given a neoantigen extracted from full genome sequencing of tumors (29, 30) and a target TCR, one could estimate the binding probability of the TCR to such a neoantigen. To test for that, we applied ERGO to neoantigen binding prediction; we used a positive dataset of cancer neoantigen peptides and their matching TCRs, published by Zhang et al. (31), *(Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs , p.5)*

Excerpt: The united IEDB and MIRA datasets were downloaded from https://github.com/mnielLab/netTCR. Unfortunately, the authors did not publish the IEDB train data separated from the MIRA test data, thus we had to evaluate ERGO in another train/test partition. We used 80% of the IEDB and MIRA data for training and the rest of it (20%) for testing. Additional “C” prefix and “F” suffix were added to each TCR sequence. The MIRA data was containing new test TCRs (but was not evaluated with new test peptides), therefore we compare NetTCR results with ERGO TPP-II scores (Supplementary Table S3). *(Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs , p.9)*

Excerpt: One important application of our model would be to identify binding TCRs specific to one or more of the peptides from a large data set of irrelevant TCRs obtained, for instance, by repertoire sequencing. We simulated this task by selecting the three most common peptides in the IEDB (GILGFVFTL, GLCTLVAML and NLVPMVATV) which are also part of the MIRA data. Subsequently, we predicted binding of all TCRs in the MIRA data to each of those three peptides, using two different models: one trained on the IEDB data with internal negative data and another trained with additional negative data (derived from TCR sequencing projects and eluted peptide ligands, for details see methods). Subsequently we evaluated how the models could separate positive TCRs binding one of the three peptides from the negative TCRs.  *(NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks , p.7)*

Excerpt: [Image] *(NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks , p.8)*

Object Group

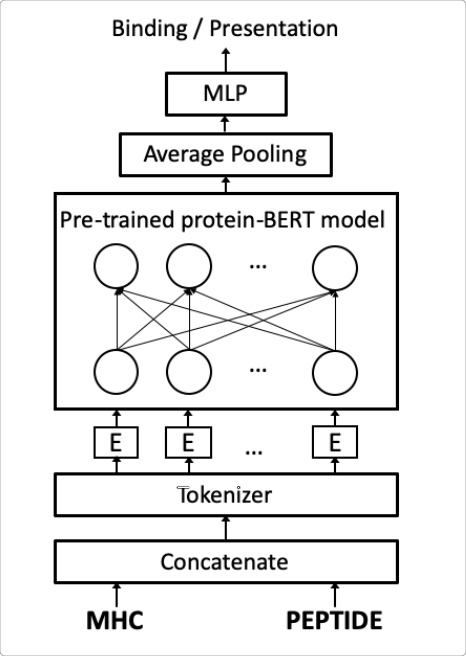
Textbox:

Excerpt: The inputs to the model are peptides and MHC alleles, which are represented as sequences of amino acids. To encode these amino acid sequences, we use a continuous vector representation called embedding (Collobert et al., 2011). Embeddings, when used as the underlying input representations, have been shown to boost the per- formance of various Natural Language Processing (NLP) tasks as they capture the semantic meanings of words in sentences. The in- put, a sequence of amino acids, can be treated as a sentence where the individual words are the amino acids. *(MHCAttnNet: predicting MHC-peptide bindings for MHC alleles classes I and II using an attention-based deep neural model , p.3)*

Textbox:

Object Group

Textbox:

Excerpt: [Image] *(BERTMHC, p.3)*

Object Group

Excerpt: [Image] *()*

Excerpt: To avoid model overfitting and overestimation of model performance, the entire data set was partitioned into 5 sets prior to model training. Prior to partitioning, TCR beta chain (TCRb) CDR3 sequences were compared to each other using blastp and TCRs sharing more than 90% sequence identity, determined by blastp, were kept in the same data partition. Otherwise data points were assigned to partitions randomly.  *(NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks , p.3)*

Excerpt: All models were evaluated using a stratified 5-fold cross-validation strategy. With this validation strategy we obtained the receiver operating characteristic (ROC) curve, precision-recall (PR) curve, the balanced accuracy, the area under the ROC curve (AUC) value and the average precision value for each classifier as implemented in the scikit-learn library (23). Models that had poor AUC (<0.7) or poor average precision (<0.35) values were excluded from the model collection. *(Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires , p.3)*

Excerpt: growing evidence suggests that T cell receptors specific to a common target share common properties (17, 18). Increasing amounts of data are now available linking TCR sequences to their cognate targets. The predictive power of the model proposed here is in line with these observations.  *(NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks , p.12)*