

WHAT IS CLAIMED IS:

1. A method, comprising:
 - receiving a first input comprising a target protein sequence having a feature of interest;
 - receiving a second input comprising a candidate protein;
 - applying a first machine learning model to the target protein sequence to generate fragments of interest from the target protein sequence;
 - applying the first machine learning model to the candidate protein to generate fragments from the candidate protein that corresponds to the fragments of interest from the target protein sequence;
 - applying a second machine learning model to the fragments of interest and the fragments from the candidate protein, wherein applying the second machine learning model comprises generating an encoded representation in a multidimensional space for each of the fragments of interest and the fragments from the candidate protein;
 - generating a similarity score between the target protein sequence and the candidate protein based on a similarity between the fragments of interest from the target protein sequence and the fragments from the candidate protein;
 - generating a hierarchical scale of similarity between the target protein sequence and a plurality of candidate proteins comprising the candidate protein according to the feature of interest and the similarity score; and
 - selecting candidate proteins from the plurality of candidate proteins based on the hierarchical scale, wherein higher scores on the hierarchical scale indicate candidate proteins or the more similar substitute candidate proteins inputs for each of the interest inputs.
2. The method of claim 1, wherein the first machine learning model is configured to generate the fragments based on splitting the target protein sequence and the candidate protein into functional domain fragments.

3. The method of any of claims 1 and 2, wherein the second machine learning module comprises a plurality of fully-connected layers.

4. The method of any of claims 1-3, wherein the second machine learning module comprises a plurality of convolutional neural network layers.

5. The method of any of claims 1-4, wherein the second machine learning module comprises a plurality of recurrent neural network layers configured to identify the beginning and end of functional domains.

6. The method of any of claims 3-5, where the layers are connected with direct connections or residual neural networks connections.

7. The method of claim 1, wherein the computational algorithm to split comprises a module to allow the input to be divided into a defined size.

8. The method of claim 1, wherein the first machine learning model is configured to generate the fragments of interest and the fragments from the candidate protein with different sizes or functional domains.

9. The method of claim 1, wherein applying the second machine learning model comprises encoding an amino-acid representation of the candidate protein in a multidimensional space given a local context of the candidate protein.

10. The method of claim 9, wherein encoding the amino-acid representation comprises applying a sub-model having a plurality of fully-connected layers, a plurality of convolutional neural networks layers, or a plurality of recurrent neural networks layer to compress the amino-acid representation given the local context of the candidate proteins.

11. The method of claim 10, where the layers are connected with direct connections or residual neural networks connections.

12. The method of claim 9, further comprising training the sub-model to predict a probability of the amino acid beginning or ending at a certain position in the candidate protein given the local context.

13. The method of claim 9, further comprising training the sub-model using all known or predicted protein sequences.

14. The method of any of claims 1-13, wherein the second machine learning model comprises a sub-model to encode the candidate proteins and the target protein in a multidimensional space given a protein sequence and a compressed positional representation information.

15. The method of claim 14, wherein the sub-model comprises a plurality of fully-connected layers, a plurality of convolutional neural networks layers, or a plurality of recurrent neural networks layer to compress the amino-acid representation given the local context of the candidate proteins.

16. The method of claim 15, where the layers are connected with direct connections or residual neural networks connections.

17. The method of claim 14, further comprising training the sub-model to predict a protein structure, and a function of protein, and a contact-map of protein;

18. The method of any of claims 1-17, wherein generating a similarity score between the target protein sequence and each candidate protein comprises applying a third machine learning model to generate the similarity score between the target protein and the one or more candidate proteins.

19. The method of Claim 18, wherein the third machine learning model comprises a plurality of fully-connected layers, a plurality of convolutional neural networks layers, or a

plurality of recurrent neural networks layer to compare two multidimensional protein representations.

20. The method of claim 18, wherein the similarity score is a hierarchical number that contains information of the protein sequence, structure and function.

21. The method of 20, further comprising training the third machine learning to predict similarity of proteins using the SCOP hierarchical information.

22. The method of claim 1, where the module ranker orders the candidates from the best rated to the worst.

23. A method of use a computational system implemented by one or more computers to:

receiving a plurality of inputs, each input is a protein sequence where one or more are animal protein and the rest of the inputs are plant-based, animal-free candidate proteins;

processing each of the plurality of plant-based, animal-free candidate proteins inputs with a computational algorithm to split the protein in fragments of interest;

processing each of the plurality of inputs with an artificial intelligence model to get an encoded representation in a multidimensional space;

processing each of the plurality of inputs to generate a similarity score between the animal protein inputs and the plant-based, animal-free candidate proteins;

generating a hierarchical scale of similarity between animal protein inputs and plant-based, animal-free candidate proteins; and

selecting the higher scores or the more similar substitute candidate proteins inputs for each of the interest inputs.

24. The method of use of claim 23, wherein the plant-based, animal-free candidate proteins are split into functional domains and/or into a defined size.

25. The method of use of any of claims 23 and 24, wherein the plant-based, animal-free candidate proteins and the animal-origin protein are encoded using a trained artificial intelligence model, given an encoded representation of proteins and/or fragments.

26. The method of use of any of claims 23-25, wherein the plant-based, animal-free candidate proteins previously encoded in a multidimensional space are compared using a hierarchical similarity metric.

27. The method of use of any of claims 23-26, wherein the plant-based, animal-free candidate proteins previously compared using a hierarchical similarity metric are ranked from the best rated to the worst.

28. The method of use of claim 20, wherein the best plant-based, animal-free candidate proteins are selected given a number of max candidates or if the score exceeds a threshold.

29. The method of use of claim 20, wherein the best selected plant-based, animal-free candidate proteins are bioinformatically simulated and/or synthesized in the laboratory to verify the activity and fulfill the desired function.

30. A system comprising a processor and a memory including instruction to program the processor to perform the method of any of Claims 1-29.

31. The method of any of Claims 1-29, wherein the at least one fragment of interest is a target antifungal protein.

32. The method of any of Claims 1-29, further comprising identifying, based on substitute candidate proteins, alternative antifungal proteins to a target antifungal protein, wherein the at least one fragment of interest comprises a feature of the target antifungal protein.

33. The method of any of Claims 1-29, further comprising identifying, based on substitute candidate proteins, alternative enzymes to a target enzyme, wherein the at least one fragment of interest comprises a feature of the target enzyme.