(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 September 2024 (26.09.2024)

WIPO | PCT

(10) International Publication Number
WO 2024/194208 A1

(51) International Patent Classification:
G16B 40/20 (2019.01) *A61K 39/00* (2006.01)
G16B 20/00 (2019.01) *C07K 14/725* (2006.01)
G16B 5/00 (2019.01)

(21) International Application Number:
PCT/EP2024/057046

(22) International Filing Date:
15 March 2024 (15.03.2024)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
2303920.9 17 March 2023 (17.03.2023) GB

(71) Applicant: **ACHILLES THERAPEUTICS UK LIMITED** [GB/GB]; 245 Hammersmith Road, London, W6 8PW (GB).

(72) Inventors: **O'BRIEN, Hugh**; c/o Achilles Therapeutics UK Limited, 245 Hammersmith Road, London W6 8PW (GB). **SALM-HORSTMAR, Maximilian Prinz zu**; c/o Achilles Therapeutics UK Limited, 245 Hammersmith Road, London W6 8PW (GB).

(74) Agent: **MEWBURN ELLIS LLP**; Aurora Building, Counterslip, Bristol BS1 6BX (GB).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:
— *of inventorship (Rule 4.17(iv))*

Published:
— *with international search report (Art. 21(3))*
— *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*

(54) Title: PREDICTION OF IMMUNOGENICITY

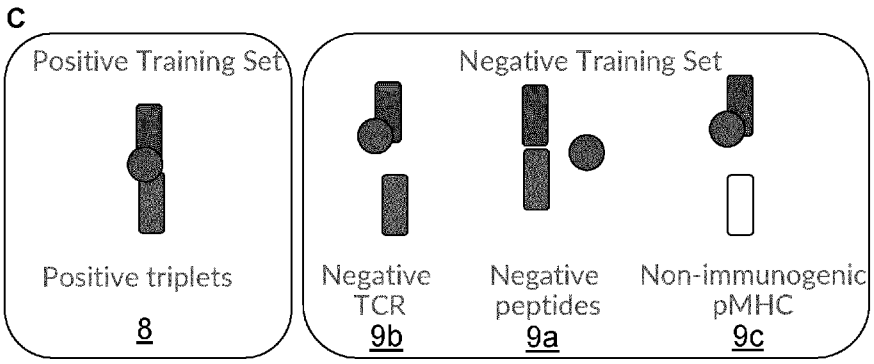


Fig. 1

(57) Abstract: Methods of predicting whether an antigen is likely to be immunogenic are provided, the method comprising using a machine learning model trained to predict a score representing the probability that the antigen is immunogenic in the context of a candidate MHC molecule and a candidate TCR, wherein the machine learning model has been trained using training data comprising amino acid sequences or information derived therefrom for three different types of negative peptide-MHC-TCR triplets Related methods and products are also described.

WO 2024/194208 A1

PREDICTION OF IMMUNOGENICITY

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 964998.

FIELD OF THE DISCLOSURE

The present disclosure relates to methods for predicting whether an antigen is likely to be immunogenic. The present disclosure also relates to methods and compositions for the treatment of diseases which make use of such methods.

BACKGROUND TO THE DISCLOSURE

T lymphocytes (T cells) are essential mediators of immune response to pathogens as well as malignant cells. Activation of T cell requires recognition of the antigens in the context of Major Histocompatibility Complex (MHC) molecules. Thus, successful recognition of antigens by the immune system requires presentation of the antigens by Major Histocompatibility Complex (MHC) molecules, and specific binding of the peptides bound to MHC molecules by a T-cell receptor (TCR).

Tremendous progress has been made in the prediction of peptide-MHC binding and presentation, with algorithms such as NetMHCpan (Reynisson et al., 2020) and MHCflurry (O'Donnell et al., 2020) achieving high performance predictions of peptide-MHC binding from sequence alone, thanks to increasing amounts of data available from *in vitro* binding assays as well as immunopeptidomics (which detect antigen presentation rather than just binding). However, when using these tools to identify peptides that may be immunogenic, very high error rates are still observed, because these predictions do not capture the TCR recognition part that is essential to immunogenicity.

In recent years, a number of methods have been proposed to attempt to predict immunogenicity, taking into account both the MHC and the TCR part of the immunogenicity process (see e.g. PMTnet (Lu et al. 2021) and Imrex (Moris et al., 2020)).

SUMMARY OF THE INVENTION

The present inventors have identified that existing methods for predicting immunogenicity of a peptide, while seemingly achieving good accuracies on specific datasets, suffer from a lack of generalisability. Thus, the inventors recognised that there was a need for improved methods for predicting whether an antigen is likely to be immunogenic. The inventors postulated that such methods would benefit from taking into account both the TCR and MHC molecule sequence when considering a particular peptide, and further that the generalisability problem was at least in part due to biases and a lack of diversity in the negative data used to train the models. Thus, the present inventors have developed a new method for predicting whether an antigen is likely to be immunogenic that addresses one or more of the problems of prior art approaches. The method

uses a machine learning approach to predict immunogenicity from the sequence of a peptide, MHC molecule and TCR, where the machine learning model has been trained using three different types of native training data that each capture a different aspect of the antigen-recognition process. The present inventors demonstrated that this approach outperformed all prior art approaches at predicting immunogenicity of peptides. The effect was particularly stringent when the methods were evaluated using test data specifically filtered to exclude data present in the training data of the models (i.e. specifically designed to test the generalisability of the methods).

Thus, according to a first aspect, there is provided a method of predicting whether an antigen is likely to be immunogenic, the method comprising: obtaining, by a processor, a triplet of sequences comprising: an amino acid sequence of a peptide encoding the antigen, an amino acid sequence of a candidate MHC molecule or a part thereof, and an amino acid sequence of a candidate T cell receptor (TCR) beta chain and/or alpha chain or a part thereof; and providing, by said processor, the triplet of sequences or information derived therefrom as inputs to a machine learning model trained to predict a score representing the probability that the antigen is immunogenic in the context of the candidate MHC molecule and the candidate TCR. The machine learning model is a machine learning model that has been trained using training data comprising amino acid sequences or information derived therefrom for negative peptide-MHC-TCR triplets comprising: a. a first set of one or more peptide-MHC-TCR triplets each comprising: (i) a TCR-MHC pair comprising an MHC molecule and a TCR chain or chains known to bind the MHC molecule (positive TCR-MHC pair), and (ii) a peptide not known to interact with the TCR-MHC pair; b. a second set of one or more peptide-MHC-TCR triplets each comprising: (i) a peptide-MHC pair comprising an MHC molecule and a peptide known to bind the MHC molecule (positive peptide-MHC pair), and (ii) a TCR chain or chains not known to interact with the peptide-MHC pair, wherein the peptide-MHC pair has been previously found to interact with a TCR (immunogenic positive peptide-MHC pair); and c. a third set of one or more peptide-MHC-TCR triplets each comprising: (i) a peptide-MHC pair comprising an MHC molecule and a peptide known to bind the MHC molecule (positive peptide-MHC pair), and a TCR chain or chains not known to interact with the peptide-MHC pair, wherein the peptide-MHC pair has been previously found to not be immunogenic (non-immunogenic positive peptide-MHC pair).

The use of a negative training set that captures different aspects of the biology of immunogenicity results in a model that has improved performance in predicting immunogenicity, particularly when confronted with molecules (especially peptides) that are not part of the data set on which it has been trained (improved generalisability). In this context, the present inventors postulated that a training data set that captures the following aspects would be particularly beneficial: (i) peptides and MHC molecules can effectively bind without the resulting complex being immunogenic, (ii) the same TCR and MHC molecules can interact to trigger an immune reaction in the presence of some peptides but not others, and (iii) peptides and MHC molecules can effectively bind to result in a

complex that has the potential to be immunogenic given the right TCR. Thus, the method results in a more robust prediction. Robustness / generalisability is an extremely important aspect of the performance of a model for predicting immunogenicity. Indeed, in real clinical situations the diversity of peptides and TCR sequences is extremely high, and a model that performs well on a restricted (and biased) test set may in fact perform poorly on "real" data. This is problematic not only because it results in bad predictions, but also because without a proper assessment of the model with data outside of its training distribution, one may not be aware that the model's predictions are unreliable. By contrast, the inventors have described herein models that are designed to generalise well, and have rigorously benchmarked these against other models to verify that this is the case. Even at comparative performance on the respective sets described in the original publications for comparative models, the data shown herein demonstrates that in a realistic rigorous benchmarking setting the performance of immunogenicity prediction models in the prior art is in fact lower than expected, and lower than the models of the present disclosure. The method of the present aspect may have one or more of the following features.

The terms "peptide" and "antigen" may be used interchangeably to refer to the peptide that comprises the sequence of the antigen. A peptide may be considered likely to be immunogenic if the probability is above a predetermined threshold. The predetermined threshold may be identified using test data comprising triplets with known immunogenicity status, for example by selecting a probability threshold that results in a desired sensitivity and/or specificity of classification of triplets as immunogenic vs non-immunogenic.

The first, second and/or third sets of negative peptide-MHC-TCR triplets may have been derived from amino acid sequences or information derived therefrom for a plurality of positive peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response. The positive peptide-MHC-TCR triplets may also be comprised in the training data used to train the machine learning model. Thus, the machine learning model may have been trained using training data further comprising amino acid sequences or information derived therefrom for a plurality of positive peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response. The positive peptide-MHC-TCR triplets used to derive the first, second and/or third sets of negative peptide-MHC-TCR triplets may be a subset (including a complete subset) of the positive peptide-MHC-TCR triplets used to train the machine learning model. In other words, the training data may comprise positive triplets and negative triplets derived from the positive triplets.

The TCR chain or chains in the second set may have been selected from a database or reference dataset. The TCR chain or chains in the second set may have been selected from a source other than the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response

(positive peptide-MHC-TCR triplets). The TCR chain or chains and the peptide-MHC pairs in the second set may not form a triplet that is present in the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets). The peptide-MHC pairs in the second set may have been selected from the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets). The second set of negative triplets may be generated using TCR sequences from a database or other data source, rather than by shuffling the TCR chains from the positive data. This increases the diversity of the negative set, widens the distribution of TCRs that the model is trained to recognise, and reduces the risk of false negative labelling where TCRs (or at least TCRs with specific CDR3 sequences) are able to recognise multiple peptide-MHC complexes. This in turn improves the performance of the trained model on any test data set, but particularly on test datasets that do not comprise subsets of the training data.

The peptides in the first set may have been selected from a database or reference dataset. The peptides in the first set may have been randomly selected from a reference proteome. The peptides in the first set may have been selected from a source other than the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets). The peptides and the TCR-MHC pairs in the first set may not form a triplet that is present in the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets). The TCR-MHC pairs in the first set may have been selected from the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets). The first set of negative triplets may have been generated using peptide sequences from a database or other data source, rather than by shuffling the peptides from the positive data. This increases the diversity of the negative set, and improves the performance of the trained model on any test data set, particularly on test datasets that do not comprise peptides similar to those in the training data. In other words, this results in a model with higher and/or more robust performance.

The TCR chain or chains in the third set may have been selected from the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets). The third set of negative triplets may have been generated using peptide-MHC combinations which have been found to be non-reactive in one or more immunogenicity experiments. The third set of negative triplets may have been generated using TCR sequences

sampled from the positive triplets. This captures the fact that peptide-MHC pairs that do form stable complexes may be non-reactive, even when paired with TCR sequences that are known to be reactive. This helps the model to learn differences between peptide-MHC binding and immunogenicity in the context of TCRs that are otherwise known to bind peptide-MHC complexes.

The training data may comprise a ratio of negative triplets to positive triplets of at least 100:1, at least 150:1, between 100:1 and 300:1, preferably between 150:1 and 250:1, or around 200:1. The use of higher negative to positive ratios in the training data increases the diversity in the dataset and provides a more realistic classification situation. Indeed, when assessing the immunogenicity of a peptide, for example in a clinical situation, a large number of candidate TCRs are likely to be considered, whether from a database or reference dataset or from TCR repertoire sequencing from a patient. Thus, at each training iteration, the model may take as input a triplet sampled from the training data, which may be a negative triplet or a positive triplet, but where negative triplets are more likely to be sampled than positive triplets (e.g. 100, 150 or 200 times more likely to be sampled). A loss function may be evaluated based on the prediction for the triplet, and this may be used to update the parameters of the model as known in the art.

The training data may comprise similar proportions of negative triplets from the first, second and third sets. For example, the training data may comprise approximately as many negative triplets in the first, second and third sets. For example, approximately 33% of the total number of negative triplets may belong to the first set, approximately 33% may belong to the second set, and approximately 33% may belong to the third set. Variations around these proportions are envisaged. For example, each of the first, second and third sets of negative triplets may respectively represent between 10 and 40% of the total number of negative triplets, provided that the total does not exceed 100%. The negative triplets may comprise additional types of negative triplets, in which case each of the first, second and third sets of negative triplets may represent at least 10%, at least 20% or at least 30% of the total number of negative triplets. The negative triplets may not comprise additional types of negative triplets, such all negative triplets belong to one of the first, second or third sets.

The machine learning model may take as input an amino acid sequence comprising a part of the variable region of one or more chains of a TCR, or information derived therefrom. The machine learning model may take as input an amino acid sequence comprising one or more CDRs of one or more chains of a TCR, or information derived therefrom. The machine learning model may take as input an amino acid sequence comprising or consisting of the CDR3 sequence of one or more chains of a TCR, or information derived therefrom. The machine learning model may take as input an amino acid sequence comprising or consisting of the sequence of the CDR3 region of the alpha and/or beta chain of a TCR, or information derived therefrom. The machine learning model may take as input an amino acid sequence comprising or consisting of the sequence of the CDR3 region of the beta chain of a TCR.

The machine learning model may take as input the triplet of amino acid sequences and produce an encoding for each sequence. The machine learning model may take as input an encoding for each sequence of a triplet of amino acid sequences. The amino acid sequences may be encoded using encoding schemes selected from: a predetermined token for each amino acid and optionally a padding character, one-hot-encoding, an encoding using a substitution matrix, an encoding using an embedding matrix, and an encoding using physicochemical descriptors. One or more of the amino acid sequences may be encoded as fixed length strings with a token for each amino acid and a padding character. The TCR sequence may be encoded as a fixed length string. The peptide sequence may be encoded as a fixed length string. The MHC sequence may be encoded as a sequence or pseudosequence with fixed length. A fixed length string for the TCR sequence may have a length of 26 tokens. A fixed length string for the peptide may have a length of at least 7, 8, 9, 10, 11, 12 or 12 tokens, such as e.g. a length of 16 tokens. The MHC molecule may also be encoded as a fixed length string. A fixed length string for the MHC molecule may have a length of 34 tokens. Reference to a “fixed length” may refer to the input having a predetermined length, such that a candidate sequence that has a length shorted than the predetermined length may be input with padding characters, and a candidate sequence that has a length longer than the predetermined length may be input as a pseudosequence (e.g. selecting a predetermined subset of the positions in the sequence) and/or a set of sequences of length equal to the predetermined length (e.g. by tiling). A suitable predetermined length for the TCR sequence (respectively, the peptide) may be the maximum length (or maximum length with a frequency above a predetermined threshold, e.g. maximum length that occurred in at least 1% of a predetermined data set) observed for a TCR sequence (respectively, peptide sequence) in a data set of paired peptide-TCR sequences (or triplets), such as e.g. the maximum length observed in the training data.

The machine learning model may be a deep learning model. The machine learning model may comprise one or more natural language processing models. The machine learning model may comprise a first encoder or pair of encoders for encoding the TCR sequence, and a second encoder for encoding the peptide and MHC sequences. The first and/or second encoders may have been pretrained prior to training the machine learning model using the training data comprising the negative triplets. The machine learning model may have been trained using the training data comprising the negative triplets with the parameters of the first and/or second encoders maintained to their pretrained values. Alternatively, the training of the machine learning model using the training data comprising the negative triplets may have included fine-tuning the parameters of one or more of the encoders. Thus, the machine learning model may be trained using the training data comprising the negative triplets with the parameters of the first and/or second encoders frozen, or the training may include fine-tuning of the parameters of one or more of the encoders. The first encoder or pair of encoders may comprise a single encoder taking as input a TCR beta chain or a part thereof. The first encoder or pair of encoders may comprise a single encoder taking as input a part of a TCR beta chain comprising or consisting of the CDR3

region. The first encoder or pair of encoders may comprise a single encoder taking as input the concatenation of a TCR beta chain or a part thereof and a TCR alpha chain or a part thereof. The first encoder or pair of encoders may comprise a pair of encoders taking as input respectively a TCR beta chain or a part thereof and a TCR alpha chain or a part thereof. A part of a TCR chain may comprise or consist of the CDR3 region of the respective chain. The first encoder or pairs of encoders may have been trained in a self-supervised manner to encode TCR sequences or parts thereof. Training in a self-supervised manner can be used, e.g. a random masking task (also referred to as a masked language modelling task). The first encoder or pair of encoders may have been trained in a self-supervised manner using random masking. The first encoder may have been pretrained as part of a machine learning model comprising the encoder and one or more fully connected layers. The one or more fully connected layers may be configured to reconstruct the input of the model from the output of the encoder. When a pair of encoders are used, each encoder may have been independently trained in this manner. When a pair of encoders are used, the encoders may have been trained separately using training data comprising TCR beta chain sequences or parts thereof and TCR alpha chain sequences or parts thereof, respectively. Alternatively, a first encoder of the pair of encoders may have been trained using training data comprising sequences for a first TCR chain or parts thereof, then the second encoder of the pair of encoders may have been trained using training data comprising sequences for the second TCR chain or parts thereof and the weights of the first encoder as starting weights for the training. In other words, the second encoder may be trained by transfer learning using the first encoder. The first encoder can be the TCR beta chain encoder (as TCR beta chain sequences are typically available in large amounts for training than TCR alpha chain sequences).

The second encoder may take as input a peptide sequence and an MHC sequence. The second encoder may have been trained as part of a model trained to predict whether the peptide is likely to bind the MHC molecule, whether the peptide is likely to be presented by the MHC molecule, and/or whether the peptide and MHC molecule are likely to form a stable complex. The second encoder may have been trained as part of a model trained to predict the binding affinity between a peptide sequence and an MHC molecule corresponding to the MHC sequence. This may be referred to as binding affinity prediction. The binding affinity may be predicted as a normalised binding affinity metric with a value between 0 and 1. The second encoder may have been trained as part of a model trained to classify pairs comprising a peptide sequence and an MHC sequence between a first class comprising peptide-MHC pairs known to bind to each other and be presented on the surface of cells, and a second class comprising peptides-MHC pairs that are not expected to bind to each other and be presented on the surface of cells. This may be referred to as training for eluted ligand prediction. The use of a pretrained encoder for encoding the peptide-MHC (second encoder) as part of a model that performs eluted ligand prediction is particularly advantageous as it results in a pretrained model that can reliably produce encodings that distinguish peptides that would be very unlikely to be presented as peptide-MHC complexes. This

is by contrast to TCR binding models of the prior art which are not as good at identifying those peptides. The second encoder may have been trained as part of a model trained using training data comprising peptide-MHC pairs known to bind to each other on the basis of the peptide being an eluted ligand for the MHC molecule from which the MHC sequence is derived (e.g. as determined using mass spectrometry identification of eluted ligands from samples with known MHC molecules, also referred to as "eluted ligand" data or "immunopeptidomics" data). Such peptides may be expected to have strong binding affinity to their cognate MHC molecule (such as e.g. <500 nM). The second encoder may have been trained as part of a model trained using training data comprising peptide-MHC pairs known to bind to each other, such as peptide-MHC molecules with a binding affinity lower than a predetermined threshold (e.g. <500 nM) as determined using any experimental method known in the art. The second encoder may have been trained as part of a model trained using training data comprising peptide-MHC pairs that are not expected to bind to each other, such as e.g. peptides having very weak binding affinity (e.g. >30,000 nM) as determined using any experimental method known in the art, such as e.g. immunopeptidomics. For example, the data used in O'Donnell et al., 2020 and Reynisson et al. 2020 comprises negative mass spectrometry data from eluted ligand experiments. The second encoder may have been trained as part of a model trained to predict a metric indicative of the stability of a complex comprising the peptide and MHC molecule corresponding to the MHC sequence. This may be referred to as binding stability prediction. The metric indicative of the stability of a complex comprising the peptide and MHC molecule corresponding to the MHC sequence may be a normalised metric with a value between 0 and 1. The second encoder may take as input a peptide sequence and an MHC sequence, and may produce as output a probability that the peptide and MHC sequence bind to each other (eluted ligand prediction), a normalised binding affinity metric, or a normalised metric indicative of the stability of the complex comprising the peptide and MHC molecule corresponding to the MHC sequence. Normalised metrics may be metrics normalised to take values between 0 and 1. The second encoder may have been trained as part of a model that has been: (i) pretrained to predict whether the peptide is likely to bind the MHC molecule, and/or whether the peptide is likely to be presented by the MHC molecule; and (ii) pre-trained, optionally after step (i), for predicting whether the peptide and MHC molecule are likely to form a stable complex. At step (i) the second encoder may have been trained as part of a model that has been pretrained to predict whether the peptide is likely to bind the MHC molecule, then further trained using transfer learning to predict whether the peptide is likely to be presented by the MHC molecule. A model trained to predict whether the peptide and MHC molecule are likely to form a stable complex may be a model configured to take as input a peptide and MHC sequence or information derived therefrom and produce as output a metric indicative of the stability of a complex comprising the peptide and MHC molecule corresponding to the MHC sequence. The metric may be a half-life or scaled (also referred to as "normalised") half-life.

A peptide and MHC sequence may be provided as a concatenated sequence or as separate input sequences. The probability that the peptide and MHC sequence bind to each other may represent a probability that the peptide and MHC sequence pair is identified in an eluted ligand experiment (this may also be referred to as the probability that the peptide-MHC pair is an eluted ligand). The second encoder may comprise a first input branch that encodes the peptide sequence and a second input branch that encodes the MHC sequence. Each input branch may be individually referred to as an "encoder". When these encoders are transformer based, they may be referred to as "self-attention" models. Optionally, the input to the first and/or second branches may itself be the output of an encoder, such as e.g. a model that has been trained for general purpose encoding of proteins such as protGPT2 (Ferruz et al., 2022). Alternatively, the input to the first and/or second branches may be amino acid sequences or pseudosequences for the peptide and/or MHC molecule. In the embodiments exemplified herein, the input to the first and second branches are a peptide amino acid sequence and a MHC amino acid pseudosequence, respectively. In other words, an additional encoding model has not been found to be necessary to obtain a benefit over the prior art. The output of the first and second input branches may be concatenated and input into a further model that is configured to learn from both encoded sequences. The further model may also be seen as an encoder in that it encodes the information in the pair of peptide and MHC sequences. When the further model is a transformer based model this may be referred to as a "cross attention" model. The first and second input branches and the further model may together form the second encoder. The output of the further model may be provided as input to one or more fully connected layers configured to predict whether the peptide is likely to bind the MHC molecule / the probability that the peptide and MHC sequence bind to each other / classify pairs comprising a peptide sequence and an MHC sequence between a first class comprising peptide-MHC pairs known to bind to each other (such as e.g. on the basis of the peptide being an eluted ligand for the MHC molecule from which the MHC sequence is derived) and a second class comprising peptides-MHC pairs that are not expected to bind to each other. Thus, the second encoder may be pretrained as part of a model comprising the first and second input branches, the further model (together forming the second encoder) and the one or more fully connected layers. In other words, after pretraining of a model comprising the first and second input branches, the further model and the one or more fully connected layers, the one or more fully connected layers may be removed to obtain the second encoder used in the machine learning model.

The encoders may each independently be selected from: transformer-based encoders, autoencoders, and recurrent neural network encoders such as long-short-term memory (LSTM) networks, and/or wherein the encoders are transformer-based encoders. An autoencoder may be an autoencoder with one or more convolutional layers. The present inventors have found transformer-based encoders to perform particularly well.

The machine learning model may further comprise a deep learning block that takes as input the concatenated outputs of the first and second encoders, and produces as output the probability that the antigen is immunogenic in the context of the candidate MHC molecule and the candidate TCR. The deep learning block may comprise a first block that learns from the combined outputs of the first and second encoders. The deep learning block may comprise a second block comprising one or more fully connected layers producing a single numerical output and optionally a sigmoid activation function. The first block may comprise a deep artificial neural network model and/or a natural language processing model. The natural language processing model may be a transformer-based model. The present inventors tested the use of both a deep ANN and a transformer-based model for the first block, and found both to perform satisfactorily. A transformer-based model was found to perform slightly better and was thus selected in the particular models exemplified.

The machine learning model may have been trained by fine tuning a peptide-MHC immunogenicity prediction model, wherein a peptide-MHC immunogenicity model is a machine learning model that has been trained to take as input a doublet of sequences comprising an amino acid sequence of a peptide encoding the antigen, and an amino acid sequence of a candidate MHC molecule or a part thereof, or information derived from the doublet of sequences, and provide as output a score representing the probability that the antigen is immunogenic in the context of the candidate MHC molecule. The peptide-MHC immunogenicity model may have been trained using training data comprising amino acid sequences or information derived therefrom for (i) positive peptide-MHC doublets comprising a peptide and MHC sequences that have been experimentally demonstrated to form an immunogenic complex; and (ii) negative peptide-MHC doublets comprising: (a) a first set of one or more peptide-MHC doublets each comprising: a MHC molecule selected from the positive peptide-MHC doublets and a peptide sequence not known to interact with the selected MHC molecule, optionally a randomly sampled peptide sequence, and (b) a second set of one or more peptide-MHC doublets each comprising: a peptide-MHC pair comprising an MHC molecule and a peptide known to bind the MHC molecule (positive peptide-MHC pair), wherein the peptide-MHC pair has been previously found to not be immunogenic (non-immunogenic positive peptide-MHC pair). The peptide-MHC immunogenicity model may have any of the features described herein in relation to the machine learning model. In particular, the peptide-MHC immunogenicity model may comprise a peptide-MHC encoder having any of the features of a peptide-MHC encoder (second encoder) described herein. The peptide-MHC immunogenicity model may comprise a peptide-MHC encoder, the output of which is provided to a classification head (e.g. comprising a transformer layer and classification block). The machine learning model may comprise the peptide-MHC encoder of the peptide-MHC immunogenicity model and the classification block of the peptide-MHC immunogenicity model, and a TCR encoder (first encoder). The outputs of the first and second encoders may be concatenated, provided as input to a transformer layer, the output of which is provided to a classification block which is initialised using the trained weights of the peptide-immunogenicity model. The same architecture can be used without initialisation of the

classification block weights using weights from a peptide-MHC immunogenicity model. However, the use of such an initialisation step was found to advantageously enable the training of the machine learning model to benefit from exposure to a broader variety of training data (particularly in relation to peptide diversity) than would be possible using training of the immunogenicity (classification) block using triplet data alone. In embodiments, obtaining the triplet of sequences comprises obtaining an amino acid sequence of a peptide encoding the antigen, an amino acid sequence of a candidate MHC molecule or a part thereof, and an empty TCR sequence vector as candidate TCR, and the score predicted by the machine learning model represents the probability that the antigen is immunogenic in the context of the candidate MHC molecule and an unknown TCR.

The method may further comprise: (i) repeating one or more times the steps of: obtaining, by said processor, a triplet of sequences comprising: an amino acid sequence of a peptide encoding the antigen, an amino acid sequence of a candidate MHC molecule or a part thereof, and an amino acid sequence of a candidate T cell receptor (TCR) beta chain and/or alpha chain or a part thereof; and providing, by said processor, the triplet of sequences or information derived therefrom as inputs to the machine learning model trained to predict a score representing the probability that the antigen is immunogenic in the context of the candidate MHC molecule and the candidate TCR, wherein each triplet of sequences differs in the amino acid sequence of the candidate MHC molecule or part thereof, and/or in the amino acid sequence of the candidate T cell receptor (TCR) beta chain and/or alpha chain or part thereof, thereby obtaining a plurality of respective probabilities that the antigen is immunogenic; and (ii) selecting the highest of the plurality of probabilities as the probability that the antigen is immunogenic. The method may further comprise identifying the antigen from a sample. The method may comprise performing the method of any preceding embodiment using one or more candidate MHC molecules and/or one or more candidate TCR molecules identified from a sample. The sample may be a sample from which the antigen has been identified or a related sample. The sample may have been previously obtained from a subject. The subject may be a human subject. The subject may be a mammalian subject. The subject may be a subject who has been diagnosed as having cancer or being likely to have cancer. The sample may be tumour sample. A related sample may be a sample previously obtained from the same subject from which another sample has been previously obtained. A related sample may be a tumour sample (such as e.g. a tumour biopsy or sample comprising circulating tumour DNA or circulating tumour cells) or a normal sample (such as e.g. a blood sample). Identifying the antigen from the sample may comprise analysing DNA and/or RNA sequence data from the sample. Identifying the antigen from the sample may comprise obtaining DNA and/or RNA sequence data from the sample. Identifying candidate MHC and/or TCR molecules may comprise analysing DNA and/or RNA sequence data from the sample. Identifying candidate MHC and/or TCR molecules may comprise obtaining DNA and/or RNA sequence data from the sample.

According to a second aspect, there is provided a method of providing a tool for predicting whether an antigen is likely to be immunogenic, the method comprising: (i) obtaining, by a processor, a training dataset comprising amino acid sequences or information derived therefrom for a plurality of peptide-MHC-TCR triplets, each triplet comprising an amino acid sequence of a peptide encoding the antigen, an amino acid sequence of a candidate MHC molecule or a part thereof, and an amino acid sequence of a candidate T cell receptor (TCR) beta chain and/or alpha chain or a part thereof; and (ii) training, using said training data, a machine learning model that predicts the probability that an antigen is immunogenic in the context of a candidate MHC molecule and a candidate TCR provided as a triplet of sequences or information derived therefrom as input to the machine learning model. The plurality of peptide-MHC-TCR triplets comprise: a. a first set of one or more peptide-MHC-TCR triplets each comprising: (i) a TCR-MHC pair comprising an MHC molecule and a TCR chain or chains known to bind the MHC molecule (positive TCR-MHC pair), and (ii) a peptide not known to interact with the TCR-MHC pair; b. a second set of one or more peptide-MHC-TCR triplets each comprising: (i) a peptide-MHC pair comprising an MHC molecule and a peptide known to bind the MHC molecule (positive peptide-MHC pair), and (ii) a TCR chain or chains not known to interact with the peptide-MHC pair, wherein the peptide-MHC pair has been previously found to interact with a TCR (immunogenic positive peptide-MHC pair); and c. a third set of one or more peptide-MHC-TCR triplets each comprising: (i) a peptide-MHC pair comprising an MHC molecule and a peptide known to bind the MHC molecule (positive peptide-MHC pair), and a TCR chain or chains not known to interact with the peptide-MHC pair, wherein the peptide-MHC pair has been previously found to not be immunogenic (non-immunogenic positive peptide-MHC pair).

The method of the present aspect may have any one or more of the following features. The method according to the present aspect may have any of the features disclosed in relation to the first aspect. In particular, references to features of the trained model in relation to the first aspect may be interpreted as active steps of training the model in relation to the present aspect. The method according to the present aspect may further comprise performing the method of any embodiment of the first aspect. Thus, also envisaged herein is a method comprising providing and using a tool as described herein. Further, any features of the model architecture (e.g. configuration of the various parts of the model, characteristics of inputs and outputs, etc.), training or training data apply equally to the present aspect. For example, the method may comprise obtaining the first, second and/or third sets of negative peptide-MHC-TCR triplets using amino acid sequences or information derived therefrom for a plurality of positive peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response. Obtaining the training data may further comprise obtaining amino acid sequences or information derived therefrom for a plurality of positive peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response. Obtaining the second set of triplets may

comprise selecting the TCR chain or chains in the second set from a database or reference dataset. Obtaining the second set of triplets may comprise selecting the TCR chain(s) in the second set, wherein selection is not from the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets), or not from the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets). Obtaining the second set of triplets may comprise selecting the peptide-MHC pairs in the second set from the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets). Obtaining the first set of triplets may comprise selecting the peptides in the first set from a database or reference dataset, such as by randomly selecting peptides from a reference proteome. Obtaining the first set of triplets may comprise selecting the peptides, wherein the selecting is not from the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets). Obtaining the first set of triplets may comprise selecting the TCR-MHC pairs in the first set from the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets). Obtaining the third set may comprise selecting the TCR chain or chains in the third set from the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets).

Obtaining the training data may comprise obtaining positive triplets and obtaining the first, second and third sets (negative triplets) such that the training data comprises a ratio of negative triplets to positive triplets of at least 100:1, at least 150:1, between 100:1 and 300:1, preferably between 150:1 and 250:1, or around 200:1.Thus, obtaining the training data may comprise obtaining at least 100, at least 150, between 100 and 300, between 150 and 250, or about 200 times more negative triplets than positive triplets in the training data.

The machine learning model may comprise a first encoder or pair of encoders for encoding the TCR sequence, and a second encoder for encoding the peptide and MHC sequences. The method may comprise pretraining the encoders prior to training the machine learning model using the training data comprising the negative triplets. Training the machine learning model may comprise training the model with the parameters of the encoders maintained to their pretrained values. Pretraining the encoders may comprise training the first encoder or pairs of encoders in a self-supervised manner to encode TCR sequences or parts thereof, optionally using random masking. The method may comprise pretraining the second encoder as part of a model trained to predict

whether the peptide is likely to bind the MHC molecule and be presented by cells, trained to predict the binding affinity between the peptide and MHC molecule, and/or trained to predict the stability of a complex comprising the peptide and MHC molecule corresponding to the MHC sequence. The method may comprise pretraining the second encoder: (a) as part of a model trained to classify pairs comprising a peptide sequence and an MHC sequence between a first class comprising peptide-MHC pairs known to bind to each other and a second class comprising peptides-MHC pairs that are not expected to bind to each other, (b) as part of a model trained to predict a binding affinity metric for the peptide-MHC pair, and/or (c) as part of a model trained to predict a stability metric for the peptide-MHC pair.

According to a third aspect, there is provided a method of identifying one or more tumour-specific peptides that are likely to be immunogenic, the method comprising: obtaining the amino acid sequence of one or more candidate tumour-specific peptides derived from one or more tumour-specific mutations previously identified in a tumour; and determining whether the one or more candidate peptides are likely to be immunogenic using the method of any of embodiment of the first aspect. The method may further comprise selecting one or more of the tumour-specific peptides as peptides likely to be immunogenic using one or more criteria applying to the results of the step of determining whether the one or more candidate peptides are likely to be immunogenic. The one or more criteria may be selected from: selecting peptides with a probability above a predetermined threshold, and selecting a predetermined number or proportion of tumour-specific peptides with the highest probability amongst a plurality of peptides for which immunogenicity was predicted.

According to a fourth aspect, there is provided a method of characterising an immunogenic composition comprising a plurality of candidate peptides or sequences encoding a plurality of candidate peptides, the method comprising: determining whether the one or more candidate peptides are likely to be immunogenic using the method of any embodiment of the first aspect, and identifying which one or more of the candidate peptides are likely to be immunogenic by applying one or more predetermined criteria to the results of the determining. The immunogenic composition may comprise a plurality of candidate peptides. The immunogenic composition may be a peptide vaccine. The immunogenic composition may be a composition for use in obtaining a population of cells that display the candidate peptides, and/or a composition for use in obtaining a population of cells that are reactive to one or more of the candidate peptides.

According to a fifth aspect, there is provided a method of designing or providing an immunotherapy for a subject that has been diagnosed as having cancer, the method comprising: obtaining a set of one or more candidate neoantigens for the subject, wherein the one or more candidate neoantigens were identified using a process comprising analysing one or more samples from the subject comprising tumour genetic material; and designing an immunotherapy that targets one or more of the neoantigens identified, wherein the designing comprises identifying at least one

peptide encoding at least one of the candidate neoantigens that is immunogenic using the method of any embodiment of the first aspect.

The methods according to the present aspect may have any one or more of the following optional features.

The one or more neoantigens may be clonal neoantigens. The immunotherapy that targets the one or more of the neoantigens may be an immunogenic composition, a composition comprising immune cells or a therapeutic antibody. The method may further comprise producing one or more peptides selected from the identified peptides. The method may further comprise producing one or more sequences encoding one or more peptides selected from the identified peptides. The method may further comprise producing an immunotherapy using one or more peptides selected from the identified peptides. The method may further comprise identifying one or more cancer neoantigens for the subject, wherein the one or more candidate neoantigens are identified using a process comprising analysing one or more samples from the subject comprising tumour genetic material. Also provided are methods for expanding a T cell population for use in the treatment of cancer in a subject using a neoantigen identified as immunogenic using a method described herein.

Also described are compositions comprising a population of T cells obtained or obtainable by such a method. Also described are compositions comprising a neoantigen peptide, a sequence encoding a neoantigen peptide, a neoantigen peptide specific immune cell, or an antibody that recognises a neoantigen peptide, for use in the treatment or prevention of cancer in a subject, wherein said neoantigen peptide has been identified using the methods described herein. Also described are uses of said products and compositions in the manufacture of a medicament for use in the treatment or prevention of cancer in a subject, and methods of treating a subject that has been diagnosed as having cancer, the method comprising administering an immunotherapy that has been provided using the methods described herein, or a composition as described herein.

According to a sixth aspect, there is provided a method of treating a subject that has been diagnosed as having cancer, the method comprising administering an immunotherapy that has been provided using the method of any embodiment of the fifth aspect. The method may comprise providing the immunotherapy for the subject using the method of any embodiment of the fifth aspect.

According to a further aspect, there is provided a system comprising: a processor; and a computer readable medium comprising instructions that, when executed by the processor, cause the processor to perform the steps of any method described herein, such as a method according to any embodiment of the first, second, third, fourth or fifth aspects above.

According to a further aspect, there is provided one or more computer readable media comprising instructions that, when executed by one or more processors, cause the one or more processors to

perform the steps of any method described herein, such as a method according to any embodiment of the first, second, third, fourth or fifth aspects above.

