CLAIMS

What is claimed is:

1. A system including numerous processors coupled to memory, the memory loaded with computer instructions to implement a deep convolutional neural network-based variant pathogenicity classifier comprising:

a positional frequency or weights matrix, collectively abbreviated PFM, generator applied to sequence groups of primates and mammals, to generate a primate PFM, and a mammal PFM;

an input processor that accepts a variant amino acid sequence of at least 51 amino acids with a target variant amino acid flanked upstream and downstream by at least 25 amino acids in each direction, wherein a single nucleotide variant produces the target variant amino acid;

a supplemental data allocator that

allocates a reference amino acid sequence of at least 51 amino acids with a target reference amino acid flanked upstream and downstream by at least 25 amino acids in each direction, aligned with the variant amino acid sequence; and

allocates primate and mammal PFMs produced by the PFM generator, aligned with the reference amino acid sequence;

a deep convolutional neural network,

trained on input including at least variant amino acid sequences of at least 51 amino acids, reference amino acid sequences of at least 51 amino acids, and primate and mammal PFMs to generate a pathogenicity score for the variant amino acid sequence,

that generates the pathogenicity score of the variant amino acid sequence accepted by the input processor as benign or pathogenic, using input including the based on evaluating at least the 51 amino acids of the variant amino acid sequence, the 51 amino acids of the allocated reference amino acid sequence, and the allocated PFMs; and

an output processor that reports at least a pathogenicity score for the variant amino acid sequence generated by the deep convolutional neural network.

2. The system of claim 1, further configured to classify the single nucleotide variant as benign or pathogenic based on the pathogenicity score.

3. The system of any of claims 1-2, wherein the deep convolutional neural network, in parallel, accepts and processes

the variant amino acid sequence,

the allocated reference amino acid sequence,

the allocated primate PFM, and

the allocated mammal PFM.

4. The system of any of claims 1-3, further configured to train to classify, as pathogenic, a single nucleotide variant that produces a target variant amino acid from a target reference amino acid that is conserved in reference amino acid sequences across primates and mammals.

5. The system of any of claims 4, wherein the conservation represents functional significance of the target reference amino acid and is determined from the PFMs.

6. The system of claim 1, further comprising:

a first secondary structure subnetwork trained to predict a three-state secondary structure at amino acid locations within a protein sequence, wherein three structure states distinguish among secondary structures that are helix, sheet, and coil, using as input at least a vertebrae PFM and an amino acid sequence of at least 51 amino acids with the target variant amino acid flanked upstream and downstream by at least 25 amino acids in each direction;

a second solvent accessibility subnetwork trained to predict a three-state solvent accessibility at amino acid locations within a protein sequence, wherein three accessibility states distinguish among solvent accessibility that is exposed, buried, and intermediate, using as input at least a vertebrae PFM and an amino acid sequence of at least 51 amino acids with the target variant amino acid flanked upstream and downstream by at least 25 amino acids in each direction; and

the supplemental data allocator that further

allocates reference state classifications produced by the first and second subnetworks for the reference amino acid sequence; and

allocates variant state classifications produced by the first and second subnetworks for the variant amino acid sequence;

wherein the deep convolutional neural network

is trained on input further including reference and variant state classifications generate the pathogenicity score for the variant amino acid sequence, and

scores the variant amino acid sequence as benign or pathogenic.

7. The system of claim 6, further configured to classify the single nucleotide variant as benign or pathogenic based on the pathogenicity score.

8. The system of any of claims 6-7, wherein the deep convolutional neural network accepts, in parallel, as input at least

the variant amino acid sequence,

the allocated reference amino acid sequence,

the allocated variant secondary structure state classification,

the allocated reference secondary structure state classification,

the allocated variant solvent accessibility state classification,

the allocated reference solvent accessibility state classification,

the allocated primate PFM, and

the allocated mammal PFM.

9. The system of any of claims 6-8, further configured to:

use batch normalization layers, REL non-linearity layers, and dimensionality altering layers to pre-process the variant amino acid sequence, the allocated reference amino acid sequence, the allocated primate PFM and the allocated mammal PFM, to generate pre-processed characterizations;

sum the pre-processed characterizations and concatenate resulting sums with the allocated variant secondary structure state classification, the allocated reference secondary structure state classification, the allocated variant solvent accessibility state classification, and the allocated reference solvent accessibility state classification to produce a concatenated input; and

process the concatenated input through a dimensionality altering layer and accept the processed concatenated input to initiate residual blocks of the deep convolutional neural network.

