CLAIMS

What is claimed is:

1. A method of training a machine learning model to predict binding affinities between antibodies and a target, the method comprising:

using at least one computer hardware processor to perform:

training at least one language model to encode amino acid sequences;

obtaining training data using a candidate amino acid sequence of a candidate antibody, wherein the candidate antibody binds to the target with a candidate binding affinity; and

training the machine learning model to predict the binding affinities between the antibodies and the target using the at least one trained language model and the obtained training data.

2. The method of claim 1, wherein training the at least one language model to encode the amino acid sequences comprises training the at least one language model to predict masked amino acids in at least one amino acid sequence.

3. The method of claim 1, wherein training the at least one language model comprises training a protein language model using protein training data, the protein training data comprising amino acid sequences for individual protein domains.

4. The method of claim 3, wherein the amino acid sequences for the individual protein domains include at least 1 million amino acid sequences, at least 10 million amino acid sequences, at least 15 million amino acid sequences, at least 20 million amino acid sequences, at least 25 million amino acid sequences, or at least 30 million amino acid sequences.

5. The method of claim 1, wherein training the at least one language model comprises training a heavy chain language model using heavy chain training data, the heavy chain training data comprising amino acid sequences of antibody heavy chains.

6. The method of claim 5, wherein the amino acid sequences of the antibody heavy chains include at least 1 million amino acid sequences, at least 10 million amino acid sequences, at least 25 million amino acid sequences, at least 50 million amino acid sequences, at least 75 million amino acid sequences, at least 100 million amino acid sequences, at least 150 amino acid sequences, at least 200 million amino acid sequences, at least 250 million amino acid sequences, or at least 270 million amino acid sequences.

7. The method of claim 1, wherein training the at least one language model comprises training a light chain language model using light chain training data, the light chain training data comprising amino acid sequences of antibody light chains.

8. The method of claim 7, wherein the amino acid sequences of the antibody light chains include at least 1 million amino acid sequences, at least 10 million amino acid sequences, at least 20 million amino acid sequences, at least 30 million amino acid sequences, at least 40 million amino acid sequences, at least 50 million amino acid sequences, at least 60 million amino acid sequences or at least 70 million amino acid sequences.

9. The method of claim 1, wherein training the at least one language model comprises training a paired heavy-light chain language model using paired heavy-light chain training data, the paired heavy-light chain training data comprising pairs of antibody heavy chain amino acid sequences and antibody light chain amino acid sequences.

10. The method of claim 9, wherein the pairs of the antibody heavy chain amino acid sequences and the antibody light chain amino acid sequences include at least 1 million pairs, at least 10 million pairs, at least 15 million pairs, at least 20 million pairs, at least 25 million pairs, or at least 30 million pairs.

11. The method of claim 9, wherein training the paired heavy-light chain model comprises providing, as input to the paired heavy-light chain model, a concatenation of an amino acid sequence of an antibody heavy chain and an amino acid sequence of an antibody light chain.

12. The method of claim 1, wherein the training data comprises a plurality of amino acid

sequences, and wherein obtaining the training data using the candidate amino acid sequence comprises introducing mutations into the candidate amino acid sequence of the candidate antibody to obtain the plurality of amino acid sequences.

13. The method of claim 12, wherein the plurality of amino acid sequences includes amino acid sequences of antibody light chains and amino acid sequences of antibody heavy chains.

14. The method of claim 13, wherein training the machine learning model to predict the binding affinities between the antibodies and the target comprises:

training a first machine learning model using the amino acid sequences of the antibody heavy chains to predict a first set of binding affinities between the antibodies and the target; and

training a second machine learning model using the amino acid sequences of the antibody light chains to predict a second set of binding affinities between the antibodies and the target.

15. The method of claim 1, wherein training the machine learning model to predict the binding affinities between the antibodies and the target comprises fine-tuning the at least one trained language model using the training data to predict the binding affinities between the antibodies and the target.

16. The method of claim 15, wherein fine-tuning the at least one trained language model comprises:

processing the training data using the at least one trained language model to obtain an intermediate output; and

training at least one regression model using the intermediate output to predict the binding affinities between the antibodies and the target.

17. The method of claim 15, wherein fine-tuning the at least one trained language model comprises:

processing the training data using the at least one trained language model to obtain an intermediate output; and

training a probabilistic model using the intermediate output to predict the binding

affinities between the antibodies and the target.

18. The method of claim 17,

wherein processing the training data using the at least one trained language model to obtain the intermediate output comprises processing an amino acid sequence included in the training data using the at least one trained language model to obtain vector representations of amino acids in the amino acid sequence, and

wherein the method further comprises:

concatenating the vector representations of the amino acids;

performing principal component analysis to reduce dimensions of the vector representations to obtain reduced vector representations; and

training the probabilistic model using the reduced vector representations.

19. A system, comprising:

at least one computer hardware processor; and

at least one non-transitory computer-readable storage medium storing processor-executable instructions that, when executed by the at least one computer hardware processor, cause the at least one computer hardware processor to perform a method of training a machine learning model to predict binding affinities between antibodies and a target the method comprising:

training at least one language model to encode amino acid sequences;

obtaining training data using a candidate amino acid sequence of a candidate antibody, wherein the candidate antibody binds to the target with a candidate binding affinity; and

training the machine learning model to predict the binding affinities between the antibodies and the target using the at least one trained language model and the obtained training data.

20. At least one non-transitory computer-readable storage medium storing processor-executable instructions that, when executed by at least one computer hardware processor, cause the at least one computer hardware processor to perform a method of training a machine learning model to predict binding affinities between antibodies and a target the method comprising:

training at least one language model to encode amino acid sequences;

obtaining training data using a candidate amino acid sequence of a candidate antibody, wherein the candidate antibody binds to the target with a candidate binding affinity; and

training the machine learning model to predict the binding affinities between the antibodies and the target using the at least one trained language model and the obtained training data.

21. A method for designing proteins for binding to a target, the method comprising:

obtaining an amino acid sequence of a candidate protein wherein the candidate protein binds to the target with a candidate binding affinity;

determining, for antibodies in a set of antibodies, probabilities that binding affinities between the antibodies and the target are greater than the candidate binding affinity, the proteins in the set of proteins having different amino acid sequences, the proteins including a first protein having a first amino acid sequence, and the probabilities including a first probability that a first binding affinity between the first protein and the target is greater than the candidate binding affinity, wherein determining the first probability comprises:

processing the first amino acid sequence of the first protein using a trained machine learning model to obtain a first output indicative of the first binding affinity between the first protein and the target; and

determining the first probability using the first output indicative of the first binding affinity between the first protein and the target; and

identifying a subset of the set of proteins based on the determined probabilities that the binding affinities are greater than the candidate binding affinity.