According to a further aspect, there is provided a computer program comprising code which, when the code is executed on a computer, causes the computer to perform the steps of any method described herein, such as a method according to any embodiment of the first, second, third, fourth or fifth aspects above.

As the skilled person understands, the complexity of the operations described herein (due at least to the complexity of, and the amount of data and computation that is typically required for training of deep learning models as used herein, as well as the size (number of parameters and computations) of such models) are such that they are beyond the reach of a mental activity. For example, models as demonstrated herein may be trained using over 40,000 different triplets of sequences identified experimentally as positive (each having lengths of up to multiple dozens of amino acids that are individually encoded), from which over 8 million negative triplets are generated and used to train a model. Further, training of models as demonstrated herein involves determining the value of 700,000 to 800,000 parameters. Thus, unless context indicates otherwise (e.g. where sample preparation or acquisition steps are described), all steps of the methods described herein are computer implemented.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 illustrates schematically the interaction between a T cell and an antigen presenting cell (APC) presenting a peptide in the context of a MHC class I molecule (A), the concept of predicting whether an antigen is likely to be immunogenic (B), and the training data used to obtain tools that can predict whether an antigen is likely to be immunogenic according to the present disclosure.

Figure 2 is a flowchart illustrating schematically a method of predicting whether an antigen is likely to be immunogenic and a method of providing a tool according to the disclosure.

Figure 3 is a flowchart illustrating schematically a method of designing or providing an immunotherapy for a subject.

Figure 4 shows an embodiment of a system for predicting whether an antigen is likely to be immunogenic and/or for providing an immunotherapy.

Figure 5 shows schematically the process of obtaining training data used in an example of the present disclosure.

Figure 6 shows schematically a model architecture that can be used to predict immunogenicity according to an example of the present disclosure.

Figure 7 shows a ROC (receiver operating characteristic) curve illustrating the performance of a model as described herein. The performance of the model was assessed on a test set comprising

827626 triplets of which 4148 are positive triplets. Continuous line=method of the disclosure ("Genesis") (AUC=0.758), dashed line=chance prediction.

Figure 8 shows ROC (receiver operating characteristic) curves illustrating the performance of a model as described herein, compared to prior art models predicting CDR3-pMHC interactions. A. The performance of the models was assessed on a test set comprising 44590 triplets of which 1829 are positive triplets, restricted to triplets not contained in any model's training datasets. pmtNet (Lu et al., 2021) AUC=0.565, imrex (Moris et al., 2020) AUC=0.548, ERGO (Springer et al., 2021) AUC=0.590, model of the present disclosure AUC=0.633. B. The performance of the models was assessed on a test set comprising 439887 triplets of which 789 are positive triplets, restricted to epitope sequences not contained in any model's training dataset. pmtNet AUC=0.502, imrex AUC=0.525, ERGO AUC=0.512, model of the present disclosure (Genesis) AUC=0.737. Dashed lines=chance prediction.

Figure 9 shows a ROC (receiver operating characteristic) curve illustrating the performance of a model as described herein, compared to prior art models predicting peptide-MHC interactions. The performance of the models was assessed on a test set comprising 11267 doublets of which 217 are positive doublets, restricted to epitope sequences not contained in any model's training datasets. bigMHC (Albert et al., 2022) AUC=0.800, DeepAttentionPan (Jin et al., 2021) AUC=0.727, NetMHCpan (Reynisson et al., 2020) elution prediction AUC=0.534 and affinity prediction AUC=0.728, MHCflurry (O'Donnell et al., 2020) AUC=0.635, IEDB (tools.iedb.org/analyze/html_mhcibinding20071227/mhc_binding.html) AUC=0.673, Prime (Schmidt e al., 2021) AUC=0.627, Genesis AUC=0.779. Dashed line=chance prediction.

Figure 10 shows ROC (receiver operating characteristic) curves illustrating the performance of a model as described herein, compared to prior art models predicting peptide-MHC interactions. The comparative models are the same as on Figure 9 but the test dataset for each subplot only contains epitope sequences not contained in the training sets of the particular models being compared. A. NetMHCpan elution prediction AUC=0.600, NetMHCpan binding affinity prediction AUC=0.792, model of the present disclosure AUC=0.791. Test set comprising 12144 doublets including 374 positive doublets. B. MHCflurry AUC=0.690, Genesis AUC=0.789. Test set comprising 12149 doublets including 368 positive doublets. C. DeepAttentionPan AUC=0.797, model of the present disclosure AUC=0.798. Test set comprising 11527 doublets including 414 positive doublets. D. Prime AUC=0.667, model of the present disclosure AUC=0.803. Test set comprising 12003 doublets including 282 positive doublets. E. bigMHC AUC=0.819, Genesis AUC=0.791. Test set comprising 12029 doublets including 349 positive doublets. Dashed lines=chance prediction.

Figure 11 shows ROC curves illustrating the performance of a model as described herein, using different ratios of the number of negative and positive triplets in the data used for training and testing.

Figure 12 shows schematically a model architecture that can be used to predict immunogenicity according to an example of the present disclosure. A. peptide-MHC only based immunogenicity prediction, trained with peptide-MHC immunogenicity data. B. peptide-MHC optional TCR immunogenicity prediction, trained with combination of peptide-MHC and peptide-MHC-TCR immunogenicity data, or peptide-MHC-TCR immunogenicity prediction, trained with peptide-MHC-TCR immunogenicity data. C. Pretraining of pMHC transformer encoder. E. Pretraining of TCR encoder.

Figure 13 shows results of stability prediction using a pMHC transformer encoder (A) and immunogenicity prediction using a pMHC transformer encoder trained using a stability prediction task and comparative models (B). B shows Receiver Operator Characteristic (ROC) and Precision Recall curves, with the AUC and average precision (AP), respectively, indicated for each model. The number of observations (doublets) used to obtain these results is indicated as N=2601 of which 951 are positive doublets.

Figure 14 shows results of TCR specificity prediction using methods of the disclosure and comparative methods. A. Comparison of TCR specificity prediction performance between methods of the disclosure (Genesis) and comparative method (NetTCR). Precision Recall and ROC curves, respectively with average precision (AP) and AUC provided for each model. B. Comparison of TCR specificity prediction performance between methods of the disclosure (Genesis) and comparative method (STAPLER). Precision Recall curves for each fold of the cross-validation with average curve in bold. Average precision (AP) provided for each model.

Figure 15 shows the results of an analysis of the effect of including TCR information when predicting immunogenicity of a pMHC complex.

DETAILED DESCRIPTION

In describing the present invention, the following terms will be employed, and are intended to be defined as indicated below.

The disclosure relates at least in part to the prediction of immunogenicity of antigens.

The term “immunogenicity” refers to the ability of an antigen peptide to bind an MHC molecule for presentation of the antigen and recognition of the peptide-MHC complex by a T cell receptor on a T cell. This recognition process underlines the triggering of an immune reaction, also referred to as “cellular immune reaction”. The terms “antigen”, “peptide” and “antigen peptide” are used interchangeably to refer to a peptide that is potentially immunogenic. Thus, such a peptide can also be referred to as a candidate antigen peptide. **Figure 1A** illustrates the process of antigen recognition. An MHC molecule 3 is expressed on the surface of a cell 2 (which can be an antigen presenting cell, APC, or any other cell such as e.g. a cancer cell). The MHC molecule displays a peptide 6. The complex formed by the MHC molecule 3 and the peptide 6 is recognised by a T cell receptor 5 expressed on the surface of a T cell 4. The present disclosure provides methods to

predict whether a peptide is likely to be immunogenic, as illustrated on **Figure 1B**. The methods use the sequence of the peptide 6, the sequence of at least a part of the MHC molecule 3, and the sequence of at least a part of the TCR molecule 5 (also referred to herein as a "triplet"), as inputs to a machine learning model 10. The machine learning model is trained to use this information to predict whether the peptide, MHC and TCR will interact or not. The method can also be seen a predicting whether a triplet will interact, or whether any member of the triplet will interact with the other members to trigger an immune response.

MHC (major histocompatibility complex) molecules are cell-surface proteins encoded by the human leukocyte antigen (HLA) gene complex, and which are an important part of the adaptive immune system. MHC molecules are typically classified as "class I" or "class II". Class I MHC molecules present peptides from inside the cells for recognition by T cell receptors as will be explained further below. Class I MHC molecules are normally expressed on the surface of all cells. The peptides presented are typically produced from digested proteins produced in the proteasomes, and are typically about 8-11 amino acids in length. There are 3 types of MHC class I molecules (A, B and C), each encoded by a separate gene. Class II MHC molecules present antigens from outside the cells for recognition by T cells. Class II MHC molecules are primarily found on antigen-presenting cells such as dendritic cells, mononuclear phagocytes, some endothelial cells, thymic epithelial cells and B cells. There are 6 types of MHC class II molecules: DP, DM, DOA, DOB, DQ and DR, each encoded by a separate gene. The HLA locus is highly polymorphic and therefore many different alleles exist for each gene. The process of HLA typing refers to determining which alleles of each of one or more HLA genes is present in a sample or subject. The term "MHC sequence" as used herein refers to the amino acid sequence of an MHC molecule or part thereof. The MHC molecule may be a class I MHC molecule or a class II MHC molecule. The models described herein are typically trained using a single class of MHC molecule, and used for prediction for this class.

The term "TCR sequence" as used herein refers to the amino acid sequence of a T cell receptor or part thereof. A T cell receptor is a membrane anchored protein expressed on the surface of T cells. A T cell receptor comprises a pair of protein chains that together form a binding moiety that recognises a cognate antigen. The TCR chains are expressed in a complex with constant T cell coreceptor chains CD3 (illustrated as reference numeral 7 in Figure 1A), comprising a CD3γ chain, a CD3δ chain, and two CD3ε chains in mammals. The constant chains associate with the T cell receptor and the constant ζ-chain to form the TCR complex, which together is able to generate a signal upon antigen binding to the T cell receptor. As illustrated on Figure 1A, the TCR 5 is a heterodimeric protein, comprising two highly variable chains 5a and 5b, the α and β chains (in the majority of T cells), or the alternative γ and δ chains (in a minority of T cells). Each chain comprises two extracellular domains: a variable region (or variable domain; 50a, 50b) and a constant region (or constant domain, proximal to the cell membrane; 50a', 50b'), a transmembrane region and a

short cytoplasmic tail. The variable regions together bind to a peptide (antigen) 6, within the context of a MHC (major histocompatibility complex) molecule 3 in the case of αβ TCRs. Each variable domain contains three hypervariable regions referred to as the complementarity-determining regions (CDRs, respectively referred to as CDR1, CDR2 and CDR3 on each of the chains), which together form an antigen binding site. A TCR sequence may comprise the complete sequence of one or both chains of a TCR, or a part of one or both chains. In the context of the present disclosure, a TCR typically comprises α and β chains, and thus a TCR sequence comprises the sequence of at least a part of a TCR α chain and/or a at least a part of a TCR β chain. Most of the information currently available about TCR repertoire has been obtained by bulk-sequencing on single chain repertoires, mostly the β chain repertoire. Paired chain information (i.e. information about the sequence of both the α and β chain of a TCR) as well as information about the sequence of the α chain repertoire is not as extensive at present. Therefore, the models described herein may use only TCR sequences from the β chain. This may enable models to be trained in an optimal manner (by using large amounts of data for better constraining of the model) and applicable to a wider range of situations (where α chain sequence may not be available). However, as the skilled person understands, it is also possible to make use of α chain sequences when available, as described further herein.

The term “sequence data” refers to information that is indicative of the presence of genomic material (DNA or RNA) or proteomic material in a sample that has a particular sequence. Thus, sequence data may comprise one or more nucleotide sequences and/or one or more amino acid sequences. Such information may be obtained using sequencing technologies, such as e.g. next generation sequencing (NGS), for example whole exome sequencing (WES), whole genome sequencing (WGS), whole transcriptome sequencing (RNAseq) or sequencing of captured genomic loci (targeted or panel sequencing). When NGS technologies are used, the sequence data may comprise a count of the number of sequencing reads that have a particular sequence. Sequence data may be mapped to a reference sequence, for example a reference genome, using methods known in the art (such as e.g. Bowtie (Langmead et al., 2009)). Thus, counts of sequencing reads or equivalent non-digital signals may be associated with a particular location or locus (where the “location” refers to a location in the reference genome or transcriptome to which the sequence data was mapped). Further, a location may contain a mutation, in which case counts of sequencing reads or equivalent non-digital signals may be associated with each of the possible variants (also referred to as “alleles”) at the particular location. The process of identifying the presence of a mutation at a particular location in a sample is referred to as “variant calling” and can be performed using methods known in the art (such as e.g. general purpose NGS variant callers such as the GATK HaplotypeCaller, gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller or tools specifically designed for immune sequences such as IgBLAST, www.ncbi.nlm.nih.gov/igblast/, [Ye et al., 2013]). Genomic sequence data may be converted to amino acid sequences by translating coding regions *in silico* (directly from an

mRNA sequence or from identified coding regions in a genomic sequence), as known in the art. This may be performed for example to obtain a peptide from a coding sequence comprising a coding mutation (a mutation that alters the sequence of amino acid encoded by a DNA or RNA sequence).

The term "peptide" is used in the normal sense to mean a series of residues, typically L-amino acids, connected one to the other typically by peptide bonds between the a-amino and carboxyl groups of adjacent amino acids. The term includes modified peptides and synthetic peptide analogues. In particular, peptides as used herein may include one or more non-canonical amino acids (also referred to as “nonstandard amino acids” or “modified amino acids”).

According to the present disclosure, the probability that a peptide is immunogenic in the context of a candidate MHC sequence and a candidate TCR sequence (which can also be interpreted as the probability that the peptide, candidate MHC molecule and candidate TCR molecule interact with each other) is predicted using one or more machine learning models. The term “machine learning model” refers to a mathematical model that has been trained to predict one or more output values based on input data, where training refers to the process of learning, using training data, the parameters of the mathematical model that result in a model that can predict outputs values that satisfy an optimality criterion or criteria. In the case of supervised learning, training typically refers to the process of learning, using training data, the parameters of the mathematical model that result in a model that can predict outputs values that with minimal error compared to comparative (known) values associated with the training data (where these comparative values are commonly referred to as “labels”). The term “machine learning algorithm” or “machine learning method” refers to an algorithm or method that trains and/or deploys a machine learning model. The machine learning model may comprise one or more natural language processing models. The machine learning model may comprise one or more encoders trained to encode one or more of the sequences in a triplet. Encoding refers to processing of input data to generate a latent representation of the input data, from which a prediction can be made or the input data can be reconstructed. An encoder may be selected from: transformer-based encoders, autoencoders, and recurrent neural network encoders such as long-short-term memory (LSTM) networks. Transformer-based encoders include bidirectional encoders trained by masked language modeling, such as the BERT model described in Devlin et al. (BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805). In embodiments, the machine learning model comprises one or more artificial neural network (ANNs, also referred to simply as “neural network” (NN)). ANNs are typically parameterized by a set of weights that are applied to the inputs of each of a plurality of connected neurons in order to obtain a weighted sum that is fed to an activation function to produce the neuron’s output. The parameters of an NN can be trained using a method called backpropagation through which connection weights are adjusted to compensate for errors found in the learning process, in combination with a weight updating procedure such as stochastic

gradient descent. An ANN may be a deep neural network, i.e. a neural network comprising more than one layer (also referred to as "hidden layer") between the input layer and the output layer. In embodiments, a machine learning model comprises an ensemble of models whose predictions are combined. Alternatively, a machine learning model may comprise a single model. The training data may comprise, for a plurality of triplets: the primary structure of the peptides (i.e. amino acid sequence) or features derived from said primary structure by encoding as described further herein; and a status label identifying a triplet as a triplet that has been experimentally determined to interact (positive triplet), or a triplet that is not expected to interact (negative triplet). **Figure 1C** illustrates the types of triplets included in the training data according to the present disclosure. This comprises one or more positive triplets 8, and a first, second and third sets of negative triplets 9a, 9b, 9c. Triplets in the first set of negative triplets 9a comprise an MHC-TCR pair that is known to interact, and a peptide that is not known to interact with this MHC-TCR pair. This may also be referred to as a "negative peptide" set. Triplets in the second set of negative triplets 9b comprise a peptide-MHC pair that is known to bind and is expected to be immunogenic given the right TCR, and a TCR that is not known to interact with this peptide-MHC pair. This may also be referred to as a "negative TCR" set. Triplets in the third set of negative triplets 9c comprise a peptide-MHC pair that is known to interact and is also known not to be immunogenic, and a TCR sequence. This may also be referred to as a "non-immunogenic pMHC" set. Advantageously, the training data may comprise or consists of data that relates to triplets identified in an organism that is the same as that from which the triplets for which immunogenicity is to be predicted originates. The training data may comprise data for at least 10,000, at least 20,000, at least 30,000 or at least 40,000 unique positive triplets. The training data may comprise data for at least 100 times, at least 150 times, or at least 200 times more negative triplets than positive triplets. For example, the training data may comprise data for at least 8 million negative triplets. The training data may be divided between a training set and a test set. The training set may comprise data for at least 10,000, at least 15,000, at least 25,000 or at least 35,000 unique positive triplets. The test set may comprise data for at least 1000, at least 2000, at least 3000, or at least 4000 unique positive triplets. The training of the model may be performed using cross-validation, as known in the art, wherein a model is trained multiple times using a subset of the training set and evaluated using the remaining subset of the training set, then performance of the multiple models obtained is combined when evaluating the model.

A "sample" as used herein may be a cell or tissue sample, a biological fluid, an extract (e.g. a DNA extract obtained from the subject), from which genomic material can be obtained for genomic analysis, such as genomic sequencing (e.g. whole genome sequencing, whole exome sequencing). The sample may be a cell, tissue or biological fluid sample obtained from a subject (e.g. a biopsy). Such samples may be referred to as "subject samples". In particular, the sample may be a blood sample, or a tumour sample, or a sample derived therefrom. The sample may be one which has been freshly obtained from a subject or may be one which has been processed

and/or stored prior to genomic analysis (e.g. frozen, fixed or subjected to one or more purification, enrichment or extraction steps). The sample may be a cell or tissue culture sample. As such, a sample as described herein may refer to any type of sample comprising cells or genomic material derived therefrom, whether from a biological sample obtained from a subject, or from a sample obtained from e.g. a cell line. In embodiments, the sample is a sample obtained from a subject, such as a human subject. The sample is preferably from a mammalian (such as e.g. a mammalian cell sample or a sample from a mammalian subject, such as a cat, dog, horse, donkey, sheep, pig, goat, cow, mouse, rat, rabbit or guinea pig), preferably from a human (such as e.g. a human cell sample or a sample from a human subject). Further, the sample may be transported and/or stored, and collection may take place at a location remote from the genomic sequence data acquisition (e.g. sequencing) location, and/or any computer-implemented method steps described herein may take place at a location remote from the sample collection location and/or remote from the genomic data acquisition (e.g. sequencing) location (e.g. the computer-implemented method steps may be performed by means of a networked computer, such as by means of a “cloud” provider).

A “normal sample” or “germline sample” refers to a sample that is assumed not to comprise tumour cells or genetic material derived from tumour cells. A germline sample may be a blood sample, a tissue sample, or a purified sample such as a sample of peripheral blood mononuclear cells from a subject. Similarly, the terms “normal”, “germline” or “wild type” when referring to sequences or genotypes refer to the sequence / genotype of cells other than tumour cells. A germline sample may comprise a small proportion of tumour cells or genetic material derived therefrom, and may nevertheless be assumed, for practical purposes, not to comprise said cells or genetic material. In other words, all cells or genetic material may be assumed to be normal and/or sequence data that is not compatible with the assumption may be ignored.

The terms “tumour-specific mutation”, “somatic mutation” or simply “mutation” are used interchangeably and refer to a difference in a nucleotide sequence (e.g. DNA or RNA) in a tumour cell compared to a healthy cell from the same subject. The difference in the nucleotide sequence can result in the expression of a protein which is not expressed by a healthy cell from the same subject. For example, a mutation may be a single nucleotide variant (SNV), multiple nucleotide variant (MNV), a deletion mutation, an insertion mutation, a translocation, a missense mutation, a translocation, a fusion, a splice site mutation, or any other change in the genetic material of a tumour cell. A mutation may result in the expression of a protein or peptide that is not present in a healthy cell from the same subject. Mutations may be identified by exome sequencing, RNA-sequencing, whole genome sequencing and/or targeted gene panel sequencing and or routine Sanger sequencing of single genes, followed by sequence alignment and comparing the DNA and/or RNA sequence from a tumour sample to DNA and/or RNA from a reference sample or reference sequence (e.g. the germline DNA and/or RNA sequence, or a reference sequence from a database). Suitable methods are known in the art.

An "indel mutation" refers to an insertion and/or deletion of bases in a nucleotide sequence (e.g. DNA or RNA) of an organism. Typically, the indel mutation occurs in the DNA, preferably the genomic DNA, of an organism. In embodiments, the indel may be from 1 to 100 bases, for example 1 to 90, 1 to 50, 1 to 23 or 1 to 10 bases. An indel mutation may be a frameshift indel mutation. A frameshift indel mutation is a change in the reading frame of the nucleotide sequence caused by an insertion or deletion of one or more nucleotides. Such frameshift indel mutations may generate a novel open-reading frame which is typically highly distinct from the polypeptide encoded by the non-mutated DNA/RNA in a corresponding healthy cell in the subject.

A “neoantigen” (or “neo-antigen”) is an antigen that arises as a consequence of a mutation within a cancer cell. Thus, a neoantigen is not expressed (or expressed at a significantly lower level) by normal (i.e. non-tumour) cells. A neoantigen may be processed to generate distinct peptides which can be recognised by T cells when presented in the context of MHC molecules. As described herein, neoantigens may be used as the basis for cancer immunotherapies. References herein to "neoantigens" are intended to include also peptides derived from neoantigens. The term "neoantigen" as used herein is intended to encompass any part of a neoantigen that is immunogenic. An "antigenic" molecule as referred to herein is a molecule which itself, or a part thereof, is capable of stimulating an immune response, when presented to the immune system or immune cells in an appropriate manner. The binding of a neoantigen to a particular MHC molecule (encoded by a particular HLA allele) results on the neoantigen being presented by said MHC molecule on the cell surface, a necessary but not sufficient condition for immunogenicity. Immunogenicity further requires recognition of the peptide-MHC complex by a T cell receptor. As used herein a “candidate neoantigen” refers to a peptide or sequence thereof that arises as a consequence of a mutation within a cancer cell, the immunogenicity of which has not yet been verified. The present disclosure provides methods to predict whether a candidate neoantigen is likely to be immunogenic, i.e. a *bona fide* neoantigen.

A “clonal neoantigen” (also sometimes referred to as “truncal neoantigen”) is a neoantigen that results from a mutation that is present in essentially every tumour cell in one or more samples from a subject (or that can be assumed to be present in essentially every tumour cell from which the tumour genetic material in the sample(s) is derived). Similarly, a “clonal mutation” (sometimes referred to as “truncal mutation”) is a mutation that is present in essentially every tumour cell in one or more samples from a subject (or that can be assumed to be present in essentially every tumour cell from which the tumour genetic material in the sample(s) is derived). Thus, a clonal mutation may be a mutation that is present in every tumour cell in one or more samples from a subject. A “sub-clonal” neoantigen is a neoantigen that results from a mutation that is present in a subset or a proportion of cells in one or more tumour samples from a subject (or that can be assumed to be present in a subset of the tumour cells from which the tumour genetic material in the sample(s) is derived). Similarly, a “sub-clonal” mutation is a mutation that is present in a subset

or a proportion of cells in one or more tumour samples from a subject (or that can be assumed to be present in a subset of the tumour cells from which the tumour genetic material in the sample(s) is derived). As the skilled person understands, a neoantigen or mutation may be clonal in the context of one or more samples from a subject while not being truly clonal in the context of the entirety of the population of tumour cells that may be present in a subject (e.g. including all regions of a primary tumour and metastasis). Thus, a clonal mutation may be "truly clonal" in the sense that it is a mutation that is present in essentially every tumour cell (i.e. in all tumour cells) in the subject. This is because the one or more samples may not be representative of each and every subset of cells present in the subject. Thus, within the context of the present disclosure, a "clonal neoantigen" or "clonal mutation" may also be referred to as a "ubiquitous neoantigen" or "ubiquitous mutation", to indicate that the neoantigen is present in essentially all tumour cells that have been analysed, but may not be present in all tumour cells that may exist in the subject. The terms "clonal" and "ubiquitous" are used interchangeably unless context indicates that reference to "true clonality" was intended. The wording "essentially every tumour cell" in relation to one or more samples or a subject may refer to at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94% at least 95%, at least 96%, at least 97%, at least 98%, or at least 99% of the tumour cells in the one or more samples or the subject.

A cancer immunotherapy (or simply "immunotherapy") refers to a therapeutic approach comprising administration of an immunogenic composition (e.g. a vaccine), a composition comprising immune cells, or an immunoactive drug, such as e.g. a therapeutic antibody, to a subject. The term "immunotherapy" may also refer to the therapeutic compositions themselves. In the context of the present disclosure, the immunotherapy typically targets a neoantigen. For example, an immunogenic composition or vaccine may comprise a neoantigen, neoantigen presenting cell or material necessary for the expression of the neoantigen. As another example, a composition comprising immune cells may comprise T and/or B cells that recognise a neoantigen. The immune cells may be isolated from tumours or other tissues (including but not limited to lymph node, blood or ascites), expanded ex vivo or in vitro and re-administered to a subject (a process referred to as "adoptive cell therapy"). Instead or in addition to this, T cells can be isolated from a subject and engineered to target a neoantigen (e.g. by insertion of a chimeric antigen receptor that binds to the neoantigen) and re-administered to the subject. As another example, a therapeutic antibody may be an antibody which recognises a neoantigen. One skilled in the art will appreciate that if the neoantigen is a cell surface antigen, an antibody as referred to herein will recognise the neoantigen. Where the neoantigen is an intracellular antigen, the antibody will recognise the neoantigen peptide-MHC complex. As referred to herein, an antibody which "recognises" a neoantigen encompasses both of these possibilities. Further, an immunotherapy may target a plurality of neoantigens. For example, an immunogenic composition may comprise a plurality of neoantigens, cells presenting a plurality of neoantigens or the material necessary for the expression of the plurality of neoantigens. As another example, a composition may comprise

immune cells that recognise a plurality of neoantigens. Similarly, a composition may comprise a plurality of immune cells that recognise the same neoantigen. As another example, a composition may comprise a plurality of therapeutic antibodies that recognise a plurality of neoantigens. Similarly, a composition may comprise a plurality of therapeutic antibodies that recognise the same neoantigen.

A composition as described herein may be a pharmaceutical composition which additionally comprises a pharmaceutically acceptable carrier, diluent or excipient. The pharmaceutical composition may optionally comprise one or more further pharmaceutically active polypeptides and/or compounds. Such a formulation may, for example, be in a form suitable for intravenous infusion.

References to "an immune cell" are intended to encompass cells of the immune system, for example T cells, NK cells, NKT cells, B cells and dendritic cells. In a preferred embodiment, the immune cell is a T cell. An immune cell that recognises a neoantigen may be an engineered T cell. A neoantigen specific T cell may express a chimeric antigen receptor (CAR) or a T cell receptor (TCR) which specifically binds a neoantigen or a neoantigen peptide, or an affinity-enhanced T cell receptor (TCR) which specifically binds a neoantigen or a neoantigen peptide (as discussed further hereinbelow). For example, the T cell may express a chimeric antigen receptor (CAR) or a T cell receptor (TCR) which specifically binds to a neo-antigen or a neo-antigen peptide (for example an affinity enhanced T cell receptor (TCR) which specifically binds to a neo-antigen or a neo-antigen peptide). Alternatively, a population of immune cells that recognise a neoantigen may be a population of T cell isolated from a subject with a tumour. For example, the T cell population may be generated from T cells in a sample isolated from the subject, such as e.g. a tumour sample, a peripheral blood sample or a sample from other tissues of the subject. The T cell population may be generated from a sample from the tumour in which the neoantigen is identified. In other words, the T cell population may be isolated from a sample derived from the tumour of a patient to be treated, where the neoantigen was also identified from a sample from said tumour. The T cell population may comprise tumour infiltrating lymphocytes (TIL).

The term "Antibody" (Ab) includes monoclonal antibodies, polyclonal antibodies, multispecific antibodies (e.g., bispecific antibodies), and antibody fragments that exhibit the desired biological activity. The term "immunoglobulin" (Ig) may be used interchangeably with "antibody". Once a suitable neoantigen has been identified, for example by a method according to the disclosure, methods known in the art can be used to generate an antibody.

An “immunogenic composition” is a composition that is capable of inducing an immune response in a subject. The term is used interchangeably with the term “vaccine”. The immunogenic composition or vaccine described herein may lead to generation of an immune response in the subject. An "immune response" which may be generated may be humoral and/or cell-mediated immunity, for example the stimulation of antibody production, or the stimulation of cytotoxic or killer

cells, which may recognise and destroy (or otherwise eliminate) cells expressing antigens corresponding to the antigens in the vaccine on their surface. The immunogenic composition may comprise one or more neoantigens, or the material necessary for the expression of one or more neoantigens. In addition, a neoantigen may be delivered in the form of a cell, such as an antigen presenting cell, for example a dendritic cell. The antigen presenting cell such as a dendritic cell may be pulsed or loaded with the neo-antigen or neo-antigen peptide or genetically modified (via DNA or RNA transfer) to express one, two or more neo-antigens or neoantigen peptides, for example 2, 3, 4, 5, 6, 7, 8, 9 or 10 neo-antigens or neo-antigen peptides. Methods of preparing dendritic cell immunogenic compositions or vaccines are known in the art.

An antigen peptide refers to a peptide that is capable of binding to an MHC molecule and interact with a TCR receptor in the context of an MHC molecule to elicit an immune response. The term “peptide” as used herein encompasses an antigen peptide and a peptide that is a candidate antigen peptide, i.e. a peptide for which immunogenicity is to be predicted for example as described herein, or a fragment thereof. By way of example, a peptide which is capable of binding to an MHC class I molecule is typically 7 to 13 amino acids in length, or more specifically 8 to 11 amino acids. When longer peptides are used, such as e.g. peptides longer than the maximal length of sequence that can be provided as input to a machine learning model as described herein (e.g. longer than 16 amino acids), immunogenicity may be predicted as described herein for one or more fragments of the peptide. Such fragments may also be referred to herein as “minimal peptides”. For example, immunogenicity may be predicted for one or more fragments of at least a minimal length (such as e.g. 7 or 8 amino acids) and/or at most a maximal length (such as e.g. 11 or 13 amino acids). A prediction for a longer peptide may be obtained as a summarised prediction over a set of fragments of the peptide. For example, the average or maximum predicted probability amongst probabilities predicted for the one or more fragments of the peptide may be taken as the probability predicted for the peptide.

A “neoantigen peptide” as described herein refers to a peptide that comprises a cancer cell specific mutation (e.g a non-silent amino acid substitution encoded by a single nucleotide variant (SNV), an indel or any other genetic alteration that results in a change in primary structure of a peptide or protein) at any residue position within the peptide. As mentioned above, in a peptide that has a length of between 7 and 13 amino acids, the amino acid substitution may be present at position 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 or 13 in a peptide comprising thirteen amino acids. In embodiments, longer peptides, for example 15-31-mers, may be used, and the mutation may be at any position, for example at the centre of the peptide, e.g. at positions 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 or 16. Such peptides can also be used to stimulate both CD4 and CD8 cells to recognise neoantigens. As used herein "treatment" refers to reducing, alleviating or eliminating one or more symptoms of the disease which is being treated, relative to the symptoms prior to treatment. "Prevention" (or prophylaxis) refers to delaying or preventing the onset of the symptoms of the disease. Prevention

may be absolute (such that no disease occurs) or may be effective only in some individuals or for a limited amount of time.

As used herein, the terms "computer system" or "computer device" includes the hardware, software and data storage devices for embodying a system or carrying out a method according to the above-described embodiments. For example, a computer system may comprise one or more processing units such as a central processing unit (CPU) and/or a graphical processing unit (GPU), input means, output means and data storage, which may be embodied as one or more connected computing devices. Preferably the computer system has a display or comprises a computing device that has a display to provide a visual output display (for example in the design of the business process). The data storage may comprise RAM, disk drives or other computer readable media. The computer system may include a plurality of computing devices connected by a network and able to communicate with each other over that network. For example, a computer system may be implemented as a cloud computer. The term "computer readable media" includes, without limitation, any non-transitory medium or media which can be read and accessed directly by a computer or computer system. A computer readable medium may be a tangible computer readable medium. A computer readable medium may be realized as a plurality of discrete tangible computer readable media. The media can include, but are not limited to, magnetic storage media such as floppy discs, hard disc storage media and magnetic tape; optical storage media such as optical discs or CD-ROMs; electrical storage media such as memory, including RAM, ROM and flash memory; and hybrids and combinations of the above such as magnetic/optical storage media.

Prediction of immunogenicity

The present disclosure provides methods for predicting immunogenicity of antigens, and methods for providing a tool for predicting the immunogenicity of one or more antigens. As the skilled person understands, the methods described herein can provide a prediction of immunogenicity of a candidate antigen in the context of a particular MHC molecule and TCR molecule. Therefore, the methods described herein can also be seen as providing a prediction of immunogenicity of a triplet comprising an antigen, an MHC molecule and a TCR molecule, and/or providing a prediction of whether a TCR molecule can recognise a candidate peptide-MHC complex. In other words, the methods can be used primarily with the aim of characterising any member of the triplet, including but not limited to the peptide.

An illustrative method for providing a tool and/or predicting the immunogenicity of one or more antigens will be described by reference to **Figure 2**. At step 10, one or more triplets of sequences is/are obtained for a peptide. A triplet of sequences corresponds to a peptide-MHC-TCR triplet and comprises at least part of the sequence of the peptide, the MHC molecule, and the TCR molecule in the triplet. Step 10 may comprise step 10a of providing one or more candidate MHC sequences, and one or more candidate TCR sequences, and step 10b of providing the peptide sequence to

be analysed. The peptide sequence to be analysed may be a predicted peptide sequence derived from a cancer specific mutation identified in a sample or patient. The one or more candidate MHC sequences may be obtained or may have been previously obtained from a database, computing device or user interface. The one or more MHC sequences may comprise or consist of sequences that are associated with a particular patient or sample. The step of obtaining the one or more MHC sequences may comprise identifying one or more MHC sequences that are present in a patient or sample (a process referred to as "HLA typing"). Methods for HLA typing are known in the art and include e.g. flow cytometry-based methods and methods based on sequencing data such as Polysolver (Shukla et al. 2015) and OptiType (Szolek et al. 2014). The one or more candidate MHC sequences may be sequences of HLA alleles determined not to have been lost in a sample or patient (a process called "HLA Loss of Heterozygosity", HLA LOH). Methods for identifying HLA LOH are known in the art and include e.g. the method described in WO2019/012296. The patient / sample may be the same patient / sample from which the peptide sequence to be analysed has been derived. The one or more candidate TCR sequences may be obtained or may have been previously obtained from a database, computing device or user interface. The one or more TCR sequences may comprise or consist of sequences that are associated with a particular patient or sample. The step of obtaining the one or more TCR sequences may comprise identifying one or more TCR sequences that are present in a patient or sample (for example by TCR repertoire sequencing, also referred to as "TCRseq"). The one or more TCR sequences may be an empty TCR sequence vector. In other words, an empty TCR sequence vector can be used as input to the machine learning model as described further below. In such embodiments, the score predicted by the machine learning model (as will be described further below) represents the probability that the antigen is immunogenic in the context of the candidate MHC molecule and an unknown TCR. By contrast, when a single candidate TCR sequence is provide as part of the triplet, the score predicted by the machine learning model (as will be described further below) represents the probability that the antigen is immunogenic in the context of the candidate MHC molecule and the particular candidate TCR sequence. The one or more candidate MHC sequences may comprise the sequence of a part of a MHC molecule, such as e.g. the sequence of a part of an MHC corresponding to the peptide binding groove, or the sequence of a complete MHC molecule chain. The sequence of the part of MHC molecule may include the whole sequence corresponding to the binding groove or a portion thereof. The one or more candidate TCR sequences may comprise the sequence of a part of a TCR molecule. For example, the one or more candidate TCR sequences may each comprise the sequence of one or more CDR regions of a TCR. As another example, the one or more candidate TCR sequences may each comprise the sequence of one or more chains of a TCR. The one or more CDR regions typically comprise at least the CDR3 region of the TCR β chain. The one or more CDR regions may comprise the CDR3 region of the TCR α chain. Any one or more of the CDR1β, CDR2β, CDR1α and CDR2α regions may also be represented. Where multiple non connected sequences (e.g. a plurality of CDR sequences, sequences from the two

TCR chains, etc) are provided, these can be concatenated into a single sequence, or they can be processed and input into the model separately. Similarly, the peptide and MHC molecule sequence can be concatenated into a single sequence, or they can be processed and input into the model separately. The one or more triplets may be obtained by combining the peptide sequence with a candidate MHC sequence and a candidate TCR sequence. When a plurality of candidate TCR and/or MHC sequences are provided, a plurality of triplets may be obtained each comprising the peptide and a unique combination of candidate MHC and TCR sequences.

