CLAIMS

What is claimed is:

1. A method to reduce overfitting of a neural network-implemented model that processes sequences of amino acids and accompanying position frequency matrices (PFMs), the method including:

generating supplemental training example sequence pairs, labelled benign, that include a start location, through a target amino acid location, to an end location, wherein each supplemental training example sequence pair:

matches the start location and the end location of a missense training example sequence pair; and

has identical amino acids in a reference and an alternate sequence of amino acids;

inputting with each supplemental training example sequence pair a supplemental training PFM that is identical to the PFM of the missense training example sequence pair at the matching start and end location; and

training the neural network-implemented model using the benign supplemental training example sequence pairs, the supplemental training PFMs, the missense training example sequence pairs, and the PFMs of the missense training example sequence pairs at the matching start and end locations;

whereby training influence of the supplemental training PFMs is attenuated during the training.

2. The method of claim 1, wherein the supplemental training example sequence pairs match the start location and the end location of pathogenic missense training example sequence pairs.

3. The method of claim 1, wherein the supplemental training example sequence pairs match the start location and the end location of benign missense training example sequence pairs.

4. The method of claim 1, further including: modifying the training of the neural network-implemented model to cease using the supplemental training example sequence pairs and the supplemental training PFMs after a predetermined number of training epochs.

5. The method of claim 1, further including: modifying the training of the neural network-implemented model to cease using the supplemental training example sequence pairs and the supplemental training PFMs after five training epochs.

6. The method of claim 2, further including: a ratio of the supplemental training example sequence pairs to the pathogenic missense training example sequence pairs is between 1:1 and 1:8.

7. The method of claim 3, further including: a ratio of the supplemental training example sequence pairs to the benign missense training example sequence pairs is between 1:1 and 1:8.

8. The method of claim 1, further including: using, in creating the supplemental training PFMs, amino acid locations from data for non-human primates and non-primate mammals.

9. A system including one or more processors coupled to memory, the memory loaded with computer instructions to reduce overfitting of a neural network-implemented model that

processes sequences of amino acids and accompanying position frequency matrices (PFMs), the instructions, when executed on the processors, implement actions comprising:

generating supplemental training example sequence pairs, labelled benign, that include a start location, through a target amino acid location, to an end location, wherein each supplemental training example sequence pair:

matches the start location and the end location of a missense training example sequence pair; and

has identical amino acids in a reference and an alternate sequence of amino acids;

inputting with each supplemental training example sequence pair a supplemental training PFM that is identical to the PFM of the missense training example sequence pair at the matching start and end location; and

training the neural network-implemented model using the benign supplemental training example sequence pairs, the supplemental training PFMs, the missense training example sequence pairs, and the PFMs of the missense training example sequence pairs at the matching start and end locations;

whereby training influence of the supplemental training PFMs is attenuated or counteracted during the training.

10. The system of claim 9, wherein the supplemental training example sequence pairs match the start location and the end location of pathogenic missense training example sequence pairs.

11. The system of claim 9, wherein the supplemental training example sequence pairs match the start location and the end location of benign missense training example sequence pairs.

12. The system of claim 9, further implementing actions comprising: modifying the training of the neural network-implemented model to cease using the supplemental training example sequence pairs and the supplemental training PFMs after a predetermined number of training epochs.

13. The system of claim 9, further implementing actions comprising: modifying the training of the neural network-implemented model to cease using the supplemental training example sequence pairs and the supplemental training PFMs after five training epochs.

14. The system of claim 10, further implementing actions comprising: a ratio of the supplemental training example sequence pairs to the pathogenic missense training example sequence pairs is between 1:1 and 1:8.

15. The system of claim 11, further implementing actions comprising: a ratio of the supplemental training example sequence pairs to the benign missense training example sequence pairs is between 1:1 and 1:8.

16. The system of claim 9, further implementing actions comprising: using, in creating the supplemental training PFMs, amino acid locations from data for non-human primates and non-primate mammals.

17. A non-transitory computer readable storage medium impressed with computer program instructions to reduce overfitting of a neural network-implemented model that processes

sequences of amino acids and accompanying position frequency matrices (PFMs), the instructions, when executed on a processor, implement a method comprising:

generating supplemental training example sequence pairs, labelled benign, that include a start location, through a target amino acid location, to an end location, wherein each supplemental training example sequence pair:

matches the start location and the end location of a missense training example sequence pair; and

has identical amino acids in a reference and an alternate sequence of amino acids;

inputting with each supplemental training example sequence pair a supplemental training PFM that is identical to the PFM of the missense training example sequence pair at the matching start and end location; and

training the neural network-implemented model using the benign supplemental training example sequence pairs, the supplemental training PFMs, missense training example sequence pairs, and the PFMs of the missense training example sequence pairs at the matching start and end locations;

whereby training influence of the supplemental training PFMs is attenuated during the training.

18. The non-transitory computer readable storage medium of claim 17, wherein the supplemental training example sequence pairs match the start location and the end location of pathogenic missense training example sequence pairs.

19. The non-transitory computer readable storage medium of claim 17, wherein the supplemental training example sequence pairs match the start location and the end location of benign missense training example sequence pairs.

20. The non-transitory computer readable storage medium of claim 17, implementing the method further comprising: modifying the training of the neural network-implemented model to cease using the supplemental training example sequence pairs and the supplemental training PFMs after a predetermined number of training epochs.

21. The non-transitory computer readable storage medium of claim 17, implementing the method further comprising: modifying the training of the neural network-implemented model to cease using the supplemental training example sequence pairs and the supplemental training PFMs after five training epochs.

22. The non-transitory computer readable storage medium of claim 18, implementing the method further comprising: a ratio of the supplemental training example sequence pairs to the pathogenic missense training example sequence pairs is between 1:1 and 1:8.

23. The non-transitory computer readable storage medium of claim 19, implementing the method further comprising: a ratio of the supplemental training example sequence pairs to the benign missense training example sequence pairs is between 1:1 and 1:8.

24. The non-transitory computer readable storage medium of claim 17, implementing the method further comprising: using, in creating the supplemental training PFMs, amino acid locations from data for non-human primates and non-primate mammals.