10. The system of any of claims 6-8, wherein the deep convolutional neural network comprises groups of residual blocks.

11. The system of any of claims 9-10, wherein the deep convolutional neural network is parameterized by a number of residual blocks, a number of skip connections, and a number of residual connections without non-linear activations.

12. The system of any of claims 6-8 and 10-11, wherein the deep convolutional neural network comprises dimensionality altering layers that reshape spatial and feature dimensions of a preceding input.

13. The system of any of claims 6-12, further configured to train to classify, as pathogenic, a single nucleotide variant that produces a target variant amino acid from a target reference amino acid that is conserved in aligned reference amino acid sequences across primates, mammals, and vertebrates.

14. The system of any of claims 8-9, wherein conservation is represented in the PFMs , which indicates functional significance of the target reference amino acid.

15. The system of any of claims 6-14, further configured to train to classify, as pathogenic, a single nucleotide variant that causes different secondary structures between a variant amino acid sequence and a reference variant amino acid sequence.

16. The system of any of claims 6-15, further configured to train to classify, as pathogenic, a single nucleotide variant that causes different solvent accessibilities between a variant amino acid sequence and a reference variant amino acid sequence.

17. The system of any of claims 6-15, wherein a PFM represents conservation of amino acids in a human protein sequence across aligned protein sequences of other species by determining, on a location-by-location basis, frequency of occurrence an amino acid in the human protein sequence across the aligned protein sequences of the other species.

18. The system of any of claims 6-17, wherein the first secondary structure subnetwork comprises groups of residual blocks.

19. The system of any of claims 6-8 and 12-17, wherein the first secondary structure subnetwork is parameterized by a number of residual blocks, a number of skip connections, and a number of residual connections without non-linear activations.

20. The system of any of claims 6-17 or 19, wherein the first secondary structure subnetwork comprises dimensionality altering layers that reshape spatial and feature dimensions of a preceding input.

21. The system of any of claims 6-8, 12-17 and 19-20, wherein the second solvent accessibility subnetwork comprises groups of residual blocks.

22. The system of any of claims 6-8, 12-17 and 19-21, wherein the second solvent accessibility subnetwork is parameterized by a number of residual blocks, a number of skip connections, and a number of residual connections without non-linear activations.

23. The system of any of claims 6-8, 12-17 and 19-22, wherein the second solvent accessibility subnetwork comprises dimensionality altering layers that reshape spatial and feature dimensions of a preceding input.

24. The system of any of claims 9-11, 20-19 and 21-22, wherein each residual block comprises at least one batch normalization layer, at least one rectified linear unit, abbreviated ReLU layer, at least one dimensionality altering layer, and at least one residual connection.

25. The system of any of claims 6-9-11, 20-21 and 21-22, wherein each residual block comprises two batch normalization layers, two ReLU non-linearity layers, two dimensionality altering layers, and one residual connection.

26. The system of any of claims 6-8, 12-17 and 19-25, wherein the deep convolutional neural network, the first secondary structure subnetwork, and the second solvent accessibility subnetwork each comprise an ultimate classification layer.

27. The system of claim 26, wherein the ultimate classification layer is a sigmoid-based layer.

28. The system of claim 26, wherein the ultimate classification layer is a softmax-based layer.

29. The system of any of claims 26-28, further configured to ablate ultimate classification layers of the first secondary structure subnetwork and the second solvent accessibility subnetwork for collaboration with the deep convolutional neural network.

30. The system of any of claims 6-8 and 12-29, further configured to, during training of the deep convolutional neural network, further train the first secondary structure subnetwork and the second solvent accessibility subnetwork on pathogenicity classification, including back propagating errors to the subnetworks and updating subnetwork weights.

31. The system of any of claims 6-8, 12-17 and 19-30, wherein the second solvent accessibility subnetwork comprises at least one atrous convolution layer.

32. The system of any of claims 6-8, 12-17 and 19-31, wherein the variant amino acid sequence and the reference amino acid sequence share flanking amino acids.

33. The system of any of claims 6-8, 12-17 and 19-32, further configured to use one-hot encoding to encode inputs to the deep convolutional neural network.