At step 12, each of the sequences in the triplet (or the plurality of triplets) is encoded to obtain information derived from the respective sequences. Each type of sequence (peptide, TCR, MHC) may be encoded using a respective encoding scheme. The same encoding scheme may be used for at least the peptide and the MHC sequence. The same encoding scheme may be used for all sequences. The peptide and MHC sequences can be concatenated and encoded together, or they can be encoded individually. Encoding uses a predetermined encoding scheme. Suitable encoding schemes include: the use of predetermined tokens for each amino acid (e.g. a number for each amino acid and optionally a further number for a padding character, all of which can be consecutive integers, such as e.g. numbers from 0 to 20), the use of predetermined tokens for multiple amino acids (e.g. doublets or triplets of amino acids, or even full sequences can be encoded with a single token, at the extreme it is even possible for a single token to be used to represent e.g. an MHC molecule, where each token may e.g. correspond to an MHC allele), one-hot encoding, the use of substitution matrices (such as e.g. BLOSUM) , and the use of a set of physicochemical descriptors for each amino acid. Examples of encoding schemes for amino acid sequences are provided in ElAbd *et al.* 2020. Advantageously, the encoding scheme may be one that encodes every amino acid of an input sequence separately. This may increase the resolution of the model, enabling it to learn from interactions between all amino acid positions. In embodiments, the peptide amino acid sequence is encoded using: (a) a predetermined amino acid encoding scheme to produce an amino acid sequence embedding, and (b) a fixed or learned encoding scheme applied to hydrophobicity values associated with each of the amino acids in the sequence. The amino acid sequence embedding and hydrophobicity-based embedding can be combined to obtain an encoded sequence for the peptide using a positional encoding scheme, where the hydrophobicity based encodings are used as positional encodings. The amino acid sequence embedding and hydrophobicity-based embedding can be combined by summing the amino acid sequence embedding and the hydrophobicity based embedding. The hydrophobicity values can be Kyte-Doolittle values. For example, the following hydrophobicity values (corresponding to the Kyte-Doolittle scale) can be used for canonical amino acids: A: 1.8; C: 2.5 D: -3.5; E: -3.5; F: 2.8; G: -0.4; H: -3.2; I: 4.5; K: -3.9; L: 3.8; M: 1.9; N: -3.5; P: -1.6; Q: -3.5; R: -4.5; S: -0.8; T: -0.7; V: 4.2; W: -0.9; and Y: -1.3. Note that other hydrophobicity scales can be used, such as e.g. the Engelman scale, the Eisenberg scale, and the Hopps-Woods scale. A fixed

encoding scheme can be any encoding scheme known in the art (e.g. one hot encoding, or use of the hydrophobicity values themselves). A learned encoding scheme can be any encoding that is specified through the use of an encoding block (e.g. a fully connected layer) that is trained when training the model that takes the peptide sequence as input (e.g. the peptide-MHC module as described further below).

At step 14, a probability that the peptide is immunogenic is predicted. This may comprise step 14A of, for each triplet, inputting the encoded sequences into a model trained to predict the probability that the peptide is immunogenic in the context of the candidate MHC sequence and the candidate TCR sequence in the triplet. Step 14 may further comprise step 14B of determining a probability of the peptide being immunogenic based on the results of step 14A. For example, the maximum probability amongst the probabilities obtained for all of the one or more triplets may be selected as the probability that the peptide is immunogenic.

A method of obtaining a model for use in step 14 will now be described. The method may comprise step 10' of providing training data for training the model. The training data comprises the sequence of a plurality of peptide-MHC-TCR triplets (i.e. a plurality of triplets of sequences), and a status label for each triplet indicating whether the triplet is a positive triplet (comprising sequences for a peptide, MHC and TCR molecules that are known to interact) or a negative triplet (comprising sequences for a peptide, MHC and TCR molecules that are not expected to interact). The status labels may be set to any binary value, such as e.g. 0 for negative triplets and 1 for positive triplets. Any binary label may be used. The values may eventually be converted to values of "0" or "1" to train the model to predict a value between 0 and 1 being as close as possible to 1 for positive triplets and a value as close as possible to 0 for negative triplets. Alternatively, the values may not be used as such (i.e. they may simply be used as a class label) and the model may be trained to predict a value between 0 and 1 that is the probability of a triplet belonging to a first class (e.g. positive triplets), rather than a second class (e.g. negative triplets). Step 10' may comprise step 10'a of obtaining sequences for a plurality of positive triplets. Sequences for positive triplets may be obtained from one or more databases, computing devices or user interfaces. Sequences for positive triplets may be sequences of peptide-MHC-TCR triplets that have been experimentally demonstrated to interact. The positive triplets may all be from the same organism. For example, the positive triplets may all comprise human sequences. All positive triplets are associated with a "positive" status label. Step 10' may further comprise step 10b' if obtaining a first, second and third set of negative triplets, each comprising one or more triplets associated with a "negative" status label. The first set of triplets each comprise: (i) a TCR-MHC pair comprising an MHC molecule and a TCR chain or chains known to bind the MHC molecule (positive TCR-MHC pair), and (ii) a peptide not known to interact with the TCR-MHC pair. The positive TCR-MHC pair may be selected from the positive triplets. The peptide may be selected from any collection of peptides including the peptides in the positive triplets, or peptides extracted from a reference dataset (e.g. a reference

proteome), for example from a database. Peptide sequences from a reference dataset may comprise sequences selected from a collection of protein sequences from the same organism as the organism from which one or more positive triplets have been identified. For example, the human proteome (as defined in e.g. Uniprot, www.uniprot.org) may be used as a reference sequence. The peptide sequences may be selected randomly or may be selected from a set of proteins comprising the peptide sequences in the positive triplets. Advantageously, the peptide sequences may be selected independently from the peptide sequences in the positive triplets. For example, the peptide sequences may be randomly selected from a reference proteome. The second set of triplets each comprise: (i) a peptide-MHC pair comprising an MHC molecule and a peptide known to bind the MHC molecule (positive peptide-MHC pair), and (ii) a TCR chain or chains not known to interact with the peptide-MHC pair, wherein the peptide-MHC pair has been previously found to interact with a TCR (immunogenic positive peptide-MHC pair). The positive peptide-MHC pair may be selected from the positive triplets. The TCR sequences may be selected from the positive triplets, or independently from the positive triplets. For example, the TCR sequences may be selected from a database or reference dataset. For example, the TCR sequences may be selected from a TCR sequence database such as e.g. VDJdb. The TCR sequences are typically from the same organism as at least one or more of the positive triplets. The third set of triplets each comprise: (i) a peptide-MHC pair comprising an MHC molecule and a peptide known to bind the MHC molecule (positive peptide-MHC pair), and a TCR chain or chains not known to interact with the peptide-MHC pair, wherein the peptide-MHC pair has been previously found to not be immunogenic (non-immunogenic positive peptide-MHC pair). The positive peptide-MHC pair in the third set of negative triplets cannot be obtained from the positive triplets (as all of these are immunogenic). Therefore, the third set of negative triplets cannot be obtained by resampling the positive dataset, which is a common way to obtain negative data in the art. Instead, the positive peptide-MHC pairs of the third set of negative triplets are peptide-MHC pairs that have been experimentally shown to form a complex that is not immunogenic. These may be identified experimentally as part of the method, and/or may be obtained from one or more databases, computing devices or user interface.

At step 12', all of the sequences are encoded substantially as described in relation to step 12. At step 14A', the TCR sequence of one or more triplets in the training data is used to pretrain a model to encode TCR sequences (also referred to herein as "TCR encoder", "TCR encoding module" or "self-supervised TCR block"). This may instead or additionally use TCR sequences that are not comprised in triplets, such as e.g. sequences from TCR repertoires or databases. In other words, the pretraining of the TCR model does not require the use of TCR sequences that are parts of triplets. The model is able to learn the "language" of TCR sequences based solely on TCR sequences themselves. In other words, the model learns the sequence features that characterise a "real" TCR sequence. Thus, step 14A' may further comprise obtaining one or more further TCR sequences, such as e.g. from a TCR sequence database. The model obtained at step 14A' may

be referred to as a "TCR model". The TCR model may comprise an encoder or a pair of encoders each encoding a different part of a TCR molecule (e.g. a part of the alpha and beta chains, respectively). The model may be trained in a self-supervised manner, for example using masked language modelling as described in Devlin et al. (Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805). The TCR model may be trained as an encoder only model comprising an encoder (e.g. transformer-based) and a fully connected layer (or layers) for reconstructing the input from the output of the encoder. The fully connected layer(s) may be removed after training such that only the encoder is used as part of the full machine learning model. When the TCR sequence comprises sequence from pairs of chains, a single TCR model may be trained using as input the concatenation of the sequences in the pair. Alternatively, separate TCR models may be trained for the respective chains (i.e. taking the sequence from the respective chain as input), and the outputs of these models may be combined.

At step 14A'', the peptide and MHC sequences are used to train a model comprising an encoder, the output of which is used by the model to determine whether a peptide-MHC pair is likely to be a positive or negative doublet (i.e. a doublet determined experimentally to bind or not bind to each other and/or to form a peptide-MHC complex that is presented on the surface of cells, such as e.g. by identifying the peptide as an eluted ligand of the MHC molecule and/or by determining the binding affinity of the peptide to the MHC molecule), and/or to determine whether the peptide and MHC molecule are likely to form a stable complex. This can comprise classifying the doublets between a negative and a positive class, for example producing a probability that the peptide-MHC is in the positive class (classification task), or predicting a scaled binding affinity or metric indicative of the stability of the complex (regression task). The scaled binding affinity or metric indicative of the stability of the complex can each be treated as equivalent to a probability of the doublet being a positive doublet. This model may be referred to as a "peptide-MHC model", "pMHC module" or "pMHC encoder", and is trained to predict the probability of binding, presentation and/or stability between a peptide and an MHC molecule on the basis of their sequence.

Training of the peptide MHC model can comprise training a model that takes a doublet of sequences (or information derived therefrom, by encoding using a predetermined encoding scheme) as input and produces as output a prediction of the binding affinity between the peptide sequence and a MHC molecule corresponding to the MHC sequence in the doublet. This can also be referred to as "binding affinity" prediction. This can be trained as a regression task (e.g. predicting a binding affinity metric value) or as a classification task (e.g. predicting whether the doublet has a binding affinity within each one of a plurality of classes associated with different respective ranges of binding affinities). In embodiments, the model is trained to provide as output a prediction of the binding affinity between the peptide sequence and a MHC molecule

corresponding to the MHC sequence in the doublet as a continuous value (i.e. the model is trained for a regression task), such as a scaled binding affinity metric.

Training of the peptide MHC model can comprise training a model that takes a doublet of sequences (or information derived therefrom, by encoding using a predetermined encoding scheme) as input and produces as output a classification between a first class comprising peptide-MHC pairs known to bind to each other and be presented on the surface of cells, and a second class comprising peptides-MHC pairs that are not expected to bind to each other and be presented on the surface of cells. This can also be referred to as "eluted ligand" prediction because the model can be trained using training data from immunopeptidomics (i.e. eluted ligand information). The model may be a model that has been previously pretrained for binding affinity prediction, and is fine-tuned for eluted ligand prediction. This may also be referred to as "transfer learning", i.e. a trained binding affinity prediction model can be trained for eluted ligand prediction using transfer learning.

Training of the peptide MHC model can comprise training a model that takes a doublet of sequences (or information derived therefrom, by encoding using a predetermined encoding scheme) as input and produces as output a metric indicative of the stability of a complex comprising the peptide and MHC molecule corresponding to the MHC sequence. This can be referred to as "stability" prediction. The model may be a model that has been previously pretrained for binding affinity and/or eluted ligand prediction, and is fine-tuned for stability prediction. Thus, the peptide-MHC model may have been trained using one or more transfer learning steps in which the model is pretrained for a first task and fine-tuned for one or more further tasks. The tasks can each be selected from binding affinity prediction, eluted ligand prediction and stability prediction. The metric indicative of stability can be a half-life or scaled half-life. For example, the metric indicative of stability of a complex comprising the peptide and MHC molecule corresponding to the MHC sequence can be a scaled half-life measured using a method as described in Rasmussen, M. et al. 2016. The metric indicative of stability can be a metric indicative of relative stability compared to one or more reference peptides (e.g. a percentage ELISA signal compared to a reference). For example, the metric indicative of stability can be a metric obtained using a NeoScreen assay as described in Lie-Andersen et al. 2023. The metric indicative of stability can be metric obtained through a TR-FRET assay as described in Gurung et al. 2023. The metric indicative of stability can be a metric indicative of the temperature yielding half-maximal denaturation of the complex. This can be a metric measured using UV-cleavable peptide/HLA class I complexes and differential scanning fluorimetry as described in Blaha et al. 2019. The metric indicative of stability can be a metric indicative of the presence of stable complexes on TAP knockout cells, as described in Kaseke et al. 2021. In embodiment, the model is trained to provide as output the value of a metric that quantifies the stability of a complex comprising the peptide and MHC molecule corresponding to the MHC sequence (i.e. the model is trained for a regression task), such as a scaled half-life or

metric correlating with half-life. Alternatively, the model can be trained as a classification task (e.g. predicting whether the doublet has a binding stability within each one of a plurality of classes associated with different respective ranges of a metric that quantifies the stability of a complex comprising the peptide and MHC molecule corresponding to the MHC sequence).

Training of the peptide-MHC model (comprising the peptide-MHC encoder that will be used in the full immunogenicity model) at step 14A'' can comprise training a model that takes as input peptide and MHC sequences for one or more or all of the following tasks (each of which can be performed as described above): determine whether a peptide-MHC pair is likely to bind to each other (binding affinity prediction), determine whether a peptide-MHC is likely to form a peptide-MHC complex that is presented on the surface of cells (eluted ligand prediction), and determine whether the peptide and MHC molecule are likely to form a stable complex (stability prediction). Each task can be set up as a regression task or a classification task. The binding affinity prediction task can be set up as a regression task. This is advantageous as continuous binding affinity data is often available and as such regression more completely captures the range of possible binding affinities between peptides ad MHC molecules. The eluted ligand prediction can be set up as a classification task. This is advantageous as eluted ligand data is typically binary. The stability prediction task can be set up as a regression task, particularly when continuous binding stability data is available for training, such as e.g. measured half lives or metrics indicative of half-life of peptide-MHC complexes. Training for any of the tasks can be performed as a transfer learning step after training for another one of the tasks. Thus, training of the peptide-MHC model can comprise training the model for a first task, and using the trained weights as a starting point to train the model for a second task. The trained weights after this step can optionally be further used as a starting point to train the model for a third task. The first, second and third tasks can be selected from any of the above. In other words, the peptide-MHC model can be trained for binding affinity prediction, eluted ligand prediction and/or stability prediction in any order. In embodiments, such as e.g. as demonstrated in the examples, the peptide-MHC model is trained for binding affinity prediction, then eluted ligand prediction, then optionally stability prediction. Alternatively, the peptide-MHC model can be trained only for eluted ligand or binding affinity prediction. Each of the training steps of the peptide-MHC model may comprise training the model to produce an output between 0 and 1. In the context of a classification task, this can be the probability of a peptide-MHC belonging to a first class. In the context of a regression class, this can be a normalised binding affinity or stability metric. The model can be configured such that the first class is a class associated with positive peptide-MHC pairs. Positive peptide-MHC pairs in this context are pairs that are presented by the MHC, bind the MHC and/or form a stable complex with the MHC molecule. In other words, positive peptide-MHC pairs can be pairs where the peptide is an eluted ligand for the MHC, the peptide has a binding affinity with the MHC molecule sufficient to be considered a binder, and/or the peptide and MHC molecule form a stable complex. The model can be configured such that the normalised binding affinity or stability metric is such that positive peptide-MHC pairs are associated with scores

closer to 1 than negative peptide-MHC pairs. Thus, the model can be trained using binding affinity and/or stability training data that has been normalised to be between 0 and 1 and to be such that values closer to 1 are indicative of higher binding affinity / stability than values closer to 0. As the skilled person understands, the reverse is also possible, for example the model can be trained to predict the probability of a peptide not being an eluted ligand, and a normalised binding affinity and/or stability metric between 0 and 1 where values closer to 0 are indicative of higher binding affinity / stability than values closer to 1. This advantageously means that the same classification / regression head can be fine-tuned for all tasks. For example, a fully connected network (e.g. a 2-layer network) can be used as a classification head for the eluted ligand prediction task, then used in a transfer learning step as a regression head for the binding affinity or stability prediction tasks (and vice versa).

Training of the peptide-MHC model may instead or in addition to peptide-MHC pairs from positive triplets, use sequences from peptide-MHC complexes that are known to bind to each other but that are not necessarily part of the positive triplets in the training data obtained at step 10'. In other words, the peptide-MHC model may be trained to predict peptide-MHC binding independently from immunogenicity. As such, step 14A'' may comprise obtaining one or more further paired peptide and MHC molecule sequences, for example from one or more databases. Thus, the full training data set may comprise: (a) the triplet data obtained at step 10' (used to train the whole immunogenicity prediction model as will be described further below, and optionally also to pretrain the TCR and/or peptide-MHC models); and optionally (b) TCR sequence data obtained at step 14A' (used to pretrain the TCR model) and/or (c) peptide-MHC sequence data obtained at step 14A'' (used to train the peptide-MHC model). The MHC sequence used as input for the peptide-MHC model may be a pseudosequence. Schemes for obtaining pseudosequences for MHC molecules are described e.g. in O'Donnell et al. 2020 and Jurtz et al. 2017. These use specific positions selected by multiple sequence alignment as evolutionary conserved. The term "MHC sequence" is used throughout this disclosure to encompass both a "real" sequence (i.e. a part or whole amino acid sequence of an MHC molecule), and a pseudosequence, unless context indicates otherwise. Similarly, the TCR sequence used as input for the TCR model may be a pseudosequence and the term "TCR sequence" is used throughout this disclosure to encompass both a "real" sequence (i.e. a part or whole amino acid sequence of a TCR molecule), and a pseudosequence, unless context indicates otherwise. In embodiments, the TCR sequence used as input for the TCR model is a "real" TCR sequence. A pseudosequence may be a string of characters derived from a sequence and which includes information about selected positions in the sequence. For example, a pseudosequence for an MHC sequence may comprise the amino acids at a plurality of selected positions in the sequence, the plurality of positions being selected based on the expected relevance of the positions. Expected relevance of the positions may be derived from e.g. alignment of MHC sequences across a plurality of species. Thus, the selected positions may be positions that are conserved across a plurality of species. Instead or in addition

to evolutionary conserved residues, residues may be chosen as those at specific positions in a multiple sequence alignment that are believed to be important to the function of the molecule, such as e.g. residues of an MHC molecule that are thought to contact a cognate peptide. This may be based on predicted or experimentally determined structural information instead of or in addition to evolutionary conservation. At step 14A'", the full immunogenicity model may be trained using the triplet data obtained at step 10'. Steps 14A' and 14A'' may be referred to as "pretraining", while step 14A''' may be referred to as "training". The training may comprise determining the parameters of the full immunogenicity model using training data and the parameters of any parts of the pretrained models included in the full immunogenicity model. The parameters of any parts of the pretrained models included in the full immunogenicity model can be frozen (i.e. fixing those parameters to the values obtained in pretraining for some or all of the training process), or fine-tuned (i.e. used as starting parameters, which are further trained as part of the full immunogenicity model, for the immunogenicity prediction task). Parameters of the models may also be referred to as "weights". The model may comprise the encoder or pair of encoders of the pretrained TCR model, and the encoder of the pretrained peptide-MHC model. The model may combine the outputs of the encoders (e.g. by concatenation), and further process the combined output in an immunogenicity prediction block. This further processing allows the model to extract information associated with the combination of the TCR and peptide-MHC sequences and transform this to a probability (a score between 0 and 1). By contrast, each of the pretrained encoders learn information associated with TCR sequence (for the TCR model encoder(s)) and peptide-MHC sequence and binding (for the peptide-MHC model encoder). The immunogenicity prediction block may comprise a natural language processing model such as e.g. a model comprising a transformer block. The immunogenicity prediction block may further comprise e.g. one or more fully connected layers and an activation function to obtain a single value between 0 and 1 that can be interpreted as the probability that the TCR, peptide and MHC molecule will bind (and hence that the peptide is immunogenic in the context of the TCR and MHC molecule). During training of the full immunogenicity model, the weights of the pretrained encoders may be fixed (also referred to as "frozen") and the remaining parameters of the model may be trained (e.g. parameters of the immunogenicity prediction block). Alternatively, the weights of the pretrained encoders may be fine-tuned during training of the full immunogenicity model.

Training the full immunogenicity model at step 14A''' may comprise a first step in which a peptide-MHC immunogenicity model is trained, and a second step in which the full immunogenicity model comprising this peptide-<HC immunogenicity model is fine tuned. A peptide-MHC immunogenicity model is a machine learning model that has been trained to take as input a doublet of sequences comprising an amino acid sequence of a peptide encoding the antigen, and an amino acid sequence of a candidate MHC molecule or a part thereof, or information derived from the doublet of sequences, and provide as output a score representing the probability that the antigen is immunogenic in the context of the candidate MHC molecule. This model comprises the previously

trained peptide-MHC model encoder, but does not comprise the TCR model encoder(s). The peptide-MHC immunogenicity model may be trained using training data comprising amino acid sequences or information derived therefrom for (i) positive peptide-MHC doublets comprising a peptide and MHC sequences that have been experimentally demonstrated to form an immunogenic complex; and (ii) negative peptide-MHC doublets comprising: a. a first set of one or more peptide-MHC doublets each comprising: (i) a MHC molecule selected from the positive peptide-MHC doublets and a peptide sequence not known to interact with the selected MHC molecule, optionally a randomly sampled peptide sequence, and; b. a second set of one or more peptide-MHC doublets each comprising: (i) a peptide-MHC pair comprising an MHC molecule and a peptide known to bind the MHC molecule (positive peptide-MHC pair), wherein the peptide-MHC pair has been previously found to not be immunogenic (non-immunogenic positive peptide-MHC pair). Indeed, the present inventors have found that such a peptide-MHC immunogenicity model already improves on prior art models for immunogenicity prediction, even without inclusion of the TCR encoder (although inclusion of the TCR encoder and training of the resulting full model using the three types of negative triplets as described herein was found to yield even greater benefits).

Thus, the present disclosure also provides a computer-implemented method of predicting whether an antigen is likely to be immunogenic, the method comprising:

obtaining a doublet of sequences comprising: an amino acid sequence of a peptide encoding the antigen, and an amino acid sequence of a candidate MHC molecule or a part thereof; and

providing the triplet of sequences or information derived therefrom as inputs to a machine learning model trained to predict a score representing the probability that the antigen is immunogenic in the context of the candidate MHC molecule (peptide-MHC immunogenicity model), wherein the machine learning model has been trained using training data comprising amino acid sequences or information derived therefrom for:

(i) positive peptide-MHC doublets comprising a peptide and MHC sequences that have been experimentally demonstrated to form an immunogenic complex; and

(ii) negative peptide-MHC doublets comprising: a. a first set of one or more peptide-MHC doublets each comprising: a MHC molecule selected from the positive peptide-MHC doublets and a peptide sequence not known to interact with the selected MHC molecule, optionally a randomly sampled peptide sequence, and; c. a second set of one or more peptide-MHC doublets each comprising: a peptide-MHC pair comprising an MHC molecule and a peptide known to bind the MHC molecule (positive peptide-MHC pair), wherein the peptide-MHC pair has been previously found to not be immunogenic (non-immunogenic positive peptide-MHC pair).

Similarly, also provided herein is a computer-implemented method of providing a tool for predicting whether an antigen is likely to be immunogenic, the method comprising:

(1) obtaining a training dataset comprising amino acid sequences or information derived therefrom for a plurality of peptide-MHC doublets, each doublet comprising an amino acid sequence of a peptide encoding the antigen, and an amino acid sequence of a candidate MHC molecule or a part thereof, wherein the plurality of peptide-MHC doublets comprise:

(i) positive peptide-MHC doublets comprising a peptide and MHC sequences that have been experimentally demonstrated to form an immunogenic complex; and

(ii) negative peptide-MHC doublets comprising: a. a first set of one or more peptide-MHC doublets each comprising: a MHC molecule selected from the positive peptide-MHC doublets and a peptide sequence not known to interact with the selected MHC molecule, optionally a randomly sampled peptide sequence, and; b. a second set of one or more peptide-MHC doublets each comprising: a peptide-MHC pair comprising an MHC molecule and a peptide known to bind the MHC molecule (positive peptide-MHC pair), wherein the peptide-MHC pair has been previously found to not be immunogenic (non-immunogenic positive peptide-MHC pair); and

(2) training, using said training data, a machine learning model that predicts the probability that an antigen is immunogenic in the context of a candidate MHC molecule provided as a doublet of sequences or information derived therefrom as input to the machine learning model.

The methods above using doublets of sequences can have any of the features described herein in relation to a model that takes a triplet of sequences as input, to the extent that they are applicable to a doublet model. Further, these methods can also be used in the context of any of the methods described herein in relation to a model that takes a triplet of sequences as input, such as methods of identifying one or more tumour-specific peptides that are likely to be immunogenic, methods of characterising an immunogenic composition comprising a plurality of candidate peptides or sequences encoding a plurality of candidate peptides, methods of designing or providing an immunotherapy for a subject that has been diagnosed as having cancer.

For example, the peptides in the first set of one or more peptide-MHC doublets can have been selected from a database or reference dataset. For example, optionally the peptides in the first set may have been randomly selected from a reference proteome. The peptide-MHC pair in the second set may have been experimentally identified as not immunogenic. The training data can comprise a ratio of negative doublets to positive doublets of at least 100:1, at least 150:1, between 100:1 and 300:1, preferably between 150:1 and 250:1, or around 200:1.

The machine learning model can take as input the doublet of amino acid sequences and produce an encoding for each sequence. Alternatively, the machine learning model can take as input an encoding for each sequence of a doublet of amino acid sequences. The amino acid sequences can be encoded using encoding schemes selected from: a predetermined token for each amino acid and optionally a padding character, one-hot-encoding, an encoding using a substitution matrix, an encoding using an embedding matrix, and an encoding using physicochemical descriptors. The one or more of the amino acid sequences can be encoded as fixed length strings

with a token for each amino acid and a padding character. The peptide sequence can be encoded as a fixed length string. The MHC sequence can be encoded as a pseudosequence with fixed length. The machine learning model can be a deep learning model. The machine learning model can comprise one or more natural language processing models. The machine learning model can comprise an encoder (referred to as "second encoder" in the context of triplet-based models) for encoding the peptide and MHC sequences. The encoder may have been pretrained prior to training the machine learning model using the training data comprising the negative doublets. The machine learning model may have been trained using the training data comprising the negative doublets with the parameters of the encoder maintained to their pretrained values. Alternatively, the training of the machine learning model using the training data comprising the negative doublets can include fine-tuning the parameters of the encoder. The encoder can take as input a peptide sequence and an MHC sequence. The encoder can have been trained as part of a model trained to predict whether the peptide is likely to bind the MHC molecule, whether the peptide is likely to be presented by the MHC molecule, and/or whether the peptide and MHC molecule are likely to form a stable complex. The encoder can have been trained as part of a model trained to predict the binding affinity between a peptide sequence and a MHC molecule corresponding to the MHC sequence, trained to classify pairs comprising a peptide sequence and an MHC sequence between a first class comprising peptide-MHC pairs known to bind to each other and be presented on the surface of cells, and a second class comprising peptide-MHC pairs that are not expected to bind to each other and be presented on the surface of cells, and/or trained to predict a metric indicative of the stability of a complex comprising the peptide and MHC molecule corresponding to the MHC sequence. The encoder can have been trained as part of a model that has been: (i) pretrained to predict whether the peptide is likely to bind the MHC molecule, and/or whether the peptide is likely to be presented by the MHC molecule; and (ii) fine-tuned for predicting whether the peptide and MHC molecule are likely to form a stable complex. At step (i) the encoder may have been trained as part of a model that has been pretrained to predict whether the peptide is likely to bind the MHC molecule, then further trained using transfer learning to predict whether the peptide is likely to be presented by the MHC molecule. A model trained or fine-tuned to predict whether the peptide and MHC molecule are likely to form a stable complex can be a model configured to take as input a peptide and MHC sequence or information derived therefrom and produce as output a metric indicative of the stability of a complex comprising the peptide and MHC molecule corresponding to the MHC sequence. The metric can be a half-life or scaled half-life. The encoder can be selected from: transformer-based encoders, autoencoders, and recurrent neural network encoders such as long-short-term memory (LSTM) networks. The encoder can be a transformer-based encoder.

The machine learning model (peptide-MHC immunogenicity model) can comprise the peptide-MHC encoder and further comprises a deep learning block that takes as input the output of the encoder, and produces as output the probability that the antigen is immunogenic in the context of the candidate MHC molecule. The deep learning block can comprise a first transformer block, and

a second block comprising one or more fully connected layers producing a single numerical output and optionally a sigmoid activation function.

The method can comprise (i) repeating one or more times the steps of: obtaining a doublet of sequences, and providing the doublet of sequences or information derived therefrom as inputs to the machine learning model, wherein each doublet of sequences differs in the amino acid sequence of the candidate MHC molecule or part thereof, thereby obtaining a plurality of respective probabilities that the antigen is immunogenic; and (ii) selecting the highest of the plurality of probabilities as the probability that the antigen is immunogenic. The method can further comprise identifying the antigen from a sample, and/or performing the method using one or more candidate MHC molecules identified from a sample wherein the sample is optionally a sample from which the antigen has been identified or a related sample.

At step 16', the trained model may be provided to a user for use in predicting whether a peptide is likely to be immunogenic. The model may optionally be evaluated for example to quantify its prediction accuracy, sensitivity and specificity or derived values such as e.g. a ROC (receiver operating characteristic) curve or AUROC, area under the ROC curve, also known as AUC, area under the curve, as known in the art. The models described herein may have an AUC of at least 0.7, advantageously at least 0.75 when evaluated on an independent test set. An independent test set may be a test set that is not related to the training set, such as e.g. a data set that does not comprise any part of the training data, that is not a subset of the training data, or that is not primarily made of data that is also in the training data. The models described herein may have an AUC of at least 0.6 when evaluated using a test set comprising no peptides present in the training set used to train the model.

Applications

The above methods find applications in the context of designing immunotherapies, particularly immunotherapies that use peptides or sequences encoding peptides to generate or promote an immune response. Indeed, peptides that are predicted to be likely to be immunogenic (in general or in the context of a specific set of candidate MHC molecules and/or candidate TCR molecules, such as e.g. based on the TCR repertoire and/or MHC alleles identified to be present in a sample or patient) are more promising candidates for inclusion in the immunotherapy. In particular, the above methods may be used to provide cancer immunotherapies that target cancer-specific antigens (also referred to herein as "cancer neoantigens", or simply "neoantigens"). As the skilled person understands, a cancer-specific antigen may be truly specific to cancer cells (in the sense that it is only expressed by the genome of cancer cells), or may be practically specific to cancer cells (in the sense that it is expressed by cancer cells at a significantly higher level than by normal cells). The cancer neoantigens may be clonal neoantigens. Thus, also described herein are

methods of providing an immunotherapy for a subject, the method comprising identifying and optionally producing one or more peptides that comprise cancer neoantigens predicted to be immunogenic, wherein the identifying is based on data from one or more samples from the subject and further comprises predicting whether one or more candidate cancer neoantigen peptides are likely to be immunogenic using a method as describe herein. An example of such a method will be described by reference to Figure 3.

The methods described herein further find applications in the context of identifying peptides or TCR sequences that are likely to be responsible for an observed reactivity in a sample. For example, a plurality of peptides may be tested for immunogenicity by detecting T cell activation *in vitro*, using assays that detect "reactivity" of a T cell population to the one or more peptides. The "reactivity" of a T cell population to one or more antigens, such as e.g. tumour antigens, refers to the presence and/or magnitude of an activation response of one or more cells in the T cell population in response to exposure to said one or more antigens. The activation of T cells can be assessed by detecting the presence of one or more markers of activation (e.g. IL2RA/CD25), the secretion of one or more cytokines (e.g. IFNg, TNFa), and/or the proliferation of T cells. Assays for measuring T cell activation are known in the art. In such cases, the methods of the present disclosure may be used to identify which one or ones of the plurality of peptides may have triggered the reactivity, and/or which one or ones of the TCRs expressed by the population of T cells may have been responsible for the detected reactivity. This information can be used to design immunotherapies, such as e.g. immunotherapies that use immunogenic peptides or sequences encoding such peptides and/or immunotherapies that use modified T cells that express a predetermined TCR.

Figure 3 illustrates schematically an exemplary method of designing or providing an immunotherapy. At optional step 310, one or more samples comprising tumour genetic material and one or more germline samples are obtained from a subject. The subject may be a subject that has been diagnosed as having cancer, and may be (but does not need to be) the same subject for which the immunotherapy is provided. At step 312, a list of candidate neoantigens is obtained using methods known in the art, for example as described in WO 2022/207925, WO 2016/16174085, Landau et al. (2013), Lu et al. (2018), Leko et al. (2019), Hundal et al. (2019), and others. The neoantigens may be clonal neoantigens. Methods to identify clonal neoantigens are known in the art and include the methods described in WO 2022/207925, WO 2016/16174085, Landau et al. (2013), Roth et al. (2014), McGranahan et al. (2016). The clonal neoantigens may in particular be identified using a method as described in WO 2022/207925. The list may comprise a single neoantigen, or a plurality of neoantigens. Preferably, the list comprises a plurality of neoantigens. At optional step 313, one or more TCR sequences are obtained, for example by TCR sequencing of one or more samples from the subject. Alternatively, candidate TCR sequences may be obtained from a database or other data sources. At step 314, an immunotherapy that

targets at least one (and optionally a plurality) of the candidate neoantigens is designed. Designing such an immunotherapy comprises identifying one or more candidate peptides for each of the candidate clonal neoantigens (step 314A). For example, a plurality of peptides may be designed for at least one of the candidate clonal neoantigens, which differ in their lengths and/or the location of a sequence variation that characterises the neoantigen compared to the corresponding germline peptide. At step 314B, the one or more peptides identified are analysed to determine whether they are likely to be immunogenic using a method as described herein, and optionally one or more additional properties such as their expression in the subject's tumour, expression in reference samples or datasets, similarity to a corresponding normal peptide, manufacturability (e.g. as described in application PCT/EP2023/055383), etc. At step 314C, one or more of the peptides are selected for production based on at least some of the results of step 314B.

At step 316, the selected peptides (or sequences encoding said peptides) may be obtained. Peptides with selected sequences may be obtained using any method known in the art but they are preferably obtained using chemical synthesis. Methods for obtaining sequences that encode peptides of interest are known. For example, tandem minigenes may be obtained which encode the selected one or more peptide. At step 318, an immunotherapy may be produced using at least some of the one or more peptides or sequences encoding said peptides produced at step 316. The immunotherapy may comprise the one or more peptides (e.g. in the case of an immunogenic composition such as a synthetic long peptide vaccine), sequences encoding said peptides (e.g. in the case of a DNA or RNA vaccine) or may comprise molecules or cells that have been obtained using the selected peptides (e.g. in the case of therapeutic antibodies that selectively bind the candidate peptides, or immune cells that specifically recognise the candidate peptides). In the illustrated embodiment, the immunotherapy comprises cells that have been obtained using the selected peptides. Methods of producing an immunotherapy comprising cells that have been obtained using neoantigen peptides are known in the art, for example as described in WO 2022/207925, WO 2016/16174085, McGranahan et al. (2016), Lu et al. (2018), Leko et al. (2019), Robbins et al. (2013). At optional step 320, the immunotherapy may be administered to a subject, which is preferably the subject from which the samples used to identify the neoantigens have been obtained. An example of producing an immunotherapy comprising a T cell population selectively enriched with T cells that recognise one or more neoantigens, preferably clonal neoantigens, will be described. At step 318A, a population of T cells may be obtained. The T cells may be obtained from the subject to be treated, but do not need to be. The T cells may be obtained from a tumour sample, from a blood sample, or from any other tissue sample. At step 318B, a population of antigen presenting cells (e.g. dendritic cells) may be obtained. For example, a population of dendritic cells may be derived from mononuclear cells (e.g. peripheral blood mononuclear cells, PBMCs) from the subject to be treated. At step 318C, the population of dendritic cells may be pulsed with the selected peptides. At step 318D, the T cell population may be selectively expanded

using the population of pulsed dendritic cells. Additional expansion factors such as e.g. cytokines or stimulating antibodies may be used.

Thus, the disclosure provides a method of providing an immunotherapy for a subject that has been diagnosed as having cancer, the method comprising: optionally identifying one or more cancer neoantigens for the subject, and designing an immunotherapy that targets one or more of the cancer neoantigens, wherein the designing comprises performing the method of the first aspect for one or more candidate peptides comprising the one or more of the cancer neoantigens. The method may have any one or more of the following features. The immunotherapy that targets the one or more of the cancer neoantigens may be an immunogenic composition, a composition comprising immune cells or a therapeutic antibody. The immunogenic composition may comprise one or more of the candidate peptides (such as e.g. a neoantigen peptide or protein or a cell displaying the neoantigen). The composition comprising immune cells may comprise T cells, B cells and/or dendritic cells. The composition comprising a therapeutic antibody may comprise one or more antibodies that recognise at least one of the one or more of the candidate peptides. An antibody may be a monoclonal antibody. The immunogenic composition may comprise one or more nucleic acids encoding the one or more peptides, or a construct comprising such a nucleic acid.

Designing an immunotherapy that targets one or more of the cancer neoantigens identified may comprise designing one or more candidate peptides for each of the one or more neoantigens targeted, each peptide comprising at least a portion of a neoantigen targeted. The method may further comprise obtaining the one or more candidate peptides. The method may further comprise testing the one or more candidate peptides for one or more further properties. Further testing may be performed *in vitro* or *in silico*. For example, the one or more peptides may be tested for immunogenicity, propensity to be displayed by MHC molecules (optionally by specific MHC molecule alleles, where the alleles may have been chosen depending on the MHC alleles expressed by the subject), ability to elicit proliferation of a population of immune cells, etc. A plurality of the one or more peptides may be tested simultaneously for immunogenicity, and upon determining that the plurality of peptides are able to elicit an immune reaction, the methods described herein may be used to identify a subset (including a complete subset) of the one or more peptides that is likely to have caused the observed immune reaction. Thus, also described herein are methods for characterising a plurality of peptides and compositions comprising such peptides (including immunotherapies), and methods for selecting peptides from a plurality of peptides, the methods comprising determining that the plurality of peptides is able to elicit an immune reaction by *in vivo* or *in vitro* testing (preferably *in vitro testing* or using results previously obtained from *in vitro* tests, e.g. from a clinical trial database or other data source), and identifying one or more of the plurality of peptides that are likely to elicit an immune reaction using a method as described herein. The one or more of the plurality of peptides may be peptides amongst the plurality of

peptides that have a probability of being immunogenic as determined used a method as described herein that satisfies one or more criteria. The one or more criteria may be selected from: a probability above a predetermined threshold, a probability in the top 1, 5, 10, 15 or 20 of ranked probabilities for the plurality of peptides, or a probability in the top 1, 5, 10, 15, 20, 30, 40 or 50% of ranked probabilities for the plurality of peptides. The one or more identified peptides may be used to design an immunotherapy that targets said peptides.

