CLAIMS

Claim 1. A computer-implemented method for determining whether a pair of protein chains comprising a first chain and a second chain are likely to form a functional antigen-binding protein, the method comprising:

providing a query pair of sequences comprising the sequence of the first protein chain and the sequence of the second protein chain as input to a deep learning model configured to take as input a pair of protein chain sequences and to produce as output a score indicative of the probability that the pair of protein chains will form a functional antigen-binding protein, wherein the deep learning model comprises an encoder module and a classifier module, and wherein the deep learning module has been trained using paired training sequences from known antigen-binding proteins.

Claim 2. A computer-implemented method of identifying an antigen-binding protein comprising a first protein chain and a second protein chain, the method comprising:

providing a query first protein chain, and

identifying a second protein chain by:

providing one or more candidate second protein chain sequences; and

determining whether the one or more candidate second chain sequences are likely to form a functional antigen-binding protein with the query first protein chain by:

providing as input to a deep learning model one or more query pair of sequences each comprising (i) the sequence of the first protein chain and (ii) a candidate second protein chain sequence, wherein the deep learning model is configured to take as input one or more pairs of protein chain sequences and to produce as output a score indicative of the probability that each pair of protein chains will form a functional antigen-binding protein or information derived therefrom, wherein the deep learning model comprises an encoder module and a classifier module, and wherein the deep learning module has been trained using paired training sequences from known antigen-binding proteins, optionally wherein identifying a second protein chain comprises providing a plurality of candidate second chain sequences and the information derived from the score comprises a ranking of the pairs of protein chain sequences wherein pairs of protein sequences that are more likely to form a functional antigen-binding protein are ranked higher than pairs of protein sequences that are less likely to form a functional antigen-binding protein.

Claim 3. The method of claim 1 of claim 2, wherein the antigen-binding protein comprises:

(i) a heavy-light chain pair, wherein the first chain is a heavy chain or a light chain, and the second chain is a light chain or a heavy chain, optionally wherein the first chain is a heavy chain and the second chain is a light chain; or

(ii) an αβ chain pair, wherein the first chain is a β chain or an α chain, and the second chain is an α chain or a β chain, optionally wherein the first chain is a β chain and the second chain is an α chain; or

(ii) a γδ chain pair, wherein the first chain is a δ chain or a γ chain, and the second chain is a γ chain or a δ chain, optionally wherein the first chain is a δ chain and the second chain is a γ chain.

Claim 4. The method of any preceding claim, wherein the encoder module comprises:

(i) one or two encoders that have been pretrained using training sequences from unpaired protein chains from known antigen-binding proteins, or a decoder that has been pretrained using training sequences comprising paired and unpaired protein chains from known antigen-binding proteins; and/or

(ii) one or two encoders of a sequence-to-sequence model or a decoder of a generative language model, optionally wherein the sequence-to-sequence model or generative language model is a transformer-based model, or

one or two transformer-based encoder or decoder models; and/or

(iii) one or two encoder models of a sequence-to-sequence model, wherein the encoder models have been trained using masked language modelling, optionally wherein training the encoder models comprises training the model to replace randomly masked positions of training amino acid sequences and/or wherein 15% of positions are masked during training and/or wherein masked positions are replaced with mask tokens, a random amino acid, or the original amino acid at the position, or wherein the encoder module comprises a decoder model of a generative language model, wherein the decoder model has been trained using causal language modelling; and/or

(iv) one or two encoders that have been pretrained using training sequences comprising unpaired first and second protein chains from known antigen-binding proteins, optionally wherein the training sequences comprise at least 1 million, at least 2 million, at least 5 million, at least 10 million, at least 20 million or at least 50 million individual sequences, or wherein the encoder module comprises a decoder that has been pretrained using training sequences comprising unpaired first and second protein chains from known antigen-binding proteins and paired first and second protein chains from known antigen-binding proteins,

optionally wherein the training sequences comprise at least 500 million, 600 million or 700 million individual sequences and/or at least 1, 1.5 or 2 million paired sequences.

Claim 5. The method of any preceding claim, wherein the encoder module comprises two copies of an encoder that has been pretrained using training sequences comprising unpaired first and second protein chains from known antigen-binding proteins.

Claim 6. The method of any of claims 1 to 4, wherein the encoder module comprises an encoder that has been pretrained using training sequences comprising the concatenation of a sequence from a first chain and a sequence from a second protein chain from known antigen-binding proteins, wherein the concatenated sequences comprise sequences from unpaired protein chains, or wherein the encoder module comprises a decoder that has been pretrained using training sequences each comprising a single chain or the concatenation of a sequence from a first chain and a sequence from a second protein chain from a known antigen-binding protein.

Claim 7. The method of any of claims 1 to 5, wherein the deep learning model further comprises a cross-attention module that takes as input the output of the encoder module, and produces an output used by the classifier module to provide the core indicative of the probability that the pair of protein chains will form a functional antigen-binding protein.

Claim 8. The method of claim 7, wherein the cross-attention module comprises one or more cross-attention blocks, wherein each cross-attention block comprises a self-attention layer and a cross attention layer, and/or wherein the cross-attention module comprises a plurality of cross-attention blocks.

Claim 9. The method of claim 8, wherein the cross-attention module comprises one or more cross-attention blocks comprising a cross-attention layer and a self-attention layer, and each cross-attention layer and self-attention layer comprises a plurality of attention heads.