34. The system of any of claims 6-8, 12-17 and 19-33, wherein the deep convolutional neural network, the first secondary structure subnetwork, and the second solvent accessibility subnetwork are trained on one or more training servers.

35. The system of any of claims 6-8, 12-17 and 19-34, wherein the trained deep convolutional neural network, the first trained secondary structure subnetwork, and the trained second solvent accessibility subnetwork are deployed on one or more production servers that receive input sequences from requesting clients.

36. The system of claim 35, wherein the production servers process the input sequences through at least one of the deep convolutional neural network, the first secondary structure subnetwork, and the second solvent accessibility subnetwork to produce outputs that are transmitted to the clients.

37. Use of the system of any of claims 6-8, and 12-17 and 19-36, to perform a per-gene enrichment analysis which confirms or rejects pathogenicity of variants that have been determined to be pathogenic, comprising:

for a particular gene sampled from a cohort of individuals with a genetic disorder,

applying the deep convolutional neural network-based variant pathogenicity classifier to identify candidate variants in the particular gene that are pathogenic;

determining a baseline number of mutations for the particular gene based on summing observed trinucleotide mutation rates of the candidate variants, and multiplying the sum with a transmission count and a size of the cohort;

applying the deep convolutional neural network-based variant pathogenicity classifier to identify *de novo* missense variants in the particular gene that are pathogenic; and

comparing the baseline number of mutations with a count of the *de novo* missense variants, and based on an output of the comparison confirming that the particular gene is associated with the genetic disorder and that the *de novo* missense variants are pathogenic.

38. Use of the system as in claim 37, further preforming the comparison using a statistical test that produces a p-value as the output.

39. Use of the system as in any of claims 37-38, further comparing the baseline number of mutations with the count of the *de novo* missense variants, and based on the output of the comparison confirming that the particular gene is not associated with the genetic disorder and that the *de novo* missense variants are benign.

40. Use of the system as in claim 39, wherein the genetic disorder is autism spectrum disorder (abbreviated ASD).

41. Use of the system as in claim 39, wherein the genetic disorder is developmental delay disorder (abbreviated DDD).

42. Use of the system of any of claims 6-8, 12-17 and 19-36, to perform a genome-wide enrichment analysis which confirms pathogenicity of variants that have been determined to be pathogenic comprising:

applying the deep convolutional neural network-based variant pathogenicity classifier to classify a first set of *de novo* missense variants that are pathogenic in a plurality of genes sampled from a cohort of healthy individuals;

applying the deep convolutional neural network-based variant pathogenicity classifier to classify a second set of *de novo* missense variants that are pathogenic in the plurality of genes sampled from a cohort of individuals with a genetic disorder; and

comparing respective counts of the first and second sets, and based on an output of the comparison confirming that the second set of *de novo* missense variants are enriched in the cohort of individuals with the genetic disorder and therefore pathogenic.

43. Use of the system of claim 42, further configured to perform the comparison using a statistical test that produces a p-value as the output.

44. Use of the system of any of claims 39-42, wherein the genetic disorder is DDD.

45. Use of the system of any of claims 39-42, wherein the comparison is further parametrized by respective cohort sizes.

46. Use of the system of any of claims 42-45, further configured to compare the respective counts of the first and second sets, and based on the output of the comparison confirming that the second set of *de novo* missense variants are not enriched in the cohort of individuals with the genetic disorder and therefore benign.

47. Use of the system of any of claims 42-46, wherein the genetic disorder is ASD.

48. Use of the system of any of claims 42-46, wherein the genetic disorder is DDD.

49. A deep convolutional neural network-based method of variant pathogenicity scoring, the method including:

applying a positional frequency or weight matrix collectively abbreviated PFM, generator, which runs on at least one of numerous processors, to at least sequence groups of primates and mammals to generate a primate PFM and a mammal PFM;

accepting a variant amino acid sequence of at least 51 amino acids with a target variant amino acid flanked upstream and downstream by at least 25 amino acids in each direction, wherein a single nucleotide variant produces the target variant amino acid;

allocating a reference amino acid sequence of at least 51 amino acids with a target reference amino acid flanked upstream and downstream by at least 25 amino acids in each direction, aligned with the variant amino acid sequence;

allocating primate and mammal PFMs aligned with the reference amino acid sequence;

applying a deep convolutional neural network,

trained on input including at least variant amino acid sequences of at least 51 amino acids reference amino acid sequences of at least 51 amino acids, and primate and mammal PFMs to generate a pathogenicity

score for the variant amino acid sequence;

to score the variant amino acid sequence accepted by the input processor as benign or pathogenic, based on evaluating at least the 51 amino acids of the variant amino acid sequence, the 51 amino acids of the allocated reference amino acid sequence, the allocated reference and variant state classifications, and the allocated PFMs; and

to report at least the pathogenicity score for the variant amino acid sequence.