A method of designing, providing or characterising an immunotherapy may further comprise producing the immunotherapy. The method may further comprise obtaining a population of dendritic cells that has been pulsed with one or more of the candidate peptides. The immunotherapy may be a composition comprising T cells that recognise at least one of the one or more of the neoantigens identified. The composition may be enriched for T cells that target at least one of the one or more of the neoantigens identified. The method may comprise obtaining a population of T cells and expanding the population of T cells to increase the number or relative proportion of T cells that target at least one of the one or more of the neoantigens identified. The method may further comprise obtaining a T cell population. A T cell population may be isolated from the subject, for example from one or more tumour samples obtained from the subject, or from a peripheral blood sample or a sample from other tissues of the subject. The T cell population may comprise tumour infiltrating lymphocytes. T cells may be isolated using methods which are well known in the art. For example, T cells may be purified from single cell suspensions generated from samples on the basis of expression of CD3, CD4 or CD8. T cells may be enriched from samples by passage through a Ficoll-paque gradient. The method may further comprise expanding the T cell population. For example, T cells may be expanded by ex vivo culture in conditions which are known to provide mitogenic stimuli for T cells. By way of example, the T cells may be cultured with cytokines such as IL-2 or with mitogenic antibodies such as anti-CD3 and/or CD28. The T cells may be co-cultured with antigen-presenting cells (APCs), which may have been irradiated. The APCs may be dendritic cells or B cells. The APCs, for example dendritic cells, may have been pulsed with the candidate peptides (containing one or more of the identified neoantigens) as single stimulants or as pools of stimulating neoantigen peptides. Expansion of T cells may be performed using methods which are known in the art, including for example the use of artificial antigen presenting cells (aAPCs), which provide additional co-stimulatory signals, and autologous PBMCs which present appropriate peptides. The APCs may be pulsed with peptides containing neoantigens as discussed herein as single stimulants, or alternatively as pools of stimulating neoantigens.

Also described herein is a method for expanding a T cell population for use in the treatment of cancer in a subject, the method comprising: identifying one or more neoantigen peptides that are likely to be immunogenic using a method as described herein; obtaining a T cell population comprising a T cell which is capable of specifically recognising one of the neoantigen peptides;

and co-culturing the T cell population with a composition comprising the neoantigen peptide. The method may have one or more of the following features. The T cell population obtained may be assumed to comprise a T cell capable of specifically recognising one of the neoantigen peptides. The method preferably comprises identifying a plurality of neoantigen peptides. The neoantigen peptides may comprise one or more clonal neoantigens. The T cell population may comprise a plurality of T cells each of which is capable of specifically recognising one of the plurality of neoantigen peptides, and co-culturing the T cell population with a composition comprising the plurality of neoantigen peptides. The co-culture may result in expansion of the T cell population that specifically recognises one or more of the neoantigen peptides. The expansion may be performed by co-culture of a T cell with the one or more neoantigen peptides and an antigen presenting cell. The antigen presenting cell may be a dendritic cell. Thus, the expansion may be a selective expansion of T cells which are specific for the neoantigen peptides. The expansion may further comprise one or more non-selective expansion steps. Thus, also described herein is a composition comprising a population of T cells obtained or obtainable by a method as described above.

Thus, the disclosure also provides a T cell composition comprising a T cell population selectively enriched with T cells that recognise one or more neoantigens, preferably clonal neoantigens, wherein the T cell population has been selectively enriched using peptides that have been produced using any of the methods described herein.

In a T cell composition as described herein the expanded population of neoantigen-reactive T cells may have a higher activity than the population of T cells which have not been expanded, as measured by the response of the T cell population to restimulation with a neoantigen peptide. Activity may be measured by cytokine production, and wherein a higher activity is a 5-10 fold or greater increase in activity.

References to a plurality of neoantigens may refer to a plurality of peptides or proteins each comprising a different tumour-specific mutation that gives rise to a neoantigen. Said plurality may be from 2 to 250, from 3 to 200, from 4 to 150, or from 5 to 100 tumour-specific mutations, for example from 5 to 75 or from 10 to 50 tumour-specific mutations. Each tumour-specific mutation may be represented by one or more neoantigen peptides. In other words, a plurality of neoantigens may comprise a plurality of different peptides, some of which comprise a sequence that includes the same tumour-specific mutation (for example at different positions within the sequence of the peptide, or within peptides of varying lengths). Thus, the one or more selected peptides obtained at step 216 (or any method comprising selecting peptides using the methods described herein) may comprise from 2 to multiple hundred peptides, such as e.g. between 2 and 400 peptides, between 2 and 300 peptides, between 2 and 250 peptides, between 2 and 200 peptides, between 10 and 400 peptides, between 10 and 300 peptides, between 10 and 250 peptides, between 10 and 200 peptides, between 50 and 400 peptides, between 50 and 300 peptides, between 50 and

250 peptides, or between 50 and 200 peptides. In particular, the one or more selected peptides may comprise up to a maximum number of peptides that is set by the capacity of a synthesis process or a step thereof, such as for example the number of wells in a reaction plate used for a single synthesis run or a multiple thereof. For example, when 96 wells plates are used the number of selected peptides may be set to a maximum of 96, 192, 288, or 384. Instead or in addition to this, in the context of peptides comprising a different tumour-specific mutation that gives rise to a neoantigen, the number of peptides selected may be set to a maximum corresponding to the number of tumour-specific mutations that give rise to a neoantigen identified in a subject, or to the number of different peptides of a predetermined length that comprise said tumour-specific mutations. For example, as many as 1,000 to 10,000 peptides comprising one or more coding mutations may be identified and peptides comprising each of said mutations may be selected using the methods described herein. A T cell population that is produced in accordance with the present disclosure will have an increased number or proportion of T cells that target one or more neoantigens that are represented in peptides selected using the methods described herein. That is to say, the composition of the T cell population will differ from that of a "native" T cell population (i.e. a population that has not undergone the expansion steps discussed herein), in that the percentage or proportion of T cells that target a neoantigen that is produced as described herein will be increased. The T cell population according to the disclosure may have at least about 0.2, 0.3, 0.4, 0.5, 0 6, 0 7, 0 8, 0 9, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 or 100% T cells that target a neoantigen for which a peptide is produced as described herein.

The immunotherapies described herein may be used in the treatment of cancer. Thus, the disclosure also provides a method of treating cancer in a subject comprising administering an immunotherapeutic composition as described herein to the subject.

Suitably, in any embodiment of any aspect described herein, the cancer may be ovarian cancer, breast cancer, endometrial cancer, kidney cancer (renal cell), lung cancer (small cell, non-small cell and mesothelioma), bladder cancer, gastric cancer, oesophageal cancer, colorectal cancer, cervical cancer, endometrial cancer, brain cancer (gliomas, astrocytomas, glioblastomas), melanoma, merkel cell carcinoma, clear cell renal cell carcinoma (ccRCC), lymphoma, small bowel cancers (duodenal and jejunal), leukemia, pancreatic cancer, hepatobiliary tumours, germ cell cancers, prostate cancer, head and neck cancers, thyroid cancer and sarcomas. For example, the cancer may be lung cancer, such as lung adenocarcinoma or lung squamous-cell carcinoma. As another example, the cancer may be melanoma. The cancer may be bladder cancer. The cancer may be head and neck cancer. In embodiments, the cancer may be selected from melanoma, merkel cell carcinoma, renal cancer, non-small cell lung cancer (NSCLC), urothelial carcinoma of the bladder (BLAC) and head and neck squamous cell carcinoma (HNSC) and microsatellite

instability (MSI)-high cancers. In some embodiments, the cancer is non-small cell lung cancer (NSCLC). In any embodiment of any aspect, the subject may be human.

Treatment using the compositions and methods of the present disclosure may also encompass targeting circulating tumour cells and/or metastases derived from the tumour. Treatment according to the present disclosure targeting one or more neoantigens, preferably clonal neoantigens, may help prevent the evolution of therapy resistant tumour cells which may occur with standard approaches such as chemotherapy, radiotherapy, or non-specific immunotherapy. The methods and uses for treating cancer described herein may be performed in combination with additional cancer therapies. In particular, the immunotherapies (including but not limited to T cell compositions) described herein may be administered in combination with immune checkpoint intervention, co-stimulatory antibodies, chemotherapy and/or radiotherapy, targeted therapy, cancer vaccines or monoclonal antibody therapy. 'In combination' may refer to administration of the additional therapy before, at the same time as or after administration of the immunotherapy (e.g. T cell composition) as described herein.

The invention also provides a method for producing an immunotherapeutic composition, the method comprising predicting whether one or more candidate peptides each comprising a neoantigen are likely to be immunogenic, selecting one or more peptides from the candidate peptides based on the predicting, and producing an immunotherapeutic composition that targets the neoantigen(s).

Also described herein is a composition comprising a neoantigen peptide, neoantigen peptide specific immune cell, or an antibody that recognises a neoantigen peptide, for use in the treatment or prevention of cancer in a subject, wherein said neoantigen peptide has been identified using the methods described herein. Also described herein is a composition comprising a neoantigen peptide, neoantigen peptide specific immune cell, or an antibody that recognises a neoantigen peptide, wherein said neoantigen peptide has been produced using the methods described herein. Also described herein is a neoantigen peptide, immune cell which recognises a neoantigen peptide, or antibody which recognises a neoantigen peptide, for use in the treatment or prevention of cancer in a subject, wherein said neoantigen peptide has been produced using the methods described herein. Also described herein is the use of a neoantigen peptide, immune cell which recognises a neoantigen peptide, or antibody which recognises a neoantigen peptide, in the manufacture of a medicament for use in the treatment or prevention of cancer in a subject, wherein said neoantigen peptide has been produced using the methods described herein. Also described herein is a method of treating a subject that has been diagnosed as having cancer, the method comprising administering an immunotherapy that has been provided using the methods described herein, or a composition as described herein.

Systems

Figure 4 shows an embodiment of a system for predicting whether a peptide is likely to be immunogenic, for designing an immunotherapy, and/or for characterising an immunogenic composition according to the present disclosure. The system comprises a computing device 1, which comprises a processor 101 and computer readable memory 102. In the embodiment shown, the computing device 1 also comprises a user interface 103, which is illustrated as a screen but may include any other means of conveying information to a user such as e.g. through audible or visual signals. The computing device 1 is communicably connected, such as e.g. through a network 6, to sequence data acquisition means 3, such as a sequencing machine, and/or to one or more databases 2 storing sequence data. The one or more databases may additionally store other types of information that may be used by the computing device 1, such as e.g. reference sequences, parameters, etc. The computing device may be a smartphone, server, tablet, personal computer or other computing device. The computing device is configured to implement a method for predicting whether a peptide is likely to be immunogenic, for designing an immunotherapy, and/or for characterising an immunogenic composition, as described herein. In alternative embodiments, the computing device 1 is configured to communicate with a remote computing device (not shown), which is itself configured to implement a method of predicting whether a peptide is likely to be immunogenic, for designing an immunotherapy, and/or for characterising an immunogenic composition, as described herein. In such cases, the remote computing device may also be configured to send the result of the method to the computing device. Communication between the computing device 1 and the remote computing device may be through a wired or wireless connection, and may occur over a local or public network such as e.g. over the public internet or over WiFi. The sequence data acquisition 3 means may be in wired connection with the computing device 1, or may be able to communicate through a wireless connection, such as e.g. through a network 6, as illustrated. The connection between the computing device 1 and the sequence data acquisition means 3 may be direct or indirect (such as e.g. through a remote computer). The sequence data acquisition means 3 are configured to acquire sequence data from nucleic acid samples, for example genomic DNA samples or RNA samples extracted from patient samples, such as e.g. tumour and/or normal samples (e.g. to identify tumour-specific mutations that can give rise to neoantigens and/or to identify HLA alleles present in the sample), samples comprising T cells purified from fluid and/or tissue samples (such as e.g. peripheral blood, spleen, lymph node, tumour tissue, or any other type of sample comprising B cells or T cells). In some embodiments, the sample may have been subject to one or more preprocessing steps such as DNA/RNA purification, fragmentation, library preparation, target sequence capture (such as e.g. exon capture and/or panel sequence capture). Any sample preparation process that is suitable for use in the determination of a T cell receptor sequence or repertoire, and/or for the identification of mutations and/or for the identification of HLA alleles present in a subject may be used within the context of the present invention. The sequence data acquisition means is preferably a next generation

sequencer. The sequence data acquisition means 3 may be in direct or indirect connection with one or more databases 2, on which sequence data (raw or partially processed) may be stored. The sequence data acquisition means 3 may instead or in addition be configured to acquire sequence data from peptide samples (such as e.g. by mass spectrometry). The sequence data acquisition means 3 may be located in a physically separate location from the computing device 1.

The following is presented by way of example and is not to be construed as a limitation to the scope of the claims.

EXAMPLES

These examples describe the training and benchmarking of machine learning models (hereafter referred to as the "Genesis" model) for predicting whether an antigen is likely to be immunogenic.

Data

In the present examples, immunogenicity prediction models were trained using a training data set comprising positive triplets (triplets of peptide-MHC-TCR amino acid sequences that have been found to be reactive in previous studies) and negative triplets (triplets of peptide-MHC-TCR amino acid sequences that are not expected to be reactive), as further described below. Models trained as described below are referred to as "Genesis" models.

Positive triplets. **Figure 5** illustrates the positive training dataset used to train the Genesis models. The training dataset is comprised of positive triplets comprising peptide-MHC-CDR3-β amino acid sequences that have been found to be reactive in previous studies. All of the data used is publicly available. In particular, the following datasets were used: VDJdb (vdjdb.cdr3.net, Goncharov et al., 2022), McPAS-TCR (Tickotsky et al., 2017, friedmanlab.weizmann.ac.il/McPAS-TCR), IEDB (www.iedb.org, Vita et al., 2019), and pMTnet (github.com/tianshilu/pMTnet, Lu et al., 2021). Models that also make use of CDR3-α sequences as input have also been investigated but are not illustrated here as currently public sources of paired CDR3-α sequences are more limited. The network architecture illustrated in these examples is flexible to format changes to the inputs.

The data was filtered to remove all non-human data, to remove entries with missing data (e.g. missing HLA allele), and to remove entries with modified amino acids in the TCR (such as e.g. selenocysteine or pyrrolysine, as the encoding scheme used in the exemplified implementation only handles the 21 normal proteinogenic amino acids – although different schemes are possible including extensions of the scheme used in these examples to include representations for one or more modified amino acids). Depending on the specific filters used, slightly different numbers of data points may be kept. The data used (including positive and negative triplets obtained as described above) contained approximately 7 million triplets in total, of which about 40,000 were positive triplets, comprising approximately 2.5 million unique peptides (including negative peptides

sampled as explained below), approximately 2 million unique TCRs and 122 human HLA sequences. A specific example included 7,394,684 triplets, including 37,163 positive triplets, 2,405,462 unique peptides, 2,113,003 unique TCRs and 122 human HLAs.

Negative triplets. Three types of negative data points were used:

1. Negative TCR: Positive pMHC complexes are matched with random CDR3-β sequences from TCRdb (Chen et al., 2021), a database with millions of TCR sequences. This widens the distribution of TCRs the model has been trained to recognise. These are "real" peptide-MHC interactions with the "wrong" TCR. The TCR sequences are sampled at each generation of the training process, using the chosen negative to positive ratio and a predetermined proportion of each of the three types of negative triplets to determine the number of TCRs to be sampled. For example, using a positive to negative ratio of 1:200, 66 TCRs are sampled for each positive triplet (using equal proportions of each of the three types of negative triplets – although other proportions are possible).

2. Negative Peptides: Positive MHC-CDR3 pairs are matched with random proteins from the human proteome extracted from the Consensus CDS database (www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi). These are sampled as n-mers between 8-15 amino acids in length, by selecting a random CCDS identifier and selecting a random start location within the sequence corresponding to the identifier, and checking that the sequence selected is not present in the positive peptide list (in which case another sequence is sampled instead). These are "real" MHC-TCR interactions with the "wrong" peptide. As above, the peptides are sampled at each generation of the training process, using the chosen negative to positive ratio and a predetermined proportion of each of the three types of negative triplets to determine the number of peptides to be sampled. For example, using a positive to negative ratio of 1:200, 66 peptides are sampled for each positive triplet (when using equal proportions of each of the three types of negative triplets).

3. Non-immunogenic pMHCs: Peptide-MHC combinations which have been found to be non-reactive in immunogenicity screening experiments were paired with CDR3 sequences from the positive triplet dataset to generate negatives which would be likely to form pMHC complexes. These are "real" peptide-MHC interactions that are not immunogenic, paired with "real" TCRs randomly selected. As above, these triplets are obtained by pairing TCRs sampled with replacement from the positive set and negative peptide-MHC pairs sampled with replacement from a set of negative peptide-MHC pairs (described below), using the chosen negative to positive ratio and a predetermined proportion of each of the three types of negative triplets to determine the number of triplets to be sampled for each positive triplet (e.g. sampling 66 TCRs for each positive triplet). 6084 negative peptide-MHC pairs were obtained from: (i) the TESLA study (Wells et al., 2020) and (ii) data from other peptidomics studies (Gfeller et al., 2022 – data in this paper combines data from multiple immunopeptidomics studies). 11,807 negative pMHCs were provided by the Hadrup lab (raw sequencing data available in Holm et al., 2022) which were used as holdout test

samples. The negative TCR set are pMHCs that are immunogenic given the right TCR, whereas the non-immunogenic pMHCs are not immunogenic, and capture the fact that there can be peptide-MHC binding without necessarily TCR recognition. Additionally, these negatives allow a comparison of the model of the present disclosure to other immunogenicity models that only look at the pMHC. The 'negative TCR' set only has immunogenic pMHCs so these models see them as one positive data point per pMHC (whereas they are in fact negative from an immunogenicity point of view). Those models are all very good at recognising random negative peptides, since the likeliness of them being presented by the MHC is low. When only using random peptides, most prior art p-MHC models can achieve high prediction performances (e.g. ROC AUC ~0.9), as does the model of the present disclosure. However, when using a more realistic set the performance of these models breaks down (as shown below).

In addition to a diverse set of negative sources, a high negative to positive ratio was used (200:1) for training. Other values of the negative to positive ratio are envisaged. For example, ratios between 10:1 and 250:1, preferably between 100:1 and 200:1 may be used. Ratios of approximately 200:1 have been found to be particularly advantageous. The effect of the negative to positive ratio on the performance of the model was investigated by testing a plurality of values between 1:1 and 300:1. For each value, a model was trained using 5 different initialisations of the negative datasets, and the mean prediction of models trained with the same negative to positive ratio was used to evaluate the models.

Train:Test split. From the positive set, cancer related peptides (in all datasets used, i.e. VDJdb, IEDB, McPAS-TCR, pMTNet as described above) were kept for the holdout test set. This comprised 4148 triplets comprising 450 unique peptides. The remaining 37,163 triplets were used for the training dataset. This ensured that no peptides appeared in both datasets, as well as providing challenging and realistic test scenarios for the models (where models are evaluated on peptides obtained from sources that may not necessarily be represented in the training data). Both had negatives independently generated at a rate of 200:1, equally proportioned between the three negative variants.

Stability prediction data. 28,198 half-life measurements were provided by the authors of the NetMHCstabpan paper (Rasmussen, M. et al. 2016) and one additional study of yellow fever vaccine epitopes (Stryhn, A. et al., 2020). The raw data is scaled in the same manner as demonstrated in the NetMHCstabpan paper, $S = 2^{-t0/hl}$, where S is the converted value and t0 is the conversion constant (hl=half life). The constant was set to 1, which was determined to be a reasonable pan-allotype value in Rasmussen, M. et al. 2016 and which was confirmed in the inventor's early developmental experiments. Negative entries were added from the data in O'Donnell et al., 2020 where the measurement value was over 20,000nM. 1000 negatives samples per MHC allele represented in the positive set were sampled. Performance was assessed by Pearson's correlation coefficient to the actual half-life values after 5-fold cross validation. The

scaled stability as described above is a number between 0 and 1, with 0 indicating a half-life of 0 hours.

In the present examples, immunogenicity prediction models were trained using training data set comprising positive doublets (doublets of peptide-MHC amino acid sequences that have been found to be reactive in previous studies) and negative doublets (doublets of peptide-MHC amino acid sequences that are not expected to be reactive), as further described below. Models trained as described below are also referred to as "Genesis" models (labelled as "peptide MHC immunogenicity prediction").

Positive doublets. Training for immunogenicity prediction uses positive epitopes from IEDB, VDJdb, TESLA, McPAS-TCR and the PRIME model training sets. Positive epitopes are peptide-MHC complexes that are immunogenic.

Negative doublets. Training for immunogenicity prediction uses two types of negative doublets (i) negative epitopes from IEDB (Vita et al. 2018), VDJdb (Goncharov et al. 2022) TESLA (Wells et al. 2020), McPAS-TCR (Tickotsky et al. 2017) and the PRIME model training sets (Gfeller et al. 2023); and (ii) negative peptides. Negative peptides are additional negatives that were sampled from the human proteome using the consensus coding sequence project (CCDS; Pruitt et al. 2009), as explained above in relation to "Negative Peptides" (i.e. MHC from positive doublets are matched with random proteins from the human proteome extracted from the Consensus CDS database www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi). Negative epitopes are peptide-MHC complexes that are not immunogenic (i.e. peptide-MHC complexes that have been demonstrated experimentally to form but to not be immunogenic). This is different from peptide-MHC doublets that are assumed to be unlikely to be immunogenic because they are unlikely to form a pMHC complex (as is the case for negative doublets that are obtained by random pairing of a peptide and MHC).

Importantly, the use of the above set of positive and negative doublets means that the model is truly trained for immunogenicity prediction (assuming no TCR knowledge). This is different from predicting peptide-MHC complex formation (i.e. binding affinity and/or eluted ligand status) as a proxy to immunogenicity. Further, the use of the two types of negatives ensures that the model can learn peptide-MHC sequence features that result in a lack of immunogenicity for different biological reasons including a lack of peptide-MHC binding, peptide presentation by MHC, and lack of recognition by / activation of TCR despite the peptide-MHC complex forming.

Methods

Implementation. As explained further below, the models exemplified here are designed as a modular immunogenicity model, trained in an iterative manner using different transfer learning techniques. The model is broken up into different reusable modules depending on the end prediction goal. All models were developed in Python 3.8.11 using PyTorch 1.12. Training was

performed on NVIDIA Tesla T4 GPUs with 16GB of VRAM. All models were trained with the Adamax optimiser to minimise binary cross entropy loss.

Model Architecture. The model architecture for the Genesis model is illustrated on **Figure 6**. The chosen model architecture is a language modelling approach, with all sequence embedding learned during the pre-training steps. The model utilises an end-to-end transformer approach (Vaswani et al., 2017), with separate transformer-based modules (602, 604) handling the CDR3 (606) and pMHC (608a, 608b) inputs. Each section of the model is modular, with the number of layers and configuration of layers (in particular, number of attention heads, encoding dimensions and linear dimensions) being tuneable parameters during the cross validation (i.e. a plurality of models are trained with different numbers of layers and configuration of layers, and the performance of the models are compared by cross-validation). The model comprises a pretrained encoder (602) that is trained to encode TCR sequence data, input as a CDR3-β sequence encoded using tokens representing each amino acid (606). The TCR encoder takes the form of an encoder-only transformer network with a fully connected layer at the head to reconstruct the input. For the final model the fully connected layer is removed, and the output of the transformer encoder layers is fed into the combined model. The model further comprises a pretrained peptide-MHC (pMHC) encoder (604) that is trained to encode peptide-MHC sequences as part of a model trained to predict the probability of a given pMHC being a true eluted ligand (results on Figures 7-11) or trained to predict a normalised binding affinity of a given peptide-MHC then further trained using transfer learning to predict the probability of a given pMHC being a true eluted ligand (results on Figures 13-15). The architecture of the pMHC encoder is a combined self and cross attention model. This takes the form of a pair of transformer-based input branches which encode the peptide and HLA sequences separately, before concatenating the attended sequences for input into a final set of transformer layers attending to the whole peptide-MHC sequence. This model is pretrained by use of a fully connected layer, trained to predict eluted ligands (either from scratch or as a fine tuning / transfer learning step from a model trained to predict binding affinity), which is removed in the final model. The fully connected layer is trained for a classification task (when predicting eluted ligands) or for a regression task (when predicting binding affinity or stability). The pretrained peptide-MHC encoder takes as inputs a peptide sequence encoded using tokens representing each amino acid (608a) and an MHC sequence encoded using tokens representing each amino acid (608b). In the model exemplified, the MHC sequence was provided as a pseudosequence comprising the amino acids at 34 positions selected based on a multiple sequence alignment of MHC class I molecules across a plurality of species, as described in O'Donnell et al., 2020 and Jurtz et al. 2017. In O'Donnell et al., the authors represented MHC class I alleles by the amino acids at 37 positions from a global multiple sequence alignment, comprising 34 peptide-contacting positions identified in Jurtz et al. 2017 (which describes the NetMHCpan 4.0 pseudosequence), and 3 additional positions that were selected to differentiate several pairs of alleles that shared identical 34-mer NetMHCpan 4.0 pseudosequences and were present above a threshold in the

training data set used by the authors. These additional positions were not found to be necessary in the present case as those alleles were not common in the training data. However, the 37 amino acids pseudosequence of O'Donnell et al. 2020 may also be used. Full length sequences of MHC molecules may also be used. Both the pMHC and TCR models take in tokens and learn encodings with an embedding depth of 128. The inputs of both encoders are the respective amino acid sequences converted to numbers 0-20 (20 amino acids plus a padding character). The inputs of all encoders have a fixed length (respectively 26 amino acids for the TCR, 16 amino acids for the peptide, and 34 amino acids for the MHC). Other lengths are possible, up to and including models that take a full MHC sequence and/or a full TCR chain or chains sequence as inputs. The TCR and MHC sequences are provided as sequences of length 26 and 34, respectively. The peptide sequences may be provided with variable lengths. When a peptide sequence with a length below 16 amino acids is used the peptide sequence is middle padded to keep the start and end of the peptide sequence at the start and end of the input string. Other padding schemes are envisaged. Other encoding schemes are possible, such as one-hot encoding, BLOSUM encoding (see e.g. Eddy, 2004; or other encodings using substitution matrices), or physicochemical character-based schemes such as the principal components score Vectors of Hydrophobic, Steric, and Electronic properties (VHSE8, see Mei et al., 2005), or encoding of amino acids with Atchley factors. For example, both the use of BLOSUM-50 encodings and Atchley factors (Atchley et al., 2005) were tested (data not shown) and found to perform satisfactorily, albeit not as well as the simple amino acid tokenisation scheme used in this example. Other input lengths are also possible. The outputs of both encoders are concatenated (610) and provided to a transformer head block (612), the output of which is fed into an immunogenicity prediction block (614) comprising a series of fully connected layers, producing a single output value that is input into a sigmoid function, generating a value between 0 and 1. The parameters of all modules, pMHC encoder, TCR encoder and immunogenicity prediction, (e.g. number of layers, transformer heads and encoding dimensions) are trainable parameters and models trained with a variety of parameters (e.g. 2, 4 or 6 layers, 6 or 8 heads, 72 or 128 encoding dimensions) were compared by cross-validation. In models also making use of the sequence of the CDR3- α, this can be appended to the sequence of the CDR3-β provided as input to the TCR encoder (602). Alternatively, a third encoder similar to the TCR encoder 602 can be pretrained in a similar manner, using CDR3-α sequences. The outputs of the pre-trained CDR3-α and CDR3-β can then be concatenated in a similar manner as explained above. Availability of paired CDR3-α - CDR3-β sequences is currently sufficient to pretrain a combined encoder or a pair of self-supervised TCR encoders. However, the availability of triplet data comprising both CDR3-α and CDR3-β sequences for the TCR is currently much lower. Therefore, only a CDR3-β encoder was included in the present model.

Other architectures are envisaged and may similarly benefit from the training scheme described herein. For example, the first encoder and second encoders may be selected from autoencoders and LSTM models, and the outputs of these may be fed directly into a deep neural network. The

present inventors have found end-to-end transformer-based architectures comprising a first and second transformer-based encoders followed by a transformer head block and an immunogenicity prediction block to perform particularly well.

Variants of the above models were also created using components of the model. **Figure 12** shows schematically the model architecture of these variants. **Figure 12A** shows the architecture of a peptide-MHC only based immunogenicity prediction model, trained with peptide-MHC immunogenicity data. **Figure 12B** shows the architecture of a peptide-MHC optional TCR immunogenicity prediction, trained with combination of peptide-MHC and peptide-MHC-TCR immunogenicity data. The model takes as input triplets comprising a peptide, MHC and TCR if available, and a peptide, MHC and empty TCR vector if no TCR information is available. **Figure 12B** also shows the architecture of a peptide-MHC-TCR immunogenicity prediction model, trained with peptide-MHC-TCR immunogenicity data, and optionally pre-trained using peptide-MHC immunogenicity data. **Figure 12C** shows the architecture of a p-MHC stability, elution and/or binding affinity prediction model used to pretrain the pMHC transformer encoder that is used in each of the models on Figures 12A-B. **Figure 12D** illustrates the pretraining of the TCR encoder used in the models of Figure 12B using a masked language modelling task.

Variants of the above model were also created in which peptide hydrophobicity information was included as part of the inputs to the peptide-MHC model. In particular, this was performed by adding to the embeddings of the peptide in the peptide-MHC module additional information related to the known hydrophobicity of each amino acid. In other words, each peptide amino acid can have an additional fixed embedding added matching its known hydrophobicity at the initial encoding step before being fed into the first transformer layer, similar to positional encoding embedding. This provides additional prior knowledge of similarities between amino acids to the model prior to training. Indeed, hydrophobicity or other physiochemical features can be embedded in the model embeddings if deemed necessary for a particular application where such features are found to be important. Genesis is capable of interpreting these features as additional embeddings. The known hydrophobicity values used were the Kyte-Doolittle scale values. Thus, the first transformer layer of the peptide-MHC module was provided as input, for the peptide sequence, an amino acid sequence embedding (of dimension equal to sequence length x embedding depth, here 72) and an amino acid sequence Kite-Doolittle embedding (of dimension equal to sequence length x embedding depth, here 72). Two versions were tested, one using a fixed embedding for the Kyte-Doolittle values, and one using a learned embedding. The latter at least showed a small improvement in terms of AUROC and average precision when tested on the CEDAR dataset (data not shown).

Training. The full model is trained in three steps. In a first training step, the TCR transformer encoder is trained in an unsupervised manner (using a random masking approach as described in Devlin et al., 2018) using data from TCRdb (Chen et al., 2021). 7 million CDR3-β sequences were

used in the present examples. In a second training step, the pMHC transformer encoder is pre-trained separately to perform eluted ligand prediction (or binding affinity prediction and eluted ligand prediction) for a given peptide-MHC pair. This training was performed using a combination of the MHCFlurry 2.0 (O'Donnell et al., 2020) and netMHCpan 4.1 (Reynisson et al., 2020) datasets. Following steps 1 and 2 (note that steps 1 and 2 can be performed in any order), the weights for both encoders are frozen for the remaining of the training, and the final layer from both encoders is removed. In a third step, the full immunogenicity model is trained. Output from the encoders representing the TCR and combined peptide-MHC are concatenated before being input into a final series of transformer layers followed by a series of fully connected layers to produce a single output value. Following a sigmoid function this is treated as the predicted probability of binding and immunogenicity.

In versions of the model that only use peptide-MHC information for immunogenicity prediction (**Figure 12A**), the pMHC transformer encoder is pre-trained separately to perform one or all of eluted ligand prediction, binding affinity prediction and stability prediction as explained above (i.e. in the same way as when the pMHC encoder is pretrained for use in the full model – see **Figure 12B**). Then, the fully connected network is removed and a transformer block and classification head (2 layers fully connected network) are added and trained for immunogenicity prediction (i.e. they convert the classification token to an immunogenicity probability). In other words, the final transformer block outputs from the pMHC module are provided as the input to a new set of immunogenicity prediction layers. The pMHC encoder weights are frozen during training of the immunogenicity prediction layers.

In some of the data shown below, the inventors investigated the use of pMHC encoder models that are pretrained for stability prediction, either alone or as a fine-tuning step following binding affinity and/or eluted ligand prediction tasks. In such cases, step 2 comprises training or fine-tuning the p-MHC encoder for stability prediction.

In the results shown on Figures 13 to 15, the pMHC encoder is trained initially to perform a binding affinity prediction task, followed by an eluted ligand (EL) prediction task and finally fine-tuned on a pMHC stability task. The same fully connected network is used for all tasks as a classification/regression block (although the weights are of course updated at each training step), which is made possible by the use of normalised binding affinity and stability metrics (both of which are scores between 0 and 1, with 1 indicating higher affinity / more stable context, which is compatible with a classification output for eluted ligand prediction being a probability that the peptide is presented by the MHC, i.e. is an eluted ligand). Note that any order of these tasks may be used, for example the model may be trained initially to perform an eluted ligand prediction, then a binding affinity prediction, then a stability prediction. Alternatively, the model may be trained initially to perform an eluted ligand prediction, then a stability prediction, then a binding affinity prediction. All results shown in figures 13-15 use training initially for binding affinity prediction, then

for eluted ligand prediction, and (if used) for stability prediction. Each further training step uses the previously trained weights as a "warm start", i.e. as starting point for further training of the model. Model architecture was selected based on 5-fold cross validation performance on the EL task. pMHC encoders trained for stability prediction from scratch (i.e. no BA/EL pretraining) were also investigated. The design presented here (see Figure 12C) is an encoder only protein language model with separate transformer branches for both the epitope and HLA pseudosequences, which are then concatenated along with a classification token. Each transformer block is made up of 4 encoder layers. For pMHC training tasks the classification token is fed into a fully connected network as the input. MHC pseudosequences are based on the MHCFlurry 2.0 alignment describes in O'Donnell et al., 2020. Amino acid sequences are encoded by branch independent embedding layers with a depth of 72.

For the binding affinity training step of the pMHC module the binding affinity portion of the training set used to train the MHCFlurry 2.0 in O'Donnell et al., 2020 was used. This initially comprised of 219,596 affinity measurements. This was filtered for only human quantitative measurement data, resulting in a dataset of 99,761 measurements. Binding affinity measurements, in IC50 values, were scaled to between 1-0 using the formula $x = 1 - \log(\text{binding affinity})/\log(50, 000)$, where x is the nanomolar affinity and affinities are capped at 50,000 nM. The model shown on Figure 12C is trained for binding affinity prediction using this data (regression task, i.e. the fully connected network illustrated is a regression block).

For the eluted ligand training step, a combination of the eluted ligand data in O'Donnell et al., 2020 and Reynisson et al., 2020 was used as explained above.

Genesis was also benchmarked for TCR specificity prediction (results on **Figure 14**) using peptide:HLA:TCR triplets as input. For these experiments, the pMHC prediction module is unchanged to process the peptide:HLA paired input, but the immunogenicity classification head is fine-tuned with the addition of TCR based inputs. The results on Figure 14 use the model labelled as Genesis_BA_EL_STAB_IM, i.e. the model trained for peptide-MHC based (i.e. doublet-based) immunogenicity prediction using a peptide-MHC module trained for binding affinity, eluted ligand and stability prediction. Only the immunogenicity prediction portion (i.e. the last transformer block and the classification block, on Figure 12A) were fine tuned and the peptide-MHC encoding block weights were frozen. For these benchmarking experiments full length TCRs with both alpha and beta chains were used to compare to existing state-of-the-art models; however, other configuration such as CDR3-beta only input is also possible with Genesis. As explained above and as illustrated on Figure 12D, a set of TCR encoding modules each consisting of a 2-layer transformer encoder were pre-trained in a self-supervised manner using a random masking approach. Separate encoders were trained for encoding TCR-alpha and TCR-beta chains, respectively, because this enabled flexible use of one or both encoders in a final model (e.g. a final model can be built using only beta chain data, or using both beta chain and alpha chain data). A combined transformer

taking as input concatenated alpha and beta chain data, each chain preceded by a special token indicating chain type, could have been trained instead. TCR-beta chains were generated from datasets available from a sequencing dataset of 666 available from Adaptive Biotechnologies (Emerson et al. 2017). Full length amino acid sequences were reconstructed using the V and J allele annotations, with sequences define in IMGT/GENE-DB (Giudicelli, V., Chaume, D. & Lefranc, 2005). Entries containing ambiguous residues or non-standard amino acids were removed. 15,363,111 unique TCR-beta sequences were used to train the base model, with 5% randomly selected to use as a validation set for hyperparameter optimisation. For alpha chain inputs a specific encoder was trained by fine-tuning the beta chain encoder with the alpha chain pre-training set from the STAPLER model (Kwee et al. 2023), resulting in 46,207 unique alpha chains after processing. The alpha chain encoder was trained as a fine-tuning task from the pretrained beta chain encoder because beta chain data is typically available in large amounts than alpha chain data. This approach therefore reduces the risk of overfitting for the alpha chain encoder. It would have also been possible to train the alpha chain encoder from scratch. Both models were trained using negative log likelihood loss of reconstructing the original sequence from the masked input using a projection back to amino acid space using a fully connected layer. The final combined model is constructed by combining the outputs of the pMHC and TCR modules, separated by special separation tokens. As with the pMHC immunogenicity model, the combined outputs of the preceding encoders (and a class token) are provided to a classification head. The classification head is fine-tuned from the pMHC immunogenicity model rather than being trained from scratch. For the TCR specificity task versions of Genesis were trained with either encoder weights frozen as with the pMHC task, or the entire model was fine-tune including the encoders. The use of the classification head fine-tuned from the pMHC immunogenicity model is advantageous in that the pMHC model can be trained with different data from the full immunogenicity model (e.g. data in which only peptide-MHC but not TCR information is available, which can be from different assays than triplet data). Therefore, training the triplet model (full immunogenicity model) by transfer learning from the trained p-MHC immunogenicity model means that the training of the full model (and in particular the classification head that performs the immunogenicity prediction from the pMHC and TCR encoder outputs) can reflect a broader set of peptides than would be possible if training only using triplet data. As explained above, the pMHC and TCR encoder weights can be frozen or fine-tuned during training of the full immunogenicity model. The present inventors found that the former is particularly advantageous when the triplet data is less diverse than the data on which the peptide-MHC encoder was trained (which is often the case) as it enables the model to have better generalisability to unseen epitopes (peptides). In cases where diverse triplet data is available or a model that is particularly good for a type of epitopes for which triplet data is available, the latter (fine tuning of the classification head and peptide-MHC encoder and/or TCR encoder) can be advantageous.