Claim 10.The method of any of claims 7 to 9, wherein each self-attention layer comprises one or more attention heads that attend to the output of a first encoder taking as input the first chain sequence, and/or wherein each cross-attention block further comprises a residual connection between the output of a first encoder and the output of the self-attention layer; and/or

wherein each cross-attention layer comprises one or more attention heads that attend to: (i) the output of the self-attention layer or the output of a first encoder taking as input the first chain sequence, and (ii) the output of a second encoder taking as input the second chain sequence, and/or wherein each cross-attention block further comprises a residual connection between the output of a first encoder and the output of the cross-attention layer.

Claim 11. The method of any preceding claim, wherein the classification module comprises a softmax layer that produces a score between 0 and 1 that can be interpreted as the probability that a pair of input protein chains will form a functional antigen-binding protein, and/or wherein the classification module comprises one or more of: a dimensionality reduction layer, a regularisation mechanism, and a layer with an activation function, optionally wherein a dimensionality reduction layer comprises an attention pooling layer and/or wherein a regularisation layer comprises a dropout mechanism, and/or wherein an activation function is independently selected from Tanh, Leaky ReLU, SmeLU, GeLU Swish and ReLU; and/or

wherein the method further comprises providing training data comprising training first and second protein chain sequences from known antigen-binding proteins, and

training the deep learning model to take as input a pair of protein chain sequences and to produce as output a score indicative of the probability that the pair of protein chains will form a functional antigen-binding protein, wherein the deep learning model comprises an encoder module and a classifier module.

Claim 12. The method of any preceding claim, wherein the deep learning model takes as input a plurality of query pairs of chain sequences and produces as output a respective score indicative of the probability that each query pair of protein chains will form a functional antigen-binding protein, and/or a ranking of the plurality of query pairs of chain sequences such that query pairs of protein sequences that are more likely to form a functional antigen-binding protein are ranked higher than query pairs of protein sequences that are less likely to form a functional antigen-binding protein; and/or wherein the paired training sequences from known antigen-binding proteins comprise paired training heavy and light chain sequences from single B cell sequencing data, and/or wherein the training data further comprises a negative set comprising randomly paired first and second protein chain sequences,
optionally wherein the randomly paired first and second protein chain sequences are obtaining by re-pairing the paired training sequences from known antigen-binding proteins and/or by randomly pairing unpaired training sequences and/or wherein the negative set comprises paired first and second protein chain sequences from antigen-binding proteins previously determined to be non-functional and/or wherein the paired training sequences are associated with a first label and the randomly paired training sequences are associated with a second label.

Claim 13. The method of any preceding claim, wherein:
(i) providing a pair of protein chain sequences as input to the deep learning model comprise encoding each of the protein chain sequences using a predetermined encoding scheme, optionally wherein each amino acid is individually encoded or wherein sequences are encoded

using tokens that each correspond to an individual k-mer, and/or wherein each protein chain sequence is preceded or followed by a special token indicating whether the protein chain sequence is a first or second protein chain; and/or

(ii) providing a query first protein chain or a query first protein chain of a query pair comprises:

obtaining the sequence of the query first protein chain from a user through a user interface, from a computing device, from a sequence acquisition means or a computing device associated with a sequence acquisition means, from a database or other computer readable medium; and/or

sequencing a sample comprising genetic material encoding for an antigen-binding molecule comprising the query sequence, optionally wherein obtaining the query sequence comprises performing B cell bulk sequencing of a sample comprising B cells, T cell bulk sequencing of a sample comprising T cells, or bulk sequencing of a sample comprising any other cells expressing an antigen-binding molecule comprising the query sequence, or genetic material derived therefrom, such as a B cell receptor library or a T cell receptor library; and/or

obtaining a sample comprising B cells, T cells or other cells expressing an antigen-binding molecule comprising the query sequence, or genetic material derived therefrom, such as a B cell receptor library or T cell receptor library; and/or

wherein providing one or more candidate second protein chain sequences comprises obtaining the sequences of the candidate second protein chains from a user through a user interface, from a computing device, from a sequence acquisition means or a computing device associated with a sequence acquisition means, from a database or other computer readable medium; and/or wherein the one or more candidate second protein chain sequences are known second chain protein sequences or simulated second chain protein sequences.

Claim 14. The method of any preceding claim, further comprising providing one or more identified second protein chains, a part thereof or information derived therefrom, and/or one or more scores indicative of the probability that one or more query pairs form a functional antigen-binding protein or information derived therefrom to a user through a user interface; and/or

further comprising predicting a score indicative of the probability that each one or more query pairs comprising respective candidate second protein chain sequences will form a functional antigen-binding protein and identifying a candidate second protein chain sequence by applying one or more criteria on the score, optionally wherein the one or more criteria are selected from: the score being above a predetermined cutoff, the score being the

highest predicted score of a set of candidate second protein chain sequences, and the score being in a predetermined top percentile of the predicted probability of a set of candidate second protein chain sequences; and/or

wherein the score is a probability that the pair of protein chains will form a functional antigen-binding protein.

Claim 15. A method of providing antigen-binding protein chain pairings for a plurality of query sequences comprising a first chain sequence, the method comprising: performing the method of any of claims 2 to 15 for each of the query sequences, optionally wherein the plurality of query sequences are heavy or light chain sequences obtained by bulk B cell repertoire sequencing.

Claim 16. A method of providing an antigen-binding protein having a desired property, the method comprising:

providing one or more query sequences comprising a first chain sequence, wherein at least one of the one or more query sequences is likely to have the desired property, and

identifying a second chain sequence for each of the one or more query sequences using the method of any of claims 2 to 15.

Claim 17. 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 16.

Claim 18. 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 16.