50. The method of claim 49, further comprising:

applying a first secondary structure subnetwork, which runs on numerous processors coupled to memory, to predict a three-state secondary structure at amino acid locations within a protein sequence, wherein three structure states distinguish among secondary structures that are helix, sheet, and coil, using as input at least a vertebrae PFM and an amino acid sequence of at least 51 amino acids with the target variant amino acid flanked upstream and downstream by at least 25 amino acids in each direction;

applying a second solvent accessibility subnetwork, which runs on numerous processors coupled to memory, to predict a three-state solvent accessibility at amino acid locations within a protein sequence, wherein three accessibility states distinguish among solvent accessibility that is exposed, buried, and intermediate, using as input at least a vertebrae PFM and an amino acid sequence of at least 51 amino acids with the target variant amino acid flanked upstream and downstream by at least 25 amino acids in each direction ;

allocating reference state classifications produced by the first and second subnetworks for the reference amino acid sequence;

allocating variant state classifications produced by the first and second subnetworks for the variant amino acid

sequence;

applying the deep convolutional neural network:

trained on further input that includes the reference state classifications produced by the first and second subnetworks for the reference amino acid sequence and the reference state classifications produced by the first and second subnetworks for the variant amino acid sequence;

to score the variant amino acid sequence the variant amino acid sequence based on further input of the reference and variant state classifications.

51. A non-transitory computer readable storage medium impressed with computer program instructions for variant pathogenicity scoring that, when executed on a processor, implement a method comprising:

;

applying a positional frequency or weight matrix, collectively abbreviated PFM, generator, which runs on at least one of numerous processors, to three sequence groups of primates, mammals, and vertebrates excluding primates and mammals to generate a primate PFM, a mammal PFM, and a vertebrate PFM;

accepting a variant amino acid sequence of at least 51 amino acids with a target variant amino acid flanked upstream and downstream by at least 25 amino acids in each direction, wherein a single nucleotide variant produces the target variant amino acid;

allocating a reference amino acid sequence of at least 51 amino acids with a target reference amino acid flanked upstream and downstream by at least 25 amino acids in each direction, aligned with the variant amino acid sequence;

allocating primate, mammal, and vertebrate PFMs aligned with the reference amino acid sequence;

applying a deep convolutional neural network

trained on input including at least variant amino acid sequences of at least 51 amino acids, reference amino acid sequences of at least 51 amino acids, and primate and mammal PFMs to generate a pathogenicity score for the variant amino acid sequence; and

to report at least the pathogenicity score for the variant amino acid sequence.

52. A non-transitory computer readable storage medium of claim 51, further impressed with computer program instructions that implement the method comprising:

applying a first secondary structure subnetwork, which runs on numerous processors coupled to memory, to predict a three-state secondary structure at amino acid locations within a protein sequence, wherein three structure states distinguish among secondary structures that are helix, sheet, and coil, using as input at least a vertebrae PFM and an amino acid sequence of at least 51 amino acids with the target variant amino acid flanked upstream and downstream by at least 25 amino acids in each direction;

applying a second solvent accessibility subnetwork, which runs on numerous processors coupled to memory, to predict a three-state solvent accessibility at amino acid locations within a protein sequence, wherein three accessibility states distinguish among solvent accessibility that is exposed, buried, and intermediate, using as input at least a vertebrae PFM and an amino acid sequence of at least 51 amino acids with the target variant amino acid flanked upstream and downstream by at least 25 amino acids in each direction ;

allocating reference state classifications produced by the first and second subnetworks for the reference amino acid sequence;

allocating variant state classifications produced by the first and second subnetworks for the variant amino acid sequence;

applying the deep convolutional neural network:

trained on further input that includes the reference state classifications produced by the first and second subnetworks for the reference amino acid sequence and the reference state classifications produced by the first and second subnetworks for the variant amino acid sequence;

to score the variant amino acid sequence based on further input of the reference and variant state classifications.