Benchmarking. To compare the models, the holdout test dataset was used. The presently developed models (Genesis) were compared against two different types of models: 1. TCR models: this set of models considers the interaction of a peptide with the TCR in an attempt to provide a more specific measure of immunogenicity. Note that only a subset of these consider all parts of the triplet (i.e. peptide, MHC and TCR – see Table 1 below). 2. pMHC models: this set of models only considers the peptide and MHC portions of the input. These models provide a single prediction for each pMHC. By contrast, as the present model provides prediction for TCR-pMHC, it was used to process every triplet comprising a pMHC candidate (i.e. pMHC candidate paired with a plurality of candidate TCR sequences in the test set, and the maximum score obtained for the pMHC was taken as the immunogenicity estimate). Each model produced a score between 0 and 1 representing the likelihood of binding / immunogenicity. The models were compared by obtaining receiver operating characteristic (ROC) curves for classifying each triplet in the test set as a positive *vs.* negative triplet.

The following models were compared: 1. TCR models: pmtNet (Lu et al., 2021), imrex (Moris et al., 2020), and ERGO (Springer et al., 2021). Only these models are arguably comparable to those of the present disclosure. Table 1 below summarises the main features of these models as far as they are relevant to the comparison with the model of the present disclosure. 2. pMHC models: bigMHC (Albert et al., 2022), DeepAttentionPan (Jin et al., 2021), IEDB (tools.iedb.org/immunogenicity/, Calis et al., 2013), MHCflurry (O'Donnell et al., 2020), NetMHCpan 4.1 elution prediction (Reynisson et al., 2020), NetMHCpan binding affinity prediction (Reynisson et al., 2020), Prime 2.0 (Schmidt et al., 2021; Gfeller et al., 2023). These models are not directly comparable to the models of the present disclosure as they solve a simpler task that does not necessarily map to immunogenicity (since peptides can bind to MHC molecules without the complex being able to interact with a TCR and trigger an immune reaction).

Model	Model input	Model output	Model architecture
Genesis	Sequences of peptide, MHC, TCR CDR3beta. Amino acids encoded as tokens (0-20, one for each amino acid and a padding character). Other encoding schemes envisaged such as one-hot encoding, substitution matrices and physicochemical descriptors.	Probability of binding and immunogenicity.	4 components: (i) pretrained transformer-based encoder trained in self-supervised manner – other encoder architectures envisaged, (ii) pretrained transformer-based encoder trained as part of model for peptide-MHC binding prediction (elution likelihood) – other architectures envisaged,(iii) transformer block taking concatenated encoded data as input, (iv) immunogenicity block taking transformer block output as input, comprising fully connected layers producing a single value fed into a sigmoid function.

PMTnet (Lu et al. 2021)	Sequences of CDR3b, peptide and MHC. pMHC model input is based on netMHCpan (pseudosequences comprising selected key residues, encoded with BLOSUM). TCR model input is CDR3beta sequence encoded with physicochemical descriptors.	Single variable between 0 and 1, percentile rank of the predicted binding strength between the TCR and the pMHC, with respect to a pool of 10,000 randomly sampled TCRs (background distribution) against the same pMHC	3 components: (i) trained TCR autoencoder, (ii) pMHC model is a long-short term memory (LSTM) model trained to predict whether the peptide binds the MHC (categorical output), (iii) deep neural network combining encodings of (i) and (ii). The prediction deep neural network uses a differential learning scheme: two identical copies of the model, one is fed a true binding pair of TCR and pMHC and one is fed a negative pair with the same pMHC in each training cycle. Loss function taking both outputs into account – weights are the same in both models.
Imrex (Moris et al., 2020)	Amino acid sequences of the TCR's CDR3 region and the epitope. Sequences are converted into an interaction map: a set of 4 single channel 2D pseudoimages comprising a pixel for each pair of amino acids from the CDR3 and the peptide sequences, the value of the pixel being the absolute difference between the value of a physicochemical property for the amino acids in the pair.	Value between 0 and 1 for an epitope-TCR pair.	This interaction map is fed into a convolutional neural network (image recognition model) that produces a single output value.
ERGO (Springer et al., 2021)	CDR3beta and peptide sequence. CDR3beta amino acid sequences are one-hot encoded for the autoencoder or encoded using an encoding matrix	Categorical: TCR-peptide binding / no binding classification.	LSTM (peptide) and LSTM or autoencoder (TCR) produce encodings of the peptide and CDR3 (respectively), and these are fed into a multilayer perceptron trained to output 1 if the TCR and peptide bind and 0 otherwise.

	initialised with 10 random values for each amino acid. Peptide sequences encoded using encoding matrix.		
NetTCR (Montemurro et al., 2021)	CDR3 alpha and beta or beta only, and peptide sequence. The CDR3 and peptide sequences are encoded using the BLOSUM50 matrix.	Single number representing TCR-peptide binding probability.	The encoded sequences are passed independently through a 1D convolutional layer and a max-pooling layer. The extracted features are then concatenated and fed into a dense layer with 32 hidden units.

Table 1. Comparison of design of model of the present disclosure and prior art.

Model	**Training data positives**	**Training data negatives**
Genesis	37163 positive triplets combined from multiple datasets.	Negative TCR: Positive pMHC complexes matched with random CDR3-b from database (not just shuffle of positive data); Negative peptides: Positive MHC-CDR3 pairs are matched with random proteins from the human proteome. Non-immunogenic pMHCs: Peptide-MHC combinations which have been found to be non-reactive in immunogenicity screening experiments were paired with CDR3s from the positive triplet dataset to generate negatives which would be likely to form pMHC complexes. 200:1 negative to positive ratio, i.e. approx. 7.5 million negative triplets.
PMTnet (Lu et al. 2021)	32607 positive triplets	Random mixing: 10:1 negative to positive ratio obtained by creating negative pairs by random mismatching of TCR and pMHC of the positive triplets. Training scheme necessarily presents both a positive and a negative triplet.
Imrex (Moris et al., 2020)	19,842 unique (human only, MHC-1 only, limited length) CDR3-epitope pairs from VDJdb.	Compared two different methods for generating negative observations: 1) shuffling the known positive pairs (pairing CDR3 with peptide randomly sampled from positive set) and 2) sampling uniformly from a negative TCR CDR3beta reference repertoire set and pairing with peptides from

		positive dataset. Positive to negative ratio kept to 50:50 in all cases.
ERGO (Springer et al., 2020)	Pairs of peptides and TCRs from two datasets: over 20000 TCRb sequences matching over 300 unique epitope peptides, and over 40000 TCR sequences and 200 cognate epitopes. Number of unique pairs unknown.	Pair peptides from positive set with 5 randomly selected TCRs from positive set.
NetTCR (Montemurro et al., 2021)	9204 unique CDRβ-peptide pairs	387,598 negative data points of TCRs explicitly found not to be positive to any of 19 peptides screened.

Table 1 (continued). Comparison of design of model of the present disclosure and prior art.

Only Genesis and PMTNet consider triplets (peptide-MHC-TCR sequences). All other TCR models in Table 1 only use the peptide and TCR sequences. Many were trained (and tested in their original publications) using data from single HLA alleles or clusters, which reduces their generalisability.

The model according to the present disclosure (Genesis) was used by providing a prediction for every possible peptide-MHC-TCR combination and taking the highest score as the output for that peptide. Models using pMHCs as the input only see a given entry once without considering an aspect of the TCR specificity. In other words, there will be only one prediction from pMHC models for each peptide-MHC pair. By contrast, for peptide-MHC-TCR models, multiple predictions may be obtained for triplets including a respective TCR sequence. This may enable such models to take into account the TCR repertoire associated with a sample / patient, when making predictions of immunogenicity. In the present model, every peptide in the test dataset would appear in multiple triplets at least because of the way in which the negative triplets in the test data were generated. The pMHC based models produce a general immunogenicity prediction (i.e. one that is not TCR specific and hence cannot be patient specific), whereas Genesis can include an additional patient specific component by taking the TCR repertoire into account.

For benchmarking Genesis for pMHC immunogenicity prediction (results on **Figure 13B**) two variants were compared. One with the pMHC module trained up until the EL task and another with the full training including the stability prediction task in order to assess the utility of this additional transfer learning to immunogenicity. Genesis was compared to pMHC binding/elution models netMHCpan 4.1 (Reynisson et al. 2020) and MHCFlurry 2.0 (O'Donnell et al. 2020), the pMHC stability model netMHCstabpan (Rasmussen et al. 2016), along with the pMHC immunogenicity models PRIME 2.0 (Gfeller et al. 2023) and BigMHC (Albert et al. 2023). The holdout test set was compiled by combining the cancer specific dataset CEDAR (Kosaloglu-Yalcın, Z. et al. 2022) and the holdout MANAFEST assay dataset of 16 cancer patients from Albert et al. 2023. CEDAR was searched for t-cell assays in humans only. The full set was filtered for pMHC pairs contained in the

training set for any of the comparison models. This resulted in a Genesis training set of 9101 pMHC (2673 positives) and a test set of 3551 pMHCs (951 positives).

Genesis was also benchmarked for TCR specificity prediction (results on **Figure 14**) using peptide:HLA:TCR triplets as input. The pMHC prediction module is unchanged to process the peptide:HLA paired input, but the immunogenicity classification head is fine-tuned with the addition of TCR based inputs. The model was compared to NetTCR-2.1 (Montemurro, A., Jessen, L. E. & Nielsen, M., 2022) and STAPLER (Kwee et al. 2023). STAPLER is also a transformer-based model, reading full length TCR alpha and beta chains along with the epitope of interest as a sequence of tokens. The STAPLER model is trained with a mixture of masked language modelling and fine-tuning tasks for epitope-TCR pairs. The comparison task with Genesis is performed using the data provided from the author's GitHub repository consisting of a fine-tuning training set and a holdout test set consisting of positive triplets from VDJdb, along with a subset from their other sources and externally sampled negative TCRs. Positives (in the holdout test set) were a mixture of seen and unseen epitopes. The authors identified internal shuffling of the VDJdb TCRs to be a source of data-leakage (i.e. the authors found that there is an underlying similarity within VDJdb which explains some of the performance that they obtained when just doing random pairing), indicating this to be a more representative test of generalisable performance (i.e. the STAPLER model did not generalise well to unseen epitopes after removing the source of bias associated with VDJdb similarity between test and fine tuning data). This provided a training set of 23,410 triplets and a test set of 3,372 triplets (562 positive). NetTCR 2.1 is a sequence-based 1-D convolutional neural network that takes as inputs the amino acid sequences of the six TCR CDR loops, and the peptide sequence. The sequences are encoded using the BLOSUM50 encoding scheme then encoded sequences are processed independently by different convolutional blocks. The output of the convolutional layers are max-pooled across the sequence length dimension and concatenated, then processed with a final hidden layer and an output layer with a single neuron outputting a binding score of the input peptide and TCR. The model was trained with combined data from IEDB, VDJdb, McPAS and 10X Genomics Single Cell Immune Profiling of four donors, filtered to only keep data points with both CDR3 α- and β-chains and V/J gene annotations, remove any cross-reactive TCRs, restrict the data to TCRs with CDR3α/β lengths in a range from 6 to 20 amino acids, and only keep peptides with at least 100 positive TCRs. The performance of the NetTCR 2.1 model as described in the original publication is very specific to a set of 5 epitopes that are well studied (with multiple hundreds of TCRs each), on which the authors did all their experiments.

Finally, Genesis was benchmarked for TCR assisted immunogenicity prediction (results on **Figure 15**). Even if TCR specificity on unseen epitopes is still difficult with existing datasets and architectures, TCR data when available could be used in ranking pMHC complexes for immunogenicity using a compatible framework. In other words, given a completely unseen peptide, it is still difficult to predict the specific triplets that the peptide will be involved in (e.g. exact TCR

sequences that are likely to recognise the peptide). However, the experiments above showed that the present methods can provide good predictions of whether the peptide will be reactive. In this set of experiments, the inventors tested whether the TCR information could help with this task, and found that it does. In other words, the results on Figure 15 show that having the TCR information helps with the prediction of whether a peptide will be reactive, and can therefore be useful in ranking / prioritising neoantigens. The pMHC-TCR version of Genesis was compared to the pMHC only version to assess if TCR data could be useful in improving immunogenicity prediction when available. For this experiment only the CDR3-beta chains were used due to shortage in datasets with full paired chain information, particularly for holdout pMHCs unseen by other models. As explained above, training data was composed of positive triplets from the same sources as the pMHC only experiments, filtered for datasets with known reactive CDR3-beta sequences. This resulted in a training set of 37,970 positive triplets covering 1,253 unique epitopes. 3 distinct types of negative data were generated. First, negative wild type peptides were sampled from the CCDS and paired with TCRs taken from the positive set. Second, positive pMHCs were paired with negative TCRs from the background distribution used in the TCR encoder pre-training. Finally, negative immunogenic pMHCs from the PRIME and TESLA datasets were paired with TCRs from the positive fraction to create presented but non-immunogenic triplets. These negative types were produced at equal proportions, with the total negative to positive ration of 100:1. 5-fold cross validation was used to fit training parameters. The holdout test set was comprised of positives taken from CEDAR, filtered for t-cell assays of neoepitopes in humans and with beta chain CDR3 sequences available and negatives from the same search without the CDR3 requirement. Negative samples were paired with the positive TCR set to ensure difference between the positive and negative datasets could not be detected by a distribution shift in the TCR repertoire alone. Any epitopes present in both the training and test set were removed from the training set for both Genesis and the comparison models. To maximise the possible test data BigMHC was compared to as the best performing other model from the pMHC-only immunogenicity prediction. NetMHCpan-el was also included to compare against a presentation-only model. This produced a reduced test set of 115 positive pMHCs with at least one known reactive TCR and 1957 negative pMHCs paired with the positive fraction's TCRs. Genesis-TCR scores were aggregated per pMHC by taking the maximum predicted score.

Results

Figure 7 shows the receiver operating characteristic (ROC) curve of our model alone at classifying every triplet in the test set. The results show an area under the ROC curve (AUC) of 0.758, which is well above random and an exceptional performance considering the complexity of the problem, the small amount of triplet training data available, and the fact that the test set was specifically selected to be different from the training set (comprising no peptides that appeared in the training set).

Figure 8 shows a comparison between the performance of the present model and prior art models performing a similar task. Figure 8A shows the results when assessed using a test dataset filtered to ensure that it does not contain any triplet appearing in the training set of any other model being compared. Figure 8B shows the results when assessed using a test dataset filtered to ensure a hard epitope separation (i.e. peptides that also appear in the training set are excluded from the test set) between every model's training set and the benchmark (test) set. Note that in many prior art papers, such a hard epitope separation requirement is not applied, resulting in artificially inflated performance. Indeed, it is known that TCR binding prediction models perform poorly on out-of-distribution epitopes (see Deng et al., 2022). The models described herein (Genesis) outperform all prior art approaches in both cases. With the hard epitope split, competing approaches produce near chance levels of prediction (AUC close to 0.5). By contrast, the approach described herein combining a language model with targeted negative data generation improves on the prior art and manages to provide informative (non-random) prediction even with this extremely challenging test set.

The performance of the model described herein was also compared to models that have been trained to perform peptide-MHC (pMHC) binding predictions. This is a comparatively simpler task, at least in part because the phenomenon to be modelled is simpler and the amount of training data available for p-MHC models is significantly higher than for TCR-peptide-MHC models. However, peptide-MHC binding is only a crude approximation of immunogenicity.

The test dataset was again filtered to ensure that it does not contain any epitope that appears in the training set of any model before comparison with the model described herein. Each pMHC model provides a single immunogenicity prediction for each peptide-MHC. By contrast, the model of the present disclosure was tested using every triplet in the test set, taking the maximum binding probability as the prediction for each peptide-MHC pair. The model provides candidate CDR3s along with the immunogenicity or pMHC presentation prediction – in other words the maximum binding probability is associated with a candidate TCR CDR3 sequence (which is information that cannot be provided by the other models).

Because the prior art models are evaluated on the task for which they have been trained (whereas the model described herein was trained for the different tasks of immunogenicity prediction for a triplet), one may have expected the model described herein to perform less well than other models. **Figures 9 and 10** show that this was not the case. Indeed, the model described herein (Genesis) outperformed all the peptide-MHC models other than BigMHC, which performed very slightly better. Without wishing to be bound by theory, the present inventors believe that the performance of the BigMHC model is likely to be due at least in part due to the size and level of curation of the training data used, rather than due to the model and/or training design. Further, as explained above, all peptide-MHC models are more limited than the model of the present disclosure as they cannot provide TCR candidate binders along with an immunogenicity prediction. This information

is very valuable in the context of understanding immunogenicity in a patient specific manner. Additionally, the prediction in the context of interaction with a TCR is more physiologically relevant (i.e. it is possible that the model described herein is providing correct predictions of lack of immunogenicity in cases where peptide-MHC binding does occur – in which case the pMHC models may appear to outperform the present model but are in fact providing biologically incorrect predictions).

The results shown on Figures 7 to 9 are further summarised in Table 2 below.

Comparisons	Test data	AUC
Genesis vs. chance	827626 triplets, 4148 positives. No peptides appearing in the training data.	Genesis: 0.758
Genesis vs. CDR3-pMHC models	445980 triplets, 1829 positives. No triplets appearing in the training data of any model.	Genesis: 0.633 PMTNet: 0.565 Imrex: 0.548 ERGO: 0.590
Genesis vs. CDR3-pMHC models	439887 triplets, 789 positives. No peptides appearing in the training data of any model.	Genesis: 0.737 PMTNet: 0.502 Imrex: 0.525 ERGO: 0.512
Genesis vs. pMHC models	11267 doublets, 217 positives. No peptides appearing in the training data of any model.	Genesis: 0.779 BigMHC:0.800 DeepAttentionPan:0.727 IEDB: 0.673 MHCflurry: 0.635 NetMHCpan el: 0.534 NetMHCpan ba: 0.728 Prime: 0.627

Table 2. Summary of benchmarking results.

Figure 11 shows the negative to positive ratio used for training the model. This shows that the performance of the model initially improves with increasing negative to positive ratios, but that performance drops off at 300:1. This indicates that there is likely an optimal range of negative to positive ratio between 100:1 and 300:1, particularly around 200:1. Note that the exact values of the AUCs are not comparable between the figures provided as the datasets used for testing are different in each figure, but each figure shows a comparison between the models presented in the figure.

Peptide-MHC module performance and Peptide-MHC immunogenicity Prediction

The p-MHC sub-module of Genesis was trained to encode peptide-MHC input pairs for downstream immunogenicity prediction by training iteratively on related pMHC complex prediction tasks. Performance on these sub-tasks was assessed compared to similar models at two stages, first after the eluted ligand prediction on the MHCFlurry 2.0 benchmark dataset (O'Donnell et al.

2020) and during stability training for similarity to the NetMHCstabpan 1.0 (Rasmussen et al. 2016) results after cross validation.

Figure 13 shows results of these experiments, in particular an evaluation of the performance of stability prediction using the Genesis pMHC transformer encoder (A) and an evaluation of the performance of immunogenicity prediction using the Genesis pMHC transformer encoder trained using a stability prediction task and comparative models (B). **Figure 13A** shows a comparison of the performance of Genesis on the stability prediction task (Pearson Correlation Coefficient, PCC, between stability predicted by Genesis vs ground truth), when Genesis was trained from scratch for the stability prediction task (left bar in each pair of bars, each pair corresponding to data for a single HLA allele) or using transfer learning, fine tuning for stability prediction after training for binding affinity prediction and fine tuning for eluted ligand prediction (right bar in each pair of bars). Due to the shortage of stability data available, a holdout test set was not practical for this task. Similarly, a direct comparison with the NetMHCstabpan model was not possible since both models (Genesis and NetMHCstabpan) use all of the available data in training and NetMHCstabpan is not currently re-trainable. Nevertheless, the data on Figure 13A show a similarly high Pearson's correlation coefficient to that reported for NetMHCstabpan in Rasmussen ey al. 2016, with a best PCC of 0.88 across all available alleles. The model performed best when pretrained with the BA and EL tasks compared to being trained from scratch. A significant improvement in runtime is achieved in Genesis compared to NetMHCstab-pan. Benchmark pools of 50,000 peptide MHC pairs run through Genesis's stability module in a uniform 52s on laptop with an M1 chip and 16GB of RAM, whereas the same pools took between 2-20 minutes with NetMHCstabpan depending on peptide length and HLA diversity.

The models described above were then fine-tuned for peptide-immunogenicity prediction and extensively tested and benchmarked against comparative models for this task. The results of these experiments are shown on **Figure 13B**. Figure 13B shows the receiver-operator characteristic (ROC) and precision curves for Genesis (Genesis BA_EL_IM is the pMHC module trained for binding affinity and eluted ligand prediction, then included in a model trained for pMHC immunogenicity prediction; Genesis BA_EL_STAB_IM is a model trained using the same principle but with additional training of the pMHC module for stability prediction by fine tuning after BA and EL training) compared to its sub-modules (EL model is the pMHC module fine-tuned for eluted ligand prediction after pretraining for binding affinity prediction; Stability model is the pMHC module is the pMHC module fine-tuned for stability prediction after pre-training for eluted ligand prediction and binding affinity prediction) and other published prediction models (netmhcpan_ba and netmhcpan_el are respectively the binding affinity output and the eluted ligand output of NetMHCpan 4.1 as described in Reynisson et al. 2020; bigMHC_im is the pMHC immunogenicity model described in Albert et al. 2023; netmhcstabpan is the p-MHC stability model described in Rasmussen et al. 2016; mhcflurry is the MHCflurry 2.0 presentation prediction model as described

in O'Donnell et al. 2020; and Prime is the p-MHC presentation model PRIME 2.0 as described in Gfeller et al. 2023). Models were compared on both the area under the ROC curve (AUROC or AUC) and average precision (AP). Two versions of Genesis were compared with an EL-only and full stability trained pMHC input module. The inclusion of stability pre-training in the pMHC module resulted in the best overall performance (ROC AUC=0.628, average precision: AP=0.52). The EL pMHC module alone performed worst of all of the Genesis models, with performance similar to netMHCpan, indicating peptide elution likelihood was not a highly predictive feature on this dataset alone. Stability was an overall better individual feature, with both the Genesis stability model and netMHCstabpan outperforming all the BA/EL models. The data show that immunogenicity specific training as described herein enables improved discrimination of immunogenic from non-immunogenic pMHC pairs.

TCR Specificity Prediction

Prediction of TCR specificity from triplets with unseen epitopes has been shown to be limited in recent studies and reviews (see e.g. O'Brien et al. 2023), with results showing above chance performance explainable by data leakage from their training sets or frequency shifts (see e.g. Kwee et al. 2023). For example, Kwee et al. 2023 showed that above chance results are associated with in-dataset similarities, due to low epitope diversity (as similar TCRs are likely to bind to the same epitope). Indeed, publicly available triplet datasets typically contain large amounts of information about a very limited sets of epitopes (e.g. one CMV epitope making up 15k of the ~45k triplets in public datasets available to date – making it possible for models to easily spot and memorise similar TCRs, but poorly able to generalise beyond those few epitopes). In other words, current models are not able to give an exact prediction of a TCR for a known reactive peptide-MHC. This limits their application to use cases such as designing new TCR-based therapies such as genetically engineered T cells. This may be due at least partially to biases and lack of diversity in available triplet-based immunogenicity data. Nevertheless, the present inventors postulated that, with the beneficial features of the methods described herein, TCR information when available could still provide valuable information in the context of evaluating antigens themselves (i.e. evaluating peptide-MHC or peptde immunogenicity). To demonstrate the utility of Genesis as an architecture for immunogenicity prediction utilising the full peptide-MHC-TCR triplet we initially compare it to two recent TCR specificity models using data splits provided by the authors to provide a fair comparison. The results of these experiments are shown on **Figure 14**, and demonstrate Genesis as a model design capable of using TCR information in its predictions. In particular, the inventors showed that the Genesis models as described herein can make use of TCR information if available to help rank potential peptide-MHCs for immunogenicity. This is particularly useful in the context of developing therapies (such as immunotherapies) that are antigen based, such as vaccines, antigen-reactive T cell-based therapies etc.

Initially Genesis was compared to NetTCR 2.1 (Montemurro, A., Jessen, L. E. & Nielsen, M., 2022), a 1D convolutional neural network-based approach to TCR specificity. **Figure 14A** shows the performance (in terms of Precision-Recall curves, on the left, with average precision "AP" for each model; and in terms of ROC curves, with area under the ROC curve provided as "AUC" for each model) compared to Genesis when trained on the same datasets on the holdout 6th fold as described in Montemurro, A., Jessen, L. E. & Nielsen, M., 2022. NetTCR was retrained using the codebase provided by the authors. Two versions of Genesis were compared, one with frozen encoder weights and one with fine-tuning allowed on the whole model (labelled on Figure 14A as Genesis-TCR and Genesis-TCR-Fine tune respectively). Both versions of Genesis demonstrated good performance at this TCR specificity task, and both showed improved performance over NetTCR 2.1. The fine-tuned version of the model performed best, indicating increased fine-tuning of the encoder sections is beneficial for the TCR specificity prediction task.

Next, Genesis was compared to STAPLER (Kwee et al. 2023). **Figure 14B** show the results of these experiments. The figures show the performance of Genesis compared to the STAPLER model using the train-test split described in Kwee et al. 2023, with their "VDJdb+ with external negatives" used as the test set. **Figure 14B** shows that Genesis performs better at the cross-validation task with an improved average precision. The main difference between these two models is the inclusion of HLA inputs in Genesis, and the training approach as described above. STAPLER uses a model architecture that uses a variant of the BERT design. A masking experiment was conducted, removing the HLA inputs of Genesis and retraining it to identify the influence of this difference to performance. These HLA masking experiments show that the difference in performance between STAPLER and a truncated version of Genesis (with HLA masked) was greatly reduced by removing the HLA element of Genesis and retraining. Thus, the data shows that Genesis performs better than comparative models at the TCR specificity prediction task at least in part because the HLA information encoded in the pMHC module contributes to the performance of the prediction. This is particularly the case for unseen epitopes (i.e. epitopes that were not part of the training data used for immunogenicity training). Indeed, although the MHC component explains a small proportion of the variance in seen epitopes, it explains a lot of the variance for epitopes where no information was available in training about whether they bind to *any* TCR. This again underlines the superior power of the Genesis approach for generalisability beyond the very low diversity triplet data currently available. Generalisability to unseen epitopes is crucially important in many contexts including in the context of personalised immunotherapy, where candidate neoantigens to be evaluated with immunogenicity models are patient-specific.

TCR Assisted Immunogenicity Prediction

The previous experiments demonstrated the use of Genesis as an architecture for TCR specificity prediction tasks, showing its ability to learn information efficiently from TCR-based inputs. An additional experiment was conducted to investigate whether having candidate TCR information

would assist in predicting the immunogenicity of a pMHC complex. A reduced test set from CEDAR where paired CDR3-beta chain information was available was used as input for Genesis, with negatives constructed using pMHCs marked as negative on TCR assays in CEDAR. The results of these experiments are shown on **Figure 15,** demonstrating an improved performance of Genesis compared to comparative pMHC models NetMHCpan (eluted ligand prediction) and BigMHC (immunogenicity prediction), which improvement was even greater by including the TCR information in the model input.

Discussion

The present examples describe and demonstrate the performance of a model for predicting immunogenicity of candidate triplets comprising a peptide/epitope, a MHC molecule, and a TCR molecule (represented by a TRC CDR3 β sequence), based solely on the sequence of the members of the triplet.

Models with CDR3 inputs have been found to generalise poorly to peptides outside their training distribution (Moris et al., 2020). This problem has also been observed in immunogenicity models using only pMHC complex inputs, where models can generalise to similar peptides but poorly to data too far outside their training distribution, for example an unseen pathogen (Buckley et al., 2022). The present inventors hypothesised that this was at least in part due to the way in which negative training data points are generated. Most prior art models which attempt to use CDR3 sequence inputs to predict peptide binding or immunogenicity use random mixing to generate negative data (see e.g. Lu et al., 2021; Montemurro et al., 2021). This is achieved by randomly mixing positive peptide-MHC (pMHC) pairs with other CDR3 sequences from the positive set to generate a likely negative. A shortcoming with this approach is that it restricts the CDR3 diversity in the dataset to TCRs which have been observed in a positive sample. Additionally, CDR3s are not specific to a single pMHC complex. Mixing only positive pMHCs with CDR3s from the positive dataset results in a negative dataset with low diversity and a high rate of false negative labelling (triplets labelled as “negative” but that may in fact be immunogenic).

Therefore, in order to improve generalisability (which must underline any real clinical use), the inventors generated three types of negative data points: negative TCR, negative peptide, and non-immunogenic pMHCs. In addition to a diverse set of negative sources, a high negative to positive ratio was used (200:1). This was done to both increase the diversity in the dataset and to provide a more realistic classification situation. Indeed, in a clinical application where peptide ranking is required there may be a large number of candidate CDR3s. Alternatively a single CDR3 of interest may be known but needs to be matched to a large number of potential peptides.

The results obtained (see e.g. **Figures 7-11**) show that the models described herein were able to improve on the prior art by providing more reliable predictions of immunogenicity for peptide-MHC-TCR triplets as well as for peptide-MHC pairs, in a realistic scenario where the testing data used to evaluate the models is not biased to the data that was already used to train the model. The

results shown on **Figures 13-15** further show that best in class performance for immunogenicity prediction can be obtained using triplet based approaches as described herein (i.e. providing peptide, MHC and TCR sequence information as inputs to the prediction, and training using 3 types of negative triplets as described herein), although the approach still shows some benefits for immunogenicity prediction from doublets (i.e. providing peptide and MHC sequence information as inputs to the prediction, and training using 2 types of negative doublets as described herein). The data further shows that the use of a peptide-MHC encoder that has been pre-trained for stability prediction further enhances the performance of both doublets and triplets-based models as described herein.

There are many potential practical uses of the methods described herein, where the benefits of these methods are likely to prove particularly advantageous. For example, the models described herein can be used for ranking potential neoantigen peptides identified in a patient with cancer. In particular, when TCR sequencing data is available for the patient, a plurality of candidate neoantigen peptides could be used as input to the models described herein together with every TCR identified in the patient, and the model predictions (e.g. maximum prediction for a peptide across TCRs observed in the patient) can be used to rank neoantigens or peptides associated with the neoantigens. A similar analysis can be run with a generic distribution of representative TCRs (e.g. from a TCR sequence database), for example when no patient specific TCR sequence data is available. As another example, the models described herein can be used for retrospective analysis of peptide reactivity assays. This can be used to identify which peptides are likely to be reactive in a pool of peptides that triggered an immune reaction in a reactivity assay. This can be based on a generic or sample specific set of TCR sequences.

As another example, the models described herein can be used as pre-filters to conduct immunogenicity screening experiments. For example, in Holm et al. (2022), the authors screened blood samples from patients treated with an immune checkpoint inhibitor, using patient-specific neopeptide-MHC multimer libraries, to identify characteristics of neoantigen reactive T cells that are indicative of treatment response. The libraries were obtained by identifying neoantigens from whole exome and RNA sequence data, then selecting potential neoantigens based on predicted MHC class I binding affinity and expression. Such screens would benefit from improved predictions of likely immunogenic peptides which reflect more than just MHC-I binding. Indeed, this would enable the generation of more specific screening libraries, which may in turn enable an improved characterisation of samples for immunogenicity by avoiding exposing T cells to many peptide-MHC that they would not in fact be expected to react to, and which could dilute the peptide-MHC that they would otherwise react to and/or increase the risk of cells developing a non-reactive phenotype. In other words, in any experiment where immunogenicity of candidate neoantigens and/or reactivity of T cell samples is assessed, the use of more specific sets of candidate

neoantigens (i.e. those that are more likely to be in fact immunogenic in the presence of the “right” T cells) is expected to increase the likelihood that true reactivities would be identified.

References

Landau DA, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013 Feb 14;152(4):714-26. doi: 10.1016/j.cell.2013.01.019.

Lu YC, et al. An Efficient Single-Cell RNA-Seq Approach to Identify Neoantigen-Specific T Cell Receptors. Mol Ther. 2018 Feb 7;26(2):379-389.

Leko V, et al. Identification of Neoantigen-Reactive Tumor-Infiltrating Lymphocytes in Primary Bladder Cancer. J Immunol. 2019 Jun 15;202(12):3458-3467.

Hundal J, et al. Accounting for proximal variants improves neoantigen prediction. Nat Genet. 2019 Jan;51(1):175-179. doi: 10.1038/s41588-018-0283-9.

Roth A, et al. PyClone: statistical inference of clonal population structure in cancer. Nat Methods. 2014 Apr;11(4):396-8. doi: 10.1038/nmeth.2883.

McGranahan N, et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. Science. 2016 Mar 25;351(6280):1463-9. doi: 10.1126/science.aaf1490.

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers form language understanding." arXiv preprint arXiv:1810.04805 (2018).

Eddy, S. R. (2004). Where did the BLOSUM62 alignment score matrix come from? Nature biotechnology, 22(8), 1035-1036.

Mei H, Liao ZH, Zhou Y, Li SZ. A new set of amino acid descriptors and its application in peptide QSARs. Biopolymers. 2005;80:775–86.

Schmidt J, et al. Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. Cell Rep Med. 2021 Feb 6;2(2):100194.

Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. 2020 Jul 2;48(W1):W449-W454.

Timothy J. O’Donnell, Alex Rubinsteyn, Uri Laserson, MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing, Cell Systems, Volume 11, Issue 1, 2020, Pages 42-48.e7

Jin J, Liu Z, Nasiri A, Cui Y, Louis SY, Zhang A, Zhao Y, Hu J. Deep learning pan-specific model for interpretable MHC-I peptide binding prediction with improved attention mechanism. Proteins. 2021 Jul;89(7):866-883.

Benjamin Alexander Albert, et al. Deep Neural Networks Predict MHC-I Epitope Presentation and Transfer Learn Neoepitope Immunogenicity. bioRxiv 2022.08.29.505690

Albert, B. A. et al. Deep neural networks predict class I major histocompatibility complex epitope presentation and transfer learn neoepitope immunogenicity. Nature Machine Intelligence 1–12 (2023).

Paul R Buckley, et al. Evaluating performance of existing computational models in predicting CD8+ T cell pathogenic epitopes and cancer neoantigens. Briefing in Bioinformatics, page bbac141, April 2022.

Si-Yi Chen, Tao Yue, Qian Lei, and An-Yuan Guo. TCRdb: a comprehensive database for T-cell receptor sequences with powerful search function. Nucleic Acids Research, 49(D1):D468–D474, January 2021.

David Gfeller, et al. Predictions of immunogenicity reveal potent SARS-CoV-2 CD8+ T-cell epitopes, May 2022. Pages: 2022.05.23.492800 Section: New Results.

David Gfeller, Julien Schmidt, Giancarlo Croce, Philippe Guillaume, Sara Bobisse, Raphael Genolet, Lise Queiroz, Julien Cesbron, Julien Racle, Alexandre Harari, Improved predictions of antigen presentation and TCR recognition with MixMHCpred2.2 and PRIME2.0 reveal potent SARS-CoV-2 CD8+ T-cell epitopes, Cell Systems, Volume 14, Issue 1, 2023, Pages 72-83.e5.

Jeppe Sejerø Holm, et al. Neoantigen-specifi CD8 T cell responses in the peripheral blood following PD-L1 blockade might predict therapy outcome in metastatic urothelial carcinoma. Nature Communications, 13(1):1935, April 2022.

Tianshi Lu, et al. Deep learning-based prediction of the T cell receptor–antigen binding specificiy. Nature Machine Intelligence, 3(10):864–875, October 2021.

Alessandro Montemurro, et al. NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCR alpha and beta sequence data. Communications Biology, 4(1):1–13, September 2021. Number: 1

Pieter Moris, et al. Current challenges for epitope-agnostic TCR interaction prediction and a new perspective derived from image classification September 2020. Pages: 2019.12.18.880146. Section: New Results.

Ashish Vaswani, et al. Attention Is All You Need. arXiv:1706.03762 [cs], December 2017. arXiv: 1706.03762.

Wells et al. Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. Cell, 183(3):818–834.e13, October 2020.

Goncharov, M., Bagaev, D., Shcherbinin, D. *et al.* VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2. *Nat Methods* **19**, 1017–1019 (2022).

Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. Bioinformatics. 2017 Sep 15;33(18):2924-2929.

Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. 2019 Jan 8;47(D1):D339-D343.

Lu, T., Zhang, Z., Zhu, J. *et al.* Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nat Mach Intell* **3**, 864–875 (2021).William R. Atchley, Jieping Zhao, Andrew D. Fernandes, and Tanja Drüke. Solving the protein sequence metric problem. PNAS 102 (18) 6395-6400. April 25, 2005.

Calis JJA, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, Sette A, Kesmir C, Peters B. 2013. Properties of MHC class I presented peptides that enhance immunogenicity. PloS Comp. Biol. 8(1):361.

Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun. 2022 Jul 27;13(1):4348.

Lihua Deng, et al. Performance comparison of TCR-pMHC prediction tools reveals a strong data dependency. Bioarxiv v1 posted 24 November 2022. bioRxiv 2022.11.24.517666

ElAbd, H., Bromberg, Y., Hoarfrost, A. *et al.* Amino acid encoding for deep learning applications. *BMC Bioinformatics* **21**, 235 (2020)

Shukla SA, Rooney MS, Rajasagi M, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat Biotechnol.* 2015;33(11):1152-1158

Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics. 2014 Dec 1;30(23):3310-6

Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J Immunol. 2017 Nov 1;199(9):3360-3368

Stryhn, A. et al. A Systematic, Unbiased Mapping of CD8+ and CD4+ T Cell Epitopes in Yellow Fever Vaccinees. Frontiers in Immunology 11 (2020).

Rasmussen, M. et al. Pan-Specific Prediction of Peptide-MHC Class I Complex Stability, a Correlate of T Cell Immunogenicity. Journal of Immunology (Baltimore, Md.: 1950) 197, 1517–1524 (2016).

Kosaloglu-Yalcın, Z. et al. The Cancer Epitope Database and Analysis Resource (CEDAR). Nucleic Acids Research gkac902 (2022).

Emerson, R. O. et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effect on the T cell repertoire. Nature Genetics 49, 659–665 (2017).

Giudicelli, V., Chaume, D. & Lefranc, M.-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Research 33, D256–261 (2005).

Kwee, B. P. Y. et al. STAPLER: Efficient learning of TCR-peptide specificity prediction from full-length TCR-peptide data (2023). www.biorxiv.org/content/10.1101/2023.04.25.538237v1.

Montemurro, A., Jessen, L. E. & Nielsen, M. NetTCR-2.1: Lessons and guidance on how to develop models for TCR specificity predictions. Frontiers in Immunology 13, 1055151 (2022).

Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E. & Friedman, N. McPAS- TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. Bioinformatics 33, 2924–2929 (2017).

Pruitt, K. D. et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Research 19, 1316–1323 (2009).

O'Brien, H. et al. Breaking the performance ceiling for neoantigen immunogenicity prediction. Nature Cancer 4, 1618–1621 (2023).

Lie-Andersen O, et al. Impact of peptide:HLA complex stability for the identification of SARS-CoV-2-specific CD8+T cells. Front Immunol. 2023 May 18;14:1151659.

Gurung, H.R., Heidersbach, A.J., Darwish, M. et al. Systematic discovery of neoepitope–HLA pairs for neoantigens shared among patients and tumor types. Nat Biotechnol (2023).

Dylan T. Blaha, et al; High-Throughput Stability Screening of Neoantigen/HLA Complexes Improves Immunogenicity Predictions. Cancer Immunol Res 1 January 2019; 7 (1): 50–61.

Kaseke C, et al. HLA class-I-peptide stability mediates CD8+ T cell immunodominance hierarchies and facilitates HLA-associated immune control of HIV. Cell Rep. 2021 Jul 13;36(2):109378.

All references cited herein are incorporated herein by reference in their entirety and for all purposes to the same extent as if each individual publication or patent or patent application was specifically and individually indicated to be incorporated by reference in its entirety.

The specific embodiments described herein are offered by way of example, not by way of limitation. Various modifications and variations of the described compositions, methods, and uses of the technology will be apparent to those skilled in the art without departing from the scope and spirit of the technology as described. Any sub-titles herein are included for convenience only and are not to be construed as limiting the disclosure in any way.

The methods of any embodiments described herein may be provided as computer programs or as computer program products or computer readable media carrying a computer program which is arranged, when run on a computer, to perform the method(s) described above.

Unless context dictates otherwise, the descriptions and definitions of the features set out above are not limited to any particular aspect or embodiment of the invention and apply equally to all aspects and embodiments which are described. Throughout the specification and claims, the following terms take the meanings explicitly associated herein, unless the context clearly dictates otherwise. The phrase “in one embodiment” as used herein does not necessarily refer to the same embodiment, though it may. Furthermore, the phrase “in another embodiment” as used herein does not necessarily refer to a different embodiment, although it may. Thus, as described below, various embodiments of the invention may be readily combined, without departing from the scope or spirit of the invention. It must be noted that, as used in the specification and the appended claims, the singular forms “a,” “an,” and “the” include plural referents unless the context clearly dictates otherwise. Ranges may be expressed herein as from “about” one particular value, and/or to “about” another particular value. When such a range is expressed, another embodiment includes from the one particular value and/or to the other particular value. Similarly, when values are expressed as approximations, by the use of the antecedent “about,” it will be understood that the particular value forms another embodiment. The term “about” in relation to a numerical value is optional and means for example +/- 10%. Throughout this specification, including the claims which follow, unless the context requires otherwise, the word “comprise” and “include”, and variations such as “comprises”, “comprising”, and “including” will be understood to imply the inclusion of a stated integer or step or group of integers or steps but not the exclusion of any other integer or step or group of integers or steps. Other aspects and embodiments of the invention provide the aspects and embodiments described above with the term “comprising” replaced by the term “consisting of” or “consisting essentially of”, unless the context dictates otherwise. “and/or” where used herein is to be taken as specific disclosure of each of the two specified features or

components with or without the other. For example “A and/or B” is to be taken as specific disclosure of each of (i) A, (ii) B and (iii) A and B, just as if each is set out individually herein.

The features disclosed in the foregoing description, or in the following claims, or in the accompanying drawings, expressed in their specific forms or in terms of a means for performing the disclosed function, or a method or process for obtaining the disclosed results, as appropriate, may, separately, or in any combination of such features, be utilised for realising the invention in diverse forms thereof.

CLAIMS

1. A computer-implemented method of predicting whether an antigen is likely to be immunogenic, the method comprising:

obtaining a triplet of sequences comprising: an amino acid sequence of a peptide encoding the antigen, an amino acid sequence of a candidate MHC molecule or a part thereof, and an amino acid sequence of a candidate T cell receptor (TCR) beta chain and/or alpha chain or a part thereof; and

providing the triplet of sequences or information derived therefrom as inputs to a machine learning model trained to predict a score representing the probability that the antigen is immunogenic in the context of the candidate MHC molecule and the candidate TCR,

wherein the machine learning model has been trained using training data comprising amino acid sequences or information derived therefrom for negative peptide-MHC-TCR triplets comprising:

a. a first set of one or more peptide-MHC-TCR triplets each comprising: (i) a TCR-MHC pair comprising an MHC molecule and a TCR chain or chains known to bind the MHC molecule (positive TCR-MHC pair), and (ii) a peptide not known to interact with the TCR-MHC pair;

b. a second set of one or more peptide-MHC-TCR triplets each comprising: (i) a peptide-MHC pair comprising an MHC molecule and a peptide known to bind the MHC molecule (positive peptide-MHC pair), and (ii) a TCR chain or chains not known to interact with the peptide-MHC pair, wherein the peptide-MHC pair has been previously found to interact with a TCR (immunogenic positive peptide-MHC pair); and

c. a third set of one or more peptide-MHC-TCR triplets each comprising: (i) a peptide-MHC pair comprising an MHC molecule and a peptide known to bind the MHC molecule (positive peptide-MHC pair), and a TCR chain or chains not known to interact with the peptide-MHC pair, wherein the peptide-MHC pair has been previously found to not be immunogenic (non-immunogenic positive peptide-MHC pair).

2. The method of claim 1, wherein the first, second and/or third sets of negative peptide-MHC-TCR triplets have been derived from amino acid sequences or information derived therefrom for a plurality of positive peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response, and/or

wherein the machine learning model has been trained using training data further comprising amino acid sequences or information derived therefrom for a plurality of positive peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response.

3. The method of claim 2, wherein the TCR chain or chains in the second set have been selected from a database or reference dataset, and/or wherein the TCR chain or chains in the second set

have not been selected from the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets), and/or wherein the TCR chain or chains and the peptide-MHC pairs in the second set do not form a triplet that is present in the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets), and/or wherein the peptide-MHC pairs in the second set have been selected from the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets).

4. The method of claim 2 or claim 3, wherein the peptides in the first set have been selected from a database or reference dataset, optionally wherein the peptides in the first set have been randomly selected from a reference proteome, and/or wherein the peptides in the first set have not been selected from the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets), and/or wherein the peptide and the TCR-MHC pairs in the first set do not form a triplet that is present in the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets), and/or wherein the TCR-MHC pairs in the first set have been selected from the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets).

5. The method of any of claims 2 to 4, wherein the TCR chain or chains in the third set have been selected from the plurality of peptide-MHC-TCR triplets each comprising a peptide, an MHC molecule and a TCR chain or chains that are known to interact with each other to induce an immune response (positive peptide-MHC-TCR triplets).

6. The method of any of claims 2 to 5, wherein the training data comprise a ratio of negative triplets to positive triplets of at least 100:1, at least 150:1, between 100:1 and 300:1, preferably between 150:1 and 250:1, or around 200:1.

7. The method of any preceding claim, wherein the machine learning model takes as input an amino acid sequence comprising a part of the variable region of one or more chains of a TCR, or information derived therefrom, and/or wherein the machine learning model takes as input an amino acid sequence comprising one or more CDRs of one or more chains of a TCR, or information

derived therefrom, and/or wherein the machine learning model takes as input an amino acid sequence comprising the CDR3 sequence of one or more chains of a TCR, or information derived therefrom, optionally wherein the machine learning model takes as input an amino acid sequence of the CDR3 region of the alpha and/or beta chain of a TCR, or information derived therefrom, preferably wherein the model takes as input an amino acid sequence comprising or consisting of the sequence of the CDR3 region of the beta chain of a TCR.

8. The method of any preceding claims, wherein the machine learning model takes as input the triplet of amino acid sequences and produces an encoding for each sequence, or wherein the machine learning model takes as input an encoding for each sequence of a triplet of amino acid sequences, optionally wherein the amino acid sequences are encoded using encoding schemes selected from: a predetermined token for each amino acid and optionally a padding character, one-hot-encoding, an encoding using a substitution matrix, an encoding using an embedding matrix, and an encoding using physicochemical descriptors.

9. The method of claim 8, wherein one or more of the amino acid sequences are encoded as fixed length strings with a token for each amino acid and a padding character, optionally wherein the TCR sequence is encoded as a fixed length string, and/or wherein the peptide sequence is encoded as a fixed length string, and/or wherein the MHC sequence is encoded as a pseudosequence with fixed length.

10. The method of any preceding claim, wherein the machine learning model is a deep learning model, and/or wherein the machine learning model comprises one or more natural language processing models.

11. The method of any preceding claim, wherein the machine learning model comprises a first encoder or pair of encoders for encoding the TCR sequence, and a second encoder for encoding the peptide and MHC sequences.

12. The method of claim 11, wherein the encoders have been pretrained prior to training the machine learning model using the training data comprising the negative triplets, optionally wherein the machine learning model has been trained using the training data comprising the negative triplets with the parameters of the encoders maintained to their pretrained values or wherein the training of the machine learning model using the training data comprising the negative triplets included fine-tuning the parameters of one or more of the encoders.

13. The method of claim 11 or claim 12, wherein the first encoder or pair of encoders comprises a single encoder taking as input a TCR beta chain or a part thereof, preferably a part comprising or consisting of the CDR3 region, or wherein the first encoder or pair of encoders comprises a single

encoder taking as input the concatenation of a TCR beta chain or a part thereof and a TCR alpha chain or a part thereof, or wherein the first encoder or pair of encoders comprises a pair of encoders taking as input respectively a TCR beta chain or a part thereof and a TCR alpha chain or a part thereof,
preferably wherein a part of TCR chain comprises or consists of the CDR3 region of the respective chain.

14. The method of any of claims 11 to 13, wherein the first encoder or pairs of encoders have been trained in a self-supervised manner to encode TCR sequences or parts thereof, optionally wherein the first encoder or pairs of encoders have been trained in a self-supervised manner using random masking.

15. The method of any of claims 11 to 14, wherein the second encoder takes as input a peptide sequence and an MHC sequence, and wherein the second encoder has been trained as part of a model:
(i)trained to predict whether the peptide is likely to bind the MHC molecule, whether the peptide is likely to be presented by the MHC molecule, and/or whether the peptide and MHC molecule are likely to form a stable complex, and/or
(ii) trained to predict the binding affinity between a peptide sequence and a MHC molecule corresponding to the MHC sequence, trained to classify pairs comprising a peptide sequence and an MHC sequence between a first class comprising peptide-MHC pairs known to bind to each other and be presented on the surface of cells, and a second class comprising peptides-MHC pairs that are not expected to bind to each other and be presented on the surface of cells, and/or trained to predict a metric indicative of the stability of a complex comprising the peptide and MHC molecule corresponding to the MHC sequence, optionally wherein the metric indicative of the stability of a complex comprising the peptide and MHC molecule corresponding to the MHC sequence is a normalised metric with a value between 0 and 1 and/or wherein the binding affinity is a normalised binding affinity metric with a value between 0 and 1.

16. The method of any of claims 11 to 15, wherein the second encoder has been trained as part of a model that has been:
(i) pretrained to predict whether the peptide is likely to bind the MHC molecule, and/or whether the peptide is likely to be presented by the MHC molecule; and
(ii) pre-trained, optionally after step (i), for predicting whether the peptide and MHC molecule are likely to form a stable complex.

17. The method of claim 16, wherein at step (i) the second encoder has been trained as part of a model that has been pretrained to predict whether the peptide is likely to bind the MHC molecule,

then further trained using transfer learning to predict whether the peptide is likely to be presented by the MHC molecule.

18. The method of any of claims 15 to 17, wherein a model trained to predict whether the peptide and MHC molecule are likely to form a stable complex is a model configured to take as input a peptide and MHC sequence or information derived therefrom and produce as output a metric indicative of the stability of a complex comprising the peptide and MHC molecule corresponding to the MHC sequence, optionally wherein the metric is a half-life or scaled half-life.

19. The method of any of claims 11 to 18, wherein the encoders are each independently selected from: transformer-based encoders, autoencoders, and recurrent neural network encoders such as long-short-term memory (LSTM) networks, and/or wherein the encoders are transformer-based encoders.

20. The method of any of claims 11 to 19, wherein the machine learning model further comprises a deep learning block that takes as input the concatenated outputs of the first and second encoders, and produces as output the probability that the antigen is immunogenic in the context of the candidate MHC molecule and the candidate TCR, optionally wherein the deep learning block comprises a first block that learns from the combined outputs of the first and second encoders, and a second block comprising one or more fully connected layers producing a single numerical output and optionally a sigmoid activation function, optionally wherein the first block comprises a deep artificial neural network model and/or a natural language processing model.

21. The method of any preceding claim, further comprising:

(i) repeating one or more times the steps of:

obtaining a triplet of sequences comprising: an amino acid sequence of a peptide encoding the antigen, an amino acid sequence of a candidate MHC molecule or a part thereof, and an amino acid sequence of a candidate T cell receptor (TCR) beta chain and/or alpha chain or a part thereof, and

providing the triplet of sequences or information derived therefrom as inputs to the machine learning model trained to predict a score representing the probability that the antigen is immunogenic in the context of the candidate MHC molecule and the candidate TCR,

wherein each triplet of sequences differs in the amino acid sequence of the candidate MHC molecule or part thereof, and/or in the amino acid sequence of the candidate T cell receptor (TCR) beta chain and/or alpha chain or part thereof,

thereby obtaining a plurality of respective probabilities that the antigen is immunogenic; and

(ii) selecting the highest of the plurality of probabilities as the probability that the antigen is immunogenic.

22. The method of any preceding claim, wherein the machine learning model has been trained by fine tuning a peptide-MHC immunogenicity prediction model, wherein a peptide-MHC immunogenicity model is a machine learning model that has been trained to take as input a doublet of sequences comprising an amino acid sequence of a peptide encoding the antigen, and an amino acid sequence of a candidate MHC molecule or a part thereof, or information derived from the doublet of sequences, and provide as output a score representing the probability that the antigen is immunogenic in the context of the candidate MHC molecule.

23. The method of claim 22, wherein the peptide-MHC immunogenicity model has been trained using training data comprising amino acid sequences or information derived therefrom for (i) positive peptide-MHC doublets comprising a peptide and MHC sequences that have been experimentally demonstrated to form an immunogenic complex; and (ii) negative peptide-MHC doublets comprising:

a. a first set of one or more peptide-MHC doublets each comprising: a MHC molecule selected from the positive peptide-MHC doublets and a peptide sequence not known to interact with the selected MHC molecule, optionally a randomly sampled peptide sequence, and;

b. a second set of one or more peptide-MHC doublets each comprising: a peptide-MHC pair comprising an MHC molecule and a peptide known to bind the MHC molecule (positive peptide-MHC pair), wherein the peptide-MHC pair has been previously found to not be immunogenic (non-immunogenic positive peptide-MHC pair).

23. The method of any preceding claim, wherein:

obtaining the triplet of sequences comprises obtaining an amino acid sequence of a peptide encoding the antigen, an amino acid sequence of a candidate MHC molecule or a part thereof, and an empty TCR sequence vector as candidate TCR; and

the score predicted by the machine learning model represents the probability that the antigen is immunogenic in the context of the candidate MHC molecule and an unknown TCR.

24. The method of any preceding claim, further comprising identifying the antigen from a sample, and/or performing the method of any of claims 1 to 23 using one or more candidate MHC molecules and/or one or more candidate TCR molecules identified from a sample wherein the sample is optionally a sample from which the antigen has been identified or a related sample.

25. A computer-implemented method of providing a tool for predicting whether an antigen is likely to be immunogenic, the method comprising:

(i) obtaining a training dataset comprising amino acid sequences or information derived therefrom for a plurality of peptide-MHC-TCR triplets, each triplet comprising an amino acid sequence of a

peptide encoding the antigen, an amino acid sequence of a candidate MHC molecule or a part thereof, and an amino acid sequence of a candidate T cell receptor (TCR) beta chain and/or alpha chain or a part thereof, wherein the plurality of peptide-MHC-TCR triplets comprise:

a. a first set of one or more peptide-MHC-TCR triplets each comprising: (i) a TCR-MHC pair comprising an MHC molecule and a TCR chain or chains known to bind the MHC molecule (positive TCR-MHC pair), and (ii) a peptide not known to interact with the TCR-MHC pair,

b. a second set of one or more peptide-MHC-TCR triplets each comprising: (i) a peptide-MHC pair comprising an MHC molecule and a peptide known to bind the MHC molecule (positive peptide-MHC pair), and (ii) a TCR chain or chains not known to interact with the peptide-MHC pair, wherein the peptide-MHC pair has been previously found to interact with a TCR (immunogenic positive peptide-MHC pair), and

c. a third set of one or more peptide-MHC-TCR triplets each comprising: (i) a peptide-MHC pair comprising an MHC molecule and a peptide known to bind the MHC molecule (positive peptide-MHC pair), and a TCR chain or chains not known to interact with the peptide-MHC pair, wherein the peptide-MHC pair has been previously found to not be immunogenic (non-immunogenic positive peptide-MHC pair); and

(ii) training, using said training data, a machine learning model that predicts the probability that an antigen is immunogenic in the context of a candidate MHC molecule and a candidate TCR provided as a triplet of sequences or information derived therefrom as input to the machine learning model.

26. A method of identifying one or more tumour-specific peptides that are likely to be immunogenic, the method comprising:

obtaining the amino acid sequence of one or more candidate tumour-specific peptides derives from one or more tumour-specific mutations previously identified in a tumour; and

determining whether the one or more candidate peptides are likely to be immunogenic using the method of any of claims 1 to 24,

optionally wherein the method further comprises selecting one or more of the tumour-specific peptides as peptides likely to be immunogenic using one or more criteria applying to the result of the determining.

27. A method of characterising an immunogenic composition comprising a plurality of candidate peptides or sequences encoding a plurality of candidate peptides, the method comprising:

determining whether the one or more candidate peptides are likely to be immunogenic using the method of any of claims 1 to 24, and

identifying which one or more of the candidate peptides are likely to be immunogenic by applying one or more predetermined criteria to the results of the determining.

28. A method of designing an immunotherapy for a subject that has been diagnosed as having cancer, the method comprising:

obtaining a set of one or more candidate neoantigens for the subject, wherein the one or more candidate neoantigens were identified using a process comprising analysing one or more samples from the subject comprising tumour genetic material; and

designing an immunotherapy that targets one or more of the neoantigens identified, wherein the designing comprises identifying at least one peptide encoding at least one of the candidate neoantigens that is immunogenic using the method of any of claims 1 to 24.

29. A system comprising:

a processor; and

a computer readable medium comprising instructions that, when executed by the processor, cause the processor to perform the steps of the method of any of claims 1 to 28.

30. One or more computer readable media comprising instructions that, when executed by one or more processors, cause the one or more processors to perform the steps of the method of any of claims 1 to 28.

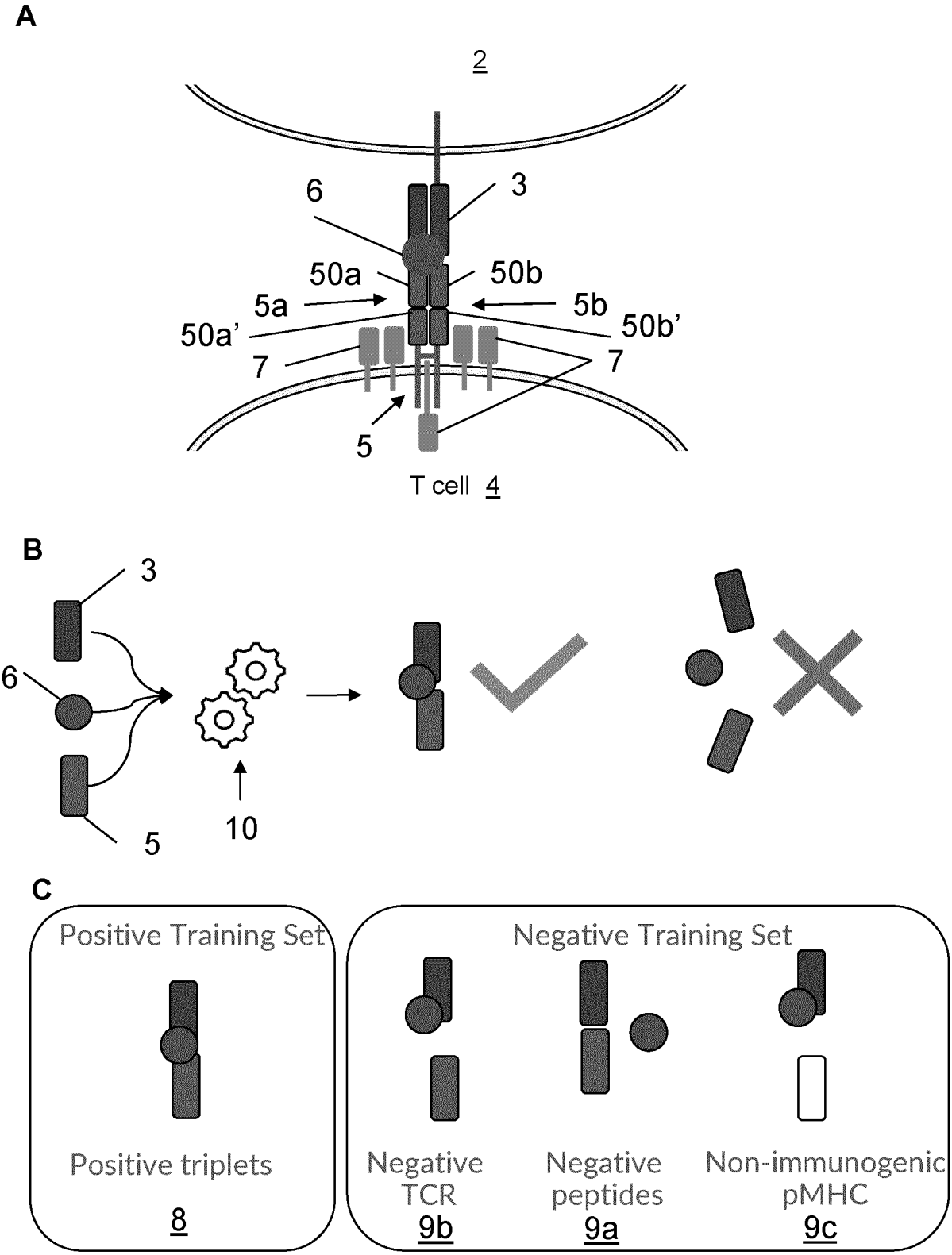

A
2
6
3
50a
50b
5a
5b
50a'
50b'
7
7
5
T cell 4
B
3
6
5
10
C
Positive Training Set
Positive triplets
8
Negative Training Set
Negative TCR
9b
Negative peptides
9a
Non-immunogenic pMHC
9c

Fig. 1

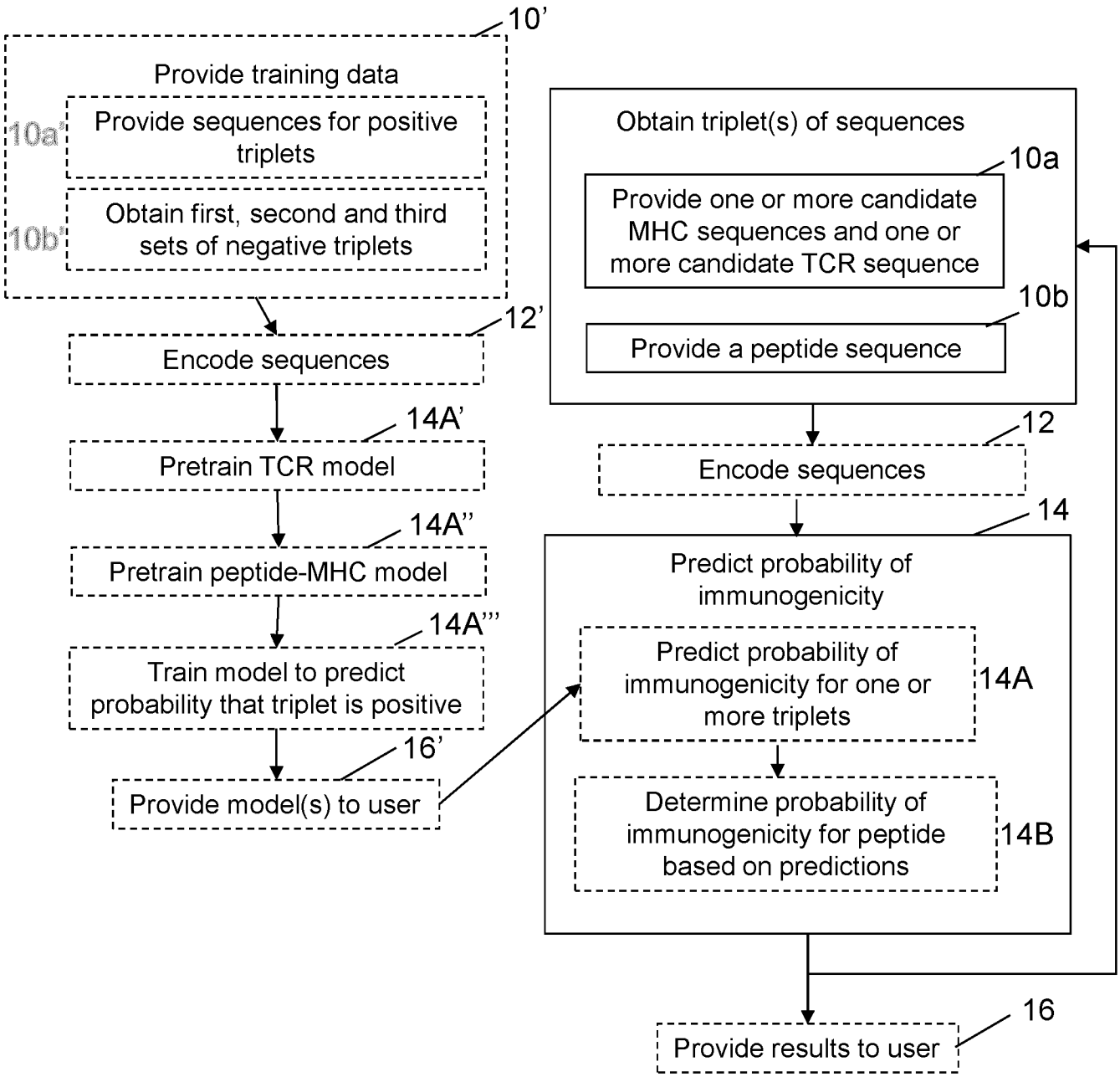

10'
Provide training data
10a'
Provide sequences for positive triplets
10b'
Obtain first, second and third sets of negative triplets
12'
Encode sequences
14A'
Pretrain TCR model
14A''
Pretrain peptide-MHC model
14A'''
Train model to predict probability that triplet is positive
16'
Provide model(s) to user
Obtain triplet(s) of sequences
10a
Provide one or more candidate MHC sequences and one or more candidate TCR sequence
10b
Provide a peptide sequence
12
Encode sequences
14
Predict probability of immunogenicity
Predict probability of immunogenicity for one or more triplets
14A
Determine probability of immunogenicity for peptide based on predictions
14B
16
Provide results to user

Fig. 2

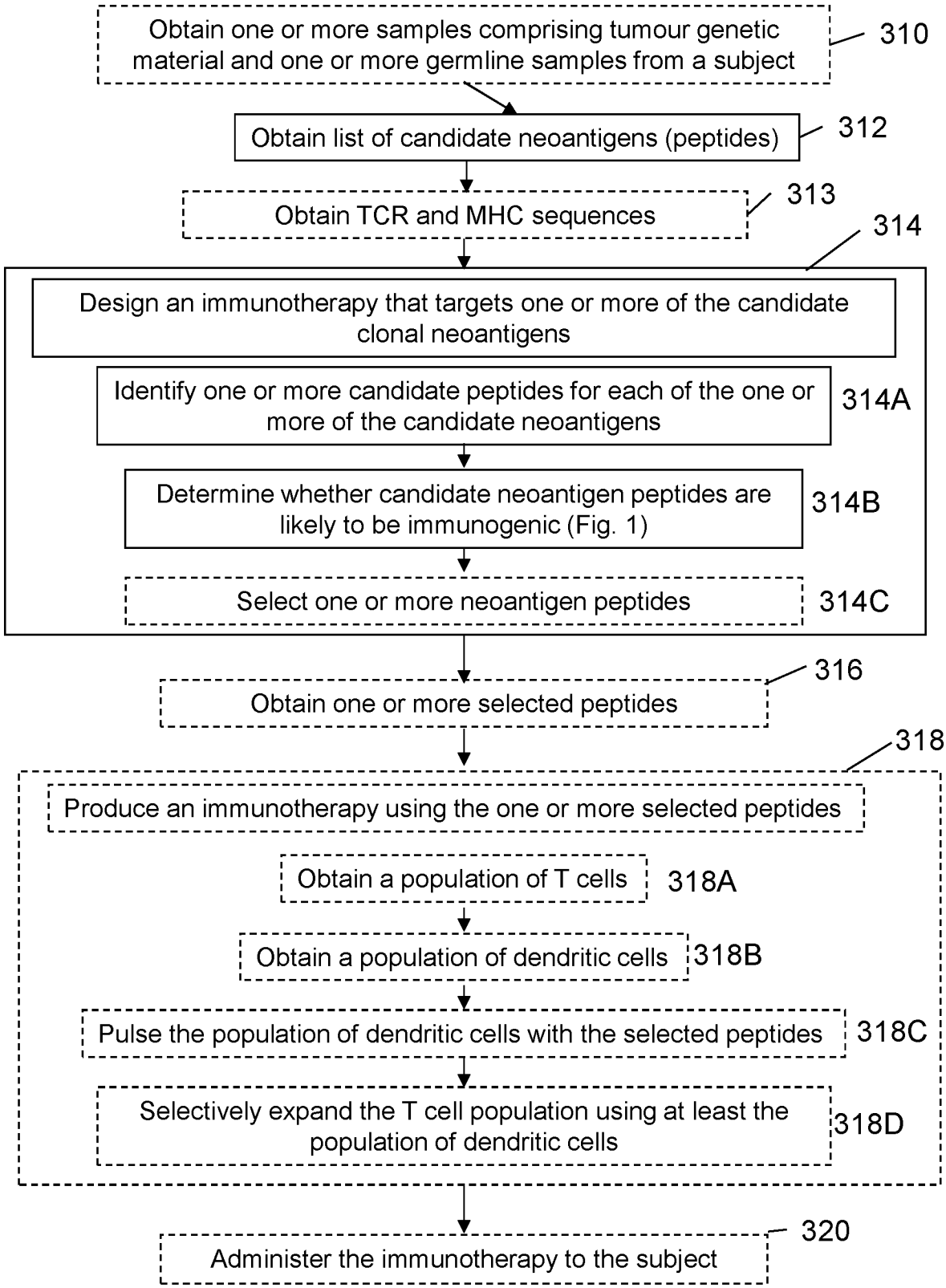

Obtain one or more samples comprising tumour genetic material and one or more germline samples from a subject
310
Obtain list of candidate neoantigens (peptides)
312
Obtain TCR and MHC sequences
313
314
Design an immunotherapy that targets one or more of the candidate clonal neoantigens
Identify one or more candidate peptides for each of the one or more of the candidate neoantigens
314A
Determine whether candidate neoantigen peptides are likely to be immunogenic (Fig. 1)
314B
Select one or more neoantigen peptides
314C
Obtain one or more selected peptides
316
318
Produce an immunotherapy using the one or more selected peptides
Obtain a population of T cells
318A
Obtain a population of dendritic cells
318B
Pulse the population of dendritic cells with the selected peptides
318C
Selectively expand the T cell population using at least the population of dendritic cells
318D
Administer the immunotherapy to the subject
320

Fig. 3

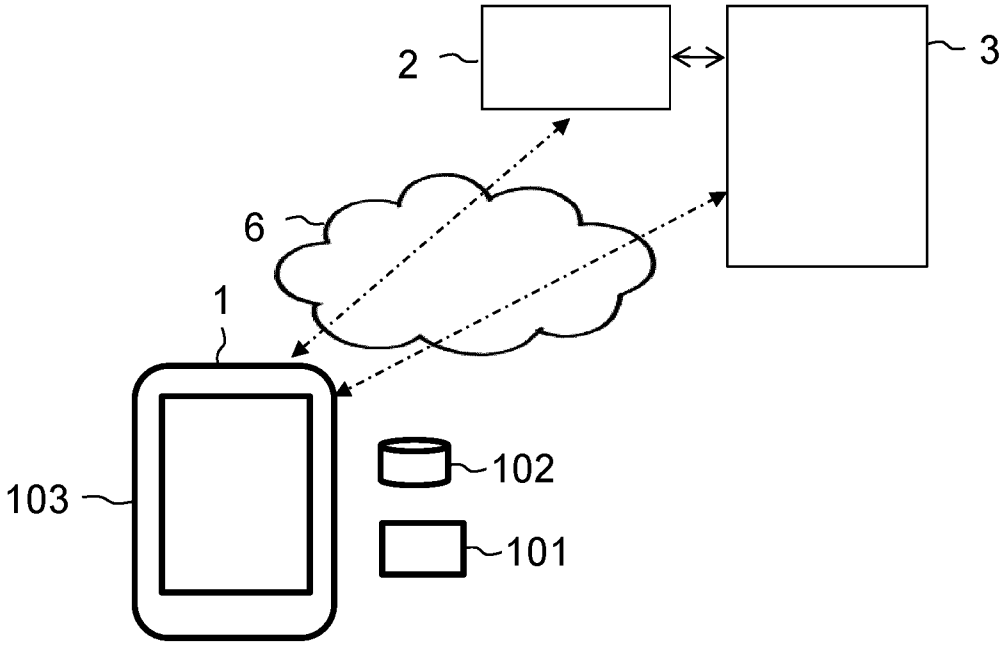


Fig. 4

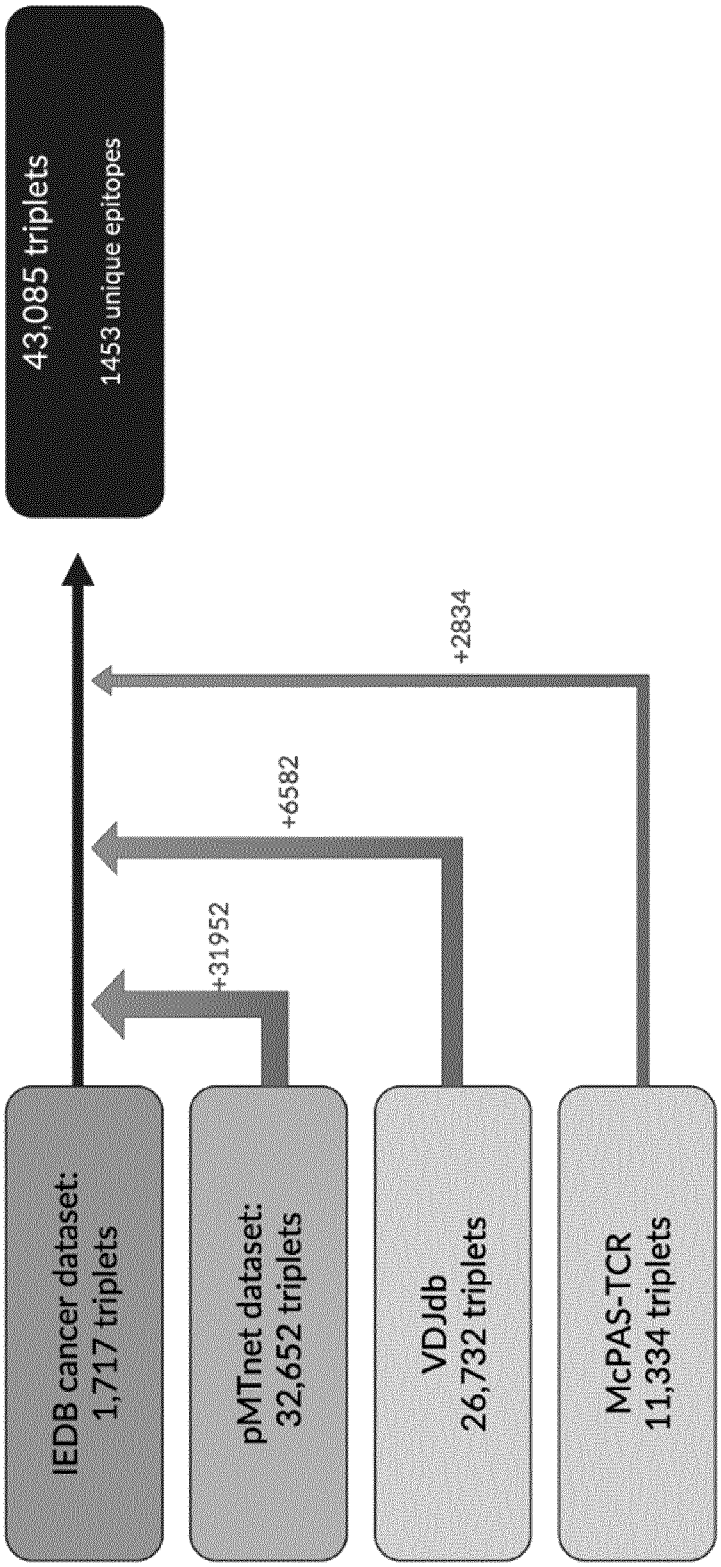

43,085 triplets
1453 unique epitopes
+2834
+6582
+31952
IEDB cancer dataset:
1,717 triplets
pMTnet dataset:
32,652 triplets
VDJdb
26,732 triplets
McPAS-TCR
11,334 triplets

Fig. 5

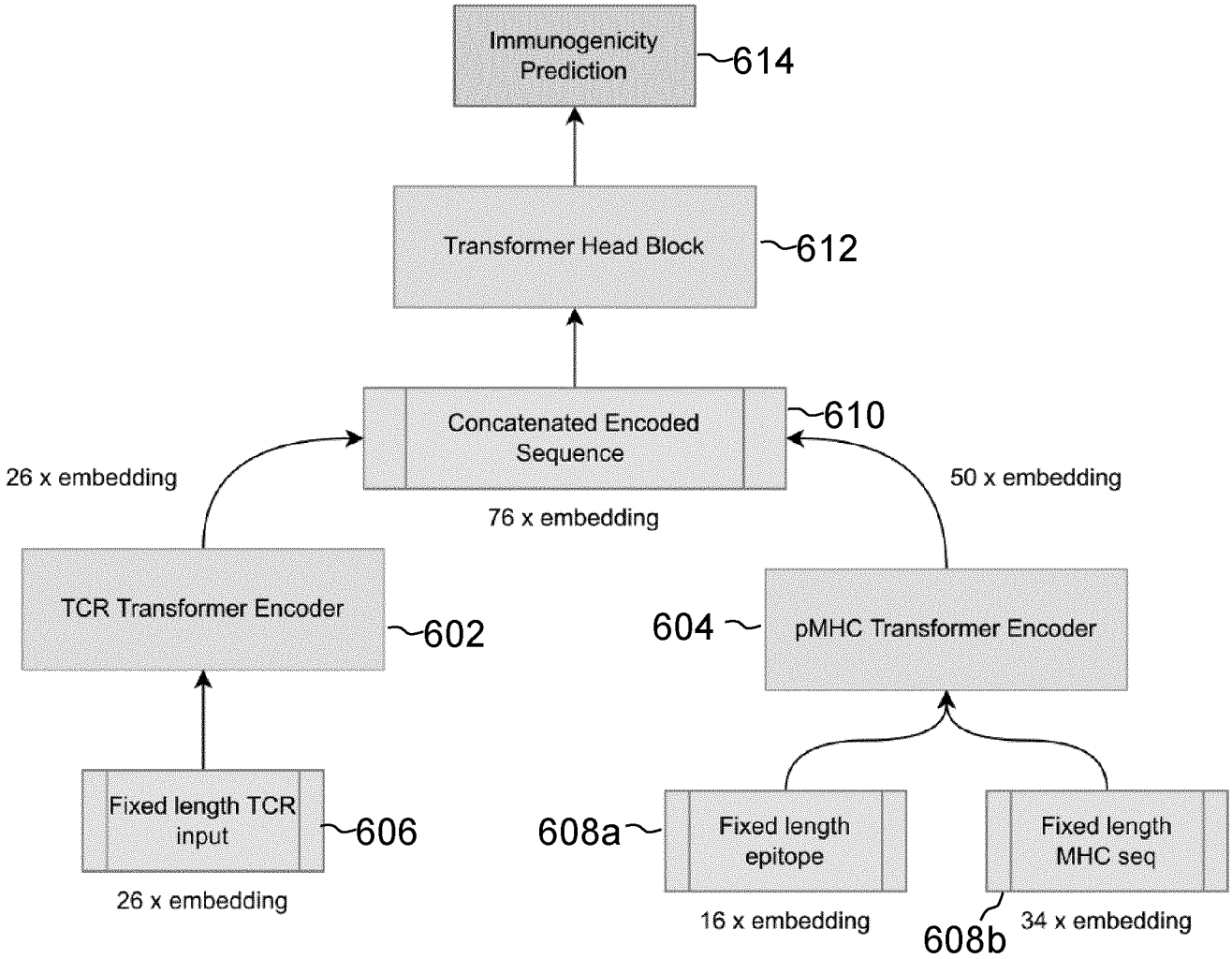

Immunogenicity Prediction
614
Transformer Head Block
612
Concatenated Encoded Sequence
610
26 x embedding
50 x embedding
76 x embedding
TCR Transformer Encoder
602
604
pMHC Transformer Encoder
Fixed length TCR input
606
608a
Fixed length epitope
Fixed length MHC seq
26 x embedding
16 x embedding
608b
34 x embedding

Fig. 6

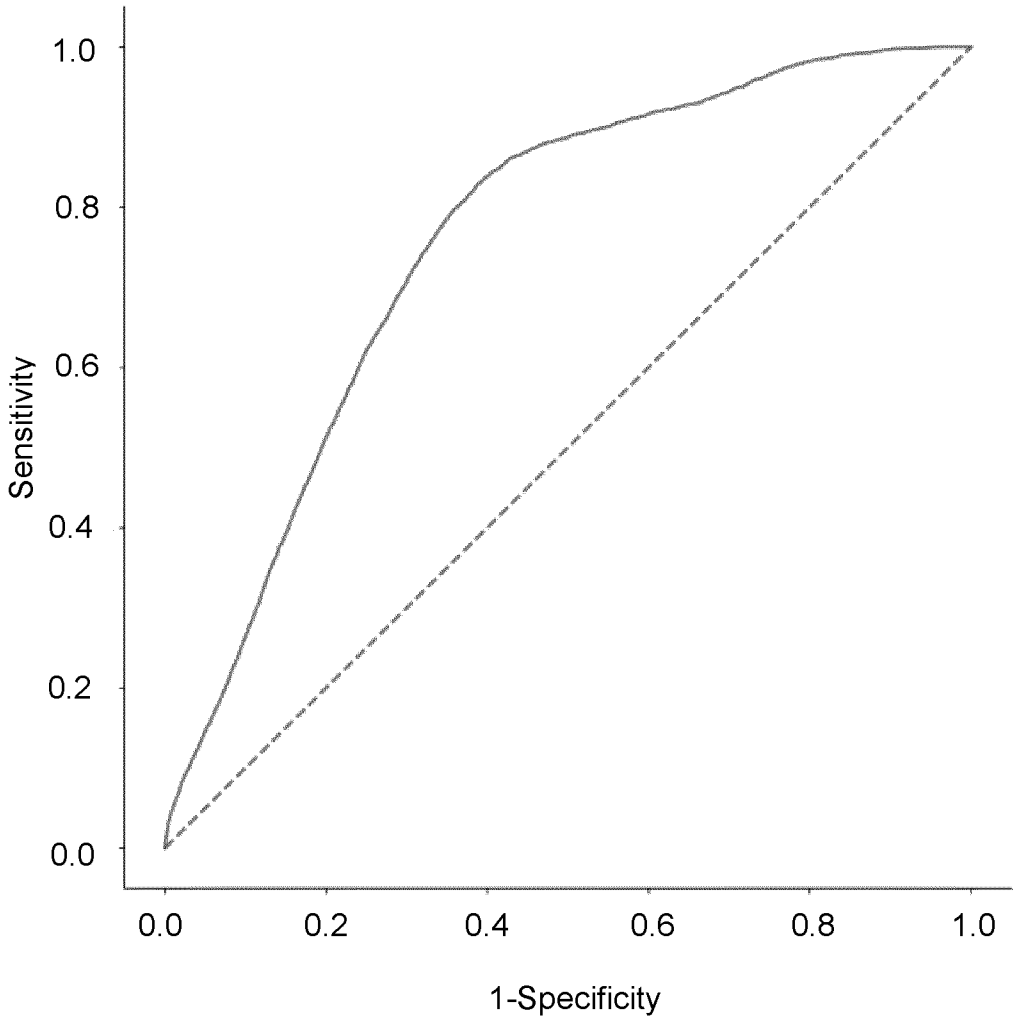


Fig. 7

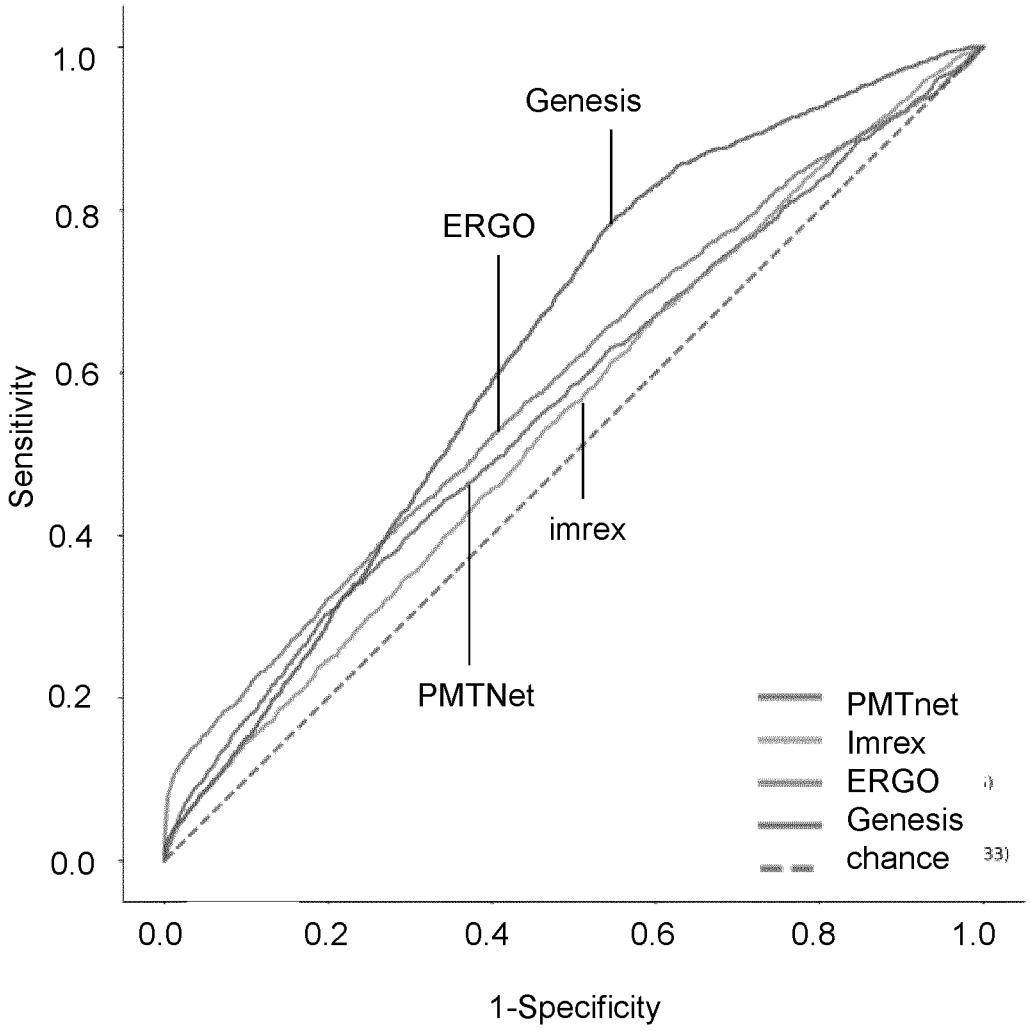

1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1-Specificity
Genesis
ERGO
imrex
PMTNet
PMTnet
Imrex
ERGO
Genesis
chance

Fig. 8A

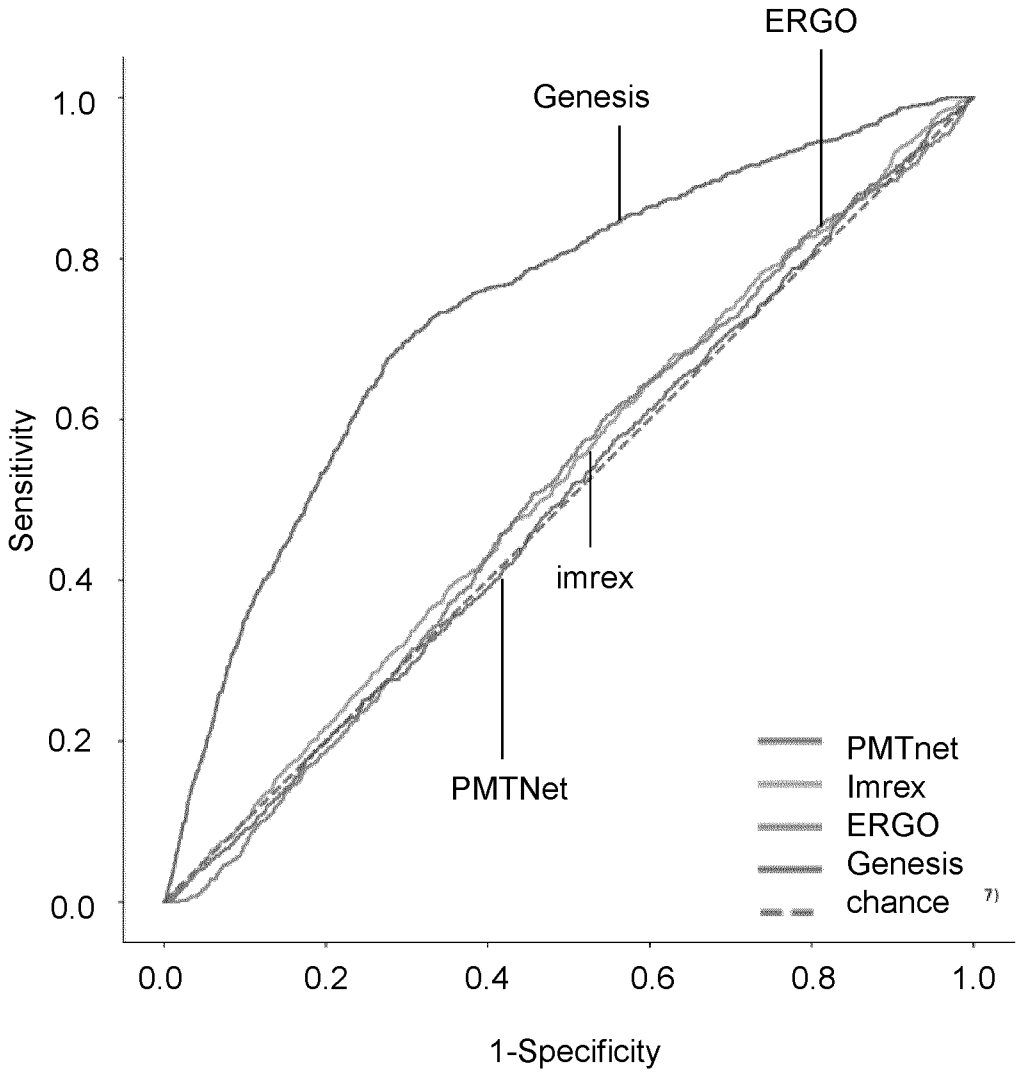

ERGO
Genesis
imrex
PMTNet
PMTnet
Imrex
ERGO
Genesis
chance
Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0

Fig. 8B

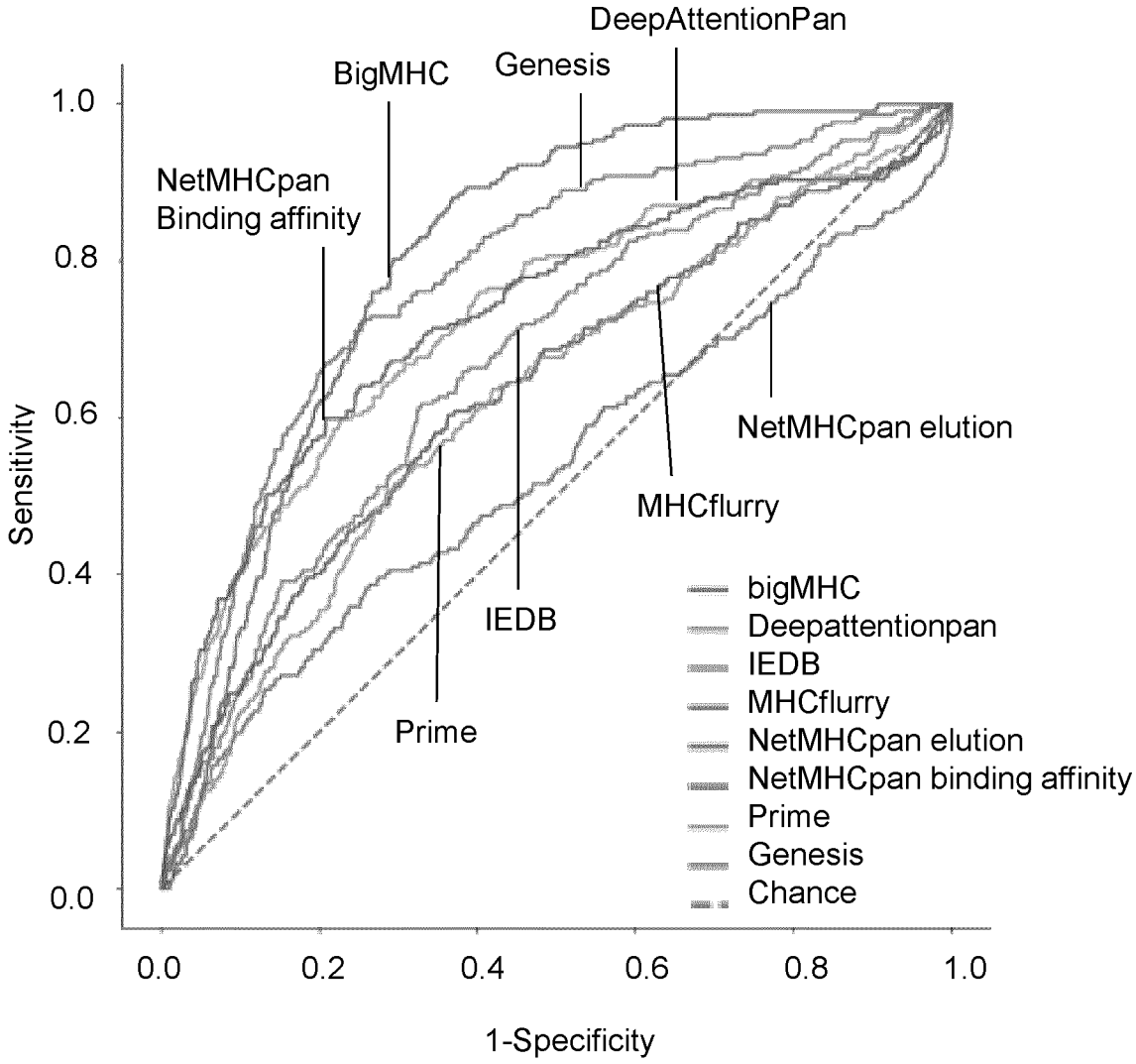

DeepAttentionPan
BigMHC
Genesis
NetMHCpan
Binding affinity
NetMHCpan elution
MHCflurry
IEDB
Prime
Sensitivity
1-Specificity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
bigMHC
Deepattentionpan
IEDB
MHCflurry
NetMHCpan elution
NetMHCpan binding affinity
Prime
Genesis
Chance

Fig. 9

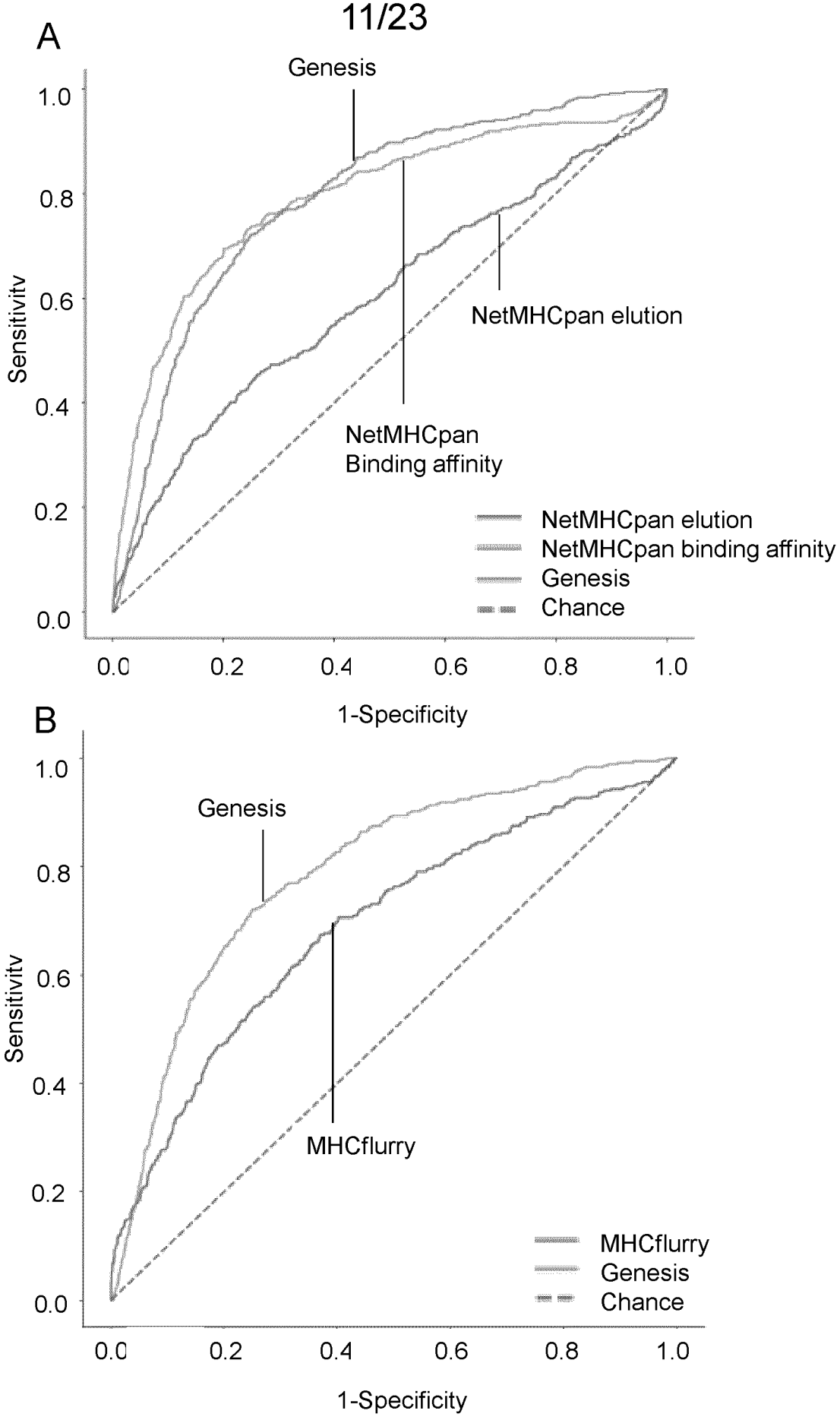
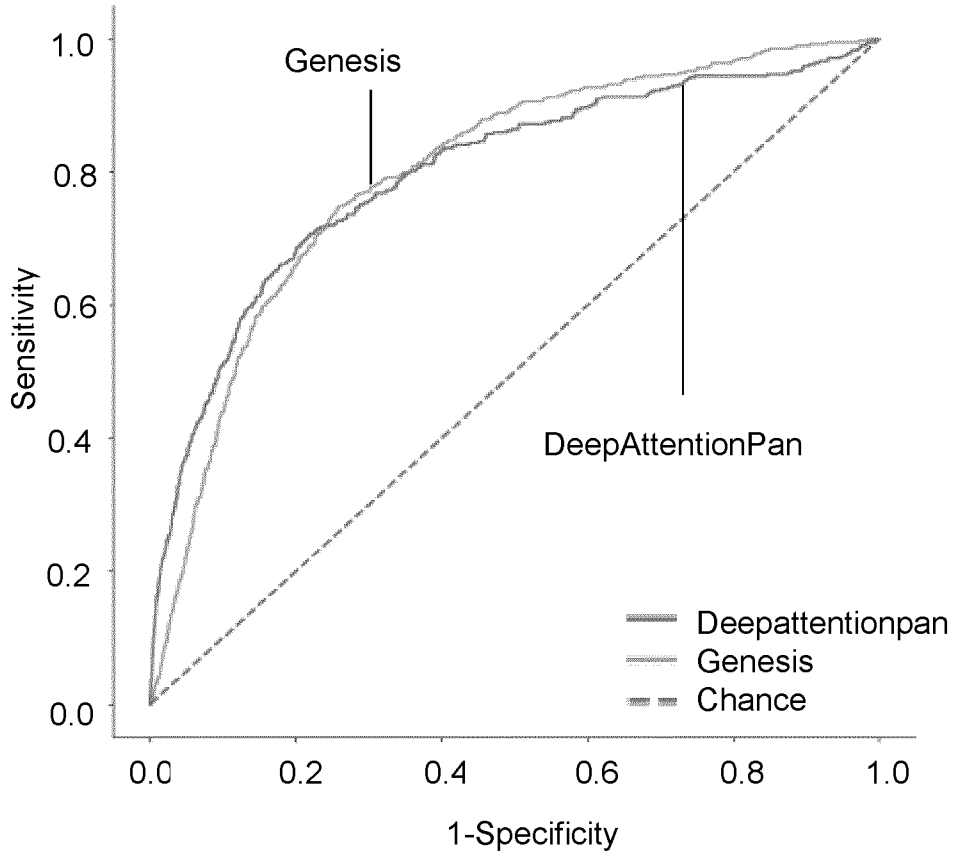

A
Genesis
NetMHCpan elution
NetMHCpan
Binding affinity
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1-Specificity
NetMHCpan elution
NetMHCpan binding affinity
Genesis
Chance
B
Genesis
MHCflurry
Sensitivity
1-Specificity
MHCflurry
Genesis
Chance

Fig. 10

C



D

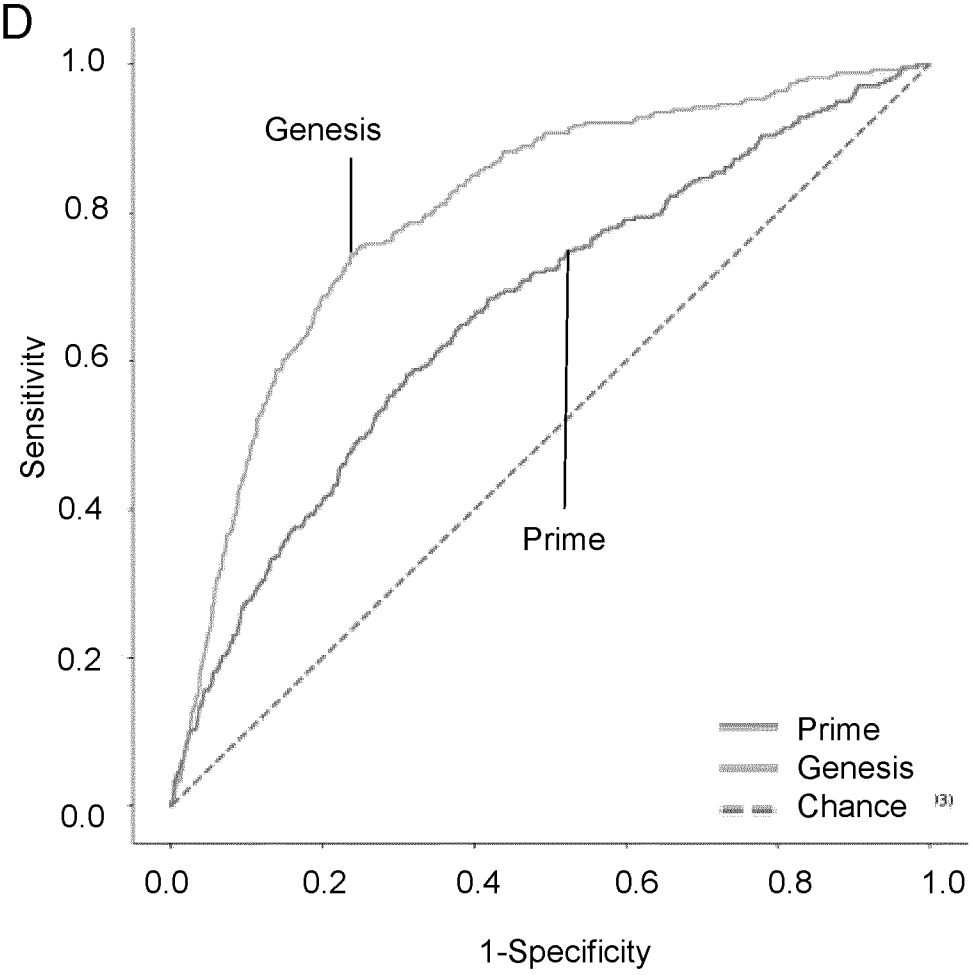


Fig. 10 (Continued)

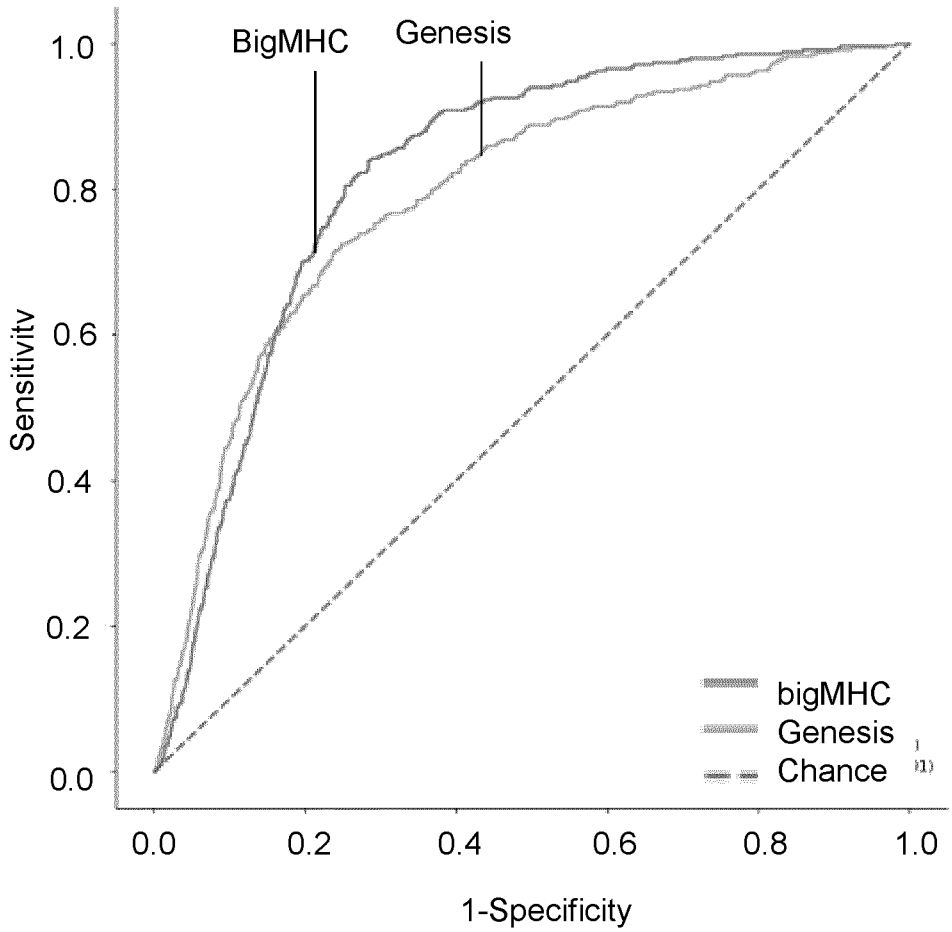

BigMHC
Genesis
1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1-Specificity
bigMHC
Genesis
Chance

Fig. 10E

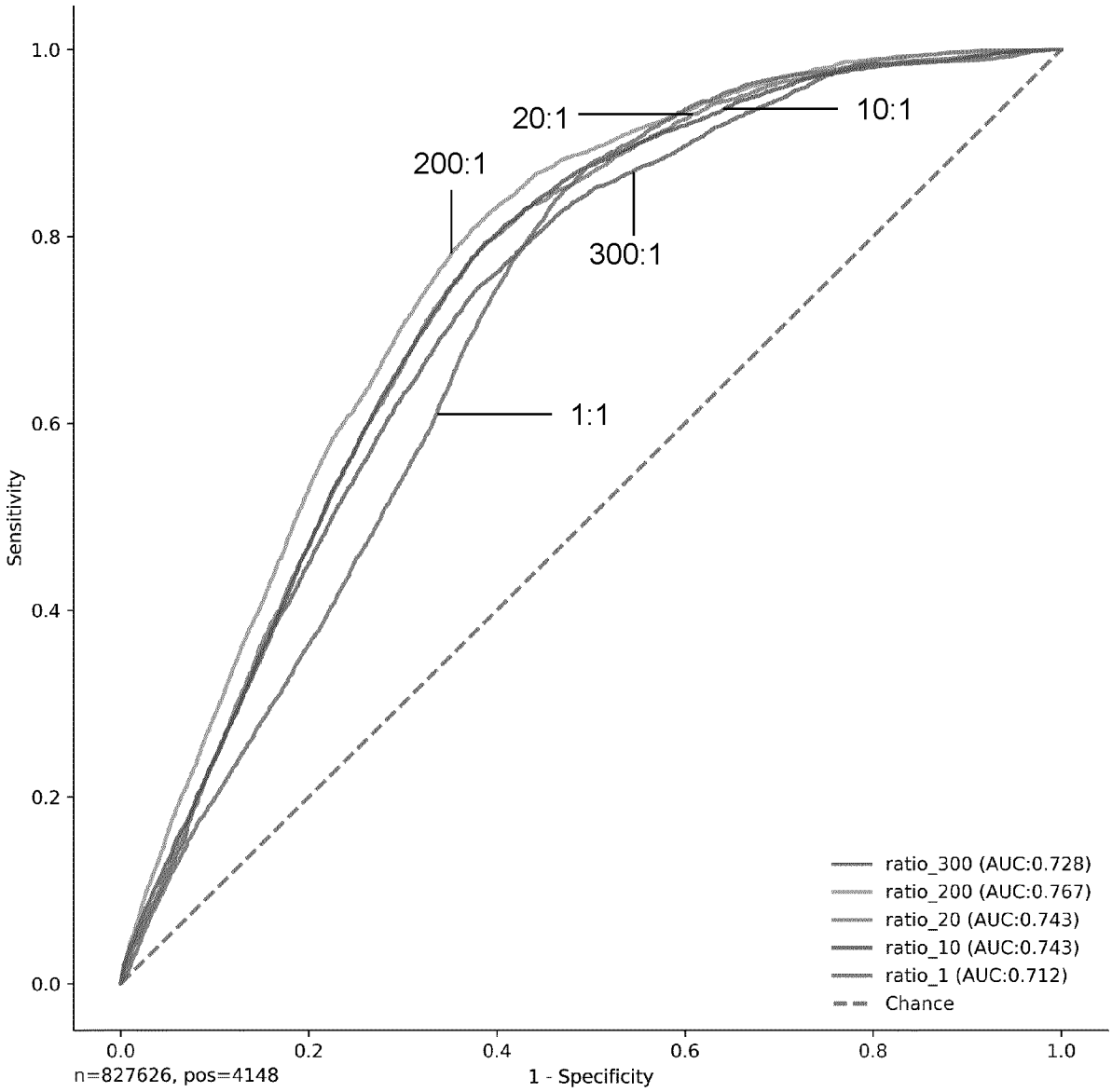

1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1 - Specificity
n=827626, pos=4148
20:1
10:1
200:1
300:1
1:1
ratio_300 (AUC:0.728)
ratio_200 (AUC:0.767)
ratio_20 (AUC:0.743)
ratio_10 (AUC:0.743)
ratio_1 (AUC:0.712)
Chance

Fig. 11

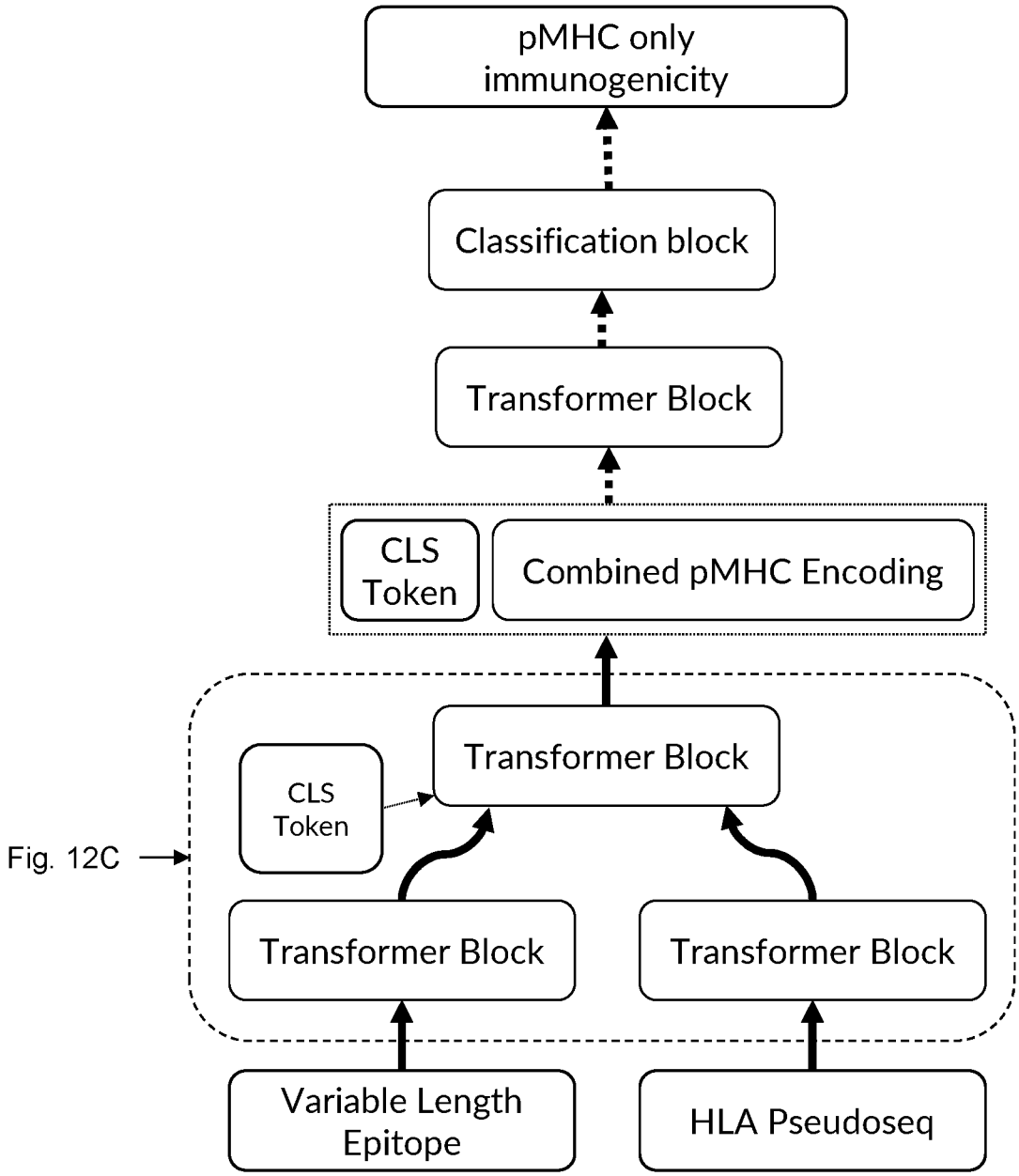

pMHC only immunogenicity
Classification block
Transformer Block
CLS Token
Combined pMHC Encoding
Transformer Block
CLS Token
Fig. 12C
Transformer Block
Transformer Block
Variable Length Epitope
HLA Pseudoseq

Fig. 12A

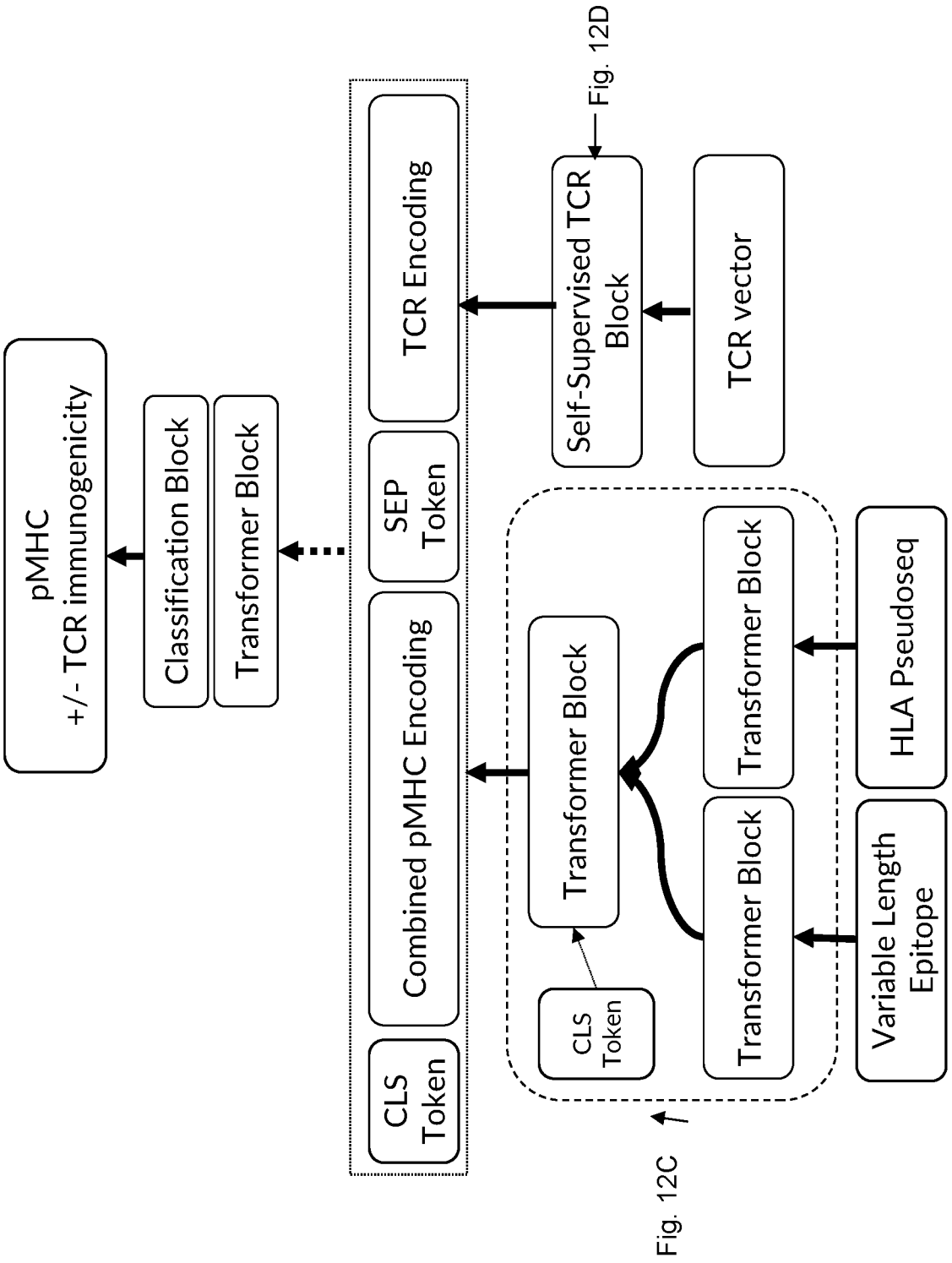

pMHC
+/- TCR immunogenicity
Classification Block
Transformer Block
CLS Token
Combined pMHC Encoding
SEP Token
TCR Encoding
Fig. 12D
Self-Supervised TCR Block
TCR vector
CLS Token
Transformer Block
Transformer Block
Transformer Block
Variable Length Epitope
HLA Pseudoseq
Fig. 12C

Fig. 12B

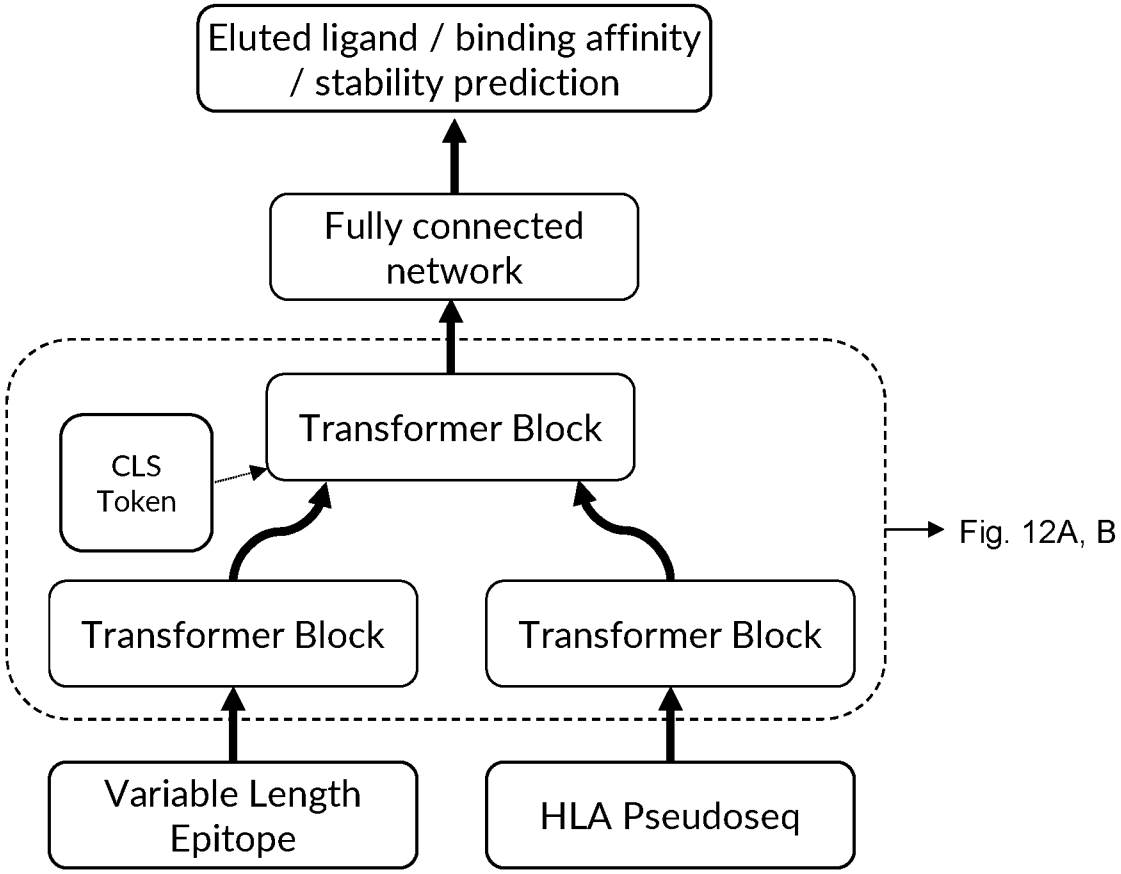


Fig. 12C

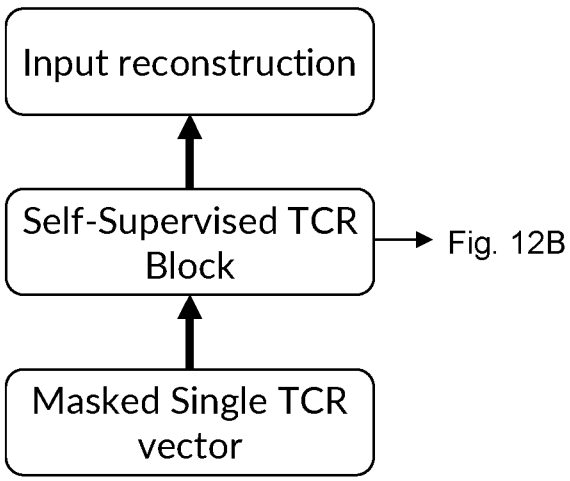


Fig. 12D

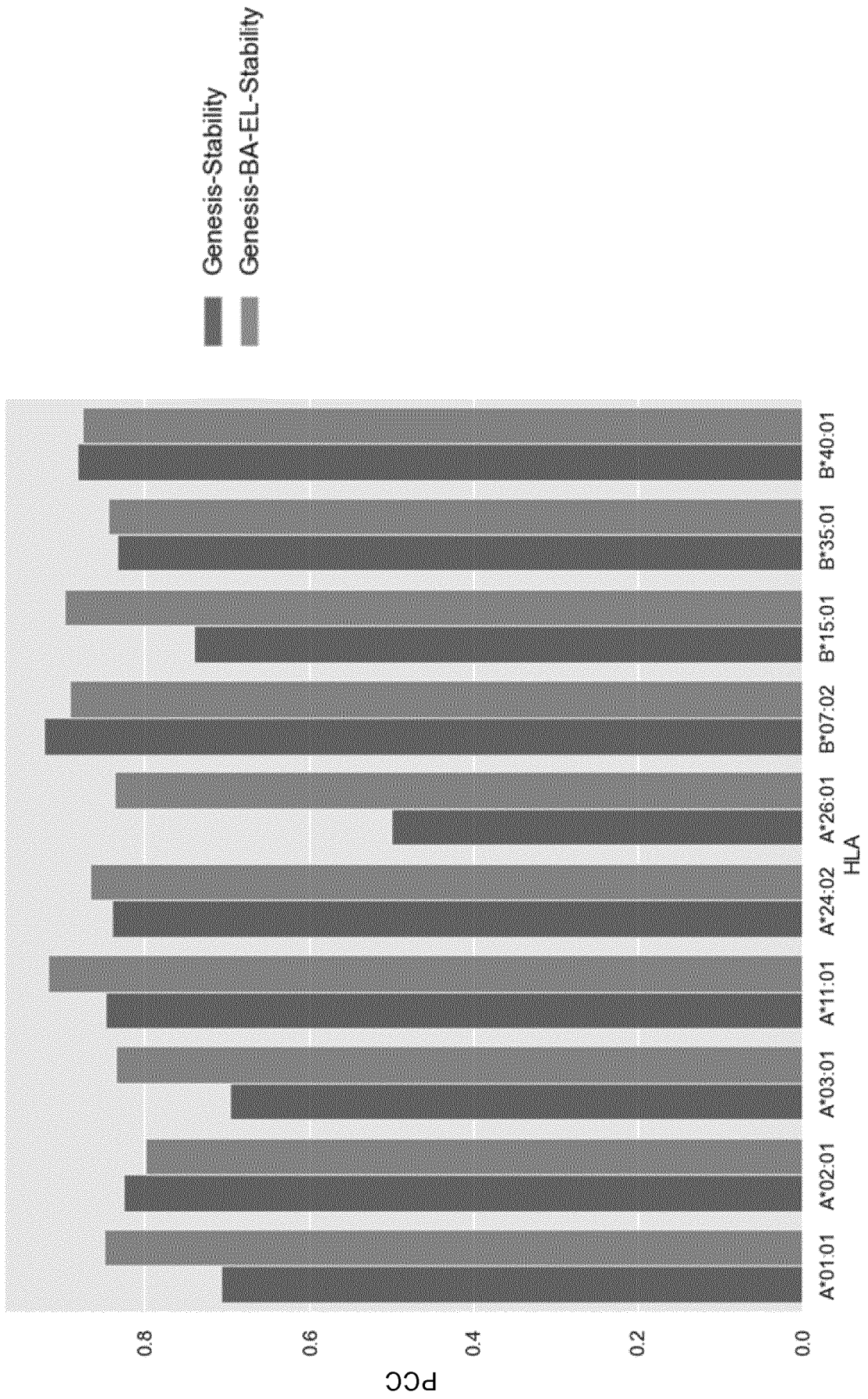

Genesis-Stability
Genesis-BA-EL-Stability
PCC
0.0
0.2
0.4
0.6
0.8
HLA
A*01:01
A*02:01
A*03:01
A*11:01
A*24:02
A*26:01
B*07:02
B*15:01
B*35:01
B*40:01

Fig. 13A

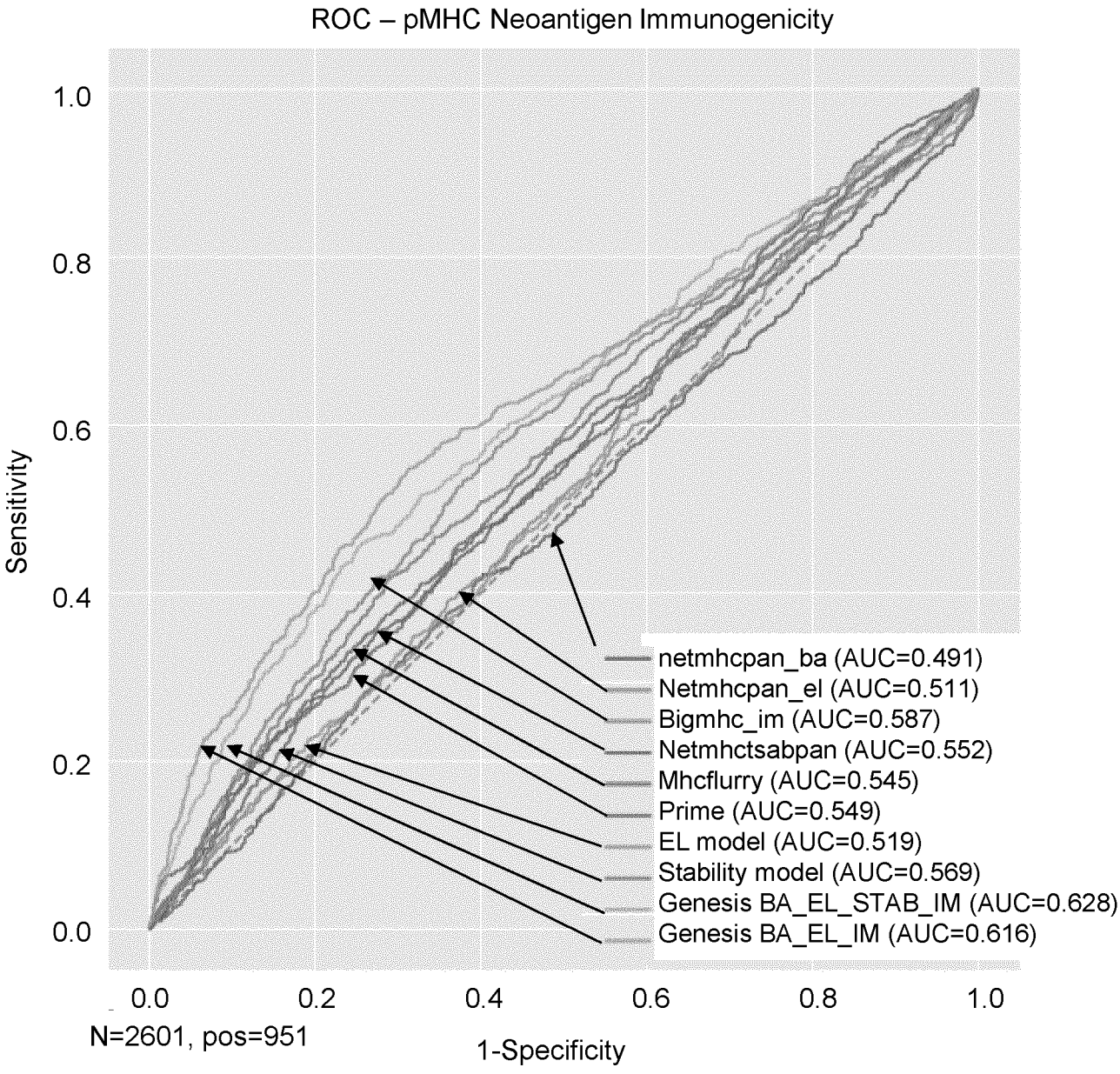

ROC – pMHC Neoantigen Immunogenicity
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1-Specificity
N=2601, pos=951
netmhcpan_ba (AUC=0.491)
Netmhcpan_el (AUC=0.511)
Bigmhc_im (AUC=0.587)
Netmhctsabpan (AUC=0.552)
Mhcflurry (AUC=0.545)
Prime (AUC=0.549)
EL model (AUC=0.519)
Stability model (AUC=0.569)
Genesis BA_EL_STAB_IM (AUC=0.628)
Genesis BA_EL_IM (AUC=0.616)

Fig. 13B-1

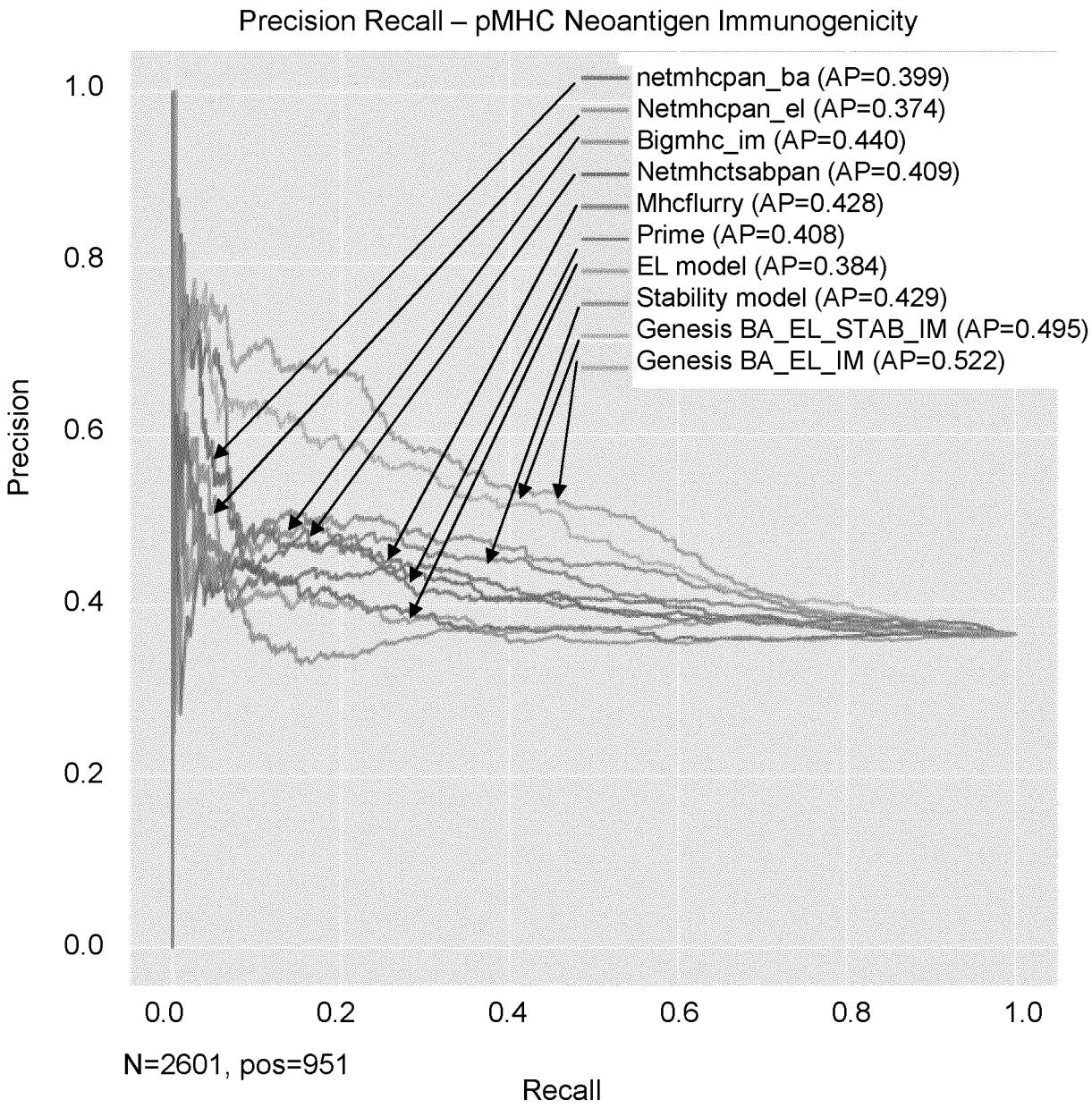


Fig. 13B-2

Sensitivity

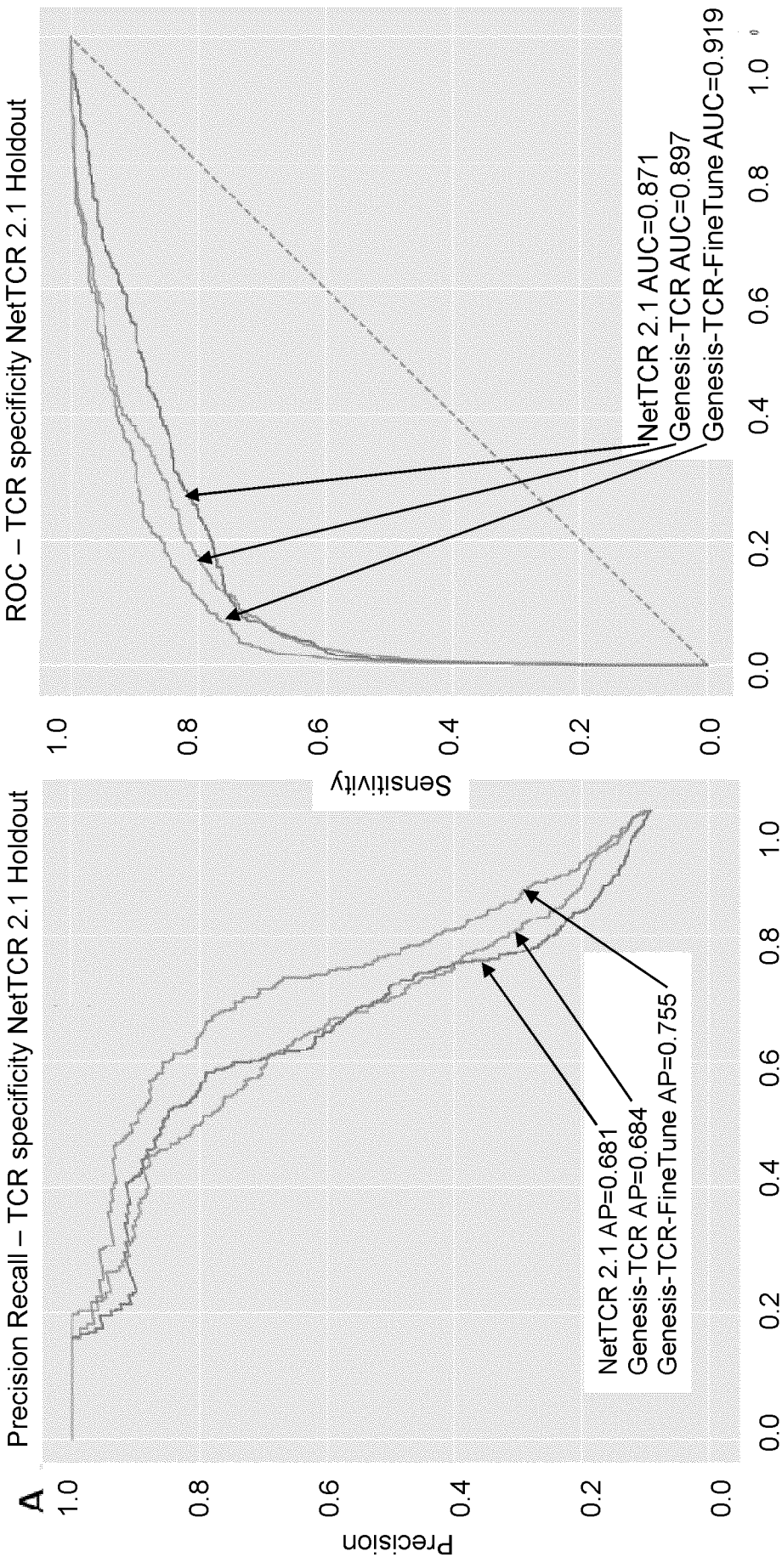

A
Precision Recall – TCR specificity NetTCR 2.1 Holdout
Precision
Recall
NetTCR 2.1 AP=0.681
Genesis-TCR AP=0.684
Genesis-TCR-FineTune AP=0.755
ROC – TCR specificity NetTCR 2.1 Holdout
Sensitivity
1-Specificity
NetTCR 2.1 AUC=0.871
Genesis-TCR AUC=0.897
Genesis-TCR-FineTune AUC=0.919

Fig. 14A

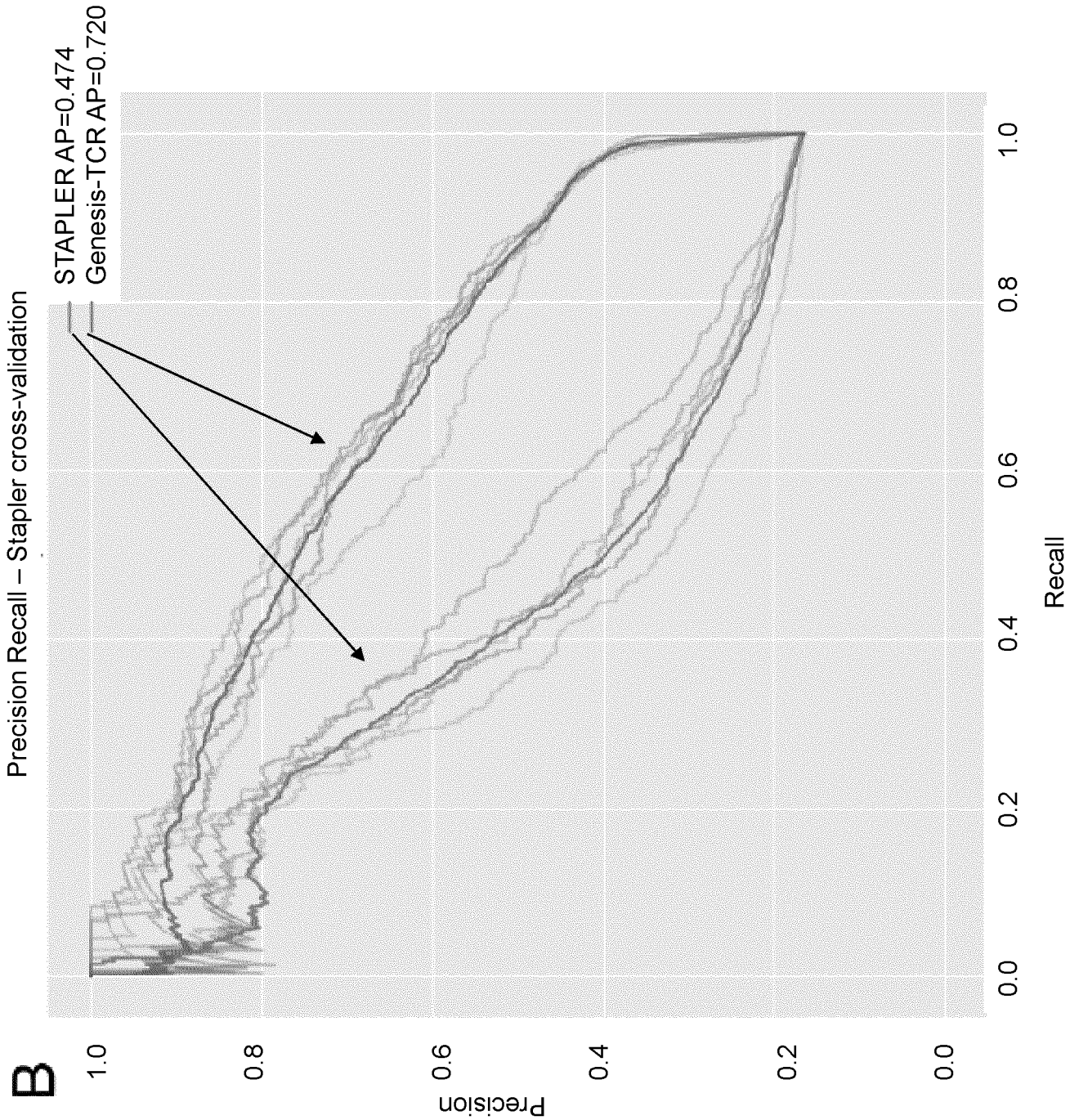

B
Precision Recall – Stapler cross-validation
STAPLER AP=0.474
Genesis-TCR AP=0.720
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

Fig. 14B

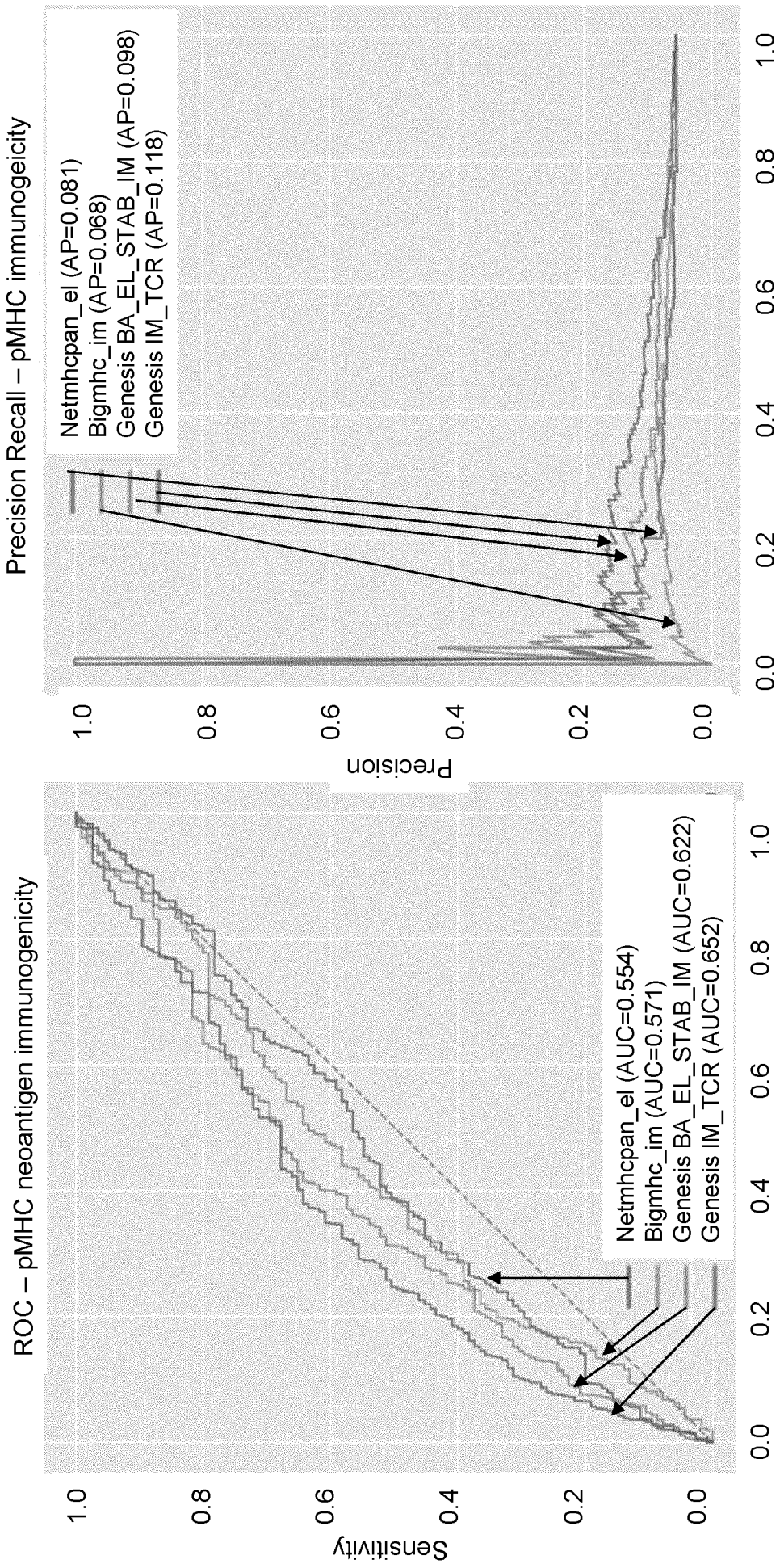

Precision Recall – pMHC immunogeicity
Netmhcpan_el (AP=0.081)
Bigmhc_im (AP=0.068)
Genesis BA_EL_STAB_IM (AP=0.098)
Genesis IM_TCR (AP=0.118)
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
ROC – pMHC neoantigen immunogenicity
Netmhcpan_el (AUC=0.554)
Bigmhc_im (AUC=0.571)
Genesis BA_EL_STAB_IM (AUC=0.622)
Genesis IM_TCR (AUC=0.652)
Sensitivity
1-Specificity
N=2072 pos=115

Fig. 15

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2024/057046

A. CLASSIFICATION OF SUBJECT MATTER

INV. G16B40/20 G16B20/00 G16B5/00 A61K39/00 C07K14/725
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G16B C07K A61K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2022/122690 A1 (LIU KAI [US] ET AL) 21 April 2022 (2022-04-21) whole document, in particular: claim 1, 17, 19, [0186],[0187],[0191]	1-31
A	WO 2020/132235 A1 (MERCK SHARP & DOHME [US]) 25 June 2020 (2020-06-25) the whole document	1-31

-/--

[x] Further documents are listed in the continuation of Box C.

[x] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
25 June 2024	10/07/2024

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Eberhardt, Anja

3

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JINGCHENG WU ET AL: "DeepHLApan: A Deep Learning Approach for Neoantigen Prediction Considering Both HLA-Peptide Binding and Immunogenicity", FRONTIERS IN IMMUNOLOGY, [Online] vol. 10, 1 November 2019 (2019-11-01), page 2559, XP055682961, DOI: 10.3389/fimmu.2019.02559 [retrieved on 2024-06-06] the whole document	1-31
A	WEILONG ZHAO ET AL: "Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes", PLOS COMPUTATIONAL BIOLOGY, [Online] vol. 14, no. 11, 8 November 2018 (2018-11-08), page e1006457, XP055682482, DOI: 10.1371/journal.pcbi.1006457 [retrieved on 2024-06-06] the whole document	1-31
A	WO 2020/132586 A1 (NEON THERAPEUTICS INC [US]) 25 June 2020 (2020-06-25) the whole document	1-31

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2024/057046

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2022122690 A1	21-04-2022	AU 2021308081 A1	17-11-2022
		BR 112023000827 A2	07-02-2023
		CA 3180799 A1	20-01-2022
		CN 115997254 A	21-04-2023
		EP 4182924 A1	24-05-2023
		IL 299801 A	01-03-2023
		JP 2023534283 A	08-08-2023
		KR 20230042048 A	27-03-2023
		US 2022122690 A1	21-04-2022
		WO 2022016125 A1	20-01-2022
WO 2020132235 A1	25-06-2020	US 2022076783 A1	10-03-2022
		WO 2020132235 A1	25-06-2020
WO 2020132586 A1	25-06-2020	AU 2019404547 A1	22-07-2021
		BR 112021012278 A2	14-12-2021
		CA 3124457 A1	25-06-2020
		CN 113474840 A	01-10-2021
		EP 3899954 A1	27-10-2021
		IL 284195 A	31-08-2021
		JP 7236543 B2	09-03-2023
		JP 2022518355 A	15-03-2022
		JP 2023071806 A	23-05-2023
		KR 20210130705 A	01-11-2021
		SG 11202106678P A	29-07-2021
		US 2020279616 A1	03-09-2020
		US 2022199198 A1	23-06-2022
		WO 2020132586 A1	25-06-2020